

THIAGO ALVES ELIAS DA SILVA

UMA METODOLOGIA DE AUXÍLIO AO DIAGNÓSTICO DE DOENÇAS DE
MAMA A PARTIR DE TERMOGRAFIAS DINÂMICAS

Tese apresentada ao Programa de
Pós-Graduação em Computação da
Universidade Federal Fluminense,
como requisito parcial para obtenção
do Grau de Doutor. Área de
Concentração: Computação Visual.

Orientadora: Prof^ª. Dr^ª. Aura Conci

Niterói
2015

THIAGO ALVES ELIAS DA SILVA

UMA METODOLOGIA DE AUXÍLIO AO DIAGNÓSTICO DE DOENÇAS DE
MAMA A PARTIR DE TERMOGRAFIAS DINÂMICAS

Tese apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Doutor. Área de Concentração: Computação Visual.

Aprovado em novembro de 2015.

BANCA EXAMINADORA

Prof^a. D.Sc. Aura Conci, UFF

Prof^a. D.Sc. Débora Christina Muchaluat Saade, UFF

Prof. D.Sc. Anselmo Antunes Montenegro, UFF

Prof. D.Sc. Anselmo Cardoso de Paiva, UFMA

Prof. D.Sc. Reinaldo Rodriguez Ramos, Universidad de la Habana

Prof. D.Sc. Alair Augusto Sarment Moreira Damas dos Santos, UFF

Niterói
2015

*Dedico ao meu filho Arthur e minha amada esposa Alana
Vale Elias.*

AGRADECIMENTOS

Inicialmente, agradeço a Deus por me permitir trilhar esse caminho e superar os obstáculos que surgiram ao longo desta caminhada. Também agradeço a minha orientadora, Prof^a. Dr^a. Aura Conci, pelos ensinamentos e, muitas vezes, pela paciência. Sem ela, a conclusão deste trabalho não seria possível. Ao Instituto Federal de Educação, Ciência e Tecnologia do Piauí que, através do DINTER firmado com a UFF, proporcionou que eu realizasse o sonho de concluir o doutorado. Aos médicos do Hospital Universitário Antônio Pedro, Dr. Alair Augusto Sarmet M. D. dos Santos, Dr. Renato Bravo, Dr^a Cristina Asvolinsque Pantaleão Fontes, Dr. Amandio Roberto Pereira Henrique, Dr^a Maria Lúcia de Oliveira Santos e Dr. Alberto Domingues Vianna, que contribuíram para o projeto desenvolvido no Visual Lab e que deu origem a esta tese. Agradeço aos meus novos e antigos amigos de doutorado, em especial Tiago Bonini, que compartilhou seu conhecimento nos meus primeiros dias no Visual Lab; ao Lincoln que, com sua calma, foi um dos responsáveis pela aquisição das imagens utilizadas nesta tese; aos amigos Stephenson, Giomar, Rafael Marques e Francieric pelas preciosas dicas de programação; aos amigos Rogério Batista e Ritomar pelas conversas descontraídas durante os estágios presenciais na UFF; além de todos os demais colegas da UFF e do IFPI que, direta ou indiretamente, contribuíram para que eu concluísse este trabalho. Agradeço aos meus pais, Edvaldo e Lúcia, e minhas irmãs, Thaline e Thaís, pelo amor incondicional e pelos incentivos e orações nos momentos mais difíceis; a Fátima e ao meu afilhado Lucas pelo apoio; a minha amada esposa Alana Vale Elias que esteve sempre ao meu lado e compreendeu os meus momentos de angústia. Por fim, agradeço, principalmente, ao meu filho Arthur que, na inocência do seu primeiro aninho de vida, foi o meu maior incentivador.

RESUMO

O câncer de mama é o que mais acomete as mulheres em todo o mundo. Assim, técnicas de rastreamento estão em constante desenvolvimento a fim de detectar a doença em estágios cada vez mais iniciais. A análise de imagens de termografia dinâmica surge como mais uma ferramenta no auxílio ao diagnóstico de doenças da mama. Nesta tese, foram utilizadas imagens térmicas capturadas pelo protocolo dinâmico onde, após o resfriamento das mamas induzido por uma corrente de ar, 20 imagens sequenciais com intervalos de 15 segundos entre elas foram capturadas durante o processo de retorno ao equilíbrio térmico com o ambiente da paciente. As imagens estão armazenadas no Database for Mastology Research with Infrared Image - DMR-IR, acessível no endereço eletrônico <http://visual.ic.uff.br/dmi>. Após a obtenção das imagens térmicas sequenciais de cada paciente, foram extraídas as matrizes de temperatura e novas imagens quantizadas em escala de cinza foram geradas. Em seguida, as regiões de interesse (Regions of Interest - ROIs) das imagens em escala de cinza foram segmentadas e, a partir delas, também as ROIs das matrizes de temperatura foram obtidas. Destas, foram extraídas características diversas (estatística, baseadas em agrupamento, comparação de histogramas, geometria fractal, índices de diversidade e estatística espacial). Destas características, foram geradas séries temporais que foram decompostas em 9 subséries de diferentes cardinalidades. As melhores características foram selecionadas e classificadas pela Máquina de Vetores de Suporte (MVS), utilizando-se o método de validação cruzada *Leave-One-Out*. Obteve-se para uma amostra de 64 mamas (32 saudáveis e 32 com alguma anomalia), uma acurácia de 100%, selecionando-se 6 características. Foram realizados testes com 12 novas mamas que não participaram da etapa de desenvolvimento da metodologia. Com elas, também o método alcançou acurácia de 100%. Ao final, foi proposta uma metodologia de sistema de auxílio ao diagnóstico por computador (CAD) baseado em termografias sequenciais.

Palavras-chave: Termografia dinâmica, CAD, Câncer de mama, detecção de doenças de mama.

ABSTRACT

Breast cancer is the one that most affects women worldwide. Thus, disease screening techniques are constantly being developed to detect the disease at early stages. The dynamic thermography image analysis emerges as one tool to aid in the diagnosis of breast diseases. Thermal images captured by the dynamic protocol were used in this thesis where, after the cooling of the breasts by air stream, 20 sequential images with intervals of 15 seconds between them were taken during the process of returning the patient's body to thermal equilibrium with the environment. The images are stored in the Database for Research Mastology with Infrared Image - DMR-IR, accessible on the website <http://visual.ic.uff.br/dmi>. After obtaining the sequential thermal images of each patient, their temperature matrices were extracted and new grayscale images were generated. Then the regions of interest (ROIs) of the images in grayscale were segmented and from them also the ROIs of the temperature matrices were determined. Features were extracted from the ROIs (they are based on statistical, on clustering, on histogram comparison, on fractal geometry, on diversity indices and on spatial statistics). Time series which were broken down into 9 subsets of different cardinalities were generated from these features. The best features were selected and classified by the Support Vector Machine (SVM), using the Leave-One-Out Cross-Validation method. An accuracy of 100% was obtained for a sample of 64 breast (composed of 32 healthy and 32 with some abnormality). Tests were performed with 12 new breasts that did not participate in the methodology development stage. The method also achieved 100% accuracy with them. At the end a methodology of computer-aided diagnosis (CAD) system was proposed based on sequential thermography.

Keywords: Dynamic thermography, CAD, breast cancer, detection of breast diseases.

LISTA DE ILUSTRAÇÕES

1.1 - Distribuição proporcional dos dez tipos de câncer mais incidentes estimados para 2014, e válidos para 2015. Fonte: INCA (2014)	14
1.2 - Termografia de paciente com mama esquerda doente.....	17
3.1 - Representação da vizinhança do pixel	38
4.1 - Resumo da metodologia	50
4.2 - Captura de 20 quadros sequenciais	51
4.3 - Exemplo de extração da matriz de temperatura	52
4.4 - Conversão da matriz de temperatura para imagem em escala de cinza	53
4.5 - Exemplo de segmentação realizada pelo segmentador de Marques (2012).....	54
4.6 - Limites superior e inferior da mama. Fonte: Netter (2011).....	55
4.7 - Exemplo de mamas segmentadas manualmente.....	56
4.8 - Determinação da ROI da matriz de temperatura a partir da ROI em escala de cinza.....	56
4.9 - Representação da extração de características. (a) uma característica por ROI; (b) baseada em comparação entre ROIs.....	57
4.10 - Séries temporais das médias de temperatura de algumas mamas doentes e saudáveis	58
4.11 - ROI em tons de cinza com pixels agrupados em nove grupos.....	60
4.12 - Extração das características de textura a partir dos histogramas da soma e da diferença das ROIs em escala de cinza de uma dada mama.	61
4.13 - Extração dos índices de diversidade a partir dos histogramas da soma e da diferença das ROIs de uma dada mama.	63

4.14 - Exemplo de janela considerada no cálculo da lacunaridade em azul e janela não considerada no cálculo da lacunaridade em vermelho.	65
4.15 - Séries temporais de duas mamas (uma sadia e outra doente) com diferentes cardinalidades.	68
4.16 - Séries temporais do grupo baseado em comparação de histogramas. Exemplo com duas mamas.	71
4.17 - Exemplo de curva ROC. Fonte: Borchardt (2013)	76
5.1 - Maiores acurácias por grupo de características.....	78
5.2 - Quantidade vezes que cada característica foi selecionada dentre as diferente cardinalidades das séries temporais.....	80
5.3 - Resumo da proposta de metodologia de CAD	90

LISTA DE TABELAS

1.1 - Recomendações de exame de rastreio do câncer de mama por faixa etária. Fonte: Adaptado de Unbar et al. (2012).....	16
2.1 - Resumo comparativo entre os trabalhos	28
3.1 - Equivalência de características entre matriz de co-ocorrência e histogramas da soma e da diferença. Fonte: Unser (1986).....	39
3.2 - Valores do índice de diversidade de McIntosh mínimo e máximo para população de 100 indivíduos. Fonte: McIntosh (1967).....	42
4.1 - Cardinalidade das séries temporais geradas para cada característica	67
4.2 - Relação entre as duas etapas de extração de características.	72
5.1 - Resultado da seleção de características para cada tipo das séries.....	79
5.2 - Resultado da classificação para cada uma das séries após a seleção de características.....	81
5.3 - Resultado da classificação de 12 novos casos utilizando a metodologia proposta.....	85
5.4 - Comparação com outros trabalhos.....	86
A.1 - Mamas com anomalias.....	101
A.2 - Mamas saudáveis.....	103
B.1 - Resultado da classificação a partir das características das séries temporais formadas pela média e pelo desvio padrão das ROIs sequenciais das matrizes de temperatura.	105
B.2 - Resultado da classificação a partir das características das séries temporais formadas pelo valor do maior centroide das ROIs sequenciais das matrizes de temperatura.	106
B.3 - Resultado da classificação a partir das características das séries temporais formadas pelas comparações de histogramas das ROIs sequenciais das imagens em escala de cinza.	107

B.4 - Resultado da classificação a partir das características das séries temporais formadas pelos descritores de textura horizontal das ROIs sequenciais das imagens em escala de cinza.	108
B.5 - Resultado da classificação a partir das características das séries temporais formadas pelos descritores de textura vertical das ROIs sequenciais das imagens em escala de cinza.	109
B.6 - Resultado da classificação a partir das características das séries temporais formadas pelos índices de diversidade horizontais das ROIs sequenciais das imagens em escala de cinza.	110
B.7 - Resultado da classificação a partir das características das séries temporais formadas pelos índices de diversidade verticais das ROIs sequenciais das imagens em escala de cinza.	111
B.8 - Resultado da classificação a partir das características das séries temporais formadas pelas comparações de spatiogram's das ROIs sequenciais das imagens em escala de cinza.	112
B.9 - Resultado da classificação a partir das características das séries temporais formadas pelos valores da lacunaridade das ROIs sequenciais das imagens em escala de cinza.	113
B.10 - Resultado da classificação a partir das características das séries temporais formadas pelos valores de todas as características das ROIs sequenciais.....	114

LISTA DE ABREVIATURAS E SIGLAS

CAD	:	Computer-Aided Diagnosis;
DMR-IR	:	Database for Mastology Research with Infrared Image;
HUAP	:	Hospital Universitário Antônio Pedro;
INCA	:	Instituto Nacional de Câncer;
MVS	:	Máquina de Vetores de Suporte;
ROC	:	Receiver Operating Characteristic;
ROI	:	Region of Interest;
WEKA	:	Waikato Environment for Knowledge Analysis.

SUMÁRIO

CAPÍTULO 1 - INTRODUÇÃO.....	14
1.1 Motivação.....	18
1.2 Objetivos.....	18
1.3 Contribuição.....	18
1.4 Organização da Tese.....	19
CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA.....	20
CAPÍTULO 3 - REFERENCIAL TEÓRICO.....	31
3.1 Termografia Mamária.....	31
3.1.1 Termografia Dinâmica.....	33
3.2 Diagnóstico Auxiliado pro Computador.....	33
3.3 Características de Imagens.....	34
3.3.1 Estatística Simples.....	35
3.3.2 Medidas Baseadas em Histogramas.....	35
3.3.2.1 Comparação de Histogramas.....	35
3.3.2.1.1 Interseção.....	36
3.3.2.1.2 Qi-quadrado.....	36
3.3.2.1.3 Distância de Bhattacharyya.....	37
3.3.2.2 Descritores de Textura de Unser.....	37
3.3.3 Índices de Diversidade Ecológica.....	40
3.3.3.1 Índice de Simpson.....	40
3.3.3.2 Índice de Berger-Parker.....	41
3.3.3.3 Índice de McIntosh.....	41
3.3.4 Medida Baseada em Comparação de Spatiograms.....	42
3.3.5 Geometria Fractal - Lacunaridade.....	44
3.4 Classificação.....	46
CAPÍTULO 4 - METODOLOGIA.....	49
4.1 Seleção dos Dados Térmicos do Banco de Dados.....	49
4.2 Extração das Matrizes de Temperatura.....	52
4.3 Geração de Imagens em Escala de Cinza.....	52
4.4 Segmentação das Regiões de Interesse.....	53
4.5 Extração de Características.....	56
4.5.1 Etapa de Extração de Características das ROIs.....	58

4.5.1.1 Características Baseadas em Estatística Simples.....	59
4.5.1.2 Característica Baseada em Agrupamento.....	59
4.5.1.3 Características Baseadas em Comparação de Histogramas.....	60
4.5.1.4 Características Baseadas em Textura.....	61
4.5.1.5 Características Baseadas em Índices de Diversidade.....	62
4.5.1.6 Característica Baseada em Estatística Espacial.....	64
4.5.1.7 Característica Baseada em Lacunaridade.....	64
4.5.2 Geração das Séries Temporais.....	65
4.5.3 Extração de Características das Séries Temporais.....	69
4.6 Seleção de Características.....	71
4.7 Classificação.....	73
4.8 Formas de Avaliação dos Resultados.....	74
CAPÍTULO 5 - RESULTADOS E DISCUSSÕES.....	77
5.1 Resultados por Grupo de Características.....	77
5.2 Seleção de Características.....	78
5.3 Resultados da Classificação Após a Seleção de Características.....	81
5.4 Discussão.....	83
5.5 Teste da Metodologia com Novos Casos.....	85
5.6 Comparação com Outros Trabalhos.....	86
5.7 Proposta de Metodologia para um Sistema de Diagnóstico Auxiliado por Computador.....	90
CAPÍTULO 6 - CONCLUSÕES.....	93
6.1 Trabalhos Futuros.....	94
REFERÊNCIAS.....	96
APÊNDICE A - DIAGNÓSTICOS DAS PACIENTES.....	101
APÊNDICE B - RESULTADOS DA CLASSIFICAÇÃO POR GRUPO DE CARACTERÍSTICAS.....	105

CAPÍTULO 1 - INTRODUÇÃO

O câncer de mama é o tipo mais comum de câncer entre as mulheres e, apesar dos avanços recentes no campo da medicina, ainda existem algumas limitações inerentes às técnicas de rastreio atualmente utilizadas (Gerasimova *et al.*, 2014).

No Brasil, a última estimativa que ocorreu em 2014, válida também para 2015, aponta para a ocorrência de aproximadamente 576 mil casos novos de câncer. O câncer de mama feminino ocupa a terceira posição dentre os casos de câncer na população brasileira, com cerca de 57 mil mulheres acometidas pela doença. Dentre as mulheres, o câncer de mama ocupa a primeira posição, seguidos pelo de cólon e reto. Na figura 1.1, pode-se verificar o ranqueamento da incidência de diferentes tipos de câncer em homens e mulheres (INCA, 2014).


Localização primária	casos	%			Localização primária	casos	%
Próstata	68.800	22,8%		Homens	Mama Feminina	57.120	20,8%
Traqueia, Brônquio e Pulmão	16.400	5,4%			Cólon e Reto	17.530	6,4%
Cólon e Reto	15.070	5,0%			Colo do Útero	15.590	5,7%
Estômago	12.870	4,3%			Traqueia, Brônquio e Pulmão	10.930	4,0%
Cavidade Oral	11.280	3,7%			Glândula Tireoide	8.050	2,9%
Esôfago	8.010	2,6%			Estômago	7.520	2,7%
Laringe	6.870	2,3%			Corpo do Útero	5.900	2,2%
Bexiga	6.750	2,2%			Ovário	5.680	2,1%
Leucemias	5.050	1,7%			Linfoma não Hodgkin	4.850	1,8%
Sistema Nervoso Central	4.960	1,6%			Leucemias	4.320	1,6%

Figura 1.1 - Distribuição proporcional dos dez tipos de câncer mais incidentes estimados para 2014, e válidos para 2015. Fonte: INCA (2014)

Ainda em relação a figura 1.1, observa-se que para os anos de 2014 e 2015, mais de 20% dos casos de câncer entre as mulheres brasileiras aparecerá nas mamas, representando um total de 57.120 novos casos.

Na esfera mundial, o câncer de mama também é o tipo de câncer que mais acomete as mulheres, tanto em países em desenvolvimento quanto em países desenvolvidos. Cerca de 1,67 milhões de casos novos dessa doença foram detectados no ano de 2012, em todo o mundo, o que representa 25% de todos os tipos de câncer diagnosticados nas mulheres. Suas taxas de incidência variam entre as diferentes regiões do mundo, com as maiores taxas em 2012 na Europa Ocidental (96/100 mil) e as menores taxas na África Central e na Ásia Oriental (27/100 mil). Alguns fatores de risco para o desenvolvimento do câncer de mama são bem conhecidos, como: envelhecimento, fatores relacionados à vida reprodutiva da mulher, história familiar de câncer de mama,

consumo de álcool, obesidade, sedentarismo e exposição à radiação ionizante (INCA, 2014).

Para um melhor prognóstico, minimizando a agressividade do tratamento instituído, e diminuição da mortalidade, o diagnóstico do câncer de mama em fase precoce assume papel decisivo. O termo prevenção primária em epidemiologia, é utilizado para designar medidas capazes de impedir o aparecimento de um agravo à saúde; já prevenção secundária, são medidas que impedem o desenvolvimento do estágio seguinte das doenças (Batiston *et al.*, 2009). Devido à sua etiopatogenia complexa e multifatorial, até o momento, não existe prevenção primária efetiva para o câncer de mama (Warner, 2011), sendo então, os esforços voltados para medidas buscando a prevenção secundária. A sobrevida da doença em estágios iniciais pode alcançar 98% em 5 anos (Damin, 2012).

O rastreo do câncer de mama tem como objetivo a detecção de tumores assintomáticos, a fim de reduzir a mortalidade pela doença e aumentar a sobrevida após o diagnóstico. Questões específicas incluem determinar quem deve ser examinado (a estratificação de risco, a idade para começar a triagem, a idade para parar) e qual método deve ser usado para a seleção. À medida que mais informações se tornem disponíveis sobre a ocorrência de resultados de diagnósticos falso-positivos e com o tratamento cada vez mais eficaz para o câncer da mama, a relação entre benefícios e malefícios de rastreo desta doença têm recebido cada vez mais atenção.

Na década de 70, realizou-se o primeiro estudo populacional prospectivo, controlado e randomizado sobre o rastreo do câncer de mama e a taxa de mortalidade. Este estudo publicado por Shapiro (1977), demonstrou uma redução de 25% na mortalidade no grupo de mulheres submetidas ao rastreamento mamográfico. Estes resultados incentivaram novos estudos e estimularam a sociedade médica a recomendar o rastreo mamográfico (Urban *et al.*, 2012).

Atualmente, uma variedade de modalidades de exames por imagem, bem como o autoexame e o exame clínico da mama têm sido desenvolvidos para o rastreo do câncer da mama. A mamografia continua a ser o principal meio de rastreo do câncer de mama; a ultrassonografia é comumente usada para complementação do diagnóstico de uma anormalidade vista na mamografia e para esclarecer características de uma lesão; a ressonância magnética (MR) no rastreamento do câncer de mama é emergente, sendo utilizada em combinação com a mamografia para pacientes de alto risco (Drukteinis *et al.*, 2013).

Urban *et al.* (2012) compilaram dados de diversos estudos para que recomendações de exames de rastreio do câncer de mama fossem apresentadas por faixa etária. Na ausência de dados probatórios, as recomendações refletiram o consenso da comissão de especialistas de três entidades: Colégio Brasileiro de Radiologia e Diagnóstico por Imagem (CBR), Sociedade Brasileira de Mastologia (SBM) e Federação Brasileira das Associações de Ginecologia e Obstetrícia (FEBRASGO). A tabela 1.1 resume as recomendações da pesquisa.

Tabela 1.1 - Recomendações de exame de rastreio do câncer de mama por faixa etária.

Fonte: Adaptado de Unbar *et al.* (2012)

Exame	Faixa etária das mulheres		
	Abaixo de 40 anos	Entre 40 e 69 anos	Acima de 70 anos
Mamografia	Apenas em mulheres com alto risco para câncer de mama.	Recomenda-se para todas as mulheres, com a periodicidade anual.	Recomenda-se para mulheres com expectativa de vida maior que 7 anos ou após um resultado anormal do rastreamento.
Ultrassonografia	Apenas em mulheres com alto risco para câncer de mama nas quais o rastreamento por ressonância magnética não pode ser realizado por qualquer razão.	Apenas em mulheres com alto risco, especialmente aquelas em que o rastreamento por ressonância magnética não pode ser realizado por qualquer razão; ou em mulheres com tecido mamário denso.	Não mencionam
Ressonância Magnética	Apenas em mulheres com alto risco para câncer de mama.	Apenas em mulheres com alto risco para câncer de mama.	Não mencionam.

Ainda com relação à tabela 1.1, observa-se que para mulheres abaixo de 40 anos, os exames de rastreio são recomendados apenas para aquelas com alto risco de câncer. Esses riscos são mutações genéticas, histórico familiar, histórico de irradiação no tórax entre os 10 e 30 anos de idade, histórico pessoal de neoplasia lobular, carcinoma invasor de mama ou de ovário, dentre outros. Para as mulheres entre 40 e 69 anos, a mamografia é recomendada com periodicidade anual. Para as mulheres com idade

acima de 70 anos, a mamografia é recomendada apenas para aquelas com expectativa de vida maior que 7 anos ou que tenham recomendação de serem submetidas a tratamentos e exames mais invasivos, no caso de um resultado anormal do rastreio.

A termografia mamária apresenta-se como mais uma ferramenta de rastreio de doenças da mama. Ela se baseia no princípio de que tecidos mamários com alguma doença comportam-se termicamente diferentes de tecidos saudáveis. A figura 1.2 apresenta a termografia de uma paciente doente. É possível perceber que a mama esquerda (doente) tem temperaturas mais elevadas que a mama direita (saudável).

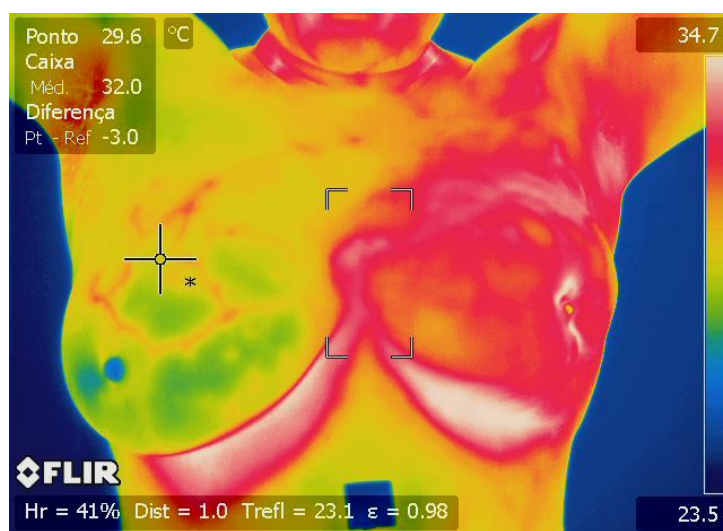


Figura 1.2 - Termografia de paciente com mama esquerda doente

Ao contrário da figura 1.2, nem sempre a diferença de temperatura entre as mamas saudáveis e doentes é tão evidente. Muitas vezes, a variação de temperatura normal é bastante sutil nos estágios iniciais da malignidade, e alterações podem passar despercebidas pela interpretação visual direta das termografias. Também mamas normais com o aumento da vascularização são geralmente mais quentes e podem ser mal interpretadas como anormais. Portanto a interpretação subjetiva de termografias muitas vezes leva a falsos diagnósticos. Por isso são necessários sistemas de detecção automática (Francis *et al.*, 2014).

O desenvolvimento de técnicas de processamento computacional de imagens térmicas apresenta-se como um ponto importante para o sucesso do rastreio de doenças na mama a partir de termografias. Além da termografia estática, como mostrada na figura 1.2, a análise de termografias dinâmicas apresenta-se como uma alternativa na tentativa de evidenciar as diferenças térmicas entre mamas saudáveis e mamas doentes. Nesta

tese, é proposta uma metodologia de auxílio à detecção de anomalias na mama a partir da análise de termografias dinâmicas, conforme detalhada nas seções seguintes.

1.1. MOTIVAÇÃO

Mesmo com os avanços dos tratamentos que reduziram a mortalidade por câncer de mama ao longo das últimas décadas, esta doença continua a ser o que mais acomete as mulheres em todo o mundo (INCA, 2014). A probabilidade de cura do câncer de mama está fortemente influenciada a sua descoberta no início da doença. A eficácia de métodos de rastreio de anomalias na mama em estágios iniciais apresenta-se como fator determinante para um tratamento adequado. O desenvolvimento de métodos que auxiliem profissionais de saúde no rastreio de doenças na mama é a principal motivação para o desenvolvimento desta tese.

A termografia mamária apresenta-se como mais uma ferramenta de auxílio ao diagnóstico de anomalias na mama. A vantagem da utilização deste método reside no fato do exame ser inofensivo, não invasivo, e não expor a paciente à radiação ionizante. Já a termografia dinâmica é usada para detectar variações de temperatura de tecidos sob um estresse térmico (Gerasimova *et al.*, 2013). Tais variações evidenciam as diferenças entre os tecidos doente e saudável (Herman, 2013). Assim, tais fatos, apresentaram-se como fator motivador para a utilização de termografias dinâmicas para o auxílio ao diagnóstico de doenças na mama.

1.2. OBJETIVOS

O objetivo principal desta tese é desenvolver uma metodologia capaz de auxiliar na detecção de doenças na mama através da identificação de padrões em termografias dinâmicas. A metodologia se baseia na análise das séries temporais geradas a partir de características extraídas das regiões de interesse das termografias, durante o processo de reaquecimento corpóreo, após um resfriamento induzido pelo protocolo de aquisição dinâmico de imagens térmicas (Silva *et al.*, 2013).

Também é proposto uma metodologia para um sistema de auxílio ao diagnóstico por computador de anomalias de mama a partir de termografias dinâmicas.

1.3. CONTRIBUIÇÃO

A principal contribuição desta tese é o desenvolvimento de uma metodologia de auxílio à detecção de anomalias na mama a partir de termografias dinâmicas. Esta metodologia

visa auxiliar os profissionais de saúde na triagem de pacientes durante o processo de rastreio de doenças da mama, ajudando, desta forma, a detecção de doenças em estágios iniciais.

1.4. ORGANIZAÇÃO DA TESE

Esta tese está organizada em seis Capítulos. No Capítulo 2 será apresentada a revisão bibliográfica com trabalhos desenvolvidos pelo grupo de pesquisa composto pelos membros do Laboratório de Computação Visual (Visual Lab) do Instituto de Ciência da Computação da Universidade Federal Fluminense. Os trabalhos desenvolvidos pelo referido grupo têm como objetivos diretos ou indiretos desenvolver metodologias que auxiliem no rastreio de doenças de mamas a partir da análise de imagens térmicas. Além disso, serão apresentados outros trabalhos encontrados na literatura também relacionados ao desenvolvimento de métodos de detecção de anomalias de mama a partir de imagens térmicas.

Em seguida, no Capítulo 3, será apresentada a fundamentação teórica necessária para o desenvolvimento da metodologia proposta nesta tese. No Capítulo 4, será discutida a metodologia proposta nesta tese para a detecção de anomalias de mama a partir da análise de termografias dinâmicas. No Capítulo 5, serão apresentados e discutidos os resultados alcançados pela metodologia proposta. Finalmente, no Capítulo 6, serão apresentadas as conclusões e trabalhos futuros que poderão ser desenvolvidos a partir da metodologia e resultados alcançados por esta tese.

CAPÍTULO 2 - REVISÃO BIBLIOGRÁFICA

Esta tese é parte de um projeto desenvolvido pelo grupo de pesquisa que compõe o Laboratório de Computação Visual (Visual Lab) do Instituto de Ciência da Computação da Universidade Federal Fluminense (<http://visual.ic.uff.br>). O projeto baseia-se no processamento e análise de imagens médicas, e tem como um dos objetivos auxiliar o diagnóstico de doenças de mama a partir de termografias. Diversos trabalhos já foram desenvolvidos pelo grupo com o objetivo direto ou indireto de auxiliar o diagnóstico de doenças de mama através de termografias. Nos parágrafos seguintes serão apresentados tais trabalhos. São eles: Serrano *et al.* (2010), Motta *et al.* (2010), Resmini *et al.* (2012), Marques *et al.* (2013), Borchart (2013), Galvão (2015), Oliveira *et al.* (2015), Araújo *et al.* (2015) e Silva (2015). Em seguida, serão discutidos outros trabalhos encontrados na literatura que propõem metodologias de auxílio à detecção de anomalias de mama a partir de termografias, independente do protocolo de aquisição utilizado. Serão discutidas, quando possível, as metodologias adotadas, resumindo o procedimento de extração e análise das características, as técnicas utilizadas de classificação das mamas e os resultados e conclusões alcançadas.

