UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
COORDENADORIA DO CURSO DE CIÊNCIA DA COMPUTAÇÃO

JOÃO VICTOR COSTA GOMES

ESTUDO DE EXTRAÇÃO DE CARACTERÍSTICAS MORFOLÓGICAS PARA CLASSIFICAÇÃO DE MASSAS MAMÁRIAS

SÃO LUÍS
2015

JOÃO VICTOR COSTA GOMES

ESTUDO DE EXTRAÇÃO DE CARACTERÍSTICAS MORFOLÓGICAS PARA CLASSIFICAÇÃO DE MASSAS MAMÁRIAS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Geraldo Braz Júnior

SÃO LUÍS

2015

JOÃO VICTOR COSTA GOMES

ESTUDO DE EXTRAÇÃO DE CARACTERÍSTICAS MORFOLÓGICAS PARA CLASSIFICAÇÃO DE MASSAS MAMÁRIAS

Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal do Maranhão, como parte dos requisitos necessários para obtenção do grau de Bacharel em Ciência da Computação.

Monografia defendida e aprovada em: 12 de Janeiro de 2015

Banca examinadora:

Pr. Dr. Geraldo Braz Júnior

Orientador

UFMA

Prof. Dr. João Dallyson Sousa de Almeida

UFMA

Prof. Dr. Tiago Bonini Borchartt

UFMA

“Portanto, quer comais, quer bebais ou façais outra coisa qualquer, fazei tudo para a glória de Deus.”

(1 Coríntios 10:31)

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelo dom da vida e pela sua graça em minha vida;
Pelo apoio recebido da minha família, em especial, da minha mãe e do meu pai;
Pela orientação e dedicação do meu orientador Geraldo Braz Júnior;
Pela contribuição e ajuda dos meus professores, amigos e colegas de curso nessa jornada difícil, porém gratificante.

RESUMO

O câncer de mama representa uma das principais causas de morte entre as mulheres no mundo ocidental. É responsável, também no Brasil, por milhares de mortes e novos casos ao ano. A probabilidade de cura aumenta consideravelmente com o diagnóstico precoce, podendo assim, evitar maiores danos à saúde da mulher. Com isso, ferramentas computacionais são desenvolvidas a fim de auxiliar o médico especialista a detectar lesões com o padrão maligno ainda em estágio inicial que estejam pouco visíveis em imagens mamográficas. Entretanto, ainda há certa dificuldade em se detectar lesões por imagens devido à particularidade da anatomia da mama feminina e também em reconhecer se a lesão apresenta um padrão maligno. Para isso, este trabalho se concentra em fazer um estudo visando reconhecer padrões segundo a forma geométrica de uma determinada região da mama. O objetivo deste trabalho é a extração de características geométricas como: comprimento do raio, densidades e momentos da imagem e classificar os indivíduos de acordo com seu grupo. Para classificação massa ou não massa, os resultados são promissores.

Palavras-chave: Câncer de Mama, Geometria Côncava, Análise Geométrica.

ABSTRACT

Breast cancer represents one of the leading causes of death among women in the western world. It is also responsible in Brazil for thousands of deaths and new cases per year. The likelihood of cure increases considerably with early diagnosis, and thus, avoid further damage to women's health. Thus, computational tools are developed to assist the specialist and can detect lesions with malignant pattern still at an early stage that are barely visible on mammography. However, there is still some difficulty in detecting lesions in images due to the particularity of the female breast anatomy and also to recognize if that presents a malignant lesion pattern. Therefore, this work focuses on geometric study to recognize patterns according to the geometrical shape of a particular breast region. The objective of this work is the extraction of geometric features such as: geometrical characteristics related to the radius length, density and image moments and classify individuals according to their group. To classification mass or no mass, the results are promising.

Keywords: Breast Cancer, Concave Geometry, Geometric Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos países ocidentais, o câncer de mama representa uma das principais causas de morte entre as mulheres. As estatísticas indicam aumento de sua frequência tanto nos países desenvolvidos quanto naqueles em desenvolvimento. Segundo a Organização Mundial da Saúde (OMS), nas décadas de 1960 e 1970, registrou-se, nos estudos de câncer de base populacional de diversos continentes, um aumento de dez vezes em suas taxas de incidência ajustadas por idade (INCA, 2012). Tem-se documentado também o aumento no risco de mulheres migrantes de áreas de baixo risco para áreas de alto risco. Nos Estados Unidos, a Sociedade Americana de Cancerologia indica que cada dez mulheres tem a probabilidade de desenvolver câncer de mama durante toda a sua vida (INCA, 2012).

Segundo o Ministério da Saúde (Brasil, 2003), no Brasil, o câncer de mama é o que mais causa morte entre as mulheres. Anualmente, são registradas por volta de 10 mil mortes decorrentes desse tipo de câncer. Ele é o principal a atingir a população feminina, sendo responsável por cerca de 40 mil novos casos de câncer de mama ao ano.

A detecção precoce do câncer de mama seguido do tratamento efetivo têm comprovadamente reduzido a mortalidade em várias séries de estudos. No Brasil, ainda 60% dos tumores malignos da mama são diagnosticados em estados avançados. Portanto é notório que diante dos números atuais, esforços não devem ser poupados no desenvolvimento de estratégias de diagnóstico precoce (prevenção secundária), já que a prevenção primária dessa neoplasia, que é a tentativa de evitar o contato ou modificar a ação de agentes que induzem a carcinogênese[1], ainda não é uma realidade para os casos de câncer de mama esporádicos (sem fator de risco conhecido), que constituem o tipo mais frequente desta neoplasia. O diagnóstico precoce consiste em identificar lesões em fases iniciais em mulheres com algum sinal de câncer de mama (nódulo, retração do mamilo, etc.) e o rastreamento que é a aplicação sistemática de um exame, em populações assintomáticas, para identificar mulheres com anormalidades sugestivas de câncer.

A mamografia, entre os métodos de diagnóstico por imagem, é o mais utilizado para o *screening* e diagnóstico do câncer de mama. Vários estudos comprovam a eficácia da mamografia em detectar lesões pequenas e impalpáveis (*screening*) ou em estágios iniciais.

1 É a formação do câncer, em geral se dá lentamente, podendo levar vários anos para que uma célula cancerosa prolifere e dê origem a um tumor visível.

Entretanto, sua sensibilidade diminui consideravelmente (estimada em 81% a 94%, decai para 54 a 58% em algumas séries) entre as mulheres com menos de 40 anos (SANTOS, 2012). Limitações tais como a alta densidade das mamas jovens, gravidez e amamentação, processos inflamatórios, uso de próteses e mamas operadas ou irradiadas não invalidam o método, mas exigem conhecimentos na hora de solicitar o exame. A mamografia digital, apesar do alto custo, aumenta a taxa de detecção de câncer em mamas densas. No Brasil, a rotina mais frequente é fazer o exame de *screening* anualmente entre os 40 e 50 anos de idade. Todavia, a presença de histórico familiar de câncer ou antecedentes de doenças proliferativas da mama altera esta rotina e o início do *screening* ocorre de forma precoce (por volta dos 35 anos de idade).

1.1 Trabalhos Relacionados

Com a motivação de buscar novas tecnologias para auxiliar no diagnóstico do câncer de mama, encontram-se, na literatura, várias metodologias que são desenvolvidas a fim de detectar lesões e posteriormente classificá-las em benignas ou malignas. Nesta seção destacam-se os trabalhos que têm objetivos relacionados ao diagnóstico do câncer de mama através de ferramentas computacionais.

A técnica *Scalar Feature Selection* (SFS) utilizada por MELO (2010) teve a finalidade de selecionar um conjunto de características que permita obter melhor classificação dos achados mamográficos. Neste trabalho ainda, foi realizada uma comparação dos métodos supervisionado e não supervisionado de classificação. Para a classificação supervisionada, foram empregadas diferentes arquiteturas de redes neurais de propagação direta. Para a classificação não supervisionada, foi utilizado o algoritmo *k-Means.* Obtendo no método supervisionado uma acurácia de 86,19% com um conjunto de seis características dos *clusters* de microcalcificações.

Em HOLSBACK (2012) foi proposta a mineração de dados para o diagnóstico do câncer de mama baseado na seleção de variáveis, baseado na análise de amostras de célula da mama de pacientes. O método proposto pode auxiliar o médico no diagnóstico do câncer de mama utilizando o menor número de variáveis com a maior acurácia possível. Aplicado ao

WBCD (*Wisconsin Breast Cancer Database*), o método proposto apresentou acurácia de 98,09%, retendo uma média de 17,24 variáveis.

O Índice de Biodiversidade *Shannor-Wiener* por SOUSA (2011), geralmente aplicado para medir a biodiversidade em um ecossistema, foi aplicado para descrever padrões de regiões de imagem de mama. E por fim foi utilizado a Máquina de Vetores de Suporte para classificar regiões em massa ou não massa. Essa metodologia obteve uma acurácia máxima de 99,95%.

Em SALES (2013) técnicas de Processamento de Imagem foram usadas para preparar as mamografias e, em seguida, o nível de simetria entre a mama esquerda e a direita foi medido com coeficiente de correlação cruzada e distância euclidiana. O índice de *Getis-Ord* na sua forma geral foi usado para extrair características das imagens para treinar uma Máquina de Vetores de Suporte que classificou regiões das mamografias em lesão e não lesão. A metodologia, de modo geral, apresentou 80,11% de sensibilidade, 84,41% de especificidade e 84,38% de acurácia.

Em SILVA et al. (2006) foi utilizada a textura como característica para a segmentação por agrupamento com *K-means*. Cada estrutura identificada pelo *K-means* é descrita usando características de textura e geometria para formar a sua assinatura, usadas para classificação com K-NN, obtendo especificidade igual a 85,13% e sensibilidade igual a 81,81%.

Em MARTINS et al. (2006) foi descrito uma metodologia para classificação de tecidos da mama em normal, benigno ou maligno, através de matrizes de coocorrência e redes neurais Bayesianas. A partir de cada amostra são obtidas diversas matrizes de coocorrência, as quais são utilizadas no cálculo de medidas estatísticas de textura. Uma rede neural Bayesiana é usada para avaliar a eficacia dessas medidas em classificar cada amostra de tecido. A metodologia obteve uma taxa de acerto de 86.84%.

Em ROCHA (2014) foi descrito uma metodologia que utiliza textura e aprendizado de máquina para descriminar padrões malignos e benignos. Além disso, foi ampliado o conceito de Índice de Diversidade, através do uso da informação de coocorrência de espécies, com o propósito de aumentar a eficiência da extração de características de textura. Assim, foram utilizadas técnicas *Local Binary Pattern*, Função K de Ripley e os índices Shannon, Mchintosh, Simpson, Gleason e de Meninhick. Na classificação foi utilizado Máquina de Vetores de Suporte com o objetivo de classificar massas malignas e benignas. O melhor

resultado foi obtido utilizando a função K de Ripley com 92,2% de acurácia, 92,96% de sensibilidade e 91,26% de especificidade.

Em JUNIOR (2014) foi avaliada a extração de características usando as abordagens de análise de diversidade, geoestatística e geométrica para a classificação das regiões suspeitas detectadas usando a Máquina de Vetores de Suporte como classificador. Foi utilizada a geometria côncava para extrair características. Os melhor resultado teve como taxa de sensibilidade de 97,30%.

1.2 Objetivos

O objetivo geral deste trabalho consiste no estudo de características geométricas aplicadas a neoplasias mamárias extraídas de mamografias digitalizadas através do desenvolvimento de uma metodologia que realiza a extração de características geométricas a fim de determinar um padrão discriminatório e posteriormente realizar dois testes de classificação, sendo o primeiro a classificação massa e não massa e o segundo a classsificação maligna ou benigna.

1.2.1 Objetivos Específicos

- Fazer um estudo do desempenho geral e individual das medidas geométricas utilizadas nesse trabalho na classificação observando as variáveis acurácia, sensibilidade e especificidade de cada medida geométrica.
- Estudar e aplicar a geometria côncava para descrição de forma para obter melhor eficiência na extração de características.
- Utilizar o classificador *Random Forest* (Subseção 2.6.1) para classificar os nódulos malignos e benignos a partir das características geométricas extraídas de cada nódulo.
- Construir uma metodologia que ofereça ao especialista uma segunda opinião na distinção de regiões extraídas de mamografias.

