Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Pós-Graduação em Engenharia de Eletricidade

Fernando Soares Sérvulo de Oliveira

Classificação de Tecidos da Mama em Massa e Não-Massa usando Índice de Diversidade Taxonômico e Máquina de Vetores de Suporte

SÃO LUÍS
2013

FERNANDO SOARES SÉRVULO DE OLIVEIRA

CLASSIFICAÇÃO DE TECIDOS DA MAMA EM MASSA E NÃO-MASSA USANDO ÍNDICE DE DIVERSIDADE TAXONÔMICO E MÁQUINA DE VETORES DE SUPORTE

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão como parte dos requisitos necessários para obtenção do título de Mestre em Engenharia de Eletricidade na área de concentração Ciência da Computação.

Orientador: Prof. Dr. Anselmo Cardoso de Paiva.
Co-orientador: Prof. Dr. Aristófanes Corrêa Silva.

SÃO LUÍS-MA
2013

Oliveira, Fernando Soares Sérvulo de.

Classificação de tecidos da mama em massa e não-massa usando índice de diversidade taxonômico e máquina de vetores de suporte / Fernando Soares Sérvulo de Oliveira – São Luís, 2013.

70 f.

Impresso por computador (fotocópia).
Orientador: Anselmo Cardoso de Paiva.
Co-Orientador: Aristófanes Corrêa Silva.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2013.

1. Classificação de tecidos de mama – Mamografia. 2. Máquina de vetores de suporte. I. Título.

CLASSIFICAÇÃO DE TECIDOS DA MAMA EM MASSA E NÃO MASSA USANDO ÍNDICE DE DIVERSIDADE TAXONÔMICO E MÁQUINA DE VETORES DE SUPORTE

Fernando Soares Sérvulo de Oliveira

Dissertação aprovada em 20 de fevereiro de 2013.

Prof. Anselmo Cardoso de Paiva, Dr.
(Orientador)

Prof. Aristófanes Corrêa Silva, Dr.
(Co- Orientador)

Profa. Alcione Miranda dos Santos, Dra.
(Membro da Banca Examinadora)

Prof. Leandro Augusto Frata Fernandes, Ph.D.
(Membro da Banca Examinadora)

À DEUS, à minha esposa e aos meus pais.

AGRADECIMENTOS

À DEUS, por me abençoar com saúde, força e sabedoria por mais uma conquista na minha vida. Obrigado Senhor!

À minha esposa, Jacira, pelo carinho, compreensão, dedicação, orações e seu amor.

Aos meus pais, Fátima e Elesbão, e minha avó, Maria, pela atenção, carinho e educação.

Aos meus orientadores, Dr. Anselmo Cardoso de Paiva e Dr. Aristófanes Corrêa Silva, pela confiança, sabedoria, dedicação, paciência e conselhos.

Aos professores, Msc. Geraldo Braz e Msc. Simara Rocha, pela ajuda e atenção nos testes deste trabalho.

Aos meus colegas do LABPAI, em especial Darlan, Oseas, Wener, Joberth, Thiago e também de outros colegas que fizeram parte desse laboratório, Leonardo Barros, Leonardo Dorneles e Péterson, obrigado pela ajuda e compartilhamento de ideias.

Ao grupo do VisualLab-UFF da professora Drª. Aura Conci e seus alunos, Lincoln, Rafael, Tiago, Roger, João Paulo, João Gabriel, Edgar e Giomar, pela experiência adquirida no período de intercâmbio.

Ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão e a todos os professores do programa pelo aprendizado transmitido.

À CAPES, pelo apoio financeiro durante o período do mestrado.

Aos meus amigos Raimundo Neto e Mauro, o primeiro pelo incentivo, conselhos e ajuda para entrar no mestrado. E o segundo, pelo conhecimento, convivência boa, risadas e dividir o mesmo apartamento.

Fica o meu MUITO OBRIGADO A TODOS!

“A nossa maior glória não reside no fato de
nunca cairmos, mas sim em levantarmo-nos
sempre depois de cada queda.”

Confúcio

RESUMO

O câncer de mama é o segundo tipo de câncer mais frequente no mundo e de difícil diagnóstico. Distintos Sistemas de Detecção e Diagnóstico Auxiliados por Computador (Computer Aided Detection/Diagnosis) têm sido utilizados para auxiliar especialistas da área da saúde com a indicação de áreas suspeitas de difícil percepção ao olho humano, assim ajudando na detecção e diagnóstico de câncer. Este trabalho propõe uma metodologia de discriminação e classificação de regiões extraídas da mama em *massa* e *não-massa*. O banco de imagens Digital Database for Screening Mammography (DDSM) é usado neste trabalho para aquisição das mamografias, onde são extraído as regiões de *massa* e *não-massa*. Na descrição da textura da região de interesse são utilizados os Índices de Diversidade Taxonômica (Δ) e Distinção Taxonômica (Δ^*), provenientes da ecologia. O cálculo destes índices é baseado nas árvores filogenéticas, sendo aplicados neste trabalho na descrição de padrões em regiões das imagens da mama com duas abordagens de regiões delimitadoras para análise da textura: círculo com anéis e máscaras internas com externas. Para classificação das regiões em *massa* e *não-massa* é utilizado o classificador Máquina de Vetores de Suporte (MVS). A metodologia apresenta resultados promissores para a classificação de *massas* e *não-massas*, alcançando uma acurácia média de 99,67%.

Palavras-chave: Mamografia, Árvores Filogenéticas, Índice de Diversidade Taxonômica, Índice de Distinção Taxonômica, Máquina de Vetores de Suporte.

ABSTRACT

Breast cancer is the second most common type of cancer in the world and difficult to diagnose. Distinguished Systems Aided Detection and Diagnosis Computer have been used to assist experts in the health field with an indication of suspicious areas of difficult perception to the human eye, thus aiding in the detection and diagnosis of cancer. This dissertation proposes a methodology for discrimination and classification of regions extracted from the breast *mass* and *non-mass*. The Digital Database for Screening Mammography (DDSM) is used in this work for the acquisition of mammograms, which are extracted from the regions of *mass* and *non-mass*. The Taxonomic Diversity Index (Δ) and the Taxonomic Distinctness (Δ^*) are used to describe the texture of the regions of interest, originally applied in ecology. The calculation of those indices is based on phylogenetic trees, which applied in this work to describe patterns in regions of the images of the breast with two regions bounding approaches to texture analysis: circle with rings and internal with external masks. Suggested in this work to be applied in the description of patterns of regions in breast imaging approaches circle with rings and masks as internal and external boundaries regions for texture analysis. Support Vector Machine (SVM) is used to classify the regions in *mass* or *non-mass*. The proposed methodology provides successful results for the classification of *masses* and *non-mass*, reaching an average accuracy of 99.67%.

Keywords: Mammography, Phylogenetic trees, Taxonomic Diversity Index, Taxonomic Distinctness Index, Support Vector Machine.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	**Acurácia**
ACS	**American Cancer Society**
A1	**Árvore 1**
A2	**Árvore 2**
A3	**Árvore 3**
CA	**Círculos e Anéis**
CAD	**Computer Aided Detection**
CADx	**Computer Aided Diagnosis**
CC	**Crânio Caudal**
DDSM	**Digital Database for Screening Mammography**
DP	**Desvio Padrão**
DS	**Índice de Distinção Taxonômica**
DV	**Índice de Diversidade Taxonômica**
DWT	**Discrete Wavelet Transformation**
ECM	**Exame Clínico das Mamas**
ES	**Especificidade**
FLD	**Fisher Linear Discriminant**
FN	**Falso Negativo**
FP	**Falso Positivo**
ICA	**Independent Component Analysis**
IE	**Máscara Interna e Externa**
INCA	**Instituto Nacional do Câncer**
J1	**Junção do Índice de Diversidade e Distinção Taxonômica**
LBP	**Local Binary Pattern**
MAC	**Média das Acurácias**
MIAS	**Mammograms Image Analysis Society**
MVS	**Máquina de Vetores de Suporte**
MLO	**Médio Lateral Oblíqua**

M3	**Filtro da Média 3x3**
M5	**Filtro da Média 5x5**
OMS	**Organização Mundial da Saúde**
PCA	**Principal Components Analysis**
Q5	**Níveis de Quantização**
RMI	**Ressonância Magnética por Imagem**
US	**Ultrassonografia**
Δ^*	**Valor do Índice de Distinção Taxonômica**
Δ	**Valor do Índice de Diversidade Taxonômica**

1 INTRODUÇÃO

Um grande problema de saúde pública que ocorre há muitos anos é o câncer, sendo apresentada em mais de cem tipos diferentes, entre homens e mulheres, onde a característica comum a todas é o crescimento desordenado de células que invadem tecidos e órgãos (INCA, 2013a). Cânceres não tratados podem causar a morte (ACS, 2013a). A Organização Mundial da Saúde (OMS) estimou que, no ano 2030, podem-se esperar 27 milhões de casos incidentes de câncer, 17 milhões de mortes por câncer e 75 milhões de pessoas vivas, anualmente, com câncer. O maior efeito desse aumento vai incidir em países de baixa e média renda (INCA, 2013b).

No Brasil, as estimativas para o ano de 2012, válidas também para o ano de 2013, apontam a ocorrência de aproximadamente 518.510 casos novos de câncer, o que reafirma a magnitude do problema do câncer no país. Removendo-se das previsões os casos de câncer de pele não-melanoma, estima-se um total de 385 mil casos novos. Os tipos mais incidentes para o sexo masculino serão os cânceres de pele não-melanoma, próstata, pulmão, cólon, reto e estômago. Já os cânceres de pele não-melanoma, mama, colo do útero, cólon, reto e glândula tireoide serão os mais incidentes para o sexo feminino (INCA, 2013b).

O câncer que possui a maior taxa de mortalidade entre as mulheres e o segundo mais frequente mundialmente é o de mama. Em 2012 foram esperados para o Brasil 52.680 casos novos de câncer da mama, com um risco estimado de 52 casos a cada 100 mil mulheres. Sendo que as regiões com maior incidência entre as mulheres são Sudeste (69/100 mil), Sul (65/100 mil), Centro-Oeste (48/100 mil) e Nordeste (32/100 mil). Na região Norte, este é o segundo tipo de tumor mais incidente (19/100 mil). No Maranhão, o número de casos novos estimado para 2012 foi de 460 para cada 100 mil mulheres. Considerando apenas a capital, São Luís, foram registrados 190 novos casos para cada 100 mil mulheres (INCA, 2013c).

A taxa bruta de mortalidade por câncer de mama no Brasil apresentou aumento de 34%, passando de 9,74 em 1998 para 13,05 mortes por 100 mil mulheres em 2010 (INCA, 2013d). Esse aumento se deve a alguns fatores como: o envelhecimento da população, a mudança do perfil reprodutivo, a exposição a poluentes, o sedentarismo, a obesidade, dentre outros.

A melhor forma de prevenção conhecida é o diagnóstico precoce, que diminui a mortalidade e aumenta a eficácia do tratamento. Uma das formas de prevenção é o exame de mamografia. Este exame é analisado por especialistas (radiologistas) que são médicos capazes de diagnosticar doenças a partir de imagens. Essa etapa é sensível, pois pode haver diferentes interpretações entre os especialistas em relação a um mesmo exame. Também é uma tarefa repetitiva que requer um nível de atenção muito grande sobre os mínimos detalhes.

A sensibilidade da mamografia varia entre 46% e 88% e é dependente dos seguintes fatores: tamanho e localização da lesão, densidade do tecido mamário, idade da paciente, qualidade do exame e habilidade de interpretação do radiologista (INCA, 2002). A única maneira de saber com certeza se uma região for câncer é fazer uma biópsia, isto significa uma amostra de células ou tecido retirado da área e observadas através de um microscópio (ACS, 2013b).

Por essas razões, nas últimas décadas tem surgido um grande interesse no desenvolvimento e uso de técnicas de processamento de imagens digitais de mamografias com o objetivo principal de aumentar a precisão do diagnóstico e servindo como uma segunda opinião para o especialista. Essas técnicas em conjunto têm sido utilizadas para desenvolver sistemas CAD/CADx (Computer-Aided Detection/Computer-Aided Diagnostic). Os sistemas CAD são sistemas que auxiliam na detecção de anormalidades, mas não realizam qualquer tipo de diagnóstico sobre as mesmas. Os sistemas CADx, por sua vez, quando aplicados em imagens de mamografia, classificam as estruturas detectadas em duas classes: *massa* e *não-massa* (CARVALHO, 2012).

Neste trabalho, é apresentado uma metodologia CADx para a classificação de tecidos da mama nas classes *massa* e *não-massa*. As *massas* são todas as regiões da mamografia que correspondem a uma neoplasia maligna ou benigna, e *não-massas* são todas as regiões que não são neoplasias. Neste trabalho é proposta a utilização do índice de diversidade taxonômica (Δ) e do índice de distinção taxonômica (Δ^*) em árvores filogenéticas para extração de características de textura de regiões da mamografia. Estes índices são usados juntos com técnicas de processamento de imagens digitais e reconhecimento de padrões, em especial Máquina de Vetores de Suporte, para a classificação de regiões pré-segmentadas da imagem mamográfica em: *massa* e *não-massa*.

O objetivo deste trabalho é avaliar o desempenho dos índices taxonômicos a partir da geração de modelos de árvores filogenéticas como método de extração de características de textura de regiões de interesse em imagens mamográficas e depois discriminar as regiões em *massa* e *não-massa*. Com isso, deseja-se contribuir na extração de textura em imagens médicas para a área de sistemas e metodologias CAD/CADx.

1.1 Trabalhos Relacionados

Diversos trabalhos de pesquisa têm sido desenvolvidos pela grande necessidade de metodologias eficientes para auxiliar na detecção e diagnóstico do câncer mamário. Os sistemas ou metodologias CADx tem alta dependência do método de extração de características e do classificador. A literatura disponível traz trabalhos relacionados que tratam do mesmo objetivo abordado pelo método proposto, ou seja, desenvolver métodos computacionais que possam auxiliar o especialista na tarefa de classificação de lesões em imagens mamográficas.

O sistema proposto por Berbar *et al.* (2012) apresenta a classificação de regiões de interesse de imagens mamográficas em *normal* e *anormal*. As imagens utilizadas foram da base *Digital Database for Screening Mammography* (DDSM) (HEATH *et al.*, 2001). As regiões de interesse foram dimensionadas com 256 x 256 pixels. Para a extração de características de textura, foram utilizadas 6 características estatísticas (média, desvio padrão, suavidade, assimetria, energia e entropia) e *Local Binary Pattern* (LBP). A classificação foi realizada através da Máquina de Vetores de Suporte (MVS) e *K-Nearest Neighbors* (K-NN). O trabalho apresenta resultados com acurácia de 98,43%. Usando estatística e K-NN obteve 95,10% de acurácia, e 97,06% usando LBP e K-NN, com a junção de estatística e LBP com K-NN foi conseguido 97,25% de acurácia. A acurácia máxima atingida foi de 98,63% para o uso de estatística e LBP com MVS.

Carvalho *et al.* (2012) apresenta um sistema para classificação de tecidos da mama em *massa* e *não-massa* em imagens mamográficas. As imagens foram adquiridas da base DDSM. As características de textura foram extraídas utilizando o índice de diversidade de McIntosh nas regiões de interesse da mama. O índice foi calculado a partir de três abordagens: Histograma, *Gray Level Co-Occurrence Matrix* (GLCM) e *Gray Level Run Length Matrix*

(GLRLM). A acurácia máxima alcançada foi de 99,75%, sensibilidade de 99,47% e 100% de especificidade para abordagem GLRLM.

A metodologia apresentada por Costa et al.(2011) descreve a classificação tecidos da mama em *massa* e *não-massa* a partir de imagens mamográficas. As imagens adquiridas foram da base DDSM. Na etapa de pré-processamento, a equalização do histograma foi aplicado nas regiões de interesse. Na etapa seguinte, as características foram extraídas com *Principal Components Analysis* (PCA), *Gabor wavelet* e o modelo codificação eficiente baseado em *Independent Component Analysis* (ICA). Para classificação foi utilizado o MVS. A abordagem modelo codificação eficiente teve melhor resultado com acurácia de 90,07%.

O trabalho apresentado por Jasmine *et al.* (2011) descreve um sistema de classificação automática de câncer de mama em imagens mamográficas. As imagens usadas foram da base *Mammograms Image Analysis Society* (MIAS) (SUCKLING *et al*., 1994). As características foram extraídas do coeficiente de contourlet usando Transformada de Contourlet Não Subamostrada. O trabalho apresenta média da acurácia máxima de 98,61% para classificar as regiões em *normal* e *anormal*, e 88,05% para *maligno* e *benigno* com a utilização do MVS.

Junior *et al.* (2009) desenvolveu uma metodologia para classificar tecidos da mama em *normal* e *anormal* através de imagens mamográficas. As imagens utilizadas foram da base DDSM. Na etapa de pré-processamento, a equalização do histograma global foi aplicado nas regiões de interesse. Depois as características de textura foram extraídas a partir de seis resoluções diferentes para quantização (2^8, 2^7, 2^6, 2^5, 2^4, 2^3) com aplicação do Índice de Moran e Coeficiente de Geary em cada resolução. A classificação foi realizada através do MVS. Neste trabalho, foram encontradas acurácia máxima de 99,39%, sensibilidade de 100% e especificidade de 98,94% para *massa* e *não-massa*. No passo seguinte foi encontrada acurácia máxima de 88,31%, sensibilidade de 84,78% e especificidade de 93,55% para as massas nas classes *malignas* e *benignas*.

É apresentado por Meenalosini e Janet (2012) uma metodologia para detecção de massas em imagens mamográficas. As imagens utilizadas foram da base MIAS e DDSM. A etapa após a aquisição das imagens desse trabalho é o pré-processamento, onde foram removidos ruídos de digitalização e componentes de alta frequência das imagens de mamografia com o filtro da média, depois foram removidas as etiquetas do filme e marcas de raio-x e no fim dessa etapa de pré-processamento, as intensidade das mamografias foram

normalizadas em faixa fixa. Na segunda etapa, foi segmentada toda a região da massa na mamografia aplicando primeiro a equalização do histograma adaptativo para aumentar o contraste de cada pixel em relação à sua vizinhança local que produz um melhor contraste para todos os níveis da imagem, em seguida foram produzidos os pixels de alarme que são determinados pela limiarização do histograma na imagem de contraste melhorado, que depois esses pixels alarme podem ser a semente de crescimento da região. A etapa seguinte de extração de característica de textura foi utilizada a *Spatial Gray Level Dependence* (SGLD), onde as medidas usadas foram contraste, energia, homogeneidade e correlação. A última etapa foi a classificação das massas em *normal* e *anormal* utilizando MVS, que obteve 95,2% de sensibilidade e 94,4% de especificidade.