Serrano *et al.* (2010) apresentou um estudo para o diagnóstico de doenças da mama em fase inicial (cistos, fibroadenomas, tumores, etc.). O trabalho foi baseado na análise de assimetria entre as mamas da paciente usando imagens térmicas. Inicialmente, cada imagem térmica das mamas foi segmentada, utilizando uma região quadrada de mesmo tamanho. O coeficiente de Hurst foi utilizado para caracterizar a textura da mama. Para o cálculo, uma janela móvel percorreu a imagem pixel a pixel. Foram utilizadas janelas quadradas de lados w igual a 5, 7, 9, 11, 13 e 15 pixels. Os valores dos coeficientes de Hurst para cada janela foram usados para formar um vetor de característica. Este procedimento foi realizado de duas maneiras: primeiro, utilizando a imagem de cada mama e, em seguida, usando uma nova imagem obtida pela diferença de temperatura nas posições simétricas verticalmente de cada mama direita e esquerda. A classificação das mamas saudáveis e doentes foi realizada e o melhor resultado foi obtido pela Classificação Via Regressão (CVR), obtendo a área sob a curva ROC (Receiver Operating Characteristic) de 0,95.

Motta *et al.* (2010) propuseram um método automático para a segmentação das regiões de interesse (ROI) em termogramas frontais da mama. Na metodologia, as seguintes

técnicas de processamento de imagens foram utilizadas: morfologia matemática, detecção de contornos de Canny e técnica de limiarização adaptativa em imagens em tons de cinza.

Resmini *et al.* (2012), após segmentarem as ROIs das termografias e dividi-las em quatro quadrantes, extraíram características a fim de classificá-la como doentes ou saudas. A classificação é feita por paciente e não por mama. Foram extraídas características de estatística simples (média, desvio padrão, diferença entre o maior e o menor tom de cinza da imagem e a quantização do último bin do histograma em uma posterização de 10 novas divisões), dimensão fractal de Higuchi usando cinco escalas para cada quadrante e característica geoestatística (índice de Moran e coeficiente de Geary). Foram utilizados dois bancos de imagem: um com 24 pacientes com alguma doença e 4 saudáveis e outro com 24 pacientes com alguma doença e 10 saudáveis. A classificação utilizou o MVS. Os autores utilizaram a validação cruzada, porém não detalharam em quantas partes a amostra foi dividida. Os melhores resultados foram alcançados com o banco de 34 imagens (24 doentes e 10 saudáveis), obtendo uma sensibilidade de 80%, especificidade de 91,7%, acurácia de 88,23% e área sob a Curva ROC de 0,858.

Marques *et al.* (2013) apresentaram um avaliador de comparação de resultados de diferentes metodologias de segmentações automáticas de imagens biomédicas. Os autores demonstraram que as medidas mais utilizadas (por exemplo, a distância de Hausdorff e funções de distância orientada) não são adequadas para aplicação biomédica. A pesquisa apresentou uma nova técnica para definir o grau de correção da região de interesse que serve de base para comparações e validação de trabalhos de segmentação de imagens biomédicas.

Borchardt (2013) analisou imagens térmicas disponíveis num banco de dados do VisualLab. Todas as imagens utilizadas no trabalho foram adquiridas pelo protocolo estático. O autor utilizou apenas as imagens frontais. Técnicas automáticas e manuais foram utilizadas para segmentar as regiões de interesse, que corresponde às mamas das pacientes. A técnica proposta baseia-se na diferença absoluta entre os valores das características extraídas das mamas direita e esquerda de cada paciente. As características utilizadas foram baseadas em medidas estatísticas, em histograma, na dimensão fractal de Higuchi e em três de geoestatística: Coeficiente de Geary, Índice de Moran e função K de Ripley. No total, foram extraídas 40 características. Foi utilizada uma amostra de 51 termografias, sendo 37 de pacientes com alguma doença

na mama e 14 pacientes saudáveis. As características mais relevantes foram selecionadas com o uso de algoritmo genético, sendo selecionadas 15 características. A classificação dessas características foi realizada usando a MVS. Para avaliar a classificação realizada pela MVS foram utilizadas as medidas de especificidade, sensibilidade, acurácia, área sobre a curva ROC e o índice de Youden. O MVS foi utilizado com seus parâmetros default e utilizou-se Leave-One-Out Cross-Validation. O melhor resultado apresentado pelo método foi uma especificidade de 79%, sensibilidade de 92%, acurácia de 88%, área sob a curva ROC de 0,852 e índice de Youden de 0,705.

Galvão (2015) desenvolveu um método de registro de imagens para os termogramas adquiridos pelo protocolo dinâmico. Para isto, utilizou-se da transformação de Möbius para o mapeamento entre os termogramas. O método elaborado foi testado em um conjunto heterogêneo de 23 pacientes. O índice de melhora foi superior a 80% em quase todas as pacientes, em somente duas, esse índice foi inferior a 75%. Entretanto, pela média e desvio padrão dos resultados, houve melhoria em todas as pacientes, para a maioria das avaliações numéricas de desempenho empregadas.

Oliveira *et al.* (2015) propuseram uma metodologia para a segmentação automática de imagens térmicas capturadas lateralmente. Usando técnicas de processamento de imagens como limiarização de histogramas, clusterização, detecção e refinamento de bordas, foi feita a extração da região de interesse (ROI).

Araújo *et al.* (2015) propuseram uma metodologia para reconstrução da geometria 3D da mama adaptada ao protocolo de aquisição de imagens térmicas. A abordagem consistiu em três etapas: na primeira foi realizada a calibração de dois Kinects fixados em um aparato desenvolvido; na segunda etapa foi realizado um refinamento como forma de melhorar o alinhamento (registro) das nuvens de pontos adquiridas pelos sensores. A última foi a reconstrução da superfície do objeto virtual. As validações foram feitas comparando as mamas reais de voluntárias e fantasmas capturadas por um escâner laser com as formas obtidas pela metodologia desenvolvida, e com medidas adquiridas diretamente nas mamas das voluntárias. A metodologia apresentou uma média da diferença entre as áreas das superfícies de 3,55%, 0,93 de Coeficiente de similaridade Dice, e a média da diferença das distâncias entre os mamilos foi de 3,51%. Silva (2015) desenvolveu uma metodologia para a detecção de anomalias da mama a partir de termografias dinâmicas. Ele utilizou as temperaturas das 20 imagens sequenciais de cada paciente para a criação de séries temporais que foram analisadas

em etapas seguintes. A região de interesse para a geração das séries temporais foi segmentada manualmente e é composta pelas duas mamas da paciente. Esta segmentação manual aconteceu apenas na primeira imagem da sequência de 20 imagens. Todas as 19 restantes pertencentes à sequência, por terem sido registradas em relação à primeira, se utilizaram da máscara gerada da primeira imagem para a determinação das suas respectivas regiões de interesse. Em sua metodologia, o autor classifica a paciente como saudável ou doente, não sendo possível informar em qual das mamas existe a anomalia. Todas as imagens sequenciais de uma dada paciente, como exposto anteriormente, passaram por uma etapa de registro, antes que suas temperaturas fossem extraídas para a geração das séries temporais. Não se avaliou numericamente a qualidade do registro realizado na metodologia. Para a geração da série, a região de interesse foi dividida em pequenas regiões quadradas de 11x11 pixels, e a temperatura máxima de cada uma dessas regiões foi observada em todas as imagens térmicas sequenciais da paciente gerando, assim, as séries temporais. Tais séries foram analisadas por técnicas de aprendizagem de máquina supervisionada e não supervisionada. O algoritmo k-means foi aplicado sobre o conjunto de séries temporais de temperatura, formando k grupos ($2 \leq k \leq 10$). Índices de avaliação de agrupamento foram aplicados para avaliar os resultados dos agrupamentos construídos pelo algoritmo k-mean, para cada valor de k. Os valores obtidos foram tratados como características e submetidos à etapa de classificação supervisionada. O autor utilizou ferramentas de mineração de dados para escolher os melhores algoritmos de seleção e classificação, além de otimização de parâmetros. Para cada um dos 10 índices de validação de agrupamento, variou-se o número de grupos gerados pelo k-means entre 2 e 10 grupos, perfazendo um total de 90 características extraídas de cada paciente. Destas 90 características, após a etapa de seleção de características, foram consideradas 7 características para compor a metodologia. Foi utilizado um banco com 70 pacientes, sendo 35 saudáveis e 35 com alguma anomalia na mama. Dentre os 39 algoritmos de classificação testados, *K-Star* apresentou o melhor resultado com acurácia de 100% com a técnica de validação *Leave-One-Out Cross Validation*. Um conjunto de 6 novas pacientes (3 com e 3 sem anormalidades de mama), fora do conjunto das 70 pacientes, foi utilizado na avaliação da metodologia. Os dados dessas 6 pacientes eram totalmente desconhecidos para os modelos de classificação, pois em momento algum estiveram no conjunto de treinamento de tais modelos. Alcançou-se uma acurácia de apenas 83,33% para os 6 novos casos.

Além dos trabalhos desenvolvidos pelo grupo de pesquisa que compõe o Visual Lab discutidos até então, outros trabalhos encontrados na literatura que também apresentam metodologias de detecção de anomalias de mama a partir de termografias serão discutidos nos parágrafos seguintes.

Na pesquisa de Ohashi e Uchida (2000), as termografias utilizadas foram adquiridas a partir do protocolo dinâmico. As pacientes foram submetidas a 2 minutos de resfriamento através de um ventilador elétrico e em seguida, durante o processo de retorno ao equilíbrio térmico, imagens com intervalos de 15 segundos foram capturadas. Os autores não determinam quantas imagens foram capturadas. Foram geradas novas termografias de subtração a partir da segunda imagem pela primeira de uma sequência. Não foi realizado o registro das imagens antes da geração das termografias de subtração. Os critérios para o diagnóstico do câncer de mama foram os seguintes: i) um ponto quente assimétrico entre as mamas, ii) exagero anormal assimétrico dos padrões vasculares, iii) diferenças de temperatura significativas no padrão de calor (um alto índice térmico), iv) padrões de calor positivos nas termografias dinâmicas e v) existência de uma região quente na termografia de subtração. O diagnóstico é realizado visualmente, uma vez que nenhum método computacional é descrito no texto. Foram analisadas 728 pacientes, todas com câncer. A taxa de verdadeiros positivos foi de 82%, em contrapartida a taxa de falsos positivos foi acima de 40%.

Schaefer *et al.* (2009) analisam termografias adquiridas pelo protocolo estático e baseiam-se na diferença de características entre as duas mamas das pacientes. Diagnostica-se a paciente, não identificando em qual mama encontra-se a anomalia. Eles utilizaram as seguintes características para a classificação das termografias: cálculo da diferença absoluta entre as médias de temperatura das duas mamas da paciente e a diferença absoluta entre o desvio padrão das temperaturas das mamas. Foram utilizados os momentos geométricos M01 e M10, que descrevem essencialmente o centro de gravidade das regiões da mama, bem como a distância (tanto na direção X e Y) do centro de gravidade a partir do centro geométrico da mama. Para todas as quatro características foram calculadas as diferenças absolutas dos valores entre as mamas esquerda e direita. Eles também construíram histogramas normalizados para as duas regiões de interesse e utilizaram a correlação cruzada entre os dois histogramas como uma característica. A partir do histograma diferença foi calculado o valor absoluto do seu valor máximo, o número de compartimentos

superiores a um determinado limiar (ajustado para 0,01 empiricamente nas suas experiências), o número de cruzamentos de zero, a energia e a diferença das partes positiva e negativa do histograma. A partir da matriz de co-ocorrência foram extraídas a homogeneidade, a energia, o contraste e a simetria, além dos quatro primeiros momentos. Os autores não informam os ângulos e as distâncias utilizados para a construção da matriz de co-ocorrência. Foram extraídas também a entropia e a entropia conjunta. Foram extraídas a diferença máxima dos valores do espectro de Fourier e a distância desta posição de máximo do espectro ao centro do gráfico. No total, foram extraídas 38 características das termografias das mamas que descrevem a assimetria entre elas. Foram normalizadas cada característica para o intervalo [0, 1] a fim de chegar a unidades comparáveis entre os descritores. Como resultado do experimento, eles utilizaram um banco com 146 termografias (29 pacientes com câncer e 117 saudáveis). Foi utilizada a lógica Fuzzy para a classificação das pacientes. Para validação dos resultados da classificação, utilizou-se *10-fold Cross-Validation*. Os melhores resultados obtiveram 79,53% de acurácia, 79,86% de sensibilidade e 79,49% de especificidade.

No trabalho desenvolvido por Wishart *et al.* (2010), foram utilizadas termografias obtidas pelo protocolo dinâmico. Durante 5 minutos, uma corrente de ar fria é direcionada para as mamas das pacientes e, durante este período, 250 termografias das mamas são adquiridas por uma câmera térmica. Foram utilizadas imagens de 106 pacientes. Destas, 65 são diagnosticadas com câncer e 41 com lesões benignas. Características de assimetria térmica e diferenças de temperaturas foram extraídas das imagens sequenciais. Os autores não determinaram as características que foram extraídas das imagens. Para a classificação, um método de inteligência artificial desenvolvido pelos autores chamado *NoTouch BreastScan* foi utilizado. Os autores não esclareceram quantas características exatamente foram extraídas e nem detalharam a técnica de classificação. A metodologia obteve uma sensibilidade de 70% e especificidade de 48%.

Acharya *et al.* (2012) utilizaram 50 termografias de mama, sendo 25 saudáveis e 25 de mulheres com câncer, coletadas no Hospital Geral de Cingapura. Foi utilizado o protocolo estático para a aquisição das imagens. Cada paciente aguardou por 15 minutos na sala de exames para estabilização da temperatura e diminuição da taxa metabólica. Como características, foram extraídas a homogeneidade, energia, entropia, os quatro primeiros momentos calculados da matriz de co-ocorrência, segundo

momento angular, contraste, média, ênfase em primitivas curtas, ênfase em primitivas longas, uniformidade do nível de cinza e percentagem de primitivas. No total, foram extraídas 16 características de textura, mas apenas 4 características (primeiro momento, terceiro momento, percentagem de primitivas e uniformidade do nível de cinza) foram selecionadas, devido a sua significância clínica. Para a classificação, foi utilizado a MVS com vetor de característica composto pelas 4 características extraídas. Foi utilizada *3-fold cross-validation* para a validação dos resultados. A metodologia proposta obteve uma acurácia de 88,1%, sensibilidade de 85,71% e especificidade de 90,48%.

Krawczyk e Schaefer (2014) apresentaram uma abordagem para analisar termografias de mama com base na extração de características das imagens capturadas pelo protocolo estático. Foi segmentada manualmente a região das mamas direita e esquerda das imagens térmicas frontais de cada paciente. As regiões das mamas foram convertidas para uma representação de coordenadas polares a fim de facilitar a fase de extração de características. As características extraídas das imagens fornecem indicações sobre a assimetria entre as regiões das mamas esquerda e direita. O sistema classificador múltiplo apresentado baseia-se em uma hibridação de três técnicas de inteligência computacional: redes neurais ou máquinas de vetores como classificadores de base, um fusor neural para combinar os classificadores e uma medida fuzzy para avaliar a diversidade do conjunto e remoção de classificadores a partir do conjunto. Foram extraídas as seguintes características: 4 características de estatística simples, 4 características de momento, oito características baseadas em histograma, 8 características extraídas a partir da matriz de co-ocorrência e 2 baseadas no espectro de Fourier. Também aplicou-se um filtro Laplaciano para aumentar o contraste e calculou-se um outro subconjunto de características (8 baseadas na matriz de co-ocorrência e 2 baseadas no espectro de Fourier) a partir das imagens resultantes. Foi utilizado um conjunto de 146 termografias das quais 29 casos diagnosticados com câncer, enquanto os outros 117 casos diagnosticados como saudáveis. A mesma amostra foi utilizada para o treinamento e teste, porém os autores não detalharam como realizaram a validação cruzada. Os melhores resultados obtiveram uma sensibilidade de 81,96%, especificidade de 90,80% e acurácia de 89,03%.

Francis *et al.* (2014) em sua pesquisa, propõem uma alteração do protocolo estático de aquisição de imagens térmicas de mama a fim de melhorar a acurácia dos índices de detecção do câncer de mama. No protocolo proposto, as imagens são capturadas de

maneira rotacional e as mamas passam por um processo de resfriamento. As pacientes são despidas até a cintura e ficam em repouso antes da captura das imagens, a fim de estabilizar a atividade metabólica basal. Elas são deitadas com os seios voltados para baixo e dentro de um orifício feito na mesa de aquisição, de tal forma que os seios fiquem pendurados livremente em uma câmara escura localizada abaixo da mesa. Dentro da câmara, um braço robótico gira no sentido horário, em torno da mama suspensa. Uma câmera infravermelha é fixada na sua extremidade, capturando imagens da mama a partir de diferentes pontos de vista. Os efeitos de curvatura das imagens foram negligenciados. São capturadas 12 imagens ao redor da mama. Em seguida, a câmara sofre um resfriamento de 2 a 3 graus centígrados e outras 12 imagens são capturadas. No final, uma última imagem do mamilo é capturada, perfazendo um total de 25 imagens. Foram utilizados no trabalho 36 pacientes, sendo 24 sadias e 12 com câncer de mama. De cada termografia, foram extraídas características de estatística de primeira ordem (média, variância, assimetria e curtose) e características de textura propostas por Haralick *et al.* (1973). Foi utilizado a MVS como classificador. A técnica de *Leave-One-Out Cross-Validation* foi utilizada para validar o resultado do classificador. A sensibilidade, especificidade e acurácia obtidas com a análise das imagens antes do resfriamento das mamas foi de 83,3%. Já para as imagens capturadas após o resfriamento, a sensibilidade foi de 66,7%, a especificidade foi de 75% e a acurácia foi de 70,8%.

Gerasimova *et al.* (2014) propõem uma análise multifractal de termografias dinâmicas como um método eficiente para identificar mulheres com risco de câncer de mama. Usando um método multiescala baseada em Wavelet para analisar as flutuações temporais da temperatura da pele da mama obtidas a partir de um grupo de pacientes com câncer de mama e algumas voluntárias com mamas saudáveis, os autores demonstraram que a complexidade multifractal de flutuações de temperatura observadas nas mamas saudáveis é perdida em mamas com tumor maligno. Participaram do estudo 47 pacientes, sendo 33 com diagnóstico de câncer e 14 saudáveis. A classificação foi realizada visualmente através de gráficos e tabelas. Os autores obtiveram uma sensibilidade de 76%, especificidade de 86% e acurácia de 86,9%.

Dourado Neto e Lima (2015) utilizaram termografias adquiridas pelo protocolo estático entre 2005 e 2013 no Hospital das Clínicas da Universidade Federal de Pernambuco, Brasil. A câmera utilizada foi a FLIR S45 com uma resolução de 320x240 pixels e

sensibilidade de $0,08^{\circ}\text{C}$. As voluntárias para o estudo eram pacientes de 40 anos de idade, a idade inicial recomendada para testes de triagem. De um total de 84 pacientes, 20 foram diagnosticados com tumor maligno (câncer), 22 com tumores benignos e 14 com cisto confirmado pela biópsia, e 28 foram diagnosticados como normal (sem alterações) confirmado por exames de mamografia e ultrassonografia. Foram extraídas duas características das matrizes de temperatura segmentadas: a primeira foi a diferença entre as temperaturas máximas de cada uma das mamas e a segunda característica foi a diferença entre o mínimo dos dois máximos anteriormente encontrados (um para cada mama) e o máximo dos dois valores mínimos. Para a classificação, foram testados cinco métodos de aprendizagem supervisionada implementados no MATLAB. São eles: classificação em árvore, MVS, Análise discriminante, Naive Bayes e K vizinhos mais próximos. Para validar os resultados, utilizou-se a validação *Leave-One-Out Cross-Validation*. Os melhores resultados foram obtidos com a classificação em árvore com acurácia de 89,29%, sensibilidade de 75% e especificidade de 93,75%.

Os trabalhos apresentados demonstram o potencial da termografia no auxílio ao diagnóstico de doenças na mama. Foram apresentadas pesquisas que utilizaram termografias obtidas tanto pelo protocolo estático quanto pelo protocolo dinâmico para o auxílio à detecção de anomalias na mama. Buscou-se com isso não restringir apenas às pesquisas com o protocolo dinâmico (foco desta tese), e sim, apresentar de maneira ampla os principais trabalhos relacionados ao uso de termografias no auxílio à detecção de doenças de mamas e seus respectivos resultados.

A tabela 2.1 apresenta um comparativo entre os principais trabalhos apresentados nesta seção que utilizam imagens térmicas no auxílio à detecção de anomalias de mama.

Tabela 2.1 - Resumo comparativo entre os trabalhos

Autor	Protocolo Aquisição Imagens	Amostra	Número de Característica	Classificador
Ohashi e Uchida (2000)	Estático Dinâmico	728 (todas doentes)	4	- Análise visual feita pelo médico.
Schaefer <i>et al.</i> (2009)	Estático	146 (117 saudáveis e 29 doentes)	38	- Lógica Fuzzy

Wishart <i>et al.</i> (2010)	Dinâmico	106 (41 saudáveis e 65 doentes)	Não informado	Sistema Proprietário denominado <i>Sentinel BreastScan</i>
Resmini <i>et al.</i> (2012)	Estático	34 (10 saudáveis e 24 doentes).	11	- MVS
Acharya <i>et al.</i> (2012)	Estático	50 (25 saudáveis e 25 doentes)	16	- MVS
Borchardt (2013)	Estático	51 (14 saudáveis e 37 doentes)	15	- MVS
Krawczyk e Schaefer (2014)	Estático	146 (117 saudáveis e 29 doentes)	36	- Rede Neural - MVS
Francis <i>et al.</i> (2014)	- Estático rotacional - Dinâmico	36 (24 saudáveis e 12 doentes)	18	- MVS
Gerasimova <i>et al.</i> (2014)	- Dinâmico	47 (14 saudáveis e 33 doentes)	1	- Análise visual por meio de gráficos e tabelas.
Dourado Neto e Lima (2015)	Estático	84 (28 saudáveis e 56 com alguma alteração na mama)	2	- Árvore de decisão
Silva (2015)	Dinâmico	70 (35 saudáveis e 35 doentes)	7	- K-Star

A partir da tabela 2.1, é possível perceber que a maioria dos trabalhos que utilizam apenas o protocolo dinâmico para a aquisição das imagens térmicas não usam métodos computacionais para a classificação das mamas em saudáveis ou doentes. Apenas Wishart *et al.* (2010) utilizaram uma técnica de inteligência artificial, porém não detalharam a metodologia. Também Silva (2015) utilizou uma metodologia de classificação automática através do classificador *k-Star*. A metodologia proposta nesta tese pretende, assim como no trabalho de Silva (2015), desenvolver uma metodologia automática de auxílio à detecção de anomalias de mama a partir da análise de termografias dinâmicas. Porém, ao contrário dos trabalhos anteriores, a metodologia proposta possui dois níveis de extração de características. No primeiro, as características são extraídas das regiões de interesse das imagens sequenciais em escala

de cinza ou das matrizes de temperatura. No segundo nível, as características são extraídas das séries temporais formadas pelas características extraídas na etapa anterior.

CAPÍTULO 3 - REFERENCIAL TEÓRICO

Neste capítulo serão apresentados os principais conceitos utilizados para o desenvolvimento da metodologia proposta nesta tese. São apresentados os conceitos de termografia mamária estática e dinâmica, ferramentas de auxílio ao diagnóstico por computador, além de características de imagens. Essas características foram agrupadas da seguinte forma: grupo de características de estatística simples formado pela média e desvio padrão das temperaturas; grupo de características baseadas em histogramas, formado pelas características de comparação de histogramas (interseção (Lee *et al.*, 2005), qui-quadrado (Cai *et al.*, 2011) e distância de Bathacharyya (Bhattacharyya, 1946)) e descritores de textura de Unser (Unser, 1986); característica baseada em comparação de spatiograms (Birchfield e Rangarajan, 2005); grupo de características baseadas em índices de diversidade (Magurran, 1988 p.07), formado pelos índices de Simpson, Berger-Parker e McIntosh e, por fim, característica baseada em geometria fractal (lacunaridade) (Du e Yeo, 2002). Também serão apresentados os conceitos sobre classificação de imagens.

3.1 TERMOGRAFIA MAMÁRIA

A termografia é o registro da distribuição de temperatura de um corpo utilizando a radiação infravermelha emitida pela sua superfície. Câmeras especiais são usadas para detectar a energia infravermelha emitida pela superfície dos corpos e, juntamente com outros parâmetros, é usada para calcular a temperatura real do objeto (Kennedy *et al.*, 2009).

Os avanços nas tecnologias de câmeras de infravermelho tem aumentado o interesse na aplicação do sistema na área médica, especialmente na detecção do câncer de mama (EtehadTavakol *et al.*, 2010).

Como as alterações anatômicas no tecido da mama podem ser precedidas por alterações fisiológicas tais como um aumento anormal da atividade metabólica e de circulação vascular, a termografia mamária tem sido considerada como um método promissor de auxílio ao rastreio do câncer da mama (Gerasimova *et al.*, 2013).

O comportamento térmico diferenciado das mamas com câncer pode ser explicado devido ao processo de angiogênese patológico, que é o surgimento de novos vasos sanguíneos para o suprimento de sangue do tumor (Kennedy *et al.*, 2009). Nesse

processo há uma diminuição de células do músculo liso dos vasos sanguíneos tornando-os incapazes de realizarem a vasoconstrição, que normalmente aconteceria (Ziyad e Iruela-Arispe, 2011). Outras explicações seriam o óxido nítrico, que é uma substância vasodilatadora que também é produzida localmente pelas células cancerígenas; e o processo de inflamação causado pelo câncer (Kennedy *et al.*, 2009).

A termografia pode apresentar vantagens no rastreio de doenças de mama em mulheres mais jovens, devido ao exame de mamografia não ser indicado para mulheres com mamas densas (Ng, 2009).

Segundo Moghbel e Mashohor (2011), 44% das pessoas com uma termografia anormal irá desenvolver câncer da mama detectável no prazo de 5 anos e em 70% das pacientes, o câncer será detectável por mamografia um ano após ter sido detectada pela termografia.

A termografia pode ser usada associada à mamografia. Ela pode destacar áreas de mama que representam algum risco, de modo que a mamografia dessa área ou da mama poderia ser examinada mais detalhadamente. Também um sistema CAD poderia incorporar a mamografia, termografia e dados fisiológicos da paciente aumentando a detecção do câncer em quase 95% (Keyserlingk *et al.*, 2000).

A aquisição de imagem térmica pode ser categorizada em relação ao comportamento do corpo sob a transferência de calor, e classificada em estática ou dinâmica. Também pode ser categorizada de acordo com a repetição de captura e classificada como única, sequencial ou acompanhada. Na aquisição estática, a paciente fica em equilíbrio térmico com o meio ambiente. Já na aquisição dinâmica é observada a recuperação da temperatura da pele após um estresse térmico (por exemplo, arrefecimento ou aquecimento) ou químico. Na aquisição única, as imagens da paciente são capturadas em um único instante de tempo. Este tipo de aquisição é adequado para a identificação de áreas de pontos quentes e frios, além de ser possível medir as assimetrias na distribuição da temperatura da pele. No modo sequencial, uma série de imagens é obtida sequencialmente ao longo de um período de tempo, que pode ser configurado na câmera, anteriormente. No modo acompanhado de captura, as imagens são adquiridas separadas por um intervalo de tempo substancial (cada 3 ou 6 meses) de modo a monitorar o progresso de uma doença ou detectá-la (Borchardt *et al.*, 2013).

Nesta tese, as imagens foram capturadas a partir de um protocolo dinâmico sequencial (Silva *et al.*, 2013), ou seja, as pacientes sofreram um resfriamento induzido através de um ventilador elétrico na região das mamas e, durante o processo de retorno ao

equilíbrio térmico com o meio ambiente, imagens sequenciais foram capturadas em intervalos de 15 segundos durante 5 minutos, perfazendo um total de 20 imagens térmicas sequenciais de cada paciente. Na seção seguinte é apresentada uma discussão sobre o uso de imagens adquiridas pelo protocolo dinâmico para o auxílio na detecção de doenças da mama.

3.1.1 TERMOGRAFIA DINÂMICA

Termografias isoladas contêm informações escassas. Alternativas a estas termografias estáticas demonstram bons resultados, em particular, a chamada termografia dinâmica. Testes dinâmicos geralmente incluem a exposição do organismo a impactos externos. O estímulo ao frio ou ao calor do organismo induz a vasoconstrição e vasodilatação, que são reflexos autônomos em estruturas normais, aumentando assim o contraste térmico devido a diferenças na vascularização das diferentes regiões da pele (Vainer, 2005). Esta abordagem é mais rápida do que a abordagem estática, que requer um controle rigoroso das condições ambientais e maior tempo (às vezes horas) para que a paciente se estabilize termicamente às condições da sala de captura das imagens (Herman, 2013).

O interesse pela termografia dinâmica na identificação do câncer de mama aumentou devido ao rápido desenvolvimento de tecnologias de imagem térmica e de novas câmeras digitais termográficas com resolução de temperatura mais elevada ($0,08^{\circ}\text{C}$ ou menor), combinados com o aumento do conhecimento da angiogênese e produção de óxido nítrico do tecido cancerígeno causando perturbações vasomotoras locais em comparação com tecidos normais (Gerasimova *et al.*, 2013).

A área da termografia dinâmica foi proposta como uma evolução da termografia estática. Quanto à repetição de captura, e como explicado anteriormente, ela pode ser sequencial, que corresponde a aquisição de várias imagens térmicas no tempo a uma dada taxa de quadros. Por meio da análise das variações de temperatura no tempo, informações sobre a perfusão local são obtidas, as quais podem ter valor diagnóstico, uma vez que tumores são conhecidos por apresentar vascularização anormal significativa, em comparação com o tecido normal (Anbar *et al.*, 2001).

3.2 DIAGNÓSTICO AUXILIADO POR COMPUTADOR

Sistemas de Auxílio ao Diagnóstico por Computador (CAD - *Computer-Aided Detection or Diagnosis*) são projetados para melhorar a capacidade de profissionais de

saúde na tomada de decisão. Eles são atualmente empregados em muitos aspectos da prática clínica, tais como detecção do câncer de mama, doenças vasculares, dentre outras (Moghbel e Mashohor, 2011).