1.3 Organização do Restante do Trabalho

O restante deste trabalho está organizado em mais quatro capítulos:

- O Capítulo 2 apresenta a fundamentação teórica que serve de base para o desenvolvimento da metodologia deste trabalho. São explanados assuntos relacionados ao diagnóstico do câncer de mama, às técnicas do processamento de imagens, à detecção de contornos e à classificação;
- O Capítulo 3 apresenta as etapas da metodologia deste trabalho como: a obtenção das imagens mamográficas, o pré-processamento, a extração de características e a classificação. Apresenta também a seleção de parâmetros para que resultassem em melhor eficiência na classificação e detecção de contornos côncavos;
- O Capítulo 4 apresenta os resultados obtidos utilizando a metodologia proposta utilizando as imagens mamográficas que foram propostas para todo o desenvolvimento e estudo geométrico dos nódulos;
- O Capítulo 5 apresenta a conclusão dos resultados obtidos e as possíveis melhorias para alcançar melhores resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados os fundamentos teóricos que foram utilizados para elaborar esta metodologia. Os temas que são abordados nas seções seguintes são: o câncer de mama, o processamento de imagens bem como suas etapas utilizadas nesse trabalho: pré-processamento, detecção de contornos côncavos, extração de características baseada em geometria côncava e classificação.

2.1 Câncer de Mama

O câncer é o nome dado a um conjunto de mais de 100 doenças que tem em comum o crescimento desordenado (maligno) de células que invadem os tecidos e órgãos, podendo espalhar-se (metástase) para outras regiões do corpo. Dividindo-se rapidamente estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores (acúmulos de células cancerosas) ou neoplasias malignas. Por outro lado, um tumor benigno significa simplesmente uma massa localizada de células que se multiplicam vagarosamente e se assemelham ao seu tecido original, raramente constituindo um risco de vida (INCA, 2014).

O câncer de mama é provavelmente o tipo de câncer mais temido pela população feminina, devido a sua alta frequência e, sobretudo, aos seus efeitos psicológicos, que afetam a percepção da sexualidade e a própria imagem pessoal. Este é relativamente raro antes dos 35 anos de idade, mas acima dessa faixa etária sua incidência cresce rapidamente e progressivamente (INCA, 2014).

O câncer de mama considerado esporádico, ou seja, sem associação, com o fator hereditário, representa mais de 90% dos casos de câncer de mama em todo mundo. Dados clínicos, epidemiológicos e experimentais têm demonstrado que o risco de desenvolvimento de câncer de mama esporádico está fortemente relacionado à produção de esteroides sexuais. Condições endócrinas moduladas pela função ovariana, como menarca precoce, menopausa e gestação tardias, assim como a utilização de estrógenos exógenos, são componentes relevantes do risco de desenvolvimento do câncer de mama. Em sinergismo com os fatores hormonais, estudos observacionais indicam comportamento humano relacionado ao estilo de vida como a inatividade física e os descuidos com a dieta tipo obesidade ou alcoolismo,

podem contribuir para o aumento da incidência do câncer de mama em todo mundo (SOARES et al., 2013).

Por outro lado, as neoplasias mamárias do tipo hereditário correspondem a 5% a 10% dentre os casos de câncer de mama, sendo este o grupo muito relacionado a alterações de genes supressores de tumor com os genes BRCA 1 e BRCA 2 e o P53. A prevalência de mutação deletéria[2] no gene BRCA 1 é de 1/800 na população geral, sendo mais frequente nas descendentes de judeus asquenazes. Mulheres portadoras de mutações nesses genes têm o risco estimado que varia de 56% a 85% de desenvolver o câncer de mama durante sua vida, tendendo a apresentá-lo mais precocemente (SOARES et al., 2013).

Várias condições são reconhecidas como capazes de aumentar ou diminuir a chance de desenvolvimento do câncer de mama como os descritos na Tabela 2.1:

Quadro 2.1: Fatores de risco do câncer de mama

Risco muito elevado (RR = 3.0)
Mãe ou irmã com câncer de mama na pré-menopausa Antecedente de hiperplasia epitelial atípica ou neoplasia lobular in situ Suscetibilidade genética comprovada (mutação de BRCA1-2)
Risco moderado (1.5 < RR < 3.0)
Mãe ou irmã com câncer de mama na pós-menopausa Nuliparidade Antecedente de hiperplasia epitelial sem atipia ou macrocistos apócrinos
Risco pouco elevado (1.0 < RR < 1.5)
Menarca precoce (=12 anos) Menopausa tardia (=55 anos) Primeira gestação a termo depois de 34 anos de idade Obesidade Dieta gordurosa Sedentarismo Terapia de reposição hormonal por mais de 5 anos Ingestão alcoólica excessiva Exposição da mama a radiações ionizantes

Fonte: Diagnóstico e Tratamento do Câncer de Mama, 2001.

2 Mutação que provoca uma modificação em determinada informação (gene) de forma que o novo alelo produzido a partir dela cause prejuízo ao organismo.

Através do crescente volume de imagens médicas digitais produzidas em hospitais. As atividades relacionadas à aquisição, gerenciamento e segmentação de imagens têm exigido esforços de pesquisadores e profissionais na informatização dos sistemas hospitalares, a fim de estudar e obter métodos computacionais auxiliando os profissionais da saúde no diagnóstico. Através disso, houve a possibilidade de desenvolver ferramentas para o diagnóstico auxiliado por computador (*Computer-Aided Diagnosys* – CAD, 2005).

Os sistemas CAD têm como objetivo auxiliar e aumentar a precisão do diagnóstico do médico, através do uso de resultados do computador como referência, como, por exemplo, a indicação de áreas suspeitas da imagem. Este auxilio é importante, pois o diagnóstico do especialista está sujeita às variações pessoais (como fadiga visual e distração). Assim, os sistemas CAD têm mostrado que podem melhorar o desempenho dos diagnósticos oferecendo uma segunda opinião ao especialista médico e podem auxiliar no rastreamento precoce do câncer de mama.

2.2 Processamento de Imagens

Uma imagem digital pode ser definida como uma função bidimensional $f(x , y)$, onde x e y são coordenadas espaciais, e a amplitude de f para qualquer par de coordenadas (x , y) é chamada de intensidade ou nível de cinza da imagem neste ponto. Quando x, y e o valor da amplitude de f são finitos, em quantidades discretas a imagem é digital. A imagem digital é composta de um número finito de elementos, sendo que cada um possui uma localização e um valor. Esses elementos são denominados de *pixels* (pontos). *Pixel* é o termo mais amplamente usado para denotar os elementos da imagem digital. O processamento de imagens digitais engloba processos cujas entradas são imagens digitais que, a partir de técnicas computacionais, geralmente são transformadas em outras imagens digitais. Com isso, pode-se obter melhorias nos aspectos estruturais de cada imagem para: a interpretação visual humana, para extração de características e para o reconhecimento de objetos em particular. O processamento de imagens surgiu da necessidade de codificar, transmitir e decodificar imagens digitais por cabos de transmissão entre pontos distantes (GONZALEZ, 2010).

Da aquisição à extração de informações, existem várias etapas a serem a executadas, para que as informações obtidas sejam consistentes. Dependendo do objetivo final, para cada etapa,

podem ser utilizado um ou mais algoritmos que compõe todo o trabalho. Geralmente o processamento de imagens segue a seguinte metodologia conforme a Figura 2.1.

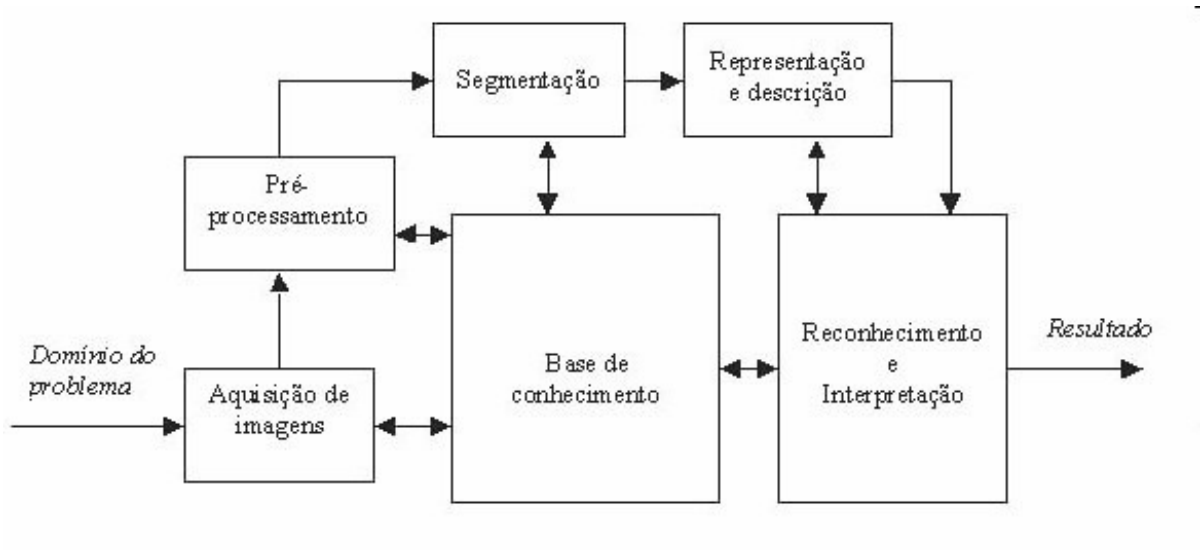

Figura 2.1: Etapas do processamento de imagens. Fonte: adaptado de (GONZALEZ; WOODS, 2010).

As etapas apresentadas na Figura 2.1 são: a aquisição de imagens, o pré-processamento, a segmentação, representação e descrição e reconhecimento e interpretação. O fluxo de informações segue a ordem apresentada na Figura 2.1, ou seja, a saída de informações de uma etapa vai para a entrada de outra etapa.

2.3 Pré-processamento

A imagem adquirida pode conter alguns ruídos, contraste e/ou brilho inadequados. O objetivo do pré-processamento é melhorar a qualidade da imagem para ser processada nas etapas posteriores. As operações lineares ou não lineares efetuadas nesta etapa trabalham diretamente com os valores de intensidade dos pixels. A técnica utilizada nesse trabalho foi a equalização de histograma.

2.3.1 Equalização do Histograma

A equalização de histograma é uma técnica em que se redistribuem os valores de tons de cinza dos pixels de uma imagem para que o percentual de pixels de qualquer nível de cinza seja quase o mesmo, de forma a obter um histograma mais uniforme (MARQUES, 1999). A função utilizada para equalizar o histograma é chamada de função de distribuição acumulada.

Seja r a variável que representa os níveis de cinza da imagem a ser aprimorada. Assume-se que r pode ser normalizado no intervalo $[0, L-1]$. Com $r = 0$ representando preto e $r = L-1$ representado branco. Após isso, considera-se a formulação discreta e atribui-se valores de pixels no intervalo $[0, L-1]$. Sendo assim, a transformação se dá da seguinte forma:

$$s = T(r), 0 \leq r \leq L-1 \tag{2.1}$$

A qual produz um nível s para todo valor de pixel r na imagem original. Assume-se então que a função de transformação *T(r)* satisfaz as seguintes condições:

1. *T(r)* é de valor único e monotonicamente crescente no intervalo $0 \leq r \leq L-1$.
2. $0 \leq T(r) \leq L-1$ para $0 \leq r \leq L-1$

Os níveis de cinza em uma imagem podem ser considerados como variáveis aleatórias no intervalo [0, L – 1]. Assim, pode-se obter um descritor fundamental que é a função densidade de probabilidade. Portanto, p_r (r) e p_s (s) denotam a função densidade de probabilidade das variáveis aleatórias de r e s. Como neste caso utiliza-se variáveis discretas para descrever os níveis de cinza das imagens que foram utilizadas, a probabilidade de ocorrência do nível de cinza r_k em uma imagem é dada por:

$$p_r(r_k) = \frac{n_k}{n}, \; k = 0, 1, 2, \ldots L-1 \tag{2.2}$$

Sendo n o número total de pixels na imagem, n_k o número total de pixels que tem o nível de cinza r_k , e L o número total de níveis de cinza possíveis em uma imagem. Então a versão discreta da transformação é dada por:

$$s_k = T(r_k) = \sum_{j=0}^{k} p_r(r_j) = \sum_{j=0}^{k} \frac{n_j}{n}, \quad k = 0,1,2,\ldots,L-1 \tag{2.3}$$

Assim, a saída processada é obtida pelo mapeamento de cada pixel com nível r_k na imagem de entrada correspondendo com o pixel com nível s_k na imagem de saída.

