

**UNIVERSIDADE FEDERAL DO PIAUÍ – UFPI
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**CLASSIFICAÇÃO AUTOMÁTICA DE MASSAS EM IMAGENS MAMOGRÁFICAS
USANDO ÍNDICES DE DIVERSIDADE FUNCIONAL.**

Marcelo de Sousa Damasceno

**PICOS – PIAUÍ
2017**

Marcelo de Sousa Damasceno

**CLASSIFICAÇÃO AUTOMÁTICA DE MASSAS EM IMAGENS MAMOGRÁFICAS
USANDO ÍNDICES DE DIVERSIDADE FUNCIONAL.**

Monografia submetida ao Curso de Bacharelado de Sistemas de Informação como requisito parcial para obtenção de grau de Bacharel em Sistemas de Informação.

Orientador: Prof.^a Ma. Patricia Medyna Lauritzen de Lucena Drumond

FICHA CATALOGRÁFICA
Serviço de Processamento Técnico da Universidade Federal do Piauí
Biblioteca José Albano de Macêdo

D155c Damasceno, Marcelo de Sousa
 Classificação automática de massas em imagens / Marcelo de
 Sousa Damasceno.– 2017.
 CD-ROM : il.; 4 ¾ pol. (34f.)
Trabalho de Conclusão de Curso (Curso Bacharelado em Sistemas
de Informação) – Universidade Federal do Piauí, Picos, 2017.
Orientador(A): Prof.^a Ma. Patrícia Medyna Lauritzen de Lucena
Drumond.

1. Processamento Digital de Imagens. 2.Câncer de Mama.
3. Índice de Diversidade Funcional. I. Título.

CDD 005.118

CLASSIFICAÇÃO DE MASSAS EM IMAGENS MAMOGRÁFICAS USANDO ÍNDICES
DE DIVERSIDADE FUNCIONAL

MARCELO DE SOUSA DAMASCENO

Monografia aprovada como exigência parcial para obtenção do grau de
Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI. 19 de janeiro de 2017



Prof.^a. Ma. Patrícia Medyna Lauritzen de Lucena Drumond
Orientador



Prof.^a. Ma. Patricia Vieira da Silva Barros
Membro



Prof. Dr. Antonio Oséas de Carvalho Filho
Membro

Aos meus pais, amigos e em memória do meu avô,
José Alves de Sousa, que Deus o tenha seu Zeca.

AGRADECIMENTOS

A Prof^a Ma. Patricia Medyna Lauritzen de Lucena Drumond pelo incentivo, simpatia, confiança e presteza no auxílio às atividades e discussões sobre o andamento e normatização desta Monografia de Conclusão de Curso.

Especialmente aos professores Antonio Oséas de Carvalho Filho e Alcilene Dalília de Sousa, pela oportunidade que me foi dada na ICV, por todo suporte e conhecimento que me repassaram.

A minha mãe Célia Maria Silva de Sousa Damasceno, por todo sacrifício e suor derramado para educar seus filhos.

Aos meus irmãos Maurício, Marcia Caroline e Karine por todos os conselhos e momentos de alegrias que passamos.

Aos demais idealizadores, coordenadores e funcionários da Universidade Federal do Piauí.

Aos colegas de classe pela espontaneidade e alegria na troca de informações e materiais numa rara demonstração de amizade e solidariedade.

Ao meu amigo e irmão Laurentino Moura por todo apoio.

E, finalmente, a DEUS pela oportunidade e pelo privilégio que nos foram dados em compartilhar tamanha experiência e, ao frequentar este curso, perceber e atentar para a relevância de temas que não faziam parte, em profundidade, das nossas vidas.

A todos vocês muito obrigado!

“Se não queres que ninguém saiba, não o faças.”

Proverbio Chinês

RESUMO

O câncer de mama atualmente é o mais comum em pacientes do sexo feminino e o segundo com maior taxa de mortalidade, ocasionado principalmente pelo diagnóstico e tratamento tardios. Vários sistemas de detecção auxiliados por computador (*Computer Aided Detection - CAD*), são desenvolvidos com o intuito de auxiliar os especialistas da área, automatizando o processo de detecção e diagnóstico. Este trabalho apresenta uma nova metodologia de extração de características e classificação de massas em imagens mamográficas. A nova metodologia consiste em três etapas, sendo a primeira, a aquisição de imagens do banco de imagens *Digital Database for Screening Mammography (DDSM)*. Na segunda etapa é realizada a extração de características de textura utilizando os índices de diversidade funcional (*Functional Diversity - FD*). Por fim, na terceira etapa, utiliza-se o classificador Máquina de Vetor de Suporte (*Support Vector Machine - SVM*), para a classificação das regiões de interesse (*Regions of interest - ROIs*) em massas e não massas. Nos resultados obtidos alcançou-se uma média de sensibilidade de 92,7%, média de especificidade de 96,24%, média de acurácia de 94,7% e área média das curvas ROC de 0,922.

Palavras-Chave: Processamento Digital de Imagens; Câncer de Mama; Índice de Diversidade Funcional;

ABSTRACT

Breast cancer is more common in female patients and the second with the highest mortality rate, which is caused mainly by the late diagnosis and treatment. The most systems of computer aided detection (CAD), are developed in order to assist the specialists, by automating the process of detection and diagnosis. This paper presents a new method of extraction of characteristics and classification of masses in mammography images. The new methodology consists of three steps, the first being the acquisition of images from Digital Database for Screening Mammography (DDSM). In the second step is performed texture features extraction using the functional diversity indices (FD). Finally, in the third step, the use Support Vector Machine (SVM) classifier, for the classification of regions of interest (ROIs) in masses and not masses. The results obtained has been an average of 92.7%, average of 96.24% specificity, accuracy average of 94.7%, and average area of ROC curves of 0.922.