No contexto do rastreio de câncer de mama, um sistema CAD deve ser visto como uma ferramenta auxiliar, para a melhoria do diagnóstico da doença (Calas *et al.*, 2012). Ele é uma ferramenta clinicamente comprovada que ajuda na detecção ou diagnóstico de câncer, auxiliando profissionais de saúde na detecção ou classificação de tipos de câncer em estágios iniciais (Tang *et al.*, 2014).

Com a evolução dos computadores, melhores câmeras infravermelhas e melhores algoritmos de classificação, diferentes abordagens de sistemas CAD foram propostos para imagens térmicas. Para o rastreio de doenças na mama, apesar dos resultados variarem entre abordagens, observa-se a eficácia de sistemas CAD (Moghbel e Mashohor, 2011).

O desempenho de um sistema CAD para o auxílio ao diagnóstico do câncer depende da forma de aquisição dos dados e dos sistemas de classificação e extração de características (Tang *et al.*, 2014). Na seção seguinte são apresentadas as características que servirão de base na etapa de classificação proposta na metodologia apresentada nesta tese.

3.3 CARACTERÍSTICAS DE IMAGENS

Em sistemas de análise de imagens, visão computacional e reconhecimento de padrões, necessita-se, muitas vezes, reconhecer elementos ou objetos de interesse nas imagens. Isso requer o estabelecimento de parâmetros quantificáveis para poder ser realizado computacionalmente. Esses parâmetros dependem de valores como cor, posição, textura e outros que caracterizem o objeto (Conci *et al.*, 2008 p.197).

Para realizar qualquer tipo de reconhecimento por classificação, é necessário ter inicialmente uma descrição o mais detalhada possível dos padrões que se deseja reconhecer. Dessa descrição serão projetadas formas possíveis de se extrair as informações quantitativas a serem obtidas analisando-se as imagens. Tais informações consistem nas características que serão extraídas das imagens (Conci *et al.*, 2008 p.197).

Nos parágrafos seguintes, serão apresentados grupos de características que auxiliam na descrição das imagens térmicas de mama.

3.3.1 ESTATÍSTICA SIMPLES

Para a extração de informações de conjuntos de dados, muitas vezes são usados métodos estatísticos.

A média para um total de n valores de temperatura, com a temperatura de cada elemento denotado por t_i , com $i = 1, \dots, n$, é definida na Equação 3.1.

$$\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i \quad (3.1)$$

O desvio padrão representa o quanto de variação existe em relação à média. Um baixo desvio padrão indica que os valores estão próximos da média. Já um alto desvio padrão indica que os dados estão espalhados em relação à média.

O desvio padrão é definido a partir da equação 3.2.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (t_i - \bar{t})^2} \quad (3.2)$$

3.3.2 MEDIDAS BASEADAS EM HISTOGRAMAS

Define-se histograma de uma imagem como sendo uma quantificação do número de pixels para cada valor de intensidade de uma dada imagem. Em outras palavras, é uma representação gráfica das intensidades de uma imagem (Gonzalez e Woods, 2007). Ele pode ser visto como uma distribuição discreta de probabilidade, pois o número de pixels normalizados para um determinado nível de cinza indica a probabilidade de se encontrar um pixel com aquele valor de cinza na imagem (Pedrini e Schwartz, 2008 p.104).

Como uma característica estatística, o histograma não é sensível a translação e rotação dos objetos (Datta *et al.*, 2008). Ele é amplamente utilizado na segmentação de imagens, registro, rastreamento, e como fonte de extração de característica para a classificação de imagens.

3.3.2.1 COMPARAÇÃO DE HISTOGRAMAS

Comparação de histogramas é uma forma de se extrair informações de imagens. A distância entre dois histogramas é a principal forma de avaliação para mensurar a

semelhança ou diferença das suas propriedades estatísticas correspondentes (Ma *et al.*, 2010). Encontrar a distância ou semelhança entre dois histogramas é uma questão importante na classificação de padrões e recuperação de imagens.

Há duas metodologias para medir a distância entre histogramas: vetorial e probabilística. Na abordagem vetorial, um histograma é tratado como um vetor de dimensão fixa. Assim, medidas como interseção podem ser usadas para mensurar distâncias. Esse tipo de medida tem sido usada na indexação e recuperação de imagens. A abordagem de natureza probabilística se baseia no fato de que uma medida de histograma fornece a base para uma estimativa empírica da função de densidade de probabilidade (pdf). O cálculo da distância entre duas pdf's pode ser considerado o mesmo que calcular a probabilidade Bayesiana. Um exemplo seria a distância de Bhattacharyya entre populações estatísticas, que pode assumir valores entre 0 e 1 (Cha e Srihari, 2002).

Nas seções seguintes, serão apresentadas três medidas de comparação de histogramas: Interseção entre histogramas, qui-quadrado e distância de Bhattacharyya.

3.3.2.1.1 INTERSEÇÃO

Dado um par de histogramas, $H(I)$ e $H(I')$, de imagens I e I' , respectivamente, cada uma contendo n níveis de cinza, define-se a interseção dos histogramas normalizados a partir da equação 3.3 (Lee *et al.*, 2005).

$$H(I) \cap H(I') = \sum_{j=1}^n \min(H_j(I), H_j(I')) \quad (3.3)$$

onde H_j é o j -ésimo elemento do histograma e n é o número de elementos do histograma.

Para duas imagens, quanto maior for o valor de interseção entre os histogramas, mais semelhante é o par de imagens consideradas.

3.3.2.1.2 QUI-QUADRADO

A distância qui-quadrado vem do teste do Qui-quadrado estatístico, que é usada para testar a qualidade do ajuste entre uma distribuição de frequência observada e uma distribuição teórica (Cai *et al.*, 2011). Considerando $H(I)$ e $H(I')$ os histogramas

normalizados das imagens I e I' , a distância qui-quadrado entre os dois histogramas é definida pela equação 3.4.

$$d(H(I), H(I')) = \sum_{j=1}^n \frac{(H_j(I) - H_j(I'))^2}{H_j(I)} \quad (3.4)$$

onde n é o número de elementos dos histogramas e H_j é o j -ésimo elemento.

3.3.2.1.3 DISTÂNCIA DE BHATTACHARYYA

É uma medida de distância entre populações estatísticas (Kailath, 1967). A distância de Bhattacharyya varia entre 0 e 1. No caso de comparação de histogramas de imagens, valores baixos indicam alta semelhança entre as imagens e altos valores indicam pequena semelhança. Uma combinação perfeita é 0 e uma incompatibilidade total é 1. Considerando dois histogramas normalizados $H_j(I)$ e $H_j(I')$ extraídos a partir de imagens de tamanho $N \times M$, com H_j sendo o j -ésimo elemento do histograma, a distância de Bhattacharyya (D_b) entre $H_j(I)$ e $H_j(I')$ é dada pela equação 3.5 (Bhattacharyya, 1946):

$$D_b = \sqrt{1 - \sum_{j=1}^n \sqrt{H_j(I)H_j(I')}} \quad (3.5)$$

onde n é o número de elementos dos histogramas.

3.3.2.2 DESCRITORES DE TEXTURA DE UNSER

Na análise de imagens, uma das tarefas mais complexas está na definição de um conjunto de características capazes de descrever de maneira efetiva cada região contida em uma imagem, de modo a ser utilizado em processos de mais alto nível, tal como a classificação de padrões. Uma abordagem natural está em recorrer às características utilizadas pelos humanos na interpretação de informações visuais. A textura encontra-se entre as características empregadas pelo sistema visual humano, contendo informações sobre a distribuição espacial e a variação de luminosidade, além de descrever o arranjo estrutural das superfícies e relações entre regiões vizinhas (Pedrini e Schwartz, 2008 p.287).

Existe uma grande quantidade de definições e métodos de análise de textura encontrados na literatura (Pedrini e Schwartz, 2008 p.288). Dentre os vários métodos, destaca-se a abordagem estatística proposta por Unser (1986). Ela se baseia nos histogramas da soma e da diferença e é uma alternativa para abordagem proposta por Haralick *et al.* (1973), onde 14 medidas são propostas para extrair informações texturais úteis das matrizes de co-ocorrência espaciais. Porém, a principal desvantagem de utilizar essas matrizes é a grande exigência de memória para armazená-las. Por vezes tem-se situações paradoxais onde as matrizes de co-ocorrência utilizadas para a caracterização da textura são mais volumosas do que as imagens originais a partir das quais são derivadas. É também evidente que, devido à sua grande dimensionalidade, tornam-se muito sensíveis à dimensão das amostras de textura em que são estimadas (Unser, 1986).

Já Unser (1986) sugere extrair os descritores de textura propostos por Haralick *et al.* (1973) a partir dos histogramas da soma e da diferença, a fim de simplificar e reduzir o custo computacional. Quando considera-se, por exemplo, uma imagem de 256 tons de cinza, esta possuirá a sua matriz de co-ocorrência composta por $256 \times 256 = 65.536$ elementos, com muitos valores nulos. Caso esta matriz seja utilizada para a análise de textura, nem sempre todos os cálculos serão necessários. Também o ganho no tempo computacional e exigência de memória podem ser consideravelmente inferiores. Por exemplo, considerando uma imagem quantizada em 32 níveis de cinza, a utilização do método baseado nos histogramas da soma e da diferença permitirá uma computação oito vezes mais rápida do que o cálculo dos descritores baseados nas matrizes de co-ocorrência e demandará oito vezes menos memória (Unser, 1986).

Os histogramas da soma e da diferença observam o relacionamento de cada pixel com seus vizinhos de distância d . Eles dependem das distâncias dx e dy , ou seja, para cada pixel de posição (i, j) da imagem, são realizadas somas ou diferenças com os pixels localizados na posição $(i+dx, j+dy)$. A figura 3.1 apresenta o pixel de posição (i, j) e o respectivo pixel de posição $(i+dx, j+dy)$.

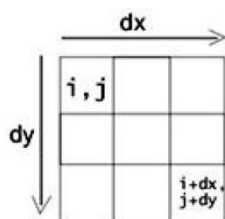


Figura 3.1 - Representação da vizinhança do pixel

Assim, para imagens de 8 bits, o histograma da diferença possui elementos possíveis dos tons -255 a 255 e o histograma da soma de 0 a 510 elementos, sendo cada elemento definido pelas equações 3.6 e 3.7, respectivamente.

$$S(d_i, d_j) = (i + d_x, j + d_y) + (i, j) \quad (3.6)$$

$$D(d_i, d_j) = (i + d_x, j + d_y) - (i, j) \quad (3.7)$$

onde os d_x, d_y indicam como os tons dos pixels serão somados ou subtraídos para a elaboração dos histogramas. Após a criação dos histogramas, esses serão normalizados de modo a se transformarem em probabilidade (Conci *et al.*, 2008 p.291).

Segundo Baraldi e Parmiggiani (1995), apenas 6 medidas estatísticas propostas por Haralick *et al.* (1973) apresentam maior relevância. São elas: variância, segundo momento angular (ou energia), correlação, entropia, contraste e homogeneidade.

A tabela 3.1 apresenta a equivalência entre as 6 características destacadas por Baraldi e Parmiggiani (1995) computados a partir das matrizes de co-ocorrência e calculadas a partir de seus histogramas da soma e da diferença equivalentes. $\hat{P}_s(i)$ é o i -ésimo elemento normalizado do histograma da soma, $\hat{P}_d(j)$ é o j -ésimo elemento normalizado do histograma da diferença e μ é o valor da média.

Tabela 3.1 - Equivalência de características entre matriz de co-ocorrência e histogramas da soma e da diferença. Fonte: Unser (1986).

Característica	Matriz co-ocorrência	Histogramas da Soma e Diferença
Variância	$f_1 = \sum_i \sum_j (i - \mu)^2 \cdot \hat{P}(i, j)$	$= \frac{1}{2} (\sum_i (i - 2\mu)^2 \cdot \hat{P}_s(i) + \sum_j j^2 \cdot \hat{P}_d(j))$
Energia	$f_2 = \sum_i \sum_j \hat{P}(i, j)^2$	$= \sum_i \hat{P}_s(i)^2 \cdot \sum_j \hat{P}_d(j)^2$
Correlação	$f_3 = \sum_i \sum_j (i - \mu) \cdot (j - \mu) \cdot \hat{P}(i, j)$	$= \frac{1}{2} (\sum_i (i - 2\mu)^2 \cdot \hat{P}_s(i) - \sum_j j^2 \cdot \hat{P}_d(j))$
Entropia	$f_4 = \sum_i \sum_j -\hat{P}(i, j) \cdot \log(\hat{P}(i, j))$	$= -\sum_i \hat{P}_s(i) \cdot \log(\hat{P}_s(i)) - \sum_j \hat{P}_d(j) \cdot \log(\hat{P}_d(j))$
Contraste	$f_5 = \sum_i \sum_j (i - j)^2 \cdot \hat{P}(i, j)$	$= \sum_j j^2 \cdot \hat{P}_d(j)$
Homogeneidade	$f_6 = \sum_i \sum_j \frac{1}{1 + (i - j)^2} \cdot \hat{P}(i, j)$	$= \sum_j \frac{1}{1 + j^2} \cdot \hat{P}_d(j)$

Medida de heterogeneidade, a variância apresenta valores altos quando os tons de cinza desviam do nível de cinza médio. Energia, ou segundo momento angular, expressa a uniformidade de uma textura. Em texturas ásperas, a energia apresenta valores próximos de um, o máximo possível para essa medida. A correlação mede a dependência linear entre os tons de cinza presentes em uma imagem. Valores altos, próximos de um, indicam a existência de relação linear entre os pares de nível de cinza. A medida entropia expressa a desordem contida na textura. O contraste caracteriza-se pela diferença entre os tons de cinza. Baixo contraste ocorre quando há pequena diferença entre os níveis de cinza dos pixels localizados em uma região contígua da imagem. A homogeneidade assume valores altos quando a textura apresenta pequenas variações de níveis de cinza entre pares de pixels. Tal medida apresenta correlação inversa com a medida de contraste discutida anteriormente (Pedrini e Schwartz, 2008 p.294-295).

3.3.3 ÍNDICES DE DIVERSIDADE ECOLÓGICA

As medidas de diversidade levam em conta dois fatores: a riqueza de espécies, que é o número de espécies, e uniformidade (conhecido também como equidade), que é a distribuição de indivíduos dentre as espécies de uma amostra (Magurran, 1988 p.34). Comunidades com a mesma riqueza podem se diferenciar pela uniformidade (McIntosh, 1967). Destacam-se os seguintes índices de diversidade: índice de Simpson, índice de Berger-Parker e índice de McIntosh.

3.3.3.1 ÍNDICE DE SIMPSON

O índice de Simpson é referido como uma medida de dominância, uma vez que é ponderada para as abundâncias das espécies mais comuns, ao invés de fornecer uma medida da riqueza de espécies. Ele expressa a probabilidade de dois indivíduos escolhidos aleatoriamente a partir de uma comunidade infinitamente grande pertencerem a espécies diferentes. É definido a partir da equação 3.8.

$$Ds = \sum p_i^2 \quad (3.8)$$

onde p_i é igual à probabilidade da ocorrência de um indivíduo na espécie i .

A fim de calcular o índice de forma apropriada para uma comunidade finita é utilizada a equação 3.9.

$$D_s = \sum \left(\frac{n_i(n_i - 1)}{N(N - 1)} \right) \quad (3.9)$$

onde n_i é igual ao número de indivíduos na espécie i e N é igual ao número total de indivíduos (Magurran, 1988 p.39).

3.3.3.2 ÍNDICE DE BERGER-PARKER

O índice de Berger-Parker também é uma medida de dominância intuitivamente simples e fácil de calcular. Este índice expressa a importância proporcional das espécies mais abundantes (Magurran, 1988). Ele é representado pela equação 3.10.

$$d = \frac{N_{max}}{N} \quad (3.10)$$

onde N_{max} é igual ao número de indivíduos da espécie mais abundante. É uma das medidas de diversidade mais satisfatórias disponíveis (Magurran, 1988 p.41).

3.3.3.3 ÍNDICE DE MCINTOSH

McIntosh (1967) propôs que uma comunidade poderia ser descrita como um ponto num hipervolume dimensional S e que a distância euclidiana do conjunto a partir da origem poderia ser usado como uma medida de diversidade. Esta distância é conhecida como U e é calculada como na equação 3.11.

$$U = \sqrt{\sum_{i=1}^S n_i^2} \quad (3.11)$$

onde S é igual ao número de espécies e n é igual ao número de indivíduos em cada uma das espécies. Uma vantagem do índice proposto por McIntosh (1967), é que deriva de uma medida de similaridade do qual é um caso especial, e é uma representação mais natural e familiar de pontos num sistema de coordenadas.

Quando amostras de tamanho diferentes são comparadas, o cálculo do índice de diversidade de McIntosh é independente de N (número total de indivíduos da amostra).

A vantagem é expressar a diversidade observada como uma proporção da diversidade máxima absoluta, conforme equação 3.12 (Rocha, 2014).

$$Mc = \frac{N - U}{N - \sqrt{N}} \quad (3.12)$$

O índice pode variar de 0, se existir apenas uma espécie, até 1 no caso de uma diversidade máxima. A partir da tabela 3.2, McIntosh (1967) ilustra os valores do índice de diversidade máxima e mínima em porcentagem imaginando uma população de 100 indivíduos e variando o número de espécies de 1 a 100.

Tabela 3.2 - Valores do índice de diversidade de McIntosh mínimo e máximo para população de 100 indivíduos. Fonte: McIntosh (1967)

Número de Espécies	Valor Mínimo do índice	Valor Máximo do índice
1	0,00	0,00
2	0,01	0,32
4	0,03	0,55
5	0,04	0,61
10	0,09	0,75
20	0,20	0,86
50	0,53	0,95
75	0,80	0,98
90	0,95	0,99
100	1,00	1,00

3.3.4 MEDIDA BASEADA EM COMPARAÇÃO DE SPATIOGRAMS

Segundo Birchfield e Rangarajan (2005), histograma espacial ou spatiogram é um histograma em que cada elemento é espacialmente ponderado pela média e covariância das coordenadas dos pixels que contribuem para aquele bin. Ele não capta apenas uma informação sobre a ocorrência do intervalo da função, como um histograma, mas também informação sobre o domínio (espacial). Então, um histograma é apenas um spatiogram de ordem zero.

Os histogramas são eficientes para calcular e permitir a comparação entre regiões de imagens correspondentes sem especificamente levar em conta a relação de posição entre elas. Os spatiograms também permitem realizar comparações entre regiões de

imagens, porém, eles levam em conta informações espaciais sobre a geometria das regiões. Em comparação com as matrizes de co-ocorrência, spatiograms capturaram as posições globais dos pixels ao invés de suas relações aos pares (Birchfield e Rangarajan, 2005).

Spatiogram pode ser entendido como uma generalização do histograma que colhe informações espaciais. O spatiogram de uma imagem é representado como um conjunto, de acordo com as equações 3.13 e 3.14.

$$S = \{s_u\}_{u=1\dots n} \quad (3.13)$$

$$s_u = \langle h_u, \mu_u, \Sigma_u \rangle \quad (3.14)$$

onde n é o número de bins. h_u , μ_u e Σ_u são, respectivamente, a frequência (ou probabilidade), a média das coordenadas e a covariância das coordenadas dos pixels pertencentes ao u -ésimo bin, que podem ser computados pelas equações 3.15, 3.16 e 3.17.

$$h_u = \frac{1}{N} \sum_{i=1}^N \delta_{iu} \quad (3.15)$$

$$\mu_u = \frac{1}{\sum_{j=1}^N \delta_{ju}} \sum_{i=1}^N x_i \delta_{iu} \quad (3.16)$$

$$\Sigma_u = \frac{1}{\sum_{j=1}^N \delta_{ju}} \sum_{i=1}^N (x_i - \mu_u)(x_i - \mu_u)^T \delta_{iu} \quad (3.17)$$

onde N é o número total de pixels. x_i é um vetor $[x_i, y_i]^T$ representando a informação espacial do i -ésimo pixel. δ_{iu} é uma função delta de Kronecker, que será igual a 1 se o i -ésimo pixel cair no u -ésimo bin, e igual a 0, caso contrário (Gong *et al.*, 2009).

Segundo Birchfield e Rangarajan (2005), a similaridade entre dois spatiograms pode ser calculada como a soma ponderada da similaridade entre dois histogramas, conforme equação 3.18.

$$\rho(h, h') = \sum_{b=1}^m \psi_b \rho_n(n_b, n'_b) \quad (3.18)$$

Para um spatiogram de ordem zero, ou seja, um histograma, $\psi_b = 1$. Para um spatiogram de segunda ordem, define-se ψ_b como a probabilidade de que $\bar{\mathbf{x}}_b$ seja elaborado a partir de uma distribuição Gaussiana descrito por $(\bar{\mathbf{x}}'_b, \Sigma'_b)$, multiplicado pela probabilidade no sentido inverso (equação 3.19).

$$\psi_b = \eta \exp \left\{ -\frac{1}{2} (\mu_b - \mu'_b)^T \hat{\Sigma}_b^{-1} (\mu_b - \mu'_b) \right\} \quad (3.19)$$

onde η é a constante de normalização Gaussiana e $\hat{\Sigma}_b^{-1} = (\Sigma_b^{-1} + (\Sigma'_b)^{-1})$. Nota-se que o valor dentro da soma é a média das duas distâncias de Mahalanobis, uma entre x e x' e a outro entre x' e x .

Já a similaridade entre os bins dos histogramas pode ser calculada usando qualquer uma das diversas métricas de comparação, tais como a intersecção, definida como na equação 3.20.

$$\rho_n(n_b, n'_b) = \frac{\min(n_b, n'_b)}{\sum_{j=1}^m n_j} \quad (3.20)$$

3.3.5 GEOMETRIA FRACTAL - LACUNARIDADE

A dimensão fractal pode ser vista como uma medida da irregularidade de muitos processos físicos. No entanto, diferentes conjuntos fractais podem compartilhar a mesma dimensão fractal e possuir diferentes texturas (Mandelbrot, 1983).

Para discriminar essas texturas, dimensão fractal sozinha seria inútil. Mandelbrot (1983) introduziu o termo lacunaridade para descrever a característica de fractais da mesma dimensão com diferentes aparências de textura.

Segundo Barros Filho e Sobreira (2008), lacunaridade pode ser definida como uma medida da dimensão fractal complementar ou o desvio de uma estrutura geométrica da sua invariância translacional. Ela permite distinguir padrões espaciais através da análise de sua distribuição em diferentes escalas de lacunas (vazios) (Plotnick *et al.*, 1996). Lacuna em uma imagem pode ser entendida como pixels com um valor específico (por exemplo, pixels de valor 1 em imagens binárias) ou de um certo intervalo de valores (em imagens em tons de cinza). Quanto maior for a lacunaridade de um padrão

espacial, maior será a variabilidade das suas lacunas em uma imagem e mais heterogênea será.

Lacunaridade está relacionada com a distribuição de tamanhos de clareiras: objetos homogêneos possuem baixa lacunaridade, porque todos os tamanhos de clareiras são os mesmos, ao passo que os objetos com alta lacunaridade são heterogêneos. Deve-se notar que os objetos que são homogêneos ao longo de uma pequena área podem ser heterogêneos quando examinados ao longo de uma grande área e vice-versa. Portanto, lacunaridade pode ser considerada uma medida dependente da escala (Gefen *et al.*, 1983).

Existem muitos algoritmos para definir, quantificar e calcular a lacunaridade, como por exemplo os de Dong (2000) e Gefen *et al.* (1983). Porém, os métodos mais populares baseiam-se na utilização do algoritmo *Gliding-box* proposto por Allain e Cloitre (1991). Esse processo analisa as flutuações da função de distribuição de massa ao longo da imagem. Para isso, utiliza janelas de tamanhos variáveis que “percorrem” o conjunto em análise, posicionando-se em cada ponto deste. A partir das diferentes translações das janelas, calculam-se os momentos de distribuição de massa, que são usados na expressão de Lacunaridade. De forma geral, para cada tamanho da janela utilizada no algoritmo tem-se um valor diferente que quantifica uma distribuição lacunar na imagem, e todos esses valores podem ser plotados em um gráfico log-log que mostra a Lacunaridade do conjunto como uma função do tamanho da janela. A curva derivada nesse processo vai apresentar características que refletem as propriedades geométricas do padrão em estudo (Fernandes e Conci, 2004).

Esse método proposto por Allain e Cloitre (1991) pode ser aplicado a objetos do espaço euclidiano de dimensão n ($n = 1, 2, 3, \dots$), com o cuidado da caixa que será transladada pela imagem ser definida com a mesma dimensão do objeto.

Ao contrário de outros estudos espaciais de amostragem, aqui as caixas deslizantes se sobrepõem. O algoritmo *Gliding-box* quando aplicado às imagens binárias (imagens com apenas 1 bit) conta apenas os pixels 1. Isso ocorre porque cada pixel em uma imagem binária pode ter apenas um dos dois valores possíveis (0 ou 1). Considerando que, em imagens em tons de cinza, um pixel pode assumir 2^8 valores, neste caso, mede-se a massa M de cada caixa através da intensidade média dos pixels por caixa, que é a diferença entre o valor máximo ($g(i, j)$) e mínimo ($b(i, j)$) de intensidade em cada caixa de tamanho r , conforme equação 3.21 (Du e Yeo, 2002).

$$n(i, j) = g(i, j) - b(i, j) \quad (3.21)$$

Então, $n(i, j)$ é a contribuição de M da (i, j) ésima caixa deslizante.

A caixa de tamanho $r \times r$ desliza sobre a imagem de tamanho $W \times W$ (sendo $r < W$) até que toda a área analisada seja coberta. O número de caixas deslizantes com tamanho r e massa M é definido como $n(M, r)$. A distribuição de probabilidade $Q(M, r)$ é obtida pela divisão $n(M, r)$ pelo número total de caixas, conforme equação 3.22 (Du e Yeo, 2002).

$$Q(M, r) = \frac{n(M, r)}{(W - r + 1)^2} \quad (3.22)$$

Logo, a lacunaridade em escala é definida pelo desvio médio quadrático da função de probabilidade $Q(M, r)$ dividido pelo quadrado de sua média, conforme equação 3.23.

$$\Lambda(r) = \frac{\sum_M M^2 Q(M, r)}{[\sum_M M Q(M, r)]^2} \quad (3.23)$$

3.4 CLASSIFICAÇÃO

A classificação de imagens consiste em determinar em que classe uma dada região da imagem se enquadra. A classificação depende das características que discriminam cada padrão dos demais. Para realizar a classificação, torna-se necessário utilizar alguma técnica de decisão. Em outras palavras, denomina-se classificação ou reconhecimento o processo de atribuição de um rótulo a um objeto, baseado nas características extraídas do objeto que são representadas por seus descritores. Existem, normalmente, duas fases em um ambiente de reconhecimento: a fase de aprendizado e a fase de interpretação. A fase de aprendizado envolve o conhecimento dos conjuntos de objetos que podem estar presentes na imagem. A fase de interpretação pressupõe a análise de novas imagens (Conci *et al.*, 2008 p.259-260).

Segundo Pedrini e Schwartz (2008), a classificação de padrões visa determinar um mapeamento que relacione as propriedades extraídas de amostras com um conjunto de rótulos, apresentando a restrição de que as amostras com características semelhantes

devem ser mapeadas ao mesmo rótulo. Quando se atribui um mesmo rótulo a amostras distintas, diz-se que tais elementos pertencem a uma mesma classe.

As técnicas de classificação podem ser divididas em duas abordagens (Haykin, 1999):

- Supervisionada: quando o processo de classificação considera classes previamente definidas. Para que os parâmetros que caracterizam cada classe sejam obtidos, uma etapa denominada treinamento deve ser executada anteriormente à aplicação do algoritmo de classificação. Tais parâmetros são obtidos a partir de amostras previamente identificadas (Pedrini e Schwartz, 2008 p.398).
- Não-supervisionada: quando não se dispõe de parâmetros ou informações coletadas previamente à aplicação do algoritmo de classificação. Nesse caso, todas as informações de interesse devem ser obtidas a partir das próprias amostras a serem rotuladas (Pedrini e Schwartz, 2008 p.398).

Existe uma infinidade de algoritmos de classificação supervisionada e não-supervisionada. Nesta tese, optou-se pela utilização da Máquina de Vetores de Suporte (MVS), que é um classificador supervisionado, para a classificação das mamas em saudáveis ou doentes.

A Máquina de Vetores de Suporte (MVS) é um método de aprendizagem supervisionado que é utilizado para estimar uma função de classificação de dados de entrada em duas classes (Vapnik, 1998). Ela foi introduzida por Vapnik e Chervonenkis (1971) e despertou grande interesse desde a sua introdução. Este método tem apresentado resultados, muitas vezes, melhores que muitos outros algoritmos de aprendizagem (Mitchell, 1997). Diversos autores discutem as máquinas de vetores de suporte, como no trabalho de Lorena e Carvalho (2007), e também a sua utilização na classificação de imagens médicas, como nos trabalhos de Acharya *et al.* (2012), Resmini (2012), Borchardt (2013), Rocha *et al.* (2014) e Francis *et al.* (2014).

Em situações reais, é difícil encontrar aplicações cujos dados sejam linearmente separáveis, e isso se deve a diversos fatores, entre eles a presença de ruídos e *outliers* nos dados ou à própria natureza do problema, que pode ser não linear (Lorena e Carvalho, 2007).

As MVSs lidam com problemas não lineares mapeando o conjunto de treinamento do espaço original (X), para um novo espaço de maior dimensão (C). Seja $\Phi: X \rightarrow C$ um mapeamento, a escolha apropriada de Φ faz com que o conjunto de treinamento

mapeado em C possa ser separado por uma MVS linear. Essa classificação realizada pelas MVSs baseia-se na separação por hiperplanos (Borchardt, 2013).

CAPÍTULO 4 - METODOLOGIA

Neste capítulo serão apresentadas as etapas propostas para a classificação de mamas doentes e saudáveis a partir de termografias dinâmicas. A figura 4.1 apresenta um resumo da metodologia que se baseia nas seguintes etapas: seleção das imagens térmicas sequenciais do banco de dados, extração das matrizes de temperatura das imagens térmicas, geração das imagens em escala de cinza a partir das matrizes de temperatura, segmentação das regiões de interesse (*region of interest* - ROI), extração de características das ROIs, geração das séries temporais correspondentes, geração das subséries com diferentes cardinalidades, extração de características das séries temporais geradas nas etapas anteriores, seleção de características, classificação e, por fim, o cálculo das métricas de avaliação da etapa de classificação.