Nithya e Santhi (2011) descreve uma metodologia para classificar tecidos da mama em *normal* e *anormal* através de imagens mamográficas. As imagens utilizadas foram da base DDSM. Para a extração de características de textura foi utilizada a GLCM com as quais foram calculadas as medidas de *Haralick* (correlação, energia, entropia, homogeneidade e soma do quadrado da variância). A classificação foi realizada com uma Rede Neural de três camadas: na primeira camada (entrada) com 5 unidades, na segunda camada (oculta) atingindo uma acurácia máxima de 96%.

Nithya e Santhi (2012) apresenta uma metodologia para classificação de tecidos da mama a partir de imagens mamográficas nas classes *normal* e *anormal*. As imagens utilizadas foram da base DDSM. As características de textura foram extraídas com medidas estatísticas (variância, desvio padrão, média, moda, alcance e suavidade) e média da intensidade dos pixels na horizontal e vertical (variância, desvio padrão, média, modo, alcance e suavidade). A classificação foi realizada com uma Rede Neural de três camadas: na camada de entrada com "n" unidades, uma da camada oculta e uma da camada de saída. Neste trabalho, a acurácia máxima é de 92% e 98% utilizando as características estatísticas e a média da intensidade dos pixels, respectivamente.

Nunes *et al.* (2010) mostra uma metodologia para detecção de massas em imagens mamográficas. As imagens utilizadas foram da base DDSM. As regiões de interesse foram extraídas através do algorítmo de *K-means* e a técnica *Template Matching*. Em seguida, na extração de características de textura foi utilizado o índice de diversidade de Simpson aplicado nas regiões de interesse da mama. O índice foi calculado a partir de três abordagens independentes: global, círculos e anéis. Na classificação em *massa* e *não-massa* foi utilizada a

MVS. Neste trabalho, foi alcançada uma acurácia de 83,94%, sensibilidade de 83,24% e especificidade de 84,14%.

Segundo Sarfraz *et al.* (2012) apresenta uma metodologia para detecção de massas em imagens mamográficas. As imagens utilizadas foram obtidas na base MIAS. As regiões de interesse foram recortadas com tamanho de 50 x 50 pixels. O algoritmo de PCA integrado com *Fisher Linear Discriminant* (FLD) foi usado para a redução de dimensionalidade e extração de características. A classificação foi realizada através do algoritmo K-NN. O melhor resultado alcançado foi 93,06% de acurácia para PCA-FLD, contra 78.47% para PCA e 47.92% para FLD.

Com base em Shanthi e Bhaskaran (2012) é proposto a detecção e classificação de câncer de mama em imagens mamográficas. As imagens usadas foram da base MIAS. As regiões de interesse foram extraídas com a segmentação automática através da técnica intuitiva de clusterização *Fuzzy C-Means*. Para a extração do vetor de características de textura, foi utilizado o histograma e a *Discrete Wavelet Transformation* (DWT), a partir da qual, foi calculada a GLCM (14 características) e Energia. A classificação foi realizada com *Self-Adaptive Resource Allocation Network* (SRAN). Neste trabalho, a acurácia máxima para classificar as massas em *normal*, *benigno* e *maligno* foi de 96,05%. Já a média das acurácias foi de 92,06%.

Sousa (2011) mostra uma metodologia para classificação de tecidos da mama em *massa* e *não-massa* por meio das imagens mamográficas. As imagens trabalhadas foram da base DDSM. Na extração de características de textura foi utilizado o índice de diversidade de Shannon aplicado nas regiões de interesse da mama. O índice foi calculado a partir de quatro abordagens independentes: global, circular, anelar e direcional. Para a classificação foi utilizada a MVS. A abordagem direcional teve melhor resultado com acurácia média de 99,71%, sensibilidade média de 99,71% e especificidade média de 99,78%.

Os trabalhos relacionados citados acima mostram que as metodologias baseadas em características de textura descrevem bem padrões em imagens, onde as medidas estatísticas são bastante utilizadas como mostra em nove desses trabalhos. E o processo de reconhecimento de padrões com a utilização da MVS apresenta resultados promissores para auxílio na detecção de câncer de mama, onde foi citada em quatro trabalhos a sua superioridade em relação a outros classificadores pelo valor da acurácia. Podemos analisar

que o processo de extração de medidas de textura através dos índices de diversidade (McIntosh, Simpson e Shannon citados) tem apresentado potencial na aplicação em imagens mamográficas. Verificamos que a classificação de regiões de interesse de mamografias em *massa* e *não-massa* é uma etapa decisiva nas metodologias de detecção de câncer de mama. Neste trabalho pretende-se apresentar melhorias na descrição de padrões de textura das imagens de mamografia com aplicação do índice de diversidade para a MVS discriminar bem as classes *massa* e *não-massa*.

1.2 Organização do Trabalho

O restante deste trabalho está constituído em mais quatro capítulos, descritos resumidamente a seguir.

No Capítulo 2, é exposta a fundamentação teórica necessária para a compreensão da metodologia proposta. Neste capítulo são descritos a extração de textura através dos Índices Taxonômicos (Diversidade e Distinção), a técnica de classificação denominada Máquina de Vetores de Suporte e métricas de validação dos resultados.

No Capítulo 3 é apresentada a metodologia utilizada para realizar a classificação das regiões de interesse extraídas de imagens mamográficas em *massa* e *não-massa*, utilizando a extração de características de textura com os índice taxonômicos e a classificação através da Máquina de Vetores de Suporte.

No Capítulo 4 são expostos e discutidos os resultados alcançados por meio da metodologia proposta. Por fim, o Capítulo 5 apresenta a conclusão deste trabalho, mostrando a eficiência dos métodos utilizados e oferecendo sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica necessária para o entendimento da metodologia proposta neste trabalho. Aborda-se o câncer de mama (Seção 2.1), o diagnóstico de câncer de mama (Seção 2.2), técnicas de processamento de imagens (Seção 2.3), realce de contraste (Seção 2.3.1), filtro da média (Seção 2.3.2), quantização uniforme (Seção 2.3.3), textura (Seção 2.3.4), índice de diversidade (Seção 2.4), diversidade filogenética (Seção 2.4.1), Índices Taxonômicos (Seção 2.4.2), classificação e reconhecimento de padrão (Seção 2.5), Máquina de Vetores de Suporte (Seção 2.5.1) e as métricas de validação de resultados (Seção 2.6).

2.1 Câncer de Mama

Segundo o INCA (2013a), o câncer de mama é um tumor maligno formado por um grupo de células cancerosas que podem crescer em tecidos vizinhos ou se espalhar (metástase) para áreas distantes do corpo. Esse tipo de câncer é cerca de 100 vezes menos comum entre homens do que entre as mulheres (ACS, 2013c).

Desenvolver câncer depende de alguns fatores de risco, como (ACS, 2013d):

- Envelhecimento: cerca de 1 em 8 cânceres de mama invasivos são encontrados em mulheres com menos de 45 anos, enquanto cerca de 2 de 3 cânceres de mama invasivos são encontrados em mulheres com 55 anos ou mais.
- Histórico familiar: o risco de câncer de mama é maior entre as mulheres cujos parentes próximos de sangue têm esta doença.
- Histórico pessoal: uma mulher com câncer de mama tem de 3 a 4 vezes de risco de desenvolver outro câncer em outra região da mama.
- Raça e etnia: as mulheres brancas são mais propensas a desenvolver o câncer de mama do que as mulheres afro-descendentes. No entanto, mulheres afro-descendentes estão mais sujeitas a morrer deste câncer.

- Tecido mamário denso: as mulheres que têm mais tecidos glandulares e menos tecidos adiposos, apresentam maior risco de desenvolver câncer de mama. Este fator também pode tornar mais difícil para os médicos detectar problemas em mamografias.
- Doença mamária benigna prévia: mulheres com certas condições benignas da mama pode ter um risco maior de ter câncer.
- Período menstrual: mulheres que tiveram mais ciclos menstruais, pelo fato de ter começado antes dos 12 anos de idade ou menopausa avançado (depois dos 55 anos) têm um risco ligeiramente maior de câncer de mama. Isso pode ser devido à exposição aos hormônios estrogênio e progesterona.
- Outros fatores: consumo de bebidas alcoólicas, ao fumo e à obesidade após a menopausa.

Ter um fator de risco, ou mesmo vários, não significa que a pessoa vai ter a doença. A maioria das mulheres que têm um ou mais fatores de risco de câncer de mama nunca desenvolvem a doença, enquanto muitas mulheres com câncer de mama não têm fatores de risco aparente. Mesmo quando uma mulher com fatores de risco desenvolve câncer de mama, é difícil saber o quanto esses fatores podem ter contribuído (ACS, 2013e).

Segundo ACS (2013f), a forma de detecção precoce do câncer aumenta as chances do mesmo poder ser diagnosticado em um estágio inicial e tratado com sucesso. Sendo que as mais eficazes são o exame clínico e a mamografia. O exame clínico das mamas (ECM), realizado por um médico ou enfermeira treinados, pode detectar tumor de até 1 (um) centímetro, se superficial. Deve ser feito uma vez por ano pelas mulheres entre 40 e 49 anos. A mamografia permite a identificação de lesões em fase inicial, muito pequenas (medindo milímetros). Deve ser realizada a cada dois anos por mulheres entre 50 e 69 anos, ou segundo recomendação médica (INCA, 2013d).

2.2 Diagnóstico de Câncer de Mama

A ultrassonografia (US) é um método diagnóstico amplamente difundido em nosso meio, utilizado como adjuvante à mamografia em casos de achado clínico ou mamográfico anormal, ou como primeira escolha em situações especiais, como na gravidez, lactação, mulheres jovens e durante os estados inflamatórios da mama. Na presença de lesões mamográficas, a US auxilia não só a caracterização e coleta de biópsias, mas também é capaz

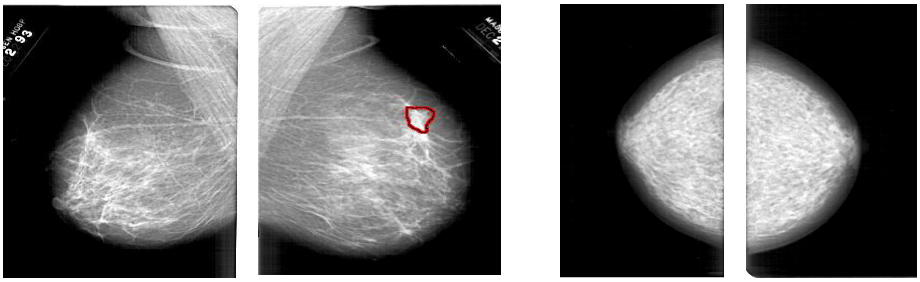
de identificar lesões adicionais em 14% das mulheres com mamas densas (NASTRI *et al.*, 2011).

Para algumas mulheres com alto risco de câncer de mama, o rastreio com Ressonância Magnética por Imagem (RMI) é recomendada junto com uma mamografia anual. RMI geralmente não é recomendada como ferramenta de triagem, por si só, porque embora seja um teste sensível, ainda pode perder alguns tipos de câncer que a mamografia detectaria. RMI pode também ser utilizada em outras situações, tais como para examinar melhor áreas suspeitas encontradas por uma mamografia. A RMI pode ser usada em mulheres que já foram diagnosticadas com câncer de mama para determinar melhor o tamanho real do câncer e para procurar por outros tipos de câncer na mama (ACS, 2013g).

O exame de mamografia é utilizado para detectar e avaliar anormalidades na mama, tanto em mulheres que não têm queixas ou sintomas quanto em mulheres que têm sintomas de doenças mamárias (ACS, 2013h). Segundo Kopans (2000), a radiografia da mama é o recurso mais usado para reduzir a mortalidade através do rastreamento de mulheres assintomáticas.

As mamografias de rotina (rastreamento) são mais aplicadas para procurar por câncer em mulheres que não apresentam nenhum sintoma. Sendo esta a melhor forma de detecção precoce antes do paciente ou médico possam notá-las ou apalpá-las, caso exista alguma alteração nas mamas. Já para mulheres que tenham notado na mama a presença de nódulos ou outros sintomas, a mamografia se torna de diagnóstico. Segundo INCA (2002), a sensibilidade desse exame varia entre 46% e 88%.

Normalmente são feitas duas radiografias, sendo uma para cada mama, em duas projeções: Médio Lateral Oblíqua (MLO) e uma Crânio-Caudal (CC). A Figura 2.1 mostra um exemplo dessas projeções com suspeitas de nódulos definido pelo médico.



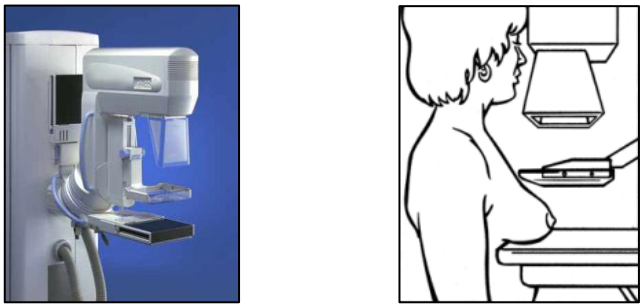
(a) MLO Esquerda (b) MLO Direita (c) CC Esquerda (d) CC Esquerda

Figura 2.1 (a) e (b) Mamografia com incidência Médio-Lateral Oblíquo em ambas as mamas, sendo que em (b) mostra delimitado em vermelho uma massa maligna (DDSM, 2013a); (c) e (d) Mamografia com

incidência Crânio-Caudal em ambas as mamas e que não apresentam região suspeita de anormalidade (DDSM, 2013b).

As imagens geradas do tecido mamário são em tons de cinza em uma folha grande de filme ou como uma imagem digital de computador, que são interpretadas por um radiologista (ACS, 2013h).

O mamógrafo é o equipamento utilizado para a realização do exame de mamografia (Figura 2.2a). Esse aparelho comprime a mama com o objetivo de fornecer as melhores imagens e, consequentemente, melhorar a capacidade de detecção ou diagnóstico (Figura 2.2b). A compressão da mama dura alguns segundos e o procedimento completo leva cerca de 20 minutos (ACS, 2013h).



(a) (b)

Figura 2.2 (a) Mamógrafo real (RADIOLOGYINFO, 2013); (b) Esquema de mamógrafo (ACS, 2013h).

Apesar de ser a melhor forma de detecção precoce de câncer de mama e ajudar para diminuição da mortalidade desta doença, a mamografia possui algumas desvantagens. Um dos fatores é a insuficiência na provação que uma área é câncer, levando uma pequena quantidade do tecido suspeito à biópsia. Outra desvantagem é a dificuldade de detectar lesões em mamografias de mulheres mais jovens porque geralmente suas mamas são densas e podem ocultar a lesão. Isso geralmente não é uma grande preocupação, pois a maioria dos cânceres de mama detectados é em mulheres mais velhas (ACS, 2013h).

Alguns casos de câncer não são diagnosticados no exame de mamografia e ainda podem apresentar uma quantidade alta de falsos-positivos (regiões normais detectadas como lesões). Assim, com a finalidade de ajudar a aumentar a precisão e eficiência da mamografia, diversas técnicas têm sido desenvolvidas, como os sistemas CAD/CADx.

2.3 Técnica de Processamento de Imagens

A formação de uma imagem digital é definida como uma função bidimensional, onde x e y são coordenadas espaciais e a amplitude de *f* em um par de coordenadas (x,y) é denominada intensidade ou nível de cinza da imagem naquele ponto (GONZALEZ & WOODS, 2002). Esses valores são finitos e discretos. Uma imagem digital é constituída por um número finito de elementos (conhecido como pixels – *picture elements*), que possuem sua localização e um valor.

O processamento de imagens digitais compreende processos cujas entradas e saídas são imagens e, além disso, engloba os processos de extração de atributos 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, 2000).

A Figura 2.3 apresenta um esquema muito utilizado para demonstrar as diversas etapas do processamento de imagem. Após a definição e delimitação do problema, seguem-se as etapas: aquisição das imagens digitais, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação. O conjunto de resultados gerados por uma etapa é utilizado pela próxima. Nem sempre esse conjunto gerado é uma imagem. Sendo que ao fim de todas as etapas, o resultado pode ser ou não representado por uma imagem digital.

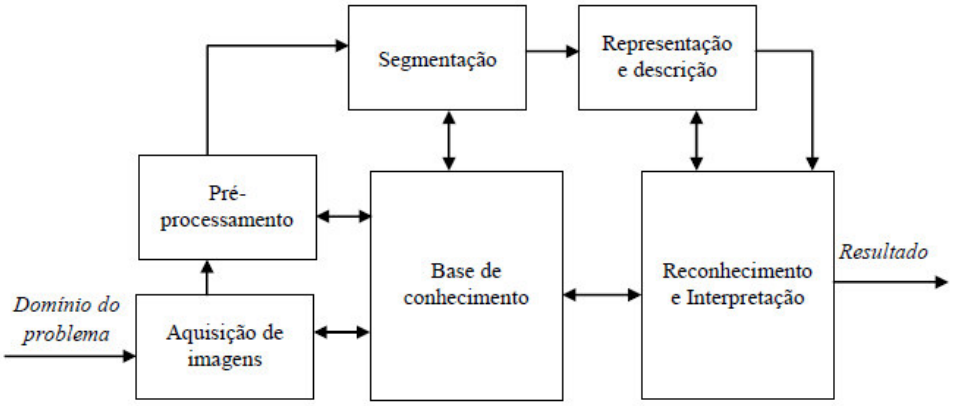


Figura 2.3 Etapas fundamentais em processamento de imagens digitais. Adaptado de Gonzalez e Woods (1992)(2002)(2008).

A aquisição de imagens é a primeira etapa no processamento de imagens digitais, onde um digitalizador converte a imagem analógica para digital.

A segunda etapa é o pré-processamento das imagens adquiridas. Esta etapa tem o objetivo de melhorar certas estruturas da imagem para aumentar as chances de sucesso dos processos seguintes. Nesta etapa podem ser aplicadas técnicas de realce e melhoramento de imagem, por exemplo: remoção de ruído, filtros morfológicos, dentre outras.

