



**Samuel Fontes  
da Costa**

**CAD para Mamografia**





**Samuel Fontes  
da Costa**

## **CAD para Mamografia**

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações, realizada sob a orientação científica do Dr. Augusto Marques Ferreira da Silva, Professor Auxiliar do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro e da Dra. Ana Maria Perfeito Tomé, Professora Associada do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro.



**o júri**

Presidente

**Prof. Dr. João Paulo Trigueiros da Silva Cunha**

Professor Associado da Universidade de Aveiro

Vogais

**Prof. Dr. Augusto Marques Ferreira Da Silva**

Professor Auxiliar da Universidade de Aveiro

**Prof. Dra. Ana Maria Perfeito Tomé**

Professora Associada da Universidade de Aveiro

**Dr. Miguel Angel Guevara Lopez**

Professor Titular da Facultad de Informática da Universidade de Ciego de Ávila de Cuba



## **Agradecimentos**

Ao Professor Augusto Silva e à Professora Ana Maria Tomé pela orientação, pela forma exigente, crítica e criativa de analisar cada fase do trabalho, pelos conhecimentos transmitidos e apoio constante durante todo o período de construção deste trabalho.

Aos meus pais, pessoas fundamentais na minha vida, que sempre me encorajaram a enfrentar os desafios, que sempre me deram tudo, e de quem muito me orgulho, o meu eterno agradecimento.

À minha querida irmã Ema por toda a ajuda e disponibilidade demonstrada e por toda a sua ternura e amor.

Ao meu querido irmão Ruben pela paciência, pelo seu humor, alegria e carinho demonstrado.

Aos meus avós pela ternura e carinho com que sempre me educaram.

Aos meus padrinhos pelo apoio, incentivo e preocupação com o sucesso na realização deste trabalho e pelo meu bem-estar.

À restante família (tios e primos) pelo apreço e força constantemente transmitida para que prosseguisse esta etapa da minha vida na conquista de um futuro melhor. Uma palavra ainda, de gratidão à minha tia Clara pela disponibilidade demonstrada na revisão e correcção ortográfica da tese.

A todos os meus colegas e amigos, em particular, ao André Reis, Carlos Morais, Luís Manuel e Ruben Ferreira pela amizade e companheirismo, e ainda à Helga e Daniela pelo apoio prestado na revisão e tradução da tese.

Ao Mestre Samuel Silva e à Mestre Ana Rita, o meu obrigado pelas dicas e ajudas na resolução de problemas durante a execução deste trabalho.

À Lara pelo amor, compreensão, ternura, carinho e por estar sempre ao meu lado nos bons momentos e nos menos bons. Para quem sabe que o meu mundo está todo nas suas mãos.

O meu obrigado a todos!





## Resumo

O cancro da mama é actualmente uma das principais causas de morte entre as mulheres.

Os sistemas de diagnóstico auxiliado por computador permitem o apoio aos radiologistas, com o objectivo de melhorar a precisão do diagnóstico, fornecendo uma “segunda opinião”.

Este trabalho apresenta uma metodologia para classificação de nódulos mamários com base em técnicas de processamento de imagens digitais para a extracção de características de textura e forma.

As imagens mamográficas digitais utilizadas foram obtidas a partir da base de dados de domínio público miniMIAS (do inglês, *Mammographic Image Analysis Society*).

Esta base de dados é constituída por um total de 322 mamogramas, onde foram recortados 67 regiões de interesse que continham massas.

O contorno das massas foram interactivamente delineados, com o suporte do algoritmo *live-wire*.

Assim, foram extraídas cinco características de forma baseadas no contorno e oito características de textura baseadas na matriz de co-ocorrência entre níveis de cinza, com o objectivo de caracterizar as imagens.

É proposta a aplicação e estudo de redes neuronais artificiais, de modo a classificar as massas em benignas e malignas.

Assim sendo, foram testadas várias redes neuronais, variando os conjuntos de entrada e as topologias das redes de modo a obter a melhor classificação. Os resultados foram comparados em termos da área Az sob a curva ROC.



## **Abstract**

Breast cancer is currently one of the leading causes of death among women. Systems designed for computer-aided diagnosis provide support to radiologists, with the aim of improving the accuracy of diagnosis by providing a "second opinion".

This thesis presents a methodology for classification of breast nodules, based on techniques for processing digital images developed for the extraction of texture and shape features.

The digital mammographic images were obtained from a public domain database, "miniMIAS".

This database as a total of 322 mammograms, from where 67 interest zones that contained masses were selected.

The contours of these masses were interactively outlined with support from a "live-wire" algorithm.

Then five shape features were computed as well as eight texture features based on the co-occurrence matrix.

We assessed the application of artificial neural networks, in order to classify the masses in benign or malignant.

Several neural networks topologies and learning algorithms were evaluated in order to obtain the best classify performance.

The results were then submitted to a comparison in terms of the Az area and under the ROC curve.