Figura 2.2: a) imagem antes da equalização de histograma b) a mesma imagem após a equalização de histograma. Fonte: (BRIDI, 2011)

2.4 Extração de Características

Esta etapa ocorre antes do reconhecimento de padrões. Tem como objetivo extrair um conjunto de dados descritivos correspondentes as características do objeto analisado. Essas características descritivas devem apresentar um bom poder de discriminação entre os indivíduos que posteriormente serão classificados na etapa de reconhecimento e interpretação. Para processamento de imagens, a entrada nesta etapa é uma imagem e a sua saída será um conjunto de dados referente àquela imagem de entrada. Esses dados gerados devem organizados de forma adequada ao classificador que será utilizado.

As características extraídas de cada indivíduo são invariantes à rotação, à translação e à escala. Pois, geralmente quando se precisam classificar indivíduos (como ocorre neste trabalho em classificar nódulos malignos e benignos), cada objeto (nódulo) analisado poderia

apresentar um tamanho, uma localização ou uma posição diferente dos demais o que poderia influenciar no resultado final e a margem de erro ser superior ao esperado.

As características escolhidas por este trabalho para caracterização de massas dividem-se em três categorias: a primeira se refere às características geométricas adquiridas através da localização do centro de massa e das medidas do raio, perímetro e área. A segunda se refere a densidade de pixels em quatro quadrantes e circunferências inscritos ao nódulo. E a terceira se refere às medidas de momentos*, Zernike Moments* e *Hu Moments*.

2.4.1 Análise Geométrica

A análise geométrica visa descrever o quão as massas são definidas em termo de circularidade, a espicularidade e a rugosidade. O estudo dessas medidas visando descrever as formas deve-se ao comportamento distinto entre nódulos malignos e nódulos benignos. Enquanto neoplasias benignas possuem uma chamada pseudocápsula que impede o tumor de crescer e invadir os tecidos normais circundantes, fazendo com que esses nódulos tenham contornos bem definidos e margens e formas arredondadas e suaves, as neoplasias malignas (sem a pseudocápsula) tendem a invadir de forma envolvente os tecidos, resultando em aspectos ultrassonográficos mal definidos, contornos irregulares e formas espiculadas (TSUI *et al*., 2010). Portanto, partindo do princípio das diferenças entre esses dois grupos de nódulos, foram calculadas medidas geométricas.

As medidas calculadas são: Circularidade, Compacidade, Desvio Padrão, Razão De Área, Rugosidade (CHIANG, CHIU, 2001); Densidade Circular e Quadrangular, *Hu Moments* (HU, 1962) e *Zernike Moments* (TEAGUE, 1980).

- Circularidade 1: $C_1 = \frac{P^2}{A}$

  – Mede o quão é circular o objeto em relação ao seu perímetro (P) tomando como referência a sua área (A).

- Circularidade 2: $C_2 = \frac{Desviomédio}{Desviopadrão}$

–Mede o quão a forma digital é similar a um círculo. Quanto maior a similaridade, maior será o valor resultante. O cálculo do desvio médio e do desvio padrão será explicitado posteriormente.

- Compacidade: $C_o = \frac{P^2}{4\pi A}$

–Mede o quão compacto é o objeto distribuído em sua área (A) ao longo de seu perímetro (P).

- Convexidade: $C_V = \frac{Áreadoobjeto}{Áreadofechoconvexodoobjeto}$

– Mede o quão é côncavo ou quão é convexo o objeto.

As medidas a seguir utilizam a distância radial normalizada que é calculada a partir da localização do centro de massa, que representa o ponto médio entre os pontos extremos do objeto, e a localização de cada pixel que estiver no contorno. A distância radial normalizada (CHIANG, CHIU, 2001) será dada por:

$$d(i) = \frac{\sqrt{\left(x(i) - X_0\right)^2 + \left(y(i) - Y_0\right)^2}}{max\left(dr(i)\right)} \tag{2.4}$$

O valor max($d(i)$) representa o valor máximo de $dr(i)$ (distância radial) e $(X_{0,} Y_0)$ representa as coordenadas do centro de massa e N representa o número de *pixels* no contorno. Assim o desvio médio d_{avg} será dado por:

$$d_{avg} = \frac{1}{N}\sum_{i=1}^{N} d(i) \tag{2.5}$$

- Desvio Padrão: $\sigma = \sqrt{\frac{1}{N-1}\left(d(i) - d_{avg}\right)^2}$
  – Mede o desvio padrão das distâncias radiais de cada objeto.
- Razão de Área: $A = \frac{1}{d_{avg} N}\sum_{i=1}^{N}\left(d(i) - d_{avg}\right)$. Onde $\forall d(i) \le d_{avg}$
  – Mede a porcentagem do tumor que está fora da região circular.
- Rugosidade: $R = \sum_{i=1}^{N}\left(d(i) - d(i+1)\right)$
  – Mede o grau de espicularidade do objeto.

- Densidade Quadrangular: $Dqi = \frac{A_{oi}}{A_{qi}}$

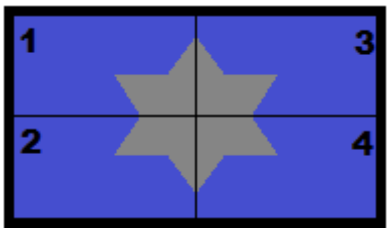

Figura 2.3: Divisão da região de interesse em quadrantes para o cálculo da densidade quadrangular.

– Divide a imagem em quadrantes e para cada quadrante é extraída a razão entre a área da parte do objeto no quadrante (A_{oi}) e a área do quadrante correspondente (A_{qi}).

- Densidade Circular: $D_{ci} = \frac{A_{oi}}{A_{ci}}$

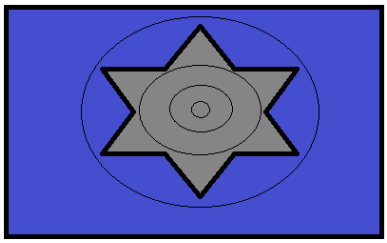

Figura 2.4: Divisão da região de interesse em círculos concêntricos para calcular os índices da densidade circular.

–Divide a imagem em circunferências concêntricas e para cada circunferência é extraída a razão entre a área da parte do objeto dentro da circunferência (A_{oi}) e a área da circunferência (A_{ci}).

2.4.2 Momentos

Momentos (ou *Moments*) descrevem o arranjo de pixels do objeto, combinando área, compacidade, irregularidade na forma e outros descritores. São descritores globais de forma e foram originalmente introduzidos na década de 1960 (HU,1962) para análise de imagens digitais. Momentos são frequentemente associados como reconhecimento de padrões estatísticos e o seu uso vem sendo bem-sucedido em muitas aplicações.

O momento cartesiano bidimensional está associado como uma ordem que se inicia a partir de um valor baixo (onde o mais baixo é zero) até ordens superiores. O momento (m_{pq}) de ordem p e q da função $I(x, y)$ é definido por:

$$m_{pq} = \int_{\infty}^{\infty} \int_{\infty}^{\infty} x^p y^q I(x, y) dxdy \tag{2.6}$$

Para imagens discretas, é usualmente aproximado para:

$$m_{pq} = \sum_{x} \sum_{y} x^p y^q I(x, y) \tag{2.7}$$

Para momentos centralizados μ_{pq} que são invariantes à translação:

$$\mu_{pq} = \sum_{x} \sum_{y} (x - x_o)^p (y - y_o)^q I(x, y) \tag{2.8}$$

onde (x_o, y_o) representa as coordenadas do centro de massa que podem ser calculadas pela relação:

$$x_o = \frac{m_{10}}{m_{00}} \; y_o = \frac{m_{01}}{m_{00}} \tag{2.9}$$

2.4.3 Momentos Invariantes

Momentos invariantes, também conhecidos como *Hu Moments* (HU, 1962), são calculados através das relações anteriores citadas como: centro de massa e momentos centralizados. Diferentemente de momentos centralizados que são invariantes apenas à translação. *Hu Moments* são invariantes à escala, à rotação e a translação. Os momentos invariantes fornecem sete índices:

$$
\begin{aligned}
h_1 &= \eta_{20} + \eta_{02} \\
h_2 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
h_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
h_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
h_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})\left((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} - \eta_{03})^2\right) + (3\eta_{21} - \eta_{03})\left(3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right) \\
h_6 &= (\eta_{20} - \eta_{02})\left((\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2\right) + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
h_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})\left((\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2\right) + (3\eta_{12} - \eta_{30})(\eta_{21} + \eta_{03})\left(3(\eta_{12} + \eta_{30})^2 - (\eta_{21} + \eta_{03})^2\right)
\end{aligned}
\tag{2.10}
$$

Onde a Equação 2.11:

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^{\gamma}} \tag{2.11}$$

representa os momentos centrais invariantes à escala e à rotação. E por fim na Equação 2.12:

$$\gamma = \frac{p+q}{2} + 1 \tag{2.12}$$

2.4.4 *Zernike Moments*

O *Zernike Moments* (TEAGUE, 1980) ou Momentos de Zernike fornecem também características que são um conjunto ortogonal de momentos invariantes à rotação. Momentos de Zernike são representados por um conjunto de polinômios complexos que formam um conjunto ortogonal completo sobre o interior do círculo unitário, isto é, $x^2 + y^2 = 1$. Para calcular os momentos de Zernike de uma imagem é necessário primeiro tomar o centro de massa como origem e as coordenadas de pixel são tomadas para o intervalo de círculo unitário $x^2 + y^2 \leq 1$. Os *pixels* que estão fora do círculo unitário não são computados. Os momentos de Zernike de uma imagem digital podem ser calculados por:

$$Z_{nl} = \frac{n+1}{\pi} \sum_x \sum_y V_{nl}(x, y) f(x, y) \tag{2.13}$$

onde $x^2 + y^2 \leq 1$, $0 \leq l \leq n$, $f(x, y)$ descreve a valores de intensidade da imagem e V_{nl} é um complexo conjugado do *Zernike Polynomial* de grau *n* e dependência angular *l*.

$$V_{nl}(x,y)=\sum_{m=0}^{\frac{n-l}{2}}(-1)^m\frac{(n-m)!}{m!\left(\frac{n-2m+l}{2}\right)!\left(\frac{n-2m-l}{2}\right)!}\left(x^2+y^2\right)^{\frac{n}{2}-m}e^{-il\theta} \tag{2.14}$$

onde, $0\le l\le n$, $n-l$ é igual, $\theta=\tan^{-1}\left(\frac{y}{x}\right)$, m o limite inferior e $i=\sqrt{-1}$.

2.5 *Alpha Shapes*

É comum que alguns nódulos mamários apresentem regiões desconexas umas das outras, o que dificulta o cálculo das medidas geométricas, pois de antemão precisa-se definir o contorno de cada nódulo. Uma solução seria computar o fecho convexo dos *pixels* existentes. Entretanto, isso afetaria negativamente no cálculo de algumas medidas como:

- Razão de Área: pois a porcentagem fora da região circular seria insignificante.
- Rugosidade: pois os contornos não seriam tão irregulares.
- Convexidade: o valor da convexidade seria igual a 1 para todos os nódulos.

Com isso, a solução adotada nesse trabalho foi a utilização do algoritmo de geometria côncava chamado *Alpha Shapes* (MUCKE, 1994). Através desse algoritmo, a partir de um conjunto de pontos pode-se obter o contorno côncavo de cada objeto da imagem a ser processada. Assumindo que há um conjunto $S\subset\mathbb{R}^d$ de n pontos num espaço *d* dimensional, pretende-se computar a forma dos *n* pontos, ou seja, traçar um contorno côncavo que englobe esse conjunto de pontos. Para controlar o grau de concavidade do contorno do conjunto *S* de pontos é utilizado o parâmetro α. Assim, para cada ponto do conjunto *S* de pontos, será englobado por uma circunferência de raio α.