A segmentação é a terceira etapa, tendo esta por objetivo, extrair da imagem apenas partes que realmente interessam para o processamento. Nesta etapa é definido um processo de particionamento da imagem em regiões desconexas, onde os elementos de uma mesma região devem ser o mais homogêneo possível e os elementos de regiões distintas o mais heterogêneo possível. A segmentação pode ser abordada de três formas: manual, semi-automática e automática.

Na segmentação manual, o processo de separação de um objeto de interesse é realizado por um especialista humano, com o uso de ferramentas que auxiliem de forma visual. Na semi-automática, o especialista passa informações a respeito do que buscar na imagem, ou onde buscar determinada característica, para um algoritmo capaz de processá-las e posterior poder realizar a segmentação. Já a segmentação automática é o tipo mais complexo que precisa ter a capacidade de separar as várias regiões da imagem em conjuntos desconexos, satisfazendo aos critérios de similaridade entre cada região.

A quarta etapa é a representação e descrição, também conhecida como extração de características. Essa etapa tem o objetivo de determinar as características que resultem em alguma informação quantitativa de interesse ou que sejam básicas para discriminação entre classes distintas. O conjunto dessas medidas compõe um vetor de características que define um padrão para uma determinada área.

A última etapa é o reconhecimento e interpretação das imagens. O reconhecimento é responsável por atribuir um rótulo a um objeto, baseado em suas características, enquanto a interpretação atribui um significado a um conjunto de objetos reconhecidos.

2.3.1 Realce de Contraste

A técnica de realce de contraste tem por objetivo aumentar a discriminação visual entre os objetos presentes na imagem, sob os critérios subjetivos do olho humano

(SAMPAIO, 2009). Duas das técnicas mais simples de transformação de intensidade em processamento de imagem são: Logarítmico e Quadrático.

O logarítmico é um tipo de realce não linear que descreve o mapeamento do contraste de partes que possuem um nível baixo de intensidade em uma faixa mais alta (GONZALEZ & WOODS, 2008). A Equação 1 expressa o comportamento dessa técnica.

$$g'(x,y) = \frac{max * log[g(x,y)+1]}{log(max)} \quad (1)$$

onde $g'(x,y)$ é o pixel de saída, $g(x,y)$ é o pixel de entrada, max é o maior nível de cinza encontrado em determinada imagem e log é a função logarítmica na base 10.

O realce quadrático é também não linear, onde o seu mapeamento tem característica diversas do anterior. Esse realce faz transformação com o aumento do contraste nas partes que possuem um nível elevado de intensidade (GONZALEZ & WOODS, 2008). O cálculo desse realce quadrático é mostrado na Equação 2.

$$g'(x,y) = \frac{(g(x,y))^2}{max} \quad (2)$$

onde $g'(x,y)$ é o pixel de saída, $g(x,y)$ é o pixel de entrada e max é o maior nível de cinza encontrado em determinada imagem.

Neste trabalho, alguns experimentos não aplicaram o realce, já em outros, teve a utilização do logarítmico ou quadrático.

2.3.2 Filtro da Média

O filtro de média é um tipo de filtro passa-baixa. O efeito de um filtro passa-baixa é de suavização da imagem e minimização dos ruídos, atenuando ou eliminando as componentes de alta frequência, porém, o efeito acaba sendo de embassamento ou borramento da imagem que acaba removendo os detalhes finos da imagem (FILHO & NETO, 1999).

A forma mais simples de implementar um filtro da média é por meio construção de uma máscara (janela) 3x3 com todos seus coeficientes iguais a 1, dividindo o resultado da convolução (operação de varredura da imagem com aplicação da máscara) por um fator de normalização, neste caso igual a 9 (largura x altura máscara = 3x3). Outra forma de máscara é

o 5x5, que tem como característica o maior grau do borramento da imagem resultante. Na etapa de melhoramento da metodologia deste trabalho, alguns experimentos não aplicaram filtro da média, já em outros, teve aplicação com máscara 3x3 ou 5x5. A

Figura **2.4** mostra exemplo dessas duas máscaras.

$$\frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix} \qquad \frac{1}{25}\begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1 & 1\\ 1 & 1 & 1 & 1 & 1\end{bmatrix}$$

Figura 2.4 Máscaras para cálculo do filtro da média: (esquerda) 3x3, (direita) 5x5 (FILHO & NETO, 1999).

O cálculo da filtragem de imagem de forma geral é dado pela expressa (GONZALEZ & WOODS, 2008):

$$g(x,y) = \frac{\sum_{s=-a}^{a}\sum_{t=-b}^{b} w(s,t) f(x+s,y+t)}{\sum_{s=-a}^{a}\sum_{t=-b}^{b} w(s,t)} \tag{3}$$

onde $g(x,y)$ é o pixel de saída, $w(s,t)$ é a máscara de tamanho $m\,x\,n$ com coeficientes, $f(x+s,y+t)$ é o pixel entrada. Sendo que $a = (m-1)/2$ e $b = (n-1)/2$.

2.3.3 Quantização Uniforme

Uma imagem digital é discretizada espacialmente em x e y, e também em amplitude (intensidade luminosa). A discretização em amplitude é conhecida como níveis de cinza e a outra denominada amostragem (GONZALEZ & WOODS, 2000).

A quantização uniforme consiste em dividir a escala de cinza da imagem em intervalos iguais, onde cada intervalo é mapeado para um valor de cinza na imagem quantizada, de modo que a escala de cinza da imagem quantizada é dada por [0, L' - 1], sendo o nível de cinza da imagem quantizada (L') menor do que da imagem original (L), ou seja, L' < L(PEDRINI & SCHWARTZ, 2008).

A expressão para calcular este mapeamento é:

$$q(i,j) = (2^b - 1)\frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \quad (4)$$

onde $q(i,j)$ é o nível de cinza do pixel (i,j) da nova imagem (quantizada), $p(i,j)$ é o nível de cinza do pixel (i,j) da imagem original, $[I_{max} - I_{min}]$ é a escala de cinza da imagem original, e b é o número de bits necessário para armazenar cada pixel da imagem quantizada.

2.3.4 Textura

A definição de textura encontrada na literatura é descrita de diversas formas. Segundo Haralick *et al.* (1973), textura é definida como a característica de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. Gonzalez e Woods (2002) afirma que a textura é intuitivamente descrita por medidas que quantificam suas propriedades de suavidade, rugosidade e regularidade. Baseado num conceito bidimensional, a textura é caracterizada pela dimensão que contém as propriedades primitivas da tonalidade e a outra corresponde aos relacionamentos espaciais entre elas.

A análise de textura é relevante em imagens digitais, uma vez que possibilita distinguir regiões da imagem que apresentam as mesmas características de padrões (CONCI, AZEVEDO, & LETA, 2008). As três abordagens principais usadas em processamento de imagens para a análise de texturas são: estrutural, espectral e estatística (GONZALEZ & WOODS, 2000).

A abordagem estrutural considera que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, conforme regras bem definida. A abordagem espectral é baseada em propriedades do espectro de Fourier, sendo utilizada principalmente na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro. A abordagem estatística define a textura como um conjunto de medidas locais extraídas do padrão, favorecendo a descrição de imagens através de regras estatísticas que regem tanto a distribuição quanto à relação entre os diferentes níveis de cinza.

Neste trabalho, propomos descrever a textura dos tecidos de regiões de imagens mamográficas através do Índice de Diversidade Taxonômica e do Índice de Distinção Taxonômica, que são medidas estatísticas.

2.4 Índice de Diversidade

A diversidade é um termo muito utilizado na área da ecologia. O seu objetivo é informar a variedade de espécies presentes em uma comunidade ou área. Segundo Magurran (2004), o conceito de comunidade é descrito como um conjunto de espécies que ocorrem em um determinado lugar e tempo. As medições como a variância e desvio padrão que são calculadas em estudos estatísticos, apresentam valores que medem a variabilidade quantitativa, enquanto que os índices de diversidade descrevem a variabilidade qualitativa (JOHNSON & KOTZ, 1988).

Para medir a diversidade temos duas componentes: a *riqueza de espécies*, que consiste no número de espécies encontradas em determinada região, e a *abundância relativa*, que é o número de indivíduos de uma determinada espécie existentes numa dada área (PIANKA, 1994). O resultado do cálculo para qualquer índice de diversidade é representado por um único valor (SANTOS, 2009). As medidas de diversidade de espécies são geralmente úteis para comparar padrões em diferentes áreas.

A forma mais simples da aplicação do índice de diversidade em imagens é quando a comunidade representa uma imagem ou região da mesma, as espécies sendo os níveis de cinza e os indivíduos sendo os pixels, como descrito em Sousa (2011). Na Figura 2.5 mostra um exemplo demonstrativo.

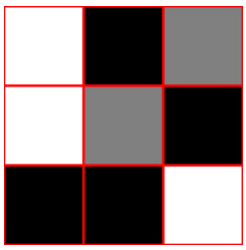


Figura 2.5 Exemplo de uma imagem que possui 3 níveis de cinza (espécies) que é o preto (P), cinza (C) e banco (B). A quantidade de pixels (indivíduos) de B é 3, C é 2 e P é 4.

2.4.1 Diversidade Filogenética

A filogenia é um ramo da biologia responsável pelo estudo das relações evolutivas entre as espécies, pela verificação dos relacionamentos entre elas, a fim de determinar possíveis ancestrais comuns (ARAÚJO, 2003). Uma árvore filogenética, ou simplesmente filogenia, é uma árvore onde as folhas representam os organismos e os nós internos

representam supostos ancestrais. As arestas da árvore denotam as relações evolutivas (ARAÚJO, 2003). Na Figura 2.6, temos um exemplo de árvore filogenética, onde verifica o relacionamento entre espécies de macacos e a espécie humana, onde podemos ver que o homem e o chimpanzé são geneticamente mais próximos que os outros pares presentes na árvore (ARAÚJO, 2003).

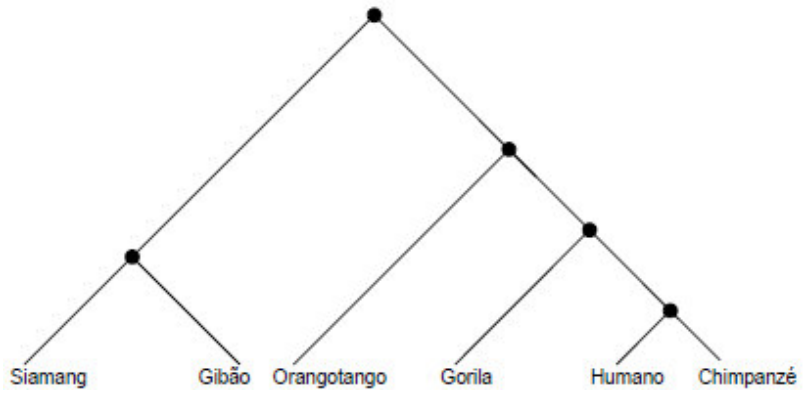


Figura 2.6 – Representação de uma árvore filogenética para alguns primatas (ARAÚJO, 2003).

De maneira geral a diversidade não pode ser medida apenas com a utilização de dados como a abundância e a riqueza de espécies e cada vez mais o parâmetro filogenético vem sendo inserido neste cálculo (CLARKE & WARWICK, 1998). A diversidade filogenética é uma medida da diversidade de uma comunidade que incorpora as relações filogenéticas das espécies (MAGURRAN, 2004). A combinação de abundância das espécies com a proximidade filogenética para gerar um índice de diversidade é denotada diversidade taxonômica (SILVA & BATALHA, 2006). Segundo Vandamme *et al.* (1996), a taxonomia é a ciência que lida com a classificação (criação de novos taxa), identificação (alocação de linhagens dentro de espécies conhecidas) e nomenclatura.

Clarke e Warwick (1998) desenvolveram um método para mensurar a diversidade taxonômica muito sensível a perturbações ambientais e apropriado para avaliar as diferenças entre comunidades (SILVA & BATALHA, 2006). Uma comunidade em que as espécies estão distribuídas em muitos gêneros deve apresentar uma diversidade maior que uma comunidade em que a maioria das espécies pertence a um mesmo gênero (MAGURRAN, 2004).

A diversidade taxonômica é baseada no conjunto das distâncias entre pares de espécies acumuladas a partir das árvores taxonômicas (RICOTTA, 2004). A Figura 2.7, apresenta uma ilustração de uma árvore taxonômica em que as folhas são as espécies e a soma da quantidade de arestas que ligam determinado par de espécies é informada pela matriz.

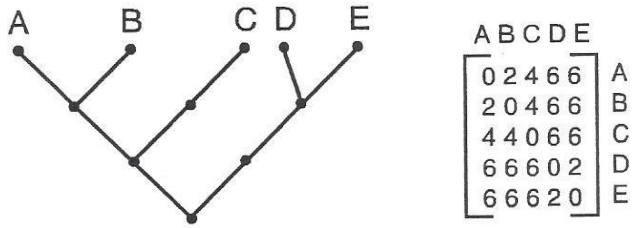


Figura 2.7 Demonstração de uma árvore taxonômica com sua matriz de distância entre espécies (RICOTTA, 2004).

Uma das formas de representar a árvore filogenética é através do cladograma, que é um diagrama representativo das relações ancestrais entre organismos. Para este trabalho foi utilizando a topologia de um cladograma mais específico, o *enraizado na forma de cladograma inclinado* (VIANA, 2007). Esse tipo de árvore é mostrado na Figura 2.8, que descreve a sequência evolutiva de alguns tetrápodes (vertebrados terrestres possuidores de quatro membros).

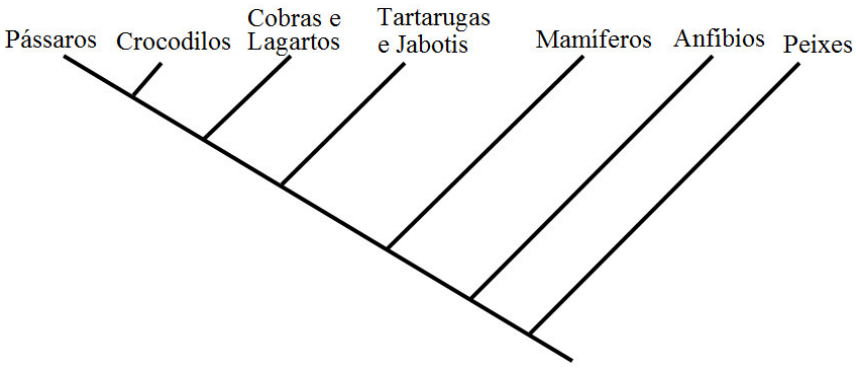


Figura 2.8 Árvore filogenética enraizada na forma de cladograma inclinado (VIANA, 2007).

2.4.2 Índices Taxonômicos

O cálculo entre dois organismos escolhidos aleatoriamente em uma filogenia existente em uma comunidade é apresentado por índices de Diversidade Taxonômica e Distinção Taxonômica (WARWICK & CLARKE, 1995). Neste trabalho são utilizados esses dois índices.

O Índice de Diversidade Taxonômica (Δ) considera a abundância das espécies e a relação taxonômica entre elas, assim, o seu valor expressa a distância taxonômica média entre quaisquer dois indivíduos, escolhidos na amostra ao acaso (GORENSTEIN, 2009).

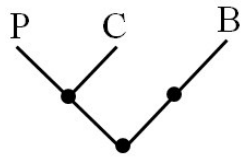
$$\Delta = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{[n(n-1)/2]} \quad (5)$$

onde x_i (i = 1, ..., s) é abundância da i-ésima espécie, n é o número total de indivíduos e ω_{ij} é a distância da espécie i à espécie j na classificação taxonômica.

Já o Índice de Distinção Taxonômica (Δ*) representa a distância taxonômica média entre dois indivíduos, com a restrição de que sejam de espécies diferentes (GORENSTEIN, 2009).

$$\Delta^* = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j} \tag{6}$$

Uma representação genérica de uma árvore taxonômica e sua matriz de distância para a imagem da Figura 2.5 é mostrada a seguir:



$$\begin{array}{c} \begin{matrix} P & C & B \end{matrix} \\ \begin{bmatrix} 0 & 2 & 4 \\ 2 & 0 & 4 \\ 4 & 4 & 0 \end{bmatrix} \begin{matrix} P \\ C \\ B \end{matrix} \end{array}$$

Figura 2.9 Exemplo representando uma imagem em uma árvore taxonômica (à esquerda) e sua matriz de distância (à direita).

Baseado na Figura 2.9, a aplicação do índice Δ e Δ* é demonstrado a seguir:

$$\Delta = \frac{w_{pc}x_p x_c + w_{pb}x_p x_b + w_{cb}x_c x_b}{n(n-1)/2} = \frac{2.4.3 + 4.4.2 + 4.3.2}{3(3-1)/2} = 26{,}67$$

$$\Delta^* = \frac{w_{pc}x_p x_c + w_{pb}x_p x_b + w_{cb}x_c x_b}{x_p x_c + x_p x_b + x_c x_b} = \frac{2.4.3 + 4.4.2 + 4.3.2}{4.3 + 4.2 + 3.2} = 3{,}08$$

2.5 Reconhecimento de Padrões

A técnica de reconhecimento de padrões é um sub-tópico da aprendizagem de máquina cujo objetivo é classificar objetos (padrões) baseado em um conhecimento *a priori* ou em informações estatísticas extraídas dos padrões. Um padrão é tudo aquilo para o qual existe uma entidade nomeável representante, geralmente, criada através do conhecimento cultural humano (LOONEY, 1997).

Essa técnica envolve dois processos: classificação e reconhecimento. O primeiro é o processo onde uma amostra de uma população qualquer é dividida em dois grupos denominados classes. E o segundo, é o processo onde uma amostra desconhecida da mesma

população é reconhecida como pertencente a uma das classes criadas. Segundo Looney (1997), a classificação pode ser realizada de duas formas: aprendizagem supervisionada e aprendizagem não-supervisionada.

No processo não-supervisionado, um conjunto de objetos representantes de uma população é examinado. Esse conjunto é dividido em subconjuntos (classes) de acordo com critérios de similaridade intra-classe e dissimilaridade extra-classe. Esse processo também é chamado de agrupamento. Já no processo supervisionado um "reconhecedor" pode ser treinado previamente para identificar a classe de qualquer objeto desconhecido da mesma população. Este trabalho usa técnica de classificação supervisionada.

Os objetos possuem determinadas propriedades que são utilizadas na distinção entre classes. Essas propriedades são utilizadas em toda a população e possibilitam que os objetos possam ser reconhecidos como pertencentes ou não a uma determinada classe. As propriedades individuais são chamadas de características (*features*). Quando existem *N* características observáveis em uma população, tem-se um vetor de características (*feature vector*). Os vetores de características são responsáveis por representar os objetos em uma população de objetos e possibilitam que o reconhecimento de padrões seja realizado.

