JOHNATAN CARVALHO SOUZA

Diagnóstico de Câncer de Mama a partir de Imagens de Mamografia 2D utilizando Descritores de Forma 3D

São Luís

2018

JOHNATAN CARVALHO SOUZA

Diagnóstico de Câncer de Mama a partir de Imagens de Mamografia 2D utilizando Descritores de Forma 3D

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da UFMA como requisito parcial para a obtenção do grau de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Anselmo Cardoso de Paiva

Coorientador: Prof. Dr. Aristófanes Corrêa Silva

São Luís

2018

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Souza, Johnatan Carvalho.
Diagnóstico de Câncer de Mama a partir de Imagens de Mamografia 2D utilizando Descritores de Forma 3D / Johnatan Carvalho Souza. – 2018.
69 f.

Coorientador(a): Aristófanes Corrêa Silva.
Orientador(a): Anselmo Cardoso de Paiva.
Dissertação (Mestrado) – Programa de Pós-graduação em Engenharia de Eletricidade/ccet, Universidade Federal do Maranhão, Centro de Ciências Exatas e Tecnologias – CCET, 2018.

1. Descritores de forma 3D. 2. Descritores de forma dentários. 3. Diagnóstico de câncer de mama. 4. Distribuição de forma. 5. Mamografia. I. Paiva, Anselmo Cardoso de. II. Silva, Aristófanes Corrêa. III. Título.

Dissertação de autoria de Johnatan Carvalho Souza, sob o título **"Diagnóstico de Câncer de Mama a partir de Imagens de Mamografia 2D utilizando Descritores de Forma 3D"**, apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Maranhão, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica, na área de concentração Ciência da Computação, aprovada em 21 de fevereiro de 2018 pela comissão examinadora constituída pelos doutores:

Prof. Dr. Anselmo Cardoso de Paiva
Orientador

Prof. Dr. Aristófanes Corrêa Silva
Coorientador

Prof. Dr. Ewaldo Eder Carvalho Santana
Membro da Banca Examinadora

Prof. Dr. Antonio Oseas de Carvalho Filho
Membro da Banca Examinadora

Prof[a]. Dra. Simara Vieira da Rocha
Membro da Banca Examinadora

Aos meus pais, família e amigos.

Agradecimentos

Agradeço em primeiro lugar a Deus, que me permitiu ir em busca dos meus sonhos por meio de muito esforço e dedicação.

Aos meus pais, Luzanira Carvalho Souza e Francisco das Chagas Souza, a quem devo tudo o que sou e tudo que conquistei, pelo apoio e amor incondicional, e ao meu irmão que sempre me faz perceber a importância da busca pelo conhecimento.

Aos meus familiares pelo apoio nos momentos de maior necessidade.

Ao professor e orientador Anselmo Cardoso de Paiva, por todo o conhecimento a mim passado que vão além da vida acadêmica, e pela paciência e confiança depositada em mim durante a orientação desde a iniciação científica até o mestrado.

Aos professores os quais tive o privilégio de ser aprendiz, pelos ensinamentos passados e auxílio durante a pesquisa. Em especial aos professores Aristófanes Correa Silva, Geraldo Braz Junior, João Dallyson Sousa de Almeida e Simara Vieira da Rocha.

Aos amigos de graduação Caio Eduardo, Giovanni Lucca, Jefferson Alves e João Otávio pelos anos de amizade e apoio na vida acadêmica.

E a todos que direta ou indiretamente contribuíram para a realização deste trabalho.

"A força não vem da vitória. Seus esforços desenvolvem suas forças. Quando você enfrenta dificuldades e decide não se entregar, isso é força."

(Arnold Schwarzenegger)

Resumo

O câncer de mama é a segunda maior causa de morte por câncer na população feminina e a quinta maior causa de morte por câncer em geral. Entretanto, sabe-se que o câncer de mama possui um melhor prognóstico e maiores chances de cura se diagnosticado em estágios iniciais. Portanto, a detecção precoce é de extrema importância e quanto mais informação estiver disponível para o especialista, maiores serão as chances de um diagnóstico correto. Este cenário justifica a necessidade do desenvolvimento de técnicas computacionais que auxiliem na detecção precoce dessa doença. Assim, o objetivo deste trabalho é apresentar um método para classificação de câncer de mama a partir de imagens de mamografia utilizando análise de forma e técnicas de reconhecimento de padrões. Para tanto, o método investiga a aplicação dos descritores *Relief Index, Average Slope, Section Area, Section Convolution*, D1dist, D2dist e D3dist. Este conjunto de descritores de forma não são tradicionais no contexto de análise de imagens médicas. Tratam-se de descritores de forma "dentários", que foram utilizados originalmente para extrair informações sobre a ecologia de espécies de mamíferos, a partir de dados coletados da morfologia de seus dentes. São realizados diversos experimentos com combinações desses descritores, onde são gerados vários vetores de características. Estes vetores são submetidos ao classificador máquina de vetores de suporte (MVS). O método proposto, utilizando estes descritores de forma, revelou resultados promissores. O melhor resultado obtido, em média, foi de 92.58% de acurácia, 92.80% de sensibilidade e 92.28% de especificidade.

Palavras-chaves: diagnóstico de câncer de mama, mamografia, distribuição de forma, descritores de forma dentários, descritores de forma 3D.

Abstract

Breast cancer is the second major cause of death by cancer in the female population and the fifth leading cause of death from cancer overall. However, it is known that breast cancer has a better prognosis and higher chances of cure if diagnosed at early stages. Therefore, an early detection is extremely important and the more information is available to the expert, the greater the chances of a correct diagnosis. This scenario justifies the need for the development of computational techniques to support the early detection of breast cancer. Therefore, the purpose of this work is to present a method for breast cancer classification from mammography images using shape analysis and pattern recognition techniques. For that, it is investigated the use of the descriptors Relief Index, Average Slope, Section Area, Section Convolution, D1dist, D2dist and D3dist. These shape descriptors are not traditional in the context of medical images analysis. They are so called "dental" shape descriptors, which have been used in dental ecology as ecometrics, characteristics of organisms that reflect a species' ecology, to analyze dental shape of mammals and reconstruct past environments. Several experiments with combinations of descriptors are performed, producing several feature vectors. Then, these vectors are submitted to the support vector machine classifier. The proposed method revealed promising results. The best result, on average, was 92.58% accuracy, 92.80% sensitivity and 92.28% specificity.

Keywords: Breast cancer diagnosis, mammography, shape distribution, dental shape descriptors, 3D shape descriptors.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

ACS	*American Cancer Society*
ACC	Acurácia
AG	Algoritmo Genético
Az	*Area Under the ROC Curve*
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
DDSM	*Digital Database for Screening Mammography*
ESP	Especificidade
FN	Falso Negativo
FP	Falso Positivo
INCA	Instituto Nacional do Câncer
MLP	*Multilayer Perceptron*
MVS	Máquina de Vetores de Suporte
RBF	*Radial Basis Function*
ROI	Região de Interesse
ROC	*Receiver Operating Characteristic*
SEN	Sensibilidade
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

Sumário

1 Introdução

Câncer é o nome dado a um conjunto de mais de 100 doenças que apresentam a característica de crescimento desordenado de células, que podem invadir tecidos e órgãos. Estas células dividem-se rapidamente e tendem a ser muito agressivas e incontroláveis, resultando na formação de tumores malignos, que podem espalhar-se para outras regiões do corpo. Existem várias causas possíveis para a formação de um câncer, que podem ser internas ou externas ao organismo, estando ambas inter-relacionadas. As causas externas referem-se ao meio ambiente e aos hábitos ou costumes próprios de um ambiente social e cultural. As causas internas são, na maioria das vezes, fatores genéticos pré-determinados, que estão ligados à capacidade do organismo de se defender das agressões externas. Os fatores externos e internos podem interagir de várias formas, o que acaba por aumentar a probabilidade de transformações malignas nas células normais (INCA, 2017).

De acordo com o Instituto Nacional de Câncer (INCA), o câncer de mama é o mais incidente entre as mulheres no mundo e no Brasil, depois do de câncer de pele não melanoma. Vale destacar que, apesar de afetar predominante mulheres, o câncer de mama também acomete homens em algumas raras situações, representando cerca de 1% dos casos. Na estimativa mais recente do INCA, realizada para o ano de 2016, foram previstos 57.960 novos casos de câncer de mama com um risco estimado de 56,20 casos a cada 100 mil mulheres (INCA, 2016).

O principal fator de risco para o câncer de mama é a idade. As taxas de incidência aumentam rapidamente até os 50 anos e, posteriormente, esse aumento ocorre de forma mais lenta. Contudo, outros fatores de riscos já estão bem estabelecidos como: aqueles que são relacionados à vida reprodutiva da mulher, histórico familiar de câncer de mama e alta densidade do tecido mamário. Além desses fatores, a exposição à radiação ionizante, mesmo em baixas doses, também é considerada um fator de risco, particularmente durante a puberdade (INCA, 2016).

O câncer de mama é considerado um câncer de bom prognóstico, entretanto suas taxas de mortalidade continuam elevadas. Isto está provavelmente associado ao fato de que a doença é frequentemente diagnosticada em estágios avançados, onde os danos são mais difíceis de serem revertidos. Em países desenvolvidos, a sobrevida média após cinco anos tem apresentado um certo aumento, cerca de 85%. Os países em desenvolvimento,

por outro lado, possuem uma taxa de sobrevida próxima a 60%. No Maranhão, esse câncer tem uma taxa de 13,97% de incidência para cada 100 mil mulheres (INCA, 2017).

Até o presente momento, ainda não é possível realizar a prevenção efetiva do câncer de mama, devido a uma variação de fatores relacionados às características genéticas de sua etiologia. Entretanto, especialistas recomendam como métodos preventivos de detecção a mamografia, para mulheres com idade entre 50 e 69 anos, e o exame clínico das mamas, anualmente a partir de 40 anos (INCA, 2017).

O exame clínico da mama (ECM) é uma das formas mais eficazes para detecção precoce do câncer de mama. Quando realizado por um especialista, o ECM pode detectar tumores de até um centímetro, se superficial. Segundo o INCA (2004), um ECM deve contemplar os seguintes passos: inspeção estática e dinâmica, palpação das axilas e palpação da mama com a paciente em decúbito dorsal. A eficiência do ECM é proporcional ao grau de habilidade e experiência do profissional para detectar anormalidades nas mamas examinadas. Ele deve ser realizado periodicamente e o médico indicará a necessidade de uma mamografia.

Segundo a ACS (2017), a mamografia é atualmente uma das melhores técnicas de detecção precoce de lesões não palpáveis na mama, pois ela possibilita a detecção visual de estruturas que podem evidenciar a presença ou ausência de câncer. Entretanto, essa avaliação tem caráter subjetivo, e exige grande habilidade do radiologista.

Nos últimos anos, diversas técnicas computacionais têm sido desenvolvidas com o propósito de auxiliar na interpretação dos exames de mamografia, identificando automaticamente estruturas que possam estar associadas a tumores, com o objetivo de melhorar a taxa de detecção precoce do câncer de mama (GIGER, 2000). Os sistemas que fornecem auxílio à detecção de lesões são conhecidos sistemas CAD (*Computer-Aided Detection*), e os que auxiliam no diagnóstico de doenças são conhecidos como sistemas CADx (*Computer-Aided Diagnosis*). Nos dias atuais estes sistemas já estão presentes em diversos centros de diagnóstico por imagem, principalmente em países desenvolvidos, como EUA e alguns países da Europa (TAYLOR et al., 2004; FENTON et al., 2007).

Os sistemas CAD e CADx fornecem ao radiologista informações adicionais que podem auxiliar na interpretação dos resultados, que em muitos casos, torna-se uma tarefa complexa devido às distorções e ruídos inerentes ao processo de aquisição de imagens. Estes sistemas são compostos das mais variadas técnicas descritas na literatura. Umas das abordagens mais utilizadas é a utilização de descritores de forma, que buscam caracterizar

ou diferenciar estruturas em imagens por meio de suas propriedades morfológicas. Existem também as abordagem que utilizam análise de textura para caracterizar os objetos de interesse, onde busca-se identificar e discriminar os padrões de textura existentes na imagem.

Neste trabalho propõe-se a investigação do uso de descritores de forma não tradicionais em processamento de imagens médicas. Tratam-se descritores de forma dentários, que são técnicas utilizadas originalmente para extrair informações sobre a ecologia de algumas espécies de mamíferos a partir de dados coletados da morfologia de seus dentes (EVANS, 2013). Estes descritores são utilizados neste trabalho com o objetivo de capturar os padrões de forma das massas em imagens de mamografias digitalizadas para posterior classificação quanto à sua natureza maligna ou benigna.

1.1 Objetivo do trabalho

O objetivo geral desta dissertação é propor um método para discriminar padrões de malignidade e benignidade de massas em imagens de mamografias, utilizando descritores de forma dentários e aprendizado de máquina.

1.1.1 Objetivos específicos

Para que o objetivo geral seja plenamente alcançado, os seguintes objetivos específicos deverão ser atingidos:

- Adaptar técnicas de realce de imagens que promovam um melhor detalhamento do formato das massas;
- Propor a utilização e implementar descritores de forma dentários para extrair características geométricas de massas em imagens de mamografia;
- Utilizar técnicas de reconhecimento de padrões para testar as características produzidas em relação à sua capacidade de discriminar massas malignas e benignas;
- Avaliar o método proposto através de experimentos, usando imagens de uma base pública de mamografias digitalizadas.

1.2 Contribuições do trabalho

O método proposto oferece uma série de contribuições para com o meio científico, podendo-se destacar as seguintes:

- Utilização de uma representação tridimensional de massas contidas em imagens de mamografia, para extração de características 3D;
- Adaptação dos descritores de forma dentários para o contexto de imagens médicas, para extração de características de massas em imagens de mamografia.

1.3 Organização do trabalho

Este trabalho apresenta a seguinte organização:

No Capítulo 2, Trabalhos Relacionados, é feita uma revisão bibliográfica dos trabalhos científicos relevantes e relacionados ao tema desta pesquisa.

No Capítulo 3, Fundamentação Teórica, são detalhados os conceitos necessários para entendimento do trabalho desenvolvido.

No Capítulo 4, Materiais e Método, o método proposto neste trabalho é descrito, onde cada etapa do processo é detalhada.

No Capítulo 5, Resultados e Discussão, são mostrados e discutidos os resultados obtidos com a aplicação do método proposto.

No Capítulo 6, Conclusão, apresentam-se as considerações finais do trabalho. Nele está contido uma revisão do que foi apresentado nesta dissertação juntamente com uma avaliação geral dos resultados obtidos e sugestões para trabalhos futuros.

2 Trabalhos Relacionados

Na literatura, existem várias produções científicas reconhecidas que tratam do mesmo problema abordado pelo método proposto, ou seja, métodos que auxiliam especialistas no diagnóstico de câncer de mama a partir de imagens de mamografia. Nestes trabalhos, a etapa de extração de características é realizada utilizando as mais diversas abordagens. As mais comuns são aquelas que utilizam análise de textura e forma (de maneira individual ou combinada) e, mais recentemente, as abordagens que utilizam técnicas de aprendizagem profunda. Neste capítulo, é feito um levantamento do estado da arte, onde são apresentadas algumas das abordagens para classificação de câncer de mama de maior relevância no meio científico.

2.1 Análise de textura

Uma das principais vantagens de se utilizar técnicas de análise textura para extração de características em imagens de mamografia é a não necessidade de uma segmentação severamente precisa. Muitas vezes, o contorno das massas não é bem definido, o que dificulta a definição exata de seus limites. Portanto, o uso de técnicas que não são grandemente afetadas por esta limitação são bastante adequadas neste aspecto. A seguir, são apresentados alguns trabalhos que utilizam técnicas de análise de textura para extração de características.

Utilizando *Gabor Filter Bank* para extração de características de textura, e MVS para classificação, Hussain et al. (2014) obtiveram um resultado de 0,87 de área sobre a curva ROC em sua metodologia para classificação de câncer de mama. Foram utilizadas 512 imagens de massa da base DDSM em seus experimentos.

