UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE ELETRICIDADE

ULYSSES SANTOS SOUSA

CLASSIFICAÇÃO DE MASSAS NA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SHANNON-WIENER

São Luís-MA
2011

ULYSSES SANTOS SOUSA

CLASSIFICAÇÃO DE MASSAS NA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SHANNON-WIENER

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão, como parte dos requisitos 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
2011

Sousa, Ulysses Santos.

Classificação de massas na mama a partir de imagens mamográficas usando índice de diversidade de shannon-wiener / Ulysses Santos Sousa – São Luís, 2011.

69 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, 2011.

1. Classificação de tecidos de mama – Mamografia – Índice de Diversidade de Shannon-Wiener. 2. Máquina de vetores de suporte. I. Título.

CDU 621.386.84:618.19-073

CLASSIFICAÇÃO DE MASSAS NA MAMA A PARTIR DE IMAGENS MAMOGRÁFICAS USANDO ÍNDICE DE DIVERSIDADE DE SHANNON-WIENER

Ulysses Santos Sousa

Dissertação aprovada em 13 de maio de 2011.

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

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

Prof. Perfilino Eugênio Ferreira Junior, Dr.
(Membro da Banca Examinadora)

Prof. Zair Abdelouahab, Ph.D.
(Membro da Banca Examinadora)

À minha esposa, aos meus pais e à minha irmã.

AGRADECIMENTOS

A Deus, pela Sua misericórdia, amor, graça e capacitação.

À minha esposa, Jerlanne, pelo companheirismo, força e paciência em todos os momentos.

Aos meus pais, Edmilson e Ana Lúcia, pela educação, amor e carinho durante todas as fases da minha vida.

À minha irmã, Luciana, por todo apoio e incentivo durante o mestrado.

Aos meus orientadores, Dr. Anselmo Cardoso de Paiva e Dr. Aristófanes Correa Silva, pela confiança, cobrança, paciência, competência e dedicação.

Ao professor Msc. Geraldo Braz Junior pela força, incentivo e dedicação que foram de grande importância para a concretização deste trabalho.

Aos meus pastores e amigos, Ronaldo e Derlange, pelo carinho, força e orações.

Ao meu amigo e irmão Jerfson por todo apoio.

Aos meus familiares que são muito importantes na minha vida.

Aos meus colegas de trabalho do Instituto Federal de Educação, Ciência e Tecnologia – Campus Buriticupu pelo apoio e motivação.

Aos meus companheiros de mestrado Alex Martins Santos e Edgar Moraes Diniz pelos momentos de ajuda e transferência mútua de conhecimento.

À CAPES pelo suporte de financeiro durante 10 meses no mestrado.

A todos que, de forma direta ou indireta, contribuíram para a elaboração deste trabalho.

“Porque o SENHOR dá a sabedoria;
da sua boca é que vem o conhecimento e o entendimento.”

(Provérbios 2:6)

RESUMO

O câncer é um dos maiores problemas de saúde mundial, sendo o câncer de mama o que mais causa óbito entre as mulheres e o segundo tipo mais freqüente no mundo. As chances de uma paciente sobreviver ao câncer de mama aumentam à medida que a doença é descoberta mais cedo. Diversos Sistemas de Detecção e Diagnóstico auxiliados por computador (*Computer Aided Detection/Diagnosis*) têm sido utilizados para auxiliar profissionais de saúde. Este trabalho apresenta uma metodologia de discriminação e classificação de regiões de tecidos de mamografias em massa e não massa. Para este propósito utiliza-se o Índice de Diversidade de Shannon-Wiener, comumente aplicado para medir a biodiversidade em um ecossistema, para descrever padrões de regiões de imagens de mama com quatro abordagens: global, em círculos, em anéis e direcional. Em seguida, utiliza-se o classificador *Support Vector Machine* para classificar estas regiões em massa e não massa. A metodologia apresenta resultados promissores para a classificação de regiões de tecidos de mamografia em massa e não massa, obtendo uma acurácia máxima de 99,85%.

Palavras-chave: Mamografia, Classificação de tecidos de mama, Índice de Diversidade de Shannon-Wiener, Máquina de Vetores de Suporte.

ABSTRACT

Cancer is one of the biggest health problems worldwide, and the breast cancer is the one that causes more deaths among women. Also it is the second most frequent type in the world. The chances of survival for a patient with breast cancer increases the sooner this disease is discovered. Several Computer Aided Detection/Diagnosis Systems has been used to assist health professionals. This work presents a methodology to discriminate and classify mammographic tissues regions in mass and non-mass. For this purpose the Shannon-Wiener's Diversity Index, which is applied to measure the biodiversity in ecosystem, is used to describe pattern of breast image region with four approaches: global, in circles, in rings and directional. After, a Support Vector Machine is used to classify the regions in mass and non-mass. The methodology presents promising results for classification of mammographic tissues regions in mass and non-mass, achieving 99.85% maximum accuracy.

Keywords: Mammography, Breast tissue classification, Shannon-Wiener Diversity Index, Support Vector Machine.

Artigos submetidos pelo autor

SOUSA, U. S., BRAZ JUNIOR, G., PAIVA, A. C., SILVA, A. C. Classification of Breast Masses in Mammographic Images through the Application of Shannon-Wiener's Diversity Index. Pattern Recognition Letter.

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS E ABREVIATURAS

ACR	*American College of Radiology*
ACS	*American Cancer Society*
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
DDSM	*Digital Database for Screening Mammography*
FN	Falso-negativo
FP	Falso-positivo
IARC	*International Agency for Research on Cancer*
INCA	Instituto Nacional do Câncer
MIAS	*Mammograms Image Analysis Society*
ROI	*Region of Interest*
SVM	*Support Vector Machine*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WHO	*World Health Organization*

SUMÁRIO

1 INTRODUÇÃO

O câncer é um dos maiores problemas de saúde mundial. O relatório da Agência Internacional para Pesquisa do Câncer - *International Agency for Research on Cancer* (IARC) (INTERNATIONAL AGENCY FOR RESEARCH ON CANCER, 2009) estimou que no ano de 2008 ocorreriam cerca de 12 milhões de novos casos de câncer, sendo que 7 milhões de pessoas iriam morrer em conseqüências dessa doença. No entanto, aproximadamente 30% das mortes causadas pelo câncer podem ser evitadas (WORLD HEALTH ORGANIZATION, 2011). Mais de 70% das mortes causadas por câncer ocorrem em países de renda baixa e média e segundo estudos da IARC as mortes devem continuar crescendo, com estimativas de 12 milhões de mortes em 2030.

O câncer, que também é conhecido como tumores malignos ou neoplasma, apresenta diferentes tipos entre homens e mulheres. Os tipos de câncer mais incidentes na população mundial são o câncer de pulmão, o câncer de mama e câncer de colo do útero. Em todo o mundo, os tipos de câncer que causam mais mortes entre homens são o de pulmão, de estômago, fígado, de cólon e reto, esôfago e próstata. Já entre as mulheres, os tipos de câncer mais comuns são o de mama, de pulmão, de estômago, de cólon e reto e de colo de útero (WORLD HEALTH ORGANIZATION, 2011).

No Brasil, segundo o Instituto Nacional do Câncer (INCA), as estimativas válidas para os anos de 2010 e 2011, indicavam que para cada ano haveria cerca de 489.270 novos casos de câncer. Em 2010, eram esperados 236.240 novos casos entre os homens e 253.030 novos casos entre as mulheres. Entre os tipos mais incidentes, à exceção do câncer de pele do tipo não melanoma, seriam o câncer de próstata e de pulmão, entre os homens e o câncer de mama e colo do útero, entre as mulheres (INSTITUTO NACIONAL DO CÂNCER, 2010).

O câncer de mama é o tipo de câncer que possui a maior taxa de óbitos entre as mulheres e é o segundo tipo de câncer mais freqüente mundialmente. Estima-se que no Brasil surgiriam 49.240 novos casos de câncer de mama em 2010, com uma proporção de 49 casos a cada 100 mil mulheres. Cerca de 22% dos novos casos de câncer em mulheres a cada ano correspondem a câncer de mama (INSTITUTO NACIONAL DO CÂNCER, 2010).

Na região Nordeste, a estimativa para 2010 era de 8.270 novos casos de câncer de mama em mulheres. Apenas no estado do Maranhão, o número de novos casos de câncer de

mama é de 390 ficando atrás apenas do número de novos casos de câncer de colo do útero, estimado em 730.

O câncer de mama é uma doença que atinge em maior número as mulheres que os homens. As causas do câncer de mama ainda não são conhecidas. As células da mama feminina estão constantemente expostas a efeitos de hormônios femininos como o estrogênio e a progesterona, promovendo o aumento dessas células (AMERICAN CANCER SOCIETY, 2010). Os homens podem ter câncer de mama, mas esta doença é cerca de 100 vezes mais comum em mulheres que em homens (AMERICAN CANCER SOCIETY, 2010).

Os diversos tipos de câncer são conseqüência de alteração no DNA, que levam a um crescimento descontrolado das células originando as massas. Essas massas são chamadas de neoplasias, que são classificadas em benignas e malignas. As neoplasias benignas se caracterizam por ter crescimento organizado e geralmente lento, enquanto os tumores benignos apresentam limites bem nítidos. A neoplasia maligna apresenta crescimento mais rápido, desorganizado, infiltrativo e capacidade de se desenvolver em outras partes do corpo (metástase). Apenas a neoplasia maligna é chamada de câncer.

Apesar do alto índice de mortalidade causado pelo câncer de mama, esta doença é considerada relativamente de bom prognóstico (INSTITUTO NACIONAL DO CÂNCER, 2010). Quanto mais cedo o câncer de mama for detectado, maiores são as chances da paciente sobreviver. O recurso mais utilizado atualmente para detecção precoce do câncer de mama é o exame de mamografia. A mamografia é um raio-X da mama que é analisado por um especialista (radiologista) com o objetivo de reconhecer lesões na mama em fase inicial que são muito pequenas (em milímetros).

No entanto, o diagnóstico baseado em mamografias é uma etapa sensível, pois pode haver diferentes interpretações entre os radiologistas em relação a um mesmo exame. Além disso, a interpretação de mamografias é uma tarefa repetitiva que necessita de um nível de atenção muito grande sobre os detalhes presentes nas estruturas da imagem de raio-X.

Devido a esses fatores, nas últimas décadas tem surgido um grande interesse no uso de técnicas de reconhecimento de padrões, processamento e análise de imagens a partir de imagens de mamografias. Essas técnicas em conjunto têm sido utilizadas para desenvolver os 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, classificam as estruturas com detectadas com anormalidade nas classes maligno ou benigno (MARTINS, 2007). Esses sistemas têm o objetivo de aumentar o grau de precisão na detecção e no diagnóstico, dando uma segunda opinião ao radiologista.

Neste sentido, o presente trabalho descreve uma metodologia CAD para classificação de tecidos da mama a partir de regiões de imagens de mamografia em duas classes: massa e não massa. A classe massa abrange as regiões que correspondem a uma neoplasia maligna ou benigna. A classe não massa abrange regiões presentes em uma imagem mamográfica com tecido normal da mama. Neste trabalho é proposta a utilização do Índice de Diversidade de Shannon-Wiener (MAGURRAN, 2004) para extração de características de regiões de imagens mamográficas. Além disso, propomos uma modificação para o cálculo deste índice e sua aplicação a este problema. O índice é utilizado em conjunto com técnicas de processamento de imagens e reconhecimento de padrões para classificação em massa e não massa de regiões pré-segmentadas de imagens de mamografia.

Devido à grande necessidade de prover metodologias eficientes ao auxílio à detecção e diagnóstico de câncer de mama, diversos trabalhos de pesquisa têm sido desenvolvidos.

Em (MOAYEDI et al., 2010) é proposta uma metodologia para classificação automática de tecidos de mamografia. Primeiramente, é realizado um pré-processamento para retirada do músculo peitoral e segmentação da região de interesse. Depois, é empregada a Transformada de *Contourlet* como um descritor de características e Algoritmo Genético para selecionar características da base. A classificação é realizada baseada em *Successive Enhancement Learning* (SEL) *Weighted Support Vector Machine* (WSVM), *Support Vector-based Fuzzy Neural Network* (SVFNN) e *Support Vector Machine* (SVM). O trabalho apresenta resultados com acurácia de 96,6%, 91,5% e 82,1% respectivamente para cada uma das técnicas de classificação utilizada, com as imagens da base *Mammograms Image Analysis Society* (MIAS) (SUCKLING, 1994).

Em (ELTOUKHY *et al.*, 2010a) é apresentada uma metodologia para diagnóstico de câncer de mama em mamografias digitais utilizando Transformada de *Curvelet*. Esta abordagem é dividida em três fases. Na primeira fase uma operação de corte é aplicada na imagem para retirar as partes pretas da imagem. Na segunda fase a Transformada de *Curvelet* é utilizada para representar as ROIs *(Regions of Interest)* em níveis de decomposição

multiescala. Na terceira fase o vetor de características é utilizado como entrada para um classificador baseado em Distância Euclidiana com o objetivo de distinguir entre tecido normal e anormal. Neste trabalho a base de imagens MIAS foi utilizada para os testes e a acurácia máxima foi 98,59%. Em (ELTOUKHY *et al.*, 2010b) é apresentado uma estudo comparativo entre Transformada de Curvelet e Transformada de *Wavelet* para extração de características. Um classificador baseado em distância euclidiana é utilizado para classificar as amostras em normal, maligna e benigna. Neste trabalho, a base de imagens MIAS é utilizada e a acurácia máxima é 94,07% e 90,05% utilizando a Transformada de *Curvelet* e de *Wavelet*, respectivamente.

Em (BRAZ JUNIOR *et al.*, 2009) é proposta uma metodologia para discriminação de regiões extraídas de mamografias em massa e não massa usando Índice de Moran e Coeficiente de Geary como descritores de textura. Estas características são utilizadas como entrada para o classificador SVM com o propósito de distinguir entre massa e não massa, e em seguida em um segundo passo, distinguir os casos classificados como massa entre maligno e benigno. Neste trabalho, a base *Digital Database of Screening Mamography* (DDSM) (HEATH *et al.*, 2000) foi utilizada para os testes. No passo inicial, referente à classificação em massa e não massa, foram encontradas acurácia máxima de 99,39%, sensibilidade de 100% e especificidade de 98,94%. E, no passo seguinte foi encontrada acurácia máxima de 88,31%, sensibilidade de 84,78% e especificidade de 93,55% para a classificação das massas nas classes maligno e benigno.

Em (FAYE *et al.*, 2009) é apresentada uma metodologia para classificação de tecidos de mama a partir de imagens de mamografias onde são utilizados coeficientes de *Wavelet* para extração de características de tecidos da mama. Uma matriz é construída pela inserção de coeficientes de *Wavelet* de cada imagem como um vetor de linhas. Um limiar é definido e o desvio padrão para cada coluna é calculado. As colunas com os desvios padrões maiores que o limiar são mantidas e utilizadas como características para classificar as mamografias. As imagens são classificadas de acordo com a menor distância euclidiana entre seus vetores e as classes de vetores médios. Foi utilizada a base MIAS, e na classificação entre massa e não massa foi encontrada a acurácia de 98,55%, enquanto na classificação entre massas malignas e benignas foi encontrada a acurácia de 98%.

Em (MARTINS *et al.*, 2009) é apresentada uma metodologia para detecção de massas em imagens mamográficas digitais utilizando o algoritmo *Growing Neural Gas* com o

intuito de segmentar a imagem, e a função K de Ripley para descrever a textura dos objetos segmentados. Uma das etapas da metodologia proposta é a classificação dos objetos segmentados em massa e não massa utilizando a SVM. Este trabalho utiliza a base DDSM e obteve acurácia de 89,30%, com uma taxa média de 0,93 falso-positivo por imagem, para a classificação de tecidos em massa e não massa. Com objetivo semelhante, (MARTINS *et al.*, 2009b) utiliza o algoritmo *K-means* para segmentação de imagens de mamografia e matriz de co-ocorrência para extrair características das estruturas segmentadas. A classificação é realizada pelo classificador SVM em massa e não massa, obtendo 85% de acurácia.

Em (NUNES *et al.*, 2010) é proposta uma metodologia para detecção de massas em imagens mamográficas digitais utilizando o algoritmo de agrupamento *K-means* e a técnica *Template Matching* para identificação de regiões suspeitas. Em seguida, utiliza o Índice de Diversidade de Simpson para obter informações da textura da região suspeita. Finalmente, as informações das texturas são utilizadas pelo classificador SVM para classificar as regiões suspeitas em duas classes: massa e não massa. Este trabalho obteve uma acurácia de 83,94%, sensibilidade de 83,24% e especificidade de 84,14%.

