UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Caio Eduardo Falcão Matos

Diagnóstico de Câncer de Mama em Imagens Mamográficas através de Características Locais e Invariantes

São Luís
2017

Caio Eduardo Falcão Matos

Diagnóstico de câncer de mama em imagens mamográficas através de características locais e invariantes

Dissertação apresentada ao Curso de Programa de Pós-Graduação em Ciência da Computação da UFMA como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Universidade Federal do Maranhão – UFMA

Programa de Pós-Graduação Ciência da Computação – PPGCC

Orientador: Prof. Dr. Geraldo Braz Junior

Coorientador: Prof. Dr. Anselmo Cardoso de Paiva

São Luís

2017

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

Matos, Caio Eduardo Falcão.
Diagnóstico de Câncer de Mama em Imagens Mamográficas através de Características Locais e Invariantes / Caio Eduardo Falcão Matos. - 2017.
86 f.

Orientador(a): Geraldo Braz Junior Anselmo Cardoso de Paiva.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luís/MA, 2017.

1. Câncer de Mama. 2. Detecção de Características Locais. 3. Mamografia. 4. Reconhecimento de Padrões. 5. Representação de Características. I. Anselmo Cardoso de Paiva, Geraldo Braz Junior. II. Título.

Este trabalho é dedicado em primeiro lugar a Deus que,
com sua imensa misericórdia, concedeu-me força e coragem
durante toda esta longa caminhada.

Agradecimentos

Primeiramente a Deus que pela sua imensa misericórdia me proporciona o dom da vida a cada manhã. Agradeço por mais esta vitória em conseguir chegar até este momento.

Aos meus pais Zenilde Falcão Matos e José de Ribamar Matos, por todo amor, sacrifício, motivação, apoio incondicional, dedicação e muitas vezes até renúncia a mim dirigida ao longo de toda a vida, no qual não poderia obter êxito sem sua ajuda.

A minha namorada Aline Klycia, por todo amor e incentivo mesmo nos momentos de tribulações e pressões vivenciados ao longo destes anos.

Ao meu orientador Geraldo por todo apoio, paciência, disponibilidade e, sobretudo conhecimento a mim transmitido.

Ao meu coorientador Anselmo que me concedeu minha primeira bolsa de projeto, pela qual ingressei no Núcleo de Computação Aplicada, onde as portas da pesquisa e desenvolvimento foram abertas a mim.

Aos professores do programa de pós-gradução em Ciência da Computação, por todo conhecimento transmitido em sala de aula.

Aos amigos de graduação e mestrado pelo total de seis anos e meio de convivência. Aos amigos Caio Belfort, Giovanni Lucca, Johnatan Carvalho, João Bandeira, Jefferson Alves, Whesley Dantas, um agradecimento especial pela contribuição em todas as disciplinas, nesta dissertação e no decorrer de todo o mestrado.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

"A persistência é o menor caminho do êxito". (Charles Chaplin)

Resumo

O câncer de mama é apontado como uma das principais causas de morte entre as mulheres. As altas taxas de mortalidade e registros de ocorrência desse câncer em todo o mundo evidenciam a importância do desenvolvimento e investigação de meios para a detecção e diagnóstico precoce dessa doença. Sistemas de Detecção e Diagnóstico auxiliados por computador (*Computer Aided Detection/Diagnosis*) vêm sendo usados e propostos como forma de auxílio aos profissionais de saúde. Este trabalho propõe uma metodologia para discriminação de padrões de malignidade e benignidade de massas em imagens de mamografia através da análise de características locais. Para tanto, a metodologia combina detectores e descritores de características locais com um modelo de representação de dados para a análise, tanto de textura quanto de geometria em regiões extraídas das mamografias. São utilizados os detectores SIFT, SURF e ORB, e descritores HOG, LBP, BRIEF e Haar Wavelet. Com as características geradas é aplicado o modelo *Bag of Features* em uma etapa de representação que objetiva prover nova representação dos dados e por conseguinte diminuir a dimensionalidade do espaço de características. Por fim, esta nova representação é classificada utilizando três abordagens: Máquina de Vetores de Suporte, *Random Forests* e *Adaptive Boosting* visando diferenciar as massas malignas e benignas. A metodologia contém resultados promissores para o diagnóstico de massas malignas e benignas fomentando que as características locais geradas pelos descritores e detectores produzem um conjunto descriminate satisfatório.

Palavras-chave: Reconhecimento de Padrões, Câncer de Mama, Mamografia, Detecção de Características, Representação de Características.

Abstract

Breast cancer is one of the leading causes of death among women over the world. The high mortality rates that cancers achieves across the world highlight the importance of developing and investigating the means for the early detection and diagnosis of this disease. Computer Detection and Diagnosis Systems (Computer Assisted Detection / Diagnosis) have been used and proposed as a way to help health professionals. This work proposes a new methodology for discriminating patterns of malignancy and benignity of masses in mammography images by analysis of local characteristics. To do so, it is proposed a combined methodology of feature detectors and descriptors with a model of data representation for an analysis. The goal is to capture both texture and geometry in areas of mammograms. We use the SIFT, SURF and ORB detectors, and the descriptors HOG, LBP, BRIEF and Haar Wavelet. The generated characteristics are coded by a bag of features model to provide a new representation of the data and therefore decrease a dimensionality of the space of characteristics. Finally, this new representation is classified using three approaches: Support Vector Machine, Random Forest, and Adaptive Boosting to differentiate as malignant and benign masses. The methodology provides promising results for the diagnosis of malignant and benign mass encouraging that as local characteristics generated by descriptors and detectors produce a satisfactory a discriminating set.

Keywords: Pattern Recognition, Breast Cancer, Mammography, Feature Detection, Feature Representation.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACR	*American College of Radiology*
BOF	*Bag of Features*
BRIEF	*Binary Robust Independent Elementary Features*
CAD	*Computer-Aided Detection*
CADx	*Computer-Aided Diagnosis*
DDSM	*Digital Database for Screening Mamography*
DOG	*Difference of Gaussian*
FAST	*Features From Accelerated Segment Test*
GDFNN	*Generalized Dynamic Fuzzy Neural Networks*
HOG	*Histogram Oriented Gradients*
ICA	*Independent Component Analysis*
INCA	Instituto Nacional do Câncer
MVS	Máquina de Vetores de Suporte
ORB	*Oriented FAST and Rotated BRIEF*
RF	*Random Forests*
SIFT	*Scale Invariant Feature Transform*
SURF	*Speeded Up Robust Features*
SVFNN	*Support Vector Based Fuzzy Neural Network*
TUC	*Texture Unit Coding*
VP	*Verdadeiro Positivo*
VP	*Verdadeiro Negativo*
VLAD	*Vector of Locally Aggregated Descriptors*

Sumário

1 Introdução

O câncer é uma das principais causas de mortes em todo o mundo, com 8,2 milhões de mortes apenas no ano de 2012 (IARC - International Agency for Research on Cancer, 2012). Mais da metade de todas as mortes por câncer são devidas principalmente aos cânceres de pulmão, estômago, fígado, colorretal e mama feminino. Cerca de 53% das mortes por câncer ocorrem em países em desenvolvimento onde o IDH - Índice de Desenvolvimento Humano é médio ou baixo (JEMAL et al., 2011). No Brasil, estima-se cerca de 596.070 novos casos de câncer para o ano de 2017, sendo 295.200 em homens e 300.870 em mulheres. O câncer de pele não melanoma será o mais incidente em ambos os sexos correspondendo a 175.760 novos casos a cada ano, sendo 80.850 em homens e 94.910 em mulheres. Em seguida, os cânceres mais incidentes para os homens serão os de próstata (61.200 novos casos/ano), pulmão (17.330), cólon e reto (16.660), estômago (12.920), cavidade oral (11.140), esôfago (7.950), bexiga (7.200), laringe (6.360) e leucemias (5.540). Entre as mulheres, as maiores incidências serão de cânceres de mama (57.960), cólon e reto (17.620), colo do útero (16.340), pulmão (10.860), estômago (7.600), corpo do útero (6.950), ovário (6.150), glândula tireoide (5.870) e linfoma não-Hodgkin (5.030) (INCA - Instituto Nacional do Câncer, 2016b).

O câncer de mama é a principal causa de mortes relacionadas ao câncer entre a população feminina, tornando-se a quinta causa de morte por câncer no mundo (522.000 óbitos) (WHO - World Health Organization, 2014). Em locais de poucos recursos, a maioria das mulheres são diagnosticadas em estágios avançados. Pacientes diagnosticados nestas condições possuem chance de sobrevivência variando de 10 a 40% (WHO - World Health Organization, 2014). Em locais onde a detecção precoce e o tratamento básico estão disponíveis e acessíveis, a taxa de sobrevivência para o câncer de mama é superior a 80% (WHO - World Health Organization, 2014). No Brasil, estima-se mais de 57 mil novos casos para o ano de 2017. Este câncer é o mais frequente nas mulheres das regiões Sudeste, Sul, Centro-Oeste e Nordeste. Na região Norte é o segundo mais incidente atrás somente do câncer de colo do útero (INCA - Instituto Nacional do Câncer, 2015).

Diversos fatores contribuem para o aumento do risco de desenvolver a doença, tais como: idade, fatores endócrinos ou relativos à história reprodutiva, fatores relativos ao ambiente ou comportamento, e fatores genéticos ou hereditários. O primeiro fator é a idade, que assim como em vários outros tipos de câncer, é o fator de maior contribuição para o aumento do risco do desenvolvimento do câncer de mama. O segundo fator relaciona-se ao estímulo ou consumo hormonal, dentre estes podemos citar: menopausa tardia, primeira gravidez após os 30 anos, nuliparidade[1] e uso de contraceptivos orais e de terapia

[1] Estado da mulher que nunca engravidou ou nunca teve filhos biológicos

de reposição hormonal pós-menopausa. No terceiro fator estão inclusos o sobrepeso e obesidade, tabagismo, exposição excessiva a radiação e ingestão de bebida alcoólica. O quarto e último fator, consiste na presença de mutações em determinados gene transmitidos na família (INCA - Instituto Nacional do Câncer, 2016a).

Uma importante iniciativa para a prevenção do câncer de mama é a descoberta precoce por meio dos programas de rastreamento (WHO - World Health Organization, 2014). O rastreamento é uma estratégia voltada a mulheres na faixa etária onde os benefícios dessas práticas sejam favoráveis. Essa estratégia consiste na aplicação sistemática de teste ou exames, como o exame de mamografia. Este exame busca em uma população assintomática a identificação de lesões em fase inicial de desenvolvimento, indicado a mulheres de 50 a 69 anos a cada dois anos (INCA - Instituto Nacional do Câncer, 2014).

A avaliação do exame mamográfico exige um grande conhecimento e habilidade do especialista em diagnosticar possíveis patologias, incluindo o câncer de mama, podendo este ser relacionado a nódulos malignos ou benignos. Nas últimas décadas, técnicas computacionais vêm sendo desenvolvidas com o propósito de detectar automaticamente estruturas que possam estar associadas ao câncer de mama em exames mamográficos, de modo a evitar procedimentos desnecessárioss(GIGER, 2000). Os sistemas de diagnóstico de doenças auxiliados por computador (CADx – *Computer-Aided Diagnosis*) são ferramentas que aplicam estas computacionais para fornecer outras informações ao especialista no diagnóstico. Estes sistemas normalmente possuem uma etapa de classificação, onde é determinado, por exemplo, a presença ou ausência de uma doença de interesse.

Uma grande parte destas técnicas fazem uso de exames de imagem para inferir algum diagnóstico através de técnicas de detecção e reconhecimento de padrões. Um padrão é descrito como uma coleção de características que definem um objeto e o distingue em relação aos demais objetos presentes numa imagem. Uma maneira de descrever um padrão consiste em representar o objeto na forma de pontos de interesse que forneçam informação consistente sobre regiões de alta energia e discriminatória (ANTANI; KASTURI; JAIN, 2002). O principal desafio é descrever esses pontos de interesse de maneira eficiente e compacta levando ainda em consideração invariância a escala, translação, rotação e ruídos (ALAHI; ORTIZ; VANDERGHEYNST, 2012).

Esta dissertação de mestrado objetiva o diagnóstico de câncer de mama aplicando um conjunto de detectores e descritores de características locais, *Scale-invariant feature transform* (SIFT) (LOWE, 2004), *Speeded Up Robust Feature* (SURF) (BAY et al., 2008), *Oriented Fast and Rotated BRIEF* (ORB) (RUBLEE et al., 2011), *Local Binary Pattern* (LBP) (OJALA; PIETIKÄINEN; HARWOOD, 1996), para a identificação de pontos e extração de características locais em regiões de massa, juntamente com o *Bag of Features* (BoF) (AVILA, 2013) para representar as características extraídas através da análise de sua importância coletiva. Cada região de massa, obtida da base *Digital Database for*

Screenning Mamography (DDSM) (HEATH et al., 1998) é submetida a este processo de caracterização com o objetivo de reconhecer os padrões de benignidade e malignidade. Três diferentes classificadores, *Adaptative Boosting* (AdaBoost) (FREUND; SCHAPIRE, 1995), *Random Forests* (BREIMAN et al., 1984) e Máquina de Vetores de Suporte (MVS) (VAPNIK, 1998), foram utilizados neste processo buscando a comparação e seleção do melhor modelo de classificação para o diagnóstico do câncer de mama.

O desenvolvimento desta dissertação objetiva contribuir para a construção de uma ferramenta automatizada de modo a auxiliar o especialista no diagnóstico de uma anomalia presente na mamografia. Este trabalho também apresenta como contribuição o ajuste de técnicas usadas comumente em *object matching* e *tracking* para a descrição e extração de padrões sobre coleções de imagens de mamografias digitais, nas quais a principal abordagem consiste na análise de textura e forma em regiões locais. Outra contribuição destes trabalho consiste na avaliação do desempenho de classificadores supervisionados no contexto voltado ao câncer de mama, para diferenciação de malignidade e benignidade.

1.1 Trabalhos Relacionados

A diferenciação do padrão de malignidade e benignidade para a classificação de massas em imagens mamográficas tem como objetivo auxiliar especialistas no diagnóstico de câncer de mama utilizando técnicas não invasivas e computacionais, reduzindo assim procedimentos como a biópsia da massa. Metodologias ou sistemas CADx são desenvolvidos para realizar o reconhecimento automático de imagens médicas. Eles possuem em sua composição métodos de extração, seleção e classificação de características. A extração de características é comumente realizada através da análise de textura e forma das regiões de massas, sendo que sua análise pode ser realizada de maneira individual ou combinada. A literatura disponível apresenta diversos trabalhos voltados a esta análise.

A seguir são apresentados alguns trabalhos desenvolvem a extração de características em imagens de mamografia baseada exclusivamente na análise de textura, objetivando o disgnóstico do câncer de mama.

Na metodologia proposta por Campos, Silva e Barros (2005), um método de extração de características de textura baseado em testes estatísticos foi utilizado para a classificação de massas em malignos, benignos ou normais. Sobre o método estatístico, é aplicada a Análise de Componentes Independentes (ICA) como técnica de extração, onde cada imagem mamográfica é representada como uma combinação linear de imagens básicas mutuamente independentes. Para a redução da dimensionalidade do espaço de características e classificação, as técnicas *forward-selection*, *Multilayer Perceptron Neural Network* (MLP) e *Probabilistic Neural Network* (PNN) foram usadas. Tento em vista a classificação do tecido mamário, é proposto em (MARTINS et al., 2006) um método de

extração e classificação baseado nas matrizes de co-ocorrência de níveis de cinza (GLCM) aplicadas para a análise de textura dos tecidos mamários. Esta abordagem evidencian o comportamento e dependência espacial entre os *pixels*[2]. As direções, 0°,45°,90°e 135°, foram aplicadas às matrizes permitindo assim, a extração de cinco características (contraste, homogeneidade, entropia, energia e momento de diferença inverso) das treze características originalmente propostas em (HARALICK; SHANMUGAM et al., 1973). As redes neurais Bayesianas são aplicadas para a classificação do tecido mamário em maligno ou benigno.

As matrizes de co-ocorrência são também objetos de investigação em (BEURA; MAJHI; DASH, 2015) como forma de extração de características em imagens mamográficas. Neste, as matrizes são combinadas a transformada discreta de *wavelet* bidimensional (2D-DWT). Como método de classificação, as redes MLP foram utilizadas.

Em (BRAZ JUNIOR et al., 2009) é apresentada uma abordagem baseada em medidas geoestatísticas para a classificação de tecidos mamários em normais e anormais quanto em malignos e benigno. O objetivo deste trabalho consiste no estudo da aplicabilidade do índice de Moran's e coeficiente de Geary's, ambas medidas geoestatísticas, como descritores de textura. As imagens mamográficas são quantizadas em intervalos de níveis de cinza $(2^8, 2^7, 2^6, 2^5, 2^4, 2^3)$ para obter diferentes informações de textura. As imagens quantizadas são submetidas a análise de textura espacial através do índice de Moran's e coeficiente de Geary's. A máquina de vetores de suporte (MVS) foi escolhida para a classificação das características geradas pelos descritores.

Moayedi et al. (2010) propõe uma abordagem para extração de características baseada nas transformadas de *contourlet*. Seis descritores sobre o resultado da transformada de *counterlet* foram aplicados (media, desvio padrão, energia, entropia, assimetria) para a extração. Na classificação das massas, são comparadas duas abordagens para a formação do modelo da MVS, uma utilizando o *Kernel Radial Basic Function* (RBF) e outra baseada nas *Fuzzy Neural Network* (FNN). Outro trabalho que utiliza o conceito de transformadas para a extração de característica de textura é (DHAHBI; BARHOUMI; ZAGROUBA, 2015), onde é proposto um método para o diagnóstico de câncer de mama baseado na transformada de *curvelet* e na teoria do momento, ambas medidas estatísticas. Duas abordagens são investigadas para calcular as transformada de curvelet sendo a primeira calculada para cada nível, CLM - *Curvelet Level Moments*, e a segunda por banda, CBM - *Curvelet Band Moments*. São aplicados quatro descritores de momento (média, variância, assimetria e curtose) que geram para cada abordagem (CLM e CBM), respectivamente, 16 e 64 características. O classificador K-Nearest Neighbour (KNN) é aplicado como método de classificação das características geradas.

A análise de dimensões fractais (FD) e assinatura fractal (FS) foram propostos em (DON et al., 2012) como descritores de textura aplicados em imagens mamográficas

[2] Abreviação de *Picture Element*

para o diagnóstico do câncer de mama. As redes MLP foram utilizadas como método de classificação. Hussain et al. (2014) utiliza *Gabor filter bank* (GBFB) para extrair pequenos padrões que sejam discriminantes em relação aos aspectos de textura da mama. As características geradas são submetidas a uma MVS, utilizando o kernel RBF para classificação.

Em (ROCHA et al., 2016) é investigada a capacidade de alguns índices de diversidades presentes na ecologia como forma de descrever a textura do tecido mamário. Neste trabalho os índices de diversidade, Shannon, McIntosh, Simpson, Gleason e Menhinick, são combinados a descritores amplamente utilizados na literatura como, matrizes de coocorrência de níveis de cinza (GLCM) e padrões binários locais (LBP). Tendo em vista a análise da textura em diversos espaços de características, são aplicados cinco quantizações (128,64,32,16,8) onde cada nível de cinza quantizado representa uma espécie do índice. As direções aplicadas as matrizes de coocorrências foram de 0°, 45°, 90°, 135°. Utilizou-se para classificação a MVS com função de base radial (RBF) como *kernel*.

A análise geométrica dos nódulos tornou-se uma abordagem amplamente aplicada como forma de extração de características para a diferenciação dos padrões de malignidade e benignidade presentes nas imagens de mamografia. A seguir, alguns trabalhos que aplicam esta abordagem para o diagnóstico do câncer de mama são apresentados.

A metodologia proposta em (GÖRGEL; SERTBAS; UCAN, 2013) baseia-se na combinação dos algoritmos *Local Seed Region Growing* (LSRG) e *Spherical Wavelet Transform* para forma o método de extração de características chamado de LSRG-SWT. Este método por ser uma variante do *Seed Region Growing* (SRG) possui como requisito a especificação de uma semente inicial que neste caso é o valor de intensidade do *pixel*. O método de reconhecimento aplicado neste trabalho foi a MVS.

Em (AYED; MASMOUDI; MASMOUDI, 2014) é proposta uma abordagem automática que realiza detecções das regiões de massas através do método *Improved Against Noise Gray Level and Local Difference* (IANGLLD). Neste método, características de geometria são extraídas, incluindo, circularidade e ângulo radial. As Redes Neurais Artificiais (ANN) são aplicadas para a classificação das massas. A abordagem aplicada em (WAJID; HUSSAIN, 2015) investiga a aplicação do LESH - *Local Energy-based Shape Histogram* como extrator de característica de geometria para o diagnóstico de câncer de mama. O método trabalha através da conversão de imagem em uma combinação de energias locais ao longo de diferentes orientações. O método de classificação utilizado foi a MVS.

A técnica de aprendizagem profunda, *Convolutional Neural Network* (CNN), é proposta em (AREVALO et al., 2015) para a classificação de massas malignas e benignas. Para tanto, características de geometria são extraídas utilizando uma arquitetura contendo duas camadas de convolução e *pooling* e uma camada conectada MLP. No estágio final, a penúltima camada da CNN é utilizada como vetor de características, sendo estes submetidos

a uma MVS para o aprendizado. Em (ABDEL-ZAHER; ELDEIB, 2016), foi utilizada outra técnica de aprendizagem profunda que visa o diagnóstico de câncer de mama. Neste trabalho, a retropropagação dos pesos presentes nas redes neurais são é utilizada para treinar uma *Deep Belief Network* (DBN). As características de forma em níveis mais abstrato são aprendidas através da DBN, sendo estas responsáveis pela classificação.

A extração de características, como apresentado anteriormente, pode ser baseada tanto através da análise de textura quanto pela geometria das massas. No entanto, muitos trabalhos aplicam uma combinação dessas duas abordagens visando o aumento no poder de diferenciação dos padrões de malignidade e benignidade. Alguns desses são apresentados na sequência.

Em (LIU; TANG, 2014) é proposta uma abordagem para o diagnóstico de câncer de mama em imagens mamográficas utilizando seleção geométrica, características de textura e um método de seleção de características baseado na integração da máquina de vetores de suporte (MVS) baseada em eliminação recursiva de características, com o método de seleção de característica denominado NMIFS - *Normalized Mutual Information Feature Selection*. Neste método, doze características de geometria são aplicadas, incluindo, compacidade, métricas de Fourier, tamanho do raio e orientação do gradiente relativa. As matrizes de co-ocorrência foram utilizadas para a análise de textura, sendo 19 características extraídas. Algumas destas são, contraste, autocorrelação, homogeneidade, variância, entre outros. A abordagem de seleção de características consiste na integração de uma variante da Máquina de Vetores de Suporte chamada *SVM-RFE* sendo esta baseada na eliminação de características recursivas, com uma seleção de características mutuamente normalizadas (NMIFS). Os métodos *Linear Discriminant Analysis* (LDA) com KNN e a MVS são investigados para a classificação.

As matrizes de co-ocorrência e os momentos de *Zernike* foram utilizados como forma de extração de características de geometria e textura em (ROUHI et al., 2015). As técnicas *Random Forest*, *Naive Bayes*, MVS e *k-Nearest Neighbor*(kNN) foram utilizados para a classificação das características geradas. A abordagem proposta em (KAUR, 2016) também utiliza as matrizes de co-ocorrência (GLCM,GLDM) para o diagnóstico do câncer de mama, em conjunto com características de geometria (Excentricidade, Área, Diâmetro equivalente, *Euler Number*, *Inertia*, *X-axis and Y-axis*) que também compõem a etapa de extração de características. A técnica de aprendizado de máquina usada para a classificação foi a CNN.

O método proposto por Abbas e Qaisar (2016), objetiva a classificação de massa em imagens mamográficas digitais, onde características de textura e geometria são extraídas através das técnicas *Local Binary Pattern Variance* (LBPV) e *Speed-up Robust Features* (SURF). A técnica de aprendizagem profunda, *deep belief network* (DBN), é utilizada para o treinamento e classificação das características invariantes combinadas. Este treinamento

é realizado através das *Restricted Boltzmann Machines* (RBMs).

Um resumo dos trabalhos relacionados desta seção é apresentado na Tabela 1, que contêm informações a respeito da(s) técnica(s) de extração de características, o classificador utilizado, a base de imagens mamográficas, a quantidade de amostras (M = maligno, B = benigno e N = Normal), e os resultados produzidos em termos de acurácia e área sobre a curva ROC(Az).

A revisão da literatura indica que os trabalhos que possuem como abordagem a extração de características locais apresentam, no geral, um desempenho superior aos demais que empregam a análise global no diagnóstico de câncer de mama. Por outro lado, sabe-se que informações como, quantidade e heterogeneidade das imagens mamográficas podem influenciar diretamente no desempenho da metodologia e consequentemente em sua precisão no diagnóstico.

Surge por tanto, o desafio de desenvolver uma metodologia baseada na análise de textura de maneira local em regiões de massa, de modo a permitir que características anteriormente desprezadas pela análise global possam, de maneira eficiente, ter sua probabilidade de malignidade e benignidade determinadas, mesmo nos casos de bases de imagens com elevado número de caso e alta heterogeneidade.