Em (DHAHBI; BARHOUMI; ZAGROUBA, 2015), foi proposta a utilização da Transformada de *Curvelet* e Teoria do Momento para descrever características de textura em mamografias. Foram utilizadas duas abordagens para calcular a Transformada de *Curvelet*, sendo a primeira calculada para cada nível (CLM - *Curvelet Level Moments*), e a segunda por banda (CBM - *Curvelet Band Moments*). A partir dos coeficientes calculados, foram extraídos os quatro momentos de primeira ordem (média, variância, assimetria e curtose). Então, o Teste T foi utilizado para seleção de características nos dois conjuntos gerados. A técnica utilizada para classificação foi o *K-Nearest Neighbor*. Os experimentos

foram realizados em dois conjuntos de imagens distintos, o primeiro, composto de 116 imagens da base mini-MIAS, e o segundo, de 2340 imagens da base DDSM. Os melhores resultados obtidos com este método de classificação de massas em maligno e benigno foram de 81,35% e 60,43% de acurácia para as imagens das bases mini-MIAS e DDSM, respectivamente.

O método para classificação de câncer de mama proposto em (BEURA; MAJHI; DASH, 2015) utilizou a Transformada Discreta de *Wavelet* Bidimensional (2D-DWT) e Matrizes de Co-ocorrência de Níveis de Cinza (GLCM) para extração de características. Para seleção de características foram utilizados os métodos estatísticos Teste T para duas amostras e Teste F. Neste trabalho foram utilizadas 115 imagens de massa da base MIAS e 250 da base DDSM. Utilizando *Back-Propagation Neural Network* (BPNN) como classificador, os resultados obtidos foram de 94,2% de acurácia para imagens da base MIAS, e 97,4% de acurácia para as imagens da base DDSM.

No trabalho proposto em (ROCHA et al., 2016) foram utilizados índices de diversidade e uma combinação de padrões locais binários (LBP) e matrizes de co-ocorrência de níveis de cinza (GLCM) para extração de características de textura. Os índices de diversidade utilizados são os de Shannon, McIntosh, Simpson, Gleason e Menhinick. Nas GLCMs foram utilizadas as direções 0°, 45°, 90° e 135°. As matrizes eram extraídas a partir dos LBPs obtidos de várias quantizações das imagens de mamografia. Foi atingido 88,41% de média de acurácia em um conjunto de 1155 imagens da base pública DDSM, utilizando como técnica de classificação a MVS.

2.2 Análise de forma

As características de forma tem grande importância no contexto de análise de nódulos cancerígenos. Isto acontece porque, na maioria dos casos, o formato de um nódulo é um dos fatores mais preponderantes para definir sua malignidade ou benignidade. Diversos trabalhos têm sido publicados em que a análise de forma foi aplicada para classificar a natureza de nódulos mamários. A seguir, são apresentados alguns destes trabalhos.

No trabalho desenvolvido por Cheikhrouhou, Djemal e Maaref (2011) foi proposta uma técnica baseada na detecção de depressões e protuberâncias chamada *Protuberance Selection* (PS). Esta técnica baseia-se na variação dos sinais das derivadas em diferentes pontos de interesse do contorno das massas. Experimentos foram realizados utilizando o

classificador MVS com 242 imagens de massas da base DDSM. Este método obteve como resultado 0,93 de área sobre a curva ROC.

Em (LIU; LIU; FENG, 2011), foi proposto um método em que vários descritores de forma são extraídos a partir do contorno obtido pela técnica de segmentação *Level Set*. Os descritores utilizados foram compacidade, descritor de Fourier, descritores de momento baseado em fronteira, descritores baseados em Comprimento Radial Normalizado (NRL) e Orientação Relativa do Gradiente (RGO). Foram utilizadas 309 imagens da base DDSM nos experimentos. O melhor resultado deste método foi obtido usando o classificador MVS, alcançando 76% de acurácia.

No trabalho proposto por Abdaheer e Khan (2011) foi desenvolvido um método de *area matching* baseado na aproximação do contorno das massas em relação a um círculo. Foram utilizadas 150 imagens de massa da base MIAS. Neste método foi utilizado um limiar para classificação das massas que avaliava o valor de *area matching* entre o contorno da massa e um círculo. O melhor resultado obtido neste trabalho foi de 94% de acurácia.

Em (GÖRGEL; SERTBAS; UCAN, 2013), foi proposto um método baseado nas técnicas *Local Seed Growing* (LSG) e *Spherical Wavelet Transform* (SWT) chamado LSG-SWT. Enquanto o algoritmo LSG foi utilizado para segmentação, o SWT foi utilizado para extração de características. A partir dos coeficientes calculados do SWT, foram extraídas diversas características de forma, tanto do SWT, quanto da ROI diretamente. Utilizando MVS como classificador, foi obtido um resultado de 93,59% de acurácia utilizando 60 imagens de mamografia (78 massas) cedidas pelo Departamento de Radiologia do Hospital da Universidade de Istambul. Foi realizado também, um experimento usando 60 imagens da base MIAS, no qual foi obtido um resultado de 91,67% de acurácia.

No método desenvolvido em (WAJID; HUSSAIN, 2015) foi utilizada a técnica *Local Energy-Based Shape Histogram* (LESH) para extração de características de forma. O LESH converte a imagem de entrada em uma combinação de energias locais ao longo de diferentes orientações. Foram utilizadas 117 imagens de massas da base INbreast, e 115 da base MIAS em experimentos separados. O classificador utilizado foi a MVS e os melhores resultados obtidos foram de 99,73% e 99,09% de acurácia para o conjunto de imagens das bases INbreast e MIAS, respectivamente.

2.3 Análise de forma e textura

Como citado anteriormente, entre as abordagens mais comuns para caracterização de massas em imagens de mamografia estão a utilização de descritores de forma e textura. Muitos métodos descritos na literatura utilizam essas abordagens em conjunto. A seguir são apresentados alguns desses trabalhos.

No trabalho proposto por Liu e Tang (2014) foi desenvolvida uma abordagem para classificação de câncer de mama em que são utilizadas diversas características de forma e textura juntamente com uma técnica para seleção das características mais relevantes, denominada SRN. Esta técnica consiste na integração de uma variante da MVS que é baseada na eliminação de características de maneira recursiva com a técnica *normalized mutual information feature selection* (NMIFS). Entre as características de forma utilizadas estão compacidade, momentos de distância normalizada, descritores de Fourier, descritores baseados em comprimento radial normalizado (NRL), e descritores baseados em orientação relativa do gradiente (RGO). A partir da GLCM, são extraídas 19 características de textura. Algumas delas são contraste, autocorrelação, homogeneidade, variância, entre outros. Foram utilizadas 826 imagens de regiões de massa da base DDSM e o melhor resultado reportado foi de 0,96 de área sobre a curva ROC utilizando o classificador MVS.

No método proposto em (ROUHI et al., 2015), as GLCMs também foram utilizadas para extração de características de textura, enquanto que para extração de características de forma, foram utilizados os momentos de Zernike, resultando em um total de 51 características. Após a etapa de extração, foi utilizado Algoritmo Genético para seleção de características. O melhor resultado reportado neste trabalho foi de 96,47% de acurácia utilizando 170 imagens da base DDSM e o MLP como técnica de classificação.

Em (VALARMATHIE; SIVAKRITHIKA; DINAKARAN, 2016), foi proposto um método para diagnóstico de câncer de mama em que foram utilizados diversos descritores de forma e textura. Alguns dos descritores de forma são compacidade, dispersão, excentricidade e convexidade. Os descritores de textura são extraídos a partir dos histogramas de intensidades e das GLCMs. Entre as medidas de textura analisadas estão correlação, entropia, energia, média e desvio padrão. Nos experimentos foram utilizadas 332 imagens da base mini-MIAS, e foi obtido um resultado de 98% de acurácia com o classificador MLP.

2.4 Aprendizagem profunda

Nos últimos anos, as abordagens de aprendizagem profunda permitiram alcançar resultados sem precedentes em uma ampla gama de problemas provenientes de diferentes campos, tais como, visão computacional, processamento de linguagem natural e reconhecimento de voz. Estas abordagem têm ganhado cada vez mais destaque nas aplicações de processamento de imagens médicas. Diversos trabalhos têm sido publicados em que são utilizadas técnicas de aprendizagem profunda para auxiliar na detecção e diagnóstico de vários tipos de cânceres, inclusive o de mama. A seguir, alguns destes trabalhos são apresentados.

Em (AREVALO et al., 2015), foi proposto um método em que uma rede neural convolucional (CNN - *Convolutional Neural Network*) era utilizada para extração de características. Os autores destacam o fato de que as arquiteturas CNN não necessitam explicitamente de uma etapa de extração de características, tendo em vista que este é um processo inerente a este tipo de abordagem. Neste trabalho, as características das massas são extraídas automaticamente pelo processo de aprendizado da CNN. Os experimentos são realizados em um conjunto de 736 imagens da base *Breast Cancer Digital Repository* (BCDR), e logo após a etapa de extração, as características são submetidas ao classificador MVS. Foi reportado um resultado de 0,86 de área sobre a curva ROC com uso deste método.

Em (KAUR, 2016), uma CNN foi utilizada para classificação de massas em imagens de mamografia. Foram utilizadas 30 imagens da base MIAS. As imagens passavam por um processo de segmentação chamado *nucleus segmentation* e em seguida eram extraídas diversas características de forma e textura. Então, o conjunto de características era submetido à CNN. Neste trabalho foi reportado um resultado de 96,66% de acurácia.

No método proposto por Abbas (2016) a técnica de aprendizagem profunda utilizada para classificação de câncer de mama foi a *Deep Belief Network* (DBN). Primeiramente, eram extraídas características das imagens de massas usando *Speed-up Robust Features* (SURF) e *Local Binary Pattern Variance* (LBPV). Foi utilizada uma arquitetura de 3 camadas de DBN e mais uma quarta camada do classificador *softmax*. O treinamento da rede é realizado usando *Restricted Boltzmann Machines* (RBMs) em uma abordagem de aprendizado supervisionado e não-supervisionado. Foi utilizado um conjunto de 600

imagens, 350 da base DDSM e 250 da base mini-MIAS. Este método obteve um resultado de 91,5% de acurácia.

2.5 Considerações sobre os trabalhos relacionados

Os trabalhos descritos neste capítulo são de grande importância para o meio científico, pois obtiveram resultados significativos na tarefa de classificação de câncer de mama utilizando as mais diversas técnicas descritas na literatura. A Tabela 1 apresenta um resumo comparativo destes trabalhos, onde são detalhadas as técnicas de extração de características, classificadores e base de imagens utilizadas, a quantidade de ROIs usada nos experimentos, destacando-se a quantidade de benignas e malignas, além da acurácia e área sobre a curva ROC (Az) obtidas.

Tabela 1 – Resumo dos trabalhos relacionados.

	Trabalho	Técnica(s)	Classificador	Base	ROIs (Ben./ Mal.)	Acurácia	Az
Textura	(HUSSAIN et al., 2014)	*Gabor Filter Bank*	MVS	DDSM	512 (256/256)	85,53%	0,87
	(DHAHBI; BARHOUMI; ZAGROUBA, 2015)	*Curvelet*	KNN	mini-MIAS	116/ (66/50)	81,35%	-
	(BEURA; MAJHI; DASH, 2015)	2D-DWT GLCM	BPNN	DDSM	250 (129/121)	97,4%	-
	(ROCHA et al., 2016)	GLCM Índices de Diversidade	MVS	DDSM	1155 (530/625)	88,31%	0,88
Forma	(CHEIKHROUHOU; DJEMAL; MAAREF, 2011)	*Protuberance Selection*	MVS	DDSM	242 (128/114)	-	0,93
	(LIU; LIU; FENG, 2011)	NRL RGO	MVS	DDSM	309 (142/167)	76%	-
	(ABDAHEER; KHAN, 2011)	*Area matching*	MIAS	MIAS	150 (71/79)	94%	-
	(GÖRGEL; SERTBAS; UCAN, 2013)	LSG-SWT	MVS	Istambul Univ.	78 (43/35)	93,59%	-
	(WAJID; HUSSAIN, 2015)	LESH	MVS	INbreast	117 (-/-)	99,73%	-
Forma e Textura	(LIU; TANG, 2014)	NRL RGO	MVS	DDSM	826 (418/408)	-	0,96
	(ROUHI et al., 2015)	GLCM *Zernike Moments*	MLP	DDSM	170 (74/96)	96,47%	0,95
	(VALARMATHIE; SIVAKRITHIKA; DINAKARAN, 2016)	GLCM Geometria	MLP	mini-MIAS	332 (-/-)	98%	0,96
Aprendizagem Profunda	(AREVALO et al., 2015)	CNN	MVS	BCDR	736 (310/426)	-	0,86
	(KAUR, 2016)	Textura Geometria	CNN	MIAS	30 (-/-)	96,66%	-
	(ABBAS, 2016)	SURF LBPV	DBN	DDSM/MIAS	600 (300/300)	91,5%	0,91

Por meio da análise dos trabalhos relacionados pode-se observar que, no geral, os trabalhos que são baseados somente em análise textura realizam experimentos com uma quantidade maior de imagens, em contraste com os trabalhos baseados em análise de forma, que possuem uma limitação maior nesse aspecto. Isto provavelmente se deve ao fato de que a maioria das abordagens baseada em forma necessita de uma segmentação

extremamente precisa, pois se trata de um fator determinante na geração de características. No entanto, determinar uma segmentação de massas de forma precisa, no contexto de imagens de mamografia, é uma tarefa bastante complexa, e continua como um grande desafio no meio científico. Apesar dessa limitação, a análise de forma é continua sendo uma poderosa ferramenta de discriminação de massas mamárias, devido ao fato de as massas malignas e benignas, no geral, possuírem diferentes características morfológicas.

Também é possível observar que, no geral, os trabalhos que combinam características de forma e textura obtém um desempenho superior à maioria dos outros trabalhos. Isto reforça a necessidade de haver, além de boas técnicas para análise de textura, técnicas para análise de forma que possam ser utilizadas com bom desempenho na caracterização de massas em imagens de mamografia, pois as características morfológicas das massas são fatores primordiais para diferenciar massas malignas de massas benignas.

2.6 Considerações finais

Neste capítulo foi feita uma revisão das produções científicas relevantes relacionadas ao tema desta dissertação. Foram apresentadas resumidamente as metodologias e os resultados obtidos nestes trabalhos. Além disso, foram feitas considerações a respeito das diferenças entre essas diversas abordagens para classificação de câncer de mama, de maneira a se obter uma visão geral do que têm sido produzido pela comunidade científica em relação a este tema.

No próximo capítulo, serão apresentados os conceitos teóricos fundamentais para o desenvolvimento deste trabalho.

3 Fundamentação Teórica

Neste capítulo serão detalhados os tópicos necessários para compreensão das técnicas utilizadas no método proposto. Serão abordados conceitos importantes sobre o câncer de mama e o exame de mamografia, além das técnicas de processamento digital de imagens utilizadas, tais como, equalização de histograma, filtro da mediana, etc. Também serão apresentadas as técnicas de reconhecimento de padrões e as métricas de desempenho utilizadas para avaliação deste trabalho.

3.1 Câncer de mama

A glândula mamária feminina, ou mama, é um órgão par, que se situa na parede anterior do tórax, na parte superior e está apoiada sobre o músculo peitoral maior. A mama feminina é composta por lobos [1], por dutos [2] e por estroma [3] (ACS, 2017). A Figura 1 ilustra a anatomia de uma mama feminina normal.

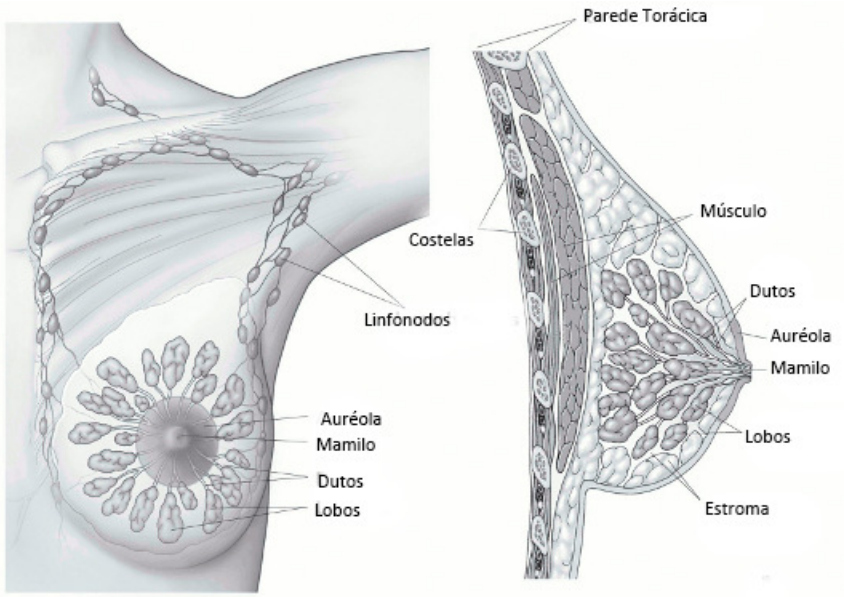

Figura 1 – Anatomia da mama feminina normal. Fonte: (ACS, 2017).

