UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ÁREA DE CIÊNCIA DA COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA
DE ELETRICIDADE

Detecção automática de massas em imagens mamográficas usando particle swarm optimization (PSO) e índice de diversidade funcional

Otilio Paulo da Silva Neto

Dissertação de Mestrado

São Luís
04 de Março de 2016

Otilio Paulo da Silva Neto

Detecção automática de massas em imagens mamográficas usando particle swarm optimization (PSO) e índice de diversidade funcional

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade, da Universidade Federal do Maranhão, como requisito para o título de Mestre em Engenharia Elétrica na área de concentração Ciência da Computação.

Orientador: Dr. Aristófanes Corrêa Silva

Co-orientador: Dr. Anselmo Cardoso de Paiva

São Luís
04 de Março de 2016

Silva Neto, Otilio Paulo da.

Detecção automática de massas em imagens mamográficas usando particle swarm optimization (PSO) e índice de diversidade funcional / Otilio Paulo da Silva Neto. – São Luís, 2016.

82 f.

Impresso por computador (fotocópia).

Orientador: Aristófanes Corrêa Silva.

Co-orientador: Anselmo Cardoso de Paiva.

Dissertação (Mestrado) – Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade, 2016.

1. Particle swarm optimization - Câncer de mama. 2. Graph clustering. 3. Índice de diversidade funcional. I. Título. CDU 004:618.19-006

Este trabalho é dedicado em memória de minha mãe, Raimunda Soares Medina, pelo exemplo de mulher e dedicação aos filhos. Ah... Raimundinha, que Deus a tenha!

"Não são as espécies mais fortes que sobrevivem nem as mais inteligentes,
e sim as mais sucetíveis às mudanças."

Charles Darwin

Agradecimentos

Primeiramente agradeço a Deus, por ter me permitido chegar onde cheguei. Pois sem ele nada disso seria possível.

Aos meus pais, Marcos Paulo e Raimunda Medina, por todo o carinho, atenção, amor, confiança, ensino e inspiração em toda a minha vida.

À minha querida esposa Dalla Cristiane, pela compreensão, paciência, carinho e amor, estando sempre ao meu lado. Obrigado, por fazer parte do meu sonho, afinal de contas esta conquista também e sua.

Aos meus maravilhosos filhos, Dalila Natanne, Paulo Otilio e Débora Ohana, pelo apoio, carinho e compreensão durante toda nossa estadia em São Luís. Obrigado meus queridos filhos.

Aos meus irmãos que de forma direta ou indiretamente contribuíram para esta realização, em especial minha irmã Josilene Medina, por ter cuidado de minha filha Dalila e de meus cães, Layla e Lork.

À minha querida sogrinha do coração, Maria do Socorro, por sempre acreditar em mim e confiar seu tesouro Dalla Cristianne, para ser minha esposa.

Ao casal que adotei como meus pais em São Luis, Francisco Costa e Abadia Costa, por todo apoio e carinho, fundamentais como família, nos apoiando em tudo e para tudo. Sentirei saudades dos finais de semanas no sítio... Meus sinceros agradecimento.

À minha tia Santidade, que adotei como tia querida, pelos seus cuidados comigo e minha família. Ah... Saudade daquela gemada gostosa.

Aos meus amigos e orientadores, Dr. Aristófanes Corrêa e Anselmo Paiva, por acreditarem em mim, desde o início ao fim do mestrado. Sempre lembrarei dos ensinamentos e das broncas... Obrigado! Eu já mais teria conseguindo sem vocês. Por sempre estarem disponíveis e dispostos a ajudar-me, meus eternos agradecimentos.

Aos meus amigos, que sempre estiveram presentes no decorrer do mestrado: Gilberto Nunes, Thiago Pinheiro, Wener Sampaio, Antônio Oséas, Stelmo Magalhães, Darlan Quitanília, Giovanio Luca, Thamila Fontenele, Valéria Príscila, Whesley Dantas, pelos esclarecimentos às minhas dúvidas e pelo apoio nos momentos de dificuldades.

Por fim, a todos que, de forma direta ou indireta, contribuíram à realização deste sonho e por toda ajuda na conclusão desta pesquisa.

Resumo

O câncer de mama hoje é configurado no senário mundial como o mais comum entre as mulheres e o segundo que mais mata. Sabe-se que diagnosticado precocemente, a chance de cura é bem significativa, por outro lado, a descoberta tardia praticamente leva a morte. A mamografia é o exame mais comum que permite a descoberta precoce do câncer, esse procedimento consegue mostrar lesões nas fases iniciais, além de contribuir para a descoberta e o diagnóstico de lesões na mama. Sistemas auxiliados por computador, têm-se mostrado ferramentas importantíssimas, no auxilio a especialistas em diagnosticar lesões. Este trabalho propõe uma metodologia computacional para auxiliar na descoberta de massas em mamas densas e não densas. Dividida em 6 fases, esta metodologia se inicia com a aquisição da imagem da mama adquirida da *Digital Database for Screening Mammography (DDSM)*. Em seguida, na segunda fase é feito o pré-processamento para eliminar e realçar as estruturas da imagem. Na terceira fase executa-se a segmentação com o *Particle Swarm Optimization* (PSO) para encontrar as regiões de interesse (ROIs) candidatas a massa. A quarta fase é a redução de falsos positivos, que se subdivide em duas partes, sendo a redução pela distância e o *graph clustering*, ambos com o objetivo de remover ROIs indesejadas. Na quinta fase são extraídas as características de textura utilizando os índices de diversidade funcional (FD). Por fim, na sexta fase, utiliza-se o classificador máquina de vetores de suporte (SVM) para validar a metodologia proposta. Os melhores valores achados para as mamas não densas, resultaram na sensibilidade de 96,13%, especificidade de 91,17%, acurácia de 93,52%, a taxe de falsos positivos por imagem de 0,64 e a acurva *Free-response Receiver Operating Characteristic (FROC)* com 0,98. Os melhores achados para as mamas densas firam com a sensibilidade de 97,52%, especificidade de 92,28%, acurácia de 94,82%, uma taxa de falsos positivos por imagem de 0,38 e a curva FROC de 0,99. Os melhores achados com todas as mamas densas e não densas, apresentaram 95,36% de sensibilidade, 89,00% de especificidade, 92,00% de acurácia, 0,75 a taxa de falsos positivos por imagem e 0,98 a curva FROC.

Palavras-chave: *Particle Swarm Optimization (PSO); Graph Clustering; Câncer de Mama; Índice de Diversidade Funcional; .*

Abstract

Breast cancer is now set on the world stage as the most common among women and the second biggest killer. It is known that diagnosed early, the chance of cure is quite significant, on the other hand, almost late discovery leads to death. Mammography is the most common test that allows early detection of cancer, this procedure can show injury in the early stages also contribute to the discovery and diagnosis of breast lesions. Systems computer aided, have been shown to be very important tools in aid to specialists in diagnosing injuries. This paper proposes a computational methodology to assist in the discovery of mass in dense and non-dense breasts. This paper proposes a computational methodology to assist in the discovery of mass in dense and non-dense breasts. Divided into 6 stages, this methodology begins with the acquisition of the acquired breast image *Digital Database for Screening Mammography (DDSM)*. Then the second phase is done preprocessing to eliminate and enhance the image structures. In the third phase is executed targeting with the *Particle Swarm Optimization (PSO)* to find regions of interest (ROIs) candidates for mass. The fourth stage is reduction of false positives, which is divided into two parts, reduction by distance and *clustering graph*, both with the aim of removing unwanted ROIs. In the fifth stage are extracted texture features using the functional diversity indicia (FD). Finally, in the sixth phase, the classifier uses support vector machine (SVM) to validate the proposed methodology. The best values found for non-dense breasts, resulted in sensitivity of 96.13%, specificity of 91.17%, accuracy of 93.52%, the taxe of false positives per image 0.64 and acurva *free-response receiver operating characteristic (FROC)* with 0.98. The best finds for dense breasts hurt with the sensitivity of 97.52%, specificity of 92.28%, accuracy of 94.82% a false positive rate of 0.38 per image and FROC curve 0.99. The best finds with all the dense and non dense breasts Showed 95.36% sensitivity, 89.00% specificity, 92.00% accuracy, 0.75 the rate of false positives per image and 0, 98 FROC curve.

Keywords: *Particle Swarm Optimization(PSO); Índice de Diversidade Funcional; Graph Clustering; Câncer de Mama; Mamografia.*

Lista de Figuras

Lista de Tabelas

Lista de Siglas

CAD/CADx Computer-Aided Detection / Diagnosis

CLAHE Contrast-Limited Adaptive Histogram Equalization

DDSM Digital Database for Screening Mammography

dpmi Desvio Padrão Médio da Imagem

FAD Diversidade Funcional Abundante

FADa Índice de Diversidade Funcional Abundante

FADe Índice de Diversidade Funcional Abundante da Especie

FADp Índice de Diversidade Funcional Atributo Pixel

FD Diversidade Funcional

FFC Fator de Forma Circular

FP/i Taxa Média de Falso Positivos por Imagem

FROC Free-response Receiver Operating Characteristic

PSO Otimização por Enxame de Partículas (*Particle Swarm Optimization*)

RFP Redução de Falsos Positivos

ROC Receiver Operating Characteristic

ROI Regiao de Interesse

SMOTE Synthetic Minority Oversampling Technique

SVM Maquina de Vetores de Suporte

Sumário

Capítulo 1

Introdução

Com o grande aumento do câncer de mama no mundo, se faz cada vez mais necessário um diagnóstico precoce na busca à cura. Este aumento se dá em decorrência de diversos fatores, tais como: alimentação, estresse, cigarro, bebidas, dentre outros (MCPHERSON et al., 2000). Em pesquisa realizada pelo Instituto Nacional do Câncer, em especial, o câncer de mama é o segundo mais comum no mundo e o mais frequente entre as mulheres. Por outro lado, sabe-se que o câncer de mama quando diagnosticado cedo, tem grande chance de cura (INCA, 2015). Com base nesse cenário, o Governo juntamente com profissionais da saúde mundial mobilizam diversas campanhas para alertar a população feminina para os riscos que causam o câncer de mama.

As causas de mortes por câncer, correspondem a cerca de 30%, devidas a cinco principais riscos comportamentais e alimentares, sendo eles: índice de massa corporal elevado; baixo consumo de frutas, verduras e legumes; ausência de atividade física; uso de álcool e tabagismo (BOYLE et al., 2008).

O mecanismo mais comum para diagnosticar o câncer de mama é a mamografia, um exame radiológico que gera uma imagem em tons de cinza da mama. O especialista analisa e identifica de forma visual onde se encontra a lesão ou as lesões (GIGER, 2000). Com o uso da mamografia observou-se uma redução na taxa de mortalidade condicionada a essa patologia (SOCIETY, 2013).

Todavia, a avaliação do exame mamográfico é subjetiva, requerendo grande experiência do radiologista. Nas últimas décadas, técnicas computacionais vêm sendo desenvolvidos com o propósito de detectar automaticamente estruturas que possam estar associadas a tumores nos exames de mamografia, visando melhorar a taxa de detecção precoce de estruturas de interesse

ligadas ao câncer de mama (GIGER, 2000).

Essas técnicas computacionais motivaram o surgimento de diversas pesquisas ao longo das últimas décadas, no sentido de desenvolver sistemas computacionais para auxiliar especialistas a desempenhar seu papel de interpretar as imagens radiológicas, que são conhecidas como sistemas CAD (*Computer Aided Detection*) e CADx (*Computer Aided Diagnosis*), e já estão presentes em diversos centros de diagnóstico por imagem, aumentando as taxas de acerto na identificação precoce de doenças graves, como o câncer de mama (FENTON et al., 2007).

A detecção de massa motiva a criação de sistemas CAD, devido ao grande foco de investigação que o tema remete. Estima-se que a sensibilidade, que é a capacidade desses sistemas detectarem verdadeiros positivos, varia em média 88% a 99% (BOYLE et al., 2008). É claro que se podem encontrar taxas mais altas de sensibilidade na literatura, como 99% (BOSE et al., 2010); (GAO et al., 2010), sistemas que apresentam valores de sensibilidade próximos a média citada acima, podem ser considerados suficientes para auxiliar o especialista.

Os sistemas CAD e CADx fornecem uma segunda opinião, auxiliando o radiologista na interpretação de resultados que em muitos casos torna-se difícil devido às distorções que este tipo de imagem sofre no seu processo de aquisição, muitas vezes as imagens geradas, apresentam certas dificulAdades como má aquisição da imagem, densidade do tecido mamário, experiência do especialista, dentre outros na obtenção da lesão (MARTINS et al., 2009).

1.1 Problemática

Sistemas CAD e CADx veem sendo peça fundamental para auxiliar especialistas na análise de exames de imagem. Devido ao fato desses sistemas apresentarem melhorias no diagnóstico do exame e nas estruturas da imagem mamográfica, tais como: a densidade; a sobreposição dos tecidos; a representação da textura e na análise da imagem. Por isso há uma necessidade de construção de técnicas cada vez mais eficientes, capazes de gerar uma quantidade reduzida de falso positivos (MEERSMAN et al., 1998).

Diante do exposto, detalhamos a problemática desta dissertação da seguinte forma:

- Desenvolver uma metodologia para detectar massa em mamografias digital, com taxas de sensibilidade superior ao padrão que é de 85%;
- Fazer uso das técnicas de processamento de imagens para o melhoramento e realce da imagem mamográfica; detecção, segmentação e discriminação de regiões suspeitas;

- Identificar um padrão à distinção entre massa e não massa, através de descritores de textura das regiões internas da mama, na busca de reduzir falsos positivos minimizando os erros, e
- Criar um modelo genérico capaz de detectar novos individuo, através das técnicas de aprendizado de máquinas.

1.2 Objetivos

Apresentamos nesta seção os objetivos gerais e específicos a serem alcançados durante o desenvolvimento deste trabalho. Cada um deles é apresentado como segue:

1.2.1 Objetivo Geral

O objetivo geral do trabalho proposto é desenvolver um método de detecção automática de massas em imagens mamográficas digitais, utilizando o Algoritmo *Particle Swarm Optmization (PSO)* e o Índice de Diversidade Funcional.

1.2.2 Objetivos Específicos

Com o intuito de alcançar o objetivo geral pretendido, fez-se necessário o cumprimento dos seguintes objetivos específicos:

- Estudar o câncer de mama;
- Estudar o pré-processamento, filtros, realce local e algoritmos de agrupamento em imagens mamográficas;
- Estudar e implementar os algoritmos de segmentação em imagens mamográficas;
- Estudar e implementar o algoritmo de otimização por enxame de partículas (PSO).
- Estudar o processo de extração de características dos elementos segmentados das mamografias através de descritores de texturas usando índices de diversidade funcional;
- Validar a metodologia deste trabalho utilizando o classificador SVM.

1.3 Contribuições Científicas

Neste trabalho podemos destacar as seguintes contribuições cientificas:

- A criação de uma metodologia automática para segmentação de massa, através da *Particle Swarm Optimization (PSO)* e índices de diversidade funcional para imagens mamográficas digitais;
- Criação de uma estratégia de clusterização automática, adaptada para mamografia digital, utilizando o PSO;
- Construção de uma técnica de união das regiões vizinhas com o uso do *Graph Clustering*;
- Utilização dos índices de Diversidade Funcional para extração de características de textura das regiões da mama.

1.4 Estrutura do Trabalho

Além deste capítulo introdutório, há mais 5 capítulos, que completam esta dissertação e estão estruturados da seguinte forma:

No Capítulo 2, é apresentado os trabalhos relacionados, que servirão de base para comparação dos resultados da metodologia.

O Capítulo 3, apresenta o estado da arte e a fundamentação teórica necessária ao entendimento e construção desta metodologia.

O Capítulo 4, serão mostradas todas as etapas de desenvolvimento desta pesquisa, iniciando-se pela aquisição da imagem Digital da *Database for Screening Mammography (DDSM)*, seguida do pré-processamento, da segmentação usando o *Particle Swarm Optimization (PSO)*, a redução de falsos positivos, a extração de características, e por fim, o reconhecimento.

No Capítulo 5, são apresentados os resultados encontrados, as discussões e os casos de sucesso e falha da metologia.

E por fim, no Capítulo 6, é apresentado as conclusões inferidas a respeito da metodologia, juntamente com os trabalhos futuros para o melhoramento desta pesquisa.

Capítulo 2

Trabalhos Relacionados

A literatura disponível traz trabalhos reconhecidamente bons, que tratam do problema da detecção de massa em imagens mamográficas digitais que é o objetivo deste trabalho. Enumeram-se a seguir alguns desses trabalhos.

2.1 Trabalhos Relacionados

A metodologia proposta por (NUNES, 2009) de detecção de massas que utiliza o algoritmo de agrupamento K-means e a técnica de *Template Matching* para segmentar as regiões suspeitas de conterem massas. Esta metodologia foi testada utilizando 650 imagens mamográficas da base de dados *Digital Database for Screening Mammography (DDSM)*. Na etapa de segmentação das regiões de interesse foram segmentadas 603 massas da amostra, o que equivale a 92,77% dos casos, e também foram selecionas 2076 não-massas. Em seguida, foram extraídas medidas de geometria e textura de cada uma dessas regiões, sendo a textura descrita através do índice de diversidade de Simpson. Por fim, as informações são submetidas a uma Maquina de Vetores de Suporte (SVM) para que as regiões suspeitas sejam classificadas em massas ou não-massas. Na etapa de treinamento e teste foi realizadas através de seis diferentes tamanhos dos conjuntos de treino/teste, sendo os seguintes: 30/70, 40/60, 50/50, 60/40, 70/30 e 80/20. A etapa de classificação atingiu em média 83,94% de acurácia, 83,24% de sensibilidade, e 84,14% de especificidade, com taxa de 0,55 falsos positivos por imagem e de 0,17 falsos negativos por imagem.

O primeiro sistema CAD comercial aprovado nos Estados Unidos, o ImageChecker®, é provavelmente um dos mais utilizados no país (HOLOGIC, 2011). Em suas configuração

padrão o ImageChecker® resulta em 88% sensibilidade, podendo chegar 90% mudando sua configuração, consequentemente aumentando levemente a taxa de falsos positivos por imagem (IMAGECHECKER., 2011). É importante observar que as taxas obtidas pelo sistema, mesmo não superando o estado da arte, o ImageChecker® é o sistema líder de mercado nos Estados Unidos.

Na literatura existem diversas metodologias que a segmentação é realizada somente sobre uma imagem. Porem, uma nova classe de estudos tem considerado o uso das incidências MLO (médio lateral oblíquo) e CC (crânio caudal) conjuntamente (intitulado ipsilateral) para detectar massas pelas diferenças entre as mesmas (ENGELAND S. VAN, 2007) (QIAN et al., 2007) (WEI et al., 2011). Um princípio similar é aplicado a metodologias que usam a visão bilateral (WU et al., 2007) (KE, 2010) (TZIKOPOULOS, 2011) (WANG et al., 2012).

BAJGER et al. (2010), apresenta um método automático para detecção de massas em mamografias para segmentação de ROIs por fusão estatística e Análise Linear Discriminante (LDA). Esta metodologia utiliza 36 imagens selecionadas a partir de um banco de imagens mamográficas proprietário e 48 imagens retiradas da base Digital Database for Screening Mammography (DDSM). Para a classificação das ROIs, o valor da área sob a curva Receiver Operating Characteristic (ROC), foi $A_z = 0{,}90$ para as imagens proprietárias e $A_z = 0{,}96$ para as imagens da DDSM.

No trabalho de HU et al. (2011), os autores desenvolveram uma técnica para detecção de massas em imagens mamográficas, combinando limiarizações adaptativas de forma global e local na segmentação em multi-resolução. Os resultados encontrados no referido trabalho, apresentaram uma taxa de 91,3% de sensibilidade, com 0,71 falsos positivos por imagem.