Tabela 1 – Resumo dos Trabalhos Relacionados

	Trabalhos	Técnica(s)	Classificador	Base	ROIs(B/M/N)	Acurácia	Az
Textura	(CAMPOS; SILVA; BARROS, 2005)	ICA	MLP PNN	MIAS	100 (50/50/200)	97,3%	-
	(MARTINS et al., 2006)	GLCM	RNB	mini-MIAS	119 (68/51)	86,84%	-
	(BRAZ JUNIOR et al., 2009)	Indices Geoestatisticos	MVS	DDSM	584 (273/311)	87,8%	-
	(MOAYEDI et al., 2010)	Contourlet	MVS	MIAS	90 (15/15/60)	96.6%	-
	(DHAHBI; BARHOUMI; ZAGROUBA, 2015)	Curvelet	KNN	mini-MIAS	252 (50/66/136)	91,27%	-
	(DON et al., 2012)	Fractal Dimension Fractal Signature	MLP	DDSM MIAS	316	98%	-
	(HUSSAIN et al., 2014)	GBFB	MVS	DDSM	512 (256/256/512)	85.53%	0.87
	(BEURA; MAJHI; DASH, 2015)	GLCM 2D-DWT	MLP	DDSM	250 (121/129/300)	97,4%	-
	(ROCHA et al., 2016)	GLCM	MVS	DDSM	1155 (530/625)	88,31%	0,88
Geometria	(GÖRGEL; SERTBAS; UCAN, 2013)	LSRG-SWT	MVS	I.U	60 (35/43)	93,59%	-
	(AYED; MASMOUDI; MASMOUDI, 2014)	IANGLLD Geometria	ANN	DDSM	100 (50/50)	98,9%	0.98
	(AREVALO et al., 2015)	CNN	MLP	BCDR	736 (310/426)	-	0,86
	(WAJID; HUSSAIN, 2015)	LESH	MVS	INbreast	396 (117/279)	99,0%	-
	(ABDEL-ZAHER; ELDEIB, 2016)	DBN	RNB	WBCD	690	99,68%	-
Geometria e Textura	(LIU; TANG, 2014)	GLCM Geometria	MVS	DDSM	826 (408/418)	-	0.9615
	(ROUHI et al., 2015)	GLCM Zernike Moments	MVS	DDSM	263 (93/170)	96,47%	0,95
	(ABBAS; QAISAR, 2016)	LBPV SURF	RBM	DDSM mini-MIAS	600 (300/300)	91,5%	0,91
	(KAUR, 2016)	GLCM Geometria	RNC	MIAS	30	96,66%	-

1.2 Objetivos

Este trabalho tem como objetivo geral o desenvolvimento de uma metodologia para a diferenciação dos padrões de malignidade e benignidade de massas em imagens de mamografias, de maneira a contribuir para o diagnóstico do câncer de mama.

Para tanto, propõem-se utilizar técnicas para a detecção (SIFT ((LOWE, 2004)), SURF ((BAY et al., 2008)), ORB ((RUBLEE et al., 2011))) e extração (LBP ((OJALA; PIETIKÄINEN; HARWOOD, 1996)), *Histograms of Oriented Gradients* ((LOWE, 2004)), Wavelet de Haar ((BAY et al., 2008)), *Binary Robust Independent Elementary Features* (RUBLEE et al., 2011)) de características de textura de maneira local, juntamente com um modelo de representação de dados (AVILA, 2013) e técnicas de reconhecimento de padrões ((BREIMAN et al., 1984), (FREUND; SCHAPIRE, 1995), (VAPNIK, 1998)).

Destacam-se como objetivos específicos deste trabalho:

- Adaptar as técnicas SIFT, SURF, ORB, comumente usadas em *object matching* e *tracking* para detecção e extração de características em regiões de imagens mamográficas;
- Realizar a análise de textura local através da combinação do descritor LBP juntamente com o detector SIFT como forma de extração de características locais;
- Análise e estudo da dimensionalidade do modelo de representação de características (vocabulário) no contexto das imagens mamográficas;
- Estruturar um modelo de representação de características tendo em vista o melhor desempenho das técnicas de reconhecimento de padrões;
- Análise e utilização de técnicas de reconhecimento de padrões, mais especificamente *Random Forests*, *Adaptative Boosting* e Máquina de Vetores de Suporte, para a diferenciação de padrões;
- Avaliação da metodologia proposta através de experimentos utilizando uma base pública de imagens mamográficas.

1.2.1 Contribuições

Como contribuições deste trabalho, pode-se destacar:

- Adaptação e avaliação de descritores locais como método de distinção de padrões em imagens mamográficas;
- Avaliação do desempenho do modelo de representação através da variação de sua dimensionalidade no âmbito das imagens mamográficas;

- Avaliação comparativa do desempenho de diferentes métodos de aprendizado de máquina para o diagnóstico de câncer de mama.

1.3 Organização do Trabalho

Este trabalho está organizado em cinco capítulos. No Capítulo 2 é apresentada a fundamentação teórica necessária para o desenvolvimento deste trabalho. Serão descritas as técnicas de detecção e descrição de características locais (Scale Invariant Feature Transform, Speed Up Robust Feature, Oriented FAST and Rotate BRIEF e Local Binary Pattern) para a análise da textura/geometria e técnicas de classificação (*Random Forest*, *Adaptative Boosting*, Máquina de vetores de Suporte).

Em seguida, no Capítulo 3, serão descritos as etapas da metodologia proposta tendo em vista o diagnóstico do câncer de mama, a partir de imagens de mamografias.

No Capítulo 4, os resultados obtidos através da metodologia proposta são apresentados e discutidos. Finalmente, no Capítulo 5, conclusões a respeito deste trabalho são apresentadas, evidenciando a eficiência dos métodos utilizados, bem como sugestões para trabalhos futuros.

2 Fundamentos Teóricos

Para melhor compreensão das técnicas e conceitos utilizados nesta dissertação, um conjunto de fundamentos teóricos são apresentados neste capítulo. São abordados inicialmente conceitos voltados ao câncer de mama e os principais exames utilizados para o diagnóstico, incluindo o exame de mamografia. Em seguida, conceitos relacionados a processamento de imagens, representações de características visuais utilizando descritores locais, dentre eles: *Scale-Invariant Feature Transform*, *Speeded Up Robust Features*, *Oriented FAST and Rotated BRIEF* e *Local Binary Pattern*, métodos de classificação e diferenciação de padrão utilizados *Random Forest*, *Adaptative Boosting* e Máquina de Vetores de Suporte.

2.1 O Câncer de Mama

A mama ou glândula mamária feminina é um órgão que situa-se na parede anterior do tórax, na parte superior e está apoiada sobre o músculo peitoral maior. É composta por lobos (glândulas produtoras de leite), por dutos (pequenos tubos que transportam o leite dos lobos ao mamilo) e por estroma (tecido adiposo e tecido conjuntivo que envolve os dutos, lobos, vasos sanguíneos e vasos linfáticos) (ACS - American Cancer Society, 2016), conforme ilustrado na Figura 1.

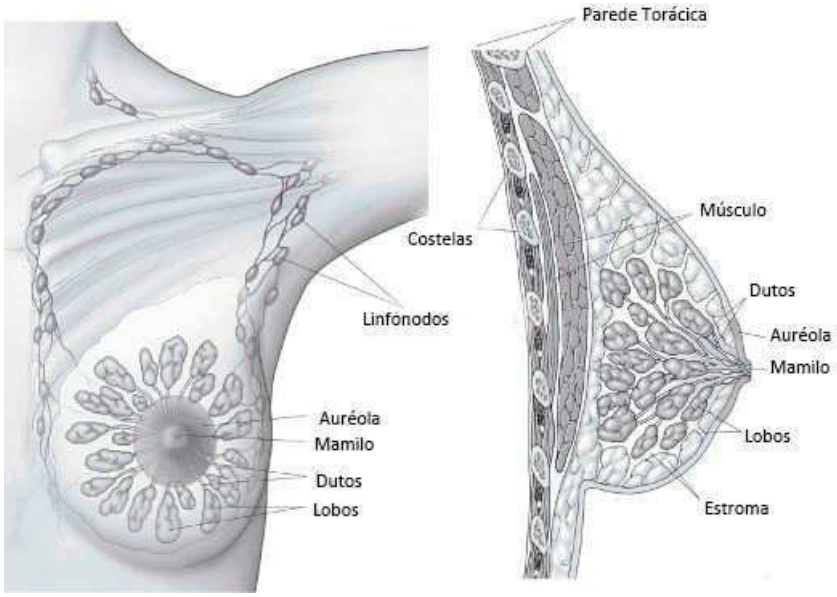


Figura 1 – Anatomia da mama feminina. Fonte: (ACS - American Cancer Society, 2016)

O câncer de mama origina-se quando as células da mama começam a crescer fora de controle, formando uma neoplasia ou tumor. Esta proliferação anormal de células é causada por mutações genéticas onde o código genético destas células é alterado, perdendo

assim sua função natural, e passando a se reproduzir de maneira exagerada. A produção demasiada de células exige que no local exista uma grande quantidade de nutrientes, desta forma novos vasos sanguíneos são criados em torno das células para alimenta-las. Este processo é denominado de angiogênese e ocorre em situações como cicatrizações, onde novos vasos sanguíneos são criados. Contudo, na neoplasia a angiogênese torna-se fator contribuinte para a manutenibilidade das células mutantes, permitindo assim a formação de tecidos e consequentemente tumores (MC - Medicina Celular, 2013).

As neoplasias são divididas em benignas e malignas, tendo como principal diferenciação a agressividade e velocidade de multiplicação. As neoplasias benignas crescem lentamente e são menos agressivas, ou seja, os tecidos próximos não sofrem grandes danos. Enquanto as neoplasias malignas aumentam de tamanho rapidamente e comprometem as células próximas na busca por nutrientes (MC - Medicina Celular, 2013).

Os casos de câncer de mama sofreram um aumento progressivo nos últimos anos. Segundo a estimativa nacional de incidência do câncer de mama em mulheres brasileiras, apresentada pelo Instituto Nacional do Câncer (INCA), em apenas cinco anos (2009 a 2014) o número de casos no país aumentou 13,4%, representando uma taxa de aumento de aproximadamente 2% ao ano (INCA - Instituto Nacional do Câncer, 2015). Para o ano de 2017, estima-se 57.960 novos casos (INCA - Instituto Nacional do Câncer, 2016b).

O câncer de mama pode afetar também os homens, contudo ainda é considerado como uma doença incomum, representando cerca de 1% de todos os cânceres de mama, o que corresponde a menos de 1% de todos os cânceres que ocorrem em homens, sendo responsável somente por menos de 0,1% das mortes. No ano de 2005, foram estimados 1.690 novos casos, com 460 casos fatais (NOGUEIRA; MENDONÇA; PASQUALETTE, 2014).

As causas do câncer de mama são objetos de estudos em âmbito mundial. Portanto, alguns determinados grupos de mulheres possuem maior propensão a doença. Estas mulheres possuem características em comum denominadas fatores de risco. Alguns fatores são considerados atuantes no desenvolvimento da doença, tais como:

- Idade: mulheres com idade acima de 50 anos devido as alterações biológicas do próprio envelhecimento;
- Fatores endócrinos ou história reprodutiva: estão relacionados principalmente aos estímulos endógenos[1] ou exógenos[2], tendo seu risco elevado de acordo com a exposição. Dentre estes fatores podemos incluir: história de menarca[3] precoce, menopausa tardia (após os 55 anos), primeira gravidez após os 30 anos, nuliparidade[4], uso

[1] Fatores que afetam o indivíduo de dentro para fora, por exemplo, os hormônios.
[2] Fatores externos que influenciam algo.
[3] Idade da primeira menstruação menor que 12 anos
[4] Ausência de gestação.

de contraceptivos orais (estrogênio-progesterona) e terapia de reposição hormonal pós-menopausa;

- Fatores comportamentais ou ambientais: incluem a ingestão de bebida alcoólica, sobrepeso e obesidade na pós-menopausa, exposição à radiação e tabagismo;

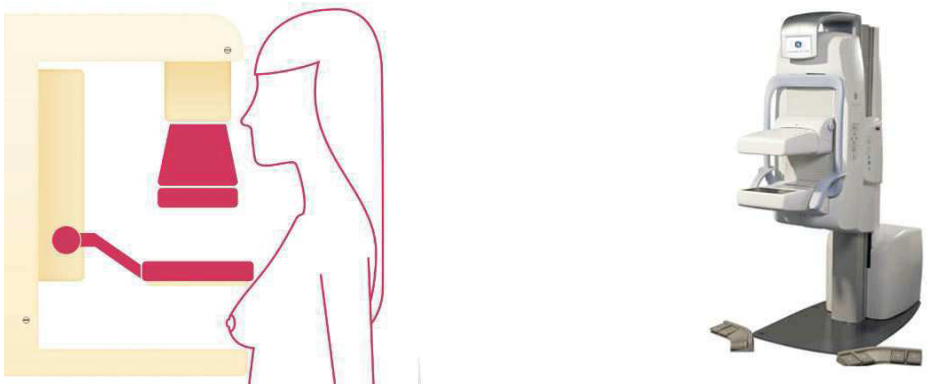
- Fatores genéticos: estes fatores relacionam-se à presença de mutações em determinados genes, de maneira especial nos genes BRCA1 e BRCA2 (ANTONIOU et al., 2003). O câncer de mama de caráter hereditário corresponde de 5% a 10% do total de casos (ADAMI, 2008). Mulheres que tem um histórico familiar de câncer de mama ou pelo menos um caso de câncer de ovário ou câncer de mama em homem em parente consanguíneo, podem ter predisposição genética e são consideradas de maior risco para a doença.

O diagnóstico precoce e o rastreamento são estratégias desenvolvidas para a detecção precoce do câncer de mama. Tais abordagens visam contribuir para a redução do estágio de apresentação do câncer, tornando o tratamento mais eficiente. No diagnóstico precoce destaca-se a importância da educação da mulher e profissionais de saúde no reconhecimento dos sinais e sintomas da doença. A autopalpação das mamas visa a descoberta casual de pequenas alterações mamárias de maneira casual sem nenhuma recomendação de técnica específica. Entretanto, orienta-se sempre a aplicação deste diagnóstico quando a mulher se sentir confortável para tal (seja no banho, no momento da troca de roupa ou em outra situação do cotidiano). Por outro lado, a autopalpação não substitui o exame físico realizado por profissional de saúde (enfermeiro ou médico) qualificado (INCA - Instituto Nacional do Câncer, 2014). O rastreamento consiste na oferta do exame de rastreio, normalmente a mamografia, a mulheres contidas no fator de risco da idade ou a mulheres que procuram as unidades de saúde. A mamografia é o exame que possui eficácia comprovada na redução do índice de mortalidade por câncer de mama.

A radiografia da mama ou mamografia permite a detecção precoce do câncer, evidenciando lesões ainda que em fases iniciais. Tornou-se o principal método de rastreamento e investigação, em mulheres, tanto nos aspectos relacionados ao câncer de mama, quanto para a maioria das alterações clínicas mamárias. Contudo, esta forma de diagnóstico ainda apresenta-se como um método com um elevado custo na iniciativa privada. Portanto, o Ministério da Saúde recomenda a realização periódica deste exame, apenas nos casos de exames clínicos suspeitos e em mulheres que apresentam situação de alto risco, com idade superior a 40 anos (Instituto Nacional do Câncer, 2002).

O rastreamento mamográfico mostra-se como um efetivo método para a redução do índice de mortalidade pelo câncer de mama em mulheres. Este método possibilita o diagnóstico precoce, contribuindo no desenvolvimento do tratamento e por conseguinte elevando a probabilidade de sucesso.

Ao realizar o exame mamográfico a paciente tem sua mama comprimida por um aparelho de raio X apropriado, denominado mamógrafo. Neste aparelho é incidida uma radiação ionizante (feixes de Raios-X), normalmente, em duas direções: vertical e angular. As Figuras (2a) e (2b) ilustram o procedimento para realização do exame e o aparelho mamográfico, respectivamente.



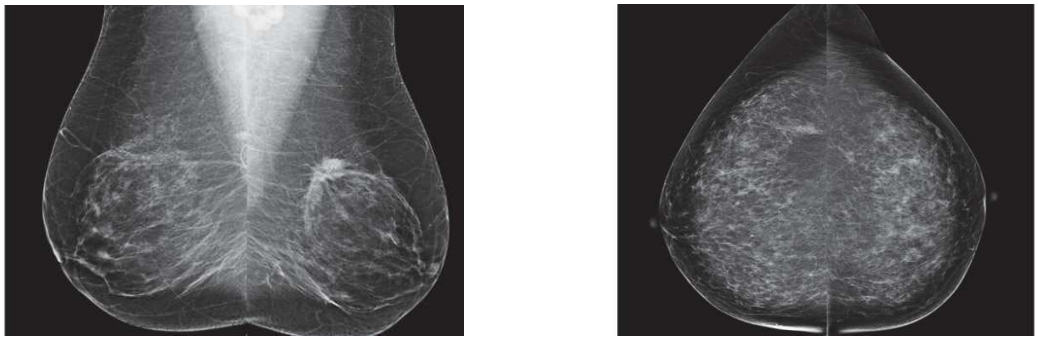
(a) (b)

Figura 2 – Mamógrafos. (a) Exemplo de visão angular (NBCF - Nacional Breast Cancer Foundation, 2016). (b) Mamógrafo real (GE - General Electric Company, 2016).

A compressão da mama realizada pelo mamógrafo busca fornecer imagens com maiores informações da estrutura interna, elevando a capacidade de diagnóstico. Este método também evita a superexposição de tecidos considerados mais finos, localizados anteriores a mama. Alguns cuidados devem ser tomados para o melhor resultado do exame, tais como o posicionamento correto durante a mamografia, que pode aumentar a probabilidade de identificar tumores de mama invasivos. Normalmente, o médico faz duas ou mais radiografias de cada mama, que é comprimida no aparelho para que fique com uma espessura mais uniforme. O processo facilita a descoberta de nódulos minúsculos.

O resultado do exame mamográfico consiste em uma imagem em tons de cinza do tecido mamário impresso em um filme radiográfico[5] (Figura 3), que é lida ou interpretada por um radiologista. A interpretação de uma imagem mamográfica é uma tarefa extremamente difícil e desafiadora, onde o especialista busca, através apenas da visão e experiência, estruturas ou objetos suspeitos que indiquem alguma patologia na mama. Alguns casos de câncer de mama produzem modificações difíceis de serem percebidas no exame. Portanto, o radiologista deve ter em mãos o exame anterior, para a comparação com o atual. Este processo auxilia o especialista na identificação de possíveis mudanças na estrutura da mama. A Figura 3 demonstra um exemplo de resultado do exame mamográfico nas indecências médio-lateral (visão vertical) (Figura 3a) e crânio-caudal (visão angular) (Figura 3b) de ambas as mamas.

[5] É composto por uma ou duas camadas de emulsão fotográficas (mistura de gelatina fotográfica com uma suspensão de cristais de haleto de prata) unidas a uma base.



(a) Médio-Lateral (b) Crânio-Caudal

Figura 3 – Resultado do exame mamográfo em diferentes visões (incidências). Fonte: (RSNA - Radiological Society of North America, 2013)

Os tipos mais comuns de anormalidades visíveis em imagens de mamografia são: calcificações, massas e distorção de arquitetura[6] (HEATH et al., 1998).

As massas, que são as neoplasias, são representadas nos exames como regiões densas, de tamanho e formato variável, sendo classificadas em circunscritas, espiculadas e mal definidas. Uma massa é definida como um aglomerado de células que formam densas regiões. Este aglomerado pode ser causado por câncer de mama, bem como por condições benignas. Características como, forma, tamanho e disposição das margens são determinantes para estabelecer a malignidade das massas. A Figura 4 mostra o exemplo de uma mamografia onde é possível identificar uma massa.

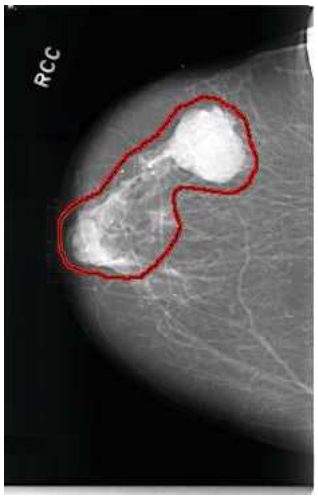
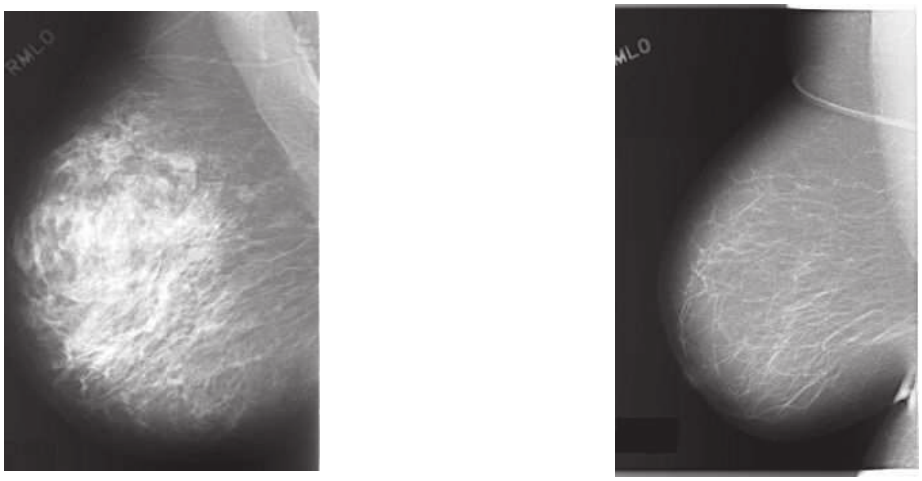


Figura 4 – Identificação de uma massa em uma mamografia. Fonte: (HEATH et al., 1998)

A densidade radiológica da mama é regulada de acordo com a quantidade de tecido glandular presente na mama. Normalmente, mulheres mais jovens apresentam elevadas taxas deste tecido, tornando as mamas mais densas e firmes. Desta forma, o exame mamográfico possui baixa sensibilidade em mamas densas. A Figura 5 apresenta exemplos

[6] Caracteriza-se pela presença de linhas finas ou espiculadas que se irradiam a partir de um ponto na mama.

de mama densa e não densa. Por esta razão, metodologias diferentes são propostas para o diagnóstico do câncer de mama de acordo com o nível de densidade da mama no intuito de facilitar o diagnóstico.



(a) (b)

Figura 5 – Exemplo de mama e suas densidades. (a) Mama densa, (b) Mama não densa Fonte: (HEATH et al., 1998)

A mamografia digital e a mamografia padrão utilizam o mesmo processo para aquisição das imagens, diferenciando-se apenas na forma de armazenamento e visualização. As imagens mamográficas tradicionais são gravadas em grandes folhas de filme fotográfico, enquanto as imagens digitais são eletronicamente armazenadas e podem ser visualizadas em um dispositivo de visualização apropriado.

As mamografias digitais permitem, de modo geral, operações como ampliação e brilho para ajudar na compreensão da imagem. Outra possibilidade viável é utilizar técnicas de processamento de imagens na variação do contraste entre os tecidos mostrados na imagem, auxiliando o médico na detecção de certas áreas mais claramente. O compartilhamento das imagens digitais é facilitado devido a rápida transmissão entre hospitais para proporcionar a análise de diversos especialistas em câncer de mama, buscando outros pontos de vista a respeito do diagnóstico.

Como desvantagem, os mamógrafos digitais e os sistemas embarcados possuem um elevado custo quando comparados com os mamógrafos tradicionais. Portanto o equipamento tradicional ainda possui maior predominância nos hospitais.

A mamografia digital e sistemas de auxilio computacional à detecção e diagnóstico de câncer de mama(CAD/CADx - *Computer-Aided Detection and Diagnosis*) tornam-se excelentes alternativas para suprir as deficiências do exame mamográfico convencional onde este, pode gerar uma grande quantidade de ruídos dificultando o diagnóstico.

2.2 Representações de Características Visuais

Informações visuais contidas em locais específicos nas imagens são representadas como características visuais que permitem através de sua análise o reconhecimento de objetos, segmentação e classificação de regiões da imagem, entre outras tarefas. A primeira etapa para o procedimento de análise das imagens consiste na extração de características visuais, onde utilizando operações a nível de *pixels* objetiva a extração de propriedades visuais de uma determinada região da imagem. De acordo com as regiões de interesse, características podem ser extraídas de forma global ou local.

Características locais são obtidas da análise de pequenas regiões das imagens, de maneira a procurar sub-padrões que se repetem e são importantes para o reconhecimento do objeto presente na imagem. A análise global busca a identificação de características globais significantes, sendo estas calculadas sobre toda a imagem. As abordagens globais e locais são baseadas em propriedades como cor, textura, forma e bordas/contornos.