Com o conjunto de vetores de características com grande quantidade de variáveis existe a possibilidade de serem definidos valores correlatados ou redundantes que podem sobrecarregar o classificador e induzi-lo ao erro. Para isso, os vetores de características são pré-processados (algoritmo stepwise ou PCA), com o objetivo de eliminar as características desnecessárias.

Após a eliminação, atribui-se um rótulo para cada vetor de característica. O rótulo é a determinação *a priori* de uma classe a partir do conhecimento humano, que neste trabalho foi definido uma rotulação binária, ou seja, "1" para classe *massa* e "-1" para *não-massa*. Depois os vetores de características são divididos no conjunto de treinamento e teste. O primeiro subconjunto gerado pelo classificador é uma assinatura única para cada rótulo contido dentro do conjunto de amostras. Essa assinatura representa as características que melhor representam distinção entre as classes e será especialmente útil no processo de reconhecimento de padrão.

No segundo subconjunto, o classificador atribui um rótulo a cada amostra de teste conforme o conhecimento prévio obtido na etapa de treinamento, ainda que o objeto não

pertença a nenhuma das classes. Com isso, a tentativa de reconhecimento de padrão de um amostra deve ser feita necessariamente sobre amostras da mesma população comparados ao do treinamento, garantindo que os padrões gerados na etapa de treinamento sejam válidos para a etapa de teste.

Neste trabalho utiliza-se como classificador a Máquina de Vetores de Suporte para realizar o reconhecimento de padrões de tecidos de mama extraídos de imagens mamográficas.

2.5.1 Máquina de Vetores de Suporte

Segundo Vapnik (1998), a Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) 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 (

Figura 2.10b). Em Santos (2002) é definida a margem como a distância entre os pontos de dados, de ambas as classes, mais próximos ao hiperplano.

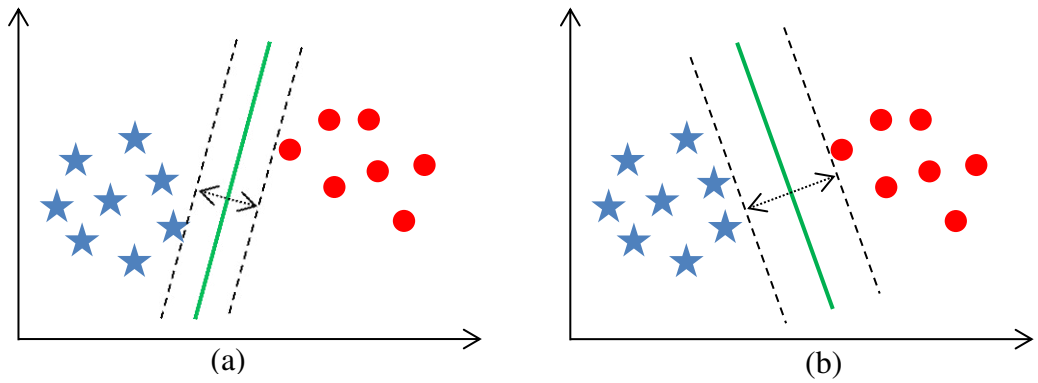


Figura 2.10 O conjunto de estrelas pertencem a uma classe e o conjunto de círculos a outra classe. A margem de separação entre as classes é definida pelas retas tracejadas. Em (a) mostra um hiperplano (em verde) para o conjunto de treinamento bidimensional com margem pequena e em (b) um hiperplano de margem máxima.

Os algoritmos de treinamento das MVS possuem forte influencia da teoria de otimizacao e de aprendizagem estatística. Segundo Cristianni e Shawe-Taylor (2000) as MVS vem demonstrando sua superioridade em comparação a outros classificadores em uma grande variedade de aplicações.

A MVS tem como ideia principal a construção de um hiperplano ótimo de separação entre as classes, onde é baseado em um conjunto de pontos denominados “vetores de suporte” (Figura 2.11). Por hiperplano, entende-se um subespaço planar (n-1)-dimensional contido em um espaço planar n-dimensional, onde n é um valor inteiro, positivo e finito. Na situação em que as duas classes não são separáveis, a MVS é capaz de encontrar um hiperplano através do uso de conceitos pertencentes à teoria da otimização. A Figura 2.12 mostra em duas dimensões os hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (em verde), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

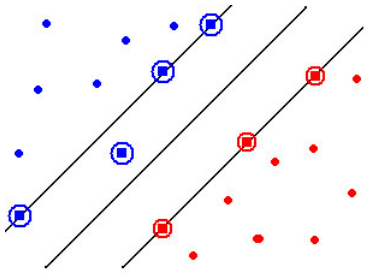


Figura 2.11 Vetores de suporte (destacados por círculos) (JUNIOR, 2008).

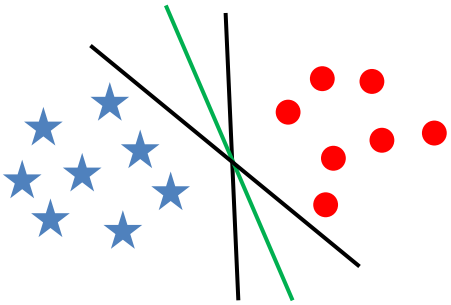


Figura 2.12 Separação de duas classes (estrelas e círculos) através de hiperplanos, onde o reta verde (em duas dimensões) é o hiperplano ótimo.

Dado o conjunto de amostras de treinamento (x_i, y_i), com $x_i \in R^n$, $i = 1, 2, \ldots, n$, onde x_i pertence a uma das duas classes identificadas pelo rótulo $y_i \in \{-1,1\}$. O objetivo é estimar uma função $f: \mathcal{R}^n \longrightarrow \{-1, +1\}$, que separe corretamente os exemplos de testes em duas classes distintas. A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando por valores de w e b tais que a seguinte equação seja satisfeita:

$$y_i = \big((w.x_i) + b\big) > 1 \tag{7}$$

onde w é o vetor normal ao hiperplano de decisão e b é o corte ou distancia da função f em relação à origem. Os valores ótimos de w e b serão encontrados ao minimizar a Equação 8, de acordo com a restrição dada pela Equacao 7 (CHAVES, 2006).

$$\Phi(w) = \frac{w^2}{2} \tag{8}$$

MVS exige solução para o seguinte problema de otimização:

$$\Phi(\mathrm{w}, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{9}$$

$$y_i((w.x_i) + b) + \xi_i \geq 1 \tag{10}$$

onde C é uma penalidade (parâmetro a ser escolhido pelo usuário) para a função Φ, ξ_i é a variável de folga que suaviza as restrições dada pela Equação 10, e N é o número de amostras de entrada.

Através da teoria dos multiplicadores de *Lagrange*, chega-se à Equação 11. O objetivo então passa a ser encontrar os multiplicadores de *Lagrange* α_i ótimos que satisfaçam a Equação 12 (CHAVES, 2006):

$$w(a) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \left(x_i, y_j\right) \tag{11}$$

$$\sum_{i=1}^{n} \alpha_i \, y_i = 0, 0 < \alpha_i < C \tag{12}$$

Somente os pontos onde a restrição da Equação 7 seja exatamente igual à unidade têm correspondentes $\alpha \neq 0$. Esses pontos são chamados de vetores de suporte. 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 ao invés do espaço de entrada original. Essa construção depende do cálculo de uma função K de núcleo de um produto interno (HAYKIN, 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 13 mostra o resultado da Equação 11 com a utilização de um núcleo K.

$$w(a)\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \,\mathrm{K}(x_i, y_j) \tag{13}$$

O kernel utilizado neste trabalho foi a função de base radial, que é definido pela equação (CHIH-CHUNG & CHIH-JEN, 2001):

$$\mathrm{K}(x_i, y_j) = exp\left(-\gamma\left\|x_i - x_j\right\|^2\right) \tag{14}$$

onde $\gamma > 0$ é um parâmetro que também é definido pelo usuário.

2.6 Métricas de Validação de Resultados

Em problemas ligados a área de saúde, a estrutura básica dos testes de classificação é para determinar quão bem um teste discrimina a presença ou ausência de uma doença. Nesse tipo de problemas, existe a presença de uma variável preditora (resultado do teste) e uma variável resultante (a presença ou ausência da doença).

Na avaliação de um sistema de reconhecimento de padrões relacionado à área médica existem quatro possíveis situações em relação ao diagnóstico:

- teste é positivo e o paciente tem a doença - Verdadeiro Positivo (VP);
- teste é positivo, mas o paciente não tem a doença - Falso Positivo (FP);
- teste é negativo e o paciente tem a doença - Falso Negativo (FN);
- teste é negativo e o paciente não tem a doença - Verdadeiro Negativo (VN).

A avaliação desses sistemas é comum medir-se o desempenho da metodologia calculando-se algumas estatísticas sobre os resultados dos testes para avaliar o desempenho do classificador, como Sensibilidade (SE), Especificidade (ES) e Acurácia (AC) (BLAND, 2000).

A sensibilidade define a proporção de verdadeiros-positivos identificados no teste. Indica quão bom é o teste para identificar indivíduos com presença de massa:

$$\mathrm{SE} = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \tag{15}$$

A especificidade define a proporção de verdadeiros-negativos identificados no teste. Indica quão bom é o teste para identificar indivíduos que não apresentam massa:

$$\mathrm{ES} = \frac{\mathrm{VN}}{\mathrm{VN+FP}} \tag{16}$$

A acurácia define a taxa de casos identificados corretamente e o número total de casos:

$$\mathrm{AC} = \frac{\mathrm{VP+VN}}{\mathrm{VP+VN+FP+FN}} \tag{17}$$

Uma forma de avaliar visualmente o desempenho entre vários experimentos é através do *Gráfico de barras com desvio padrão*. Esse gráfico é mostrado na Figura 2.13, onde as acurácias médias são descritos no eixo Y, os tipos de experimentos no eixo X. Cada experimento possui uma média de acurácia (ponto em vermelho) e desvio padrão (barra vertical acima do ponto vermelho é positivo e abaixo é negativo). A interpretação desse gráfico descreve que o melhor experimento é aquele que está mais próximo de 100% na média da acurácia e menor desvio padrão pra $\pm$ (Experimento J). Nesse trabalho foi utilizado esse gráfico para avaliar qual experimento entre vários tem melhor desempenho.

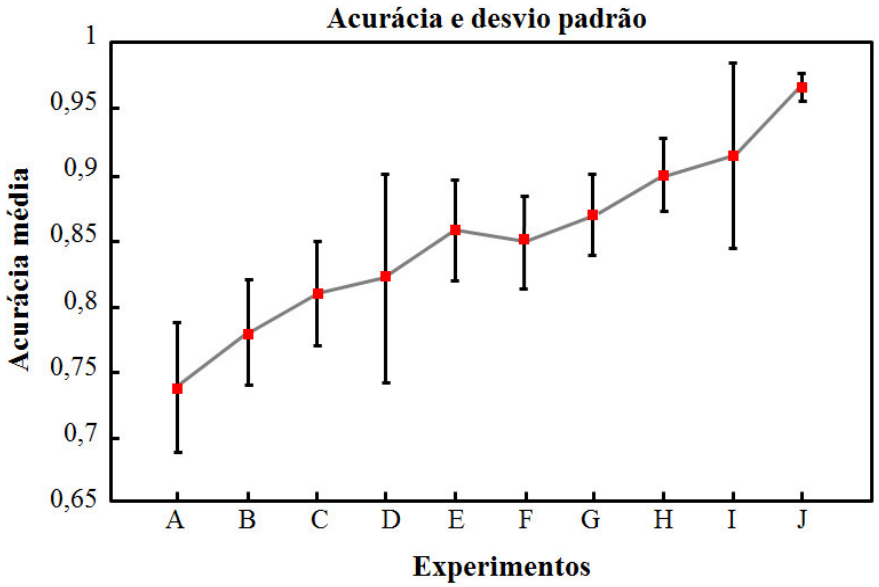


Figura 2.13 Gráfico de barras com desvio padrão. No eixo Y são valores das acurácias médias e no eixo X são diferentes experimentos. Adaptado de Moayedi *et al.* (2010).

3 MÉTODOS

Neste capítulo são descritas as etapas utilizadas na metodologia proposta para classificação de regiões de tecidos da mama em *massa* e *não-massa* (Seção 3.1). Primeiramente, é mostrada a base de imagens utilizada nos experimentos (Seção 3.1.1). Depois é descrito a etapa de pré-procesamento das imagens, com os passos de realce e melhoramento (Seção 3.1.2). Em seguida, relatamos os passos da extração de características das amostras, utilizando os índices Taxonômicos (Seção 3.1.3). Posteriormente classificamos as mesmas, utilizando o classificador MVS (Seção 3.1.4), a metodologia é finalizada avaliando os resultados (Seção 3.1.5).

3.1 Metodologia Proposta

A Figura 3.1 apresenta a sequência de etapas usadas na metodologia proposta neste trabalho.

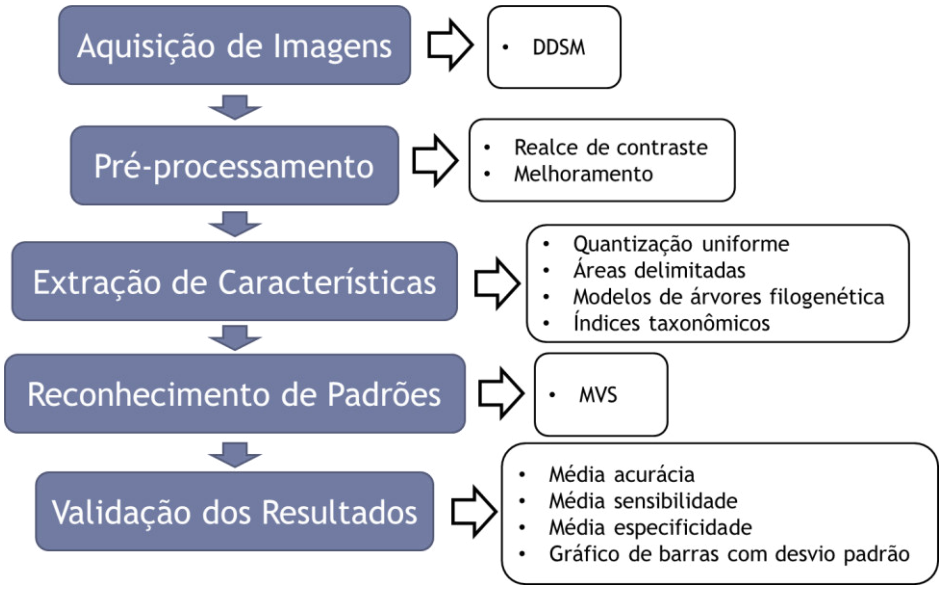


Figura 3.1 Metodologia proposta.

Na primeira fase é feita a aquisição de imagens, onde são obtidos exames de mamografias normais e não normais, a partir dos quais é feita a extração das regiões de interesse, de tecidos de *massa* e *não-massa*. Em seguir, as regiões de interesse adquiridas são submetidas a um pré-processamento através da aplicação de um algoritmo de realce

(logarítmico ou quadrático) ou sem e/ou um processo de melhoramento com filtros da média. Na etapa seguinte, é feita a extração de características para cada região de interesse, usando os índices Taxonômicos com diferentes níveis de *quantizações*.

Os vetores de características gerados na etapa anterior são encaminhados para etapa de reconhecimento de padrões, realizada com o classificador MVS. Neste processo, uma parte dos vetores de características é utilizada na etapa de *treinamento*, para gerar um *modelo* de classificação. O restante é, então, *classificado* através deste *modelo*. Os resultados da classificação são usados na última etapa que é a *validação*.

3.1.1 Aquisição das Imagens

Neste trabalho foram utilizadas imagens de mamografias digitalizadas do banco de dados *DDSM - Digital Database for Screening Mammography* (HEATH, BOWYER, KOPANS, MOORE, & KEGELMEYER, 2001), que está disponível gratuitamente na Web (DDSM, 2013c). A base DDSM possui 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, tais como a idade e a densidade da mama. Nas imagens que apresentam áreas não normais (*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*. O contorno da lesão está codificado em *chain code* (MORSE, 2000).

Foram utilizadas 300 regiões de interesse de tecidos *não-massa* e 300 regiões de interesse de tecidos com *massas* (150 benignas e 150 malignas, sendo somente massas não havendo qualquer calcificação), totalizando 600 amostras. A mesma quantidade de amostras selecionadas de forma aleatória para cada uma das classes foi usada com o intuito de não interferir na etapa de classificação. Nas amostras de massa, foram eliminados primeiramente as que tinham descrito no *overlay* informações sobre calcificação (pontos brancos na imagem) que possam prejudicar na descrição da textura e depois eliminados de forma visualmente as amostras que não apresentaram *pixels* internamente ao contorno, ou seja, possuíam “buracos” dentro da lesão. Então, as amostras de massa foram extraídas das mamografias a partir do contorno da lesão, através da aplicação do menor retângulo alinhados aos eixos do sistema de

coordenadas e que circunscreve a região (*axis aligned bounding box*). Os pixels externos ao contorno e internos a caixa delimitadora tiveram suas intensidades substituídas pelo valor -1, para distinguir áreas que não constituem massa e evitar que sejam inseridas nas etapas seguintes. Assim, apenas os pixels da região interna ao contorno são processados nas etapas de pré-processamento e extração de características. Essas amostras já foram utilizadas em Carvalho (2012) e Junior (2008), a Figura 3.2 mostra um exemplo de *massa* e *não-massa* retiradas de mamografias diferentes. As amostras de *não-massas* foram retiradas manualmente na forma de retângulo de tamanhos diferentes e em localizações aleatórias da mamografia (não são selecionadas as regiões do fundo da imagem com seus rótulos e músculo peitoral (MLO)) que estão classificadas nos casos normais na base DDSM.

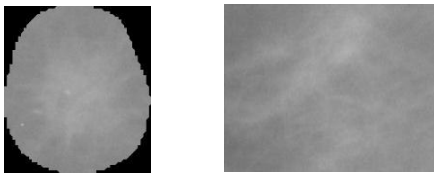


Figura 3.2 As duas imagens são regiões de mamografia diferentes, onde na esquerda mostra uma massa benigna (DDSM, 2013d) e na direita uma não-massa (DDSM, 2013e).