Keywords: Digital image processing; Breast cancer; Functional Diversity index;

LISTA DE ILUSTRAÇÕES

Figura 1 - Tipos de anomalias em imagens mamográficas.....	17
Figura 2 - Mamógrafo.....	17
Figura 3 - Resultados das projeções de uma mamografia	18
Figura 4 - Etapas necessárias para o processamento de imagens.....	19
Figura 5 - Distâncias e direções para uma nova matriz de coocorrência.....	21
Figura 6 - Construção da matriz de coocorrência.....	21
Figura 7 - Cálculo da medida funcional.....	23
Figura 8 - Correlação entre os termos da biologia e o trabalho.....	23
Figura 9 - Etapas da metodologia proposta nesse trabalho.....	28
Figura 10 - Etapa de extração de características.....	29

LISTA DE TABELAS

Tabela 1 - Resultados Obtidos em múltiplos classificadores.....	31
Tabela 2 - Comparação da metodologia com os trabalhos relacionados.....	32

LISTA DE ABREVIATURAS E SIGLAS

A	Acurácia
CAD	Computer Aided Detection
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DDSM	Digital Database for Screening Mammography
E	Especificidade
FD	Functional diversity
IARC	International Agency for Research on Cancer
MD	Matriz de Distâncias
INCA	O Instituto Nacional de Câncer José Alencar Gomes da Silva
PDI	Processamento digital de Imagens
ROC	Receiver Operating Characteristics
ROI	Region of interest
S	Sensibilidade
SVM	Support Vector Machine
VC	Vetor de Características
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1	INTRODUÇÃO	14
1.1	OBJETIVOS.....	15
1.2	ORGANIZAÇÃO DO TRABALHO.....	15
2	FUNDAMENTAÇÃO TEÓRICA	16
2.1	CANCER DE MAMA.....	16
2.1.1	Mamografia	16
2.2	PROCESSAMENTO DE IMAGENS	18
2.2.1	Textura	20
2.2.1.1	<i>MATRIZ DE COOCORRÊNCIA</i>	<i>20</i>
2.2.1.1	<i>MEDIDA DE DIVERSIDADE FUNCIONAL.....</i>	<i>21</i>
2.3	RECONHECIMENTO DE PADRÕES E VALIDAÇÃO DE RESULTADOS	24
2.3.1	Máquina de vetor de suporte.....	24
2.3.2	Métricas de desempenho.....	24
2.4	TRABALHO RELACIONADOS	26
3	METODOLOGIA	28
3.1	AQUISIÇÃO DE IMAGENS	28
3.2	EXTRAÇÃO DE CARACTERÍSTICAS.....	29
3.3	CLASSIFICAÇÃO	30
3.4	VALIDAÇÃO DOS RESULTADOS.....	30
4	RESULTADOS E DISCUSSÕES.....	31
5	CONSIDERAÇÕES FINAIS	33
6	REFERÊNCIAS BIBLIOGRÁFICAS	34

1 INTRODUÇÃO

O câncer é considerado umas das principais causas de morbidade¹ e mortalidade que afetam a população em todo o mundo. Segundo relatório *World cancer report (2014)* da *International Agency for Research on Cancer (IARC)*, no ano de 2012 foram registrados 14 milhões de novos casos, 8 milhões de mortes causados pelo câncer e 32.6 milhões de pessoas convivem com a doença após cinco anos do seu diagnóstico. O indicativo para o ano de 2025 é que o número de incidências ultrapasse os 20 milhões (STEWARD; WILD, 2014).

Os principais responsáveis por esses dados estatísticos alarmantes e que vem crescendo nos últimos anos, ainda são fatores de riscos externos como o consumo excessivo de álcool, tabaco, alimentos processados, sedentarismo, obesidade ou qualquer item associado a um estilo de vida desequilibrado (STEWARD; WILD, 2014).

No Brasil, segundo o Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA, 2016), estipula-se que no biênio 2016-2017, ocorrerão 420 mil novos casos de câncer, sendo que o tipo mais frequente nos homens será o câncer de próstata (61 mil) e nas mulheres o de mama (58 mil).

O câncer de mama é o tipo que apresenta maior taxa de incidência e mortalidade na população feminina em todo o mundo, sendo também, o segundo mais frequente dentre os demais tipos de câncer. Estima-se que 1,7 milhões de novos casos e 522 mil mortes relacionadas a essa neoplasia², foram registrados no ano de 2012. (STEWARD; WILD, 2014). Apesar de ser um câncer de bom prognóstico, quando diagnosticado cedo, os números de óbitos ligados ao câncer de mama no Brasil continuam elevados (INCA, 2016).

Atualmente a mamografia é a principal ferramenta utilizada na detecção precoce do câncer de mama. O exame consiste em uma radiografia da mama gerada através de um aparelho de raios X denominado mamógrafo. As imagens geradas pelo exame são analisadas por um especialista que visualmente detecta as lesões (GIGER, 2000).

No entanto a análise das imagens mamográficas, além de ser uma atividade

¹Refere-se ao conjunto dos indivíduos que adquiriram doenças num dado intervalo de tempo

² Tumor Maligno ou Benigno

repetitiva, requer grande habilidade do especialista responsável, que por limitações da visão humana, acaba não identificando lesões milimétricas, que pode resultar em um diagnóstico incorreto.

Partindo dessa problemática, estão sendo desenvolvidos sistemas de detecção auxiliado por computador (*Computer Aided Detection Systems - CAD*). CADs são sistemas que utilizam técnicas de processamento, análise e reconhecimento de padrões de imagens, com o principal objetivo de detectar e diagnosticar anomalias em exames médicos, fornecendo uma segunda opinião ao especialista e usufruindo da velocidade e precisão dos sistemas computacionais.

1.1 OBJETIVO

O objetivo deste trabalho é apresentar uma nova metodologia para extração de características e classificação de imagens de mamografia segmentadas em massa e não massa. Nesta nova metodologia é proposta a utilização do Índice de Diversidade Funcional (PETCHEY; GASTON,2002) para a caracterização da imagem, além de sua adaptação para trabalhar em conjunto com as técnicas de processamento digital de imagens (PDI) e reconhecimento de padrões.

1.2 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado em 5 (cinco) capítulos. No Capítulo 2 são apresentados os embasamentos teóricos essenciais para o entendimento da metodologia abordada no trabalho. O Capítulo 3 discute as etapas de desenvolvimento da pesquisa, iniciando pela aquisição das imagens segmentadas na base DDSM, seguida pela extração de características de textura utilizando o Índice de Diversidade Funcional, e por fim, a classificação utilizando SVM. O Capítulo 4 mostra os resultados obtidos com a metodologia proposta. No Capítulo 5 são apresentadas as conclusões a respeito da metodologia apresentada, bem como os trabalhos futuro para otimização desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os conceitos teóricos essenciais para que o leitor se familiarize e compreenda a teoria utilizada na metodologia proposta neste trabalho. Aqui são abordados conceitos de câncer de mama, mamografia, PDI, descritores de textura, índice de diversidade funcional, classificação e validação.

2.1 CÂNCER DE MAMA

O câncer de mama é uma doença causada pela perda do controle da divisão celular, o que ocasiona um crescimento descontrolado (maligno) de células da mama. Essas células espalham-se nos tecidos e órgãos adjacentes, formando tumores (aglomerado de celular cancerosas) ou neoplasias malignas que podem espalhar-se para outras regiões do corpo ocasionando a metástase³ (INCA, 2016).

A detecção precoce tornou-se o principal meio de combate contra câncer de mama e suas altas taxas de mortalidade. Quanto mais cedo for diagnosticado maiores são as chances de sobrevivência do paciente. Para este fim são utilizados exames como a mamografia, que são capazes de detectar lesões pré-cancerígenas ou a própria neoplasia em estágio inicial no órgão de origem, antes do processo de metástase.