A diferença básica entre características locais e globais refere-se ao nível de especialização da região a ser analisada, visto que a análise da imagem inteira ou apenas de regiões especificas representam, respectivamente, a extração de características globais e locais. Historicamente, um dos primeiros estudos voltado ao reconhecimento de padrões através da identificação de pontos de interesse locais, surgiu no âmbito militar onde imagens de satélite da *National Aeronautics and Space Administration* (NASA) foram utilizadas para o reconhecimento de objetos (KRIG, 2016). Desde então muitos trabalhos começaram a surgir para a resolução dessa problemática. Trabalhos como (MORAVEC, 1980) e (HARRIS; STEPHENS, 1988) foram os precursores na identificação de pontos que representassem características locais nas imagens.

Características locais consistem em padrões contidos na imagem que diferenciam-se de sua vizinhança imediata. Estas características são associadas a alterações de uma ou mais propriedades das imagens, normalmente textura, cor ou intensidade (TUYTELAARS; MIKOLAJCZYK, 2008). Tais características surgiram como alternativa para resolução de problemas recorrentes dos métodos globais como: mudanças de iluminação, oclusão, *viewport*, entre outros (AVILA et al., 2011). A identificação dessas características é realizada através da análise de vizinhança em regiões específicas da imagem. Esta análise consiste em operações entre *pixels*. Um ponto de interesse é definido como sendo a representação de uma característica local. Diversos termos são utilizados para esta representação como, ponto de interesse, pontos chaves, pontos característicos, entre outros. Neste trabalho utilizaremos o termo ponto de interesse.

Um conjunto de pontos de interesse pode ser usado como uma representação robusta da imagem permitindo o reconhecimento de objetos ou cenas sem a necessidade de segmentação.

Na Seção 2.2.1 introduzimos conceitos básicos de operadores de detecção (*detection operators*) de pontos de interesse e alguns dos principais descritores locais (*local descriptors*).

2.2.1 Descritores Locais

A detecção de características consiste na identificação de informações consideradas significativas na imagem, por exemplo: bordas, cantos, pontos de interesse, arestas, entre outros. A repetibilidade é a principal propriedade dos algoritmos de detecção de características locais, ou seja, dado duas diferentes imagens que contenham informações do mesmo objeto, uma elevada porcentagem das características comuns devem ser detectadas em ambas as imagens (TUYTELAARS; MIKOLAJCZYK, 2008). Diversos detectores de características locais são propostos e (MOREELS; PERONA, 2007).

Após a identificação dos pontos de interesse nas imagens, algumas medidas podem ser calculadas das regiões no entorno do ponto para a formação de um descritor local. Diversos descritores são propostos na literatura (TUYTELAARS; MIKOLAJCZYK, 2008). A seguir são apresentados os descritores *Scale-Invariant Feature Transform* (LOWE, 2004), *Speeded Up Robust Features* (BAY et al., 2008), *Oriented FAST and Rotated BRIEF* (RUBLEE et al., 2011) e *Local Binary Pattern* (OJALA; PIETIKÄINEN; HARWOOD, 1996) que são muito utilizados para o reconhecimento visual.

2.2.1.1 Scale-Invariant Feature Transform (SIFT)

Características locais presentes nas imagens demonstram em muitos trabalhos serem eficientes abordagens na comparação e reconhecimento de objetos ou cenas ((LOWE, 1999), (HARRIS; STEPHENS, 1988), (GOOL; MOONS; UNGUREANU, 1996)). Muitas abordagem foram propostas para extrair características locais invariantes a escala (MIKOLAJCZYK; SCHMID, 2004), sendo uma dessas o Scale-Invariant Feature Transform (SIFT) (LOWE, 2004).

O SIFT consiste de uma metodologia que transforma informações invariantes a escala presentes nas imagens em características locais. Para gerar um conjunto de características, são necessários quatro estágios: (1) Detecção do Espaço de Escala, (2) Localização dos Pontos de Interesse, (3) Atribuição da Orientação e (4) Descrição dos Pontos de Interesse. No primeiro estágio os pontos de interesse são detectados utilizando uma metodologia em cascata de filtragem que utiliza eficientes algoritmos para a identificação dos candidatos. A detecção das localizações invariantes a mudanças de escalas é realizada por uma busca de características estáveis, ou seja, que não possuem variação considerável em diferentes escalas. Utiliza-se uma função contínua conhecida como Espaço de Escala (WITKIN, 1984).

O espaço de escala de uma imagem é definido como uma função, $L(x, y, \sigma)$, sendo

esta produzida por uma operação de convolução de uma gaussiana, $G(x, y, \sigma)$, em uma imagem, $I(x, y)$, como ilustra a Equação 2.1

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{2.1}$$

onde $*$ é o operador de convolução nas coordenadas x,y e $G(x, y, \sigma)$ é definido de acordo com a Equação 2.2

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}}. \tag{2.2}$$

onde σ representa o nível de suavização da gaussiana G.

Visando uma eficiente detecção de pontos de interesse estáveis, o SIFT utiliza uma função de diferença de Gaussianas (DOG), $D(x, y, \sigma)$, que pode ser calculada a partir da diferença entre duas gaussianas em escalas diferentes, σ, sendo uma delas influenciada por um multiplicador k como ilustrado na Equação 2.4.

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y). \tag{2.3}$$

O fator k é responsável pelo nível de suavização da imagem. Substituindo a Equação 2.1 na Equação 2.3 obtém-se a diferença de Gaussianas definida pela Equação 2.4

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma). \tag{2.4}$$

Uma visão geral da diferença de Gaussianas (DOG) é ilustrada na Figura 6. Inicialmente a imagem, $I(x, y)$, é incrementalmente convoluída com o filtro da Gaussiana,

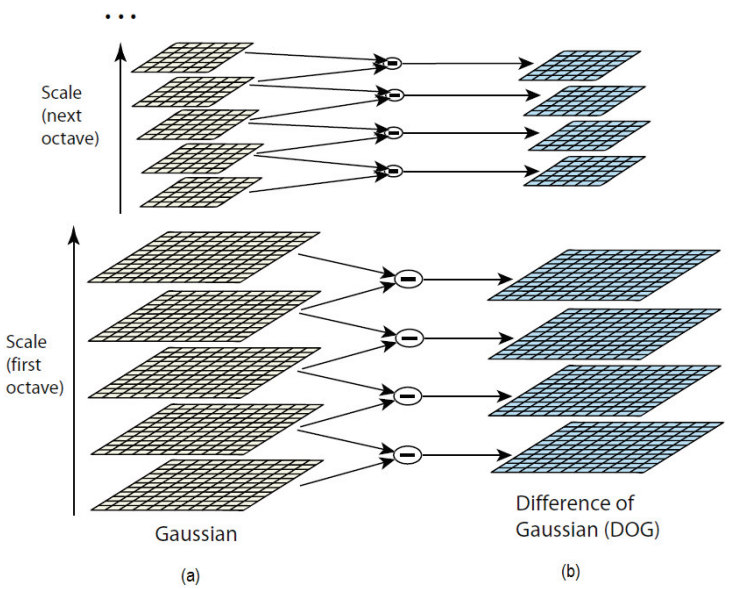


Figura 6 – Processo da Diferença de Gaussianas em uma imagem.(a) Representa o conjunto de imagens geradas após o processo de convolução da Gaussiana em diferentes resoluções.(b) Resultado da subtração de imagens com resoluções próximas (LOWE, 2004).

$G(x, y, \sigma)$, produzindo um conjunto de novas imagens em diferentes níveis de suavização chamado de oitavas ou iterações. Um conjunto de quatro imagens gaussianas são agrupadas em cada oitava, a Figura 6(a) ilustra esse processo. Em seguida imagens com níveis de suavização próximos, ou seja que possuem diferentes σ contidas na oitava são subtraídas produzindo uma imagem resultante denominada diferença de Gaussianas (DOG), como apresentado na Figura 6(b). Ao final do processo um novo valor é atribuído ao fator k e multiplicado a variável σ produzindo um novo conjunto de oitava em um nível de suavização diferente, a DOG é novamente calculada e o processo é repetido formando uma pirâmide Gaussiana. A quantidade oitavas a cada iteração foi empiricamente determinada como 3 (LOWE, 2004).

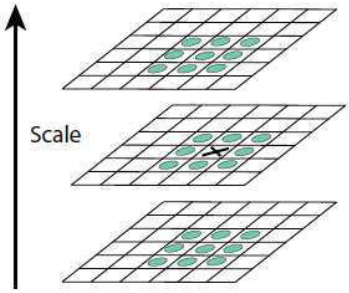


Figura 7 – Processo de descoberta do mínimo e máximo local da DOG. A descoberta é realizada com a comparação de um *pixel* (representado por X na imagem) com outros 26 vizinhos em regiões 3x3 da imagem atual com as imagem em níveis de suavização adjacentes (LOWE, 2004).

Uma vez que as DOG são encontradas no primeiro estágio, a descoberta do mínimo e máximo local das DOG é necessária pois objetiva encontrar e localizar candidatos a pontos representativos nas imagens em cada nível de suavização, etapa esta presente no segundo estágio. O processo de descoberta ocorre através da comparação de intensidade de *pixel* utilizando uma janela de tamanho 3x3. O elemento central nesta janela têm sua intensidade comparada com os 26 vizinhos próximos, sendo oito contidos na mesma DOG e nove localizados na mesma posição tanto na DOG de nível superior quanto inferior. Como apresentado na Figura 7 (LOWE, 2004). Com a descoberta dos pontos candidatos, em seguida é necessário o ajuste das informações de localização, escala e posicionamento. Essas informações permitem filtrar pontos em baixos contrastes, que possuem alta sensibilidade a ruídos, ou mal localizado em uma borda. Utiliza-se a expansão de Taylor (BROWN; LOWE, 2002) para, no espaço escala da função, $D(x, y, \sigma)$, determinar a interpolação da localização do máximo, ou seja, diminuir a quantidade de pontos candidatos que representam a mesma característica. A análise é realizada para cada ponto candidato:

$$D(x) = D + \frac{\partial D^T}{\partial x} x + \frac{1}{2} x^T \frac{\partial^2 D}{\partial x^2} x, \tag{2.5}$$

onde D e suas derivadas são avaliados em cada amostra de ponto e $x = (x, y, \sigma)^T$

corresponde ao deslocamento a partir desse ponto. A localização do máximo é determinada tomando a derivada dessa função em relação a x e definindo-a como zero, segundo (BROWN; LOWE, 2002).

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x} \tag{2.6}$$

O valor da função, $D(x)$, é útil para rejeitar os extremos instáveis, com baixo contraste (LOWE, 2004). Isto pode ser obtido substituindo a Equação 2.6 na Equação 2.5, como demonstrado na Equação 2.7.

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\hat{x} \tag{2.7}$$

Todos os extremos com o valor de $|D(x)|$ menor que $0,03$ são descartados (considerando que os valores de *pixels* da imagem estão normalizado entre 0 e 1) (LOWE, 2004).

Este método de rejeição de pontos de interesse baseado no baixo contraste não é suficiente, pois a função de diferença de gaussianas é sensível a ruídos presentes nas bordas em caso de bordas mal definidas. Para a resolução dessa deficiência utiliza-se uma matriz hessiana (HARRIS; STEPHENS, 1988).

O terceiro estágio corresponde a atribuição de orientação nos pontos anteriormente identificados e filtrados. A orientação baseia-se nas propriedades locais da imagem, onde os descritores dos pontos de interesse *(keypoint descritor)* podem ser representados em relação a orientação, tornando-os invariante a rotação da imagem. Este processo ocorre da seguinte maneira, para cada amostra de imagem, $L(x, y)$, em uma escala, a magnitude do gradiente, $m(x, y)$, e a orientação, $\theta(x, y)$ são pré-calculados utilizando a diferença entre *pixels* como demonstrado nas Equações 2.9 e 2.8.

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2} \tag{2.8}$$

$$\theta(x,y) = tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))). \tag{2.9}$$

O quinto e último estágio corresponde ao descritor de pontos de interesse. Para cada nível de suavização gaussiana na imagem, o descritor divide quatro regiões de 4x4 em torno do ponto de interesse com o objetivo de analisar as orientações das regiões e tornar o descritor invariante a rotação em relação ao ponto. Uma ponderação é utilizada na função Gaussiana com σ igual a metade da largura da janela do descritor, neste caso 2,

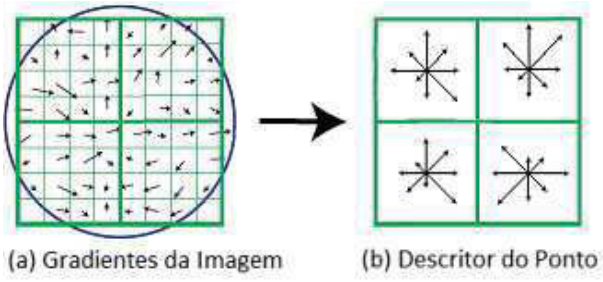


Figura 8 – Processo para a formação do descritor de pontos de interesse. (a) Cálculo do Histograma de Orientações (HoG), (b) Descritor de pontos de interesse (*keypoint descritor*) (LOWE, 2004).

é usada para atribuir um peso à magnitude de cada ponto. Esta ponderação, representada pelo círculo, e a divisão das regiões, são apresentadas na Figura 8(a). O descritor de pontos corresponde a um histograma de orientações sendo este, formado a partir dos gradientes das orientações dos pontos contidos nas regiões adjacentes ao ponto de interesse, ou seja, o descritor é obtido por um vetor que contêm todos os valores de orientação das regiões próximas como ilustrado na Figura 8 b. A figura mostra um vetor de histograma de orientações **2x2** onde cada elemento possui **8** orientações, totalizando, $2 * 2 * 8$, um vetor com **32** características para cada ponto de interesse. Em Lowe (2004), a melhor configuração aplicada no contexto de imagens naturais foi utilizando amostras de 16x16 e um descritor de 4x4 totalizando, $4 * 4 * 8$, **128** características.

2.2.1.2 Speeded Up Robust Features (SURF)

O *Speeded Up Robust Features* (SURF) foi originalmente proposto em (BAY et al., 2008) como um detector e descritor de características locais em imagens digitais para o reconhecimento de objetos.

O detector SURF é baseado em matrizes Hessianas ((LINDEBERG, 1998), (MIKOLAJCZYK; SCHMID, 2001)) é denominado *Fast-Hessian Detector* devido a seu bom desempenho computacional em relação ao tempo de execução e acurácia. O conceito de imagens integrais é utilizado pelo detector para diminuir o custo computacional no processo de identificação dos pontos de interesse. Uma imagem integral, $I_\Sigma(x)$, de localização $x = (x, y)^T$ representa a soma de todos os *pixels* de uma imagem de entrada I com uma região retangular formada pela localização x sobre a imagem original como ilustrado na Equação 2.10.

$$I_\Sigma(x) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{i \leq y} I(i, j) \tag{2.10}$$

As medidas de localização e escala dos pontos de interesse são calculadas através do determinante da matriz Hessiana para ambas as medidas. Considere um ponto $p = (x, y)$ em uma imagem I, a matriz Hessiana $H(p, \sigma)$ em p na escala σ é definida de acordo com

a Equação 2.11

$$H(p,\sigma) = \begin{bmatrix} L_{xx}(p,\sigma) & L_{xy}(p,\sigma) \\ L_{xy}(p,\sigma) & L_{yy}(p,\sigma) \end{bmatrix} \tag{2.11}$$

onde $L_{xx}(p,\sigma)$ corresponde a convolução da gaussiana de segunda ordem, $\frac{\partial^2}{\partial x^2}g(\sigma)$, com a imagem I em um ponto p de coordenadas x, y, as mesmas definições aplicam-se para $L_{xy}(p,\sigma)$ e $L_{yy}(p,\sigma)$.

Ao contrário do SIFT (LOWE, 2004), que utiliza aproximações com DOG, o detector SURF realiza aproximações com um elemento denominado *box filter* e imagens integrais como apresentado na Figura 9.

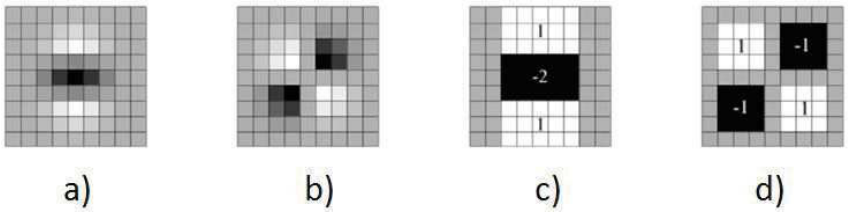


Figura 9 – Processo de simplificação das representações SURF. (a) e (b) Correspondem a aplicação da derivada parcial de segunda ordem da gassiana nas direções (xx e xy). Já (c) e (d) correspondem a aproximação usando os conceitos de imagem integral e *box filter*. As áreas em cinza são iguais a 0. Fonte: (BAY et al., 2008).

O *box filter* pode ser aplicado diretamente na imagem original em vários tamanhos de janela, sendo que ao aplica-lo o espaço de escala é escalado de acordo com o tamanho da janela ao contrário do SIFT que reduz iterativamente o tamanho da imagem, a Figura 10 demonstra este processo.

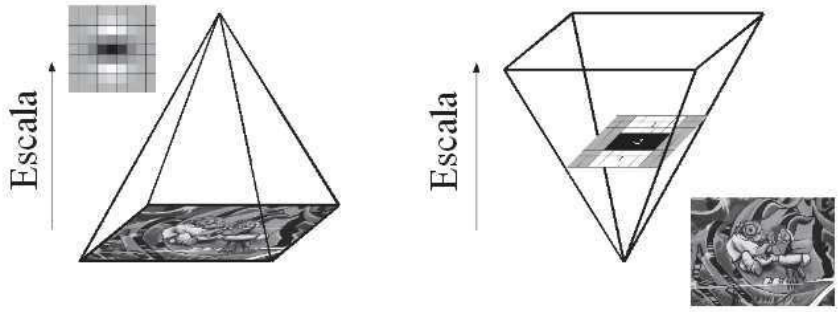


Figura 10 – Em vez de reduzir iterativamente o tamanho da imagem (esquerda), o uso de imagens integrais permite a escalabilidade do filtro a custo constante (à direita). Fonte: (BAY et al., 2008).

O espaço de escala é dividido em oitavas sendo que uma oitava representa um conjunto de filtros obtidos do processo de convolução da imagem de entrada com um filtro de tamanho incrementado. Uma oitava abrange um fator de escala igual a 2 (BAY et al., 2008). Cada oitava é subdividida em um número constante de níveis de escala, onde a diferença mínima de escala entre duas escalas próximas é dada por uma distância l_0,

podendo esta ser positiva ou negativa em relação a direção da derivação. Considerando um filtro de tamanho 9x9, a distância l_0 é igual a 3. Para a construção do espaço de escala inicia-se aplicando um filtro de tamanho 9x9, em seguida incrementa-se 6 *pixels* para cada eixo aplicando-o novamente. Este processo garante que filtros de diversos tamanhos (9x9, 15x15, 27x27 dentre outros) sejam aplicados para obter diferentes informações em várias escalas.

Os pontos de interesse são localizados através da aplicação do cálculo do determinante máximo das matrizes Hessianas, sendo que estas são interpoladas em escala e espaço de imagem com o método proposto em (LOWE, 1999). A interpolação do espaço de escala é tarefa importante devido as primeiras camadas de cada oitava possuírem uma grande diferença em relação a escala.

O descritor SURF aplica uma distribuição de intensidade na vizinhança do ponto de interesse detectado de forma semelhante ao descritor SIFT (LOWE, 2004). O descritor possui três etapas principais: atribuição de orientação, extração e indexação rápida para correspondência. A primeira etapa consiste na fixação de uma região circular no entorno do ponto de interesse para posterior análise das informações nela contida. O principal objetivo desta etapa é tornar os pontos de interesse invariantes a rotação de forma que as orientações calculadas sejam reproduzíveis. Tendo em vista este objetivo, primeiramente são calculadas as respostas de Wavelet de Haar das direções x e y dentro da região circular formada no entorno do ponto de interesse em uma determinada escala s. Tanto a fase de amostragem quanto o tamanho das *wavelets* são dependentes da escala. Em seguida as imagens integrais são novamente utilizadas para uma rápida filtragem. A direção predominante é estimada pelo cálculo do somatório de todas as respostas, *wavelets*, dentro de uma região de tamanho $\frac{\pi}{3}$. As respostas verticais e horizontais de *wavelets* são somadas para a formação de um vetor de orientação local sendo que o vetor mais longo sobre todas as janelas define a orientação do ponto de interesse.

Na etapa de extração de características para a formação do descritor, inicialmente é criando uma região quadrada centralizada no entorno do ponto de interesse com orientação já atribuída. Esta região ou janela possui tamanho igual a vinte vezes a escala, s, naquele ponto e são divididas em sub-regiões de tamanho 4x4, preservando assim importantes informação espaciais. Para cada sub-região são aplicadas as Wavelet de Haar nas direções horizontais, d_x, e verticais, d_y, sendo estas calculadas em relação ao ponto de interesse. O vetor de características gerado pelo descritor consiste no somatório entre as direções e seus valores absolutos, onde $v = (\sum dx, \sum dy, \sum |dx|, \sum |dy|)$. Concatenando os somatórios, como resultado, o vetor descritor final possui tamanho 64. A Figura 11 exemplifica este processo.

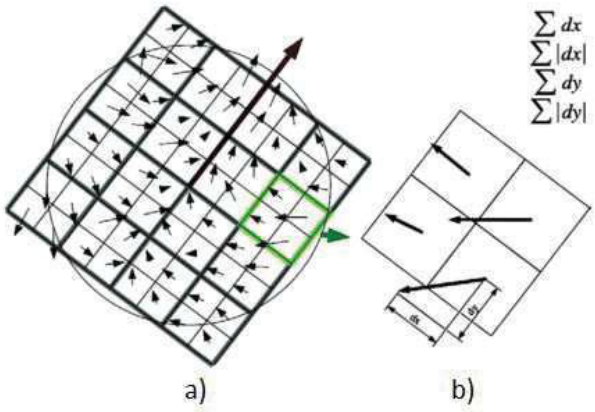


Figura 11 – Formação do descritor SURF. (a) Subdivisão da área em torno do ponto de interesse em sua orientação. (b) Aplicação das Wavelet de Haar nas sub-regiões. Fonte: (BAY et al., 2008).

2.2.1.3 Oriented FAST and Rotated BRIEF (ORB)

Os primeiros trabalhos relacionados a encontrar pontos de interesse na imagem, consideravam os cantos formados por elementos na imagem. O detector de cantos proposto em (HARRIS; STEPHENS, 1988) é um dos precursores dessa abordagem. O *Oriented FAST and Rotated BRIEF* (ORB) (RUBLEE et al., 2011) é composto por uma combinação do *Oriented Features From Accelerated Segment Test*(oFAST)(ROSTEN; DRUMMOND, 2006) e *Rotate Binary Robust Independent Elementary Features* (rBRIEF) (CALONDER et al., 2010).

O FAST e suas variantes (ROSTEN; DRUMMOND, 2006), incluindo o oFAST, são métodos de identificação de pontos de interesse em imagens ou vídeos. Tais pontos normalmente são encontrados nos "cantos" da imagem, ou seja, nos *pixels* que possuem um alto gradiente em relação aos seus *pixels* vizinhos. São muito utilizados em sistemas de tempo real como, *Parallel Tracking* e *Mapping* (KLEIN; MURRAY, 2007).

Para a identificação dos pontos, o FAST realiza o processo em seis etapas: (1) Seleciona um *pixel* p, considerando I_p como a intensidade em níveis de cinza; (2) Determina um limiar (*threshold*) de intensidade T, normalmente utiliza-se 20% do valor do *pixel* (ROSTEN; DRUMMOND, 2006); (3) Considera um círculo de 16 *pixels* ao redor de p (Círculo Bresenham (BRESENHAM, 1977) com um raio igual a 3 como apresentado na Figura 12. Os N *pixels* vizinhos devem ter seus valores de intensidade acima ou abaixo do valor do *pixel* central p, para que p seja considerado um ponto de interesse. (4) Compara a intensidade dos *pixels* I_1, I_5, I_9, I_{13} com a soma resultante entre I_p e T, se pelo menos três dos quatros valores *pixels* não estiverem acima ou abaixo, então o ponto não é considerado interessante ou ponto de "canto", logo é desconsiderado. Caso haja pelo menos três *pixels* com valores acima ou abaixo da soma, então realiza o mesmo processo de verificação para todos os outros 12 *pixels* vizinhos restantes. (6) Repete o processo para todos os *pixels* da

imagem(VISWANATHAN, 2009).