4.1 SELEÇÃO DOS DADOS TÉRMICOS NO BANCO DE DADOS

A primeira etapa da metodologia consiste na seleção dos dados térmicos de pacientes saudáveis e de pacientes diagnosticadas com anomalias na mama. Os dados foram obtidos do *Database for Mastology Research with Infrared Image* - DMR-IR, acessível no endereço eletrônico <http://visual.ic.uff.br/dmi>. Este banco de imagens é destinado ao gerenciamento e recuperação de informações de exames de mama e de dados clínicos de pacientes voluntárias. As informações armazenadas no banco de dados foram adquiridas no Hospital Universitário Antônio Pedro (HUAP) em Niterói.

A aquisição destes dados e a pesquisa foram aprovados pelo Comitê de Ética em Pesquisa do HUAP em 04/06/2012 e está registrado na Plataforma Brasil, sob o número de Certificado de Apresentação para Apreciação Ética 01042812.0.0000.5243 do Ministério da Saúde do Brasil.

Os dados utilizados nesta tese foram adquiridos pelo protocolo dinâmico. Neste protocolo, as mamas das pacientes sofreram um resfriamento induzido por uma corrente de ar e a captura dos dados inicia-se durante o processo de retorno ao equilíbrio térmico. O resfriamento acontece até que a região central do tórax atinja a temperatura média de 30,5°C ou tenham se passado 5 minutos de resfriamento. Após isso, a corrente de ar é interrompida e são capturados todos os dados térmicos de uma cena a cada 15 segundos durante 5 minutos, perfazendo um total de 20 quadros sequenciais. Assim, segundo o protocolo dinâmico de aquisição de dados térmicos,

para cada paciente, foram adquiridos $Q(i)$ quadros, onde i varia de 1 até 20. Este protocolo foi desenvolvido por Silva *et al.* (2013) a partir de testes e experimentos com diversos outros protocolos.

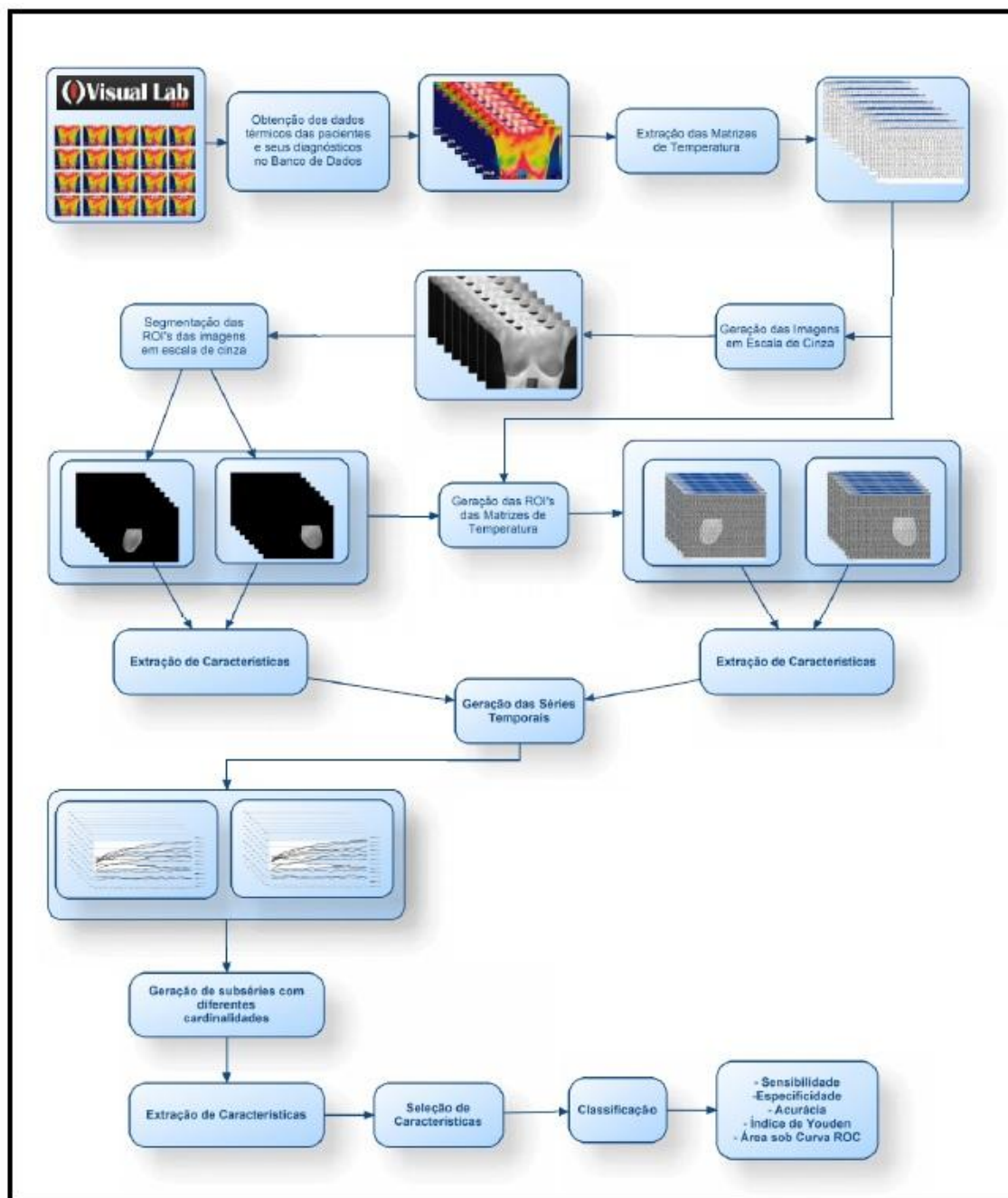


Figura 4.1 - Resumo da metodologia

Na figura 4.2, pode-se observar os 20 quadros sequenciais capturados de uma dada paciente. Em todos os quadros, as pacientes permaneceram com os braços sobre a cabeça e posicionadas frontalmente para a câmera térmica.

Para o desenvolvimento da metodologia, foram utilizados os quadros sequenciais de 64 mamas, sendo 32 doentes e 32 saudáveis. Nesta tese, foram consideradas mamas

saudáveis aquelas em que os achados na mamografia de rastreio foram classificados como BI-RADS de categoria 1 ou 2. Por outro lado, foram consideradas mamas doentes aquelas que possuíam alguma anomalia. As anomalias presentes nas mamas utilizadas nesta tese são: fibroadenomas, nódulos, tumor filodes de mama, carcinoma residual, carcinoma papilífero, carcinoma ductal infiltrante e carcinoma lobular *in situ*. O Apêndice A contém a descrição das anomalias de cada mama considerada doente, bem como a categoria BI-RADS das mamas saudáveis.

Também foram classificados os quadros sequenciais de mais 12 mamas que não participaram da etapa de desenvolvimento da metodologia, sendo 6 saudáveis e 6 doentes.

Para cada mama, 20 quadros sequenciais foram analisados, perfazendo um total de 1.280 quadros na etapa de desenvolvimento da metodologia e 240 na etapa de classificação das 12 mamas.

O processo de captura dos dados iniciou-se em 2012 e, no período de desenvolvimento desta tese, estavam disponíveis para serem utilizados os dados de 64 mamas. Após o desenvolvimento da metodologia, os dados de mais 12 mamas foram disponibilizados e utilizados como teste da metodologia proposta nesta tese.

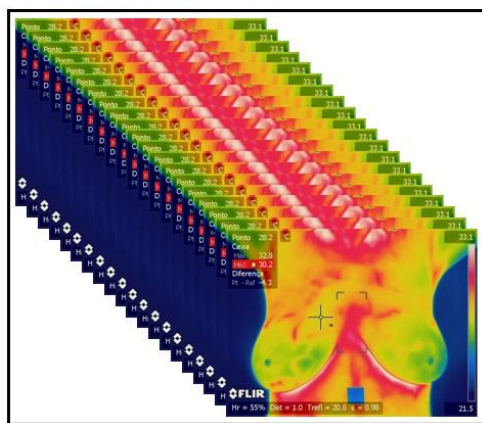


Figura 4.2 - Captura de 20 quadros sequenciais

A câmera térmica utilizada para a captura das cenas é da marca FLIR modelo SC620 (FLIR, 2013), que possui uma sensibilidade menor que 0,04°C e faixa de captura padrão de -40°C até 500°C. Ela possui 640 x 480 sensores térmicos, que realizam a captura simultaneamente. A câmera possui a capacidade de armazenamento de cenas sequenciais num intervalo mínimo de 10 segundos por até 24 horas.

4.2 EXTRAÇÃO DAS MATRIZES DE TEMPERATURA

Denomina-se matriz de temperatura o conjunto de temperaturas em graus Celsius capturado pela câmera térmica e apresentado em forma de arquivo texto.

As matrizes de temperaturas foram extraídas de cada uma das cenas térmicas. Assim, foram analisadas 64 mamas, com uma sequência de 20 cenas cada.

A figura 4.3 representa a extração da matriz de temperatura de uma dada paciente. Nela, é possível perceber a representação das temperaturas em pseudocores à esquerda e à direita a matriz que representa as temperaturas em graus Celsius.

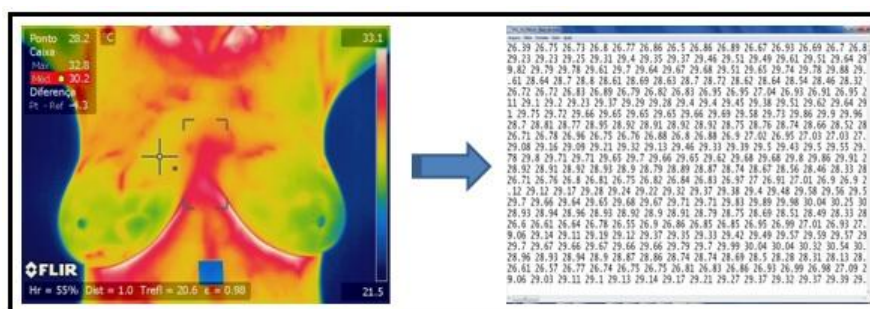


Figura 4.3 - Exemplo de extração da matriz de temperatura

Para a extração das matrizes, foi utilizado o aplicativo *QuickReport* (FLIR, 2009) disponibilizado pela FLIR, fabricante da câmera. Foi necessária a utilização do *QuickReport* já que as cenas geradas pela câmera térmica possuem um formato proprietário padrão da fabricante da câmera, denominado JPEG radiométrico.

Esta etapa foi necessária pois algumas das características se baseiam nas temperaturas em graus Celsius e não nas cenas quantizadas em escala de cinza. Pretende-se, assim, aumentar a precisão das informações extraídas das mamas, uma vez que, ao quantizar as matrizes para a geração das cenas em escala de cinza, informações poderiam ser perdidas.

4.3 GERAÇÃO DE IMAGENS EM ESCALA DE CINZA

Com as matrizes de temperatura extraídas de cada quadro sequencial das pacientes, foram geradas imagens em escala de cinza. A figura 4.4 ilustra uma imagem em escala de cinza gerada a partir da matriz de temperatura equivalente.

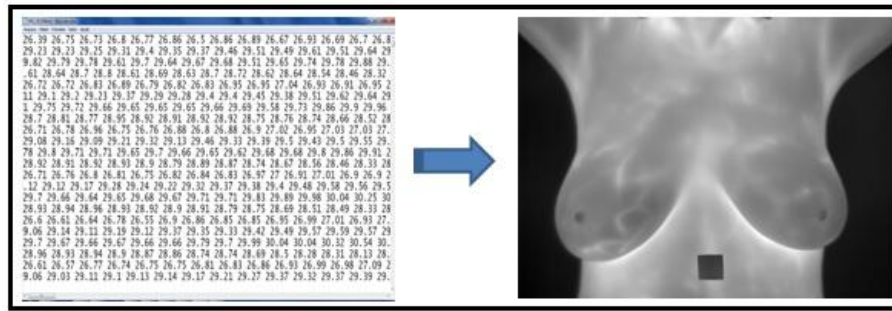


Figura 4.4: Conversão da matriz de temperatura para imagem em escala de cinza

Na etapa seguinte da metodologia houve a segmentação da região de interesse (ROI - *Region of Interest*), tornando necessária a geração das imagens em escala de cinza, uma vez que as análises realizadas nesta tese baseiam-se nas mamas das pacientes. Mesmo as análises que usam matrizes de temperatura utilizaram as ROIs em escala de cinza como indicativo de qual temperatura será considerada, a fim de determinar a ROI nas matrizes.

A equação 4.1 mostra como os tons de cinza foram gerados a partir das temperaturas de cada ponto.

$$value.val[i] = ((255 * (value.val[i] - min)) / (max - min)) \quad (4.1)$$

onde $value.val[i]$ é o valor do tom de cinza que foi atribuído ao pixel da imagem e min e max correspondem ao menor e maior valor de temperatura dentre todas as consideradas no desenvolvimento da metodologia. Cada valor de temperatura da matriz é utilizado na transformação para obtenção do tom de cinza que foi atribuído ao pixel correspondente da imagem em escala de cinza.

Para a obtenção do maior e menor valor de temperatura, foram percorridas todas as 1.280 matrizes das mamas utilizadas no desenvolvimento dessa metodologia. Assim, garantiu-se que as imagens fossem geradas na mesma escala e fosse possível a comparação entre elas.

4.4 SEGMENTAÇÃO DAS REGIÕES DE INTERESSE

ROI (Region of Interest) é uma área da imagem total que apresenta peculiaridades importantes a serem analisadas. A extração da ROI de uma imagem acontece por meio do processo de segmentação que a delimita em relação à imagem total.

Inicialmente, tentou-se segmentar as ROIs utilizando o segmentador automático de Marques (2012), porém dois motivos fizeram com que ele não se mostrasse adequado ao nosso trabalho. O primeiro é o fato desta segmentação basear-se nas regiões mais quentes encontradas nas imagens para delimitar as pregas inframamárias. Este fato, em muitos casos, não se confirmou uma vez que, como as pacientes são submetidas a um resfriamento através de uma corrente de ar, antes do início da captura das imagens, nem sempre as temperaturas mais elevadas se localizavam na posição das pregas inframamárias. Assim, acredita-se que este seja o motivo de apenas cerca de 10% das imagens de testes terem sido segmentadas corretamente por esse segmentador. O segundo fator determinante para que o segmentador desenvolvido não fosse adequado, é fato do autor considerar como região de interesse, além da região de formação do câncer (i.e. a mama), regiões de possível disseminação da doença. Assim, ele considera como ROI a união da região axilar com a região composta pela glândula mamária e o sistema vascular sanguíneo e linfático, conforme pode ser observado na figura 4.5.

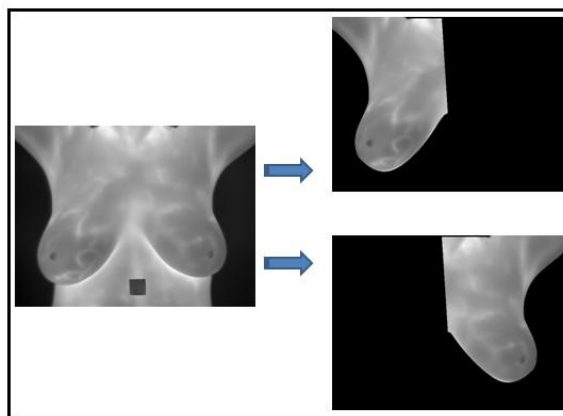


Imagem 4.5 - Exemplo de segmentação realizada pelo segmentador de Marques (2012).

No contexto desta tese, foram consideradas como região de interesse as áreas das imagens correspondentes às mamas das pacientes. A glândula mamária pode variar entre pessoas, porém, em média, localiza-se horizontalmente entre a segunda e terceira costela até a sexta ou sétima costela (Conceição, 2006 p.33). Tais limites podem ser observados na figura 4.6. Em alguns casos, a ptose da mama pode apresentar-se de forma que o conjunto mamilo-auréola pode ultrapassar a dobra inframamária, chegando até a altura do umbigo. Nesses casos, a segmentação considerou toda a mama, independente do grau de ptose da paciente. Já verticalmente, a mama limita-se entre a borda do osso esterno e a linha axilar média (Conceição, 2006 p.33).

Nesta tese, como apenas a região da mama é considerada como região de interesse, todas as 1.280 imagens (64 mamas, cada uma com 20 imagens sequenciais) utilizadas na etapa de desenvolvimento da metodologia e as 240 imagens (12 mamas, cada uma com 20 imagens sequenciais) utilizadas na etapa de teste foram segmentadas manualmente usando a ferramenta GIMP, que é um *software* livre de edição de imagens. Durante este processo, três pessoas foram responsáveis pelas segmentações. Duas delas foram responsáveis pelas segmentações manuais, já a terceira verificava se as segmentações foram feitas de forma adequada. Caso não estivessem, as segmentações eram repetidas. Todas as imagens segmentadas estão disponibilizadas no site do Visual Lab, no endereço eletrônico <http://visual.ic.uff.br/proeng/thiagoelias/>.

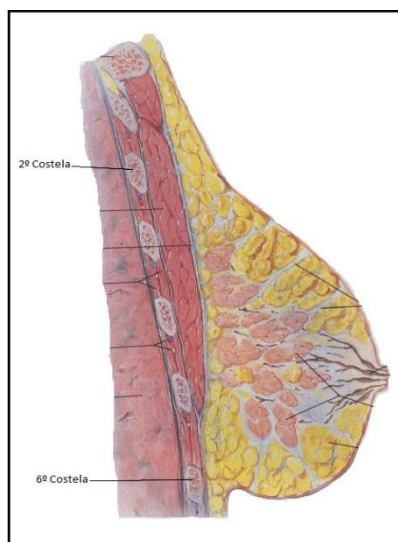


Figura 4.6 - Limites superior e inferior da mama.

Fonte: Netter (2011 p.167)

A figura 4.7 apresenta um exemplo de segmentação da ROI realizado na imagem em escala de cinza de forma manual. Foram utilizados para a segmentação os limites superior, inferior e laterais discutidos anteriormente. Além das referências ósseas, foram utilizadas referências visuais externas para determinar a região de interesse.

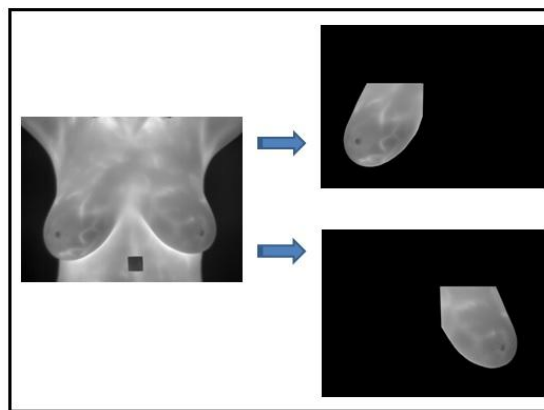


Figura 4.7 - Exemplo de mamas segmentadas manualmente

Além da ROI das imagens em escala de cinza, também foi determinada a ROI das matrizes de temperatura. Para isto, a partir da correspondência entre cada pixel das imagens em escala de cinza e cada ponto das matrizes de temperatura sequenciais, foi possível determinar ambas as ROIs. A figura 4.8 exemplifica a utilização da ROI em escala de cinza na determinação da ROI da matriz de temperatura. Apenas as temperaturas correspondentes à mama da paciente são consideradas.

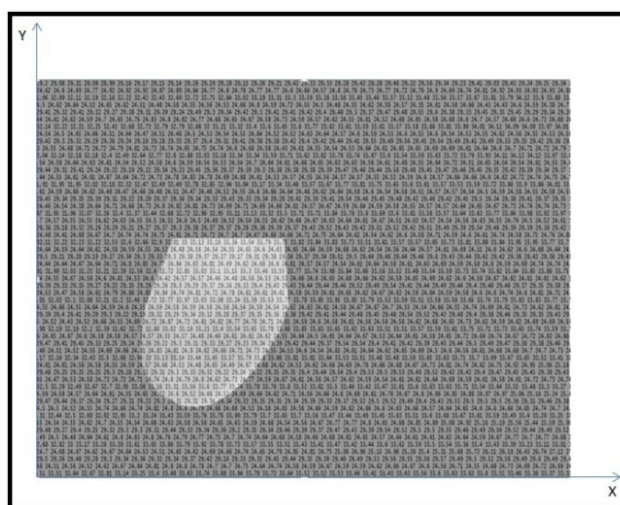


Figura 4.8 - Determinação da ROI da matriz de temperatura a partir da ROI em escala de cinza

4.5 EXTRAÇÃO DE CARACTERÍSTICAS

Esta etapa foi dividida em duas outras etapas, a serem explicadas nos próximos parágrafos.

A primeira etapa corresponde à extração de características das ROIs quantizadas em escala de cinza ou das ROIs das matrizes de temperatura. Cada mama, conforme explicado nas seções anteriores, possui $Q(i)$ imagens ou matrizes, onde i varia de 1 até

20. Assim, as características foram extraídas de cada uma das ROIs. Para algumas características, foram obtidos 20 valores, ou seja, um valor para cada imagem i , enquanto que, para as características baseadas em comparação de ROIs, foram obtidos 19 valores. A figura 4.9 apresenta um esquema da extração de características nesses dois casos.

A figura 4.9(a) exemplifica o caso da característica depender de uma única ROI, enquanto que a figura 4.9(b) o caso em que a característica depende de duas ROIs consecutivas: da ROI i e da ROI $i+1$. Os valores de todas as características servirão de base para a etapa seguinte de extração de característica das séries temporais.

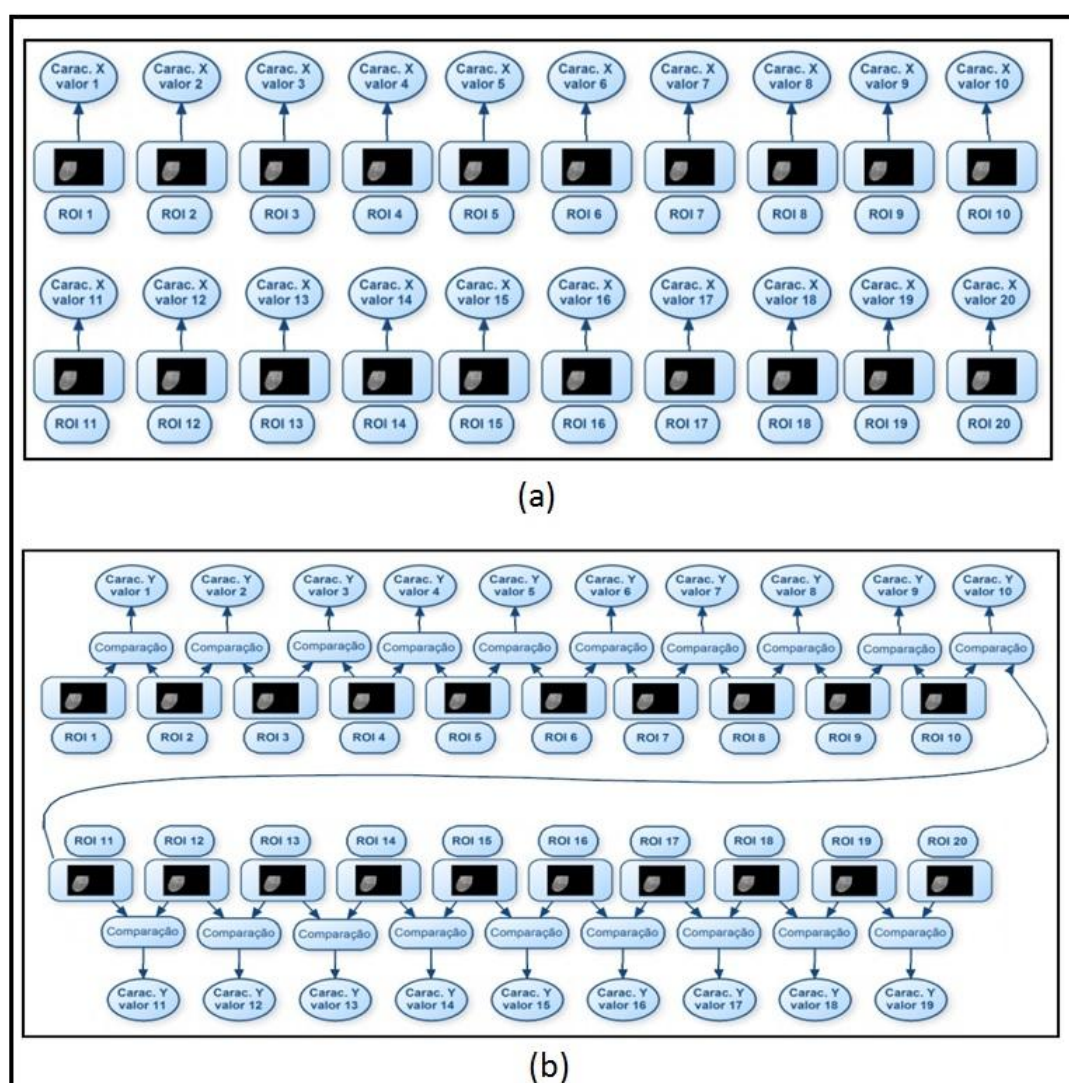


Figura 4.9 - Representação da extração de características. (a) uma característica por ROI; (b) baseada em comparação entre ROIs

A segunda etapa corresponde à extração de características das séries temporais geradas a partir dos valores das características extraídas na etapa anterior. Para a construção das séries, foram considerados os valores das características de cada mama. A figura 4.10 apresenta um exemplo das séries temporais da característica temperatura média de algumas ROIs de temperatura. Foram utilizadas linhas em vermelho para representar as séries das mamas com diagnóstico de câncer e linhas em azul para representar as séries das mamas saudáveis. No eixo horizontal são representadas as 20 ROIs de temperatura capturadas pelo protocolo dinâmico. Já no eixo vertical, estão representados os valores das temperaturas médias destas ROIs.

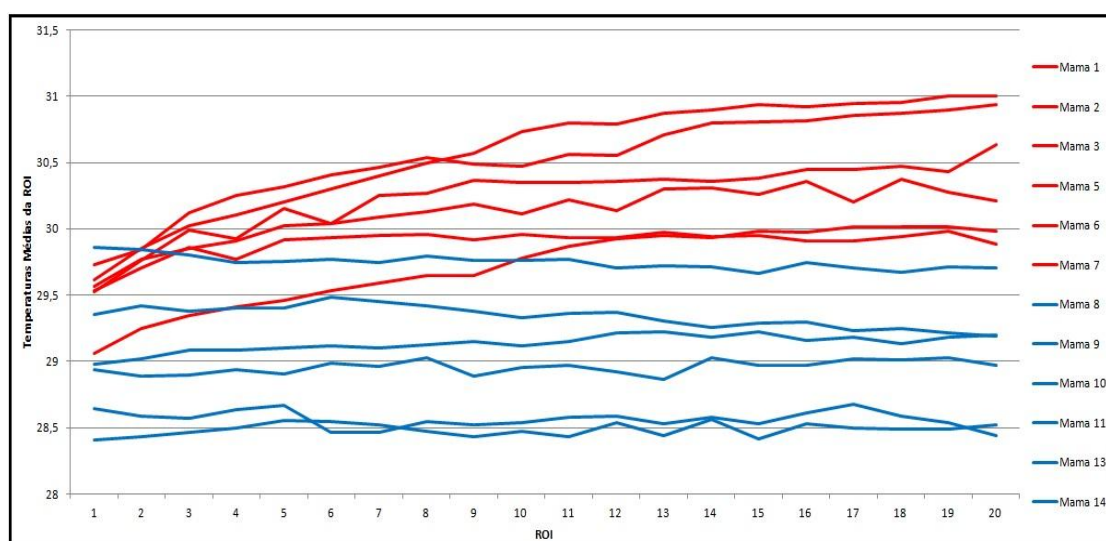


Figura 4.10 - Séries temporais das médias de temperatura de algumas mamas doentes e saudáveis

As características extraídas das séries temporais formaram o vetor de características que foi utilizado na etapa de classificação.

Nas próximas duas seções serão explicadas detalhadamente as duas formas de extração de características.

4.5.1 ETAPA DE EXTRAÇÃO DE CARACTERÍSTICAS DAS ROIs

As características extraídas nesta tese foram organizadas em grupos. Cada grupo é formado por uma ou mais características, conforme pode ser observado nas seções seguintes.

Tais características serviram, conforme exposto na seção anterior, para a construção das séries temporais da etapa seguinte.

4.5.1.1 CARACTERÍSTICAS BASEADAS EM ESTATÍSTICA SIMPLES

Este grupo de características é formado pela média das temperaturas e pelo desvio padrão das ROIs de temperatura.

Para cada uma das 20 ROIs das matrizes de temperatura sequenciais de uma dada mama foi calculada a média de temperatura e o desvio padrão, gerando 20 valores para cada uma das características.

4.5.1.2 CARACTERÍSTICA BASEADA EM AGRUPAMENTO

Esta característica foi extraída das ROIs de temperatura. Para a extração desta característica, usou-se o algoritmo de agrupamento K-means proposto por MacQueen (1967), implementado na biblioteca *OpenCV* através da função *kmeans()*, onde o agrupamento inicia-se com a escolha do número de sementes. Nesta tese, foram utilizadas 9 sementes a fim de se chegar a 9 grupos. A escolha deste número deu-se a partir de testes que demonstraram que o número de 9 sementes proporciona os melhores resultados. Foram testadas as quantidades de 3, 6, 9, 12, e 15 sementes. Alguns parâmetros da função *kmeans()* do *OpenCV* foram configurados para atender as necessidades da pesquisa. O primeiro deles foi a forma da escolha das 9 sementes iniciais para a inicialização do algoritmo. Esta escolha foi feita a partir da proposta de Arthur e Vassilvitskii (2007) que consiste em um algoritmo de aproximação para o problema k-means NP-difícil. Isto garantiu que, independente do número de vezes que a função fosse executada, sempre as mesmas sementes seriam selecionadas.