Em (VERMA *et. al.*, 2009) é apresentada uma nova técnica para classificação de lesões de tecidos de mama em benigno e maligno. Esta técnica introduz o conceito de *soft clusters* dentro de uma camada de rede neural e as combina com mínimos quadrados para otimização dos pesos da rede neural. Para alimentar a rede neural quatro descritores de características BI-RADS juntamente com a idade da paciente e um valor de característica de sutileza. Nesta técnica foram utilizadas amostras da base DDSM e os resultados encontrados são a acurácia de 93%, sensibilidade de 93,88% e especificidade de 92,16% incluindo todos os *clusters*. Nos testes realizados sem os clusters fracos foram encontrados acurácia de 94%, sensibilidade de 97,83% e especificidade de 90,74%.

Ainda com o objetivo de classificar lesões de tecidos de mama em maligno e benigno, em (VANI *et al.*, 2010) é realizado um estudo para identificar as melhores características que distinguem entre massas malignas e benignas. Logo após é utilizado um classificador baseado em *Extreme Learning Machine (ELM)* para realizar a classificação. É realizado um estudo extenso para determinar a quantidade de neurônios para alcançar os melhores resultados na etapa de classificação. Neste trabalho são utilizadas imagens da base MIAS e os resultados encontrados nos testes identificaram que a melhor quantidade de

neurônios na camada oculta foi 30, com acurácia de 91%, sensibilidade de 90% e especificidade de 98%.

Os trabalhos relacionados indicam que metodologias baseadas em características de textura e reconhecimento de padrões apresentam resultados promissores para auxílio à detecção de câncer de mama a partir de mamografias. Vários trabalhos apresentam a taxa de acurácia acima de 90%. No entanto, ainda é necessário identificar técnicas que permitam melhorar estes resultados. Verificamos que a classificação de ROIs da mamografia em massa e não massa é uma etapa crucial nas metodologias de detecção de câncer de mama e que há um potencial a ser explorado em medidas que descrevam a textura e sejam baseadas em geoestatística e padrões de pontos.

Este trabalho tem como objetivo principal apresentar uma metodologia para classificação de regiões de imagens de mamografias em massa e não massa utilizando o Índice de Diversidade de Shannon-Wiener para descrição de textura. Assim, pretende-se apresentar uma inovação na extração de textura de imagens médicas e contribuir para a literatura da área de sistemas e metodologias CAD/CADx. Além disso, conforme a qualidade obtida pelos resultados deste trabalho, a metodologia aplicada poderá ser incorporada em uma ferramenta para a área médica, auxiliando o especialista com uma segunda opinião durante a etapa de detecção de massas em imagens de mamografias.

O restante deste trabalho está organizado conforme a seguir. No Capítulo 2 é apresentada a fundamentação teórica necessária para a compreensão do mesmo. Nesta seção aborda-se a estrutura da mama feminina, câncer de mama, exame de mamografia, Técnicas de Processamento de Imagens, Estatística Espacial para Extração de Textura, Índice de Diversidade de Shannon-Wiener, método de classificação *Support Vector Machine* – Máquina de Vetores de Suporte (SVM) e técnicas de validação de resultados.

O Capítulo 3 apresenta a metodologia utilizada neste trabalho para realizar a classificação das regiões extraídas de imagens de mamografias em massa e não massa, utilizando a extração de características com o Índice de Diversidade de Shannon-Wiener e a classificação utilizando Máquina de Vetores de Suporte.

O Capítulo 4 apresenta e discute os resultados obtidos através da metodologia proposta. No Capítulo 5 a conclusão deste trabalho, apresentando a eficiência dos métodos utilizados e apresentando sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica necessária para a compreensão da metodologia utilizada neste trabalho. Aborda-se a estrutura da mama feminina, o câncer de mama, o exame de mamografia, as técnicas de processamento de imagens, a estatística espacial para extração de textura, o Índice de Diversidade de Shannon-Wiener, o método de classificação *Support Vector Machine* – Máquina de Vetores de Suporte (SVM) e as técnicas de validação de resultados.

2.1 Mama feminina

A mama em uma mulher adulta possui forma semi-esférica ou cônica. Com a base localizada no músculo peitoral, sua forma, tamanho, consistência e aspecto geral são muito variáveis. A mama feminina é composta principalmente por lobos, lóbulos, ductos e estroma. Existem entre 15 e 20 lobos de tecido glandular do tipo túbulo-alveolar na mama, divididos por tecido conjuntivo fibroso e tecido adiposo (BIAZÚS, 2000). Os lobos contêm vários lóbulos, onde se encontram as pequenas glândulas responsáveis pela produção de leite. O leite produzido pelas glândulas flui até o mamilo através de pequenos tubos chamados de ductos. O estroma é formado por tecido adiposo e tecido conjuntivo que fica ao redor dos ductos e lóbulos, vasos sanguíneos e vasos linfáticos. A Figura 1 apresenta o esquema de uma mama.

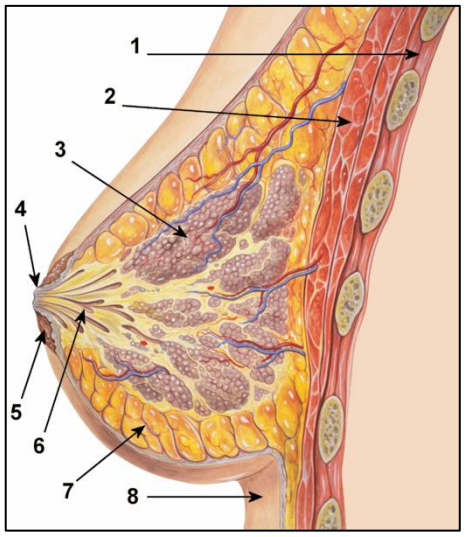


Figura 1 – Esquema de uma mama de mulher adulta. 1 – Caixa torácica. 2 – Músculo peitoral. 3 – Lóbulos. 4 – Mamilo. 5 – Aréola. 6 – Ductos. 7 – Tecido adiposo. 8 – Pele. FONTE: (WIKIMEDIA COMMONS, 2010)

2.2 Câncer de mama

O câncer de mama é uma doença originada pela multiplicação anormal das células da mama, formando um tumor maligno (INSTITUTO NACIONAL DO CÂNCER, 2010). Um tumor maligno é um grupo de células cancerosas que podem invadir tecidos adjacentes ou se disseminar para outras áreas do corpo (metástase). Embora seja possível que homens desenvolvam esta doença, o câncer de mama acomete quase que exclusivamente as mulheres (AMERICAN CANCER SOCIETY, 2010a). Raramente mulheres com menos de 35 desenvolvem o câncer de mama, em contrapartida, a partir desta idade a incidência desta doença cresce de forma rápida e progressiva.

Como a maioria das doenças, o câncer de mama possui alguns fatores de risco. Um fator de risco é qualquer coisa que afete a chance de uma pessoa ter uma doença. Um desses fatores de risco é o envelhecimento. Apenas 12,5% dos casos de câncer de mama invasivo são em mulheres abaixo de 45 anos, enquanto que cerca de 66,67% dos casos são em mulheres de 55 anos ou mais (AMERICAN CANCER SOCIETY, 2010a).

Fatores genéticos também podem contribuir para que uma mulher possua câncer de mama. A chance de desenvolver câncer de mama em mulheres que apresentam mutações nos genes BRCA1 e BRCA2 antes do 70 anos de idade é de 85% (INSTITUTO NACIONAL DO CÂNCER, 2010).

Outros fatores de risco que se destacam (AMERICAN CANCER SOCIETY, 2010a) são:

- Histórico familiar de câncer de mama – o risco é maior entre as mulheres que possuem parentes próximos com a doença;
- Histórico pessoal de câncer de mama – 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 – mulheres brancas são mais propensas a desenvolver a doença que mulheres afro-descendentes. No entanto, mulheres afro-descendentes estão mais sujeitas a morrer por causa da doença;
- Tecido mamário denso – por ter mais tecidos glandulares e menos tecidos adiposos, aumenta o risco de câncer de mama. Além disso, é mais difícil para os médicos detectar o câncer na mama.

- Doença mamária benigna prévia – mulheres que apresentaram uma doença benigna têm maior chance de ter câncer de mama.
- Período menstrual – mulheres que tiveram mais ciclos menstruais, pelo fato de ter começado antes dos 12 anos de idade ou passaram da menopausa em idade mais avançada (depois dos 55 anos) têm um risco ligeiramente maior de desenvolver câncer de mama. Isso pode estar relacionado à exposição aos hormônios estrogênio e progesterona.

Outros fatores de risco estão relacionados ao abuso do consumo de bebidas alcoólicas, ao fumo, à obesidade após a menopausa. É importante também que a mulher que se submete à reposição hormonal tenha acompanhamento médico.

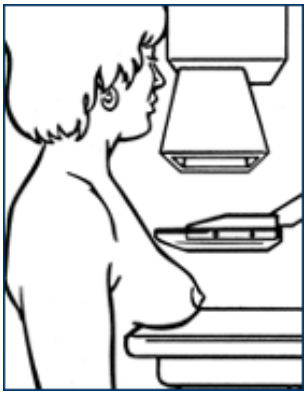
A detecção precoce do câncer de mama é a principal forma de diminuir as taxas de mortalidade causada por esta doença. Quanto mais cedo esta doença for detectada, maiores são as chances da paciente ser curada. O objetivo da detecção precoce é detectar lesões pré-cancerígenas ou câncer em estágio inicial ainda localizado no órgão de origem, sem ter ocorrido a metástase.

Como forma da doença ser detectada precocemente, recomenda-se que toda mulher com 40 anos ou mais procure um posto de saúde para ser examinada por um profissional da saúde uma vez por ano. Para mulheres com idade entre 50 e 69 anos, é recomendado que as mesmas façam o exame de mamografia a cada dois anos, pois o risco de câncer aumenta com a idade (MINISTÉRIO DA SAÚDE, 2004). Mulheres com idade a partir de 35 anos que tenham histórico de familiar de câncer de mama ou de ovário devem realizar o exame clínico e a mamografia uma vez por ano.

2.2.1 Mamografia

A mamografia é um raio-X da mama que produz uma imagem de alta resolução das estruturas internas da mama, com a finalidade de permitir a detecção do câncer de mama. Este exame é realizado geralmente para procurar por câncer em mulheres que não apresentam nenhum sintoma. Neste caso o exame é denominado mamografia de rotina ou de rastreamento.

O aparelho onde o exame de mamografia é realizado é chamado de mamógrafo. Neste aparelho a mama é comprimida com o objetivo de fornecer as melhores imagens, melhorando a capacidade de detecção/diagnóstico, como na Figura 2. A compressão da mama dura alguns segundos e o procedimento completo leva cerca de 20 minutos (AMERICAN CANCER SOCIETY, 2010a).

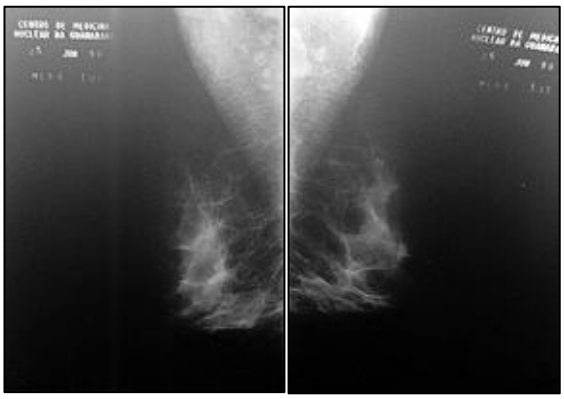


(a)

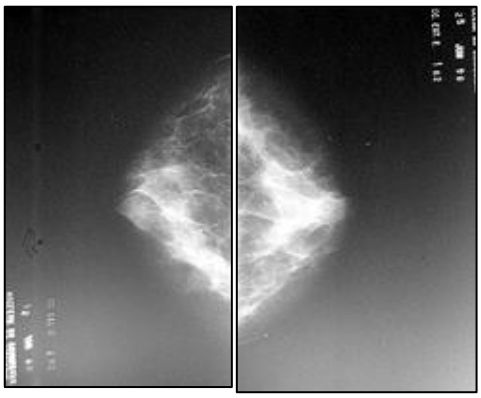
(b)

Figura 2 – Mamógrafos. (a) Esquema: Fonte: (AMERICAN CANCER SOCIETY, 2010b). (b) Mamógrafo real. Fonte: (MAMOWEB, 2011)

Normalmente são feitas duas radiografias no exame de mama, sendo uma para cada mama. Em cada mama são feitas duas projeções: uma médio lateral oblíqua (MLO) e uma crânio-caudal (CC). A Figura 3 exemplifica essas duas projeções.



(a)



(b)

Figura 3 – Mamografia com Incidência Médio-Lateral Oblíqua (ambas as mamas); (b) Mamografia com Incidência Crânio-Caudal (ambas as mamas). Fonte: (MAMOWEB, 2011).

A visão médio-lateral oblíqua é considerada a mais eficiente, pois mostra uma maior quantidade de tecido mamário incluindo estruturas mais profundas do quadrante súpero-externo e do prolongamento axilar. A visão crânio-caudal tem o objetivo de incluir o

material póstero-medial completo, de forma que complemente a médio-lateral oblíqua (MAMOWEB, 2011).

O exame de mamografia produz uma imagem em escala de cinza em uma folha grande de filme ou uma imagem de computador que é lida e interpretada pelo radiologista (AMERICAN CANCER SOCIETY, 2010a). O radiologista pode encontrar dois tipos de anormalidades na imagem de mamografia: calcificações e massas. As calcificações são depósitos minerais minúsculos que aparecem dentro do tecido da mama na forma de pequenas manchas brancas na imagem. Essas calcificações podem ser ou não causadas por câncer. Uma massa, que pode ocorrer com ou sem calcificação, é o outro de tipo de anormalidade encontrada em mamografias. As massas podem ser elementos não cancerígenos (cistos e tumores sólidos), mas também podem ser câncer. Normalmente as massas que não são cistos devem passar por uma biópsia.

Apesar de ser a melhor forma de detecção precoce de câncer de mama e contribuir para diminuição do índice de mortalidade em decorrência desta doença, a mamografia apresenta algumas desvantagens. Uma delas é o fato de não ser suficiente para provar que uma área anormal é câncer, necessitando que uma pequena quantidade do tecido suspeito seja submetida à biópsia. Outra desvantagem é o fato de ser difícil detectar lesões em mamografias de mulheres mais jovens porque geralmente suas mamas são densas e podem ocultar a lesão. Isso pode ocorrer também com mulheres grávidas ou que estejam amamentando (AMERICAN CANCER SOCIETY, 2010a). No entanto, isso geralmente não é uma grande preocupação, pois a maioria dos cânceres de mama detectados é em mulheres mais velhas.

O exame de mamografia tem sido muito importante para detecção precoce de câncer de mama, no entanto alguns casos de câncer não são diagnosticados e há ainda uma quantidade relativamente grande de falsos-positivos (regiões normais detectadas como lesões). Com o intuito de ajudar a aumentar a precisão e eficiência da mamografia, diversas técnicas têm sido desenvolvidas, dentre elas os sistemas CAD/CADx.

2.3 Processamento de Imagens Digitais

Em (GONZALEZ e WOODS, 2002), uma imagem pode ser 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. Tem-se uma Imagem Digital quando *x, y* e os valores de amplitude de *f* são finitos e discretos. Cada imagem digital é composta por um número finito de elementos, os quais têm uma localização e um valor particular. Esses elementos são conhecidos como *pixels* (*Picture elements*).

O Processamento de Imagens Digitais refere-se ao processamento de imagens por meio de um computador digital, ou seja, abrange processamentos cuja entrada e saída são imagens, além disso, engloba a extração de atributos a partir de imagens e também o reconhecimento de objetos individuais (GONZALEZ e WOODS, 2002). Um dos objetivos principais do processamento de imagens digitais é melhorar o aspecto visual de certas características estruturais à percepção humana, fornecendo informações suficientes para a sua interpretação.

O processamento de imagens segue diversas etapas. A Figura 4 apresenta um esquema clássico dessas etapas. Após a definição e delimitação do problema, normalmente as etapas a serem seguidas são: aquisição das imagens digitais, pré-processamento, segmentação, representação e descrição, reconhecimento e interpretação. O conjunto de dados gerados por uma etapa é utilizado pela próxima, no entanto, nem sempre o conjunto de dados gerados é uma imagem. Logo, ao fim de todas as etapas o resultado pode ser ou não representado por uma imagem digital.