Com a abordagem apresentada anteriormente, as *regiões de interesse* extraídas possuem tamanhos diferentes (Figura 3.2) para conservar o máximo de informação de textura presente nos tecidos de *massas*. Isto não prejudicou o processo de extração de características, uma vez que, segundo Magurran (2004), os índices de diversidade taxonômicos apresentam a vantagem de serem independentes do esforço amostral (número de indivíduos encontrados).

3.1.2 Pré-Processamento

Após a aquisição das imagens, as mesmas foram submetidas a pré-processamento com o objetivo de realçar suas características. O uso do realce de contraste não-linear logarítmico (Seção 2.3.1) foi considerado porque sua aplicação mapeia uma faixa de baixos valores de intensidade luminosa para faixas maiores, tornando visíveis partes da imagem que se encontravam muito escuras. Já o realce quadrático (Seção 2.3.1) que também é não-linear, tem o mapeamento que aumenta o contraste de feições claras. Outro passo trabalhado foi a de melhoramento, com a aplicação do filtro da média (máscara 3x3 ou 5x5) (Seção 2.3.2) que tem como principal característica a suavização da imagem ao reduzir as altas frequências e apresenta como principal vantagem a redução de ruídos e como desvantagem o borramento da imagem.

Nesta etapa, foram realizados experimentos com uso do realce logarítmico ou quadrático e outros sem. No processo seguinte alguns experimentos utilizaram o filtro da média 3x3 ou 5x5.

3.1.3 Extração de Características

Depois do pré-processamento das regiões de interesse, as mesmas são encaminhadas à fase de extração de características de textura. Nesta fase, foram realizados experimentos com a aplicação da técnica de quantização uniforme (Seção 2.3.3) e sem esta aplicação. As amostras foram quantizadas em 256, 128, 64, 32 e 16 níveis de cinza. Para descrição da textura dos objetos foram utilizados os Índices de Diversidade Taxonômica (Seção 2.4.2) e Distinção Taxonômica (Seção 2.4.2). Como estes índices são baseados na distância filogenética (contabilização do número de arestas) a partir da arquitetura de determinada árvore, foram desenvolvidas três formas de árvores para este trabalho, mais detalhes nas Seções 3.1.3.5, 3.1.3.6 e 3.1.3.7. Os outros requisitos necessários para a geração da árvore são as espécies (níveis de cinza) e os indivíduos (pixels) adquiridos com base nas quatro abordagens que foram: circulares (Seção 3.1.3.1), anéis (Seção 3.1.3.2), máscaras internas (Seção 3.1.3.3) e externas (Seção 3.1.3.4). Essas abordagens foram utilizadas para encontrar padrões de textura que melhor descreva *massa* e *não-massa*.

3.1.3.1. Abordagem em Círculos

Esta abordagem tem o objetivo de descobrir padrões de diversidade nas áreas próximas à borda da região e nas mais internas. Foram extraídos *n* círculos concêntricos e sobrepostos, com diferentes raios, partindo do centro de massa da imagem como mostra a equação a seguir:

$$Cx = \frac{\sum_{i=1}^{n} x_i}{n} \qquad Cy = \frac{\sum_{i=1}^{n} y_i}{n} \tag{18}$$

onde, x_i e y_i são as coordenadas dos pixels que incluem a massa, ou seja, valores diferentes de -1, e n é o número total de pixels encontrado.

O tamanho de cada raio i é definido pela equação:

$$R_i = \frac{i}{6R_0} \tag{19}$$

onde, R_0 é tamanho do raio que circunscreve a amostra, ou seja, toda a dimensão da região de interesse, e R_i são os raios menores para $i = 1, 2, \ldots, n$.

Neste trabalho os melhores resultados obtidos foram com 5 círculos, ou seja, $n = 4$, partindo do raio R_0. Um exemplo das áreas circulares é mostrado na Figura 3.3. Baseados nesta abordagem são criadas as árvores que serão usadas para calcular os pesos necessários para aplicação dos índices Taxonômicos. O número de variáveis geradas com esta abordagem é um total de 25 (5 círculos e 5 quantizações) para cada árvore utilizada (Seções 3.1.3.5, 3.1.3.6 e 3.1.3.7).

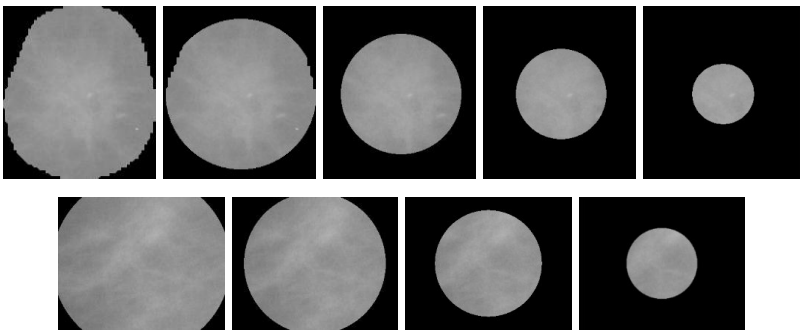


Figura 3.3 Primeira linha é a ROI anormal e na segunda linha é a ROI normal em círculos.

3.1.3.2. Abordagem em Anéis

Esta abordagem é similar à abordagem em círculo. São utilizados 4 anéis para esta abordagem que são adquiridos por quatro raios consecutivos simultaneamente, onde somente os pixels dentro da região anelar são considerados. Um exemplo dessas regiões é mostrado a seguir:

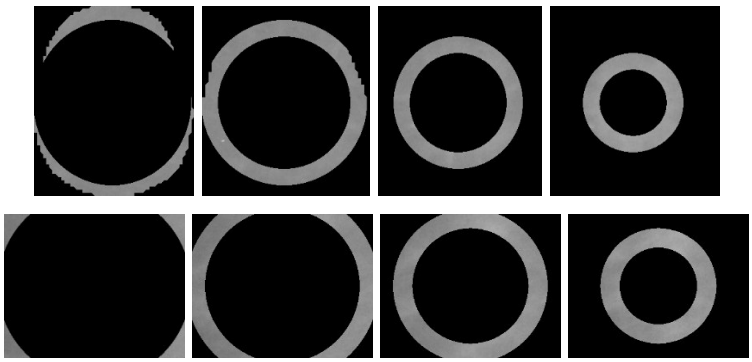


Figura 3.4 Primeira linha é a ROI anormal e na segunda linha é a ROI normal em anéis.

O número de variáveis adquiridas com esta abordagem é um total de 20 (4 anéis e 5 quantizações) para cada árvore utilizada (Seções 3.1.3.5, 3.1.3.6 e 3.1.3.7).

3.1.3.3. Abordagem com Máscara Interna

A proposta desta abordagem é baseada na mesma ideia de círculos, em descobrir padrões de diversidade nas áreas perto da borda e internamente. Essas regiões foram geradas a partir de máscaras, que são imagens binárias. A primeira máscara foi criada com a binarização da ROI original, a segunda com base na diminuição da escala em relação a primeira pelo centro de massa e as sucessores máscaras foram adquiridas a partir das suas anteriores na sequencia até a mais interna. O esquema do procedimento para geração das máscaras e consequentemente suas áreas de interesse é apresentado na Figura 3.5.

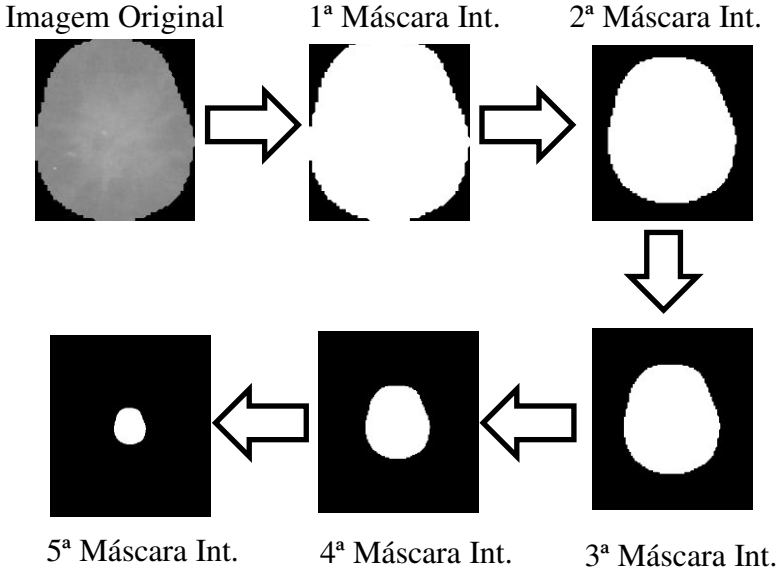


Figura 3.5 Procedimento da criação de 5 máscaras internas.

Neste trabalho foi definido o valor de 20% na diminuição da escala, pois nos testes foi verificado que os melhores resultados foram obtidos utilizando-se 5 máscaras de imagem com essa proporção de escalonamento, sendo que a 1ª máscara não foi escalada (imagem original binarizada). A Figura 3.6 exibe um exemplo das áreas geradas para *massa* e *não-massa*. O número de variáveis conseguidas com esta abordagem é um total de 25 (5 máscaras internas e 5 quantizações) para cada árvore utilizada (Seções 3.1.3.5, 3.1.3.6 e 3.1.3.7).

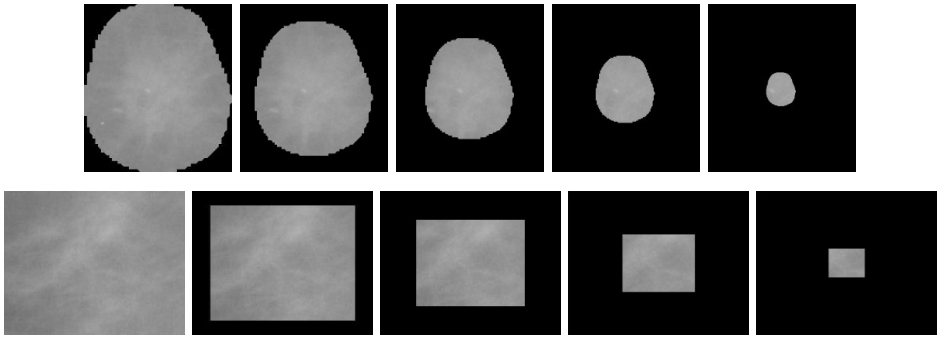


Figura 3.6 A primeira linha é a ROI anormal e na segunda linha é a ROI normal em máscaras internas.

3.1.3.4. Abordagem com Máscara Externa

Esta abordagem é similar à abordagem com máscara interna. As máscaras externas são formadas pela diferença entre as máscaras internas, onde a primeira foi criada pela diferença entre a primeira e segunda da interna, a segunda entre a terceira com a segunda máscara interna. As demais máscaras externas foram originadas da sequencia das diferenças até o último par das máscaras internas. Na Figura 3.7 são demostrados os passos para criação das máscaras externas.

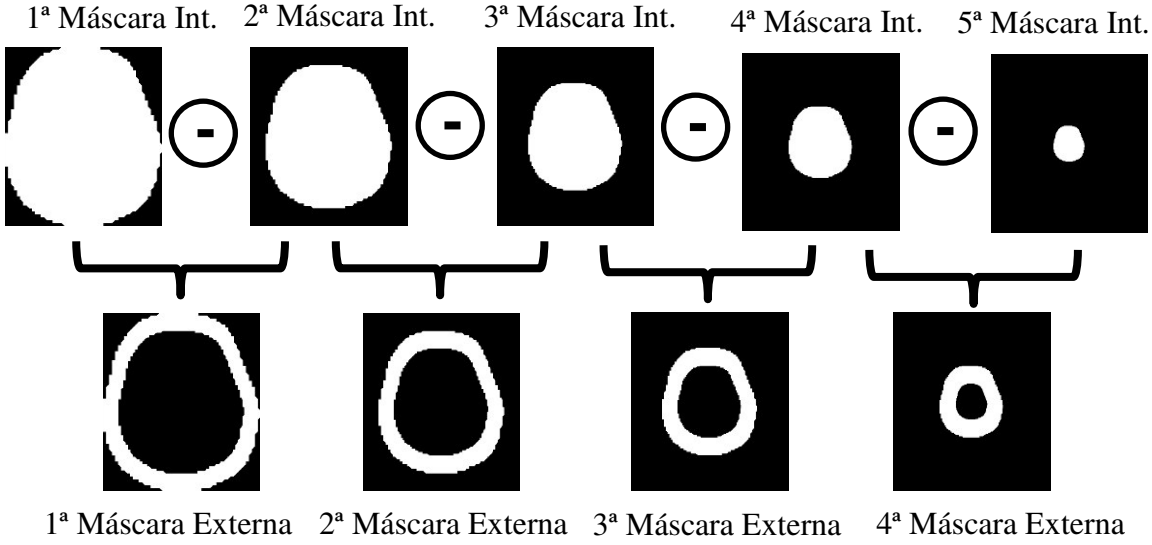


Figura 3.7 Procedimento da criação de 5 máscaras externas.

Foram utilizadas 4 máscaras internas para esta abordagem. O número de variáveis geradas com esta abordagem é um total de 20 (4 máscaras internas e 5 quantizações) para cada árvore utilizada (Seção 3.1.3.5, 3.1.3.6 e 3.1.3.7. Um exemplo das áreas geradas para *massa* e *não-massa* é mostrada na Figura 3.8.

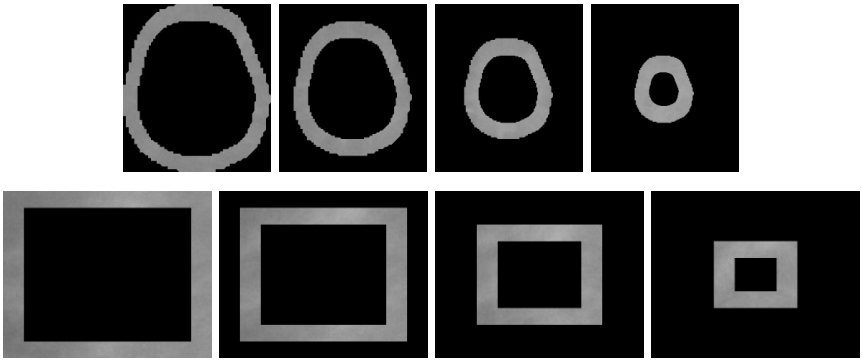


Figura 3.8 A primeira linha é a ROI anormal e na segunda linha é a ROI normal em máscaras externas.

3.1.3.5. Árvore 1 – Árvore Enraizada na Forma de Cladograma Inclinado

Após a geração das abordagens descritas nas seções anteriores são criadas as árvores para cada uma das áreas delimitadas. Os índices Taxonômicos (Δ e Δ*) são baseados nas árvores filogenéticas que possuem três fatores essenciais para aplicação: número de espécies, número de indivíduos e a estrutura de ligação das espécies (quantidade de arestas). Para representar as imagens, utilizamos o modelo de árvore enraizada na forma de cladograma inclinado (Seção 2.4.1). Na Figura 3.9 é mostrada uma árvore, onde as espécies são os níveis de cinza totalizando em 256 (as mamografias utilizadas são de 8 bits) e os indivíduos são pixels da imagem.

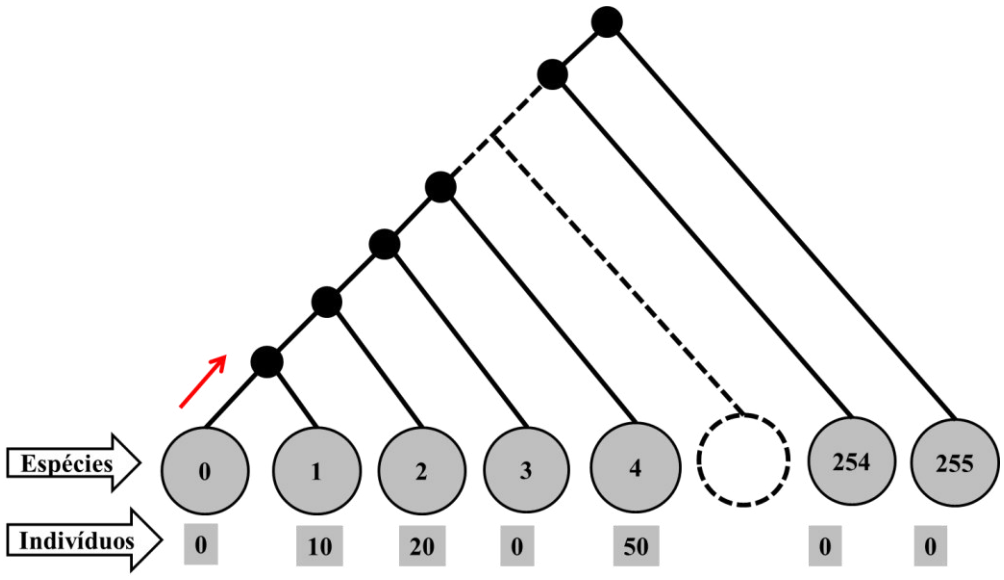


Figura 3.9 Árvore 1: árvore enraizada na forma de cladograma inclinado.

A relação entre as espécies da Árvore 1 é feita no sentido da esquerda para direita (seta vermelha). Assim, a primeira relação é entre a espécie 0 (zero) e 1 (um) que possui duas arestas ligando as mesmas (ω_{01}), como mostra em (a) na Figura 3.10, e também é informado

o cálculo do denominado das equações 5 e 6. Na segunda relação são três arestas (ω_{02}) que ligam as espécies 0 (zero) e 2 (Figura 3.10b). Em seguida é a combinação entre 0 (zero) e 3, com quatro arestas (ω_{03}) (Figura 3.10c). Depois é encontrado cinco arestas (ω_{04}) entre 0 (zero) e 4 (Figura 3.10d). Então, o último relacionamento da espécie 0 (zero) é com a espécie 255, que possui 256 arestas (ω_{0255}).