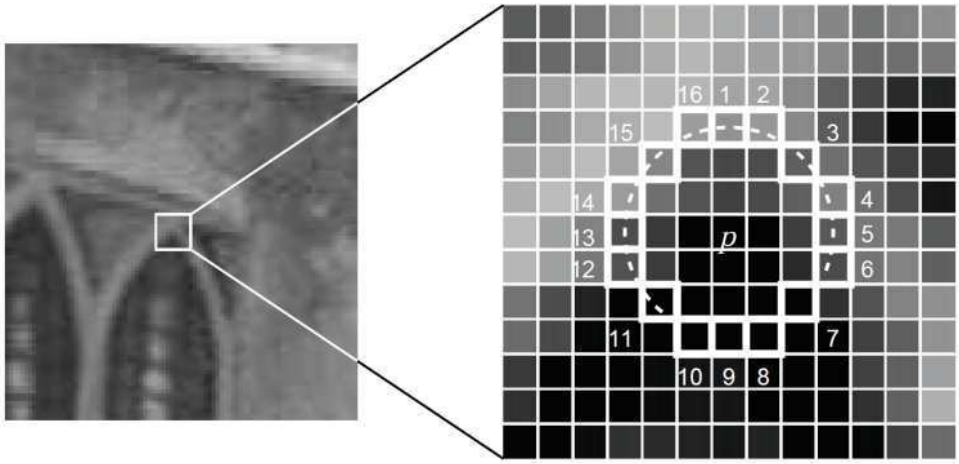


Figura 12 – Demonstra o processo de verificação de um ponto de interesse em ma imagem. Fonte (ROSTEN; DRUMMOND, 2006).

O FAST é amplamente utilizado devido a sua rápida detecção dos pontos pertencentes aos "cantos" (*corner*) nas imagens. Contudo, ele não possui uma componente de orientação desses cantos. Para suprir esta necessidade foi criado o *Oriented FAST* que utiliza uma medida simples e eficaz de orientação, chamada de intensidade do centroide (ROSIN, 1999). Os momentos padrões (*standard moments*) podem ser utilizados para determinar a orientação de cantos. Estes momentos são definidos de acordo com a Equação 2.12

$$m_{pq} = \sum_{x,y} x^p y^q I\left(x, y\right) \tag{2.12}$$

onde $I(x, y)$ representa a intensidade no ponto x, y. O centroide é então calculado a partir dos momentos padrões como apresentado na Equação 2.13.

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{2.13}$$

Após determinar o centroide é formado um vetor $\overrightarrow{OC}$, onde O é o centro de um canto e C é o centroide. A orientação é obtida através do cálculo do ângulo θ definido na Equação 2.14

$$\theta = atan2\left(m_{01}, m_{10}\right) \tag{2.14}$$

onde $atan2$ corresponde a uma versão da função arco tangente aplicada com os momentos padrões.

BRIEF é um descritor proposto em (CALONDER et al., 2010) que utiliza testes binários simples entre pixels em um conjunto de imagens suavizadas. Seu desempenho é semelhante ao SIFT (LOWE, 1999), em muitos aspectos, incluindo robustez à iluminação e distorção de perspectiva (RUBLEE et al., 2011). No entanto, é muito sensível a rotação no plano.

O descritor BRIEF é um vetor binário de descritores obtidos a partir de um conjunto de testes de intensidade como apresentados na Equação 2.16. Primeiramente são obtidos os pontos de interesse usando o FAST e então para cada teste binário (Equação 2.15) é realizado um processo de várias subamostragens considerando janelas de dimensões S x S que contenham um ponto de interesse. Em seguida, suavizações são aplicadas em todas as janelas, através de uma distribuição Gaussiana. A seguir são realizadas comparações de intensidades entre os pontos da subamostragem e o ponto de interesse, caso a intensidade do ponto de interesse seja menor que o elemento na subamostragem então é retornado o valor 0, caso contrário retorna-se 1. Desta forma, são construindo um vetor binário de características.

Um teste binário τ é definido na Equação 2.15

$$\tau(p, x, y) := \begin{cases} 0, & : p(x) < p(y) \\ 1, & : p(x) \geq p(y) \end{cases} \tag{2.15}$$

onde $p(x)$ é a intensidade do ponto de interesse x em uma imagem suavizada do conjunto p e $p(y)$ é a intensidade dos demais pontos. As características são definidas como um vetor de n testes binários de acordo com a Equação 2.16

$$f_n(p) := \sum_{1 \leq i \leq n} 2^{i-1} \tau(p; x_i, y_i) \tag{2.16}$$

onde n define o número de dimensões do vetor binário.

O BRIEF é extremamente sensível a mudanças na rotação do plano, mesmo em poucos graus seu desempenho é prejudicado. Em (RUBLEE et al., 2011) é proposto um método que utiliza a orientação dos pontos de interesse calculada pelo FAST para determinar a orientação das subamostragens. Para qualquer conjunto de n testes binários de localizações (x_i, y_i), definido pela matriz $2 \times n$

$$S = \begin{pmatrix} x_1, ..., x_n \\ y_1, ..., y_n \end{pmatrix}. \tag{2.17}$$

Utilizando a orientação θ da subamostragem e a correspondente matriz de rotação R_θ, é construída a versão orientada S_θ de S:

$$S_\theta = R_\theta S. \tag{2.18}$$

Desta forma tem-se a versão modificada do operador BRIEF, denominada rBRIEF - *Rotation-Aware* BRIEF, definida pela Equação 2.19

$$g_n(p, \theta) := f_n(p) | (x_i, y_i) \in S_\theta. \tag{2.19}$$

O rBRIEF tem melhorias significativas na variância e na correlação em relação ao BRIEF padrão (RUBLEE et al., 2011).

2.2.1.4 Local Binary Pattern (LBP)

O *Local Binary Pattern* (LBP) consiste em um descritor visual amplamente utilizado para a classificação na área de visão computacional. Originalmente em (HE; WANG, 1990) é introduzido um modelo de análise de imagem baseado em unidade de textura, onde uma textura é considerada como conjunto de unidades essenciais. O LBP foi proposto em Ojala, Pietikäinen e Harwood (1996) e surge como uma especificação do modelo de textura espectral sendo definido como um operador não parametrizado capaz de descrever a estrutura espacial local de uma imagem, mostrando ser um poderoso método para a diferenciação de características de textura.

A Equação 2.20 demonstra o cálculo do LBP,

$$LBP(x_c, y_c) = \sum_{n=0}^{n-1} S(i_n - i_c)2^n \tag{2.20}$$

onde, n corresponde a quantidade de *pixels* vizinhos do *pixel* central (x_c, y_c), i_c é o valor de nível de cinza do *pixel* central, i_n é o valor de nível de cinza de cada *pixel* vizinho e $S(x)$ e uma função que resulta 1 se $x \geq 0$ e 0, caso contrário.

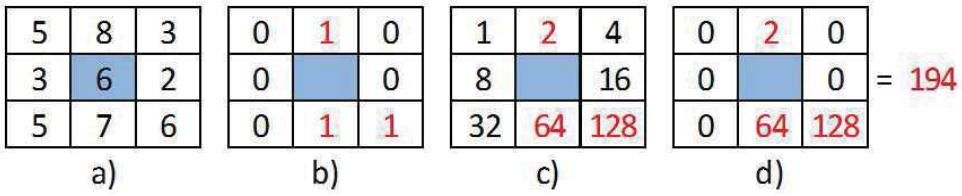


Figura 13 – Cálculo do LBP para uma unidade de textura. Adaptado de Ojala, Pietikäinen e Harwood (1996).

A Figura 13 apresenta um exemplo do cálculo LBP. Considerando uma janela de tamanho 3x3 (Figura 13(a)) os valores dos níveis de cinza dos *pixels* vizinhos são subtraídos dos valores do nível de cinza do *pixel* central, resultando em uma matriz binária composta por 0 ou 1, dependendo da diferença dos *pixels* analisados (Figura 13(b)). Em seguida, os valores dos *pixels* de vizinhança da matriz binária são multiplicados pelos pesos dados aos correspondentes *pixels* da matriz de pesos (Figura 13(c)). Finalmente, os valores dos *pixels* são somados para obter o valor da unidade de textura que neste exemplo é 194 (Figura 13(d)).

2.3 Modelo Bag of Words

Diversos métodos de indexação e recuperação de informações em documentos foram desenvolvidos (FALOUTSOS; OARD, 1998). Características como robustez e eficiência são indispensáveis para tais métodos devido ao volume de documentos a ser analisado em tempo real. Tais documentos, normalmente possuem alguma distribuição ou padrões de palavras,

e assim podem ser resumidos de forma compacta por uma contagem de suas palavras. O modelo de representação de dados conhecido como *Bag of Words* (BoW) (HARRIS, 1954) realiza essa tarefa de contagem e agrupamento. Este modelo é utilizado principalmente nas áreas de Visão Computacional, Processamento de Imagens e Representação Textual sendo que para cada área surge variações deste modelo, *Bag of Visual Word* (Visão Computacional), *Bag of Words* (Representação Textual) e *Bag of Features* (Processamento de Imagens). Nesta seção, abordaremos o modelo de representação *Bag of Features* de forma mais aprofundada, apresentando seus formalismos matemáticos.

2.3.1 Bag of Features

O modelo *Bag of Feature* (BoF) ou *Bag-of-Visual-Words* (BoVW) é inspirado no modelo tradicional BoW, onde uma imagem é representada como um histograma de ocorrências de características ou "palavras visuais". A construção dessas características no BoF é definida em etapas sequenciais de codificação e *pooling* (AVILA, 2013). A etapa de codificação objetiva quantificar características locais obtidas da imagem em relação ao vocabulário visual, associando descritores locais, SIFT *descriptors*, extraídos da imagem (características) como elementos mais representativos da imagem. O vocabulário visual normalmente é construído a partir da aplicação de um algoritmo de clusterização, geralmente K-Means (HARTIGAN; WONG, 1979), em um conjunto de amostras dos descritores locais extraídos, onde cada característica resultante corresponde ao centroide obtido pela clusterização. O tamanho do vocabulário está diretamente vinculado a quantidade de *clusters* a serem gerados. Em (YANG et al., 2007) são propostos vários experimentos onde conclui-se que um vocabulário que possui tamanho entre 100 e 500 é considerado bem discriminativo. A etapa de *pooling* resume as características obtidas em um único vetor de características com o objetivo de representar toda a imagem.

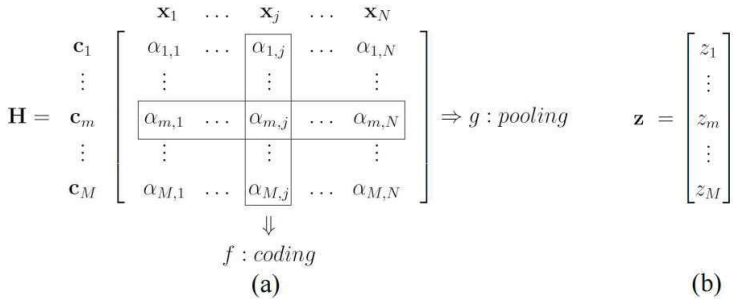


Figura 14 – (a) Representação das funções de codificação e *pooling* na matriz **H**.(b) vetor de representação final **z**. Fonte (AVILA, 2013).

Em (AVILA, 2013) é formalizado matematicamente o modelo BoF. Seja X um conjunto desordenado de descritores locais extraídos de uma imagem, $X = \{x_j\}, j \in \{1, ..., N\}$, onde $x_j \in R^D$ é o vetor de descritores locais e N é o número de descritores locais na imagem. Seja C um vocabulário visual obtido por um algoritmo de aprendizado

não supervisionado (por exemplo, o algoritmo de agrupamento k-means). Considere $C = \{c_m\}, m \in \{1, ..., M\}$ onde $c_m \in R^D$ é uma palavra visual e M corresponde a quantidade de palavras visuais. $z \in R^M$ é um vetor que representa um histograma do vocabulário formado. Este vetor final é utilizado para a classificação de novas palavras.

A Figura 14 apresenta a matriz H, que representa as funções de codificação e *pooling* do modelo, onde as colunas X representam o conjunto de características locais, previamente extraídas da imagem, e as linhas C o conjunto de palavras. Cada $\alpha_{m,j}$ quantifica uma similaridade entre x_j e a palavra visual c_m. Essa medida de similaridade é baseada na distância, normalmente de Hamming (HAMMING, 1950). A Figura 15, ilustra o processo de modo geral.

A etapa de codificação pode ser modelada pela função de ativação f :

$$\begin{aligned} &f : R^D \rightarrow R^M, \\ &x_j \rightarrow f(x_j) = \alpha_j = \{\alpha_{m,j}\}, \quad m \in \{1, ..., M\} \end{aligned} \tag{2.21}$$

onde, $\alpha_{m,j}$ é componente m do vetor codificado α_j. A etapa de *pooling* é realizada após a codificação e pode ser representada pela função g:

$$\begin{aligned} &g : R^N \rightarrow R, \\ &\alpha_j = \{\alpha_{m,j}\}, j \in \{1, ..., N\} \rightarrow g(\{\alpha_j\}) = z. \end{aligned} \tag{2.22}$$

O vetor z, a representação final da imagem, é construído através da sequência de codificações e *pooling*. O objetivo do BoW é descobrir quais operadores f e g oferecem o melhor desempenho da classificação, gerando melhores vetores de características. A Figura 15, ilustra todo o processo (AVILA, 2013).

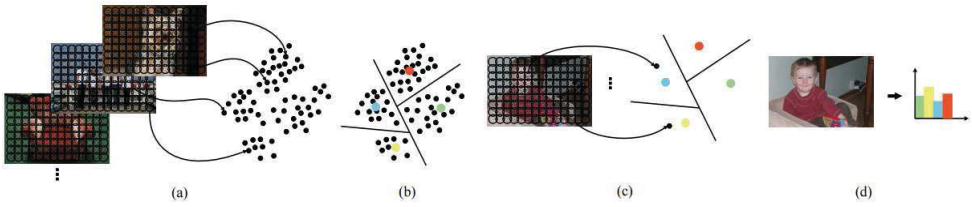


Figura 15 – Processo de elaboração do vocabulário visual.(a) Etapa de busca e identificação das palavras visuais.(b) Escolha dos centroides e clusterização. (c) Criação do vocabulário visual a partir dos centroides. (d) Contagem de ocorrências de palavras visuais/características encontradas na imagem através de um histograma. Fonte: (AVILA, 2013).

2.3.1.1 Estimação de Tamanho do Vocabulário

A construção de um vocabulário no âmbito textual possui tamanho relativamente fixo, delimitado pela quantidade de palavras. No contexto relacionado a imagens/vídeos a dimensão do vocabulário não possui esta mesma delimitação. Um conjunto de palavras

visuais são extraídas e formam um vocabulário visual, cujo tamanho é controlado pelo número de *clusters* de pontos de interesse no processo de *clustering*.

A generalização e o poder de distinção são características fundamentais para determinar o tamanho correto de um vocabulário. Por exemplo, em um vocabulário pequeno as características das palavras visuais não possuem um grande poder de distinção, pois pontos de interesse dissimilares podem ser mapeados para a mesma palavra visual.

A medida que o tamanho do vocabulário é incrementado, as características se tornam mais discriminantes, porém menos generalizáveis e tolerantes a ruídos. Outro fator que sofre grande influência em relação ao tamanho do vocabulário é o custo computacional, onde ambos são proporcionais, ou seja, quanto maior o vocabulário maior o esforço computacional para processá-lo.

O vocabulário de palavras visuais, normalmente, necessita de alguma técnica de agrupamento para a construção de um vocabulário. O K-Means é o *clusterizador* mais utilizado para a formação do vocabulário no modelo BoF, devido sua baixa complexidade. Sua principal desvantagem ocorre devido a necessidade prévia de estabelecer a quantidade de agrupamentos (*clusters*) a serem formados. Esta desvantagem é também compartilhada pelo vocabulário, visto que utiliza o *clusterizador*.

Portanto, a estimação do número de *clusters* é necessária para a melhor adaptação do modelo BoF de acordo com cada aplicação. Diversos trabalhos ((ZHOU; ZHOU; HU, 2013), (SUN et al., 2012) (YANG et al., 2007), (YANG; NEWSAM, 2010), (XU et al., 2010), (PENG et al., 2016)), realizam esta estimação mediante comparação entre diversos valores de números de *clusters* em um determinado intervalo fechado. Normalmente estes trabalhos ilustram graficamente os resultados obtidos para cada quantidade de *clusters* aplicada. Contudo, alguns trabalhos estimam automaticamente esta quantidade, como em Naldi et al. (2011) onde foi desenvolvido o algoritmo *Ordered Multiple Runs of K-Means* (OMRK) que consiste em uma abordagem exaustiva, que executa o K-Means varias vezes em busca do agrupamento que minimize um índice denominado de (*Simplified Silhouette* (Vendramin, Campello e Hruschka (2010)) relativo a validação dos grupos. Esta variante do K-Means possui como entrada o maior valor possível de *clusters* diferentemente da versão padrão. Outra metodologia que busca esta estimação automática é apresentada em (KRISHNA; MURTY, 1999) onde através da aplicação dos conceitos presentes nos algoritmos genéticos (GA) busca encontrar a quantidade ótima de *clusters* a serem criados em um determinado conjunto de dados.

2.4 Reconhecimento de Padrões

Um padrão é definido por Looney (1997) como sendo uma entidade nomeável e representativa obtida através do conhecimento cultural humano. Outra definição pode ser

encontrada em (GONZALEZ; WOODS, 2010) onde um padrão consiste em um arranjo de vetores de características (descritores), sendo que um conjunto de padrões que possuem propriedades semelhantes formam uma classe de padrões.

O reconhecimento de padrões pode ser estruturado em dois processos segundo Looney (1997): classificação, onde classes são definidas como amostras de uma população particionada em grupos; e reconhecimento, amostras desconhecidas que possuem propriedades semelhantes, são identificadas como pertencentes a uma das classes criadas. As técnicas de classificação podem ser divididas em três principais abordagens (HAYKIN, 2001) (PEDRINI; SCHWARTZ, 2008):

- Supervisionada: onde são fornecidas classes previamente definidas no qual um padrão comportamental será identificado. Uma etapa de treinamento, responsável pela identificação das características de cada classe, antecede a aplicação dos algoritmos de classificação.
- Não supervisionada: são fornecidas classes que não possuem características previamente conhecidas, logo, todas as informações a respeito das classes serão obtidas a partir das próprias amostras. Contudo, assim como na classificação supervisionada, as amostras que compartilharem propriedades semelhantes devem receber o mesmo rótulo.
- Por reforço: onde há uma interação com o ambiente, recebendo gratificações em caso de acerto, ou penalidades em caso de erro. Este tipo de classificação utiliza o princípio da tentativa e erro, sendo tendenciado a escolha das classes das quais geram maiores recompensas.

A classificação de padrões é tarefa desempenhada por diversas técnicas de aprendizado de máquinas (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Surge portanto a necessidade de determinar quais classificadores, e seus parâmetros, são mais eficientes de acordo com cada problemática. Ferramentas como *Waikato Environment for Knowledge Analysis* (WEKA) e Auto-WEKA buscam suprir esta necessidade.

Neste trabalho foram empregadas as técnicas de classificação supervisionadas: *Adaptative Boosting*, *Random Forests* e Máquina de Vetores de Suporte, sendo as duas primeiras através da ferramenta auto-WEKA, para realizar o reconhecimento de padrões presentes no tecido da mama (massas) e classifica-las em maligno ou benigno, contribuindo assim no diagnóstico do câncer de mama.

2.4.1 Waikato Environment for Knowledge Analysis

O *Waikato Environment for Knowledge Analysis* (WEKA) é uma ferramenta projetada para auxiliar a aplicação das técnicas de aprendizagem de máquina. Para isso,

diversas técnicas de clusterização, seleção e classificação de características, são apresentadas ao usuário de forma totalmente abstrata através de uma *Application Programming Interface* (API) "amigável", ocultando assim os formatos de entrada e saída (GARNER et al., 1995). Tipicamente seus módulos são distribuídos em três etapas: processamento do conjunto de dados de entrada, aprendizagem de máquina e processamento do conjunto de saída. A primeira, baseia-se na extração de informações a respeito do conjunto de dados a serem processados, como por exemplo: divisão dos dados em treino e teste, filtragem dos dados, portabilidade dos dados, entre outros. A segunda etapa consiste nas implementações dos algoritmos de aprendizado de máquina, que normalmente tomam um conjunto de dados e produzem uma saída baseada em um conjunto de regras predefinidas em cada algoritmo. Por fim, a última etapa concentra a saída dos algoritmos e realiza uma avaliação mediante um conjunto de dados de teste (GARNER et al., 1995).

Esta ferramenta concentra uma grande variedade de classificadores incluindo, *Random Forests*, *Adaptative Boosting*, *Bagging*, *NaiveBayes*, *Multilayer Perceptron*, entre outros.

Observando esta variedade de técnicas de aprendizado de máquinas existentes e a grande quantidade de parâmetros para cada técnica. Foi proposto por Thornton et al. (2013) o Auto-WEKA, que é uma ferramenta baseada no WEKA. Esta ferramenta objetiva identificar, de maneira automatizada, as melhores técnicas de aprendizagem de máquina e configurações de parâmetros de acordo com cada tipo de aplicação. Um dos objetivos desta ferramenta é a estimação automática de parâmetros de um determinado classificador.

De maneira mais formal, o Auto-WEKA visa os problemas de classificação, onde a aprendizagem é uma função $f : XY$ com Y finito. Um algoritmo de aprendizagem A mapeia um conjunto $d_1, ..., d_n$ de dados, $d_i = (x_i, y_i) \in XxY$, de treinamento para esta função, que é frequentemente expressa através de um vetor de parâmetros do modelo. A maioria dos algoritmos de aprendizagem permite a mudança nos parâmetros, λ, que alteram o funcionamento do próprio algoritmo. Esses parâmetros (λ) são normalmente otimizados em um loop externo, que avalia o desempenho de cada configuração de parâmetros usando validação cruzada.

Segundo Thornton et al. (2013), os parâmetros são, normalmente, contínuos, ou seja, possuem um espaço com alta dimensionalidade onde pode-se explorar a correlação entre diferentes configurações de parâmetros. Os parâmetros são subdivididos em dois tipos, condicional e ativo. O *Sequential Model-Based Algorithm Configuration* (SMAC) é utilizado como forma de estimação automática, onde através da aplicação de um modelo *Random Forests* e uma função de ponderação, este determina os valores ótimos para os dois tipos de parâmetros.

A ideia principal do SMAC é progressivamente melhorar a estimativa destes parâmetros, analisando-os um de cada vez, elevando assim o custo computacional para

executar este processo.

2.4.2 Adaptive Boosting

A estratégia de *Boosting* consiste na combinação de diversos *Weak Classifiers* [7] para a construção de um classificador robusto. O *Boosting* funciona de forma iterativa onde a cada iteração é gerado um classificador simples por meio de um algoritmo-base que utiliza diferentes versões do conjunto de dados de treinamento. Essas versões por sua vez, são obtidas através da variação do peso associado a cada um dos exemplos. Desta forma, diferentes versões ponderadas do conjunto de dados são obtidas. Ao decorrer das iterações, o *Boosting* combina os diversos classificadores parciais para formar um único classificador possivelmente mais robusto. Todos os algoritmos derivados do Boosting adotam estes mesmos procedimentos, diferenciando-os apenas em dois pontos no algoritmo: a forma de ponderação do conjunto de dados em cada iteração e o modo como os classificadores parciais irão se combinar (DUARTE, 2009)(FREUND; SCHAPIRE, 1995).

O *Adaptative Boosting*, ou AdaBoost, é um meta-algoritmo para o aprendizado de máquina que baseia-se na estratégia de Boosting onde este pode ser utilizado em conjunto com qualquer algoritmo de aprendizado de máquina de forma a incrementar a robustez e consequentemente melhorando o desempenho de um conjunto de dados. O fluxo de operações do AdaBoost inicia-se com a chamada de um algoritmo-base em várias iterações t, onde $t \in [1...T]$. Em cada iteração t, é atualizada a distribuição de pesos do conjunto de treinamento de forma a penalizar os pesos dos exemplos classificados incorretamente. Considerando um conjunto de dados de treinamento $\{(x_i, y_i)\}_{i=1}^{n}$ onde $x_i \in X$ e $y_i \in \{-1, +1\}\}$. Inicialmente todos os pesos são iguais, dada uma distribuição de pesos do conjunto de dados na iteração t, D_t. A cada iteração os exemplos classificados incorretamente sofrem penalidade e tem seus pesos aumentados. Em seguida o algoritmo-base deve encontrar um classificador apropriado, h_t, que minimize o erro ponderado de treinamento, ϵ_t, definido na Equação 2.23

$$\epsilon_t = \sum_{\forall i | y_i \neq h_t(x_i)} D_t(i). \tag{2.23}$$

O AdaBoost define um parâmetro α_t para cada classificador h_t, sendo este responsável por medir a importância daquele classificador no final da execução. Em (FREUND; SCHAPIRE, 1995) é provado que o valor ótimo, da variável α_t, que minimiza o erro de

[7] São classificadores simples que possuem desempenho um pouco melhor do que um classificador aleatório. Estes portanto, são menos precisos e complexos em relação a classificadores como: *Naive Bayes* (MURPHY, 2006), *Neural Networks* ou MVS.

treinamento dado de acordo com a Equação 2.24

$$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t}). \tag{2.24}$$

Logo a distribuição de pesos, D_t, pode ser atualizada segundo a Equação 2.25

$$D_{t+1} = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t} \tag{2.25}$$

onde Z_t é uma constante de normalização que realiza a transformação da distribuição D_{t+1} em uma distribuição de probabilidade é definida na Equação 2.26

$$Z_t = \sum_{i=1}^{n} D_t(i) e^{-\alpha_t y_i h_t(x_i)} \tag{2.26}$$

Por fim, o classificador final H, é composto por todos os classificadores parciais gerados em todas as T iterações, sendo que cada um possuiu a capacidade de voto igual a α_t. Este classificador resultante é determinado pela expressão ilustrada na Equação 2.27

$$H(x) = sinal(\sum_{t=i}^{T} \alpha_t h_t(x)), \tag{2.27}$$

onde $sinal(x)$ é dado por

$$sinal(x) = \begin{cases} -1, & x < 0 \\ +1, & x \geq 0 \end{cases} . \tag{2.28}$$

Caso o parâmetro α_t encontrado em cada iteração for igual a zero, ou muito pequeno, o classificador h_t é rejeitado, pois o mesmo não influencia na formação do classificador final. Quando este processo ocorre o algoritmo é interrompido e o classificador final é formado pelos classificadores anteriormente gerados.

