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

JOHNATAN CARVALHO SOUZA

DIFERENCIAÇÃO DOS PADRÕES DE MALIGNIDADE E BENIGNIDADE DE MASSAS EM IMAGENS DE MAMOGRAFIA USANDO DESCRITORES DE FORMA E MÁQUINA DE VETORES DE SUPORTE

SÃO LUÍS
2016

JOHNATAN CARVALHO SOUZA

DIFERENCIAÇÃO DOS PADRÕES DE MALIGNIDADE E BENIGNIDADE DE MASSAS EM IMAGENS DE MAMOGRAFIA USANDO DESCRITORES DE FORMA E MÁQUINA DE VETORES DE SUPORTE

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.

Orientadora: Prof. Dra. Simara Vieira da Rocha

SÃO LUÍS

2016

Souza, Johnatan Carvalho

Diferenciação dos padrões de malignidade e benignidade de massas em imagens de mamografia usando descritores de forma e máquina de vetores de suporte/ Johnatan Carvalho Souza. – São Luis, 2016.

43f.

Orientadora: Simara Vieira da Rocha

Monografia (Graduação) – Universidade Federal do Maranhão, Curso de Ciência da Computação, 2016.

1. Diagnóstico de Câncer de Mama 2. Descritores de Forma 3. Relief Index 4. Average Slope

CDU 004.932:618.19-006.6

JOHNATAN CARVALHO SOUZA

DIFERENCIAÇÃO DOS PADRÕES DE MALIGNIDADE E BENIGNIDADE DE MASSAS EM IMAGENS DE MAMOGRAFIA USANDO DESCRITORES DE FORMA E MÁQUINA DE VETORES DE SUPORTE

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.

Aprovada em: 04 / 05 / 2016

BANCA EXAMINADORA

Prof. Dra. Simara Vieira da Rocha (Orientadora)
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão – UFMA

Prof. Dr. Geraldo Braz Junior
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão - UFMA

Prof. Ms. Carlos Eduardo Portela Serra de Castro
Centro de Ciências Exatas e Tecnologia - CCET
Universidade Federal do Maranhão - UFMA

À minha família, por acreditar e investir em meu potencial. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a motivação para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinho nessa caminhada.

AGRADECIMENTOS

Primeiramente a Deus que permitiu que tudo isso acontecesse, ao longo de minha vida, e não somente nestes anos como universitário, mas que em todos os momentos é o maior mestre que alguém pode ter.

Aos meus pais, a quem devo tudo o que tenho e sou hoje, pelo amor, incentivo e apoio incondicional.

Ao meu irmão, por me fazer perceber todos os dias a importância da busca pelo conhecimento.

À professora e orientadora Simara, cuja contribuição, com seus conselhos e esclarecimentos, foi primordial para este trabalho.

Ao professor Anselmo, pelos direcionamentos dados no desenvolvimento deste trabalho, pela orientação e suporte nos assuntos acadêmicos, e também pela oportunidade de trabalhar em um dos melhores laboratórios de processamento de imagens do país.

Aos meus amigos do NCA, com os quais pude contar em quase todos os momentos. Em especial, meus amigos da graduação Caio Eduardo, Giovanni, João, Jefferson e Wendell, pela amizade e companheirismo durante esses anos, e meus amigos com quem trabalhei na iniciação científica, Thamila e Rodrigo que muito me auxiliaram durante esse projeto.

Agradeço também a todos os professores por me proporcionar o conhecimento não apenas técnico e científico, mas também a manifestação do caráter e afetividade da educação no processo de formação profissional.

"Que os vossos esforços desafiem as impossibilidades, lembrai-vos de que as grandes coisas do homem foram conquistadas do que parecia impossível."

(*Charles Chaplin*)

RESUMO

O câncer de mama é o tipo de câncer mais comum entre as mulheres no mundo e no Brasil, depois do de pele não melanoma, respondendo por cerca de 25% dos casos novos a cada ano. É a quinta maior causa de morte por câncer em geral e a causa mais frequente de morte por câncer em mulheres. O câncer de mama, se identificado em estágios iniciais, quando as lesões possuem menos que um centímetro de diâmetro, apresenta prognóstico mais favorável e elevado percentual de cura. Por essa razão, uma detecção precoce é de suma importância para aumentar as chances de cura do paciente, e quanto mais informações o médico dispuser, mais preciso será o diagnóstico. O objetivo deste trabalho é investigar a aplicação de técnicas que descrevem a forma de massas, em imagens de mamografias, juntamente com reconhecimento de padrões, para caracterizar os padrões malignos e benignos dessas massas, provendo ao especialista informações adicionais, afim de auxiliá-lo no momento do diagnóstico.

A metodologia aplicada neste trabalho baseia-se em técnicas de processamento de imagens e reconhecimento de padrões. Foram utilizados os descritores de forma *Relief Index* e *Average Slope*. Os cálculos desses descritores são baseados nas características morfológicas das lesões, sendo aplicadas neste trabalho para caracterização das massas. Após essa etapa, foi realizado o reconhecimento de padrões pela Máquina de Vetores de Suporte (MVS). Os resultados obtidos foram: 83,73% de média de sensibilidade, 78,87% de média de especificidade e 81,18% de media de acurácia.

Palavras-chave: Diagnóstico de Câncer de Mama. Descritores de Forma. *Relief Index*. *Average Slope*.

ABSTRACT

Breast cancer is the most common type of cancer amongst women in the world and in Brazil, second to nonmelanoma skin cancer, representing over 25% of new cases each year. It is the fifth leading cause of death from cancer in women. Breast cancer, if diagnosed at an earlier stage, when the tumors are no bigger than one centimeter in diameter, has better prognosis and a high percentage of cure. Therefore, an early detection is extremely important to increase the pacient's chances of cure, and the more information is available to the physician, the more accurate diagnosis is. The purpose of this work is to investigate the usage of techniques that describe shape of tumors, in mammograms, alongside with pattern recognition, to generate characteristics that classify tumors as malignant or benign, in order to provide phisicians with additional information to increase the accuracy of diagnoses.

The methods presented in this work are based on image processing techniques and pattern recognition, in which the Relief Index and Avarage Slope shape descriptors were used to generate characteristics from the mammograms. The math behind these descriptors is based on the morphology of the tumors. The data provided by these descriptors were submitted to the Support Vector Machine (SVM) to classify the tumors in malignant and benign. The results were given as follow: 83,73% of average sensitivity, 78,87% of average specificity and 81,18% of average accuracy.

Keywords: Breast cancer diagnosis. Shape descriptors. Relief Index. Average Slope.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CAD	Computer-Aided Detection
CADx	Computer-Aided Diagnosis
DDSM	Digital Database for Screening Mammography
FN	Falso Negativo
FP	Falso Positivo
INCA	Instintuto Nacional do Cãncer
MVS	Máquina de Vetores de Suporte
ROI	Region of Interest
SVM	Support Vector Machine
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
2D	Bidimensional
3D	Tridimensional
α	Alfa
CGAL	Computational Geometry Algorithms Library

SUMÁRIO

1 INTRODUÇÃO

Câncer é o nome dado a um conjunto de mais de 100 doenças que têm em comum o crescimento desordenado de células, que invadem tecidos e órgãos. Dividindo-se rapidamente, estas células tendem a ser muito agressivas e incontroláveis, determinando a formação de tumores malignos, que podem espalhar-se para outras regiões do corpo. As causas de câncer são variadas, podendo ser externas ou internas ao organismo, estando inter-relacionadas. As causas externas referem-se ao meio ambiente e aos hábitos ou costumes próprios de uma sociedade. As causas internas são, na maioria das vezes, geneticamente pré-determinadas, e estão ligadas à capacidade do organismo de se defender das agressões externas (INCA, 2016).

O câncer de mama, de acordo com o Instituto Nacional de Câncer (INCA, 2015), será o tipo de câncer mais incidente entre mulheres no Brasil no ano de 2016. A previsão é de 57.960 novos casos desse tipo de câncer e com um risco estimado de 56,20 casos a cada 100 mil mulheres. Além disso, a incidência do câncer de mama tende a aumentar cerca de 1% todos os anos. O principal fator de risco para o câncer de mama é a idade. As taxas de incidência aumentam rapidamente até os 50 anos e, posteriormente, esse aumento ocorre de forma mais lenta. Contudo, outros fatores de riscos já estão bem estabelecidos como: aqueles que são relacionados à vida reprodutiva da mulher (idade da primeira gestação acima do 30 anos, anticoncepcionais orais, menopausa tardia e terapia de reposição hormonal), histórico familiar de câncer de mama e alta densidade do tecido mamário (razão entre o tecido glandular e o tecido adiposo da mama). Além desses, a exposição à radiação ionizante, mesmo em baixas doses, também é considerada um fator de risco, particularmente durante a puberdade, segundo mostram alguns estudos.

A prevenção primária dessa doença ainda não é totalmente possível em decorrência da variação de fatores e das características genéticas que estão envolvidas na sua etiologia. Uma nova característica de rastreamento factível para países com dificuldades orçamentárias tem sido estudada, e, até o momento, a mamografia, para mulheres com idade entre 50 e 69 anos e o exame clínico das mamas anualmente a partir dos 40 anos, é recomendada como método efetivo para a detecção precoce. A amamentação, a prática de atividade física e a alimentação saudável com a manutenção do peso corporal estão associadas a um menor risco de desenvolver esse distúrbio (INCA, 2016).