O câncer de mama é o crescimento descontrolado de células da mama que adquiriram características anormais resultando na formação de uma neoplasia ou tumor. Este comportamento anormal das células é causado por mutações em seu material genético,

[1] glândulas produtoras de leite.
[2] tubos que transportam o leite dos lobos ao mamilo.
[3] tecido adiposo e tecido conjuntivo que envolve os dutos e lobos, vasos sanguíneos e vasos linfáticos.

que ocorrem geralmente devido a fatores determinados pela própria genética do indivíduo. Fatores externos, tais como, hábitos sociais ou culturais, dieta e meio ambiente, podem também associar-se aos fatores genéticos do indivíduo e potencializar as chances de desenvolver a doença. Apesar de ocorrer predominantemente em mulheres, o câncer de mama também pode acometer homens em ocasiões raras. Segundo o INCA (2016), os casos de câncer de mama registrados em homens representam 1% do total.

Os tumores, ou neoplasias, são divididos em benignos e malignos de acordo com seu comportamento biológico. A principal diferença entre os tipos de cânceres é o nível de agressividade e velocidade de multiplicação das células. Geralmente, os tumores benignos crescem lentamente e provocam menos dano aos tecidos vizinhos. Os tumores malignos, por sua vez, possuem como característica principal o crescimento agressivo e descontrolado das células, provocando danos aos tecidos vizinhos e podendo espalhar-se para outras partes do corpo. Este fenômeno é chamado de metástase (NATIONAL BREAST CANCER FOUNDATION, 2017).

Estima-se que o tumor da mama duplique de tamanho a cada 4 meses aproximadamente. Na fase inicial, quando o tumor é impalpável, têm-se a impressão de crescimento lento, pois as dimensões das células são mínimas. Entretanto, depois que o tumor torna-se palpável, percebe-se facilmente a duplicação. Quando não tratado, o tumor desenvolve metástase, agravando mais ainda o quadro do paciente. Os órgãos mais comuns em que ocorrem a metástase do câncer de mama são os linfonodos, pulmões, ossos, fígado e cérebro. O tempo de sobrevida dos pacientes, após o descobrimento do tumor pela palpação, é geralmente de 3 a 4 anos (INCA, 2016).

As massas costumam aparecer como regiões densas de tamanho e formato variáveis, e podem ser classificadas de acordo com suas bordas em micro-lobuladas, obscurecidas, mal-definidas, circunscritas e espiculadas, e também de acordo com seu formato em ovais, circulares, lobuladas ou irregulares (NUNES; SILVA; PAIVA, 2009). Devido às características de crescimento, o formato das massas é um dos fatores mais importantes na diferenciação dos padrões malignos e benignos. As Figuras 2 e 3 ilustram massas classificadas de acordo com suas bordas e formato, respectivamente.

A detecção do câncer de mama em estágios avançados é um dos fatores mais agravantes e que dificultam as opções de tratamento. Atualmente, as chances de cura do câncer de mama são relativamente altas se a doença for diagnosticada precocemente. Para mulheres em situação de risco, cuja idade é maior ou igual a 40 anos, a forma mais eficaz

Figura 2 – Classificação das massas de acordo com o aspecto de suas bordas. Fonte: (NUNES; SILVA; PAIVA, 2009).

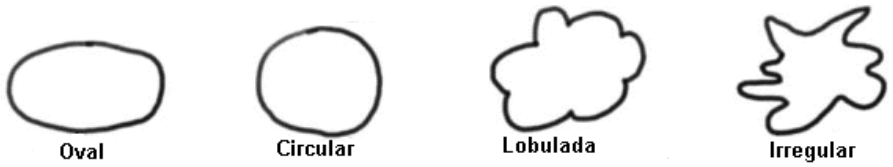

Figura 3 – Classificação das massas de acordo com sua forma. Fonte: (NUNES; SILVA; PAIVA, 2009).

para a detecção precoce do câncer de mama ainda é a mamografia, pois ela permite que o especialista possa identificar lesões muito pequenas, ainda em fase inicial (INCA, 2016).

De acordo com o INCA (2016), a forma mais eficaz para a detecção precoce do câncer de mama para mulheres em situação de risco ainda é a mamografia, pois ela permite que o especialista possa identificar lesões muito pequenas, ainda em fase inicial.

3.1.1 Exame de mamografia

A mamografia, ou radiografia da mama, é o exame mais indicado para detecção precoce do câncer de mama em mulheres com idade igual ou superior a 40 anos. Este exame é capaz de mostrar lesões ainda em estágios iniciais, quando os tumores possuem milímetros de diâmetro. É o principal exame de rastreamento do câncer de mama e é indicado por diversos especialistas e organizações de saúde no Brasil e no mundo (INCA, 2016; ACS, 2017).

O exame de mamografia é realizado em um aparelho de raios X específico, chamado mamógrafo. Neste aparelho, os feixes de raios X são incididos sobre a mama que está comprimida em ângulos que fornecem uma melhor visualização das estruturas internas, para aumentar as chances de um diagnóstico preciso. Como ainda não existem sistemas capazes de posicionar a mama mecanicamente, o posicionamento é realizado manualmente por técnicos altamente especializados, a fim de obter a melhor visualização possível da mama (INCA, 2016). A Figura 4 ilustra a realização deste exame em um aparelho mamógrafo.

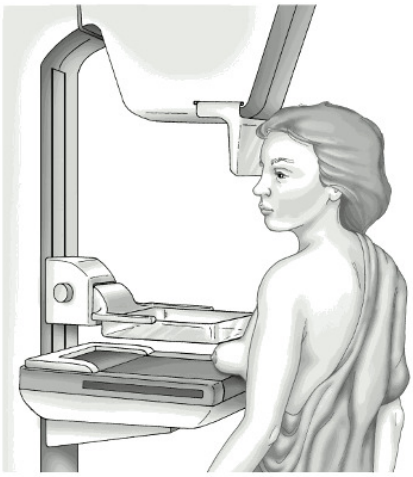

Figura 4 – Realização de um exame de mamografia. Fonte: (ACS, 2017).

O resultado do exame de mamografia é uma imagem em tons de cinza do tecido mamário (Figura 5) que é interpretada por um radiologista. Entretanto, essa tarefa é bastante complexa e exige grande habilidade e experiência do profissional, pois em alguns casos de câncer de mama, são produzidas alterações difíceis de serem percebidas. Além disso, a aparência da mama em uma mamografia varia muito de mulher para mulher. Por esses motivos, o radiologista deve ter em mãos as imagens do exame anterior para a comparação com o atual. Isto auxilia o especialista a encontrar alterações e, possivelmente, detectar um câncer de mama.

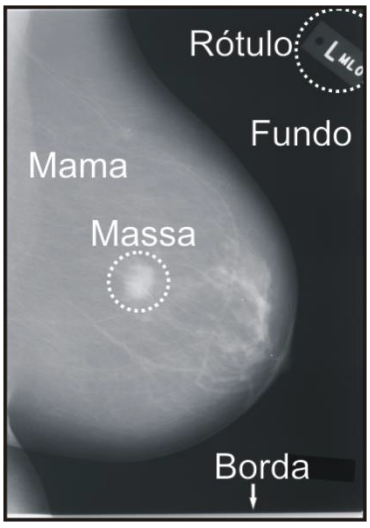

Figura 5 – Exemplo de uma imagem de mamografia com seus principais elementos. Fonte: (SAMPAIO et al., 2011).

Os tipos de anormalidade mais comuns encontrados nos exames de mamografia são as calcificações, as massas, e as distorções de arquitetura (linhas finas ou espiculadas que

se irradiam a partir de um ponto na mama) (HEATH et al., 1998). As massas, ou tumores, aparecem nas imagens de mamografia como regiões densas de tamanho e formato variáveis. Tratam-se de aglomerados de células que se unem de maneira mais densa em relação aos tecidos vizinhos. Estes aglomerados podem ser causados por condições benignas ou malignas (câncer).

Como visto anteriormente, as massas podem ser classificadas quanto a suas bordas em micro-lobuladas, obscurecidas, mal-definidas, circunscritas e espiculadas. Entre as características determinantes para definição da probabilidade de malignidade da massa estão: tamanho, forma e disposição das margens. Isto reforça a necessidade de haver técnicas para diferenciação de padrões malignos e benignos baseados em forma. A Figura 6 ilustra uma imagem de mamografia com uma massa bastante visível.

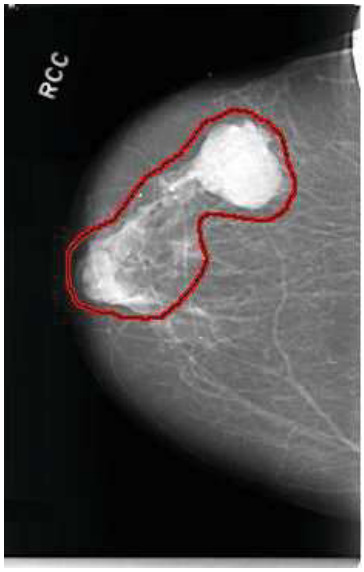

Figura 6 – Exemplo de uma imagem de mamografia com uma massa bastante visível. Fonte: (HEATH et al., 1998).

Além da mamografia convencional, onde a imagem é gravada em grandes folhas de filme fotográfico, existe também a mamografia digital, nas quais as imagens são capturadas eletronicamente e visualizadas em um monitor de alta resolução apropriado. As mamografias digitais permitem operações como ampliação e alteração de brilho e contraste para auxiliar na interpretação da imagem. Outra vantagem da mamografia digital é o compartilhamento facilitado devido à rápida transmissão de imagens digitais. Uma desvantagem dos mamógrafos digitais em relação aos convencionais é o seu alto custo, o que resulta no mamógrafo tradicional ainda estar presente com maior predominância nos hospitais.

As imagens de mamografia digitais, ou digitalizadas, juntamente com os sistemas de auxílio computacional à detecção e diagnóstico tornam-se poderosas ferramentas no combate ao câncer de mama, pois combinadas à habilidade experiencia do especialista, são as maneiras mais efetivas para a detecção precoce desta doença. Desta forma, tornam-se cada vez mais necessários os esforços para contribuir com técnicas que auxiliem o desenvolvimento de sistemas robustos de detecção e diagnóstico.

3.2 Técnicas de pré-processamento de imagens

Técnicas de pré-processamento ou realce de imagens são de grande importância para aprimorar o desempenho de algoritmos de segmentação, extração de características e reconhecimento de padrões. Nesta seção são apresentadas as técnicas de melhoramento de imagens utilizadas no método proposto.

3.2.1 Equalização de histograma

O histograma é utilizado como base para muitas técnicas de realce no domínio espacial da imagem. O histograma de uma imagem digital com intensidades variando entre 0 e $L-1$ é dado pela função discreta $f(r_k) = n_k$, onde r_k é a k-ésima intensidade e n_k é o número de *pixels* com intensidade r_k. A manipulação eficiente do histograma de uma imagem pode melhorar características específicas de uma imagem (GONZALEZ; WOODS, 2010).

A equalização de histograma é uma das técnicas de realce que se baseia no histograma da imagem. Equalizar um histograma significa obter a máxima variância do histograma, o que resulta em uma imagem com melhor contraste. O contraste é uma medida qualitativa que está relacionada à distribuição dos níveis de cinza em uma imagem.

Sendo $h(r_k)$ o histograma da imagem S, então o histograma acumulado de $h(r_k)$ é definido por:

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

para $r_k = 1, ..., L-1$. O histograma da imagem equalizada é obtido por:

$$T(r_k) = round\left(\frac{L-1}{MN}H(r_k)\right) \tag{2}$$

onde M e N são as dimensões da imagem. A imagem digital equalizada é obtida por *pixel* por $I = T(r_k)$.

3.2.2 Filtro da mediana

O filtro da mediana é um filtro não-linear que é utilizado para suavizar, e principalmente, remover ruídos ou distorções em uma imagem. Este filtro, que é conhecido por ser tão simples quanto seu nome sugere, substitui cada *pixel* da imagem pelo valor da mediana dos níveis de cinza na vizinhança daquele *pixel* (GONZALEZ; WOODS, 2010).

Seja S_{xy} o conjunto de coordenadas em uma janela retangular de tamanho $m \times n$, centralizada no ponto (x, y). O filtro calcula a mediana da imagem $g(x, y)$ na área definida por S_{xy}. Desta forma, a nova imagem f em qualquer ponto (x, y) é simplesmente a mediana obtida usando os *pixels* na região definida por S_{xy}. Em outras palavras:

$$f(x, y) = \underset{(s,t) \in S_{xy}}{mediana}\{g(s, t)\} \tag{3}$$

3.2.3 *Contrast-limited adaptive histogram equalization* (CLAHE)

Geralmente, técnicas de realce globais costumam melhorar não só os aspectos de interesse na imagem mas também ruídos. O *Contrast-Limited Adaptive Histogram Equalization* (CLAHE) é uma técnica de realce local em imagens. A vantagem do CLAHE em relação a outros ajustes de contraste é que ele evita que possíveis ruídos sejam realçados. Este algoritmo consiste em dividir a imagem em regiões contextuais e aplicar a equalização de histograma em cada região individualmente. Isso resulta em um maior equilíbrio na distribuição de níveis de cinza e torna caraterísticas ocultas nas imagens mais visíveis.

O CLAHE limita a amplificação por recorte do histograma em um valor pré-definido antes de calcular o Função de Distribuição Cumulativa (histograma acumulado). Isto limita a inclinação do mesmo e, consequentemente, a função de transformação. Este valor de corte do histograma depende do tamanho da vizinhança utilizada. Os valores de corte

geralmente são de 3 a 4 vezes o valor médio do histograma (ZUIDERVELD, 1994). A parte do histograma que excede o limite de corte é distribuída igualmente em todas as faixas do histograma, como ilustrado na Figura 7.

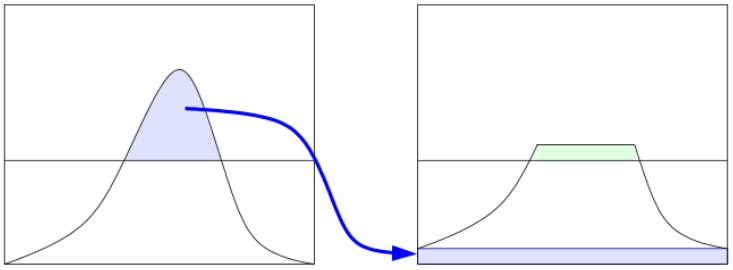

Figura 7 – Ilustração da redistribuição de histograma no CLAHE. Fonte: (ZUIDERVELD, 1994).

Existem várias funções de transformação de tons de cinza que podem ser utilizadas no CLAHE. Neste trabalho foi utilizado método uniforme:

$$g = [g_{max} - g_{min}]P(f) + g_{min} \tag{4}$$

onde g é o novo nível de cinza do pixel, os valores g_{min} e g_{max} são, respectivamente, as variáveis de menor e maior nível de cinza na vizinhança, e $P(f)$ é a função de distribuição cumulativa.

3.3 Quantização

O processo de quantização é utilizado para obter a representação de uma imagem com L níveis de cinza possíveis, com $L = 2^b$, sendo b o número de *bits* usados para armazenar o valor do *pixel*. Dada uma imagem com L níveis de cinza, se houver necessidade de quantizá-la para L' níveis de cinza, onde $L' < L$, utiliza-se a quantização uniforme, que consiste em dividir a escala de cinza da imagem em intervalos iguais para que sejam mapeados para novos valores de cinza na imagem quantizada, de maneira que a escala de cinza da imagem seja dada por $[0, L' - 1]$ (GONZALEZ; WOODS, 2010).