# Lista de Figuras

2.1	Equipamento utilizado na Ultra-Sonografia . . . . .	4
2.2	Equipamento utilizado na Ressonância Magnética . . . . .	5
2.3	Estrutura Interna do Peito . . . . .	6
2.4	Incidências obtidas girando o Mamógrafo . . . . .	7
2.5	Esquema de um Mamógrafo . . . . .	8
2.6	Características de a)Forma e b)Margem em que se dividem os vários tipos de massas. Adaptado de [1] . . . . .	10
3.1	Convenção utilizada para os eixos em Processamento de Imagens . . . . .	17
3.2	Variações angulares utilizadas na obtenção da matriz SGLD para uma distância $d=1$ . . . . .	21
3.3	Cálculo matriz SGLD onde a)Imagem 4x4 pixel com 4 valores de nível de cinza, b)Construção da matriz SGLD, onde $(i, j)$ indica os pares de níveis de cinza possíveis da imagem em questão e c)Cálculo das 4 matrizes SGLD quando $d=1$	22
3.4	Neurónio proposto por <i>McCulloch e Pitts</i> . . . . .	25
3.5	Constituição da Célula Neuronal . . . . .	26
3.6	Camadas de uma RNA genérica . . . . .	28
3.7	Esquema de funcionamento do algoritmo <i>backpropagation</i> . . . . .	29
4.1	Diagrama de blocos da estrutura do esquema CAD implementado . . . . .	31
4.2	Mamograma original e extracção da ROI de uma massa benigna circunscrita	33
4.3	Diagrama de blocos da Extracção de Características . . . . .	33
4.4	a) Segmentação de uma ROI com a técnica <i>Live-wire</i> b) Aplicação da função <i>poly2mask</i> . . . . .	34
4.5	Diagrama de blocos dos procedimentos a efectuar no vector de características antes da entrada nas RNAs . . . . .	35
4.6	Método <i>Leave-One-Out</i> . Adaptado de [2] . . . . .	37
4.7	Definição de verdadeiro e falso positivos e negativos. Adaptado de [2] . . . . .	40
4.8	Curva ROC ( <i>Receiver Operating Characteristics Curve</i> ) . . . . .	42
5.1	Curvas ROC dos resultados obtidos com a rede MLP para a topologia a) 5/2/1, b) 8/3/1 e c) 13/3/1 . . . . .	46
5.2	Avaliação das características . . . . .	48



# Lista de Tabelas

4.1	Composição da base de dados (B: Benigno, M: Maligno). . . . .	32
5.1	Treinos realizados com o algoritmo <i>Perceptron</i> . . . . .	44
5.2	Treinos realizados com o algoritmo <i>Pocket</i> . . . . .	45
5.3	Topologias utilizadas para a rede neuronal MLP . . . . .	45
5.4	Resultados obtidos para cada conjunto de dados com a rede MLP . . . . .	46
5.5	Resultados obtidos com a técnica <i>Leave-One-Out</i> modificada . . . . .	47





# Glossário

**ACR** *Colégio Americano de Radiologia.*

**Az** *Área sob a curva ROC.*

**BI-RADS** *Breast Imaging Reporting and Data System.*

**CAD** *Computer-Aided Diagnosis.*

**FDA** *Food and Drug Administration.*

**GLCM** *Grey Level Co-occurrence Matrix.*

**OMS** *Organização Mundial de Saúde.*

**MIAS** *Mammographic Image Analysis Society.*

**MLP** *Multi-Layer Perceptron.*

**RBF** *Radial Basis Function.*

**RNA** *Redes Neurais Artificiais.*

**ROC** *Receiver Operating Characteristic.*

**ROI** *Região de Interesse.*

**SGLD** *Spatial Grey Level Dependence.*

**SOM** *Mapas Auto-Organizáveis.*



# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Objectivos . . . . .	2
1.2	Estrutura do Trabalho . . . . .	2
<b>2</b>	<b>Enquadramento</b>	<b>3</b>
2.1	O Cancro da Mama . . . . .	3
2.2	Morfologia do Peito . . . . .	5
2.3	Factores de Risco . . . . .	5
2.4	Mamografia . . . . .	6
2.4.1	O Sistema Mamográfico . . . . .	7
2.4.2	Mamografia Digital vs Mamografia Tradicional . . . . .	8
2.4.3	Lesões Detectadas na Mamografia . . . . .	9
2.5	Classificação BI-RADS . . . . .	12
<b>3</b>	<b>Diagnóstico Auxiliado por Computador</b>	<b>15</b>
3.1	Casos de Estudo . . . . .	15
3.2	Imagem Digital . . . . .	17
3.3	Extracção de Características . . . . .	18
3.3.1	Características de Forma . . . . .	18
3.3.2	Características de Textura . . . . .	19
3.4	Redes Neurais Artificiais . . . . .	25
3.4.1	Breve Cronologia . . . . .	25
3.4.2	Estrutura do Neurónio Biológico . . . . .	25
3.4.3	Funcionamento do Neurónio Artificial . . . . .	26
3.4.4	Caracterização das RNAs . . . . .	27
3.4.5	O algoritmo <i>Backpropagation</i> . . . . .	29
<b>4</b>	<b>Material e Métodos</b>	<b>31</b>
4.1	Base de dados Mamográfica . . . . .	31
4.2	Seleção de ROI . . . . .	32
4.3	Extracção de Características . . . . .	32
4.3.1	Características de Forma . . . . .	33
4.3.2	Características de Textura . . . . .	34
4.4	Análise e Tratamento dos dados . . . . .	35
4.4.1	Tratamento dos <i>Outliers</i> . . . . .	35
4.4.2	Normalização . . . . .	36

4.4.3	Separação dos conjuntos de treino e teste . . . . .	36
4.5	Redes Neurais Artificiais . . . . .	36
4.5.1	Estrutura Linear . . . . .	36
4.5.2	Estrutura Multicamada . . . . .	38
4.6	Medidas de Desempenho . . . . .	39
4.6.1	Verdadeiro e Falso, Positivos ou Negativos . . . . .	40
4.6.2	Sensibilidade, Especificidade e Precisão . . . . .	40
4.6.3	Curvas ROC . . . . .	41
<b>5</b>	<b>Resultados e discussão</b>	<b>43</b>
5.1	Descrição dos Dados de Entrada . . . . .	43
5.2	Classificadores Lineares . . . . .	43
5.3	Classificadores Não Lineares . . . . .	44
5.4	Avaliação das Características de Forma . . . . .	47
<b>6</b>	<b>Conclusões</b>	<b>49</b>
6.1	Trabalhos Futuros . . . . .	50
	<b>Referências Bibliográficas</b>	<b>53</b>
	<b>Apêndice A</b>	<b>55</b>