O trabalho de LIU et al. (2011), apresenta um sistema para detecção automática de massas em mamografias digitais. Este sistema combina duas técnicas, a de múltiplas camadas Concêntricas e a Região de Faixa Estreita baseada em Contornos Ativos para segmentar as regiões suspeitas de conter lesões. Para a extração de características de textura das ROIs foi utilizado o Padrão Binário Local Completo (*Complete Local Binary Patterns - CLBP*), para serem classificadas pela SVM. O método foi avaliado com um conjunto de 231 imagens, contendo 245 massas. Dessas imagens, 125 contendo 133 massas foram utilizadas para treinar a SVM. As imagens restantes foram utilizadas para testar o desempenho. O resultado apresentou uma sensibilidade de 76,8% e uma taxa de 1,36 falsos positivos por imagem.

No trabalho proposto por SAMPAIO et al. (2011) utilizando redes neurais celulares para

segmentar as imagens mamográficas e gerar as ROIs, combina características de forma (excentricidade, circularidade, densidade circular, desproporção circular e densidade) e de textura (Função K de Ripley, índices de Moran e Geary) para descrever as ROIs. As características extraídas foram classificadas em massa e não massa através da SVM, apresentando como resultados uma sensibilidade de 80%, 0,84 falsos positivos por imagem e 0,2 falsos negativos por imagem.

AL MUTAZ et al. (2011a) propuseram na sua metodologia a detecção de massas em mamografias digitais através de estatísticas de segunda ordem. Para a extração de características de texturas das ROIs segmentadas, utilizou matrizes de Co-ocorrência dos Níveis de Cinza (Gray Level Co-ocorrency Matrix - GLCM), que são adquiridas de quatro orientações espaciais (0^o, 45^o, 90^o e 135^o), com a distância de dois pixels e três diferentes tamanhos de janelas (8x8, 16x16 e 32x32). Os resultados apresentados na classificação através de Redes Neurais Artificiais (RNA), mostraram que a GLCM em 0^o, 45^o, 90^o e 135^o com uma janela de tamanho 8x8 produziram os melhores resultados, atingindo uma sensibilidade de 91,67% e uma especificidade de 84,17%.

Em AL MUTAZ et al. (2011b), a metodologia de classificação automática de massas e não massas utiliza os descritores de Haralick derivados das GLCM dos níveis de cinza das ROIs para extrair as características de textura. As imagens sofreram uma quantização para 16 níveis de cinza. Este processo visa reduzir os dados e consequentemente o tempo de processamento. Logo após, para cada nível de cinza, foi atribuído um número de código baseado na frequência de cada nível de cinza, transformando a imagem original em uma imagem codificada. Em seguida, as características de textura de Haralick são extraídas. A metodologia apresenta uma acurácia de 95,85% com o classificador Analise de Discriminantes Lineares (LDA).

No trabalho de ERICEIRA et al. (2010), foi desenvolvido um sistema CAD para detecção de massas através de registro de pares de mamografias, utilizando como descritor espacial o variograma cruzado e SVM. Este trabalho apresentou como melhor resultado a sensibilidade de 100%, a especificidade de 95,34% e acurácia de 96%.

Em MOREIRA et al. (2013) é proposta uma metodologia para auxiliar o processo de detecção de massas em imagens da mama. Esta metodologia divide-se em quatro etapas: melhoramento da imagem; segmentação baseada em grafo; redução de falsos positivos e descrição e classificação. Para encontrar os limiares utilizados na etapa de redução de falsos positivos, foram utilizadas oito imagens que mais representam todas as imagens presentes na base para

extrair as ROIs referentes à região com nódulo e então foi utilizada a média da intensidade do desvio padrão encontrado nessas regiões como limiares. Os valores obtidos para estes limiares foram 180 para a média e 37,8 para o desvio padrão. O limiar utilizado para o índice de circularidade foi 0,45, sendo estimado empiricamente. Após a fase de redução de falsos positivos, a metodologia apresentou 85% de sensibilidade e uma taxa média de falsos positivos de 6,67%.

A proposta de (BRAZ, 2014), remete a detecção de regiões de massas em mamografias digitalizadas com a metodologia que envolve aspectos relacionados a necessidade de encontrar regiões suspeitas e descrevê-las de maneira discriminatória. Esse trabalho busca avaliar a extração de características usando as abordagens de análise de diversidade, geoestatística e geométrica de maneira a obter uma classificação de regiões suspeitas usando a SVM como classificador. Os resultados encontrados nesse trabalho, são promissores e obteve-se alta sensibilidade e baixa taxa média de falsos positivos ao usar geometria côncava para extrair características. O melhor resultado para a base MIAS obteve sensibilidade equivalente a 97,30% com 0,3333 falso positivos médio por imagem e área Free-Response Receiver Operating Characteristic (FROC) equivalente a 0,89%. O melhor resultado usando a base DDSM obteve sensibilidade de 91,63% com 0,013 falso positivos médio por imagem, com AFROC equivalente a 0,86%.

O método de DONG et al. (2015) para classificação de ROIs, adquiridas da base DDSM, utiliza código em cadeia para indicar as ROIs. Suas estruturas internas são realçadas através de *Rough Set* (conjunto áspero). A convolução de Campos Vetoriais são utilizados para extrair 32 características das ROIs. Estas características são utilizadas na etapa de treinamento/classificação onde o desempenho dos classificadores *Random Forest*, SVM, SVM genética, PSO, PSO-MVS e árvores de decisão são comparados. O melhor desempenho do método foi utilizando o classificador Random Forest, apresentando uma acurácia de 93,24%, sensibilidade de 94,78% e especificidade de 91,76%. A curva ROC apresentou o valor de 0,95.

O trabalho de SAMPAIO et al. (2015), apresenta uma metodologia computacional para auxiliar especialistas na descoberta de massas mamárias, com base na densidade da mama. O primeiro passo da metodologia é a detecção do tipo de densidade da mama (densa ou não densa). De posse desta informação, é possível optar por uma sequência de processos e configurações para realizar a detecção de massas em mamografias de acordo com a sua densidade. Em seguida, uma etapa de melhoramento é aplicada sobre a imagem, para a remoção de objetos externos à mama e redução de ruídos. Na etapa de segmentação, encontra-se as regiões da imagem que provavelmente contenham massas, esta etapa utilizou um micro Algoritmo Genético para criar

uma máscara de proximidade de textura e selecionar regiões suspeitas de conter lesão. Para reduzir o número de regiões suspeitas, utilizou-se duas etapas de redução de falsos positivos. A primeira redução de falsos positivos utiliza o *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* e um ranking de proximidade de textura extraídos das regiões de interesse da mama. Na segunda redução de falsos positivos as regiões resultantes têm as suas texturas e formas analisadas pela combinação de árvores Filogenéticas e descritores geométricos, Padrões Binários Locais e SVM. Um Micro Algoritmo Genético foi utilizado para escolher as regiões suspeitas que geram os melhores modelos de treinamento e maximizam a classificação de massas e não-massas usados na SVM. Os melhores resultados obtidos produziram uma sensibilidade de 94,02%, especificidade de 82,28% e acurácia de 84,08%, com uma taxa de 0,85 falsos positivos por imagem com uma área sob a Free-Response Receiver Operating Characteristic (curva FROC) de 1,13 nas análises de mamas não densas. Para mamas densas obteve-se uma sensibilidade de 89,13%, especificidade de 88,61% e acurácia de 88,69%, com uma taxa de 0,71 falsos positivos por imagem com uma área sob a curva FROC de 1,47.

Percebe-se a importância das técnicas de segmentação e extração de características dos trabalhos apresentados, especialmente na análise de forma e textura, assim como a utilização da classificação por Máquina de Suporte de Vetores (SVM) como classificador, por apresentar bons resultados. Um fato negativo encontrado nos referidos trabalhos é o simples fato de não utilizarem um processo de otimização durante a etapa de segmentação. Neste trabalho, é proposto uma metodologia de detecção de massas utilizando o PSO para otimizar as regiões durante o processo de segmentação e os índices de diversidade funcional para a etapa de classificação. A Tabela 2.1 resume a comparação dos trabalhos relacionados.

Tabela 2.1: Comparação dos resultados dos trabalhos relacionados. As métricas de desempenho são medidas por: Sensibilidade (Sen.); Especificidade (Esp.); Acurácia (Acu.); área sobe a curva ROC (ROC); falsos positivos por imagem (FP/i), área sobe a curva FROC, e tamanho da amostra (Amostra).

Trabalho	Base	Sen.	Esp.	Acu.	ROC	FP/i	FROC	Amostra
NUNES (2009)	DDSM	83,24	84,14	83,94	—	0,55	—	650
IMAGECHECKER. (2011)	Privada	88,00	—	—	—	—	—	—
(BAJGER et al., 2010)	Privada	—	—	—	0,90	—	—	36
	DDSM	—	—	—	0,96	—	—	48
HU et al. (2011)	MIAS	91,30	—	—	—	0,71	—	170
LIU et al. (2011)	DDSM	76,80	—	—	—	1,36	—	125
SAMPAIO et al. (2011)	DDSM	80,00	—	—	—	0,84	—	623
AL MUTAZ et al. (2011a)	DDSM	91,67	84,17	—	—	—	—	120
AL MUTAZ et al. (2011b)	DDSM	—	—	95,85	—	—	—	60
ERICEIRA et al. (2010)	DDSM	100,00	95,34	96,0	—	—	—	620
MOREIRA et al. (2013)	MIAS	85,00	—	—	—	6,67	—	74
BRAZ (2014)	MIAS	97,30	—	—	—	0,33	—0,89	74
	DDSM	91,63	—	—	—	0,01	0,86	621
DONG et al. (2015)	DDSM	94,78	91,76	93,24	0,95	—	—	200
MARTINS et al. (2015)	DDSM	98,60	98,85	98,88	—	—	—	600
SAMPAIO et al. (2015) Não Densas	DDSM	94,02	82,28	84,08	—	0,85	1,13	388
SAMPAIO et al. (2015) Densas	DDSM	89,13	88,61	88,69	—	0,71	1,47	233

Capítulo 3

Fundamentação Teórica

Nesta seção será apresentada a base teórica que fundamenta este trabalho, onde será dado um enfoque geral de modo a familiarizar o leitor com a teoria que será utilizada na solução do problema proposto.

3.1 Câncer de Mama

O câncer de mama se inicia quando há um erro no processo de divisão celular, criando-se múltiplas células alteradas de forma desordenada. Estas células alteradas alastram-se pelos tecidos adjacentes, gerando aglomerados de células modificadas (tumoral), obstruindo veias e vasos linfáticos, podendo se alastrar pela corrente sanguínea em outros órgãos ocasionando a metástases. A Figura 3.1 ilustra o processo de metástase.

3.1.1 Tipos de Anomalias

Os tipos de anomalias detectadas em uma imagem mamográfica, podem ser observados na Figura 3.2, e os achados referente a esta imagem, podem ser descritos como:

- Massa: Qualquer opacidade com algum contorno arredondado e definido segundo a forma e a densidade;
- Microcalcificações: pequenos depósitos de cálcio classificados de acordo com sua morfologia e distribuição; e
- Distorção de arquitetura: espiculações em uma região da mama ou uma retração focal do contorno parenquimatoso denso.

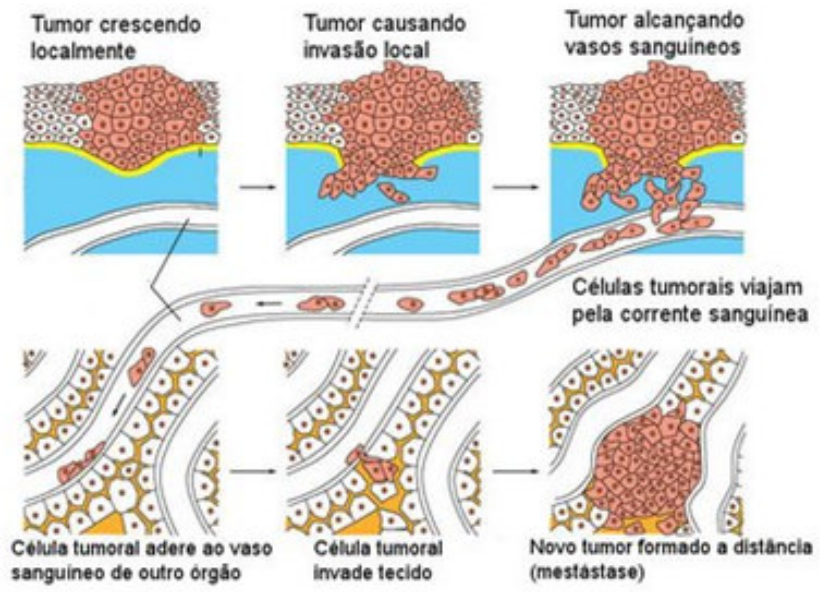

Figura 3.1: Ilustração do processo de metástases das células cancerígenas. (SAÚDE-MEDICINA, 2015).

Durante a realização desta pesquisa, investigaremos apenas as massa como escopo do trabalho, sendo desconsideradas as microcalcificações e as distorções arquiteturais.

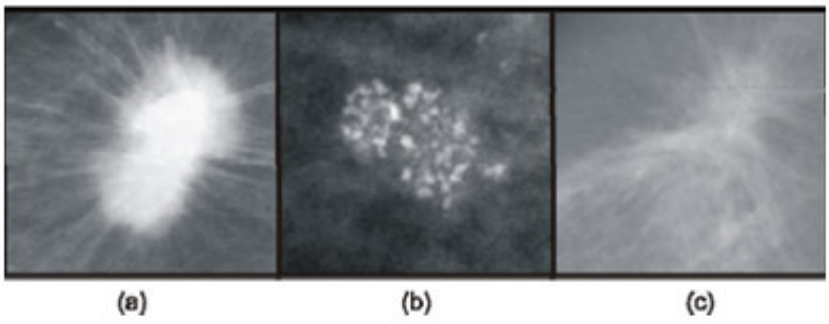

Figura 3.2: Ilustração dos tipos de anomalias em uma imagem mamográfica: (a) Massa espiculada; (b) Microcalcificações; (c) Distorção de arquitetura. Fonte: (HEATH et al., 2001)

Dentre as diversas neoplasias existentes, o Colégio Americano de Radiologia em conjunto com outros órgãos criou o BI-RADS (ACR., 2003) uniformizando o laudo médico, padronizando os termos utilizados, estabelecendo categorias de avaliação fina e sugerindo condutas a cada uma delas.

A Tabela 3.1 criada pela BI-RADS, nos mostra as seis categorias dos graus de malignidade dos achados clínico.

Tabela 3.1: Categorização BI-RADS dos achados suspeitos

Categoria	Interpretação	VPP[2]	Conduta
0	Inconclusivo		Exame adicional
1	Exame normal	0%	Controle anual a partir dos 40 anos
2	Achados benignos, como calcificações, linfonodos intramamários, cistos, etc	0%	Controle anual a partir dos 40 anos
3	Provavelmente benigno	$< 2\%$	Controle semestral
4 (A,B e C)[1]	Suspeito	$> 2\%$ e $<90\%$	Biópsia
5	Provavelmente Maligno	$> 95\%$	Biópsia
6	Lesão maligna – biopsada ou diagnosticada não submetida a terapia intensiva	100%	

[1](A) Menor, (B) Médio e (C) Maior [2]VPP: Valor Preditivo Positivo

3.1.2 Mamografia

A mamografia é um exame (procedimento) que resulta em uma imagem da mama, capaz de detectar o câncer na mama de forma precoce, sendo o um exame capaz de reduzir a mortalidade através do rastreamento de mulheres assintomáticas (KOPANS, 2007).

O objetivo da mamografia é gerar 4 imagens de alta resolução em tons de cinza das estruturas internas da mama, com o intuito de detectar anomalias. As principais anomalias visualizadas a partir de uma mamografia, em geral, são calcificações e massas, descritas anteriormente na Seção 3.1.1. A Figura 3.3 mostra uma imagem gerada por uma mamografia.

Figura 3.3: Ilustração de uma imagem gerada por uma mamografia com marcações. Fonte: (SAMPAIO et al., 2011).

A mamografia deve ser realizada por um aparelho de raio x específico, conhecido como mamógrafo. Este procedimento visa radiografar as duas mamas, por isso na mamografia deve ser

realizada duas projeções para cada mama, sendo essas, a médio-lateral obliqua (MLO) e a crânio caudal (CC). A primeira projeção permite visualizar do alto da axila para baixo, incluindo a prega infla-mamária e o músculo peitoral, estendendo-se obliquamente sobre a mama. A segunda projeção permite visualizar a região póstero-medial da mama, complementando a visão da projeção médio-lateral obliqua. A Figura 3.4 demonstra a realização de um exame de mamografia.

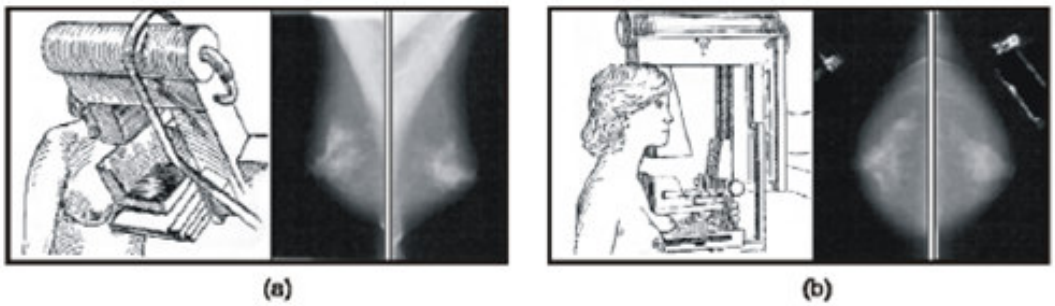

Figura 3.4: Realização e resultado das projeções de uma mamografia: a) Projeção médio lateral obliqua (MLO); b) Projeção crânio caudal (CC). Fonte: (SAMPAIO et al., 2011).

3.2 Processamento de Imagens

O processamento de imagens, consiste em um conjunto de técnicas para capturar, representar e transformar imagens com auxílio do computador (PEDRINE e SCHWARTZ, 2008). Didaticamente é dividido em etapas, conforme apresenta GONZALES e WOODS (2010).

Na aquisição utiliza-se algum mecanismo para capturar ou gerar as imagens que se deseja processar. Estas imagens podem ser obtidas através de equipamentos de captura como câmeras e radares ou através de simulações por computador.

O melhoramento tem a finalidade de aumentar a qualidade da imagem, proporcionando uma melhora através da redução de ruídos, melhoramento do contraste e recorte do objeto de interesse da imagem de fundo, dentre outros. Esta etapa na maioria dos casos é indispensável à eficiência das etapas posteriores.

A segmentação é extremamente importante, pois através dela podemos isolar os objetos de interesse, permitindo assim trabalharmos com estes elementos individualmente, proporcionando resultados mais satisfatórios na determinação destes objetos, dando prosseguimento à etapa de representação na análise da imagem.

A representação tem por finalidade extrair da região segmentada um conjunto descritivo de

características mensuráveis, que representem de forma significativa os objetos. Estas características variam muito de acordo com o que se pretende analisar, mas podem incluir perímetro, cor dos *pixels*, geometria, dentre outros.

Por fim, na classificação, as características obtidas de cada imagem são analisadas para se chegar a alguma conclusão sobre a imagem. Nesta etapa podemos inferir conceitos a respeito dos objetos analisados e agrupá-los em categorias.

3.2.1 Pré-processamento

O pré-processamento, visa melhorar a qualidade da imagem aplicando técnicas para atenuação de ruídos, correção de contrastes ou brilho e suavização de determinadas propriedades da imagem (PEDRINE e SCHWARTZ, 2008).

3.2.1.1 Filtro de Média

O filtro de média é uma técnica muito usada no processamento de imagens para reduzir o ruído, melhorando sua qualidade (GONZALES e WOODS, 2010).

Dada uma imagem I e uma janela J de tamanho NxN, centralizada no pixel p_{ij}, o resultado do filtro de média sobre p_{ij} é a média aritmética dos valores dos *pixels* contidos em J. Quanto maior for o valor de N, mais influência o p_{ij} transformado sofrerá, e isto pode resultar no efeito de borramento da imagem.

