
Universidade Federal de Pernambuco

Centro de Tecnologia e Geociências

Programa de Pós-Graduação em Engenharia Mecânica

Jessica Hipolito de Vasconcelos

# Investigações sobre métodos de classificação para uso em termografia de mama

Recife

2017

Jessica Hipolito de Vasconcelos

# INVESTIGAÇÕES SOBRE MÉTODOS DE CLASSIFICAÇÃO PARA USO EM TERMOGRAFIA DE MAMA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Mecânica, PPGEM, da Universidade Federal de Pernambuco como parte dos requisitos necessários à obtenção do grau de Mestre em Engenharia Mecânica.

Área de concentração: Processos e Sistemas Térmicos

Orientadora: Profª Drª Rita de Cássia Fernandes de Lima

Coorientador: Prof. Dr. Wellington Pinheiro dos Santos

Recife

2017

Catalogação na fonte
Bibliotecária Maria Luiza de Moura Ferreira, CRB-4 / 1469

V331i Vasconcelos, Jessica Hipolito de.
Investigações sobre métodos de classificação para uso em termografia de mama / Jessica Hipolito de Vasconcelos. - 2017.
92 folhas, il., gráfs., tabs.

Orientadora: Profª Drª Rita de Cássia Fernandes de Lima.
Coorientador: Prof. Dr. Wellington Pinheiro dos Santos.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CTG. Programa de Pós-Graduação em Engenharia Mecânica, 2017.
Inclui Referências.

1. Engenharia Mecânica. 2. Termografia. 3. Câncer de mama. 4. Processamento de imagens. 5. Aprendizado de máquina. I. Lima, Rita de Cássia Fernandes de (Orientadora). II. Santos, Wellington Pinheiro dos (Coorientador). III. Título.

UFPE

621 CDD (22. ed.) BCTG/2017-150

10 de março de 2017.

"INVESTIGAÇÕES SOBRE MÉTODOS DE CLASSIFICAÇÃO PARA USO EM TERMOGRAFIA DE MAMA"

JESSICA HIPOLITO DE VASCONCELOS

ESTA DISSERTAÇÃO FOI JULGADA ADEQUADA PARA OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA MECÂNICA

ÁREA DE CONCENTRAÇÃO: PROCESSOS E SISTEMAS TÉRMICOS

APROVADA EM SUA FORMA FINAL PELO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA MECÂNICA/CTG/EEP/UFPE

_____

Profª Drª RITA DE CÁSSIA FERNANDES DE LIMA
ORIENTADORA/PRESIDENTE

_____

Prof. Dr. WELLINGTON PINHEIRO DOS SANTOS
CO-ORIENTADOR

_____

Prof. Dr. CEZAR HENRIQUE GONZALEZ
COORDENADOR DO PROGRAMA

BANCA EXAMINADORA:

_____

Profª Drª RITA DE CÁSSIA FERNANDES DE LIMA (UFPE)

_____

Prof. Dr. WELLINGTON PINHEIRO DOS SANTOS (UFPE)

_____

Profª Drª ANA LÚCIA BEZERRA CANDEIAS (UFPE)

_____

Profª Drª GUILHERME VILAR (UFRPE)

## AGRADECIMENTOS

Agradeço aos meus pais, meu irmão e a toda minha família que tanto se esforçaram e me ajudaram a chegar até aqui, são minha motivação para que eu lute e tente sempre, sem pensar em desistir.

Agradeço a minha amiga Ariadne que sempre esteve presente, nos bons e nos maus momentos, que se tornou minha família em Recife e me ajudou a superar todas as coisas ruins que aconteceram no último ano. Foi meu apoio em todas as perdas, nunca vou esquecer.

Agradeço a minha orientadora, Professora Rita, que confiou na minha capacidade, me incentivou e me orientou sempre que necessário. Agradeço ao meu coorientador, Professor Wellington, pelas dicas e ideias inovadoras no desenvolvimento do trabalho.

Agradeço as minhas amigas da graduação que levarei para sempre na minha vida. Meninas, obrigada pela amizade, sorrisos compartilhados, conselhos e por todo o incentivo no desenvolvimento deste trabalho.

Agradeço aos colegas do LABTERMO por toda a ajuda, pelos conhecimentos repassados, e por me acolherem tão bem.

Ao Programa de Pós-Graduação em Engenharia Mecânica pela oportunidade e a todos os professores do mestrado que contribuíram de algum modo a minha formação.

Ao CNPq pelo apoio financeiro dado, o qual tornou possível a realização deste trabalho.

*“O fardo é proporcional às forças, como a recompensa será proporcional à resignação e à coragem.”*

Allan Kardec

## RESUMO

Estudos recentes mostram que a termografia vem se mostrando bastante promissora como ferramenta auxiliar na tarefa de detectar o câncer de mama precocemente, o que é fator fundamental para aumentar as chances de cura do paciente. Tumores pequenos podem ser detectados pelos termogramas por causa da elevada atividade metabólica das células cancerígenas, o que leva a um aumento de temperatura no local e que é captado pela termografia. As referidas variações na temperatura assim como as alterações vasculares podem estar entre os primeiros sinais de anormalidade na mama. A técnica é um procedimento de diagnóstico não invasivo, indolor, com ausência de qualquer tipo de contato com o corpo do paciente, além de não emitir qualquer tipo de radiação, sendo então um procedimento confortável e seguro. A termografia é realizada utilizando câmeras de infravermelho sensíveis e um *software* que permite a interpretação de imagens de alta resolução. O presente trabalho tem como objetivo analisar métodos de classificação de imagem digital por infravermelho (IR) de mama e avaliar os resultados obtidos com o objetivo de investigar a viabilidade do uso de imagens IR para a detecção do câncer de mama. Inicialmente, a imagem termográfica é obtida e processada. Em seguida, procede-se à extração de características, que se baseia nas faixas de temperatura obtidas a partir do termograma, determinando-se assim os dados de entrada para o processo de classificação. Foram avaliados sete classificadores e utilizados 233 termogramas de pacientes do Ambulatório de Mastologia do Hospital das Clínicas da Universidade Federal de Pernambuco. Obtiveram-se como resultado, 93,42% de acurácia, 94,73% de sensibilidade e 92,10% de especificidade para a Classe Câncer em uma análise binária (Câncer x Não-Câncer) e para uma análise multiclasse (Maligno, Benigno, Cisto e Normal), 63,46% de acurácia, 80,77% de sensibilidade e 86,54% de especificidade para a Classe Maligno.



Palavras-chave: Termografia. Câncer de Mama. Processamento de imagens. Aprendizado de Máquina.

**ABSTRACT**

Recent studies have stated that the thermography technique has shown to be very promising as an auxiliary tool in the task of early detection of breast cancer, which is a fundamental factor to increase the chances of cure of the patient. Small tumors can be detected by thermograms because of the high metabolic activity of cancer cells, which leads to an increase in temperature on the spot and is captured in thermography. Such variations in temperature as well as vascular changes may be among the first signs of abnormality in the breast. The technique is a non-invasive, painless diagnostic procedure without any type of contact with the patient's body, besides not emitting any type of radiation, and is therefore a comfortable and safe procedure. Thermography is performed using sensitive infrared cameras and software that allows the interpretation of high resolution images. The present work aims to analyze methods of digital image classification of breast infrared images (IR) and to evaluate the results obtained with the purpose of investigating the feasibility of the use of IR images for the detection of breast cancer. Initially, the thermographic images were obtained and processed. Then, the next step is the feature extraction and it is based on the several temperature ranges obtained from the thermogram, determining the input data for the classification process. Seven classifiers were evaluated and used 233 thermograms of patients from the Mastology Outpatient Clinic of the Hospital das Clínicas of the Federal University of Pernambuco. Finally, 93.42% of accuracy, 94.73% of sensitivity and 92.10% of specificity were obtained for the Cancer Class in a binary analysis (Cancer versus Non-cancer) for a multiclass analysis (Malignant, Benign, Cyst and Normal), the obtained results for the Malignant Class were 63.46% of accuracy, 80.77% of sensitivity and 86.54% of specificity.

Keywords: Thermography. Breast Cancer. Image Processing. Machine Learning.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE SÍMBOLOS

| | |
|---|---|
| $FN_i$ | Falso negativo para i-ésima classe |
| $FP_i$ | Falso positivo para i-ésima classe |
| $k$ | Kernel |
| **T** | Matriz de temperaturas |
| $VN_i$ | Verdadeiro negativo para i-ésima classe |
| $VP_i$ | Verdadeiro positivo para i-ésima classe |
| $v_n$ | Vetor sintético |

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| HC | Hospital das Clínicas |
| IBK | Instance Based Learner |
| INCA | Instituto Nacional do Câncer |
| IR | Infravermelho |
| KNN | K-Nearest Neighbours |
| LEMD | Imagem lateral externa da mama direita |
| LEME | Imagem lateral externa da mama esquerda |
| LIMD | Imagem lateral interna da mama direita |
| LIME | Imagem lateral interna da mama esquerda |
| MLP | Multilayer Perceptron |
| ROI | Region of Interest |
| SMO | Sequential Minimal Optimization |
| SVM | Support Vector Machine |
| T1 | Imagem frontal da mama (paciente com as mãos na cintura) |
| T2 | Imagem frontal das mamas (paciente com as mãos levantadas segurando a barra do aparato mecânico) |
| TCLE | Termo de Consentimento Livre e Esclarecido |
| UFPE | Universidade Federal de Pernambuco |
| WEKA | Waikato Enviroment for Knowledge Analysis |

**SUMÁRIO**

# 1 INTRODUÇÃO

## 1.1 MOTIVAÇÃO

O câncer, também chamado de neoplasia, é o nome dado a um grupo de doenças que tem como característica principal a proliferação celular excessiva e descontrolada. Ele ocorre quando uma célula normal do corpo fica fora de controle e passa a se proliferar de maneira desenfreada, ou seja, essa proliferação se perdura mesmo depois que o estímulo inicial que a ocasionou tenha cessado (MALZYNER & CAPONERO, 2013).

De acordo com Tepperwin (2002), no organismo humano existem 200.000 células cancerosas entre as 60 bilhões que o constituem. As células cancerosas tem uma capacidade reprodutora menor que as células normais. Diariamente são regeneradas 100 milhões de células sãs no corpo humano, ou seja, o número de células cancerosas é insignificante e inofensivo a princípio. Porém, quando o sistema imunológico do organismo não as destroem, estas se multiplicam até formarem um tumor.

O tumor é caracterizado por um aumento de volume do tecido, podendo ser de origem benigna ou maligna. O câncer é um tipo de tumor maligno, porém nem todo câncer tem a forma de tumor. Existem casos em que as células ficam dispersas (leucemia), e casos que surgem em forma de ulceração (tumores do estômago) (MALZYNER & CAPONERO, 2013).

Segundo Jorde et al. (2004), evidências mostram que uma em quatro mortes é causada pelo câncer e que mais da metade da população vai ser diagnóstica com câncer em algum momento de suas vidas. Ainda de acordo com o referido autor, esse aumento da frequência no número de câncer se deve ao fato do aumento relativo da população mais velha. Além disso, as causas do câncer podem ser uma mistura de componentes do meio ambiente e alterações genéticas que ocorrem nos tecidos.

O câncer de mama é um dos mais comuns entre mulheres no Brasil e no mundo, sendo o que mais leva a óbito mulheres com idade entre 45-55 anos. Em 2012, o câncer de mama representou 25% do total de casos de câncer do mundo, com aproximadamente 1,7 milhões de novos casos naquele ano. Sendo então a quinta causa de morte por câncer considerando ambos os sexos (522.000 mortes) e a causa mais frequente de morte do sexo

feminino (WHO, 2012). Só no Brasil, são estimados que em 2016 existam 57.960 casos novos, o que representa uma taxa de incidência de 56,2 casos por 100.000 mulheres (INCA, 2015).

Ainda de acordo com o INCA, o câncer de mama pode ser considerado um câncer de bom prognóstico, se diagnosticado precocemente e tratado oportunamente. Porém, o câncer de mama apresenta taxas de mortalidade elevadas no Brasil, isto porque, na maioria dos casos, a doença é diagnosticada em estágios avançados. As altas taxas de mortalidade e diagnóstico tardio devem-se em parte à cobertura irregular dos exames de rastreamento, e falta de conscientização da população e dos profissionais de saúde (BIM et al., 2010).

As principais técnicas utilizadas na detecção do câncer de mama são: a mamografia, a ultrassonografia e a ressonância magnética. Além delas, existem outras tecnologias que têm sido estudadas nas mamas, tais como a tomografia por emissão de pósitrons, a espectroscopia, a tomografia computadorizada, a tomossíntese e a ultrassonografia com contraste (CHALAS & BARROS, 2007). Ainda segundo os autores, a mamografia é a mais importante técnica de imagem para as mamas, tratando-se de um método de rastreamento populacional do câncer de mama em mulheres assintomáticas, e é também a primeira técnica de imagem indicada para avaliar a maioria das alterações clínicas mamárias.

Uma outra técnica tem surgido com o objetivo de auxiliar na detecção de anomalias mamárias: a termografia. Mahmoudzadeh et al. (2015) afirmam que apesar de ser um processo ainda não tão difundido, a hipótese da termografia de mama como auxiliar no diagnóstico do câncer vem sendo proposta há mais de 50 anos. Como dito, a detecção na fase inicial do câncer é fator fundamental para aumentar as chances de cura do paciente. A termografia mostrou eficiência nessa identificação de tumores na fase inicial e também em tecidos densos, como é o caso de pacientes jovens (HEAD et al., 2000). Os tumores pequenos podem ser detectados pelos termogramas por causa da elevada atividade metabólica das células cancerígenas, o que leva a um aumento de temperatura no local e nas vizinhanças e que é captado na termografia (SCHAEFER; NAKASHIMA; ZAVISEK, 2008). Estas variações na temperatura e alterações vasculares podem estar entre os primeiros sinais de anormalidade na mama.

Para o uso de termogramas como auxiliar no diagnóstico do câncer de mama, normalmente são necessárias técnicas de processamento de imagens e visão computacional. A classificação de imagens é comumente a fase final desse processamento e a responsável por

informar o possível diagnóstico da paciente. Levando em consideração esse fato, este trabalho busca investigar métodos de classificação aplicados à termografia de mama, avaliando quem melhor classifica corretamente a presença das anormalidades mamárias.

Esta dissertação faz parte do projeto "**Análise da viabilidade do uso de câmera termográfica como ferramenta auxiliar no diagnóstico de câncer de mama em hospital público localizado em clima tropical**", aprovado pelo Comitê de Ética da Universidade Federal de Pernambuco (UFPE), com registro no Ministério da Saúde CEP/CCS/UFPE Nº 279/05 e em andamento desde novembro de 2005 sob a coordenação da Prof.ª Rita de Cássia Fernandes de Lima.

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

O objetivo geral deste trabalho é analisar métodos de classificação de imagens termográficas de mama e avaliar os resultados obtidos com a finalidade de averiguar a viabilidade do uso da termografia infravermelha para detecção do câncer de mama.

### 1.2.2 Objetivos específicos

- Avaliar sete classificadores de imagens com uma base de dados ampliada em relação à proposta por Queiroz (2016), para a classificação binária (câncer x não-câncer) e multiclasse (maligno, benigno, cisto e normal) por meio da técnica *one-versus-one;*
- Extrair novas características para as amostras avaliadas, que servirão de entrada para os classificadores analisados;
- Avaliar qual o melhor método de classificação e base de dados para o diagnóstico das anomalias de mama.

## 1.3 ORGANIZAÇÃO DO TRABALHO

A presente dissertação está dividida em cinco capítulos. No Capítulo 1 foram introduzidas às informações básicas sobre o câncer, o câncer de mama, a importância para o diagnóstico precoce, e os tipos de exames utilizados para identificar anomalias mamárias.

Além disso, a termografia infravermelha foi também apresentada como ferramenta auxiliar na detecção do câncer de mama.

O Capítulo 2 é composto por uma fundamentação teórica abordando conceitos da termografia, como acontece o processamento de imagens por infravermelho, além de noções sobre a classificação de imagens digitais e sobre alguns classificadores de imagens.

No Capítulo 3 é apresentada a metodologia utilizada. Esta foi realizada seguindo os seguintes passos: aquisição dos termogramas; construção da base de dados; segmentação automática; extração de características e classificação.

No Capítulo 4 são mostrados os resultados obtidos com a segmentação automática, seleção de características e métodos de classificação utilizados. Os resultados são apresentados e discutidos de acordo com o que foi estabelecido nos objetivos.

No quinto e último capítulo são apresentadas as conclusões dos resultados obtidos, levando em consideração os objetivos propostos inicialmente. E como continuidade da presente dissertação, também são sugeridas sugestões para trabalhos futuros.

# 2 FUNDAMENTAÇÃO TEÓRICA

## 2.1 TERMOGRAFIA DE MAMA

A termografia infravermelha é uma técnica a partir da qual é possível visualizar o calor irradiado pelo corpo, por meio do registro da radiação infravermelha emitida e que se encontra em uma parte do espectro magnético na qual a visão humana não é capaz de identificar (MARINS et al., 2015). É uma técnica que vem sendo aplicada nos mais diversos campos da engenharia, na indústria e também na medicina como auxílio no diagnóstico de algumas patologias, como por exemplo, nos distúrbios de mama.

Para a detecção do câncer de mama, existe uma diversidade de técnicas convencionais, como a ressonância magnética, o ultrassom e a mamografia, sendo esta última a mais utilizada. Pretende-se firmar a termografia como uma técnica complementar às mais tradicionais. Segundo Rastghalam e Pourghassem (2016), apesar de a mamografia ser a mais utilizada no diagnóstico do câncer de mama, ela conta com algumas desvantagens, entre elas o fato de ser um exame doloroso, e de submeter a paciente à radiação ionizante, além de ser impróprio para mulheres com implantes, com mamas densas, ou que estejam em processo de terapia hormonal. Nesses casos, o método de ultrassom pode ser utilizado. O ultrassom é uma técnica de imagem não invasiva que é útil para determinar o tipo e a forma da massa ou do cisto, porém, a interpretação desta imagem depende da experiência e do conhecimento do médico.

A termografia de mama vem surgindo como uma alternativa promissora na tarefa de auxiliar na detecção precoce do câncer de mama. Ela é um procedimento de diagnóstico não-invasivo, indolor, com ausência de qualquer tipo de contato com o corpo da paciente, além de não emitir qualquer tipo de radiação, sendo então um procedimento confortável e seguro. Ela é realizada utilizando câmeras de infravermelho sensíveis e um *software* que permite a retirada de dados para o processamento de imagens de alta resolução (NG, 2009). Desse modo, a termografia de mama pode ser usada como uma ferramenta para a detecção precoce, como um instrumento de prevenção, e com a função de detectar alguma anormalidade na mama.

Com o avanço da tecnologia, as câmeras termográficas que antes tinham baixa resolução, foram substituídas por câmeras de infravermelho altamente sensíveis. Além disso,

ocorreram avanços nas técnicas de processamento de imagem, aumentando então o interesse pelo uso da termografia no diagnostico do câncer de mama (TAVAKOL; CHANDRAN; RABBANI, 2013).

A interpretação das imagens é a parte mais importante do exame de termografia de mama, onde as imagens infravermelho (IR) são analisadas e podem ser detectados padrões térmicos atípicos. Porém, de acordo com Francis et al. (2014), essa interpretação de termogramas de mama convencionais é altamente subjetiva, em razão de imagem incompleta das mamas causada por algoritmos de segmentação de imagem ineficientes.

Segundo Mahmoudzadeh et al. (2015), assim como as técnicas tradicionais, termografia de mama não é uma ferramenta de diagnóstico autônoma, para a confirmação da interpretação das imagens em IR normalmente usam-se os demais procedimentos como a mamografia, ultrassom ou ressonância magnética. A termografia pode identificar os sinais anormais que a mamografia e outras tecnologias estruturais não seriam capazes de perceber. Ela é especificamente útil durante os estágios iniciais do crescimento do tumor, que ainda não seriam reconhecidos pela mamografia. A combinação do exame clínico, mamografia e a termografia tem um maior potencial de fornecer um diagnóstico correto e precoce, aumentando a chance de sobrevivência da paciente (NG, 2009).

A análise das imagens termográficas é feita através de um sistema que contém a aquisição de imagens, da extração, da segmentação, do reconhecimento de padrões e, por fim, da interpretação das imagens. A segmentação das imagens é a etapa de suma importância nesse processo, pois a forma, o tamanho e as fronteiras das regiões mais quentes do que o resto da imagem, auxiliam a determinar os recursos que são utilizados para a detecção das anormalidades. Isto contribui para distinguir qual é o tipo de tumor detectado, porém é necessária uma segmentação precisa para melhor eficiência (GOLESTANI; TAVAKOL; NG, 2014). Após a segmentação, ocorre a extração de características, seguida pela classificação da imagem, que mostra o resultado da detecção da lesão suspeita.

Estudos recentes como o de  Rastghalam e Pourghassem (2016), Francis et al. (2014) e Mahmoudzadeh et al. (2015), são exemplos de pesquisas que vêm investigando cada vez mais maneiras de conseguir uma melhor eficiência e maior confiabilidade no uso da termografia como método auxiliar na detecção do câncer de mama.

## 2.2 PROCESSAMENTO DA IMAGEM TERMOGRÁFICA

Uma imagem digital é considerada uma função bidimensional f(x, y) de intensidade luminosa que possui uma matriz correspondente cujos índices de linhas e colunas nos eixos x e y representam um ponto (elemento). Os pontos (elementos) dessa matriz são chamados de *pixels* (GONZALEZ e WOODS, 2008). Logo, no caso das imagens termográficas digitais, a matriz associada é uma matriz de temperaturas **T**(x, y), em que cada elemento da mesma corresponde à temperatura de um ponto.

Um termograma é composto também por um mapa de cores que representa um conjunto de cores específicas que correspondem a um determinado valor de temperatura nas coordenadas x e y (ROBERTO E SOUZA, 2014). As cores que representam um termograma são chamadas de pseudocores e podem variar conforme a paleta de cores utilizada. O uso dessas pseudocores codifica em cores, a temperatura em cada região da mama, onde cada cor representa um valor de temperaturas (SANTOS, 2012; RESMINI et al., 2012).

Sendo assim, como o termografia está associada a uma matriz de temperaturas T, cada *pixel* corresponde a um valor de temperatura dessa matriz **T**, valor este que permanece inalterado independente de qualquer mudança no mapa de pseudocores, o que torna possível o processamento das imagens a partir da matriz de temperaturas. (QUEIROZ, 2016).

### 2.2.1 Segmentação da Região de Interesse

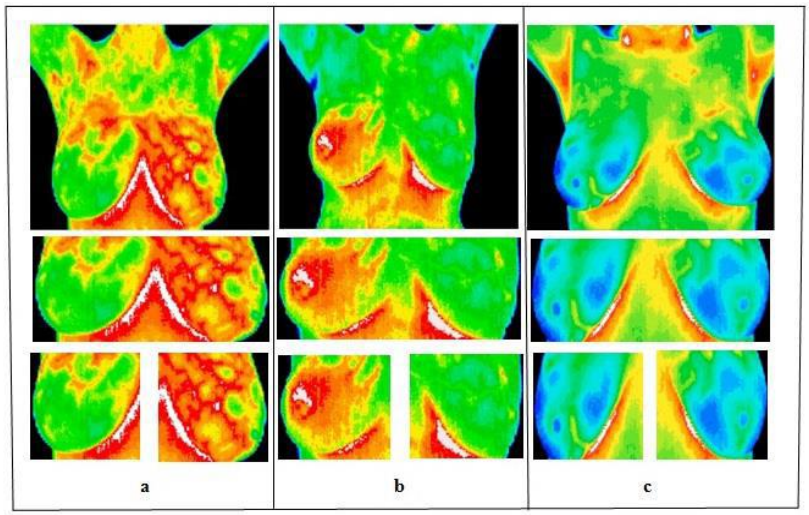
A segmentação da região de interesse (ROI - *Region of Interest*) consiste em separar as áreas da imagem que são importantes para determinada tarefa. Em tal segmentação, a região de maior relevância em um estudo é extraída, destacando-a de todo o resto. Em um termograma, por exemplo, Borchartt (2013) cita que essa segmentação acontece separando toda a região da mama e os gânglios linfáticos de regiões próximas, como as axilas. Já na tomografia computadorizada, Dougherty (2009) afirma que a segmentação é utilizada para a detecção de órgãos, assim como é utilizada também em imagens de ressonância magnética com o objetivo de ressaltar o tecido patológico a partir do tecido normal.