A quantização é calculada pela seguinte equação:

$$q(i,j) = (2^b - 1)\frac{p(i,j) - I_{min}}{I_{max} - I_{min}} \tag{5}$$

onde $q(i, j)$ é o nível de cinza do *pixel (i, j)* da imagem quantizada, $p(i, j)$ é o nível de cinza do *pixel (i, j)* da imagem original, $[I_{min}, I_{max}]$ são os limites inferior e superior da escala de cinza da imagem original, e b é o número de *bits* utilizados para armazenar cada *pixel* na imagem quantizada.

3.4 Análise de forma

Nesta seção serão detalhadas as estratégias e técnicas utilizadas neste trabalho para análise de forma.

3.4.1 Descritores de forma "dentários"

Os descritores detalhados nesta subseção fazem parte de um conjunto de descritores que têm sido utilizados em diversos estudos da ecologia, mais precisamente para análise da morfologia dentária de várias espécies de mamíferos. Neste contexto, estes descritores foram utilizados como métricas para comparação taxonômica e também para obter informações sobre dietas e reconstruir seus ambientes passados (BOYER, 2008; PLYUSNIN et al., 2008; UNGAR; WILLIAMSON, 2000). Em (EVANS, 2013), foram destacados a importância e os avanços obtidos com o uso desses descritores. O autor destaca o poder descritivo destas métricas e encoraja a sua utilização em outros sistemas de análise morfológica. Foi também neste trabalho que eles foram chamados pela primeira vez de descritores de forma dentários.

3.4.1.1 *Section area* e *section convolution*

Em (PLYUSNIN et al., 2008), os descritores *section area* e *convolution* são definidos e calculados a partir de cortes, ou seções, na superfície do objeto de interesse. São realizados 10 cortes no plano *x-y*, ao longo do eixo *z*. A área contida em cada corte é chamada de *section area*. Este valor de área é divido pela área total do objeto no plano *x-y* e é utilizada como descritor de forma. A Figura 8 ilustra a realização dos cortes feitos na superfície de um objeto com as áreas de cada corte.

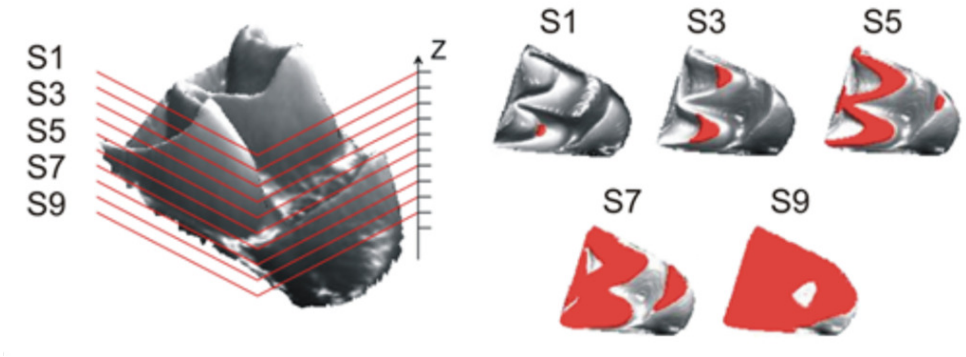

Figura 8 – Ilustração de cortes em uma superfície com as respectivas áreas obtidas. Fonte: Adaptado de (PLYUSNIN et al., 2008).

De cada corte também é calculado um outro descritor chamado convolução, ou *section convolution*. São calculadas além das áreas, os perímetros dos objetos contidos em cada corte. Então, a convolução é calculada pela Equação 6:

$$conv(S) = \frac{\sum_{i=1}^{n} P_i}{\sqrt{\sum_{i=1}^{n} A_i}} \tag{6}$$

onde n é o número de regiões ou objetos em cada corte, P_i é o perímetro, e A_i a área do i-ésimo objeto contido no corte S.

3.4.1.2 *Relief index*

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

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

$$ri = \frac{SA}{PA} \tag{7}$$

O *Relief Index* também foi definido por Boyer (2008) em um trabalho posterior como:

$$ri = ln\left(\frac{\sqrt{SA}}{\sqrt{PA}}\right) \tag{8}$$

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

Na etapa de extração de características, as duas versões deste descritor são utilizadas.

3.4.1.3 *Average slope*

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

O *Average Slope* foi calculado por meio da técnica *average maximum technique* (BURROUGH, 1986), um algoritmo muito utilizado em Sistemas de Informação Geográfica que calcula a taxa de mudança das direções vertical e horizontal de um ponto em uma superfície para estimar o grau de declive naquele ponto. Este algoritmo é tipicamente aplicado sobre uma janela de tamanho 3x3, na célula central e seus 8 vizinhos, para todos os *pixels* da imagem. A Figura 9 mostra o esquema de janela 3x3 utilizado. Os valores da célula central e seus 8 vizinhos determinam a taxa de mudança na direções vertical e horizontal. Os vizinhos são identificados como letras de a até i, sendo e a célula central, na qual o *slope* está sendo calculado.

a	**b**	**c**
d	**e**	**f**
g	**h**	**i**

Figura 9 – Esquema de janela 3x3 para o cálculo do *slope* de um ponto utilizando a técnica *average maximum technique*. Fonte: Adaptado de (PECKHAM; JORDAN, 2007).

O *slope* é comumente expresso em graus e é calculado pela técnica *average maximum technique* por meio da expressão:

$$S = arctan(\sqrt{[dz/dx]^2 + [dz/dy]^2}) * 57.29578 \tag{9}$$

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

$$[dz/dy] = ((g + 2h + i) - (a + 2b + c))/(8 * y\_cellsize) \tag{10}$$

$$[dz/dx] = ((c + 2f + i) - (a + 2d + g)/(8 * x\_cellsize) \tag{11}$$

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

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

$$avgS = \frac{\sum_{i=1}^{n} S_i}{n} \tag{12}$$

onde n representa o número de *slopes* calculados.

3.4.2 Descritores de distribuição de forma

A distribuição de forma foi proposta em (OSADA et al., 2002) para calcular assinaturas de forma de objetos e também para gerar medidas de similaridades entre objetos. Ela é calculada a partir de funções de forma, que são funções especiais que medem propriedades geométricas de uma superfície, e sua distribuição amostral é armazenada em histogramas. Esta abordagem fornece um método robusto para descrever e diferenciar objetos 3D pois é invariante a transformações geométricas, tais como, translação, rotação, espelhamento e escala.

A amostragem aleatória garante que a distribuição de forma seja indiferente a pequenas pertubações, como ruídos na superfície. Esta abordagem pode ser aplicada em modelos 2D e 3D, sejam armazenados como polígonos, malhas, *voxels*, ou qualquer outra

representação geométrica, desde que seja possível calcular a função de forma a partir dessa representação. Distância entre pontos, área e volume, neste caso apenas para modelos 3D, são exemplos de funções de forma.

Para extrair as propriedades geométricas utilizando distribuição de forma, é gerado um conjunto de pontos aleatórios na superfície do modelo. Então as funções de forma são calculadas para este conjunto de pontos. Em (OSADA et al., 2002) foram descritos um conjunto de funções de forma, entre as quais são utilizadas neste trabalho (Figura 10):

- D1: Calcula a distância entre um ponto fixo e um ponto aleatório na superfície.
- D2: Calcula a distância entre dois pontos aleatórios na superfície.
- D3: Calcula a raiz quadrada da área do triângulo formado por três pontos aleatórios na superfície.

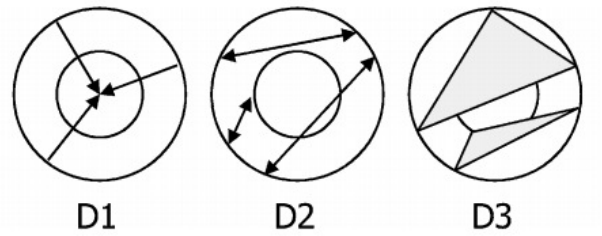

Figura 10 – Ilustração das funções de forma D1, D2 (baseadas em distâncias) e D3 (baseada em área). Fonte: Adaptado de (OSADA et al., 2002).

Na utilização convencional destes descritores para extração de características são utilizadas as frequências obtidas diretamente do histograma da distribuição. Em trabalhos recentes esta abordagem convencional foi utilizada em metodologias de classificação de câncer de pulmão e de mama (FERNANDES et al., 2016; SILVA, 2016).

Em (PLYUSNIN et al., 2008), foi proposta a utilização dos valores de média e desvio padrão da amostra obtida da função de forma D2, ao invés de todos valores do histograma. Esta modificação foi chamada de *D2dist*. Buscando investigar a aplicabilidade desta modificação na utilização das funções de forma, é utilizado neste trabalho os descritores D1dist, D2dist e D3dist, que consistem nos valores de média e desvio padrão dos descritores D1, D2 e D3, respectivamente. Por reduzir consideravelmente o espaço de características, acredita-se que esta modificação terá impactos positivos no processo de classificação.

3.5 Seleção de características

Uma situação frequente em aplicações de visão computacional é a utilização de técnicas que geram um grande número de características. Muitas das vezes, nem todas as características do conjunto são realmente necessárias para realizar a separação precisa dos indivíduos em suas classes. A inclusão de características irrelevantes ou redundantes pode, até mesmo, gerar resultados inferiores no processo de classificação (HAND; MANNILA; SMYTH, 2001).

A seleção de características é uma etapa opcional no processo de reconhecimento de padrões que tem como objetivo selecionar, entre todos os descritores, as características mais discriminativas do conjunto. Este processo pode ser visto como um procedimento de busca, em que o algoritmo utilizado deve encontrar, a partir de um conjunto de características, um subconjunto que tenha a melhor eficiência no processo de classificação. Em outras palavras, o objetivo deste processo é encontrar um subconjunto de características com menor dimensionalidade que promova um aumento na precisão, ou que não gere perdas significativas nos resultados (PEDRINI; SCHWARTZ, 2008).

A seleção de características é normalmente utilizada com o intuito de obter um conjunto de características com menos redundância e também para reduzir o custo do processamento. Neste trabalho a técnica utilizada para seleção de características foi o Algoritmo Genético.

3.5.1 Algoritmo genético

Algoritmo Genético (AG) é uma técnica de busca e/ou otimização que pode ser utilizada em uma ampla gama de problemas. No AG, a estrutura das possíveis soluções é implementada por uma representação de cromossomos. O algoritmo opera sobre uma população de cromossomos, que passa por processos evolutivos através de operações genéticas (mutação e cruzamento) e também por seleção natural. A finalidade do processo de evolução é criar indivíduos mais aptos a cada geração, ou em outras palavras, obter a solução mais satisfatória (GOLDBERG et al., 1989; HAUPT; HAUPT, 2004).

Geralmente, o AG itera por um número de gerações obedecendo um critério de parada, que pode ocorrer quando um número pré-estabelecido de gerações é atingido, ou quando as soluções param de melhorar. A comparação de cromossomos é feita em termos

de aptidão, ou *fitness*. Este nível de aptidão é obtido por meio uma função determinada pelo usuário, que indica o quão eficiente é aquela solução para o problema determinado. O resultado de saída de um AG é, normalmente, o cromossomo mais apto da população.

No problema de seleção de características, um cromossomo representa um subconjunto de características, e geralmente, é representado por uma sequência finita de 0's e 1's, onde cada elemento denota a ausência ou presença da i-ésima característica. O processo de otimização ocorre ao longo das gerações, onde são realizadas sobre a população diversas operações de cruzamento, mutação e seleção, com o objetivo de encontrar o subconjunto que represente a melhor solução.

3.6 Reconhecimento de padrões

Em (LOONEY, 1997), um padrão é definido como tudo aquilo para o qual existe uma entidade nomeável representante, geralmente criada através do conhecimento cultural humano. Em (GONZALEZ; WOODS, 2010), um padrão é definido como um conjunto de vetores de características, e que uma classe de padrões é na verdade uma família de padrões que possuem propriedades em comum.

O reconhecimento de padrões visa determinar um mapeamento que relacione as propriedades extraídas de amostras com um conjunto de rótulos (entidade nomeável representante), apresentando a restrição de que amostras com características semelhantes devem ser mapeadas ao mesmo rótulo. Os algoritmos que estabelecem este mapeamento são chamados algoritmos de classificação ou classificadores (PEDRINI; SCHWARTZ, 2008).

Segundo LOONEY (1997), o processo de reconhecimento de padrões envolve duas etapas: classificação e reconhecimento. Durante a etapa de classificação uma amostra de uma população qualquer é particionada em grupos chamados classes. Na etapa de reconhecimento uma amostra desconhecida, porém pertencente à mesma população, é reconhecida como integrante de uma das classes criadas anteriormente.

As técnicas de classificação podem ser divididas em duas abordagens principais: supervisionada e não-supervisionada. A classificação supervisionada ocorre quando o classificador considera classes pré-definidas e uma etapa de treinamento é executada antes da classificação para que os parâmetros que caracterizam cada classe sejam obtidos. Na classificação não-supervisionada não se dispõe de parâmetros ou informações coletadas

previamente à aplicação do algoritmo de classificação, e todas as informações são obtidas a partir das próprias amostras a serem rotuladas (PEDRINI; SCHWARTZ, 2008).

Neste trabalho foi utilizada a técnica de classificação supervisionada Máquina de Vetores de Suporte para o reconhecimento dos padrões maligno e benigno das massas em imagens de mamografia.

3.6.1 Máquina de vetores de suporte

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

As MVS são consideradas sistemas de aprendizagem que utilizam um espaço de hipóteses de funções lineares em um espaço multidimensional. Seus algoritmos de treinamento possuem grande influência da teoria da otimização e da aprendizagem estatística. De acordo com Cristianini e Shawe-Taylor (2000), a MVS tem mostrado sua superioridade em frente a outros classificadores em uma variedade de aplicações.

Há casos em que podem existir vários possíveis hiperplanos de separação, mas a MVS busca apenas encontrar o que maximize a margem entre os exemplos de treinamento. A Figura 11 mostra em duas dimensões, para melhor visualização, hiperplanos de separação entre duas classes linearmente separáveis. O hiperplano ótimo (linha mais escura), não somente separa as duas classes, mas mantém a maior distância possível com relação aos pontos da amostra.

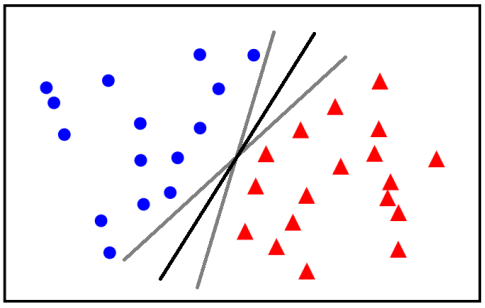

Figura 11 – Separação entre duas classes através de hiperplanos.

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

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

$$y_i((w.x_i) + b) \geq 1 \tag{13}$$

sendo w o vetor normal ao hiperplano de decisão e b o corte ou distância da função f em relação à origem, os valores ótimos de w e b serão encontrados de acordo com a restrição dada pela Equação 13 ao minimizar a seguinte equação:

$$\phi(w) = \frac{w^2}{2} \tag{14}$$

A MVS ainda possibilita encontrar um hiperplano que minimize a ocorrência de erros de classificação nos casos em que uma perfeita separação entre as duas classes não seja possível. Isso graças à inclusão de variáveis de folga, que permitem que as restrições presentes na Equação 13 sejam quebradas.

O problema de otimização passa a ser então a minimização da Equação 15, de acordo com a restrição imposta na Equação 16. C é um parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento e deve ser selecionado pelo usuário.

$$\phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{15}$$

sujeito à

$$y_i((w.x_i) + b) + \xi_i \geq 1 \tag{16}$$

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 17. O objetivo então passa a ser encontrar os multiplicadores de Lagrange a_i ótimos que satisfaçam a Equação 18 (CHAVES, 2006).

$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j (x_i, x_j) \tag{17}$$

$$\sum_{i=1}^{N} a_i y_i = 0, \qquad 0 \leq a_i \leq C \tag{18}$$

Apenas os pontos onde a restrição dada pela Equação 13 é exatamente igual à unidade têm correspondentes $a_i \neq 0$. Esses pontos são chamados de vetores de suporte, pois se localizam geometricamente sobre as margens. Tais pontos têm fundamental importância na definição do hiperplano ótimo, pois os mesmos delimitam a margem do conjunto de treinamento. Os demais pontos além da margem não influenciam decisivamente na determinação do hiperplano, enquanto que os vetores de suporte, por terem pesos não nulos, são decisivos. A Figura 12 ilustra os pontos que representam vetores de suporte.