2.1.1 Mamografia

A mamografia é um exame radiográfico que gera uma imagem da mama, com todas as suas formas e estruturas, sendo capaz de mostrar lesões milimétricas, que poderiam tornar-se possíveis tumores. Com isso este procedimento mostrou-se essencial na detecção precoce do câncer de mama.

Denomina-se mamografia de rastreio quando o exame é realizado em pacientes que não apresentam sintomas da doença. Mas caso o paciente já apresente algum sintoma ou tenha identificado alguma anomalia, o exame passa a ser chamado de mamografia de diagnóstico (NATIONAL CANCER INSTITUTE, 2016).

³ Disseminação e crescimento das células cancerosas em locais distantes da sua origem

É importante frisar que em ambas as categorias o principal objetivo da mamografia é identificar anomalias, tais como podem ser observadas na Figura 1. Figura 1(a) apresenta uma massa oval; a Figura 1(b) mostra microcalcificações e na Figura 1(c) vê-se distorções da arquitetura da mama. Massa, microcalcificações e distorções de arquitetura significa:

- Massa: Qualquer opacidade com algum contorno arredondado e definido segundo a forma (redonda, oval, microlobulada e irregular)
- Microcalcificações: Pequenos depósitos de cálcio que variam de forma e densidade; e
- Distorção de arquitetura: São espiculações em uma região da mama ou uma retração focal do contorno parenquimatoso⁴ denso

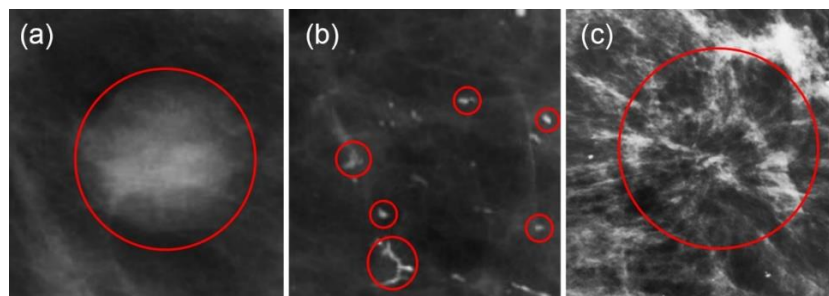


Figura 1. Tipos de anomalias em imagens mamográficas: (a) Massa Oval; (b) Microcalcificações; (c) Distorção de arquitetura; Fonte (HEATH et al., 2001).

O aparelho utilizado neste procedimento médico é chamado de mamógrafo. Para realizar o procedimento a mama é comprimida de forma a fornecer melhores imagens, e, portanto, melhor capacidade de diagnóstico (INCA, 2016), conforme mostra a Figura 2.



Figura 2. Mamógrafo; Fonte (AMERICAN CANCER SOCYETY, 2016).

⁴ Relativo ao parênquima (tecido, substância da mama).

A mamografia é realizada em duas posições gerando duas projeções para cada mama: uma médio lateral oblíqua (MLO) e uma crânio-caudal (CC). As duas projeções podem ser visualizadas na Figura 3.

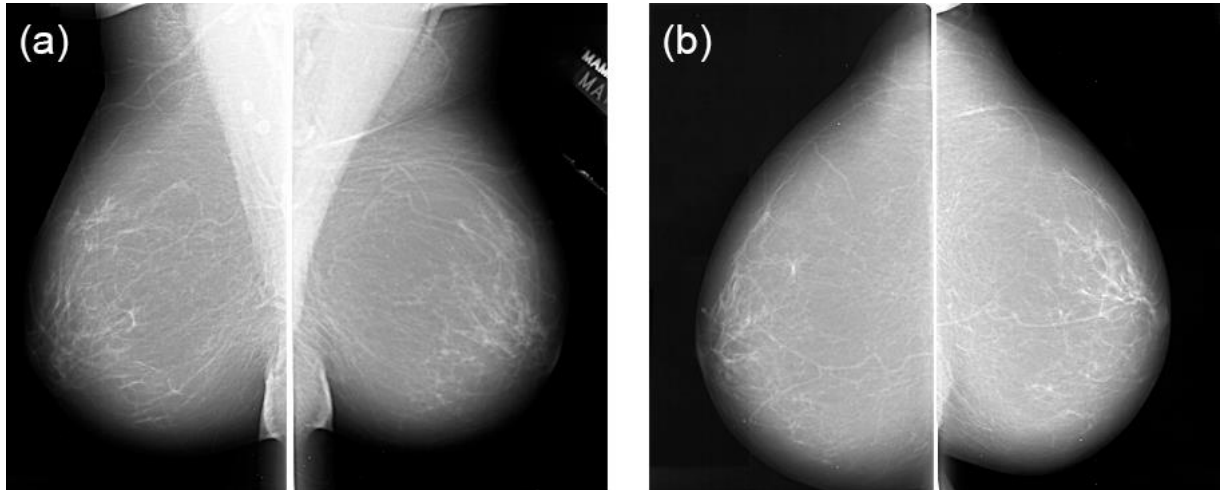


Figura 3. Resultados das projeções de uma mamografia; (a) Projeção médio lateral oblíqua (MLO); b) Projeção crânio caudal (CC); Fonte (HEATH et al., 2001).

2.2 PROCESSAMENTO DE IMAGENS

PDI consiste em um conjunto de técnicas direcionadas à manipulação de imagens utilizando um sistema computacional, tal que a entrada e a saída do processo são imagens. Essa área de pesquisa tem como objetivo o melhoramento de aspectos visuais e estruturais de uma imagem, além da extração de atributos e reconhecimento de objetos individuais presentes na mesma (GONZALES e WOODS, 2002).

Por não ser um processo simples, o processamento digital de imagens subdivide-se em várias etapas, sendo elas: A Aquisição de Imagem, Pré-Processamento, Segmentação, Extração de características e Reconhecimento e Interpretação. A Figura 4 apresenta um diagrama com etapas por ordem de realização, ou seja, iniciando com o problema, a aquisição das imagens a serem analisadas e terminando com o resultado da interpretação.