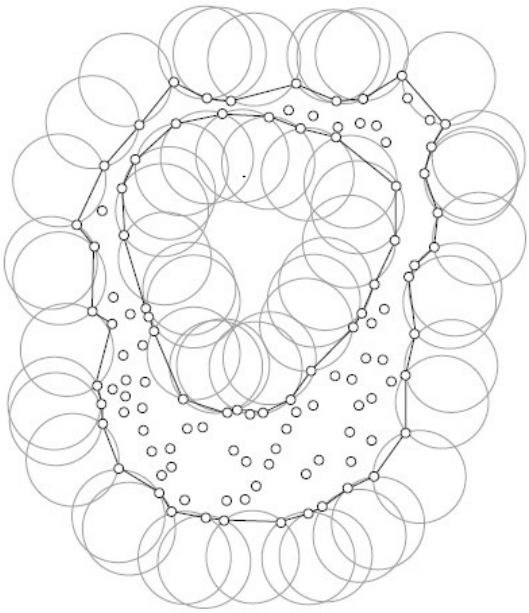

Figura 2.5: Contorno computado a partir de um conjunto de pontos representado pela linha mais escura. Fonte: (FISCHER, 2000).

Na Figura 2.5, é possível observar que os pontos pertencentes ao contorno do *Alpha Shapes* são incidentes sobre o contorno das circunferências. Ao alterar o valor de α, altera-se o tamanho da circunferência e assim altera-se a concavidade da forma computada. Para valores de α que tendem a zero, o *Alpha Shapes* degenera para o conjunto de pontos. Já para valores de α relativamente altos que tendem ao infinito, não haverá contornos internos a outros e nem contornos isolados um dos outros. Portanto, esse contorno será um fecho convexo do conjunto de pontos existentes.

Para $0 < \lambda < \infty$, seja uma circunferência aberta com raio λ. Uma 0-circunferência é um ponto e uma ∞-circunferência é um espaço aberto. Uma circunferência b é chamada de vazia se $b \cap S = 0$. Com isso, um k-simplexo Δ_T é dito α-exposto se existe uma α-circunferência vazia onde $T = \partial b \cap S$ e ∂b é a superfície da esfera (para $d = 3$) ou da circunferência (para $d = 2$) delimitadora *b*. Onde *d* representa a dimensão em que se encontra o conjunto de pontos. Δ_T é o fecho convexo de T e $T \subset S$ com $|T| = k+1 \leq d+1$, com isso Δ_T de dimensão k é chamado de *k*-simplexo. Na Figura 2.6, pode-se observar o exemplo de um *k*-simplexo α-exposto para o caso de $d = 2$.

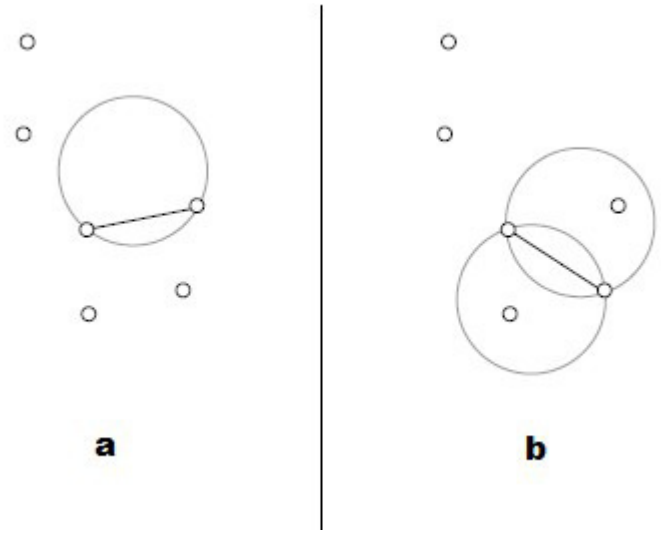

Figura 2.6: a) k-simplexo α-exposto. b) um k-simplexo que não é α-exposto. Fonte: (FISCHER, 2000).

O limite ∂S_α do *α-shape* do conjunto de pontos *S* consiste de todos os k-simplexos de *S* para $0 \leq k < d$ que são α-exposto.

$$\partial S\alpha = \Delta T \vee T \subset S, |T| \leq d \, e \, \Delta T \; \alpha - exposto. \tag{2.15}$$

Para o cálculo do α-shape, assume-se que qualquer limite ∂S_α, para qualquer valor de α, é obtido como subconjunto da triangulação de Delaunay. Assim, dado um conjunto $S \subset \mathbb{R}^d$, a triangulação de Delaunay de *S* é o complexo *DT*(*S*) consistindo de:

1. Todos os *d*-simplexos ΔT em que $T \subset S$ tal que a circunferência de *T* não contém mais nenhum ponto de *S*.
2. Todos os *k*-simplexos que sejam faces para outros simplexos em *DT*(*S*).

Portanto, para que $\Delta_T \in DT(S)$, Δ_T deve ser um α-exposto simplexo de *S*. Assim através da triangulação *DT*(*S*) obtida, para obter o *α-shapes* do conjunto de pontos, cada face da triangulação deve atender pelo menos uma das seguintes condições:

1. A circunferência que engloba a face Δ_T é vazia e tem raio menor que α, ou

2. Se Δ_T é face de outro simplexo no conjunto α-complexo, que é representado por C_α (S).

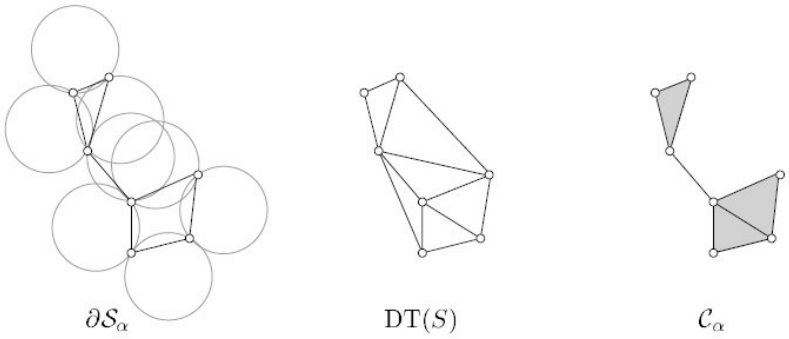

Figura 2.7: Representação do Alpha Shapes a partir da Triangulação de Delaunay.

2.6 Reconhecimento de padrões

Esta etapa tem por objetivo a classificação ou descrição de objetos (padrões) em categorias ou classes a partir das características extraídas. Os padrões podem ser entendidos como entidade, objeto ou evento que pode ser previamente definido por um nome. A classe pode ser definida como um conjunto de objetos que possuem características em comum e as características como já mencionadas anteriormente são dados que podem ser extraídos a partir de alguma medida. A partir de um conjunto de características resultantes da etapa de extração de características o classificador separa os objetos em grupos denominados por classes. E assim, de acordo com a particularidade das características em comum em cada um de seus grupos pode ser reconhecido como pertencente ou não pertencente a uma determinada classe.

2.6.1 Random Forest

Random Forest é um algoritmo de classificação formado por um conjunto de árvores estruturadas classificadoras. O algoritmo produz a classificação de acordo com o resultado independente de cada uma dessas árvores, isto é, a classificação final será dada pelo maior número de “votos” dados por árvore.

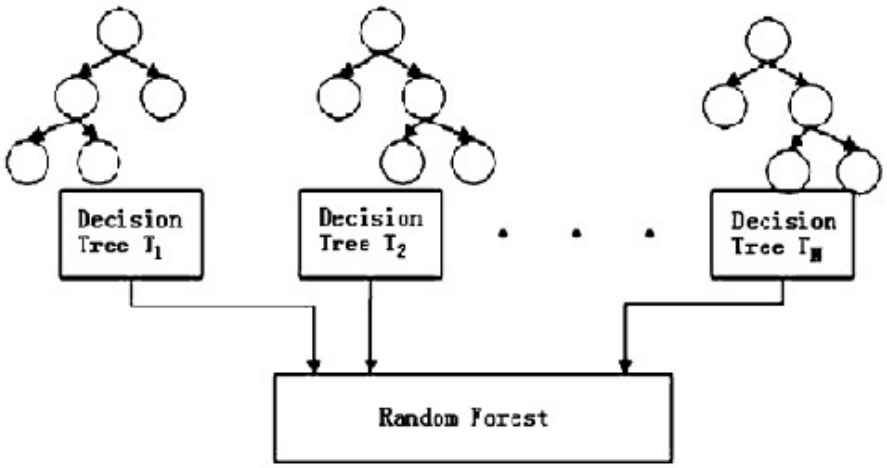

Figura 2.8: Procedimento do Random Forest. Fonte: adaptado de (BREIMAN, 1999)

O procedimento comum para todas as árvores é que, para a *k*-ésima árvore, um vetor aleatório v_k é gerado de forma independente dos outros vetores aleatórios gerados $v_1,\ldots, v_{k-1,}$ entretanto gerado com a mesma distribuição. Cada árvore é cultivada utilizando o conjunto de treinamento e v_k, resultando em um classificador $h(x, v_k)$, onde x é um vetor de entrada. O vetor v é gerado conforme as contagens em N caixas resultantes de N dardos lançados aleatoriamente nas caixas, onde N é o número de exemplos no conjunto de treinamento. Na seleção de divisão aleatória, v consiste de um número independe de inteiros aleatórios entre 1 e K. Depois de um grande número de árvores, elas votam pela classe mais popular e assim esse procedimento é chamado de *Random Forest* (BREIMAN,1999).

O crescimento de cada árvore é dado como segue:

- Se o número de casos do conjunto de treinamento é N, apresentam-se N amostras aleatórias v, mas com a substituição, a partir dos dados originais. Esta amostra será o conjunto de treinamento para o crescimento da árvore.
- Se existem M variáveis de entrada, um número $m \ll M$ é especificado de modo em que a cada nó m variáveis são selecionadas aleatoriamente fora de M e a melhor divisão sobre essas m variáveis é usada para dividir o nó. O valor de m é constante durante o crescimento da floresta.
- Cada árvore é cultivada na maior extensão possível. Não há poda.

No *Random Forest*, a taxa de erro depende de duas variáveis que são medidas de precisão dos classificadores individuais e da dependência entre eles. A interação entre eles da base para compreensão do *Random Forest*. A primeira variável é a correlação entre as árvores na floresta. Quando a correlação entre duas árvores cresce, aumenta a taxa de erro da floresta. E a outra variável é a força individual da árvore. Uma árvore com uma baixa taxa de erro é um classificador forte, assim o aumento das forças individuais das árvores diminui a taxa de erro da floresta. Reduzir o valor de *m* reduz tanto a correlação quanto a força. Usando a taxa de erro *out-of-bag* o valor de *m* pode ser facilmente encontrado na faixa. Este é o único parâmetro ajustável que influencia na sensibilidade do *Random Forest*. Estes dados *out-of-bag* são usados para obter uma estimativa imparcial de execução do erro de classificação conforme as árvores são adicionadas à floresta. São calculados a partir dos casos que são descartados durante a construção do conjunto de treinamento em comparação aos votos das árvores e a proporção em que esses votos dos casos e das árvores são diferentes é a estimativa do erro *out-of-bag*.

Outra característica importante que se pode obter do *Random Forest* são as proximidades entre os casos. As proximidades entre os casos são representadas por meio de uma matriz N x N. Depois que a árvore que é cultivada e depois da fase de treinamento. Se os casos *k* e *n* estão no mesmo nó terminal, a proximidade entre eles é igual a 1. No final normalizam-se as proximidades dividindo pelo número de árvores.

A proximidade média do caso *n* na classe *j* para outros casos do treinamento pode ser definida por:

$$P(n)=\sum_{cl=j} prox^2(n,k) \tag{2.16}$$

3 METODOLOGIA

Este capítulo apresenta a metodologia utilizada para: aquisição de imagens, o pré-processamento, obtenção das características das regiões extraídas da mamografia e classificação. A metodologia deste trabalho, exemplificada na Figura 3.1, é formada por um conjunto de etapas, semelhantes às etapas do processamento de imagens apresentadas na fundamentação teórica. As etapas da metodologia são: aquisição de imagens, pré-processamento, extração de características e classificação.