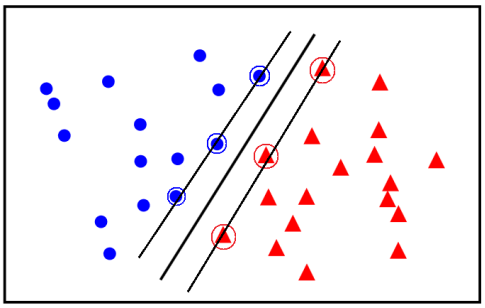

Figura 12 – Vetores de Suporte (destacado por círculos).

Para que a MVS possa classificar amostras que não são linearmente separáveis, é necessária uma transformação não-linear que transforme o espaço de entrada (dados) para um novo espaço (espaço de características). Esse espaço deve apresentar uma dimensão suficientemente grande para que por meio dele, a amostra possa ser linearmente separável. Desta forma, o hiperplano de separação é definido como uma função linear de vetores retirados do espaço de características em vez do espaço de entrada original. Essa construção depende do cálculo de uma função K de núcleo de um produto interno (HAYKIN, 2007). A função K é capaz de realizar o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos.

A Equação 19 mostra o resultado da Equação 17 com a utilização de um núcleo K.

$$L(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) \tag{19}$$

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

$$K(x_i, x_j) = exp(-\gamma \ ||x_i - x_j||^2) \tag{20}$$

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

3.6.2 Métricas de desempenho

Geralmente, após o processo de reconhecimento de padrões existe a necessidade de validar os resultados produzidos. Esta atividade tem por finalidade avaliar o desempenho do método desenvolvido por meio de uma análise estatística dos resultados.

Em problemas ligados à área da saúde, geralmente, os testes de classificação visam determinar o quão preciso é um método para indicar a presença ou ausência de uma doença. Neste tipo de problema existe uma variável preditora, referente ao resultado do teste para a doença, e uma variável resultante, que indica a presença ou ausência da doença. Desta forma, em uma amostra de casos positivos e negativos para uma determinada doença os resultados dos testes podem encaixar-se em um dos seguintes grupos:

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

A matriz de confusão permite visualizar a quantidade de casos pertencentes a cada uma das categorias (VP, FP, VN, FN). A Tabela 2 ilustra uma matriz de confusão.

Na avaliação do desempenho de classificadores para imagens médicas, geralmente são utilizadas as medidas de estatística descritiva sensibilidade (Sen), especificidade (Esp) e acurácia (Acc) (BLAND, 2015).

Tabela 2 – Matriz de confusão. Fonte: Elaborado pelo autor.

Resultado do Teste	**Possui a Doença**	**Não possui a Doença**
Positivo para doença	VP	FP
Negativo para doença	FN	VN

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

$$Acc = \frac{VP + VN}{VP + FP + VN + FN} \tag{21}$$

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

$$Sen = \frac{VP}{VP + FN} \tag{22}$$

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

$$Esp = \frac{VN}{VN + FP} \tag{23}$$

3.7 Considerações finais

Este capítulo apresentou os fundamentos teóricos usados nesta dissertação, os quais são necessários para compreensão das técnicas utilizadas e de suas aplicações no método proposto. Foram abordados temas como câncer de mama, exame de mamografia, técnicas de pré-processamento digital de imagens, descritores para análise de forma, e técnicas para reconhecimento de padrões e validação de resultados.

No próximo capítulo, serão abordados os materiais e os métodos utilizados no desenvolvimento deste trabalho, cujo objetivo é propor um método para classificação de massas em imagens de mamografia quanto à malignidade e benignidade.

4 Materiais e Método

Neste capítulo, são apresentados os materiais e as técnicas utilizadas juntamente com os procedimentos realizados no desenvolvimento deste trabalho. O método proposto consiste de cinco etapas: base de imagens, pré-processamento, extração de características, reconhecimento de padrões e validação de resultados. A Figura 13 ilustra o fluxo geral do método e nas próximas seções cada uma dessas etapas será apresentada com detalhes.

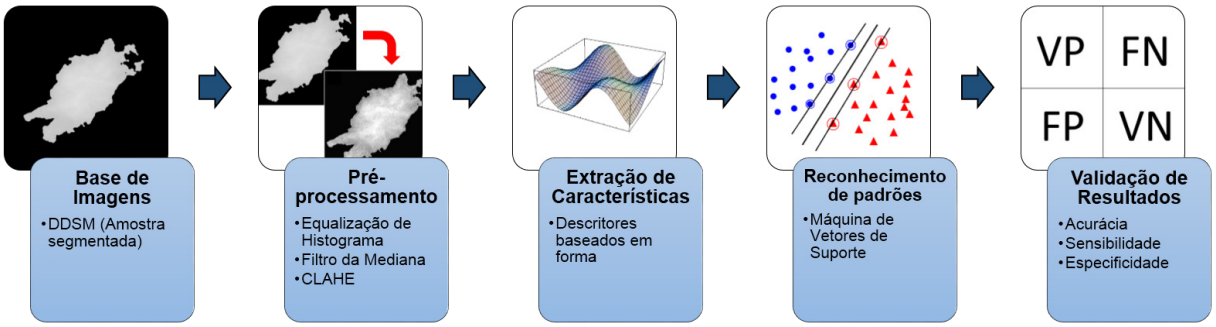

Figura 13 – Etapas do método proposto.

4.1 Base de imagens

Como o foco deste trabalho é a análise morfológica de massas em imagens de mamografia, não foram utilizadas as imagens completas, mas somente os *bounding boxes* das regiões das massas. Foram utilizadas as imagens disponibilizadas por Silva (2016), que contém regiões de massas segmentadas.

O método utilizado em (SILVA, 2016) para segmentação das massas consiste das seguintes etapas: aplicação do filtro *High-Boost*, na região demarcada pelo especialista, com detecção automática do fator de amplificação; definição dos grupos conexos; remoção dos menores grupos conexos (mantendo-se somente o maior grupo); suavização das bordas; e extração do contorno. A Figura 14 ilustra as etapas realizadas neste processo.

O conjunto contém imagens de 794 regiões de massas (395 malignas e 399 benignas) que não são afetadas por quaisquer aspectos que deformam seus contornos. Entre os aspectos deformadores de contorno citam-se: ruídos excessivos, sobreposição da região

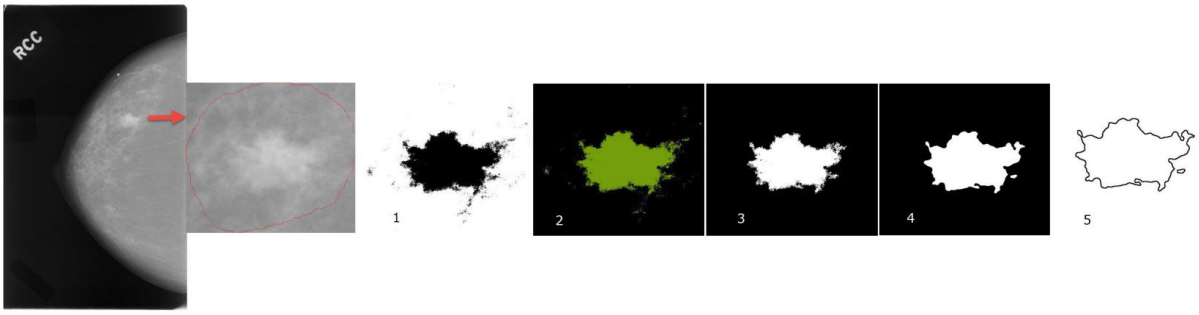

Figura 14 – Ilustração das etapas de segmentação de uma região de massa maligna.

de massa por outras estruturas da mama, e massa localizada junto ao músculo peitoral. Esta restrição é necessária devido a natureza dos descritores utilizados nestes trabalho, que são baseados exclusivamente em forma e portanto, podem ser comprometidos por estes aspectos. Essas situações causam problemas na segmentação, porque deformam severamente a fronteira das massas e, consequentemente, impedem que seja feita uma demarcação precisa de contornos. A Figura 15 ilustra algumas regiões de massa utilizadas, obtidas a partir dos contornos definidos com a segmentação.

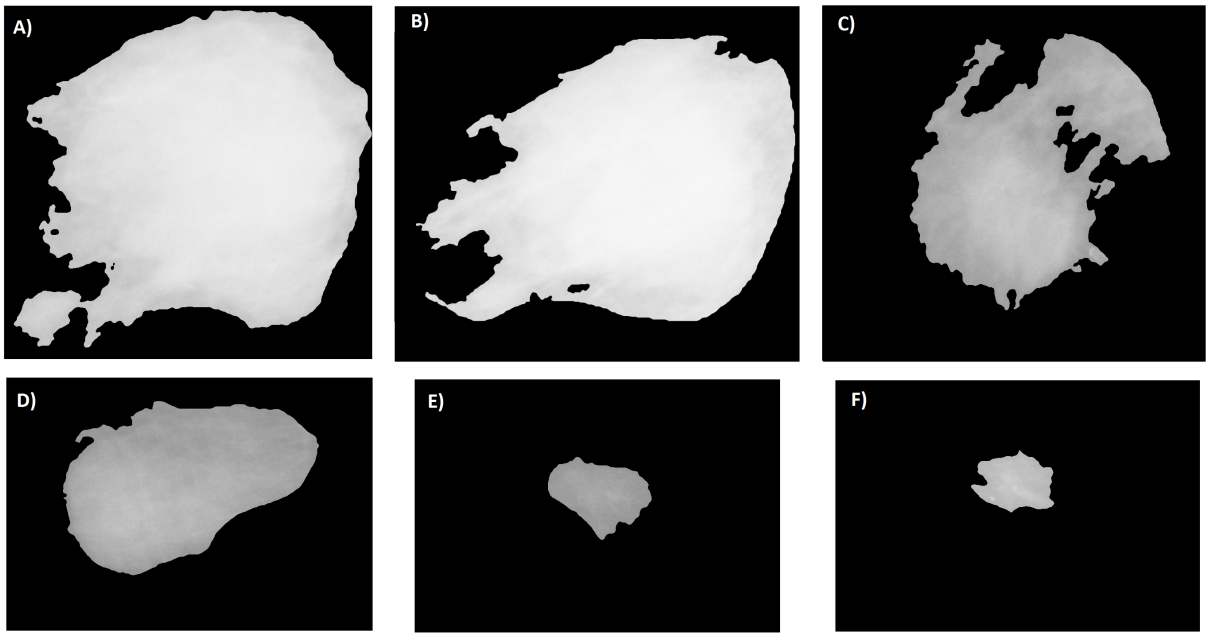

Figura 15 – Ilustração de imagens de massa: (a), (b) e (c) massas malignas; (d), (e) e (f) massas benignas. Fonte: (SILVA, 2016).

Do conjunto inicial foi necessário remover as imagens que não continham área suficiente, as quais pertenciam em sua grande maioria à classe benigna. As imagens

remanescentes consistiam de 310 benignas e 391 malignas. Desta forma, foram selecionadas 620 regiões de massa, das quais 310 são malignas e 310 são benignas. O número de amostras foi balanceado com o objetivo de promover uma melhor etapa de treinamento do classificador.

O conjunto de imagens utilizadas neste trabalho foi obtido originalmente da base DDSM - *Digital Database for Screening Mammography*, que é uma base pública de mamografias digitalizadas que contém exames de pacientes de diversas origens étnicas e raciais. A DDSM é o produto de uma colaboração entre o Hospital Geral de Massachusetts, a Universidade de Wake Forest e a Escola de Medicina da Universidade de Washington, cujo principal objetivo é fomentar a pesquisa e o desenvolvimento de técnicas computacionais para auxiliar na detecção e diagnóstico de patologias da mama (HEATH et al., 1998).

Na DDSM são disponibilizadas diversas informações dos pacientes, como idade, densidade da mama, data do exame e tipo de patologia. São disponibilizadas informações técnicas, como o tipo do digitalizador utilizado e a resolução de cada imagem. As imagens que contém regiões suspeitas possuem um arquivo de descrição da anormalidade, chamado *overlay*. Neste arquivo estão contidas informações como: a quantidade de lesões, a localização e contorno da lesão, e o seu diagnóstico. O contorno da lesão é codificado em *chain-code* e as descrições das lesões seguem o padrão do *American College of Radiology* (ACR) e são publicados em BI-RADS (*Breast Imaging Reporting and Data System*) (AMERICAN COLLEGE OF RADIOLOGY, 1998).

4.2 Pré-processamento

A etapa de pré-processamento tem o objetivo de melhorar o contraste da imagem na região da massa. Com um melhor contraste, os traços e o contorno das massas ficam mais evidentes, o que possibilita uma melhor caracterização do formato das massas. Para realizar este procedimento, foram utilizadas as técnicas de equalização de histograma, filtro da mediana e CLAHE.

A equalização de histograma é umas das principais técnicas para melhoramento de contraste. Ela permite reduzir diferenças acentuadas na imagem, e também e acentuar detalhes não visíveis anteriormente.

Após a equalização do histograma, é aplicado um filtro da mediana com uma janela de tamanho 3×3. Este filtro é utilizado com o objetivo de redução de ruídos, pois após

a equalização de histograma, alguns ruídos podem ser realçados. Além disso, este filtro produz um efeito de suavização nos *pixels* da imagem. O filtro da mediana tem se mostrado superior a outros filtros de suavização pois produz uma excelente redução de ruídos sem provocar borramento da imagem.

Após a aplicação dessas duas técnicas, percebia-se que em algumas imagens ainda existiam alguns ruídos, possivelmente porque foram realçados durante a aplicação da equalização de histograma e persistiram após o filtro da mediana. Desta forma, o CLAHE foi utilizado com o objetivo de remover o ruído persistente. Como detalhado na Seção 3.2.3, o CLAHE opera em pequenas janelas na imagem e aplica a equalização de histograma localmente. Este processo produziu melhoras visuais consideráveis, destacando ainda mais as diferenças de intensidade de *pixel* na região da massa. A aplicação do pré-processamento é ilustrada na Figura 16.

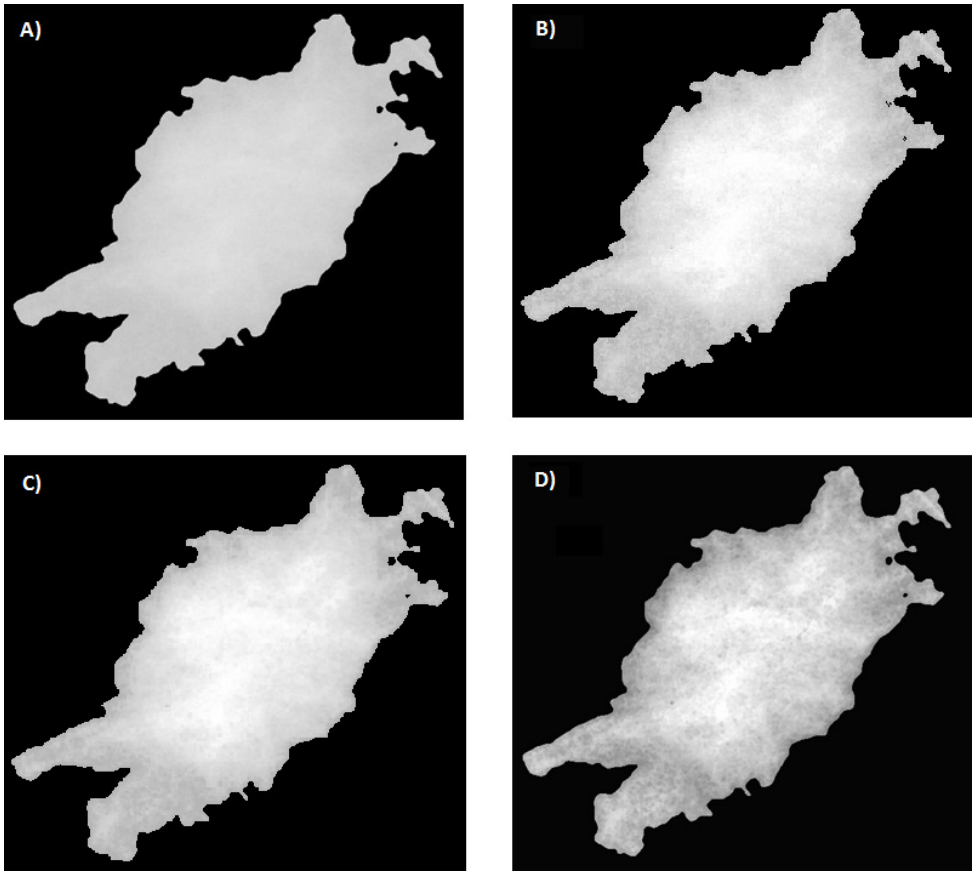

Figura 16 – Aplicação do pré-processamento em uma ROI normal (a). Após a aplicação das técnicas de equalização de histograma (b), filtro da mediana (c) e CLAHE (d), em sequência, obtém-se a ROI melhorada.