A segmentação de imagens médicas é uma etapa importante no processo de diagnóstico, isto porque nesta fase é fundamental um reconhecimento de certos padrões que são conseguidos pela correta aplicação dessa segmentação a fim de obter o diagnóstico correto. Borchartt (2013) diz que nas imagens termográficas de mama, a segmentação das mamas é desafiadora, isto porque, entre as diversas pacientes, não existe um padrão no

formato da mama, além disso, nessas imagens há um baixo contraste natural, dificultando o processo de extração da região de interesse.

Kamath et al. (2015) propôs uma segmentação automática do peito esquerdo e direito a partir de imagens de termograma de mama usando o Método de Perfil de Projeção. No qual o Método de Perfil de Projeção Horizontal é utilizado para localizar as fronteiras das mamas, detectando a dobra inframamária e as curvas da axila da mama, respectivamente. O método de Perfil de Projeção Vertical é usado para localizar os limites esquerdo e direito da imagem do termograma da mama, que detecta a forma parabólica da mama. A generalização deste método pode ser feita para vários tipos de imagens de termograma de mama, padronizando a altura, fundo e remoção do ruído presente na imagem, a Figura 1 mostra o resultado dessa segmentação.

Figura 1- Resultados da segmentação.

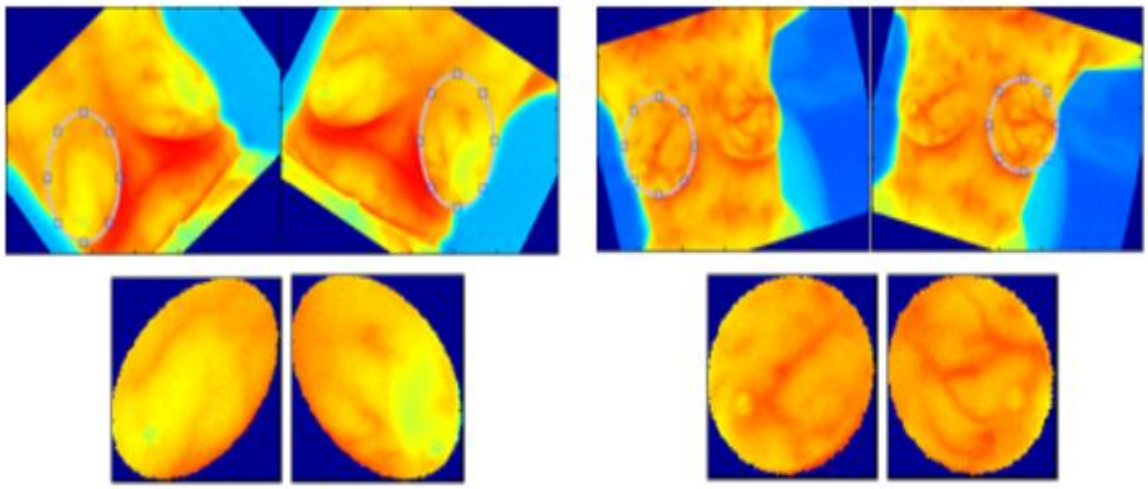



a) Mama com cisto b) Mama com tumor maligno c) Mama sem anormalidade

Fonte: KAMATH et al., 2015.

Com o objetivo de reduzir os erros de segmentação que eventualmente pode surgir numa segmentação automática, devido às variações do corpo de cada paciente a uma possível assimetria das mamas, Araújo (2014) desenvolveu uma segmentação semiautomática. Nesta, o usuário seleciona cada mama a partir de elipses que são geradas manualmente no Matlab, obtendo no final do processo, mostrado na Figura 2, duas imagens de temperaturas independentes, uma para cada mama.

Figura 2 - Processo da Segmentação Semi-automática.



Fonte: ARAÚJO, 2014.

Dessa maneira, a escolha de um método de segmentação que identifique a região de interesse corretamente é de fundamental importância, visto que, no processamento de imagens, a segmentação é a etapa inicial para a extração de informações contidas na imagem, e que vão influenciar fortemente no resultado final que a imagem fornecerá.

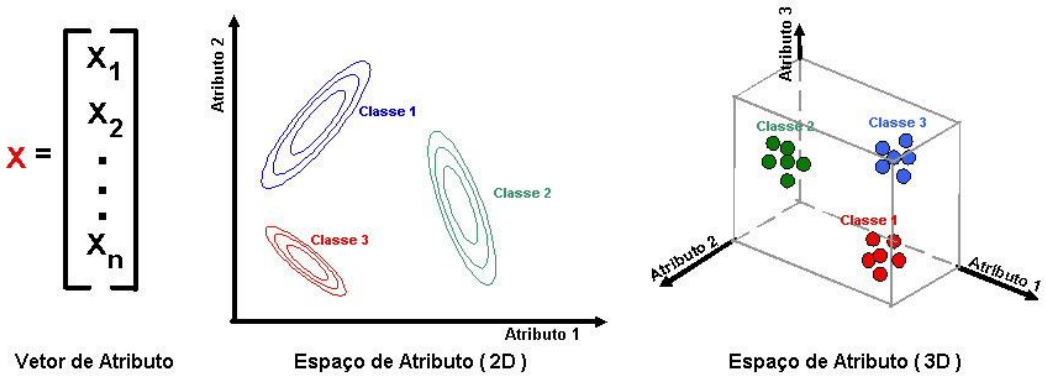
### 2.2.2 Extração de Características

A extração de características pode ser definida como a extração das informações mais relevantes de um conjunto de dados, que podem ser utilizados como entrada em um método de classificação (DUDA; HART; STORK, 2012). O objetivo principal da extração de características é eliminar os dados redundantes, a fim de reduzir sua dimensionalidade.

Apesar da escolha dessas características dependerem do tipo da imagem e da aplicação específica, Dougherty (2009) afirma que elas devem ser robustas, discriminantes, confiáveis e independentes. As robustas geralmente requerem um pré-processamento, porque elas devem ser invariantes à translação, à orientação, à escala e à iluminação. Além disso, devem ser, pelo menos, parcialmente invariantes à presença de ruído e de artefatos. Nas características discriminantes, a gama de valores para objetos em classes diferentes devem ser distintos, e preferencialmente, bem separados e não sobrepostos. As características são consideradas confiáveis se todos os objetos da mesma classe tiverem valores semelhantes. E elas são independentes se não estão correlacionadas. Logo, atendendo a todos esses critérios, segundo o autor citado, as características são apropriadamente extraídas.

De acordo com Oliveira (2009), o vetor de características obtido por meio da extração deve conter atributos relacionados às classes a serem reconhecidas, ou seja, esse vetor é formado por vários atributos que representam um objeto. Esse conjunto de características também pode ser representado no espaço de atributos 2D e 3D (Figura 3).

Figura 3 - Conjunto de características representado no espaço 1D, 2D e 3D.




Fonte: OLIVEIRA, 2009.

Sendo assim, a extração de características é um método que constrói combinações de valores que representam a parte relevante da informação para executar determinada tarefa e que são capazes de contornar problemas, como o uso de uma grande quantidade de memória e de processamento devido a uma alta variação numérica ou da sobrecarga das amostras de treinamento, diminuindo muitas vezes, sua eficiência.

### 2.2.3 Classificação

A classificação de imagens pode ser definida como a aplicação de métodos de reconhecimento de padrões em imagens, com o objetivo de detectar os objetos que a compõem. Fundamentalmente, a classificação de imagem baseia-se em determinar uma função capaz de mapear um conjunto de padrões em um determinado conjunto de classes. Processo de aprendizagem é então, o nome que se dá à etapa de estimação da função de mapeamento (NEGRI; SANT'ANNA; DUTRA, 2013).

De acordo com Campbell (2002), a classificação de imagens é uma etapa de grande importância nos campos de sensoriamento remoto, de análise de imagens e de reconhecimento de padrões. Em alguns casos, a própria classificação pode ser o objeto da análise. A classificação de imagem representa, por conseguinte, uma ferramenta significativa para o estudo de imagens digitais, podendo em certas aplicações, produzir um produto final. Em

outras, pode ser utilizada como um dos vários procedimentos analíticos aplicados em uma imagem, com o objetivo de extrair informações da mesma.

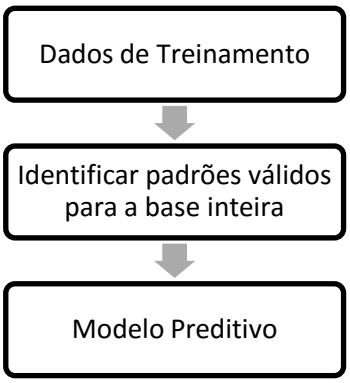
Uma característica que possibilita a distinção dos métodos de classificação de imagens é o tipo de aprendizagem. Dentre os diferentes tipos existentes na literatura, o aprendizado supervisionado e não supervisionado são os comumente utilizados (LIU; MASON, 2016):

- Classificação não supervisionada: é baseada integralmente nas estatísticas da distribuição de dados da imagem e geralmente é chamada de *clustering*. O método é, portanto, objetivo e totalmente orientado pelos dados da imagem. As classes não são conhecidas.

- Classificação supervisionada: é um método que se baseia nas estatísticas dos dados de diferentes objetos com suas classes conhecidas, e que formam o grupo de treinamento do classificador. Uma desvantagem desse tipo de classificação é que ela pode ser mal orientada por informações inadequadas ou imprecisas do grupo de treinamento. Depende do conjunto de treino, que por sua vez depende do conhecimento especialista.

Normalmente, para a descoberta de regras de classificação, os dados são divididos em dois conjuntos: um conjunto para treinamento e um conjunto para teste. A criação de um grupo de treino é essencial no processo de aprendizado da máquina. Tal processo gera uma regra de classificação que é capaz de classificar, com base nas informações conhecidas, qualquer outra amostra utilizada em casos futuros (ARAÚJO, 2014). Já o conjunto de teste é utilizado para validação do classificador, através da verificação das regras descobertas através do conjunto de treinamento. O processo está representado na Figura 4, a seguir.

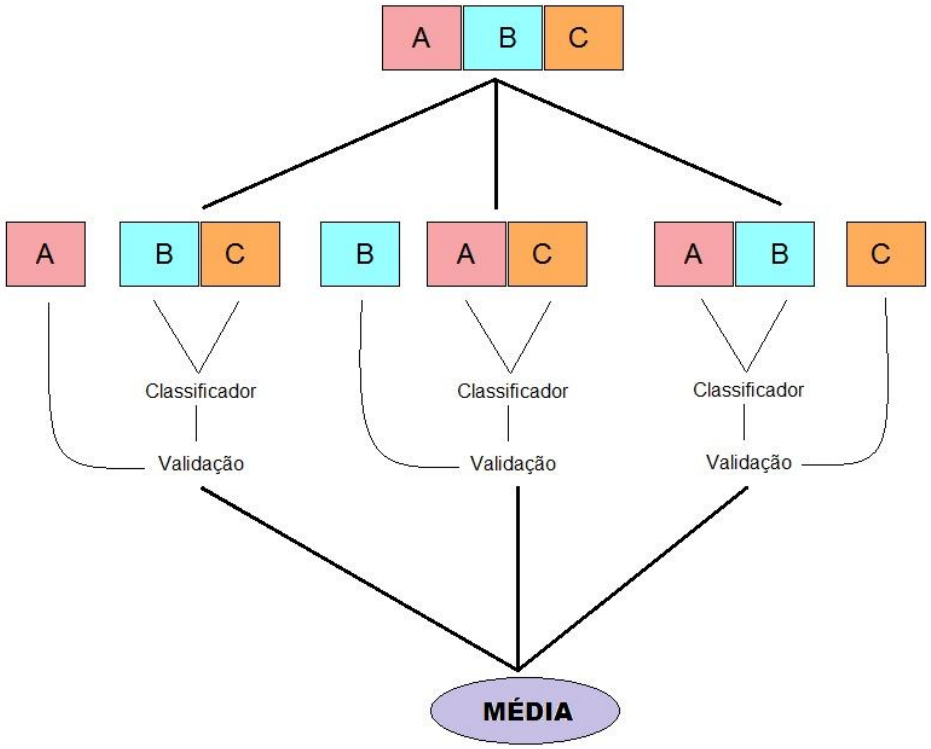
Figura 4 - Processo de aprendizado de máquina.




Outra maneira de avaliar a capacidade de generalização de um classificador é utilizar a técnica de validação cruzada. Os métodos de validação cruzada normalmente utilizados são o *k-fold* e o *leave-one-out*.

De acordo com Dua e Chowriappa (2012), no método da validação cruzada *k-fold*, todos os dados formam o conjunto de treinamento e o mesmo é dividido em *k* subconjuntos. Destes *k* subconjuntos, um subconjunto é guardado para ser utilizado na validação do modelo e os outros (*k-1*) subconjuntos restantes são utilizados no treinamento. Logo, o processo de validação cruzada é repetido *k* vezes, de maneira que todos os *k* subconjuntos sejam utilizados exatamente uma única vez como grupo de teste para validação do modelo. O resultado final é então dado pelo desempenho médio do classificador nos *k* testes. Já a validação cruzada *leave-one-out*, apesar de possuir a mesma definição do método *k-folds*, tem como diferença o número de subconjuntos *k*. O método *leave-one-out* determina que o número de subconjuntos seja igual ao número de exemplos pertencentes ao conjunto de treinamento, como pode ser visto na Figura 5.

Figura 5 - Esquema do método de validação cruzada k-fold.




Tratando-se de imagens médicas, particularmente em termografias de mamas, há uma série de trabalhos que utilizam métodos de classificação para a identificação de anomalias mamárias. Alguns exemplos de trabalhos e seus respectivos métodos de classificação podem ser encontrados na Tabela 1.

Tabela 1 - Estudos que utilizam métodos de classificação para a identificação de anomalias mamárias.

| **Autores** | **Extração de Características** | **Classificação** |
|---|---|---|
| Resmini (2011) | Medidas estatísticas simples, fractais e medidas de geoestatística. | *Support Vector Machine* (SVM) |
| Schaefer, Závišek e Nakashima (2009) | Características de estatística básica, momentos, histograma, matriz de concorrência cruzada, informação mútua e análise de Fourier. | Sistema de classificação baseado em lógica Fuzzy. |
| Borchartt (2013) | Medidas estatísticas, histograma, dimensão fractal de Higuchi e três métodos de geoestatística: Coeficiente de Geary, Índice de Moran e função K de Ripley. | *Support Vector Machine* (SVM). |
| Belfort, Motta e Silva (2014) | Extração de características a partir de funções geoestatísticas: semivariograma, semimadograma, covariograma e correlograma. | *Support Vector Machine* (SVM). |
| Araújo, Lima e Souza (2015) | Extração de características por diferença de medidas intervalares da temperatura. | Discriminante linear, Distância Euclidiana, Distância de Mahalanobis e Distância City-block. |

## 2.3 SENSIBILIDADE, ESPECIFICIDADE E MATRIZ DE CONFUSÃO

A sensibilidade representa a proporção de uma população portadora de um problema em questão com resultados positivos no teste de diagnóstico, ou seja, é a probabilidade do teste dar positivo quando o problema está presente. Já a especificidade é a proporção de uma população em estudo sem o problema determinado e cujos resultados dos testes são negativos, ou seja, é a probabilidade de o teste dar negativo, quando o problema está ausente (PINHEIRO et al., 2015).

Em exames médicos, existem dois tipos de diagnósticos em relação a uma determinada enfermidade, positivo ou negativo. Quando o resultado é positivo e o indivíduo realmente possui a doença, este é chamado de verdadeiro positivo (VP), do contrário é um falso positivo (FP). Do mesmo modo, se o resultado for negativo e o indivíduo não possuir a doença, o resultado é chamado de verdadeiro negativo (VN), caso contrário, é um falso negativo (FN).

Sendo assim, a sensibilidade para a i-ésima classe, pode ser calculada por:

$$Sensibilidade = \frac{VP_i}{VP_i + FN_i} \tag{2.1}$$

E a especificidade para a i-ésima classe, por:

$$Especificidade = \frac{VN_i}{VN_i + FP_i} \tag{2.2}$$

Quanto mais sensível e específico for um teste, maior será a sua capacidade de levar a diagnósticos corretos. Isso quer dizer que, a eficiência de um classificador pode ser calculada através dos índices de sensibilidade e especificidade para a classe estudada.

Partindo do fato de que num diagnóstico existem verdadeiros positivos, verdadeiros negativos, falsos negativos e falsos positivos, e que os verdadeiros positivos e os verdadeiros negativos são os acertos do estudo em questão, e os falsos negativos e falsos positivos são os erros do mesmo. A frequência de erros e acertos colocada numa tabela é conhecida como matriz de confusão. Considerando uma amostra composta por duas classes, onde Classe A é considerada a classe positiva e a Classe B é considerada negativa, a matriz de confusão dessa amostra é mostrada na Figura 6.

Figura 6- Modelo de uma matriz de confusão.

| | | CLASSIFICAÇÃO | | |
|---|---|---|---|---|
| | | **Classe A** | **Classe B** | |
| VERDADEIRO | **Classe A** | VP | FN | VP+FN= nº de amostras da Classe A |
| | **Classe B** | FP | VN | FP+VN= nº de amostras da Classe B |

De acordo com Goldschmidt, Bezerra e Passos (2015), a matriz de confusão de um classificador oferece um detalhamento do desempenho do modelo de classificação utilizado.

Isto acontece ao indicar, para cada classe, o número de classificações corretas em relação ao número de classificações indicadas pelo modelo.

Há quem considere que importa apenas a sensibilidade para cada classe, o que pode ser expresso também pela matriz de confusão normalizada por linha.

## 2.4 COEFICIENTE KAPPA

Uma outra maneira de mensurar o desempenho do processo de classificação é a partir do Coeficiente Kappa. O Coeficiente Kappa pode ser definido como uma medida de associação utilizada para descrever e testar o grau de concordância na classificação, ou seja, sua confiabilidade e precisão (KOTZ & JOHNSON, 1983 apud PERROCA & GAIDZINSKI, 2003).

De acordo com Cohen (1960) o Coeficiente Kappa leva em consideração todos os elementos da matriz de confusão ao invés de apenas aqueles que se situam na diagonal principal da mesma, ou seja, estima a soma da coluna e linha marginais. Ainda segundo o autor, no cálculo do Coeficiente Kappa é assumido que:

- As unidades são independentes;
- As classes são independentes e mutuamente exclusivas;
- O classificador e os pontos de referência operam de forma independente.

De acordo com Viera e Garrett (2005), o cálculo se baseia na diferença entre o valor observado em comparação com o valor esperado, ou seja, avalia o nível de concordância entre dois conjuntos de dados. O coeficiente Kappa é uma medida padronizada para se situar em uma escala de -1 a 1, onde 1 é perfeita concordância, 0 é exatamente o que seria esperado por acaso, e valores negativos sugerem que a concordância encontrada foi menor do aquela esperada por acaso. O Coeficiente Kappa pode ser calculado por (2.3).

$$k = \frac{p_0 - p_e}{1 - p_e} \tag{2.3}$$

onde:

$p_0$ é a taxa de aceitação relativa e

$p_e$ é a taxa hipotética de aceitação.

Estas variáveis podem ser calculadas pelas Equações (2.4) e (2.5).

$$p_0 = \frac{VP+VN}{VP+FN+FP+VN} \tag{2.4}$$

e

$$p_e = \frac{[(VP+FN)(VP+FP)]+[(FP+VN)(FN+VN)]}{(VP+FN+FP+VN)^2} \tag{2.5}$$

A Tabela 2 mostra a interpretação do valor de Kappa obtido.

Tabela 2 - Interpretação do Coeficiente Kappa.

| **Coeficiente Kappa** | **Nível de Concordância** |
|---|---|
| < 0 | Não existe concordância |
| 0,01-0,20 | Concordância mínima |
| 0,21-0,40 | Concordância razoável |
| 0,41-0,60 | Concordância moderada |
| 0,61-0,80 | Concordância substancial |
| 0,81-1,0 | Concordância perfeita |

Fonte: VIERA e GARRETT (2005).

## 2.5 CLASSIFICADORES

O termo classificador pode se referir a um programa de computador que implementa um procedimento específico para a classificação de imagens. Com o passar do tempo, uma série de processos de classicação foram descobertos, tornando possível utilizar um classificador específico para uma determinada tarefa. Ainda não é possível afirmar que um determinado classificador é o mais eficiente para todas as situações, visto que, as características de cada imagem e as aplicações de cada estudo variam (CAMPBELL, 2002).

Borchatt (2013) afirmou que nos trabalhos de Serrano (2010) e Borchartt et al. (2012) mais de 70 classificadores diferentes foram testados em imagens termográficas de mamas a fim de classificar anomalias mamárias nestas imagens. Ainda segundo ele, de todos os classificadores testados, os 10 que tiveram maior eficiência foram: *Support Vector Machine* (SVM), *Naive Bayes Simple* (NBS), *Rotation Forest* (RF), *Logitboost Alternating Decision Tree* (LADTree), *Classification Via Clustering* (CVC), *Naive Bayes* (NB), *Naive*

*Bayes Updateable* (NBU), *Logistic Model Trees* (LMT), *Simple Logistic* (SL), *Instance-based Classifier KStar* (KS).

A partir dos resultados conseguidos nessas pesquisas citadas o classificador*: Support Vector Machine* (SVM) foi escolhido para estudo nessa dissertação. E para critérios de estudo e comparação foram também testados, os seguintes classificadores: *Naive Bayes, Bayes Net, Multilayer Perceptron, Random Forest, Random Tree, IBK (K-nearest neighbours – KNN)* e o *Sequential Minimal Optimization* (SMO).

### 2.5.1 SUPPORT VECTOR MACHINE (SVM)

O *Support Vector Machine* (SVM) é uma técnica computacional de aprendizado para problemas de reconhecimento de padrão. Cortes e Vapnik (1995) desenvolveram o SVM a partir de uma generalização de um algoritmo desenvolvido na Rússia nos anos 60, fundamentado no uso de vetores-suporte como estratégia de aprendizado, baseado na teoria de aprendizado estatístico. Ainda segundo os autores, o SVM é caracterizado como uma técnica de classificação binária capaz de realizar uma separação ótima (com margem máxima) entre dois tipos de classes através de um hiperplano de separação.

O SVM é uma técnica de aprendizado de máquina que se fundamenta nos princípios de minimização do risco estrutural. Ao contrário do risco empírico, o estrutural pode evitar o excesso de ajuste dos dados de convergência do treinamento, aumentando consequentemente a capacidade de generalização (VAPNIK, 1999).

Segundo Ayat, Cheriet e Suen (2005), o treinamento do SVM produz uma função discriminante que minimiza o erro de treinamento, maximizando a margem que separa as classes de dados. A maximização da margem é um processo de regularização implícito que reduz a complexidade classificador. O SVM possui vetores-suporte que são um subconjunto do conjunto de dados de treinamento que definem a fronteira de decisão.