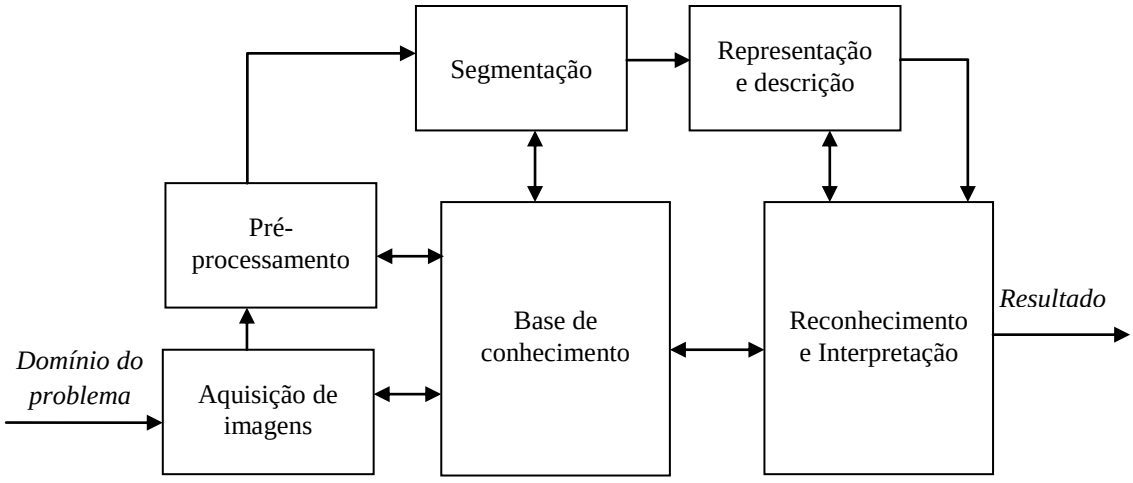


Figura 4 – Etapas fundamentais em processamento de imagens. Adaptado de (GONZALEZ e WOODS, 2002).

A aquisição de imagens é a primeira etapa no processamento de imagens. Nesta etapa a imagem pode ser obtida diretamente, como em um aparelho de raio-X digital, ou através de um digitalizador. O digitalizador é um aparelho que tem a função de ler sinais analógicos e convertê-los em sinais digitais. Por exemplo, a mamografia é impressa em uma folha de filme e é digitalizado por *scanners* especializados.

A segunda etapa é o pré-processamento das imagens adquiridas. Esta etapa tem o objetivo de melhorar a definição de certas estruturas da imagem. Nesta etapa podem ser aplicadas técnicas de realce e melhoramento de imagem, por exemplo: equalização de realce de contrastes, diminuição de ruídos, dentre outras.

A próxima etapa é a segmentação. Nesta etapa o objetivo é extrair da imagem apenas partes dela que realmente interessam para o processamento. A etapa de segmentação é responsável por separar os objetos presentes nas imagens de acordo com suas características. A segmentação pode ser definida como o processo de particionamento de regiões onde todos os elementos de uma região devem ser o mais homogêneo possível e elementos de regiões diferentes devem ser o mais heterogêneo possível. Há três abordagens diferentes de segmentação: manual, semi-automática e automática.

A segmentação manual é realizada por um especialista. Neste caso, é necessário o uso de ferramentas que possam auxiliar o especialista de forma visual para a separação da área de interesse. Na segmentação semi-automática, o especialista deve passar informações sobre o que ele deseja buscar na imagem ou onde buscar determinadas características. Um algoritmo lê e interpreta as informações passadas e busca na imagem regiões suspeitas de apresentarem tais informações. A segmentação automática é mais complexa. Neste tipo de segmentação, o especialista não passa nenhuma informação e o algoritmo deve ter habilidade suficiente para separar regiões da imagem em conjuntos desconexos obedecendo aos critérios de similaridade de cada região.

A quarta etapa do processamento de imagens é a representação e descrição dos objetos. Esta etapa também é conhecida como extração de características. As características extraídas nesta etapa são informações importantes para discriminar as classes distintas na imagem. Cada medida extraída de uma determinada área compõe um vetor de características que define um padrão para a determinada área.

A quinta etapa é o reconhecimento e interpretação das imagens. O reconhecimento atribui um rótulo para o objeto baseado em suas características, enquanto a interpretação está relacionada à atribuição de um significado a um conjunto de objetos reconhecidos.

2.3.1 Equalização de Histograma

Um histograma é uma representação da freqüência dos pixels na imagem. O histograma de uma imagem digital com níveis de cinza no intervalo $[0, \mathrm{L}-1]$ é uma função discreta $h(r_k) = n_k$, onde r_k é o k-ésimo nível de cinza e n_k é o número de *pixels* na imagem com a intensidade r_k.(GONZALEZ e WOODS, 2002). A manipulação do histograma pode ser utilizada com eficiência para melhoramento das características presentes da imagem.

Uma técnica comumente usada para o melhoramento de imagens é a Equalização de Histograma. Esta técnica procura redistribuir os valores de tons de cinza dos pixels em uma imagem, afim de que se obtenha um histograma uniforme e o número de pixels para qualquer nível de cinza seja o mesmo (MARQUES FILHO e VIEIRA NETO, 1999). Como conseqüência dessa distribuição uniforme obtém-se uma imagem com melhor contraste, o qual é uma medida qualitativa e que está relacionada à distribuição de tons de cinza em uma imagem.

Considerando $h(r_k)$ um histograma calculado em uma imagem S, então o histograma acumulado de $h(r_k)$ pode ser calculado por:

$$H(0) = h(0); H(1) = H(0) + h(1); H(r_k) = H(r_k - 1) + h(r_k)$$

para $r_k = 1, \dots, L-1$. O novo histograma é obtido por:

$$T(r_k) = round(\frac{L-1}{MN} H(r_k))$$

onde M e N são as dimensões da imagem. A imagem digital equalizada é obtida por pixel por $I = T(r_k)$. A Figura 5 apresenta a imagem original e a imagem equalizada com seus respectivos histogramas.

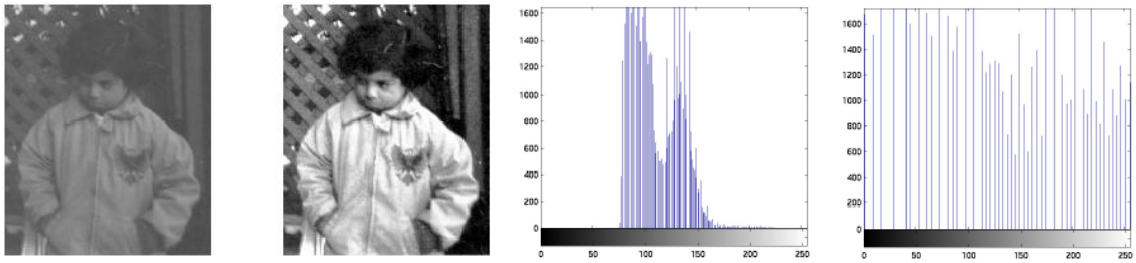


Figura 5 - Equalização de Histograma. (a) Imagem Original, (b) Imagem Equalizada, (c) Histograma da Imagem Original e (d) Histograma Equalizado. Adaptado de (MATHWORKS, 2011).

2.3.2 Textura

A análise de textura é uma aplicação relevante em análise de imagens, ela possibilita distinguir regiões da imagem que apresentam as mesmas características de padrões e as mesmas cores em determinada combinação de bandas (CONCI *et al.*, 2008). Isso torna a textura uma propriedade importante nos processos de reconhecimento, descrição e classificação de imagens.

Não há uma definição precisa para textura. Dependendo da aplicação e do método de análise das imagens digitais, existem diversas definições (CONCI *et al.*, 2008). Em (HARALICK *et al.*, 1973) a textura é definida como a característica de uma região relacionada a coeficientes de uniformidade, densidade, aspereza, regularidade, intensidade, dentre outras características da imagem. Em (GONZALEZ e WOODS, 2002) afirma-se que a textura é intuitivamente descrita por medidas que quantificam suas propriedades de suavidade, rugosidade e regularidade. A textura, de forma geral, é caracterizada de forma bidimensional, onde uma dimensão contém as propriedades primitivas ou estatísticas de tonalidades e cores enquanto a outra corresponde aos relacionamentos espaciais entre elas.

Há várias abordagens para análise de textura, sendo que as principais são a estrutural, a espectral e a estatística (GONZALEZ e WOODS, 2002). Neste trabalho foi utilizada a abordagem estatística.

A abordagem estrutural considera que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, conforme regras bem definidas. 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. Essa abordagem favorece a descrição de imagens através de regras estatísticas que regem tanto a distribuição quanto a relação entre os diferentes níveis de cinza. Há dois tipos de vertentes na abordagem estatística: a de primeira ordem (onde as características são extraídas a partir de histogramas de primeira ordem) e a de segunda ordem (onde o posicionamento relativo da ocorrência dos níveis de cinza é levado em consideração).

Dentre os métodos mais utilizados na abordagem estatística destaca-se a extração de características de textura a partir da matriz de co-ocorrência, proposta por (HARALICK *et al.*, 1973). Essa matriz contém a relação de freqüência relativa entre dois *pixels* separados por uma distância d na orientação θ, um *pixel* com tom de cinza i e o outro com tom de cinza j. As matrizes de co-ocorrência formam a base para várias medidas estatísticas conhecidas como descritores de Haralick.

Em (HARALICK *et al.*, 1973) é proposto que a informação da textura pode ser especificada usando matrizes de dependência espacial dos níveis de cinza calculados em vários ângulos (0°, 45°, 90° e 135°) e distâncias. Uma matriz de co-ocorrência será gerada para cada direção. Dessa forma, cada *pixel* terá sua informação de nível de cinza compartilhada com múltiplas análises, melhorando a determinação de sua classe.

O tamanho da matriz de co-ocorrência depende do número de tons de cinza que a imagem possui. Para que esse tamanho esteja dentro dos limites computáveis, geralmente diminui-se o número de tons possíveis em cada imagem através da quantização da imagem.

A Figura 6 apresenta um exemplo de construção de matriz de co-ocorrência com 4×5 *pixels*, 8 níveis de cinza (0 a 7) e a distância $d = 1$. Observa-se que o elemento (0, 0) da matriz de co-ocorrência de tons de cinza foi associado o valor 1. Este valor 1 representa que apenas uma única vez em toda a imagem existe a combinação [0 0] entre dois *pixels* separados a uma distância $d = 1$ e ângulo $\theta = 0°$ com o valor 0. Para a posição (0,1) da matriz de co-ocorrência, foi associado o valor 2 indicando a quantidade de vezes que existe a combinação [0 1] entre dois *pixels* obedecendo os mesmos critérios de ângulo e distância.

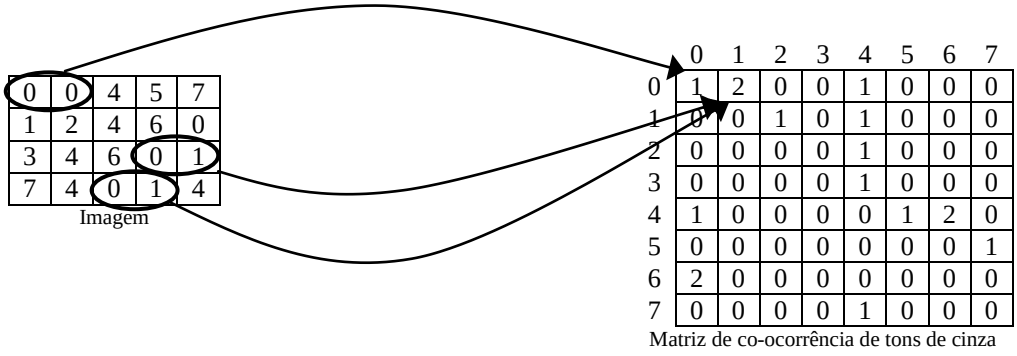


Figura 6 - Exemplo de cálculo de matriz de co-ocorrência com $\theta = 0°$ e $d = 1$. Adaptado de (CONCI *et al.*, 2008)

Após a obtenção da matriz de co-ocorrência, esta passa por uma normalização que consiste em dividir cada elemento da matriz pelo número de ocorrências totais.

2.4 Estatística Espacial para Extração de Textura

A análise espacial é um estudo quantitativo de fenômenos localizados no espaço, diferenciando-se da estatística clássica pelo uso explícito das coordenadas espaciais no processo de coleta, descrição e análise dos dados (BRAZ JUNIOR, 2008). Na análise espacial os fenômenos espaciais ou objetos de estudo estão sempre relacionados a uma determinada localização no espaço.

Em (PAIVA *et al.*, 1999) é dito que o processo de análise de pontos pode ser descrito em termos do efeito de primeira e segunda ordem. Os efeitos de primeira ordem correspondem às variações do valor médio do processo no espaço e são considerados globais ou de grande escala. Os efeitos de segunda ordem representam a dependência espacial no processo proveniente da estrutura de correlação espacial e são denominados locais ou de pequena escala.

O processo de análise de dados espaciais contém métodos de visualização, métodos exploratórios para investigar algum padrão nos dados e métodos que auxiliem a escolha de um modelo estatístico e a estimação dos parâmetros desse modelo (PAIVA *et al.*, 1999). Segundo (LEVINE, 1996), as estatísticas de segunda ordem usadas para descrever tanto pontos quanto áreas podem ser subdivididas em três categorias gerais:

- Medidas de distribuição espacial: descrevem o centro, a dispersão, a direção e a forma da distribuição de uma variável;
- Medidas de autocorrelação espacial: descrevem a relação entre as diferentes localizações para uma variável simples, indicando o grau de concentração ou dispersão;
- Medidas de associação espacial entre duas ou mais variáveis: descrevem a correlação ou associação entre variáveis distribuídas no espaço.

As medidas de autocorrelação espacial são utilizadas sempre que o valor de uma variável em um lugar do espaço está relacionado com seu valor em outros lugares no espaço. Assim, observações no espaço separadas a uma certa distância d possuem valores similares (correlação). A estatística tem o objetivo de medir o grau de associação espacial entre as observações de uma ou mais variáveis. A autocorrelação é dita positiva quando o fato observado em um lugar também é observado em seus vizinhos separados a uma certa distância. A autocorrelação é negativa quando a situação é inversa, ou seja, o fato observado em um lugar não é observado em seus vizinhos separados a uma determinada distância.

No contexto da estatística, a textura também pode ser descrita em termos de dois componentes principais associados a *pixels* (ou outra unidade): variabilidade e autocorrelação. O uso de técnicas de estatística espacial apresenta a vantagem de que os dois aspectos podem ser medidos em conjunto. Essas medidas descrevem a textura obtida de uma determinada imagem através do grau de associação espacial presente dentro dos elementos.

2.5 Índice de Diversidade

Um índice de diversidade é a medida da heterogeneidade de uma comunidade. Ele é utilizado na ecologia para medir a biodiversidade em um ecossistema. Na economia, serve para medir a distribuição sobre os setores de atividades econômicas em uma região. De forma mais geral, os índices de diversidade podem ser utilizados para estimar a diversidade de uma população na qual cada membro pertence a um único grupo ou espécie.

2.5.1 Índice de Diversidade de Shannon-Wiener

O Índice de Diversidade de Shannon-Wiener (IDSW) é utilizado neste trabalho como ponto chave para descrever padrões de regiões de imagens da mama e classificá-los como representantes das classes massa e não massa.

Este índice é derivado da teoria da informação, retratando a possibilidade de se coletar dois indivíduos aleatoriamente em uma comunidade e estes pertencerem a espécies distintas (SPELLERBERG, 2003). O cálculo do IDSW é definido por:

$$IDSW = -\sum_{i=1}^{S} p_i \ln p_i \tag{1}$$

onde p_i é a proporção total da amostra pertencente à espécie i, ou seja, $p_i = \frac{n_i}{N}$ com n_i sendo o número de indivíduos da espécie i; S é número de espécies amostradas e N é número total de indivíduos.

Espécies raras e abundantes têm pesos iguais no *IDSW*. Quanto maior for o *IDSW*, maior será a diversidade da população em estudo. Este índice pode expressar diversidade e uniformidade.

Para a utilização deste índice com o objetivo de caracterizar a textura de uma região de interesse de uma imagem (ROI) de mamografia, consideramos que os pixels representam os indivíduos da população e as suas intensidades indicam a sua espécie. Consideramos que cada ROI utilizada possui uma distribuição com S tonalidades de cinza. Assim, S é o número de tonalidades possíveis na ROI, n_i é o número total de pixels de uma tonalidade i presente na ROI e N é o número total de pixels da ROI.