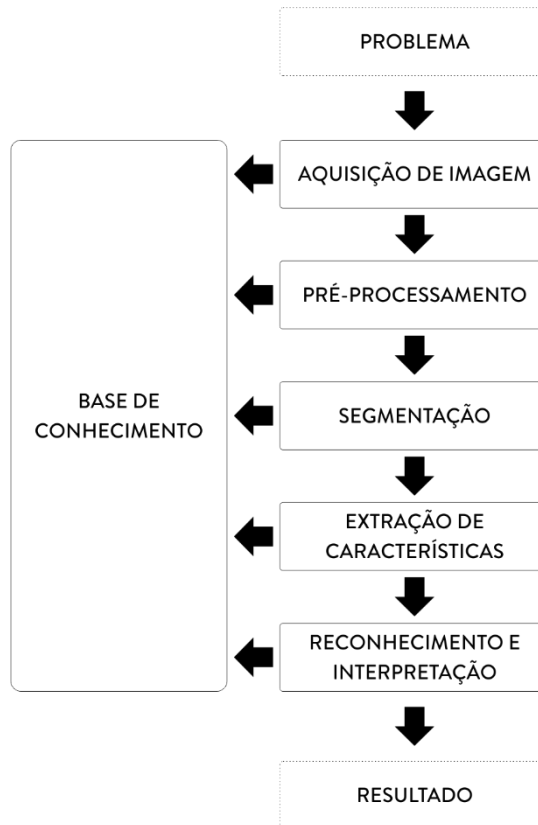


Figura 4. Etapas necessárias para o processamento de imagens; Adaptado de (GONZALES e WOODS, 2002).

Na primeira fase, aquisição da Imagem Digital, utiliza-se algum dispositivo físico que pode capturar determinada faixa de energia eletromagnética através de seus sensores e, posteriormente, digitalizar o sinal obtido. A mamografia, por exemplo, é impressa em folha de filme e digitalizado por *scanners* preparados para esse tipo de imagem.

O pré-processamento é a etapa onde são realizadas as transformações na imagem, utilizando técnicas como a equalização de realce, de contrastes, redução de ruídos, dentre outras, visando o melhoramento da imagem para as etapas posteriores.

A segmentação é responsável por dividir a imagem em regiões disjuntas, possibilitando o isolamento das ROIs. No exame de mamografia, as lesões presentes nas imagens é que são segmentadas.

A extração de características, também conhecida como representação e descrição, é a etapa responsável por extrair atributos da imagem através de descritores de forma e textura. Os dados levantados nessa etapa formam um Vetor de Características (VC) que são utilizados na discriminação das classes existentes na imagem.

Por fim, no reconhecimento e interpretação, o VC gerado na etapa anterior é analisado para que os dados adquiridos possam ser separados em diferentes categorias, por exemplo, em uma imagem mamográfica a região analisada pode ser classificada como massa ou não massa.

A metodologia proposta está inserida nas etapas de extração de características, utilizando descritores de textura e índices de diversidade funcional, e reconhecimento de padrões utilizando SVM.

2.2.1 Textura

Por possibilitar a distinção de regiões com mesmos aspectos de cores e padrões, a análise de textura é uma propriedade de grande relevância nas etapas de reconhecimento, descrição e classificação de imagens.

Para (HARALICK et al, 1973) a textura é definida como a característica ligada aos coeficientes estatísticos da imagem como a uniformidade, aspereza, regularidade, dentre outros. Para (GONZALEZ e WOODS, 2002), dentre as diversas abordagens existentes no processo de análise de textura, pode-se destacar a abordagem espectral, a estrutural e a estatística, conforme descritas a seguir:

- Abordagens Estatísticas: a textura é definida por um conjunto de medidas locais extraídas do padrão.
- Abordagens Estruturais: utilizam a ideia de que texturas são compostas de primitivas dispostas de forma aproximadamente regular e repetitiva, de acordo com regras bem definidas.
- Abordagens Espectrais: baseiam-se em propriedades do espectro de Fourier, sendo principalmente utilizadas na detecção de periodicidade global em uma imagem através da identificação de picos de alta energia no espectro.

2.2.1.1 MATRIZ DE COCORRÊNCIA

A matriz de coocorrência é um dos principais métodos utilizados na abordagem estatística, para a extração de características de textura. O foco principal da matriz de coocorrência é descrever a textura através das ocorrências de cada nível de cinza nos *pixels* da imagem considerando diversas direções (HARALICK et al, 1973).

A matriz de coocorrência considera a relação entre dois *pixels* por vez, um *pixel* é chamado de *pixel* referência e o outro de *pixel* vizinho. O *pixel* vizinho sempre está

situado na distância e angulação definidas para a contagem de ocorrências. Uma matriz de ocorrência é gerada para cada direção (HARALICK et al, 1973). A Figura 5 apresenta uma configuração para uma nova matriz de coocorrência, onde foi definida a distância de um *pixel* e quatro angulações (0° , 45° , 90° e 135°). Com isso serão geradas 4 matrizes de coocorrência, sendo que cada uma corresponderá a uma angulação.

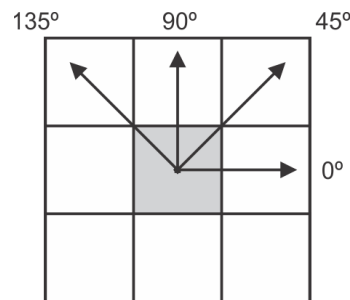


Figura 5. Distâncias e direções para uma nova matriz de coocorrência.

Na Figura 6 é ilustrado como se dá o processo de construção de uma matriz de coocorrência, a partir de uma configuração pré-definida. No exemplo estão sendo calculados as ocorrências entre dois *pixels* separados a uma distância $d=1$ e ângulo $\theta=0^\circ$.

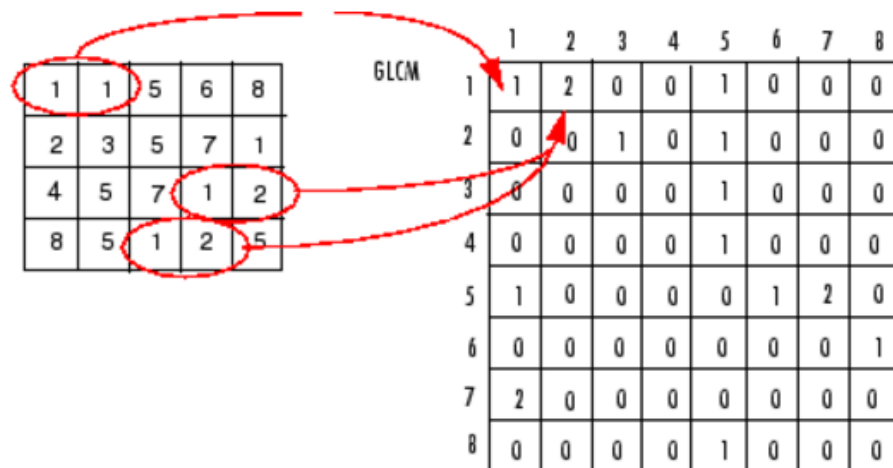


Figura 6. Construção da matriz de coocorrência.