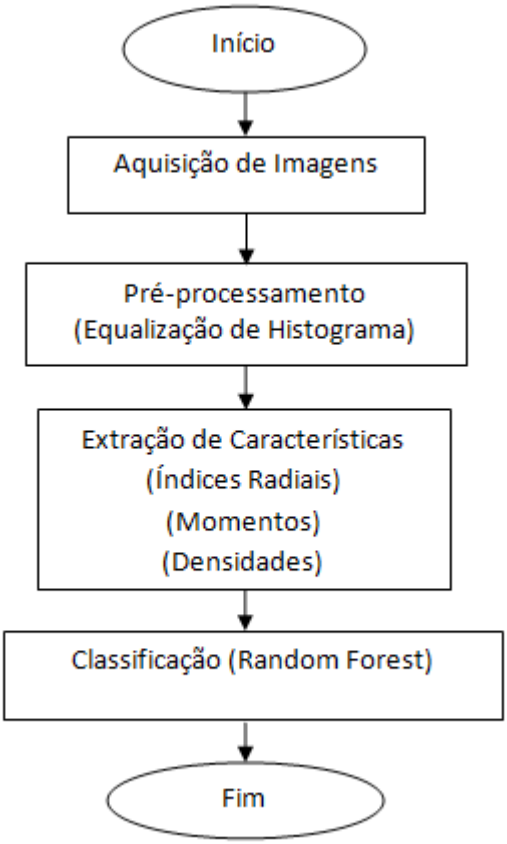

Figura 3.1: Etapas da metodologia utilizada para desenvolver o estudo geométrico de neoplasias mamárias.

A primeira etapa é a aquisição da base de imagens já com os nódulos mamários pré-segmentados manualmente por especialistas médicos. A segunda etapa é o pré-processamento

que consiste no realce da imagem e na quantização não linear. A terceira etapa tem o objetivo de computar e detectar o contorno e extrair as medidas geométricas e a quarta é a classificação utilizando o *Random Forest*. As seções seguintes descrevem mais detalhadamente os procedimentos utilizados na realização deste trabalho.

3.1 Aquisição de Imagens

As mamografias utilizadas para avaliar os resultados de cada teste foram obtidas através da base DDSM (*Digital Database for Screening Mamography*) (HEAT et al., 1998). Todas as imagens possuem a resolução 1024 x 1024, com 8 bits de profundidade e foram obtidas segundo a projeção Médio Lateral Oblíqua (MLO). Para cada respectiva imagem há um arquivo de texto explicativo com os pontos da região de interesse. Uma coluna representa a linha, uma coluna representa a coluna e a última coluna representa se o ponto pertence ao objeto ou ao fundo da imagem.

Foram selecionadas 700 regiões de interesse com a presença de lesões mamárias, sendo 340 nódulos malignos e 360 nódulos benignos pré-segmentados por um especialista médico para a classificação benigna e maligna. Dentre essas 700 regiões de interesse com a presença de lesões mamárias, foram selecionadas 260 regiões de interesse para serem consideradas como padrão massa e também obteve-se 231 regiões de interesse sem a presença de qualquer lesão, isto é, padrão não massa para a classificação massa e não massa. Com essas imagens digitais já disponibilizadas, elas foram lidas uma a uma pelo programa desenvolvido em linguagem C em conjunto com a biblioteca de processamento de imagens OpenCV para a execução dos procedimentos desse trabalho.

3.2 Pré-processamento

Esta etapa tem o objetivo de aprimorar a qualidade da imagem para as etapas subsequentes. Os procedimentos adotados nessa etapa são muito importantes para a definição dos contornos de cada imagem, entre as funções adotadas aqui está o aumento do contraste e uma melhor definição das regiões que pertencem ao fundo e das regiões que pertencem ao objeto em questão. Nesta fase foram realizadas: a equalização do histograma, para melhorar o

contraste da imagem e a quantização não linear para avaliar possíveis diferenças na distribuição dos valores de intensidade dos pixels das regiões de interesse.

3.2.1 Equalização de Histograma

A fim de melhorar a distribuição dos valores de pixels na imagem e melhorar o contraste para depois utilizar a quantização não-linear foi utilizada a equalização de histograma. Com isso, ficam mais evidentes as regiões que possuem uma densidade mais acentuada que as outras. E com essa distribuição mais igualitária entre os valores de intensidade de *pixels*, foi realizada a quantização não-linear das imagens. A Figura 3.2 mostra o procedimento da equalização de histograma.

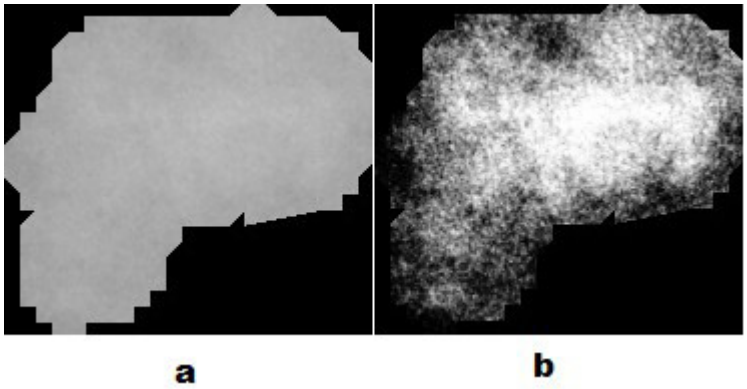

Figura 3.2: a) imagem de um nódulo antes da equalização de histograma e b) a imagem do mesmo nódulo após a equalização de histograma.

3.2.2 Quantização Não-linear

Visando as diferenças na distribuição dos valores de intensidade de *pixel*, dividiu-se cada imagem em um número determinado de faixas de valores de intensidade. Cada faixa dá origem a uma imagem com os *pixels* correspondentes a ela, ou seja, cada imagem mamográfica contendo o nódulo tem seus *pixels* distribuídos a novas imagens de acordo com seu valor de intensidade.

Para computar a distribuição de pixels da imagem original para outras imagens representando cada faixa, deve-se anteriormente calcular o número *N* de valores diferentes de

intensidade de *pixel* existentes na imagem original e depois escolher o número *f* de faixas que a imagem original deve ser redistribuída. Após isso, é criado um vetor *A* de tamanho *N* com os valores existentes dispostos em ordem crescente. Assim é calculado o quociente *q*, que representa o número de valores de intensidade de pixel do vetor *A* em cada faixa, através de uma divisão inteira da seguinte relação:

$$q = \frac{N}{f}, f \geq 1 \tag{3.1}$$

Após isso, para calcular a faixa que o pixel de valor r_b é distribuído, deve-se obter a posição *p* do valor r_b no vetor *A*. A faixa *m* que o pixel de valor r_b é computada pela seguinte relação:

$$m = \frac{p}{q} + 1 \tag{3.2}$$

onde p varia de 1 a *N* e p/q será uma divisão inteira. Caso $m > f$, os pixels correspondentes a *m* são alocados à última faixa de número *f*. A Figura 3.3 apresenta a divisão de faixas da quantização não linear.

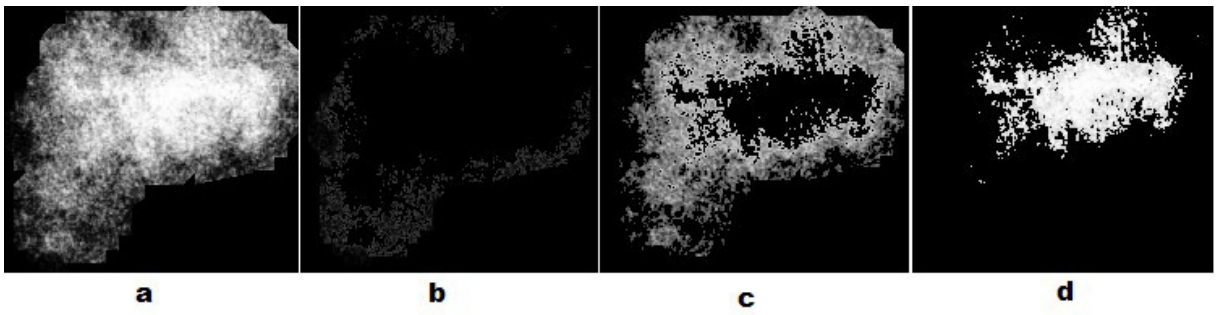

Figura 3.3: Divisão de faixas. a) Imagem original. b) Primeira faixa. c) Segunda faixa d)Terceira faixa

3.3 *Alpha Shapes*

A fim de se obter o contorno côncavo de cada objeto foi o utilizado o algoritmo *Alpha Shapes* (Seção 2.5). O contorno de cada nódulo foi computado a partir do conjunto de pontos (pixels) e o parâmetro α que controla a concavidade do contorno. Quanto maior o valor do parâmetro α, mais pontos serão englobados pelo contorno do *Alpha Shapes* e assim se tem um contorno mais próximo do fecho convexo dos pontos existentes em cada imagem e quanto menor o valor do parâmetro α, menos pontos serão englobados e assim se tem um contorno

mais côncavo ou degenerando para o conjunto de pontos. O parâmetro α foi um dos principais parâmetros utilizados para variar o seu valor a fim de se obter melhores resultados na fase de testes, pois ele influencia diretamente nos valores das medidas geométricas e assim influencia também nos resultados de classificação. A Figura 3.4 apresenta contornos computados pelo Alpha Shapes.

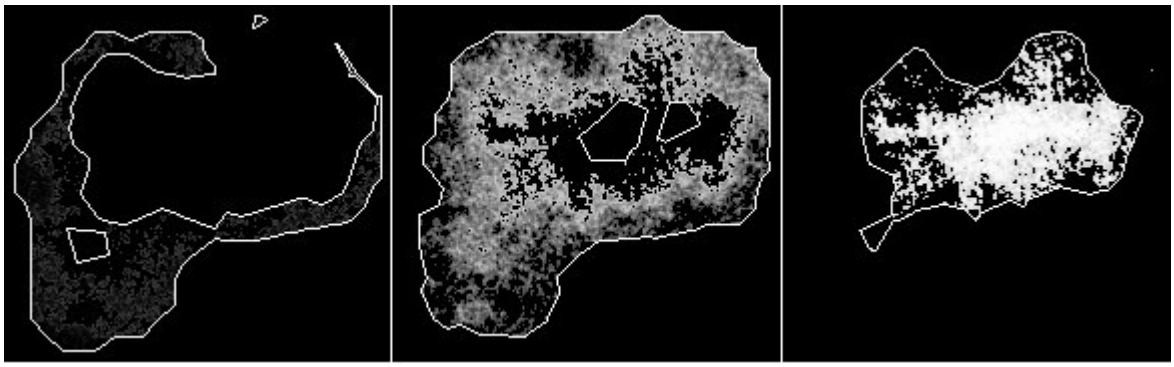

Figura 3.4: Os contornos (destacados pelas linhas brancas) computados de cada imagem representando cada faixa.

É comum que apareçam vários contornos na mesma imagem. Alguns contornos aparecem internamente a outros maiores devido a algumas pequenas regiões que não contém pontos (buracos) e outros contornos aparecem externos a outros devido a alguns conjuntos de pontos estarem disjuntos a outros conjuntos de pontos. O processo de obtenção das medidas geométricas através dos contornos computados de cada imagem é explicado posteriormente.

3.4 Medidas Geométricas

As medidas geométricas foram extraídas de cada faixa da quantização não linear e cada imagem terá um conjunto de características provenientes das medidas geométricas em cada faixa, ou seja, se o número de variáveis correspondentes às medidas geométricas é *N* e o número de faixas obtidos através da quantização não-linear é *M*, são gerados para a fase de classificação *N* x *M* variáveis para cada imagem.

Em alguns casos, é possível a existência de vários contornos para cada faixa gerada da quantização não linear. Como as medidas geométricas descrevem cada imagem representando cada faixa da quantização não linear e algumas medidas geométricas tais como: circularidade, compacidade, convexidade, desvio padrão, razão de área e rugosidade são extraídas de cada

um dos contornos. Após o cálculo dessas medidas em cada contorno, é calculada a média dessas medidas que representa os índices geométricos de cada faixa. Em contrapartida medidas como: densidades quadrangular e circular, *Hu Moments* e *Zernike Moments* são calculadas diretamente para cada faixa e não nos contornos individualmente.