A função, após determinar as sementes, escolhe cada uma das demais temperaturas aleatoriamente e as agrupa pelo critério de menor distância euclidiana ao centroide do grupo, recalculando-o à medida que novos membros são inseridos. Este procedimento foi repetido até que todas as temperaturas fossem agrupadas. Concluída esta etapa, cada elemento de um dado grupo foi reposicionado em outro grupo a fim de verificar se este novo posicionamento gerava um grau de homogeneidade interna melhor do que o grupo gerado aleatoriamente. A função obtém o grau de homogeneidade a partir da Soma de Quadrados Residuais, obtido pela equação 4.2:

$$SQRes(j) = \sum_{i=1}^{n_j} d^2(o_i(j); \bar{o}(j)) \quad (4.2)$$

onde $o_i(j)$, $\bar{o}(j)$ e n_i são respectivamente os valores do i -ésimo dado, o centro do grupo j e o número de dados agrupados deste grupo.

Configurou-se a função para realizar 1000 vezes o movimento de elementos entre grupos para tentar obter a maior homogeneidade interna, com uma precisão de 0,01.

Concluída a etapa de agrupamento de cada ROI da matriz de temperatura, identificou-se o centroide de maior valor de temperatura dentre os nove. Este valor do maior centroide foi utilizado como a característica da ROI da matriz de temperatura. Assim, ao final, cada mama obteve 20 valores, um para cada ROI.

A figura 4.11 apresenta, através da ROI em escala de cinza, os 9 agrupamentos explicados anteriormente. À esquerda, é apresentada uma ROI sem agrupamento dos pixels e à direita, a mesma ROI após passar pelo processo de agrupamento em nove grupos. Percebe-se, na imagem à direita, uma região de tom mais claro caracterizando o grupo de maior valor de centroide, ou seja, a região de maior valor médio de temperatura.

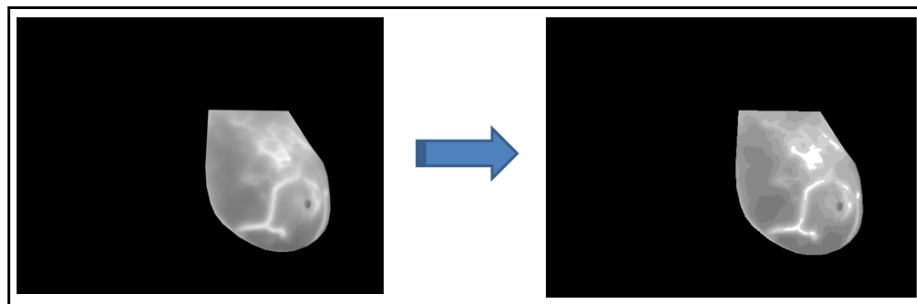


Figura 4.11 - ROI em tons de cinza com pixels agrupados em nove grupos

A média das temperaturas do grupo de maior temperatura (em branco na figura 4.11) foi utilizada como característica extraída desta ROI.

4.5.1.3 CARACTERÍSTICAS BASEADAS EM COMPARAÇÃO DE HISTOGRAMAS

Os histogramas foram construídos para cada ROI em escala de cinza variando entre 0 e 255 tons. Assim, para uma dada mama, foram construídos 20 histogramas.

Todos os histogramas foram normalizados, ou seja, os valores foram divididos pelo número de pixels de cada ROI. Esta etapa se fez necessária uma vez que as ROIs, mesmo que pertencendo à mesma mama, não necessariamente possuíam o mesmo número de pixels.

Para a extração das características baseadas em comparação de histogramas, são necessários 2 histogramas para gerar um único valor. Nesta tese, a comparação é feita entre os histogramas subsequentes, ou seja, o *iésimo* histograma foi comparado com o *iésimo+1*. Desta forma, ao final, para cada característica baseada em comparação de histograma de uma dada mama, obtiveram-se 19 valores, e não 20, como obtiveram-se nos grupos anteriores. A figura 4.9(b) representa a extração desta característica.

As características de comparação de histogramas extraídas, conforme explicadas no capítulo 2, foram: Distância de Battacharyya, Qui-quadrado e Interseção.

4.5.1.4 CARACTERÍSTICAS BASEADAS EM TEXTURA

No contexto desta tese, foram analisadas as 6 características apresentadas por Baraldi e Parmiggiani (1995) a partir dos histogramas da soma e da diferença, conforme proposto por Unser (1986), para o estudo da textura das ROIs em escala de cinza. Essas características são: energia, entropia, contraste, variância, correlação e homogeneidade. Para cada ROI em escala de cinza, foram extraídas as 6 características citadas anteriormente. A figura 4.12 representa a extração destas características. Para cada uma das 20 ROIs foram gerados os histogramas da soma e da diferença e, a partir deles, calcularam-se as 6 características.

Os elementos do histograma da soma variam de 0 a 510, enquanto que os do histograma da diferença variam de -255 a 255.

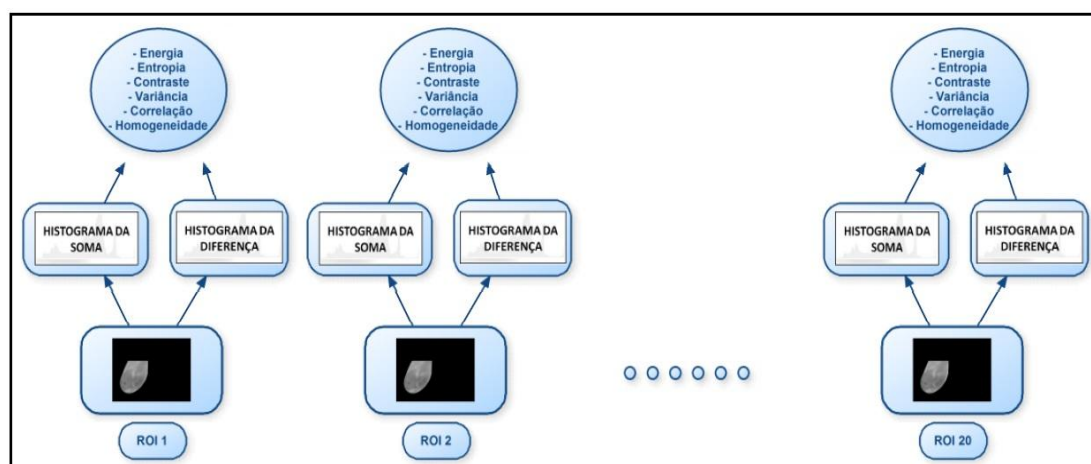


Figura 4.12 - Extração das características de textura a partir dos histogramas da soma e da diferença das ROIs em escala de cinza de uma dada mama.

Para a geração dos histogramas da soma e da diferença normalizados, foram consideradas duas vizinhanças. A primeira com $dx=1$ e $dy=0$, denominada nesta tese de análise horizontal; e a segunda com $dx=0$ e $dy=1$, denominada de análise vertical. Assim, o processo representado pela figura 4.12 foi realizado duas vezes, um para cada vizinhança (vertical e horizontal).

4.5.1.5 CARACTERÍSTICAS BASEADAS EM ÍNDICES DE DIVERSIDADE

Os índices de diversidade ecológicos também foram utilizados para analisar as 20 ROIs em escala de cinza sequenciais de cada mama desta tese. Foram utilizados os seguintes índices explicados no capítulo 2: índice de McIntosh, índice de Simpson e índice de Berger-Parker.

Nesta tese, os índices de diversidade ecológica foram calculados a partir dos histogramas da soma e da diferença, utilizando as duas vizinhanças anteriormente explicadas. Para isso, uma vez que os histogramas da soma e da diferença possuem 511 bins (histograma da soma varia de 0 a 510 e o da diferença varia de -255 a 255), cada um deles foi transformado em um novo histograma com 73 bins. Cada grupo de 7 bins dos histogramas originais foram agrupados e suas frequências somadas para dar origem ao novo bin, ou seja, as frequências dos bins de 0 a 6 do histograma da soma foram somadas para comporem o primeiro bin do novo histograma; as frequências dos elementos de 7 a 13 também foram somadas para formarem o segundo bin do novo histograma, e assim por diante até o bin 510. O mesmo foi feito para o histograma da diferença, iniciando-se de -255 até 255.

Criados os novos histogramas, passou-se a fazer um outro agrupamento. A população foi formada por todos os elementos do primeiro histograma mais todos do segundo histograma. Cada um deles foi considerado como uma espécie da população, e o número de indivíduos de cada espécie é o valor da frequência correspondente. Assim, ao final, os dois histogramas de cada ROI compuseram uma população de 146 espécies (73 geradas a partir do histograma da soma e 73 geradas a partir do histograma da diferença).

Dessas 146 espécies com seus respectivos indivíduos de cada uma das 20 ROIs de cada mama, foram calculados os índices de McIntosh, índice de Simpson e índice de Berger-Parker.

A figura 4.13 representa o processo de cálculo dos índices de diversidade a partir dos histogramas da soma e da diferença das ROIs sequenciais em escala de cinza de uma dada mama.

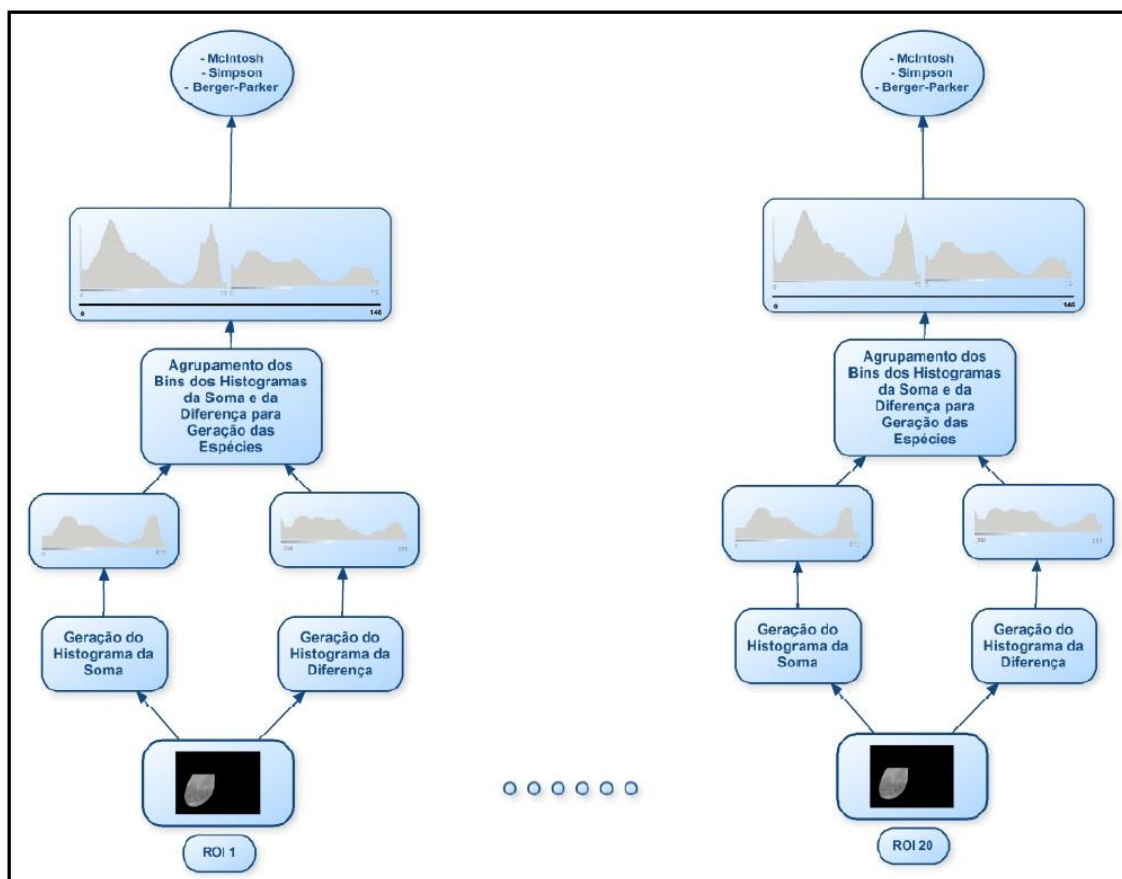


Figura 4.13 - Extração dos índices de diversidade a partir dos histogramas da soma e da diferença das ROIs de uma dada mama.

Os valores dos índices de diversidade das ROIs de uma dada mama são as características extraídas nesta etapa da metodologia.

Assim como as características da seção anterior que também se basearam em histogramas da soma e da diferença, as características baseadas em índices de diversidade também se basearam em histogramas da soma e da diferença com diferentes vizinhanças. Foram realizadas duas abordagens: na primeira abordagem foram considerados os histogramas da soma e da diferença com vizinhança $dx=1$ e $dy=0$; na segunda abordagem foram considerados os mesmos histogramas, porém com vizinhança $dx=0$ e $dy=1$.

4.5.1.6 CARACTERÍSTICA BASEADA EM ESTATÍSTICA ESPACIAL

Assim como histogramas, os Spatiograms também capturaram momentos de ordem superior. A segunda ordem de um spatiogram de um objeto é idêntico a um histograma de suas características, exceto que ele também armazena informação espacial adicional, ou seja, a média e covariância da posição espacial de todos os pixels que caem em cada subdivisão do histograma (Conaire *et al.*, 2007).

Nesta tese, foram realizadas comparações dos spatiograms das ROIs sequenciais em escala de cinza das mamas das pacientes. Assim como na comparação de histogramas, o spatiogram da i -ésima imagem foi comparado ao spatiogram da i -ésima+1 imagem. Assim, ao final, foram obtidos 19 valores de comparação de spatiograms para cada mama. A figura 4.9(b) representa a comparação dos spatiograms das ROIs sequenciais em escala de cinza de uma dada mama.

A comparação de spatiograms, por ser uma característica de estatística espacial, ou seja, levar em conta o posicionamento de cada pixel na imagem, exige que seja feito o registro das ROIs sequenciais em escala de cinza de cada mama. A técnica utilizada para o registro foi desenvolvida por Galvão (2015).

Galvão (2015) utiliza a transformação de Möbius para o mapeamento entre as imagens térmicas das pacientes. Ela é uma função pouco explorada nos métodos de registro de imagens. Inicialmente, as imagens a serem registradas são classificadas em imagem de referência e imagem sensível. Após isso, ambas as imagens são processadas de forma que o contorno do corpo da paciente seja reconhecido. Através do contorno, algumas características das imagens são identificadas. Em seguida, essas características são utilizadas nos cálculos das duas transformações realizadas. Com essas transformações, duas novas imagens são criadas, uma por cada tipo de transformação. Por fim, as imagens geradas são comparadas com a imagem de referência e a imagem que apresentou o melhor resultado é escolhida como saída do método de registro.

4.5.1.7 CARACTERÍSTICA BASEADA EM LACUNARIDADE

No contexto desta tese, apesar de existirem diversos autores que propõem métodos diferentes de calcular a lacunaridade, optou-se por utilizar o método proposto por Du e Yeo (2002), utilizando-se das ROIs sequenciais em escala de cinza das mamas das pacientes. Os detalhes deste método foram explicados no capítulo 2. O motivo pelo qual optou-se em utilizar imagens em escala de cinza ao invés das imagens binarizadas,

que possuem um custo computacional inferior, deve-se ao processo de binarização que provoca perda de informação nas imagens, comprometendo a qualidade dos resultados. Nesta tese, foram utilizadas as seguintes dimensões em pixels de janelas deslizantes: 2x2, 4x4, 8x8, 16x16, 32x32 e 64x64. Vale ressaltar que mesmo para as ROIs sequenciais em escala de cinza de uma mesma mama, o número de janelas deslizantes pode variar de uma ROI para a outra, uma vez que elas podem variar em quantidade de pixels.

As ROIs em escala de cinza das mamas possuem um formato curvo. Por este motivo, foram consideradas apenas as janelas deslizantes contidas completamente nas ROIs, ou seja, aquelas janelas que não ficaram contidas na sua totalidade nas ROIs em escala de cinza, foram desconsideradas no cálculo da lacunaridade. A figura 4.14 ilustra um exemplo de uma janela deslizante considerada no cálculo da lacunaridade, representada na cor azul; e uma janela que foi desconsiderada, representada em vermelho.

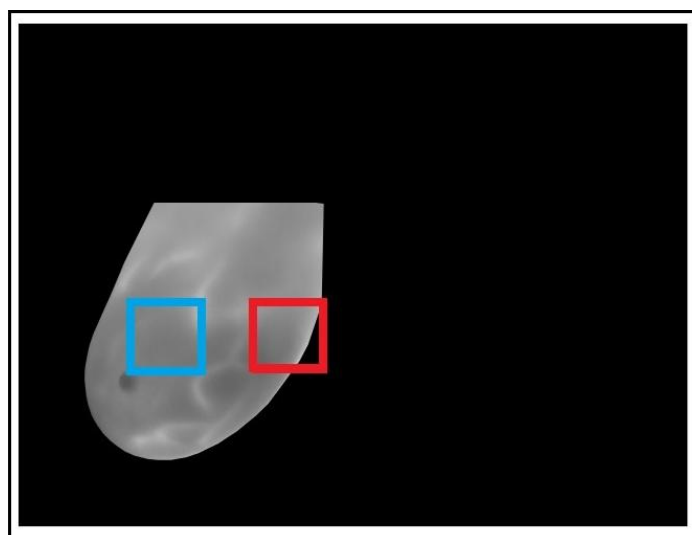


Figura 4.14 - Exemplo de janela considerada no cálculo da lacunaridade em azul e janela não considerada no cálculo da lacunaridade em vermelho.

Ao final, foi obtido um valor de lacunaridade para cada uma das 20 ROIs sequencias em escala de cinza de uma dada mama. A figura 4.9(a) representa a obtenção da lacunaridade para as 20 ROIs em escala de cinza de uma dada mama.

4.5.2 GERAÇÃO DAS SÉRIES TEMPORAIS

Uma série temporal pode ser vista como uma sequência de pontos de dados numéricos ordenados sucessivamente, ocorrendo geralmente em intervalos uniformes. Em outras

palavras, uma série temporal é uma sequência de dados obtidos em intervalos regulares de tempo durante um período específico. Na análise de uma série temporal, primeiramente deseja-se modelar o fenômeno estudado para, a partir daí, descrever o comportamento da série (Latorre e Cardoso, 2001).

Nesta tese, para cada característica extraída das ROIs sequenciais das imagens em escala de cinza ou das ROIs sequenciais das matrizes de temperatura de uma dada mama, foi gerada uma série temporal correspondente. Em outras palavras, uma dada característica foi extraída de cada uma das 20 ROIs sequenciais das imagens em escala de cinza ou das matrizes de temperatura de uma mama. Os 20 valores obtidos formaram uma série temporal que representa a variação desta característica ao longo das 20 ROIs sequenciais da mama. Além disso, utilizando-se como base a proposta de Higuchi (1988), um conjunto de subséries de cardinalidades diferentes também foi gerado e igualmente analisado numa etapa posterior.

Dada uma série unidimensional de N valores $X = \{x(1), x(2), \dots, x(N)\}$, foram geradas k novas séries X_k^m definidas pela equação 4.3.

$$X_k^m = \{x(m), x(m+k), x(m+2k), \dots, x\left(m + \left\lfloor \frac{N-m}{k} \right\rfloor k\right)\} \quad (4.3)$$

onde k e m são inteiros, k representa o intervalo discreto entre os pontos e m indica o valor inicial da série analisada sendo $m = (0, 1, \dots, (k-1))$. O comprimento de cada nova série gerada é calculada através da Equação 4.4.

$$L(m, k) = \left\{ \frac{\sum_{i=1}^{\left\lfloor \frac{N-m}{k} \right\rfloor} |x(m+ik) - x(m+(i-1)k)| (N-1)}{\left(\left\lfloor \frac{N-m}{k} \right\rfloor\right) k} \right\} / k \quad (4.4)$$

onde $(N-1)/(\left\lfloor (N-m)/k \right\rfloor)k$ é o fator de normalização da série. Nesta tese, considerou-se k variando entre 1 e 4. À medida que varia-se k , a subsérie gerada apresenta-se com uma cardinalidade diferente, ou seja, um subconjunto dos 20 valores das características extraídas na seção anterior é considerado para a construção da série. Foram geradas 10 séries com diferentes cardinalidades para cada característica proposta nesta tese, conforme pode ser observado na tabela 4.1.

Tabela 4.1 - Cardinalidade das séries temporais geradas para cada característica

Série	Valor de K	Valor de m
1	1	0
2	2	0
3	2	1
4	3	0
5	3	1
6	3	2
7	4	0
8	4	1
9	4	2
10	4	3

A figura 4.15 apresenta um exemplo das 10 possíveis subséries geradas da série temporal de uma dada característica de duas mamas, variando o "k" e o "m", conforme explicado anteriormente. A série em vermelho representa a característica extraída das ROIs sequenciais de uma mama doente. Já a série em azul representa esta mesma característica extraída das ROIs sequenciais de uma mama saudável. No eixo horizontal são representadas cada uma das ROIs sequenciais e no eixo vertical o valor da característica.

Observa-se na figura 4.15 que, quando $k=1$, todos os 20 valores obtidos da característica são usados para a construção da série temporal; quando $k=2$, dois subconjuntos dos 20 valores de característica são gerados para a construção das respectivas séries, ambos com 10 valores; quando $k=3$, três subconjuntos dos 20 valores da característica são gerados para a construção da série temporal; e por fim, quando $k=4$, quatro subconjuntos dos 20 valores da característica são gerados para a construção das séries temporais correspondentes.

O principal objetivo de variar-se o "k" e o "m", gerando novas cardinalidades para as séries, é eliminar os *outliers* que podem vir a surgir e otimizar a etapa seguinte da metodologia, que é a classificação. Isto pode ser observado na primeira das séries na figura 4.15 ($k=1$, $m=0$). Não há uma separação clara entre as séries da característica das ROIs das mamas doente e saudável. Por outro lado, observa-se que quando $k=4$, $m=3$ é clara a separação entre as séries das ROIs da mama doente e da mama saudável.

Este mesmo procedimento de geração de subséries variando a sua combinação foi repetido para todas as características extraídas das ROIs sequenciais em escala de cinza ou das ROIs das matrizes de temperatura de cada mama analisada nesta tese.

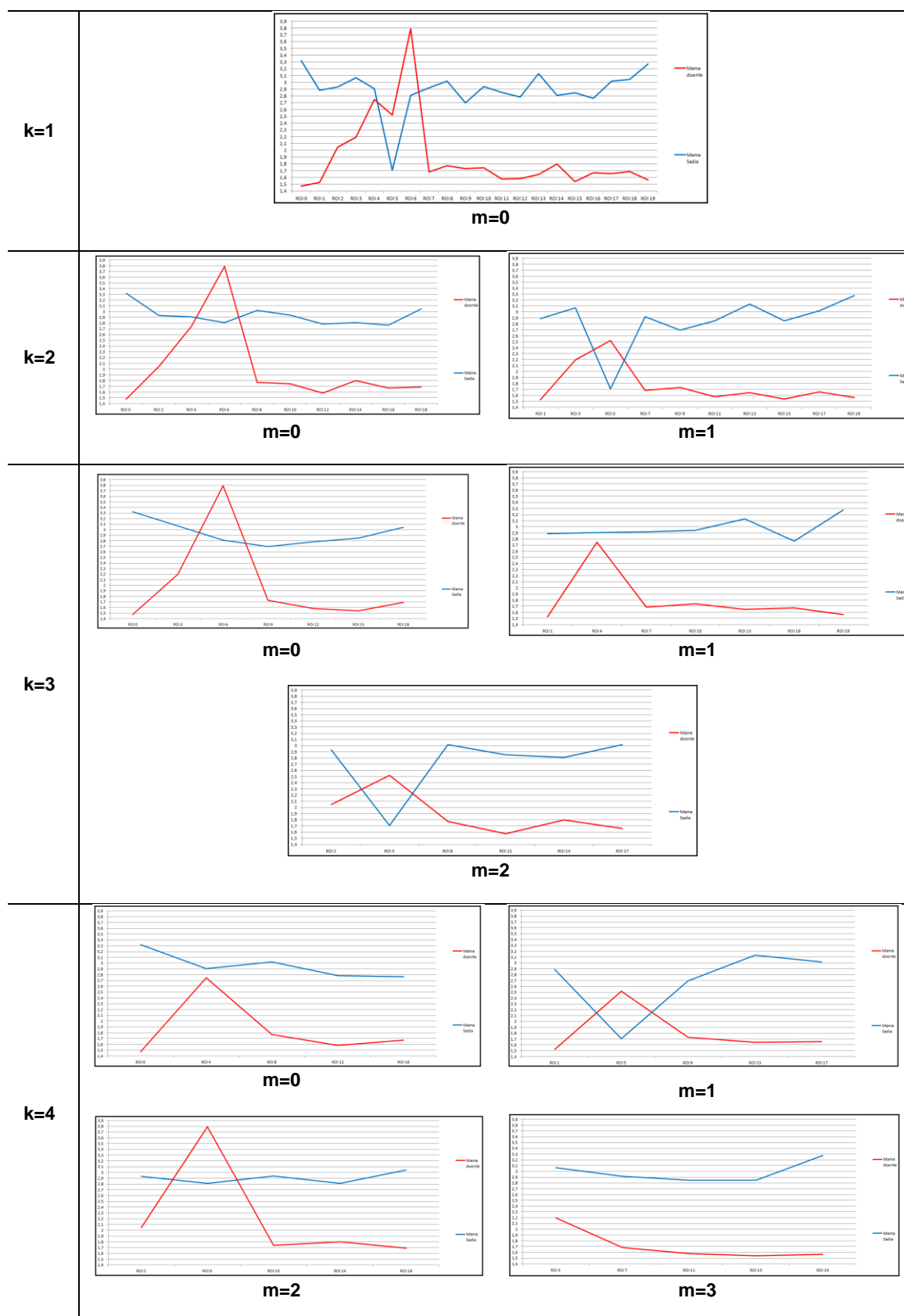


Figura 4.15 - Séries temporais de duas mamãs (uma sadia e outra doente) com diferentes cardinalidades.

4.5.3 EXTRAÇÃO DE CARACTERÍSTICAS DAS SÉRIES TEMPORAIS

Após serem criadas as séries temporais de todas as características das ROIs sequenciais de cada mama, iniciou-se o processo de extração de características dessas séries. Antes, as séries foram divididas pelos grupos de características determinados na seção 4.5.1. Depois, cada grupo foi subdividido pelas cardinalidades das séries, ou seja, a partir dos valores de k e m . Assim, grupos de séries, como por exemplo da figura 4.16, foram formados.

A figura 4.16 apresenta as diferentes formas das séries do grupo de características baseado em comparação de histogramas, explicado na seção 4.5.1, de apenas duas mamas: uma saudável e outra doente nas cores azul e vermelha, respectivamente. O grupo de característica em questão é formado por três características: comparação de histogramas baseada em interseção, qui-quadrado e distância de Bhattacharyya. Assim, para cada mama, existem 3 séries temporais representando cada uma das 3 características do grupo. As séries representadas com a linha contínua representam a interseção entre os histogramas, as séries tracejadas representam a comparação por qui-quadrado dos histogramas e as séries pontilhadas representam a distância de Bhattacharyya entre os histogramas.

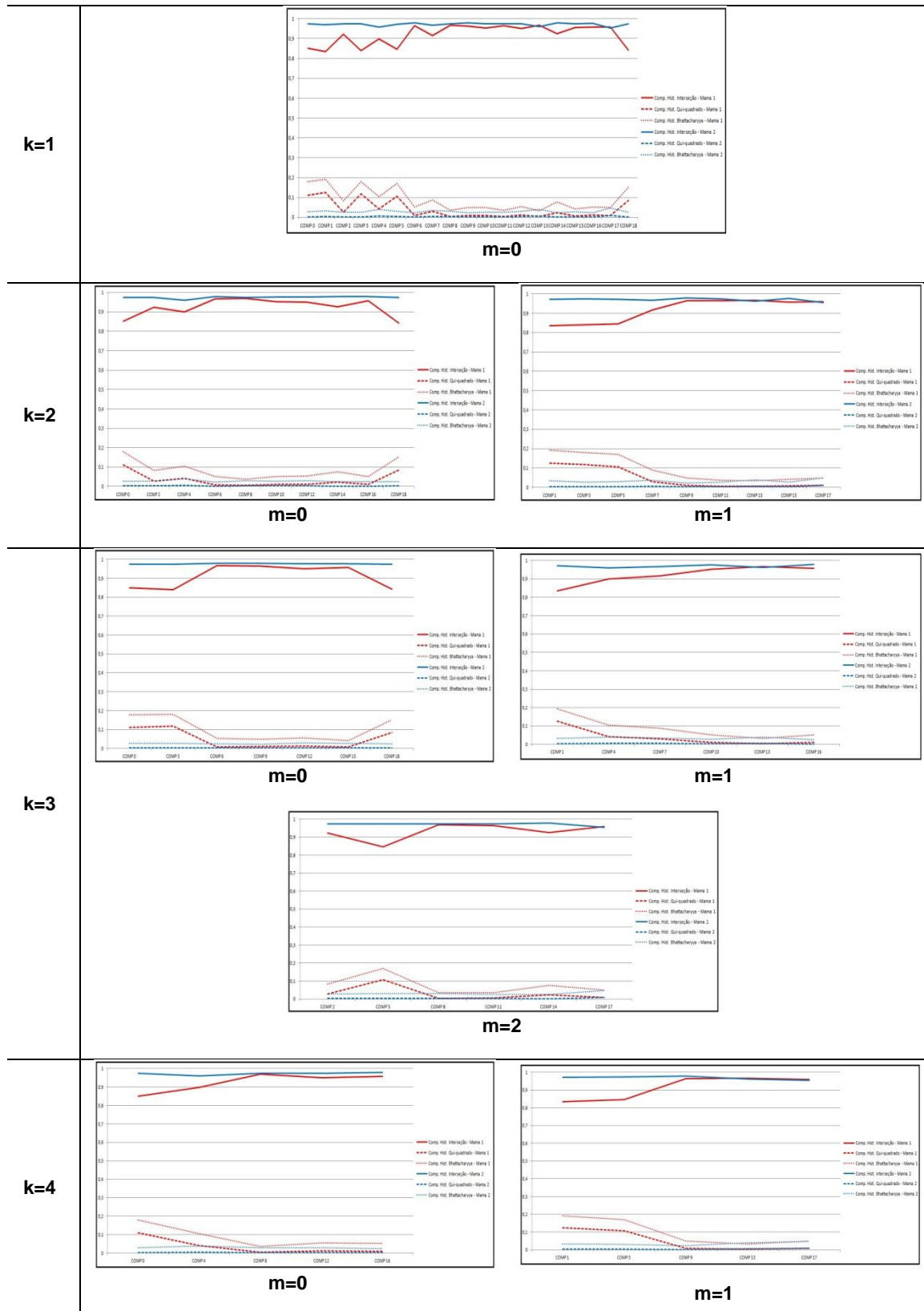
Após a organização das séries temporais por grupo de características e por forma de composição, iniciou-se a segunda etapa do processo de extração de características, que corresponde à extração de características das séries temporais.