2.2.1.2 MEDIDA DE DIVERSIDADE FUNCIONAL

Antes mesmo das ciências biológicas existirem o homem já mostrava interesse

nos padrões da distribuição de espécies nas comunidades naturais. Partindo dessa curiosidade muitas teorias surgiram ao longo dos anos formando as chamadas medidas de diversidade. As medidas tradicionais, baseavam-se apenas na quantidade de espécies e suas contribuições têm se mostrado estimativas pouco preditivas da estrutura e do funcionamento das comunidades (CIANCURSO *et al.*, 2009).

Com isso surgiram novas medidas de diversidade mais sensíveis à detecção de respostas das comunidades às mudanças no ambiente. A diversidade filogenética, por exemplo, leva em consideração as relações de parentesco entre as espécies (CIANCURSO *et al.*, 2009). Outra medida que vem sendo bastante estudada nos últimos anos é a diversidade funcional, que considera funcionamento e manutenção dos processos das comunidades baseado em suas características funcionais (PETCHEY; GASTON, 2002). Por fazer parte do escopo do trabalho é importante definir de maneira precisa o conceito de diversidade funcional.

Para (TILMAN, 2001) a diversidade funcional é o valor e a variação das espécies e de suas características que influenciam o funcionamento das comunidades. Com isso pode-se concluir que, medir a diversidade funcional significa medir a diversidade de características funcionais, que são componentes dos fenótipos dos organismos que influenciam os processos na comunidade.

Como exemplo, pode-se imaginar duas comunidades (A e B) com o mesmo número de espécies. Se todas as espécies em A forem dispersas por aves, enquanto que as em B forem dispersas por mamíferos, aves, lagartos e pelo vento, mesmo que ambas possuam o mesmo número de espécies, B será mais diversa por apresentar espécies funcionalmente diferentes no que se refere ao tipo de dispersão (CIANCURSO *et al.*, 2009).

A medida de diversidade funcional, consiste na soma dos comprimentos dos braços de um dendrograma⁵ funcional, ou seja, um dendrograma gerado a partir de uma matriz de “espécies × características funcionais”. (PETCHEY; GASTON, 2002).

⁵ Um dendrograma é um diagrama de árvore usado frequentemente para ilustrar a disposição dos clusters produzidos por agrupamento hierárquico.

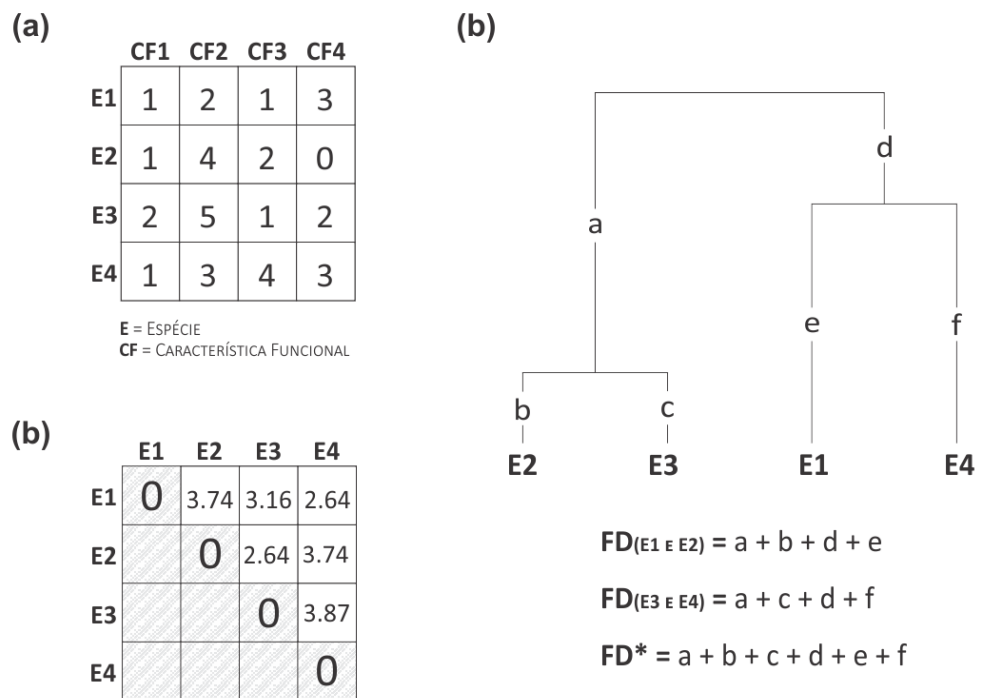


Figura 7. Cálculo da medida funcional; (a) Matriz de funcional (b) Matriz de distância obtida da matriz funcional; (c) Dendrograma gerado a partir de agrupamentos da matriz de distância e soma de suas ramificações.

Com base na Figura 7, pode-se observar que inicialmente deve-se obter a matriz funcional (espécies x característica), em seguida convertê-la para uma matriz de distâncias (MD). Feito isso, pode-se realizar agrupamentos da MD e produzir um dendrograma. Por fim, calcula-se o comprimento total de ramificações do dendrograma, achando os índices funcionais para um grupo de espécies ou de toda uma comunidade.

Para uma melhor compreensão da FD, é necessário fazer uma correlação entre a biologia e a metodologia proposta, como apresentado na Figura 8.

BIOLOGIA	METODOLOGIA
Comunidade	ROI
Espécies	Valor do pixel da ROI
Indivíduos	Quantidade de pixel de uma espécie

Figura 8. Correlação entre os termos da biologia e o trabalho.

2.3 RECONHECIMENTO DE PADRÕES E VALIDAÇÃO DE RESULTADOS

Nesta seção, é apresentado o reconhecimento de padrões, voltado à classificação de padrões, ou informações, que utilizam um conjunto de informações estatísticas extraídas ou de um conhecimento prévio.

2.3.1 Máquina de Vetor de Suporte

A SVM é um método de aprendizagem supervisionada utilizada para atribuir uma função que classifique os dados de entrada em duas classes (VAPNIK, 1998).

Segundo Scholkopf e Smola (2001), diversas SVMs possuem uma característica que consiste na redução do erro empírico de classificação, aumentando, à medida dessa redução, a margem geométrica de erro. Assim, esses métodos são denominados de *maximum margin classifiers* (classificadores de margem máxima).

O conceito básico de uma SVM é a construção de um hiperplano como superfície de decisão, com o intuito de que as margens de separação entre as classes sejam máxima (SCHOLKOPF; SMOLA, 2001).

Segundo Lorena e Carvalho (2007), hiperplano é a separação de duas regiões através de uma superfície, em um espaço multidimensional, sendo que o número de dimensões varia desde uma quantidade definida até uma quantidade infinita.

2.3.2 Métricas de Desempenho

Quatro possíveis situações podem existir na avaliação de um sistema de reconhecimento de padrões voltado à área médica, relacionado ao diagnóstico: verdadeiro positivo (VP), definido pelo teste positivo e a presença da doença no paciente; verdadeiro Negativo (VN), caracterizado pelo teste negativo e a não presença da doença no paciente; Falso Positivo (FP), onde o teste é positivo e o paciente não tem a doença; E por fim, Falso Negativo (FN), definido pelo teste negativo e a presença da doença no paciente.