Este filtro de média foi utilizado na metodologia proposta com o objetivo de reduzir os ruídos existentes nas mamografias digitais.

3.2.1.2 Contrast-Limited Adaptive Histogram Equalization

O realce local é uma maneira interessante de contrastar os *pixels* de uma imagem. O *Contrast Limited Adaptive Histogram Equalization (CLAHE)* é uma técnica de realce local de contraste que altera o tom de cinza de um pixel através da análise de sua vizinhança. O CLAHE evita o aumento de contraste em ruídos, baseando-se numa equalização de histograma adaptativa. Em resumo, cada pixel é transformado com base no histograma de um quadrado ao seu redor, onde a função de transformação é obtida através da função de Distribuição Cumulativa (FDC), encontrada em PEDRINE e SCHWARTZ (2008) de valores de pixels na vizinhança (GONZALES e WOODS, 2010).

O CLAHE limita a amplificação por recorte do histograma em um valor pré-definido antes de calcular o FDC. Isto limita a inclinação do mesmo e, por conseguinte, a função de transformação. Este valor de corte do histograma depende do tamanho da vizinhança. Os valores de corte estão entre 3 e 4 vezes o valor médio do histograma (ZUIDERVELD, 1994). É mais interessante não descartar a parte do histograma que ultrapassa o limite de corte, então pega-se a parte que excedeu e distribui-se igualmente em todas as faixas do histograma, como mostra a Figura 3.5.

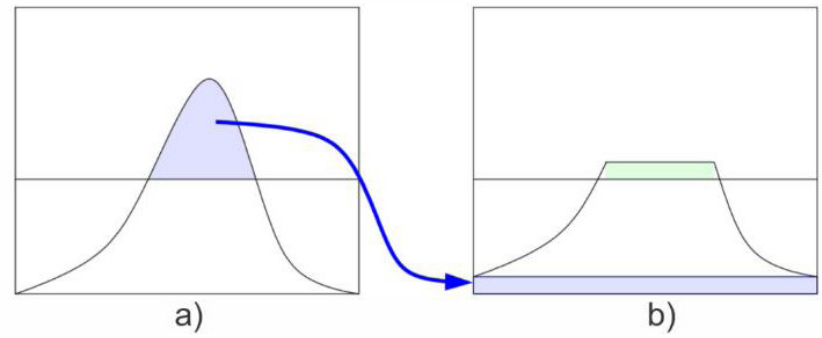

Figura 3.5: Ilustração da redistribuição do histograma de CLAHE. a) Ponto de corte, b) Redistribuição dos valores de corte do histograma. Fonte: (ZUIDERVELD, 1994).

O CLAHE, possui diversas funções de transformação dos tons de cinza (ZUIDERVELD, 1994). Para este trabalho foi utilizada a função uniforme, de acordo com a Equação 3.1.

$$g = [g_{max} - g_{min}]P(f) + g_{min}, \tag{3.1}$$

onde g é o novo valor de cinza do pixel. Os valores g_{min} e g_{max} são as variáveis que tem como valor o menor e maior valor de cinza na vizinhança, respectivamente; e $P(f)$ é a função FDC.

Esta técnica de realce local (CLAHE), foi utilizada para realçar as estruturas internas das mamas nas mamografias digitais.

3.2.2 Crescimento de Região

A técnica de crescimento de regiões consiste em agregar conjuntos de pixels vizinhos em regiões maiores, obedecendo algum critério.

Este processamento parte de um ponto inicial, chamado de semente, o qual pode ser tanto um único pixel como um conjunto de pixels. A partir desta semente agrega-se os pixels vizinhos que possuam algum atributo de similaridade (tons de cinza) aos do grupo da semente. Esse processo continua até que se atinja uma condição de parada pré-estabelecida, como, por

exemplo, um determinado nível de cinza, uma distância específica, ou quando não existem mais vizinhos a serem agregados no grupo (PAL e PAL, 1993).

Para este trabalho, o algoritmo de crescimento de região foi utilizado na etapa de segmentação, para gerar as Regiões de Interesse (ROI) das imagem, de modo a isolar as ROIs candidatas.

3.2.3 Segmentação

A segmentação é um processo de separação da imagem, sendo capaz de identificar corretamente a localização, a topologia e a forma dos objetos de forma que as informações resultantes de um sistema de análise de imagens sejam confiáveis (PEDRINE e SCHWARTZ, 2008). Neste contexto, entendemos a segmentação como uma operação que distinguir os objetos contidos na imagem.

Também pode-se afirmar que a segmentação é um processo de particionamento de imagens em regiões desconexas de maneira a buscar detectar descontinuidade ou similaridade na imagem. Podemos dizer assim, mais homogêneas possíveis entre os elementos contidos nesta e mais heterogênea possível entre as demais regiões da imagem (PEDRINE e SCHWARTZ, 2008).

Devido à grande quantidade de estruturas diferentes (músculo peitoral, rótulos identificadores, vasos sanguíneos, mamilos, nódulos, micro calcificações) que uma imagem de mamografia pode ter, a segmentação torna-se um processo indispensável para o sucesso deste trabalho, pois é através da dela que podemos montar os grupos de maior similaridade entre si.

3.2.4 Algoritmo de Otsu

O algoritmo de OTSU (1975) é uma técnica que busca encontrar um limiar ótimo baseado no histograma da imagem, dividindo a imagem em duas partes, cada uma com a maior similaridade entre si.

Partindo de uma imagem em tons de cinza com N pixels e L possíveis níveis de cinza, a probabilidade de ocorrência dos níveis de cinza i na imagem é mostrada na Equação 3.2.

$$p_i = \frac{f_i}{N} \tag{3.2}$$

onde, *fi* representa a frequência de repetição dos níveis de cinza de i, com i = *1,2,...,L*.

Os *pixels* são divididos em duas classes C_1 e C_2, com níveis de cinza *[1,2,...,t]* e *[t+1,t+2,...,L]*

respectivamente, onde as distribuições de probabilidade são mostrados nas Equações 3.3 e 3.4.

$$C_1 = \frac{p_1}{\omega_1(t)},...,\frac{p_t}{\omega_1(t)} \tag{3.3}$$

$$C_2 = \frac{p_{t+1}}{\omega_2(t)},\frac{p_{t+2}}{\omega_2(t)},...,\frac{p_L}{\omega_2(t)} \tag{3.4}$$

onde as intensidades para cada uma das classes $\omega_1(t)$ e $\omega_2(t)$ são definidas como:

$$\omega_1(t) = \sum_{i=1}^{t} p_i \tag{3.5}$$

$$\omega_2(t) = \sum_{i=t+1}^{L} p_i \tag{3.6}$$

As médias para cada uma das classes $\mu_1(t)$ e $\mu_2(t)$ são definidas com as Equações 3.7 e 3.8 a seguir:

$$\mu_1(t) = \sum_{i=1}^{t} \frac{i.p_i}{\omega_1(t)} \tag{3.7}$$

$$\mu_2(t) = \sum_{i=t+1}^{L} \frac{i.p_i}{\omega_2(t)} \tag{3.8}$$

O valor da média total da imagem μ_T é defina com as Equações 3.9 e 3.10.

$$\omega_1.\mu_1 + \omega_2.\mu_2 = \mu_T \tag{3.9}$$

$$\omega_1 + \omega_2 = 1 \tag{3.10}$$

Por fim, o algoritmo de *Otsu* definirá a variância entre as classes por meio da Equação 3.11 e usará com tom t, que representa entre todos os testados a maior variância, onde $1 \leq t \leq L$.

$$\delta_B^2 = \omega_1.(\mu_1 - \mu_T)^2 + \omega_2.(\mu_2 - \mu_T)^2 \tag{3.11}$$

$$t^* = Max\{\delta_B^2(t)\} \tag{3.12}$$

onde, t^* é o limiar ótimo encontrado pela Equação 3.12 que será utilizado para dar início aos centroides da primeira partícula do enxame do *Particle Warm Optimization* (*PSO*).

3.2.5 Otimização por Enxame de Partículas

O *PSO* é um algoritmo baseada no comportamento social de um bando de pássaros em movimento (MERWE e ENGELBRECHT, 2003), sendo estendido do modelo de (REYNOLDS e SNAPP, 1986).

Merwe e Engelbrecht (MERWE e ENGELBRECHT, 2003), dizem que dado um problema, o *PSO* mantém uma população de partículas onde cada partícula representa uma solução potencial para o problema e está associada a uma posição em um espaço de busca multidimensional, mantendo as seguintes informações:

- x_i - Posição atual da partícula.
- v_i - Velocidade atual da partícula.
- y - Melhor posição local da partícula.
- $\hat{y}$ - Melhor posição global da partícula.
- w - Valor de inércia da partícula.

Para cada iteração a velocidade da partícula é alterada conforme a Equação 3.13.

$$v_{i,k}(t+1) = wv_{i,k}(t) + c_1 r_{1,k}(t)(y_{i,k}(t) - x_{i,k}(t)) + c_2 r_{2,k}(t)(\ddot{y}_k(t) - x_{i,k}(t)), \tag{3.13}$$

onde $v_{i,k}$ representa a $k^{-ésima}$ dimensão do vetor velocidade da $i^{-ésima}$ partícula. Cada velocidade é atualizada separadamente para cada dimensão $k \epsilon i...n$. Os valores r_1 e r_2 são gerados aleatoriamente entre 0 e 1.

As constantes c_1 e c_2 regulam a aceleração, onde $0 < c_1, c_2 \leq 1$, c_1 regula a direção da melhor posição local e c_2 regula a direção da posição global *(gbest)* ou da vizinhança *(lbest)* (VAN DEN BERGH, 2006). Os termos $c_1 r_{1,k}(t)(y_{i,k}(t) - x_{i,k}(t))$ e $c_2 r_{2,k}(t)(\hat{y}_k(t) - x_{i,k}(t))$ representam as experiências passadas da partícula, sendo o primeiro associado à cognição e o segundo ao social, pois cada partícula leva em consideração a melhor solução ao seu redor (KENNEDY, 2010).

A partícula terá sua posição atualizada usando o novo vetor velocidade calculado na Equação 3.13 e usado na Equação 3.14.

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{3.14}$$

O *PSO* consiste em repetidas iterações das equações anteriores (WEI et al., 2007) e toda vez que uma partícula for melhorada através do cálculo de *fitness* (Seção 4.3.3), a mesma é atualizada, como mostra a Equação 4.10. Nesta pesquisa o PSO será utilizado para encontrar os melhores limiares, gerando assim, os grupos de regiões mais homogêneas.

3.2.6 Agrupando Grafo (Graph Clustering)

Algoritmos de agrupamento podem ser divididos em dois grupos principais: algoritmos hierárquicos e particionados. Estes algoritmos consistem em dividir o conjunto de dados em $k \leq n$ grupos, onde n é o número de elementos no conjunto de dados. Soluções são obtidas através de objetos em movimento entre os *clusters* até que um critério de parada seja satisfeito. Cada solução é avaliada por uma dada função objectivo (SCHAEFFER, 2007).

O algoritmo *Graph Clustering* é um grafo de agrupamento de vértices em grupos, com base na estrutura da extremidade do grafo. A partição do vértice resultante deve ter a propriedade de que dentro de cada *cluster* os vértices sejam altamente conectados, sendo que só existem poucas arestas entre *clusters*.

Esta abordagem adota explicitamente os conceitos da teoria dos grafos e pode fornecer as definições necessárias e o formalismo matemático, resultando em um suporte importante para a análise de modelos de grafos (SCHAEFFER, 2007). A Figura 3.6 mostra este modelo.

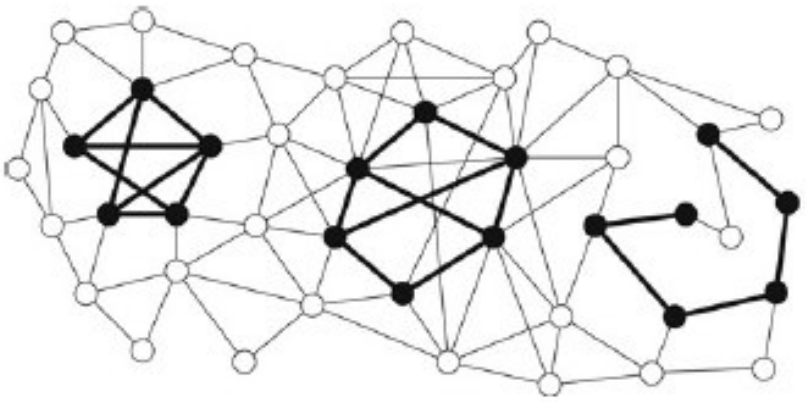

Figura 3.6: Figura ilustrativa do *Graph Clustering* com 3 subgrafos, representando cada um os *clusters* (SCHAEFFER, 2007).

No particionamento do grafo, os problemas são definidos como sendo a dificuldade de encontrar bons agrupamentos, nas configurações de ambos os grafos simples e ponderados respectivamente. Restringindo a atenção primeiro para a bi-partição de um gráfico, denotamos (S, S^c) uma partição do conjunto de vértices V de um grafo *(V, E)*. Deixe $\sum(S, S^c)$ ser a soma total dos pesos de arestas entre S e S^c, que é o custo associado com (S, S^c). O problema máximo do corte é encontrar uma partição (S, S^c) que maximiza $\sum(S, S^c)$, e o problema do corte do quociente mínimo é encontrar uma partição que minimiza $\sum(S, S^c)$ / min($\|S\|$, $\|S^c\|$).

Em problemas de partição de grafos padrão, os tamanhos dos elementos da partição são prescritos, para se minimizar o peso total das arestas que ligam os nós em subconjuntos distintos na partição. Assim se tamanhos $m_i > 0$, i = *1,..., k*, satisfazendo k $\leq n$ e $\sum m_i = n$, forem especificados, em que n é o tamanho de V, e uma partição de V em subconjuntos S_i de tamanho m_i é chamado de função que minimiza o custo. O papel de uma função de custo/rendimento é principalmente útil para a comparação de diferentes estratégias em relação a referências comuns (DONGEN e STIJN, 2001).

O *Graph clustering* será utilizado neste trabalho para unir as ROIs vizinhas, reduzindo falsos positivos, com o intuito de melhorar as estruturas das regiões candidatas a massa.

3.2.7 Descritores de Forma

Os descritores baseados em região, fazem uso dos *pixels* localizados no interior da região ou do objeto e não somente dos *pixels* que formam a bordas. (PEDRINE e SCHWARTZ, 2008). Estes descritores são utilizados quando se deseja conhecer o formato da região ou dos objetos, como por exemplo: a área; a compacidade; a circularidade; a excentricidade, dentre outros.

A circularidade, definida de acordo com a Equação 3.15, descreve o quão um objeto é circular ou não, e foi utilizada nesta pesquisa para remover regiões indesejadas.

$$C = \frac{P^2}{A}, \tag{3.15}$$

onde, P e A são o perímetro e a área da região, respectivamente, considerando os *pixels* como unidade de medida.

Nesta proposta de trabalho, adaptou-se a circularidade para definir um valor aceitável ás regiões, separando-as em candidatas e não candidatas. Esse processo é detalhado no Capítulo 4.

3.2.8 Descritores de Textura

Descritores de textura é uma tarefa complexa e encontra-se entre as características empregadas pelo sistema visual humano, contendo informações sobre a distribuição espacial e a variação de luminosidade. A textura também descreve o arranjo estrutural das superfícies e relações entre regiões vizinhas (PEDRINE e SCHWARTZ, 2008).

Na literatura é encontrado diversas definições de análises de texturas. O trabalho de TUCERYAN e JAIN (1998) apresenta uma revisão bibliográfica sobre métodos existentes.

Para a realização deste trabalho, descritores de texturas foram utilizados na extração de características das ROIs candidatas e posteriormente no processo de classificação.

3.2.8.1 Medidas de Diversidade Funcional

As ciências biológicas há muito tempo vem estudando a diversidade filogenética, e o interesse dos pesquisadores por este tema está crescendo muito nos últimos anos, em diversos campos da Ecologia e em estudos com diversos grupos taxonômicos, sugerindo que o conceito está ganhando grande importância. Em virtude da potencial relação entre a diversidade funcional e o funcionamento e manutenção dos processos das comunidades (PETCHEY e GASTON, 2006), é importante definir de maneira precisa o conceito de diversidade funcional.

TILMAN (2001) define diversidade funcional como sendo o valor e a variação das espécies e de suas características que influenciam o funcionamento das comunidades. Para facilitar melhor o entendimento, a Figura 3.7 mostra a nomenclatura da árvore filogenética de uma comunidade de plantas e suas diversidades. Essa definição é a que adotamos neste trabalho, juntamente com as medidas de diversidade funcional descritas nas seções seguintes.

3.2.8.2 Diversidade Funcional

Dessa maneira, medir a Diversidade Funcional (FD) implica em medir a diversidade de características funcionais dos indivíduos, que são componentes dos fenótipos dos organismos que influenciam os processos dentro da comunidade. Isto é, imaginemos duas comunidades (X e Y) com a mesma quantidade de espécies. Se todas as espécies em X forem dispersas por aves, enquanto que as em Y forem dispersas por mamíferos, aves, lagartos e pelo vento, apesar de ambas possuírem o mesmo número de espécies, Y será mais diversa por apresentar espécies funcionalmente diferentes no que se refere ao tipo de dispersão (CIANCIARUSO et al., 2009). Existe uma vasta literatura a respeito das características funcionais para as plantas, assim

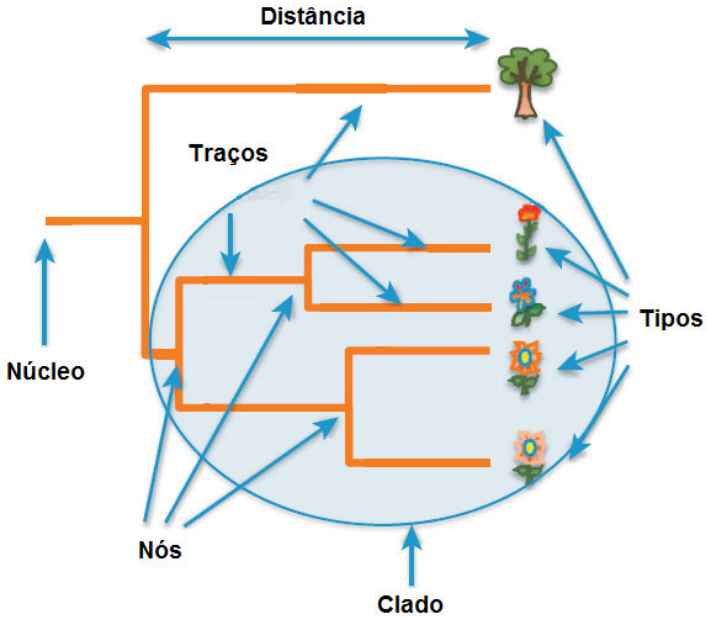

Figura 3.7: Ilustração da teoria da árvore filogenética de uma comunidade de plantas e suas diversidades. Adaptada de (WINTER et al., 2013).

como linhas de pesquisa dedicadas a testar o poder preditivo dessas características em relação a respostas ou efeitos no funcionamento das comunidades e a processos biológicos de difícil mensuração. Mais exemplos podem ser encontrado em CORNELISSEN et al. (2003) e VIOLLE et al. (2007).

Medir a FD, consiste na soma dos comprimentos dos braços de um dendrograma funcional, ou seja, um dendrograma gerado a partir de uma matriz de "espécies × características funcionais"(CIANCIARUSO et al., 2009). O cálculo da FD é o mais simples e baseia-se em fundamentos da análise de agrupamento e segue os seguintes passos:

1º Obter uma matriz funcional (espécies × características funcionais);

2º Converter a matriz funcional em uma matriz de distância;

3º Realizar o agrupamento da matriz de distância para produzir um dendrograma; e

4º Calcular o comprimento total das ramificações do dendrograma

Mensurar a diversidade funcional é estimar as diferenças entre os organismos diretamente a partir de suas características funcionais relacionadas com as hipóteses em estudo. Medir a diversidade funcional significa medir a diversidade de traços funcionais que influenciam os processos da comunidade, independentemente da filogenia dos organismos (CIANCIARUSO et al.,

2009). Essas medidas diferem na informação que contém e na maneira com que quantificam a diversidade (RICOTTA (2005), PETCHEY e GASTON (2006)). A Figura 3.8 mostra como é encontrada a FD a partir de um dendrograma.