2.6 Técnicas de Reconhecimento de Padrões

As técnicas de reconhecimento de padrões utilizam um conjunto de propriedades ou características previamente extraídas para classificar ou descrever padrões ou objetos. Segundo (LOONEY, 1997), um padrão é tudo aquilo para o qual existe uma entidade nomeável representante, geralmente, criada através do conhecimento cultural humano.

Há dois processos em reconhecimento de padrões: classificação e reconhecimento. A classificação é o processo onde uma amostra de uma população qualquer é dividida em dois grupos denominados classes. O reconhecimento é o processo onde uma amostra desconhecida da mesma população é reconhecida como pertencente a uma das classes criadas. A classificação pode ser realizada de duas formas: aprendizagem supervisionada e aprendizagem não supervisionada (LOONEY, 1997).

No processo de aprendizagem não supervisionada, um conjunto de objetos representantes de uma população é examinado e dividido em sub-conjuntos (classes). Os objetos da mesma classe devem ter o maior grau de similaridade possível, enquanto os objetos de classes diferentes devem ter o maior grau de dissimilaridade. Este processo também é conhecido como agrupamento. Em contrapartida, no processo de aprendizagem supervisionada um "reconhecedor" pode ser treinado previamente para identificar a classe de qualquer objeto desconhecido da mesma população. Neste trabalho utiliza-se o processo de aprendizagem supervisionada.

Os objetos possuem determinadas propriedades que são capazes de realizar 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.

Quando o conjunto de vetores de características é muito grande, é possível que algumas características sejam extremamente correlatas ou redundantes que podem sobrecarregar o classificador e induzi-lo ao erro. Por isso, um passo importante é pré-processar os vetores de características com o objetivo de eliminar parte dessas características redundantes ou correlatas.

Após a eliminação das características redundantes de cada objeto da população, atribui-se um rótulo para cada um. Este rótulo é a determinação de uma classe a partir de um conhecimento humano prévio. Logo após, divide-se o conjunto de vetores de características em dois subconjuntos. O primeiro subconjunto, chamado de amostras de treinamento, é utilizado no classificador no processo de treinamento. Nesse processo, uma assinatura única é gerada pelo classificador 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.

Finalmente, após o treinamento do classificador, o segundo subconjunto de objetos pode ser submetido ao classificador para que seja realizado o reconhecimento de cada objeto e sua devida classe. Este subconjunto é chamado de amostras de testes e possui elementos da população de objetos que não são conhecidos pelo classificador no processo de treinamento. No processo de teste, 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. Sendo assim, a tentativa de reconhecimento de padrão de um objeto deve ser feita necessariamente sobre objetos 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 de mamografia.

2.6.1 Máquina de Vetores de Suporte

A *Support Vector Machine* – Máquina de Vetores de Suporte (SVM) é um método de aprendizagem supervisionada usado para estimar uma função que classifique dados de entrada em duas classes. Introduzida em (VAPNIK, 1998), a idéia básica da SVM é construir um hiperplano como superfície de decisão, de forma que a margem de separação entre as classes seja máxima. A etapa de treinamento na SVM tem o objetivo de obter hiperplanos de forma que dividam as amostras, otimizando os limites de generalização.

A SVM tem se apresentado como um classificador superior quando comparado a outros classificadores em uma variedade de aplicações (CRISTIANINI e SHAWE-TAYLOR, 2000). As SVMs são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço de muitas dimensões. As SVM são uma classe de algoritmo de aprendizado baseados na Teoria de Aprendizagem Estatística.

Nos casos onde o conjunto de amostras é constituído por duas classes com padrões linearmente separáveis, um classificador SVM consegue encontrar um hiperplano baseado em um conjunto de pontos chamados de *vetores de suporte*. O vetor de suporte maximiza a distância entre as margens de separação entre as classes. O hiperplano pode ser entendido como uma superfície de separação de duas regiões em um espaço multidimensional, sendo que é possível até ter um número infinito de dimensões. Existem casos nos quais as classes não possuem padrões linearmente separáveis. No entanto, a SVM também é capaz de encontrar um hiperplano nesses casos utilizando conceitos pertencentes à teoria da otimização. A Figura 7 apresenta classes linearmente separáveis com hiperplanos separando as duas dimensões. O hiperplano ótimo (linha central) é aquele cuja distância entre os pontos da amostra é a maior possível.

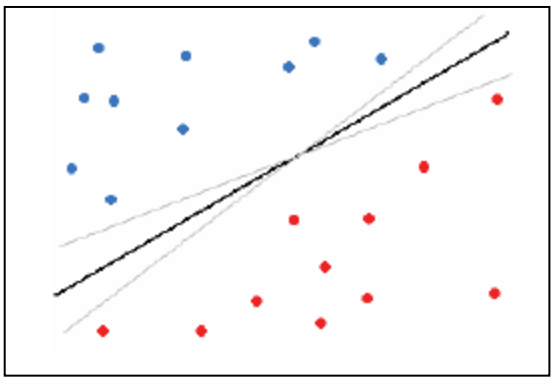


Figura 7 - Duas classes separadas através de hiperplanos

Considerando (x_i, y_i) como o conjunto de amostras de treinamento com $x_i \in \Re^n$, $y_i \in \{-1, +1\}$, $i = \{1,2, \dots, n\}$, sendo x_i o vetor de entrada, y_i o rótulo da classe e i o índice de cada ponto amostral. O objetivo é estimar uma função $f: \Re^n \longrightarrow \{-1, +1\}$, que separe corretamente os exemplos de testes em duas classes distintas.

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

$$y_i\big((w \cdot x_i) + b\big) > 1 \tag{2}$$

onde w é o vetor normal ao hiperplano de decisão e b é o corte ou distância da função f em relação à origem. A Figura 8 apresenta a interpretação geométrica de w e b sobre um hiperplano.

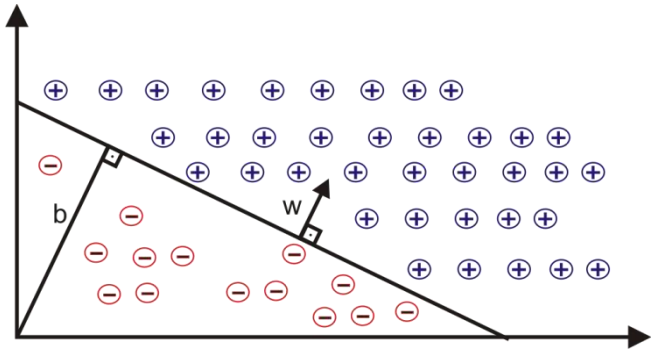


Figura 8 - Interpretação geométrica de *w* e *b* sobre um hiperplano

De acordo com a Equação 2, os valores ótimos para w e b serão encontrados pela seguinte equação (CHAVES, 2006):

$$\Phi(w) = \frac{w^2}{2} \tag{3}$$

Mesmo nos casos onde seja impossível uma perfeita separação entre as classes, a SVM possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação. Isso é possível devido a inclusão de variáveis de folga, que permitem que as restrições da Equação 2 sejam quebradas.

Nesse caso o problema de otimização passa a ser então a minimização da Equação 4 conforme a restrição imposta pela Equação 2. C é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro do treinamento e deve ser definido pelo usuário.

$$\Phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{4}$$

para

$$y_i\big((w \cdot x_i) + b\big) + \xi_i \geq 1 \tag{5}$$

Utilizando a teoria dos multiplicadores de Lagrange, chega-se à Equação 6. O objetivo então passa a ser encontrar os multiplicadores de Lagrange α_i ótimos que satisfaçam a Equação 7 (CHAVES, 2006).

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i y_i) \quad (6)$$

$$\sum_{i=1}^{N} \alpha_i y_j = 0, 0 \leq \alpha_i \leq C \quad (7)$$

Somente os pontos onde a restrição da Equação 2 seja exatamente igual à unidade têm correspondentes $\alpha \neq 0$. Esses pontos se localizam sobre as margens e são chamados de vetores de suporte. Tais pontos têm fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento.

Na Figura 9 os pontos que representam vetores de suporte são destacados. Os pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos.

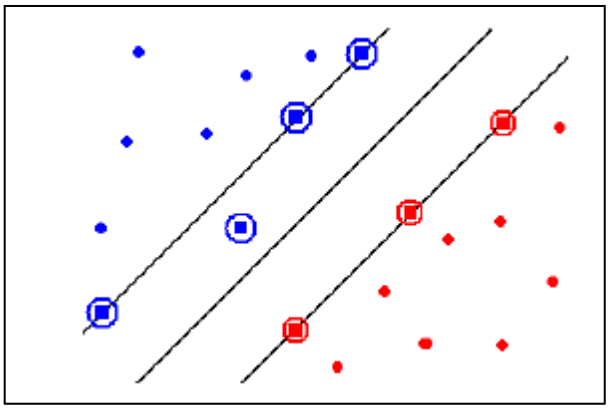


Figura 9 - Vetores de suporte (destacados por círculos)

Para que seja possível a SVM classificar amostras que não são linearmente separáveis, necessita-se de uma transformação não-linear que transforme o espaço de entrada em um novo espaço. A dimensão desse espaço deve ser suficientemente grande, e através dele, a amostra pode ser linearmente separável. Assim, o hiperplano de separação é definido como uma função linear de vetores retirados do novo espaço ao invés do espaço de entrada original. Essa construção depende do cálculo de um produto interno com uma função *K* de núcleo (HAYKIN, 2007). A função *K* pode realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 8 mostra o resultado da Equação 6 com a utilização de um núcleo.

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i y_i) \tag{8}$$

A base radial é uma importante família de funções de núcleo, sendo muito utilizada em problemas de reconhecimento de padrões. Neste trabalho utiliza-se a base radial e esta função é definida na Equação 9.

$$K(x_i x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{9}$$

2.7 Validação de resultados

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

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

Nesses 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 (S), Especificidade (E), Acurácia (A), Valor Preditivo Positivo (VPP) e Valor Preditivo Negativo (VPN) (BLAND, 2000).

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

$$S = \frac{VP}{VP + FN} \tag{10}$$

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

$$E = \frac{VN}{VN + FP} \tag{11}$$

A acurácia é a razão entre o número de casos identificados corretamente e o número total de casos:

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{12}$$

O valor preditivo positivo indica a probabilidade de um paciente com o resultado positivo ter a doença:

$$VPP = \frac{VP}{VP + FP} \tag{13}$$

O valor preditivo negativo indica a probabilidade de um paciente com o resultado negativo não ter a doença:

$$VPN = \frac{VN}{VN + FN} \tag{14}$$

A sensibilidade, especificidade, valor preditivo positivo e negativo e acurácia são utilizados para avaliar o desempenho e adequação da metodologia criada neste trabalho.

3 METODOLOGIA

Neste capítulo são descritos os procedimentos utilizados para classificação de regiões de tecidos de mama, extraídos a partir de imagens de mamografia, em massa e não massa. Em primeiro lugar é apresentada a base de imagens utilizada nos testes. Em seguida, é descrita a sequência de ações adotadas para a classificação dos tecidos obtidos, através da extração de características utilizando o Índice de Diversidade de Shannon, classificação de objetos em massa e não massa. Finalmente, é apresentada a forma de validação de resultados.

3.1 Metodologia Proposta

A sequência apresentada pela Figura 10 mostra a metodologia proposta neste trabalho.

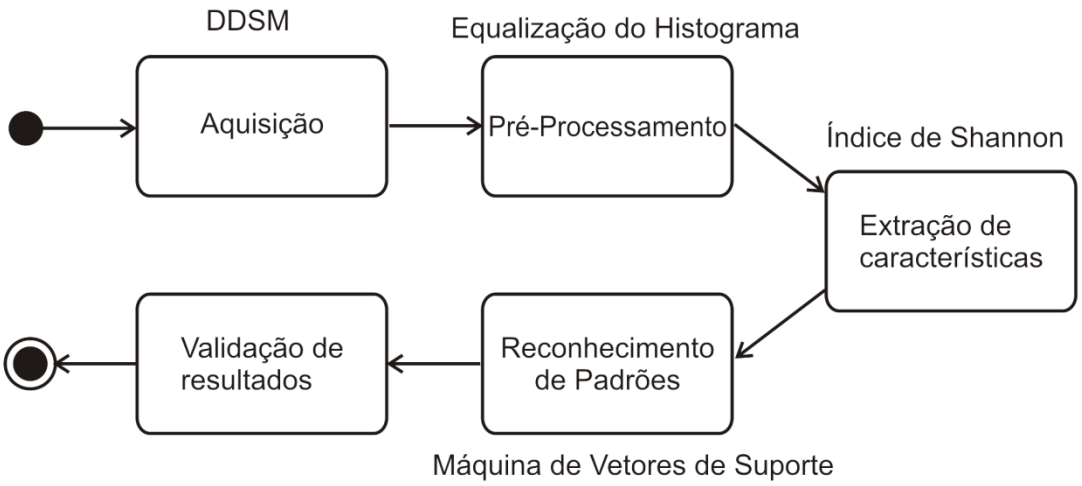


Figura 10 - Etapas da metodologia proposta neste trabalho

A primeira etapa tem o objetivo de obter as amostras de tecidos de mama necessárias para o processamento. Nesta etapa as regiões de interesse com tecidos de não massa e de massa são obtidas. Logo depois, as amostras passam pela etapa de pré-processamento que visa realçar os aspectos de textura presentes nas imagens. Após o pré-processamento, são extraídos os valores do Índice de Diversidade de Shannon para cada amostra.

As características extraídas de cada amostra formam o vetor de características e o conjunto de vetores de características é submetido ao processo de classificação supervisionada utilizando a SVM. Neste processo, um subconjunto de vetores de características é utilizado para etapa de treinamento, onde é criado um padrão sobre as medidas extraídas. O outro subconjunto restante é utilizado na etapa de testes. Este subconjunto é totalmente desconhecido da etapa de treinamento sendo utilizado para realizar os testes e validação dos resultados. Finalmente, segue a etapa de validação e comparação de resultados obtidos no reconhecimento do padrão de massa e não massa utilizando diversas métricas comumente adotadas.

3.1.1 Aquisição das Amostras

Neste trabalho foi utilizada a base de mamografias digitalizadas a partir de filmes radiográficos: *Digital Database for Screening Mammography* (DDSM). Esta base está disponível na Internet (HEATH *et al.*, 2000) e tanto a localização quanto o diagnóstico das lesões nesta base foi realizados por especialistas.

A base DDSM contém 2620 exames de pacientes de diferentes origens étnicas e raciais. Cada caso contém duas imagens de cada mama, nas projeções médio-lateral oblíqua e crânio-caudal. Juntamente com as imagens, são disponibilizadas algumas informações associadas à paciente como idade na época do exame e densidade da mama. Além disso, são disponibilizadas informações sobre a imagem, como tipo do *scanner* e resolução da imagem. Para as imagens com suspeitas de lesões, estão disponíveis informações sobre a localização, o diagnóstico e a descrição da anormalidade em um arquivo chamado arquivo de *overlay*.

As imagens da base DDSM foram adquiridas através das seguintes instituições americanas: *Massachusetts General Hospital, University of South Florida, Sandia National Laboratories, Washington University School of Medicine, Wake Forest University School of Medicine, Sacred Heart Hospital* e *ISMD, Incorporated*. As descrições das imagens de mamografias seguem os termos lexicográficos padronizados pelo ACR (*American College of Radiology*) e são publicados no BI-RADS (*Breast Imaging Reporting and Data System).*

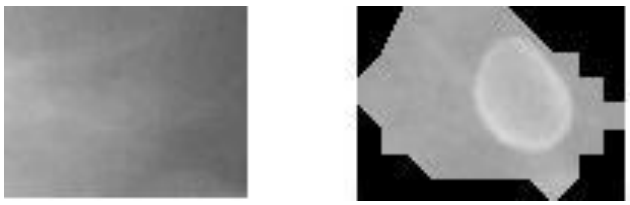
Neste trabalho foram utilizadas 3322 regiões de interesse de tecidos normais de mama e 3576 regiões de interesse de tecidos com massa (maligno e benigno), constituindo no

total 6898 regiões de interesse que foram extraídas de imagens de mamografia tanto na visão crânio-caudal como médio-lateral-oblíqua.

Além de informações sobre a paciente e a imagem, o arquivo de descrição das imagens que possuem lesões contém as seguintes informações: quantidade de lesões presentes na mamografia, tipo de lesão e o contorno que delimita a lesão. O contorno da lesão está codificado em *chain code* (MORSE, 2000) que torna possível a extração automática apenas da região de massa.