Para calcular a densidade quadrangular primeiro faz-se um *bounding box* bidimensional englobando todo o objeto e depois divide a região em quatro quadrantes e partir desses quadrantes menores criados calculam-se quatro índices de densidade quadrangular dividindo-se o número de pixels do objeto dentro do quadrante pelo número total de pixels do respectivo quadrante. E para calcular a densidade circular, primeiro é construída uma circunferência que envolva todo o objeto e depois são construídos outras três circunferências menores concêntricas à primeira circunferência, correspondendo a 1/2, 1/4 e 1/8 do raio da primeira circunferência e então as densidades circulares são calculadas a partir do número de *pixels* do objeto dentro da circunferência pelo número total de pixels dentro da circunferência e assim se tem 4 índices de densidade circular.

3.5 Ajuste de Parâmetros

Como mencionado anteriormente, nos testes desse trabalho com o objetivo de obter melhores resultados na classificação, o projeto em relação aos ajustes de parâmetros foi dividido em 3 fases. A primeira fase teve como parâmetro de teste o número de faixas, a segunda fase o valor de α para computar os contornos côncavos e a terceira fase o valor do grau do *Zernike Moments*. O melhor parâmetro encontrado em uma fase foi utilizado na fase seguinte.

Na primeira fase para encontrar o melhor resultado para o número de faixas da quantização não-linear utilizaram-se apenas algumas medidas geométricas: circularidade, compacidade, desvio padrão, convexidade, razão de área e rugosidade. Os números de faixas testados variaram de 1 a 7. O valor de α para esta fase foi ajustado em 10000 para todos os testes e o número de indivíduos testados foram de 260 incluindo imagens de nódulos malignos e benignos. Na segunda fase a fim de se obter o melhor valor de α para as imagens utilizadas neste trabalho, foram testados 21 valores de α que variaram de $0.05 \times N$ até N, sendo N representando a raiz quadrada do tamanho da imagem em pixels, incluindo os valores ótimos de α para cada imagem. Na segunda fase as mesmas medidas geométricas da

primeira fase foram utilizadas. O número de faixas foi ajustado em 3 e o número de indivíduos testados para cada teste foi de até 380. Na terceira fase o parâmetro a se ajustar foi o valor do grau do *Zernike Moments*, os valores testados variaram de 6, resultado em um vetor de característica de tamanho 16, até 13, resultando em um vetor de características de tamanho 56. O número de faixas foi três e $\alpha = 0.55 \times N$. Os resultados das classificações de cada fase são apresentados na Seção 4.1.

3.6 Classificação

Para fazer a classificação das imagens, foram reunidas para cada imagem do nódulo as características das imagens de cada faixa correspondente. A saída da extração de características foi gerada em um arquivo de texto *Attribute-Relation File Format* (ARFF) para a leitura e processamento do aplicativo *Weka* [WAIKATO, 2013] que implementa o classificador *Random Forest*. Cada arquivo contém as definições dos atributos (variáveis) utilizados na classificação e o conjunto de características que representam as medidas geométricas extraídas de cada nódulo região. Cada nódulo com suas respectivas características foram distribuídos linha por linha. Para nódulos benignos, foi atribuído o atributo “nao” indicando a ausência de câncer e para nódulos malignos, foi atribuído o atributo “sim” indicando a presença de câncer. Na classificação de indivíduos massa e não massa, o atributo “sim” indicava que o indivíduo apresentava massa e o atributo “nao” foi utilizado para indivíduos não massa. Na classificação foi utilizada a validação cruzada com 10 *folds*.

4 RESULTADOS

Os resultados foram obtidos através dos testes de cada fase em que os parâmetros de número de faixas, valor de α do *Alpha Shapes* e valor do grau do *Zernike Moments* sofreram variação a fim de se encontrar os parâmetros que ocasionassem melhores resultados na classificação. Para se avaliar o resultado dos testes, foram utilizadas três medidas estatísticas: acurácia, sensibilidade e especificidade.

A sensibilidade caracteriza como a capacidade de um teste para identificar corretamente os indivíduos onde há presença de uma determinada doença. A especificidade é a capacidade de se identificar os indivíduos onde há ausência de uma determinada doença. A acurácia é a proporção de indivíduos que foram classificados corretamente. A sensibilidade (*S*), especificidade (*E*) e acurácia (*A*) são definidas pelas Equações 4.1, 4.2 e 4.3.

$$S = \frac{VP}{VP + FN} \tag{4.1}$$

$$E = \frac{VN}{VN + FP} \tag{4.2}$$

$$A = \frac{VP + VN}{VP + VN + FN + FP} \tag{4.3}$$

Verdadeiros positivos (*VP*) são doentes (nesse caso com a presença de câncer) classificados como doentes. Verdadeiros negativos (*VN*) são indivíduos saudáveis classificados como saudáveis. Falsos positivos (*FP*) são indivíduos saudáveis classificados como doentes. E falsos negativos (*FN*) são indivíduos doentes classificados como saudáveis. Os resultados desta metodologia se subdividem em duas seções: a Seção 4.1 trata dos resultados de ajuste de parâmetros e a Seção 4.2 trata dos resultados da classificação massa e não massa e a Seção 4.3 trata dos resultados da classificação benigna e maligna.

4.1 Determinando Parâmetros

Todos os testes a fim de se encontrar esses parâmetros foram realizados na classificação de padrões malignos e benignos, pois objetivo inicial do trabalho era fazer um estudo sobre o uso de índices geométricos para diagnosticar o câncer de mama. O critério utilizado para escolher os melhores resultados foi o valor da acurácia geral de cada teste.

Na primeira fase, a fim de se encontrar o melhor valor para o número de faixas da quantização não linear, o valor de α foi fixado em α = 10000 e o número de faixas da quantização foi variado de um a sete. As medidas utilizadas na primeira fase foram: duas medidas de circularidade, compacidade, desvio padrão, convexidade, razão de área e rugosidade.

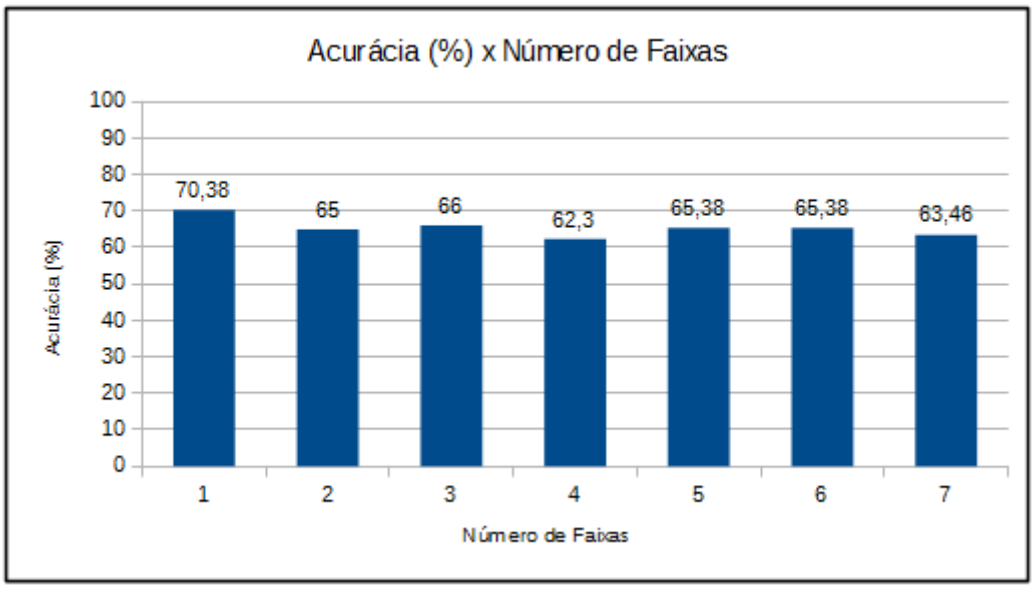

Figura 4.1: Gráfico dos resultados da acurácia pelo número de faixas.

Como se pode observar na Figura 4.1, após os testes com várias quantidades de faixas, constatou-se que sem a divisão de faixas ocasionava os melhores resultados. Entretanto, escolheu-se a divisão em três faixas para se avaliar posteriormente a distribuição dos valores de pixel nos nódulos benignos e malignos.

A segunda fase teve como objetivo encontrar o valor de α que ocasiona na melhor acurácia. Como visto na Seção 2.5, o valor de α influencia no número de pontos englobados pelo contorno côncavo e consequentemente na extração de características. Por isso valores de

α foram testados de $0{,}05 \times N$ a N, variando 0,05 no valor de α de um teste para outro. E também o valor ótimo de α para cada conjunto de pontos. Onde:

$$N = \sqrt{Altura\ da\ imagem \times Largura\ da\ imagem} \tag{4.4}$$

O número de faixas foi fixado em três e foram utilizadas as mesmas medidas da primeira fase.

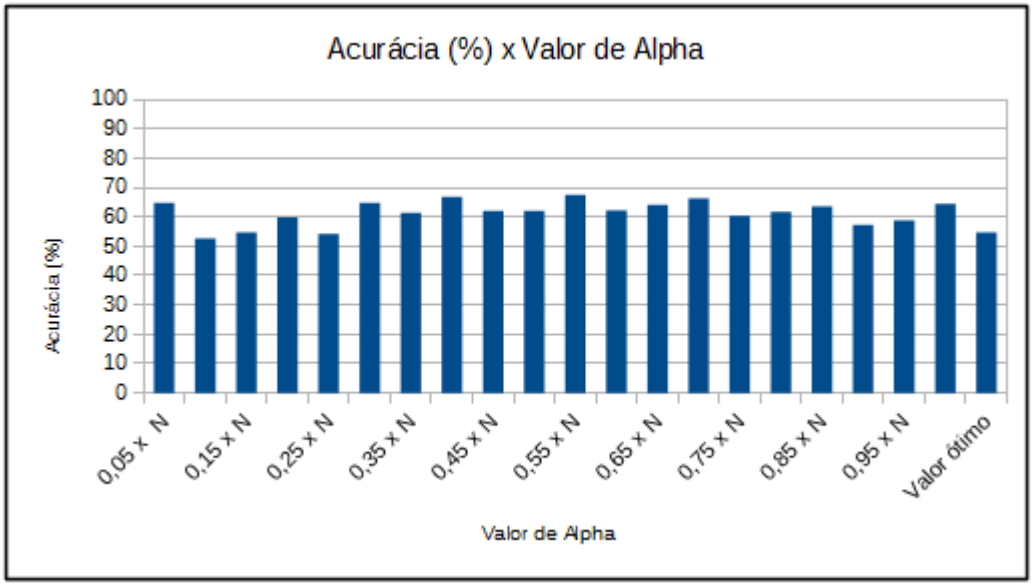

Figura 4.2: Gráfico dos resultados da acurácia pelo valor de α.

Como se pode observar na Figura 4.2, para $\alpha = 0{,}55 \times N$, obteve-se a melhor acurácia dentre 21 testes realizados nessa fase. Este parâmetro foi fixado na fase posterior. Na terceira fase foi fixado $\alpha = 0{,}55 \times N$, três faixas na divisão de faixas da quantização não linear e foram adicionadas as medidas de densidade circular e quadrangular e os descritores de momentos *Hu Moments* e *Zernike Moments*. Para melhorar a distribuição dos valores de intensidade de pixel na imagem foi adicionada a etapa de pré-processamento com a equalização de histograma. Nesta fase apenas os resultados individuais do *Zernike Moments* e do total mudam. Os resultados referentes à densidade, aos índices geométricos inclusive *Hu Moments* são iguais para todos os testes. A Figura 4.3 apresenta o gráfico com os resultados dos testes reunindo todas as medidas já apresentadas neste trabalho variando-se o valor do grau do *Zernike Moments*.