# Capítulo 1

## Introdução

Actualmente, o cancro da mama é uma das principais causas de morte entre as mulheres. De facto, segundo a *Organização Mundial de Saúde* (OMS), anualmente e à escala mundial, cerca de um milhão de mulheres descobre que tem cancro da mama. [3, 4, 5]

A nível nacional sabe-se que, com algumas diferenças regionais, o número de novos casos oscila entre 70 e 80 por cada 100000 mulheres por ano, número que tenderá a aumentar, sendo a principal causa de morte por neoplasia no sexo feminino. [3, 4, 5]

De salientar o facto, de que o cancro da mama, apesar de ser uma doença tipicamente feminina, atinge cerca de 1% da população masculina mundial[6, 7].

O diagnóstico precoce por meio do rastreio mamográfico é assim fundamental, pois aumenta as hipóteses de cura, evitando que o cancro se espalhe para outras partes do corpo, favorecendo o prognóstico, a recuperação e a reabilitação.

A mamografia é um dos métodos actuais de inquestionável importância no diagnóstico precoce pois é considerado o método mais eficaz para a detecção do cancro da mama ainda impalpável devido à sua alta sensibilidade.

Embora a mamografia seja considerada o principal método de auxílio ao diagnóstico do cancro da mama, a análise deste tipo de imagem é uma tarefa bastante difícil.

Além da dificuldade na interpretação destas lesões, devido sobretudo à grande variedade dos formatos das lesões, vários factores podem ainda influenciar a qualidade das imagens mamográficas, tais como a composição do parênquima mamário (por exemplo, as mulheres mais jovens apresentam um maior tecido glandular) ou ainda os efeitos dos vários tipos de ruído, nomeadamente o ruído radiológico e o ruído anatómico.

Distorções na análise e posterior erro na avaliação radiológica podem ter efeitos directos sobre a saúde do paciente além de que financeiramente acarretam custos elevados[8].

Veja-se por exemplo, que muitos falsos-positivos podem levar a um acréscimo de biópsias desnecessárias, exame que além de invasivo pode ser traumático para a paciente.

As ferramentas *Computer-Aided Diagnosis* (CAD) permitem o apoio aos radiologistas pois melhoraram a precisão do diagnóstico, ajudando na identificação de lesões, alertando o radiologista para que examine mais detalhadamente, padrões suspeitos detectados pelo sistema.

Estas permitem o processamento de imagens para melhorar a sua qualidade, por exemplo, aumentando o contraste, permitem segmentar, parametrizar (obter características), classificar e analisar.

Este trabalho contempla a aplicação e estudo do respectivo desempenho de 3 redes neuronais no processo de classificação das mamografias em benignas ou malignas.

Neste trabalho vamos recorrer a vários índices de mérito com destaque para a análise da curva característica de resposta do observador (curva *Receiver Operating Characteristic* (ROC)). O principal parâmetro para se proceder à comparação de desempenho entre duas curvas ROC é a sua área. Esta área pode ser vista como a capacidade do sistema ser, simultaneamente sensível e específico, sendo que, quanto maior for, melhor será a precisão diagnóstica da metodologia em análise.

## 1.1 Objectivos

As redes neuronais artificiais são frequentemente utilizados como classificadores nos esquemas CAD e o seu desempenho depende dos processos anteriores à classificação, como a aquisição do padrão mamográfico, pré-processamento e a extracção e selecção das *features*[9, 10]. Este trabalho tem como objectivo o estudo de técnicas CAD em mamografia e a implementação em Matlab[11] de um classificador baseado nas redes neuronais a partir de uma base de dados de referência bem como a respectiva validação através do desenvolvimento de estudos ROC.

## 1.2 Estrutura do Trabalho

O presente trabalho está estruturado em 6 capítulos.

- No capítulo 1, são apresentadas algumas considerações iniciais sobre o trabalho, expondo igualmente os objectivos pretendidos.
- No capítulo 2, são apresentadas aspectos principais relativos ao cancro da mama e à imagem mamográfica, abrangendo conceitos sobre a anatomia, histologia e patologias das mamas.
- O capítulo 3, expõe os métodos teóricos para descrever textura e forma dos nódulos, assim como uma abordagem teórica às redes neuronais artificiais.
- O capítulo 4, aborda os materiais e métodos utilizados assim como as várias etapas implementadas até à classificação dos achados mamográficos.
- No capítulo 5, são expostos os resultados alcançados para diversos conjuntos de treino e teste, utilizando 3 redes neuronais de diferentes topologias.
- No capítulo 6, expõem-se as conclusões do trabalho e discutem-se trabalhos futuros.

## Capítulo 2

# Enquadramento

### 2.1 O Cancro da Mama

Actualmente, o cancro da mama é uma das principais causas de morte entre as mulheres. De facto, segundo a Organização Mundial de Saúde, anualmente e à escala mundial, cerca de um milhão de mulheres descobre que tem cancro da mama. [3, 4, 5]

A nível nacional sabe-se que, com algumas diferenças regionais, o número de novos casos oscila entre 70 e 80 por cada 100000 mulheres por ano, número que tenderá a aumentar, sendo a principal causa de morte por neoplasia no sexo feminino. [3, 4, 5]

De salientar o facto, de que o cancro da mama, apesar de ser uma doença tipicamente feminina, atinge cerca de 1% da população masculina mundial[6, 7].