Segundo BLAND (2000), é possível calcular algumas estatísticas de desempenho da metodologia através dos resultados dos testes, com o intuito de avaliar o desempenho do classificador, como Sensibilidade (S), Especificidade (E), e

Acurácia (A).

A sensibilidade é definida pela proporção de VPs identificados no teste, indicando o quão bom é o teste para identificação dos indivíduos doentes. O cálculo da sensibilidade é exibido na Fórmula 1.

$$S = \frac{VP}{VP + FN} \quad (1)$$

A especificidade é definida através da proporção de VNs identificados no teste, indicando o quão bom o teste será através identificação de indivíduos não doentes. O cálculo da especificidade é exibido na Fórmula 2.

$$E = \frac{VN}{VN + FP} \quad (2)$$

A acurácia é a razão entre o número de caso identificados de maneira correta e o número total de casos na classificação.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

Vários exames dos pacientes são medidos em uma escala numérica, onde a sensibilidade e especificidade são dependentes da localização do ponto de corte (*cut off*) entre os resultados positivos e negativos.

Com o intuito de uma melhor eficiência da relação adversa entre a sensibilidade e a especificidade dos exames que apresentam resultados contínuos, utilizou-se as Curvas de Características de Operação do Receptor (*Receiver Operating Characteristic* - ROC), uma ferramenta bastante eficaz considerando a mensuração e especificação dos problemas no desempenho do diagnóstico na área médica, por permitir o estudo da variação da sensibilidade e da especificidade para diferentes valores de corte da curva (MARTINEZ, 2003).

Segundo Cohen (1960), o índice *Kappa* é definido como o coeficiente de concordância de escalas nominais que pede a proporção desta mesma concordância, após a proporção de concordância depois que ocorre a atribuição a casualidade e a mesma é retirada de consideração.

Em suma, o coeficiente Kappa leva em consideração todos os elementos da matriz de erros, em vez de daqueles que se localizam na diagonal principal da mesma, ou melhor, a estimação da soma da coluna e linha marginais.

2.4 TRABALHOS RELACIONADOS

A literatura disponível traz trabalhos que possuem o mesmo objetivo da metodologia apresentada. Enumeram-se a seguir alguns desses trabalhos.

Silva Neto (2016) desenvolveu um método de detecção automática de massas em imagens mamográficas digitais, utilizando o Algoritmo *Particle Swarm Optimization* (PSO) e o Índice de Diversidade Funcional. Os testes desta metodologia foram realizados em mamas densas e não densas. Os melhores achados para as mamas densas foram com a sensibilidade de 97,52%, especificidade de 92,28%, acurácia de 94,82%, uma taxa de falsos positivos por imagem de 0,38 e a curva FROC de 0,99. Os melhores achados com todas as mamas densas e não densas, apresentaram 95,36% de sensibilidade, 89,00% de especificidade, 92,00% de acurácia, 0,75 a taxa de falsos positivos por imagem e 0,98 a curva FROC.

Sampaio (2015) desenvolveu uma metodologia computacional para detecção automática de massas mamárias, baseando-se na densidade da mama. Na etapa de segmentação foram utilizados algoritmos genéticos, com o objetivo de criar uma máscara de proximidade de textura e selecionar regiões suspeitas em conter lesão. Para reduzir o número de regiões suspeitas erroneamente segmentadas, utilizaram-se duas etapas de redução de falsos positivos. A primeira redução de falsos positivos usa o *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN) e um *ranking* de proximidade de textura extraídos das regiões de interesse. Na segunda, as regiões resultantes têm as suas texturas e formas analisadas pela combinação de Árvores Filogenéticas e descritores geométricos, Padrões Binários Locais e SVM. Foi alcançada uma sensibilidade de 94,02%, especificidade de 82,28% e acurácia de 84,08%, com uma taxa de 0,85 falsos positivos por imagem e uma área sob a *Free-Response ROC Curve* (curva FROC) de 1,13 nas análises de mamas não densas. Para mamas densas, obteve-se uma sensibilidade de 89,13%, especificidade de 88,61% e acurácia de 88,69%, com uma taxa de 0,71 falsos positivos por imagem e uma Área sob a Curva FROC de 1,47.

Carvalho (2012) apresenta uma metodologia de classificação de regiões extraídas de mamografias em massa e não massa. É aplicado o Índice de Diversidade de *McIntosh*, comumente usados em ecologia, para descrever a textura da região de interesse. O cálculo deste índice é proposto em quatro abordagens: por meio do Histograma, da matriz de coocorrência de níveis de cinza, da matriz de comprimentos de corrida de cinza e da matriz de comprimentos de lacuna de cinza. Para classificação das regiões em massa e não massa foi utilizado o classificador SVM. A metodologia apresenta bons resultados, alcançando uma acurácia de 93,68%.

Sousa (2011) também desenvolveu uma metodologia de discriminação e classificação de regiões massa e não massa. Para isso utilizou o Índice de Diversidade de *Shannon Wiener*, comumente aplicado para medir a biodiversidade em um ecossistema, o qual descreve padrões de regiões de imagens de mama com quatro abordagens: global, em círculos, em anéis e direcional. Para classificação utiliza-se o classificador SVM. Os resultados alcançaram uma acurácia máxima de 99,8%.

Os trabalhos relacionados indicam que metodologias baseadas em descritores de textura e reconhecimento de padrões utilizando SVM, apresentam resultados promissores para auxílio à detecção de massas em imagens mamográficas. Vários trabalhos apresentam a taxa de acurácia acima de 90%. Porém, ainda é necessário descobrir novas técnicas que melhorem estes resultados. Verificamos que a classificação de ROIs da mamografia em massa e não massa é uma etapa essencial 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 índices de diversidade funcional.

3 METODOLOGIA

Neste Capítulo é apresentada a metodologia proposta neste trabalho, com as seguintes etapas: Na primeira etapa é realizada a aquisição das imagens da base *Digital Database for Screening Mammography* (DDSM); na segunda etapa é realizada a descrição de textura, extraindo-se os índices de diversidade funcional como características das ROIs; na terceira etapa, realiza-se a classificação dos candidatos em massa e não massa; e por fim, são apresentadas as formas de validação dos resultados obtidos. A Figura 9 ilustra este processo.



Figura 9. Etapas da metodologia proposta nesse trabalho.

3.1 AQUISIÇÃO DE IMAGENS

A *Digital Database for Screening Mammography* é uma base pública de imagens de mamografias. O objetivo desse banco é facilitar a pesquisa e desenvolvimento de algoritmos para sistemas CAD, além de ser uma ferramenta de ensino para formação profissional.