Em resumo, o processo de treinamento do SVM fundamenta-se em treinar um classificador da seguinte maneira: dadas ‘X’ amostras de treinamento $\{x_i, y_i\}$, com $i$ =1, 2, ..., $X$, onde $x_i$ é uma representação vetorial de um conjunto e $y_i \in \{-1,1\}$é sua classe associada, há uma distribuição de probabilidade $P(x,y)$ desconhecida, da qual serão retirados os dados de treinamento. Logo, o processo de treinamento treina o classificador de maneira que aprenda um mapeamento $x \rightarrow y$ através de exemplos de treinamento $\{x_i, y_i\}$ de forma que a máquina

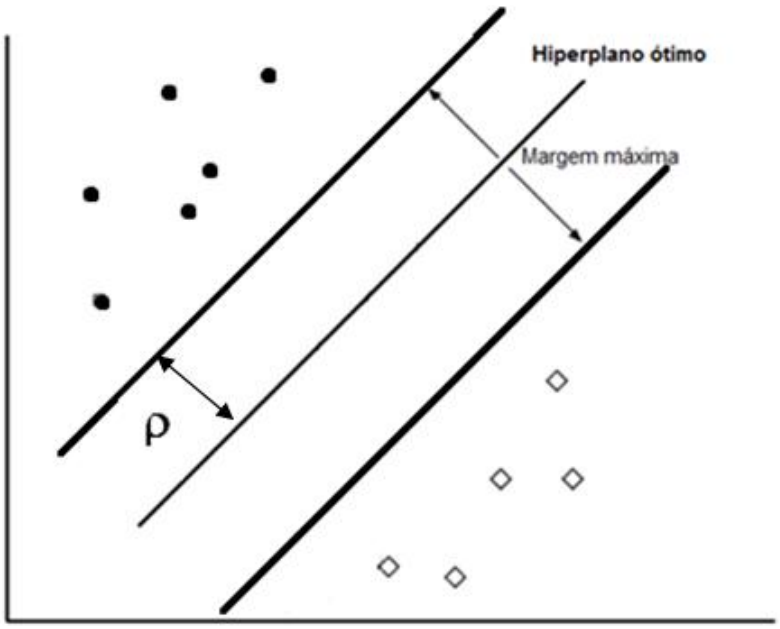
possa classificar um exemplo *(x, y)* desconhecido, mas que siga a mesma distribuição de probabilidade *P* dos dados de treinamento (NASCIMENTO et al., 2009).

SVM Linearmente Separável

O problema de classificação binária, que foi o problema de classificação inicial tratado pelo SVM, trata da classificação de duas classes, e é o problema melhor utilizado para uma questão de generalização. Nesse tipo de classificação e a partir de um conjunto de treinamento linearmente separável é possível se obter um hiperplano ótimo, mostrado na Figura 7.

Este princípio é implementado pelo SVM no qual toda a formulação matemática exposta adiante está baseada em Abe (2010).

Figura 7 - Hiperplano ótimo de separação.




Adaptada de ABE, 2010.

Considere-se o conjunto de treinamento $x_i$ (i= 1, 2, 3, ..., m), num problema de duas classes linearmente separáveis ($c_1$ e $c_2$). Neste problema, $x_i$ é o padrão de entrada para o i-ésimo exemplo associada à resposta desejada $y_i$, onde $y_i$= 1 se $x_i \in c_1$ e $y_i$= -1 se $x_i \in c_2$. A função de decisão pode ser escrita como:

$$D(x) = \boldsymbol{w}^T.\boldsymbol{x} + b \tag{2.6}$$

na qual $\mathbf{w^T.x}$ é o produto escalar entre os vetores $\mathbf{w^T}$ e **x**, onde **x** é um vetor de entrada do conjunto de treinamento e **w** é o vetor de pesos ajustáveis e b é um limiar (termo *bias*).

Logo, o processo de classificação fica:

$$\begin{cases} \boldsymbol{w^T}.\boldsymbol{x_i} + b > 0, x_i \in c_1 \ (y_i = 1) \\ \boldsymbol{w^T}.\boldsymbol{x_i} + b < 0, x_i \in c_2 \ (y_i = -1) \end{cases} \quad (2.7)$$

Uma vez que as amostras são linearmente separáveis, $\boldsymbol{w^T}.\boldsymbol{x_i} + b = 0$ não irá ocorrer. Logo, (2.7) pode ser reescrita como:

$$\begin{cases} \boldsymbol{w^T}.\boldsymbol{x_i} + b > a, x_i \in c_1 \ (y_i = 1) \\ \boldsymbol{w^T}.\boldsymbol{x_i} + b < -a, x_i \in c_2 \ (y_i = -1) \end{cases} \quad (2.8)$$

Sendo *a* uma constante maior que zero e dividindo ambos os termos das inequações (2.8) pela mesma e ajustando **w**, pode-se obter a seguinte inequação:

$$\begin{cases} \boldsymbol{w^T}.\boldsymbol{x_i} + b \quad \geq 1 \ para \ y_i = 1 \\ \qquad\qquad\quad \leq -1 \ para \ y_i = -1 \end{cases} \quad (2.9)$$

Que é equivalente a:

$$y_i(\boldsymbol{w^T}.\boldsymbol{x_i} + b) \geq 1 \quad \text{para i= 1, 2, 3, ... m.} \quad (2.10)$$

O hiperplano definido por:

$$D(x) = \boldsymbol{w^T}.\boldsymbol{x_i} + b = c \text{ para } -1 < c < 1 \quad (2.11)$$

forma o hiperplano de separação entre as duas classes. No momento que c=0, a Equação (2.11) demarca um hiperplano posicionado a meia distância entre os dois hiperplanos extremos (c= 1 e c= -1), onde a distância entre estes é denominada margem e é representada por ρ na Figura 7. Admitindo que os hiperplano D(x)= 1 e D(x)= -1 incluem uma amostra de treinamento, o hiperplano D(x)= 0 representa a melhor separação entre as amostras, ou seja, é o hiperplano ótimo, e tem uma margem máxima de $-1 < c < 1$, quer dizer, a região $-1 \leq D(x) \leq 1$, é a região de generalização da função decisão.

A distância Euclidiana de uma amostra **x** a um hiperplano D (**x**) é dada por $|D(\boldsymbol{x})| \ / \ \|\boldsymbol{w}\|$ é usada para determinar a separação ótima entre os hiperplanos (quando esta distância é máxima). Essa condição é obtida quando se minimiza ||w|| ou equivalentemente minimizando-se:

$$Q(\boldsymbol{w}) = \frac{1}{2}\|w\|^2 = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \quad (2.12)$$

onde deve existir um **w** e um b que satisfaçam (2.10). Isto acontece para assegurar que não ocorram amostras de treinamento na região de separação entre as duas classes (entre as margens).

Para a solução do problema de minimização da Equação (2.12) com a inclusão da restrição (2.10), é normalmente utilizada a técnica dos multiplicadores de Lagrange (α), dado por:

$$Q(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{m} \alpha_i\{y_i(\boldsymbol{w}^T x_i + b) - 1\} \quad (2.13)$$

em que $\alpha = (\alpha_1, \alpha_2, ..., \alpha_m)^T$ e $\alpha_i$ são os multiplicadores de Lagrange não-negativos.

A solução é dada então por um ponto de sela, que é encontrado pela minimização de Q (**w**, b, **α**) com relação a w, maximizando com relação a $\alpha_i$ ($\geq$0), e maximização ou minimização em relação a b. A solução deve satisfazer às condições de Karush-Kuhn-Tucker (ABE, 2010).

$$\frac{\partial Q(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i\, y_i \boldsymbol{x_i} = 0 \quad (2.14)$$

$$\text{logo, } \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i\, y_i \boldsymbol{x_i} \quad (2.15)$$

e,

$$\frac{\partial Q(\boldsymbol{w},b,\boldsymbol{\alpha})}{\partial b} = \sum_{i=1}^{m} \alpha_i\, y_i = 0 \quad (2.16)$$

Com as seguintes condições:

$$\alpha_i\{y_i(\boldsymbol{w}^T.\boldsymbol{x_i} + b) - 1\} = 0 \quad \text{para i= 1, 2, 3, ..., m.} \quad (2.17)$$

$$\alpha_i \geq 0 \quad \text{para i= 1, 2, 3, ..., m.} \quad (2.18)$$

Daí, é possível se obter uma equação expressa apenas em termos de **α**:

$$Q(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x_i}^T \boldsymbol{x_j} \quad (2.19)$$

E então, o problema consiste na maximização em relação a **α**, seguindo as restrições (2.16) e (2.18).

A formulação então descrita é chamada na literatura de problema dual, enquanto que o problema originado se trata de uma forma primal que apresenta maior grau de complexidade

na sua solução. Devido a esse grau de dificuldade, para sua resolução foi então utilizada a Teoria de Langrange para se obter a forma dual de mais simples resolução, ou seja, pode-se então, resolver indiretamente um problema primal a partir da resolução direta do dual (CARVALHO, 2005). De acordo com Lorena e Carvalho (2007), a forma dual é útil também quando se trata de SVM para dados não-lineares, visto que, possui restrições mais simples e permite a representação do problema em termos de produtos internos entre os dados.

Substituindo então, a condição (2.15) em (2.6), a função decisão é dada por:

$$D(x) = \sum_{i \in S} \alpha_i y_i \boldsymbol{x_i}^T \boldsymbol{x} + b \tag{2.20}$$

onde S é o conjunto de índices dos vetores de suporte, ou seja, todas as amostras de treinamento, nos quais $\alpha_i > 0$.

Das condições de Karush-Kuhn-Tucker é possível determinar o termo *bias* como $b = y_i - \mathbf{w}^T\mathbf{x}_i$, em que, uma estimativa ainda mais confiável pode ser obtida utilizando um valor médio com respeito a todos os vetores de suporte dada por:

$$b = \frac{1}{|S|} \sum_{i \in S} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i) \tag{2.21}$$

em que S é o conjunto de todos os vetores de suporte e |S| é o número de vetores de suporte que ocorre no problema.

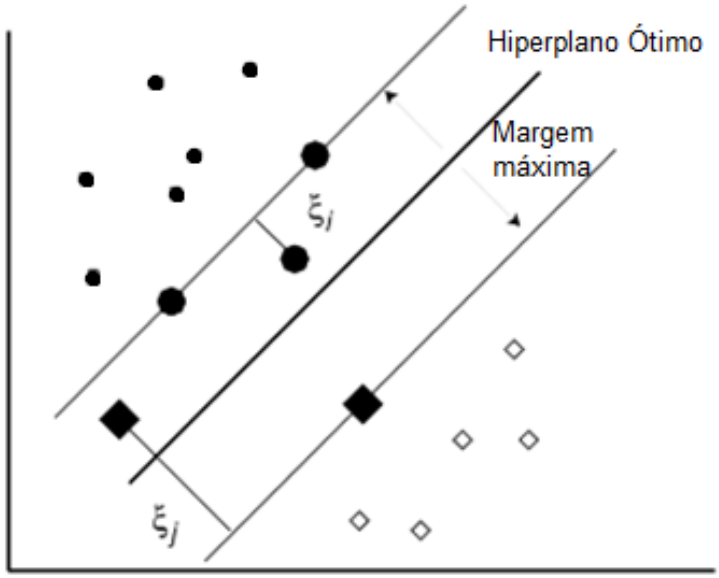
Nas formulações obtidas anteriormente, foi considerado que a amostra de treinamento era linearmente separável. Em situações reais, porém, essa condição é improvável de acontecer e diversos fatores como ruídos e *outliers* (exemplos muito distintos dos demais presentes no conjunto de dados) nos dados, ou a própria natureza do programa, fazem a amostra não ser linearmente separável (LORENA E CARVALHO, 2007).

Quando a amostra não é linearmente separável, não há uma solução possível para o SVM de margens rígidas. Logo, para lidar com conjuntos de treinamento mais gerais, alguns dados são inseridos para violar a restrição da Equação (2.10). Assim, se introduz o conceito de variável de folga (*slack variable*) $\xi_i$ ($\geq 0$). O problema de otimização primal, torna-se então:

$$\alpha_i \{ y_i (\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 + \xi_i \} = 0 \text{ para i= 1, 2, 3, ..., m.} \tag{2.22}$$

o que resulta no SVM com margens suaves. Esta condição permite que alguns pontos fiquem entre as margens, caso $0 \leq \xi_i \leq 1$, ou estejam na região que corresponde a outra classe ($\xi_i > 1$), neste caso, a amostra $x_i$ será classificada erroneamente, como mostra a Figura 8:

Figura 8 - SVM com margens suaves.




Adaptada de ABE (2010).

Levando em consideração o termo $\xi_i$, subtraindo então em relação aos dados de treinamento, a Equação (2.12) se torna:

$$Q(\boldsymbol{w}, b, \xi) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i \tag{2.23}$$

Segundo Abe (2010), C é chamado de parâmetro de margem e determina o equilíbrio entre a maximização da margem e minimização do erro de classificação. Utilizando uma abordagem lagrangeana na Equação (2.23) semelhante ao já exposto, tem-se como resultado um problema também dual e igual ao da Equação (2.20), porém agora sujeito as seguintes restrições:

$$\sum_{i=1}^{m}\alpha_i\, y_i = 0 \text{ e } C \geq \alpha_i \geq 0 \text{ para i= 1, 2, 3, ..., m.} \tag{2.24}$$

Sendo assim, uma condição complementar $(C - \alpha_i)\,\xi_i = 0$ é acrescida às condições de Karush-Kuhn-Tucker, mostradas anteriormente. Desta condição e usando a Equação (2.22), tem-se três diferentes casos para $\alpha_i$ (LORENA; CARVALHO, 2007; ABE, 2010):

- Se $\alpha_i = 0$ e $\xi_i = 0$, a amostra $x_i$ é corretamente classificada.

- Se $0 < \alpha_i < C$, então $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1 + \xi_i = 0$ e $\xi_i = 0$. Logo, $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$ e sendo assim, $\mathbf{x_i}$ é um vetor de suporte que se encontra sobre as margens e é denominados vetor livre.
- Se $\alpha_i = C$, caso $\xi_i > 0$, os pontos são erroneamente classificados. Porém, se $0 < \xi_i \leq 1$, os pontos são corretamente classificados entre as margens. Por outro lado, se $\xi_i = 0$ os pontos ficam sobre as margens.

Como resultado final, tem-se então, uma função de classificação idêntica à mostrada na Equação (2.20), porém, neste caso, as variáveis $\alpha_i$ da função decisão são determinadas pela solução de (2.19), usando as restrições impostas em (2.24).

SVM não-linear

Existe uma diversidade de casos para os quais não é possível, dividir de maneira eficiente, os padrões do conjunto de treinamento através de um hiperplano, mesmo analisando as variáveis livres. De acordo com Takahashi (2012), para estes casos usa-se uma função *kernel* adequada para fazer um mapeamento no domínio do espaço de entrada do conjunto de treinamento para um outro espaço (espaço de características).

Uma função *Kernel k* é uma função que recebe dois pontos $\mathbf{x}_i$ e $\mathbf{x}_j$ do espaço de entrada e calcula o produto escalar $\varphi^T(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ no espaço de características. Sendo o termo $\varphi^T(\mathbf{x}_i) \cdot \varphi(\mathbf{x}_j)$ que representa o produto interno dos vetores $\mathbf{x}_i$ e $\mathbf{x}_j$, o *Kernel* é dado por (HERBRICH, 2001):

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \varphi^T(\boldsymbol{x}_i).\varphi(\boldsymbol{x}_j) \text{ para i, j = 1, 2, 3, .... m.} \quad (2.25)$$

Envolvendo agora o espaço de características, a Equação (2.15) pode ser reescrita da seguinte maneira:

$$\boldsymbol{w} = \sum_{i,j=1}^{m} \alpha_i \, y_i \varphi^T(\boldsymbol{x}_i).\varphi(\boldsymbol{x}_j) \quad (2.26)$$

onde o $\varphi^T(\mathbf{x}_i)$ equivale ao padrão de entrada $x_i$ no i-ésimo exemplo.

O hiperplano ótimo pode então, ser construído no espaço de características utilizando o produto interno $k(\boldsymbol{x}_i \boldsymbol{x}_j)$, se considerar o próprio espaço de características de forma explícita. As funções φ devem então, pertencer a um domínio no qual seja possível o cálculo dos produtos internos, geralmente, se faz o uso do *teorema de Mercer* para satisfazê-las. Este,

diz que os *kernels* devem ser matrizes definidas positivamente em que $k_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, para todo i, j = 1, 2, 3 ..., m, deve ter auto vetores maiores que 0 (TAKAHASHI, 2012).

As funções *Kernel* mais clássicas são mostradas na Tabela 3:.

Tabela 3 - Funções *Kernel* clássicas.

| ***Kernel*** | **Fórmula** |
|---|---|
| Linear | $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i . \boldsymbol{x}_j$ |
| Polinomial | $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (a\, \boldsymbol{x}_i . \boldsymbol{x}_j + b)^d, a > 0$ |
| RBF | $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-a \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|^2)$ |
| Sigmoidal | $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(a\, \boldsymbol{x}_i . \boldsymbol{x}_j + r)$ |

Adaptado de CHAKRABORTY, SARKAR E MAULIK (2016).

SVM Multiclasse

O algoritmo original do SVM é adequado a problemas de classificação binária, isto é, quando há apenas duas classes a serem separadas. O classificador SVM foi modificado para trabalhar com problemas de multiclasses, ou seja, classificação com três ou mais classes (STATNIKOV et al., 2011).

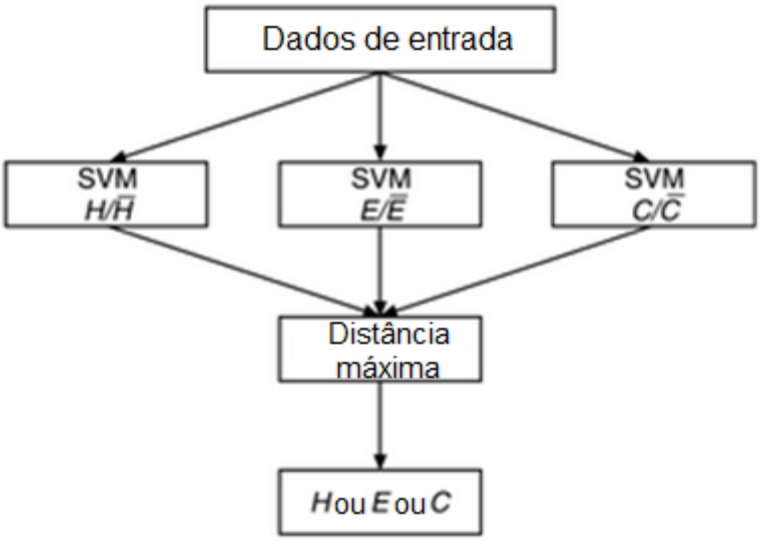
De acordo com Syed (2008), após diversos estudos, pesquisadores estenderam o método binário a um método multiclasse. Os dois principais tipos de classificadores SVM multiclasse são: um-contra-um e o um-contra-todos. Eles são uma combinação de vários classificadores binários em um solucionador multiclasse.

**Um-contra-todos**

Admitindo que existam k classes que se deseja separar. A abordagem um-contra-todos constrói k classificadores SVM binários adequados para separar uma classe das outras. Uma classe é fixada e assumida positiva e o resto k-1 classes são supostas negativas. Durante o teste, cada classificador produz um valor de decisão para os dados de teste e a classe com o maior valor de decisão positiva é considerada como a decisão final. A comparação entre os valores de decisão produzidas por diferentes SVMs ainda é válida porque os parâmetros de formação e o conjunto de dados permanecem os mesmos (MA; GUO, 2014).

Supondo três classes, H, E e C e aplicando a estratégia do um-contra-todos, tem-se consequentemente 3 classificadores SVM, no qual os novos dados estão sujeitos ao teste. O processo de classificação um-contra-todos é mostrado em resumo na Figura 9:

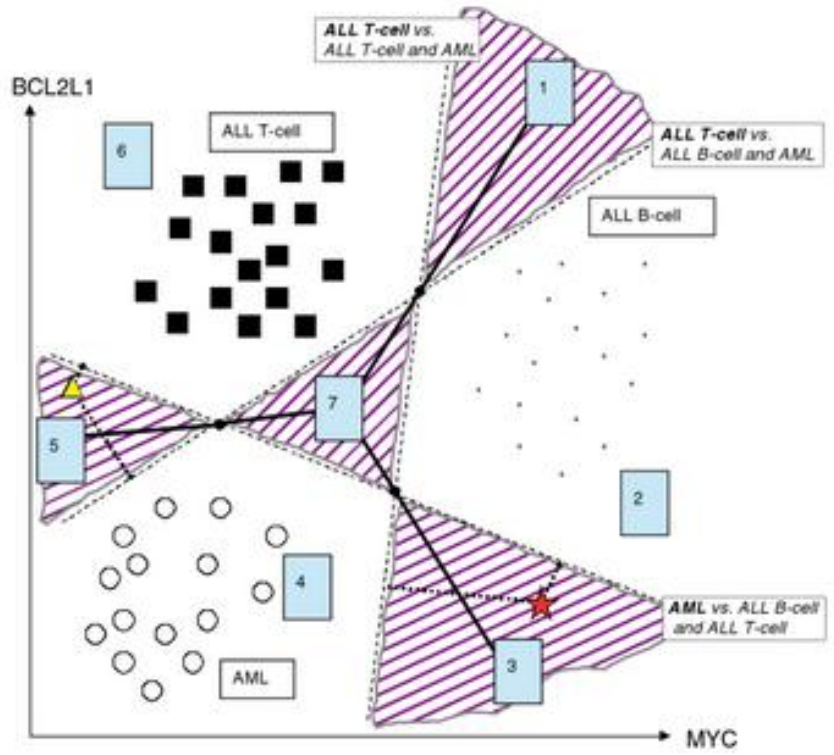
Figura 9 - Processo de classificação um-contra-todos.




Adaptado de SOMAN; LOGANATHAN; AJAY (2011).

Statnikov et al. (2011) aplicaram o classificador SVM um-contra-todos em um problema de diagnóstico com três resultados possíveis (classes): leucemia linfoblástica aguda (ALL T-cell), leucemia de células B (ALL B-cell), e leucemia mielóide aguda (AML). Logo, foram construídas três categorias do classificador SVM um-contra-todos para amostras de sangue de pacientes descritos por níveis de dois genes BC2L1 e MYC. A aplicação de três classificadores SVM binários (para separar cada classe do resto) resultou em três hiperplanos que são mostrados com linhas tracejadas na Figura 10. As regiões sombreadas da figura correspondem às situações quando dois ou nenhum dos classificadores tem votos positivos ao mesmo tempo para a sua classe. A superfície de decisão para o SVM multiclasse um-contra-todos é mostrada com uma linha em negrito sólida na Figura 10.

Figura 10 - Classificação SVM um-contra-todos.




Fonte: STATNIKOV et al. (2011).

Considerando a classificação da nova amostra denotada pela forma triangular, que se situa na região ambígua 5 mostrada na Figura 10. Essa amostra recebe votos positivos de ambos os classificadores AML e ALL de células T. No entanto, a sua distância a partir do hiperplano "ALL T-cell vs ALL B-cell e AML" é maior do que a partir do hiperplano "AML vs ALL B-cell e ALL T-cell". Por isso, esta amostra é classificada como ALL de células T. A Tabela 4 mostra a classe escolhida pelo classificador para cada região.

Tabela 4 - Decisão do classificador (votos positivos).

| Região | AML vs (ALL B-cell e ALL T-cell) | ALL T-cell vs (ALL B-cell e AML) | ALL B-cell vs (ALL T-cell e AML) | Resultado da classe |
|---|---|---|---|---|
| 1 | - | ALL T-cell | ALL B-cell | ? |
| 2 | - | - | ALL B-cell | ALL B-cell |
| 3 | AML | - | ALL B-cell | ? |
| 4 | AML | - | - | AML |
| 5 | AML | ALL T-cell | - | ? |
| 6 | - | ALL T-cell | - | ALL T-cell |
| 7 | - | - | - | ? |

Adaptado de STATNIKOVet al. (2011).

De acordo com Soman, Loganathan e Ajay (2011), um dos principais problemas da abordagem um-contra-todos é o conjunto de treinamento desbalanceado. Considerando que todas as classes têm um tamanho igual de exemplos de treino, a razão de classe assumida positiva para exemplos negativos em cada classificador indivual é 1/ K-1. Neste caso, a simetria original do problema é perdido.