Apesar de ser considerado um câncer de bom prognóstico quando diagnosticado e tratado nas fases iniciais, as taxas de mortalidade por essa doença continuam elevadas, provavel-

mente porque ela é diagnosticada em estágios avançados. A sobrevida média após cinco anos na população de países desenvolvidos tem apresentado um pequeno aumento, cerca de 85%. Entretanto, em países em desenvolvimento, a sobrevida é próxima a 60%. No Maranhão, esse câncer tem uma taxa de 13,97% para cada 100 mil mulheres (INCA, 2016).

Uma das formas mais eficazes para detecção precoce do câncer de mama é o Exame Clínico da Mama (ECM). Quando realizado por um especialista, o ECM pode detectar tumor de até um centímetro, se superficial. Segundo o INCA (2004), um ECM deve contemplar os seguintes passos: inspeção estática e dinâmica, palpação das axilas e palpação da mama com a paciente em decúbito dorsal. A eficiência do exame é proporcional ao grau de habilidade e experiência do profissional para detectar qualquer anormalidade nas mamas examinadas. Ele deve ser realizado periodicamente e o médico indicará a necessidade de uma mamografia, que por sua vez, segundo Giger e MacMahon (1996), é o exame mais indicado para detecção precoce do câncer de mama, uma vez que possibilita a detecção visual de possíveis estruturas que possam evidenciar a presença ou ausência deste tipo de câncer.

Segundo a ACS (2014) a mamografia é atualmente uma das melhores técnicas de detecção precoce de lesões não palpáveis na mama. Contudo, a avaliação do exame mamográfico é subjetiva, requerendo grande habilidade do radiologista. Nas últimas décadas, técnicas computacionais vêm sendo desenvolvidas com o propósito de detectar automaticamente estruturas que possam estar associadas a tumores nos exames de mamografia, visando melhorar a taxa de detecção precoce de estruturas de interesse ligadas ao câncer de mama (GIGER; MACMAHON, 1996) (DENGLER et al., 1993). Esses esquemas de processamento são conhecidos como sistemas CAD ("*Computer Aided Detection*") e CADx ("*Computer Aided Diagnosis*"), e já estão presentes em diversos centros de diagnóstico por imagem, principalmente em países do primeiro mundo, como EUA e alguns países da Europa (TAYLOR et al., 2004) (FANDOS-MORERA et al., 1988). Os sistemas CAD e CADx fornecem uma segunda opinião, auxiliando o radiologista na interpretação de resultados que, em muitos casos, torna-se difícil devido às distorções que esse tipo de imagem sofre no seu processo de aquisição.

1.1 OBJETIVOS

Investigar a aplicação de descritores de forma e reconhecimento de padrões para caracterizar o padrão maligno e benigno de massas em imagens de mamografia, que tem por finalidade dar ao especialista um maior suporte no diagnóstico do câncer de mama. As técnicas

descritas neste trabalho poderão ainda ser incorporadas a um sistema do tipo CADx, e portanto, corroborar para o aumento da produtividade e melhoria nas taxas de diagnósticos mais precisos.

1.1.1 Objetivos Específicos

Para que o objetivo geral seja atingido, alguns objetivos específicos devem ser alcançados:

- Estudar a viabilidade da utilização de descritores de forma para extrair características geométricas de massas em imagens de mamografia;
- Estudar a viabilidade da utilização dos descritores de forma *Relief Index* e *Average Slope*, que originalmente foram aplicados em Evans (2013) na ecologia dentária de mamíferos, para extração de características das massas em imagens de mamografia;
- Estudar métodos para implementação dos descritores de forma.
- A partir dos descritores de forma implementados, extrair características das massas nas mamografias;
- Utilizar a MVS para testar as características produzidas em relação à sua capacidade de discriminar massas em imagens de mamografia nas classes maligno e benigno.

1.2 TRABALHOS RELACIONADOS

Na literatura, têm-se vários trabalhos reconhecidos que tratam do mesmo problema abordado pelo método proposto, ou seja: uma metodologia que auxilie os especialistas no diagnóstico de câncer de mama a partir de imagens de mamografia. Há também trabalhos que utilizam descritores de forma e reconhecimento de padrões para classificação em diferentes escopos.

Em (EVANS, 2013), os descritores que são utilizados neste trabalho, o *Relief Index* e o *Average Slope* são empregados na ecologia dentária de mamíferos. A caracterização dos dentes de mamíferos, por meio destes descritores, dão muita informação a respeito da vida de um animal e seus hábitos. O descritor *Relief Index* foi utilizado para diferenciar o tipo de dieta dos animais. O *Average Slope* foi empregado para criar um formato médio de um dente. Tendo em vista que a análise geométrica é comumente um fator primordial no diagnóstico de câncer de

mama, buscou-se neste trabalho testar a viabilidade dos descritores *Relief Index* e *Average Slope* no contexto de processamento de imagens e reconhecimento de padrões para massas em imagens de mamografia, devido ao seu bom desempenho quanto à caracterização da forma de objetos.

Em (ERPEN, 2004) são mostrados alguns estudos de caso que utilizam descritores de forma para reconhecimento de padrões, mais especificamente, na área de robótica móvel, para reconhecimento de comandos localizados no ambiente, por parte de um robô. O segundo e principal estudo de caso foi direcionado para aplicações de reconhecimento de placas de automóveis, que é bastante importante para o sistema de trânsito e pode ser utilizado em várias aplicações.

Visando estudar o comportamento dos contornos dos tumores de mama, (ALMEIDA et al., 2005) apresenta um método denominado *Curvature Scale Space* (CSS) (MOKHTARIAN; MACKWORTH, 1992), para analisar e classificar as lesões tumorais baseando-se na forma das mesmas.

Em (MENECHELLI et al., 2010) foi proposto o desenvolvimento de um software para auxiliar na classificação da forma de nódulos em imagens de mamografias, a partir de descritores geométricos que caracterizem as mesmas, utilizando a rede neural artificial multi-layer Perceptron (MLP) com algoritmo de *backpropagation*. Os principais atributos descritores da forma de objetos utilizados nesta pesquisa, foram: perímetro, área da região de interesse (ROI), irregularidade e área do retângulo mínimo. O melhor resultado obtido na validação foi de aproximadamente 82%, para classificação das massas em redonda e com distorção de arquitetura, que configuram um atributo de bastante peso na análise de malignidade ou benignidade de um nódulo mamário.

Em (SOUZA; GULIATO, 2009) foi proposto um método para um sistema de apoio ao diagnóstico (CAD) em que foram utilizados os descritores de forma *Complexity Index* e *Elliptic Variance*. Ambos descritores foram utilizados como extratores de características baseados na forma do contorno de uma dada lesão. Os resultados foram considerados bons, sendo o *Complexity Index*, o que obteve o melhor resultado com valor de área da curva ROC de 0,91.

Vários descritores de forma vêm sendo utilizados como ferramenta de reconhecimento de padrões em diversas áreas e apresentando bons números, inclusive na classificação de nódulos mamários para diagnóstico de câncer de mama. No entanto, ainda é necessário identificar técnicas que permitam melhorar e consolidar estes resultados. Verifica-se que a classificação de câncer de mama quanto a sua malignidade e benignidade é ainda um problema

em aberto, e que medidas de forma se mostram muito promissoras para essa discriminação.

1.3 ORGANIZAÇÃO DO TRABALHO

Esta monografia apresenta a seguinte organização:

No Capítulo 2, Fundamentação Teórica, seguem informações importantes para o contexto e entendimento do trabalho, tais como Processamento de Imagens e Descritores de Forma.

No Capítulo 3, Metodologia, serão abordados os métodos utilizados como ponto de partida para o desenvolvimento desse trabalho, tal como a aquisição de imagens, a aplicação dos descritores de forma *Relief Index* e *Average Slope* para a extração de características e a classificação através da Máquina de Vetores de Suporte.

No Capítulo 4, Resultados e Discussão, mostra-se os resultados alcançados na aplicação da metodologia.

No Capítulo 5, Conclusão, apresenta-se a conclusão do trabalho. Nela está contida uma retrospectiva do que foi apresentado neste trabalho juntamente com uma avaliação dos resultados obtidos e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os fundamentos teóricos para compreensão da metodologia utilizada neste trabalho.

2.1 PROCESSAMENTO DE IMAGENS

Uma imagem pode ser definida como uma função bidimensional, f(x, y), onde x e y são as coordenadas espaciais, e a amplitude de f em qualquer par de coordenadas (x, y) é a intensidade ou o chamado nível de cinza da imagem em um ponto (GONZALEZ; WOODS, 2002).

O processamento de imagens digitais compreende processos cujas entradas e saídas são imagens e, além disso, engloba os processos de extração de características a partir de imagens, incluindo o reconhecimento de objetos individuais (GONZALEZ; WOODS, 2002). Um dos objetivos principais desse processamento é melhorar a informação visual para interpretação humana e os dados para percepção automática através de máquinas (GONZALEZ; WOODS, 2002).