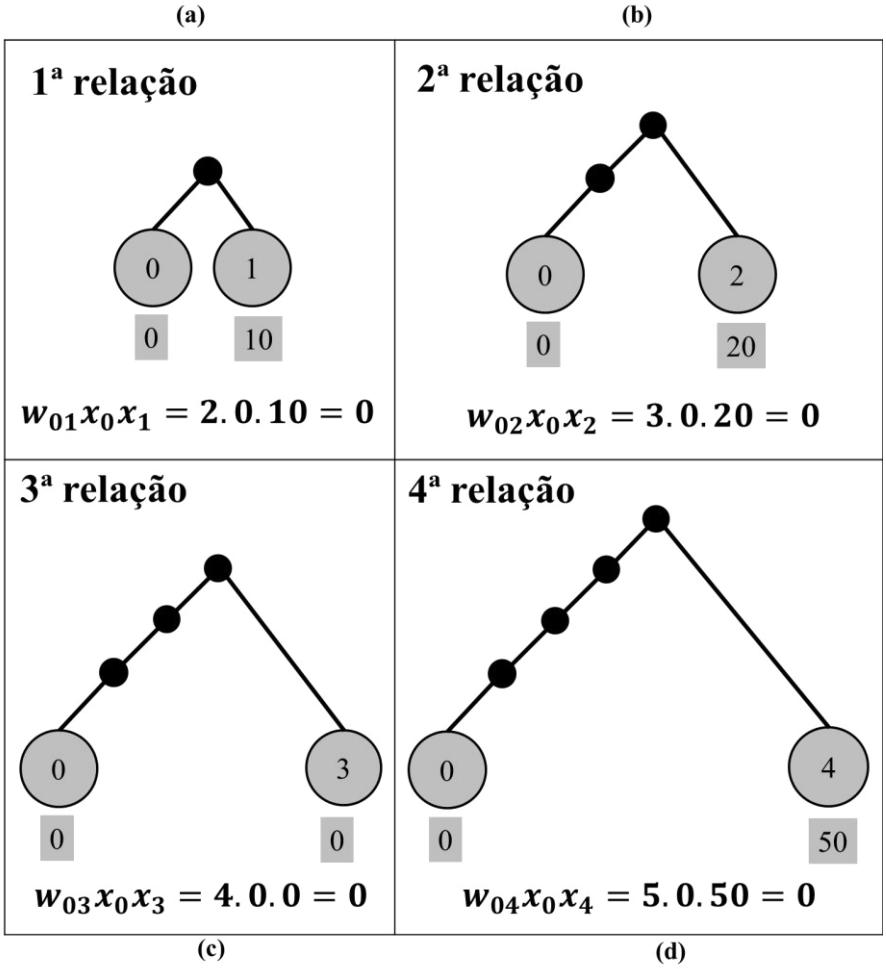


Figura 3.10 Descrição da quantidade de arestas que combina as espécies 0 (zero) com 1, 0 com 2, 0 com 3 e 0 com 4

A próxima espécie a se relacionar com as outras espécies é 1, onde a primeira relação não é feita com 0 (zero) e sim com 2 (ω_{12}= três arestas) (Figura 3.11a), como mostra na condição $i < j$ (Equação 5 e 6), ou seja, é feito a combinação de uma espécie com outras, que não seja com a já feita (ω_{10} é o mesmo que ω_{01}). Em seguida, a espécie 1 é combinado com 3 (ω_{13}= quatro arestas), como mostra na Figura 3.11b. Depois 1 é combinado com 4 através de cinco arestas (ω_{14},) (Figura 3.11c). O último relacionamento da espécie 1 é feito com a espécie 255 que possui também 256 arestas (ω_{14}).

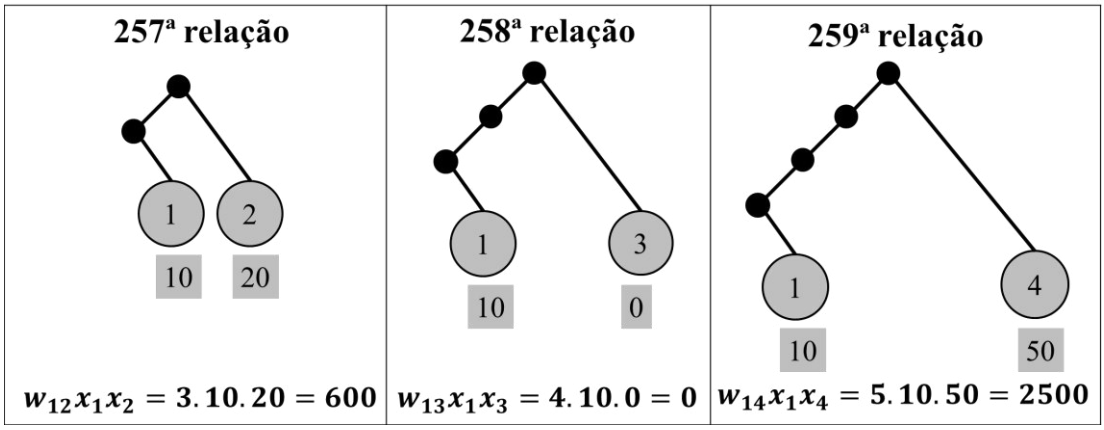


Figura 3.11 Descrição da quantidade de arestas que combina as espécies 1 com 2, 1 com 3 e 1 com 4.

O restante das combinações segue a mesma regra de não fazer relação com espécies que já foram combinadas.

3.1.3.6. Árvore 2 – Árvore Enraizada na Forma de Cladograma Inclinado Excluindo as Espécies Vazias

Seguindo a mesma lógica do cálculo dos índices com base na árvore anterior, foi desenvolvida outra arquitetura de árvore que tem como destaque a eliminação das espécies que possuem valores zero de indivíduos, resultando consequentemente na reorganização das arestas para as espécies restantes. Supondo que a árvore da Figura 3.11 tenha somente indivíduos nas espécies 1, 2 e 4, o novo modelo de árvore possui somente essas espécies e segue a mesma arquitetura, como descreve a seguir na figura:

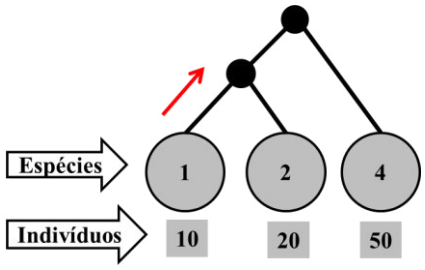
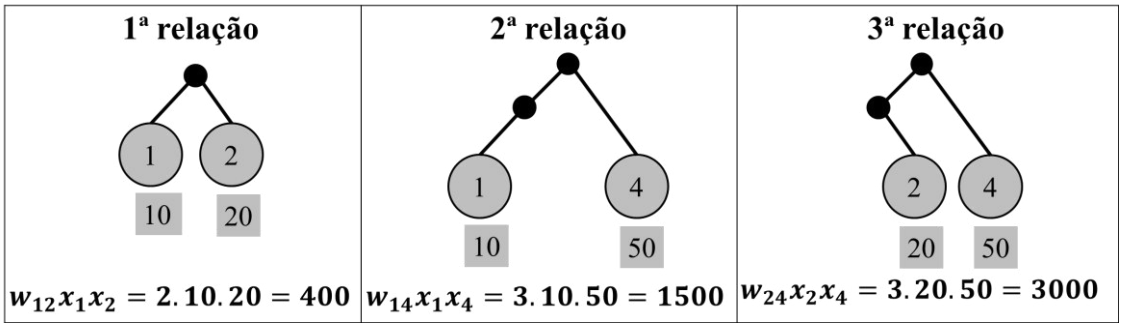


Figura 3.12 Árvore 2: modelo criado a partir da Árvore 1, com a eliminação dos níveis de cinza zero e mudança na quantidade de arestas que ligam as espécies restantes.

A combinação entre as espécies dessa árvore segue a mesma ideia da Árvore 1, como mostra na Figura 3.13.



(a) **(b)** **(c)**

Figura 3.13 Descrição da quantidade de arestas que combina as espécies 1 com 2, 1 com 4 e 2 com 4.

3.1.3.7. Árvore 3 – Árvore Enraizada na Forma de Cladograma Inclinado Modificado as Arestas

Já a terceira árvore proposta tem o mesmo processo de combinação entre as espécies da Árvore 1, onde a única diferença é na quantidade de arestas do lado direito do nó ancestral entre as espécies. Assim, a Figura 3.14 descreve o mesmo procedimento da Figura 3.10, sendo destacado em vermelho as arestas que era somente uma na Árvore 1 e a primeira combinação de uma espécie com as outras que não tem mudança (segue mesmo exemplo da Figura 3.10a e Figura 3.11a).

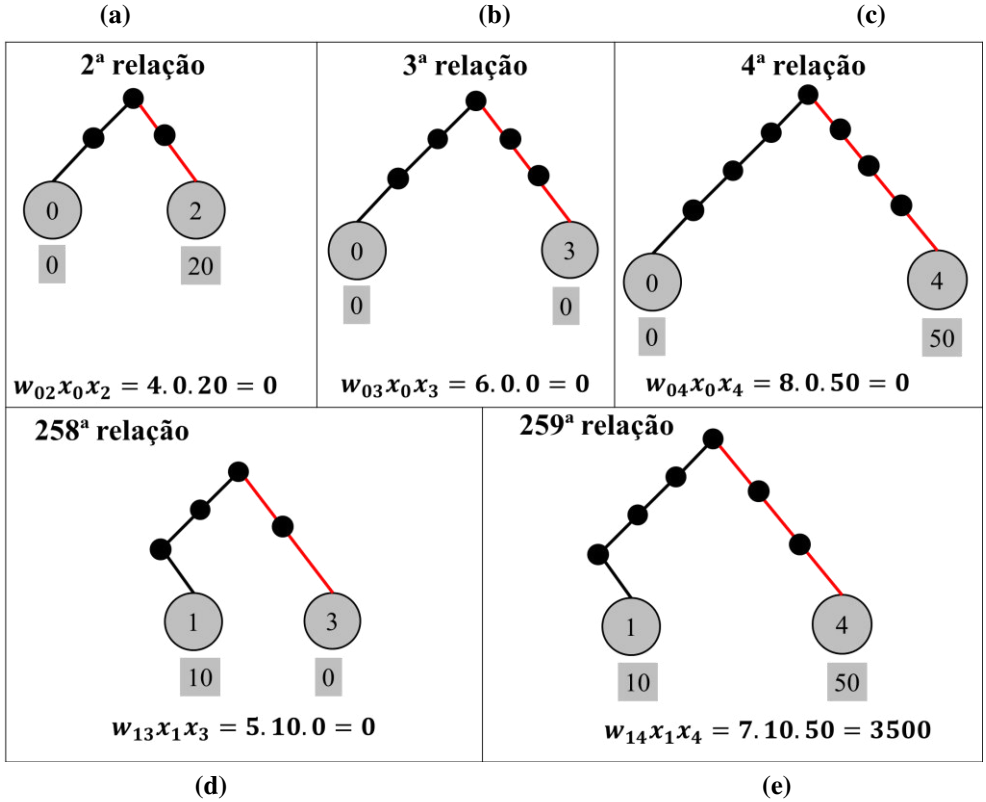


Figura 3.14 Descrição da quantidade espécies 0 com 2, 0 com 3 e 0 com 4.

3.1.4 Reconhecimento de Padrões

A última fase da metodologia consiste em classificar as amostras em *massa* e *não-massa*, utilizando um classificador (Seção 2.5.1). Os vetores de características adquiridos na fase de extração de característica são obtidos pelo cálculo dos Índices Taxonômicos a partir das árvores filogenéticas, sendo que as mesmas são calculadas considerando quatro abordagens espaciais: círculos, anéis, escalas e diferença das escalas. Esses vetores são encaminhados ao classificador supervisionado MVS.

A base de características foi formulada pelo conjunto de medidas de textura que são extraídas em cada amostra onde as mesmas possuem um rótulo a qual pertence à classe *massa* ou *não-massa*, conseguido do banco DDSM. Neste trabalho, foram geradas 324 diferentes bases de amostras (vetores de características) que representa o total de experimentos testados. Cada base foi dividida em dois grupos: base de treino e teste, com proporções de 20% e 80%, 40% e 60%, 50% e 50%, 60% e 40% e 80% e 20%, respectivamente. A divisão foi repetida aleatoriamente 5 vezes com o intuito de verificar se os acertos em todas as repetições apresentam semelhança dos valores altos e pequena diferença entre eles, assim representando que o padrão de textura discrimina bem as amostras de *massa* e *não-massa*. Todos os valores das bases que estavam no conjunto R+ (conjunto de números reais não-negativos) foram normalizados entre -1 à 1 para ajudar o classificador a convergir com maior facilidade na etapa de treinamento.

Neste trabalho foi utilizada a função de base radial (RBF) (Seção 2.5.1), que tem sido aplicada em reconhecimento de padrões. Nos trabalhos de Martins (2007), Junior (2008), Sousa (2011) e Carvalho (2012) apresenta os melhores resultados com a aplicação dessa função. Com a utilização dessa técnica foi necessário à estimação dos valores de dois parâmetros: $C \; e \; \gamma$. O parâmetro C é responsável por atribuir peso aos erros de classificação das amostras de treinamento, de modo a minimizar a ocorrência de violações do hiperplano. Já o parâmetro γ é estimado de modo que o MVS apresente a melhor eficácia para cada problema. Os valores destes parâmetros são estimados através de uma busca exaustiva realizada pelo script *grid.py*, em Phyton, pertencente ao pacote LIBSVM (CHANG & LIN, 2010) e tendo como base apenas nas amostras de treinamento.

Uma vez que as amostras foram selecionadas, foi feita a validação cruzada por *k-fold* com 5 *folds* (valor definido no script). Com isso as amostras foram divididas em 5 partes. Para

cada combinação de valores dos parâmetros, 4 partes foram utilizadas para o treinamento e 1 para a validação. Este procedimento foi repetido 5 vezes de modo que todas as partes tenham sido usadas como validação e em seguida foi calculada a média (sobre os *folds*) da taxa de acerto do conjunto de validação. Por fim, foram selecionadas as combinações de valores dos parâmetros que apresentaram as maiores taxas médias de acerto.

Na fase de treinamento é gerado um modelo que o MVS utiliza para classificar as amostras de teste. Nesta fase as amostras de teste são desconhecidas totalmente, com o desígnio de assemelhar com as condições reais de testes. Por fim, o MVS gera uma base de predições, contendo apenas os rótulos (classes) das amostras de teste depois da classificação.

3.1.5 Validação dos Resultados

Após a finalização da etapa de reconhecimento de padrões, é necessário validar os resultados e discutir prováveis melhorias. Essa metodologia usa métricas comumente empregadas em sistemas CAD/CADx e aceitas pela sociedade para a análise de desempenho de sistemas baseados em processamento de imagens. Estas métricas são sensibilidade, especificidade e acurácia (Seção 2.6). Outras métricas foram verificadas, como: desvio padrão dos cinco testes aleatórios e o *gráfico de barras com desvio padrão* para medir o desempenho entre os diferentes experimentos a partir da metodologia propostas neste trabalho.

Essas métricas tem o objetivo de medir o desempenho da metodologia como satisfatória ou não, além de ajudá-la a identificar pontos positivos e negativos para melhoria futura deste trabalho na fase de treinamento e teste.

4 RESULTADOS E DISCUSSÃO

Este capítulo mostra e discute os resultados conseguidos com a metodologia proposta por esse trabalho, obtendo assim, a classificação das regiões de interesse de imagens mamográficas consideradas em *massa* e *não-massa*. Este trabalho apresenta vários testes realizados a partir das diversas formas de combinações encontradas em duas etapas da metodologia: pré-processamento e extração de características.

Para cada fase utilizada é identificada a técnica aplicada pela sua sigla. O significado dessas siglas é: sem realce (SR), realce logarítmico (RL), realce quadrático (RQ), sem melhoramento (SM), filtro da média 3x3 (M3), filtro da média 5x5 (M5), sem quantização (SQ), 5 níveis de quantização (Q5), círculos e Anéis (CA), máscara interna e externa (IE), árvore 1 (A1), árvore 2 (A2), árvore 3 (A3), índice de diversidade taxonômica (DV), índice de distinção taxonômica (DS) e junção do índice de diversidade e distinção taxonômica (JI). A Figura 4.1 apresenta resumidamente todos os procedimentos utilizados para formação dos experimentos.

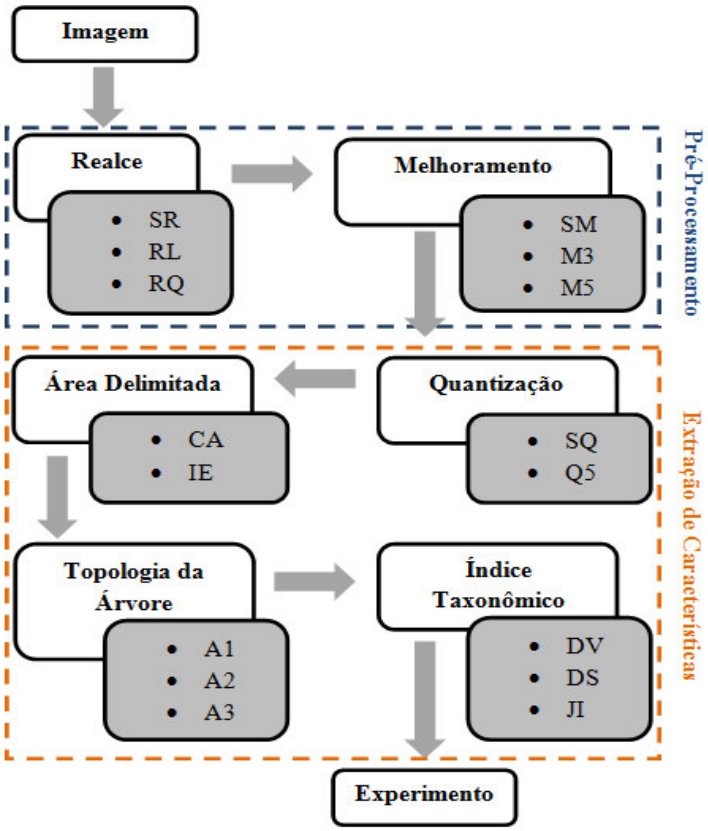


Figura 4.1 Fases utilizados para geração dos experimentos.

Foram realizadas diversas combinações de técnicas, um total de 324 experimentos, sendo: 3 processos de realce, 3 de melhoramento, 2 para a quantização, 2 para a definição das sub-regiões de interesse, 3 árvores e 3 índices.

Tendo como base a etapa de treinamento, os parâmetros C e γ (Seção 2.5.1) utilizados no núcleo radial (RBF) são estimados automaticamente para cada conjunto de amostras de treinamento, e usados durante a etapa de classificação pela MVS. Os resultados alcançados em seguida são avaliados através do processo de validação (Seção 2.6). Sendo, Acurácia Média (AM), Sensibilidade Média (SM), Especificidade Média (EM) e Desvio Padrão em relação à média da acurácia (DP). Este capítulo é encerrado com uma discussão sobre o melhor experimento conseguido com a metodologia proposta e a verificação do desempenho entre as outras pelo meio do Gráfico de Barras. O cálculo dos índices Taxonômicos foram implementados neste trabalho assim como a manipulação das imagens, com a utilização da biblioteca OpenCV (OPENCV, 2013).