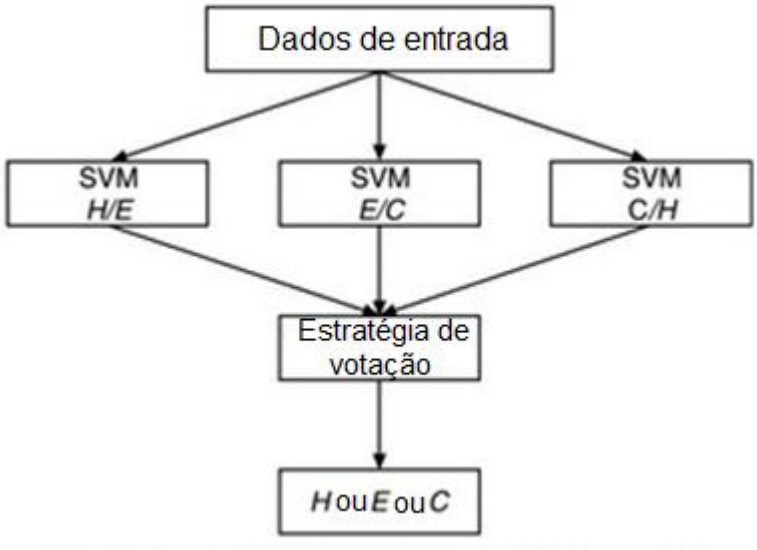
**Um-contra-um**

Outra abordagem clássica para a classificação SVM multi-classe é do um-contra-um ou decomposição em pares. Para k classes, ele avalia todas as possíveis classes emparelhadas e, assim, induz k (k-1) / 2 classificadores binários individuais. Aplicando cada classificador a um exemplo de teste, as classes são votadas e a mais votada é a classe vencedora, que é então classificada (MA; GUO, 2014).

Segundo Kressel (1999) apud Statnikov et al. (2011) um dos benefícios da abordagem multi-classe SVM um-contra-um é que, para cada par de classes resolve-se um problema de classificação binária SVM, que utiliza um número menor de objetos do que o número total de objetos nos dados de treinamento. Isso pode resultar em economia substancial no tempo computacional total em comparação com a abordagem SVM multi-classe um-contra-todos. Além do mais, alguns dos subproblemas binários podem ser separáveis e a abordagem um-contra-um pode acarretar em uma classificação mais eficiente em comparação com a abordagem um-contra-todos.

Dadas três classes H, C e E, a Figura 11 retrata o funcionamento da estratégia um-contra-um:

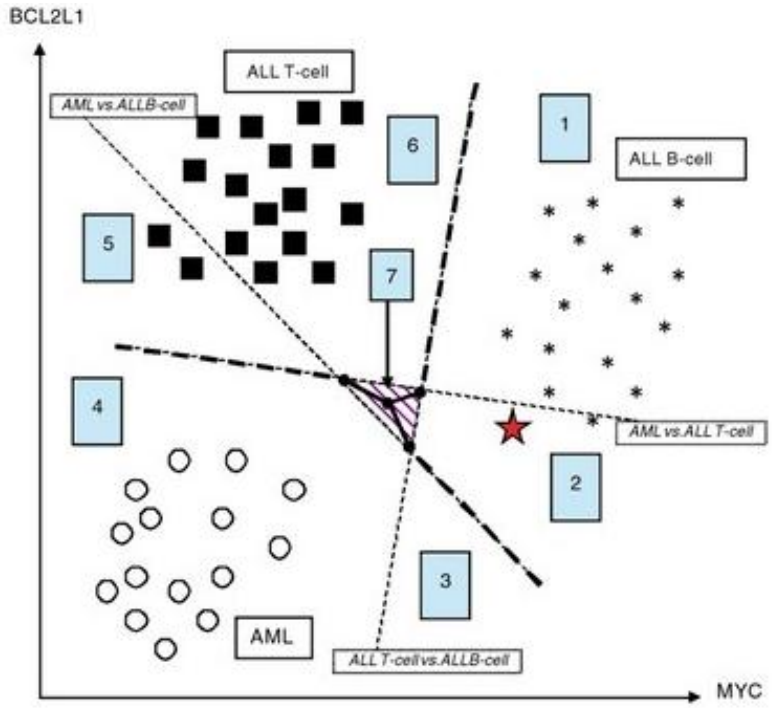
Figura 11 - Processo de classificação um-contra-um.




Fonte: SOMAN; LOGANATHAN; AJAY (2011).

Statnikov et al. (2011) utilizaram a abordagem um-contra-um para a realização do mesmo estudo da classificação de três tipos de leucemia (ALL T-cell vs ALL B-cell vs AML) para amostras de sangue de pacientes descritos por níveis de dois genes BC2L1 e MYC. A aplicação de três classificadores SVM binários resultou em três hiperplanos que mostram linhas tracejadas na Figura 12. A região sombreada da Figura 12 corresponde a situação de empate (quando todos os três classificadores votam em uma classe diferente) e a superfície de decisão da abordagem um-contra-um é mostrada pela linha em negrito.

Figura 12 - Classificação SVM um-contra-um.




Fonte: STATNIKOV et al. (2011).

A Tabela 5 apresenta a classe escolhida pelo classificador SVM um-contra-um para cada região.

Tabela 5 - Decisão do classificador.

| **Região** | **AML vs ALL B-cell** | **ALL T-cell vs ALL B-cell** | **AML vs ALL T-cell** | **Resultado da classe** |
|---|---|---|---|---|
| 1 | ALL B-cell | ALL B-cell | ALL T-cell | ALL B-cell |
| 2 | ALL B-cell | ALL B-cell | AML | ALL B-cell |
| 3 | AML | ALL B-cell | AML | AML |
| 4 | AML | ALL T-cell | AML | AML |
| 5 | AML | ALL T-cell | ALL T-cell | ALL T-cell |
| 6 | ALL B-cell | ALL T-cell | ALL T-cell | ALL T-cell |
| 7 | ALL B-cell | ALL T-cell | AML | - |

Adaptado de STATNIKOV et al., 2011

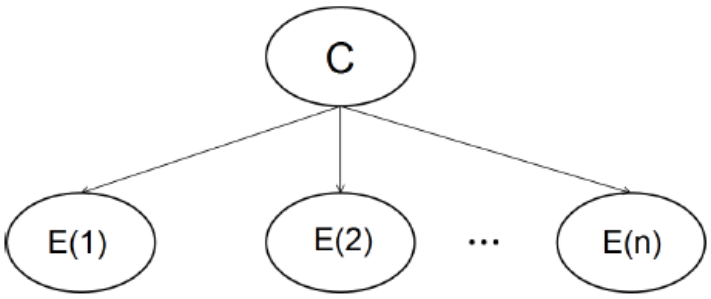
### 2.5.2 Outros classificadores

Outros exemplos de classificadores são mostrados a seguir:

**Bayes Net** – É um classificador que constrói uma rede bayesiana completa e então realiza a busca nessa rede de acordo com um algoritmo de busca qualquer. O principal parâmetro para o algoritmo BayesNet é o tipo de busca realizado. O algoritmo atribui relação entre os valores dos atributos, onde as probabilidades das classes são calculadas sem levar em conta a dependência dos possíveis valores que um atributo pode assumir (JOHN & LANGLEY, 1995).

**Naive Bayes –** É um classificador probabilístico baseado no teorema de Bayes. Os valores estimados são escolhidos com base na análise dos dados de treinamento. Neste classificador, cada uma das características contribui independentemente para a classificação. Os classificadores Navies Bayes são úteis para conjuntos de dados muito grandes. Apesar de sua simplicidade, os classificadores Naive Bayes proporcionam bom resultados para problemas complexos do mundo real. Além disso, eles exigem pequena quantidade de dados de treinamento para prever os parâmetros para a classificação (ZHANG, 2004) (KUMAR & SAHOO, 2012). De acordo com Moura (2016), um exemplo de uma rede bayesiana do tipo Naive Bayes é mostrada na Figura 13.

Figura 13 - Exemplo de rede bayesiana do tipo Naive Bayes




Fonte: MOURA, 2016.

Na figura acima as variáveis de evidência E assumem independência e todas possuem um nó pai C.

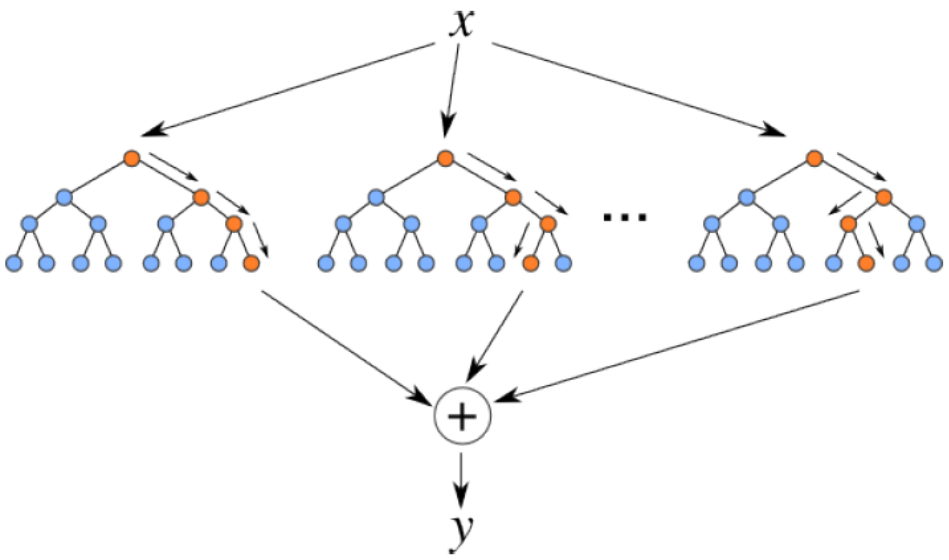
***Multilayer Perceptron* (MLP)** – O MLP é um modelo de rede neural artificial que mapeia conjuntos de dados de entrada para um conjunto de saídas apropriadas. Um MLP é composto por várias camadas de nós em um grafo direcionado, com cada camada completamente conectada à próxima. Exceto para os nós de entrada, cada nó é um neurônio (ou elemento de processamento) com uma função de ativação não-linear. O MLP utiliza uma técnica de aprendizagem supervisionada chamada *backpropagation* ou retropropagação para treinar a rede. O modelo é uma modificação do Perceptron Linear Padrão e pode distinguir os dados que não são linearmente separáveis (KUMAR & SAHOO, 2012) (SIVANANDAM; SUMATHI; DEEPA, 2006). Segundo Affonso et al. (2010), o algoritmo de *backpropagation* é um tipo de aprendizado supervisionado no qual um valor de saída é gerado e o erro é calculado, em seguida, seus valores são retropropagados para entrada, os pesos são ajustados e os valores são novamente calculados, até o conjunto de dados de saídas ser considerado apropriado.

***Sequential Minimal Optimization* (SMO)** – É um algoritmo de otimização mínima sequencial para treinar um classificador de SVM. Essa implementação substitui globalmente todos os valores em falta e transforma atributos nominais em binários. Ele também normaliza todos os atributos por padrão. Então, neste caso, os coeficientes na saída baseiam-se nos dados normalizados, não nos dados originais. Os problemas multi classe são resolvidos utilizando a classificação de pares um-contra-um (*one-versus-one*) (KEERTH et al., 2001). O algoritmo SMO divide o problema quadrático do SVM em vários subproblemas mais simples e estes podem ser resolvidos sem nenhum tipo de armazenamento de dados, reduzindo o uso de memória.

***Random Forest*** - É um classificador que compreende um conjunto de árvores como classificadores. Idealmente, o classificador Random Forest é uma randomização de variáveis aleatórias independentes e identicamente distribuídas de aprendizado fraco. O classificador utiliza um grande número de árvores de decisão individuais, todas treinadas para resolver o mesmo problema. Uma amostra é alocada àquela classe que mais frequentemente ocorre, tal como determinado por árvores individuais. O algoritmo para implementação da Random Forest foi desenvolvido por Leo Breiman e Adele Cutler. O termo veio da floresta de decisão aleatória o qual foi proposto pela primeira vez por Tin Kam Ho de Bell Labs em 1995. O método combina a ideia de "ensacamento" de Breiman e a seleção aleatória de características, introduzidas independentemente por Ho e Amit e Geman afim de construir uma coleção de

árvores de decisão com variação controlada (BREIMAN, 2001) (CHEN, 2007). A Figura 14 mostra a lógica do algoritmo Random Forest.

Figura 14 - Ilustração da lógica do algoritmo Random Forest




Fonte: LORENZETT & TELÖCKEN, 2016.

***Random Tree*** - É um conjunto de preditores de árvores que é chamado de floresta. Este classificador é uma árvore de decisão que considera apenas alguns atributos escolhidos aleatoriamente para cada nó da árvore. A classificação funciona do seguinte modo: o classificador Random Tree toma o vetor de entrada composto pelas características, classifica-o com todas as árvores da floresta e o associa a classe que recebeu a maioria dos "votos" (KALMEGH, 2015).

***Instance Based Learner*** **(IBK) ou** ***K-Nearest Neighbours*** **(KNN)** – O classificador KNN é um classificador baseado em aprendizagem de máquina e na inteligência artificial. O método consiste em algoritmos que melhoram seu desempenho e processamento através da experiência adquirida. Sua classificação é feita por analogia, isto é, a cada novo objeto que se deseja classificar verifica-se entre os dados de treinamento, quais dados mais se assemelham a esse objeto, e então é feita a classificação. Desta forma, não são criados padrões de classificação. Sendo assim, no KNN, essa classificação é realizada em função da menor distância, ou seja, a classe é determinada pelos elementos do conjunto de treinamento que estejam mais próximos do elemento desconhecido (ISHII et al., 2009).

## 2.6 WEKA

O *Waikato Environment for Knowledge Analysis* (WEKA) foi um projeto criado em 1992, que surgiu por meio da consciência da necessidade de algo que permitisse aos pesquisadores um fácil acesso às técnicas de aprendizagem de máquina. O WEKA foi previsto para não só fornecer uma caixa de ferramentas de algoritmos de aprendizagem, como também a permitir aos pesquisadores a implementação de novos algoritmos de acordo com sua necessidade. O acesso livre de usuários ao código-fonte permitiu o desenvolvimento e criação de muitos projetos que incorporam e/ou estenderam o WEKA (HALL, 2009).

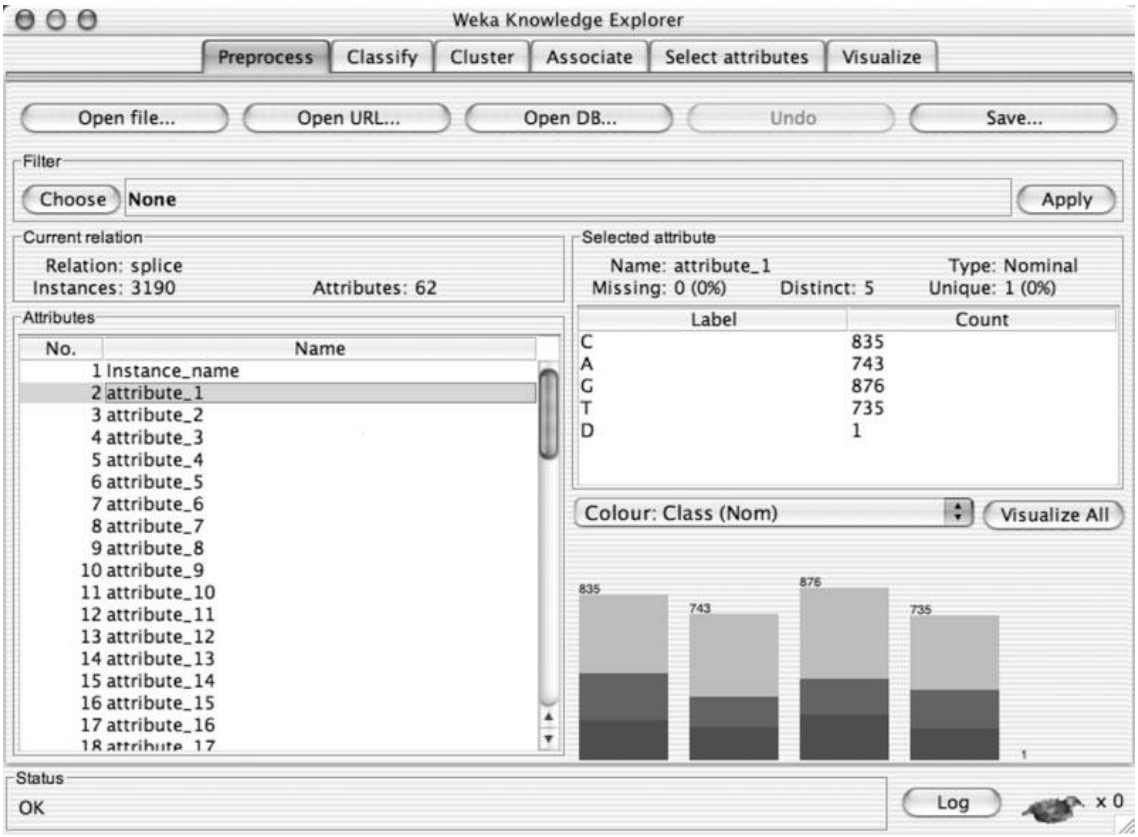
O WEKA tem como objetivo proporcionar para pesquisadores e profissionais, um conjunto abrangente de algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados. Ele permite aos usuários experimentar rapidamente e comparar diferentes métodos de aprendizado de máquina em novos conjuntos de dados. A interface do programa inclui algoritmos para regressão, classificação, agrupamento e seleção de atributos (FRANK et al., 2005). Este *software* trabalha com arquivos do tipo *arff*.

De acordo com Zhong (2011), o WEKA utiliza uma interface gráfica unificada com a tecnologia de aprendizado de máquina padrão, que dispõe de variados métodos de pré-tratamento e pós-processamento. Em tais métodos, diversos algoritmos de estudos diferentes podem ser aplicados ao conjunto de dados com a finalidade de avaliar o resultado correspondente.

### 2.6.1 WEKA Explorer

O WEKA tem como principal interface o Explorer, mostrado na Figura 15.

Figura 15 - Principal interface do WEKA.




Fonte: FRANK et al., 2004.

Ele possui um conjunto de painéis, cada um dos quais pode ser usado para realizar uma determinada tarefa (FRANK et al., 2004):

**Preprocess** - Nesse painel, ocorre a recuperação de dados do arquivo, banco de dados SQL ou URL. Após isso, os dados podem ser pré-processados usando os instrumentos de filtragem do WEKA. Além disso, nesse painel, também é mostrado um histograma da característica selecionada e algumas estatísticas sobre a mesma. Também, pode exibir os histogramas de todas as características numa janela a parte.

**Classify** – Se os dados são de um problema de classificação ou de regressão, os mesmos podem ser processados nesse painel. Este fornece algoritmos para modelos de classificação e regressão de aprendizagem além de ferramentas de avaliação para analisar o resultado do processo de aprendizagem. O WEKA utiliza as seguintes técnicas de aprendizagem de classificação e regressão: Árvore de Decisão, Classificador Bayesiano, SVM (*Support Vector Machine*), Conjunto de Regras, Regressão Logística e Linear, MLP (*MultiLayer Perceptron)* e o KNN (*Nearest Neighbors).*

**Cluster** – A terceira aba do Explorer dá acesso a algoritmos de agrupamento. Estes incluem: k-médias, misturas de distribuições normais com a estimativa da matriz de covariância diagonal, utilizando EM, além de um esquema de agrupamento hierárquico com incrementais heurísticas.

**Associate** – Neste painel, geram-se regras de associação que podem ser usadas para identificar as relações entre os grupos de atributos nos dados.

**Select Attributes –** Este painel dispõe de métodos para a identificação de subconjuntos de atributos que são preditivos no conjunto de dados. Os métodos utilizados são: *best-first search, forward selection, genetic algorithms* e *simple ranking.*

**Visualize** – A última aba mostra uma matriz de gráficos de dispersão para todos os pares de atributos dos dados. Qualquer elemento da matriz pode ser selecionado e ampliado em uma janela separada, onde se pode aumentar o zoom em subconjuntos dos dados e recuperar informações sobre pontos de dados individuais.

# 3 METODOLOGIA

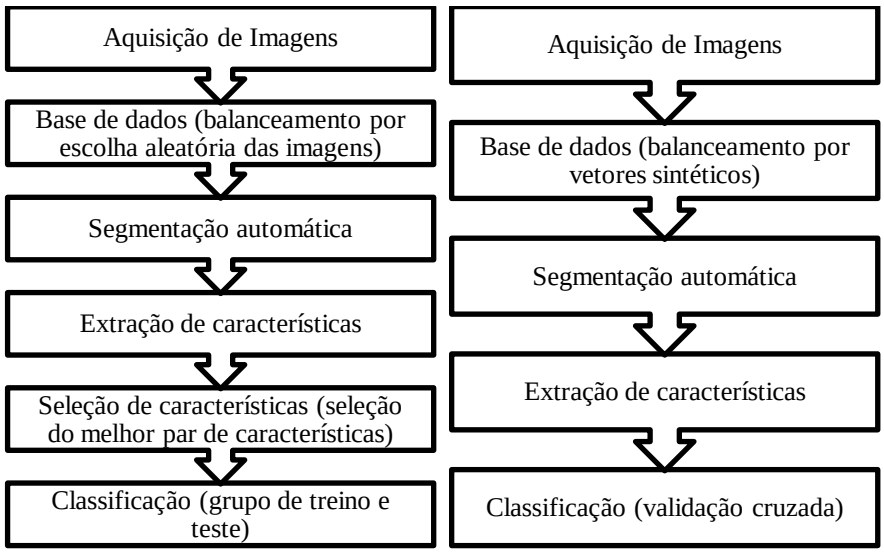
O presente capítulo apresenta as duas metodologias usadas no presente trabalho e descreve as ferramentas utilizadas em cada estágio do processo de classificação. O foco deste projeto é criar uma base de dados que associada a um classificador forneça resultados confiáveis no diagnóstico de anomalias mamárias a partir de termogramas de mamas numa classificação binária (Câncer ou Não-Câncer) e em uma classificação multiclasse (Maligno, Benigno, Cisto e Normal).

Com o intuito de se obter melhores resultados no processo de classificação, foi utilizado o Método SVM, o mesmo utilizado por Dourado Neto (2014) e Queiroz (2016), além de classificadores disponíveis no *software* WEKA. Este *software* também foi utilizado para a seleção dos atributos mais adequados, quando necessário. O WEKA trabalha com arquivos no padrão *arff* para suas tarefas, logo, para cada caso, foram construídos arquivos nesse formato para a posterior análise no *software*.

Para uma análise multiclasse usando o classificador binário SVM, foi utilizada a técnica *one-versus-one*. Tal fato permitiu uma comparação com o trabalho de Queiroz (2016) no qual foi implementada a técnica *one-versus-all*.

A Figura 16 apresenta de forma sucinta os dois tipos de metodologia utilizados na presente dissertação, onde cada etapa será detalhada ao longo do capítulo.

Figura 16 - Metodologias propostas.




(a) Metodologia 1 (b) Metodologia 2

As ferramentas utilizadas em cada etapa da presente dissertação serão discutidas a seguir.

## 3.1 AQUISIÇÃO DOS TERMOGRAMAS

Neste trabalho, foram utilizadas imagens termográficas obtidas com uma câmera de infravermelho FLIR S45 de pacientes do Ambulatório de Mastologia do Hospital das Clínicas (HC) da UFPE entre os anos de 2005 e 2014. Apenas pacientes com diagnóstico concluído através de exames clínicos tradicionais, como a ultrassonografia, a mamografia e a biópsia foram consideradas. Estas pacientes concordaram em participar da pesquisa por meio da assinatura do Termo de Consentimento Livre e Esclarecido (TCLE) que está vinculado ao projeto aprovado pelo Comitê de Ética da Universidade Federal de Pernambuco (UFPE) - Brasil, e registrado no Ministério da Saúde sob CEP/CCS/UFPE N°279/05, em novembro de 2005.