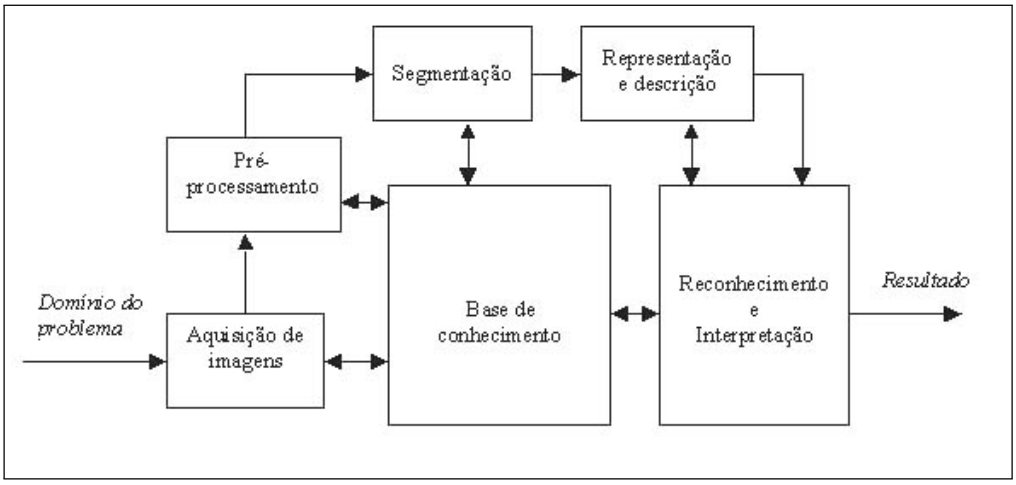
A área de processamento de imagens vem sendo objeto de crescente interesse por permitir viabilizar grande número de aplicações em duas categorias bem distintas: (1) o aprimoramento de informações pictóricas para interpretação humana; e (2) a análise automática por computador de informações extraídas de uma cena (MARQUES; VIEIRA, 1999).

A Figura 1 apresenta um esquema utilizado para demonstrar as diversas fases do processamento de imagem. Seguem-se as etapas: aquisição das imagens digitais, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação.

A primeira etapa então, é a aquisição da imagem, isto é, procedimento em que um digitalizador converte a imagem analógica para digital. Nesta etapa, há dois elementos principais: um dispositivo físico sensível a uma faixa de energia no espectro eletromagnético (como raio-X, ultravioleta, espectro visível ou raios infravermelhos), que produz na saída um sinal elétrico proporcional ao nível de energia detectado; e um digitalizador propriamente dito - converte o sinal elétrico analógico em informação digital (MARQUES; VIEIRA, 1999).

A segunda etapa é o pré-processamento das imagens adquiridas. Esta etapa tem como finalidade melhorar certos elementos da imagem para que se possa aumentar as chances de sucesso das etapas seguintes. Tipicamente envolve técnicas para o realce de contrastes,

Figura 1 – Etapas do Processamento da Imagem Digital.



Fonte: (GONZALEZ; WOODS, 2002)

remoção de ruído e isolamento de regiões cuja textura indique a probabilidade de informação alfanumérica.

A próxima etapa é a segmentação, que consiste em dividir uma imagem de entrada em partes ou objetos constituintes. É a técnica aplicada para extrair das imagens apenas as regiões que interessam no o processamento. Em geral, a segmentação automática é uma das tarefas mais difíceis no processamento de imagens digitais.

A etapa de representação e descrição, é também conhecida como extração de características. Essa etapa tem como finalidade determinar características que resultam em informação quantitativa de interesse ou que sejam básicas para discriminação entre classes distintas. O conjunto dessas medidas constituirá um vetor de características que definirão um padrão estimado para aquela região específica.

Finalmente, o último estágio envolve reconhecimento e interpretação. Reconhecimento é o processo que atribui um rótulo a um objeto, baseado na informação fornecida pelo seu descritor. A interpretação envolve a atribuição de significado a um conjunto de objetos reconhecidos.

2.2 DESCRITORES DE FORMA

Descritores de forma são representações compactas e expressivas de objetos adequadas para solucionar problemas como reconhecimento, classificação e recuperação de formas, que são tarefas computacionalmente custosas quando executadas sobre grandes quantidades de

dados. (FLORIANI; SPAGNUOLO, 2008).

Para entender a motivação do uso de descritores de forma em processamento de imagens digitais, é importante conhecer o conceito de análise de forma que pode ser apresentado como o estudo das características geométricas de um objeto com objetivo de discriminá-lo em algum contexto. Na análise automática de formas geométricas, por exemplo, o objetivo é utilizar um computador para detectar objetos que possuem formas semelhantes. Para isso, os objetos devem estar representados digitalmente. Geralmente, são utilizados modelos de representação da fronteira, similar a uma casca, do objeto. Entretanto, representações de volume também podem ser utilizadas (DELFOUR; ZOLESIO, 2001).

A partir da representação digital do objeto, é necessário que haja pelo menos uma simplificação dessa representação para que alguma comparação possa ser feita. Essas simplificações são, na maioria das vezes, denominadas descritores de forma (DELFOUR; ZOLESIO, 2001).

Por possuir aplicações em diversas áreas e demonstrando bons resultados em trabalho relacionados a processamento de imagens e reconhecimento de padrões, buscou-se neste trabalho investigar a aplicação e a viabilidade dos descritores de forma *Relief Index* e *Average Slope* para caracterizar os padrões malignos e benignos de massas em mamografias. A seguir serão apresentados os descritores propostos pela metodologia neste trabalho.

2.2.1 Relief Index

O *Relief Index*, entre outros descritores de forma, foi apresentado em (UNGAR; WILLIAMSON, 2000) como uma medida de quantidade de relevo de um objeto. O *Relief Index* foi então utilizado para descrever as superfícies oclusais dos dentes de alguns mamíferos, que segundo (UNGAR; WILLIAMSON, 2000), estão correlacionadas aos hábitos alimentares das espécies e portanto, fornecem informações de bastante relevância para o campo da ecologia.

O *Relief Index* foi definido como a razão entre a área da superfície tridimensional e a área bidimensional do objeto (M'KIRERA; UNGAR, 2003).

Este *index* é definido como:

$$ri = \frac{SA}{PA} \tag{2.1}$$

O *Relief Index* também foi definido por Boyer (2008) como:

$$ri = ln(\sqrt{SA}/\sqrt{PA}) \tag{2.2}$$

onde SA é a área da superfície 3D e PA é a área da superfície 2D.

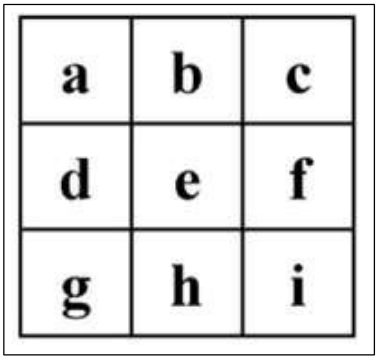
Na etapa de geração de características, as duas versões do descritor são utilizadas.

2.2.2 Average Slope

Segundo Ungar e Williamson (2000), *slope* (relevo, declive), é uma medida de relevo topográfico de uma superfície oclusal. O *Average Slope* seria portanto, o valor médio dos declives de uma região. Foi utilizado como um quantificador do desgaste proveniente do atrito nos dentes dos animais estudados. Com o uso desse descritor de forma, foi possível, mais uma vez, extrair características que estão correlacionadas a certos hábitos das espécies estudadas.

Nesse trabalho, o *Average Slope* foi computado por meio da técnica *average maximum technique* (BURROUGH, 1986), um algoritmo muito utilizado em Sistemas de Informação Geográfica que calcula a taxa de mudança das direções vertical e horizontal de um ponto em uma superfície. Essa taxa de mudança determina o *slope* naquele ponto. O algoritmo é aplicado sobre uma janela 3x3 na célula central e seus 8 vizinhos, para todos os *pixels* da imagem. A Figura 2 mostra o esquema de janela 3x3 utilizado. Os valores da célula central e seus 8 vizinhos determinam a taxa de mudança na direções vertical e horizontal. Os vizinhos são identificados como letras de *a* até *i*, sendo *e* a célula na qual o *slope* está sendo calculado.

Figura 2 – Esquema de janela 3x3 para o cálculo do *slope* de um ponto utilizando o *average maximum technique*.



Fonte: Adaptado de (PECKHAM; JORDAN, 2007)

O *slope* é comumente expresso em graus e pode ser calculado pela expressão:

$$S = arctan(\sqrt{[dz/dx]^2 + [dz/dy]^2}) * 57.29578 \tag{2.3}$$

onde $[dz/dy]$ é a taxa de mudança na direção vertical e $[dz/dx]$ na horizontal e são dados por:

$$[dz/dy] = ((g+2h+i) - (a+2b+c))/(8 * y\_cellsize) \tag{2.4}$$

$$[dz/dx] = ((c+2f+i) - (a+2d+g)/(8 * x\_cellsize) \tag{2.5}$$

com *x_cellsize* e *y_cellsize* representando os tamanhos de células nas direções horizontal e vertical e o valor 57.29578 sendo um fator que representa $180/\pi$ e que é multiplicado ao valor do *slope* para que a unidade de medida resultante seja em graus (GIS AG MAPS, 2011).