Neste trabalho, todas as amostras representantes da classe massa foram extraídas com base no contorno da massa codificado em *chain code*, porém foram utilizadas duas formas diferentes para extração de representantes de massa. Na primeira forma, foi extraído o *bounding box* que engloba a massa, ou seja, região retangular formada pelo x_{min}, x_{max}, y_{min} e y_{max} do contorno da massa. Estas regiões extraídas são utilizadas na abordagem de extração de características em círculos e em anéis abordadas nas Seções 3.1.3.2 e 3.1.3.3, respectivamente.

Na segunda forma, foram extraídas como representantes da classe massa apenas regiões internas ao contorno da massa, de acordo com o *chain code* do arquivo de *overlay*. Essas regiões são utilizadas nas abordagens global e direcional, apresentadas nas Seções 3.1.3.1 e 3.1.3.4, respectivamente. Neste processo os *pixels* da área fora do contorno da massa e presentes no *bounding box* recebem o valor -1. Em seguida, a nova imagem é armazenada em um arquivo texto onde cada linha possui a coordenada espacial do *pixel* (*x*, *y*) e seu valor correspondente. No processo de equalização de histograma e de extração de característica apenas os valores de *pixels* diferentes de -1 são incluídos. A Figura 11(a) apresenta um exemplo de não massa e a Figura 11(b) apresenta um exemplo de massa, onde aos *pixels* fora do contorno da massa foi atribuído o valor -1 e na imagem tem a cor preta.



(a) **(b)**

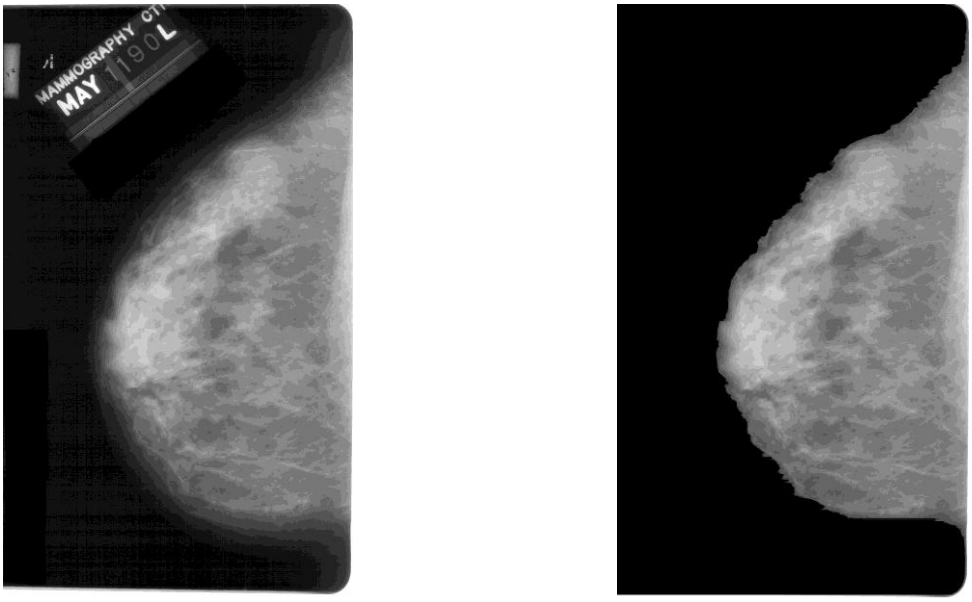
Figura 11 - Regiões extraídas da base DDSM. (a) Não massa. (b) Massa

As amostras representantes da classe não massa foram extraídas de forma aleatória a partir de imagens mamográficas que continham massas (maligna ou benigna). O

este algoritmo para este propósito deveria garantir que não fossem selecionadas regiões pertencentes aos seguintes objetos: o fundo da imagem com seus rótulos, as massas e o músculo peitoral (caso a projeção fosse médio-lateral oblíqua). Os rótulos das imagens mamográficas trazem informações como o tipo de projeção, a data do exame ou a lateral do corpo a qual pertence à mama.

Para retirada desses rótulos, foi utilizado um método simples para segmentar somente a parte correspondente à mama. Primeiro, foi criada uma cópia da imagem da mama. Depois, essa cópia foi binarizada com base em um limiar, onde: se um *pixel* tem o valor menor que o limiar, então recebe o valor 0 (zero); no caso, do valor do *pixel* ter o valor maior ou igual ao limiar, então recebe o valor 255. Neste trabalho foi adotado o valor 80 como limiar, conforme obtido em (MARTINS, 2007).

Se a imagem de mamografia é de uma mama direita o algoritmo inicializa a varredura no canto superior esquerdo da imagem. Em caso contrário, o algoritmo inicializa no canto superior direito da imagem. Quando um *pixel* de cor preta (valor 0) é encontrado durante a varredura, o valor 0 é atribuído aos *pixels* dessa linha que ainda não foram lidos. Dessa forma, as estruturas que não pertencem à mama são apagadas. A Figura 12 apresenta duas imagens de mamografia. A primeira corresponde à imagem original e a segunda corresponde à imagem após a execução do algoritmo para remoção de objetos que não pertencem à mama em si.

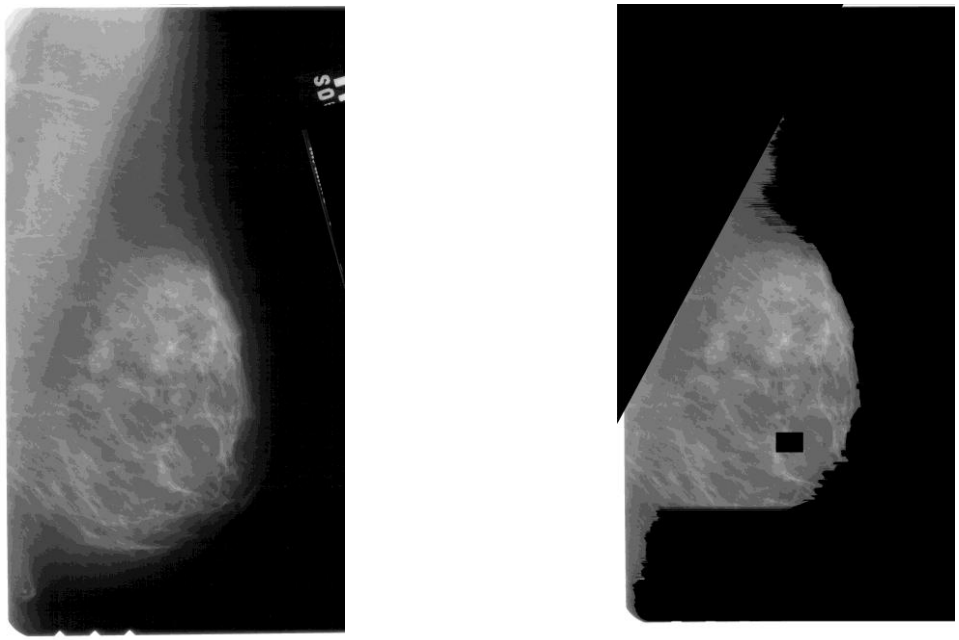


(a) (b)

Figura 12 – (a) Imagem de mamografia original. (b) Imagem de mamografia após a segmentação.

Após a retirada do fundo da imagem, todos os valores dos *pixels* internos à *bounding box* que engloba uma massa recebem o valor 0 (zero) para evitar que algum *pixel* dessa região seja selecionado como representante da classe não massa.

Nas imagens de mamografia com o tipo de projeção médio-lateral oblíqua, foi necessário que o músculo peitoral fosse removido para que nenhum *pixel* dessa área fosse extraído como representante da classe não massa. Para isso, se uma imagem representa uma mama direita, então o valor 0 é atribuído aos *pixels* correspondentes a um triângulo retângulo com o ângulo de 90° no canto superior esquerdo da imagem, e os catetos horizontal e vertical têm 2/3 do tamanho da largura e da altura da imagem, respectivamente. Se a imagem representa uma mama esquerda, então o ângulo de 90° do triângulo corresponde ao canto superior direito da imagem. A Figura 13(a) apresenta a imagem de mamografia original enquanto a Figura 13(b) apresenta a imagem de mamografia onde foi atribuído o valor 0 aos *pixels* que formam o triângulo retângulo no canto superior esquerdo e aos *pixels* internos à *bounding box* da massa na imagem, restando apenas regiões normais da mama para serem selecionadas pelo algoritmo.



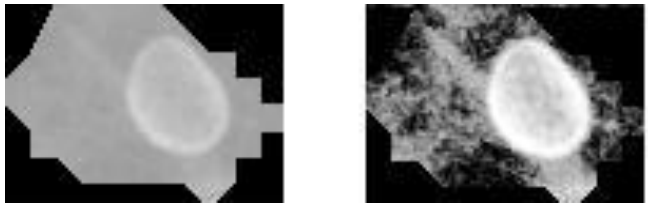
(a) (b)

Figura 13 – (a) Imagem de mamografia original com projeção médio-lateral oblíqua. (b) A mesma imagem de mamografia apenas com regiões internas à área da mama (exceto região com massa).

Então, o algoritmo para seleção de ROIs da classe não massa seleciona de forma aleatória uma região que não contenha um *pixel* com o valor 0 (zero).

3.1.2 Pré-processamento

Após a obtenção das amostras a partir da base DDSM, as mesmas passaram pelo processo de pré-processamento com o objetivo de realçar suas características. Neste processo foi utilizada a equalização do histograma, descrito na Seção 2.2.2, na região de interesse com objetivo de aumentar o contraste do objeto promovendo uma melhor descrição de sua textura. A Figura 14 apresenta um exemplo do processo de equalização do histograma.



(a) **(b)**

Figura 14 - (a) Massa original, (b) Massa após a equalização do histograma

Percebe-se um aumento da distinção visual das características. Logo, espera-se que os métodos de extração de característica possam extrair com mais facilidade as aparências dissimilares entre as classes de amostra aumentando a capacidade da acurácia geral desta metodologia.

3.1.3 Extração de Características

Após o pré-processamento das regiões de interesse, as mesmas são submetidas à etapa de extração de características de textura. Nesta etapa foi utilizado o Índice de Diversidade de Shannon-Wiener para descrever a textura dos objetos representantes de massa e não massa. A fim de obter-se melhor descrição de textura, a extração de características foi realizada em múltipla escala de tonalidades, utilizando a quantização da região de interesse em 6 níveis de cinza (2^8, 2^7, 2^6, 2^5, 2^4 e 2^3). Quatro abordagens foram utilizadas com este índice: global, circular, anéis e direcional. As próximas seções apresentam a maneira de como cada abordagem para extração de característica é utilizada neste trabalho.

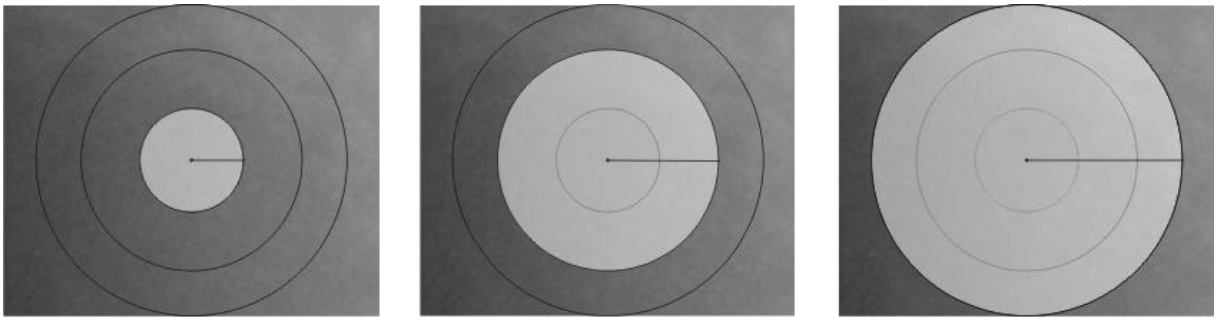
3.1.3.1 Índice de Diversidade de Shannon-Wiener com Abordagem Global

Na abordagem global, todos os pixels da região de interesse foram considerados para o cálculo do índice de diversidade. Isso significa que o valor obtido para o Índice de Diversidade de Shannon-Wiener corresponde à diversidade da região de interesse como um todo, não levando em consideração as variações locais de diversidade que possam existir entre diversas áreas da região. Nesta abordagem foi utilizada a equação original do Índice de Diversidade de Shannon-Wiener (Equação 1). Com esta abordagem, são obtidas no total 6 variáveis para cada região de interesse, sendo uma para cada nível de quantização.

3.1.3.2 Índice de Diversidade de Shannon-Wiener com Abordagem em Círculos

Nesta abordagem os valores do Índice de Shannon são extraídos em diferentes regiões da amostra. O objetivo desta abordagem é tentar descobrir padrões de diversidade entre as áreas mais próximas à borda da região examinada e as áreas mais internas. Nesta abordagem, extrai-se os valores de diversidade da região de interesse a partir de n círculos concêntricos e sobrepostos, com diferentes raios, partindo do centro da imagem.

Para definir o tamanho do raio foi utilizada a equação $R_i = i \times \frac{D}{2 \times n}$, onde $i = 1,2,\dots,n$; R_i é o raio i; n é o número de circunferências que representam as regiões de interesse e D é o tamanho do diâmetro do maior círculo inscrito na *bounding box* da amostra. Neste trabalho foi definido $n = 3$, pois nos testes preliminares os melhores resultados foram obtidos utilizando-se 3 círculos. A Figura 15 apresenta um exemplo onde as regiões de interesse são estabelecidas através de três áreas circulares, onde o valor do Índice de Diversidade de Shannon-Wiener é calculado para cada uma delas.



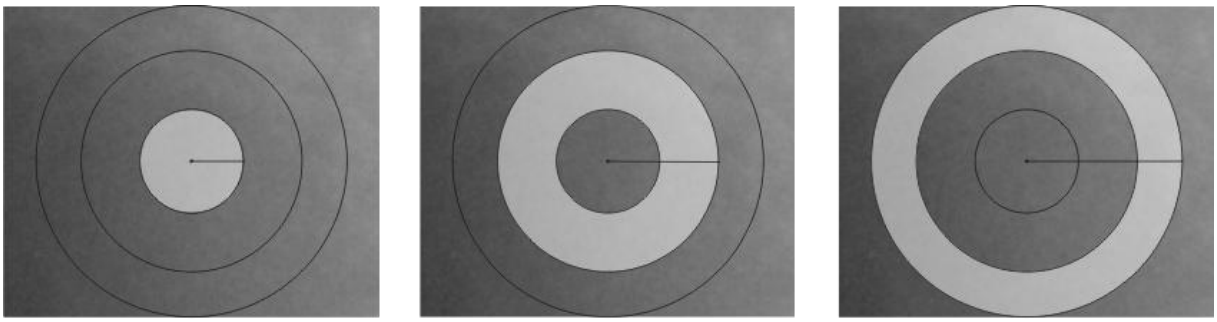
(a) (b) (c)

Figura 15 - *Pixels* da região de interesse tomados em áreas circulares (n=3).

Esta abordagem também utiliza a equação original do Índice de Diversidade de Shannon-Wiener (Equação 1) e gera um total de 18 variáveis (3 círculos × 6 quantizações).

3.1.3.3 Índice de diversidade de Shannon-Wiener com Abordagem em Anéis

Esta abordagem de extração de características em anéis é similar à abordagem circular. No entanto, nesta abordagem são utilizados dois raios consecutivos simultaneamente, levando em consideração apenas os *pixels* que estão dentro do anel formado pelos raios. Nesta abordagem foi definida a utilização de 3 anéis, pois este número apresentou os melhores resultados nos testes preliminares. A Figura 16 apresenta um exemplo onde as regiões de interesse têm os seus *pixels* tomados em três áreas em forma de anéis ($n = 3$) e o Índice de Diversidade de Shannon-Wiener é calculado em cada área formada por anéis.



(a) (b) (c)

Figura 16 - Pixels da região de interesse na abordagem em anéis

Esta abordagem também utiliza a equação original do Índice de Diversidade de Shannon-Wiener (Equação 1) e gera um total de 18 variáveis (3 anéis × 6 quantizações).