Com o objetivo de minimizar os erros gerados por oscilações de temperaturas no interior da sala ou da temperatura da paciente, foi utilizado o protocolo para exame termográfico de mama que foi proposto por Oliveira (2012), e estabelece uma adequação da sala, do paciente e da aquisição das imagens. No trabalho de Oliveira (2012), o protocolo está detalhadamente descrito. Também é usado um aparato mecânico que foi projetado por Oliveira (2012) e construído na oficina mecânica do Departamento de Engenharia Mecânica da UFPE (DEMEC/UFPE). Este aparato auxilia no posicionamento dos braços da paciente, e é formado por dois trilhos de deslocamento onde fica um carro que serve de suporte para a câmera e seu tripé, como mostra a Figura 17.

Figura 17 - Aparato mecânico.

Fonte: OLIVEIRA, 2012.

Além da temperatura e da umidade relativa da sala, são medidas as distâncias entre paciente e a câmera. A emissividade da pele humana que foi considerada como sendo 0,98 (LAHIRI et al., 2012 apud QUEIROZ, 2016). Estes parâmetros servem como parâmetro de entrada para a câmera termográfica.

A paciente deve esperar cerca de dez minutos sem tocar na mama, para que se estabeleça um equilíbrio térmico com o ambiente. É preparado então, o documento de Termo de Consentimento Livre e Esclarecido (TCLE), como solicitado pelo Ministério da Saúde do Brasil, no qual a paciente assina no final do exame caso concorde em participar da pesquisa. Neste caso, também são feitas cópias da anamnese e dos laudos de outros exames realizados, como ultrassom, mamografia, biópsia ou punção. Estes resultados são comparados com os obtidos pelos classificadores com o objetivo de validá-los. Também é medida a temperatura da paciente com um termômetro clínico.

Além disso, o procedimento de atendimento consiste em instruir que as pacientes passem determinado tempo sem a exposição à luz solar, sem a realização de exercícios físicos, sem a ingestão acima da média de alimentos e bebidas e sem se banharem a determinado tempo antes do atendimento. Tais precauções antes dos exames termográficos

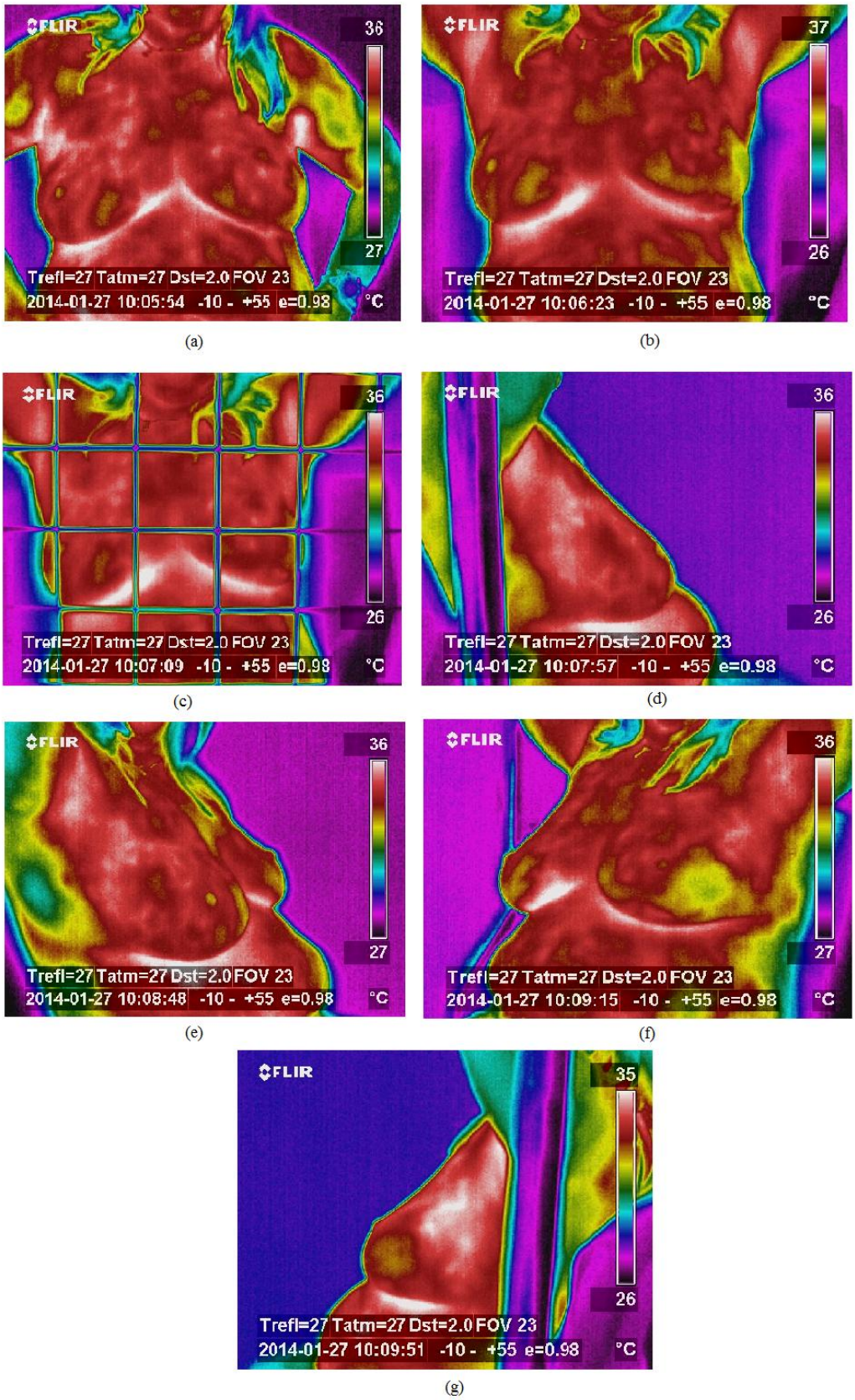
são necessárias para se obter uma melhor precisão dos valores das temperaturas medidas através das imagens por IR.

Para cada paciente são realizadas duas séries de imagens, obtendo um número mínimo de doze imagens. Na primeira série é usada uma distância (D1), com a câmera posicionada a uma distância maior do paciente. Na segunda série, a câmera é colocada mais a uma distância menor (D2), e são obtidas novas imagens com esta distância. As imagens são realizadas das maneiras mostradas a seguir e mostradas na Figura 18:

- T1 (frontal com as mãos na cintura);
- T2 (frontal com as mãos levantadas segurando uma barra localizada acima da cabeça);
- T2 com grade, lateral interna da mama direita (LIMD);
- Lateral interna da mama esquerda (LIME);
- Lateral externa da mama direita (LEMD);
- Lateral externa da mama esquerda (LEME).

Em alguns casos, é necessária a realização de imagens extras de algumas das posições citadas. No presente trabalho foram analisadas imagens do tipo T2, tanto da primeira quanto da segunda série e também foram utilizadas algumas das imagens T2 extras.

Figura 18 - Série de imagens.




(a) T1 (b) T2 frontal (c) T2 com grade (d) LEMD (e) LIME (f) LIMD (g) LEME

## 3.2 BASE DE DADOS

A amostra utilizada no presente trabalho foi composta por 233 imagens termográficas de mama que foram obtidas entre os anos de 2005 e 2014 no HC/UFPE, seguindo os critérios descritos anteriormente. Desta amostra, 43 pacientes foram diagnosticadas com tumor maligno, 78 com tumor benigno, 47 com cisto e 52 pacientes sem anormalidades na mama.

O desbalanceamento da amostra em relação à quantidade de pacientes com tumores benignos pode prejudicar o processo de classificação, tornando o classificador mais tendencioso à Classe Benigno. Sendo assim, na Metodologia 1, as imagens termográficas foram separadas em dois grupos; no primeiro, 175 imagens foram selecionadas aleatoriamente e de modo que a amostra ficasse balanceada para as quatro classes a serem analisadas. Estas imagens foram utilizadas como base de dados no grupo de treinamento de um classificador.

Na Metodologia 2, com o objetivo de manter a amostra totalmente balanceada foram criados vetores sintéticos, de modo que cada classe tivesse a mesma quantidade de amostras, tomando-se como base a classe com maior quantidade de imagens.

Após a retirada das 58 imagens para o balanceamento da amostra, a base de dados utilizada na Metodologia 1 corresponde então, a 175 pacientes, cujos diagnósticos estão informados de maneira sucinta na Tabela 6.

Tabela 6 - Base de dados balanceada para quatro classes.

| **Diagnóstico** | **Quantitativo** |
|---|---|
| Normal | 41 |
| Tumor Benigno | 47 |
| Tumor Maligno | 41 |
| Cisto | 46 |
| **Total** | **175** |

As 58 imagens foram retiradas de maneira aleatória, de forma que a base de dados tivesse o maior número possível de amostras e de modo que estas imagens retiradas tivessem amostras das quatro classes. Isto porque, essas imagens foram separadas para serem testadas e

servirem como grupo de teste, ou seja, utilizadas para validação do classificador utilizado de forma que será discutida mais adiante.

Com o intuito de realização de um método de triagem, também foi construída uma base de dados para ser usada em um classificador binário. Este classificador retornava o diagnóstico a partir do termograma, como Câncer ou Não-Câncer, considerando como Não-Câncer os pacientes dos grupos “benigno”, “cisto” e “normal”. Neste caso, a amostra com 233 imagens foi balanceada em função destas duas classes. Visto que, o número de pacientes diagnosticados com câncer normalmente é menor em relação aos outros diagnósticos, para que a base de dados ficasse balanceada, a base passou a ser composta por apenas 71 imagens. Após este balanceamento, a base de dados formada é mostrada na Tabela 7.

Tabela 7 - Base de dados usada no classificador binário.

| **Diagnóstico** | **Quantitativo** |
|---|---|
| Câncer | 35 |
| Não-Câncer | 36 |
| **Total** | **71** |

Assim como o método anterior, as 162 imagens retiradas para o balanceamento da base de dados, foram utilizadas como grupo de teste para a validação do classificador. Pode-se ressaltar que entre essas 162 imagens, também foi fundamental a presença de amostras em ambas as classes, já que, essas imagens foram utilizadas para validar o classificador.

Na Metodologia 2, foram utilizadas as 233 amostras na base de dados. Com o objetivo de manter a amostra totalmente balanceada foram criados vetores sintéticos, de modo que cada classe tivesse a mesma quantidade de amostras, tomando-se como base a classe com maior quantidade de imagens.

O procedimento foi realizado selecionando 3 vetores da classe *n*, ou seja três valores de uma mesma característica e classe, de amostras distintas, e o vetor sintético foi obtido pela seguinte fórmula:

$$\boldsymbol{v}_n = \frac{v_1+v_2+v_3}{3} \tag{3.1}$$

onde $\boldsymbol{v}_n$ é o vetor sintético. Este vetor pode ser caracterizado como um atributo de uma nova amostra.

Após esse procedimento, a amostra formada para a Metodologia 2 é mostrada na Tabela 8.

Tabela 8 - Base de dados balanceada por meio de vetores sintéticos para quatro classes.

| **Diagnóstico** | **Quantitativo** |
|---|---|
| Normal | 78 |
| Tumor Benigno | 78 |
| Tumor Maligno | 78 |
| Cisto | 78 |
| **Total** | **312** |

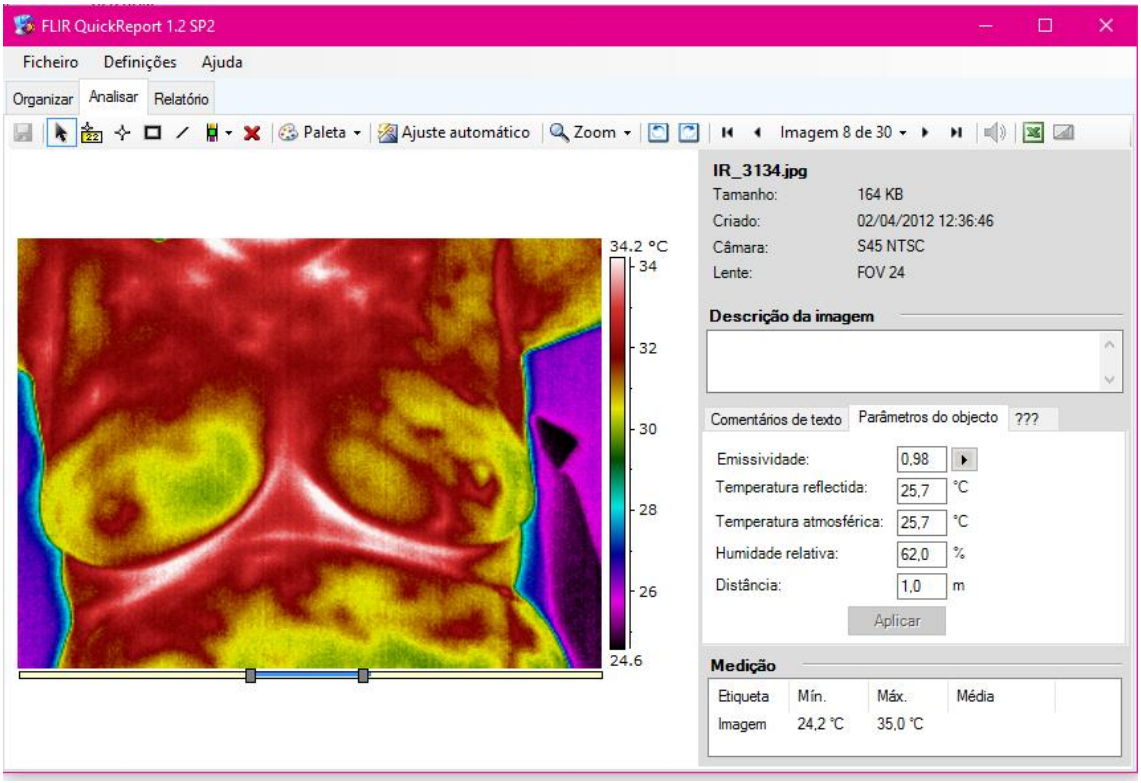
O mesmo procedimento foi utilizado na construção de vetores sintéticos para a classificação binária Câncer ou Não-Câncer. A amostra ficou então, balanceada como mostra a Tabela 9.

Tabela 9 - Base de dados balanceada por meio de vetores sintéticos para classificador binário.

| **Diagnóstico** | **Quantitativo** |
|---|---|
| Câncer | 190 |
| Não-Câncer | 190 |
| **Total** | **380** |

Para a obtenção das matrizes de temperatura a partir das imagens termográficas, com formato *jpeg*, foi usado o *software* FLIR QuickReport, que é disponibilizado pelo fabricante da câmera, uma tela do *software* está mostrada na Figura 19. Neste *software*, caso necessário, podem ser ajustados dados medidos no ato da aquisição das imagens, tais como: a temperatura refletida, a temperatura e a umidade relativa do ambiente, e a distância entre a câmera e a paciente. A emissividade do objeto também pode ser alterada.

Figura 19 - Principal interface do *software* FLIR QuickReport.




Ainda no mesmo programa, é possível exportar os valores de temperaturas de cada *pixel* em planilhas no formato *csv* para o Microsoft Excel. As matrizes obtidas têm dimensão 320x240, que é igual à resolução da câmera e podem ser processadas como uma imagem térmica digital.

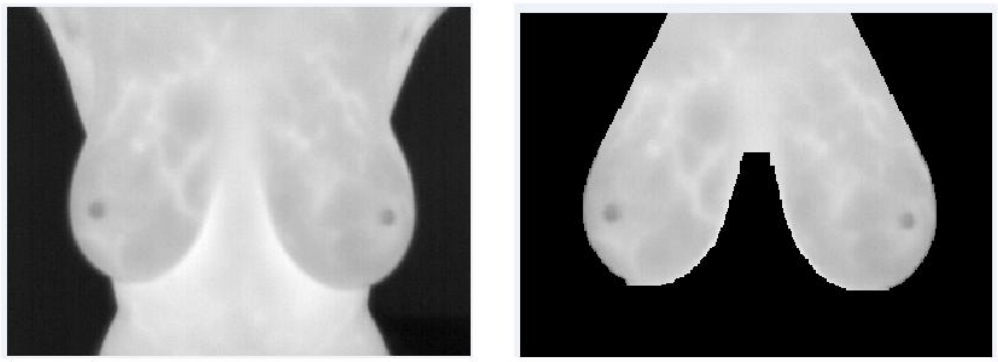
## 3.3 SEGMENTAÇÃO DA IMAGEM DIGITAL

A metodologia de segmentação da região de interesse (ROI) das imagens utilizadas nesse trabalho foi desenvolvida por Dourado Neto (2014). Este autor desenvolveu um método de segmentação automática, que foi escolhida para uma maior praticidade na segmentação. Neste método, as mamas são extraídas do restante da imagem. O autor obteve boas classificações de patologias mamárias usando tal segmentação automática.

A matriz de temperaturas **T** obtida através do *software* FLIR QuickReport foi reconstruída em uma imagem digital no Matlab, onde para cada valor de temperatura foi atribuído uma cor. A partir desta nova imagem, é realizada a segmentação.

O método proposto por Dourado Neto (2014) é um método de segmentação automática implementado na plataforma Matlab, o qual consiste em separar a mama do paciente do resto da imagem, partindo-se de uma imagem em níveis de cinza. Em suma, primeiramente a região corporal é separada do fundo da imagem. Após isso, são determinados limites superiores e inferiores, fazendo com que sejam descartadas regiões do pescoço, das axilas e a borda inferior das mamas. Um exemplo do resultado obtido pelo processo está mostrado na Figura 18:

Figura 20 - Resultado da segmentação automática.



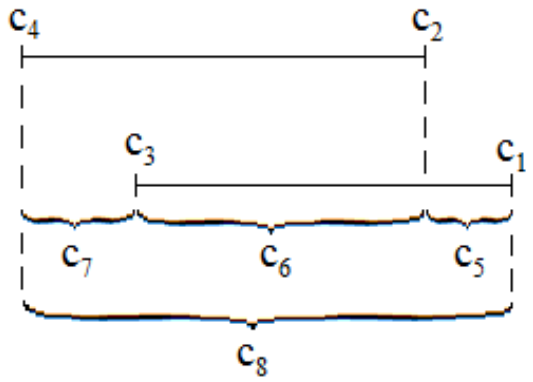
(a) Imagem original (b) Imagem segmentada

## 3.4 EXTRAÇÃO DE CARACTERÍSTICAS

Após a segmentação, a região de interesse (ROI), foi analisada com o objetivo de se extraírem as características. Estas características são necessárias para a classificação das imagens termográficas das pacientes, e são utilizadas como entrada dos classificadores.

Araújo (2014) desenvolveu um método com abordagens fundamentadas em medidas estatísticas e em medidas intervalares, tais como: a diferença das temperaturas máximas entre as mamas; a diferença das temperaturas mínimas entre as mamas; a diferença entre a máxima e a mínima temperatura obtida para as duas mamas como mostrado na Figura 21. Vinte características foram extraídas.

Figura 21 - Representação gráfica de algumas características definidas a partir das temperaturas máximas e mínimas da mama esquerda e da mama direita.




Fonte: DOURADO NETO (2014).

onde:

$c_1$ = máximo entre as temperaturas máximas da mama esquerda e da mama direita;

$c_2$ = mínimo entre as temperaturas máximas da mama esquerda e da mama direita;

$c_3$= máximo entre as temperaturas mínimas da mama esquerda e da mama direita;

$c_4$ = mínimo entre as temperaturas mínimas da mama esquerda e da mama direita.

$c_5 = c_1 - c_2$;

$c_6 = c_2 - c_3$;

$c_7 = c_3 - c_4$;

$c_8 = c_1 - c_4$.

As Características 9 e 10, definidas a seguir, são medidas estatísticas básicas de média e desvio padrão:

$c_9$ = média das temperaturas;

$c_{10}$ = desvio padrão das temperaturas.

Dourado Neto (2014) propôs um critério de robustez que tem por objetivo evitar a influência de possíveis regiões que no processo de segmentação, não foram totalmente excluídas, ele considera que:

- Na procura da temperatura máxima, apenas se consideraram os pixels com temperatura maior ou igual aos 8 pixels imediatamente vizinhos e que não estavam na borda da região segmentada;
- Na procura da temperatura mínima, apenas se consideraram os pixels com temperatura menor ou igual aos 8 pixels imediatamente vizinhos e que não estavam na borda da região segmentada.

A extração das Características 11 a 18 foi feita de forma equivalente às Características 1 a 8, porém sem utilizar o critério de robustez proposto por Dourado Neto (2014):

$c_{11}$ = máximo entre as temperaturas máximas da mama esquerda e da mama direita;

$c_{12}$ = mínimo entre as temperaturas máximas da mama esquerda e da mama direita;

$c_{13}$ = máximo entre as temperaturas mínimas da mama esquerda e da mama direita;

$c_{14}$ = mínimo entre as temperaturas mínimas da mama esquerda e da mama direita.

$c_{15} = c_{11} - c_{12}$;

$c_{16} = c_{12} - c_{13}$;

$c_{17} = c_{13} - c_{14}$;

$c_{18} = c_{11} - c_{14}$.

Em algumas amostras, quando não é necessário o uso do critério de robustez, as Características 11 a 18 são iguais as Características 1 a 8.

As Características 19 e 20 são a obliquidade (medida de assimetria entre a esquerda e a direita do histograma) e a curtose (medida de "achatamento" do histograma), respectivamente:

$c_{19}$ = obliquidade do histograma das temperaturas;

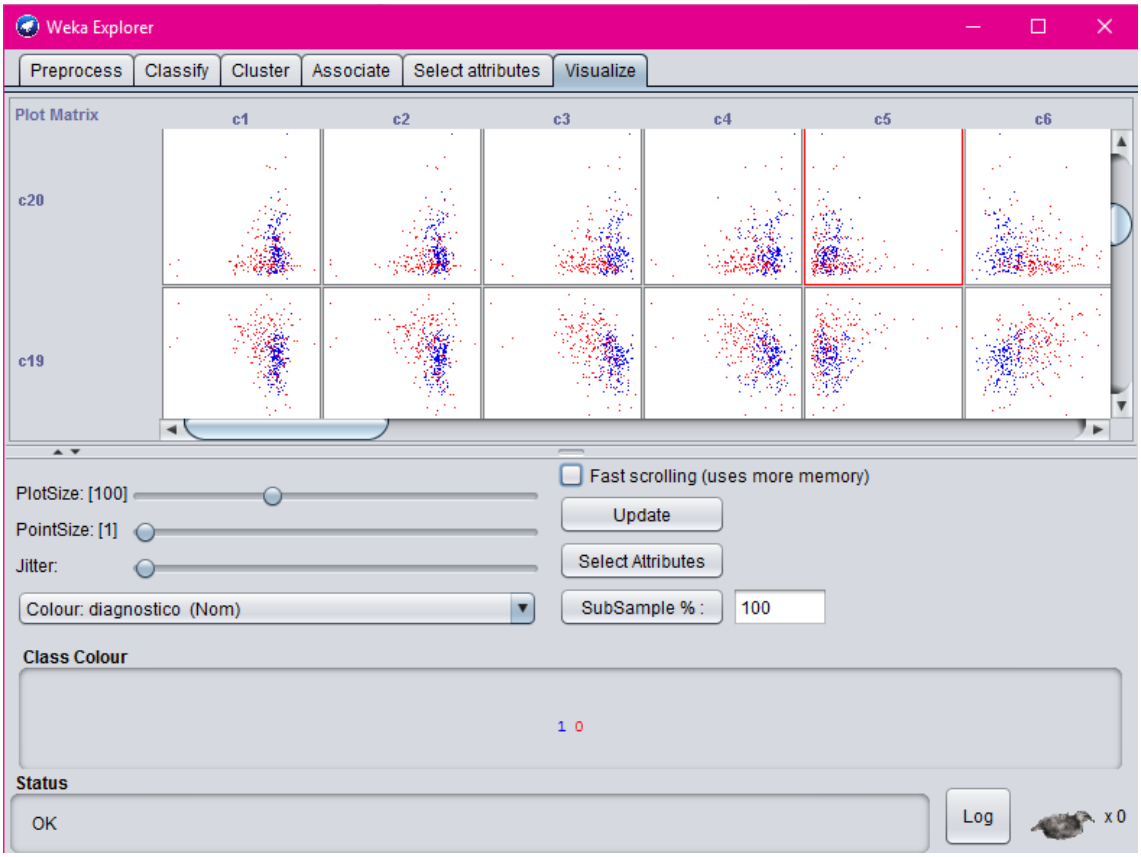
$c_{20}$ = curtose do histograma das temperaturas.