A Figura 3.9 mostra um exemplo de uma imagem 3x3 com seus valores de *pixels* e o seu respectivo dendrograma, criado a partir dessa imagem. Nas folhas do referido dendrograma estão as espécies, representadas pelos valores de *pixels*. Os grupos são representados pelas espécies que pertencem ao mesmo nó. Essa técnica é utilizada para montar os dendrogramas das imagens mamográficas, com o objetivo de extrair os índices de diversidade funcional.

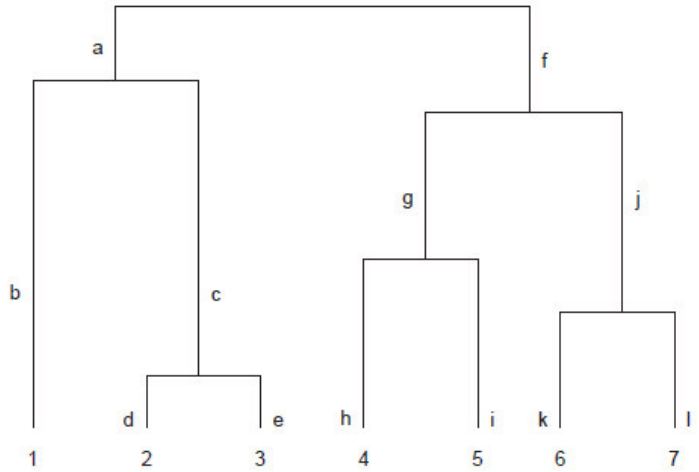

Figura 3.8: Dendrograma funcional hipotético. A FD é igual à soma de todos os braços necessários para conectar as espécies presentes em uma dada comunidade. Por exemplo, uma comunidade formada pelas espécies 1, 2, 5 e 6 terá uma FD igual a b + a + c + d + f + g + i + j + k. Fonte: (CIANCIARUSO et al., 2009)

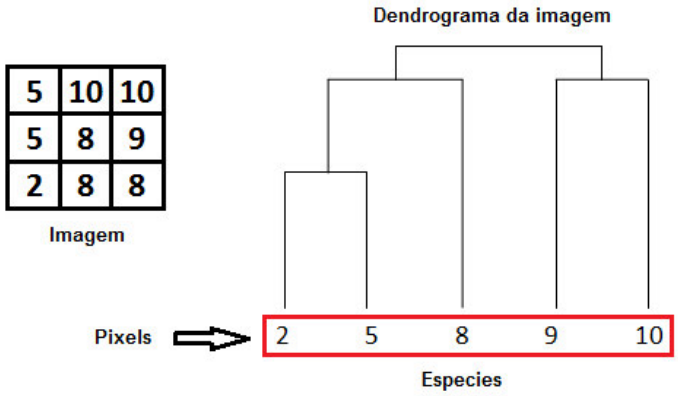

Figura 3.9: Exemplo de um Dendrograma montado a partir de uma imagem

O valor da FD calculada a partir da Figura 3.9 é igual a 8, pois basta somar todos os traços pertencentes ao dendrograma. Neste trabalho, os Nós são gerados a partir do algoritmo de Otsu, formando os grupos de uma dada comunidade.

Para melhor entendimento da FD, apresentamos na Figura 3.10 os termos da bilogia em relação aos da metodologia proposta.

Biologia	Metodologia
Comunidade	Região de interesse (Candidatos) da Mamografia
Espécies	Pixel da Região de interesse
Indivíduos	Valor do pixel
Abundância: Números de indivíduos de uma referida espécie	Número de Pixels de mesmo valor de uma Região de interesse

Figura 3.10: Comparação dos termos da Biologia e da metodologia proposta.

3.2.8.3 Diversidade Funcional Abundante

A diversidade funcional abundante (FAD) estima a dispersão pela soma das distâncias pareadas entre as espécies no espaço multidimensional, enquanto que uma medida semelhante o faz pela média dessas distâncias (HEEMSBERGEN et al., 2004). Outra medida proposta recentemente baseia-se na entropia quadrática de RAO (1982) e é semelhante às anteriores, mas permite a inclusão da abundância das espécies, como mostra a Equação 3.16.

$$FAD = \sum_{i=1}^{S-1} \sum_{j=i+j} d_{ij} p_i p_j, \tag{3.16}$$

$$d_{ij} = \frac{1}{n} \sum_{k=1}^{n} (X_{ik} - X_{jk})^2, \tag{3.17}$$

onde, d_{ij} é a distância funcional entre pares de espécies no dendrograma. Essa distância d_{ij}, pode ser calculada de várias maneiras, neste trabalho, optamos pela distância euclidiana, conforme Equação 3.17. BOTTA-DUKÁT (2005) propôs um índice de diversidade funcional baseado na entropia quadrática de RAO (1982), que incorpora tanto as abundâncias relativas das espécies e uma medida das diferenças funcionais emparelhadas entre espécies. Considere uma comunidade de S-espécies, caracterizado por abundância relativa o vetor $p = (p_1, p_2, ..., p_s)$ tal que, a soma total de p é igual a 1, como mostra a Equação 3.18:

$$\sum_{i=1}^{S} p_i = 1 \tag{3.18}$$

Na metodologia proposta, os índices de FD foram utilizados na extração de características para classificar as ROIs. Essas medidas têm a vantagem de serem matematicamente simples e bastante utilizadas (RAO (1982), BOTTA-DUKÁT (2005), RICOTTA (2005)).

3.2.9 Técnica de Sobreamostragem Minoritária Sintética

O Técnica de Sobreamostragem Minoritária Sintética (Synthetic Minority Oversampling Technique (SMOTE)) foi proposto por CHAWLA et al. (2002), e é um método baseado em sobreamostragem informativa e cria novos exemplos da classe minoritária por meio da interpolação de exemplos da classe minoritária que se encontram próximos.

Especificamente, para um subconjunto $S^{'}$ pertencente a S, considere o $k-vizinho$ mais próximo para cada exemplo x_i pertencente a $S^{'}$, para algum número inteiro especificado k; o $k-vizinho$ mais próximo é definido como elemento de $S^{'}$ se a distância euclidiana entre ele e x_i em consideração exibir uma menor magnitude ao longo do espaço n-dimensional x. Para gerar uma amostra sintética, escolhe-se aleatoriamente um dos seus $k-vizinhos$ mais próximos, em seguida, multiplica-se a diferença entre o valor correspondente por um número aleatório entre [0,1], e, por fim, adiciona este novo valor ao subconjunto $S^{'}$, conforme Equação 3.19.

$$x = x_i + (y_i - x_i) * \delta, \tag{3.19}$$

onde, $x_i \in S^{'}$ é uma amostra da classe minoritária em questão; y_i é um dos seus $k - vizinhos$ mais próximos de x_i; $y_i \in S^{'}$ e; δ é um número aleatório entre [0,1]. Como exemplo sintético gerado através da Equação 3.19, o resultado é um ponto ao longo da reta que une o ponto x_i em relação aos $k-vizinhos$ mais próximos escolhidos aleatoriamente (HE et al., 2009).

Para um entendimento melhor do SMOTE, a Figura 3.11(b) mostra como é o resultado da sobreamostragem. Em vez de selecionar aleatoriamente a partir dos mesmos casos, sinteticamente é gerado novos casos com base nos casos existentes e seus vizinhos mais próximos, em um esforço para ampliar a fronteira de decisão do modelo (WANG et al., 2015).

No exemplo da Figura 3.11(a), os triângulos são os casos da classe minoritária e os pontos azuis são os casos da classe maioritária. Novos casos são criados à classe minoritária, entre os casos existentes e seus vizinhos mais próximos, resultando na Figura 3.11(b).

O SMOTE foi utilizado neste trabalho para balancear a base de indivíduos, devido ao número de candidatos a não massa possuírem uma quantidade muito superior ao de massa.

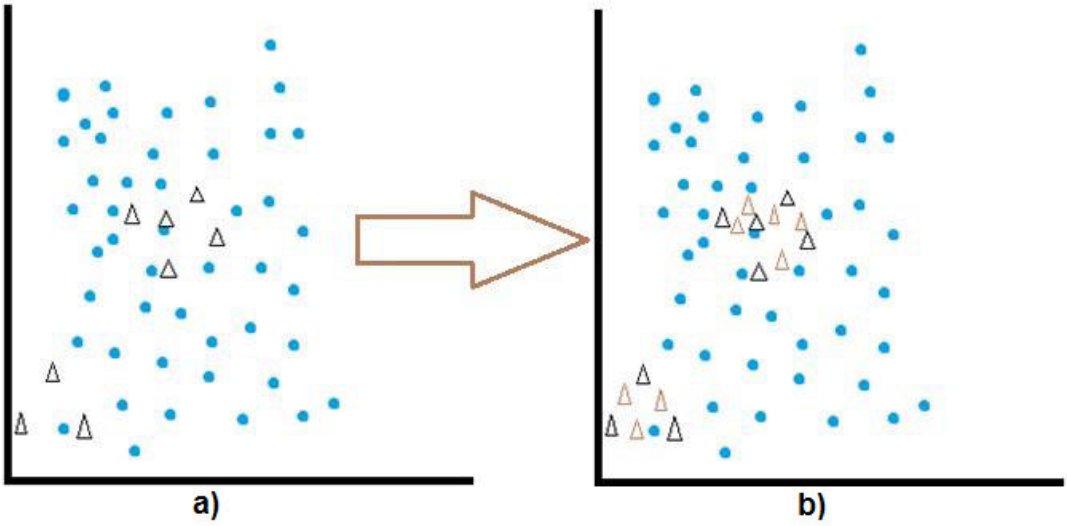

Figura 3.11: Exemplo do resultado do SMOTE: a) Amostra original; b) Amostra depois da execuçao do SMOTE. Fonte:(WANG et al., 2015)

.

3.3 Reconhecimento de Padrões e Métricas de Desempenho

Nesta seção discute-se reconhecimento de padrões, uma subárea da aprendizagem de máquina, que visa classificar informações, ou padrões, baseado em um conhecimento prévio ou em informações estatísticas extraídas dos padrões encontrados. Em seguida é apresentada as métricas utilizadas para validação da metodologia deste trabalho.

3.3.1 Máquina de Vetor de Suporte

A Máquina de Vetor de Suporte (SVM) é uma técnica de aprendizagem supervisionada usada para estimar uma função que classifique dados de entrada em duas classes (VAPNIK, 1998). Uma característica especial desta família de técnicas consiste no fato delas minimizarem o erro empírico de classificação, maximizando ao mesmo tempo a margem geométrica de erro. Com isso, essas técnicas são também conhecidas como classificadores de margem máxima (maximum margin classifiers) (SCHÖLKOPF e SMOLA, 2001).

O princípio básico da SVM é a construção de um hiperplano como superfície de decisão, cujas margens de separação entre as classes (Figura 3.12) seja máxima (SCHÖLKOPF e SMOLA, 2001). Por hiperplano entende-se uma superfície que separa duas regiões em um espaço multidimensional, onde o número de dimensões pode ser, até, infinito (LORENA e CARVALHO, 2007).

Seja o conjunto de amostras de treinamento (x_i, y_i) sendo, $x_i \in R^n$ o vetor de entrada y_i,

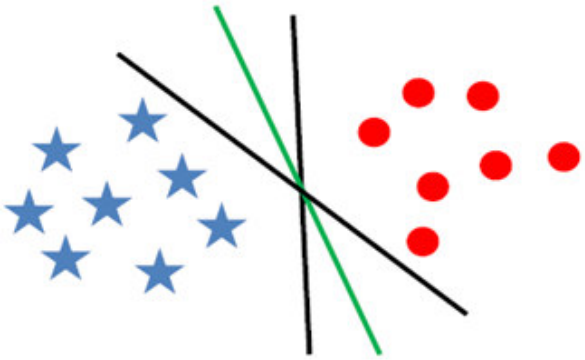

Figura 3.12: Ilustrando a separação de duas classes linearmente separáveis através de hiperplanos (LORENA e CARVALHO, 2007).

a classificação correta das amostras e $i = 1, 2,..., n$ o índice de cada ponto amostral.

O objetivo da classificação é estimar a função $f(x) : R^n \rightarrow [-1; +1]$ que separe corretamente os exemplos de teste em classes distintas, como mostra a Figura 3.12.

A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando valores tais que a seguinte relação seja satisfeita com a Equação 3.20:

$$y_i((w.x_i) + b) \geq 1, \tag{3.20}$$

onde w é o vetor normal ao hiperplano e b é o corte ou a distância da função de f em relação a origem, como mostra a Figura 3.13.

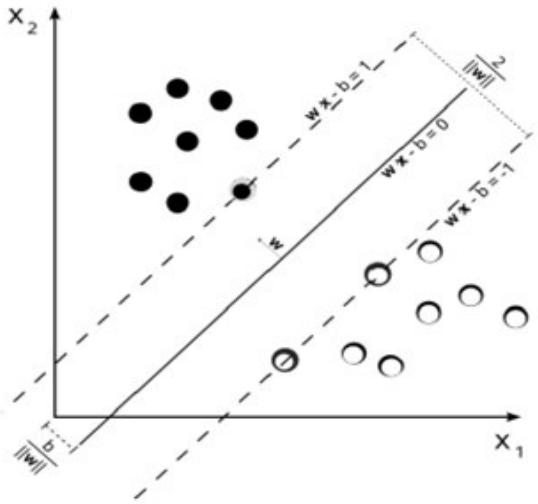

Figura 3.13: Ilustração dos hiperplanos de separação do SVM. (LORENA e CARVALHO, 2007)

Os valores ótimos de w e b serão encontrados de acordo com a minimização da seguinte Equação 3.21:

$$\Phi(w) = \frac{w^2}{2}, \tag{3.21}$$

sujeita a restrição da Equação 3.20.

A SVM 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 for possível, conforme Equação 3.22.

$$y_i((w.x_i) + b) \geq 1 \longrightarrow y_i((w.x_i) + b) + \varepsilon_i \geq 1 \tag{3.22}$$

A partir de agora o problema de otimização passa a ser então a minimização da Equação 3.23.

$$\Phi(w) = \frac{w^2}{2} + C\sum_{i=1}^{N} \varepsilon_i \tag{3.23}$$

Com a seguinte restrição da Equação 3.24.

$$y_i((w.x_i) + b) + \varepsilon_i \geq 1, \tag{3.24}$$

onde C parâmetro de treinamento que estabelece um equilíbrio entre a complexidade do modelo e o erro de treinamento, ε é a variável de folga, custo extra para os erros, permitindo a classificação de *outliers* e N representa o número de amostras da entrada.

Através da teoria dos multiplicadores de Lagrange, chega-se à Equação 3.25. O objetivo então passa a ser encontrar os multiplicadores de Lagrange α_i ótimos que satisfaçam a segunda Equação 3.26.

$$L(a) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j (x_i x_j) \tag{3.25}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, 0 \leq \alpha_i \leq C \tag{3.26}$$

Depois de resolvido o problema nas equações anteriormente, a classificação de um novo padrão envolve apenas verificar o sinal da Equação 3.27.

$$g(x) = sgn(f(x)) = sgn(\sum_{x_i} \alpha_i^* y_i x_i.x + b^*) = \{^{+1 se \sum_{x_i} \alpha_i^* y_i x_i.x + b^* > 0}_{-1 se \sum_{x_i} \alpha_i^* y_i x_i.x + b^* < 0} \tag{3.27}$$

Classificar amostras que não são linearmente separáveis é necessário uma transformação não-linear que transforme o espaço de entrada (dados) para um novo espaço (espaço de características).

A SVM remete a amostra a um novo espaço de dimensão suficientemente grande, e através dele, a amostra pode ser linearmente separável. Essa construção depende do cálculo de uma função k (LORENA e CARVALHO, 2007) de núcleo de um produto interno, adicionado a Equação 3.25, transformando-a na Equação 3.28.

$$L(a) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i x_j) \tag{3.28}$$

A função k realiza o mapeamento das amostras para um espaço de dimensão muito elevada sem aumentar a complexidade dos cálculos (LORENA e CARVALHO, 2007), transformando uma equação com restrição em outra sem restrição. Para melhor visualização a Figura 3.14 nos mostra a transformação de um espaço não-linear em um novo espaço linearmente separável.

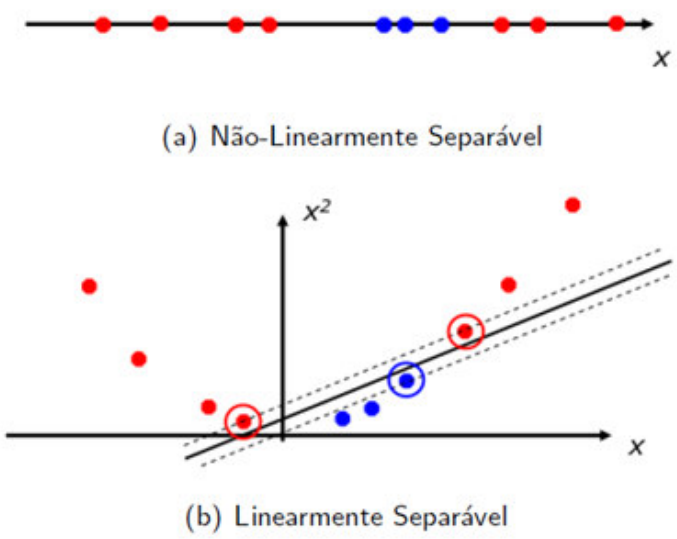

Figura 3.14: Imagem da transformação de um espaço não-linear separável em um espaço linearmente separável. (LORENA e CARVALHO, 2007)

Dessa forma, a técnica possibilita a classificação de padrões de acesso, separando-os de acordo com os vetores de suporte do hiper-plano determinados. A técnica pode ser utilizada para predição de padrões numa sequência longa, conforme demonstrado em (HIROSE et al., 2007).

Nesta pesquisa o SVM foi aplicado na classificação e validação dos resultados.

3.3.2 Validação de Resultados

Decorrido o reconhecimento de padrões, deve-se realizar a avaliação para verificar se os resultados obtidos foram satisfatórios e onde pode ser melhorado. Esta avaliação é um processo de comparação das medidas obtidas durante o reconhecimento dos padrões, utilizando o teste em estudo e um teste de referência.

Para a validação de uma metodologia, as métricas utilizadas são derivadas da matriz de confusão (Figura 3.15), sendo as medidas: Sensibilidade, Especificidade, Acurácia e Média de Falsos Positivos por Imagem as mais utilizadas na área de processamento de imagens médicas, conforme descritas abaixo:

		Previsto	
		positivos	negativos
Real	povitivos	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	negativos	**FN** Falso Positivo	**VN** Verdadeiro Positivo

Figura 3.15: Matriz de confusão.

- Sensibilidade é a proporção de verdadeiros positivos: a capacidade do sistema em predizer corretamente a condição para casos que realmente a tem. O cálculo da sensibilidade é feito com na Equação 3.29.

$$sen = \frac{VP}{VP + FN} \tag{3.29}$$

- Especificidade é a proporção de verdadeiros negativos: a capacidade do sistema em predizer corretamente a ausência da condição para casos que realmente não a tem. O cálculo da especificidade é feito com na Equação 3.30.

$$esp = \frac{VN}{VN + FP} \tag{3.30}$$

- Acurácia é a proporção de predições sem levar em consideração o que é positivo e o que

é negativo. O cálculo da acurácia é feito com a Equação 3.31.

$$acu = \frac{VP + VN}{VP + FN} \tag{3.31}$$

- Falsos Positivos por Imagem é a média de falsos positivos por imagem (BUSHBERG e BOONE, 2011) (BUSHBERG e BOONE, 2011), é a razão entre o número de falsos positivos encontrados e o total de casos avaliados de uma imagem.

$$FP/i = \frac{\sum_{i=1}^{n} i_{FP}}{n} \tag{3.32}$$

onde i é a $i^{-ésima}$ imagem analisada e n o número total de imagens. FP é a quantidade de falsos positivos da imagem i.