O entendimento desta realidade conduz à compreensão do quanto importante e decisivo pode ser a detecção precoce desta doença, tanto mais por ser inviável predizer a sua ocorrência.

O cancro da mama tem origem quando um nódulo, constituído por diversas células da mama, deixa de ter a capacidade de limitar e controlar o seu crescimento, dividindo-se de uma forma anormal, descontrolada e a grande velocidade, desencadeando o aparecimento de massas celulares, denominadas neoplasias malignas.

Estas células não guardam semelhanças com as que lhe deram origem e têm capacidade para se alastrar para outras partes do corpo.

O diagnóstico precoce por meio do rastreio mamográfico é assim fundamental, pois aumenta as hipóteses de cura, evitando que o cancro se espalhe para as outras partes do corpo, favorecendo o prognóstico, a recuperação e a reabilitação.

Assim sendo, destaca-se para o rastreio do carcinoma da mama, como medida de detecção precoce, o auto-exame, o exame clínico e as técnicas de obtenção de imagem da estrutura interna do peito.

O auto-exame é o exame das mamas efectuado pela própria mulher. Este não substitui o exame físico realizado por um profissional de saúde.

O exame clínico é um exame realizado por um profissional médico, sendo um procedimento periódico a que as mulheres se devem submeter, periodicidade essa que depende, principalmente, da idade.

Com o objectivo de melhorar as hipóteses de detecção do possível tumor, principalmente num estado ainda impalpável, usam-se técnicas de obtenção da imagem do peito, colaborando assim para um diagnóstico mais preciso e eficaz, sendo de destacar principalmente a mamografia, ressonância magnética e a ultra-sonografia.

### **Ultra-Sonografia (figura 2.1)**

Não é o método indicado no rastreio do cancro da mama, porque apresenta alta taxa de falsos-negativos devido a ser bastante limitado na detecção de microcalcificações não associadas a nódulos. A taxa de falsos-positivos também é elevada.



Figura 2.1: Equipamento utilizado na Ultra-Sonografia

Apresenta ainda limitações na avaliação das calcificações devido à elevada dependência do operador e do equipamento, não permitindo a melhor avaliação da quantidade e sua morfologia.

### **Ressonância Magnética (figura 2.2)**

A imagem por ressonância magnética mostra-se muito útil na avaliação de várias doenças mamárias, porém, o seu uso como método de rastreio não é, no entanto, aconselhado porque possui um custo elevado e uma baixa especificidade.

Actualmente, a mamografia é considerada o método mais eficaz na detecção e consequentemente será abordado com mais detalhe.





Figura 2.2: Equipamento utilizado na Ressonância Magnética

## 2.2 Morfologia do Peito

As mamas são constituídas por três tipos de tecido (adiposo, glandular e fibroso) que foram adaptados para secretar leite após o parto da mulher.

As glândulas que produzem leite são chamadas lóbulos e os tubos que interligam estes aos mamilos são designados por ductos.

Os homens também possuem lóbulos e mamilo, mas não há produção de leite devido à falta da hormona feminina estrogénio.

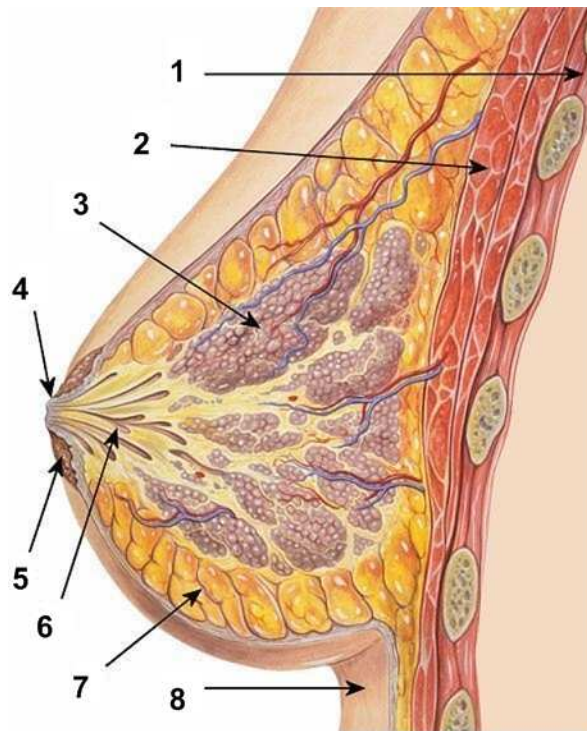
A figura 2.3 apresenta um esboço dos principais aspectos da anatomia da mama.

## 2.3 Factores de Risco

Infelizmente, não são totalmente conhecidos os factores que podem aumentar as hipóteses de contrair o cancro da mama.

Os factores que podem aumentar a probabilidade de desenvolver esta doença incluem:

- Idade, uma vez que a possibilidade de ter cancro da mama aumenta com o aumento da idade.
- Historial familiar, principalmente em parentes de 1º grau. Estima-se que entre 5% a 10% dos casos de cancro da mama são hereditários[5].
- A combinação de mutações genéticas, o estilo e hábitos de vida e até a influência ambiental podem levar à mudança dos genes nas células mamárias.
- A raça, uma vez que o cancro da mama ocorre com maior frequência em mulheres caucasianas (brancas), em comparação com as mulheres latinas, asiáticas ou afro-americanas.
- Primeira menstruação precoce (antes dos 12 anos de idade)
- Entrada na menopausa tardiamente (após os 55 anos)



Legenda: 1 - Caixa Torácica  
 2 - Músculo Peitoral  
 3 - Lóbulos  
 4 - Mamilo  
 5 - Aréola  
 6 - Ductos  
 7 - Tecido Adiposo  
 8 - Pele

Figura 2.3: Estrutura Interna do Peito

- Primeira gravidez ser depois dos 30 anos
- As mulheres que chegam à meia idade sem terem filhos

## 2.4 Mamografia

A mamografia é um dos métodos actuais não invasivos de inquestionável importância no diagnóstico precoce pois é considerado o método mais eficaz para a detecção do cancro da mama ainda impalpável devido à sua alta sensibilidade.