3.1.3.4 Índice de Diversidade de Shannon-Wiener Modificado (Abordagem Direcional)

Nos testes iniciais realizados com o Índice de Diversidade de Shannon-Wiener para extração de características nas abordagens global, em círculos e anéis, seguidos de classificação realizada pela SVM, percebeu-se que a capacidade de discriminação dessas

abordagens apresentavam dificuldades para a distinção entre as classes massa e não massa, apresentando uma especificidade muito baixa na abordagem global e sensibilidade baixa nas abordagens em círculos e em anéis. Na abordagem global para extração de características, presume-se que características locais importantes que poderiam diferenciar as duas classes podem estar sendo ignoradas. Nas abordagens em círculos e em anéis, embora os resultados preliminares tenham apresentado melhorias, suspeita-se que a divisão da região de interesse em áreas da forma que foi realizada também não foi suficiente para que características locais de textura pudessem ser detectadas. Sendo assim, com o objetivo de melhorar a capacidade de descrição do Índice de Diversidade de Shannon-Wiener e levar em consideração essas supostas características, propõe-se uma modificação. Esta modificação é baseada em uma abordagem direcional e em distâncias para substituir a diversidade das espécies (tons de cinza) por um índice que mede a diversidade de co-ocorrência de uma mesma espécie (tom de cinza) a uma distância d, medida ao longo de um ângulo, com o intuito de identificar através de autocorrelação espacial os padrões referentes a massa e não massa. O índice proposto foi denominado Índice de Diversidade de Shannon-Wiener Modificado (IDSWM).

A modificação proposta no cálculo do índice se dá conforme a Equação 15

$$h^{\theta,d} = \sum_{i=1}^{S} p_i^{\theta,d} \ln p_i^{\theta,d} \tag{15}$$

onde, $h^{\theta,d}$ é o IDSWM para direção θ e distância d, $p_i^{\theta,d}$ é a probabilidade de co-ocorrência de dois indivíduos da mesma espécie *i* distantes um do outro da distância *d* medida ao longo da direção θ.

Esta probabilidade é definida como:

$$p_i^{\theta,d} = \frac{nco_i^{\theta,d}}{N} w_d \tag{16}$$

sendo que *N* é o número total de indivíduos da amostra, $nco_i^{\theta,d}$ é a freqüência com que dois indivíduos da espécie i ocorrem na amostra a uma distância d medida ao longo da direção θ e w_d é um peso igual ao inverso da distância d entre dois indivíduos que diminui a importância da relação à medida que a distância entre os pixels observados aumenta.

Para o cálculo do $nco_i^{\theta,d}$, seja então uma ROI retangular I_G com *Px* pixels na direção horizontal, *Py* pixels na direção vertical e uma quantização em *G* níveis de cinza.

Denomina-se *Lx=1, 2,..., Px* o domínio espacial horizontal; *Ly=1, 2, ..., Py* o domínio vertical; e *G=8, 16, 32, ...,256* níveis de cinza.

A informação de co-ocorrência de pixels $nco_i^{\theta,d}$ pode ser caracterizada pelo vetor de co-ocorrência *P(i, d, θ)*, que representa a relação de dois *pixels* com nível de cinza *i* separados pela distância *d* e por uma relação angular θ na imagem. Este vetor tem tamanho igual à quantidade de tonalidades ou espécies presentes na amostra analisada. O cálculo do vetor de co-ocorrência é descrito pelas Equações 4 a 7 (adaptadas de (HARALICK *et al.*, 1973)), para as direções 0°, 45°, 90° e 135°, respectivamente, onde # denota o número de vezes que os pixels com determinado valor ocorrem como descrito anteriormente:

$$P(i,d,0°) = \#\{((k,l),(m,n)) \in f(k-m=0,|l-n|=d), f(k,l)=i, f(m,n)=i\} \quad (17)$$

$$P(i,d,45°) = \#\{((k,l),(m,n)) \in f(k-m=d,l-n=-d), f(k,l)=i, f(m,n)=i\} \quad (18)$$

$$P(i,d,90°) = \#\{((k,l),(m,n)) \in f(|k-m|=d,l-n=0), f(k,l)=i, f(m,n)=i\} \quad (19)$$

$$P(i,d,135°) = \#\{((k,l),(m,n)) \in f(k-m=d,l-n=d), f(k,l)=i, f(m,n)=i\} \quad (20)$$

O vetor de co-ocorrência é na realidade associado à diagonal da matriz de co-ocorrência. Assim, além da informação comum codificada pelo índice de Shannon, também consegue-se gerar informação de associação espacial na distribuição das espécies. A intenção desta proposta é aumentar a riqueza com que se descreve as características de distribuição de espécies na região espacial do indivíduo em estudo.

Cada imagem quantizada é submetida a uma análise de distância e direção. Esta análise é feita para cada pixel da ROI e cumulativamente representada pelo valor do índice calculado.

As características extraídas dos tecidos de mama (massa e não massa), consideradas como assinaturas de textura foram obtidas por um conjunto de quatro direções, correspondendo aos valores de θ iguais a 0°, 45°, 90° e 135°. Foi adotada ±22.5° como tolerância para medidas de ângulo. Com distância variando de 1 à 5 e tolerância de ±0.45. O número de características extraídas para cada amostra nesta abordagem foi de 5 *distâncias* × 4 *direções* × 6 imagens quantizadas, totalizando 120 características para cada amostra.

3.1.4 Classificação

A última etapa da metodologia proposta consiste em classificar os objetos nas classes massa e não massa, alimentando um classificador com os vetores obtidos nas

abordagens de extração de características global, em círculos, em anéis e direcional. Neste trabalho é utilizada a SVM com o objetivo de classificar as regiões de interesse extraídas em massa e não massa utilizando os vetores de características adquiridos na etapa anterior.

A base de imagens que é utilizada para treinamento e teste no classificador SVM é formada por todas as amostras obtidas da base DDSM. Para cada abordagem de extração de características utilizada a base de dados foi dividida em 50% para treinamento e 50% para testes de forma aleatória. Em cada amostra as medidas de textura foram extraídas e o conjunto de medidas recebeu um rótulo de acordo com a sua classe gerando uma base de características extraídas. Com a finalidade de ajudar o classificador a convergir com maior facilidade na etapa de treinamento, a base de características foi normalizada de maneira uniforme no intervalo $[-1,1]$.

Neste trabalho foi utilizado o núcleo radial que é comumente utilizado em trabalhos de reconhecimento de padrões. Trabalhos como de (BRAZ JUNIOR, 2008) e de (MARTINS, 2007) apresentam os melhores resultados utilizando-se o núcleo radial. Para utilização do núcleo radial foi necessário estimar os valores dos parâmetros C e γ.

O parâmetro C é ajustável e controla qual peso é dado na etapa de otimização à classificação incorreta de exemplos do conjunto de treinamento, estabelecendo um equilíbrio entre a complexidade do modelo e o erro do treinamento (Equação 4). Por outro lado, o parâmetro γ, também ajustável, geralmente é testado para se encontrar qual valor faz com que a SVM tenha melhor eficácia para cada problema (Equação 9).

Os valores dos parâmetros C e γ foram estimados através de uma busca exaustiva realizada pelo script em *phyton grid.py* que faz parte do pacote LIBSVM (CHANG e LIN, 2010). Esses parâmetros são obtidos com base apenas nas amostras de treinamento. Através deste script utiliza a validação cruzada tornando possível estimar os melhores parâmetros para a base que retorne como resposta o melhor percentual de acerto total sobre dados de teste e treino na validação cruzada.

Na etapa de treinamento gera-se o modelo com os vetores de suporte que é utilizado pela SVM na etapa de testes. A etapa de treinamento desconhece totalmente as amostras de teste, com a finalidade de assemelhar ao máximo com condições reais de testes. Após o modelo gerado, é possível realizar a etapa de reconhecimento de padrões com as amostras de teste separadas.

3.1.5 Validação dos Resultados

Após a etapa de reconhecimento de padrões utilizando o Índice de Shannon com as abordagens a extração de características citadas e o classificador SVM, é necessário validar os resultados e discutir possíveis melhorias. Com esta finalidade, neste trabalho são utilizadas métricas comumente usadas em sistemas CAD/CADx e aceitas pela comunidade para análise de desempenho de sistemas baseados em processamento de imagens. Tais métricas são especificidade, sensibilidade, acurácia, valor preditivo positivo e valor preditivo negativo (Seção 2.7).

O objetivo da utilização dessas métricas é medir o desempenho da metodologia utilizada neste trabalho como satisfatório ou não. Além disso, ajudam a apontar pontos positivos e negativos para melhoria futura deste trabalho.

4 RESULTADOS E DISCUSSÃO

Este capítulo tem o objetivo de apresentar e discutir os resultados obtidos com a utilização da metodologia proposta neste trabalho para classificação de regiões extraídas de imagens de mamografias em massa e não massa. São então, apresentados e discutidos os resultados obtidos utilizando quatro abordagens, cada uma utilizando o Índice de Diversidade de Shannon-Wiener calculado de maneira global, em círculos, em anéis e direcional. Em todas as abordagens foi utilizada a base de dados DDSM, conforme a Seção 3.1.1.

4.1 Resultados Obtidos

4.1.1 Abordagem Global

A abordagem global para extração de características utilizando o Índice de Diversidade de Shannon-Wiener, apresentada na Seção 3.1.3.1, gera um conjunto de 6 características (uma para cada quantização).

A partir da extração de características, a base de amostras foi organizada em dois grupos para serem utilizados no classificador SVM: base de treinamento e base de teste. Foram realizados 5 conjuntos treinamento/teste com proporção 50/50. A Tabela 1 apresenta os valores estimados para os parâmetros C e γ para cada conjunto de treinamento utilizando o núcleo radial no classificador SVM.

Tabela 1 - Parâmetros para SVM para uso com características extraídas com o Índice de Diversidade de Shannon-Wiener com abordagem global

Conjunto de Treinamento	C	γ
1	128	8
2	32768	0,5
3	32768	2
4	512	8
5	512	8

Após a estimação dos parâmetros C e γ, a etapa seguinte na metodologia proposta é a etapa de classificação e validação dos resultados. A Tabela 2 apresenta os resultados obtidos pelo classificador SVM utilizando os parâmetros acima.

Tabela 2 - Resultados obtidos na classificação entre massa e não massa utilizando o Índice de Diversidade Shannon-Wiener com abordagem global.

Conjunto de teste	VP	VN	FP	FN	S	E	VPP	VPN	A	*e*
1	1254	912	753	530	70,29%	54,77%	62,48%	63,44%	62,80%	37,20%
2	1157	1008	627	657	63,78%	61,65%	64,85%	60,54%	62,77%	37,23%
3	1238	903	781	527	70,14%	53,62%	61,31%	63,14%	62,07%	37,93%
4	1239	910	729	571	68,45%	55,52%	62,95%	61,44%	62,30%	37,70%
5	1227	897	794	531	69,79%	53,04%	60,71%	62,81%	61,58%	38,42%

O melhor resultado obtido utilizando o Índice de Diversidade de Shannon-Wiener com a abordagem global alcança a acurácia de 62,80%, onde os valores da sensibilidade e da especificidade neste caso foram, respectivamente, 70,29% e 54,77%. Entre os 5 casos de treinamento e teste realizados apresentados na Tabela 2, obteve-se uma acurácia média de 62,30%.

A quantidade de falsos-positivos (FP) e falsos-negativos (FN) apresentados na Tabela 2 mostram que muitas regiões de massa e não massa foram classificadas erroneamente. Isso demonstra que esta abordagem tem um baixo índice de confiança, como pode ser constatado através dos valores de VPP e VPN da Tabela 2.

4.1.2 Abordagem em Círculos

Na abordagem de extração de características em círculos o objetivo é tentar descobrir padrões de diversidade entre as áreas mais próximas da borda da região examinada e as áreas mais internas. Os valores dos raios de cada círculo são determinados conforme a equação $R_i = i \times \frac{D}{2 \times n}$, conforme visto na Seção 3.1.3.2. Neste trabalho foram utilizados 3 círculos concêntricos e sobrepostos onde o valor do Índice de Shannon foi calculado para cada imagem quantizada. No total, foi gerado um conjunto de 18 características (6 quantizações × 3 círculos).

Após a extração de características, a base de amostras foi organizada em dois grupos para serem utilizados pelo classificador SVM: base de treinamento e base de teste. Nesta abordagem, também foi utilizado o núcleo radial no classificador SVM e foi necessário estimar os valores para os parâmetros C e γ para cada grupo de treinamento. A Tabela 3 apresenta os valores estimados para estes parâmetros.

Tabela 3 - Parâmetros para SVM para uso com características extraídas com o Índice de Diversidade de Shannon-Wiener com abordagem em círculos.

Conjunto de treinamento	C	γ
1	32768	0,03125
2	32768	0,03125
3	32768	0,03125
4	32768	0,03125
5	2048	0,125

Com os parâmetros C e γ estimados, a próxima etapa na metodologia proposta é a de classificação e validação dos resultados. Os resultados obtidos com os parâmetros acima utilizados no classificador SVM nesta abordagem são apresentados na Tabela 4.

Tabela 4 - Resultados obtidos na classificação entre massa e não massa utilizando o Índice de Diversidade de Shannon-Wiener com abordagem em círculos.

Conjunto de teste	VP	VN	FP	FN	S	E	VPP	VPN	A	*e*
1	1255	1414	213	567	68,88%	86,90%	85,49%	71,37%	77,38%	22,62%
2	1269	1388	258	534	70,38%	84,32%	83,10%	72,21%	77,03%	22,97%
3	1259	1389	248	553	69,48%	84,85%	83,54%	71,52%	76,77%	23,23%
4	1268	1379	305	497	71,84%	81,88%	80,61%	73,50%	76,74%	23,26%
5	1263	1396	250	540	70,04%	84,81%	83,47%	72,10%	77,09%	22,91%

Conforme, pode-se observar na Tabela 4, o melhor resultado obtido nesta abordagem alcança a acurácia de 77,38%, onde a sensibilidade e a especificidade obtidas neste caso foram, respectivamente, 68,88% e 86,90%. A acurácia média entre os 5 casos apresentados na Tabela 4 é de 77,00%.

Na abordagem de extração de características em círculos, observa-se que a acurácia obteve uma significativa melhora em relação à abordagem global. No entanto, nota-se que muitas amostras ainda foram classificadas de forma errada. Este fato compromete o índice de confiança desta abordagem, como pode ser visto através dos valores de VPP e VPN na Tabela 4.

4.1.3 Abordagem em Anéis

A abordagem de extração de características em anéis é similar à abordagem em círculos, porém são utilizados dois raios consecutivos simultaneamente para formar os anéis. Os valores do raio de maior valor de cada anel também são determinados conforme a equação $R_i = i \times \frac{D}{2 \times n}$ (Seção 3.1.3.2). O raio menor é determinado pela equação $R_i = i \times \frac{D}{2 \times (n-1)}$. Nesta abordagem a área mais interna excepcionalmente corresponde a um círculo, pois utiliza apenas um raio. Neste trabalho foi definido o valor de n igual a 3, onde o valor do Índice de Shannon foi calculado para cada quantização da imagem. No total, foi gerado um conjunto de 18 características (6 quantizações × 3 círculos).

Após a extração de características, a base de amostras foi organizada em duas partes para formar a base de treinamento e a base de teste. O classificador SVM foi utilizado nesta abordagem com o núcleo radial. Os valores dos parâmetros C e γ estimados são apresentados na Tabela 5.

Tabela 5 - Parâmetros para SVM para uso com características extraídas com o Índice de Diversidade de Shannon-Wiener com abordagem em anéis.

Conjunto de treinamento	C	γ
1	128	0,5
2	32768	0,03125
3	8192	0,125
4	8192	0,125
5	8192	0,03125

Após a estimação dos parâmetros C e γ, a próxima etapa na metodologia proposta é a etapa de classificação e validação dos resultados. A Tabela 6 apresenta os resultados obtidos utilizando o classificador SVM com os parâmetros C e γ da Tabela 5.

Tabela 6 - Resultados obtidos na classificação entre massa e não massa utilizando o Índice de Diversidade de Shannon-Wiener com abordagem em anéis.

Conjunto de teste	VP	VN	FP	FN	S	E	VPP	VPN	A	e
1	1281	1373	287	508	71,60%	82,71%	81,69%	72,99%	76,94%	23,06%
2	1265	1408	236	540	70,08%	85,64%	84,27%	72,27%	77,50%	22,50%
3	1256	1461	220	512	71,04%	86,91%	85,09%	74,04%	78,77%	21,23%
4	1288	1391	260	510	71,63%	84,25%	83,20%	73,17%	77,67%	22,33%
5	1273	1418	266	492	72,12%	84,20%	82,71%	74,24%	78,02%	21,98%

Com a abordagem de extração de características em anéis, o melhor resultado obtido alcança a acurácia de 78,77%, com sensibilidade de 71,04% e especificidade de 86,91%. A taxa média de acerto entre os 5 casos apresentados é de 77,78%.