3.3.3 Curva Free Receiver Operating Characteristic

A *Free Receiver Operating Characteristic* ou simplesmente FROC, é um gráfico utilizado para representar a eficiência de um método para detectar achados clínicos em imagens, usando como medidas a sensibilidade e a quantidade de média de falsos positivos por imagem, Como mostra a Figura 3.16.

A FROC é aconselhável sua utilização quando a localização da estrutura de interesse é uma informação importante, indicando se a segmentação ou detecção pode ser mapeada em um intervalo pré-determinado, de modo a torná-lo estatisticamente mais flexível no processo de análise. Segundo mostra BRAZ (2014) em seu trabalho.

Para a construção da FROC é necessário definir as seguintes estatísticas:

1. LL = lesão localizada corretamente;
2. NL = lesão erroneamente localizada (não existente);
3. LLF = fração de lesões corretamente localizadas (LL / número de lesões); e
4. NLF = fração de lesões erroneamente localizadas (NL / número de imagens).

obedecendo a restrição de $0 \leq LLF \leq 1$ e $0 \leq NLF$.

Analisando a Figura 3.16, observa-se que o eixo Y (LLF) está relacionado à sensibilidade e seu maior valor é 1. No eixo X (NLF) relaciona-se com quantidade de falsos positivos, e seu valor cresce indefinidamente (SAMPAIO et al., 2015).

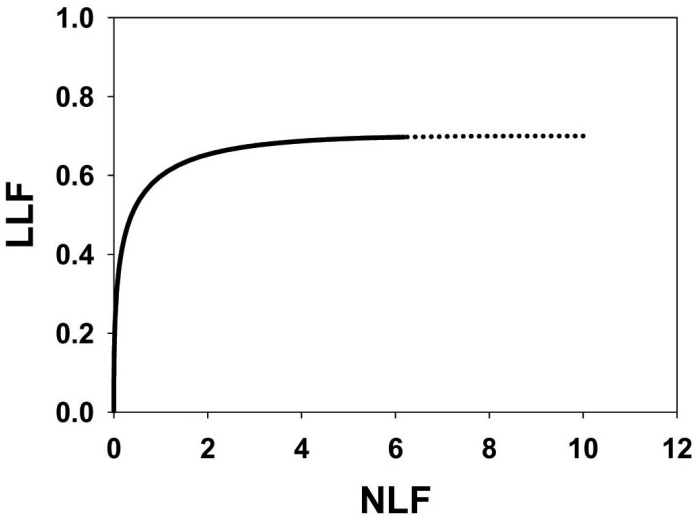

Figura 3.16: Ilustração de uma curva FROC. Fonte:(BRAZ, 2014)

.

Capítulo 4

Metodologia Proposta

Neste Capítulo é apresentada a metodologia propostas deste trabalho, com as seguintes etapas: Na primeira etapa é realizada a aquisição das imagens da base *Digital Database for Screening Mammography (DDSM)*; na segunda etapa é realizado o pré-processamento, com o intuito de remoção das estruturas indesejadas e um realce com objetivo de facilitar o processo da segmentação; na terceira etapa, realizamos a segmentação das imagens usando o algoritmo de *Otsu* e *PSO*; na quarta etapa são realizados dois processos de redução de falsos positivos, a redução por altura da imagem e o *Graph Clustering*, em seguida são encontrados os candidatos a massa e não massa; na quinta etapa é realizado a descrição de textura, extraído-se os índices de diversidade funcional como características dos indivíduos. Por fim, é executado reconhecimento e classificação. A Figura 4.1 mostra este processo.

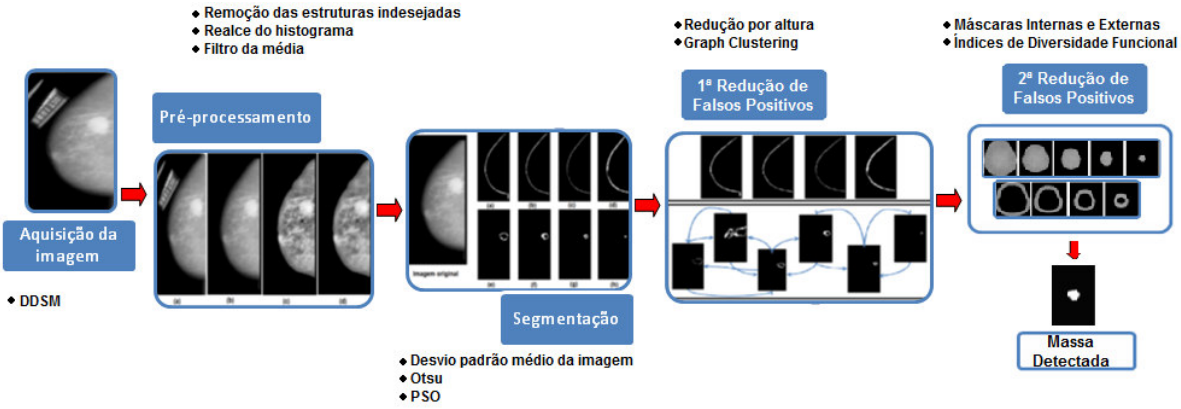

Figura 4.1: Etapas da metodologia.

4.1 Aquisição de Imagens

O banco de imagens escolhido foi o *Digital Database for Screening Mammography-DDSM* (HEATH et al., 2000) que é um banco de dados público contendo mais 2.000 casos, fornecidas gratuitamente na Internet.

Cada caso da base de dados possui quatro imagens da mama (projeções Crânio Caudal - CC e Médio Lateral Oblíquo - MLO), além de informações sobre o exame realizado e os dados das imagens. Todas as informações contidas no DDSM foram fornecidas por especialistas (HEATH et al., 2000).

Para a realização deste trabalho utilizamos 621 imagens com a existência de lesão como critério de inclusão. Essas foram as mesmas utilizadas nos trabalhos de BRAZ (2014) e SAMPAIO et al. (2011), por este motivo utilizamos a mesma base de imagens para efeito de comparação.

4.2 Pré-processamento

Antes de fazer a segmentação nas mamografias, é importante ressaltar a existência de estruturas que são indesejadas a qualquer metodologia de segmentação em mamografias digitais, tais como: ruídos, bordas, marcações e músculo peitoral. Esses oriundos na aquisição das imagens.

Para remover estas estruturas, utilizou-se a metodologia desenvolvida por (SAMPAIO et al., 2011). Após a remoção destas estruturas, é aplicado uma técnica de realce local baseado no histograma (CLAHE) e um filtro da média (Seções 3.2.1.2 e 3.2.1.1), com objetivo de realçar as estruturas internas das mamografias (GONZALEZ e WOODS, 2007), como mostra a Figura 4.2.

Após a realização destes filtros as imagens serviram de entrada para a segmentação proposta, detalhada nas próximas seções.

4.3 Segmentação das Imagens Mamográficas

Após a etapa de pré-processamento, as imagens foram submetidas ao processo de segmentação, obedecendo as etapas descritas abaixo:

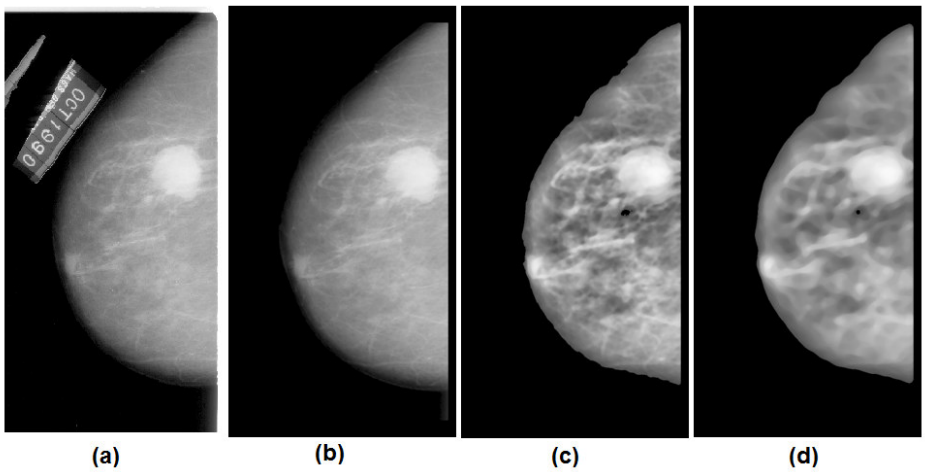

Figura 4.2: Resultado das etapas do melhoramento: (a) Imagem original; (b) Imagem sem as bordas e marcações; (c) Realce local (CLAHE); (d) Filtro da média.

4.3.1 Desvio Padrão Médio da Imagem

O desvio padrão médio da imagem (*dpmi*) é o valor do desvio padrão médio de todas as janelas da imagem, que serve de base à comparação entre os desvios padrões de cada *cluster* gerado. Neste processo divide-se a imagem em janelas de 12x12 (Figura 4.3). Este tamanho de janela foi escolhido de forma empírica, por ter apresentado os melhores resultados. Em seguida, calcula-se o desvio padrão de cada janela (δ_j), soma-se todos os desvios padrões e divide-se pela quantidade dos *clusters*, conforme as Equações 4.1 e 4.2:

$$dpmi = \frac{1}{M}\sum_{j=1}^{M}\delta_j \tag{4.1}$$

$$\delta_j = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_i - \mu_j)^2} \tag{4.2}$$

onde M é a quantidade de janelas encontradas, δ_j é o desvio padrão de cada janela, N representa o número de pixels e μ_j é a média dos elementos da janela.

4.3.2 Algoritmo de *Otsu*

O algoritmo de OTSU (1975) é executado para encontrar o primeiro limiar da imagem, dividindo esta em dois *clusters* (Figura 4.4(a)), em seguida calcula-se com a Equação 4.3 o desvio padrão de cada *cluster* gerado.

$$\delta^2 = \frac{1}{N}\sum_{j=1}^{N}(y_i - m_{ij})^2, \tag{4.3}$$

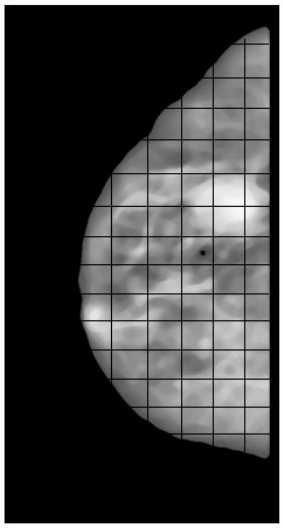

Figura 4.3: Imagem da mama dividida em janelas 12x12.

se o $\delta^2 > dpmi$, calcula-se com a Equação 4.4 o centróide g de cada *cluster* como o novo limiar para se dividir este *cluster* em dois novos *clusters* (Figura 4.4(b)). Novamente é calculado o desvio padrão deste *cluster*, se ($\delta^2 > dpmi$), então, este passo é executado de forma recursiva até que δ^2 não seja maior que *dpmi*.

$$g = \frac{1}{N} \sum_{j=1}^{N} I(w_i), \tag{4.4}$$

onde, I é a intensidade do pixel w_i e N é o número de *pixels* no grupo. Os limiares encontrados serão utilizados como centróides para o vetor de posições x_i da partícula inicial do enxame de partículas e submetido ao *PSO*, como mostra a Equação 4.5. A Figura 4.4 apresenta os *clusters* montados a partir dos limiares do vetor x_i, gerados pelo algoritmo de Otsu.

$$x_i = (m_{i1}m_{i2},...,m_{ij}m_{iNc}), \tag{4.5}$$

O PSO recebe o vetor x_i e otimiza seus valores, buscando encontrar os melhores limiares, consequentemente os melhores *clusters*. O resultado desse processo é explicado pela Figura 4.5 e detalhado na seção seguinte.

4.3.3 *Particle Swarm Optinization*

Gerado o vetor inicial x_i da primeira partícula do enxame, através do algoritmo de Otsu, o mesmo é submetido ao *PSO* para que estes valores sejam otimizados.

Para a execução do PSO, alguns parâmetros devem ser inicializados, e neste trabalho, após

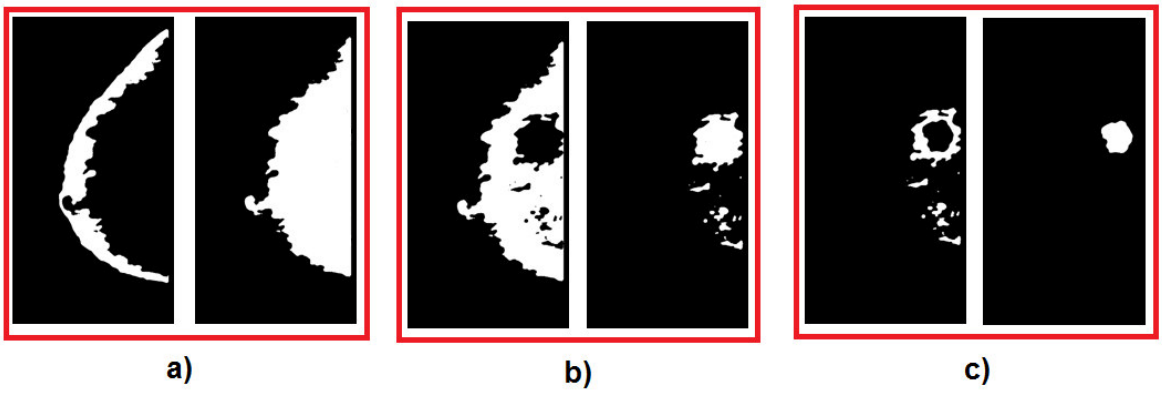

Figura 4.4: Resultado dos *clusters* gerados pelo algoritmo de Otsu: a) 1º *cluster* gerado; b) 2º *cluster* gerado a partir do anterior; c) 3º *cluster* gerado do a partir do segundo.

vários testes, os melhores valores foram:

- k = 10 (quantidade mínima do enxame);
- Dim = 2*k (dimensão da função de aptidão (*fitness*));
- $L = 255$ (número do nível de cinza);
- *SwarmSize* = 3*Dim (tamanho do enxame);
- $lb = 0$ (limite inferior de aptidão de cada partícula);
- $ub = L$ (limite superior de aptidão de cada partícula);
- *Iterations* = k*Dim (quantidade de Iterações);
- $Vmax = L$ (Velocidade máxima de cada partícula);
- $Vi = 0$ (velocidade inicial da partícula);
- $w = 0.729$ (constante de inércia da partícula);
- $c_1 = c_2 = 2.05$ (constantes de aprendizado cognitivo (c_1) e social (c_2)).

Para cada partícula, é gerado um vetor de posições aleatórias de acordo com a quantidade de partículas, Equação 4.6.

$$x_i = (rand(0,1).(ub - lb) - lb), \tag{4.6}$$

Para cada elemento do vetor x_i de cada partícula, atualiza-se o valor da velocidade, Equação 4.7:

$$v_i = wv_i + (c_1.rand(0,1).(pbest - x_i)) + (c_2.rand(0,1).(gbest - x_i)), \tag{4.7}$$

onde *pbest* representa o melhor valor de x_i local e *gbest* representa o melhor valor global de todas as partículas. Em seguida atualiza-se o valor de x_i, Equação 4.8.

$$x_i = x_i + v_i \tag{4.8}$$

Após todos os elementos do vetor de posições x_i serem atualizados, atualiza-se os valores de *pbest* e *gbest*. De posse do vetor de posições x_i, contendo todos os centróides da partícula, é calculado a menor distância euclidiana de todos os pixels da imagem em relação a todos os centróides do vetor de posições x_i. Cada pixel será agrupado ao centróide que, dada a Equação 4.9, retorne o menor valor.

$$d_{min} = \sqrt{(z_p - x_i)^2} \tag{4.9}$$

denota-se, d_{min} a menor distância do pixel z_p em relação aos centróides do vetor de posições x_i.

Por fim, é calculado o valor de aptidão *(fitness)* da partícula (Equação 4.10) dos valores atualizados do vetor de posições. Quanto menor o valor encontrado, melhor é aptidão (*fitness*) da partícula.

$$\delta^2 = (\sum_{j=1}^{N_c}[\frac{\sum_{\forall z_p \in C_{ij}} d_{min}}{|C_{ij}|}])/N_c \tag{4.10}$$

onde $|c_{ij}|$ é a quantidade de elementos do *cluster* e N_c é quantidade de *clusters* encontrado.

Decorridos estes passos, repete-se novamente até o número de *Iterations* como estabelecido nos valores dos parâmetros iniciais. Ao fim das iterações o melhor vetor de posições x_i representará o conjunto de centróides para geração dos *clusters*, consequentemente a criação das imagens agrupadas. A Figura 4.5 mostra o resultado.

4.4 Redução de Falso Positivo

Neste etapa, inicialmente aplicou-se o crescimento de região com o intuito de isolar a ROI. Em seguida, executam-se três técnicas de redução de falsos positivos, sendo a primeira um filtro de área, chamado nesta pesquisa de redução pela distância, a segunda o *Graph Clustering* e a

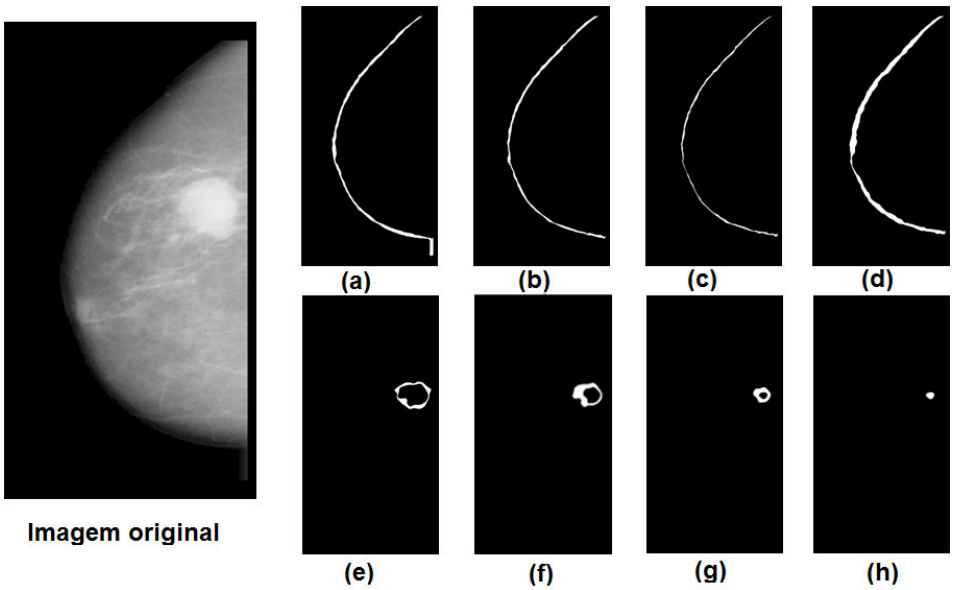

Figura 4.5: Resultado do processo de segmentação: imagem original a esquerda, as demais imagens são os *clusters* gerados.

terceira a descrição de textura. Essas reduções são detalhadas nas seções seguintes.

4.4.1 Primeira Redução de Falsos Positivos

Nesta etapa, descrevemos como foram realizadas as reduções necessárias à condução desta pesquisa. Na busca dos melhores resultados, aplicamos duas reduções de falsos positivos, detalhadas a seguir.

4.4.1.1 Redução pela Distância

Este processo remove as ROIs que possuem uma distância entre o primeiro e o último pixel da imagem, maior que 55% em relação aos da original, e, estas serão consideradas como ROIs não candidatas, sendo automaticamente descartadas. O percentual de 55% foi escolhido de forma empírica, pois foi o que apresentou melhores resultados.