De cada série temporal, foram extraídas duas características. A primeira corresponde à amplitude (A) da série, que é a diferença entre o maior e menor valor que compõe a série temporal. A segunda característica corresponde a raiz quadrada da soma ao quadrado (R) da série temporal. Elas são definidas a partir das equações 4.5 e 4.6, respectivamente.

$$A = \text{Max}(x_i) - \text{Min}(x_i) \quad (4.5)$$

$$R = \sqrt{\frac{\sum_{i=1}^n x_i^2}{n}} \quad (4.6)$$

onde x corresponde aos valores que compõem as séries temporais e n é a quantidade de valores da série.



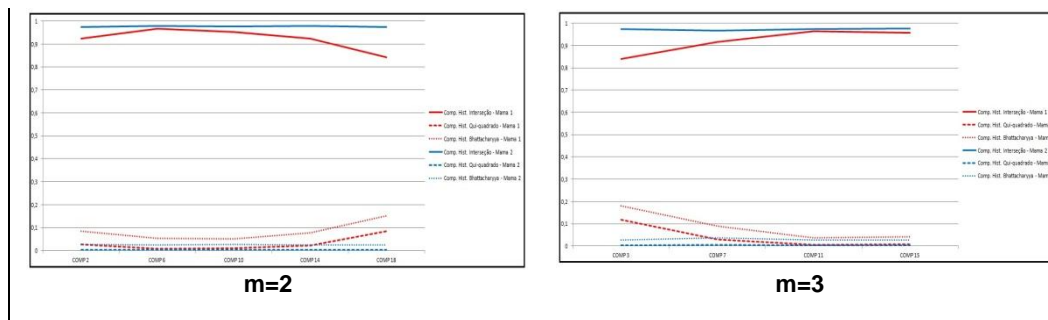


Figura 4.16 - Séries temporais do grupo baseado em comparação de histogramas. Exemplo com duas mamas.

Assim, usando como exemplo as séries temporais com $k=1$, $m=0$ do grupo de características baseadas em comparação de histogramas, exposto na figura 4.16, para cada mama, obteve-se um total de 6 características extraídas. Em outras palavras, foram extraídas duas características (amplitude e raiz quadrada da soma ao quadrado) de cada uma das séries temporais que representam cada uma das 3 comparações de histogramas das ROIs em escala de cinza.

A tabela 4.2 apresenta uma relação entre as duas etapas de extração de características. A primeira que corresponde à extração das ROIs sequenciais das imagens em escala de cinza ou das matrizes de temperatura e a segunda que corresponde à extração de características das séries temporais geradas a partir das características extraídas na etapa anterior.

Foram as características extraídas das séries temporais que compuseram os vetores usados pelo classificador na etapa de classificação.

4.6 SELEÇÃO DE CARACTERÍSTICAS

Por vezes, o processo de extração de características gera redundâncias ou características irrelevantes. Estas podem comprometer o desempenho dos algoritmos na construção de modelos de predição. Desta forma, faz-se necessária a aplicação de técnicas de seleção de características, proporcionando uma maior acurácia aos modelos gerados.

A etapa de seleção de características foi inserida no desenvolvimento da metodologia com o objetivo de aumentar a eficácia do classificador. Em outras palavras, esta etapa objetiva diminuir a dimensionalidade do número de característica, encontrando um subconjunto que seja relacionado com a classe, mas que seja pouco correlacionado entre seus elementos.

No *WEKA*, a fim de selecionar atributos, é necessário especificar um método de pesquisa e um atributo avaliador. O atributo avaliador indica como um subconjunto de atributos é avaliado. Por exemplo, eles podem ser avaliados através da construção de um modelo e avaliação da precisão deste modelo. O atributo avaliador *default* do *WEKA* e o utilizado nesta tese é o *CfsSubsetEval* (Hall, 1999), que avalia o valor de um subconjunto de características, considerando a capacidade preditiva individual de cada um juntamente com o grau de redundância entre as características. Em outras palavras, ele avalia subconjuntos que se correlacionam altamente com o valor da classe e possuem baixa correlação entre si.

O método de pesquisa indica a maneira estruturada em que o espaço de busca de possíveis subconjuntos de atributos é navegado com base na avaliação dos subconjuntos. O método padrão de pesquisa no *WEKA*, utilizado nesta tese, é o *BestFirst*.

Tabela 4.2 - Relação entre as duas etapas de extração de características.

Grupo de característica extraída das ROIs	Quant. de características de cada grupo	Quant. de características extraídas de cada série
Estatística simples	2	4
Agrupamento	1	2
Comparação de histogramas	3	6
Textura horizontal	6	12
Textura vertical	6	12
Índices de diversidade - dx=1 e dy=0	3	6
Índices de diversidade - dx=0 e dy=1	3	6
Estatística espacial	1	2
Geometria fractal	1	2
Total		52 características

4.7 CLASSIFICAÇÃO

Nesta etapa, deseja-se classificar as mamas em doentes ou sadias. Para isso, as características extraídas das séries temporais foram apresentadas como vetores de características e submetidas ao classificador MVS implementado na ferramenta WEKA (Hall *et al.*, 2009).

Antes da classificação, os valores obtidos foram normalizados entre 0 e 1 através da equação 4.7.

$$Y_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (4.7)$$

onde Y_i é o valor do atributo X_i escalado entre 0 e 1, X_{min} e X_{max} são os valores mínimos e máximos, respectivamente, obtidos a partir de toda a amostra utilizada no desenvolvimento da metodologia desta tese. A principal vantagem da normalização é evitar que os atributos em maiores intervalos numéricos dominem aqueles em intervalos numéricos menores.

Nesta tese, utilizou-se MVS com o aprendizado SVC (*Support Vector Classification*), que permite realizar a classificação em duas classes. A classificação ocorreu configurando o aprendizado de duas maneiras: *c*-SVC e *nu*-SVC, onde *c* é um fator de regularização para o espaço de características e *nu* representa um limite superior sobre a fração de erros de treinamento e um limite inferior da fração de vetores de suporte. A função de núcleo utilizada para a classificação foi a de Base Radial (*Radial Basis Function* - *RBF*), pois testes confirmaram que esta função apresentava os melhores resultados.

A técnica *Leave-One-Out Cross-Validation* foi utilizada para validar o desempenho do classificador. Ela é uma técnica de validação utilizada para assegurar que os resultados apresentados por um classificador sejam independentes do conjunto de dados utilizados. Esta técnica utiliza um único vetor de característica de uma classe como os dados de validação, e todos os vetores restantes da amostra como os dados de treinamento. Isto é repetido de modo a que cada vetor de características na amostra seja utilizado uma vez como dados de validação.

4.8 FORMAS DE AVALIAÇÃO DOS RESULTADOS

A fim de se obter métricas que sirvam de comparação da qualidade da metodologia proposta, foram calculados avaliadores sobre os resultados obtidos. São eles: sensibilidade, especificidade, acurácia e índice de Youden. Também foi calculada a área sob a Curva ROC. Tais medidas foram adotadas na tese uma vez que a grande maioria dos trabalhos que usam sistemas de apoio ao diagnóstico também utilizam tais medidas para demonstrarem a qualidade de suas metodologias.

Para calcular tais avaliadores de uma amostra que possui casos positivos e negativos de uma dada anomalia na mama, os resultados dos testes devem se enquadrar em um dos quatro grupos a seguir:

Verdadeiro Positivo (VP): o teste classificou como doente, e a mama realmente possui a anomalia.

Verdadeiro Negativo (VN): o teste classificou como saudável, e a mama realmente não possui a anomalia.

Falso Positivo (FP): o teste classificou como doente, porém a mama não possui a anomalia.

Falso Negativo (FN): o teste classificou como saudável, porém a mama possui a anomalia.

A acurácia, representada na equação 4.8 e que corresponde à taxa de classificação correta, é calculada a partir da razão entre o somatório dos verdadeiros positivo e negativo (VP + VN) pelo número total de casos da amostra (N) (Hastie *et al.*, 2009).

$$Acurácia = \frac{VP + VN}{N} \quad (4.8)$$

A sensibilidade corresponde à proporção de acerto para os casos em que a mama apresenta a doença. Assim, esta medida serve para medir a taxa de acertos para os casos de mamas doentes (Hastie *et al.*, 2009). A equação 4.9 apresenta o cálculo da sensibilidade.

$$Sensibilidade = \frac{VP}{VP + FN} \quad (4.9)$$

A especificidade mede a proporção de casos que são negativos e que foram classificados corretamente como negativos (Hastie *et al.*, 2009), conforme observado na equação 4.10.

$$Especificidade = \frac{VN}{VN + FP} \quad (4.10)$$

Proposto por Youden (1950), o índice que recebe o seu nome serve para relacionar a sensibilidade e a especificidade, e tem o objetivo de medir de forma geral a qualidade dos testes. A equação 4.11 apresenta o cálculo do índice.

$$Y = Sensibilidade + Especificidade - 1 \quad (4.10)$$

A curva ROC foi desenvolvida no contexto de detecção de sinais eletrônicos e problemas com radares. Seu objetivo era quantificar a habilidade dos operadores de radares em distinguir um sinal de um ruído. Esta habilidade era chamada de Características de Operação do Receptor (*Receiver Operating Characteristic* - ROC). Nos anos 70, as curvas ROC se disseminaram em vários ramos da pesquisa biomédica, área em que seu objetivo tornou-se basicamente auxiliar a classificação de indivíduos em doentes ou não doentes (Martinez *et al.*, 2003). Elas são uma forma de relacionar a sensibilidade e a especificidade. A figura 4.17 apresenta um exemplo de curva ROC onde o eixo vertical representa os valores da sensibilidade e o eixo horizontal a taxa de falsos positivos (1 - sensibilidade).

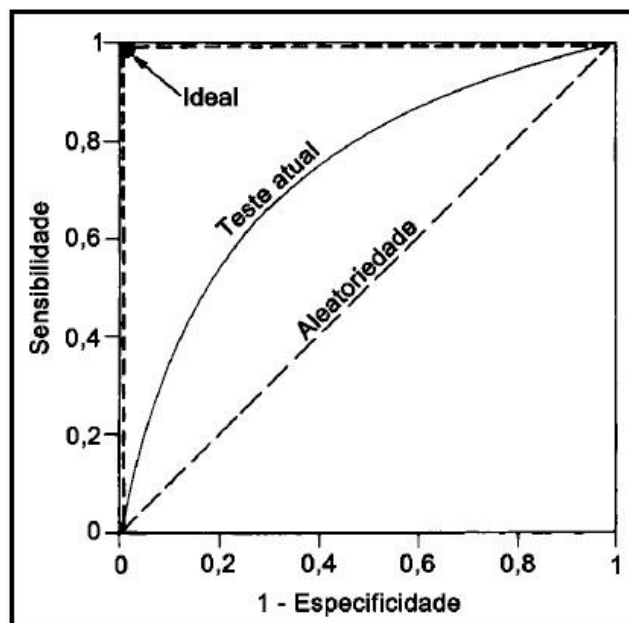


Figura 4.17 - Exemplo de curva ROC. Fonte: Borchardt (2013)

Ainda de acordo com a figura 4.17, a linha tracejada representa uma classificação onde há 50% de chance do caso ser classificado como doente ou não doente. A linha contínua representa uma classificação qualquer onde as chances de classificação correta são superiores a 50%. A acurácia de 100% é representada no canto superior esquerdo do gráfico, onde a sensibilidade seria de 100% e os falsos positivos seria zero. Um modo prático de avaliar a qualidade do classificador que se está usando é calcular a área sob a Curva ROC. Esta área indica numericamente o quão próximo da classificação perfeita está o método utilizado. O valor da área varia entre 0 e 1 e é diretamente proporcional a proximidade com a classificação perfeita e, caso o valor seja menor que 0,5, significa que uma classificação baseada na aleatoriedade daria melhores resultados (Borchardt, 2013).

CAPÍTULO 5 - RESULTADOS E DISCUSSÕES

Neste Capítulo serão apresentados e discutidos os resultados alcançados a partir da metodologia proposta no capítulo anterior. É apresentado primeiro um resumo com os melhores resultados da classificação por grupo de características. Em seguida, é realizada a seleção de características e os resultados são apresentados para cada série. Na seção seguinte, são apresentados os resultados da classificação a partir do resultado da seleção de características da seção anterior. Em seguida, são discutidos os resultados alcançados, além de serem realizados testes da metodologia proposta com novos casos de mamas que não participaram da etapa de desenvolvimento da metodologia. Também são realizadas comparações dos resultados alcançados com a metodologia proposta com outros trabalhos disponíveis da literatura relacionados com o auxílio ao diagnóstico de doenças de mamas a partir de imagens térmicas. Por fim, é feita uma proposta de metodologia para um sistema de auxílio ao diagnóstico de doenças de mama baseado em termografia dinâmica.

5.1 RESULTADOS POR GRUPO DE CARACTERÍSTICAS

Para cada grupo de características, as séries foram divididas por cardinalidade, conforme explicado no Capítulo anterior. Foram extraídas de cada série, duas características que foram submetidas ao classificador, são elas: a amplitude da série e a raiz quadrada da soma ao quadrado da série. Os valores obtidos foram normalizados entre 0 e 1. O classificador MVS implementado no WEKA foi utilizado com todos os seus parâmetros *default*. Utilizou-se o tipo de aprendizado SVC (*Support Vector Classification*), variando-o entre C-SVC e o Nu-SVC.

Os resultados detalhados de cada grupo de característica com seus valores de sensibilidade, especificidade, acurácia, índice de Youden e área sob a curva ROC para cada análise realizada são apresentados no Apêndice B.

Os melhores resultados das análises explicadas anteriormente são apresentados no gráfico da figura 5.1. Um novo grupo formado pela união de todos os demais grupos foi também considerado.

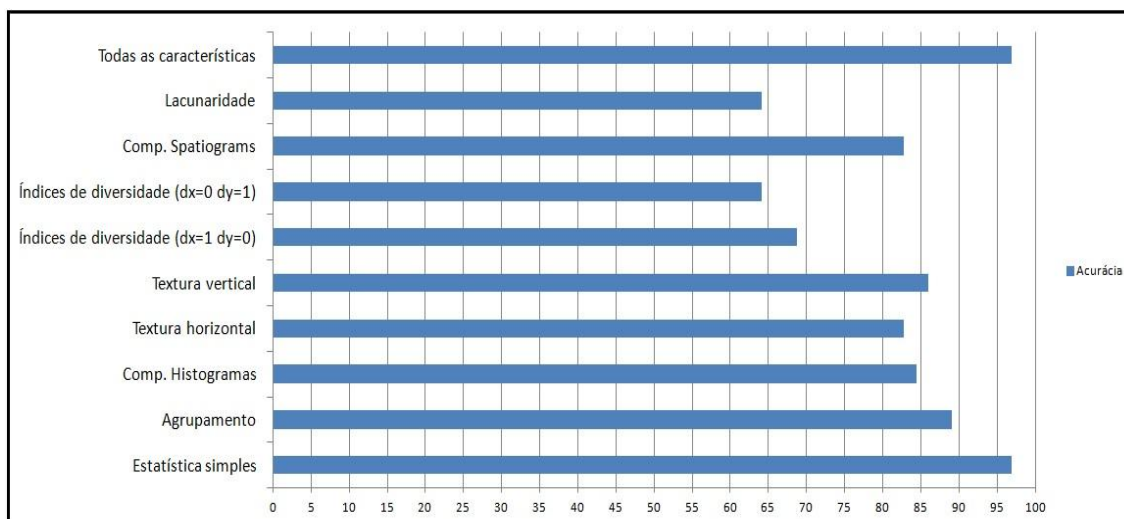


Figura 5.1 - Maiores acurácias por grupo de características

O gráfico da figura 5.1 apresenta no eixo vertical, os grupos de características considerados no Capítulo anterior, acrescido do grupo formado por todas as características. No eixo horizontal, é apresentada uma escala com os valores em porcentagem das acurácias. As barras do gráfico representam os maiores valores de acurácia alcançados variando-se o tamanho das séries e o classificador entre C-SVC e Nu-SVC para cada grupo de características.

Ainda em relação à figura 5.1, é possível perceber que as maiores acurácias foram obtidas pelas características do grupo de estatística simples e pela união de todas as características, alcançando valores próximos de 97% de acurácia. Por outro lado, grupos formados pelos índices de diversidade (horizontal e vertical) e lacunaridade não obtiveram bons resultados, alcançando acurácias entre 64% e 69%.

5.2 SELEÇÃO DE CARACTERÍSTICAS

Após a classificação utilizando os grupos da seção anterior, iniciou-se a etapa de seleção de características. Esta etapa objetiva selecionar os atributos mais relevantes para classificação em cada tipo de série. Os detalhes da seleção de características foram explicados na seção 4.6.

O resultado da seleção de características para cada uma das séries temporais pode ser observado na tabela 5.1, onde "R" representa a raiz quadrada da soma ao quadrado e "A" representa a amplitude da série.

Tabela 5.1 - Resultado da seleção de características para cada tipo de série.

Série		Quantidade de Características Seleccionadas	Características Seleccionadas
K	M		
1	0	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Comparação Spatiograms; - A Contraste horizontal; - A Contraste vertical.
2	0	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Lacunaridade; - A Interseção histogramas; - A Contraste vertical.
2	1	8	<ul style="list-style-type: none"> - R Média; - A Média; - R Energia horizontal; - A Energia horizontal; - R Contraste horizontal; - R Comparação Spatiograms; - A Interseção histogramas; - A Contraste vertical;
3	0	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Contraste horizontal; - R Lacunaridade; - A Contraste vertical.
3	1	5	<ul style="list-style-type: none"> - R Média; - A Média; - R Lacunaridade; - A Interseção histogramas; - A Energia horizontal.
3	2	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas; - A Homogeneidade horizontal; - A Contraste vertical.
4	0	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Energia horizontal; - A Energia horizontal; - R Lacunaridade.
4	1	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Entropia horizontal; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas.

4	2	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Entropia horizontal; - R Lacunaridade; - A Dist. Bhattacharyya histogramas; - A Contraste vertical.
4	3	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Interseção histogramas; - R Lacunaridade; - A Contraste vertical.

A partir dos resultados anteriormente apresentados na tabela 5.1, pode-se constatar que algumas características não foram selecionadas em nenhum dos tipos de séries temporais consideradas nesta tese. Todas as características relacionadas aos índices de diversidade; ao desvio padrão; à comparação de histogramas usando o qui-quadrado; a variância e correlação horizontais e a variância, energia, correlação, entropia e homogeneidade verticais não foram selecionadas para nenhuma das cardinalidades das séries temporais consideradas.

Ao contrário, outras características como as relacionadas à média de temperatura, agrupamento, comparação de histogramas usando a interseção, energia horizontal, contraste vertical e lacunaridade foram selecionadas em todas ou na maioria das cardinalidades das séries temporais consideradas nesta tese.

A figura 5.2 representa o resumo do exposto nos parágrafos anteriores.

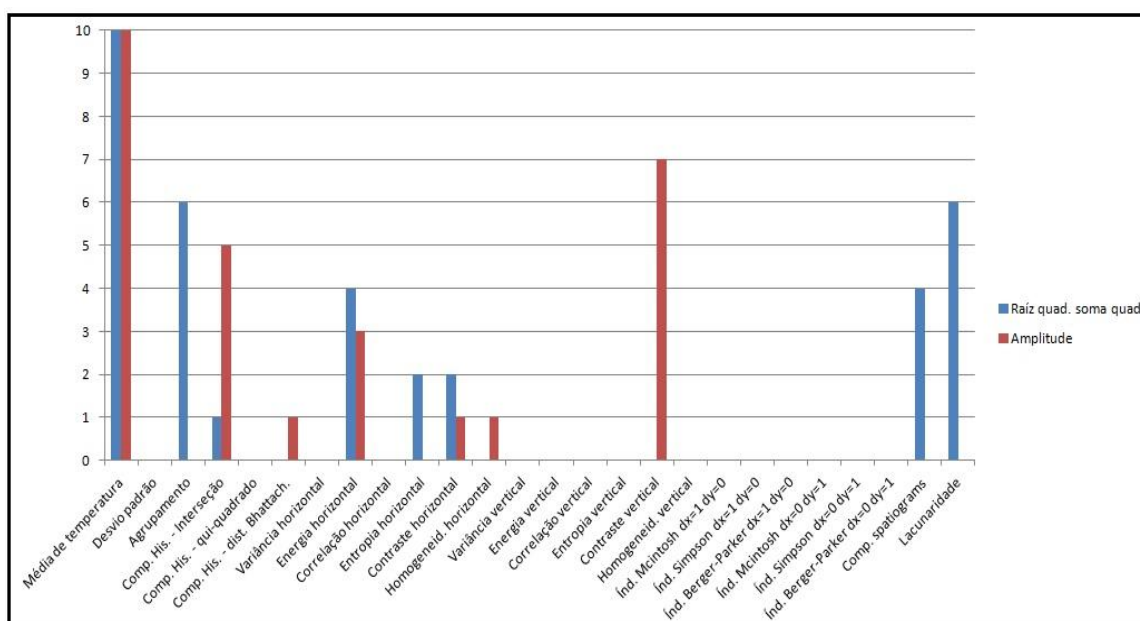


Figura 5.2 - Quantidade vezes que cada característica foi selecionada dentre as diferente séries temporais.

5.3 RESULTADOS DA CLASSIFICAÇÃO APÓS A SELEÇÃO DE CARACTERÍSTICAS

Após a etapa de seleção de características, o processo de classificação foi refeito considerando apenas as características selecionadas na etapa anterior para cada uma das séries temporais desta tese.

Os resultados podem ser verificados na tabela 5.2 a seguir. Assim como nas classificações anteriores, todos os parâmetros do MVS foram mantidos *default* variando apenas entre C-SVC e Nu-SVC.

Tabela 5.2 - Resultado da classificação para cada uma das séries após a seleção de características.

	K	M	Sens (%)	Esp (%)	Acur. (%)	Índ. Youden	Área ROC	Quant Carac. Selec.	Características Selecionadas
C-SVC	1	0	100	96,97	98,44	0,97	0,984	7	- R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Comparação Spatiograms; - A Contraste horizontal; - A Contraste vertical.
	2	0	96,97	100	98,44	0,97	0,984	6	- R Média; - A Média; - R Agrupamento; - R Lacunaridade; - A Interseção histogramas; - A Contraste vertical.
	2	1	96,43	86,11	90,63	0,83	0,906	8	- R Média; - A Média; - R Energia horizontal; - A Energia horizontal; - R Contraste horizontal; - R Comparação Spatiograms; - A Interseção histogramas; - A Contraste vertical;
	3	0	96,77	93,94	95,31	0,91	0,953	7	- R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Contraste horizontal; - R Lacunaridade; - A Contraste vertical.
	3	1	100	86,49	92,19	0,87	0,922	5	- R Média; - A Média; - R Lacunaridade; - A Interseção histogramas; - A Energia horizontal.
	3	2	100	96,97	98,44	0,97	0,984	7	- R Média; - A Média; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas; - A Homogeneidade horizontal; - A Contraste vertical.

	4	0	100	100	100	1	1	6	- R Média; - A Média; - R Agrupamento; - R Energia horizontal; - A Energia horizontal; - R Lacunaridade.
	4	1	100	94,12	96,88	0,94	0,969	6	- R Média; - A Média; - R Entropia horizontal; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas.
	4	2	100	100	100	1	1	7	- R Média; - A Média; - R Agrupamento; - R Entropia horizontal; - R Lacunaridade; - A Dist. Bhattacharyya histogramas; - A Contraste vertical.
	4	3	100	96,97	98,44	0,97	0,984	6	- R Média; - A Média; - R Agrupamento; - R Interseção histogramas; - R Lacunaridade; - A Contraste vertical.
Nu-SV C	1	0	96,77	93,94	95,31	0,91	0,953	7	- R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Comparação Spatiograms; - A Contraste horizontal; - A Contraste vertical.
	2	0	100	100	100	1	1	6	- R Média; - A Média; - R Agrupamento; - R Lacunaridade; - A Interseção histogramas; - A Contraste vertical.
	2	1	96,97	100	98,44	0,97	0,984	8	- R Média; - A Média; - R Energia horizontal; - A Energia horizontal; - R Contraste horizontal; - R Comparação Spatiograms; - A Interseção histogramas; - A Contraste vertical;
	3	0	96,97	100	98,44	0,97	0,984	7	- R Média; - A Média; - R Agrupamento; - R Energia horizontal; - R Contraste horizontal; - R Lacunaridade; - A Contraste vertical.
	3	1	96,77	93,94	95,31	0,91	0,953	5	- R Média; - A Média; - R Lacunaridade; - A Interseção histogramas; - A Energia horizontal.

	3	2	100	96,97	98,44	0,97	0,984	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas; - A Homogeneidade horizontal; - A Contraste vertical.
	4	0	100	96,97	98,44	0,97	0,984	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Energia horizontal; - A Energia horizontal; - R Lacunaridade.
	4	1	96,77	93,94	95,31	0,91	0,953	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Entropia horizontal; - R Comparação spatiograms; - R Lacunaridade; - A Interseção histogramas.
	4	2	100	100	100	1	1	7	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Entropia horizontal; - R Lacunaridade; - A Dist. Bhattacharyya histogramas; - A Contraste vertical.
	4	3	100	100	100	1	1	6	<ul style="list-style-type: none"> - R Média; - A Média; - R Agrupamento; - R Interseção histogramas; - R Lacunaridade; - A Contraste vertical.

Usando C-SVC, foram observados os melhores resultados com acurácia de 100% nas séries $k=4$, $m=0$ na qual 6 características foram selecionadas; e na série $k=4$, $m=2$, onde foram selecionadas 7 características. Já usando Nu-SVC, os melhores resultados foram alcançados com 100% de acurácia nas séries de $k=2$, $m=0$, na qual 6 características foram selecionadas; na série $k=4$, $m=2$, onde foram selecionadas 7 características; e na série $k=4$, $m=3$, onde 6 características foram selecionadas.

5.4 DISCUSSÃO

Os melhores resultados foram alcançados após a etapa de seleção de características, obtendo-se uma acurácia de 100% em 5 diferentes séries temporais com suas respectivas características selecionadas, conforme pôde-se observar na tabela 5.2.

Para a escolha da série recomendada por esta tese e suas respectivas características selecionadas, adotou-se como critério de seleção o menor número de ROIs e o menor número de características selecionadas.

Apesar da acurácia de 100% ter sido alcançada em 2 diferentes tipos de séries temporais quando utilizado C-SVC e em 3 tipos de séries quando utilizado Nu-SVC, apenas a série de $k=4$, $m=0$ utilizando C-SVC (em destaque na tabela 5.2), com suas respectivas características selecionadas, é recomendada por esta tese para a construção do modelo de auxílio ao diagnóstico de doenças de mama a partir de termografia dinâmica.

Analisando os 5 tipos de séries temporais e suas respectivas características selecionadas, que proporcionaram acurácias de 100%, em duas delas ($k=4$, $m=2$ com C-SVC e Nu-SVC) foram selecionadas 7 características. Estas não são recomendadas por esta tese uma vez que necessitam de um número de maior de características extraídas das ROIs sequenciais de cada mama.

Já a série com $k=2$, $m=0$ considerando Nu-SVC, apesar de ter alcançado a acurácia de 100% selecionando 6 características, por ser uma série temporal de $k=2$, esta é formada por um número maior de ROIs, ou seja, para a obtenção desta série, foi necessário extrair as 6 características de 10 ROIs sequenciais.

Logo, restaram apenas as séries com $k=4$, $m=0$ considerando C-SVC e $k=4$, $m=3$ considerando Nu-SVC. Em ambas, as séries temporais foram construídas a partir das características de 5 ROIs sequenciais de cada mama. Também em ambas, para a construção das séries foram extraídas apenas 6 características. As características raiz quadrada da soma ao quadrado e amplitude da série formada pela média de temperatura, a raiz quadrada da soma ao quadrado da série formada pela característica baseada em agrupamento e a raiz quadrada da soma ao quadrado da série formada pela lacunaridade são comuns às duas séries. Porém, a série de $k=4$, $m=3$ requer mais duas características. São elas: a raiz quadrada da soma ao quadrado da série formada pela comparação de histogramas baseada na interseção e a amplitude da série formada pelo contraste baseado nos histogramas da soma e da diferença verticais. Já a série de $k=4$, $m=0$ requer as características raiz quadrada da soma ao quadrado e amplitude da série formada pela energia baseada nos histogramas da soma e da diferença horizontais. Comparando os dois últimos grupos de características que diferenciam as duas séries, percebe-se que a primeira, além dos histogramas da soma e da diferença, necessita da construção dos histogramas de cada ROI para a extração das duas características. Já a segunda configuração de série ($k=4$, $m=0$) requer apenas a construção dos histogramas da soma e da diferença para a extração das duas últimas características.

Assim, recomenda-se para a construção do modelo de auxílio ao diagnóstico de doenças de mama baseado em termografia dinâmica, a extração de características das séries $k=4$, $m=0$, considerando todos os parâmetros *default* do MVS implementado na ferramenta WEKA, inclusive a variável *default* C-SVC do aprendizado. As 6 seis características recomendadas para serem extraídas das séries temporais formadas pelas características das ROIs das imagens em escala de cinza ou das ROIs das matrizes de temperatura estão em destaque na tabela 5.2.

Com essa recomendação, a metodologia proposta alcançou 100% de acurácia utilizando o classificador MVS com todos os seus parâmetros *default* e a técnica de validação *Leave-One-Out Cross-Validation*. Vale destacar que esses resultados são válidos para a base de dados utilizada no desenvolvimento da metodologia. Porém, eles demonstram uma estimativa da potencialidade da metodologia proposta na tarefa de classificação de mamas que não pertencem à base considerada.