2.4.3 Random Forest

As *Decision Tree*, ou Árvores de Decisão, consistem em uma abordagem de apoio à decisão que utiliza um grafo ou um modelo de decisões em árvore, onde cada nó interno representa um "teste" em um atributo. Esta árvore possui três tipos principais de nós, nó de Raiz , nó Interno e nó de finalização. Um exemplo é ilustrado na Figura 16. Esta abordagem tem sido extensivamente estudada no aprendizado de máquina como solução para tarefas de classificação ((BREIMAN et al., 1984), (QUINLAN, 1986), (QUINLAN, 2014)). Estes estudos demonstram o conjunto ou *ensemble* de árvores como um modelo que apresenta melhorias significativas na acurácia da classificação. Sendo que estes *ensemble* adotam um esquema de votação onde o classificador mais votado torna-se um candidato a ser o melhor método de classificação para um determinado conjunto de dados. Um

exemplo de método que utiliza as árvores de decisão é o *Bagging* ou *bootstrap aggregation* (EVERITT, 1998). A ideia principal do *Bagging* é obter uma média de modelos imparciais com o objetivo de reduzir a variança entre os mesmos. As árvores de decisão são o modelo ideal para o *Bagging* devido ao seu poder de representação da complexidade dos dados e baixo bias(DIETTERICH; KONG, 1995).

É uma modificação de outro modelo baseado em árvores de decisão chamado *Bagging* ou *bootstrap aggregation* (EVERITT, 1998)

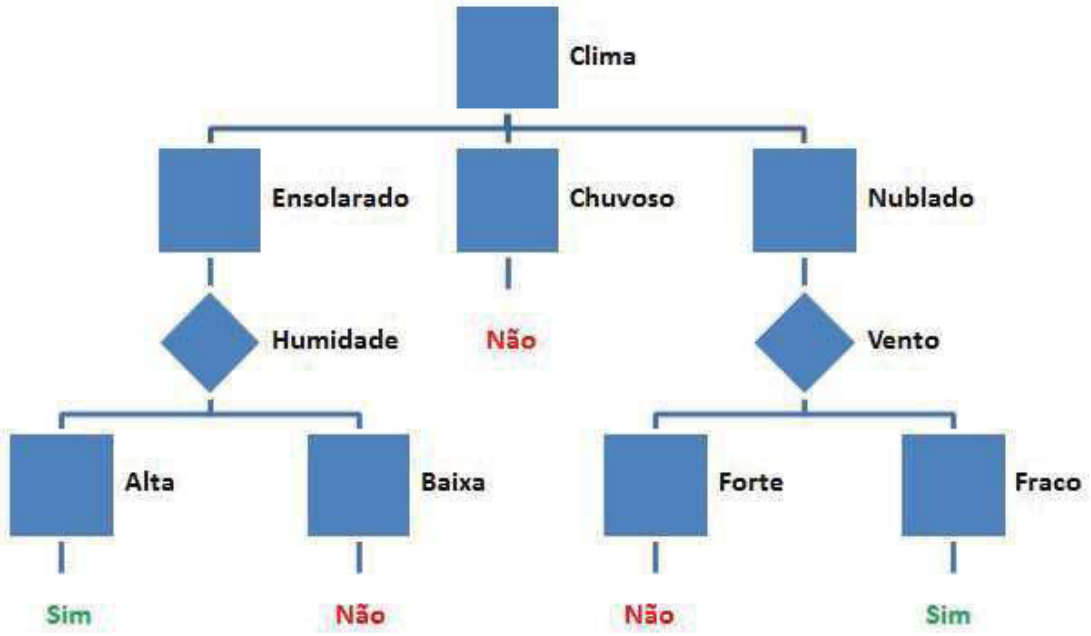


Figura 16 – Exemplo de árvore de decisão tendo como problemática à prática esportiva. Baseado em (SAFAVIAN; LANDGREBE, 1990)

O *Random Forests* (BREIMAN, 2001) é um classificador baseado principalmente em dois métodos, *Bagging* (EVERITT, 1998) e *Random Subspace* (HO, 1998). Este classificador consiste em um conjunto de árvores de classificadores estruturados $\{h(x, \theta_k), l = 1, ...\}$ onde θ_k representam vetores aleatórios independentes e igualmente distribuídos, onde cada árvore tem direito a apenas um único voto para eleger o classificador mais popular proveniente da entrada x. O crescimento de cada árvore é determinado pelas seguintes características:

1. Considerando um conjunto de treinamento N, então N amostras aleatórias constituirão o conjunto de treinamento a ser submetido como entrada para a árvore.

2. Para manter sempre um número constante de variáveis de entrada, V, então um número $v \in V$ é especificado de tal forma que em cada nó, v variáveis são selecionadas aleatoriamente.

3. Cada árvore cresce considerando a maior extensão possível, ou seja, não há poda. Após a construção de cada árvore as aproximações das amostras são calculadas para cada par de nós.

Caso algumas destas ocupem o mesmo nó terminal então sua aproximação é incrementada em um. Ao final da execução, são normalizadas dividindo-se pelo número de árvores. Estas aproximações objetivam a localização de *outliers* (HODGE; AUSTIN, 2004).

Em *Random Forests*, não há necessidade de validação cruzada ou um conjunto de teste separado para obter uma estimativa imparcial do erro do conjunto de teste. Estima-se internamente, durante a execução, onde cada árvore é construída usando uma amostra diferente dos dados originais. Esta estimativa de erro denomina-se *out-of-bag* (OOB). Por exemplo, considerando uma amostra de treinamento $z_i = (x_i, y_i)$ o OOB consiste no erro médio para cara z_i calculado usando previsões das árvores que não contêm z_i em sua respectiva amostra (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Uma definição mais formal é demonstrada a seguir:

Considerando um conjunto de classificadores $h_1(x), h_2(x), ..., h_K(x)$, e um conjunto de treino aleatório a partir da distribuição dos vetores Y, X definidos pela função de margem:

$$mg(X,Y) = av_k I(h_k(X) = Y) - max_{j \neq Y} av_k I(h_k(X) = j). \tag{2.29}$$

Esta função de margem resulta na média do número de votos em X, Y, onde a classe corretamente classificada excederá o número de votos médio em relação as demais classes. Quanto maior a margem, maior a confiança na classificação. O erro de generalização é dado por:

$$PE^* = P_{X,Y}(mg(X,Y) < 0), \tag{2.30}$$

onde os índices X, Y indicam que a probabilidade está sobre o espaço X, Y. Em *Random Forests*, $h_k(X) = h(X, \theta_k)$. Para um grande volume de árvores, aplica-se a lei da força dos números grandes, ou *Strong law of Large Numbers* nas estruturas das árvores. Esta lei é definida pelo teorema que possui a premissa de que à medida que o número de árvores aumenta, quase certamente todas as seqüências $\theta_1, ...PE^*$ convergem para

$$P_{X,Y}(P_\theta(h(X,\theta) = Y)) - max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0). \tag{2.31}$$

Logo a função de margem para uma *Random Forests* é

$$mr(X,Y) = P_\theta(h(X,\theta) = Y)) - max_{j \neq Y} P_\theta(h(X,\theta) = j). \tag{2.32}$$

2.4.4 Máquina de Vetores de Suporte

A Máquina de Vetores de Suporte (MVS) é uma técnica de aprendizagem de máquina supervisionada que busca uma classificação. A MVS foi inicialmente proposta por Boser, Guyon e Vapnik (1992) e logo tornou-se conhecida e amplamente utilizada na área de mineração de dados, devido ao seu elevado índice de acerto e sua precisão ao modelar situações não-lineares complexas. A Teoria de Aprendizado Estatístico (TAE), proposta por Vapnik (1998), constitui a base do aprendizado de máquina utilizado pela MVS. Esta teoria quantifica os registros erroneamente classificados e utiliza o princípio da indução para gerar um modelo de predição que torne o erro sobre os dados de teste mínimo, até um limite que seja considerado aceitável.

O princípio básico por trás da MVS consiste na construção de um hiperplano que desempenhe a função de uma superfície de decisão, tendo como premissa que a margem de separação entre as classes seja máxima. Desta forma, o treinamento realizado através da MVS objetiva a obtenção de hiperplanos que dividam as amostras de tal forma que os limites de generalização sejam ótimos.

Por utilizar um espaço de hipóteses de funções lineares em um espaço multidimensional, as MVS são consideradas um sistema de aprendizagem. Seus algoritmos de treinamento sofreram bastante influência das áreas da teoria de otimização e aprendizagem estatística. Desta forma, sua superioridade vem sendo comprovada frente a diversos outros classificadores em uma grande variedade de aplicações (CRISTIANINI; TAYLOR, 2000).

O classificador MVS encontra um hiperplano baseado em um conjunto de pontos denominados "vetores de suporte" dado um conjunto de amostras separáveis, normalmente composto por duas classes. Estes vetores maximizam a margem de separação entre as classes. Em um hiperplano, que é uma superfície de separação entre regiões multidimensionais, pode haver ate um número infinito de dimensões. Entretanto, mesmo para os casos de classes não separáveis, a MVS utilizando conceitos provenientes da teoria da otimização consegue encontrar um hiperplano ótimo. A Figura 17 ilustra o hiperplano de separação entre duas classes linearmente separáveis.

Considerando um conjunto de amostras de treinamento (x_i, y_i) sendo $x_i \in \Re^n$ o vetor de entrada, y_i a classificação correta das amostras e $i = 1, ..., n$ o índice de cada ponto amostral. A classificação objetiva a estimação da função $f : \Re^n \rightarrow \{\pm 1\}$, que separe corretamente os exemplos de teste em classes distintas.

A etapa de treinamento estima a função $f(x) = (w.x) + b$, procurando por valores de w e b que satisfação a relação definida na Inequação 2.33

$$y_i((w.x_i) + b) \geq 1 \tag{2.33}$$

sendo w é o vetor normal ao hiperplano de decisão e b corresponde a distância da função f em relação à origem. Tendo em vista a descoberta dos valores ótimos de w e b, a Equação

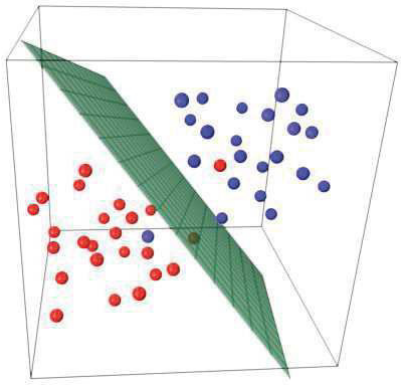


Figura 17 – Separação de duas classes através de hiperplanos

2.34 é minimizada segundo a restrição apresentada na Equação 2.33 (CHAVES, 2006).

$$\Phi(w) = \frac{w^2}{2} \tag{2.34}$$

Variáveis de folga são incluídas para que seja possível encontrar um hiperplano nos casos em que não haja uma perfeita separação entre classes. Estas variáveis permitem que as restrições da Equação 2.33 não influenciem nesta separação. A minimização da Equação 2.35 de acordo com a restrição evidenciada pela Equação 2.33, constitui o problema de otimização. C é um parâmetro de treinamento que proporciona o equilíbrio entre a complexidade do modelo e o erro de treinamento. Este deve ser informado pelo usuário

$$\Phi(w, \xi) = \frac{w^2}{2} + C \sum_{i=1}^{N} \xi_i \tag{2.35}$$

$$y_i((w.x_i) + b) + \xi_i \geq 1 \tag{2.36}$$

onde, C é uma penalidade para a função Φ, ξ é a variável de folga que objetiva a suavização das restrições presente na Equação 2.36 e N é o número de amostras de entrada.

A Equação 2.37 é baseada na teoria dos multiplicadores de Lagrange, que possui como objetivo encontrar os α_i multiplicadores de Lagrange ótimos que satisfaçam a Equação 2.38 (CHAVES, 2006).

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i x_j) \tag{2.37}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \ \leq C \tag{2.38}$$

Os pontos que sejam submetidos a restrição da Equação 2.33 e esta por sua vez é exatamente igual à unidade, o multiplicador de Lagrange nestes casos é diferente de zero, ou seja, $\alpha \neq 0$. Esses pontos são denominados vetores de suporte, pois sua localização

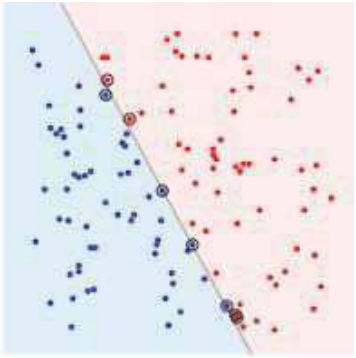


Figura 18 – Vetores de suporte (destacados por círculos)

geométrica é normalmente sobre as margens. Tais pontos têm fundamental importância na definição do hiperplano ótimo, devido a sua delimitação a margem do conjunto de treinamento. A Figura 18 exemplifica estes vetores.

Amostras linearmente não separáveis, a MVS necessita de uma transformação não-linear que transforme o espaço de entrada em um novo espaço de características. Esse novo espaço deve possuir uma dimensão suficientemente grande, tendo em visa que as amostras possam através deste espaço ser linearmente separáveis. Desta forma, o hiperplano de separação é definido então como uma função linear de vetores retirados do espaço de características ao invés do espaço de entrada original. A construção do hiperplano depende de uma função K de núcleo de um produto interno (HAYKIN, 2000). Esta função mapeia as amostras em um elevado espaço de dimensão sem aumentar a complexidade dos cálculos.

A Equação 2.39 mostra o resultado da Equação 2.37 com a utilização de um núcleo K.

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2.39}$$

A função de base radial, ou Radial Basis Function (RBF) é uma função de núcleo amplamente aplicada em problemas de reconhecimento de padrões e também utilizada neste trabalho. Esta é definida na Equação 2.40.

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2) \tag{2.40}$$

3 Metodologia

Estão contidos neste capítulo os procedimentos realizados para a construção da metodologia proposta, que objetiva a diferenciação dos padrões de malignidade e benignidade em massas, a partir de imagens de mamografias digitais. Esta metodologia proposta é composta por cinco macroetapas, que são: base de imagens, preprocessamento, extração de características locais, reconhecimento de padrões e validação. A Figura 19 ilustra o fluxo geral da metodologia.

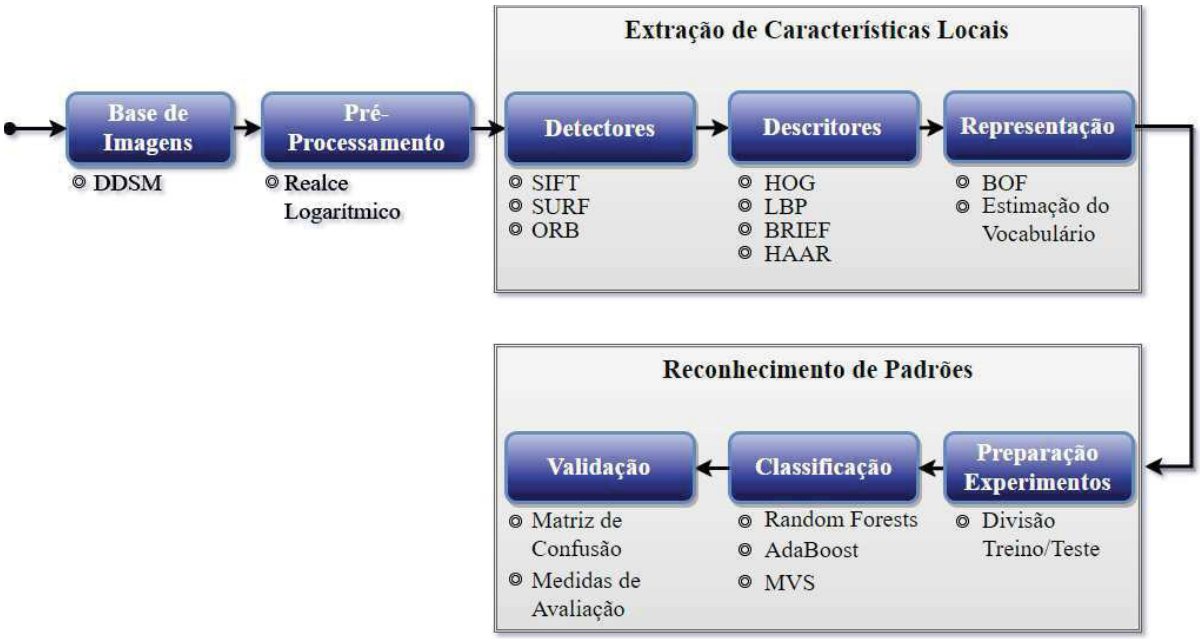


Figura 19 – Etapas da Metodologia Proposta.

3.1 Base de Imagens

O objetivo desta etapa consiste na obtenção das imagens mamográficas, sendo esta estritamente necessária para o prosseguimento da metodologia. Desta forma, utilizou-se a base pública de mamografias digitalizadas, *Digital Database for Screening Mammography* (DDSM) (HEATH et al., 1998), disponibilizada via internet.

Esta base é um projeto envolvendo quatro instituições americanas, *Massachusetts General Hospital*, *University of South Florida*, *Sandia National Laboratories* e *Washington University School of Medicine*. O objetivo principal desta base é fomentar a pesquisa e desenvolvimento de algoritmos e técnicas computacionais, como forma de auxilio na detecção e diagnóstico de patologias relacionadas a mama. A DDSM contém aproximadamente 2.500 casos, onde para cada caso há duas imagens de ambas as mamas, nas projeções crânio-

caudal e médio-lateral oblíqua juntamente com informações associadas ao paciente como: idade do paciente, classificação da densidade da mama, sensibilidade a anomalias (tipo de patologia), localização das anomalias. Em relação à imagem, informações como: nome do arquivo, tipo do digitalizador (*scanner*), quantidade de *pixels* por imagem, também são disponibilizadas (USF - University of South Florida, 2001).

É importante destacar que para as imagens que possuem áreas consideradas suspeitas, é disponibilizado um arquivo extra, chamado *overlay*, contendo a descrição da(s) lesão(ões). Neste arquivo estão presentes informações como: quantidade de lesões, localização da lesão, contorno da lesão e seu diagnóstico. A codificação *chain code* foi utilizada como forma de representação dos contornos das lesões presentes nas mamografias. As descrições das lesões seguem os termos lexicográficos[1] padronizados pelo *American College of Radiology* (ACR) e são publicados no *Breast Imaging Reporting and Data System* (BI-RADS) (ACR - American College of Radiology, 1998).

Tendo em vista que este trabalho objetiva o diagnóstico de câncer de mama através de regiões anotadas das mamografias, as regiões presentes nas imagens de mamografia que não estão inseridas na marcação do especialista, não foram utilizadas para análise. Para seleção das amostras, aplicou-se a mesma abordagem empregada em (ROCHA, 2014), onde mediante as marcações dos especialistas nas mamografias, regiões de interesse[2] (ROIs) foram extraídas contendo apenas as regiões com as massas, totalizando 1.155 ROIs. Os experimentos realizados neste trabalho utilizaram as mesmas 1.155 ROIs produzidas em (ROCHA, 2014), sendo 625 massas malignas e 530 massas benignas.

3.2 Pré-Processamento

O processamento digital de imagens é definido, segundo Gonzalez e Woods (1992) como o conjunto de técnicas computacionais que transformam uma imagem de entrada em uma saída desejada. Uma imagem digital consiste em uma função $f(x, y)$ discreta em relação as coordenadas espaciais e intensidade, onde esta pode ser considerada como uma matriz cujos índices indicam a coordenada espacial de um *pixel*. Técnicas de realce de imagens surgem com o objetivo de adequar a imagem, através da manipulação de *pixels*, para melhorar a visualização e processamento em determinadas aplicações específicas (GONZALEZ; WOODS, 1992).

Com base nestes conceitos, esta etapa busca melhorar o contraste do objeto de interesse, neste caso as massas, em relação ao fundo que possa existir nas ROIs e, desta forma, prover uma melhor descrição tanto da geometria quanto da textura das massas. Para tanto, a técnica de realce logarítmico foi utilizada para desempenho desta tarefa. O

[1] Que se refere à lexicografia ou ao lexicógrafo.
[2] *Region Of Interest*, ou Região de interesse.

realce logarítmico é definido pela Equação 3.1

$$g_t(x, y) = G \log_{10}(g(x, y) + 1) \tag{3.1}$$

onde $g_t(x, y)$ é o valor do novo nível de cinza nas coordenadas (x, y), $g(x, y)$ é valor original do nível de cinza e G é uma constante definida de acordo com a Equação 3.2

$$G = \frac{p_{max}}{\log_{10}(1 + |R|)}. \tag{3.2}$$

onde, p_{max} corresponde ao maior valor de nível de cinza contido na imagem e R representa o valor máximo do nível de cinza, normalmente 255.

A técnica de realce logarítmico consiste na aplicação de uma função logarítmica no histograma (GONZALEZ; WOODS, 1992) original da imagem. Esta função possui uma significativa inclinação na porção relativa aos níveis de cinza com baixa intensidade. Devido a esta inclinação os níveis de cinza de alta intensidade não sofrem grandes influências. Por conseguinte, esta propriedade busca realçar as informações contidas nas porções de níveis de cinza de menor intensidade, ou seja, regiões mais escuras da imagem, em detrimento de um baixo realce nas porções de níveis de cinza com maior intensidade (GONZALEZ; WOODS, 1992).

A função logarítmica, de acordo com a Equação 3.1, necessita da definição da constante G. Esta constante visa garantir que os novos valores de *pixels* estejam normalizados entre 0 e o nível de cinza máximo permitido pela imagem. Neste trabalho, o valor da constante G foi $105,97$, tendo em vista que valores de magnitude e limite máximo das ROIs são iguais a 255. Este fato ocorre devido as ROIs possuírem 8 *bits* por *pixel*. Esta constante foi aplicada para todas as imagens da base, ou seja, nas 1155 ROIs.

O resultado da aplicação do realce logarítmico nas ROIs pode ser visualizado na Figura 20.

3.3 Extração de Características Locais

O propósito desta etapa, consiste em identificar e extrair informações ou características locais contidas nas imagens anteriormente preprocessadas. Estas características contribuirão para a diferenciação dos padrões de malignidade e benignidade em massas. Esta etapa também é responsável por produzir medidas descritivas das imagens objetivando a formação de vetores de características, a serem classificados na etapa de reconhecimento de padrões (Seção 3.4). As Seções 3.3.1, 3.3.2 e 3.3.3 descrevem as etapas de detecção, descrição e representação das características juntamente com a configuração dos métodos utilizados.

Neste trabalho, foi utilizada uma série de detectores e descritores de características locais, no intuito de estabelecer uma comparação através de suas aplicações em imagens mamográficas, tendo como objetivo a diferenciação de malignidade e benignidade.

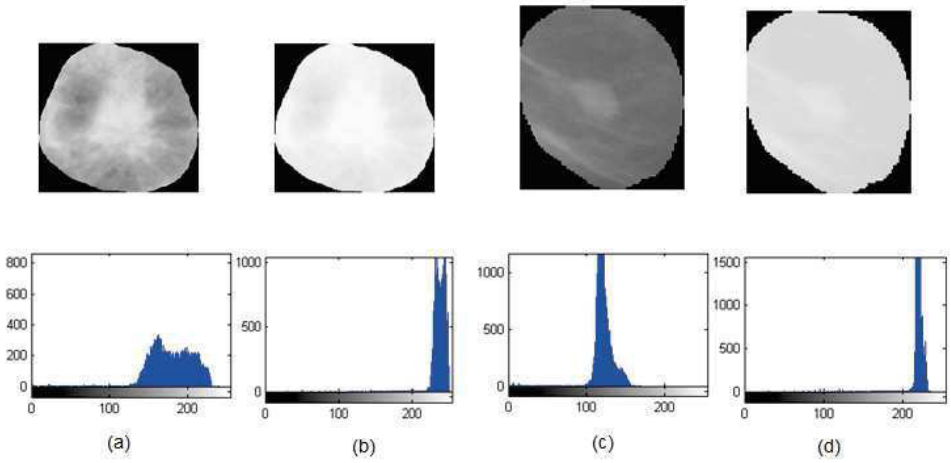


Figura 20 – Resultado da aplicação do realce logarítmico e seus respectivos histogramas. (a) e (c) ROIs originais, (b) e (d) ROIs após o realce logarítmico. Sendo que (a) corresponde a uma massa maligna e (b) massa benigna.