Esta abordagem apresenta um desempenho semelhante à abordagem de extração de características em círculos, onde a taxa de acerto média fica na casa dos 77%. Como visto na abordagem em círculos, na abordagem em anéis também houve uma grande quantidade de amostras classificadas erroneamente. Os baixos valores de VPP e VPN da Tabela 6 indicam o baixo índice de confiança desta abordagem.

4.1.4 Abordagem Direcional

A abordagem direcional para extração de características (apresentada na Seção 3.2.3.4) faz o cálculo do Índice de Diversidade de Shannon-Wiener baseado em distâncias e direções. Nesta abordagem, toma-se um *pixel* de referência e outro *pixel* alvo. Estes *pixels* estão distantes um do outro a uma distância *d* medida ao longo da direção θ. Para o cálculo do Índice de Shannon são considerados apenas os *pixels* que possuem a mesma tonalidade do *pixel* de referência. Esta abordagem gera um conjunto de 120 características (6 quantizações × 4 direções × 5 distâncias).

Após a extração de características, a base de amostras é dividida em duas partes de tamanhos iguais: base de treinamento e base de teste. Na classificação foi utilizado o classificador SVM com o núcleo radial, cujos valores dos parâmetros C e γ estimados são apresentados na Tabela 7.

Tabela 7 - Parâmetros para SVM para uso com características extraídas com o Índice de Diversidade de Shannon-Wiener com abordagem direcional.

Conjunto de treinamento	C	γ
1	32768	0,000488281
2	128	0,5
3	8192	0,03125
4	128	0,03125
5	8	0,5

Obtido os parâmetros C e γ, a próxima etapa na metodologia proposta é a classificação das amostras e a validação dos resultados. Os resultados obtidos na etapa de classificação com a SVM são apresentados na Tabela 8.

Tabela 8 - Resultados obtidos na classificação entre massa e não massa utilizando o Índice de Diversidade de Shannon-Wiener com abordagem direcional.

Conjunto de Teste	VP	VN	FP	FN	S	E	VPP	VPN	A	*e*
1	1772	1667	2	8	99,55%	99,88%	99,88%	99,52%	99,71%	0,29%
2	1762	1670	8	9	99,49%	99,52%	99,54%	99,46%	99,50%	0,50%
3	1767	1673	6	3	99,83%	99,64%	99,66%	99,82%	99,73%	0,27%
4	1782	1662	2	3	99,83%	99,87%	99,88%	99,81%	99,85%	0,15%
5	1795	1648	6	2	99,88%	99,63%	99,66%	99,87%	99,76%	0,24%

Na abordagem de extração de características direcional, o melhor resultado obtido alcança a acurácia de 99,85%, com sensibilidade de 99,83% e especificidade de 99,87%. A acurácia média entre os 5 casos apresentados na Tabela 8 é de 99,71%.

Os valores de VPP e VPN, apresentados na Tabela 8, indicam que esta abordagem possui um alto índice de confiança para classificar regiões de mamografias em massa e não massa.

Esta abordagem apresenta uma modificação no cálculo do Índice de Diversidade de Shannon-Wiener, onde apenas pares de *pixels* com a mesma tonalidade são considerados no cálculo do índice, obedecendo critérios de direção e distância (conforme apresentado na Seção 3.1.3.4). Pode-se perceber um ganho significativo da acurácia nesta abordagem em relação às três abordagens anteriores: global, em círculos e em anéis.

4.2 Resultados Finais

Esta seção tem o objetivo de discutir os principais resultados obtidos com as 4 abordagens utilizadas. Baseado na acurácia final, um resumo resultados obtidos que alcançaram a maior acurácia para cada abordagem é apresentado na Tabela 9.

Tabela 9 – Acurácia máxima obtida em cada abordagem utilizada neste trabalho.

Abordagem	Sensibilidade	Especificidade	Acurácia
Global	70,29%	54,77%	62,80%
Círculos	68,88%	86,90%	77,38%
Anéis	71,04%	86,91%	78,77%
Direcional	99,83%	99,87%	99,85%

Analisando os resultados obtidos, verifica-se que a abordagem direcional foi bastante superior às demais abordagens utilizadas neste trabalho. Considerando as taxas de sensibilidade, especificidade e acurácia, os resultados obtidos com essa abordagem foram, de forma geral, bastante estáveis. O melhor resultado com a abordagem direcional obteve sensibilidade de 99,83%, especificidade de 99,87% e acurácia de 99,85%. A acurácia da abordagem direcional apresentou uma diferença de mais de 21 pontos percentuais em relação à acurácia da abordagem em anéis.

As abordagens de extração de características em anéis e em círculos apresentaram resultados semelhantes. Conforme os valores apresentados na Tabela 10, pode-se observar que a sensibilidade e a especificidade média dessas duas abordagens apresentam valores semelhantes. Essas abordagens apresentam um desempenho regular para descrever tecidos não massa, porém para descrever tecidos de massa o desempenho cai significativamente.

Tabela 10 - Valores médios para sensibilidade, especificidade e acurácia em cada abordagem utilizada neste trabalho

Abordagem	Sensibilidade	Especificidade	Acurácia
Global	68,49%	55,72%	62,30%
Círculos	70,12%	84,55%	77,00%
Anéis	71,29%	84,74%	77,78%
Direcional	99,71%	99,71%	99,71%

A abordagem global possui o valor médio da sensibilidade bem próximo às abordagens em círculos e em anéis, no entanto o valor médio da especificidade é muito inferior, quase 10 pontos percentuais em relação ao valor da especificidade média da abordagem em círculos e em anéis. Os baixos valores encontrados na abordagem global indicam um baixo índice de confiabilidade. A Figura 17 apresenta uma breve comparação entre os valores encontrados nas 4 abordagens utilizadas neste trabalho.

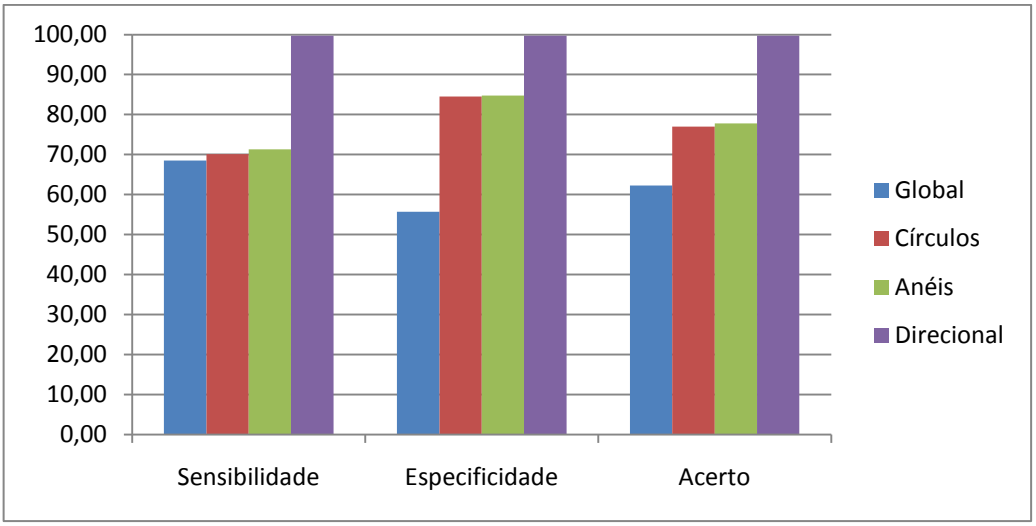


Figura 17 - Comparação entre os valores médios das 4 abordagens utilizadas neste trabalho.

A Figura 18 apresenta a aplicação da abordagem de extração global utilizando o Índice de Diversidade de Shannon-Wiener em dez amostras (5 não massas e 5 massas) selecionadas aleatoriamente na base de dados. As variáveis *c1*, *c2*, *c3*, *c4*, *c5* e *c6* do gráfico da Figura 18 representam as seis características extraídas de cada amostra, ou seja, são os índices calculados para cada quantização em 2^3, 2^4, 2^5, 2^6, 2^7 e 2^8 níveis de cinza, respectivamente. Analisando o gráfico da Figura 18, pode-se notar um indicativo da dificuldade de discriminação entre os vetores de características referentes a massa e não massa, uma vez que o gráfico demonstra que algumas amostras da classes não massa apresentam o mesmo comportamento de uma da classes massa e vice-versa.

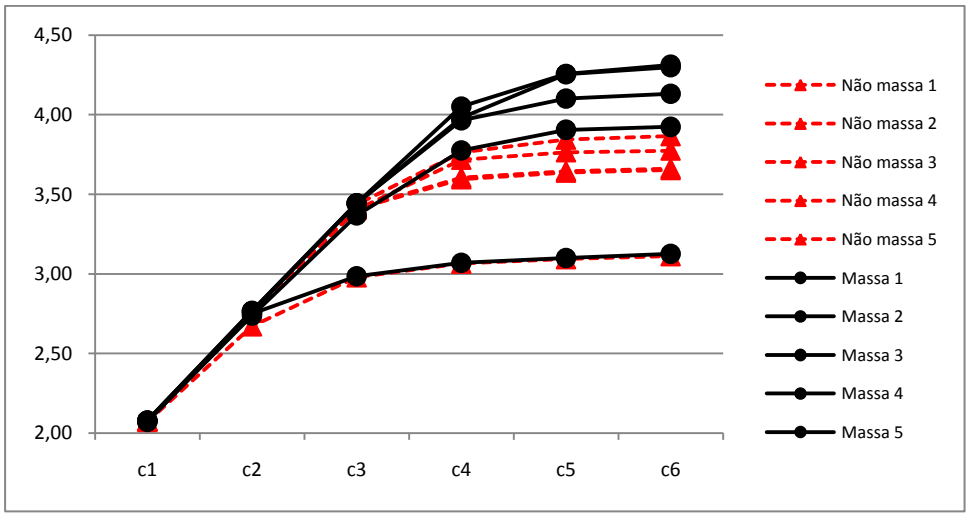


Figura 18 - Comparação entre as características extraídas com a abordagem global a partir de dez amostras selecionadas aleatoriamente. As linhas tracejadas representam a classe não massa e as linhas contínuas representam a classe massa.

O gráfico da Figura 19 apresenta a aplicação de extração de características utilizando a abordagem em círculos. Este gráfico apresenta dez amostras selecionadas aleatoriamente, sendo 5 amostras da classe massa e 5 da classe não massa . As variáveis de *c1* a *c18*, da mesma forma que o gráfico anterior, representam as 18 variáveis extraídas de cada amostra. Para cada quantização há três valores de índices calculados (um para cada círculo). Os valores *c1*, *c2* e *c3* correspondem aos índices calculados em 2^3 níveis de cinza, *c4*, *c5* e *c6* correspondem aos índices calculados em 2^4 níveis de cinza, e assim por diante.

No gráfico da Figura 19, pode-se verificar que os valores das amostras da classe não massa têm comportamentos relativamente semelhantes. O comportamento mais definido da classe não massa pode ser um indicativo de que esta abordagem é melhor para discriminar esta classe. Isto pode visto no valor da especificidade média da abordagem de 84,55%. Por outro lado, os valores das amostras da classe massa apresentam comportamento bem diferente. Isto pode ser um indicativo de que esta abordagem apresenta dificuldades para discriminar amostras de massa e refletindo em sua sensibilidade média de 70,12%.

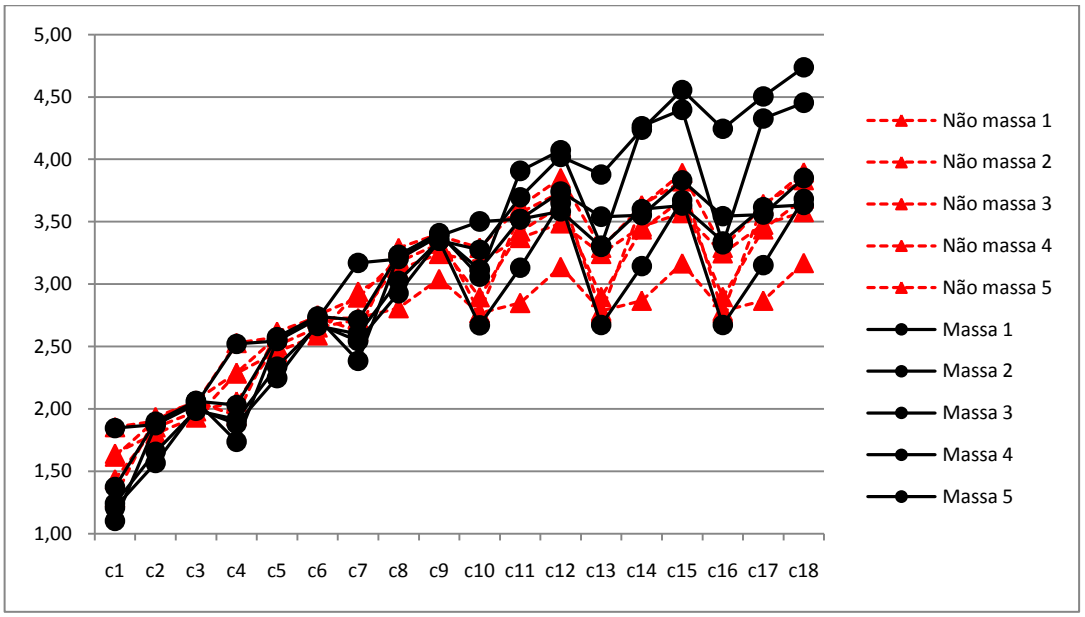


Figura 19 - Comparação entre as características extraídas com a abordagem em círculos a partir de dez amostras selecionadas aleatoriamente. As linhas tracejadas representam a classe não massa e as linhas contínuas representam a classe massa.

O gráfico da Figura 20 apresenta a aplicação de extração de características utilizando a abordagem em anéis. Neste gráfico há cinco amostras de cada classe (massa e não massa) selecionadas aleatoriamente. As variáveis de *c1* a *c18* representam as 18 variáveis extraídas de cada amostra, ou seja, são os valores do índice calculado para a quantização. Para

cada quantização há três valores de índices calculados (um para cada anel). Os valores *c1*, *c2* e *c3* correspondem aos índices calculados em 2^3 níveis de cinza, *c4*, *c5* e *c6* correspondem aos índices calculados em 2^4 níveis de cinza, e assim por diante.

No gráfico da Figura 20 pode-se perceber que os valores das características das amostras da classe massa não possuem um comportamento distinto em relação à classe não massa. Os valores da classe não massa, por sua vez, apresentam um comportamento mais semelhantes entre si. Portanto, a exemplo do gráfico da Figura 19, considera-se este gráfico como um indicativo da dificuldade de discriminação desta abordagem para amostras referentes à classe massa. Isto pode ser verificado na sensibilidade média da abordagem com valor de 71,29%.

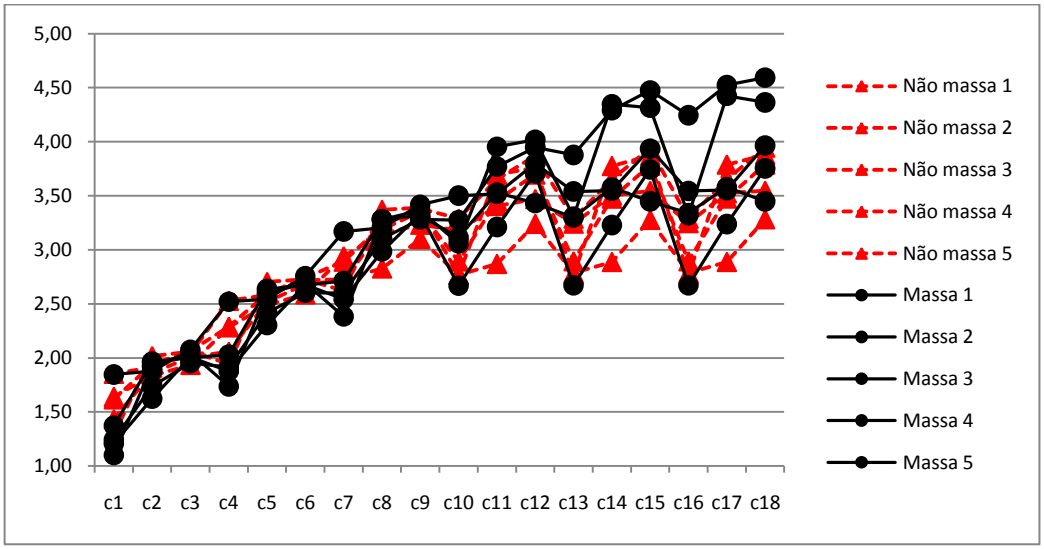


Figura 20 - Comparação entre as características extraídas com a abordagem em anéis a partir de dez amostras selecionadas aleatoriamente. As linhas tracejadas representam a classe não massa e as linhas contínuas representam a classe massa.