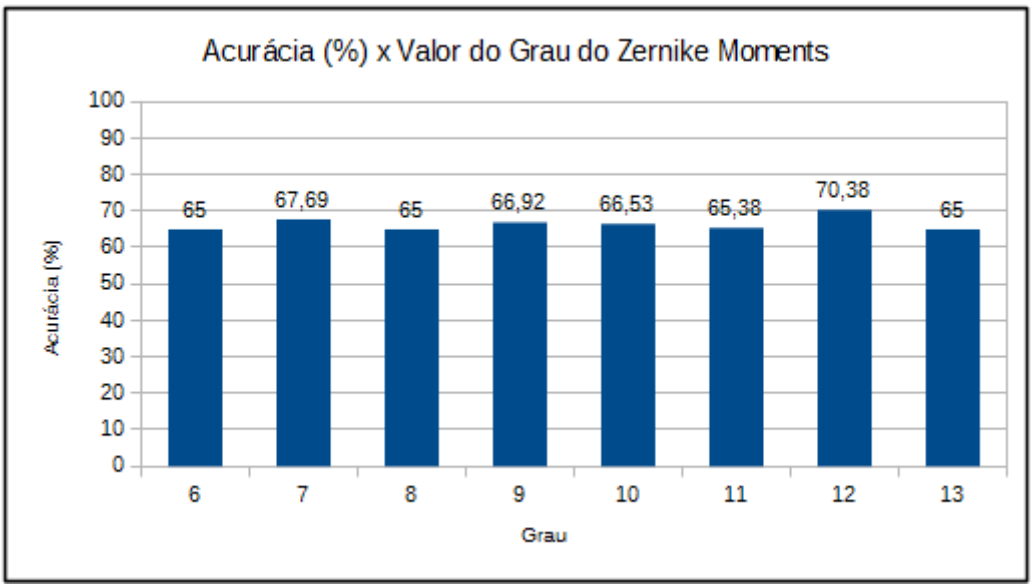

Figura 4.3: Gráfico dos resultados da acurácia pelo valor do grau do Zernike Moments.

Como se pode perceber na Figura 4.3 em todos os testes realizados com o *Zernike Moments*, a variação do grau do *Zernike Moments* pouco altera nos resultados gerais. Os testes obtiveram os melhores resultados com o grau do *Zernike Moments* ajustado em n=12, fornecendo 49 descritores no total.

4.2 Resultados Massa e Não Massa

Depois de estimados os melhores parâmetros para extração de características, foram testados a classificação de indivíduos em massa e não massa. No total, foram utilizados 490 indivíduos 260 indivíduos massa e 230 indivíduos não massa. Os parâmetros de número de faixas e valor de α foram os mesmos utilizados da terceira fase e o valor do grau do *Zernike Moments* utilizado foi grau n= 12, resultando em um vetor de características de tamanho 49.

Tabela 4.1: Resultados gerais e individuais das medidas geométricas utilizadas para classificação massa e não massa.

	Sensibilidade (%)	Especificidade (%)	Acurácia (%)	N° de índices gerados para cada faixa
Circularidade 1	93,8	93,1	93,48	1
Circularidade 2	95	91,3	93,2	1
Compacidade	93,5	93,9	96,53	1
Desvio Padrão	91,9	92,6	92,26	1
Convexidade	77,3	68,8	73,31	1
Razão de Área	98,8	97,8	98,37	1
Rugosidade	100	100	100	1
Densidade Quadrangular	95,4	93,5	94,5	4
Densidade Circular	84,6	84	84,31	4
Hu Moments	98,8	98,3	98,57	7
Zernike Moments	76,5	58,4	68,02	49
Total	100	99,1	99,59	71

Para a classificação de indivíduos massa e não massa, os resultados gerais são apresentados na Tabela 4.1. A acurácia ficou próxima a um, a sensibilidade foi um, ou seja, acertou todos os indivíduos massa e a especificidade foi próxima a um. Com relação aos resultados individuais a medida de rugosidade classificou totalmente de forma correta os indivíduos massa e não massa e em relação às outras medidas geométricas, a maioria teve acurácia, sensibilidade e especificidade acima de 90%, com exceção do Zernike Moments, das densidades circulares e da medida de convexidade.

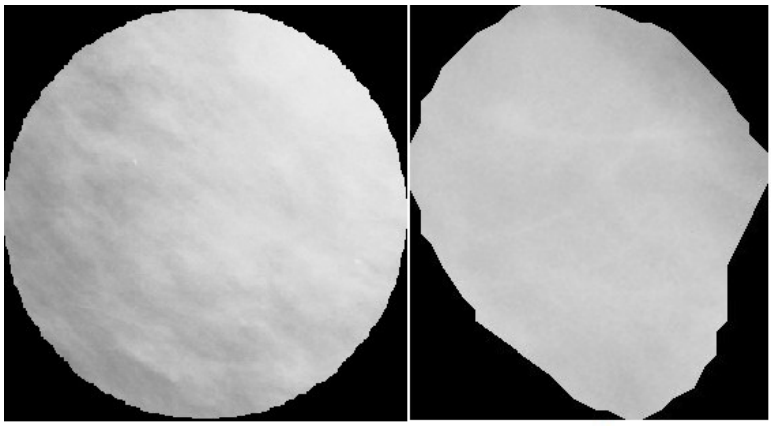

Figura 4.4: a) imagem com o padrão não massa e b) imagem com o padrão massa.

Observando o exemplo da Figura 4.4, pode-se avaliar as possíveis justificativas do resultado de cada medida geométrica utilizada para classificação de padrões massa e não massa. Em relação às medidas geométricas que obtiveram acurácia acima de 90%, o resultado se deve diferença de distribuição dos *pixels* na forma. Enquanto o recorte de regiões não massa apresenta uma distribuição de *pixels* mais compacta e regular, as regiões com padrão massa apresentam uma distribuição de *pixels* relativamente desconexa e concentrada em determinadas faixas da quantização não linear. Isso fica mais claro na Figura 4.5.

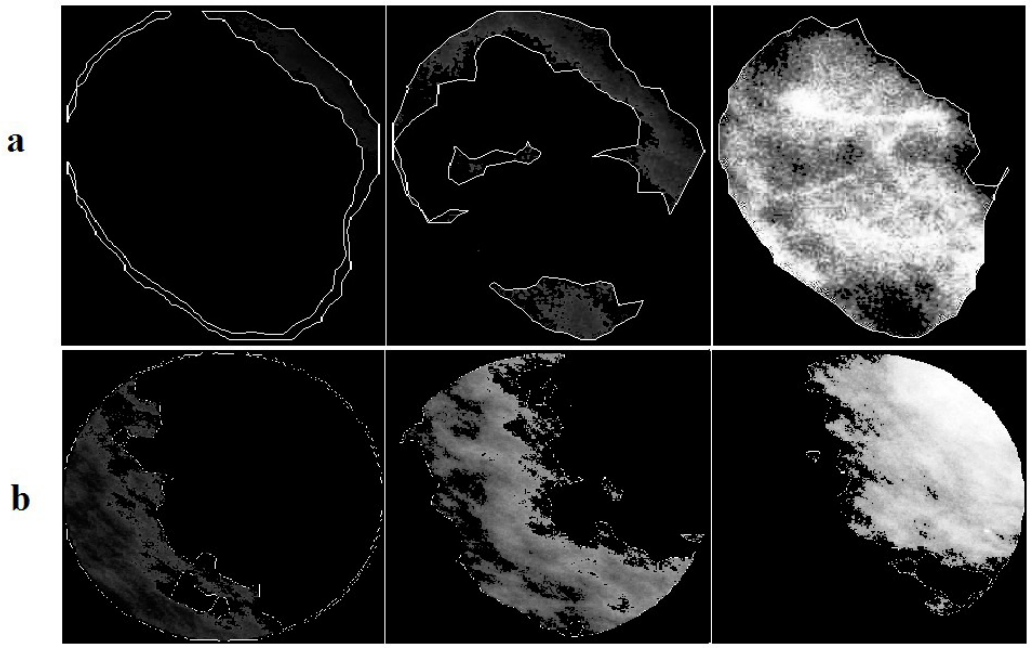

Figura 4.5: Divisão de faixas da quantização não linear: (a) Padrão Massa (b) Padrão Não Massa

Como se pode observar na Figura 4.5, as duas primeiras faixas da quantização não linear do padrão massa apresentam regiões pouco concentradas de pixels e regiões consideravelmente desconexas e a última faixa do padrão massa obteve grande concentração de pixels. O mesmo não ocorre no padrão não massa, pois a distribuição de pixels entre as faixas foi consideravelmente mais equivalente que a distribuição de pixels no padrão massa. É possível observar que cada faixa do padrão não massa apresentou uma distribuição de pixels mais compacta e a existência de poucas regiões desconexas.

Em relação às medidas geométricas que obtiveram acurácia menor que 90% tem-se o *Zernike Moments*, que não apresentou bons resultados em nenhum dos testes possivelmente devido a grande variedade de formas geométricas que as regiões massa e não massa assumem, a convexidade, que obteve acurácia de 73,31% devido a divisão dos pixels em faixas ocasionando em muitas regiões consideravelmente côncavas, e a densidade circular, que apesar de ter obtido acurácia maior que 80% não atingiu o mínimo desejado que foi de 90% devido suas densidades serem calculadas através de círculos concêntricos, isto é, devido, em algumas faixas, a distribuição de pixels em imagens de nódulos serem mais densas próximos ao centro de massa faz com que as densidades de círculos mais intrínsecos aos nódulos sejam relativamente semelhantes às densidades circulares de regiões não massa.

4.3 Resultados Benignos e Malignos

Após a classificação de regiões massa e não massa, o objetivo principal é identificar quais são as regiões que possuem o padrão benigno e as regiões com o padrão maligno. A Tabela 4.2 apresenta os resultados individuais e o resultado geral da classificação de nódulos benignos e malignos.

Tabela 4.2: Resultados gerais e individuais das medidas geométricas utilizadas para classificação de benigno e maligno.

	Sensibilidade (%)	Especificidade (%)	Acurácia (%)	N° de índices gerados para cada faixa
Circularidade 1	60,5	61,1	60,76	1
Circularidade 2	63,6	47,3	55,38	1
Compacidade	60,5	61,8	61,15	1
Desvio Padrão	59,7	51,1	55,38	1
Convexidade	72,9	65,6	69,23	1
Razão de Área	62	54,2	58,07	1
Rugosidade	53,5	45	49,23	1
Densidade Quadrangular	71,3	64,9	68,07	4
Densidade Circular	66,7	61,1	63,84	4
Hu Moments	62	47,3	54,61	7
Zernike Moments	63,6	52,7	58,07	49
Total	76,7	64,1	70,38	71

Na classificação de padrões malignos e benignos, os resultados apresentados pela Tabela 4.2 não foram bons. A acurácia geral foi de 70,38% e entre as medidas geométricas, a convexidade obteve melhor acurácia individual e *Hu Moments* obteve a pior acurácia individual que foi de 54,61%, errando a classificação de quase metade dos indivíduos. Para justificar os resultados, é importante observar o exemplo dado pela Figura 4.6 de um nódulo maligno e de um nódulo benigno.

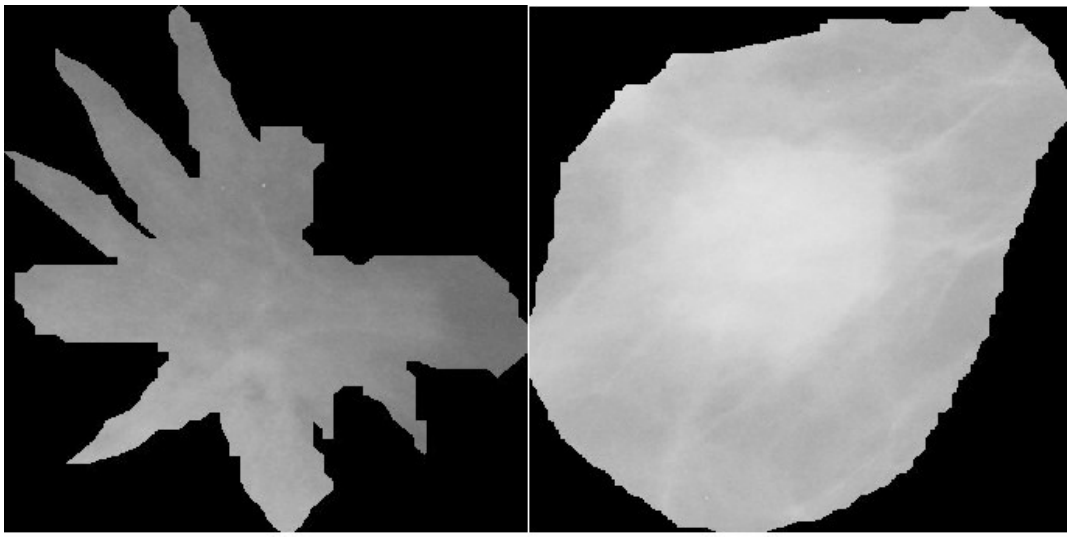

Figura 4.6: a) imagem de um nódulo benigno e b) imagem de nódulo maligno.