3.3.1 Detectores

Os detectores de características locais utilizados neste trabalho, bem como suas configurações, são apresentados a seguir.

A primeira abordagem utilizada para a identificação das características locais, foi o detector SIFT. Este detector, como discutido anteriormente (Seção 2.2.1.1), utiliza as diferenças de gaussianas (DoG) para a identificação de pontos de interesse nas imagens. O nível de suavização, σ, é um fator determinante que contribui para a qualidade de informações resultantes da DoG. Desta forma, é necessário determinar o valor inicial de σ, sendo que neste trabalho determinou-se empiricamente $\sigma = 1,6$, onde este valor preserva informações relevantes ao mesmo tempo em que elimina possíveis ruídos. Outro elemento que regula a qualidade das informações é o limiar de rejeição, onde ao contrário do nível de suavização, este busca eliminar pontos de interesse incorretamente identificados como ruídos. Este limiar varia entre 0 e 1, considerando que os níveis de cinza estejam normalizados neste intervalo. O valor de $0,03$ foi determinado para este limiar, onde este apresentou uma ótima filtragem de pontos incorretamente classificados. As oitavas determinam a quantidade de iterações a serem desenvolvidas tanto pelo detector SIFT, quanto pelo descritor SIFT (Seção 3.3.2). Neste trabalho, utilizou-se o mesmo valor aplicado em (LOWE, 2004), que consiste em 3 oitavas. A quantidade máxima de pontos de interesse não foi determinada, ou seja, o detector SIFT identificou sempre o máximo de pontos possíveis. A quantidade média de pontos identificados nas ROIs foi de 173 pontos por ROI, conclui-se então que as 1.155 imagens resultam em 199.815 pontos de interesse.

Como segunda abordagem o detector SURF foi aplicado com o mesmo objetivo. Para tanto, as medidas de localização e escala dos pontos de interesse são determinados com base na matriz hessiana. Desta forma, um limiar hessiano é utilizado para limitar a

saída desta matriz de modo a determinar quais pontos serão considerados como pontos de interesse, ou seja, quanto maior o valor deste limitador, menos pontos de interesse serão identificados, porém pontos mais significantes, teoricamente, serão identificados. Contudo, quanto menor o valor deste limiar maior a quantidade de pontos, porem perde-se em qualidade. O valor 100 foi estabelecido empiricamente como sendo um bom limitador. O detector SURF utiliza o conceito semelhante de oitavas do detector SIFT, onde este também corresponde a quantidade de iterações. A quantidade de oitavas estabelecida neste trabalho foi igual a 4.

Por fim, o detector ORB foi o terceiro e último detector aplicado. O oFAST, descrito na Seção 2.2.1.3, é responsável por identificar pontos de interesse invariantes a rotação através da análise dos "cantos" ou *corners*. Portanto, é necessário determinar a quantidade máxima de *pixels* a serem considerados como bordas. Este limiar de *corners* define a quantidade de *pixels* dos círculos a serem considerados ao analisar se os pontos encontrados pertencem ao "canto" ou não, definindo assim se são pontos de interesse. Neste trabalho o valor de 31 *pixels* foi determinado para este limiar. A quantidade máxima de pontos a serem identificados foi determinada como 500 pontos de interesse por imagem. Esta quantidade foi determinada empiricamente. Este detector aplicado ao conjunto de 1.155 ROIs identificou um total de 308.074 pontos de interesse obtendo-se uma média de 266 pontos por imagem.

3.3.2 Descritores

Com o término da etapa anterior, foram detectados e obtidos um conjunto de pontos de interesse locais por imagem, proveniente da aplicação individual de cada detector, SIFT, SURF e ORB. Cada ponto identificado representa uma característica local descriminante das lesões. Estes pontos são invariantes a rotação e escala. Desta forma, quatro descritores foram aplicados a estes conjuntos de pontos para a análise e extração de características locais, são eles: *Histogram of Gradients*, Wavelet de Haar, *BRIEF* e LBP.

O *Histogram of Gradients* é utilizado como forma de descrição dos pontos de interesse provenientes do detector SIFT. Este descritor, como apresentado na Seção 2.2.1.1, baseia-se na análise das orientações, calculadas a partir da magnitude do gradiente, dos pontos ao redor do ponto de interesse, tornando-o invariante a escala e rotação. O histograma de orientações de gradientes utilizado neste trabalho possui tamanho de $4x4$ e 8 orientações, resultando em um vetor com 128 características por ponto.

Com o mesmo objetivo, o descritor LBP foi aplicado ao conjunto de pontos do detector SIFT. Esta abordagem tem como objetivo a análise da textura em torno do ponto, tal processo ocorre da seguinte maneira: (1) Dado um conjunto de pontos de interesse, Figura 21(a), contendo informações de coordenadas, escala e orientação, uma área de 16x16 pixels e recortada da imagem de massa em torno de cada ponto de interesse, Figura

21(b), nas mesmas coordenadas, orientações e escala. Em seguida, é aplicado o método LBP para obter as características de textura apenas na região local da massa, ou seja, ao redor do ponto de interesse, Figura 21(c). Por fim, um histograma de ocorrência de níveis de cinza é acumulado com os valores resultantes do método LBP, Figura 21(d). O tamanho do vetor de característica gerado é de 256 características por ponto de interesse.

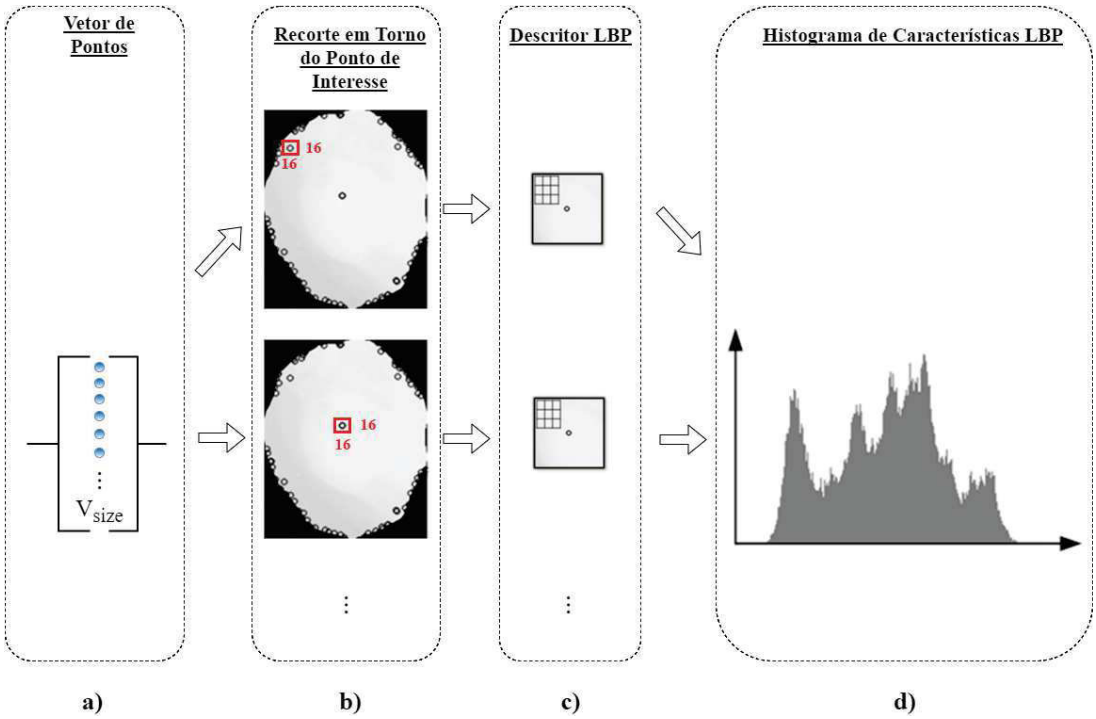


Figura 21 – Fluxo para criação do descritor LBP. (a) Conjunto de pontos de interesse identificados pelo detector SIFT.(b) Recorte da área, 16x16 *pixels*, em torno dos pontos de interesse preservando as informações de escala, orientação e coordenadas. (c) Análise local da textura com o descritor LBP. (d) Histogramas de características LBP, onde o eixo x corresponde aos níveis de cinza na imagem e o eixo y corresponde a frequência dos valores resultante do calculo LBP.

As Wavelet de Haar, como apresentado anteriormente (Seção 2.2.1.2), são responsáveis tanto por tornar os pontos de interesse, derivados do detector SURF, invariantes a rotação, quanto por determinar as direções horizontais e verticais em relação aos pontos. Nesta abordagem, uma janela quadrática de tamanho $8x8$ é aplicada e orientada em torno do ponto de interesse. Em seguida, para cada subregião de $4x4$ são calculadas as Wavelet de Haar e seus valores absolutos. Desta forma, cada subregião é calculada 4 vezes, produzindo assim $(4 * 4 * 4 = 64)$ um vetor resultante com até 64 características. Neste trabalho, utilizou-se $16x16$ como tamanho da janela e $4x4$ subregiões, produzindo assim um vetor de característica de tamanho 128.

Por fim, o descritor BRIEF foi aplicado no conjunto de pontos identificados pelo detector ORB. O BRIEF é um descritor binário (Seção 2.2.1.3) onde as características são definidas como um vetor de testes binário com tamanho fixo. Em (CALONDER et al., 2010), diferentes tamanhos $(128, 256, 512)$ foram testados. O experimento utilizando um

vetor binário com tamanho 256, foi obtido como melhor resultado. Neste trabalho, bem como em (CALONDER et al., 2010), o tamanho do vetor foi determinado como 256.

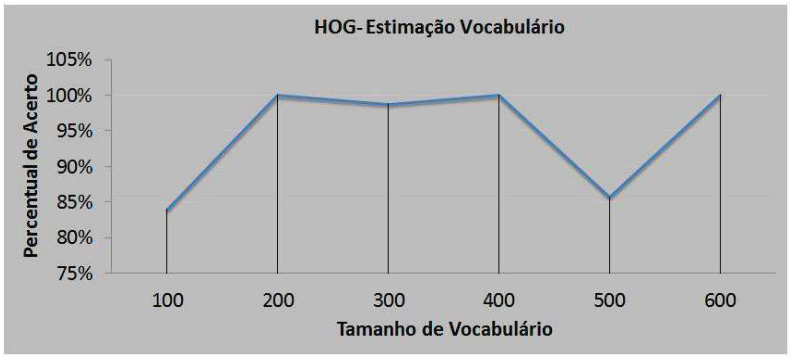
3.3.3 Representação

Mediante o término da etapa anterior (descritores), todos os vetores de características locais extraídos de cada imagem possuem vários pontos de interesse. Desta forma, a quantidade de vetores de características geradas está diretamente vinculada ao número de pontos de interesse identificados em cada imagem. Tomando o detector SIFT como exemplo, a quantidade total de pontos obtido nas 1.155 ROIs foi de 199.815 pontos de interesse, conforme apresentado na Seção 3.3.1. Observando essa quantidade, necessitou-se utilizar um modelo de representação que reduza a dimensionalidade dos dados mantendo a representatividade das características locais, de forma a não comprometer o desempenho dos classificadores presentes na etapa de reconhecimento de padrões (Seção 3.4).

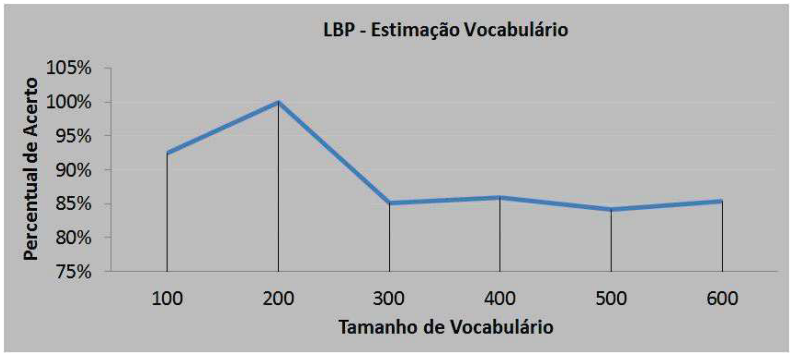
O modelo de representação adotado neste trabalho foi o *Bag of Features* (BoF). O BoF utiliza o conceito de "palavras visuais" (Seção 2.3.1) como forma de representação dos vetores de características gerados anteriormente (HoG, LBP, Wavelet de Haar e BRIEF). Buscando uma melhor representação dessas palavras, o modelo utiliza a técnica de clusterização K-Means para o agrupamento dos conjuntos de palavras visuais, onde determina-se um valor k que representa a quantidade de agrupamentos de palavras que serão criados. O conjunto desses agrupamentos forma um vocabulário de palavras, sendo este limitado pela quantidade de agrupamentos determinado pelo K-Means.

Diversas metodologias foram propostas para a estimação da quantidade ótima de agrupamentos que devem ser gerados, como apresentado na Seção 2.3.1.1. Neste trabalho, adotou-se a abordagem de estimação através da análise de gráfico. Para tanto, foi realizado um experimento com tamanhos de dicionário variando de 100 a 600, com incrementos de 100. Estes vocabulários foram aplicados somente às características provenientes dos descritores LBP e HoG, tendo como objetivo encontrar o tamanho do vocabulário adequado com melhor representatividade no contexto das imagens mamográficas. Após a construção dos vocabulários, estes são submetidos à ferramenta Autoweka (Seção 2.4.1) para determinar qual vocabulário obteve melhor acurácia foi usado o processo de validação cruzada, e por conseguinte, determinar qual tamanho de dicionário é o mais representativo. A Figura 22, ilustra o gráfico com o resultado dos tamanhos de vocabulários e seus respectivos percentuais de acertos para a base de ROIs utilizando as abordagens HOG (Figura 22a) e LBP (Figura 22b). Baseado nestes resultados, o tamanho de vocabulário 200 mostrou-se representativo em ambas as abordagens, portanto este valor foi adotado para a construção dos vocabulários dos demais descritores deste trabalho.

O modelo BoF aplicado neste trabalho segue as etapas ilustradas na Figura 23. Inicialmente as 1.155 ROIs são submetidas ao processo de detecção e descrição das



(a)



(b)

Figura 22 – Experimento estimação do tamanho do vocabulário. (a) Utilizando descritor HoG. (b) Utilizando descritor LBP.

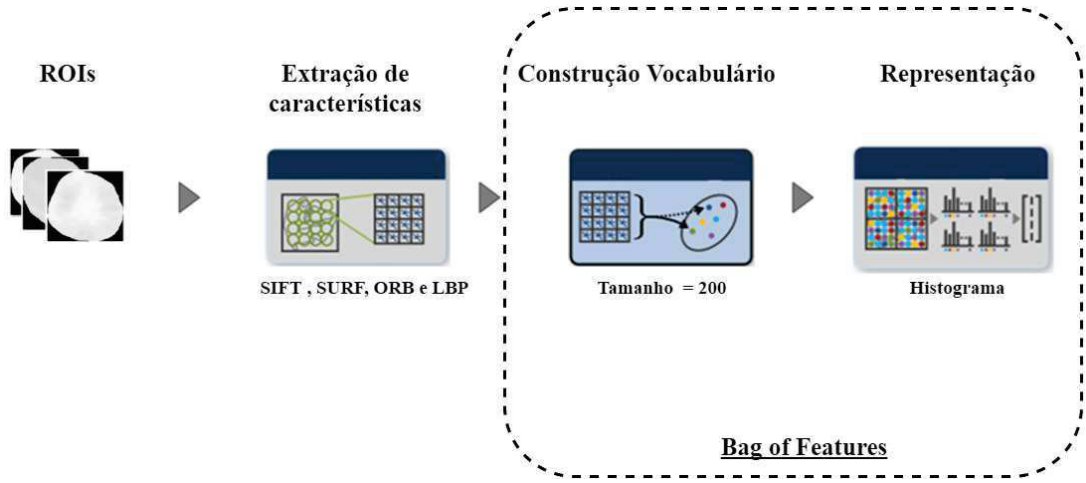


Figura 23 – Visão geral do modelo bag of features (BoF) aplicado neste trabalho.

características locais, através dos descritores anteriormente apresentados (SIFT, SURF, ORB e LBP). Em seguida, 80% do total de ROIs é utilizado para a construção do vocabulário, onde as características serão agrupadas por meio do algoritmo K-Means. A quantidade de agrupamentos ou tamanho do vocabulário foi determinada como 200 (Seção 2.3.1.1). Os agrupamentos resultantes são então comparados a todas as imagens da base utilizando a distância de Hamming, conforme apresentado na Seção 2.3.1, para a construção de um histograma de características. Por fim, o vetor de característica final

é construído baseado no histograma de característica e sua dimensão é limitada pelo vocabulário, portanto seu tamanho é de 200 características por ROI.

3.4 Reconhecimento de Padrões

A última etapa da metodologia proposta consiste em analisar se as características locais, anteriormente geradas, diferenciam padrões de malignidade e benignidade, utilizando reconhecimento de padrões (Seção 2.4). O fluxo das atividades desta etapa é ilustrado na Figura 24.

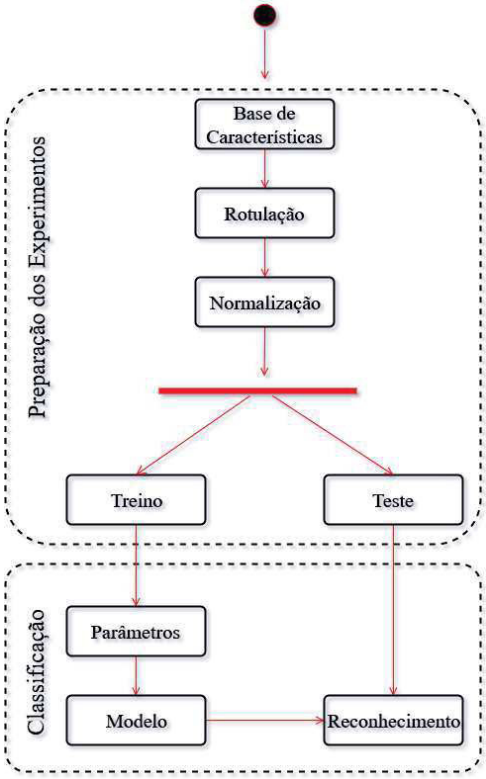


Figura 24 – Fluxo de atividades da etapa de reconhecimento.

3.4.1 Preparação dos Experimentos

Esta etapa consiste na preparação das amostras para a etapa de classificação (Seção 3.4.2). A extração de características locais provenientes da etapa anterior (Seção 3.3), resulta em um vetor de características contendo uma quantidade de características por imagem, ou seja, cada uma das 1.155 ROIs possui 200 características. Cada conjunto de características recebe uma rotulação de acordo com sua classe. Neste trabalho adotou-se os rótulos 0 e 1 como forma de representar o conjunto de características resultantes das imagens benignas e malignas, respectivamente. Após a rotulação, é formada uma base de características para cada abordagem de descritores utilizada.

Diante da diversidade de valores apresentados pelas características, necessitou-se normalizar estes valores para uma faixa comum, neste caso -1 a 1. Tal mecanismo auxilia na convergência dos classificadores na etapa de classificação, bem como padroniza a distribuição de valores das variáveis.

Tendo em visa a validação dos resultados a serem gerados na etapa de classificação, a base de características foi dividida aleatoriamente em dois grupos: base de treino e base de teste. A Tabela 2 apresenta os percentuais de treino e teste. Um total de três diferentes combinações de percentuais de treino e teste são utilizados (Tabela 2), com o objetivo de avaliar o desempenho da metodologia em relação ao gradativo aumento de amostras de treino e redução de amostras de teste.

Tabela 2 – Combinações de percentuais treino/teste na etapa de Preparação dos Experimentos.

Configuração	Treino	Teste
70/30	70%	30%
80/20	80%	20%
90/10	90%	10%

3.4.2 Classificação

Uma vez que os conjuntos de características já estão rotulados, normalizados e subdivididos em base de treino e teste, classificadores são utilizados para realizar a diferenciação de padrões de malignidade e benignidade.

Neste trabalho, para efeitos de comparação foram aplicados os classificadores MVS, *AdaBoost* e *Random Forests*, conforme descrito na Seção 2.4. Estes classificadores foram aplicados a todas as bases de características formadas a partir da etapa de extração de características (Seção 3.3) (SIFT, SURF, ORB e LBP), de acordo com as proporções anteriormente descritas na Tabela 2.

A estimação dos parâmetros dos classificadores *AdaBoost* e *Random Forests* foi realizada usando o Auto-Weka. Para o classificador *Random Forests* os parâmetros estimados pela ferramenta Auto-WEKA (Seção 2.4.1) são: a quantidade máxima de árvores (K) e a profundidade (D). A estimação do primeiro parâmetro baseia-se no intervalo de valores $1-32$. Contudo, como o parâmetro referente a profundidade das árvores não é determinado (Seção 2.4.3) sua estimação não é necessária. O *AdaBoost*, necessita da definição de um *Weak Classifier* inicial (Seção 2.4.2. A escolha deste classificador é estimada por meio do Auto-WEKA, onde são testados diversos classificadores iniciais, sendo que aquele que resulta na melhor classificação é apresentado como melhor solução.

Para a MVS foi utilizado o núcleo radial. Os valores de C e γ (Seção 2.4.4), devem ser estimados. Estes valores foram estimados através de uma busca exaustiva realizada

pelo *script* em *python* chamado *grid.py* contido no pacote LIBSVM. Este *script* busca por meio de validação cruzada a melhor combinação de parâmetros para a base que retorne o melhor percentual de acerto total, em relação aos dados de treino e teste na validação cruzada.

Por fim, ao decorrer da etapa de treinamento é gerado um modelo para cada classificador. Todos os parâmetros são estimados com base nas amostras de treinamento. Esta etapa é realizada totalmente às "cegas", ou seja, as amostras de teste são completamente desconhecidas. Desta forma, podem-se simular as condições reais de teste. Após a construção dos modelos, novas amostras poderão ser submetidas e classificadas. Vale ressaltar que, o vocabulário do BoF é utilizado apenas como forma de redução do espaço de características e não influência para na criação da base de treinamento.

3.4.3 Validação dos Resultados

O resultado da etapa de reconhecimento de padrões consiste em predições, normalmente baseada em probabilidades. Surge portanto a necessidade da validação destes resultados mediante métricas quantitativas.

Embora existam diversas medidas que avaliam o desempenho dos classificadores, a medida considerada mais significativa, consiste na avaliação do desempenho destes a partir da classificação de novos casos (conjunto de testes). Desta forma, metodologias inseridas no contexto de processamento de imagens e reconhecimento de padrões, vinculadas a área médica, costumam ter seu desempenho avaliado através de cálculos estatísticos sobre os resultados dos testes. Esta avaliação, no contexto da área da saúde, baseia-se em determinar o grau de discriminação dos testes, em relação a presença ou ausência de uma doença. Ainda neste contexto, existe a presença de uma variável preditora (resultado do teste) e uma variável resultante (presença ou ausência da doença).

Desta forma, considerando uma doença qualquer com casos positivos (possui a doença) e negativos (não possui a doença), quatro possíveis situações subdividem os resultados dos testes de classificação:

- Verdadeiro Positivo (VP): Casos positivos classificados corretamente como casos doentes;
- Verdadeiro Negativo (VN): Casos negativos classificados corretamente como casos saudáveis;
- Falso Positivo (FP): Casos negativos classificados erroneamente como casos doentes;
- False Negativo (FN): Casos positivos classificados erroneamente como casos saudáveis.

A matriz de confusão permite quantificar os casos dos testes de classificação, onde cada padrão é vinculado a uma das quatro categorias (VP, VN, FP, FN), compondo assim as medidas estatísticas de desempenho dos classificadores. A Tabela 3 apresenta a matriz de confusão.

Tabela 3 – Matriz de Confusão. Baseado em (DOWNEY et al., 1999)

	Doença	
Resultado do Teste	Presença	Ausência
Positivo	VP	FN
Negativo	FP	VN

Algumas estatísticas aplicadas para avaliar o desempenho do classificador, são derivadas da matriz de confusão. Neste trabalho, utilizou-se as estatística de Sensibilidade (S), Especificidade (E), Acurácia (Acc), F1-Score e Precisão.