O gráfico da Figura 21 apresenta os valores da abordagem direcional aplicada a dez amostras selecionadas aleatoriamente (5 de massas e 5 de não massas) na direção 0° e distâncias 1 a 5 (representadas por *c1*, *c2*, *c3*, *c4* e *c5*) no nível de tonalidade 2^8. Para as demais direções e distâncias em todos os níveis de tonalidade, as características seguem um comportamento semelhante, por isso não foram colocadas no gráfico. Os valores das amostras da classe massa e da não massa apresentam comportamentos bem distintos. O gráfico da Figura 21 é um indicativo de que a abordagem direcional possui uma maior capacidade de discriminação para as classes massa e não massa.

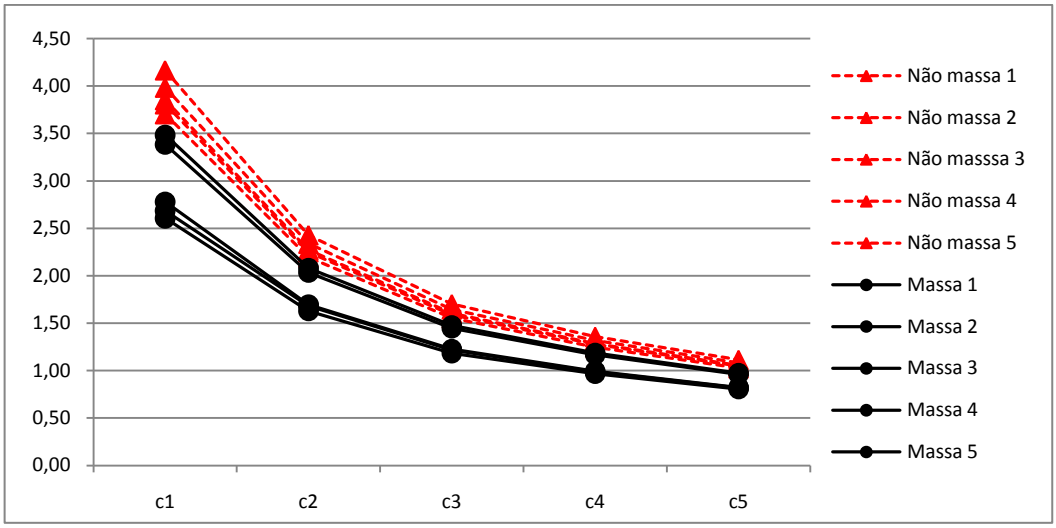


Figura 21 - Comparação entre as características extraídas com a abordagem em anéis a partir de dez amostras selecionadas aleatoriamente. As linhas tracejadas representam a classe não massa e as linhas contínuas representam a classe massa.

Além das análises dos gráficos, observou-se que nas abordagens de extração de características global, em círculos e em anéis a SVM criou funções de classificação complexas para discriminar padrões de massas e não massas. Isso pode ser notado verificando-se o número de vetores de suporte utilizados para fazer a classificação na tablea.

Tabela 11 - Número de vetores de suporte em cada abordagem.

Abordagem	Número de vetores de suporte	
	Não massa	Massa
Global	1339	1338
Círculos	890	895
Anéis	891	899
Direcional	17	17

Por outro lado, a abordagem direcional criou uma função de classificação simples, utilizando um número bem reduzido de vetores de suporte em relação às outras três abordagens.

A Tabela 12 apresenta uma breve comparação entre os resultados encontrados neste trabalho e alguns trabalhos citados na Seção 1 que realizam a classificação de tecidos de imagens de mamografia em massa e não massa. Observa-se que o resultado encontrado com a abordagem direcional alcançou uma acurácia comparável com os melhores resultados

publicados na literatura recente para classificação de tecidos de mamografia nas classes massa e não massa.

Tabela 12 – Comparação com alguns trabalhos referentes à classificação de tecidos de mama de imagens mamográficas em massa e não massa.

Trabalhos	Base de Dados	Acurácia (%)
(Moayedi et al., 2010) (SEL weighted SVM) (Etapa de classificação)	*MIAS*	*96,60*
(Moayedi et al., 2010) (SVFNN) (Etapa de classificação)	*MIAS*	*91,50*
(Moayedi et al., 2010) (kernel SVM) (Etapa de classificação)	*MIAS*	*82,10*
(Eltoukhy et al., 2010) (Etapa de classificação)	*MIAS*	*98,59*
(Faye et al., 2009) (Etapa de classificação)	*MIAS*	*98,55*
(Braz et al., 2009) (Coeficiente de Geary)	*DDSM*	*96,04*
(Braz et al., 2009) (Índice de Moran)	*DDSM*	*99,39*
(Martins et al., 2009) (Etapa de Classificação)	*DDSM*	*89,30*
(Nunes et al., 2010) (Etapa de classificação)	*DDSM*	*83,94*
Nosso método (Abordagem Global)	***DDSM***	***62,80***
Nosso método (Abordagem em Círculos)	***DDSM***	***77,38***
Nosso método (Abordagem em Anéis)	***DDSM***	***78,77***
Nosso método (Abordagem Direcional)	***DDSM***	***99,85***

A comparação entre os resultados obtidos com a utilização do Índice de Diversidade de Shannon Wiener Modificado (Abordagem Direcional) e os resultados encontrados em (BRAZ JUNIOR *et al.*, 2009) merecem destaque por terem utilizado as mesmas amostras de tecidos de mama representantes da classe massa. Neste trabalho, verifica-se que a utilização do Índice de Diversidade de Shannon-Wiener Modificado obteve uma pequena melhoria na acurácia, alçando 99,85%, enquanto o trabalho de (BRAZ JUNIOR *et al.*, 2008) obteve, utilizando o Índice de Moran, a acurácia de 99,39%.

5 CONCLUSÃO

Este trabalho apresentou a viabilidade do uso do Índice de Diversidade de Shannon-Wiener e Máquinas de Vetores de Suporte para discriminação e classificação de regiões de mamografias como massa e não massa.

A metodologia apresentou 4 abordagens com o Índice de Diversidade de Shannon-Wiener para extração de características (global, em círculos, em anéis e direcional) e comparou os resultados obtidos na classificação com a SVM, utilizando 50% das amostras na base de treinamento e a outra metade na base de teste.

Os melhores resultados no desempenho conjunto de sensibilidade, especificidade e acurácia foram obtidos com a abordagem direcional. Nesta abordagem, o Índice de Diversidade de Shannon-Wiener sofreu uma modificação com objetivo de se utilizar técnicas de extração de características de textura espacial, estes resultados evidenciaram o desempenho promissor dessas técnicas em conjunto com o classificador SVM.

Nas abordagens de extração de características global, em círculos e em anéis foi utilizado o Índice de Diversidade de Shannon-Wiener original da Equação 1, onde o índice era calculado dentro de uma determinada área conforme a abordagem. As abordagens em círculos e anéis apresentaram resultados semelhantes, porém bem inferiores aos resultados da abordagem direcional. A abordagem global foi a que obteve os piores resultados, apresentando, assim, um baixo nível de confiabilidade.

No entanto, mesmo com os resultados promissores obtidos com a abordagem direcional, se faz necessário aumentar a variabilidade das amostras de mamografia para alcançar uma metodologia robusta e genérica. Porém, 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 no Índice de Diversidade de Shannon-Wiener para descrição de texturas possam ser desenvolvidas.

Alguns aspectos deste trabalho podem ser melhorados no futuro e não puderam ser concluídos ou inclusos. O primeiro é extrair características de tecidos de imagens de mamografia da classe massa sem a necessidade de se utilizar o contorno da área da massa dos arquivos *overlay* da base DDSM. Outro aspecto que pode ser acrescentado a este trabalho é a

expansão dos resultados de classificação de massa e não massa para também classificação de massas malignas e benignas realizando o diagnóstico assistido por computador.

A fim de tentar melhorar os resultados, futuramente pode-se acrescentar à metodologia a etapa de seleção de características sobre os dados extraídos das imagens, utilizando algum método como Análise Discriminante *stepwise* ou Análise de Componentes Principais. Esta etapa tem o objetivo de eliminar as características que não contribuem ou que aumentam a probabilidade de erros na etapa de classificação.

Neste trabalho, a classificação das amostras foi feita pelo classificador SVM, porém a SVM pode ser substituída por outros classificadores com o objetivo de avaliar seus desempenhos na tarefa de reconhecimento de padrões de massa e não massa em regiões extraídas de mamografias.

Finalmente, a metodologia apresentada neste trabalho poderá integrar uma ferramenta CAD para ser aplicada em casos atuais na detecção e tratamento de câncer de mama. Da mesma forma, depois de acrescentada e validade a etapa de classificação em regiões malignas e benignas, a metodologia também poderá fazer parte de um sistema CADx com o objetivo de classificar as regiões detectadas como massa.

REFERÊNCIAS

AMERICAN CANCER SOCIETY (2010a). Breast Cancer. Disponível em: http://www.cancer.org.

AMERICAN CANCER SOCIETY (2010b). Mammograms and Other Breast Imaging Procedures. Disponível on-line em: http://www.cancer.org/Healthy/FindCancerEarly/ExamandTestDescriptions/MammogramsandOtherBreastImagingProcedures/index

BLAND, M. e OVID TECHNOLOGIES Inc. (2000). An introduction to medical statistics. Oxford University Press, v 132.

BIAZÚS, J. V. (2000). Rotinas em Cirurgia Conservadora da Mama. Porto Alegre: Artes Médicas. V. 1. 128 p.

BRAZ JUNIOR, G., PAIVA, A. C., SILVA, A. C., OLIVEIRA, A. C. M. (2009). Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM, Computers in Biology and Medicine, 39(12), 1063-1072.

BRAZ JUNIOR, G. (2008). Classificação de Regiões de Mamografias em Massa e Não Massa usando Estatística Espacial e Máquina de Vetores de Suporte. Tese de Mestrado. Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão.

CHANG, C. C. e LIN, C. J. (2010). LIBSVM – A Library for Support Vector Machines. Disponível on-line em: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

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

CONCI, A., AZEVEDO, E., LETA, F. R. (2008). Computação Gráfica: Teoria e Prática. Volume 2. Ed. Campus, Rio de Janeiro.

CRISTIANINI, N. e SHAWE-TAYLOR, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

ELTOUKHY, M. M.; FAYE, I.; SAMIR, B. B. (2010a). Breast cancer diagnosis in digital mammogram using multiscale curvelet transform, Computerized Medical Imaging and Graphics, Volume 34, Issue 4, June 2010, Pages 269-276, ISSN 0895-6111, DOI: 10.1016/j.compmedimag.2009.11.002.

ELTOUKHY, M. M.; FAYE, I.; SAMIR, B. B. (2010b). A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram, Computers in Biology and Medicine, Volume 40, Issue 4, April 2010, Pages 384-391, ISSN 0010-4825, DOI: 10.1016/j.compbiomed.2010.02.002.

FAYE, I.; SAMIR, B. B.; ELTOUKHY, M. M. (2009). Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method, Second International Conference on Computer and Electrical Engineering, v. 2, p. 318-322.

GONZALEZ, R. C. e WOODS, R. E. (2002). Digital Image Processing. 2nd ed. Prentice Hall. Upper Sadde River, New Jersey.

HARALICK, R. M., SHANMUGAM, K., DINSTEIN, I. (1973). Textural Features for Image Classification. IEEE Transactions on Systems, Man, and Cybernetics, 3(6), 610-621.

HAYKIN, S. (2007). Redes Neurais: Princípios e Prática. Ed. 2. Bookman.

HEATH, M., BOWYER, K., KOPANS, D., MOORE, R., KEGELMEYER, W. P. (2000). The Digital Database for Screening Mammography, In: Proceedings of the Fifth International Workshop on Digital Mammography, Medical Physics Publishing.

INTERNATIONAL AGENCY FOR RESEARCH ON CANCER (2009). Biennal Report 2008/2009. Disponível on-line em: http://governance.iarc.fr/SC/SC46/SC46_2Text.pdf. Último acesso: 27/03/2011.

INSTITUTO NACIONAL DO CÂNCER (2010). Estimativas 2010: Incidência de Câncer no Brasil. Disponível em: http://www.inca.gov.br.

LEVINE, N. (1996). Spatial Statistic and GIS: Software Tools to Quantify Spatial Patterns. Journal of the American Planning Association. V. 62, n. 3, p. 381-391.

LOONEY, C.G. (1997). Pattern Recognition using Neural Networks: Theory and Algorithms for Engineers and Scientists. Oxford University Press, Inc. New York, NY, USA.

MAGURRAN, A. E. (2004). Measuring Biological Diversity. Blackwell Science, Oxford.

MAMOWEB (2011). Disponível em: http://lapimo.sel.eesc.usp.br/lapimo/portal/index.html. Último acesso: 25/01/2011.

MARTINS, L. O., SILVA, A. C., PAIVA, A. C., GATTASS, M. (2009a). Detection of Breast Masses in Mammogram Images Using Growing Neural Gas Algorithm and Ripley's K Function. J. Sign. Process. Syst., 55(1), 77–90.

MARTINS, L. O., BRAZ JUNIOR, G., SILVA, A. A., PAIVA, A. C. e GATTASS, M. (2009b). Detection of Masses in Digital Mammograms using K-means and Support Vector Machine. Electronic Letters on Computer Vision and Image Analysis, 8(2): 39-50.

MARTINS, L. O. (2007). Detecção de Massas de Imagens Mamográficas através do Algoritmo Growing Neural Gas e da Função K de Ripley. Tese de Mestrado. Curso de Pós-Graduação em Engenharia de Eletricidade da Universidade Federal do Maranhão.

MARQUES FILHO, O.; VIEIRA NETO, H. (1999). Processamento Digital de Imagens, Rio de Janeiro: Brasport. ISBN 8574520098.

MATHWORKS (2011). Adjusting Pixel Intensity Values. Disponível on-line em: http://www.mathworks.com/help/toolbox/images/f11-14011.html. Último acesso: 07/02/2011.

MINISTÉRIO DA SAÚDE (2004). Controle de câncer de mama: documento de consenso. Disponível em: http://www.inca.gov.br/publicacoes/Consensointegra.pdf. Último acesso: 20/01/2011.

MOAYEDI, F., AZIMIFAR, Z., BOOSTANI, R., KATEBI, S. (2010). Contourlet-based mammography mass classification using the SVM family. Computers in Biology and Medicine, 40(4), 373-383.

MORSE, B. S. (2000). Data Structures for Image Analysis. Brigham Young University. Disponível em: http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/MORSE/data-structures.pdf. Acessado: 08/02/2011.

NUNES, A. P., SILVA, A. C., PAIVA, A. C. (2010). Detection of masses in mammographic images using geometry, Simpson's Diversity Index and SVM. International Journal of Signal and Imaging Systems Engineering, 3(1), 40-51.

PAIVA, J. A. C., RODRÍGUEZ, A. e CORREIA, V. R. M. (1999). Métodos Computacionais para Analisar Padrões de Pontos Espaciais. Instituto Nacional de Pesquisas Espaciais. Disponível em http://www.dpi.inpe.br/geopro/trabalhos/gisbrasil99/estat_pontos/

SPELLERBERG, I. F., FEDOR, P. J. (2003). A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. Global Ecology and Biogeography, 12(3), 177–179.

SUCKLING, J. *et al*. (1994). The Mammographic Image Analysis Society Digital Mammogram Database, Exerpta Medical, 1069, 375-378.

VANI, G.; SAVITHA, R.; SUNDARARAJAN, N. (2010). Classification of abnormalities in digitized mammograms using Extreme Learning Machine. Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on , vol., no., pp.2114-2117, 7-10 Dec. 2010 doi: 10.1109/ICARCV.2010.5707794.

VAPNIK, V. N. (1998). Statistical Learning Theory. Wiley New York., p. 736.

VERMA, B., McLEOD, P., KLEVANSKY, A. (2009). A novel soft cluster neural network for the classification of suspicious areas in digital mammograms, Pattern Recognition, Volume 42, Issue 9, September 2009, Pages 1845-1852, ISSN 0031-3203, DOI: 10.1016/j.patcog.2009.02.009

WORLD HEALTH ORGANIZATION (2011). World Health Organization – Cancer. Disponível on-line em: http://www.who.int/cancer/en/. Último acesso: 27/03/2011.

WIKIMEDIA COMMONS (2010). Disponível on-line em: http://commons.wikimedia.org/wiki/File:Breast.svg