Além destas características, foram adicionados ao vetor de características de cada imagem por infravermelho os respectivos dados sobre: idade do indivíduo, temperatura ambiente e umidade relativa do ar no momento da aquisição da imagem.

Na Metodologia 1, foi realizada uma seleção de atributos baseadas nas vinte características propostas. O *software* livre WEKA foi utilizado para encontrar o melhor par de características para a base proposta neste trabalho.

Para a classificação nas quatro classes propostas, uma busca do melhor par de características foi feita através do *software* WEKA. A melhor característica foi selecionada por meio de histogramas individuais de cada atributo, utilizando em conjunto a função de seleção de atributos disponível naquele *software*. Por fim, para a seleção do seu par, foi realizada uma análise visual através dos gráficos gerados na Figura 20 com a finalidade de encontrar uma configuração que melhor separasse cada classe. Para a classificação binária (Câncer ou Não-Câncer), o par de atributos foi escolhido seguindo os mesmos critérios descritos.

Figura 22 - Gráficos das possíveis combinações de atributos.




Ao contrário da Metodologia 1, onde só se utilizaram os pares de características considerados mais eficientes como entrada do grupo de treinamento e de teste do classificador, na Metodologia 2, foram utilizadas todas as vinte características extraídas, além

da temperatura inicial da paciente, da sua idade e da umidade do ambiente, com o objetivo de aumentar a precisão do classificador.

A fim de eliminar as possíveis relações lineares entre as características e apenas para critério de comparação, na Metodologia 2, também foram utilizadas as Características 5 a 8 e suas equivalentes 15 a 18, da seguinte forma:

$c_5 = c_1^2 - c_2^2$;

$c_6 = c_2^2 - c_3^2$;

$c_7 = c_3^2 - c_4^2$;

$c_8 = c_1^2 - c_4^2$;

$c_{15} = c_{11}^2 - c_{12}^2$;

$c_{16} = c_{12}^2 - c_{13}^2$;

$c_{17} = c_{13}^2 - c_{14}^2$;

$c_{18} = c_{11}^2 - c_{14}^2$.

## 3.5 CLASSIFICAÇÃO DE IMAGENS TERMOGRÁFICAS

Depois de feita a devida extração das características, estas serviram de entrada para os classificadores, tanto para o grupo usado na base de dados, quanto para o grupo de validação do classificador. Com o objetivo de avaliar o desempenho dos classificadores a respeito da classificação correta do diagnóstico das pacientes, tanto numa classificação binária, quanto numa classificação multiclasse, foram utilizados sete tipos de classificadores.

No trabalho realizado por Queiroz (2016), com uma base de dados de 98 imagens, o classificador que apresentou melhor desempenho foi o SVM com núcleo polinomial 8, utilizando a técnica *one-versus-all*. A classificação foi realizada para três classes (Benigno, Cisto e Normal). Os resultados obtidos pela autora foram: 51,58% de taxa de acerto, 66,67% de sensibilidade e 82,35% de especificidade.

Com o intuito de melhorar os resultados, o presente trabalho utilizou uma base de dados ampliada. Na Metodologia 1, foram utilizadas 175 imagens como base de dados para

um classificador SVM e a técnica *one-versus-one* para a classificação SVM multiclasse para as classes Maligno, Benigno, Cisto e Normal. No presente trabalho foi utilizado o LIBSVM, que tem como padrão a utilização do método *one-versus-one* para classificação multiclasse. O LIBSVM é uma biblioteca de aprendizado de máquina de código aberto, desenvolvida pela Universidade Nacional de Taiwan por Chih-Chung Chang e Chih-Jen Lin (CHANG & LIN, 2011). A biblioteca foi incorporada ao Matlab, possibilitando efetuar as classificações, tanto binária quanto multiclasse.

Na Metodologia 2, ainda com o objetivo de aumentar a eficiência dos classificadores, foram utilizados os classificadores do *software* WEKA: *Naive Bayes*, *Bayes Net*, *Multilayer Perceptron*, *Random Forest*, *Random Tree*, *IBK (K-nearest neighbours – KNN)* e o SMO, descritos detalhadamente no Capítulo 2 do presente trabalho. Nesta etapa, toda a amostra foi dividida em 5 subconjuntos, utilizada como grupo de treinamento e a validação do classificador foi feita por meio de validação cruzada.

Na validação cruzada, o conjunto de treinamento original é dividido em N subconjuntos, no qual um destes subconjuntos é retido e utilizado como validação do classificador, e os N-1 subconjuntos são utilizados como treinamento. O processo de validação cruzada se repete, então, N vezes, de maneira que todos os N subconjuntos sejam utilizados uma vez como dado de teste, para a validação do classificador. O resultado desse processo é dado pelo desempenho médio do classificador nos N testes. Uma maior quantidade de testes gera uma maior confiabilidade de estimativa da precisão do classificador (YATES & RIBEIRO NETO, 2013).

Finalmente, a eficiência das classificações foi calculada por meio da taxa de acerto de cada classificador, do coeficiente Kappa e dos valores de sensibilidade e especificidade para a Classe Maligno. Foram comparados os resultados obtidos para cada método para avaliar qual classificador fornece uma análise mais precisa.

# 4 RESULTADOS E DISCUSSÃO

Neste capítulo são apresentados os resultados obtidos com o desenvolvimento e aplicação das metodologias citadas no Capítulo 3 do presente trabalho. Tais resultados foram avaliados com o intuito de descobrir os melhores métodos de classificação a serem aplicados à base de dados composta por imagens termográficas de mama.

Inicialmente apresentam-se os resultados obtidos através de uma avaliação da classificação multiclasse, com às classes Maligno, Benigno, Cisto e Normal. Em seguida, foi apresentada os resultados da classificação binária, do tipo (Câncer/Não-Câncer). Foram considerados como Câncer, aqueles pacientes pertencentes ao grupo "Maligno" e como Não-Câncer, os pacientes dos grupos "Benigno", "Cisto" e "Normal".

As imagens foram classificadas de duas formas. Primeiro, com a metodologia baseada naquela desenvolvida por Dourado Neto (2014). Em seguida, a classificação foi efetuada utilizando-se vetores sintéticos para balanceamento das amostras, na segunda metodologia desenvolvida. Nos dois casos foi usada a segmentação automática, desenvolvida também pelo mesmo autor.

Foram utilizados o classificador SVM na Metodologia 1 e *Naive Bayes, Bayes Net, Multilayer Perceptron, Random Forest, Random Tree,* KNN e o SMO na Metodologia 2, para obtenção dos resultados, apresentados a seguir.

## 4.1 ANÁLISE DA CLASSIFICAÇÃO DA METODOLOGIA 1

Partindo-se das vinte características propostas por Dourado Neto (2014) e mostradas em detalhe no Capítulo 3, foi realizada uma seleção de atributos baseadas nestas características. O *software* livre WEKA foi utilizado para encontrar o melhor par de características para a base de dados proposta para esta metodologia.

Para a classificação nas quatro classes propostas, uma busca do melhor par de características foi realizada através do WEKA, com a finalidade de encontrar uma configuração que melhor separasse individualmente cada classe. Por meio deste *software*, observou-se que os atributos que apresentaram melhores resultados individualmente foram: $c_2$, $c_3$, $c_4$, $c_6$, $c_9$, $c_{13}$, $c_{14}$ e $c_{18}$, que têm seus histogramas mostrados na Figura 23.

Figura 23 - Histogramas para seleção de características com quatro classes (Azul-Maligno, Vermelho-Benigno, Verde-Cisto e Cinza-Normal).

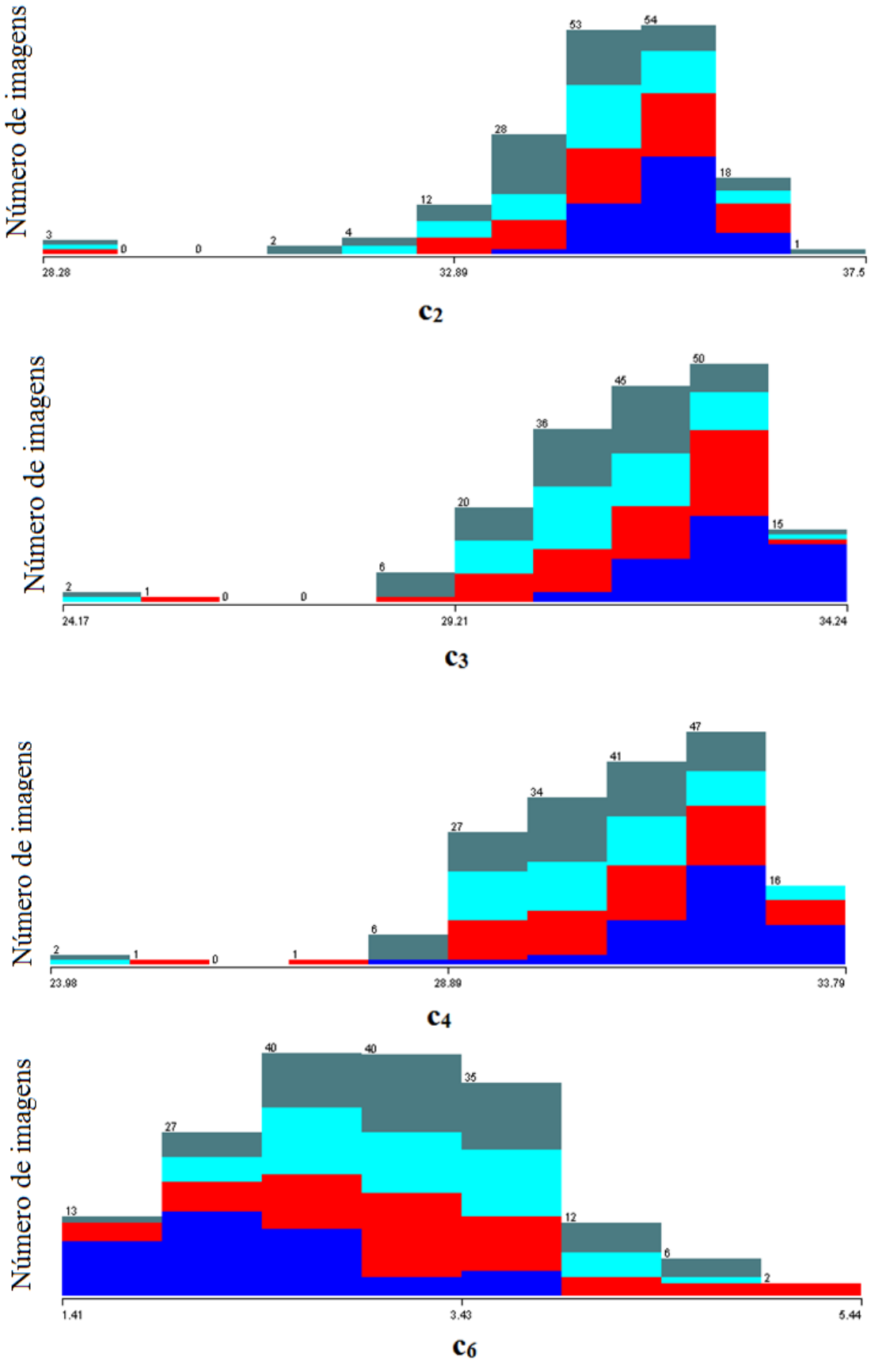
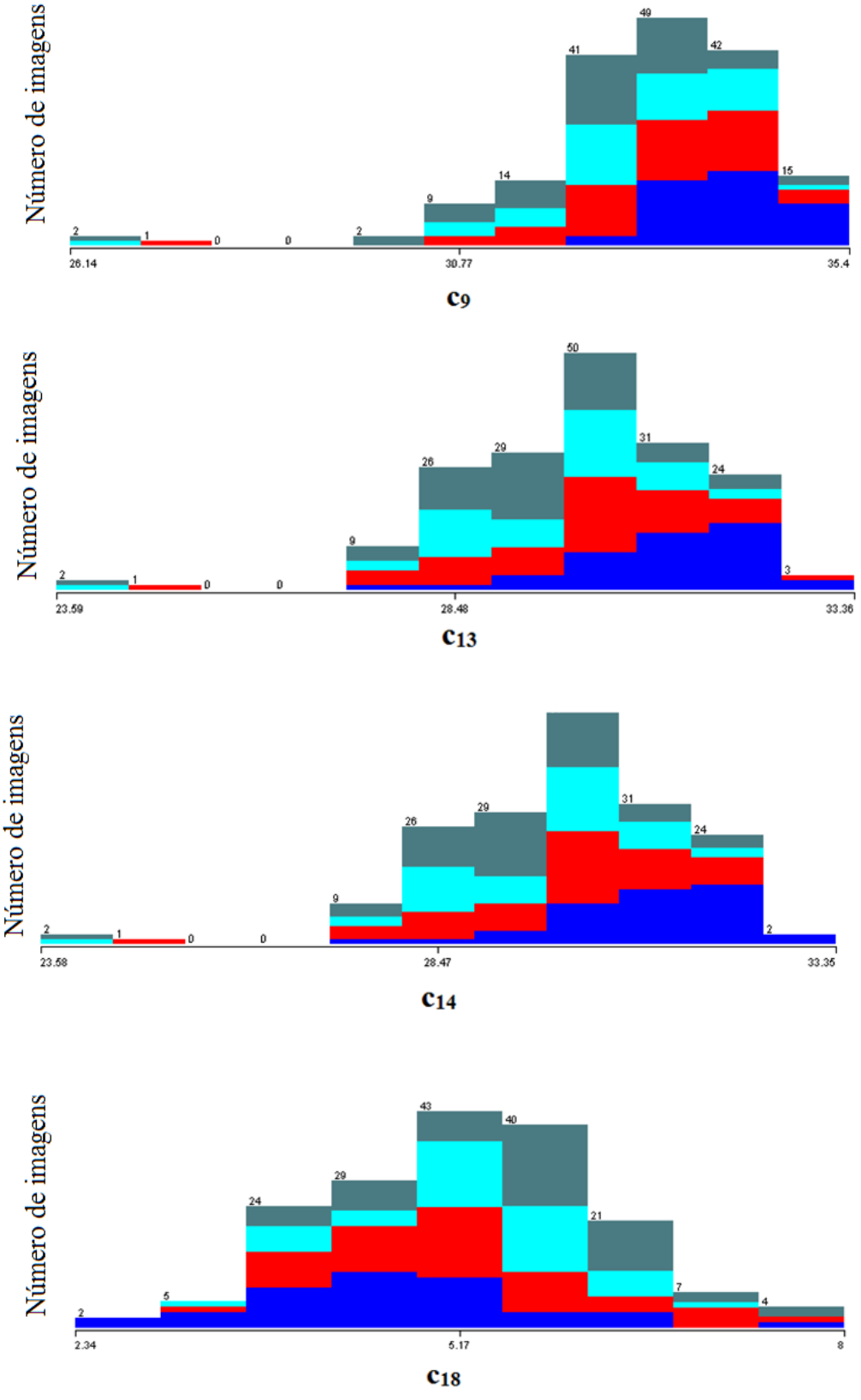


Figura 23 - Histogramas para seleção de características com quatro classes (Azul-Maligno, Vermelho-Benigno, Verde-Cisto e Cinza-Normal).



A Tabela 10 apresenta os resultados para a seleção de características por meio do WEKA, utilizando o método *Attribute Ranking* e o analisador *Information Gain Ranking Filter*.

Tabela 10 - Ranking das melhores características para quatro classes.

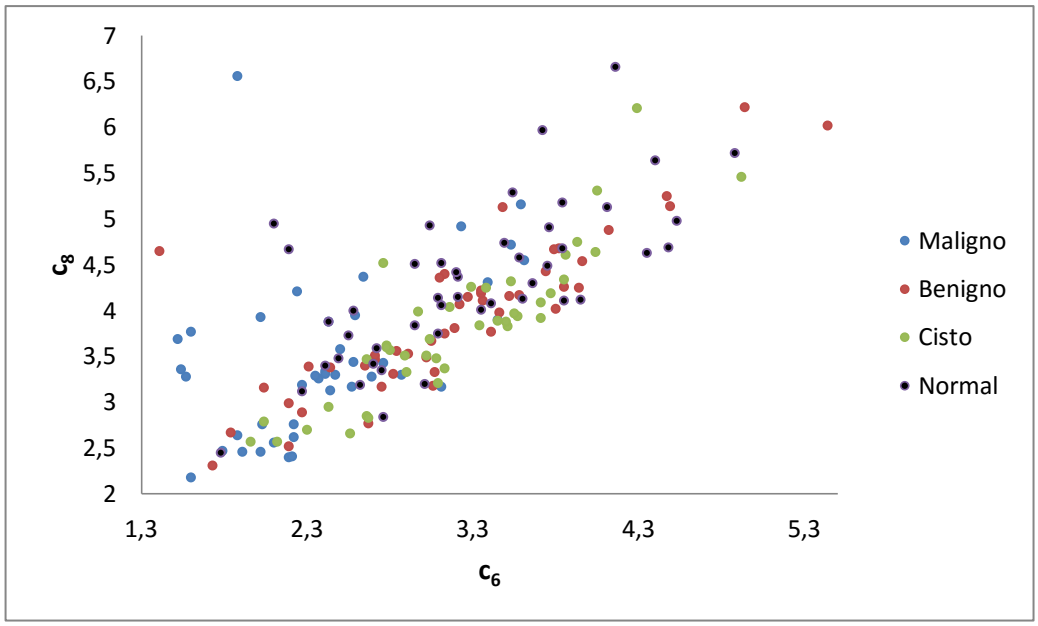
| **Característica** | ***Ranked*** |
|---|---|
| $c_6$ | 0,188 |
| $c_3$ | 0,188 |
| $c_4$ | 0,16 |
| $c_{13}$ | 0,159 |
| $c_{14}$ | 0,159 |
| $c_9$ | 0,158 |
| $c_2$ | 0,126 |
| $c_{18}$ | 0,12 |

A característica que apresentou o melhor agrupamento de cada classe e a menor sobreposição no histograma foi a selecionada. De uma análise conjunta dos histogramas com a Tabela 8, o atributo que mais se adequou ao objetivo do trabalho foi o $c_6$. Já que este atributo, além de apresentar menor sobreposição entre as classes, foi também um dos atributos que apresentaram melhores resultados na seleção das características do *software* WEKA.

A busca pela característica que melhor combina com $c_6$ foi feita de maneira visual, através de todas as combinações possíveis, buscando a configuração que melhor separasse as classes individualmente.

O par de características $c_6$ e $c_8$ foi escolhido como sendo o que gerou a discriminação mais adequadas entre as classes, fornecendo uma menor dispersão das mesmas. Porém, é possível notar na Figura 24 que a Classe Normal está sobreposta às outras classes, o que dificulta a classificação. A dificuldade de separação da Classe Normal das demais também é comentada por Araújo (2014) e Queiroz (2016).

Figura 24 - Espaço de características usado para a seleção da melhor combinação das quatro classes [$c_6$ e $c_8$].




A amostra que forma a base de dados, composta por 175 imagens, foi então analisada com base no par de características [$c_6$, $c_8$]. A Tabela 11 mostra os resultados encontrados após a classificação, com os pares de atributos selecionados, do grupo de teste composto por termogramas de 58 pacientes, utilizando a base de dados citada.

Os resultados apresentados foram obtidos através do classificador SVM, utilizando uma função linear, e o método um-contra-um. A definição dos melhores resultados foi feita buscando as maiores taxas de acertos globais e a maior sensibilidade à Classe Maligno, visto que, a presença de falsos negativos em relação a esta classe compromete o diagnóstico precoce do câncer na paciente.

Tabela 11 - Resultados da classificação multiclasse.

| Características | [$c_6$, $c_8$] |
|---|---|
| Acertos | 53,45% (31/58) |
| Sensibilidade | 100% |
| Especificidade | 64,44% |
| Acerto Maligno | 100% (2/2) |
| Acerto Benigno | 58,06% (18/31) |
| Acerto Cisto | 0% (0/1) |
| Acerto Normal | 45,83% (11/24) |

A Figura 25 mostra a matriz de confusão dos resultados encontrados.

Figura 25 - Matriz de confusão da classificação multiclasse.

| | | CLASSIFICAÇÃO | | | | |
|---|---|---|---|---|---|---|
| | | Maligno | Benigno | Cisto | Normal | Total de amostras |
| VERDA-DEIRO | Maligno | 2 | 0 | 0 | 0 | 2 |
| | Benigno | 9 | 18 | 0 | 4 | 31 |
| | Cisto | 1 | 0 | 0 | 0 | 1 |
| | Normal | 6 | 7 | 0 | 11 | 24 |
| | | | | | Total: | 58 |

Estes resultados, mostram uma melhora na classificação se comparados ao de Queiroz (2016), que utilizou uma base de dados reduzida e a classificação SVM multiclasse um-contra-todos. Outros tipos de função *Kernel* foram testadas, porém levando-se em consideração a definição comentada dos melhores resultados, a função linear foi a que apresentou resultados mais compatíveis.

Para a classificação binária (Câncer x Não-Câncer), o par de atributos foi escolhido seguindo os mesmos passos e critérios descritos anteriormente. Através do *software* WEKA, verificou-se que as características que apresentaram melhores resultados individualmente foram: $c_1$, $c_2$, $c_3$, $c_6$, $c_7$ e $c_9$, que têm seus histogramas apresentados na Figura 26.

Figura 26 - Histogramas para seleção de características com duas classes (Azul-Câncer e Vermelho-Não-Câncer).