A medida de sensibilidade de um teste é baseada pela proporção de casos doentes que foram classificados como doentes. Esta métrica indica o quão bom é o teste em identificar indivíduos doentes, dada pela Equação 3.3.

$$S = \frac{VP}{VP + FN} \tag{3.3}$$

Ao contrário da sensibilidade, a especificidade é indicada para identificar indivíduos saudáveis. Por tanto, baseia-se na proporção de pessoas saudáveis corretamente classificados como saudáveis, definida pela Equação 3.4.

$$E = \frac{VN}{VN + FP} \tag{3.4}$$

A acurácia é definida como a razão entre o número de casos da amostra que foram corretamente classificados e o número total de casos, como ilustra a Equação 3.5.

$$ACC = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.5}$$

A precisão é uma medida que representa a quantidade de casos relevantes selecionados em termos de TP e FP. O valor resultante desta métrica varia entre 0 e 1, sendo que quanto mais próximo de 1 mais precisa é a classificação. Esta métrica é definida na Equação 3.6.

$$P = \frac{VP}{(VP + FP)} \tag{3.6}$$

Outra medida estatística utilizada como forma de validação é o F1-Score. Esta medida demonstra a média harmônica entre a precisão e sensibilidade, ou seja, o quão equilibrada estão estas duas medidas. O F1-Score é definido de acordo com a Equação 3.7.

$$F1 = \frac{2 * VP}{(2 * VP + FP + FN)}. \tag{3.7}$$

4 Resultados e Discussões

O objetivo deste capítulo consiste na apresentação e discussão dos resultados obtidos pela metodologia proposta por este trabalho, tendo em vista o diagnóstico de regiões de massas extraídas de mamografias em malignas ou benignas. Os resultados de cada uma das quatro abordagens de extração de características apresentados na Seção 3.3 são discutidos individualmente, bem como a representação destas utilizando o modelo *Bag of Features*. Conforme descrito na Seção 3.4.1, em cada proporção de treino e teste, foram obtidos 5 bases aleatórias de modo a garantir a generalização da metodologia. Portanto, são apresentadas as médias de cada métrica em relação as proporções. Em seguida, os resultados obtidos da etapa de classificação são avaliados através das médias de acurácia, sensibilidade, especificidade, precisão, F1-Score e seus respectivos desvios-padrão.

Os experimentos foram subdivididos em dois grupos: abordagens que utilizam o modelo *Bag of Features* juntamente com detectores e descritores de características locais, Seção 4.2, e as que utilizam somente o descritor de características globais, Seção 4.1.

4.1 Abordagens somente com Descritores

Para fins de comparação, as abordagens a seguir não utilizam *Bag of Features* e as características extraídas são obtidas de maneira global na imagem, ou seja, os métodos são aplicados sem uma referência local (ponto de interesse) na imagem. Portanto todos os *pixels* da imagem influenciam no cálculo do descritor.

4.1.1 Usando Descritor HOG

Nesta abordagem são gerados 288 características por imagem. Os resultados alcançados por esta abordagem são apresentados na Tabela 4.

Tabela 4 – Resultados obtidos no diagnóstico de massa malignas e benignas usando o descritor *HoG* sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	80,23 ± 0,042	89,57 ± 0,069	69,18 ± 0	0,81 ± 0,056	0,79 ± 0,044
	80/20	79,39 ± 0,056	91,52 ± 0,090	65,09 ± 0	0,81 ± 0,082	0,78 ± 0,063
	90/10	72,24 ± 0,066	90,15 ± 0,098	50,94 ± 0	0,79 ± 0,058	0,78 ± 0,048
AdaBoost	70/30	78,84 ± 0,044	88,61 ± 0,072	67,29 ± 0	0,75 ± 0,111	0,71 ± 0,072
	80/20	79,91 ± 0,032	95,68 ± 0,050	61,32 ± 0	0,83 ± 0,053	0,79 ± 0,034
	90/10	72,93 ± 0,053	93,01 ± 0,076	49,05 ± 0	0,77 ± 0,099	0,71 ± 0,056
MVS	70/30	73,65 ± 0,019	78,29 ± 0,029	68,17 ± 0,041	0,74 ± 0,016	0,76 ± 0,013
	80/20	71,86 ± 0,024	80,96 ± 0,059	61,13 ± 0,045	0,71 ± 0,008	0,75 ± 0,022
	90/10	68,10 ± 0,057	82,53 ± 0,097	50,94 ± 0,074	0,63 ± 0,021	0,68 ± 0,029

O classificador *Random Forests* obteve o melhor percentual de acurácia dentre os demais classificadores, alcançando $80,23\%$ quando 70% das amostras são utilizadas para treinamento. Os percentuais de sensibilidade e especificidade foram de 89,57% e 69,18%, respectivamente.

Comparando os resultados dos classificadores utilizados nesta abordagem podemos verificar que o *Random Forests* foi superior em relação as taxas de acurácia e especificidade. Contudo, o classificador *AdaBoost* apresentou uma maior sensibilidade, ou seja, maior diferenciação em casos malignos. No entanto, este classificador obteve o pior percentual de especificidade desta abordagem, indicando que o modelo gerado não obteve uma boa representação das características. Considerando a MVS, podemos concluir que mesmo com resultados inferiores, estes motraram-se os mais estáveis desta abordagem.

4.1.2 Descritor LBP

Nesta abordagem são gerado 256 características por imagem. Os resultados obtidos por esta abordagem mediante a aplicação dos classificadores MVS, *Random Forests* e *AdaBoost* são apresentados na Tabela 5.

Utilizando o classificador *Random Forests*, esta abordagem alcançou os percentuais de 87,18%, 93,12% e 80,18% de acurácia, sensibilidade e especificidade, respectivamente. Este resultado foi obtido com a proporcionalidade de amostra de treino e teste divididas em 80%/20%. A diferença entre o maior e menor percentual de acurácia e especificidade, ultrapassou 9 pontos percentuais, demonstrando assim que a aplicação deste descritor de maneira global juntamente com o classificador *Random Forests* não produz um modelo que diferencie de forma satisfatória massa malignas e benignas.

Tabela 5 – Resultados obtidos no diagnóstico de massa malignas e benignas usando descritor LBP sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	78,09 ± 0,041	90,42 ± 0,065	63,52 ± 0	0,79 ± 0,058	0,77 ± 0,044
	80/20	87,18 ± 0,051	93,12 ± 0,089	80,18 ± 0	0,87 ± 0,064	0,87 ± 0,059
	90/10	82,93 ± 0,089	90,79 ± 0,150	73,58 ± 0	0,84 ± 0,123	0,82 ± 0,113
AdaBoost	70/30	84,26 ± 0,051	87,97 ± 0,091	79,87 ± 0	0,84 ± 0,059	0,84 ± 0,059
	80/20	81,21 ± 0,022	80,48 ± 0,041	82,07 ± 0	0,81 ± 0,021	0,81 ± 0,023
	90/10	80,34 ± 0,044	73,33 ± 0,089	88,67 ± 0	0,81 ± 0,035	0,80 ± 0,052
MVS	70/30	78,61 ± 0,025	76,91 ± 0,038	80,62 ± 0,031	0,82± 0,019	0,79 ± 0,020
	80/20	80,77 ± 0,018	79,52 ± 0,030	82,26 ± 0,020	0,84 ± 0,013	0,81 ± 0,015
	90/10	79,65 ± 0,019	80,63 ± 0,037	78,49 ± 0,027	0,81 ± 0,013	0,81 ± 0,017

Nesta abordagem, ao comparar os classificadores, observa-se que o *Random Forests* se mostrou superior aos demais classificadores em todas as taxas. Contudo, bem como na abordagem anterior, todos os classificadores apresentam uma alta variabilidade entre os percentuais de sensibilidade e especificidade, indicando assim que as características

produzidas por ambas as abordagens não são eficientes para a diferenciação de padrões de malignidade e benignidade em imagens mamográficas.

4.2 Abordagens com Bag of Features

O modelo de representação *Bag of Features* é aplicado como forma de representação em todas as abordagens apresentadas a seguir.

4.2.1 Detector SIFT e Descritor HOG com Bag of Features

A abordagem de detecção e descrição de características locais usando SIFT apresentada nas Seções 3.3.1 e 3.3.2 gera uma média de 173 pontos de interesse por imagem, onde para cada ponto são obtidas 128 características.

A Tabela 6 ilustra os resultados obtidos pela aplicação dos classificadores MVS, *Random Forests* e *AdaBoost* utilizando os parâmetros anteriormente citados no início desse capítulo.

Tabela 6 – Resultados obtidos no diagnóstico de massa malignas e benignas usando o detector SIFT e descritor HoG com Bag of Features sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	96,94 ± 0,017	100 ± 0	93,33 ± 0,040	0,97 ± 0,016	0,96 ± 0,019
	80/20	98,09 ± 0,024	100 ± 0	95,84 ± 0,054	0,98 ± 0,014	0,98 ± 0,027
	90/10	98,10 ± 0,014	98,41 ± 0	97,73 ± 0,031	0,98 ± 0,014	0,98 ± 0,015
AdaBoost	70/30	96,59 ± 0,022	100 ± 0	92,57 ± 0,052	0,96 ± 0,020	9,96 ± 0,025
	80/20	97,83 ± 0,025	100 ± 0	95,28 ± 0,056	0,97 ± 0,024	0,97 ± 0,028
	90/10	95,00 ± 0,023	100 ± 0	89,05 ± 0,054	0,95 ± 0,020	0,94 ± 0,026
MVS	70/30	94,63 ± 0,015	99,57 ± 0,002	88,80 ± 0,036	0,99 ± 0,003	0,93 ± 0,017
	80/20	94,28 ± 0,013	100 ± 0	87,54 ± 0,031	0,99 ± 0,005	0,92 ± 0,023
	90/10	94,31 ± 0,016	100 ± 0	87,54 ± 0,039	1 ± 0	0,93 ± 0,019

O melhor resultado geral foi obtido pela abordagem que utiliza o classificador *Random Forests* alcançando a acurácia de 98,10% usando 90% das amostras para treinamento e o restante para teste. Avaliando os resultados do classificador, observa-se que os percentuais de acurácia em cada configuração de treino/teste possuem uma variação por volta de até 1 ponto percentual considerando a menor e maior taxa. Os resultados obtidos pela sensibilidade e especificidade na configuração do classificador que obteve o maior acerto são de 98,41% e 97,73%, demostrando que a abordagem possui um alto índice de confiança no diagnóstico de casos malignos devido a elevada taxa de sensibilidade.

Embora os classificadores MVS e *AdaBoost* não tenham obtido resultados superiores ao *Random Forests*, em algumas proporções os mesmos mantiveram-se bem próximos, tornando seus resultados satisfatórios. O resultado de sensibilidade e especificidade, para os três classificadores, nesta abordagem, ilustra um alto poder de discriminação entre padrões,

sobretudo em casos de massas malignas, onde este está vinculado ao elevado índice de sensibilidade. O baixo desvio padrão apresentado pelas taxas de acurácia, sensibilidade e especificidade, apresentada em ambos os classificadores, demonstra que estas tentem a estar próximas da média, demonstrando a estabilidade da abordagem aplicada.

4.2.2 Detector SIFT e Descritor LBP com Bag of Features

Esta abordagem objetiva a aplicação do método de descrição LBP nos pontos gerados pelo detector SIFT, tendo em visa a análise local da textura. Tal como a abordagem anterior, que utiliza o detector SIFT, são gerados uma média de 173 pontos de interesse por imagem, entretanto o descritor LBP produz 256 características por imagem.

A Tabela 7 ilustra os resultados obtidos pela aplicação dos classificadores MVS, *Random Forests* e *AdaBoost*.

Tabela 7 – Resultados obtidos no diagnóstico de massa malignas e benignas usando o detector SIFT e descritor LBP com Bag of Features sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	97,00 ± 0,021	99,46 ± 0	94,08 ± 0,048	0,97 ± 0,020	0,96 ± 0,023
	80/20	97,14 ± 0,017	99,2 ± 0	94,71 ± 0,0400	0,97 ± 0,017	0,97 ± 0,019
	90/10	99,13 ± 0,012	100 ± 0	98,11 ± 0,027	0,99 ± 0,012	0,99 ± 0,012
AdaBoost	70/30	97,46 ± 0,022	100 ± 0	94,46 ± 0,050	0,97 ± 0,020	0,97 ± 0,024
	80/20	94,71 ± 0,011	100 ± 0	88,49 ± 0,027	0,95 ± 0,010	0,94 ± 0,012
	90/10	99,65 ± 0,004	100 ± 0	99,24 ± 0,010	0,99 ± 0,004	0,99 ± 0,004
MVS	70/30	93,77 ± 0,015	99,46 ± 0	87,04 ± 0,035	0,99 ± 0	0,92 ± 0,017
	80/20	94,63 ± 0,015	100 ± 0	88,30 ± 0,035	1 ± 0	0,93 ± 0,17
	90/10	93,79 ± 0,026	100 ± 0	86,41 ± 0,057	1 ± 0	0,92 ± 0,032

O melhor resultado para esta abordagens foi obtido com o classificador *AdaBoost*, tendo acerto total, quando 90% das amostras são utilizadas para treinamento, chegando à 99,65%. Analisando os resultados das demais configurações de treino e teste, observa-se que os percentuais de acurácia possuem uma diferença quase 5 pontos percentuais entre a maior e menor acurácia obtida. Os resultados obtidos pela sensibilidade e especificidade evidenciam a eficiência das características LBP, alcançando taxas de 100% sensibilidade e 99,24% especificidade na configuração 70%/30% respectivamente. Verifica-se também que os resultados das medidas de precisão e F1-Score e seus respectivos desvios padrão, demonstram a eficiência e estabilidade desta abordagem na diferenciação de padrões, sobretudo nos casos malignos.

Avaliando os resultados obtidos entre os três classificadores, observa-se uma elevação significativa em todas as taxas, à medida que mais amostras são utilizadas para treinamento. Embora os demais classificadores não tenham sido superiores ao *AdaBoost*, seus resultados mantiveram-se próximos, mostrando-se promissores. Comparando com os resultados de acurácia, sensibilidade e especificidade da abordagem anterior, Seção 4.2.1,

podemos observar que esta abordagem obteve resultados significativamente superiores. Uma suposição é que o descritor LBP por realizar a análise de características de textura locais, possui maior complexidade e sensibilidade em extrair padrões que diferenciem massas malignas e benignas.

4.2.3 Detector ORB e Descritor BRIEF com Bag of Features

A abordagem utilizando o detector ORB e descritor BRIEF juntamente com o modelo *Bag of Features* apresentados na Seção 3.3.1 gera uma média de aproximadamente 266 pontos de interesse por imagem, onde para cada ponto são obtidas 256 características provenientes do descritor BRIEF.

A Tabela 8 apresenta estes resultados obtidos através da aplicação dos classificadores MVS, *Random Forests* e *AdaBoost*.

Tabela 8 – Resultados obtidos no diagnóstico de massa malignas e benignas usando o detector ORB e descritor BRIEF com Bag of Features sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	88,12 ± 0,041	90,31 ± 0,075	85,53 ± 0	0,88 ± 0,045	0,88 ± 0,047
	80/20	86,83 ± 0,041	90,08 ± 0,074	83,01 ± 0	0,87 ± 0,047	0,86 ± 0,047
	90/10	84,48 ± 0,034	95,23 ± 0,056	71,69 ± 0	0,85 ± 0,049	0,84 ± 0,037
AdaBoost	70/30	86,45 ± 0,039	90,42 ± 0,069	81,76 ± 0	0,86 ± 0,045	0,86 ± 0,044
	80/20	85,71 ± 0,052	91,20 ± 0,040	79,24 ± 0	0,86 ± 0,064	0,85 ± 0,060
	90/10	82,41 ± 0,074	91,42 ± 0,123	71,69 ± 0	0,85 ± 0,049	0,84 ± 0,037
MVS	70/30	79,76 ± 0,011	76,38 ± 0,022	83,77 ± 0,009	0,84 ± 0,007	0,80 ± 0,010
	80/20	78,87 ± 0,007	77,44 ± 0,022	80,56 ± 0,024	0,82 ± 0,012	0,82 ± 0,006
	90/10	75,17 ± 0,031	78,09 ± 0,069	71,69 ± 0,018	0,76 ± 0,006	0,77 ± 0,029

O classificador *Random Forests* obteve a maior acurácia chegando a 88,12% com 70% das amostras para treinamento. A diferença entre o menor e o maior percentual não ultrapassou 4 pontos percentuais. Os percentuais de 90,31% e 85,53% de sensibilidade e especificidade respectivamente, demonstram que esta abordagem utilizando o classificador *Random Forests* possui uma considerável capacidade de distinção de padrões, sobretudo em casos malignos.

Nota-se que todos os melhores resultados foram obtidos com configuração de 70%/30%. Observa-se também que os classificadores *AdaBoost* e MVS não foram superiores ao classificador *Random Forests* nesta abordagem. Contudo, mesmo inferiores seus resultados mostram-se satisfatórios.

Comparando os resultados desta abordagem com as demais abordagens já apresentadas (Seções 4.2.1 e 4.2.2), podemos perceber a elevada diferença entre suas taxas, chegando a quase 10 pontos percentuais em alguns casos. Acredita-se que mesmo o detector ORB identificando uma maior quantidade média de pontos por imagem, ainda sim o detector

SIFT através das diferenças de gaussianas identifica pontos nos quais as características locais estão melhores representadas.

4.2.4 Detector SURF e Descritor Wavelet de Haar com Bag of Features

Este detector identifica uma média de 154 pontos de interesse por ROI. Segundo os conceitos apresentados na Seção 3.3.2, o descritor Wavelet de Haar produz um total de 64 características por imagem que são representadas em 200 palavras visuais.

A Tabela 9 apresenta os resultados obtidos por esta abordagem aplicando os classificadores MVS, *Random Forests* e *AdaBoost*.

Tabela 9 – Resultados obtidos no diagnóstico de massa malignas e benignas usando o detector SURF e descritor *Haar Wavelet* com Bag of Features sobre as amostras do DDSM

Classificador	Proporção	Acurácia	Sensibilidade	Especificidade	Precisão	F1-Score
Random Forests	70/30	88,11 ± 0,025	92,76 ± 0,043	82,38 ± 0	0,88 ± 0,029	0,87 ± 0,026
	80/20	85,62 ± 0,036	95,84 ± 0,060	73,58 ± 0	0,87 ± 0,052	0,85 ± 0,040
	90/10	78,10 ± 0,033	93,01 ± 0,052	60,37 ± 0	0,80 ± 0,054	0,77 ± 0,035
AdaBoost	70/30	86,39 ± 0,026	91,38 ± 0,046	80,50 ± 0	0,86 ± 0,031	0,86 ± 0,028
	80/20	84,32 ± 0,031	91,84 ± 0,054	75,47 ± 0	0,84 ± 0,042	0,84 ± 0,034
	90/10	77,06 ± 0,046	95,87 ± 0,068	54,71 ± 0	0,81 ± 0,084	0,75 ± 0,049
MVS	70/30	81,21 ± 0,009	80,95 ± 0,014	81,50 ± 0,025	0,84 ± 0,006	0,82 ± 0,014
	80/20	79,39 ± 0,026	83,68 ± 0,034	74,33 ± 0,020	0,79 ± 0,015	0,81 ± 0,020
	90/10	69,48 ± 0,050	81,90 ± 0,078	54,71 ± 0	0,68 ± 0,017	0,74 ± 0,036

O classificador *Random Forests* obteve como maior percentual de acurácia com valor de $88,01\%$ quando utilizados 70% das amostras para treinamento e 30% para teste. A diferença entre o maior e menor percentual de acurácia alcançou aproximadamente 10 pontos percentuais. Com relação a sensibilidade e especificidade seus percentuais máximos foram de 92,76% e 82,38% demonstrando um satisfatório poder de discriminação sobretudo em imagens malignas.

Avaliando os resultados dos classificadores nesta abordagem podemos verificar que o *Random Forests* foi superior em relação a todas as taxas máximas. Contudo, mesmo os resultados inferiores, os demais classificadores apresentaram percentuais próximos do melhor e portanto são satisfatórios. Verifica-se também que para esta abordagem, a melhor proporção foi utilizando a configuração de 70%/30%.

Esta abordagem obteve resultado inferior quando comparada as demais já apresentadas (Seções 4.2.1, 4.2.2 e 4.2.3). Acredita-se que o detector SURF, por utilizar uma estrutura de aproximação simplificada (*box filter*), não identificou pontos locais característicos para a diferenciação dos padrões de malignidade e benignidade.

4.3 Discussão

Esta seção tem por objetivo discutir os principais resultados obtidos durante os testes com as 6 abordagens utilizadas. Um resumo dos melhores resultados (segundo acerto final) obtidos para cada abordagem é apresentada na Tabela 10.

Tabela 10 – Resumo dos melhores resultados obtidos pela metodologia estudada no trabalho.

Abordagem	Acurácia			Sensibilidade			Especificidade		
	RF	AdaBoost	MVS	RF	AdaBoost	MVS	RF	AdaBoost	MVS
Com Bag of Features									
SIFT - HoG	98,10%	97,83%	94,63%	98,41%	100%	99,57%	97,73%	95,28%	88,80%
SIFT - LBP	99,13%	99,65%	95,51%	100%	100%	100%	98,11%	99,24%	90,18%
ORB - BRIEF	88,13%	86,45%	79,76%	90,31%	90,42%	76,38%	85,53%	81,76%	83,77%
SURF - Haar Wavelet	88,11%	86,39%	81,21%	92,76%	91,38%	80,95%	82,38%	80,50%	81,50%
Sem Bag of Features									
HoG	80,23%	79,91%	73,65%	89,57%	95,68%	78,29%	69,18%	61,32%	68,17%
LBP	87,18%	84,26%	80,77%	93,12%	87,97%	79,52%	80,18%	79,87%	82,26%

Analisando os resultados obtidos pelas abordagens que não utilizam o modelo BOF, verifica-se que quando aplicadas ao classificador *Random Forests*, o melhor resultado de acurácia obtida foi de 87,18% usando o Descritor LBP. Um resultado próximo foi alcançado ao utilizar o Descritor HOG atingindo 80,23%. Considerando os resultados em termos de sensibilidade, o Descritor HOG obteve o maior percentual com 95,68%. No entanto estas duas abordagens obtiveram os piores resultados de especificidade da metodologia, concluindo um baixo poder de diferenciação nos casos benignos usando estes descritores.

As abordagens ORB-BRIEF e SURF-Wavelet de Haar, que utilizam o modelo BOF, atingiram acurácias próximas quando comparadas entre os três classificadores, sendo a maior, 88,13%, obtida com o classificador *Random Forests* utilizando ORB e BRIEF como detectores e descritores.

A melhor acurácia encontrada foi de 99,65% utilizando a abordagem SIFT-LBP seguida pela abordagem SIFT-HOG, atingindo 99,10%. As melhores taxas de acerto final são alcanças pelas abordagens que utilizam o Detector SIFT.

Os resultados de sensibilidade obtidos pelas abordagens ORB-BRIEF e SURF-Wavelet de Haar evidenciam um alto poder de descrição de massa malignas, no entanto existe uma perda considerável quando utilizado para descrever massas benignas. O mesmo pode ser dito das abordagens que não utilizam características locais (Detectores) e o modelo BOF.

Os resultados de acurácia utilizando o SIFT-LBP e SIFT-HOG foram de maneira geral bastante estáveis, se for levado em consideração as elevadas taxas de sensibilidade, especificidade e acerto encontrado durante a utilização dos mesmos. Constata-se então, que estes foram os melhores resultados encontrados pela metodologia para o diagnóstico de malignidade e benignidade em imagens de mamografia.

Ainda analisando estas abordagens, verifica-se que a primeira obteve melhores resultados tanto de especificidade quanto de sensibilidade se comparados entre si. Este resultado remete ao fato de que o descritor LBP consegue descrever melhor a textura local tanto em regiões de massas malignas, quanto em regiões de massas benignas.

4.4 Comparação com outros trabalhos

A comparação de resultados produzidos por este trabalho com os apresentados na Seção 1.1 não é simples, pois conforme discutidos anteriormente, os trabalhos apresentam diferentes metodologias, bases de imagens e quantidade de amostras utilizadas assim como diferentes abordagens de avaliação da metodologia. Contudo, algumas conclusões podem ser extraídas destas comparações. Os resultados da metodologia proposta juntamente com alguns trabalhos anteriormente apresentados, são listados na Tabela 11.

Tabela 11 – Principais trabalhos relacionados ao diagnóstico de câncer de mama

Trabalho	Base	ROIs	Acurácia	Sensibilidade	Especificidade	Az
(CAMPOS; SILVA; BARROS, 2005)	MIAS	100	97,3%	96%	100%	-
(MARTINS et al., 2006)	mini-MIAS	119	86,84%	90%	71,42%	-
(BRAZ JUNIOR et al., 2009)	DDSM	584	87,8%	87,18%	94,87%	0,89
(MOAYEDI et al., 2010)	MIAS	90	96.6%	95,8%	99%	-
(DHAHBI; BARHOUMI; ZAGROUBA, 2015)	mini-MIAS	252	91,27%	-	-	-
(DON et al., 2012)	DDSM	316	96,47%	-	-	-
(HUSSAIN et al., 2014)	DDSM	512	85.53%	-	-	0.87
(BEURA; MAJHI; DASH, 2015)	DDSM	250	97,4%	100%	94,7	-
(ROCHA et al., 2016)	DDSM	1155	88,31%	85%	91,89%	0,88
(GÖRGEL; SERTBAS; UCAN, 2013)	I.U	60	93,59%	97%	91%	-
(AYED; MASMOUDI; MASMOUDI, 2014)	DDSM	100	98,9%	-	-	0.98
(WAJID; HUSSAIN, 2015)	INbreast	396	99,0%	-	-	-
(AREVALO et al., 2015)	BCDR	736	-	-	-	0,86
(ABDEL-ZAHER; ELDEIB, 2016)	WBCD	690	99,68%	100%	99,47%	-
(LIU; TANG, 2014)	DDSM	826	99,68%			0.9615
(ROUHI et al., 2015)	DDSM	263	96,47%	96,86%	95,94	0,95
(ABBAS; QAISAR, 2016)	DDSM	600	91%	92%	84,2%	0,91
(KAUR, 2016)	MIAS	30	96,66%	-	-	-
Metodologia Proposta	DDSM	1155	99,65%	100%	99,24%	0,99

Primeiramente nota-se que o desempenho da metodologia proposta neste trabalho, no geral, é superior ou igualável aos melhores resultados encontrados na literatura. A abordagem aplicada em (ABDEL-ZAHER; ELDEIB, 2016) obteve resultado ligeiramente superior aplicando a técnica de aprendizagem profunda *Deep Belief Networks*. Contudo o número de amostras aplicadas neste trabalho é quase 2 vezes menor ao empregado nos experimentos desta dissertação, extraídas da base privada WBCD o que inviabiliza a replicação dos trabalhos para a correta comparação. Em comparação a abordagem (AB-

BAS; QAISAR, 2016), que também utiliza características locais, esta metodologia obteve resultados significativamente superiores. Podemos afirmar, avaliando comparativamente os trabalhos na área que a metodologia proposta nesta dissertação demonstrou robustez no diagnóstico de câncer de mama mesmo em situações de elevadas quantidades de amostras.