O projeto DDSM é um esforço colaborativo que envolve as seguintes instituições americanas: *Massachusetts General Hospital, the University of South Florida, Sandia National Laboratories, Washington University School of Medicine, Wake Forest University School of Medicine (Departments of Medical Engineering and Radiology), Sacred Heart Hospital e ISMD, Incorporated.*

A base DDSM possui mais de 2500 estudos, sendo que cada estudo contém quatro imagens da mama (projeções Crânio Caudal - CC e Médio Lateral Oblíquo - MLO), juntamente com informações do paciente, do exame e dos equipamentos utilizados. Todas as informações contidas em cada caso foram fornecidas por especialistas (HEATH et al., 2001).

Neste trabalho foram utilizadas 1765 imagens de mamografias segmentadas,

sendo: 765 imagens com a presença de massa e 1000 sem o indicativo de anomalias. Todas as ROIs analisadas neste trabalho, foram extraídas do proposto em Silva Neto (2016).

3.2 EXTRAÇÃO DE CARACTERÍSTICAS

Para a realização desta etapa, foram utilizados como característica de medida das imagens analisadas, os índices de diversidade Funcional, conforme descrito na seção 3.2.1.2. Cada ROI possui 4 índices de diversidade funcional (FD1, FD2, FD3, FD4). Cada índice funcional está associado a uma matriz de coocorrência e sua respectiva configuração de distância entre *pixels* e direção.

Para a extração de características de textura baseado em índices de diversidade funcional, reduziu-se a imagem para 16 níveis de cinza, para reduzir o tamanho da matriz de coocorrência e reduzir o tempo de processamento. Foram calculadas as matrizes de coocorrência predefinidas para quatro angulações (0° , 45° , 90° e 135°) distância entre *pixels* $d = 8$. Os índices gerados nas etapas subsequentes são armazenados em um VC, para serem classificados posteriormente. Para facilitar a compreensão, a Figura 10 ilustra como funciona o algoritmo responsável pela etapa de extração de características.

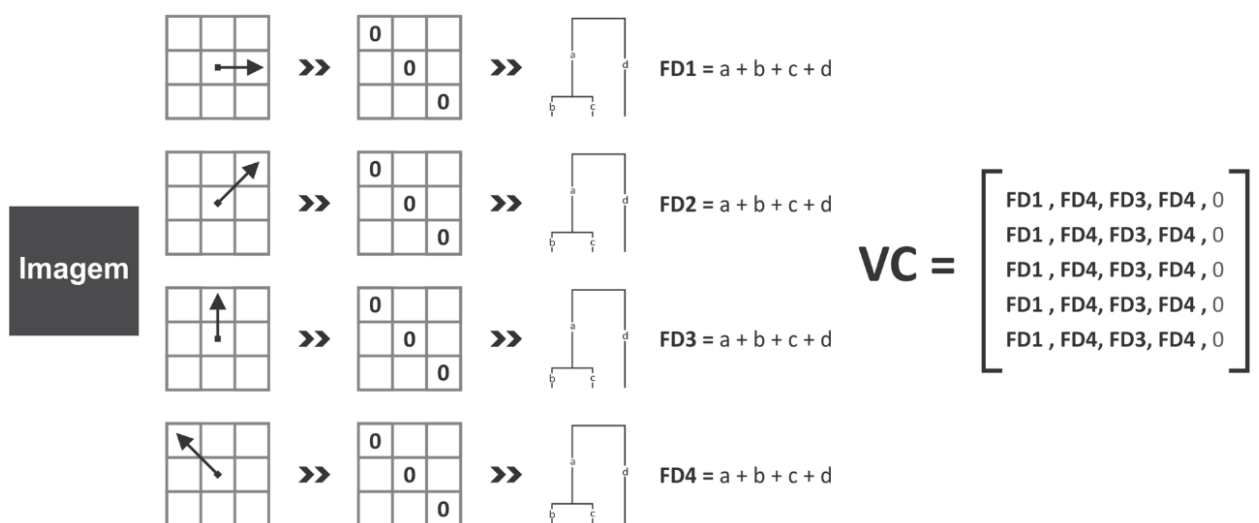


Figura 10. Etapa de extração de características.

3.3 CLASSIFICAÇÃO

A classificação foi realizada com o uso do *software Waikato Environment for Knowledge Analysis (WEKA)*, que é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Contém ferramentas para pré-processamento de dados, classificação, regressão, *clustering*, regras de associação e visualização. Ele foi desenvolvido por um grupo de pesquisadores da Universidade de Waikato, Nova Zelândia (HALL et al., 2009).

Os classificadores testados utilizaram o método *k-fold cross validation* para a obtenção dos resultados. Os dados foram divididos em 10 conjuntos, sendo 9 deles para treinamento e 1 para testes. Este processo é repetido 10 vezes, de forma que o conjunto escolhido para o teste será diferente do anterior e no final é gerada uma média dos resultados

3.4 VALIDAÇÃO DE RESULTADOS

Para a validação dos resultados, utilizou-se as métricas de avaliação baseadas em estatísticas como, Sensibilidade (S), Especificidade (E), Acurácia (A), taxa média de falso positivos por imagem (FP/i) e Curva ROC área sob a curva *Receiver-Operating Characteristics (ROC)* (Metz et al. 1086).

A sensibilidade demonstra o quanto a metodologia foi boa para detectar a presença de massas, a especificidade representa o quanto o método foi eficiente em detectar a presença de não massas, o kappa é uma medida de concordância entre observadores, a acurácia simula a taxa de acertos e a curva ROC descreve a habilidade de separar corretamente o conjunto de ROIs em duas classes baseado na fração verdadeiro positivo (especificidade) e na fração falso positivo (sensibilidade), quanto mais próximo de 1 significa que a classificação foi boa.

4 RESULTADOS E DISCUSSÕES

Neste trabalho foram analisadas e submetidas aos classificadores um conjunto de 1.765 imagens pertencentes a base pública DDSM, sendo elas 765 massas e 1000 não massas. As características extraídas foram submetidas às métricas de avaliação *Kappa* (K), área sob a curva *Receiver-Operating Characteristics* (ROC), Sensibilidade (S), Especificidade (E) e Acurácia (A) a partir do classificador SVM.

Devido aos resultados obtidos, foi necessária a aplicação das métricas de desempenho em outros classificadores para garantir a consistência dados obtidos. Os classificadores adicionais utilizados foram o *Multilayer Perceptron* (MLP); *LibLinear*; *Random Forest* (RF); *Logistic Model Trees* (LMT); REPTree; J48 e IBK.

Tabela 1. Resultados Obtidos em múltiplos classificadores.