Assume-se pois, como determinante no rastreio de mulheres assintomáticas, com o objectivo de detectar o cancro em estágios mais precoces antes que a própria mulher ou o médico possam detectar no auto-exame ou no exame clínico.

Assim sendo, a mamografia provou ser capaz de reduzir as mortes por cancro da mama, redução essa associada ao estado precoce da lesão aquando o diagnóstico.

É de consenso geral, o rastreio anual a mulheres a partir dos 50 anos.

Com base no conhecimento da anatomia, os equipamentos mamográficos actuais utilizam principalmente duas projecções de cada peito.

- A crânio caudal (CC), onde o feixe de raios-X incide na direcção do eixo da coluna vertebral e no sentido do crânio para os pés.
- A médio lateral oblíqua (MLO), onde os raios-X incidem na direcção perpendicular ou oblíqua à projecção crânio-caudal.

Além destas incidências é possível girar o mamógrafo de modo a obter incidências de qualquer direcção (ver figura 2.4).

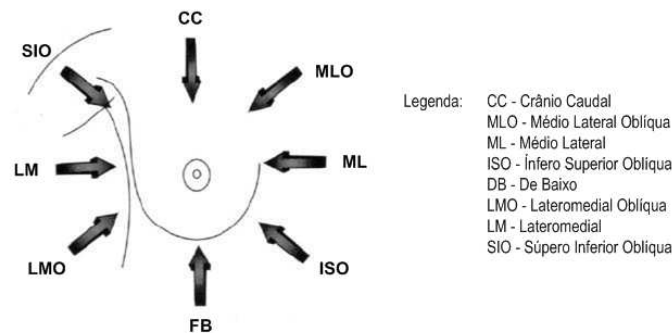


Figura 2.4: Incidências obtidas girando o Mamógrafo

O principal problema na detecção precoce são as mamas radiograficamente densas, onde em geral, o cancro é encontrado num estado mais avançado.

Assim, um aspecto importante a destacar refere-se à composição do tecido mamário que exerce influência na nitidez da imagem mamográfica (por exemplo, as mulheres mais jovens apresentam um maior tecido glandular).

### 2.4.1 O Sistema Mamográfico

A mamografia constitui uma forma particular de radiografia, que trabalha com níveis de tensões e correntes em intervalos específicos e destina-se a produzir imagens detalhadas das estruturas internas do peito a fim de diagnosticar a presença ou ausência de estruturas que possam indicar tumores.

O equipamento mamográfico possui, ao nível do tubo de raio-X e do campo da radiação, características peculiares em relação a outros sistemas de raios-X convencionais.

Enquanto nos outros sistemas é utilizado um tubo de tungsténio, o mamógrafo usa um tubo de molibdénio, pois este tem um espectro que o aproxima de um feixe monoenergético, essencial no caso da mamografia devido à constituição dos tecidos da mama.

Quanto ao campo da radiação, no mamógrafo, é um pouco maior que a metade do campo da radiação dos sistemas convencionais, uma vez que se utilizam colimadores de feixes e restritores, direccionando os feixes de raio-X, impedindo que a radiação se alastre.

Ajudam ainda a diminuir a dose de radiação ionizante em outras partes do corpo do paciente.

O mamógrafo (figura 2.5) deve ser operado com potência constante ou trifásica, cujo feixe de raios-X tem maior poder de penetração. Geralmente a tensão usada para mamografia varia de 25 a 50 kVp, valor que depende normalmente da espessura do peito (que frequentemente, depois de comprimida, fica entre três e oito cm).

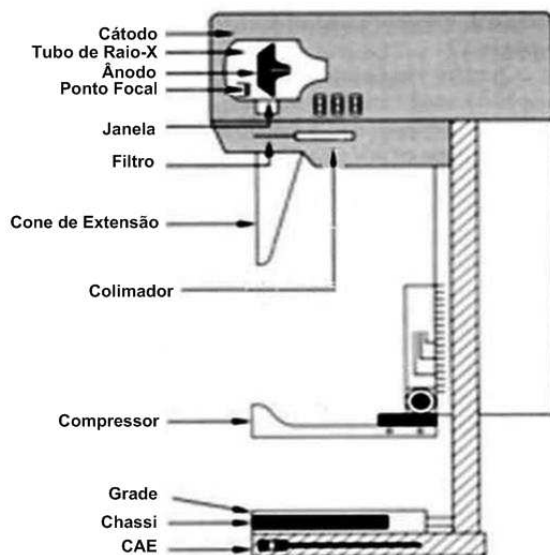


Figura 2.5: Esquema de um Mamógrafo

Além dos dispositivos comuns aos aparelhos de radiografia, o mamógrafo possui ainda um compartimento de compressão que tem como função uniformizar e comprimir a espessura do peito por uma placa de material radiotransparente até que se consiga a menor espessura possível de modo a torná-la mais acessível à penetração do feixe de raios-X, melhorando a nitidez e a resolução.

A redução da espessura do peito resulta numa menor dose de radiação dispersa, além de evitar que o paciente se movimente e prejudique a qualidade do exame.

O ponto focal é outro factor de grande importância no sistema mamográfico, pois este é determinante na capacidade de resolução espacial.

Este deve possuir dimensões reduzidas, para que estruturas de 0,3 mm de diâmetro, como as microcalcificações, possam ser possíveis de visualizar.