O *Average Slope* é então definido como a média total dos *slopes* de uma imagem.

$$as = \frac{\sum_{i=1}^{n} S_i}{n} \tag{2.6}$$

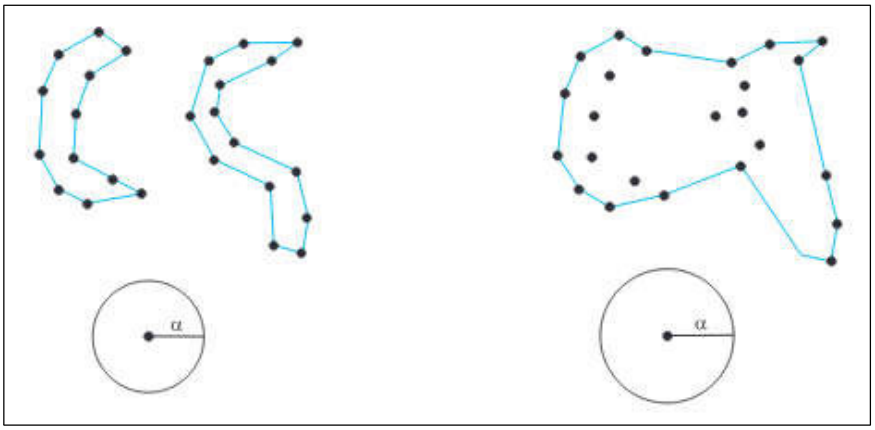
onde *n* representa o número de *pixels* da imagem.

2.3 ALPHA-SHAPES

Assuma que dado um conjunto de pontos, queremos descrever a figura formada por esses pontos. Haverão provavelmente várias interpretações da figura formada, e o α-*shape S* é uma delas. A Figura 3 ilustra o α-*shape* de um conjunto de pontos para diferentes valores de α.

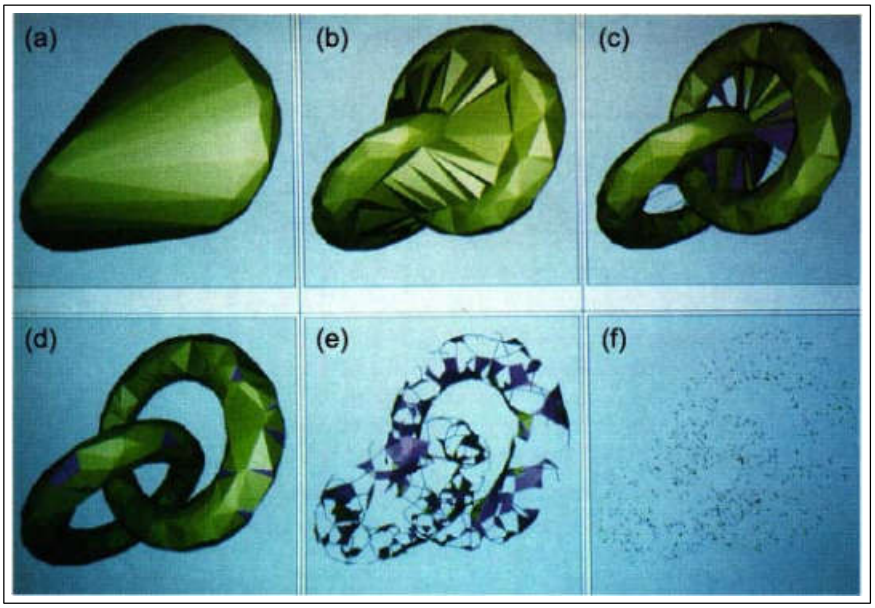
O conceito de α-*shapes* é uma generalização do fecho convexo de um conjunto de pontos. Seja X um conjunto finito de pontos em $\mathbb{R}^3$, e α um número real no intervalo $0 \leqslant \alpha \leqslant \infty$. Quando $\alpha = \infty$, o α-*shape* é idêntico ao fecho convexo de *X*. Porém, à medida que α decresce, o α-*shape* se retrai pelo aparecimento gradual de cavidades. Essas cavidades podem se unir formando túneis, e até mesmo buracos podem aparecer (EDELSBRUNNER; MÜCKE, 1994). A Figura 4 ilustra a reconstrução de um conjunto de pontos pelo α-*shape*, para diversos valores de α.

Figura 3 – α-*shapes* para diferentes valores de α em $\mathbb{R}^2$.



Fonte: (VATH, 2007)

Figura 4 – Reconstrução de uma nuvem de pontos por α-*shapes*. Na figura (a) $\alpha = +\infty$, formando o fecho convexo do conjunto de pontos. Na figura (f) $\alpha = 0$, assim restam apenas os pontos. As figuras (b), (c), (d) e (e) mostram valores intermediários do parâmetro α, sendo a figura (d) uma reconstrução obtida com um valor ideal para α.



Fonte: Adaptado de (EDELSBRUNNER; MÜCKE, 1994)

2.4 RECONHECIMENTO DE PADRÕES

Em (LOONEY, 1997), um padrão é definido como tudo aquilo para o qual existe uma entidade nomeável representante, geralmente criada através do conhecimento cultural humano. O reconhecimento de padrões visa determinar um mapeamento que relacione as propriedades ex-

traídas de amostras com um conjunto de rótulos (entidade nomeável representante), apresentando a restrição de que amostras com características semelhantes devem ser mapeadas ao mesmo rótulo. Os algoritmos que estabelecem este mapeamento são denotados como algoritmos de classificação ou classificadores (PEDRINI; SCHWARTZ, 2008).

O processo de classificação pode ser feito de duas formas, supervisionada e não supervisionada. A classificação supervisionada ocorre quando o classificador considera classes pré-definidas e uma etapa de treinamento deve ser executada para que os parâmetros que caracterizam cada classe sejam obtidos. Na classificação não-supervisionada não se dispõe de parâmetros ou informações coletadas previamente à aplicação do algoritmo de classificação, e todas as informações devem ser obtidas a partir das próprias amostras a serem rotuladas (PEDRINI; SCHWARTZ, 2008).

Neste trabalho foi utilizado o processo de classificação supervisionada MVS, para realizar o reconhecimento de padrões de tecidos da mama (massas), de modo a determinar sua natureza maligna ou benigna.

2.4.1 Máquina de Vetores de Suporte

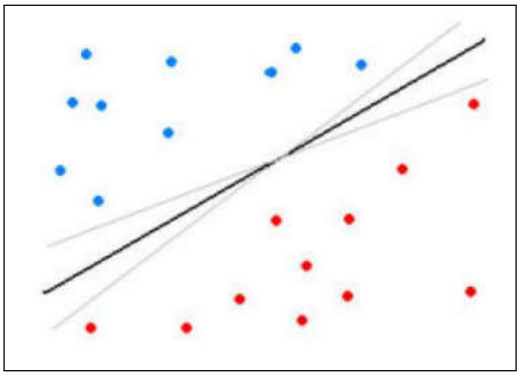
A Máquina de Vetores de Suporte (MVS) é uma técnica de aprendizagem supervisionada, usada para estimar uma função que classifique dados de entrada em duas classes. O princípio básico é a construção de um hiperplano como superfície de decisão, cuja margem de separação entre as classes seja máxima (VAPNIK, 1998). Por hiperplano entende se uma superfície de separação de duas regiões em um espaço multidimensional, em que o número de dimensões pode ser, até, infinito.

A Figura 5 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

Há casos em que podem existir vários possíveis hiperplanos de separação, mas MVS busca apenas encontrar o que maximize a margem entre os exemplos de treinamento.

Seja o conjunto de amostras de treinamento (x_i, y_i) sendo, $x_i \in \mathbb{R}^n$ o vetor de entrada, y_i classificação correta das amostras e i = 1, 2,..., n o índice de cada ponto amostral. O objetivo da classificação é estimar a função $f(x) : \mathbb{R}^n \rightarrow \{\pm 1\}$ separe corretamente os exemplos de teste em classes distintas.

Figura 5 – Separação entre duas classes através de hiperplanos



Fonte: (NASCIMENTO, 2012)

A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando valores tais que a seguinte relação seja satisfeita:

$$y_i((w.x_i) + b) \geqslant 1 \tag{2.7}$$

sendo w o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação à origem, os valores ótimos de w e b serão encontrados de acordo com a restrição dada pela Equação 2.7 ao minimizar a seguinte equação:

$$\phi(w) = \frac{w^2}{2} \tag{2.8}$$

O MVS ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não seja possível. Isso graças à inclusão de variáveis de folga, que permitem que as restrições presentes na Equação 2.7 sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 2.9, de acordo com a restrição imposta na Equação 2.10. C é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento e deve ser selecionado pelo usuário.

$$\phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{2.9}$$

sujeito à

$$y_i((w.x_i)+b)+\xi_i \geqslant 1 \tag{2.10}$$

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 2.11 O objetivo então passa a ser encontrar os multiplicadores de Lagrange a_i ótimos que satisfaçam a Equação 2.12 (CHAVES, 2006).

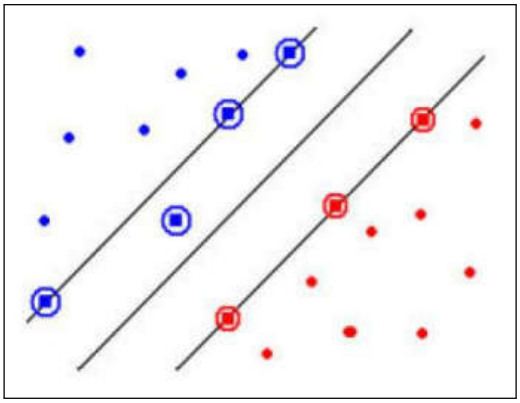
$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j (x_i, x_j) \tag{2.11}$$

$$\sum_{i=1}^{N} a_i y_i = 0, \qquad 0 \leqslant a_i \leqslant C \tag{2.12}$$

Apenas os pontos onde a restrição dada pela Equação 2.7 é exatamente igual à unidade têm correspondentes $a_i \neq 0$. Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos têm fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento. A Figura 6 destaca os pontos que representam os vetores de suporte.

Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

Figura 6 – Vetores de Suporte (destacado por círculos).



Fonte: (NASCIMENTO, 2012)

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessária uma transformação não-linear que transforme o espaço de entrada (dados) para um novo espaço (espaço de características).

Esse espaço deve apresentar dimensão suficientemente grande e, através dele, a amostra pode ser linearmente separável. Dessa maneira, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características em vez do espaço de entrada original. Essa construção depende do cálculo de uma função K de núcleo de um produto interno (HAYKIN; ENGEL, 2001). A função K pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 2.13 mostra o resultado da Equação 2.11 com a utilização de um núcleo K.

$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \tag{2.13}$$

Uma importante família de funções de núcleo é a função de base radial, muito utilizada em problemas de reconhecimento de padrões e também empregada neste trabalho. A função de base radial é definida por:

$$K(x_i, x_j) = exp(-\gamma \, ||x_i - x_j||^2) \tag{2.14}$$

onde $\gamma = 1/\sigma^2$, sendo σ a variância.

2.5 VALIDAÇÃO DOS RESULTADOS

Na etapa de validação dos resultados produzidos procura-se avaliar o desempenho da metodologia por meio de uma análise estatística dos resultados dos testes. Na análise de imagens médicas, geralmente utilizam-se as medidas de estatística descritiva sensibilidade (S), especificidade (E) e acurácia (A) (BLAND, 2000). Essas métricas são calculadas a partir de quatro situações em relação ao diagnóstico:

- VP – Verdadeiro Positivo: o teste é positivo e o paciente tem a doença;
- FP – Falso Positivo: o teste é positivo, mas o paciente não tem a doença;
- VN – Verdadeiro Negativo: o teste é negativo e o paciente não tem a doença;

- FN - Falso Negativo: o teste é negativo, mas o paciente tem a doença;

A acurácia corresponde a taxa de casos classificados corretamente sobre o numero total de casos:

$$A = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.15}$$

A sensibilidade define a proporção de pessoas com a doença de interesse que têm o resultado do teste positivo. Indica quão bom é o teste para identificar indivíduos doentes:

$$S = \frac{VP}{VP + FN} \tag{2.16}$$

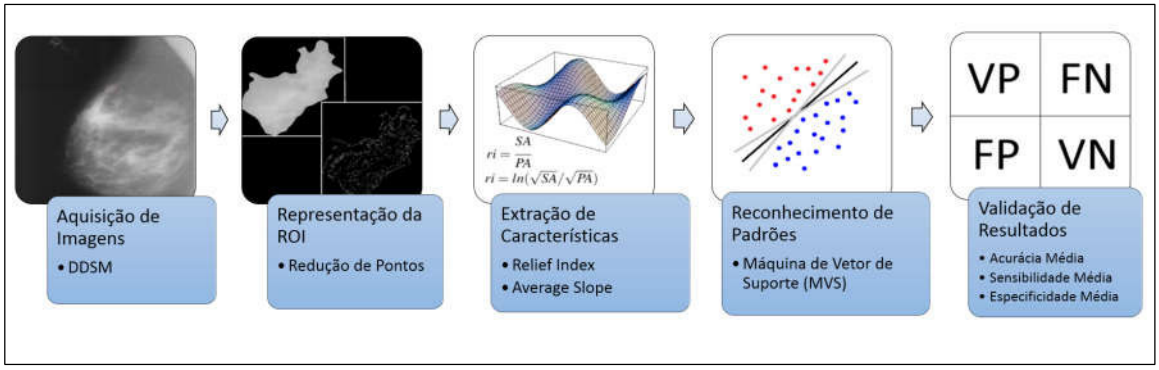
A especificidade define a proporção de pessoas sem a doença de interesse que tem o resultado do teste negativo. Indica quão bom é o teste para identificar indivíduos não doentes:

$$E = \frac{VN}{VN + FP} \tag{2.17}$$

3 METODOLOGIA

Neste capítulo são descritas as etapas realizadas na metodologia deste trabalho, que tem por objetivo classificar massas em imagens de mamografia por padrões de malignidade e benignidade. A Figura 7 mostra o fluxo das atividades realizadas. Nas próximas seções cada uma dessas etapas será descrita detalhadamente, mostrando-se suas sub-etapas, os problemas encontrados e as pesquisas realizadas durante o trabalho para o alcançar o objetivo do trabalho.

Figura 7 – Etapas da metodologia proposta.



Fonte: Elaborado pelo autor.

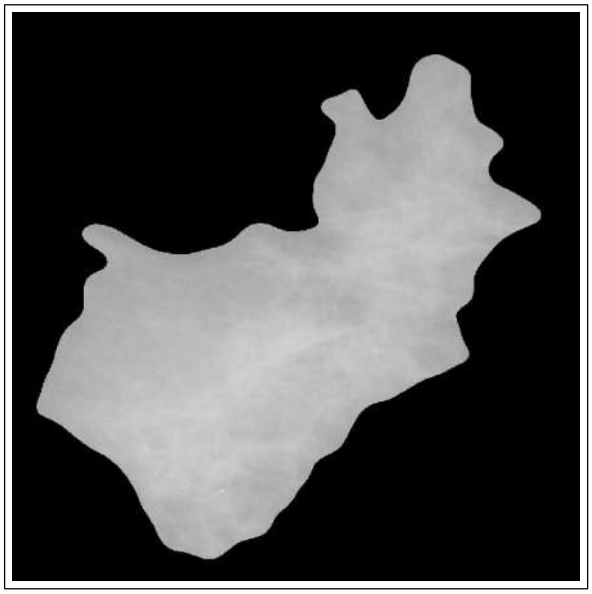
3.1 AQUISIÇÃO DE IMAGENS

Esta é a etapa em que foram obtidas as amostras de mamografias utilizadas nos testes da metodologia proposta. Foi utilizada a base pública de mamografias digitalizadas DDSM (*Digital Database for Screening Mammography*), disponível na internet (HEATH et al., 1998). A base é formada por 2620 exames de pacientes de diferentes origens étnicas e raciais. Cada exame contém duas imagens de cada mama, nas projeções médio-lateral oblíqua e crânio-caudal. Além disso, são disponibilizadas informações sobre a paciente, tal como a idade e a densidade da mama. Junto com as imagens que apresentam áreas suspeitas (massas) é fornecido um arquivo de descrição de lesão (*overlay*), contendo a quantidade de lesões presentes na mamografia, a localização da lesão, o tipo de lesão, o contorno da lesão e seu diagnóstico.

Dado que o objetivo deste trabalho é diagnosticar possíveis casos de câncer de mama, de todo o contexto da imagem de mamografia, o objeto de interesse se resume a região da massa na mama. Portanto, os descritores estudados neste trabalho são aplicados somente sobre a região de interesse (ROI). A Figura 8 mostra um exemplo de ROI utilizado no processo.

Tendo em vista que o foco deste trabalho é mostrar a aplicabilidade dos descritores *Relief Index* e *Average Slope* para classificação de massas, não foi feito nenhum processo de segmentação de imagens neste trabalho. A seleção da amostra foi feita sobre uma base previamente segmentada pelo grupo de pesquisa do Núcleo de Computação de Aplicada da Universidade Federal do Maranhão.

Figura 8 – Exemplo de região de interesse.



Fonte: Elaborado pelo autor.

Na etapa de geração de características foram utilizados um total de 118 imagens, 61 benignas e 57 malignas, contendo os bounding boxes das regiões de massas. A escolha das imagens foi realizada a partir de uma análise visual das ROIs em que buscou-se identificar as regiões de massa cujas fronteiras se encontravam melhor definidas e portanto, com maior potencial de discriminação geométrica.

3.2 REPRESENTAÇÃO DA IMAGEM

Esta etapa tem como propósito geral criar uma representação da imagem que melhor se adeque a uma determinada metodologia. Como apresentado na Seção 1.1, buscou-se neste trabalho investigar a aplicação dos descritores de forma *Relief Index* e *Average Slope*, que avaliam o relevo de superfícies tridimensionais (EVANS, 2013). A necessidade da etapa de representação nessa metodologia é justificada, especificamente, pela busca de uma solução para um problema

encontrado no cálculo do descritor *Relief Index*.

Como descrito na Subseção 2.2.1, o cálculo do *Relief Index* depende basicamente de dois parâmetros: (1) a área da superfície 2D e (2) a área da superfície 3D. A área da superfície 2D pode ser facilmente calculada a partir do contorno encontrado da ROI. Entretanto, o processo para calcular a área da superfície 3D apresenta um maior nível de complexidade. A seguir é apresentado o método utilizado neste trabalho para o cálculo da área da superfície 3D, juntamente com a justificativa para a utilização da etapa de representação de imagem.