5.5 TESTE DA METODOLOGIA COM NOVOS CASOS

A fim de testar a metodologia proposta nesta tese, realizou-se a classificação de 12 novas mamas que não participaram da etapa de desenvolvimento da metodologia. Desses 12 novos casos, 6 possuem são consideradas doentes e 6 são saudáveis. A tabela 5.3 apresenta o resultado da classificação:

Tabela 5.3 - Resultado da classificação de 12 novos casos utilizando a metodologia proposta.

Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob a Curva ROC
100%	100%	100%	1	1

A partir da tabela 5.3, observa-se uma acurácia de 100% também para as 12 novas mamas classificadas a partir da metodologia proposta nesta tese.

Segundo a metodologia, considerou-se a série de $k=4$, $m=0$, sendo analisadas 5 ROIs das imagens em escala de cinza ou das matrizes de temperatura, dentre as 20 sequenciais de cada mama capturadas pelo protocolo dinâmico. Analisou-se a primeira, a quinta, a nona, a décima terceira e a décima sétima ROIs das imagens em escala de cinza ou ROIs das matrizes de temperatura. Foram extraídas as 6 características seguintes das séries temporais: raiz quadrada da soma ao quadrado e amplitude das

séries formadas pelas médias de temperatura das ROIs, a raiz quadrada da soma ao quadrado das séries formadas pela característica baseada em agrupamento, a raiz quadrada da soma ao quadrado e a amplitude das séries formadas pela energia horizontal e a raiz quadrada da soma ao quadrado da série formada pelos valores da lacunaridade. O classificador utilizado foi o MVS com todos os seus parâmetros *default*.

Assim como as ROIs das mamas utilizadas na etapa de desenvolvimento da metodologia, todas as ROIs das 12 novas mamas estarão disponíveis no endereço eletrônico <http://visual.ic.uff.br/proeng/thiagoelias/>.

5.6 COMPARAÇÃO COM OUTROS TRABALHOS

A tabela 5.4 apresenta uma comparação com os principais trabalhos relacionados com auxílio ao diagnóstico de doenças de mama a partir de imagens térmicas.

Buscou-se com este comparativo, analisar os principais trabalhos relacionados, independente do protocolo de aquisição adotado.

Tabela 5.4 - Comparação com outros trabalhos.

Autor	Amostra	Número de Características	Validação	Avaliação	
Ohashi e Uchida (2000)	728 (todas doentes)	4	-	Esp.	-
				Sens.	-
				Acur.	82%
				ROC	-
				Youd	-
Schaefer <i>et al.</i> (2009)	146 (117 saudáveis e 29 doentes)	38	10-fold Cross-Validation	Esp.	79,5%
				Sens.	79,9%
				Acur.	79,5%
				ROC	-
				Youd	0,59

Wishart et al. (2010)	106 (41 saudáveis e 65 doentes)	Não informado	Não informado	Esp.	48%
				Sens.	70%
				Acur.	
				ROC	
				Youd	0,18
Resmini et al. (2012)	34 (10 saudáveis e 24 doentes).	11	Validação Cruzada	Esp.	91,7%
				Sens.	80%
				Acur.	88,2%
				ROC	0,858
				Youd	0,72
Acharya et al. (2012)	50 (25 saudáveis e 25 doentes)	16	3-fold Cross-Validation	Esp.	90,5%
				Sens.	85,7%
				Acur.	88,1%
				ROC	-
				Youd	0,76
Borchartt (2013)	51 (14 saudáveis e 37 doentes)	15	Leave-One-Out Cross-Validation	Esp.	78,6%
				Sens.	91,7%
				Acur.	88,2%
				ROC	0,852
				Youd	0,71
Krawczyk e Schaefer (2014)	146 (117 saudáveis e 29 doentes)	36	Validação Cruzada	Esp.	90,8%
				Sens.	82%
				Acur.	89%
				ROC	-
				Youd	0,73
Francis et al. (2014)	36 (24 saudáveis e 12 doentes)	18	Leave-One-Out Cross-Validation	Esp.	83,3%
				Sens.	83,3%
				Acur.	83,3%
				ROC	-
				Youd	0,67

Gerasimo-va et al. (2014)	47 (14 saudáveis e 33 doentes)	1	-	Esp.	86%
				Sens.	76%
				Acur.	86,9%
				ROC	-
				Youd	0,62
Dourado Neto e Lima (2015)	84 (28 saudáveis e 56 com alguma alteração na mama)	2	Leave-One-Out Cross-Validation	Esp.	93,7%
				Sens.	75%
				Acur.	89,3%
				ROC	-
				Youd	0,69
Silva (2015)	70 (35 saudáveis e 35 doentes)	7	Leave-One-Out	Esp.	100%
				Sens.	100%
				Acur.	100%
				ROC	1
				Youd	1
Metodologia desta Tese	64 (32 saudáveis e 32 doentes)	6	Leave-One-Out	Esp.	100%
				Sens.	100%
				Acur.	100%
				ROC	1
				Youd	1

Comparando os resultados dos trabalhos relacionados na tabela 5.4, observa-se que a metodologia proposta nesta tese alcançou o melhor resultado. Apenas o trabalho proposto por Silva (2015) alcançou resultados equivalentes ao desta tese. Sendo assim, nos próximos parágrafos serão feitas comparações entre a metodologia proposta nesta tese e a metodologia proposta por Silva (2015).

A metodologia de Silva (2015) propõe uma classificação por paciente, ou seja, não é possível determinar em qual mama existe a anomalia. Também, em uma das primeiras etapas da metodologia, é necessário realizar o registro das 20 ROIs sequenciais de cada paciente, para depois gerar a série temporal de temperatura. A etapa de registro é considerada custosa para o processo proposto pelo autor. As séries temporais foram

formadas pelos maiores valores encontrados em porções quadradas da imagem de 11x11 pixels.

Também o fato da segmentação da ROI acontecer apenas na primeira imagem da sequência de vinte imagens, compromete a eficácia da metodologia. Apesar das demais ROIs estarem registradas em relação à primeira, foi utilizada apenas a inspeção visual como métrica para avaliar a qualidade do registro proposto na metodologia. Em outras palavras, não se tem a garantia de que a ROI da primeira imagem corresponde exatamente às demais ROIs sequenciais podendo, assim, estar sendo considerado para a geração das séries temporais regiões não correspondentes às mamas das pacientes, como o tórax ou fundo.

Na metodologia de Silva (2015), foram selecionadas 7 características mais relevantes dentre as 90 extraídas inicialmente das 20 ROIs sequenciais de cada paciente. Este número é maior do que o número de características propostas nesta tese, que é de 6 características.

Na metodologia proposta nesta tese, a classificação é feita por mama, ou seja, é possível determinar em qual das duas mamas da paciente encontra-se a anomalia. Também não são necessárias todas as 20 ROIs sequenciais capturadas pelo protocolo dinâmico. Apenas 5 ROIs com intervalo de 1 minuto de captura entre elas foram necessárias para compor a metodologia. Não é necessário a realização do registro das ROIs, uma vez que a metodologia proposta não se baseia em características de estatística espacial. De cada uma das 5 ROIs de uma dada mama, foram extraídas apenas 4 informações: a média de temperatura, o valor do maior centroide, a lacunaridade e a energia obtida a partir dos histogramas da soma e da diferença considerando a vizinhança horizontal. Destas 4 informações, foram geradas as séries temporais e extraídas 6 novas características que foram submetidas ao classificador. São elas: a raiz quadrada da soma ao quadrado e a amplitude da série formada pelas médias de temperatura, a raiz quadrada da soma ao quadrado da série formada pelos valores dos maiores centroides, a raiz quadrada da soma ao quadrado da série formada pelos valores da lacunaridade e a raiz quadrada da soma ao quadrado e a amplitude da série formada pela energia horizontal.

5.7 PROPOSTA DE METODOLOGIA PARA UM SISTEMA DE DIAGNÓSTICO AUXILIADO POR COMPUTADOR

A partir dos estudos e testes realizados, faz-se uma proposta de metodologia para um sistema de auxílio ao diagnóstico de doenças de mama baseado em termografia dinâmica. O figura 5.3 apresenta um diagrama que resume a proposta.

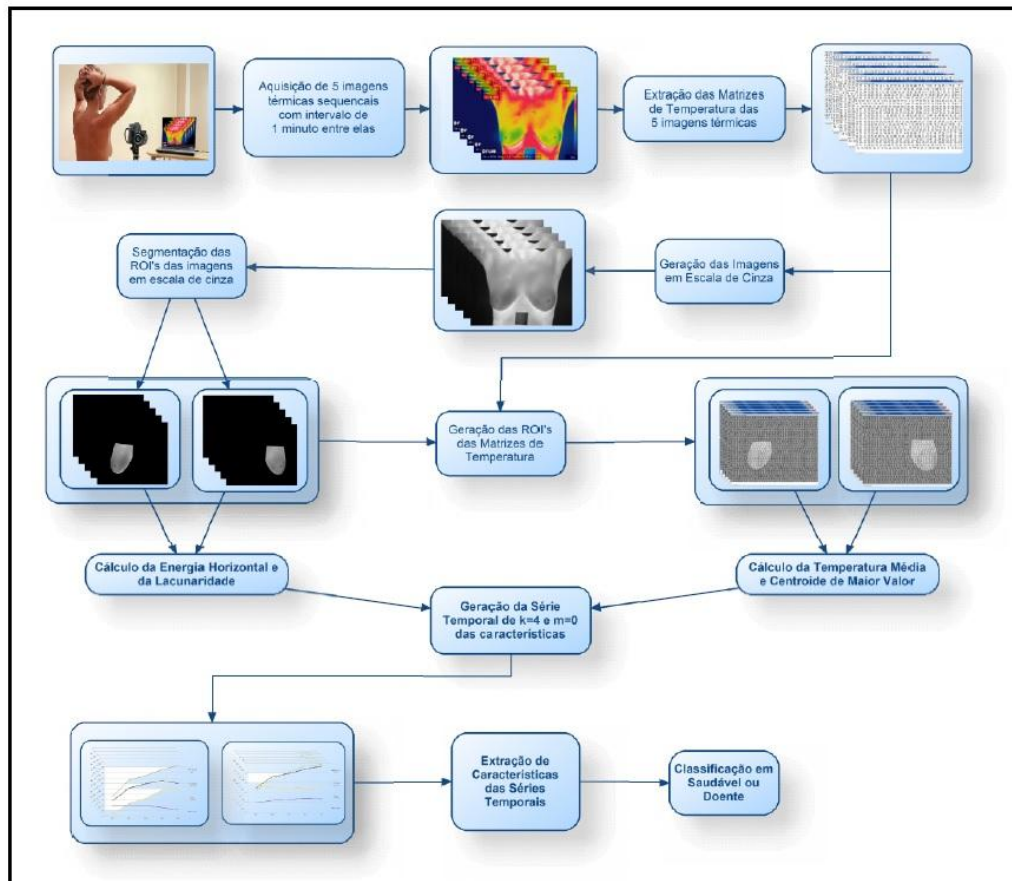


Figura 5.3 - Resumo da proposta de metodologia de CAD

Observa-se a partir da figura 5.3, as seguintes etapas da proposta de metodologia de um sistema de diagnóstico auxiliado por computador:

I - A primeira etapa consiste na aquisição das imagens térmicas a partir do protocolo dinâmico. Após a região das mamas ser resfriada por uma corrente de ar por no máximo 5 minutos, é feita a captura de 5 imagens sequenciais com intervalo de 1 minuto entre elas.

II - A segunda etapa consiste na extração das 5 matrizes de temperatura das imagens capturadas.

III - O passo seguinte consiste na geração de imagens quantizadas em 256 tons de cinza a partir das respectivas matrizes de temperatura. Para isto, é necessário que cada uma

das imagens seja gerada com a mesma escala. Assim, os mesmos valores mínimo e máximo devem ser adotados no cálculo da quantização de todas as imagens. Nesta tese, recomenda-se que o valor mínimo adotado seja de 15 graus Celsius e o valor máximo de 40 graus Celsius. Chegou-se a estes valores a partir da observação das temperaturas de todas as termografias utilizadas nesta tese, que não foram superiores a 40°C e nem inferiores a 15°C.

IV - A quarta etapa consiste na segmentação das ROIs a partir das imagens quantizadas em escala de cinza. Cada ROI é equivalente à mama direita ou à mama esquerda de cada uma das 5 imagens sequenciais da paciente. Assim, ao final da segmentação, tem-se um total de 10 ROIs, 5 ROIs sequenciais da mama direita e 5 ROIs sequenciais da mama esquerda.

V - Além da determinação das ROIs sequenciais das imagens quantizadas em escala de cinza, é necessária a determinação das ROIs das matrizes de temperatura. Isso se faz necessário uma vez que algumas características serão extraídas tanto das ROIs em escala de cinza quanto das ROIs das matrizes de temperatura. Para isso, as ROIs em escala de cinza são utilizadas como máscara das matrizes de temperatura, fazendo com que apenas as temperaturas correspondentes à região das mamas sejam consideradas.

VI - A sexta etapa consiste na extração de características das ROIs sequenciais das imagens em escala de cinza e das ROIs das matrizes de temperatura. De cada ROI em escala de cinza, é calculada a energia horizontal e a lacunaridade. O processo de cálculo de ambas as características foi apresentado no capítulo 4 desta tese. Da mesma forma, das ROIs sequenciais das matrizes de temperatura, é calculada a média de temperatura e o centroide de maior valor.

VII - Com as 4 características extraídas das ROIs sequenciais de cada mama, são geradas as séries temporais correspondentes. Em outras palavras, cada mama terá 4 séries temporais referentes à variação ao longo das ROIs sequenciais de cada uma das 4 características.

VIII - Geradas as séries, inicia-se o processo de extração de característica das séries temporais. Nesta etapa são extraídas as seguintes características das séries: a raiz quadrada da soma ao quadrado e a amplitude das séries geradas a partir da característica energia horizontal, a raiz quadrada da soma ao quadrado das séries geradas pela lacunaridade, a raiz quadrada da soma ao quadrado e a amplitude das séries geradas pela característica temperatura média e a raiz quadrada da soma ao quadrado das séries geradas pela característica centroide de maior valor. São essas 6

novas características extraídas das séries temporais que comporão o vetor de característica submetido ao classificador MVS.

IX - A última etapa consiste na classificação de cada uma das mamas em doente ou saudável. A classificação é feita a partir do vetor de característica normalizado composto pelas 6 características extraídas na etapa anterior. O classificador utilizado é o MVS implementado no WEKA com todos os parâmetros *default*.

A proposta de metodologia para um sistema CAD apresentada nesta tese possui como uma de suas vantagens, quando comparada a outros trabalhos, o diagnóstico por mama, e não por paciente. Em outras palavras, é possível identificar qual das duas mamas da paciente apresenta anomalia, aumentando assim a precisão do diagnóstico. A outra vantagem, é o pequeno número de imagens sequenciais necessárias, que é de 5 imagens capturadas com intervalo de 1 minuto entre elas. Mais uma vantagem da proposta é o pequeno número de características extraídas das ROIs sequenciais e, posteriormente, das séries temporais. Apenas 4 características são extraídas das ROIs das imagens em escala de cinza ou das ROIs das matrizes de temperatura. Tais características formam as séries temporais e, destas, 6 características são extraídas.

Assim, ao final, apenas 6 características de cada mama são submetidas ao classificador MVS para a classificação em saudável ou doente. O classificador MVS implementado no WEKA é utilizado com todos os seus parâmetros *default*. Ele vem demonstrando sua superioridade frente a outros classificadores em uma grande variedade de aplicações (Cristianini e Shawe-Taylor, 2000).

CAPÍTULO 6 - CONCLUSÕES

O câncer de mama é o câncer que mais acomete as mulheres em todo o mundo. O diagnóstico da doença nos primeiros anos é decisivo para a sua cura.

Nesta tese, foi apresentada uma metodologia promissora de auxílio na detecção de doenças na mama. A partir da metodologia baseada em termografias dinâmicas, é possível ter indícios para que exames complementares sejam realizados a fim de detectar possíveis doenças. A metodologia apresenta-se como uma ferramenta que auxiliaria na triagem inicial de pacientes, não substituindo os exames que atualmente são utilizados para o rastreio de doenças da mama.

A metodologia proposta baseou-se na análise do comportamento das mamas das pacientes durante o processo de reaquecimento corpóreo, após sofrerem um resfriamento induzido por uma corrente de ar. Esta análise partiu da extração de características das séries temporais formadas pelas características extraídas das ROIs sequenciais de cada mama ao longo do processo de reaquecimento corpóreo. Durante o desenvolvimento da metodologia, foi utilizada uma base de dados com 64 mamas, sendo 32 saudáveis e 32 com diagnóstico de doença na mama. Para cada mama, foram analisadas as 20 ROIs sequenciais capturadas a partir do protocolo dinâmico.

Os melhores resultados e recomendados por esta tese atingiram uma acurácia de 100% considerando apenas 5 ROIs sequenciais de cada mama com intervalo de captura de 1 minuto entre elas. Este fato evidencia a eficiência do método, uma vez que não é necessária a utilização das 20 ROIs sequenciais. As características extraídas dessas ROIs foram temperatura média, valor do maior centroide, energia horizontal e lacunaridade. A partir destas 4 características, foram extraídas 6 novas características de suas respectivas séries temporais, que foram submetidas ao classificador MVS. As características extraídas das séries temporais foram as seguintes: a raiz quadrada da soma ao quadrado e a amplitude da série formada pelas médias de temperatura, a raiz quadrada da soma ao quadrado da série formada pelos centroides de maior valor, a raiz quadrada da soma ao quadrado e a amplitude da série formada pelas energias horizontais e a raiz quadrada da soma ao quadrado da série formada pelas lacunaridades. O classificador utilizado foi o MVS implementado no WEKA com todos os seus parâmetros *default*. A técnica de validação utilizada foi *Leave-One-Out Cross-Validation*. Vale destacar que esses resultados são válidos para a base de dados

utilizada no desenvolvimento da metodologia. Porém, eles demonstram uma estimativa da potencialidade da metodologia proposta na tarefa de classificação de mamas que não pertencem à base considerada.

Em seguida, 12 novas mamas que não participaram da etapa de desenvolvimento da metodologia foram classificadas. Destas, 6 são saudáveis e 6 possuem diagnóstico de doença de mama. A classificação dessas novas mamas também alcançou uma acurácia de 100%, demonstrando a boa potencialidade da metodologia em classificar mamas que não pertencem à base de dados considerada durante o desenvolvimento da metodologia proposta nesta tese.

Também foram comparados os resultados obtidos pela metodologia proposta nesta tese com outros trabalhos descritos na literatura que propõem métodos de auxílio ao diagnóstico de doenças da mama a partir de imagens térmicas. Como foi possível evidenciar, os resultados alcançados nesta tese mostram-se superiores.

A principal contribuição desta pesquisa foi o desenvolvimento de uma metodologia de auxílio ao diagnóstico de doenças de mama a partir de imagens térmicas adquiridas pelo protocolo dinâmico. A metodologia requer um pequeno número de imagens térmicas sequenciais (apenas 5), não sendo necessária a aquisição das 20 imagens sequenciais, inicialmente propostas pelo protocolo dinâmico. Isto demanda menos espaço no banco de dados para o armazenamento das imagens térmicas. Também a metodologia propõe um pequeno número de característica extraídas das ROIs sequenciais das mamas.

6.1 TRABALHOS FUTUROS

Diante do tempo e esforço demandado para a segmentação manual das 1.520 ROIs sequenciais das imagens em escala de cinza (1280 ROIs das 64 mamas usadas na etapa de desenvolvimento da metodologia e 240 ROIs das mamas usadas na etapa de teste da metodologia), sugere-se como trabalho futuro o desenvolvimento de um segmentador automático das ROIs sequenciais obtidas pelo protocolo dinâmico. Assim como explicado na seção 4.4, o segmentador automático desenvolvido por Marques (2012) não se mostrou adequado à metodologia proposta nesta tese. A primeira explicação é o fato do segmentador basear-se nas regiões mais quentes das imagens para determinar as pregas inframamárias. Como as pacientes utilizadas nesta tese passaram por um processo de resfriamento da região das mamas, nem sempre a região mais quente encontrava-se nas pregas inframamárias causando, assim, um erro de segmentação na

maioria das imagens utilizadas. O segundo motivo da não utilização do segmentador automático é o fato da ferramenta segmentar, além da região das mamas, as regiões de possível disseminação da doença, que corresponde à região axilar e do sistema vascular sanguíneo e linfático, conforme observado na figura 4.5. Já nesta tese, foi considerada como região de interesse apenas as glândulas mamárias, como pode ser observado na figura 4.7.

As aquisição das imagens térmicas acontece semanalmente no Hospital Universitário Antônio Pedro, em Niterói. Assim, sugere-se como trabalho futuro o teste da metodologia proposta nesta tese com um banco de dados formado por um número maior de mamas.

A metodologia proposta é capaz de auxiliar no diagnóstico de doenças na mama, dando indícios de qual mama da paciente existe uma anomalia. Porém, não é possível indicar em qual região da mama a anomalia se localiza. Assim, sugere-se como trabalho futuro o desenvolvimento de técnicas que, além de auxiliar na indicação da mama possivelmente doente, indique a região da mama em que se localiza a anomalia.

O *Database for Mastology Research with Infrared Image* - DMR-IR, acessível no endereço eletrônico <http://visual.ic.uff.br/dmi>, armazena, além de imagens térmicas de mamas, imagens de outros exames, como a mamografia, além de dados clínicos de pacientes voluntárias. Assim, sugere-se como trabalho futuro a análise de imagem de diferentes tipos de exames para o auxílio ao diagnóstico de doenças da mama. Além disso, um outro trabalho futuro seria a associação da análise de imagens e dados clínicos das pacientes para o auxílio ao diagnóstico de doenças da mama.

Assim, observa-se que ainda há muitos desdobramentos a partir da pesquisa aqui realizada. Diversos novos trabalhos podem ser desenvolvidos a partir de termografias dinâmicas a fim de melhorar ou avançar na metodologia proposta nesta tese.

REFERÊNCIAS

- Acharya, U.R., Ng, E.Y.K., Tan, J.H., Sree, S.V., 2012. Thermography Based Breast Cancer Detection Using Texture Features and Support Vector Machine. *Journal of Medical Systems* 36, 1503–1510. doi:10.1007/s10916-010-9611-z
- Allain, C., Cloitre, M., 1991. Characterizing the lacunarity of random and deterministic fractal sets. *Physical Review* 44, 3552–3558.
- Anbar, M., Milesescu, L., Naumov, A., Brown, C., Button, T., Carty, C., AlDulaimi, K., 2001. Detection of cancerous breasts by dynamic area telethermometry. *IEEE engineering in medicine and biology magazine* 20, 80–91.
- Araújo, F.A., Conci, A., Gonçalves, M., Oliveira, T., 2015. Metodologia para Reconstrução da Geometria da Mama usando Sensores de Profundidade. Presented at the XXXV Congresso da Sociedade Brasileira de Computação - XV Workshop de Informática Médica, Recife, Pernambuco.
- Arthur, D., Vassilvitskii, S., 2007. K-means++: The Advantages of Careful Seeding, in: *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA, pp. 1027–1035.
- Baraldi, A., Parmiggiani, F., 1995. An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. *IEEE Transactions on Geoscience and Remote Sensing* 33, 293–304.
- Barros Filho, M.N., Sobreira, F.J.A., 2008. Accuracy of Lacunarity Algorithms in Texture Classification of High Spatial Resolution Images from Urban Areas, in: *XXI Congress of the International Society for Photogrammetry and Remote Sensing-ISPRS*. Beijing, China, pp. 417–422.
- Batiston, A.P., Tamaki, E.M., Santos, M.L. de M. dos, Cazola, L.H. de O., 2009. Método de detecção do câncer de mama e suas implicações. *Cogitare Enfermagem* 14, 59–64. doi:10.5380/ce.v14i1.14103
- Bhattacharyya, A., 1946. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)* 7, 401–406.
- Birchfield, S.T., Rangarajan, S., 2005. Spatiograms versus histograms for region-based tracking, in: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 1158–1163.
- Borchardt, T.B., 2013. Análise de imagens termográficas para a classificação de alterações na mama (Tese de doutorado). UFF, Niterói.
- Borchardt, T.B., Conci, A., Lima, R.C.F., Resmini, R., Sanchez, A., 2013. Breast thermography from an image processing viewpoint: A survey. *Signal Processing* 93, 2785–2803.
- Cai, X., Xiao, B., Wang, C., Zhang, R., 2011. Quadratic-chi similarity metric learning for histogram feature. Presented at the First Asian Conference on Pattern Recognition (ACPR), pp. 47–51.
- Calas, M.J.G., Gutfilen, B., Pereira, W.C. de A., 2012. CAD and mammography: why use this tool? *Radiologia Brasileira* 45, 46–52. doi:10.1590/S0100-39842012000100011
- Carvalho, P.M. de S., Paiva, A.C. de, Silva, A.C., 2012. Classification of Breast Tissues in Mammographic Images in Mass and Non-mass Using McIntosh's Diversity Index and SVM, in: *Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 482–494.
- Cha, S.H., Srihari, S.N., 2002. On measuring the distance between histograms. *Pattern Recognition* 35, 1355–1370.
- Conaire, C.O., O'Connor, N.E., Smeaton, A.F., 2007. An Improved Spatiogram Similarity Measure for Robust Object Localisation, in: *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*. pp. 1–1069–1–1072.
- Conceição, J.C.J., 2006. *Ginecologia Fundamental*. Atheneu Editora, Rio de Janeiro.

- Conci, A., Azevedo, E., Leta, F.R., 2008. *Computação Gráfica: Teoria e Prática*, Segunda. ed. Campus, Rio de Janeiro.
- Cristianini, N., Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA.
- Damin, A.P.S., 2012. Rastreamento do câncer de mama: quais as tendências atuais? *Clinical & Biomedical Research* 32, 129–130.
- Datta, R., Joshi, D., Li, J., Wang, J.Z., 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys* 40, 5:1–5:60.
- Dong, P., 2000. Test of a new lacunarity estimation method for image texture analysis. *International Journal of Remote Sensing* 3369–3373.
- Dourado Neto, H.M., Lima, R.C.F., 2015. Automated breast cancer detection via thermography, in: *Encontro Nacional de Engenharia Biomecânica - ENEBI*. Uberlândia - MG.
- Drukeinis, J.S., Mooney, B.P., Flowers, C.I., Gatenby, R.A., 2013. Beyond mammography: new frontiers in breast cancer screening. *The American Journal of Medicine* 126, 472–479. doi:10.1016/j.amjmed.2012.11.025
- Du, G., Yeo, T.S., 2002. A novel lacunarity estimation method applied to SAR image segmentation. *IEEE Transactions on Geoscience and Remote Sensing* 2687–2691.
- EtehadTavakol, M., Lucas, C., Sadri, S., Ng, E., 2010. Analysis of Breast Thermography Using Fractal Dimension to Establish Possible Difference between Malignant and Benign Patterns. *Journal of Healthcare Engineering* 27–44.
- Fernandes, J.L., Conci, A., 2004. Uso da Lacunaridade no Controle de Qualidade Visual de Padrões Industriais, in: *III Congresso Nacional de Engenharia Mecânica*(CD-Rom). Belém, PA.
- FLIR, 2013. FLIR SC620 Thermal Imaging Cameras [WWW Document]. URL <http://www.flir.com.hk/cs/display/?id=41413> (accessed 9.14.13).
- FLIR, 2009. FLIR QuickReport Software [WWW Document]. URL <http://www.flir.com.br/instruments/display/?id=60093> (accessed 4.9.13).
- Francis, S.V., Sasikala, M., Bhavani Bharathi, G., Jaipurkar, S.D., 2014. Breast cancer detection in rotational thermography images using texture features. *Infrared Physics & Technology* 67, 490–496. doi:10.1016/j.infrared.2014.08.019
- Galvão, S., 2015. Registro de imagens térmicas intrapaciente adquiridas dinamicamente (Tese de Doutorado). UFF, Niterói.
- Gefen, Y., Meir, Y., Mandelbrot, B.B., Aharony, A., 1983. Geometric Implementation of Hypercubic Lattices with Noninteger Dimensionality by Use of Low Lacunarity Fractal Lattices. *Physical Review Letters* 50, 145–148. doi:10.1103/PhysRevLett.50.145
- Gerasimova, E., Audit, B., Roux, S.G., Khalil, A., Argoul, F., Naimark, O., Arneodo, A., 2013. Multifractal analysis of dynamic infrared imaging of breast cancer. *EPL- Europhysics Letters* 104, 68001. doi:10.1209/0295-5075/104/68001
- Gerasimova, E., Audit, B., Roux, S.G., Khalil, A., Gileva, O., Argoul, F., Naimark, O., Arneodo, A., 2014. Wavelet-based multifractal analysis of dynamic infrared thermograms to assist in early breast cancer diagnosis. *Front Physiol* 5, 176. doi:10.3389/fphys.2014.00176
- Gong, L., Wang, T., Liu, F., Chen, G., 2009. A Lie group based spatiogram similarity measure, in: *IEEE International Conference on Multimedia and Expo - ICME*. pp. 582–585. doi:10.1109/ICME.2009.5202563
- Gonzalez, R.C., Woods, R.E., 2007. *Digital Image Processing*, 3 edition. ed. Prentice Hall, Upper Saddle River, N.J.
- Hall, M.A., 1999. Correlation-based Feature Subset Selection for Machine Learning (Tese de Doutorado). Department of Computer Science, The University of Waikato, Hamilton, New Zealand.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 10–18. doi:10.1145/1656274.1656278

- Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics* SMC-3, 610–621. doi:10.1109/TSMC.1973.4309314
- Hastie, T., Tibshirani, R., Friedman, J., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA.
- Haykin, S., 1999. *Neural networks: a comprehensive foundation*, 2nd ed. Prentice Hall, New Jersey.
- Herman, C., 2013. The role of dynamic infrared imaging in melanoma diagnosis. *Expert review of dermatology* 8, 177–184. doi:10.1586/edm.13.15
- Higuchi, T., 1988. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena* 31, 277–283. doi:10.1016/0167-2789(88)90081-4
- INCA, 2014. *Estimativa 2014: Incidência de Câncer no Brasil*. Instituto Nacional de Câncer José Alencar Gomes da Silva, Rio de Janeiro: Divisão de Tecnologia da Informação.
- Kailath, T., 1967. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology* 15, 52–60. doi:10.1109/TCOM.1967.1089532
- Kennedy, D.A., Lee, T., Seely, D., 2009. A comparative review of thermography as a breast cancer screening technique. *Integrative Cancer Therapies* 8, 9–16. doi:10.1177/1534735408326171
- Keyserlingk, J.R., Ahlgren, P.D., Yu, E., Belliveau, N., Yassa, M., 2000. Functional infrared imaging of the breast. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society* 19, 30–41.
- Krawczyk, B., Schaefer, G., 2014. A hybrid classifier committee for analysing asymmetry features in breast thermograms. *Applied Soft Computing, Hybrid intelligent methods for health technologies* 20, 112–118. doi:10.1016/j.asoc.2013.11.011
- Latorre, M. do R.D. de O., Cardoso, M.R.A., 2001. Time series analysis in epidemiology: an introduction to methodological aspects. *Revista Brasileira de Epidemiologia* 4, 145–152. doi:10.1590/S1415-790X2001000300002
- Lee, S.M., Xin, J.H., Westland, S., 2005. Evaluation of image similarity by histogram intersection. *Color Research & Application* 30, 265–274. doi:10.1002/col.20122
- Lorena, A.C., Carvalho, A.C.P.L.F. de, 2007. Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada* 14, 43–67.
- MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations, in: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. pp. 281–297.
- Magurran, A.E., 1988. *Ecological Diversity and Its Measurement*, 1st ed. Princeton University Press, Princeton, New Jersey.
- Mandelbrot, B.B., 1983. *The Fractal Geometry of Nature*. Henry Holt and Company.
- Marques, R. de S., 2012. *Segmentação Automática das Mamas em Imagens Térmicas* (Dissertação de Mestrado). UFF, Niterói.
- Marques, R.S., Machado, D.A., Giraldo, G., Novotny, A., Machenry, T., Conci, A., 2013. A new algorithm for efficient computation of Hausdorff distance in evaluation of digital image segmentation, in: *Workshop on Distance Geometry and Applications (DGA13)*. 1-5, Manaus, Amazonas.
- Martinez, E.Z., Louzada-Neto, F., Pereira, B.B., 2003. A Curva ROC para testes diagnósticos. *Cadernos Saúde Coletiva* 1, 7–31.
- Ma, Y., Gu, X., Wang, Y., 2010. Histogram similarity measure using variable bin size distance. *Computer Vision and Image Understanding* 114, 981–989. doi:10.1016/j.cviu.2010.03.006
- McIntosh, R.P., 1967. An Index of Diversity and the Relation of Certain Concepts to Diversity. *Ecology* 48, 392–404. doi:10.2307/1932674
- Mitchell, T.M., 1997. *Machine Learning*, 1st ed. McGraw-Hill, Inc., New York, NY, USA.