Na Figura 4.5(a,b,c e d) são apresentados alguns exemplos destas ROIs. As imagens mostradas na Figura 4.5 serão submetidas ao procedimento de redução pela distância, que é a distância euclidiana d entre o primeiro ponto (x_1, y_1) até o último ponto (x_2, y_2) da ROI, de acordo com a Equação 4.11.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{4.11}$$

4.4.1.2 Redução com Graph Clustering

Graph Clustering é o processo de agrupar os vértices do grafo em *clusters* levando em consideração a estrutura das arestas dos grafos. Neste trabalho, o *Graph Clustering* é utilizado para unir as ROIs vizinhas, montando um grafo, a partir da união dessas ROIs. Para isso, adotamos algumas definições: a vizinhança será definida em 3x3, o grafo G será formado a partir das ROI de uma imagem, verificando sua vizinhança com todas as ROIs existentes do *cluster* original, e por fim, o grafo G será direcionado.

Após essas definições, analisamos a vizinhança da ROI atual, se um pixel qualquer desta ROI, possuir no mínimo dois pixels vizinhos em outra ROI, podemos dizer que estas ROIs são adjacentes, e, portanto deve existir uma aresta em G ligando os vértices correspondentes.

Após todo o processo, os nós do grafo G que possuírem mais de duas ligações, devem ser removidos e todos os nós que não tenham nenhuma ou no máximo duas ligações devem permanecer, pois esses nós darão origem as ROIs candidatas. Cada nó do grafo G resultante desse processo, representa uma ROI (Figura 4.6a). A partir de um nó qualquer do grafo G, é calculado seu valor de fator de forma circular (FFC) (Equação 4.12), da seguinte forma:

- Se FFC for menor que 10%: o nó será descartado e será escolhido outro nó, e o processo se repete.
- Se FFC for maior que 10%: verifica-se suas adjacências, realizando as uniões do nós. Após cada união, o FFC é calculado novamente, caso este, seja maior que 10% a união é válida e será verificada (caso exista) a próxima adjacência. Se o FFC, após cada união, resultar em um valor menor que 10%, esta união não será válida, e o nó que foi unido será descartado.

$$FFC = \frac{4.\pi.A}{P^2}, \tag{4.12}$$

onde A e P correspondem a área e o perímetro respectivamente de cada ROI. Adotamos o percentual de 10% para o FFC, em decorrência dos testes, e este percentual foi o que apresentou melhores resultados.

Por fim, restarão apenas as ROIs que não tiveram ligações e a ROI resultante da união das ROIs vizinhas, (Figura 4.6b).

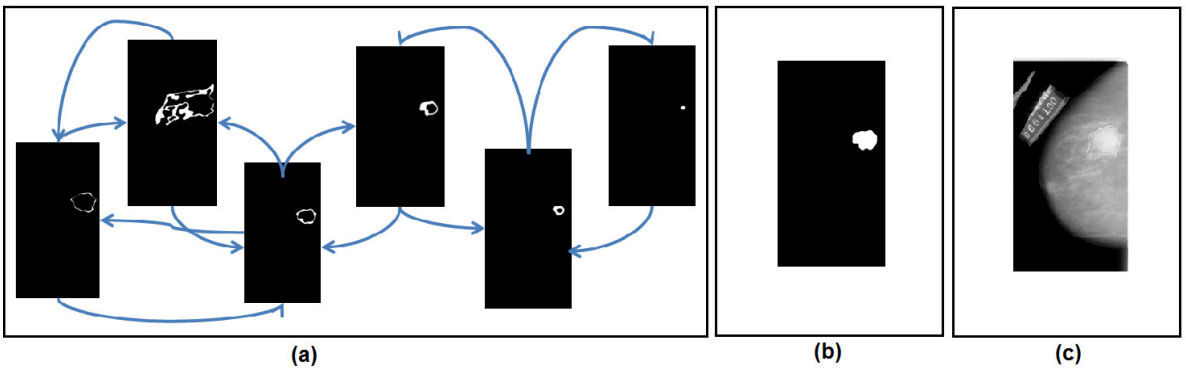

Figura 4.6: Resultado do *Graph Clustering*: (a) Grafo gerado pelas ROIs vizinhas; (b) Imagem final gerada da união das ROIs vizinhas; (c) Imagem original com a marcação do especialista.

4.4.2 Segunda Redução de Falsos Positivos

Após a primeira redução de falso positivos, aplicamos a segunda redução em cada região que será individualmente caracterizada usando os índices de diversidade funcional, Seção 3.2.8.2, permitindo a separação dos candidatos a massa ou não massa.

4.4.2.1 Extração de Características

Foram extraído como características de medidas das imagens, os índice de diversidade Funcional, conforme descritos na Seção 3.2.8.2. Cada ROI possui as seguintes características: i) FADa: índice de diversidade funcional abundante; II) FADe: índice de diversidade funcional abundante da Especie; III) FADp: índice de diversidade funcional atributo pixel. Todos os índices, são extraídos das máscaras internas e externas da imagem original de 8 bits, e das quantizações de 7 e 6 bits, permitindo extrair características mais detalhada das regiões. Estes passos serão descritos detalhadamente nas próximas seções.

4.4.2.2 Máscaras Internas e Externas

Inicialmente é necessário prepararmos as sub-regiões de cada ROI para que os índices de diversidade funcional sejam extraídos. Dessa maneira a análise de textura será mais vantajosa, pois analisará regiões distintas de cada ROI.

A geração das máscaras internas inicia-se pelo centro da ROI. Essa máscara nada mais é que uma redução da ROI original preservando-se o mesmo centro da original. Essas máscaras foram geradas com as reduções, conforme descritas na Figura 4.7.

Na geração das máscaras externas, o princípio é a diferença de duas internas, consecutivas de mesmo centro. Isto é feito para se obter mais detalhes de outras regiões da mama. A

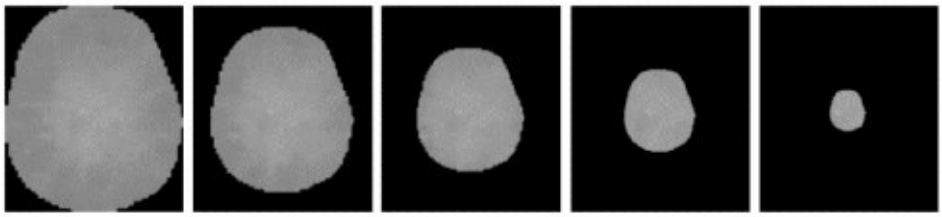

Figura 4.7: Exemplo de máscaras internas. Da esquerda para direita: ROI original, ROI 80%, ROI 60%, ROI 40% e ROI 20%. Fonte: (SAMPAIO et al., 2015).

Figura 4.8 mostra uma exemplo de máscara externa.

Figura 4.8: Exemplo de máscaras externas. Da esquerda para direita: diferença da ROI original com a 80%, diferença da ROI 80% com a 60%, diferença da ROI 60% com a 40%, diferença da ROI 40% com a 20%. Fonte: (SAMPAIO et al., 2015).

Após a geração das máscaras internas e externas, os vetores de características são extraídas através dos índices de diversidades funcional descritos abaixo.

4.4.2.3 Criação do Dendrograma

Nesta etapa é criado o dendrograma a partir dos pixel pertencentes a uma ROI. Na construção, utilizou-se o algoritmo de Otsu para separar os grupos de *pixels* a partir da similaridade dos tons de cinza, formando as comunidades. Cada nó representa os grupos separados pelo algoritmo de Otsu. Para exemplificar melhor, veja os números 2 e 5 na Figura 3.9 da Seção 3.2.8.2, eles pertencem ao mesmo grupo, e representam os valores de *pixel* da ROI. Os demais números pertencentes aos outros nós, formam os grupos restantes.

4.4.2.4 Extração de Características Usando Índice de Diversidade Funcional Abundante (FADa)

O FADa foi extraído da ROI através da Equação 3.16, levando-se em consideração a abundância da quantidade de *pixels* da mesma espécie no dendrograma..

Para calcular a distância d_{ij} entre as os pares de espécies (*pixels*) da ROI, utilizou-se a Equação 3.17, onde, o valor de X na referida equação é a posição da especie no dendrograma

e a abundância p para este caso é a quantidade de cada pixel de mesmo valor conforme Equação 3.18.

4.4.2.5 Extração de Características Usando Índice de Diversidade Funcional Abundante da Especie (FADe)

O FADe foi extraído da ROI através da Equação 3.16, levando-se em consideração a abundância da quantidade de *pixel* da mesma espécie no dendrograma.

A distância d_{ij} entre as posições das espécies (valor do *pixel*) da ROI é calculada pela Equação 3.17, considerando X o valor da espécie (valor do pixel) no dendrograma e a abundância p, neste caso é a quantidade de *pixels* de mesmo valor (Equação 3.18).

4.4.2.6 Extração de Características Usando Índice de Diversidade Funcional Abundante pixel (FADp)

O FADp é extraído da ROI usando a Equação 3.16, levando-se em consideração a abundância como sendo a soma total dos valores de pixel da mesma espécie no dendrograma.

A d_{ij} (Equação 3.17) é a distância entre as posições das espécies (valor do *pixel*) da ROI, sendo X o valor da espécie (valor do pixel) no dendrograma e a abundância p, neste caso é a soma total dos valores de *pixels* de mesmo valor, conforme Equação 3.18.

4.5 Reconhecimento de Padrões

Nesta etapa, classifica-se cada região em massa ou não massa. Para isso, foi utilizado o reconhecimento de padrões baseado na textura, adquirido na extração de características (Seção 4.4.2.1), juntamente com o SVM (Seção 3.3.1) para classificar as regiões.

4.5.1 Base de Treinamento e Teste

Com as regiões segmentadas, é gerada uma base de indivíduos, formando um vetor de características com cada rótulo pertencente a sua classe (massa ou não massa). Para realizar o reconhecimento, primeiro normaliza-se os valores das variáveis para uma melhor convergência do SVM. Em seguida realiza os seguintes passos:

1. A base de características é balanceada através do SMOTE (Seção 3.2.9).

2. Faz-se a separação da base em treino e teste de forma aleatória, com os percentuais de 80% e 20%, respectivamente;

3. A base de treino, com os 80% é submetidos ao SVM (Seção 3.3.1) 5 (cinco) vezes para encontrar o melhor modelo;

4. Encontrado o melhor modelo, o mesmo é testado com os 20% para validação da metodologia.

Por fim, para validar os resultados do reconhecimento e da metodologia são utilizados as medidas de sensibilidade, especificidade, acurácia, taxa média de falso positivos por imagem (FP/i) e a curva FROC.

Capítulo 5

Resultados e Discussões

Neste capítulo, é apresentado os resultados e discussões da metodologia empregada neste trabalho. Serão apresentados também alguns estudos de casos relatando os sucesso e falhas encontrados na pesquisa.

5.1 Aquisição e Pré-processamento das Mamas

Nesta etapa, foram utilizadas 388 mamas não densas e 233 mamas densas adquiridas da DDSM, totalizando 621 mamas. Essas foram selecionadas com o critério de densidade, conforme a especificação do especialista no arquivo *overlay* e possuindo pelos menos uma lesão de massa. O critério de escolha dessas imagens, foi o fato de que os trabalhos de (SAMPAIO et al., 2011) e (BRAZ, 2014) usaram as mesmas imagens, assim, poderíamos fazer uma comparação mais fidedigna.

Em seguida, as imagens foram submetidas ao processamento de remoção das estruturas indesejadas (Seção 3.2.1), logo depois aplicou-se o realce local do histograma (Seção 3.2.1.2) para melhorar o contraste, e por fim, o filtro da média (Seção 3.2.1.1) suavizando a imagem.

No processo de detecção das regiões candidatas a massa, inicialmente utilizou-se o *dpmi* (Seção 4.3.1) para encontrar o desvio padrão médio da imagem que servirá de base para gerar os grupos de cada imagem. Após encontrado o dpmi, o algoritmo de *Otsu* (Seção 4.3.2) separa os grupos a partir de um limiar, gerando o vetor de limiares. Por fim, o PSO (Seção 4.3.3) recebe esse vetor como partícula inicial e otimiza, gerando novos valores e consequentemente as ROIs candidatas.

5.2 Resultado com as Mamas Não Densas

Após todo o processo descrito acima, de posse das ROIs pré-candidatas, já segmentadas, executa-se uma redução de falsos positivos com o intuito de descartar regiões, que de fato não são massas.

Nessa etapa, das 388 mamas não densas, foram geradas 61.556 ROIs. Após os processos de redução de falsos positivos (Seção 4.4), a quantidade das ROIs candidatas se reduziram para 16.501. Dessas, 1.659 foram consideradas massa e 14.842 não massa. Foram perdidas apenas 13 massas, correspondendo a uma taxa de 2,33% de perda. A Figura 5.1 apresenta os resultados encontrados com a referida redução. Notadamente, observa-se que a redução de falsos positivos atingiu um bom desempenho, chegando a reduzir 73,19% dos candidatos, além de atingir um percentual de 96,13% de acerto, 0,64 de FP/i e 0,98 de área sob a curva FROC.

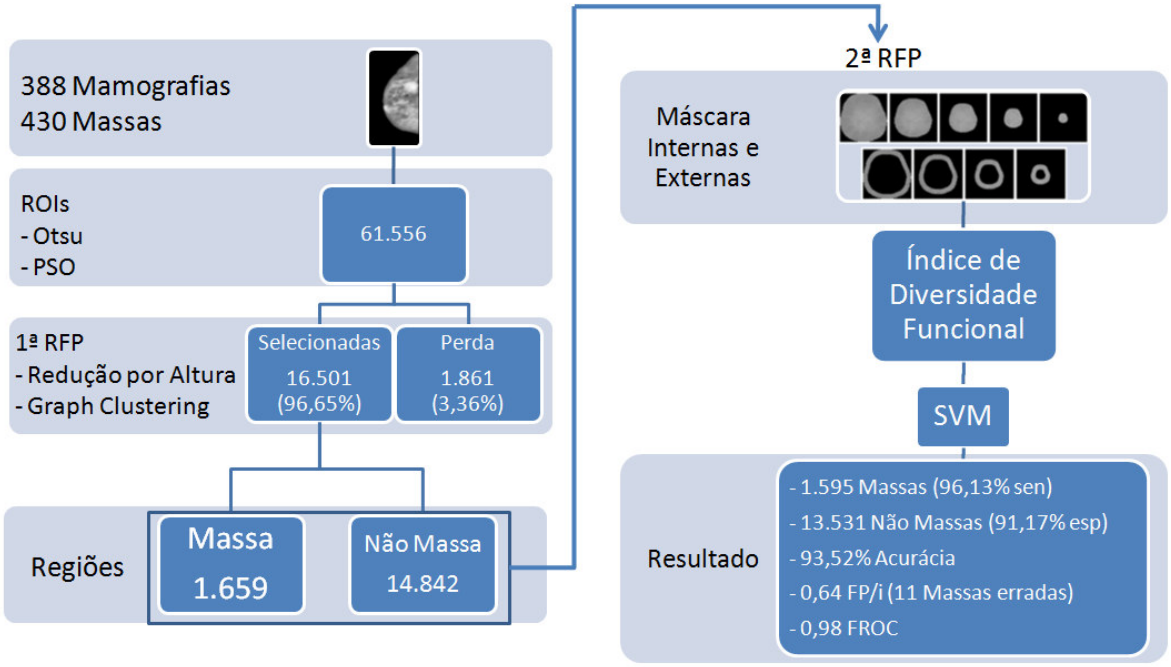

Figura 5.1: Resultado da metodologia com as mamas não densas.

5.2.1 Resultado das Mamas Não Densas Depois das Reduções de Falsos Positivos

Após as etapas de redução, inicia-se a fase de teste dos candidatos a massa ou não massa, para isso foi utilizado a análise de textura com os índices de FD (Seção 3.2.8.2). Em seguida, utilizou-se o SMOTE (Seção 3.2.9) para o balanceamento da base e submetida ao SVM (Seção 3.3.1)para classificar as ROIs candidatas.

A Tabela 5.1 apresenta o desempenho dos testes para diferentes tipo de análise, utilizando os vetores de características extraídos de mamas não densas da DDSM. percebe-se também, que a combinação de análise de textura dos índices de FD juntamente com o SMOTE apresentou o melhor desempenho com 96,13% de sensibilidade, 91,17% de especificidade e 93,52% de acurácia.

Tabela 5.1: Resultado do desempenho da classificação da metodologia nas mamas não densas, comparada com outras técnicas.

Testes	Sensibilidade	Especificidade	Acurácia
Haralick	60,00	70,58	70,12
Índice de FD sem Máscaras	93,62	64,55	65,15
Índice de FD com Máscaras	67,86	73,46	73,22
Índice de FD com Máscaras+SMOTE	96,13	91,17	93,52

Todos os testes foram realizados 5 (cinco) vezes e no final foi calculado a média aritmética de cada um deles, como mostra a Tabela 5.2. Analisando a referida tabela, podemos inferir que, em todos os testes os valores de sensibilidade, especificidade e acurácia se comportaram de maneira satisfatória em todos os casos, mesmo no pior caso (20/80) os valores foram bons, mostrando que a técnica é promissora. Podemos afirmar ainda, que o teste 80/20, apresentou valores significativos, mostrando que a metodologia é eficiente.

Tabela 5.2: Testes do desempenho da classificação da metodologia nas mamas não densas.

Treino/Teste	Sensibilidade	Especificidade	Acurácia
20/80	84,46	79,67	81,93
40/60	91,48	85,50	88,33
60/40	94,73	89,10	91,76
80/20	96.13	91.17	93.52

Outra análise realizada para avaliar o desempenho da metodologia, foi a curva FROC (Seção 3.16), que apresentou o valor de 0,98. Esse resultado permite dizer que a metodologia obteve um bom desempenho.

5.3 Resultado com as Mamas Densas

Nesta etapa foram utilizadas 233 mamas densas da DDSM. Dessas foram geradas 26.585 ROIs, após a etapa da segmentação. Portanto, a redução de falsos positivos (Seção 4.4) se faz necessária devido a grande quantidade de ROIs geradas após a segmentação.

Para essa etapa, também foram usadas as reduções por altura da imagem para remover as

regiões indesejadas e o *Graph Clustering* para unir as ROIs vizinhas. Na etapa da segmentação, foram geradas 26.585 ROIs pré-candidatas, depois da redução de falsos positivos restaram apenas 7.205, correspondendo a uma redução de 72,90%. Resultando em 765 massas e 6.440 não massas. A Figura 5.2 apresenta os resultados encontrados com a referida redução.

A perda de 15 massas durante o processo, mostra que a metodologia é eficiente, chegando a atingir 97,52% de acerto, 0,38 de FP/i e 0,98 de área sob a curva FROC..

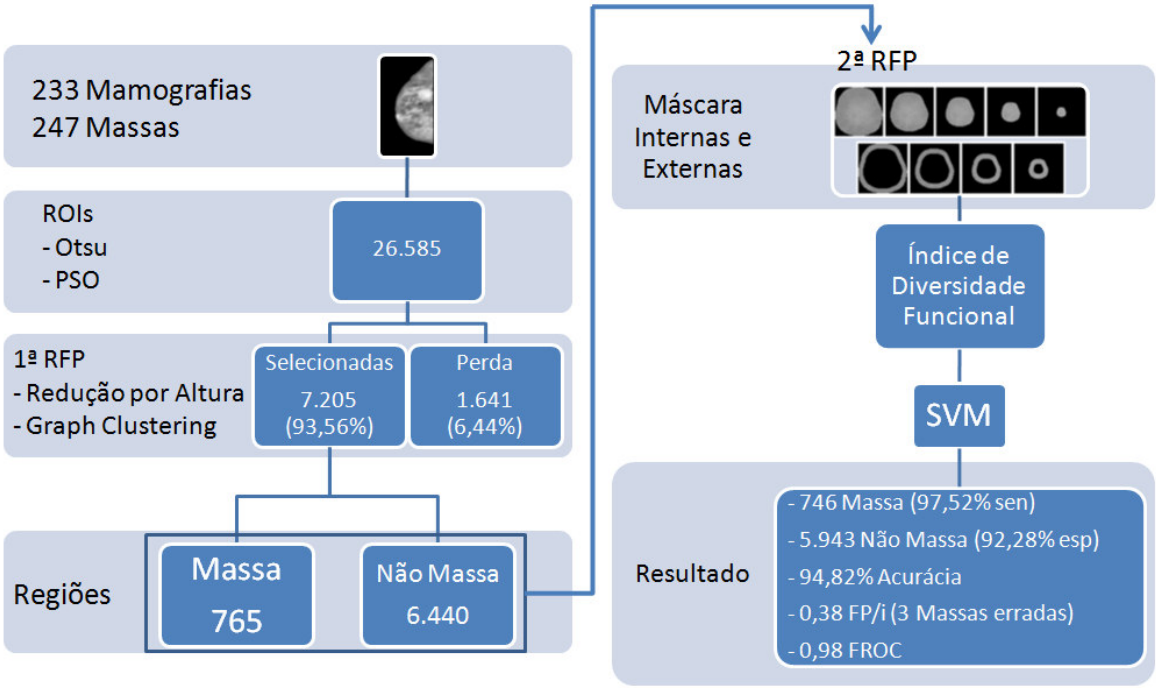

Figura 5.2: Resultado da metodologia com as mamas densas.