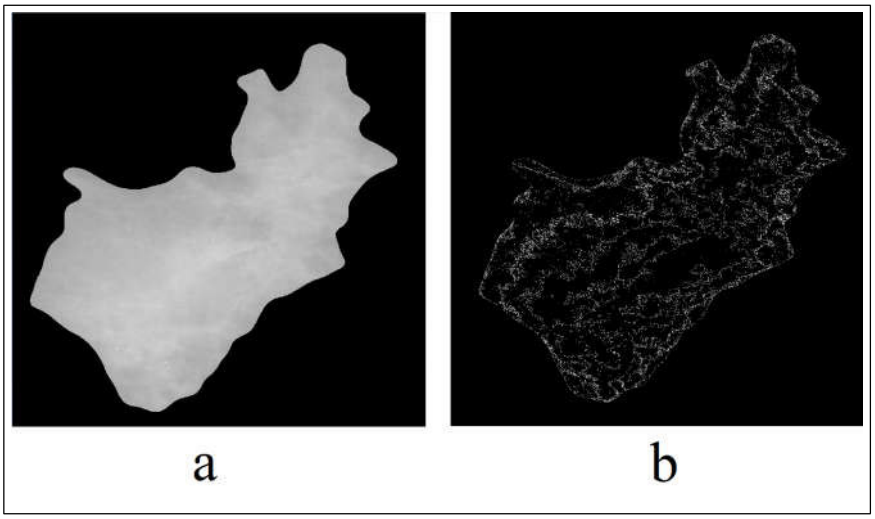
Para reconstruir uma superfície 3D, é necessário o conjunto de pontos no espaço tridimensional que definem essa superfície. Como as imagens de mamografia possuem apenas duas dimensões, o conjunto de pontos que definem a superfície foram definidos como segue: para cada *pixel p* da imagem de mamografia foi criado um ponto tridimensional *P* equivalente no qual as coordenadas *x, y* e *z* são compostas pelas coordenadas espaciais do *pixel p* na imagem e pela amplitude, ou nível de cinza, desse *pixel*, gerando consequentemente um número de pontos igual ao número de *pixels* na imagem.

O número de pontos gerados por esse procedimento se mostrava com frequência demasiado, o que ocasionava constantes estouros de memória na execução do algoritmo do cálculo de área utilizando α*-shapes*. Por essa razão, foi necessário buscar um método alternativo para representar e reconstruir a superfície 3D da ROI.

A representação utilizada consiste em uma versão simplificada das imagens através de uma redução do número de pontos gerados. Essa redução é feita da seguinte maneira: Os *pixels* na imagem da ROI são divididos por faixas de intensidade, de acordo com o seus níveis de cinza. Sobre cada faixa de *pixel*, é aplicado um procedimento para retirar somente os contornos das formas encontradas em cada faixa. A representação final é a união dos *pixels* que representam os contornos encontrados em cada uma dessas faixas. A Figura 9 mostra uma comparação entre uma ROI com todos os *pixels* e uma ROI com os *pixels* reduzidos, pela Figura 10 pode-se observar um exemplo da separação de uma ROI em 5 faixas de intensidade de pixel. A Figura 11 ilustra os contornos das regiões divididas por faixas e Figura 12 mostra a união sendo realizada faixa a faixa.

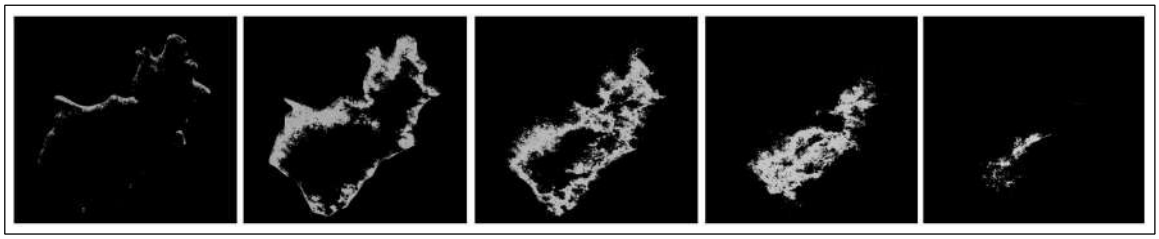
Após a realização de testes para diversos números de faixas, observou-se de forma empírica que a representação da imagem, pelos contornos das formas encontradas em 20 faixas de níveis de cinza, era capaz de reconstruir as superfície das ROI de forma satisfatória, mesmo com a perda de informação resultante pela redução do número de pontos. Como visto na Seção

Figura 9 – Uma região de interesse (a) ao lado de sua representação correspondente (b).



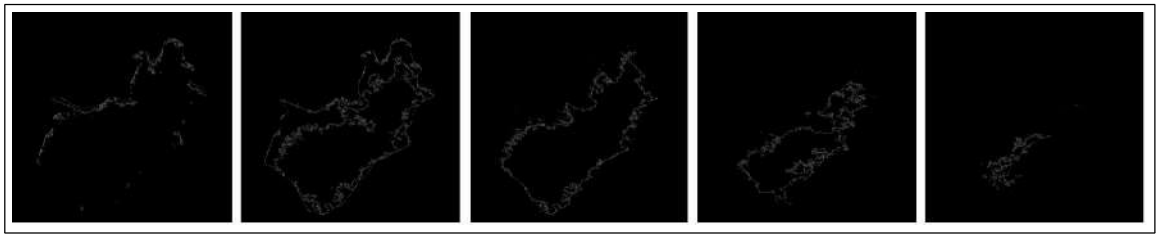
Fonte: Elaborado pelo autor.

Figura 10 – Exemplo da divisão dos *pixels* de uma ROI em 5 faixas.



Fonte: Elaborado pelo autor.

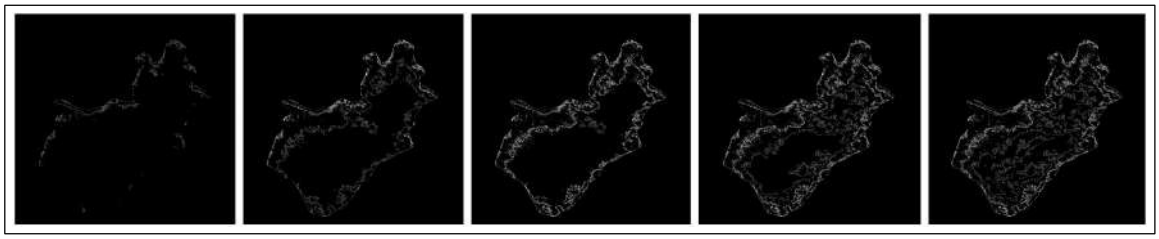
Figura 11 – Contornos da regiões para cada faixa de intensidade na Figura 10.



Fonte: Elaborado pelo autor.

2.3, um α-*shape* é capaz de, na maioria dos casos, reconstruir uma superfície de maneira ótima a partir do valor de um α ideal. A reconstrução da superfície e a estimação do α ótimo foram feitas utilizando-se a implementação de α-*shapes* da biblioteca de geometria computacional CGAL (*Computational Geometry Algorithms Library*). Após o processo de redução dos pontos, o procedimento para calcular as áreas das superfície 3D eram bem sucedidos, eliminando assim, o problema dos estouros de memória.

Figura 12 – União dos contornos obtidos faixa a faixa.



Fonte: Elaborado pelo autor.

3.3 EXTRAÇÃO DE CARACTERÍSTICAS

Essa etapa tem como objetivo produzir medidas descritivas das imagens, as quais formarão os vetores de características que serão usados na etapa de classificação. Os procedimentos empregados neste trabalho encontram-se detalhados na sequência.

3.3.1 Extração de Caraterísticas usando Relief Index

O processo de extração de características usando o *Relief Index* pode ser dividido em três fases: (1) cálculo da área da superfície 2D, (2) cálculo da área da superfície 3D e (3) cálculo da razão entre os valores de área 2D e 3D.

Na primeira fase, a área da superfície 2D é facilmente calculada a partir do contorno da região interesse. A segunda fase consiste em utilizar α*-shapes* 3D para reconstrução da superfície da massa a partir do conjunto de pontos mapeados dos *pixels* da região de interesse.

Primeiramente, os pixels da ROI são reduzidos pelo procedimento descrito na Seção 3.2. A partir da representação criada, são gerados os pontos que definem a superfície da massa. Então, utiliza-se o α*-shapes* para reconstruir a área da superfície tridimensional – agora sem os estouros de memória – e finalmente, calcula-se área da superfície a partir dessa reconstrução.

Munido dos valores de área de superfície 2D e 3D, finaliza-se a última fase da extração de características com *Relief Index*. Como visto na Subseção 2.2.1, o descritor é calculado de duas maneiras em estudos diferentes. De acordo com (M'KIRERA; UNGAR, 2003), o cálculo do *Relief Index* é realizado pela razão entre as áreas, e segundo (BOYER, 2008), pelo logaritmo da razão entre as raízes quadradas das áreas. Nesta metodologia buscou-se estudar a aplicação das duas versões do descritor, gerando desta forma, duas novas características.

3.3.2 Extração de Características usando Average Slope

A extração de características com o descritor *Average Slope* consiste em calcular a média dos valores de *slope* para cada *pixel* da ROI.

O cálculo do *slope* é feito como descrito na subseção 2.2.2, onde uma máscara 3x3 percorre todos os *pixels* da ROI, calculando os valores de *slope*, e armazenando-os em um vetor. Após o cálculo dos *slopes* de cada pixel, a nova característica é gerada a partir da média dos valores armazenados na vetor.