4.3 Extração de características

Esta etapa tem como objetivo produzir medidas descritivas das massas que formarão os vetores de características. Estes vetores são utilizados posteriormente na etapa de classificação. A extração de características é realizada utilizando as técnicas descritas na Seção 3.4. A seguir, são detalhadas as adaptações utilizadas para extração de características de massas com essas técnicas.

4.3.1 Representação 3D das massas

Alguns dos descritores de forma abordados nesta seção foram desenvolvidos para objetos 3D. Entretanto, como os objetos de estudo deste trabalho são massas em imagens de mamografia, que são imagens 2D, foi necessário criar uma representação 3D para as mesmas. Neste trabalho, cada ROI [1], ou região de interesse, contém uma única massa, da qual é gerada uma superfície 3D para extração de características.

Para construir uma superfície 3D, é necessária uma nuvem de pontos 3D que represente essa superfície. Como as imagens de mamografia possuem apenas duas dimensões, foi utilizada a seguinte estratégia. Com exceção dos *pixels* que compõem o fundo (intensidade igual a zero), foi criado para cada *pixel* p na ROI um ponto tridimensional P equivalente, no qual as coordenadas x, y e z são, respectivamente, as coordenadas espaciais de p (x_p e y_p) e a amplitude ou nível de cinza de p. Portanto, o número de pontos tridimensionais gerados para a representação 3D é igual ao número de *pixels* da ROI (com exceção dos *pixels* de fundo). A Figura 17 mostra um exemplo de superfície 3D gerada a partir da imagem de uma massa utilizando este método.

Com a representação 3D detalhada acima, os descritores de forma utilizados neste trabalho são apresentados nas próximas seções.

4.3.2 Descritores de forma dentários

Durante os experimentos com os descritores *section area* e *section convolution* buscou-se avaliar diferentes representações da forma. Mais precisamente, das formas associadas aos agrupamentos formados por quantizações adicionais. Deste modo, a partir

[1] *region of interest* (região de interesse).

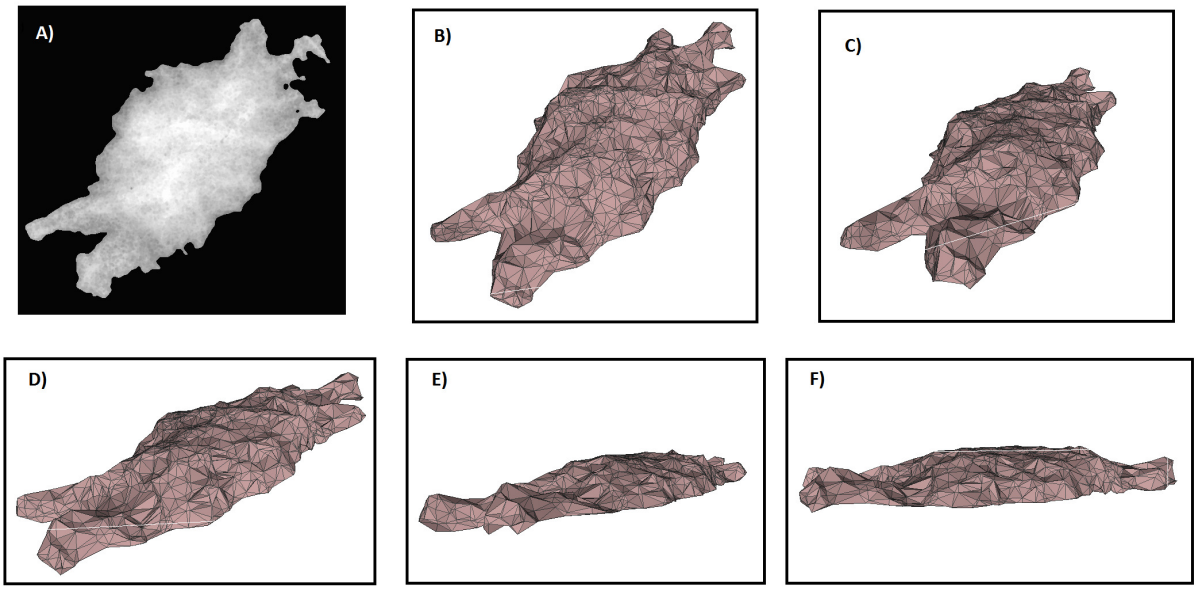

Figura 17 – Ilustração do modelo de representação 3D de uma ROI 2D: (a) ROI; (b-f) Diferentes perspectivas da representação 3D de (a). Fonte: Elaborado pelo autor.

da ROI melhorada pela etapa de pré-processamento, foram geradas as quantizações de 6 e 7 *bits* (64 e 128 níveis de cinza máximos) para extração de características adicionais. Os descritores *section area* e *section convolution* são calculados para a ROI melhorada e para as quantizações geradas utilizando o valor padrão de 10 cortes na superfície (assim como em (PLYUSNIN et al., 2008)). As características são obtidas de cada um dos cortes realizados, como detalhado na Seção 3.4.1.1. Desta forma, é gerada 1 característica por corte. Como são utilizadas 3 imagens (ROI melhorada e 2 quantizações), e em cada imagem são realizados 10 cortes, são produzidas 30 características para cada descritor.

Os cortes são realizados no modelo 3D das massas ao longo do eixo z. Os limites superior e inferior do intervalo de cortes são, respectivamente, o maior e o menor nível de cinza na região da massa. Este intervalo é dividido igualmente pelo número de cortes realizados (neste caso, 10) para determinar a posição de cada corte.

A estratégia utilizada para realização dos cortes na superfície das massas é baseada na utilização da técnica de limiarização por *threshold*. Conforme explicado na Seção 3.4, a representação tridimensional das massas consiste na utilização dos valores de intensidade dos *pixels* como coordenadas do eixo z. Desta forma, a aplicação de um limiar L produz o mesmo resultado de um corte no eixo z quando $z = L$. A Figura 18 ilustra a realização dos cortes a partir de uma ROI pré-processada.

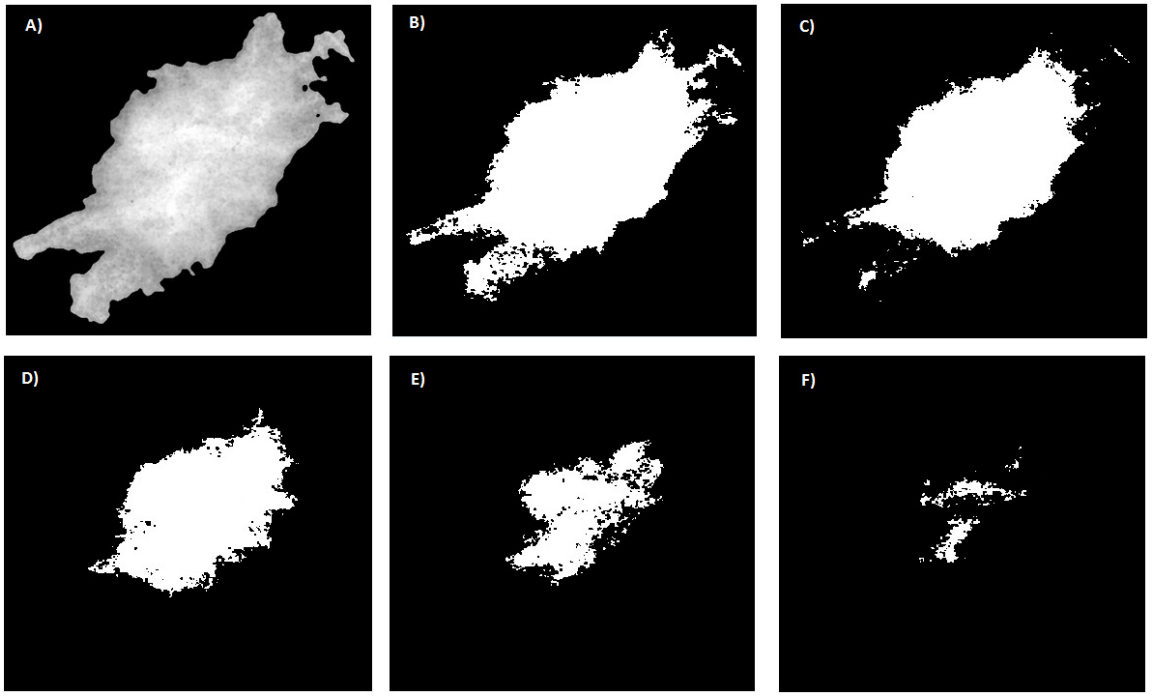

Figura 18 – Resultados do procedimento de cortes na superfície para os descritores *section area* e *convolution*: (a) ROI pré-processada; (b) corte número 2; (c) corte número 4; (d) corte número 6; (e) corte número 8; (f) corte número 10.

No cálculo do descritor *relief index* foi utilizada a implementação de *alpha shapes* da biblioteca CGAL (DA; SÉBASTIEN; YVINEC, 2017) para reconstruir as superfícies 3D das massas e assim calcular suas áreas. As áreas 2D são facilmente obtidas calculando-se as áreas dos contornos das massas. Então, as duas versões do *relief index* são calculadas utilizando as Equações 7 e 8, gerando duas características.

O descritor *average slope* produz uma característica, que é obtida pela média dos declives (*slopes*) de todos os pixels da região da massa. Os declives são calculados utilizando a técnica *average maximum technique*, detalhada na Seção 3.4.1.3, com uma janela de tamanho 3×3.

4.3.3 Descritores de distribuição de forma

Os descritores de distribuição de forma D1dist, D2dist e D3dist são calculados a partir das distribuições amostrais das funções de forma D1, D2 e D3. Essas distribuições são obtidas a partir de pontos gerados aleatoriamente na superfície do objeto, que podem ser feito em 2D, no contorno do objeto, ou em 3D, na superfície externa do objeto. Portanto,

como descrito na Seção 3.4.2, é possível calcular as distribuições das funções de forma D1, D2 e D3 para ambas as representações da forma das massas (2D e 3D).

Para cada função de forma são gerados pontos aleatórios dos quais serão calculadas informações particulares. A quantidade de pontos gerada para cada imagem é igual a um terço do número de pontos no seu respectivo vetor de contorno. No D1dist são armazenadas as distâncias entre pontos aleatórios e um ponto fixo, que neste caso é o centroide do objeto. No D2dist são armazenadas as distâncias entre dois pontos aleatórios na superfície. E no D3dist são armazenadas as raízes quadradas das áreas dos triângulos formados por conjuntos de três pontos aleatórios na superfície. A Figura 19 ilustra uma região de massa sendo utilizada nos descritores D1, D2 e D3. Na ilustração pode-se observar pontos que foram gerados aleatoriamente no contorno da massa para que sejam armazenadas suas distâncias (D1 e D2) e os áreas formadas (D3).

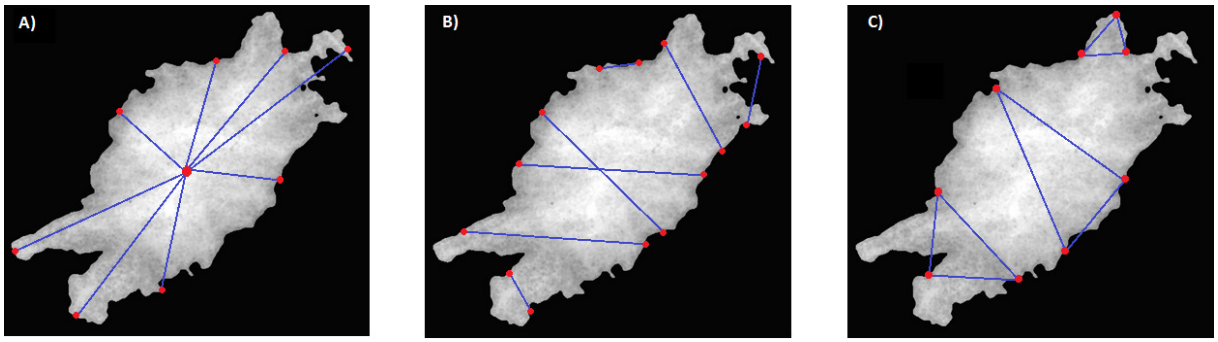

Figura 19 – Ilustração de pontos aleatórios no contorno da massa para utilização dos descritores D1dist (a), D2dist (b) e D3dist (c) na superfície bidimensional.

Após armazenar a distribuição das funções de forma, são calculados seus valores de média e desvio padrão, produzindo, para cada uma das funções, 4 características (2 para a superfície 2D, e 2 para a superfície 3D).

4.4 Reconhecimento de padrões

Nesta seção são apresentadas as técnicas utilizadas na etapa de reconhecimento de padrões. Por meio de experimentos de classificação, esta fase visa analisar se as características produzidas são capazes de diferenciar os padrões malignos e benignos das massas mamárias.

4.4.1 Seleção de características

Esta etapa foi utilizada no experimento com todas as características geradas, com o objetivo de eliminar as variáveis redundantes e, desta forma, melhorar a eficiência do processo de classificação.

Para efetuar este processo, foi utilizado o Algoritmo Genético apresentado na Seção 3.5.1. Este algoritmo busca a melhor solução do problema por meio de cruzamentos e mutações realizadas ao longo de um número determinado de gerações. Espera-se que o conjunto de características resultante possa conter menos variáveis redundantes, e que isso resulte em maiores taxas de acerto.

4.4.2 Classificação

Como descrito na Seção 3.6, o objetivo desta etapa é classificar cada imagem de massa como maligna ou benigna, a partir dos vetores de características produzidos na etapa anterior, utilizando o algoritmo de classificação MVS.

Os experimentos de classificação seguem o fluxo ilustrado pela Figura 20. Após a geração da base de características, incia-se o processo de reconhecimento de padrões pela normalização dos dados. Este processo tem como objetivo padronizar a distribuição dos valores das características para uma faixa de valores comuns, como -1 a 1. A normalização dos dados também é utilizada para auxiliar o classificador a convergir com maior facilidade durante a etapa de treinamento.

Após a normalização, a base de características é dividida em dois grupos: base de treino e base de teste. Nesta etapa são utilizadas diversas proporções para divisão da base de características, como mostrado na Tabela 3. Em cada proporção foram repetidos 5 experimentos com divisão aleatória, com o objetivo de verificar a consistência dos resultados. Em outras palavras, a repetição dos experimentos busca analisar se a função de classificação sofre grandes variações a cada experimento.

Tabela 3 – Proporções de treino/teste utilizadas nos experimentos de classificação.

Proporção	Treino	Teste
50/50	50%	50%
60/40	60%	40%
70/30	70%	30%
80/20	80%	20%

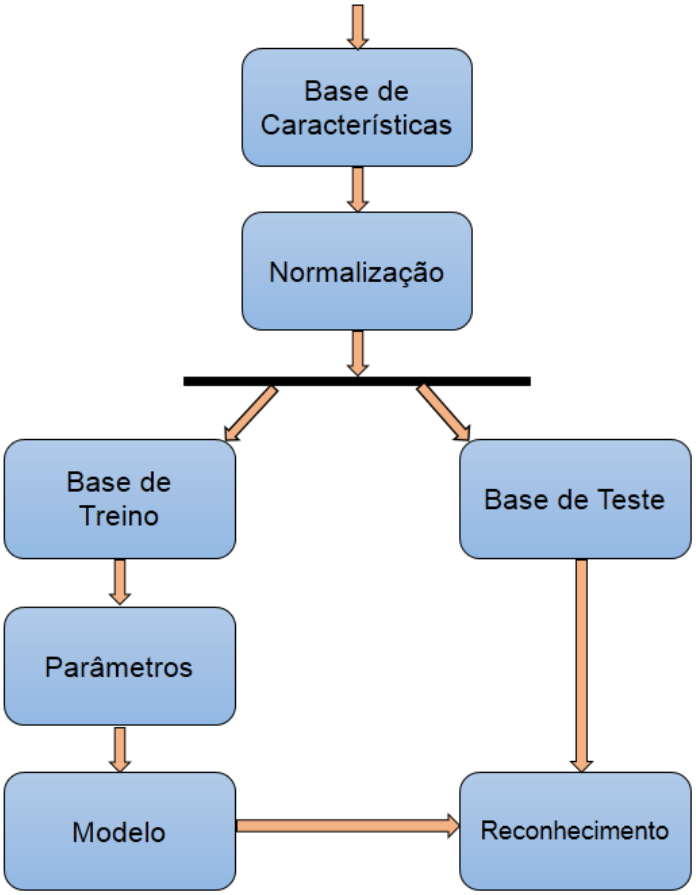

Figura 20 – Fluxo de atividades na etapa de reconhecimento de padrões.