- Moghbel, M., Mashohor, S., 2011. A review of computer assisted detection/diagnosis (CAD) in breast thermography for breast cancer detection. *Artificial Intelligence Review* 39, 305–313. doi:10.1007/s10462-011-9274-2
- Motta, L.S., Conci, A., Lima, R.C.F., Diniz, E.M., 2010. Automatic Segmentation on Thermograms in Order to Aid Diagnosis and 2D Modeling, in: *Anais CSBC 2010*. Presented at the XXX Congresso da Sociedade Brasileira de Computação - X Workshop de Informática Médica, Belo Horizonte, Minas Gerais, pp. 1610–1619.
- Nascimento, L.B., de Paiva, A.C., Silva, A.C., 2012. Lung Nodules Classification in CT Images Using Shannon and Simpson Diversity Indices and SVM, in: *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition*. pp. 454–466. doi:10.1007/978-3-642-31537-4_36
- Netter, F.H., 2011. *Atlas de Anatomia Humana*, 5th ed. Elsevier/Medicina Nacionais.
- Ng, E.Y.-K., 2009. A review of thermography as promising non-invasive detection modality for breast tumor. *International Journal of Thermal Sciences* 48, 849–859. doi:10.1016/j.ijthermalsci.2008.06.015
- Ohashi, Y., Uchida, I., 2000. Applying dynamic thermography in the diagnosis of breast cancer. *IEEE Engineering in Medicine and Biology Magazine* 19, 42–51. doi:10.1109/51.844379
- Oliveira, J.P.S., Conci, A., Perez, M.G., Andaluz, V.H., 2015. Segmentation of infrared images: A new technology for early detection of breast diseases, in: *IEEE International Conference on Industrial Technology (ICIT)*. Sevilla, Espanha, pp. 1765–1771. doi:10.1109/ICIT.2015.7125353
- Pedrin, H., Schwartz, W.R., 2008. *Análise de imagens digitais: princípios, algoritmos e aplicações*. Thomson Learning.
- Plotnick, R.E., Gardner, R.H., Hargrove, W.W., Prestegard, K., Perlmutter, M., 1996. Lacunarity analysis: A general technique for the analysis of spatial patterns. *Phys. Rev. E* 53, 5461–5468. doi:10.1103/PhysRevE.53.5461
- Resmini, R., Borchardt, T.B., Conci, A., Lima, R.C.F., 2012. Extraction of features based on geostatistical measure using thermal images of breast. Presented at the VII Congresso Nacional de Engenharia Mecânica, São Luís, MA.
- Rocha, S.V. da, Braz Junior, G., Silva, A.C., Paiva, A.C. de, 2014. Texture analysis of masses in digitized mammograms using Gleason and Menhinick diversity indexes. *Revista Brasileira de Engenharia Biomédica* 30, 27–34. doi:10.4322/rbeb.2014.008
- Rocha, S.V., 2014. *Diferenciação do Padrão de Malignidade e Benignidade de Massas em Imagens de Mamografias Usando LBP, Geoestatística e Índice de Diversidade (Tese de Doutorado)*. UFMA, São Luís.
- Schaefer, G., Závisek, M., Nakashima, T., 2009. Thermography based breast cancer analysis using statistical features and fuzzy classification. *Pattern Recognition, Digital Image Processing and Pattern Recognition Techniques for the Detection of Cancer* 42, 1133–1137. doi:10.1016/j.patcog.2008.08.007
- Serrano, R.C., Conci, A., Zamith, M., Lima, R.C.F., 2010. About the feasibility of Hurst coefficient in thermal images for early diagnosis of breast diseases, in: *PACAM XI -11th Pan-American Congress of Applied Mechanics*. Foz do Iguaçu, Paraná.
- Shapiro, S., 1977. Evidence on screening for breast cancer from a randomized trial. *Cancer* 39, 2772–2782.
- Silva, L.F., 2015. *Uma Análise Híbrida para Detecção de Anomalias da Mama usando Séries Temporais de Temperatura (Tese de Doutorado)*. Universidade Federal Fluminense - UFF, Niterói - RJ.
- Silva, L.F., Marques, R.S., Carvalho, G.S., Santos, L.M.O., Fontes, C.A.P., Santos, A.A.S.M.D., Conci, A., 2013. Protocolo de captura de imagens térmicas da mama para construção de um banco público de exames, in: *IV Encontro Nacional de Engenharia Biomecânica - ENEBI*. Vitória - ES, pp. 104–105.

- Tang, J., Agaian, S., Thompson, I., 2014. Computer-Aided Detection or Diagnosis (CAD) Systems. *IEEE Systems Journal* 8, 907–909. doi:10.1109/JSYST.2014.2317378
- Unser, M., 1986. Sum and Difference Histograms for Texture Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*, 118–125. doi:10.1109/TPAMI.1986.4767760
- Urban, L.A.B.D., Schaefer, M.B., Duarte, D.L., Santos, R.P. dos, Maranhão, N.M. de A., Kefalas, A.L., Canella, E. de O., Ferreira, C.A.P., Peixoto, J.E., Chala, L.F., Costa, R.P., Francisco, J.L.E., Martinelli, S.E., Amorim, H.L.E. de, Pasqualette, H.A., Pereira, P.M.S., Junior, C., De, H.S.A., Sondermann, V.R., 2012. Recomendações do Colégio Brasileiro de Radiologia e Diagnóstico por Imagem, da Sociedade Brasileira de Mastologia e da Federação Brasileira das Associações de Ginecologia e Obstetrícia para rastreamento do câncer de mama por métodos de imagem. *Radiologia Brasileira* 45, 334–339. doi:10.1590/S0100-39842012000600009
- Vainer, B.G., 2005. FPA-based infrared thermography as applied to the study of cutaneous perspiration and stimulated vascular response in humans. *Physics in Medicine and Biology* 50, R63–94. doi:10.1088/0031-9155/50/23/R01
- Vapnik, V., Chervonenkis, A., 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications* 16, 264–280.
- Vapnik, V.N., 1998. *Statistical Learning Theory*, 1 edition. ed. Wiley-Interscience, New York.
- Warner, E., 2011. Breast-Cancer Screening. *New England Journal of Medicine* 365, 1025–1032. doi:10.1056/NEJMcp1101540
- Wishart, G.C., Campisi, M., Boswell, M., Chapman, D., Shackleton, V., Iddles, S., Hallett, A., Britton, P.D., 2010. The accuracy of digital infrared imaging for breast cancer detection in women undergoing breast biopsy. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology* 36, 535–540.
- Youden, W.J., 1950. Index for rating diagnostic tests. *Cancer* 3, 32–35.
- Ziyad, S., Iruela-Arispe, M.L., 2011. Molecular mechanisms of tumor angiogenesis. *Genes & Cancer* 2, 1085–1096. doi:10.1177/1947601911432334

APÊNCIA A - DIAGNÓSTICOS DAS PACIENTES

As Tabelas A.1 e A.2 contêm, respectivamente, os dados clínicos das mamas doentes e saudáveis utilizadas nesta tese, conforme descrito no Capítulo 4.

Tabela A.1 - Mamas com anomalias

Mama	Descrição da anomalia
138	Fibroadenoma em quadrante inferior externo com 0,5x 0,6cm.
179	Carcinoma ductal infiltrante grau 3 com focos multifocais, representando comprometimento de margem cirúrgica.
180	Carcinoma ductal infiltrante com 7cm em união dos quadrantes superiores.
181d	Carcinoma Ductal infiltrante grau 3 com 2 cm em União dos quadrantes superiores.
181e	Carcinoma Ductal infiltrante grau 3 com 2cm em união dos quadrantes externos e nódulo de 1,5 cm em quadrante superior externo.
192	Carcinoma ductal infiltrante localmente avançado com cerca de 5-10cm, inicialmente em união de quadrantes inferiores, porém atualmente lesão ulcerada ao redor do complexo aréolo-mamilar (CAM).
198	Carcinoma residual, correspondendo a 60% de celularidade em leito tumoral que 2,5x2,0 cm, sendo 2% representado por carcinoma intra-ductal angio-linfática.
202	Carcinoma ductal in situ, localmente avançado, com tamanho inicial de 10 cm em quadrante superior externo, submetida a quimioterapia e radioterapia neoadjuvantes até maio/2013. Atualmente com espessamento de 3,5 cm em quadrante superior externo.
204	Descarga papilar de líquido seroso em mama direita. Resultado da peça cirúrgica após retirada da mama: Carcinoma ductal in situ de baixo grau.
209	Carcinoma ductal infiltrante grau 2, nódulo com cerca de 4cm em união dos quadrantes internos.
210	Carcinoma ductal in situ de alto grau , união dos quadrantes superiores com 1,1-2 cm.
213	Carcinoma papilífero no quadrante superior externo, nódulo com 1,5cm.
240	Carcinoma ductal infiltrante a derme, com graduação histológica Nottingham II (grau arquitetural 2, nuclear 2 índice mitótico 3).
241	Carcinoma ductal infiltrante, Nottingham grau II (grau histopatológico=3, grau nuclear=2, índice mitótico=1).
245	Carcinoma ductal infiltrante, grau histológico I de Nottingham (formação tubular=2; grau nuclear=2; contagem mitótica=1).

249	Tumor filoide de mama, de baixo potencial de malignidade.
250	Carcinoma lobular in situ.
251	Carcinoma infiltrante, de tipo especial, grau histopatológico nal de Nottingham I (formação tubular 1; pleomorfismo nuclear 1; índice mitótico 1)
321	Carcinoma lobular in situ.
255	Carcinoma ductal infiltrante , grau histológico de Nottingham final II (arranjo tubular =3; grau nuclear =2, índice mitótico=1).
257	Carcinoma ductal infiltrante , grau histológico I de Nottingham (formação tubular =2; grau nuclear=2; contagem mitótica=1), comprometendo toda a derme e hipoderme.
259	Carcinoma ductal infiltrante.
322	Carcinoma papilífero no quadrante superior externo
261	Carcinoma invasivo de tipo não especial. Grau 3 de Nottingham (arranjo tubular=3, pleomorfismo nuclear =3; índice mitótico = 2).
263	Esteatonecrose com área de fibrose hialina e calcificação.
356	Carcinoma Ductal infiltrante
264	Tumor filoide maligno com componente estromal maligno heterólogo tipo Lipossarcoma.
270	Carcinoma infiltrante do tipo não especial, grau histológico final de Nottingham II. Presença de carcinoma ductal in situ de baixo grau, do tipo sólido, correspondendo a cerca de 5% do total da neoplasia.
357	Fibroadenoma em quadrante superior externo
378	Carcinoma ductal infiltrante grau 3.
380	Carcinoma ductal infiltrante grau 2.
382	Carcinoma ductal infiltrante.

Tabela A.2 - Mamas saudáveis

Mama	Resultado da mamografia de rastreio
33	BI-RADS Categoria 1
34	BI-RADS Categoria 1
35	BI-RADS Categoria 2
36	BI-RADS Categoria 1
37	BI-RADS Categoria 2
38	BI-RADS Categoria 1
39	BI-RADS Categoria 1
40	BI-RADS Categoria 1
41	BI-RADS Categoria 2
42	BI-RADS Categoria 2
43	BI-RADS Categoria 2
44	BI-RADS Categoria 2
45	BI-RADS Categoria 2
46	BI-RADS Categoria 1
47	BI-RADS Categoria 2
48	BI-RADS Categoria 2
49	BI-RADS Categoria 1
50	BI-RADS Categoria 2
51	BI-RADS Categoria 1

52	BI-RADS Categoria 1
53	BI-RADS Categoria 2
54	BI-RADS Categoria 2
55	BI-RADS Categoria 1
56	BI-RADS Categoria 1
57	BI-RADS Categoria 2
58	BI-RADS Categoria 2
59	BI-RADS Categoria 2
60	BI-RADS Categoria 2
61	BI-RADS Categoria 1
62	BI-RADS Categoria 1
63	BI-RADS Categoria 1
64	BI-RADS Categoria 2

APÊNCIA B - RESULTADOS DA CLASSIFICAÇÃO POR GRUPO DE CARACTERÍSTICAS

As Tabelas a seguir apresentam os resultados da classificação para cada grupo de características considerando a variação dos valores de k e m, bem como a variação entre C-SVC e Nu-SVC. Os melhores resultados de cada grupo estão em destaque.

Tabela B.1 - Resultado da classificação a partir das características das séries temporais formadas pela média e pelo desvio padrão das ROIs sequenciais das matrizes de temperatura.

	K	m	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	100,00%	88,89%	93,75%	0,89	0,938
	2	0	100,00%	88,89%	93,75%	0,89	0,938
	2	1	100,00%	86,49%	92,19%	0,87	0,922
	3	0	96,77%	93,94%	95,31%	0,91	0,953
	3	1	100,00%	84,21%	90,63%	0,84	0,906
	3	2	100,00%	91,43%	95,31%	0,91	0,953
	4	0	100,00%	91,43%	95,31%	0,91	0,953
	4	1	100,00%	88,89%	93,75%	0,89	0,938
	4	2	100,00%	86,49%	92,19%	0,87	0,922
	4	3	100,00%	91,43%	95,31%	0,91	0,953
	4	3	100,00%	91,43%	95,31%	0,91	0,953
Nu-SVC	1	0	100,00%	88,89%	93,75%	0,89	0,938
	2	0	100,00%	88,89%	93,75%	0,89	0,938
	2	1	100,00%	91,43%	95,31%	0,91	0,953
	3	0	100,00%	94,12%	96,88%	0,94	0,969
	3	1	100,00%	88,89%	93,75%	0,89	0,938
	3	2	100,00%	91,43%	95,31%	0,91	0,953
	4	0	100,00%	91,43%	95,31%	0,91	0,953
	4	1	100,00%	94,12%	96,88%	0,94	0,969
	4	2	100,00%	88,89%	93,75%	0,89	0,938
	4	3	100,00%	88,89%	93,75%	0,89	0,938
	4	3	100,00%	88,89%	93,75%	0,89	0,938

Tabela B.2 - Resultado da classificação a partir das características das séries temporais formadas pelo valor do maior centroide das ROIs sequenciais das matrizes de temperatura.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	84,85%	87,10%	85,94%	0,72	0,859
	2	0	84,85%	87,10%	85,94%	0,72	0,859
	2	1	87,10%	84,85%	85,94%	0,72	0,859
	3	0	87,10%	84,85%	85,94%	0,72	0,859
	3	1	81,82%	83,87%	82,81%	0,66	0,828
	3	2	84,85%	87,10%	85,94%	0,72	0,859
	4	0	79,41%	83,33%	81,25%	0,63	0,813
	4	1	90,32%	87,88%	89,06%	0,78	0,891
	4	2	90,32%	87,88%	89,06%	0,78	0,891
	4	3	90,32%	87,88%	89,06%	0,78	0,891
Nu-SVC	1	0	90,32%	87,88%	89,06%	0,78	0,891
	2	0	90,32%	87,88%	89,06%	0,78	0,891
	2	1	87,50%	87,50%	87,50%	0,75	0,875
	3	0	84,38%	84,38%	84,38%	0,69	0,844
	3	1	82,35%	86,67%	84,38%	0,69	0,844
	3	2	82,35%	86,67%	84,38%	0,69	0,844
	4	0	79,41%	83,33%	81,25%	0,63	0,813
	4	1	87,10%	84,85%	85,94%	0,72	0,859
	4	2	87,50%	87,50%	87,50%	0,75	0,875
	4	3	84,85%	87,10%	85,94%	0,72	0,859

Tabela B.3 - Resultado da classificação a partir das características das séries temporais formadas pelas comparações de histogramas das ROIs sequenciais das imagens em escala de cinza.

	K	m	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	64,52%	63,64%	64,06%	0,28	0,641
	2	0	59,38%	59,38%	59,38%	0,19	0,595
	2	1	77,78%	70,27%	73,44%	0,48	0,734
	3	0	59,38%	59,38%	59,38%	0,19	0,594
	3	1	91,30%	73,17%	79,69%	0,65	0,797
	3	2	70,59%	57,45%	60,94%	0,28	0,609
	4	0	58,62%	57,14%	57,81%	0,16	0,578
	4	1	81,48%	72,97%	76,56%	0,55	0,766
	4	2	77,78%	60,87%	65,63%	0,39	0,656
	4	3	66,67%	56,52%	59,38%	0,23	0,594
Nu-SVC	1	0	55,17%	54,29%	54,69%	0,10	0,547
	2	0	51,52%	51,61%	51,56%	0,03	0,516
	2	1	88,46%	76,32%	81,25%	0,65	0,813
	3	0	60,00%	62,07%	60,94%	0,22	0,609
	3	1	95,83%	77,50%	84,38%	0,73	0,844
	3	2	77,78%	70,27%	73,44%	0,48	0,734
	4	0	42,42%	41,94%	42,19%	-0,15	0,422
	4	1	92,00%	76,92%	82,81%	0,69	0,828
	4	2	63,33%	61,76%	62,50%	0,25	0,625
	4	3	55,56%	54,05%	54,69%	0,10	0,547

Tabela B.4 - Resultado da classificação a partir das características das séries temporais formadas pelos descritores de textura horizontal das ROIs sequenciais das imagens em escala de cinza.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	60,00%	58,82%	59,38%	0,19	0,594
	2	0	62,96%	59,46%	60,94%	0,22	0,609
	2	1	69,57%	60,98%	64,06%	0,31	0,641
	3	0	68,00%	61,54%	64,06%	0,30	0,641
	3	1	72,73%	61,90%	65,63%	0,35	0,656
	3	2	67,86%	63,89%	65,63%	0,32	0,656
	4	0	70,37%	64,86%	67,19%	0,35	0,672
	4	1	75,00%	65,00%	68,75%	0,40	0,688
	4	2	62,96%	59,46%	60,94%	0,22	0,609
	4	3	65,00%	56,82%	59,38%	0,22	0,594
Nu-SVC	1	0	74,29%	79,31%	76,56%	0,54	0,766
	2	0	73,53%	76,67%	75,00%	0,50	0,75
	2	1	74,19%	72,73%	73,44%	0,47	0,734
	3	0	75,68%	85,19%	79,69%	0,61	0,797
	3	1	69,44%	75,00%	71,88%	0,44	0,719
	3	2	76,92%	68,42%	71,88%	0,45	0,719
	4	0	83,87%	81,82%	82,81%	0,66	0,828
	4	1	72,41%	68,57%	70,31%	0,41	0,703
	4	2	74,07%	67,57%	70,31%	0,42	0,703
	4	3	77,42%	75,76%	76,56%	0,53	0,766

Tabela B.5 - Resultado da classificação a partir das características das séries temporais formadas pelos descritores de textura vertical das ROIs sequenciais das imagens em escala de cinza.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	65,71%	68,97%	67,19%	0,35	0,672
	2	0	71,88%	71,88%	71,88%	0,44	0,719
	2	1	67,65%	70,00%	68,75%	0,38	0,688
	3	0	69,70%	70,97%	70,31%	0,41	0,703
	3	1	66,67%	71,43%	68,75%	0,38	0,688
	3	2	61,76%	63,33%	62,50%	0,25	0,625
	4	0	76,92%	68,42%	71,88%	0,45	0,719
	4	1	68,57%	72,41%	70,31%	0,41	0,703
	4	2	64,29%	61,11%	62,50%	0,25	0,625
	4	3	70,00%	59,09%	62,50%	0,29	0,625
Nu-SVC	1	0	75,86%	71,43%	73,44%	0,47	0,734
	2	0	85,71%	77,78%	81,25%	0,64	0,813
	2	1	81,48%	72,97%	76,56%	0,55	0,766
	3	0	80,00%	76,47%	78,13%	0,57	0,781
	3	1	72,73%	74,19%	73,44%	0,47	0,734
	3	2	92,59%	81,08%	85,94%	0,74	0,859
	4	0	86,21%	80,00%	82,81%	0,66	0,828
	4	1	78,13%	78,13%	78,13%	0,56	0,781
	4	2	73,33%	70,59%	71,88%	0,44	0,719
	4	3	81,48%	72,97%	76,56%	0,55	0,766

Tabela B.6 - Resultado da classificação a partir das características das séries temporais formadas pelos índices de diversidade horizontais das ROIs sequenciais das imagens em escala de cinza.

	K	m	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	3,13%	3,13%	3,13%	-0,93	0,031
	2	0	33,33%	44,90%	42,19%	-0,21	0,422
	2	1	12,50%	37,50%	31,25%	-0,50	0,313
	3	0	32,00%	38,46%	35,94%	-0,29	0,359
	3	1	53,33%	51,02%	51,56%	0,04	0,516
	3	2	26,67%	42,86%	39,06%	-0,30	0,391
	4	0	52,94%	51,06%	51,56%	0,04	0,516
	4	1	41,67%	48,08%	46,88%	-0,10	0,469
	4	2	3,13%	3,13%	3,13%	-0,93	0,031
	4	3	6,25%	6,25%	6,25%	-0,87	0,063
Nu-SVC	1	0	63,64%	64,52%	64,06%	0,28	0,641
	2	0	70,83%	62,50%	65,63%	0,33	0,656
	2	1	66,67%	62,16%	64,06%	0,29	0,641
	3	0	68,97%	65,71%	67,19%	0,35	0,672
	3	1	64,52%	63,64%	64,06%	0,28	0,641
	3	2	66,67%	64,71%	65,63%	0,31	0,656
	4	0	68,97%	65,71%	67,19%	0,35	0,672
	4	1	72,00%	64,10%	67,19%	0,36	0,672
	4	2	68,75%	68,75%	68,75%	0,38	0,688
	4	3	61,54%	57,89%	59,38%	0,19	0,594

Tabela B.7 - Resultado da classificação a partir das características das séries temporais formadas pelos índices de diversidade verticais das ROIs sequenciais das imagens em escala de cinza.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	0,00%	0,00%	0,00%	-1,00	0
	2	0	10,00%	14,71%	12,50%	-0,75	0,125
	2	1	0,00%	0,00%	0,00%	-1,00	0
	3	0	3,13%	3,13%	3,13%	-0,93	0,031
	3	1	12,90%	15,15%	14,06%	-0,71	0,141
	3	2	0,00%	0,00%	0,00%	-1,00	0
	4	0	21,05%	37,78%	32,81%	-0,41	0,328
	4	1	6,06%	3,23%	4,69%	-0,90	0,047
	4	2	3,03%	0,00%	1,56%	-0,97	0,016
	4	3	0,00%	0,00%	0,00%	-1,00	0
Nu-SVC	1	0	56,67%	55,88%	56,25%	0,13	0,563
	2	0	52,38%	51,16%	51,56%	0,04	0,516
	2	1	59,38%	59,38%	59,38%	0,19	0,594
	3	0	44,00%	46,15%	45,31%	-0,09	0,453
	3	1	64,00%	58,97%	60,94%	0,23	0,609
	3	2	56,00%	53,85%	54,69%	0,10	0,547
	4	0	58,33%	55,00%	56,25%	0,13	0,563
	4	1	61,54%	57,89%	59,38%	0,19	0,594
	4	2	69,57%	60,98%	64,06%	0,31	0,641
	4	3	61,54%	57,89%	59,38%	0,19	0,594

Tabela B.8 - Resultado da classificação a partir das características das séries temporais formadas pelas comparações de spatiograms das ROIs sequenciais das imagens em escala de cinza.

	K	m	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	71,43%	60,47%	64,06%	0,32	0,641
	2	0	66,67%	56,52%	59,38%	0,23	0,594
	2	1	70,83%	62,50%	65,63%	0,33	0,656
	3	0	61,90%	55,81%	57,81%	0,18	0,578
	3	1	80,00%	76,47%	78,13%	0,57	0,781
	3	2	66,67%	60,00%	62,50%	0,27	0,625
	4	0	62,50%	57,50%	59,38%	0,20	0,594
	4	1	82,76%	77,14%	79,69%	0,60	0,797
	4	2	50,00%	50,00%	50,00%	0,00	0,500
	4	3	54,55%	50,94%	51,56%	0,06	0,516
Nu-SVC	1	0	55,56%	57,14%	56,25%	0,13	0,563
	2	0	70,37%	64,86%	67,19%	0,35	0,672
	2	1	75,68%	85,19%	79,69%	0,61	0,797
	3	0	61,11%	64,29%	62,50%	0,25	0,625
	3	1	78,38%	88,89%	82,81%	0,67	0,828
	3	2	55,17%	54,29%	54,69%	0,10	0,547
	4	0	66,67%	64,71%	65,63%	0,31	0,656
	4	1	81,25%	81,25%	81,25%	0,63	0,813
	4	2	48,48%	48,39%	48,44%	-0,03	0,484
	4	3	54,17%	52,50%	53,13%	0,07	0,531

Tabela B.9 - Resultado da classificação a partir das características das séries temporais formadas pelos valores da lacunaridade das ROIs sequenciais das imagens em escala de cinza.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	57,45%	70,59%	60,94%	0,28	0,609
	2	0	54,90%	69,23%	57,81%	0,24	0,578
	2	1	59,57%	76,47%	64,06%	0,36	0,641
	3	0	59,18%	80,00%	64,06%	0,39	0,641
	3	1	59,09%	70,00%	62,50%	0,29	0,625
	3	2	54,00%	64,29%	56,25%	0,18	0,563
	4	0	56,60%	81,82%	60,94%	0,38	0,609
	4	1	56,10%	60,87%	57,81%	0,17	0,672
	4	2	61,54%	68,00%	64,06%	0,30	0,641
	4	3	55,77%	75,00%	59,38%	0,31	0,594
Nu-SVC	1	0	52,63%	53,85%	53,13%	0,07	0,531
	2	0	50,00%	50,00%	50,00%	0,00	0,5
	2	1	46,51%	42,86%	45,31%	-0,10	0,453
	3	0	52,78%	53,57%	53,13%	0,06	0,531
	3	1	54,29%	55,17%	54,69%	0,10	0,547
	3	2	47,37%	46,15%	46,88%	-0,06	0,469
	4	0	45,24%	40,91%	43,75%	-0,13	0,438
	4	1	55,26%	57,69%	56,25%	0,13	0,563
	4	2	52,94%	53,33%	53,13%	0,06	0,531
	4	3	45,16%	45,45%	45,31%	-0,09	0,453

Tabela B.10 - Resultado da classificação a partir das características das séries temporais formadas pelos valores de todas as características das ROIs sequenciais.

	K	M	Sensibilidade	Especificidade	Acurácia	Índice de Youden	Área sob Curva ROC
C-SVC	1	0	85,71%	77,78%	81,25%	0,64	0,813
	2	0	85,19%	75,68%	79,69%	0,61	0,797
	2	1	95,83%	77,50%	84,38%	0,73	0,844
	3	0	84,62%	73,68%	78,13%	0,58	0,781
	3	1	92,59%	81,08%	85,94%	0,74	0,859
	3	2	96,00%	79,49%	85,94%	0,76	0,859
	4	0	84,00%	71,79%	76,56%	0,56	0,766
	4	1	96,00%	79,49%	85,94%	0,76	0,859
	4	2	95,45%	73,81%	81,25%	0,69	0,813
	4	3	91,30%	73,17%	79,69%	0,65	0,797
Nu-SVC	1	0	96,67%	91,18%	93,75%	0,88	0,938
	2	0	96,55%	88,57%	92,19%	0,85	0,922
	2	1	96,77%	93,94%	95,31%	0,91	0,953
	3	0	90,00%	85,29%	87,50%	0,75	0,875
	3	1	93,55%	90,91%	92,19%	0,85	0,922
	3	2	100,00%	94,12%	96,88%	0,94	0,969
	4	0	100,00%	86,49%	92,19%	0,87	0,922
	4	1	96,77%	93,94%	95,31%	0,91	0,953
	4	2	100,00%	94,12%	96,88%	0,94	0,969
	4	3	96,88%	96,88%	96,88%	0,93	0,969