Classificadores	FP/i	K	ROC	S (%)	E (%)	A (%)
J48	0,032	0.9343	0,971	96,9	96,7	96,8
RandomForest	0,012	0.9746	0,996	99	98,6	98,8
MLP	0	1	1	100	100	100
LibLinear	0	1	1	100	100	100
SVM	0	1	1	100	100	100
LMT	0	1	1	100	100	100
REPTree	0,034	0.932	0,98	96,5	96,8	96,7
IBK	0,003	0.9942	0,997	99,7	99,7	99,7

Os resultados apresentados na Tabela 1 demonstram que o conjunto de classificadores obteve resultado satisfatório, com todas as medições de S, E e A superiores a 95%. Com ênfase aos classificadores MLP, *LibLinear*, SVM, LMT, que obtiveram êxito de 100% na classificação entre massas e não-massas. Também podemos destacar o classificador IBK, que sendo o segundo maior resultado, obteve 99,7% de Sensibilidade, 99,7% de Especificidade, 99,7% de Acurácia, 0.9942 no índice Kappa e 0,997 de área sobre a curva ROC.

Não tendo acesso aos códigos e especificação dos exames utilizados, a análise comparativa entre os resultados da metodologia proposta e trabalhos relacionados ilustrados na Tabela 2, passa a ser meramente ilustrativa.

Tabela 2. Comparação da metodologia com os trabalhos relacionados.

Trabalhos	Base	FP/i	S (%)	E (%)	A (%)
Silva Neto (2016)	DDSM	0,75	95,36	89	92
Sampaio (2015)	DDSM	0,85	94,02	82,28	84,08
Carvalho (2012)	DDSM	-	90,1	-	93,68
Sousa (2011)	DDSM	-	91,5	94	99,85
Metodologia Proposta	DDSM	0,003	99,7	99,7	99,7

O a metodologia de Silva Neto (2016) que trabalha com detecção de massas utilizando *Particle Swarm Optimization* (PSO) e classificação utilizando Índice de diversidade Funcional, apresentou resultados inferiores nas métricas de desempenho em relação à metodologia proposta, demonstrando assim que o trabalho traz contribuições relevantes para o estudo do tema.

Observa-se na que a metodologia proposta apresenta valores significativos em relação aos trabalhos relacionados, que usam a base DDSM, ocupando um lugar de destaque, pois os valores encontrados são bastante promissores.

5 CONSIDERAÇÕES FINAIS

O câncer mama é um dos tipos mais comuns de câncer, o diagnóstico precoce dessa doença é essencial. Técnicas de processamento de imagens vem sendo usadas na construção de sistemas, que tem como objetivo auxiliar o especialista no diagnóstico, desta forma tornando seu trabalho menos cansativo e diminuindo as chances de erros.

Este trabalho apresentou uma nova metodologia para classificação de massas e não massas em imagens de mamografia, usando descritores de textura baseados em índice de diversidade funcional. A partir dos resultados obtidos pode-se concluir que a metodologia apresentada atingiu com êxito o seu objetivo.

Para trabalhos futuros pretende-se testar a metodologia proposta em novas bases de imagens, outros contextos e doenças. Além disso implementar a metodologia com outras técnicas de aprendizado de máquina como *deep learning*.

REFERÊNCIAS BIBLIOGRÁFICAS

AMERICAN CANCER SOCIETY. **Mammograms and Other Breast Imaging Tests**. 2016. Disponível em: <

<http://www.cancer.org/healthy/findcancerearly/examandtestdescriptions/mammogramandotherbreastimagingprocedures/mammograms-and-other-breast-imaging-procedures-what-is-mammogram>>. Acesso em: 07 dez. 2016.

COHEN, J. A **Coefficient of Agreement for Nominal Scales**. Educational and Measurement. Vol XX, Nº 1, p. 37-46, 1960.

GIGER, M. L. (2000). **Computer-aided diagnosis of breast lesions in medical images**. Computing in Science Engineering, páginas v. 2, p.39–45.

GONZALES, R. e WOODS, R. (2010). **Processamento Digital de Imagens**. 3a. ed. São Paulo: Pearson Prentice Hall.

HARALICK, R.M., K. Shanmugan, and I. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man, and Cybernetics, Vol. SMC-3, 1973, pp. 610-621.

HEATH, M., BOWYER, K., KOPANS, D., MOORE, R., e KEGELMEYER, W. P. (2001). The digital database for screening mammography. In Proceedings of the Fifth International Workshop on Digital Mammography, páginas 212–218. Medical Physics Publishing.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). Estimativa 2016: incidência de câncer no Brasil. Rio de Janeiro, 2016. Disponível em: <<http://www.inca.gov.br/estimativa/2016/estimativa-2016-v11.pdf>>. Acesso em: 07 dez. 2016.

Instituto Nacional de Câncer José Alencar Gomes da Silva (INCA). **Diretrizes para a detecção precoce do câncer de mama no Brasil**. Rio de Janeiro, 2015. Disponível em: <http://www1.inca.gov.br/inca/Arquivos/livro_deteccao_precoce_final.pdf>. Acesso em: 07 dez. 2016.

M. ERVIK, F. LAM, J. FERLAY, L. MERY, I. SOERJOMATARAM, F. BRAY (2016). Cancer Today. Lyon, France: International Agency for Research on Cancer. Cancer Today. Available from: <http://gco.iarc.fr/today>, accessed [05/december/2016].

MARTINEZ, E.Z.; LOUZADA-NETO, F.; PEREIRA, B.B. A curva ROC para testes diagnósticos. *Cad. Saúde Colet.*, v.11, p.7-31, 2003

METZ C. E. "**ROC methodology in radiologic imaging**". Invest. Radiol, v. 21, n.9, p. 720-33, 1986.

NATIONAL CANCER INSTITUTE (2016). **Mamograms**. Disponível em: <<https://www.cancer.gov/types/breast/mammograms-fact-sheet>>. Acesso em: 09 dez. 2016.

PETCHEY, O.L. & GASTON, K.J. 2002. Functional Diversity (FD), species richness, and community composition. *Ecol. Lett.* 5(3):402-411.

STEWART, B. W.; WILD, C. P. (Ed.) *World Cancer Report 2014*. Lyon, France: International Agency for Research on Cancer; Geneva: World Health Organization, 2014.

WEKA - Machine Learning Group at the University of Waikato. Disponível em: <<http://www.cs.waikato.ac.nz/ml/weka/>>. Acesso em 26 de fev. de 2016.




TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA
“JOSÉ ALBANO DE MACEDO”

Identificação do Tipo de Documento

- () Tese
() Dissertação
() Monografia
() Artigo

Eu, **Marcelo de Sousa Damasceno**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação **Classificação automática de massas em imagens mamográficas usando índices de Diversidade Funcional** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI 03 de março de 2017.


Assinatura