4.1 Resultados Obtidos

Esta seção apresenta e discute os resultados com objetivo de avaliar a metodologia proposta. De todos os 324 experimentos realizados, 9 deles são descritos com mais detalhes sendo apresentados o pior e o melhor resultado, baseado na média das acurácias dos testes, em cada árvore abordada e índice taxonômico. Para facilitar o entendimento e a procura, os experimentos foram identificados por códigos que representam a junção de siglas descritas na Figura 4.1. A catalogação dos experimentos realizados a serem detalhados nos tópicos seguintes foi resumida na Tabela 4.1.

Tabela 4.1 Resumo dos melhores e piores resultados dos experimentos.

		MELHORES RESULTADOS					PIORES RESULTADOS				
Exp.	CÓDIGO	Tr|Te (%|%)	AM (%)	DP	SM (%)	EM (%)	Tr|Te (%|%)	AM (%)	DP	SM (%)	EM (%)
1	RL-M5-SQ-IE-A1-DV	80|20	99,17	0,59	99,02	99,33	20|80	98,42	0,68	99,08	97,78
2	RL-M5-SQ-IE-A1-DS	50|50	99,20	0,3	99,47	98,93	20|80	98,13	0,53	98,44	97,82
3	RL-M5-SQ-IE-A1-JI	40|60	99,39	0,23	99,65	99,12	50|50	98,47	0,73	99,05	97,89
4	RL-M5-SQ-IE-A2-DV	40|60	98,83	0,5	99,66	98,03	80|20	97,83	0,46	99,32	96,52
5	RL-M5-Q5-IE-A2-DS	80|20	99,33	0,7	99,65	99,93	20|80	98,88	0,35	99,42	98,34
6	RL-M5-Q5-IE-A2-JI	80|20	99,50	0,46	99,69	99,38	20|80	97,04	1,24	98,37	95,80
7	RL-M5-Q5-IE-A3-DV	80|20	99,50	0,46	100,00	99,02	20|80	97,72	2,63	98,17	97,29
8	RL-M5-Q5-IE-A3-DS	80|20	99,50	0,75	99,67	99,34	20|80	99,17	0,49	100,00	98,36
9	RL-M5-Q5-IE-A3-JI	80|20	99,67	0,75	100,00	99,25	40|60	98,44	1,05	99,20	97,71

A base de características foi dividida em dois grupos para serem utilizados no classificador MVS: base de treinamento e base de teste. Essa divisão foi repetida 5 vezes aleatoriamente dos vetores de características da base com a execução do script *subset.py* do pacote LIBSVM. As proporções para treino e teste são (Treino | Teste): (20% | 80%), (40% | 60%), (50% | 50%), (60% | 40%) e (80% | 20%). Essas proporções foram utilizadas para avaliar o desempenho da metodologia, com o aumento do treinamento e diminuição do teste. Posteriormente, foi usada a função base radial (RBF) no classificador MVS, com a estimação dos parâmetros C e γ, pela aplicação do script *grid.py* sobre as bases de treinamento. Após a estimação, o processo seguinte na metodologia proposta é a etapa de classificação e validação dos resultados.

No experimento 1, de código RL-M5-SQ-IE-A1-DV da Tabela 4.1, constituí a seguinte combinação de técnicas: realce logarítmico, filtro média 5x5, nenhuma quantização, máscara interna e externa, árvore 1, índice de diversidade taxonômico (Δ). Gera um conjunto de 9 características (9 Áreas x 1 Índice). Esse experimento obteve a melhor acurácia média de 99,17% para a proporção (80% | 20%). O desvio padrão entre os 5 casos testados foi de 0,59, significando que os valores variaram muito pouco em relação a média da acurácia. A sensibilidade média e especificidade média foram respectivamente, 99,02% e 99,33. No pior resultado para esse experimento, a configuração (20% | 80%) apresenta 0,75% abaixo da acurácia média em relação a melhor, com sensibilidade média de 0,06% acima e especificidade média de 1,55% abaixo. O desvio padrão foi de 0,68.

No experimento 2, com código RL-M5-SQ-IE-A1-DS, difere do anterior somente pela mudança do índice (distinção taxonômica) (Tabela 4.1) e o total de características também é 9. A acurácia média alçando foi de 99,20% para a configuração (50% | 50%), sendo a sensibilidade média de 99,47%, 98,93% de especificidade média e 0,3 no valor do desvio padrão para o melhor resultado. No pior caso desse experimento, apresenta acurácia média, sensibilidade média e especificidade média valores baixos em comparação ao melhor, 1,07%, 1,03% e 1,11% respectivamente, encontrado na proporção (20% | 80%). Sendo que o desvio padrão foi de 0,53.

A diferença dos experimentos anteriores para o 3, de código RL-M5-SQ-IE-A1-JI também está no índice (junção dos vetores de características com índice de diversidade e distinção taxonômica), mostrado na Tabela 4.1. O vetor de característica desse experimento tem o total de 18 (9 Áreas x 2 Índices). A melhor taxa média de acertos foi de 99,39%,

sensibilidade média de 99,65%, especificidade média de 99,12% e com variação em relação à média da acurácia de 0,23, conseguido na configuração (40% | 60%). O pior resultado desse experimento foi à diminuição de 0,92%, 0,6% e 1,23% para média acurácia, sensibilidade média e especificidade média na proporção (50% | 50%), respectivamente com base no melhor. O desvio padrão foi de 0,73.

No código RL-M5-SQ-IE-A2-DV (Experimento 4) mostrado na Tabela 4.1, utiliza a árvore 2. Esse experimento totaliza um conjunto de 9 características (9 Áreas x 1 Índice). Sua melhor média de acurácia alcança 98,83%, sensibilidade média de 99,66%, especificidade média de 98,03% e apresenta desvio padrão de 0,5, para proporção (40% | 60%). E no pior resultado foi constatado menor valor de 1%, 0,34% e 1,51% para acurácia média, sensibilidade média e especificidade média encontrada em (80% | 20%), respectivamente, em relação ao melhor. Seu desvio padrão foi de 0,46.

A combinação de técnicas do código RL-M5-Q5-IE-A2-DS (Experimento 5) exibido na Tabela 4.1, constitui no uso dos 5 níveis de quantização e do índice de distinção taxonômica, totalizando um conjunto de 45 características (5 Quantizações x 9 Áreas x 1 Índice). O melhor resultado nesse experimento foi de 99,33% de acurácia média, 99,65% de sensibilidade média, 99,93% de especificidade média e 0,7 de desvio padrão, na configuração (80% | 20%). Em seu menor resultado atingido, obteve-se uma diminuição de 0,45%, 0,23%, 1,59% para média da acurácia, sensibilidade média e especificidade média, respectivamente, encontrada em (20% | 80%). Nesse caso o desvio foi de 0,35.

O experimento 6, com código RL-M5-Q5-IE-A2-JI (Tabela 4.1), apresenta um vetor de 90 características (5 Quantizações x 9 Áreas x 2 Índice). A melhor média de acurácia atingido foi de 99,50%, sensibilidade média de 99,69%, especificidade média de 99,38% e desvio padrão de 0,46, na configuração (80% | 20%), nesse experimento. Já o pior resultado desse experimento comparado com o melhor, apresenta diminuição de 2,46%, 1,32% e 3,58%, respectivamente para a taxa média de acertos, sensibilidade média e especificidade média, adquiridos em (20% | 80%). E o desvio padrão constatado foi de 1,24.

No experimento 7, de código RL-M5-Q5-IE-A3-DV mostrado na Tabela 4.1, foi testado com árvore 3. O vetor de características nesse experimento é formado por 45 características (5 Quantizações x 9 Áreas x 1 Índice). Seu maior valor alcançado na média da acurácia foi de 99,50%, 100% de sensibilidade média, 99,02% de especificidade média e

apresentando desvio padrão de 0,46, para a configuração (80% | 20%). E no pior resultado conseguido, constatou-se uma queda de 1,78%, 1,83% e 1,73% respectivamente para a taxa média de acertos, sensibilidade média e especificidade média, em relação ao maior valor, encontrado em (20% | 80%). E o desvio padrão foi de 2,63.

Com a mudança no índice para distinção taxonômica do experimento anterior, o experimento 8, de código RL-M5-Q5-IE-A3-DS (Tabela 4.1) apresenta também um conjunto 45 características (5 Quantizações x 9 Áreas x 1 Índice). O melhor resultado nesse experimento foi de 99,50% para acurácia média, 99,67% de sensibilidade média, 99,34% de especificidade média e desvio padrão de 0,75, para (80% | 20%). E no pior resultado, verificou-se uma queda de 0,33% para acurácia média, aumento de 0,33% na sensibilidade média e diminuição de 0,98% na especificidade média, baseado no melhor resultado, sendo encontrado em (20% | 80%). E o desvio padrão foi de 0,49.

E o experimento 9 da última linha da Tabela 4.1 que consta o código RL-M5-Q5-IE-A3-JI, apresenta na sua combinação de técnicas utilizadas, a junção dos vetores de características com índice de diversidade e distinção taxonômica. Esse experimento tem um total de 90 características (5 Quantizações x 9 Áreas x 2 Índice). A melhor média de acurácia atingido foi de 99,67%, sensibilidade média de 100%, especificidade média de 99,25% e desvio padrão de 0,75, na configuração (80% | 20%), nesse experimento. Já o pior resultado desse experimento comparado com o melhor, apresenta diminuição de 1,23%, 0,8% e 1,54%, respectivamente para a taxa média de acertos, sensibilidade média e especificidade média, adquiridos em (40% | 60%). E o desvio padrão verificado foi de 1,05.

4.2 Discussão dos Resultados

Esta seção tem por objetivo discutir os principais resultados dos experimentos utilizados que compõem as várias combinações de técnicas. Nos experimentos descritos na Tabela 4.1 foi constatado que os melhores resultados apresentam valores acima de 98% na taxa média de acertos, 99% para sensibilidade média e 98% para especificidade média. Sendo que o maior valor alçando entre todos os experimento foi de 99,67% para a média da acurácia, que constituí pelas seguintes combinações de técnicas (código RL-M5-Q5-IE-A3-JI): realce logarítmico, filtro média 5x5, 5 níveis de quantização, máscara interna e externa, árvore 3, índice de diversidade(Δ) e distinção (Δ^*) taxonômica.

Um destaque verificado na Tabela 4.1 é a presença das técnicas de realce logarítmico, filtro da média 5x5 e as máscaras interna e externa, em todos os experimentos. A justificativa para a primeira técnica está no aumento dos valores de cinza baixos que permite deixar a textura das regiões de *não-massa* mais homogêneas do que as com *massa*. Na segunda, tem a maior eliminação de ruído. Finalmente, com a terceira, não é perdida nenhuma informação da imagem.

A árvore 3 apresentou valores maiores com sua aplicação devido a maior diversidade entre as espécies, ou seja, as espécies estão mais distintas filogeneticamente (muitos níveis hierárquicos na árvore) (Seção 3.1.3.7).

Como o melhor resultado alcançado (Experimento 9) foi a junção dos dois índice (Δ e Δ*), é preciso dar ênfase ao índice de distinção taxonômica (Δ*), por obter-se a acurácia média muito próxima (99,50%), em relação a anterior, por meio de um vetor de 45 características contra 90.

A Figura 4.2 demostra visualmente a avaliação entre os experimentos da Tabela 4.1. Os melhores resultados são representados por "M" seguido por um número que pertence ao experimento e os piores por "P". Nessa avaliação pode-se verificar que M6, M7, M8 e M9 apresentaram os melhores resultados com valores de suas médias de acurácia ≥ 99,5% (pontos em vermelho) e nos desvios padrões não passaram de ± 1 (barras verticais acima e abaixo dos pontos em vermelhos). O M3 merece ser destacado por apresentar o menor desvio padrão em comparação com os outros experimentos e está entre 99% e 99,5% na média da acurácia.

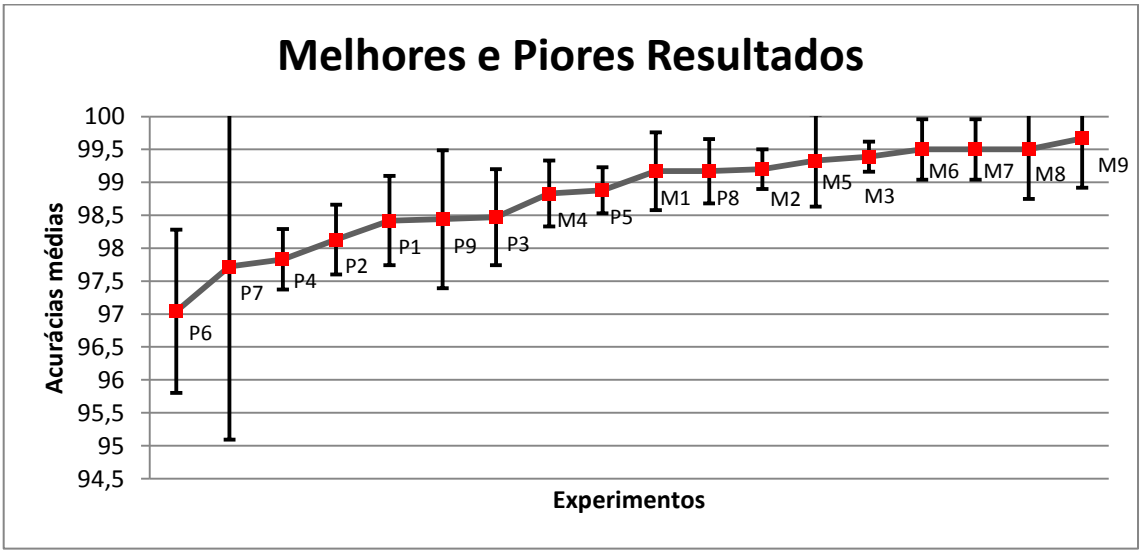


Figura 4.2 Gráfico de barra com desvio padrão dos melhores e piores resultados.

4.2.1 Comparação com outros trabalhos relacionados

A Tabela 4.2 apresenta uma breve comparação entre os resultados encontrados neste trabalho e alguns trabalhos citados na Seção 1.1, que realizam a classificação de regiões extraídas de mamografias em *massa* e *não-massa*.

Tabela 4.2 Comparação com alguns trabalhos referentes à classificação de tecidos extraídos de mamografias em *massa* e *não-massa*.

Trabalhos	Base de Dados	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
(SARFRAZ *et al.*, 2012)	MIAS	93,06	__	__
(BERBAR *et al.*, 2012)	DDSM	98,63	__	__
(NITHYA *et al.*, 2012)	DDSM	98,00	__	__
(NITHYA *et al.*, 2011)	DDSM	96,00	__	__
(JASMINE *et al.*, 2011)	MIAS	98,61	__	__
(SHANTHI *et al.*, 2012)	MIAS	92,06	__	__
(MEENALOSINI, *et al.*, 2012)	MIAS e DDSM	__	95,20	94,40
(SOUSA, 2011)	DDSM	99,71	99,71	99,71
(NUNES *et al.*, 2010)	DDSM	83,94	83,24	84,14
(JUNIOR *et al.*, 2009)	DDSM	99,39	100,00	98,94
(COSTA *et al.*, 2011)	DDSM	90,07	__	__
(CARVALHO, 2012)	DDSM	99,75	99,47	100,00
Abordagem proposta (Δ com Δ*)	**DDSM**	**99,67**	**100,00**	**99,25**
Abordagem proposta (Δ)	**DDSM**	**99,50**	**100,00**	**99,02**
Abordagem proposta (Δ*)	**DDSM**	**99,50**	**99,67**	**99,34**

Os resultados na abordagem proposta com a junção de Δ com Δ*, como pode ser analisada na Tabela 4.2, obteve uma pequena melhoria em comparação com os de (Junior *et al.*, 2009), alcançando uma acurácia média de 99,67% e especificidade média de 99,25%, enquanto o outro trabalho obteve 99,39% de acurácia e 98,94% de especificidade. Outra vantagem dessa abordagem em relação ao de (Junior *et al.*, 2009) é a respeito da menor quantidade de características, sendo 90 contra 240. Então, a metodologia proposta (Tabela 4.2) alcançou uma acurácia comparável com os melhores resultados encontrados na literatura recente para classificação de tecidos de mamografia nas classes *massa* e *não-massa*.

5 CONCLUSÃO

Este trabalho apresentou uma boa funcionalidade com o uso dos Índices de Diversidade (Δ) e Distinção (Δ*) Taxonômicos, junto com a Máquina de Vetores de Suporte para discriminação e classificação de regiões mamográficas em *massa* e *não-massa*.

A metodologia foi aplicada em um total de 324 experimentos, sendo constituídos pela combinação de técnicas na etapa de pré-processamento e extração de características mostradas na Figura 4.1. Depois, comparou os resultados obtidos na classificação com a MVS, utilizando 5 conjuntos diferentes de distribuição de amostras para treinamento e teste.

Os resultados obtidos demonstraram o desempenho próspero das técnicas de extração de textura pelos índices taxonômicos em união com MVS. Os melhores resultados na avaliação da acurácia, sensibilidade e especificidade foram com aplicação da técnica de real logarítmico, filtro da média 5x5, 5 níveis de quantização, área delimitada com as máscaras internas e externas, topologia de árvore tipo 3 e a junção dos índices Δ e Δ*. Sendo que a utilização das máscaras para delimitar uma área da região de interesse das imagens de mamografia, contribui para o aproveitamento de todas as informações da imagem, o que demonstra superioridade nos resultados em comparação com as áreas de círculos com anéis. Outro fator importante encontrado nos resultados foi a criação do modelo de árvore tipo 3 (Seção 3.1.3.7), pois constatou que quanto mais níveis hierárquicos e distantes (ou próximas) as espécies estiverem em uma árvore filogenética, maior diversidade encontrado em uma comunidade. Ou seja, a utilização dessa árvore teve melhor contribuição na discriminação ente *massa* e *não-massa*.

Nos experimentos com as mesmas combinações de técnicas do melhor resultado, com exceção da mudança do índice, apresentaram resultados muito semelhantes, com uma diferença inferior nos resultados insignificativa, tanto com o Índice de Diversidade Taxonômica, como para o de Distinção Taxonômica.