5.3.1 Resultado das Mamas Densas Depois das Reduções de Falsos Positivos

Terminada a etapa de redução, inicia-se a fase de classificação das mamas densas, para isso foi utilizado a análise de textura com os índices de FD (Seção 3.2.8.2). Em seguida submetidos ao SVM (Seção 3.3.1) para classificar as ROIs candidatas a massa ou não massa. Os testes são mostrados na Tabela 5.4.

O melhor resultado encontrado pela metodologia, foi a combinação de análise de textura usando os índices de FD juntamente com o SMOTE (Seção 3.2.9). Os resultados estão apresentado na Tabela 5.3, juntamente com outras técnicas para comparação.

Os testes seguiram a mesma metodologia empregada nas mamas não densas, sendo repetidos 5 (cinco) vezes e no final foi calculado a média aritmética de cada um deles, como mostra a

Tabela 5.4. Fazendo uma análise dos dados, podemos aferir que, em todos os testes os valores de sensibilidade, especificidade e acurácia apresentaram valores bem satisfatórios, até mesmo no pior caso (20/80). Podemos afirmar ainda, que o teste 80/20, apresentou valores promissores, mostrando que a metodologia é eficiente.

Tabela 5.3: Resultado do desempenho da classificação da metodologia nas mamas densas em relação a outras técnicas.

Testes	Sensibilidade	Especificidade	Acurácia
Haralick	67.86	72.22	72.04
Índice de FD sem Máscaras	89.47	72.90	73.32
Índice de FD com Máscaras	96.36	71.46	72.06
Índice de FD com Máscaras+SMOTE	97.52	92.28	94.82

Tabela 5.4: Testes do desempenho da classificação da metodologia nas mamas densas.

Treino/Teste	Sensibilidade	Especificidade	Acurácia
20/80	86.87	79.63	83.15
40/60	93.68	86.91	90.23
60/40	96.50	89.81	93.03
80/20	97.52	92.28	94.82

Encerrando os testes da metodologia nas mamas densas, utilizamos como métricas de avaliação os valores de sensibilidade, especificidade e acurácia. Na Tabela 5.3 é apresentado o desempenho dos testes para diferentes tipo de análise, utilizando os vetores de características extraídos das mamas densas da DDSM. Observa-se, que a combinação da análise de textura dos índices de FD juntamente com o SMOTE apresentou melhor desempenho com 97.52% de sensibilidade, 92.28% de especificidade e 94.82% de acurácia.

Por fim, foi analisada a área sob a curva FROC para avaliar o desempenho da metodologia, obtendo o valor de 0,98. Esse resultado permite dizer que a metodologia obteve um bom desempenho.

5.4 Resultados com as Mamas Densas e não Densas

Foi realizado um teste com todas as mamas densas e não densas, perfazendo um total de 621 mamas. Nesse teste também foram usadas as redução pela distância para remover as regiões indesejadas e o *Graph Clustering* para unir as ROIs vizinhas. Na etapa da segmentadas, foram geradas 88.141 ROIs pré-candidatas, depois dessas reduções de falsos positivos restaram apenas 23.706, correspondendo a uma redução de 73.10%. Restaram ao final 2.424 massas e 21.282

não massas.

Após os processos de redução, foram perdidas 28 massas, o que mostra a eficiência da metodologia, atingir uma taxa de acerto de 95.36%, e 0,75 de FP/i e 0,98 de área sob a curva FROC. A Figura 5.3 apresenta os resultados encontrados com a referida redução.

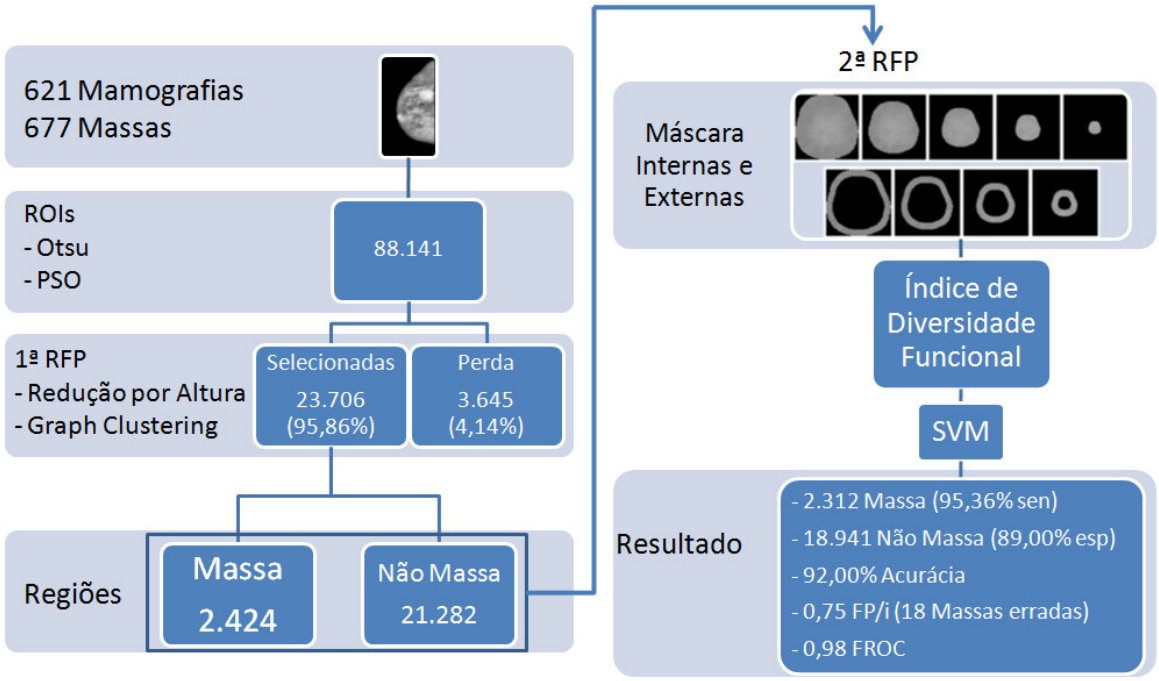

Figura 5.3: Resultado da metodologia com as mamas densas e não densas.

5.4.1 Resultado das Mamas Densas e Não Densas Depois das Reduções de Falsos Positivos

Terminada as etapas de redução, foram utilizados a análise de textura com os índices de FD, o SMOTE para balanceamento da base. Elas foram submetidos ao SVM para classificar as ROIs candidatas. Os resultados podem ser vistos na Tabela 5.6

Na Tabela 5.5, é apresentado o melhor resultado desse teste, e o de outras técnicas para servir de comparação. Portanto, verifica-se mais uma vez que a metodologia proposta conseguiu valores satisfatórios, mostrando ser eficiente.

Os testes obedeceram os mesmos critérios empregados nas mamas densas e não densas. Foram realizados 5 (cinco) vezes e no final é calculado a média aritmética de cada um deles. Analisando os dados da Tabela 5.6, podemos verificar que, nos testes realizados, as métricas de sensibilidade, especificidade e acurácia do teste 80/20, apresentaram os melhores valores e bem próximos aos dos testes anteriores (mamas densas e não densas).

Tabela 5.5: Resultado do desempenho da classificação da metodologia nas mamas densas e não densas em relação a outras técnicas.

Testes	Sensibilidade	Especificidade	Acurácia
Haralick	60.00	68.89	70.04
Índice de FD sem Máscaras	90.41	63.91	64.32
Índice de FD com Máscaras	66.36	70.36	70.16
Índice de FD com Máscaras+SMOTE	95,36	89,00	92,00

Tabela 5.6: Testes do desempenho da classificação da metodologia nas mamas densas e não densas.

Treino/Teste	Sensibilidade	Especificidade	Acurácia
20/80	83.39	78.18	80.66
40/60	90.61	83.75	87.03
60/40	93.36	87.08	90.08
80/20	95.36	89.00	92.00

Para finalizar os testes da metodologia em todas as mamas densas e não densas, utilizamos as mesmas métricas de avaliação dos testes anteriores, e os valores encontrados foram 95,36% de sensibilidade, 89,00% de especificidade e 92,00 de acurácia.

Concluindo a análise utilizamos a curva FROC para avaliar o desempenho da metodologia, e o valor encontrado foi de 0,98, mostrado que a metodologia é promissora.

5.5 Estudo de Casos

Para uma melhor compreensão da metodologia, nesta seção será apresentado alguns casos específicos, com o intuito de exemplificar os testes realizados ao longo da pesquisa.

5.5.1 Primeiro Caso: Sucesso na Detecção da Massa na Mama Não Densa

O primeiro exemplo de caso de sucesso apresentado é a imagem A_1006_1.LEFT_CC, que desde o início do processo até o fim, apresentou resultados satisfatórios em todas as etapas da metodologia. A Figura 5.4 mostra o processo que se inicia com a remoção das estruturas indesejadas da imagem original (a), logo após a remoção dessas estruturas (b), é aplicado a técnica de realce local baseado no histograma CLAHE (c), e o filtro da média (d), com objetivo de realçar as estruturas internas da mama.

Realizado o pré-processamento, a imagem da Figura 5.4(d) é submetida à etapa de segmentação para serem extraídas as regiões candidatas. Nesta etapa, foram geradas 132 ROIs, mas com as reduções de falsos positivos, foram reduzidas para 36 candidatas. A Figura 5.5

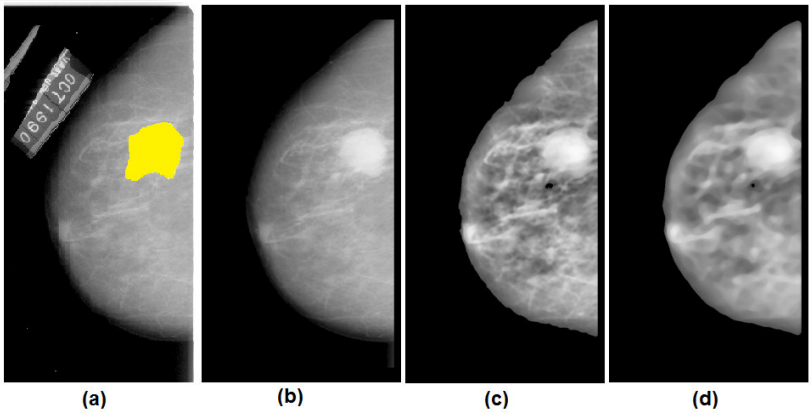

Figura 5.4: Pré-processamento da imagem A_1006_1.LEFT_CC: (a) Imagem original com a marcação em amarelo; (b) Imagem sem as bordas e marcações; (c) Realce do CLAHE; (d) Filtro da média.

mostra o resultado final da metodologia.

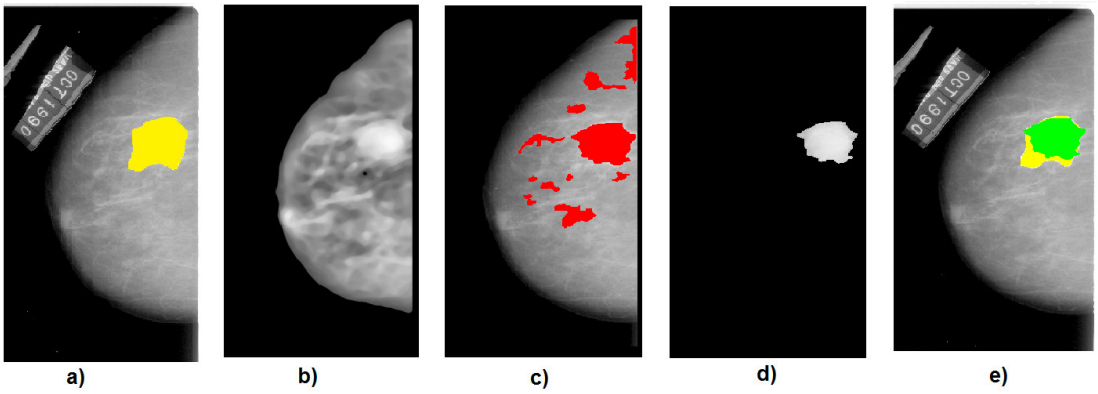

Figura 5.5: Resultado da metodologia aplicado a imagem A_1006_1.LEFT_CC: (a) Imagem original com a marcação em amarelo; (b) Imagem pré-processada; (c) Imagem com as regiões segmentadas em vermelho; (d) Imagem da massa segmentada; (e) Imagem da marcação do especialista com a massa segmentada em verde.

5.5.2 Segundo Caso: Falha na Detecção da Massa na Mama Não Densa

No segundo caso, apresenta-se uma falha na detecção da massa em uma mama não densa, durante o processo de segmentação. A Figura 5.6 mostra duas imagens, sendo a primeira, a imagem original com a localização da lesão feita pelo especialista (a), e a segunda, a imagem com as regiões segmentadas pela metodologia, juntamente com a marcação.

Se observarmos melhor esta imagem, nota-se que a lesão é bem pequena, como se fosse uma microcalcificação, portanto, para este caso a metodologia falhou em detectar a leão.

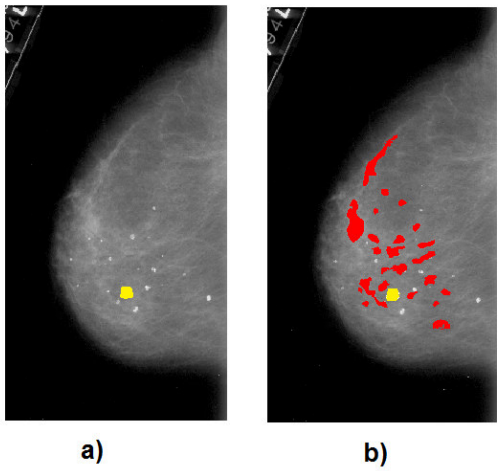

Figura 5.6: Resultado da metodologia aplicado a imagem A_1512_1.LEFT_CC: (a) Imagem original; (b) Imagem com o resultado da metodologia sobreposta a imagem original com a marcação do especialista.

5.5.3 Terceiro Caso: Sucesso na Detecção da Massa na Mama Densa

O Terceiro caso de sucesso é a imagem A_1036_1.LEFT_CC, referente a uma mama densa, que também se comportou de maneira satisfatória durante as etapas da metodologia. A Figura 5.7 mostra a imagem original com a marcação do especialista (a), as ROIs candidatas geradas pela segmentação em vermelho (b), a massa detectada pela metodologia (c) e uma imagem ilustrando a original com a marcação e a massa na cor verde sobreposta.

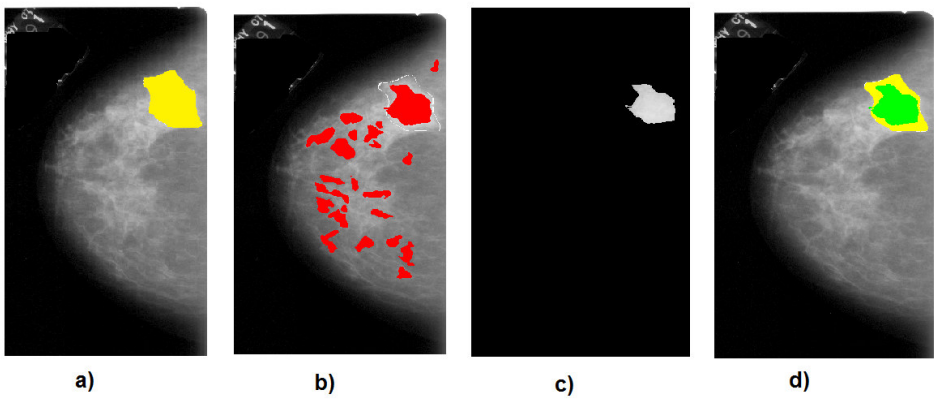

Figura 5.7: Resultado da metodologia aplicado a imagem A_1036_1.LEFT_C: (a) Imagem original com a marcação do especialista em amarelo; (b) Imagem com as regiões segmentadas em vermelho; (c) Imagem da massa detectada pela metodologia (d) Imagem com o resultado da metodologia sobreposta a imagem original com a marcação.

Analisando a região marcada com a região da massa detectada, pode-se afirmar que, a metodologia encontrou praticamente toda a estrutura da lesão em questão, pois a mesma está ocupando quase toda a região em destaque.

5.5.4 Quarto Caso: Falha na Detecção da Massa na Mama Densa

Para exemplificar um caso de falha da metodologia, apresentamos na Figura 5.8 o resultado da imagem A_1512_1.LEFT_CC, pelo fato que a massa foi perdida desde a fase da segmentação (c), pois não foi encontrado nenhuma região candidata a massa. Analisando melhor, pode-se observar na imagem original sem a marcação (a), que a região é de difícil visibilidade. Se observarmos a imagem com a marcação em amarelo (b), e olharmos a mesma região na imagem original (a), realmente veremos que a região seria a menos provável para marcarmos como lesão, mas mesmo assim, a metodologia encontrou regiões próximas a massa (c).

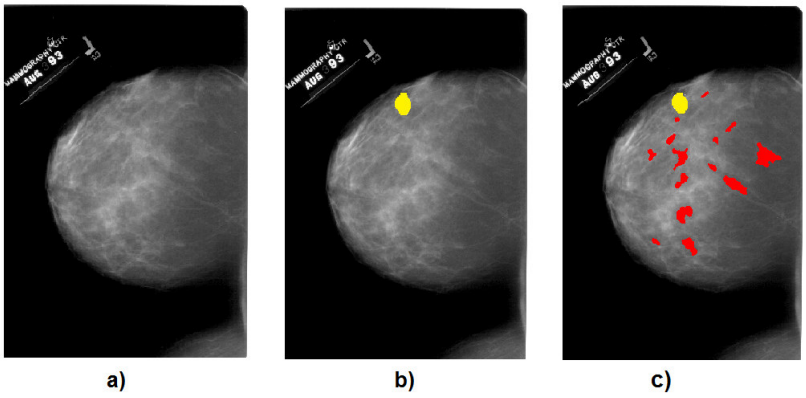

Figura 5.8: Resultado da metodologia aplicado a imagem A_1512_1.LEFT_CC: (a) Imagem original; (b) Imagem original com a marcação em amarelo (c) Imagem com o resultado da metodologia em vermelho, sobreposta a imagem original com a marcação.

5.6 Resumo dos Resultados

Em resumo, observa-se que a metodologia proposta tem um maior desempenho nas análises de mamas densas. Isto pode ser explicado devido aos parâmetros do PSO terem se adequado melhor às densas, pois todos os testes foram realizados com os mesmos parâmetros apresentados na Seção 3.2.5. Os melhores valores encontrados foram: 97,52% de sensibilidade, 92,28% de especificidade e 92,00 de acurácia.

Os índices de diversidade Funcional (FD), como descritores de textura, conseguiram caracterizar bem as massas e não massas, permitindo diferenciá-las. Nos testes realizados, a FD apresentou melhor desempenho para as mamas densas.