3.4 RECONHECIMENTO DE PADRÕES

Na etapa anterior, foram gerados vetores de características das amostras através do cálculo dos descritores *Relief Index* e *Average Slope*. O processo de classificação tem como objetivo analisar os padrões encontrados na etapa de extração de características utilizando o classificador MVS. De posse da base de características gerada, é necessário normalizar os diferentes valores das características para uma faixa de valores comuns como -1 a 1. Esse mecanismo ajuda o classificador a convergir com maior facilidade na etapa de treinamento, e também padroniza a distribuição de valores das variáveis, as quais podem assumir diferentes domínios (BRAZ, 2008).

A base de características foi dividida randomicamente em dois grupos: base de treino e base de teste. Os percentuais usados neste trabalho para treino e teste foram respectivamente: 50/50, 60/40, 70/30 e 80/20. Para cada proporção foram realizadas 5 repetições do teste de forma aleatória. Como foi usado o núcleo radial do MVS, cada experimento teve os parâmetros de custo C e grau de complexidade da função de mapeamento γ. Os valores desses parâmetros são estimados através de busca exaustiva realizada pelo *script* em *python grid.py*, pertencente ao pacote LIBSVM (CHANG; LIN, 2011). Este script busca, através de validação cruzada, a melhor combinação de parâmetros para a base, retornando o melhor percentual de acerto total sobre as amostras de treino e teste.

Durante a etapa de treinamento é gerado o modelo que o MVS utiliza para classificar as amostras de teste. O mecanismo de classificação, que desconhece as amostras de teste, busca se assemelhar com condições reais de teste, assim com o modelo gerado se torna possível realizar a etapa de reconhecimento de padrões com as amostras de teste separadas.

3.5 VALIDAÇÃO DE RESULTADOS

A etapa de validação dos resultados tem como objetivo avaliar o desempenho da metodologia pelo nível de satisfatoriedade, e também para discriminar seus pontos positivos e negativos visando alcançar melhorias em trabalhos futuros.

A validação é feita pelo uso de métricas que são comumente utilizadas em sistemas CAD e CADx, e aceitas pela sociedade para a análise de desempenho de sistemas baseados em processamento de imagens. São elas: Acurácia, Sensibilidade e Especificidade.

4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos pelos testes da metodologia proposta para classificação de massas em de imagens de mamografia nas classes maligno e benigno utilizando os descritores de forma apresentados neste trabalho.

Para a realização dos testes, a partir da etapa de extração de características, a base de amostras foi organizada em dois grupos para serem utilizados no classificador MVS: grupo de treino e grupo de teste, com proporções de 50% e 50%, 60% e 40%, 70% e 30% e 80% e 20%, respectivamente. Para cada proporção foram realizadas 5 repetições do teste de forma aleatória. Todos os valores das bases que estavam no conjunto $\mathbb{R}+$ (conjunto de números reais não-negativos) foram normalizados entre -1 a 1 para ajudar o classificador a convergir com maior facilidade na etapa de treinamento. Do conjunto total de resultados obtidos foram a analisadas as médias das 5 repetições de cada proporção.

Como mencionado na Seção 3.1 foram selecionadas 118 ROIs previamente segmentadas, sendo 61 contendo regiões de nódulos benignos e 57 contendo regiões de nódulos malignos. Para cada nódulo foram feitos 3 experimentos em que diferentes combinações dos descritores *Average Slope* e *Relief Index* eram utilizadas, como apresentado nas Tabelas 1, 2 e 3.

Os resultados obtidos são apresentados em termos de média das 5 repetições de cada proporção. São apresentadas as médias de acurácia, de sensibilidade e de especificidade, com seus respectivos desvios padrão, destacando-se o melhor resultado em cada experimento.

A Tabela 1 apresenta os resultados do experimento 1 onde utilizou-se a combinação dos descritores *Average Slope* e *Relief Index* segundo M'kirera e Ungar (2003). O melhor resultado foi obtido na proporção 50/50 com média de acurácia de 80,00%, média de sensibilidade de 75,93% e média de especificidade de 84,28%.

Na Tabela 2 são apresentados os resultados obtidos pelo experimento 2 onde foi utilizada a combinação dos descritores *Average Slope* e *Relief Index* segundo Boyer (2008). O

Tabela 1 – Experimento 1: Resultado da classificação com MVS dos descritores *Relief Index* segundo M'kirera e Ungar (2003) e *Average Slope*.

Proporção	Acurácia (%)		Sensibilidade (%)		Especificidade (%)	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
50/50	**80,00**	**0,05**	**75,93**	**0,14**	**84,28**	**0,07**
60/40	76,96	0,10	78,07	0,16	77,55	0,15
70/30	74,12	0,07	75,93	0,11	72,46	0,17
80/20	76,52	0,12	82,54	0,23	70,54	0,15

Tabela 2 – Experimento 2: Resultado da classificação com MVS dos descritores *Relief Index* segundo Boyer (2008) e *Average Slope*.

Proporção	Acurácia (%)		Sensibilidade (%)		Especificidade (%)	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
50/50	80,70	0,02	84,15	0,05	77,58	0,05
60/40	79,13	0,06	80,94	0,08	77,74	0,08
70/30	**81,18**	**0,05**	**83,73**	**0,12**	**78,87**	**0,05**
80/20	79,13	0,08	86,51	0,15	73,21	0,09

melhor resultado foi obtido na proporção 70/30 com média de acurácia de 81,18%, média de sensibilidade de 83,73% e média de especificidade de 78,87%.

Os resultados obtidos pelo experimento 3 em que utilizou-se a combinação dos descritores *Average Slope* e *Relief Index*, em suas duas versões, são apresentados na Tabela 3. O melhor resultado foi obtido na proporção 70/30 com média de acurácia de 80,59%, média de sensibilidade de 81,39% e média de especificidade de 79,29%.

Tabela 3 – Experimento 3: Resultado da classificação com MVS dos descritores *Relief Index* segundo M'kirera e Ungar (2003) e Boyer (2008) e *Average Slope*.

Proporção	Acurácia (%)		Sensibilidade (%)		Especificidade (%)	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
50/50	79,65	0,07	80,79	0,08	80,10	0,17
60/40	75,22	0,09	80,14	0,14	72,67	0,21
70/30	**80,59**	**0,08**	**81,39**	**0,10**	**79,29**	**0,09**
80/20	79,13	0,13	78,93	0,10	78,81	0,18

Na Tabela 4 são apresentadas as médias gerais de cada experimento, ou seja, as médias em todas as proporções de treino e teste para cada um dos experimentos, com seus respectivos desvios padrão.

Tabela 4 – Resultados geral dos experimentos (média) com seus respectivos desvios padrão.

Experimento	Acurácia (%)		Sensibilidade (%)		Especificidade (%)	
	Média	Desvio Padrão	Média	Desvio Padrão	Média	Desvio Padrão
1	76,90	2,09	78,12	2,70	76,21	5,32
2	**80,03**	**0,92**	**80,31**	**1,98**	**77,72**	**2,16**
3	78,65	2,05	83,83	0,91	76,85	2,95

Analisando os resultados gerados pelo experimentos mostrados nas Tabelas 1 e 2, foi possível observar que com exceção do valor de especificidade para proporção 50/50, o

experimento em que utilizou-se o *Average Slope* e o *Relief Index* segundo Boyer (2008) (Tabela 2), em média, obteve melhor desempenho. Acredita-se que a superioridade dessa abordagem seja dada pelo fato de a função logarítmica apresentar os resultados em um domínio de valores que é melhor aproveitado pelo classificador MVS. Os desvios-padrão das médias de acurácias foram, nessa ordem, de 2,09% e 0,92% para o experimento da Tabela 1 e da Tabela 2. Observou-se também que a diferença entre as médias de sensibilidade e especificidade dos 2 experimentos foi consideravelmente grande. Uma possível explicação para esta diferença seria a não realização de técnicas de melhoramento na etapa de pré-processamento, que tem por objetivo realçar os aspectos mais importantes em uma imagem. Estes aspectos por sua vez, podem ser determinantes para discriminar as classes benigno e maligno.

A partir da Tabela 4 é possível notar que os resultados apresentados obtiveram valores desvio padrão pequenos, o que demonstra pouca variabilidade entre as diferentes execuções dos experimentos, indicando assim uma certo nível consistência dos resultados nessa metodologia. É possível observar que experimento 2 além de ter obtido os melhores resultados em média, obteve também os menores valores de desvio padrão entre os demais experimentos, com exceção do desvio padrão da sensibilidade que foi o segundo menor. Porém, em um sistema CADx é importante que o equilíbrio entre sensibilidade e especificidade seja mantido, e apesar dos melhores resultados em média, o experimento 2 não apresenta valores de sensibilidade e especificidade bem balanceados (80,31% e 77,72% respectivamente). Pelo fato de os nódulos malignos, na maioria dos casos, apresentarem uma geometria mais complexa (formato espiculado), acredita-se que a sensibilidade, que avalia a taxa de acertos dos casos doentes, obtém valores superiores à especificidade porque os descritores de forma são capazes de descrever melhor os objetos que possuem geometria mais diferenciadas.

5 CONCLUSÕES E TRABALHOS FUTUROS

Os elevados índices de morte e registros de ocorrências de câncer de mama no Brasil e no mundo demonstram a importância dos trabalhos de pesquisa que buscam produzir recursos para um diagnóstico precoce dessa doença.

O uso de ferramentas computacionais para o auxílio ao diagnóstico e detecção tem evoluído em técnicas, áreas de abrangência e também em interesse por parte da comunidade científica. Como consequência dessa evolução, a utilização dessas ferramentas torna-se cada vez mais frequente para auxiliar especialistas na detecção e diagnóstico de doenças, tornando-as cada vez mais aplicáveis e mais presentes em seu cotidiano.