Outros dispositivos do mamógrafo são o chassi mamográfico e o colimador que tem a função de evitar a radiação ionizante noutras partes do corpo da paciente.

## 2.4.2 Mamografia Digital vs Mamografia Tradicional

A mamografia dita convencional é aquela cuja apresentação, visualização e arquivo é feita sob a forma de películas (também conhecida como “chapas de raio X”).

Com a crescente informatização dos hospitais e a necessidade de obtenção de melhores exames,

de forma mais eficiente e rápida, surgiu uma nova tecnologia que também tem como princípio a emissão de raios-X, mas que produz uma imagem mamográfica digitalizada (informação na forma binária), em vez da revelação numa película de filme.

O *Colégio Americano de Radiologia* (ACR) (do inglês, *American College of Radiology*), num estudo realizado no período de 2003 a 2005[12], compararam a eficácia do diagnóstico da mamografia digital com a mamografia tradicional, observando uma superioridade da mamografia digital na análise de mamas densas, representando assim, benefícios principalmente nos exames de rastreio nas mulheres jovens, já que estas apresentam maior densidade mamária. A mamografia digital apresenta vantagens devido à possibilidade de utilizar adicionalmente aplicações da ciência da computação no campo da medicina.

As principais vantagens são :

1. Permite o armazenamento dos exames mamográficos em bancos de dados informatizados, evitando a perda dos mesmos;
2. Tem uma maior gama dinâmica e como tal, permite em toda a latitude de exposição obter boas imagens;
3. O processo de aquisição da imagem não fica sujeito a agentes químicos, proporcionando imagens mais estáveis;
4. Permite a transmissão das imagens mamográficas para, por exemplo, um centro de especialistas;
5. Pode-se tirar partido de uma série de métodos computacionais que permitem otimizar digitalmente e em qualquer altura, por exemplo, o contraste da imagem;
6. Reduz o tempo geral na obtenção dos resultados, logo menos tempo de espera;
7. Reduz os custos com os filmes, películas e agentes químicos;
8. Em relação à exposição permite uma menor dose de radiação ao cliente, assim como aumenta a vida útil do tubo de raios-X.

### 2.4.3 Lesões Detectadas na Mamografia

Os tipos de lesões que aparecem com grande frequência nas mamografias são as microcalcificações e as massas.

De notar, que este trabalho foca essencialmente nos achados mamográficos correspondentes às massas.

#### **Massas**

Massas são lesões tridimensionais que podem representar o sinal de presença de cancro da mama.

Estes são descritos pela sua localização, tamanho, forma, características de margem, atenuação de raios-X, efeitos sobre o tecido circundante, bem como outros achados associados, tais como, distorções arquitecturais, alterações cutâneas, anormalidades na aréola e mamilo, etc.

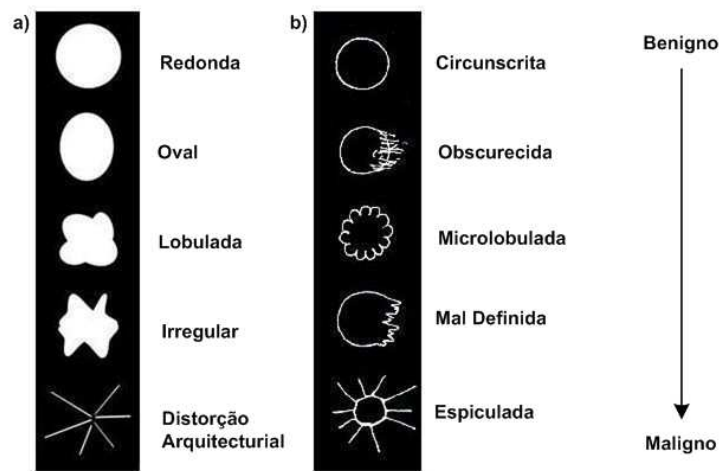


Figura 2.6: Características de a)Forma e b)Margem em que se dividem os vários tipos de massas. Adaptado de [1]

O grau de malignidade aumenta em função da ordem representada na figura 2.6. Dependendo dos critérios morfológicos da massa, a probabilidade de malignidade pode ser estabelecida.

A localização da massa pode ser estabelecida pelo exame físico, caso a massa seja palpável. Caso contrário, a sua localização pode ser determinada pelas diferentes vistas mamográficas apresentadas na figura 2.4.

É importante realçar que as massas numa mamografia podem não corresponder a um nódulo palpável.

A localização pode levantar suspeitas de malignidade, pois o cancro da mama, tende a desenvolver-se nas zonas periféricas da mama.

O tamanho por si só, não pode prever malignidade, contudo, o tamanho de uma massa maligna é indicativo da sua progressão.

Os nódulos podem apresentar cinco tipos de características de forma, sendo elas: redonda, oval, lobulada, irregular e distorção arquitectural[1].

As descrições de cada uma destas formas são automaticamente identificadas pela figura 2.6 a).

A distorção arquitectural não é tecnicamente uma massa, já que não há qualquer massa visível. Esta pode ser identificada pela distorção na arquitectura normal da massa.

A margem é a fronteira de uma massa, sendo considerado um dos mais importantes critérios para determinar se a massa é susceptível de ser benigna ou maligna.

Existem cinco tipos de margens[1], tal como definido pelo *Breast Imaging Reporting and Data System* (BI-RADS): circunscrita, obscurecida, microlobulada, mal definida e espiculada (ver figura 2.6 b)).

As margens circunscritas são bem definidas, existindo uma transição nítida entre a lesão e os tecidos circundantes.

A margem obscurecida ocorre quando o tecido circundante esconde a borda verdadeira da lesão.