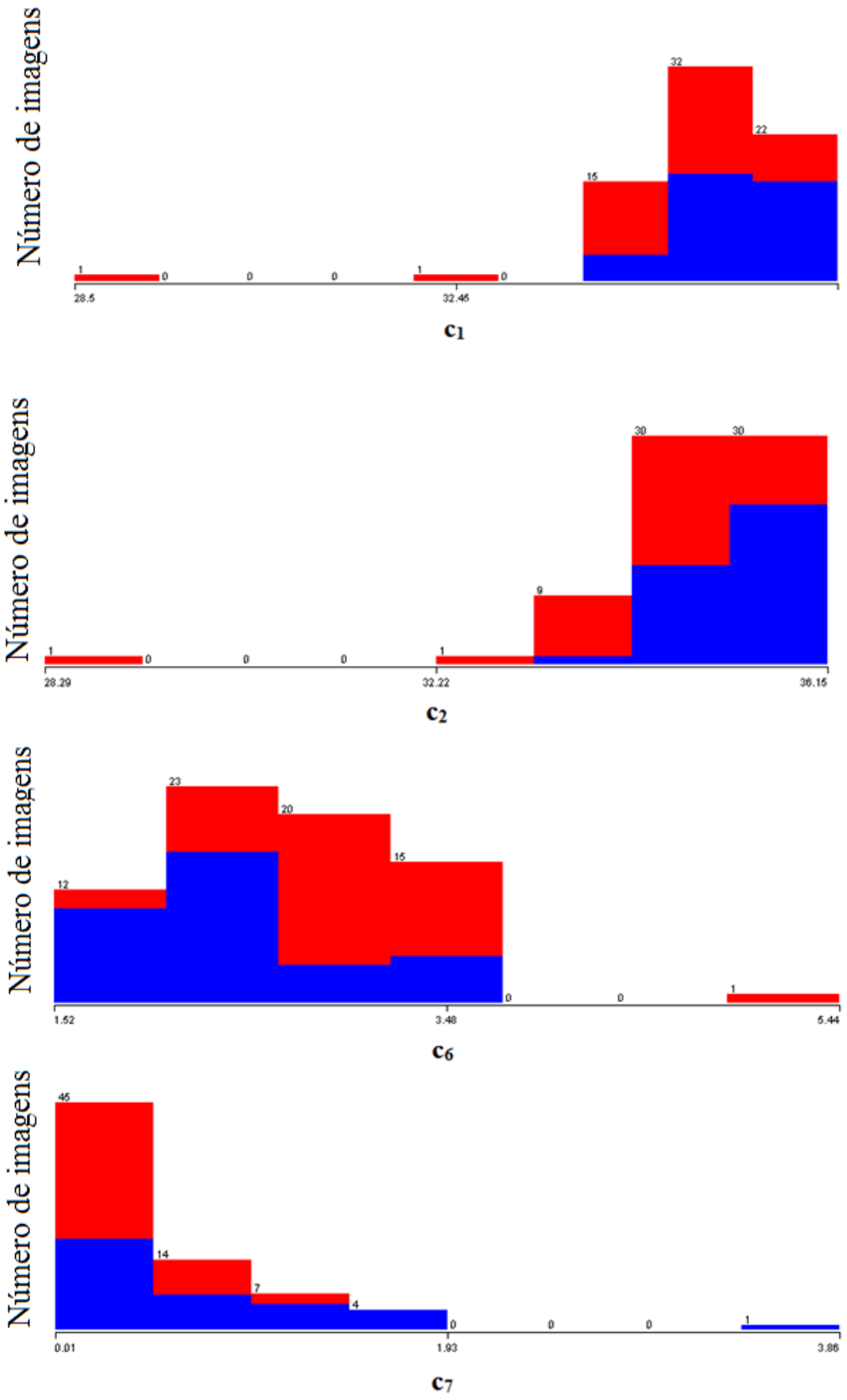
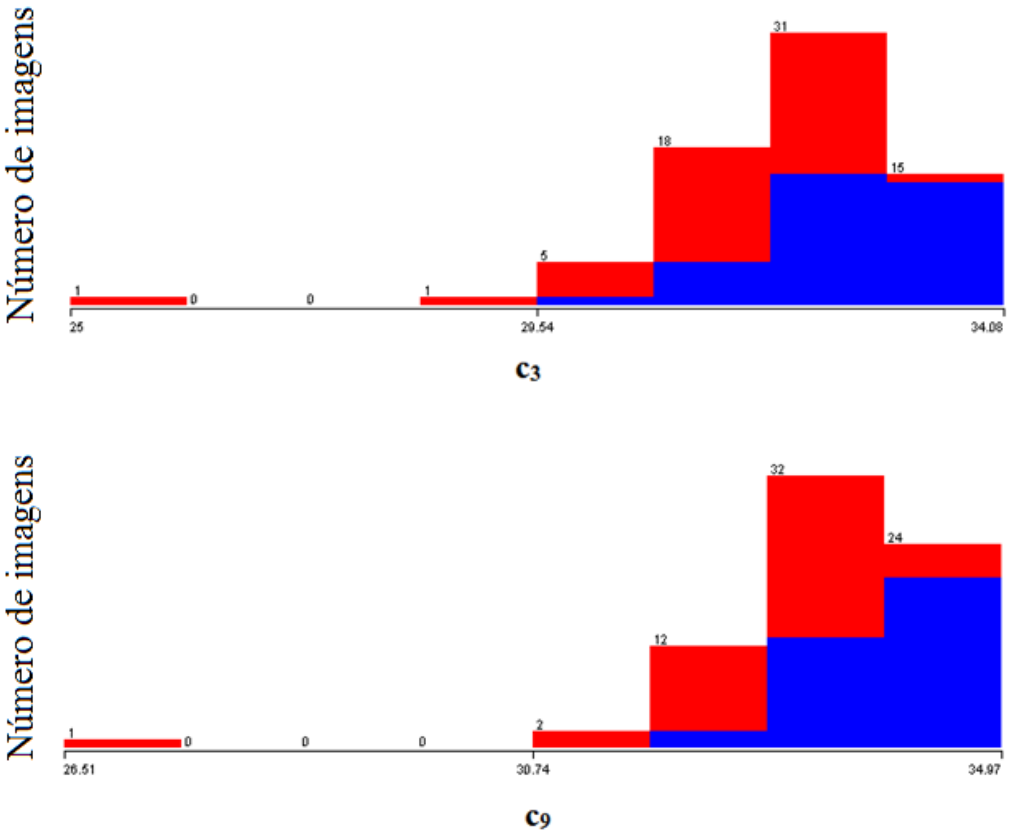


Figura 26 - Histogramas para seleção de características com duas classes (Azul-Câncer e Vermelho-Não-Câncer).




A Tabela 12 mostra, do mesmo modo, os resultados para a seleção de características obtidas por meio do WEKA, utilizando o método *Attribute Ranking* e o analisador *Information Gain Ranking Filter*.

Tabela 12 - Ranking das melhores características para duas classes.

| **Característica** | ***Ranked*** |
|---|---|
| $c_3$ | 0,231 |
| $c_6$ | 0,183 |
| $c_9$ | 0,182 |
| $c_7$ | 0,161 |
| $c_1$ | 0,137 |
| $c_2$ | 0,137 |

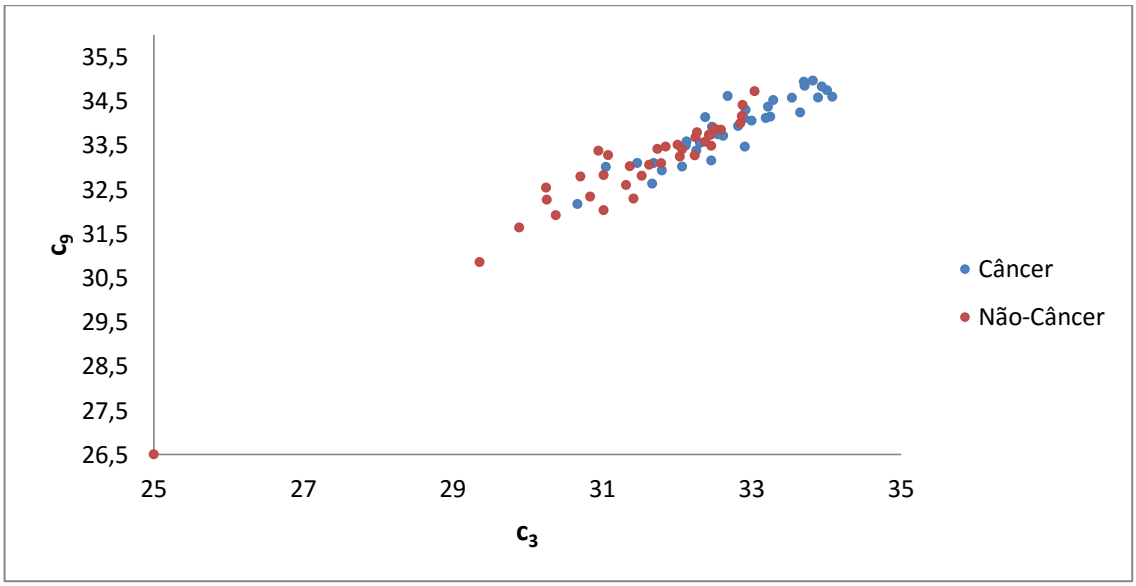
Usando a mesma metodologia utilizada para a análise multiclasse, foi selecionada a característica que apresentou o melhor agrupamento de cada classe no histograma. E de uma análise conjunta dos histogramas com a Tabela 10, a característica que mais se adequou ao objetivo do trabalho foi a $c_3$.

A busca pela característica que melhor combina com $c_3$ foi realizada do mesmo modo, o modo visual, através de todas as combinações possíveis, buscando a configuração que melhor separasse as classes individualmente, neste caso, a Classe Câncer da Classe Não-Câncer.

O par de características $c_3$ e $c_9$ foi escolhido como sendo o que gerou a discriminação mais adequada entre as classes, fornecendo uma menor dispersão das mesmas, como mostra a Figura 2

7:

Figura 27 - Espaço de características usado para a seleção da melhor combinação das duas classes [$c_3$ e $c_9$].




A amostra que forma a base de dados, composta por 71 imagens, foi então utilizada com base no par de características [$c_3$, $c_9$]. A Tabela 13 mostra os resultados encontrados após a classificação, com os pares de atributos selecionados, para o grupo de teste formado por 162 imagens.

Tabela 13 - Resultados da classificação binária.

| Características | [$c_3$, $c_9$] |
|---|---|
| Acertos | 83,33% (135/162) |
| Sensibilidade | 87,50% |
| Especificidade | 83,12% |
| Acerto Câncer | 87,5% (7/8) |
| Acerto Não-Câncer | 83,12% (128/154) |

A Figura 28 apresenta a matriz de confusão obtida nesta classificação binária.

Figura 28 - Matriz de confusão da classificação binária.

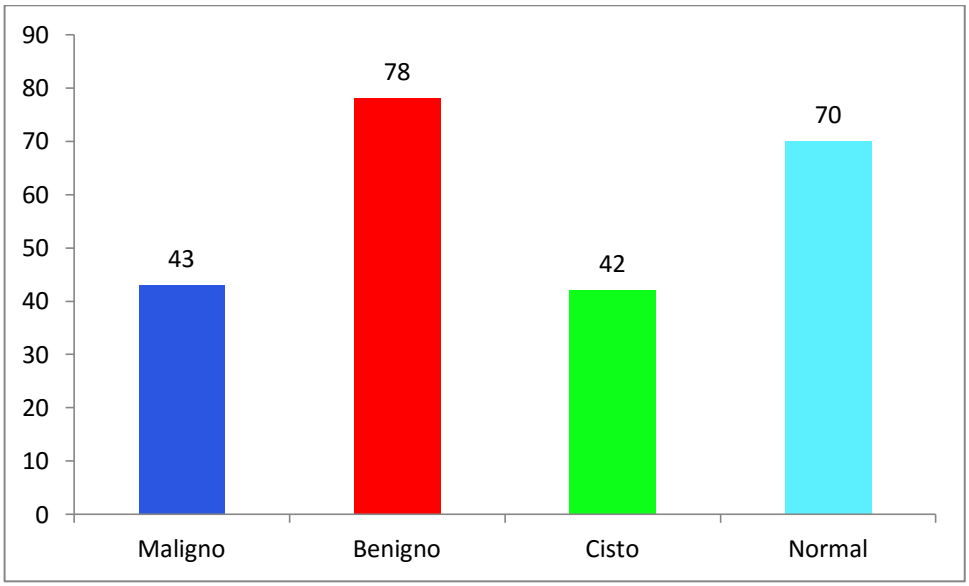
| | | CLASSIFICAÇÃO | | |
|---|---|---|---|---|
| | | Câncer | Não-Câncer | Total de amostras |
| VERDADEIRO | Câncer | 7 | 1 | 8 |
| | Não-Câncer | 26 | 128 | 154 |
| | | | Total: | 162 |

Os resultados apresentados também foram obtidos através do classificador SVM, porém neste caso, foi usada uma função de núcleo polinomial de ordem 3. A definição dos melhores resultados foi realizada do mesmo modo, fazendo uma busca das maiores taxas de acertos globais e a maior sensibilidade à classe Câncer. Aqui, outros tipos de função *Kernel* também foram testados, como a função linear, além de variações do grau do polinômio da função *Kernel* polinomial de 2 a 8. Porém, para essa análise binária e levando em consideração a definição dos melhores resultados, a função polinomial de grau 3 foi a que apresentou os melhores resultados.

## 4.2 ANÁLISE DA CLASSIFICAÇÃO DA METODOLOGIA 2

A amostra inicial composta por 233 imagens é distribuída nas quatro classes analisadas da maneira mostrada pela Figura 29:

Figura 29 - Histograma da amostra inicial em quatro classes.




Com o intuito de balancear a amostra sem diminuir, consideravelmente, o número de imagens que formariam a base de treinamento do classificador, foram utilizados vetores sintéticos para o balanceamento daquela base.

Como mostrado na Figura 29, a população máxima do histograma é de 78 amostras. A criação dos vetores sintéticos teve como objetivo fazer com que todas as classes, tenham esse mesmo valor de amostras.

A Tabela 14 mostra a quantidade de vetores sintéticos necessários para cada classe, de modo que todas as classes tenham o mesmo número de amostras.

Tabela 14 - Número de vetores sintéticos utilizados.

| **Classe** | **Quantidade de vetores sintéticos** |
|---|---|
| Maligno | 35 |
| Benigno | 0 |
| Cisto | 36 |
| Normal | 8 |

Apesar de garantir o balanceamento total e aumento do número de amostras da base de dados, o balanceamento por meio de vetores sintéticos poderia ocasionar uma diminuição

na eficiência do processo de classificação, visto que, as novas amostras, ou seja, as amostras sintéticas poderiam ser associadas a classes errôneas.

Após o balanceamento das imagens, o arquivo do tipo *arff* foi montado e o *software* WEKA foi utilizado para classificar as 312 imagens balanceadas por meio de vetores sintéticos. A Tabela 15 mostra os resultados encontrados para uma classificação multiclasse, nas quatro classes analisadas no presente trabalho, utilizando os classificadores citados anteriormente.

Tabela 15 - Resultados da classificação multiclasse através do WEKA

| **Classificador** | **Taxa de Acerto** | **Sensibi-lidade** | **Especifi-cidade** | **Acertos Maligno** | **Acertos Benigno** | **Acertos Cisto** | **Acertos Normal** | **Kappa** |
|---|---|---|---|---|---|---|---|---|
| BayesNet | 46,47% | 73,08% | 65,19% | 73,08% | 12,82% | 47,43% | 52,56% | 0,2863 |
| NaiveBayes | 48,08% | 85,90% | 55,70% | 85,90% | 10,25% | 46,15% | 50% | 0,3077 |
| MLP | 59,29% | 75,64% | 88,11% | 75,64% | 41,02% | 64,10% | 56,41% | 0,4573 |
| **SVM** | **63,46%** | **80,77%** | **86,54%** | **80,77%** | **44,87%** | **71,79%** | **56,41%** | **0,5128** |
| RandomForest | 62,82% | 78,21% | 85,44% | 78,21% | 33,33% | 70,51% | 69,23% | 0,5043 |
| RandomTree | 49,09% | 66,67% | 74,41% | 66,67% | 26,92% | 53,85% | 48,72% | 0,3205 |
| Ibk (KNN) | 62,18% | 84,62% | 85,91% | 84,62% | 26,92% | 83,33% | 48,72% | 0,4957 |

Na Figura 30 são apresentadas as matrizes de confusão com respectivo classificador utilizado, dos resultados expostos na Tabela 15.

Figura 30 - Matrizes de confusão da classificação multiclasse através do WEKA.

Bayes Net

CLASSIFICAÇÃO

| VERDADEIRO | | Maligno | Benigno | Cisto | Normal | Total de amostras |
|---|---|---|---|---|---|---|
| | Maligno | 57 | 9 | 5 | 7 | 78 |
| | Benigno | 27 | 10 | 23 | 18 | 78 |
| | Cisto | 13 | 10 | 37 | 18 | 78 |
| | Normal | 7 | 14 | 16 | 41 | 78 |
| | | | | | Total: | 312 |

Figura 30 - Matrizes de confusão da classificação multiclasse através do WEKA.

Naive Bayes

| | | CLASSIFICAÇÃO | | | | |
|---|---|---|---|---|---|---|
| | | Maligno | Benigno | Cisto | Normal | Total de amostras |
| VERDADEIRO | Maligno | 67 | 4 | 4 | 3 | 78 |
| | Benigno | 31 | 8 | 21 | 18 | 78 |
| | Cisto | 20 | 9 | 36 | 13 | 78 |
| | Normal | 15 | 11 | 13 | 39 | 78 |
| | | | | | Total: | 312 |

MLP

| | | CLASSIFICAÇÃO | | | | |
|---|---|---|---|---|---|---|
| | | Maligno | Benigno | Cisto | Normal | Total de amostras |
| VERDADEIRO | Maligno | 59 | 7 | 3 | 9 | 78 |
| | Benigno | 9 | 32 | 28 | 9 | 78 |
| | Cisto | 1 | 18 | 50 | 9 | 78 |
| | Normal | 7 | 13 | 14 | 44 | 78 |
| | | | | | Total: | 312 |

SMO

| | | CLASSIFICAÇÃO | | | | |
|---|---|---|---|---|---|---|
| | | Maligno | Benigno | Cisto | Normal | Total de amostras |
| VERDADEIRO | Maligno | 63 | 9 | 3 | 3 | 78 |
| | Benigno | 10 | 35 | 20 | 13 | 78 |
| | Cisto | 3 | 16 | 56 | 3 | 78 |
| | Normal | 8 | 14 | 12 | 44 | 78 |
| | | | | | Total: | 312 |

Random Forest

| | | CLASSIFICAÇÃO | | | | |
|---|---|---|---|---|---|---|
| | | Maligno | Benigno | Cisto | Normal | Total de amostras |
| VERDADEIRO | Maligno | 61 | 7 | 1 | 9 | 78 |
| | Benigno | 12 | 26 | 25 | 15 | 78 |
| | Cisto | 2 | 13 | 55 | 8 | 78 |
| | Normal | 9 | 8 | 7 | 54 | 78 |
| | | | | | Total: | 312 |

Figura 30 - Matrizes de confusão da classificação multiclasse através do WEKA.

Random Tree

CLASSIFICAÇÃO

| VERDADEIRO | Maligno | Benigno | Cisto | Normal | Total de amostras |
|---|---|---|---|---|---|
| Maligno | 52 | 6 | 9 | 11 | 78 |
| Benigno | 17 | 21 | 22 | 18 | 78 |
| Cisto | 6 | 20 | 42 | 10 | 78 |
| Normal | 10 | 17 | 13 | 38 | 78 |
| | | | | Total: | 312 |

Ibk (KNN)

CLASSIFICAÇÃO

| VERDADEIRO | Maligno | Benigno | Cisto | Normal | Total de amostras |
|---|---|---|---|---|---|
| Maligno | 66 | 3 | 3 | 6 | 78 |
| Benigno | 14 | 25 | 29 | 10 | 78 |
| Cisto | 1 | 9 | 65 | 3 | 78 |
| Normal | 6 | 14 | 20 | 38 | 78 |
| | | | | Total: | 312 |

Neste caso, não houve uma seleção de atributos, foram utilizados todos os vinte atributos citados no Capítulo 3, além da idade do indivíduo, da temperatura ambiente e da umidade relativa do ar no momento da aquisição da imagem. A validação do classificador foi realizada através de validação cruzada com cinco subconjuntos. Para todos os classificadores, com exceção do SMO, foram usados os parâmetros padrões do WEKA. Nesse último, foi utilizado função *Kernel* polinomial de grau 3.

Nesta metodologia, para a escolha dos melhores resultados, além de buscar as maiores taxas de acertos globais e a maior sensibilidade à Classe Maligno, também se analisou o coeficiente Kappa, e o utilizou como critério de desempate do classificador mais eficiente. As baixas taxas de acertos encontradas nas Classes Benigno e Normal devem-se ao fato de ainda se ter uma sobreposição entre as classes.

Para critério de comparação, a fim de eliminar as possíveis combinações lineares entre as características, foram utilizadas também as características citadas no Capítulo 3, e propostas para esse fim. A Tabela 16 mostra as taxas de acertos e os respectivos coeficientes Kappa, para critérios de comparação.

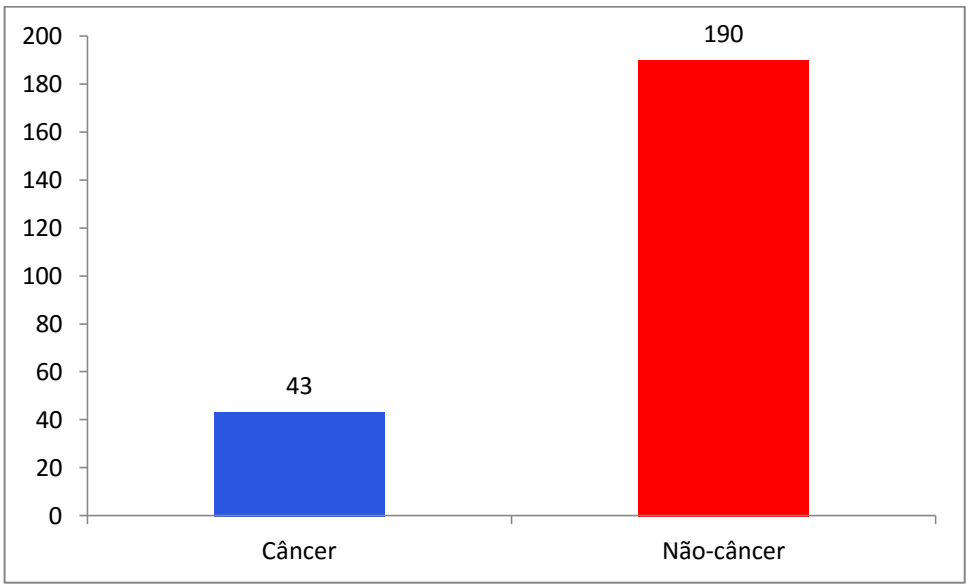
Tabela 16 - Resultados da classificação multiclasse através do WEKA e eliminando as possíveis combinações lineares.

| Classificador | Taxa de Acerto | Kappa |
|---|---|---|
| BayesNet | 47,11% | 0,295 |
| NaiveBayes | 47,75% | 0,303 |
| MLP | 57,37% | 0,431 |
| **SVM** | **63,78%** | **0,517** |
| RandomForest | 60,58% | 0,474 |
| RandomTree | 48,71% | 0,316 |
| Ibk (KNN) | 62,18% | 0,496 |

O uso de características sem dependência linear apresentou pouca variação nos resultados de taxa de acerto e Coeficiente Kappa para uma classificação multiclasse. Isso mostra que uma possível dependência linear entre as características, nesse caso, não afeta fortemente o processo de classificação.

Para a construção da base de dados da classificação binária, utilizou-se do mesmo procedimento para o balanceamento da amostra por meio de vetores sintéticos. A Figura 31 mostra o histograma das 233 amostras distribuídas nas Classes Câncer e Não-câncer.

Figura 31 - Histograma da amostra inicial em duas classes.




Como se pode observar, a Classe Não-Câncer é a amostra com maior frequência já que engloba os diagnósticos de “tumor benigno”, “cisto” e “normal”, composta por 190

imagens. A Tabela 17 mostra o número de vetores sintéticos necessários para o balanceamento da amostra.

Tabela 17 - Número de vetores sintéticos utilizados.

| **Classe** | **Quantidade de vetores sintéticos** |
|---|---|
| Câncer | 147 |
| Não-câncer | 0 |

Do mesmo modo para que a classificação multiclasse, após o balanceamento da base de dados, as 380 imagens resultantes foram classificadas através do *software* WEKA. A validação do classificador também fez o uso da validação cruzada com 5 subconjuntos. A Tabela 18 mostra os resultados conseguidos.

Tabela 18 - Resultados da classificação binária através do WEKA.

| **Classificador** | **Taxa de Acerto** | **Sensibilidade** | **Especificidade** | **Kappa** |
|---|---|---|---|---|
| BayesNet | 80,26% | 86,84% | 73,68% | 0,6053 |
| NaiveBayes | 81,05% | 91,58% | 70,53% | 0,6211 |
| MLP | 90,26% | 92,63% | 87,89% | 0,8053 |
| **SVM** | **93,42%** | **94,73%** | **92,10%** | **0,8684** |
| RandomForest | 90,79% | 90,52% | 91,05% | 0,8158 |
| RandomTree | 83,95% | 84,74% | 83,16% | 0,6789 |
| Ibk (KNN) | 88,68% | 96,31% | 81,05% | 0,7737 |

Assim como na classificação multiclasse, com o objetivo de eliminar as possíveis combinações lineares entre as características, foram utilizadas também as características citadas no Capítulo 3, e propostas para esse fim, e seus resultados utilizados para critérios de comparação. A Tabela 19 mostra as taxas de acertos e os respectivos coeficientes Kappa, para critérios de comparação.