Como mencionado anteriormente na Seção 2, os nódulos benignos e malignos possuem comportamentos biológicos diferentes, pois enquanto o padrão benigno tende apresentar um crescimento lento, organizado e limitado, o padrão maligno tende apresentar um crescimento desordenado e invasivo (INCA, 2004). E através dos comportamentos distintos faz com que geralmente nódulos benignos apresentem formas mais regulares e nódulos malignos apresentassem formas relativamente irregulares. Entretanto, como se pode observar na Figura 4.6, este caso não obedece à regra. À esquerda pode-se observar que o nódulo benigno possui forma relativamente mais irregular que o nódulo maligno que está à direita. Nota-se que o nódulo maligno na Figura 4.6 tem o aspecto mais circular, compacto e conexo que são características que se esperam que os nódulos benignos apresentem. Outra possível justificativa para que os resultados não fossem satisfatórios é o fato dos nódulos serem segmentados manualmente pelos especialistas médicos. Isso abre a possibilidade de haver erros quanto ao recorte das regiões que ainda não foram invadidas pela neoplasia, devido à composição fisiológica da mama ser diferente para cada mulher e a possibilidade de essa estrutura sofrer alterações de acordo com a condição de saúde que a mulher se encontra (DUARTE, 2006).

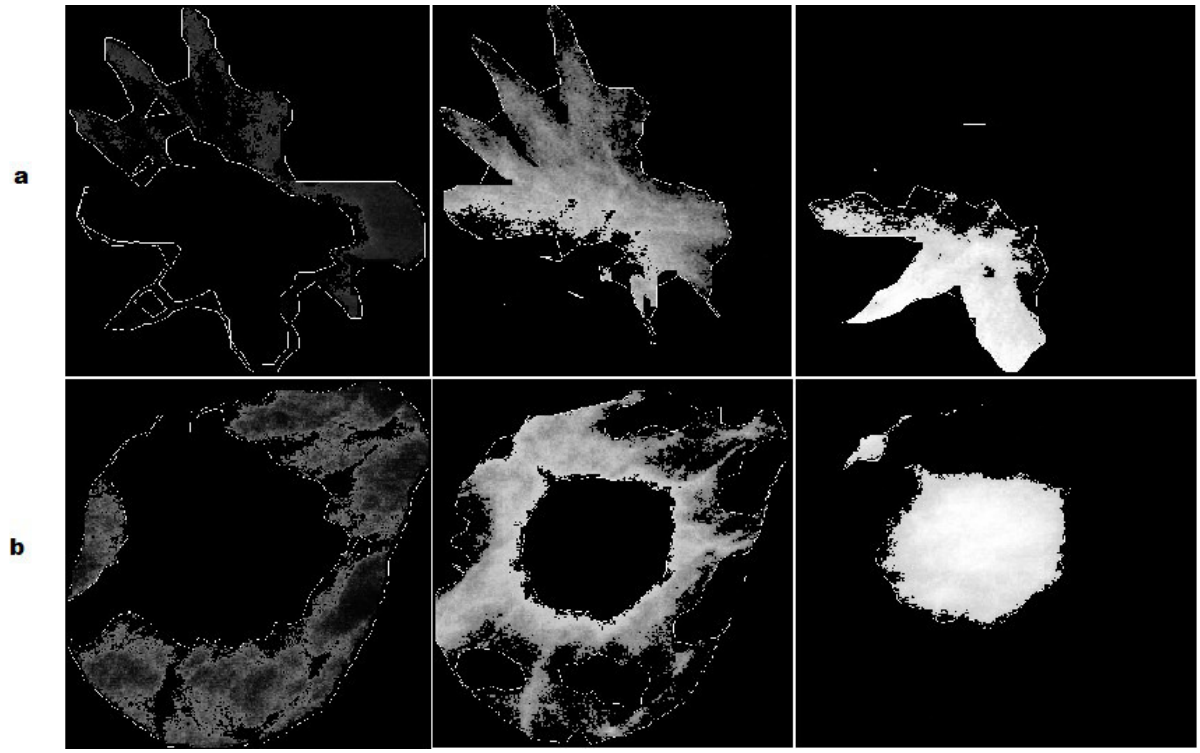

Figura 4.7: Divisão de faixas da quantização não linear: (a) Padrão Benigno (b) Padrão Maligno

Como se pode observar na Figura 4.7. A divisão de faixas dos padrões benigno e maligno apresentam comportamentos semelhantes. Ambos têm os *pixels* com os menores valores de intensidade próximo às bordas e os *pixels* com os maiores valores próximo ao centro de massa. É notório também que, em ambos os casos, as regiões próximo ao centro de massa são relativamente mais densas e mais compactas que as regiões próximo às bordas. Contribuindo assim, para que os resultados não obtivessem acurácia esperada de 90%.

Na Tabela 4.3 pode-se observar a comparação dos resultados dos trabalhos relacionados (Seção 1.1) com os resultados da metodologia proposta por este trabalho.

Tabela 4.3: Comparação do desempenho dos trabalhos relacionados e da metologia proposta por este trabalho

Trabalho	Sensibilidade	Especificidade	Acurácia	Objetivo
MELO (2010)	–	–	86,19%	Encontrar achados mamográficos
HOLSBACK (2012)	–	–	98,09%	Diagnóstico de câncer de mama
SOUZA (2011)	–	–	99,95%	Descrever padrões de região de imagem de mama.
SALES (2013)	80,11%	84,41%	84,38%	Classificação lesão e não lesão
SILVA et al. (2006)	81,81%	85,13%	–	Identificação de massas
MARTINS et al. (2006)	–	–	86,84%	Classificação benigno e maligno
ROCHA (2014)	92,96%	91,26%	92,2%	Classificação maligno e benigno
JUNIOR (2014)	97,30%	–	–	Detecção e diagnóstico de lesões
Metodologia Proposta	100%	99,1%	99,59%	Classificação massa e não massa
Metodologia Proposta	76,7%	64,1%	70,38%	Classificação benigno e maligno

5 CONCLUSÃO

Este trabalho apresenta o desenvolvimento de uma metodologia para extração de características de nódulos mamários para posteriormente classificá-los em massa ou não massa e depois classificá-los em maligno ou benigno.

Foram utilizadas imagens mamográficas da base DDSM (*Digital Database for Screening Mamography*) (HEAT et al., 1998). Sendo dois grupos: o primeiro grupo representava as imagens que não continham qualquer massificação e o segundo grupo que continha neoplasias malignas e benignas.

A adição da equalização de histograma contribuiu para melhora de distribuição dos valores de *pixels* e consequentemente na divisão de faixas da equalização não linear que foi importante para o estudo geométrico em cada região de cada imagem mamográfica. Os resultados na classificação de padrões massa e não massa atingiram acurácia de 93,48% e apenas a convexidade, a densidade circular e o *Zernike Moments* não atingiram individualmente acurácia acima de 90%. A metodologia provou ser eficaz em reconhecer os padrões massa e não massa. Para a classificação de neoplasias malignas e benignas, os resultados com acurácia atingindo 70,38% ainda não foram satisfatórios devido à variedade de formas geométricas que nódulos podem apresentar. Para este caso, a medida de convexidade obteve os melhores resultados individuais atingindo acurácia de 69,23% e sensibilidade de 72,9%. De forma geral, a sensibilidade foi maior que a especificidade, ou seja, houve mais acertos em reconhecer os nódulos malignos que reconhecer os nódulos benignos. Desse modo, as contribuições que esse trabalho apresenta são:

1. Implementação de uma metodologia de extração de características segundo a geometria, densidade e momentos de imagem;
2. Uma metodologia promissora no reconhecimento de padrões massa e não massa; E
3. A possibilidade de se utilizar índices geométricos com a adição de outros parâmetros para reconhecer padrões malignos e benignos;

5.1 Trabalhos Futuros

O desenvolvimento de uma metodologia para detecção de massas e posteriormente alertar ao médico se aquela dada massificação possui um padrão maligno pode contribuir para diagnosticar o câncer de mama ainda em estágio inicial e consequentemente aumentando a

probabilidade de cura. Com isso, melhorias são necessárias a fim de que este trabalho possa ser utilizado como auxílio ao médico. As melhorias são:

- Utilizar técnica de segmentação *MeanShift* após o recorte médico e o pré-processamento;
- Adicionar outros índices como a análise de textura na extração de características;
- Adicionar outros índices geométricos como: excentricidade, desproporção circular e solidez na extração de características;
- Adicionar a correlação de histograma na extração de características;

REFERÊNCIAS

BRASIL. (2003). Ministério da Saúde., disponível em: http://portalsaude.saude.gov.br/. Acesso em: 02 dez. De 2014.

BREIMAN, L.; CUTLER, ADELE. *Random Forest.* Disponível em: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html. Acesso em: 31 out. 2014.

DUARTE, D. L. *A Mama em Imagens.* Rio de Janeiro: Guanabara/Koogan, 2006.

GONZALEZ, R., & WOODS, R. *Digital Image Processing. 3. ed.* Person Prentice Hall. 2010.

HEATH, M.; BOWYER, K.; KOPANS, M.; KEGELMAYER, W.P. Digital Database for Screening Mamography. Exerpta Medica International, Disponivel em: http://marathon.csee.usf.edu/Mammography/Database.html. Acesso em: 02 jan. 2015.

HOLSBACK. N. "Método de Mineração de Dados para Diagnóstico do Câncer de Mama Baseado na Seleção de Variáveis". Dissertação de Mestrado. Universidade Federal do Rio Grande do Sul. 2012.

Instituto Nacional do Câncer (INCA). Atlas de mortalidade por câncer. Disponível em: <http://mortalidade.inca.gov.br/Mortalidade/prepararModelo05.action>. 2014. Acesso em: 02 nov. 2014.

JUNIOR. G.B. "Detecção de Regiões de Massas em Mamografias usando índices de Diversidade, Geoestatística e Geometria Côncava". Tese de Doutorado. Universidade Federal do Maranhão. 2014.

MARTINS, L.O.; SANTOS, A.M.; SILVA, A.C.; PAIVA, A.C. "Classificação de Tecidos Normais, Benignos e Malignos Utilizando Matrizes de Coocorrência e Redes Neurais Bayesianas em Imagens de Mamografia". JIM 2006 - I Jornada de Informática do Maranhão. Universidade Federal do Maranhão. 2006.

MUCKE, H. E. *Three-dimensional alpha shapes*. ACM Trans. Graph, v. 13. 43–72, 1994.

NIXON, M., & AGUADO, A. *Feature Extraction & Image Processing.* Elsevier. 2008

PAIVA, A.C.; SILVA, A.C.; JUNIOR, G.B.; OLIVEIRA, A.C.M. “Identificação de Massas em Mamografias usando Textura, Geometria e Algoritmos de Agrupamento e Classificação”. VI Workshop de Informática Médica - WIM2006. Universidade Federal do Maranhão. 2006.

ROCHA. S.V. “Diferenciação do Padrão de Malignidade e Benignidade de Massas em Imagens de Mamografias Usando Padrões Locais Binários, Geoestatística e Índice de Diversidade”. Tese de Doutorado. Universidade Federal do Maranhão. 2014.

SALES, A. M. V.; SILVA, A.C.; PAIVA, A.C. “Detecção de Lesões em Mamografias Através da Assimetria das Mamas e Extração de Características com Índice de Getis-Ord”. Cad. Pesq. São Luís, v. 20, n. 3. Universidade Federal do Maranhão. 2013.

SOUSA. U.S. “Treinamento De Redes Neurais Artificiais Utilizando Algoritmos Genéticos Em Plataforma Distribuída”. Monografia. Universidade Federal do Maranhão. 2011

TSUI. P.H.; LIAO, Y.Y.; CHANG, C.C.; KUO, W.H.; CHANG; YEH, C.H.“*Classification of Benign and Malignant Breast Tumors by 2-D Analysis Based on Contour Description and Scatterer Characterization”*.IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 29, N. 2. Fev. 2010.

VIEIRA, S.; SOARES, L.F.M; JUNIOR, J.; LUSTOSA, A.; BARBOSA, C.N.M.; BRITO, L.X.E.; FERREIRA, M.A.T. Oncologia Básica. 1. ed. Fundação Kixote. 2012.