As margens lobuladas possuem pequenos círculos ondulantes ao longo da borda da massa.

As margens mal definidas, tal como o próprio nome indica, possuem margens mal definidas e dispersas. A probabilidade de neoplasia maligna é neste caso elevada.

As margens espiculadas são marcadas por linhas finas radiantes, que correspondem às projecções fibrosas que se estendem a partir do nódulo.

A atenuação de raios-X é uma descrição da densidade da massa.

De um modo geral, o cancro da mama apresenta frequentemente uma densidade elevada.

## Calcificações

Calcificações são achados comuns e importantes numa mamografia.

Estas podem aparecer com ou sem uma lesão associada e as suas morfologias e distribuições formam indícios quanto à sua etiologia, bem como se estão associadas a benignidade ou malignidade.

De um modo geral e atendendo ao tamanho, as microcalcificações são associadas a malignidade e as macrocalcificações a processos benignos.

O problema com esta regra geral, é que não existe uma barreira que permita a distinção entre micro e macro.

A maioria dos radiologistas adopta que calcificações com 0.5mm ou menos possuem alta probabilidade de ter associado cancro e calcificações de 2mm ou maiores são típicas de benignidade. O número de calcificações é igualmente indicador de tumores benignos e malignos.

Um número menor que quatro calcificações por si só, raramente vai levar à detecção de cancro da mama.

Contudo, como todos os critérios na análise mamográfica não são discriminatórios, um número de duas ou três calcificações poderão merecer uma desconfiança maior se apresentarem morfologias preocupantes.

Em relação à análise das calcificações quanto à sua morfologia, esta pode ser dividida em:

1. Calcificações tipicamente benignas que correspondem à maioria dos achados:
  - Calcificações cutâneas: pequenas, de paredes finas, com centro radioluscente.
  - Calcificações vasculares: em linhas paralelas, associadas a estruturas tubulares.
  - Calcificações grosseiras ou “em pipoca”: grandes (2-3 mm de diâmetro), lobulares, típicas dos fibroadenomas.
  - Calcificações grandes: geralmente maiores que 1 mm de diâmetro, seguindo uma distribuição ductal, convergindo para o mamilo.
  - Calcificações redondas: podem ser múltiplas e variar em tamanho, geralmente formadas em ácinos ou lóbulos.
  - Calcificações puntiformes: redondas ou ovais, com menos de 0,5 mm de diâmetro e margens bem definidas.
  - Calcificações com centro radioluscente: redondas ou ovais, com centro radioluscente e paredes mais espessas do que as calcificações em “casca de ovo”.
  - Calcificações em “casca de ovo”: arredondadas com paredes finas.
  - Calcificações em “leite de cálcio”: são arredondadas na incidência crânio-caudal e ficam

lentiformes na incidência em perfil absoluto.

Calcificações em fios de sutura: lineares ou tubulares, com nós, frequentemente aparentes.

Calcificações distróficas: forma irregular, geralmente aparecem em mamas que sofreram trauma.

2. Calcificações de carácter intermediário de suspeita:

Amorfas: são tão ténues e diminutas que não é possível determinar a sua morfologia.

Calcificações heterogéneas grosseiras: geralmente são bilaterais e apresentam-se em múltiplos focos.

3. Calcificações com alta probabilidade de malignidade:

Calcificações finas e pleomórficas: alta capacidade de variação em tamanho e forma, geralmente são menores que 0,5mm em diâmetro (quanto maior a variação, maior o risco de malignidade).

Calcificações finas e lineares: são finas, lineares ou curvilíneas, descontínuas e têm menos que 0,5mm de espessura.

Quanto à sua distribuição, são classificadas em:

1. Agrupadas: quando cinco ou mais calcificações ocupam um volume menor que  $1\text{cm}^3$ .
2. Lineares: arrançadas em linha nas duas incidências.
3. Segmentares: sugere acompanhar o ducto e seus lóbulos.
4. Regionais: distribuídas em grande volume de tecido, não respeitando a distribuição ductal.
5. Difusas: distribuídas randomicamente nas mamas.

## 2.5 Classificação BI-RADS

Com o intuito de uniformizar a avaliação em relação às calcificações baseando-se principalmente na sua morfologia e distribuição (características baseadas nos estudos de Getty e de Swets de 1988 e 1991)[13, 14] e propor condutas adequadas de acordo com categorização dos achados, em 1992, o Colégio Americano de Radiologia (ACR)[12] publicou a 1ª edição do BI-RADS (do inglês, *Breast Imaging Reporting and Data System*).

Desde 1992, com a colaboração de outras entidades tais como, o Instituto Nacional de Cancro (*National Cancer Institute*), o Centro para Prevenção e Controle de Doenças (*Centers for Disease Control and Prevention*), da FDA (*Food and Drug Administration*), do Colégio Americano de Cirurgiões (*American College of Surgeons*) e do Colégio Americano de Patologistas (*College of American Pathologists*), têm vindo a aperfeiçoar os critérios na classificação das calcificações mamárias nas diversas categorias de avaliação final.

As principais modificações da última edição foram a subdivisão da categoria 4 em A, B e C, de acordo com o grau de suspeita, e a inserção da categoria 6, que inclui os casos nos quais já foi feita biópsia com resultado maligno, mas a lesão não foi totalmente retirada ou tratada por quimioterapia, radioterapia ou hormonoterapia.



É agora pedido aos radiologistas que usem as categorias fixas de avaliação do BI-RADS na sua interpretação dos mamogramas.

Investigadores da Universidade Duke(1995)[15, 16, 17, 18] usaram descritores do BI-RADS como entrada nas redes neuronais artificiais distinguindo lesões malignas de benignas. Os resultados mostraram o potencial de usar as características do BI-RADS no CAD.