5 Conclusão

Os padrões de malignidade e benignidade presentes nas massas podem possuir características semelhantes, tornando difícil, porém necessária sua diferenciação. Este trabalho apresentou a viabilidade da utilização de características locais e técnicas de aprendizado de máquina para a diferenciação dos padrões de malignidade e benignidade de massas em imagens mamografias.

A metodologia realizou experimentos com 6 abordagens de extração de características utilizando os detectores SIFT, SURF, ORB e descritores HOG, LBP, BRIEF, Haar Wavelet e comparou os resultados obtidos na etapa de classificação utilizando 3 combinações diferentes de distribuição de amostras para treinamento e teste assim como três técnicas de aprendizado, *Random Forests*, *AdaBoost* e MVS.

As técnicas de extração de características locais obtiveram resultados promissores quando aplicadas em conjunto com os classificadores *Random Forests*, *AdaBoost* e MVS. Os melhores resultados para todas as abordagens foram encontrados quando utilizaram o *Random Forests* e *AdaBoost* como classificadores.

De maneira geral, os resultados obtidos pela abordagem que utiliza o SIFT como detector e LBP como descritor (SIFT-LBP) obteve os melhores resultados de sensibilidade, especificidade e acurácia. A utilização deste detector e descritor em conjunto proporciona a análise de textura local, através do LBP, de modo a permitir o diagnóstico eficiente do câncer de mama. Estes resultados mostraram-se promissores quando comparados as demais pesquisas no estado da arte.

A utilização da base pública DDSM neste trabalho, que tem como característica a alta heterogeneidade, proporcionou comprovar a robustez da metodologia. Portanto pode-se concluir que os resultados apresentados por este trabalho evidenciam o poder das características locais no diagnóstico do câncer de mama.

Desta forma, espera-se que novas abordagens surjam baseadas neste tipo de característica e por conseguinte neste trabalho, de modo que sejam sanadas algumas de suas limitações. Durante o processo de execução deste trabalho, surgiram possibilidade de trabalhos que não puderam ser inclusos e que poderão ser trabalhados na forma de estudos futuros.

Dentre os trabalhos futuros estão:

- Utilizar outro método de representação de características e criação de vocabulário, como o *Vector of Locally Aggregated Descriptors* (VLAD). A diferença desse método consiste na capacidade de codificar não só as semelhanças totais das palavras visuais,

mas também todas as semelhanças parte a parte;

- Investigar técnicas automáticas de estimação do tamanho do vocabulário;
- Investigar outras técnicas de agrupamento para formação de dicionário como Quality Threshold e MeanShift;
- Investigar e adaptar outras variações do LBP para a extração de características de textura local, como LBP circular, *Completed Local Binary Pattern* (CLBP), *Multi-block* LBP, entre outros;
- Investigar técnicas de aprendizagem profunda como forma de substituição dos classificadores aplicados neste trabalho, de maneira a avaliar o desempenho dessas novas técnicas na tarefa de reconhecimento de padrões de massa malignas e benignas em imagens de mamografia

Referências

ABBAS, Q.; QAISAR. Deepcad: A computer-aided diagnosis system for mammographic masses using deep invariant features. *Computers*, v. 5, n. 4, p. 28, 2016. ISSN 2073-431X. Citado 4 vezes nas páginas 18, 19, 72 e 73.

ABDEL-ZAHER, A. M.; ELDEIB, A. M. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, Elsevier Ltd, v. 46, p. 139–144, 2016. ISSN 09574174. Citado 3 vezes nas páginas 18, 19 e 72.

ACR - American College of Radiology. *Breast Imaging Reporting and Data System*. Silver Spring, MD, EUA: American College of Radiology press, 1998. Citado na página 53.

ACS - American Cancer Society. *What is breast cancer?* 2016. Acessado em: 13/12/2016. Disponível em: <http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-what-is-breast-cancer>. Citado 2 vezes nas páginas 7 e 22.

ADAMI, H.-O. *Textbook of cancer epidemiology*. Washinton DC,USA: Oxford University Press, USA, 2008. Citado na página 24.

ALAHI, A.; ORTIZ, R.; VANDERGHEYNST, P. Freak: Fast retina keypoint. In: IEEE. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. Providence, Ilha de Rhode, EUA, 2012. p. 510–517. Citado na página 14.

ANTANI, S.; KASTURI, R.; JAIN, R. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern recognition*, Elsevier, v. 35, n. 4, p. 945–965, 2002. Citado na página 14.

ANTONIOU, A. et al. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *The American Journal of Human Genetics*, Elsevier, v. 72, n. 5, p. 1117–1130, 2003. Citado na página 24.

AREVALO, J. et al. Convolutional neural networks for mammography mass lesion classification. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, v. 2015-November, p. 797–800, 2015. ISSN 1557170X. Citado 3 vezes nas páginas 17, 19 e 72.

AVILA, S. *Extended bag-of-words formalism for image classification*. Tese (Doutorado) — Université Pierre et Marie Curie-Paris VI, 2013. Citado 5 vezes nas páginas 8, 14, 20, 40 e 41.

AVILA, S. et al. Bossa: Extended bow formalism for image classification. In: IEEE. *2011 18th IEEE International Conference on Image Processing*. Meeting Center, Bruxelles, Bélgica, 2011. p. 2909–2912. Citado na página 28.

AYED, N. G. B.; MASMOUDI, A. D.; MASMOUDI, D. S. A New Automated CAD System for Classification of Malignant and Benign Lesions. *Asian Journal of Information Technology*, v. 13, n. 9, p. 477–484, 2014. Citado 3 vezes nas páginas 17, 19 e 72.

BAY, H. et al. Speeded-up robust features (surf). *Computer vision and image understanding*, Elsevier, v. 110, n. 3, p. 346–359, 2008. Citado 7 vezes nas páginas 7, 14, 20, 29, 33, 34 e 36.

BEURA, S.; MAJHI, B.; DASH, R. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. *Neurocomputing*, Elsevier, v. 154, p. 1–14, 2015. ISSN 18728286. Citado 3 vezes nas páginas 16, 19 e 72.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: ACM. *Proceedings of the fifth annual workshop on Computational learning theory*. Pittsburgh,USA, 1992. p. 144–152. Citado na página 49.

BRAZ JUNIOR, G. et al. Classification of breast tissues using Moran ' s index and Geary ' s coefficient as texture signatures and SVM. *Computers in Biology and Medicine*, Elsevier, v. 39, n. 12, p. 1063–1072, 2009. ISSN 0010-4825. Citado 3 vezes nas páginas 16, 19 e 72.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Citado na página 47.

BREIMAN, L. et al. *Classification and regression trees*. Flórida, EUA: CRC press, 1984. Citado 3 vezes nas páginas 15, 20 e 46.

BRESENHAM, J. A linear algorithm for incremental digital display of circular arcs. *Communications of the ACM*, ACM, v. 20, n. 2, p. 100–106, 1977. Citado na página 36.

BROWN, M.; LOWE, D. G. Invariant features from interest point groups. In: *Electronic Proceedings of the 13th British Machine Vision Conference*. College Park, Washington DC, EUA: [s.n.], 2002. Citado 2 vezes nas páginas 31 e 32.

CALONDER, M. et al. Brief: Binary robust independent elementary features. *Computer Vision–ECCV 2010*, Springer, p. 778–792, 2010. Citado 4 vezes nas páginas 36, 37, 57 e 58.

CAMPOS, L. F. A.; SILVA, A. C.; BARROS, A. K. Diagnosis of Breast Cancer in Digital Mammograms Using Independent Component Analysis and Neural Networks. p. 460–469, 2005. Citado 3 vezes nas páginas 15, 19 e 72.

CHAVES, A. d. C. F. *Extração de Regras Fuzzy para Máquinas de Vetores Suporte (SVM) para Classificação em Múltiplas Classes*. Tese (Doutorado) — PUC-Rio, 2006. Citado na página 50.

CRISTIANINI, N.; TAYLOR, J. An introduction to support vector machine and other kernel-based learning. *Methods*, 2000. Citado na página 49.

DHAHBI, S.; BARHOUMI, W.; ZAGROUBA, E. Breast cancer diagnosis in digitized mammograms using curvelet moments. *Computers in Biology and Medicine*, v. 64, p. 79–90, 2015. ISSN 00104825. Citado 3 vezes nas páginas 16, 19 e 72.

DIETTERICH, T. G.; KONG, E. B. *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Corvallis,Oregon, USA, 1995. Citado na página 47.

DON, S. et al. A new approach for mammogram image classification using fractal properties. *Cybernetics and Information Technologies*, v. 12, n. 2, p. 69–83, 2012. Citado 3 vezes nas páginas 16, 19 e 72.

DOWNEY, T. J. et al. Using the receiver operating characteristic to asses the performance of neural classifiers. In: IEEE. *Neural Networks, 1999. IJCNN'99. International Joint Conference on*. Washington,DC,USA, 1999. v. 5, p. 3642–3646. Citado 2 vezes nas páginas 9 e 63.

DUARTE, J. C. *O algoritmo Boosting at Star e suas aplicações*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, 2009. Citado na página 45.

EVERITT, B. *The Cambridge dictionary of statistics*. Cambridge, United Kingdom: Cambridge University Press, 1998. Citado na página 47.

FALOUTSOS, C.; OARD, D. W. *A survey of information retrieval and filtering methods*. Washington,DC, USA, 1998. Citado na página 39.

FREUND, Y.; SCHAPIRE, R. E. A desicion-theoretic generalization of on-line learning and an application to boosting. In: VITANYI, P. (Ed.). *Computational Learning Theory: Second European Conference, EuroCOLT '95 Barcelona, Spain, March 13–15, 1995 Proceedings*. Berlin, Heidelberg: Springer, 1995. p. 23–37. ISBN 978-3-540-49195-8. Citado 3 vezes nas páginas 15, 20 e 45.

GARNER, S. R. et al. Weka: The waikato environment for knowledge analysis. In: CITESEER. *Proceedings of the New Zealand computer science research students conference*. Christchurch , New Zealand, 1995. p. 57–64. Citado na página 44.

GE - General Electric Company. *Mammograph Discovery NM750b*. 2016. Acessado em: 13/12/2016. Disponível em: <http://www3.gehealthcare.com/en/products/categories/nuclear_medicine/discovery_nm-ct_750b#>. Citado 2 vezes nas páginas 7 e 25.

GIGER, M. L. Computer-aided diagnosis of breast lesions in medical images. *Computing in Science & Engineering*, AIP Publishing, v. 2, n. 5, p. 39–45, 2000. Citado na página 14.

GONZALEZ, R.; WOODS, R. *Digital Image Processing*. Boston, EUA: Addison-Wesley, 1992. (Addison-Wesley world student series). ISBN 9780201508031. Citado 2 vezes nas páginas 53 e 54.

GONZALEZ, R. C.; WOODS, R. E. *Processamento digital de imagens. tradução: Cristina yamagami e leonardo piamonte*. São Paulo, Brasil: Pearson Prentice Hall, São Paulo, 2010. Citado na página 43.

GOOL, L. V.; MOONS, T.; UNGUREANU, D. Affine/photometric invariants for planar intensity patterns. In: *Computer Vision—ECCV'96*. Cambridge,United Kingdom: Springer, 1996. p. 642–651. Citado na página 29.

GÖRGEL, P.; SERTBAS, A.; UCAN, O. N. Mammographical mass detection and classification using Local Seed Region Growing-Spherical Wavelet Transform (LSRG-SWT) hybrid scheme. *Computers in Biology and Medicine*, Elsevier, v. 43, n. 6, p. 765–774, 2013. ISSN 00104825. Citado 3 vezes nas páginas 17, 19 e 72.

HAMMING, R. W. Error detecting and error correcting codes. *Bell System technical journal*, Wiley Online Library, v. 29, n. 2, p. 147–160, 1950. Citado na página 41.

HARALICK, R. M.; SHANMUGAM, K. et al. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, Ieee, n. 6, p. 610–621, 1973. Citado na página 16.

HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: CITESEER. *Alvey vision conference*. Manchester, England, 1988. v. 15, p. 50. Citado 4 vezes nas páginas 28, 29, 32 e 36.

HARRIS, Z. S. Distributional structure. *Word*, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. Citado na página 40.

HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, JSTOR, p. 100–108, 1979. Citado na página 40.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning 2nd edition*. New York, USA: Springer, 2009. Citado na página 48.

HAYKIN, S. S. Redes neurais artificiais: princípio e prática. *2ª Edição, Bookman*, São Paulo, Brasíl, 2000. Citado na página 51.

HAYKIN, S. S. *Neural networks: a comprehensive foundation*. Beijing Shi, China: Tsinghua University Press, 2001. Citado na página 43.

HE, D.-C.; WANG, L. Texture unit, texture spectrum, and texture analysis. *IEEE transactions on Geoscience and Remote Sensing*, IEEE, v. 28, n. 4, p. 509–512, 1990. Citado na página 39.

HEATH, M. et al. Current status of the digital database for screening mammography. In: *Digital mammography*. Netherlands: Springer, 1998. p. 457–460. Citado 5 vezes nas páginas 7, 15, 26, 27 e 52.

HO, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 20, n. 8, p. 832–844, 1998. Citado na página 47.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. *Artificial intelligence review*, Springer, v. 22, n. 2, p. 85–126, 2004. Citado na página 48.

HUSSAIN, M. et al. Effective extraction of gabor features for false positive reduction and mass classification in mammography. *Applied Mathematics and Information Sciences*, v. 8, n. 1 L, p. 397–412, 2014. ISSN 19350090. Citado 3 vezes nas páginas 17, 19 e 72.

IARC - International Agency for Research on Cancer. World cancer fact sheet. *Geneva, Switzerland: World Health Organization*, 2012. Citado na página 13.

INCA - Instituto Nacional do Câncer. *Ações e Programas no Brasil - Controle do Câncer de Mama*. 2014. Acessado em: 05/10/2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/acoes_programas/site/home/nobrasil/programa_controle_cancer_mama/deteccao_precoce>. Citado 2 vezes nas páginas 14 e 24.

INCA - Instituto Nacional do Câncer. *Estimativa 2016: incidência de câncer no Brasil*. 2015. Acessado em: 05/10/2016. Disponível em: <http://www.inca.gov.br/estimativa/2016/>. Citado 2 vezes nas páginas 13 e 23.

INCA - Instituto Nacional do Câncer. *Fatores de Risco Câncer de Mama*. 2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama/fatores_de_risco_1>. Citado na página 14.

INCA - Instituto Nacional do Câncer. *INCA estima que haverá 596.070 novos casos de câncer em 2016/2017*. 2016. Disponível em: <http://www2.inca.gov.br/wps/wcm/connect/agencianoticias/site/home/noticias/2015/estimativa_incidencia_cancer_2016>. Citado 2 vezes nas páginas 13 e 23.

Instituto Nacional do Câncer. Prevenção e controle de câncer. *Revista Brasileira de Cancerologia*, v. 48, n. 3, p. 317–332, 2002. Citado na página 24.

JEMAL, A. et al. Global cancer statistics. *CA: a cancer journal for clinicians*, Wiley Subscription Services, Inc., v. 61, n. 2, p. 69–90, 2011. Citado na página 13.

KAUR, P. Mammogram image nucleus segmentation and classification using convolution neural network classifier. *International Journal of Advance Research, Ideas and Innovations in Technology*, v. 2, n. 5, p. 1–12, 2016. Citado 3 vezes nas páginas 18, 19 e 72.

KLEIN, G.; MURRAY, D. Parallel tracking and mapping for small ar workspaces. In: IEEE. *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*. DoubleTree Hotel,San Jose, CA, USA, 2007. p. 225–234. Citado na página 36.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. *Supervised machine learning: A review of classification techniques*. 2007. Citado na página 43.

KRIG, S. *Global and Regional Features*. Cham: Springer International Publishing, 2016. 75–114 p. ISBN 978-3-319-33762-3. Citado na página 28.

KRISHNA, K.; MURTY, M. N. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 29, n. 3, p. 433–439, 1999. Citado na página 42.

LINDEBERG, T. Feature detection with automatic scale selection. *International journal of computer vision*, Springer, v. 30, n. 2, p. 79–116, 1998. Citado na página 33.

LIU, X.; TANG, J. Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. *IEEE Systems Journal*, 2014. ISSN 19379234. Citado 3 vezes nas páginas 18, 19 e 72.

LOONEY, C. G. *Pattern recognition using neural networks: theory and algorithms for engineers and scientists*. London, England: Oxford University Press, Inc., 1997. Citado 2 vezes nas páginas 42 e 43.

LOWE, D. G. Object recognition from local scale-invariant features. In: IEEE. *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Kerkyra, Grécia, 1999. v. 2, p. 1150–1157. Citado 3 vezes nas páginas 29, 35 e 37.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, Springer, v. 60, n. 2, p. 91–110, 2004. Citado 11 vezes nas páginas 7, 14, 20, 29, 30, 31, 32, 33, 34, 35 e 55.

MARTINS, L. D. O. et al. Classification of Normal , Benign and Malignant Tissues Using Co-occurrence Matrix and Bayesian Neural Network in Mammographic Images. 2006. Citado 3 vezes nas páginas 15, 19 e 72.

MC - Medicina Celular. *Neoplasia*. 2013. Acessado em: 13/12/2016. Disponível em: <http://www.medicinacelular.com.br/aulasdownload/aula13/13c_neoplas.mansia.htm>. Citado na página 23.

MIKOLAJCZYK, K.; SCHMID, C. Indexing based on scale invariant interest points. In: IEEE. *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. Florida, EUA, 2001. v. 1, p. 525–531. Citado na página 33.

MIKOLAJCZYK, K.; SCHMID, C. Scale & affine invariant interest point detectors. *International journal of computer vision*, Springer, v. 60, n. 1, p. 63–86, 2004. Citado na página 29.

MOAYEDI, F. et al. Contourlet-based mammography mass classification using the SVM family. *Computers in Biology and Medicine*, 2010. ISSN 00104825. Citado 3 vezes nas páginas 16, 19 e 72.

MORAVEC, H. P. *Obstacle avoidance and navigation in the real world by a seeing robot rover*. Virgínia,USA, 1980. Citado na página 28.

MOREELS, P.; PERONA, P. Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision*, Springer, v. 73, n. 3, p. 263–284, 2007. Citado na página 29.

MURPHY, K. P. Naive bayes classifiers. *University of British Columbia*, 2006. Citado na página 45.

NALDI, M. C. et al. Efficiency issues of evolutionary k-means. *Applied Soft Computing*, Elsevier, v. 11, n. 2, p. 1938–1952, 2011. Citado na página 42.

NBCF - Nacional Breast Cancer Foundation. *Mammogram*. 2016. [Acessado em: 13/12/2016]. Disponível em: <http://www.nationalbreastcancer.org/diagnostic-mammogram>. Citado 2 vezes nas páginas 7 e 25.

NOGUEIRA, S. P.; MENDONÇA, J. V. de; PASQUALETTE, H. A. P. Câncer de mama em homens. *Rev. bras. mastologia*, v. 24, n. 4, 2014. Citado na página 23.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, Elsevier, v. 29, n. 1, p. 51–59, 1996. Citado 5 vezes nas páginas 8, 14, 20, 29 e 39.

PEDRINI, H.; SCHWARTZ, W. R. *Análise de imagens digitais: princípios, algoritmos e aplicações*. Connecticut, USA: Thomson Learning, 2008. Citado na página 43.

PENG, X. et al. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, Elsevier, 2016. Citado na página 42.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, New York,EUA, v. 1, n. 1, p. 81–106, 1986. Citado na página 46.

QUINLAN, J. R. *C4. 5: programs for machine learning*. Amsterdam, Holanda: Elsevier, 2014. Citado na página 46.

ROCHA, S. V. d. *Diferenciaçao do Padrao de Malignidade e Benignidade de Massas em Imagens de Mamografias Usando Padroes Locais Binários, Geoestatıstica eÍndice de Diversidade*. Tese (Doutorado) — PhD thesis, Universidade Federal do Maranhão, 2014. Citado na página 53.

ROCHA, S. V. d. et al. Texture analysis of masses malignant in mammograms images using a combined approach of diversity index and local binary patterns distribution. *Expert Systems with Applications*, v. 66, p. 7–19, 2016. Citado 3 vezes nas páginas 17, 19 e 72.

ROSIN, P. L. Measuring corner properties. *Computer Vision and Image Understanding*, Elsevier, v. 73, n. 2, p. 291–307, 1999. Citado na página 37.

ROSTEN, E.; DRUMMOND, T. Machine learning for high-speed corner detection. In: *Computer Vision–ECCV 2006*. Graz, Austria: Springer, 2006. p. 430–443. Citado 3 vezes nas páginas 7, 36 e 37.

ROUHI, R. et al. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, v. 42, n. 3, p. 990–1002, 2015. ISSN 09574174. Citado 3 vezes nas páginas 18, 19 e 72.

RSNA - Radiological Society of North America. *Mammography Screening Intervals May Affect Breast Cancer Prognosis*. 2013. [Acessado em: 13/12/2016]. Disponível em: <http://press.rsna.org/timssnet/media/pressreleases/pr_target.cfm?ID=710>. Citado 2 vezes nas páginas 7 e 26.

RUBLEE, E. et al. Orb: An efficient alternative to sift or surf. In: IEEE. *2011 International conference on computer vision*. Barcelona, Espanha, 2011. p. 2564–2571. Citado 6 vezes nas páginas 14, 20, 29, 36, 37 e 38.

SAFAVIAN, S. R.; LANDGREBE, D. A survey of decision tree classifier methodology. 1990. Citado 2 vezes nas páginas 8 e 47.

SUN, H. et al. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 9, n. 1, p. 109–113, 2012. Citado na página 42.

THORNTON, C. et al. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In: ACM. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, USA, 2013. p. 847–855. Citado na página 44.

TUYTELAARS, T.; MIKOLAJCZYK, K. Local invariant feature detectors: a survey. *Foundations and trends in computer graphics and vision*, Now Publishers Inc., Boston, v. 3, n. 3, p. 177–280, 2008. Citado 2 vezes nas páginas 28 e 29.

USF - University of South Florida. *Digital Mammography Home Page*. 2001. [Acessado em: 15/12/2016]. Disponível em: <http://marathon.csee.usf.edu/Mammography/Database.html>. Citado na página 53.

VAPNIK, V. N. *Statistical learning theory*. New York, USA: Wiley New York, 1998. v. 1. Citado 3 vezes nas páginas 15, 20 e 49.

VENDRAMIN, L.; CAMPELLO, R. J.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, Wiley Online Library, v. 3, n. 4, p. 209–235, 2010. Citado na página 42.

VISWANATHAN, D. G. *Features from Accelerated Segment Test (FAST)*. 2009. Citado na página 37.

WAJID, S. K.; HUSSAIN, A. Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. *Expert Systems with Applications*, Elsevier Ltd, v. 42, n. 20, p. 6990–6999, 2015. ISSN 09574174. Citado 3 vezes nas páginas 17, 19 e 72.

WHO - World Health Organization. *WHO position paper on mammography screening*. Geneva: World Health Organization press, 2014. Citado 2 vezes nas páginas 13 e 14.

WITKIN, A. Scale-space filtering: A new approach to multi-scale description. In: IEEE. *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84*. San Diego, California, USA, 1984. v. 9, p. 150–153. Citado na página 29.

XU, S. et al. Object classification of aerial images with bag-of-visual words. *IEEE Geoscience and Remote Sensing Letters*, IEEE, v. 7, n. 2, p. 366–370, 2010. Citado na página 42.

YANG, J. et al. Evaluating bag-of-visual-words representations in scene classification. In: ACM. *Proceedings of the international workshop on Workshop on multimedia information retrieval*. Bavaria, Germany, 2007. p. 197–206. Citado 2 vezes nas páginas 40 e 42.

YANG, Y.; NEWSAM, S. Bag-of-visual-words and spatial extensions for land-use classification. In: ACM. *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. San Jose, CA, USA, 2010. p. 270–279. Citado na página 42.

ZHOU, L.; ZHOU, Z.; HU, D. Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition*, Elsevier, v. 46, n. 1, p. 424–433, 2013. Citado na página 42.