A função de núcleo utilizada neste trabalho foi a função de base radial, também chamada *kernel* RBF. Devido a natureza aleatória da divisão da base, foram estimados em cada experimento os parâmetros C e γ, que representam o custo, e o grau de complexidade da função de mapeamento, respectivamente. Os parâmetros foram estimados com o *script* em *python grid.py*, que é fornecido no pacote LIBSVM (CHANG; LIN, 2011). Utilizando validação cruzada, o *script grid.py* busca a combinação dos parâmetros C e γ que retorne o melhor percentual de acerto para a base de características utilizada.

Ao final da etapa de treinamento é gerado o modelo de classificação, que consiste dos vetores de suporte utilizados pela MVS para classificar as amostras de teste. A base de testes é completamente desconhecida do modelo de classificação gerado. Desta forma, é possível simular condições reais de diagnóstico, em que as amostras de teste representariam indivíduos novos.

4.5 Validação de resultados

Após a etapa de reconhecimento de padrões é necessário validar os resultados obtidos. Este processo é realizado utilizando as métricas descritas na Seção 3.6.2: acurácia, sensibilidade e especificidade. A etapa de validação dos resultados tem como objetivo avaliar o desempenho do método proposto e também discriminar seus pontos positivos e negativos, para que se possa buscar melhorias em trabalhos futuros.

4.6 Considerações finais

Este capítulo descreveu, minuciosamente, o método proposto para classificação de massas em imagens de mamografia. Foram detalhadas cada uma das etapas que compõem o método, e apresentadas as adaptações empregadas para utilização dos descritores de forma propostos nesta dissertação.

No próximo capítulo serão apresentados os resultados obtidos pelo método proposto. Será feita uma discussão acerca dos valores obtidos, e também uma comparação com os trabalhos descritos na literatura, com o objetivo de contextualizar a relevância da pesquisa realizada neste trabalho.

5 Resultados e Discussões

Neste capítulo, são apresentados os resultados obtidos pelo método proposto. Destaca-se que o propósito dos experimentos realizados é investigar a aplicabilidade dos descritores de forma explorados nesta dissertação, para que se possa contribuir com medidas discriminativas dos padrões de malignidade e benignidade de massas em imagens de mamografia. Em outras palavras, o objetivo dos experimentos realizados é identificar o conjunto de características que mais contribua para o diagnóstico de câncer de mama. Além disso, é feita uma discussão a respeito dos resultados obtidos, e também uma análise comparativa com os trabalhos apresentados no Capítulo 2.

A apresentação dos resultados inicia-se pelos experimentos realizados com cada grupo de descritores individualmente. Em seguida, uma seção é dedicada para apresentar as taxas de acerto obtidas com a combinação de descritores. E por fim, são apresentados os resultados utilizando todo o conjunto de características (com seleção de variáveis e também sem seleção de variáveis).

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

5.1 Experimentos com descritores individuais

A Tabela 4 mostra os resultados obtidos utilizando o descritor *Relief Index*. Conforme descrito na Seção 4.3 foram utilizadas as duas formas do descritor, gerando portanto duas características. A maior taxa de acerto foi obtida na proporção 80/20 com médias de 68,87% de acurácia, 73,87% de sensibilidade e 63,87% de especificidade.

Tabela 4 – Resultados utilizando o descritor *Relief index* (2 características).

Treino/Teste	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**
50/50	62,39 ± 5,37	60,65 ± 27,66	64,13 ± 15,93
60/40	65,73 ± 5,02	68,55 ± 23,41	62,90 ± 15,88
70/30	65,70 ± 2,92	63,66 ± 8,14	67,74 ± 4,05
80/20	**68,87 ± 4,65**	**73,87 ± 12,00**	**63,87 ± 5,53**

A Tabela 5 mostra os resultados obtidos utilizando o descritor *Average Slope*. Como detalhado na Seção 4.3, este descritor produz apenas uma característica e obteve, no

seu melhor caso, médias de 69,68% de acurácia, 61,94% de sensibilidade e 77,42% de especificidade na proporção 70/30 para treino e teste.

Tabela 5 – Resultados utilizando o descritor *Average Slope* (1 característica).

Treino/Teste	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**
50/50	67,03 ± 1,10	58,71 ± 7,85	75,35 ± 4,82
60/40	68,15 ± 2,93	60,32 ± 8,58	75,97 ± 5,38
70/30	**69,68 ± 3,21**	**61,94 ± 12,31**	**77,42 ± 5,47**
80/20	69,03 ± 4,01	64,52 ± 7,29	73,55 ± 2,94

A Tabela 6 mostra os resultados obtidos utilizando os descritores *Section Area* e *Section Convolution*. Como descrito na Seção 4.3, foram utilizadas nestes descritores, além da imagem padrão, as imagens obtidas a partir das quantizações de 6 e 7 bits. Desta forma, foi gerado um conjunto de 60 características. Os melhores resultados foram alcançados na proporção 60/40 com valores de 74,52%, 83,39% e 65,65% de médias de acurácia, sensibilidade e especificidade, respectivamente.

Tabela 6 – Resultados utilizando os descritores *Section Area* e *Section Convolution* nas quantizações 6, 7 e 8 bits (60 características).

Treino/Teste	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**
50/50	72,71 ± 1,71	80,65 ± 2,04	64,77 ± 2,95
60/40	**74,52 ± 2,90**	**83,39 ± 5,99**	**65,65 ± 3,64**
70/30	71,94 ± 2,33	80,65 ± 2,31	63,23 ± 2,79
80/20	66,94 ± 2,25	81,94 ± 5,64	51,94 ± 7,08

A Tabela 7 mostra os resultados obtidos utilizando os descritores de distribuição de forma D1dist, D2dist e D3dist em suas abordagens 2D e 3D. Foi gerado um total de 12 características para este experimento. O melhor resultado foi obtido na proporção 50/50 com médias de acurácia, sensibilidade e especificidade de 83,10%, 86,06% e 80,13%, respectivamente.

5.2 Experimentos com combinações de descritores

Durante a fase de testes foram feitos diversos experimentos com várias combinações de descritores. Nesta seção é destacada a combinação que produziu o melhor resultado. A combinação consiste dos descritores *Section Convolution*, e D1dist, D2dist e D3dist nas

Tabela 7 – Resultados utilizando os descritores de distribuição de forma D1dist, D2dist e D3dist em suas abordagens 2D e 3D (12 características).

Treino/Teste	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
50/50	**83,10 ± 2,22**	**86,06 ± 4,45**	**80,13 ± 2,75**
60/40	81,45 ± 1,49	83,06 ± 3,22	79,84 ± 1,89
70/30	82,58 ± 1,50	86,45 ± 2,86	78,71 ± 1,14
80/20	78,06 ± 1,85	85,48 ± 1,89	70,65 ± 4,08

abordagens 2D e 3D. Este conjunto de descritores produz um total de 42 características, e entre todas as combinações testadas, foi a que obteve as melhores taxas de acerto. O melhor resultado desta combinação foi obtido na proporção 70/30, sendo alcançado, respectivamente, 92,15%, 91,40% e 92,90% de acurácia, sensibilidade e especificidade médias, como mostrado na Tabela 8.

Tabela 8 – Resultado da combinação dos descritores *Section Convolution*, e D1dist, D2dist e D3dist nas abordagens 2D e 3D (42 características).

Treino/Teste	Acurácia (%)	Sensibilidade (%)	Especificidade (%)
50/50	90,71±1,97	90,71±1,85	90,71±3,20
60/40	90,08±1,44	88,55±2,76	91,61±1,83
70/30	**92,15±1,06**	**91,40±1,44**	**92,90±1,32**
80/20	90,16±2,32	90,97±5,11	89,35±3,52

5.3 Experimentos com todos os descritores

Nesta seção são apresentados os resultados obtidos utilizando todos os descritores abordados neste trabalho. Inicialmente são apresentados os resultados para o conjunto completo de descritores sem seleção de características (Tabela 9). A combinação de todos os descritores gera um vetor com 75 características, e o melhor resultado foi obtido na proporção 60/40, com 91,29% de acurácia, 91,13% de sensibilidade e 91,45% de especificidade médias.

Neste trabalho foi também explorada a utilização do Algoritmo Genético para seleção de características. Na Tabela 10 são mostrados os resultados desse experimento. A função de aptidão utilizada foi a acurácia do classificador MVS. Os parâmetros utilizados foram: tamanho da população 40, número de gerações 90, probabilidade de *crossover* 65% e probabilidade de mutação 5%. Estes parâmetros foram selecionados empiricamente após

Tabela 9 – Resultados utilizando todo o conjunto de descritores sem seleção de características (75 características).

Treino/Teste	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**
50/50	89,68 ± 1,69	89,68 ± 3,09	89,68 ± 1,53
60/40	**91,29 ± 0,50**	**91,13 ± 1,66**	**91,45 ± 1,01**
70/30	89,89 ± 1,86	90,11 ± 2,85	89,68 ± 2,49
80/20	89,19 ± 1,37	92,26 ± 2,28	86,13 ± 2,51

diversos testes realizados. Em cada teste buscava-se determinar o conjunto mais viável de parâmetros, observando os seguintes aspectos: tempo de execução do algoritmo, número de características selecionadas, e resultados obtidos com o conjunto selecionado. Tendo em vista estes aspectos, chegou-se à definição dos parâmetros. Após a seleção das variáveis mais relevantes, o conjunto de características foi reduzido a 21 características. A partir deste novo conjunto, foi obtido como melhor resultado 92,58% de acurácia, 92,80% de sensibilidade e 92,28% de especificidade médias na proporção 80/20.

Tabela 10 – Resultados do experimento com todo o conjunto de descritores utilizando Algoritmo Genético para seleção de características (21 características).

Treino/Teste	**Acurácia (%)**	**Sensibilidade (%)**	**Especificidade (%)**
50/50	91,55±1,10	90,18±0,75	92,97±2,71
60/40	91,53±0,88	91,31±1,69	91,80±2,20
70/30	90,43±1,22	87,56±2,40	93,28±0,91
80/20	**92,58±1,89**	**92,80±3,14**	**92,28±2,77**

5.4 Discussão

Um dos primeiros aspectos que pode ser observado ao analisar os resultados apresentados nas Seções 5.1, 5.2 e 5.3 é que as taxas de acerto são mais altas quando são utilizados descritores combinados. Os descritores *Relief Index* e *Average Slope*, quando utilizados em experimentos de maneira individual (Tabela 4 e Tabela 5), produziram como melhor resultado 68,87% e 69,68% de médias de acurácia, respectivamente. Estes valores foram os mais baixos entre os experimentos realizados e não são considerados como satisfatórios. Uma das possíveis causas para a baixa taxa de acerto com estes descritores pode ser a produção de características não-discriminativas para o problema específico deste trabalho. Outra possível causa seria o fato de o processo de classificação com os

descritores *Relief Index* e *Average Slope* ter sido realizado com um número muito pequeno de características (2 e 1, respectivamente).

Nos experimentos em que os descritores utilizados produziam um número maior de características, foi observado uma certa melhora nas taxas de acerto. Quando foram utilizados os descritores *Section Area* e *Section Convolution* (Tabela 6), por exemplo, foi obtido como melhor resultado, em média, 74,52% de acurácia, com 83,39% de sensibilidade e 65,65% de especificidade. Destaca-se que houve uma grande melhora nas taxas de sensibilidade, o que indica uma melhor capacidade para identificar indivíduos doentes.

Entre os experimentos realizados com grupos individuais de descritores o que mais se destacou foi aquele em que se utilizou os descritores de distribuição de forma (Tabela 7). Em termos de taxas de acerto e consistência de valores, esta combinação foi a que produziu os resultados mais promissores. Os resultados se mostraram balanceados em todas as proporções, apresentando desvios-padrão relativamente baixos.

Com o objetivo de alcançar as maiores taxas de acerto, foi realizado uma grande quantidade de testes, que consistiam de experimentos de classificação com diversas combinações de descritores. Buscava-se encontrar o conjunto de características mais discriminativas para classificar massas malignas e benignas. O conjunto que produziu as maiores taxas de acerto foi o que era composto pelos descritores *Section Convolution* em conjunto com os descritores de distribuição de forma D1dist, D2dist e D3dist (Tabela 8). Neste conjunto de características foram obtidos como melhores resultados médias de 92,15% de acurácia, 91,40% de sensibilidade e 92,90% de especificidade. Ao analisar estes resultados é possível observar uma menor variação dos valores obtidos entre as proporções de treino/teste utilizadas. Além disso, a capacidade de identificar indivíduos doentes e não-doentes foi aumentada expressivamente em relação aos experimentos anteriores. Os resultados obtidos com esta combinação possibilitaram, ainda, a produção de um artigo científico, que foi aceito e apresentado no evento *7th Latin American Conference on Networked and Electronic Media* - LACNEM 2017.

No experimento em que foi utilizado todo o conjunto de descritores (Tabela 9), sem seleção de características, foram obtidos bons resultados, próximos à faixa de 90%. Entretanto, foi observada uma pequena diminuição nas taxas de acerto, em relação ao que já tinha sido alcançado com a melhor combinação de descritores. As melhores taxas de acerto para o conjunto completo de características foram 91,29% de acurácia, 91,13% de sensibilidade e 91,45% de especifidade médias.

O melhor resultado deste trabalho foi obtido utilizando todos os descritores, porém com o Algoritmo Genético para seleção das características mais discriminantes (Tabela 10). A etapa de seleção de características mostrou grande eficiência na determinação das variáveis mais relevantes, de modo que o conjunto de características foi reduzido, de 75 para 21 características, e as taxas de acerto ainda obtiveram, em média, uma melhora de 2 pontos percentuais. No geral, os resultados se mantiveram balanceados e consistentes em todas as proporções, alcançando valores acima de 90%. Além disso, os valores de desvios-padrão continuaram baixos, mantendo-se, em média, inferiores a 2%.

5.5 Comparação com outros trabalhos

Nesta seção é apresentada uma análise comparativa dos resultados obtidos pelo método proposto em relação aos trabalhos relacionados, mencionados no Capítulo 2. Destaca-se que realizar uma comparação rigorosa de resultados acaba se tornando uma tarefa difícil, devido ao fato de os trabalhos possuírem diferentes metodologias, nas quais são utilizadas diversas bases de imagens com diferentes quantidades de amostras, além de utilizarem diferentes métricas de avaliação. Entretanto, é possível elaborar alguns comentários comparativos em relação a esses trabalhos. A Tabela 11 apresenta um resumo desta comparação, onde são destacados os valores de média de acurácia e de área sobre a curva ROC (Az) obtidos em cada trabalho.

Em relação aos trabalhos que utilizam somente textura, é possível observar que a performance do método proposto foi superior a quase todos os trabalhos, em termos de acurácia e área sobre a curva ROC. A exceção foi o trabalho de Beura, Majhi e Dash (2015), que apesar de ter atingido uma taxa de acurácia maior, utilizou um número de imagens equivalente a menos da metade utilizada em nossos experimentos. Entretanto, deve-se mencionar que o trabalho de Rocha et al. (2016) utilizou um grande número de imagens e obteve taxas de acerto próximas do método proposto. Como mencionado na Seção 2.5, as abordagens que utilizam textura não possuem como limitação a necessidade de uma segmentação rigorosa em suas imagens, quando comparadas à abordagens que utilizam somente forma. Deste modo, acaba sendo possível (e viável) para estas abordagens utilizar um número maior de imagens.

Entre os trabalhos com abordagens baseadas somente em forma, alguns obtiveram taxas de acerto maiores que as do método proposto. Entretanto, a quantidade de amostras

Tabela 11 – Comparação do método proposto com os trabalhos relacionados.