O *Graph Clustering* merece um destaque, pois conseguiu unir as ROIs vizinhas, melhorando as estruturas das regiões, consequentemente a classificação das massas.

De modo geral, o desempenho da metodologia, se apresentou muito bem, pois nas mamas

densas conseguiu uma taxa de acerto de 97,52% e nas mamas não densas 96,13%. Dos 621 casos, contendo 677 massas, foram encontradas com esta metodologia 649 lesões de massa, gerando uma taxa média de acerto de 95,36%.

5.7 Comparação da Metodologia com os Trabalhos Relacionados

Nesta seção é apresentado uma comparação entre os resultados dos trabalhos relacionados descritos no Capítulo 2 com os valores obtidos pela metodologia proposta. Esse comparativo está sintetizado na Tabela 5.7, onde constam as métricas de desempenhos medidas através da sensibilidade (Sen.), especificidade (Esp.), acurácia (Acu.), área sob a curva ROC, a média de falsos positivos por imagem (FP/i), área sob a curva FROC e o tamanho da amostra.

Observa-se na Tabela 5.7 que a metodologia proposta apresenta valores significativos em relação aos trabalhos relacionado. Para uma comparação bem sucedida é necessário que a base de imagens e a amostra sejam as mesma. Com base nesse preceito, comparamos esta metodologia com os trabalhos de BRAZ (2014) e SAMPAIO et al. (2011), pois foi utilizada a mesma base de imagens e a mesma quantidade de amostra, e o valor de sensibilidade ficou superior a dos referidos trabalhos citados, atingindo uma taxa de sensibilidade de 97,52%.

Outra comparação que podemos fazer é uma análise dos resultados encontrados nesta pesquisa, podendo inferir que a metodologia ocupa um lugar de destaque em relação aos trabalhos relacionados, que usam a base DDSM, pois os valores encontrados são bastante promissores.

Concluímos ainda, que em virtude dos resultados encontrados, a metodologia proposta é promissora e atingiu seu objetivo de detectar automática massas em imagens mamográficas.

Tabela 5.7: Comparação dos resultados encontrados pela metodologia proposta com os valores dos trabalhos relacionados. As métricas de comparação de desempenho são medidas por: Sensibilidade (Sen.); Especificidade (Esp.); Acurácia (Acu.); curva ROC (ROC); área sob a curva FROC (FROC); e falsos positivos por imagem (FP/i).

Trabalho	Base	Amostra	Sen.	Esp.	Acu	ROC	FP/i	FROC
NUNES (2009)	DDSM	650	83,24	84,14	83,94	—	0,55	—
IMAGECHECKER. (2011)	Privada	—	88,00	—	—	—	—	—
(BAJGER et al., 2010)	Privada	36	—	—	—	0,90	—	—
	DDSM	48	—	—	—	0,96	—	—
HU et al. (2011)	MIAS	170	91,30	—	—	—	0,71	—
LIU et al. (2011)	DDSM	125	76,80	—	—	—	1,36	—
SAMPAIO et al. (2011)	DDSM	623	80,00	—	—	—	0,84	—
AL MUTAZ et al. (2011a)	DDSM	120	91,67	84,17	—	—	—	—
AL MUTAZ et al. (2011b)	DDSM	60	—	—	95,85	—	—	—
ERICEIRA et al. (2010)	DDSM	620	100,00	95,34	96,0	—	—	—
MOREIRA et al. (2013)	MIAS	74	85,00	—	—	—	6,67	—
BRAZ (2014)	MIAS	74	97,30	—	—	—	0,33	—0,89
	DDSM	621	91,63	—	—	—	0,01	0,86
DONG et al. (2015)	DDSM	200	94,78	91,76	93,24	0,95	—	—
MARTINS et al. (2015)	DDSM	600	98,60	98,85	98,88	—	—	—
SAMPAIO et al. (2015) Não Densas	DDSM	1049	94,02	82,28	84,08	—	0,85	1,13
SAMPAIO et al. (2015) Densas	DDSM	678	89,13	88,61	88,69	—	0,71	1,47
Metodologia proposta (Não Densas)	DDSM	388	96,13	91,17	93,52	—	0,64	0,98
Metodologia proposta (Densas)	DDSM	233	97,52	92,28	94,82	—	0,38	0,98
Metodologia proposta (Densas e Não Densas)	DDSM	621	95,36	89,00	92,00	—	0,75	0.98

Capítulo 6

Conclusão

Concluímos este trabalho, afirmando que a metodologia proposta atingiu seu objetivo de detectar massas automaticamente (com a precisão de 97,52) em mamografias digitais, usando *particle swarm optimization* (PSO) e índices de diversidade funcional.

Esta metodologia é adequada para o banco de imagens DDSM, na busca de detectar automaticamente regiões que possivelmente contenham massas. Este trabalho, também pode ser utilizado com outras bases, porém, os parâmetros devem ser modificados para atendê-las de forma satisfatória.

A segmentação possui vários parâmetros, especialmente os do PSO, que é a principal técnica utilizada na metodologia, além de outros que, de forma empírica foram determinados. Nessa etapa os resultados foram bastante promissores, mas foram encontrados alguns problemas, como a geração de muitos falsos positivos, além de uma grande quantidade de ROIs pré-candidatas, por esse motivo foram realizadas as etapas de redução de falsos positivos para reduzir a quantidade de não massas.

Essa redução se deve em especial ao *Graph Clustering*, pois foi através dele que se conseguiu unir as ROIs vizinhas e descartar boa parte das não massas, sem que a perda comprometesse a pesquisa, atingindo uma taxa de 73,19% de redução no melhor caso, e uma taxa de acerto de 96,65% para mamas não densas.

Os melhores achados deste trabalho, foram obtidos utilizando as mamas densas com os índices de diversidade funcional, juntamente com o balanceamento da base utilizando SMOTE. Os valores encontrados foram, 97,52% de sensibilidades, 92,28% de especificidade, 94,82% de acurácia, 0,38 falsos positivos por imagem e 0,98 de área sob a curva FROC.

Comparando os resultados desta metodologia, com os trabalho de BRAZ (2014) e SAMPAIO

et al. (2011), deve-se ressaltar que todos os valor de sensibilidade, especificidade e acurácia foram melhores utilizando a base DDSM. Analisando os demais trabalhos, verifica-se que a metodologia proposta conseguiu superar grande parte deles, o que mostra a sua eficiência.

Por fim, a metodologia proposta pode auxiliar com segurança uma ferramenta de detecção auxiliada por computador, proporcionando ao especialista uma segunda opinião na detecção precoce do câncer de mama.

6.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se melhorar esta metodologia proposta em alguns pontos específicos.

Primeiramente na segmentação, ajustando os parâmetros do PSO para detectar melhora as regiões suspeitas da mama. Incluir novas técnicas de redução de falsos positivos, aumentando as taxas de acertos encontradas.

Melhorar a etapa de extração das características, utilizando novas medidas de textura e forma.

Adaptar a metodologia proposta para gerar os modelos de treinamento e teste, através do algoritmo genético, proporcionando modelos mais robustos que possam generalizar e predizer melhores resultados, a partir de novos casos de entrada.

Estender este trabalho para o diagnóstico da lesão quanto à malignidade.

Testar a metodologia proposta com a base de imagens MIAS, para comparar os resultados com a DDSM, no intuito de investigar para qual das bases se adéqua melhor.

6.2 Trabalho Publicado

Como resultado desta pesquisa, foi apresentado o seguinte trabalho:

- Congresso (autor)

  Neto, Otilio Paulo S.; Carvalho, Oseas; Sampaio, Wener; Corrêa, Aristófanes; Paiva, Anselmo. "Automatic segmentation of masses in digital mammograms using particle swarm optimization and graph clustering." Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on. IEEE, 2015.

Referências

ACR. (2003). of american college of r. bi-rads: Atlas - mammography. reston, va: Preston white drive, american college of radiology.

AL MUTAZ, M. A., DRESS, S., e ZAKI, N. (2011a). Detection of masses in digital mammogram using second order statistics and artificial neural network. *International Journal of Computer Science & Information Technology (IJCSIT)*, 3(3):176–186.

AL MUTAZ, M. A., DRESS, S., e ZAKI, N. (2011b). Masses detection in digital mammogram by gray level reduction using texture coding method. *jip*, 1:1.

BAJGER, M., MA, F., WILLIAMS, S., e BOTTEMA, M. (2010). Mammographic mass detection with statistical region merging. In *Digital Image Computing: Techniques and Applications (DICTA), 2010 International Conference on*, páginas 27–32. IEEE.

BOSE, J. S. C., KARNAN, M., e SIVAKUMAR, R. (2010). Detection of masses in digital mammograms. *International Journal of Computer and Network Security.*, páginas v. 2, n:2, p 78.

BOTTA-DUKÁT, Z. (2005). Rao's quadratic entropy as a measure of functional diversity based on multiple traits. *Journal of vegetation science*, 16(5):533–540.

BOYLE, P., LEVIN, B., et al. (2008). *World cancer report 2008.* IARC Press, International Agency for Research on Cancer.

BRAZ, J. G. (2014). *Detecção de regiões de massas em mamografias usando índices de diversidade, geoestatística e geometria côncava.* PhD thesis, Universidade Federal do Maranhão, Programa de Pós-Graduação em Engenharia de Eletricidade. São Luis - MA.

BUSHBERG, J. T. e BOONE, J. M. (2011). The essential physics of medical imaging.

CHAWLA, N. V., BOWYER, K. W., HALL, L. O., e KEGELMEYER, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, páginas 321–357.

CIANCIARUSO, M. V., SILVA, I. A., e BATALHA, M. A. (2009). Diversidades filogenética e funcional: novas abordagens para a ecologia de comunidades. *Biota Neotropica*, 9(3):93–103.

CORNELISSEN, J., LAVOREL, S., GARNIER, E., DIAZ, S., BUCHMANN, N., GURVICH, D., REICH, P., TER STEEGE, H., MORGAN, H., VAN Der HEIJDEN, M., et al. (2003). A handbook of protocols for standardised and easy measurement of plant functional traits worldwide. *Australian journal of Botany*, 51(4):335–380.

DONG, M., LU, X., MA, Y., GUO, Y., MA, Y., e WANG, K. (2015). An efficient approach for automated mass segmentation and classification in mammograms. *Journal of digital imaging*, páginas 1–13.

DONGEN, V. e STIJN, M. (2001). Graph clustering by flow simulation. Disponivel em: <http://dspace.library.uu.nl/bitstream/handle/1874/848/full.pdf?sequence=1>. Acessado em: 10/11/2015.

ENGELAND S. VAN, K. N. (2007). Combining two mammographic projections in a computer aided mass detection method. *Medical Physics*, 34(3):898–905.

ERICEIRA, D., CORRÊA SILVA, A., e CARDOSO DE PAIVA, A. (2010). Detecção de regiões suspeitas em mamografias digitais utilizando descrição espacial com variograma cruzado. In *XII Congresso Brasileiro de Informática em Saúde–CBIS2010*.

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. (2007). Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, 356(14):1399–1409.

GAO, X., WANG, Y., LI, X., e TAO, D. (2010). On combining morphological component analysis and concentric morphology model for mammographic mass detection. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):266–273.

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

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

GONZALEZ, R. C. e WOODS, R. E. (2007). *Digital image processing.* 3a ed. Prentice Hall Upper Saddle River, NJ.

HE, H., GARCIA, E., et al. (2009). Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284.

HEATH, M., BOWYER, K., KOPANS, D., MOORE, R., e KEGELMEYER, P. (2000). The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, páginas 212–218. Citeseer.

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

HEEMSBERGEN, D., BERG, M., LOREAU, M., VAN Hal, J., FABER, J., e VERHOEF, H. (2004). Biodiversity effects on soil processes explained by interspecific functional dissimilarity. *Science*, 306(5698):1019–1020.

HIROSE, S., SHIMIZU, K., KANAI, S., KURODA, Y., e NOGUCHI, T. (2007). Poodle-l: a two-level svm prediction system for reliably predicting long disordered regions. página 2046–2053.

HOLOGIC (2011). (2011) R2 imagechecker® digital cad. Disponivel em: 2011. <http://www.hologic.com/pt/breast-screening/imagechecker/>. Acessado em 20/07/2011.

HU, K., GAO, X., e LI, F. (2011). Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms. *Instrumentation and Measurement, IEEE Transactions on*, 60(2):462–472.

IMAGECHECKER. (2011). (2011) r2 imagechecker® cad algorithm. Disponivel em: <http://www.hologic.com/pt/breast-screening/imagechecker/>. Acessado em 20/07/2011.

INCA, D. S. J. A. G. (2015). Instituto Nacional de Câncer José Alencar Gomes da Silva. Disponível em: <www2.inca.gov.br>. Acessado em: 05/11/2015.

KE, L.; MU, N. K. Y. (2010). Mass computer-aided diagnosis method in mammogram based on texture features. *In: IEEE. 3rd International Conference on Biomedical Engineering and Informatics (BMEI). Yantai, China, 2010.*, páginas v. 1, p. 354–357.

KENNEDY, J. (2010). Particle swarm optimization. In *Encyclopedia of Machine Learning*, páginas 760–766. Springer.

KOPANS, D. B. (2007). *Breast imaging*. Lippincott Williams & Wilkins.

LIU, X., XU, X., LIU, J., e FENG, Z. (2011). A new automatic method for mass detection in mammography with false positives reduction by supported vector machine. In *Biomedical Engineering and Informatics (BMEI), 2011 4th International Conference on*, volume 1, páginas 33–37. IEEE.

LORENA, A. C. e CARVALHO, A. P. (2007). Uma introducão as support vector machines. *Revista de Informatica Teorica e Aplicada*, 14(2):43–67.

MARTINS, L. d. O., SÉRVULO, F. S., CARVALHO Filho, A. O. d., SILVA, A. C., PAIVA, A. C. d., e GATTASS, M. (2015). Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and svm. *Computers in biology and medicine*, 57:42–53.

MARTINS, L. d. O., SILVA, A. C., PAIVA, A. C. d., e GATTASS, M. (2009). Detection of breast masses in mammogram images using growing neural gas algorithm and ripley's k function. *Journal of Signal Processing Systems*, 55(1-3):77–90.

MCPHERSON, K., STEEL, C., e DIXON, J. (2000). Breast cancer—epidemiology, risk factors, and genetics. *Bmj*, 321(7261):624–628.

MEERSMAN, D., SCHEUNDERS, P., e DYCK, D. V. (1998). Detection of microcalcifications using non-linear filtering. in: Signal processing ix: theories and applications: proceedings of eusipco. *Rhodes,Greece: Typorama Publications*, 4(1):2465–2468.

MERWE, D. V. d. e ENGELBRECHT, A. P. (2003). Data clustering using particle swarm optimization. In *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 1, páginas 215–220. IEEE.

MOREIRA, A. d. S., BRAZ, G. J., ROCHA, S. V., SILVA, A. C., e PAIVA, A. C. (2013). Detecção de massas em imagens da mama usando índices de diversidade e algoritmos de segmentação em grafo. *Cad. Pesq., São Luís-MA.*, páginas v. 20, n. especial, julho 2013.

NUNES, A. P. (2009). Detecção de massas em imagens mamográficas usando índice de diversidade de simpson e máquina de vetores de suporte. Dissertação de mestrado, Departamento de Engenharia Elétrica - Universidade Federal do Maranhão, São Luís-MA.

OTSU, N. (1975). A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27.

PAL, N. R. e PAL, S. K. (1993). A review on image segmentation techniques. *Pattern recognition*, 26(9):1277–1294.

PEDRINE, H. e SCHWARTZ, W. R. (2008). *William Robson. Analise de Imagens Digitais: Principios, Algoritmos e Aplicações.* Thomson Learnig, São Paulo.

PETCHEY, O. L. e GASTON, K. J. (2006). Functional diversity: back to basics and looking forward. *Ecology letters*, 9(6):741–758.

QIAN, W., SONG D. LEI, M., R., S., e EIKMAN, E. (2007). Computeraided mass detection based on ipsilateral multiview mammograms. *Academic radiology.*, páginas v. 14, n. 5, p. 530–538.

RAO, C. R. (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43.

REYNOLDS, R. J. e SNAPP, B. R. (1986). The competitive effects of partial equity interests and joint ventures. *International Journal of Industrial Organization*, 4(2):141–153.

RICOTTA, C. (2005). Through the jungle of biological diversity. *Acta biotheoretica*, 53(1):29–38.

SAÚDE-MEDICINA (2015). Processo da metástase. Disponivel em: <http://www.saudemedicina.com/metastase/>. Acessado em 30/01/2015.

SAMPAIO, W. B., DINIZ, E. M., Silva, A. C., DE PAIVA, A. C., e GATTASS, M. (2011). Detection of masses in mammogram images using cnn, geostatistic functions and svm. *Computers in biology and medicine*, 41(8):653–664.

SAMPAIO, W. B. d., SILVA, A. C., DE PAIVA, A. C., e GATTASS, M. (2015). Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, lbp and svm. *Expert Systems with Applications*, 42(22):8911–8928.

SCHAEFFER, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1):27 – 64.

SCHÖLKOPF, B. e SMOLA, A. J. (2001). Learning with kernels: Support vector machines, regularization, optimization, and beyond (adaptive computation and machine learning). the mit press, december 2001.

SOCIETY, A. A. C. (2013). Learn about breast cancer.

TILMAN, D. (2001). Functional diversity. *Encyclopedia of biodiversity*, 3(1):109–120.

TUCERYAN, M. e JAIN, A. (1998). Texture analysis. the handbook of pattern recognition and computer vision. *River Edge.*

TZIKOPOULOS, S.; MAVROFORAKIS, M. G. H. D. N. T. S. (2011). A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry. *Academic Radiology, Elsevier.*, páginas v. 102, n. 1, p. 47–63.

VAN DEN BERGH, F. (2006). *An analysis of particle swarm optimizers.* PhD thesis, University of Pretoria.

VAPNIK, V. N. (1998). Statistical learning theory, volume 2. wiley new york.

VIOLLE, C., NAVAS, M.-L., VILE, D., KAZAKOU, E., FORTUNEL, C., HUMMEL, I., e GARNIER, E. (2007). Let the concept of trait be functional! *Oikos*, 116(5):882–892.

WANG, R., LEE, N., e WEI, Y. (2015). A case study: Improve classification of rare events with sas® enterprise miner™. Disponivel em: <http://support.sas.com/resources/papers/proceedings15/3282-2015.pdf>. Acessado em: 08/10/2015.

WANG, X., LI, L., XU, W., LIU, W., LEDERMAN, D., e ZHENG, B. (2012). Improving performance of computer-aided detection of masses by incorporating bilateral mammographic density asymmetry: an assessment. *Computer Methods and Programs in Biomedicine, Elsevier.*, páginas v. 19, n. 3, p. 303–310.

WEI, J., CHAN, H.-P., ZHOU, C., WU, Y.-T., SAHINER, B., HADJIISKI, L. M., ROUBIDOUX, M. A., e HELVIE, M. A. (2011). Computer-aided detection of breast masses: Four-view strategy for screening mammography. *Medical physics.*, páginas v. 38, p. 1867.

WEI, K., ZHANG, T., SHEN, X., e LIU, J. (2007). An improved threshold selection algorithm based on particle swarm optimization for image segmentation. In *Natural Computation, 2007. ICNC 2007. Third International Conference on*, volume 5, páginas 591–594. IEEE.

WINTER, M., DEVICTOR, V., e SCHWEIGER, O. (2013). Phylogenetic diversity and nature conservation: where are we? *Trends in Ecology & Evolution*, 28(4):199–204.

WU, Y., WEI, J., HADJIISKI, L., SAHINER, B., ZHOU, C., GE, J., SHI, J., ZHANG, Y., e CHAN, H. (2007). Bilateral analysis based false positive reduction for computer-aided mass detection. *Medical physics, NIH Public Access.*, páginas v. 34, n. 8, p. 3334.

ZUIDERVELD, K. (1994). Contrast limited adaptive histogram equalization. In *Graphics gems IV*, páginas 474–485. Academic Press Professional, Inc.