Tabela 19 - Resultados da classificação binária através do WEKA e eliminando as possíveis combinações lineares

| **Classificador** | **Taxa de Acerto** | **Kappa** |
|---|---|---|
| BayesNet | 80,79% | 0,6158 |
| NaiveBayes | 81,84% | 0,6368 |
| MLP | 89,47% | 0,7895 |
| **SVM** | **93,15%** | **0,8632** |
| RandomForest | 91,05% | 0,8211 |
| RandomTree | 83,42% | 0,6684 |
| Ibk (KNN) | 89,47% | 0,7895 |

A Figura 32 mostra as matrizes de confusão dos resultados obtidos na classificação binária, mostrados na Tabela 19, com os classificadores indicados.

Figura 32 - Matrizes de confusão da classificação binária através do WEKA.

BayesNet

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 165 | 25 | 190 |
| Não-Câncer | 50 | 140 | 190 |
| | | Total: | 380 |

NaiveBayes

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 174 | 16 | 190 |
| Não-Câncer | 56 | 134 | 190 |
| | | Total: | 380 |

MLP

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 176 | 14 | 190 |
| Não-Câncer | 23 | 167 | 190 |
| | | Total: | 380 |

SMO

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 180 | 10 | 190 |
| Não-Câncer | 15 | 175 | 190 |
| | | Total: | 380 |

Random Forest

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 172 | 18 | 190 |
| Não-Câncer | 17 | 173 | 190 |
| | | Total: | 380 |

Random Tree

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 161 | 29 | 190 |
| Não-Câncer | 32 | 158 | 190 |
| | | Total: | 380 |

Ibk (KNN)

| | Câncer | Não-Câncer | Total de amostras |
|---|---|---|---|
| Câncer | 183 | 7 | 190 |
| Não-Câncer | 36 | 154 | 190 |
| | | Total: | 380 |

Entre os classificadores avaliados, destacam-se o classificador SVM utilizando uma função *Kernel* polinomial de grau 3, tanto para a classificação binária, quanto multiclasse. A Tabela 20, mostra resumidamente uma comparação entre os resultados obtidos através da base de dados da Metodologia 1 e dos melhores resultados obtidos na Metodologia 2.

Tabela 20 - Comparação entre os resultados obtidos na Metodologia 1 e na Metodologia 2

| | **Metodologia 1** | | **Metodologia 2** | |
|---|---|---|---|---|
| | Binário | Multiclasse | Binário | Multiclasse |
| Taxa de Acerto | 83,33% | 53,45% | 93,42% | 63,46% |
| Sensibilidade | 87,5% | 100% | 94,73% | 80,77% |
| Especificidade | 83,12% | 64,44% | 92,10% | 86,54% |

A escolha dos melhores resultados foi realizada levando em consideração as maiores taxas de acerto global e a sensibilidade à Classe Maligno. O Coeficiente Kappa foi decisivo nessa escolha.

Para a classificação multiclasse, nos melhores resultados obtidos, foi atingida uma sensibilidade de 80,77 % para à Classe Maligno, taxa de acerto de 63,46% e coeficiente Kappa de 0,5128, que pela Tabela 2 a classificação pode ser considerada como tendo uma concordância moderada. Já para a classificação binária foi conseguido uma taxa de acerto de 93,42%, uma sensibilidade de 94,73% para a Classe Câncer e coeficiente Kappa de 0,8684, que pode ser considerado como sendo uma classificação com concordância perfeita, de acordo com a Tabela 2, validando assim a escolha dos classificadores propostos e apresentados na presente dissertação.

A Tabela 21 mostra os resultados da Metodologia 2 em comparação a trabalhos anteriores do mesmo grupo de pesquisa.

Tabela 21 – Método proposto comparado ao estado da arte dos trabalhos desenvolvidos no grupo de pesquisa.

| | **Segmen-tação** | **Classifi-cador** | **Base de dados** | **Acertos** | **Sensibi-lidade** | **Especifi-cidade** | **Classes** |
|---|---|---|---|---|---|---|---|
| Dourado Neto (2014) | Automática | SVM | 234 | 79,49% | 67,44% | 82,20% | Câncer e Não-Câncer. |
| **Metodo-logia 2** | **Automática** | **SVM** | **380** | **93,42%** | **94,73%** | **92,10%** | **Câncer e Não-Câncer.** |
| Araújo (2014) | Semi-automática | Distância Euclidiana | 50 | 84% | 85,7% | 86% | Maligno, Benigno e Cisto. |
| Queiroz (2016) | Automática | SVM *one-vs-all* | 98 | 51,58% | 66,67% | 82,35% | Maligno, Benigno, Cisto e Normal. |
| **Metodo-logia 2** | **Automática** | **SVM *one-vs-one*** | **312** | **63,46%** | **80,77%** | **86,54%** | **Maligno, Benigno, Cisto e Normal.** |
| **Metodo-logia 2** | **Automática** | **SVM *one-vs-one*** | **234** | **68,38%** | **82,05%** | **87,18%** | **Maligno, Benigno e Cisto.** |

O uso dos vetores sintéticos, como explicado anteriormente, poderia comprometer o processo de classificação. Porém, neste caso, o balanceamento total e aumento do número de amostras utilizando estes vetores, tornou o processo de classificação mais eficiente, se comparado a Metodologia 1 e também a resultados anteriores do grupo de pesquisa, corroborando assim, seu uso para essa aplicação.

Na última linha da Tabela 21, foram obtidos resultados através Metodologia 2 com apenas três classes (Maligno, Benigno e Cisto) para critério de comparação com o trabalho realizado por Araújo (2014). Os resultados obtidos mostraram uma menor taxa de acerto em comparação com o do autor citado, porém obteve-se sensibilidade e especificidade semelhantes aos resultados do autor. Essa diferença significativa na taxa de acerto pode ser creditada ao tipo de segmentação realizada.

# 5 CONCLUSÕES E TRABALHOS FUTUROS

Tendo em vista os resultados obtidos, pode-se sugerir que a termografia pode ser usada como uma ferramenta auxiliar para a detecção de anomalias mamárias. A partir de uma segmentação automática, ocorreu uma extração de características fundamentada em medidas intervalares e medidas estatísticas que permitiram a classificação dos termogramas de mama em relação às anomalias.

Os classificadores utilizados mostraram desempenhos que validam o uso de imagens termográficas de mama para diagnóstico e para rastreamento. Numa abordagem para uma análise binária (Câncer/Não-Câncer), os resultados mais expressivos, apresentaram altas taxas de acertos e sensibilidade para a Classe Maligno, de 93,42% e de 94,73%, respectivamente. Para um diagnóstico envolvendo as quatro classes consideradas (Maligno, Benigno, Cisto e Normal), as taxas de acerto e sensibilidade obtiveram resultados menores em comparação à classificação binária. Tais taxas foram de 63,46% e 80,77%, respectivamente. Credita-se esse resultado a uma sobreposição entre as quatro classes, sobretudo devido à inserção das pacientes normais, o que provavelmente dificultou a classificação. Apesar disso, os resultados encontrados foram satisfatórios.

Considerando os resultados, a metodologia desenvolvida pode ser uma ferramenta de grande valia para rastreamento e/ou detecção precoce de câncer de mama, principalmente em pacientes não indicados para mamografia. O desbalanceamento da amostra em relação à quantidade de pacientes com tumores benignos pode prejudicar o processo de classificação, tornando o classificador mais tendencioso. O uso de vetores sintéticos permitiu um balanceamento da amostra, além de um aumento do número de amostras na base de dados.

O estudo demonstrou que uma base de dados ampliada e totalmente balanceada gerou melhores classificações, em relação a trabalhos anteriores do grupo de pesquisa. Além disso, o estudo mostrou também que o classificador SMO do *software* WEKA conseguiu resultados mais expressivos quanto ao diagnóstico de anomalias mamárias: tanto para uma abordagem binária, quanto para uma abordagem multiclasse.

Como já citado, estudos nessa área mostram que o classificar SVM é o que se mostra mais eficiente na classificação das imagens termográficas de mama. O SMO, que obteve melhores resultados no presente trabalho, nada mais é que um algoritmo para treinamento do SVM, indicando que os resultados obtidos estão de acordo com a literatura.

Como trabalhos futuros, pode-se sugerir:

- Testar e analisar o uso de funções geoestatísticas para extração de novas características relevantes das imagens por IR;
- Analisar técnicas de redução da dimensionalidade para eliminar ruídos no processo de classificação das patologias mamárias ao se incluir a Classe Normal na amostra de treino;
- Analisar e aperfeiçoar o uso de classificadores multiclasses e de validação cruzada;
- Averiguar se o uso da segmentação automática e segmentação semiautomática tem influência sobre a classificação da imagem;
- Gerenciar o uso do *software* livre WEKA através de arcabouço computacional, para agilizar o processo de extração de características;
- Analisar uma possível combinação de classificadores efetuada através de processos de otimização usando algoritmos genéticos e/ou teoria evolutiva;
- Estudar sobre o desempenho dos classificadores SVM em função dos Kernels, uma vez que o SVM vem se mostrando o melhor classificador;
- Explorar Métodos Bayesianos e baseados em Árvores.

# REFERÊNCIAS


ABE, S. **Support Vector Machines for Pattern Classification.** Springer Science & Business Media, 2010.

AFFONSO, E. T. F.; SILVA, A. M.; SILVA, M. P.; RODRIGUES, T. M. D.; MOITA, G. F. **Uso redes neurais MultiLayer Perceptron (MLP) em sistema de bloqueio de websites baseado em conteúdo.** Mecánica Computacional, v. XXIX, p. 9075-9090, 2010.

ARAÚJO, M. C. DE. **Uso de imagens termográficas para classificação de anormalidades de mama baseado em variáveis simbólicas intervalares**. p. 162. Tese (Doutorado) - Universidade Federal de Pernambuco, Recife, 2014.

ARAÚJO, M. C.; LIMA, R. C. F.; SOUZA, R. M. C. R. **Uso de imagens termográficas para classificação de anormalidades de mama**. Biblioteca Digital Brasileira de Computação, 2015.

AYAT, N.E.; CHERIET, M.; SUEN, C.Y. **Automatic model selection for the optimization of SVM kernels.** Pattern Recognition, v. 38, p. 1733 – 1745, 2005.

BELFORT, C. N. S.; MOTTA, S. de A. C. S.; SILVA, A. C. **Detecção de lesões decorrentes do câncer de mama em imagens termográficas utilizando funções geoestatísticas e SVM.** V Jornada de Informática do Maranhão, 2014.

BIM, C. R.; PELLOSO, S. M.; Carvalho, M. D. de B.; PREVIDELLI, I. T. S. **Diagnóstico precoce do câncer de mama e colo uterino em mulheres do município de Guarapuava, PR, Brasil.** Revista Escola Enfermagem, Universidade de São Paulo, v. 44, n. 4, p. 940-946, 2010.

BORCHARTT, T. B. **Análise de Imagens Termográficas para a Classificação de Alterações na Mama**. p. 115. Tese (Doutorado) - Universidade Federal Fluminense, Rio de Janeiro, 2013.

BORCHARTT, T. B.; RESMINI, R.; MOTTA, L. S.; CLUA, E. W.; CONCI, A.; VIANA, M. J.; SANTOS, L. C.; LIMA, R. C.; SANCHEZ, A. **Combining approaches for early diagnosis of breast diseases using thermal imaging.** International Journal of Innovative Computing and Applications, v. 4, n. 3-4, p.163–183, 2012.

BREIMAN, L. **Random Forests**. Machine Learning, v. 45, p. 5-32, 2001.

CAMPBELL, J. B. **Introduction to Remote Sensing**. 3rd ed. Taylor & Francis, 2002.

CARVALHO, B. P. R. de. **O estado da arte em métodos para reconhecimento de padrões: Support Vector Machine.** Congresso Nacional de Tecnologia da Informação e Comunicação (SUCESU), Belo Horizonte, MG, 2005.

CHAKRABORTY, D.; SARKAR, A.; MAULIK, U. **A new isotropic locality improved kernel for pattern classifications in remote sensing imagery.** Spatial Statistics, v. 17, p. 71–82, 2016.

CHALA, L. F.; BARROS, N. de. **Avaliação das mamas com métodos de imagem**. Radiologia Brasileira, v. 40, n. 1, p. IV-VI, 2007.

CHEN, C.H. **Image Processing for Remote Sensing**, CRC Press, 2007.

COHEN, J. **A Coeficient of Agreement for Nominal Scales.** Educational and Measurment. v. XX, n. 1, p. 37-46, 1960.

CORTES, C.; VAPNIK, V. **Support-Vector Networks.** Machine Learning, v. 20, p. 273-297, 1995.

DOUGHERTY, G. **Digital Image Processing for Medical Applications**. 1. ed. Cambridge University Press, New York, 2009.

DOURADO NETO, H. M. **Segmentação e análise automáticas de termogramas : um método auxiliar na detecção do câncer de mama um método auxiliar na detecção do câncer de mama.** p. 99. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Recife, 2014.

DUA, S.; CHOWRIAPPA, P. **Data Mining for Bioinformatics**. CRC Press, 2012.

DUDA, R. O.; HART, P. E., STORK, D. G. **Pattern Classification**. 2nd ed. John Wiley & Sons, 2012.

FRANCIS, S. V.; SASIKALA, M.; BHARATHI, G. B.; JAIPURKAR, S. D. **Breast cancer detection in rotational thermography images using texture features**. Infrared Physics & Technology, v. 67, p. 490-496, 2014.

FRANK, E.; HALL, M.; HOLLMAN, G.; KIRKLY, R.; PFAHRINGER, B; WITTEN, I. H.; TRIGG, L. **WEKA: A Machine Learning Workbench for Data Mining.** The Data Mining and Knowledge Discovery Handbook, p. 1305-1314, 2005.

FRANK, E.; HALL, M.; TRIGG, L.; HOLMES, G.; WITTEN, I. H. **Data mining in bioinformatics using Weka.** Bioinformatics Applications Note, v. 20, n. 15, p. 2479–2481, 2004.

GOLDSCHMIDT, R; BEZERRA, E., PASSOS, E. **Data Mining: Conceitos, Técnicas, Algoritmos, Orientações e Aplicações.** 2. ed. Elsevier Brasil, 2015.

GOLESTANI, M.; TAVAKOL, M. E.; NG, E.Y.K. **Level set method for segmentation of infrared breast thermograms.** Clinical Science, v. 13, p. 241–251, 2014.

GONZALEZ, R. C.; WOODS, R. E.; EDDINS, E. L. **Digital Image Processing Using Matlab**. 3rd ed, New Jersey: Pearson Prentice Hall, 2008.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B. **The WEKA Data Mining Software: An Update.** SIGKDD Explorations, v. 11, p. 10-18.

HEAD, J.; WANG, F.; LIPARI, C.; ELLIOTT, R. **The important role of infrared imaging in breast cancer**. IEEE Eng. Med. Biol. Mag., v. 19, p. 52–57, 2000.

HERBRICH, R. **Learning Kernel Classifiers: Theory and Algorithms**. MIT Press, 2001.

INCA. **Instituto Nacional de Câncer**. Estimativa 2016: Incidência do Câncer no Brasil. Rio de Janeiro, Brasil: Ministério da Saúde, 2015.

ISHII, N.; HOKI, Y.; OKADA, Y.; BAO, Y. **Nearest Neighbor Classification by Relearning, Intelligent Data Engineering and Automated Learning**. v. 5788, p. 42-49, 2009.

JOHN, G. H.; LANGLEY, P. **Estimating Continuous Distributions in Bayesian Classifiers**. Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995.

JORDE, L. B.; CAREY, J. C.; BAMSHAD, M. J.; WHITE, R. L. **Genética Médica**, 3 ed., Elsevier Brasil, 2004.

KALMEGH, S. R. **Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data**. International Journal of Emerging Technology and Advanced Engineering, v. 5, p. 507-517, 2015.

KAMATH, D.; KAMATH, S.; PRASAD, K.; RAJAGOPAL, K. V. **Segmentation of Breast Thermogram Images for the Detection of Breast Cancer – A Projection Profile Approach**. Journal of Image and Graphics, v. 3, n. 1, p. 47-51, 2015.

KEERTHI, S. S.; SHEVADE, S. K.; BHATTACHARYYA, C.; MURTHY, K. R. K. **Improvements to Platt's SMO Algorithm for SVM Classifier Design**. Neural Computation, v. 13, p. 637-649, 2001.

KOTZ S., JOHNSON N. L. **Encyclopedia of statistical sciences**. New York: John Wiley& Sons. v.4, p.352-354, 1983.

KRESSEL, U. H.-G. **Advances in kernel methods table of contents**. MIT Press Cambridge, p. 255-268, 1999.

KUMAR, Y.; SAHOO, G. **Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA**. Information Technology and Computer Science, v. 7, p. 43-49, 2012.

LAHIRI, B.B., BAGAVATHIAPPAN S., JAYAKUMAR T., PHILIP, J. **Medical applications of infrared thermography: A review.** International Journal Infrared Physics and Technology. p. 222-232, 2012.

LIU, J. G.; MASON, P. J. **Image Processing and GIS for Remote Sensing: Techniques and Applications.** Wiley Blackwell, 2016.

LORENA, A. C.; CARVALHO, A. C. P. L. F. **Uma Introdução às Support Vector Machines.** Revista de Informática Teórica e Aplicada, v. 14, n. 2, p. 43–67, 2007.

MA, Y.; GUO, G. **Support Vector Machines Applications**. Springer Science & Business Media, 2014.

MAHMOUDZADEH, E.; MONTAZERI, M. A.; ZEKRI, M.; SADRI, S. **Extended hidden Markov model for optimized segmentation of breast**. Infrared Physics & Technology, v. 72, p. 19–28, 2015.

MALZYNER, A.; CAPONERO, R. **Câncer e Prevenção**. 1 ed. São Paulo, MG Editores, 2013.

MARINS, J. C. B.; FERNÁNDEZ-CUEVAS, I.; ARNAIZ-LASTRAS, J.; FERNANDES, A. A.; SILLERO-QUINTANA, M. **Aplicaciones de la termografía infrarroja en el deporte. Una revisión.** Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte. v. 15, n. 60, p. 805-824, 2015.

MOURA, G. B. **Rede Probabilísticas Fuzzy Naive Bayes**. p. 79. Dissertação (Mestrado) – Universidade Federal de Santa Catarina, 2016.

NASCIMENTO, R. F. F.; ALCÂNTARA, E. H. d.; KAMPEl, M.; STECH, J. L.; NOVO, E. M. L. M.; FONSECA, L. M. G. **O algoritmo *Support Vector Machines* (*SVM*): avaliação da separação ótima de classes em imagens CCD-CBERS-2**. Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, p. 2079-2086, Natal, Brasil, 2009.

NEGRI, R. G.; SANT'ANNA, S. J. S.; DUTRA, L. V. **Aplicação de Modelos de Aprendizado Semissupervisionado na Classificação de Imagens de Sensoriamento Remoto**. Revista de Informática Teórica e Aplicada, v. 20, n. 2, p. 32-54, 2013.

NG, E.Y.K. **A review of thermography as promising non-invasive detection modality for breast tumor.** International Journal of Thermal Sciences, v. 48, p. 849–859, 2009.

OLIVEIRA, J. de A.; DUTRA, L. V.; RENNÓ, C. D.; SANTOS, P. S. **Extração de Atributos de Forma para Classificação de Imagens de Alta Resolução do Satélite HRC/CBERS-2B.** Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, p. 7015-7022, Natal, Brasil, 2009.

OLIVEIRA, M. M. **Desenvolvimento de protocolo e construção de um aparato mecânico para padronização da aquisição de imagens termográficas de mama**. p. 104. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, 2012.

PERROCA, M. G., GAIDZINSKI, R. R. **Avaliando a confiabilidade interavaliadores de um instrumento para classificação de pacientes - coeficiente Kappa.** Revista Escola Enfermagem, Universidade de São Paulo, v. 37, n. 1, p. 72-80, 2003.

PINHEIRO, J. I. D.; CUNHA, S. B.; CARVAJAL, S. R.; GOMES, G. C. **Estatística básica: A arte de trabalhar com dados**. 2. ed. Elsevier Brasil, 2015.

QUEIROZ, K. F. F. C. Q. **Desenvolvimento e implementação de uma ferramenta computacional de uso médico para análise de imagens termográficas.** p. 103. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Recife, 2016.

RASTGHALAM, R.; POURGHASSEM, H**. Breast cancer detection using MRF-based probable texture feature and decision-level fusion-based classification using HMM on thermography images**. Pattern Recognition, v. 51, p. 176–186, 2016.

RESMINI, R**. Análise de imagens térmicas da mama usando descritores de textura**. 81 p. Dissertação (Mestrado) – Universidade Federal Fluminense, Niterói, RJ, 2011.

RESMINI, R.; BORCHARTT, T. B.; CONCI, A.; LIMA, R. C. F. **Auxílio ao Diagnóstico Precoce de Patologias da Mama Usando Imagens Térmicas e Técnicas de Mineração de Dados**. Computer on the Beach, p. 305-314, 2012.

ROBERTO, J. V. B.; SOUZA, B. B. **Utilização da termografia de infravermelho na medicina veterinária e na produção animal.** Journal of Animal Behaviour and Biometeorology, v. 2, n.3, p.73-84, 2014.

SANTOS, D. A. DOS. **Uma técnica baseada em imagens para correção da postura de pacientes na aquisição de termografias**. p. 85. Dissertação (Mestrado) – Universidade Federal Fluminense, 2012.

SCHAEFER, G.; NAKASHIMA, T. **Strategies for Addressing Class Imbalance in Ensemble Classification of Thermography Breast Cancer Features.** IEEE Congress on Evolutionary Computation, p. 2362–2367, 2015.

SCHAEFER, G.; NAKASHIMA, T.; ZAVISEK, M. **Analysis of breast thermograms based on statistical image features and hybrid fuzzy classification**. Springer, p. 753–762, 2008.

SERRANO, R. C. **Viabilidade do uso do coeficiente de Hurst e da lacunaridade no diagnóstico precoce de patologias da mama**. 76 p. Dissertação de Mestrado, Universidade Federal Fluminense, Niterói, RJ, 2010.

SIVANANDAM, S. N.; SUMATHI, S.; DEEPA, S. N., **Introduction to Neural Networks using Matlab 6.0**. New Delhi: Tata McGraw-Hill, 2006.

SOMAN, K.P.; LOGANATHAN, R.; AJAY, V. **Machine Learning with SVM and Other Kernel Methods.** New Delhi, PHI Learning, 2011.

STATNIKOV, A.; ALIFERIS, C. F.; HARDIN, D. P.; GUYON, I. **A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and methods.** World Scientific, 2011.

SYED, M. R; SYED, S. N. **Handbook of Research on Modern Systems Analysis and Design Technologies and Applications**. IGI Global, 2008.

TAKAHASHI, A. **Máquinas de Vetores-Suporte Intervalar**. p. 61. Tese (Doutorado) - Universidade Federal do Rio Grande do Norte, Natal, 2012.

TAVAKOL, M. E.; NG, E.Y.K.; CHANDRAN, V.; RABBANI, H. **Separable and non-separable discrete wavelet transform based texture features and image classification of breast thermograms.** Infrared Physics & Technology, v. 61, p. 274–286, 2013.

TEPPERWEIN, K. **O que a doença quer dizer**. São Paulo, Editora Ground, 2002.

VAPNIK, V.N. **The Nature of Statistical Learning Theory**. 2nd ed., USA: Springer, 1999.

VIERA, A. J.; MD; GARRETT, J. M. **Understanding Interobserver Agreement: The Kappa Statistic.** Family Medicine Journal, v. 37, n. 5, p. 360-363, 2005.

WHO. **World Health Organization**. International Agency for Research on Cancer. Globocan, 2012.

YATES, R. B.; RIBEIRO NETO, B. **Recuperação de informação: conceitos e tecnologia das máquinas de busca.** 2nd ed. Bookman Editora, 2013.

ZHANG, H. **The Optimality of Naive Bayes.** American Association for Artificial Intelligence, 2004.

ZHONG, X. **The research and application of web log mining based on the platform weka.** Procedia Engineering, v. 15, p. 4073 – 4078, 2011.