	Trabalho	Técnica(s)	Classificador	Base	ROIs (Ben./ Mal.)	Acurácia	Az
Textura	(HUSSAIN et al., 2014)	*Gabor Filter Bank*	MVS	DDSM	512 (256/256)	85,53%	0,87
	(DHAHBI; BARHOUMI; ZAGROUBA, 2015)	*Curvelet*	KNN	mini-MIAS	116/ (66/50)	81,35%	-
	(BEURA; MAJHI; DASH, 2015)	2D-DWT GLCM	BPNN	DDSM	250 (129/121)	97,4%	-
	(ROCHA et al., 2016)	GLCM Índices de Diversidade	MVS	DDSM	1155 (530/625)	88,31%	0,88
Forma	(CHEIKHROUHOU; DJEMAL; MAAREF, 2011)	*Protuberance Selection*	MVS	DDSM	242 (128/114)	-	0,93
	(LIU; LIU; FENG, 2011)	NRL RGO	MVS	DDSM	309 (142/167)	76%	-
	(ABDAHEER; KHAN, 2011)	*Area matching*	MIAS	MIAS	150 (71/79)	94%	-
	(GÖRGEL; SERTBAS; UCAN, 2013)	LSG-SWT	MVS	Istambul Univ.	78 (43/35)	93,59%	-
	(WAJID; HUSSAIN, 2015)	LESH	MVS	INbreast	117 (-/-)	99,73%	-
Forma e Textura	(LIU; TANG, 2014)	NRL RGO	MVS	DDSM	826 (418/408)	-	0,96
	(ROUHI et al., 2015)	GLCM *Zernike Moments*	MLP	DDSM	170 (74/96)	96,47%	0,95
	(VALARMATHIE; SIVAKRITHIKA; DINAKARAN, 2016)	GLCM Geometria	MLP	mini-MIAS	332 (-/-)	98%	0,96
Aprendizagem Profunda	(AREVALO et al., 2015)	CNN	MVS	BCDR	736 (310/426)	-	0,86
	(KAUR, 2016)	Textura Geometria	CNN	MIAS	30 (-/-)	96,66%	-
	(ABBAS, 2016)	SURF LBPV	DBN	DDSM/MIAS	600 (300/300)	91,5%	0,91
	Método Proposto	**Todos os descritores (Com Seleção de Carac.)**	**MVS**	**DDSM**	**620 (310/310)**	**92,58%**	**0,92**
		***Section Convolution* D1dist, D2dist, D3dist**	**MVS**	**DDSM**	**620 (310/310)**	**92,15%**	**0,92**

utilizadas neste trabalho é expressivamente maior do que as utilizadas em todos os outros trabalhos, sendo de 2 vezes até quase 8 vezes maior. Um número maior de imagens aumenta a heterogeneidade da amostra, o que leva a experimentos sob circunstâncias mais realistas.

A combinação de características de textura e forma é, teoricamente, uma das abordagens mais poderosas para classificação de massas em imagens de mamografia, pois possibilita a análise dessas regiões sob mais de uma perspectiva. Quando bons descritores de forma e textura são combinados, geralmente é produzido um conjunto de características com alto poder discriminativo, pois as características geradas complementam-se umas as outras. Apesar de ter obtido taxas de acerto inferiores aos trabalhos que utilizaram esta poderosa combinação, o número de imagens utilizas no método proposto só não foi superior ao do trabalho de Liu e Tang (2014). Destaca-se que foi possível obter um resultado próximo ao destes autores utilizando exclusivamente descritores de forma, e reforça-se que o objetivo deste trabalho não é superar todas as metodologias descritas na literatura, mas contribuir com a apresentação de novos descritores de forma no contexto de classificação de câncer de mama.

Nos últimos anos, as técnicas de aprendizagem profunda têm revolucionado o cenário de aplicações de visão computacional, superando diversas metodologias tradicionais. No contexto específico de metodologias para diagnóstico de câncer de mama, acredita-se que o potencial dessas técnicas ainda não tenha sido completamente explorado. Desta forma, ainda foi possível para o método proposto, que trata-se de uma metodologia tradicional, obter resultados superiores aos métodos de aprendizagem profunda descritos na literatura.

5.6 Considerações finais

Neste capítulo foram apresentados e discutidos os resultados dos experimentos realizados no desenvolvimento do método proposto. Também foi feito um comparativo com os trabalhos relacionados, apresentados no Capítulo 2, como forma de analisar a relevância da pesquisa desenvolvida.

No geral, os resultados obtidos podem ser considerados promissores, sendo comparáveis aos melhores trabalhos descritos na literatura. As maiores taxas de acerto foram obtidas com a utilização do algoritmo genético para seleção das características mais relevantes, onde foi possível alcançar valores médios de acurácia, sensibilidade e especificidade superiores a 92%.

No próximo capítulo são apresentas as conclusões a respeito do trabalho desenvolvido nesta dissertação. Também serão apontadas algumas de suas limitações, bem como apresentadas sugestões para trabalhos futuros.

6 Conclusão

O câncer de mama é segunda maior causa de morte entre as mulheres em todo o mundo. As altas taxas de mortalidade evidenciam a importância do desenvolvimento de sistemas de auxílio ao diagnóstico de câncer de mama, tendo em vista que o diagnóstico precoce possibilita maiores chances de cura. Neste contexto, técnicas que auxiliem o desenvolvimento de sistemas CAD e CADx são de grande contribuição para o meio científico.

Nesta dissertação foi apresentado um método de classificação de massas em imagens de mamografia para diagnóstico de câncer de mama. O objetivo deste trabalho foi apresentar a viabilidade da utilização de descritores de forma dentários e descritores de distribuição de forma para a discriminação dos padrões de malignidade e benignidade das massas mamárias.

Foram apresentados os descritores *Relief Index, Average Slope, Section Area, Section Convolution*, D1dist, D2dist e D3dist que até então foram utilizados somente em estudos relacionados à ecologia, nos quais eram aplicados para analisar a morfologia dentária de várias espécies de mamíferos. Neste trabalho, entretanto, este conjunto de descritores foi utilizado para extração de características de forma de massas mamárias, com objetivo de de identificar e diferenciar os padrões de malignidade e benignidade.

Após a etapa de extração de características, o conjunto gerado foi submetido a uma série de experimentos de classificação na etapa de reconhecimento de padrões, no qual foi utilizado o classificador MVS. O melhor resultado foi obtido ao utilizar o Algoritmo Genético para seleção automática das características mais relevantes, onde foram alcançadas médias de 92,58% de acurácia, 92,80% de sensibilidade e 92,28% de especificidade.

Os resultados alcançados neste trabalho são considerados promissores, sendo comparáveis aos melhores trabalhos descritos na literatura. A partir dos valores obtidos foi possível demonstrar que a utilização de descritores de forma dentários e descritores de distribuição de forma é bastante viável para caracterização de massas mamárias e, portanto, poderão ser utilizados em metodologias CADx para diagnóstico de câncer de mama.

Apesar de ter obtido bons resultados no geral, o método proposto possui alguns aspectos que podem ser melhorados, entre os quais citam-se: a realização de experimentos

com imagens de uma única base, a seleção não-automática de parâmetros nos descritores *section area* e *section convolution*, e a utilização de um único classificador na etapa de reconhecimento de padrões.

Conforme apresentado no Capítulo 2, existe um grande interesse da comunidade científica em pesquisas na área médica, devido a relevância deste tema. Desta forma, espera-se que os frutos da pesquisa realizada neste trabalho possam ser utilizados no futuro em novas metodologias, de modo que sejam sanadas algumas de suas limitações. Assim, como sugestão para trabalhos futuros enumeram-se:

1. Realizar experimentos com imagens de outras bases, e se possível com um número maior de amostras;
2. Combinar descritores de textura com os descritores apresentados neste trabalho, para criar um método robusto de classificação;
3. Investigar métodos para definir, de maneira automática, a melhor quantidade de cortes nos descritores *Section Area* e *Section Convolution*;
4. Investigar a aplicação de novos descritores de forma dentários, como *Orientation Patch Count* e *Dirichlet Normal Energy*;
5. Realizar testes com outras técnicas de aprendizado de máquina, como abordagens de aprendizem profunda, ou outros classificadores como *Random Forests, Multilayer Perceptron* e AdaBoost;
6. Investigar a aplicação dos descritores de forma dentários e descritores de distribuição de forma em metodologias para classificação de outros tipos de cânceres, tais como câncer de pulmão e próstata.

Referências

ABBAS, Q. Deepcad: A computer-aided diagnosis system for mammographic masses using deep invariant features. *Computers*, Multidisciplinary Digital Publishing Institute, v. 5, n. 4, p. 28, 2016.

ABDAHEER, M.; KHAN, E. An automatic and simple breast tumor classification using area matching. In: IEEE. *2011 International Conference on Image Information Processing (ICIIP)*. Shimla, India, 2011. p. 1–5.

ACS. *American Cancer Society - Breast Cancer*. 2017. Último Acesso: 07/12/2017. Disponível em: <https://www.cancer.org/cancer/breast-cancer.html>.

AMERICAN COLLEGE OF RADIOLOGY. *Breast imaging reporting and data system*. Silver Spring, MD. EUA: American College of Radiology, 1998.

AREVALO, J.; GONZÁLEZ, F. A.; RAMOS-POLLÁN, R.; OLIVEIRA, J. L.; LOPEZ, M. A. G. Convolutional neural networks for mammography mass lesion classification. In: IEEE. *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. Milan, Italy, 2015. p. 797–800.

BEURA, S.; MAJHI, B.; DASH, R. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. *Neurocomputing*, Elsevier, v. 154, p. 1–14, 2015.

BLAND, M. *An introduction to medical statistics*. Oxford, UK: Oxford University Press, 2015.

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

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

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

CHAVES, A. d. C. F. Extração de regras fuzzy para máquinas de vetores suporte (svm) para classificação em múltiplas classes. *Rio de Janeiro*, 2006.

CHEIKHROUHOU, I.; DJEMAL, K.; MAAREF, H. Protuberance selection descriptor for breast cancer diagnosis. In: IEEE. *3rd European Workshop on Visual Information Processing (EUVIP) 2011*. Paris, France, 2011. p. 280–285.

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

DA, T. K. F.; SÉBASTIEN, L.; YVINEC, M. 3d alpha shapes. In CGAL User and Reference Manual. CGAL Editorial Board, 4.10 edition, 2017.

DHAHBI, S.; BARHOUMI, W.; ZAGROUBA, E. Breast cancer diagnosis in digitized mammograms using curvelet moments. *Computers in biology and medicine*, Elsevier, v. 64, p. 79–90, 2015.

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

FENTON, J. J.; TAPLIN, S. H.; CARNEY, P. A.; ABRAHAM, L.; SICKLES, E. A.; D'ORSI, C.; BERNS, E. A.; CUTTER, G.; HENDRICK, R. E.; BARLOW, W. E. et al. Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, Mass Medical Soc, v. 356, n. 14, p. 1399–1409, 2007.

FERNANDES, V. P.; KANEHISA, R. F.; JR, G. B.; SILVA, A. C.; PAIVA, A. C. de. Lung nodule classification based on shape distributions. In: ACM. *Proceedings of the 31st Annual ACM Symposium on Applied Computing*. Pisa, Italy, 2016. p. 84–86.

GIGER, M. L. Computer-aided diagnosis of breast lesions in medical images. *Computing in Science & Engineering*, AIP Publishing, v. 2, n. 5, p. 39–45, 2000.

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

GOLDBERG, D. E. et al. *Genetic algorithms in search optimization and machine learning*. Boston, MA: Addison-Wesley Reading, 1989.

GONZALEZ, R. C.; WOODS, R. E. *Processamento digital de imagens*. Nova Jersey, EUA: Pearson Prentice Hall, 2010.

GÖRGEL, P.; SERTBAS, A.; UCAN, O. N. Mammographical mass detection and classification using local seed region growing–spherical wavelet transform (lsrg–swt) hybrid scheme. *Computers in biology and medicine*, Elsevier, v. 43, n. 6, p. 765–774, 2013.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. Cambridge, MA: MIT Press, 2001.

HAUPT, R. L.; HAUPT, S. E. *Practical genetic algorithms*. New Jersey, NY: John Wiley & Sons, 2004.

HAYKIN, S. *Redes neurais: princípios e prática*. Porto Alegre, Brasil: Bookman Editora, 2007.

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

HUSSAIN, M.; KHAN, S.; MUHAMMAD, G.; AHMED, I.; BEBIS, G. Effective extraction of gabor features for false positive reduction and mass classification in mammography. *Appl. Math*, v. 8, n. 1L, p. 397–412, 2014.

INCA. *Consenso para o Controle do Câncer de Mama*. 2004. Último Acesso: 04/12/2017. Disponível em: <http://www.inca.gov.br/publicacoes/Consensointegra.pdf>.

INCA. *Instituto Nacional do Câncer - Câncer - Tipo - Mama*. 2016. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama. Último acesso: 10/11/2017.

INCA. *Instituto Nacional do Câncer - O que é o câncer*. 2017. Disponível em: http://www2.inca.gov.br/wps/wcm/connect/cancer/site/oquee. Último acesso: 10/11/2017.

KAUR, P. Mammogram image nucleus segmentation and classification using convolution neural network classifier. 2016.

LIU, X.; LIU, J.; FENG, Z. Mass classification in mammography with morphological features and multiple kernel learning. In: IEEE. *5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE) 2011*. Wuhan, China, 2011. p. 1–4.

LIU, X.; TANG, J. Mass classification in mammograms using selected geometry and texture features, and a new svm-based feature selection method. *IEEE Systems Journal*, IEEE, v. 8, n. 3, p. 910–920, 2014.

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

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

NATIONAL BREAST CANCER FOUNDATION. *National Breast Cancer - Breast Cancer*. 2017. Último Acesso: 18/12/2017. Disponível em: <http://www.nationalbreastcancer.org/breast-tumors>.

NUNES, A. P.; SILVA, A. C.; PAIVA, A. C. de. Detection of masses in mammographic images using simpson's diversity index in circular regions and svm. In: *Machine Learning and Data Mining in Pattern Recognition*. Leipzig, Alemanha: Springer, 2009. p. 540–553.

OSADA, R.; FUNKHOUSER, T.; CHAZELLE, B.; DOBKIN, D. Shape distributions. *ACM Transactions on Graphics (TOG)*, ACM, v. 21, n. 4, p. 807–832, 2002.

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

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

PLYUSNIN, I.; EVANS, A. R.; KARME, A.; GIONIS, A.; JERNVALL, J. Automated 3d phenotype analysis using data mining. *PLoS One*, Public Library of Science, v. 3, n. 3, p. e1742, 2008.

ROCHA, S. V. da; JUNIOR, G. B.; SILVA, A. C.; PAIVA, A. C. de; GATTASS, M. Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution. *Expert Systems with Applications*, Elsevier, v. 66, p. 7–19, 2016.

ROUHI, R.; JAFARI, M.; KASAEI, S.; KESHAVARZIAN, P. Benign and malignant breast tumors classification based on region growing and cnn segmentation. *Expert Systems with Applications*, Elsevier, v. 42, n. 3, p. 990–1002, 2015.

SAMPAIO, W. B.; DINIZ, E. M.; SILVA, A. C.; PAIVA, A. C. D.; GATTASS, M. Detection of masses in mammogram images using cnn, geostatistic functions and svm. *Computers in Biology and Medicine*, Elsevier, v. 41, n. 8, p. 653–664, 2011.

SILVA, T. F. d. B. *Diferenciação do Padrão de Malignidade e Benignidade de Massas em Mamografias Utilizando Características Geométricas e Máquina de Vetor de Suporte.* Dissertação de Mestrado — Universidade Federal do Maranhão, Programa de Pós-Graduação em Ciência da Computação. São Luis - MA, 2016.

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

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

VALARMATHIE, P.; SIVAKRITHIKA, V.; DINAKARAN, K. Classification of mammogram masses using selected texture, shape and margin features with multilayer perceptron classifier. *Biomedical Research*, Biomedical Research, 2016.

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

WAJID, S. K.; HUSSAIN, A. Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. *Expert Systems with Applications*, Elsevier, v. 42, n. 20, p. 6990–6999, 2015.

ZUIDERVELD, K. Contrast limited adaptive histogram equalization. In: ACADEMIC PRESS PROFESSIONAL, INC. *Graphics gems IV*. San Diego, EUA, 1994. p. 474–485.