Neste trabalho buscou-se investigar a utilização dos descritores de forma *Relief Index* e *Average Slope*, que originalmente foram empregados no contexto de ecologia dentária de mamíferos para identificar padrões alimentares. Nesta metodologia, estes descritores foram aplicados na análise da forma de massas em imagens de mamografia. Fazendo uso do classificador MVS, objetivou-se apresentar métodos de suporte ao diagnóstico de câncer de mama.

Os melhores resultados em termos de média foram: 83,73% de sensibilidade, 78,87% de especificidade e 81,18% de acurácia no teste em que utilizou-se a combinação dos descritores *Average Slope* e *Relief Index* segundo Boyer (2008) na proporção 70/30.

Apesar de não serem considerados satisfatórios para utilização em sistemas CADx, os resultados obtidos podem ser considerados promissores pelo fato de que estes ainda podem ser melhorados pelo uso de mais descritores e também pelo uso de técnicas de pré-processamento de imagem.

Os resultados obtidos mostraram que a análise da forma como parâmetro para a diferenciação entre os padrões malignos e benignos de massa em imagens de mamografias possui grande relevância, dado que as massas em mamografia, na maioria dos casos, possuem características geométricas distintas.

Entre as limitações desta metodologia destaca-se o tamanho da amostra sobre a qual os experimentos foram realizados, em que foram utilizadas apenas 118 ROIs. Também destaca-se que a escolha da amostra foi feita manualmente levando em consideração a análise visual das ROIs. A escolha da amostra foi feita desta maneira pelo fato de que um dos objetivos específicos deste trabalho era testar a aplicabilidade dos descritores de forma aqui apresentados, no contexto de processamento de imagens e reconhecimento de padrões para extração de características de massas, justificando a pequena quantidade ROIs selecionadas. No entanto, para que os resultados

sejam melhor validados, sugere-se que em trabalhos futuros os testes com esses descritores sejam realizados sobre diferentes amostras, com um maior número de regiões de interesse e escolhidas de forma aleatória.

Propõe-se também que em trabalhos futuros um número de maior de descritores de forma seja utilizado e que estes sejam aplicados juntamente com descritores de textura, visando gerar um conjunto de características com melhor capacidade de discriminar classes de massas em imagens de mamografia.

REFERÊNCIAS

ACS. **Learn about Breast Cancer**. 2014. Último Acesso: 05/04/2016. Disponível em: <http://www.cancer.org/cancer/index>.

ALMEIDA, C. W. D. de; POEL, J. van der; VIDAL, L.; BATISTA, H. L. E. d. A. Análise de formas baseada no método da curvature scale space para tumores de câncer de mama. **Proceedings of the Brazilian Computer Society–SBC–WIM 2005**, p. 1–4, 2005.

BLAND, M. **An introduction to medical statistics**. New York: Oxford Medical Publications, 2000.

BOYER, D. M. Relief index of second mandibular molars is a correlate of diet among prosimian primates and other euarchontan mammals. **Journal of Human Evolution**, Elsevier, v. 55, n. 6, p. 1118–1137, 2008.

BRAZ, G. **Classificaçao de Regioes de Mamografias em Massa e Nao Massa usando Estatıstica Espacial e Máquina de Vetores de Suporte**. Dissertação (Mestrado) — Universidade Federal do Maranhão, 2008.

BURROUGH, P. A. Principles of geographical information systems for land resources assessment. Taylor & Francis, 1986.

CHANG, C.-C.; LIN, C.-J. Libsvm: A library for support vector machines. **ACM Trans. Intell. Syst. Technol.**, ACM, New York, NY, USA, v. 2, n. 3, p. 27:1–27:27, maio 2011. ISSN 2157-6904. Disponível em: <http://doi.acm.org/10.1145/1961189.1961199>.

CHAVES, A. C. F. **EXTRAÇÃO DE REGRAS FUZZY PARA MÁQUINAS DE VETOR DE SUPORTE (SVM) PARA CLASSIFICAÇÃO EM MÚLTIPLAS CLASSES**. Tese (Doutorado) — PUC-RIO, 2006.

DELFOUR, M. C.; ZOLESIO, J. P. **Shapes and geometries : analysis, differential calculus, and optimization**. Philadelphia: SIAM, Society for Industrial and Applied Mathematics, 2001. (Advances in design and control). ISBN 0-89871-489-3. Disponível em: <http://opac.inria.fr/record=b1098580>.

DENGLER, J.; BEHRENS, S.; DESAGA, J. F. Segmentation of microcalcifications in mammograms. **Medical Imaging, IEEE Transactions on**, IEEE, v. 12, n. 4, p. 634–642, 1993.

EDELSBRUNNER, H.; MÜCKE, E. P. Three-dimensional alpha shapes. ACM Transactions On Graphics, 1994.

ERPEN, L. R. C. **Reconhecimento de padrões em imagens por descritores de forma**. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul. Instituto de Informática, 2004.

EVANS, A. R. Shape descriptors as ecometrics in dental ecology. **Hystrix, The Italian Journal of Mammalogy**, v. 24, n. 1, p. 133–140, 2013.

FANDOS-MORERA, A.; PRATS-ESTEVE, M.; TURA-SOTERAS, J.; TRAVERIA-CROS, A. Breast tumors: composition of microcalcifications. **Radiology**, v. 169, n. 2, p. 325–327, 1988.

FLORIANI, L. D.; SPAGNUOLO, M. (Ed.). **Shape Analysis and Structuring**. Springer, 2008. ISBN 978-3-540-33264-0. Disponível em: <http://dx.doi.org/10.1007/978-3-540-33265-7>.

GIGER, M.; MACMAHON, H. Image processing and computer-aided diagnosis. **Radiologic Clinics of North America**, v. 34, n. 3, p. 565–596, 1996.

GIS AG MAPS. **Understanding Neighborhood Slope Angle**. 2011. Último Acesso: 07/03/2016. Disponível em: <http://www.gisagmaps.com/neighborhood-slope/>.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2002.

HAYKIN, S. S.; ENGEL, P. **Redes neurais: Princípios e Prática**. Porto Alegre: Bookman, 2001.

HEATH, M.; BOWYER, K.; KOPANS, D.; JR, P. K.; MOORE, R.; CHANG, K.; MUNISH-KUMARAN, S. Current status of the digital database for screening mammography. In: **Digital mammography**. Dordrecht: Springer, 1998. p. 457–460.

INCA. **Consenso para o Controle do Câncer de Mama**. 2004. Último Acesso: 30/03/2016. Disponível em: <http://www.inca.gov.br/publicacoes/Consensointegra.pdf>.

INCA. **Estimativa Incidência Câncer 2016**. 2015. Último Acesso: 18/02/2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2015/estimativa_incidencia_cancer_2016>.

INCA. **O que é o câncer**. 2016. Último Acesso: 17/02/2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee>.

LOONEY, C. Pattern recognition using neural networks: Theory and algorithms for engineers and scientists. Oxford University Press, Inc. New York, NY, USA, 1997.

MARQUES, O.; VIEIRA, H. **Processamento digital de imagens**. Rio de Janeiro: Brasport, 1999.

MENECHELLI, R. C.; RIBEIRO, P. B.; SCHIABEL, H. Desenvolvimento de um software classificador da forma de nódulos mamográficos segmentados utilizando a rede neural artificial multi-layer perceptron (mlp). In: **VI Workshop de Visão Computacional.** Presidente Prudente, Brasil: [s.n.], 2010.

M'KIRERA, F.; UNGAR, P. S. Occlusal relief changes with molar wear in pan troglodytes troglodytes and gorilla gorilla gorilla. **American Journal of Primatology**, Wiley Online Library, v. 60, n. 2, p. 31–41, 2003.

MOKHTARIAN, F.; MACKWORTH, A. K. A theory of multiscale, curvature-based shape representation for planar curves. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, IEEE, n. 8, p. 789–805, 1992.

NASCIMENTO, L. B. **Classificação de Nódulos Pulmonares em Maligno e Benigno utilizando os Índices de Diversidade de Shannon e de Simpson**. Dissertação (Mestrado) — Universidade Federal do Maranhão, 2012.

PECKHAM, R. J.; JORDAN, G. **Digital terrain modelling: development and applications in a policy support environment**. Budapest: Springer Science & Business Media, 2007.

PEDRINI, H.; SCHWARTZ, W. R. **Análise de Imagens Digitais: Princípios, Algorítmos e Aplicações.** São Paulo: Thomson Learning, 2008.

SOUZA, L. B. L. de; GULIATO, D. Extratores de características aplicados à classificação de tumores de mama. 2009.

TAYLOR, P.; CHAMPNESS, J.; GIVEN-WILSON, R.; POTTS, H.; JOHNSTON, K. An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms. **The British journal of radiology**, v. 77, n. 913, p. 21, 2004.

UNGAR, P.; WILLIAMSON, M. Exploring the effects of tooth wear on functional morphology: a preliminary study using dental topographic analysis. **Palaeontologia electronica**, v. 3, n. 1, p. 1–18, 2000.

VAPNIK, V. N. **Statistical Learning Theory.** New York: Wiley, 1998.

VATH, B. **Uniao de bolas, eixo medial e deformaçoes no espaço tridimensional**. Dissertação (Mestrado) — PUC–Rio, 2007.