Apesar dos resultados obtidos da análise das árvores filogenéticas e dos índices taxonômicos serem promissores, se faz necessário aumentar a quantidade e variabilidade das amostras de mamografia para alcançar uma metodologia robusta e genérica. Entretanto, considerando-se o tamanho e o reconhecimento da base de mamografias utilizada neste

trabalho pelo meio acadêmico, pode-se concluir que os resultados obtidos indicam que novas abordagens baseadas em Índice Taxonômico (Δ e Δ*) para descrição de texturas possam ser ampliadas.

Este trabalho pode ser melhorado futuramente. Uma primeira extensão é a extração características de textura pelos índices taxonômicos, onde as espécies seriam pares de pixels adquiridos pela Matriz de Co-ocorrência de Níveis de Cinza (GLCM). Outra situação seria ampliar os resultados da classificação de *massa* e *não-massa* para também classificação de massas malignas e benignas, por exemplo, da análise de uma árvore filogenética que melhor descreva a discriminação dessas classes. Sugerimos, também, o uso do PCA (Principal Component Analysis) para reduzir características, visando alcançar melhores resultados.

O classificador MVS utilizado neste trabalho pode ser substituído por outros classificadores com o objetivo de avaliar seu desempenho na tarefa de reconhecimento de padrões de *massa* e *não-massa* em regiões extraídas de mamografias.

Por fim, a metodologia apresentada neste trabalho poderá integrar uma ferramenta CADx a ser aplicada em casos reais e atuais na detecção e tratamento de câncer de mama. Essa metodologia poderá fazer parte de um sistema CAD no intuito de classificar regiões detectadas como suspeitas em *massa* e *não-massa*.

6 REFERÊNCIAS

ACS - American Cancer Society. **Breast Cancer**, 2013a. Disponível em: <http://www.cancer.org/cancer/cancerbasics/what-is-cancer>. Acesso em 17 de Jan. de 2013.

ACS - American Cancer Society. **Breast biopsy**, 2013b. Disponível em: <http://www.cancer.org/treatment/understandingyourdiagnosis/examsandtestdescriptions/mammogramsandotherbreastimagingprocedures/mammograms-and-other-breast-imaging-procedures-breast-bx>. Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Breast Cancer**, 2013c. Disponível em: <http://www.cancer.org/cancer/breastcancerinmen/detailedguide/breast-cancer-in-men-key-statistics>. Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Breast Cancer**: Causes, Risk Factors, and Prevention Topics, 2013d. Disponível em: <http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-risk-factors>. Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Breast Cancer**: Early Detection, 2013e. Disponível em: <http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-risk-factors>. Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Breast Cancer**: Early Detection - the importance of finding breast cancer early, 2013f. Disponível em: <http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-importance-of-finding-early>. Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Magnetic resonance imaging**, 2013g. Disponível em: <http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-a-c-s-recs-m-r-i> Acesso em 15 de Jan. de 2013.

ACS - American Cancer Society. **Breast Cancer**: Early Detection - Mammograms, 2013h. Disponível em: <http://www.cancer.org/cancer/breastcancer/moreinformation/breastcancerearlydetection/breast-cancer-early-detection-acs-recs-mammograms>. Acesso em 15 de Jan. de 2013.

ARAÚJO, G. S. **Filogenia de proteomas**. 2003, 73 f. Dissertação (Ciência da Computação) - Universidade Federal do Mato Grosso do Sul, MS, 2003.

BERBAR, M. A.; REYAD, Y. A.; HUSSAIN, M. Breast Mass Classification using Statistical and Local Binary Pattern Features. In: **16th International Conference on Information Visualisation**, IV, 2012, Montpellier, France, v. 4, pp. 486 - 490, Jul. 11-13, 2012. Disponível em: <http://dblp.uni-trier.de/db/conf/iv/iv2012.html>. Acesso em 29 de Nov. de 2012.

BLAND, M. **An introduction to medical statistics**. Oxford: Oxford University Press, 2000.

CARVALHO, P. M. **Classificação de tecidos da mama a partir de imagens mamográficas em massa e não massa usando índice de diversidade de mcintosh e máquina de vetores de suporte**. 2012, 75 f. Dissertação (Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís, MA, 2012.

CARVALHO, P. M.; PAIVA, A. C.; SILVA, A. C. Classification of breast tissues in mammographic images in mass and non-mass using mcintosh's diversity index and svm. In: **Machine Learning and Data Mining in Pattern Recognition - 8th International Conference**, MLDM 2012, Berlin, Germany, v. 7376, pp. 482-494, Jul. 13-20, 2012. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/conf/mldm/mldm2012.html>. Acesso em: 10 de Dez. de 2012.

CHANG, C. C.; LIN, C. J. **LIBSVM:** A library for support vector machines. Disponível em: <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>. Acesso em 19 de janeiro de 2012.

CHAVES, A. C. F. **Extração de Regras Fuzzy para Máquinas de Vetor de Suporte (SVM) para Classificação em Múltiplas Classes**. 2006. PhD thesis - Pontifícia Universidade Católica do Rio de Janeiro, 2006.

CLARKE, K. R.; WARWICK, R. M. A taxonomic distinctness index and its statistical properties. In: **Journal of Applied Ecology**,vol. 35, n. 4, pp. 523-531, Aug. 1998. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2664.1998.3540523.x/abstract>. Acesso em 1 de Jul. de 2012.

CONCI, A.; AZEVEDO, E.; LETA, F. R. **Computação Gráfica***: Teoria e Prática*. vol. 2. Rio de Janeiro: Campus, 2008.

CORTES, C.; VAPNIK, V. Support-vector network. In: **Machine Learning**, v. 20, n. 3, pp. 273-297, 1995. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/journals/ml/ml20.html>. Acesso em 26 de Jun. 2012.

COSTA, D. D.; CAMPOS, L. F.; BARROS, A. K. Classification of breast tissue in mammograms using efficient coding. In: **BioMedical Engineering Online**, v. 10, n. 1, pp. 1-

14, Jun. 2011. Diponível em: <http://www.biomedical-engineering-online.com/content/10/1/55>. Acesso em 13 de Nov. 2012.

CRISTIANNI, N.; SHAWE-TAYLOR, J. **An introduction to support vector machines and other kernel-based learning methods**. New York: Cambridge University Press, 2000.

DDSM. **Digital database for screening mammography**, 2013a. Disponível em: <http://marathon.csee.usf.edu/Mammography/DDSM/thumbnails/cancers/cancer_06/case1187/A-1187-1.html>. Acesso em 13 de Jan. de 2013.

DDSM. **Digital database for screening mammography**, 2013b. Disponível em: < http://marathon.csee.usf.edu/Mammography/DDSM/thumbnails/normals/normal_07/case4555/D-4555-1.html>. Acesso em 13 de Jan. de 2013.

DDSM. **University of South Florida Digital Mammography**, 2013c. Disponível em: <http://marathon.csee.usf.edu/Mammography/Database.html>. Acesso em 13 de Jan. de 2013.
FILHO, O. M.; NETO, H. V. **Processamento Digital de Imagens**. Rio de Janeiro: Brasport, 1999.

DDSM. **Digital database for screening mammography**, 2013d. Disponível em: <http://marathon.csee.usf.edu/Mammography/DDSM/thumbnails/cancers/cancer_04/case1010/A-1010-1.html>. Acesso em 13 de Jan. de 2013.

DDSM. **Digital database for screening mammography**, 2013e. Disponível em: <http://marathon.csee.usf.edu/Mammography/DDSM/thumbnails/normals/normal_01/case0009/A-0009-1.html>. Acesso em 13 de Jan. de 2013.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. Tradução de Roberto Marcondes Cesar Junior e Luciano da Fontoura Costa. São Paulo: Edgard Blucher ltda, 2000.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 2nd ed. Massachusetts: Addison Wesley, 1992.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 2nd ed. New Jersey: Prentice Hall, 2002.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 3nd ed. New Jersey: Prentice Hall, 2008.

GORENSTEIN, M. R. **Diversidade de espécies em comunidades arbóreas: aplicação de índices de distinção taxonômica em três formações florestais do estado de são paulo**. Dissertação (Recursos Naturais) - Universidade de São Paulo, Piracicaba, SP, 2009.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEIN, I. Textural Features for Image Classification. In: **IEEE Transactions on Systems, Man, and Cybernetics**, vo. 3, n. 6, pp. 610-621, Nov. 1976. Diponível em : <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4309314&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4309314>. Acesso em 20 de Fev. de 2012.

HAYKIN S. **Redes Neurais:** *Princípios e prática*. 2nd ed. Porto Alegre: Bookman, 2001.

HEATH, M.; BOWYER, K.; KOPANS, D.; MOORE, R.; KEGELMEYER, W. P. The Digital Database for Screening Mammography. In:Yaff Mj, Ed. **Proceedings of the Fifth International Workshop on Digital Mammography**. Wiscosin: Medical Physics Publishing, pp. 212-218, Dec. 2001.

INCA - Instituto Nacional de Câncer. **O que é câncer?**, 2013a. Disponível em: <http://www1.inca.gov.br/conteudo_view.asp?id=322>. Acesso em 17 de Jan. de 2013.

INCA - Instituto Nacional do Câncer. **Estimativas 2012: Introdução**, 2013b. Disponível em: <http://www.inca.gov.br/estimativa/2012/index.asp?ID=2>. Acesso em 15 de Jan de 2013.

INCA - Instituto Nacional do Câncer. **Estimativas 2012: Síntese de Resultados e Comentários**, 2013c. Disponível em: <http://www.inca.gov.br/estimativa/2012/index.asp?ID=5>. Acesso em 15 de Jan de 2013.

INCA - Instituto Nacional do Câncer. **Atlas de Mortalidade por Câncer**, 2013d. Diponível em: <http://mortalidade.inca.gov.br/Mortalidade/prepararModelo07.action>. Aceso em 24 de Fev. de 2013.

INCA - Instituto Nacional de Câncer. **Detecção Precoce**, 2013d. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/deteccao_precoce>. Acesso em 15 de Jan. de 2013.

INCA, Prevenção e controle do câncer: normas e recomendações do INCA. **Revista Brasileira de Cancerologia**, Rio de Janeiro, v. 48, n. 3, pp. 317-332, jul./set. 2002.

JASMINE, L. J.; BASKARAN, S.; GOVARDHAN, A. An automated mass classification system in digital mammograms using contourlet transform and support vector machine. In:

International Journal of Computer Applications, New York, USA, v. 31, n. 9, pp. 54-61, Oct. 2011. Diponível em: <http://www.ijcaonline.org/archives/volume31/number9/3857-5375>. Acesso em 29 de Nov. de 2012.

JOHNSON, N. L.; KOTZ, S. **Encyclopedia of statistical science**. New York: John Wiley, 1988.

JUNIOR, G. B. **Classificação de Regiões de Mamografias em Massa e Não Massa usando Estatística Espacial e Máquina de Vetores de Suporte**. 2008. 99 f. Dissertação (Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luís, MA, 2008.

JUNIOR, G. B., PAIVA, A. C.; SILVA, A. C.; OLIVEIRA, A. C. M. Classification of breast tissues using moran's index and geary's coefficient as texture signatures and svm. In: **Computers in Biology and Medicine**, New York, USA, v. 39, n. 12, pp. 1063-1072, Dec. 2009. Diponível em: <http://dl.acm.org/citation.cfm?id=1655433>. Acesso em 13 de Nov. de 2012.

KOPANS, D. B. **Imagem da Mama**.Porto Alegre: MEDSI, 2000.

LOONEY, C. G. **Pattern Recognition using Neural Networks**: Theory and Algorithms for Engineers and Scientists. New York: Oxford University Press, 1997.

MAGURRAN, A. E. **Measuring Biological Diversity**. Oxford: Blackwell Science Ltd, 2004.

MARTINS, L. O. **Detecção de massas de imagens mamográficas através do algoritmo growing neural gas e da função k de ripley**. 2007. 106 f. Dissertação (Engenharia de Eletricidade) - Universidade Federal do Maranhão, São Luis, MA, 2007.

MEENALOSINI, S.; JANET, J. Computer Aided Diagnosis of Malignancy in Mammograms. In: **European Journal of Scientific Research**. v. 72, n. 3, pp. 360-368, Mar. 2012. Diponível em: <http://www.europeanjournalofscientificresearch.com/ISSUES/EJSR_72_3.htm>. Acesso em 29 de Nov. de 2012.

MOAYEDI, F.; AZIMIFAR, Z.; BOOSTANI, R.; KATEBI, S. Contourlet-based mammography mass classification using the SVM family. In: **Computers in Biology and Medicine**, New York, USA, v. 40, n. 4, pp. 373-383, Apr. 2010. Diponível em: <http://www.sciencedirect.com/science/article/pii/S0010482510000041>. Acesso em 13 de Nov. de 2012.

MORSE, B. S. Data Structures for Image Analysis. In: **Brigham Young University**, 2000. Diponível em: <http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/data-structures.pdf>. Acesso em 13 de jan. de 2012.

NASTRI, C. O.; MARTINS, W. P.; LENHARTE, R. J. Ultrassonografia no rastreamento do câncer de mama. In:**Femina**, Rio de Janeiro, Brasil, v. 39, pp. 97-102, Fev. 2011. Diponível em: < http://www.febrasgo.org.br/arquivos/femina/Femina2011/fevereiro/Femina_v39n2_97-102.pdf>. Acesso em 13 Jan. de 2013.

NITHYA, R.; SANTHI, B. Classification of Normal and Abnormal Patterns in Digital Mammograms for Diagnosis of Breast Cancer. In: **International Journal of Computer Applications**, New York, USA, v. 28, n. 6, pp. 21-25, Aug. 2011. Diponível em: <http://www.ijcaonline.org/archives/voulme28/number6/3391-4707>. Acesso em 29 Nov. de 2012.

NITHYA, R.; SANTHI, B. Mammogram Analysis Based on Pixel Intensity Mean Features. In: **Journal of Computer Science**, v. 8, n. 3, pp. 329-332, Jan. 2012. Diponível em: <http://thescipub.com/issue-jcs/8/3>. Acesso em 29 Nov. de 2012.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. Detection of Masses in Mammographic Images using Geometry, Simpson's Diversity Index and SVM. In: **International Journal of Signal and Imaging Systems Engineering**, v. 3, n. 1, pp. 40-51, Aug. 2010. Diponível em: <http://www.inderscience.com/info/inarticletoc.php?jcode=ijsise&year=2010&vol=3&issue=1>. Acesso em 13 Nov. de 2012.

OPENCV - Open Source Computer Vision. **Open Source Computer Vision Library**, 2013. Disponível em: <http://opencv.org/>. Acesso em 13 jan. de 2013.

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

PIANKA, E. R. **Evolutionary Ecology**. New York: HarperCollins, 1994.

RADIOLOGYINFO. **What does the equipment look like?**. Disponível em: <http://www.radiologyinfo.org/en/info.cfm?pg=mammo>. Acesso em 13 de jan. de 2013.

RICOTTA, C. A parametric diversity measure combining the relative abundances and taxonomic. In: **A Journal of Conservation Biogeography**, vol. 10, n. 2, pp. 143-146, Feb. 2004. Diponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1366-9516.2004.00069.x/abstract>. Acesso em 11 Dez. de 2012.

SAMPAIO, W. B. **Detecção de massas em imagens mamográficas usando redes neurais celulares, funções geoestatísticas e máquina de vetores de suporte**. 2009, 103 f. Dissertação (Engenharia Elétrica) - Universidade Federal do Maranão, São Luis, MA, 2009.

SANTOS, E. M. **Teoria e aplicação de support vector machines à aprendizagem e reconhecimento de objetos baseado na aparência**. 2002. 111 f. Dissertação (Informática). Universidade Federal da Paraíba, Campina Grande, PA, 2002.

SANTOS, V. K. **Uma generalização da distribuição do índice de diversidade generalizado por good com aplicação em ciências agrárias**. 2009. 56 f. Dissertação (Biometria e Estatística Aplicada) - Universidade Federal de Pernambuco, Recife, PE, 2009.

SARFRAZ, M.; ABU-AMARA, F.; ABDEL-QADER, I. A computer aided detection framework for mammographic images using fisher linear discriminant and nearest neighbor classifier. In: **J. Biomedical Science and Engineering**, v. 5, n. 6, pp. 323-329, Jun. 2012. Diponível em: <http://www.scirp.org/journal/PaperInformation.aspx?paperID=19717>. Acesso em 21 Nov. de 2012.

SHANTHI, S.; BHASKARAN, M. V. Computer aided detection and classification of mammogram using self-adaptive resource allocation network classifier. In: **Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering**, Tamilnadu, India, pp. 284-289, Mar. 2012. Diponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6208359>. Acesso em 29 Nov. de 2012.

SILVA, I. A.; BATALHA, M. A. Taxonomic distinctness and diversity of a hyperseasonal savanna in central Brazil. In: **A Journal of Conservation Biogeography**, v. 12, n. 6, pp. 725-730, Nov. 2006. Diponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1472-4642.2006.00264.x/abstract>. Acesso em 10 Nov. de 2012.

SOUSA, U. S. **Classificação de massas na mama a partir de imagens mamográficas usando o índice de diversidade de shannon-wiener**. 2011. 69 f. Dissertação (Engenharia Elétrica) - Universidade Federal do Maranão, São Luis, MA, 2011.

SUCKLING J., PARKER J., DANCE D. R., ASTLEY S., HUTT I., BOGGIS C. R. M., RICKETTS I., STAMAKIS E., CERNEAZ N., KOK S-L., TAYLOR P., BETAL D., SAVAGE J. The Mammographic Image Analysis Society digital mammogram database. In: **Proceedings of the 2nd International Workshop on Digital Mammography**, v. 1069, pp. 375–378, York: Elsevier, 1994.

VANDAMME, P.; POT, B.; GILLIS, M.; DE VOS, P.; KERSTERS, K.; SWINGS, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. In: **Microbiology and**

Molecular Biology Reviews, v. 60, n. 2, pp. 407-438, Jun. 1996. Diponível em: <http://mmbr.asm.org/content/60/2/407.abstract>. Acesso em 2 Nov. de 2012.

VAPNIK, V. N. **Statistical learning theory**. New York: Wiley, 1998.

VIANA, G. V. **Técnicas para construção de árvores filogenéticas**. 2007. 176 f. Dissertação (Ciência da Computação) - Universidade Federal do Ceará, Fortaleza, CE, 2007.

WARWICK, R. M.; CLARKE, K. R. New "biodiversity" measures reveal a decrease in taxonomic distinctness with increasing stress. In: **Marine Ecology Progress Series**, v. 129, pp. 301-305, Dec. 1995. Diponível em: <http://www.int-res.com/abstracts/meps/v129/> Acesso em 1 de Jul. de 2012.