O laúdo mamográfico descreve inicialmente a densidade radiológica das mamas, ressaltando que a maior densidade do parênquima mamário reduz a sensibilidade do método, contudo não altera a categoria BI-RADS.

Após a descrição da densidade, descrevem-se nódulos ( detalhando forma, margens e densidade), calcificações e outros achados.

Finalmente, conclui-se o diagnóstico expressando a categoria final e propondo a conduta adequada a essa mesma categoria:

- Categoria 0 : inconclusivo, necessitando de exame adicional ( ultra-sonografia, magnificação ou compressão localizada)
- Categoria 1 : mamografia normal, benigna
- Categoria 2 : mamografia com achados radiológicos benignos

Nestas duas últimas categorias recomenda-se o controlo anual a partir dos 40 anos.

- Categoria 3 : mamografia com achados provavelmente benignos com seguimento mamográfico em 6 meses
- Categoria 4(A,B,C) : mamografia com achados radiológicos suspeitos de malignidade
- Categoria 5 : mamografia com achados radiológicos altamente suspeitos de malignidade

Nestas duas categorias indica-se uma investigação cito-histológica (biópsia).

- Categoria 6 : lesão já biopsiada e diagnosticada como maligna, mas não retirada ou tratada

À medida que a padronização do BI-RADS seja aceite mais geralmente pelos radiologistas, a variabilidade das classificações diminui e assim os diagnósticos auxiliados por computador baseados no BI-RADS ficarão mais robustos.



## Capítulo 3

# Diagnóstico Auxiliado por Computador

A análise de imagens mamográficas é um problema para os radiologistas devido principalmente ao baixo contraste das mamografias, à possibilidade de algumas estruturas suspeitas ficarem sobrepostas e à fadiga visual por parte do radiologista[19].

Diagnosticar torna-se assim num trabalho muito complexo e dependente de vários factores que muitas vezes levam a cometer erros aquando o diagnóstico, tendo como consequência a submissão do paciente a novos exames desnecessários, que além de invasivos podem ser traumáticos e, financeiramente acarretam custos elevados.

### 3.1 Casos de Estudo

Com o objectivo de auxiliar na difícil tarefa de avaliar e analisar estruturas de interesse em imagem mamográficas digitais para posterior diagnóstico, diversos trabalhos utilizando técnicas de processamento de imagens, reconhecimento de padrões, inteligência artificial vêm sendo desenvolvidos como parte de esquemas CAD.

Mudigonda e Rangayyan em 2000[8], estudaram a classificação de massas benignas e malignas de mamogramas, calculando características baseadas no gradiente e textura.

Os autores propuseram calcular o gradiente e textura baseadas em matrizes de co-ocorrência entre níveis de cinza, de modo a avaliar a informação de textura por regiões do centro da massa comparada com as informações de textura presentes nas margens da massa.

Para o cálculo das medidas baseadas no gradiente foi necessário extrair a intensidade dos pixels de dentro do limite da massa para o exterior ao longo da direcção perpendicular em todos os pontos do contorno.

Foram utilizadas um total de 54 imagens das quais 39 pertencentes à base de dados *Mammographic Image Analysis Society* (MIAS), sendo as restantes de uma base de dados local.

A melhor classificação foi de 82,1% de acerto total com uma área  $Az = 0.85$  sob a curva ROC. O método usado para a classificação é baseado em probabilidades a partir da distância de Mahalanouis.

A precisão usando classificação Jack-Knife obteve 74,4% de acerto total com um  $Az$  de 0.67. Características baseadas no gradiente obtiveram um  $Az = 0.6$  com a base de dados MIAS e um  $Az=0.76$  para a base de dados combinada, utilizando o primeiro método de classificação.

Os correspondentes valores obtidos usando a classificação Jack-Knife, observou um  $Az=0.52$  e um  $Az=0.73$ , respectivamente, para a base de dados MIAS e para a base de dados combinada.

Túlio André e Rangaraj Rangayyan em 2003[20] propuseram uma aproximação que usa redes neurais artificiais com o objectivo de classificar massas em mamogramas.

O contorno de 108 massas foram manualmente delimitadas e representadas por modelos poligonais para análise da forma.

Foram calculadas três medidas do factor de forma baseada no contorno e 14 características de textura baseadas na matriz de co-ocorrência de níveis de cinza.

As características de textura foram calculadas usando as margens da massa definidas como tiras que cercam a massa da *Região de Interesse* (ROI) e não a região inteira.

Esta aproximação segue o trabalho de Mudigonda[8], no qual usaram as margens das massas para melhorar a discriminação entre massas benignas e malignas quando comparadas com as massas medidas obtidas usando toda a região.

Foram utilizadas várias combinações de características e várias topologias para a classificação com o *perceptron*.

Os resultados obtidos foram comparados em termos da área  $Az$ , sob a curva ROC. Valores de  $Az=0.75$  até 0.99 foram obtidos com as características de forma e de 0.48 até 0.63 para características de textura.

Ribeiro (2006)[21] apresentou uma metodologia para classificação de nódulos mamários por contorno.

Através das imagens mamográficas digitalizadas foram recortadas regiões de interesse de onde foram extraídas características de textura, intensidade e geométricas com o objectivo de caracterizar os padrões de contorno dos nódulos.

As características serviram de entrada para duas redes neurais, uma *Multi-Layer Perceptron* (MLP) e *Mapas Auto-Organizáveis* (SOM), obtendo a melhor classificação para a rede neuronal MLP, com 20 neurónios de entrada (características seleccionadas através da rede SOM e gaussianas), 40 neurónios na primeira camada intermédia, 20 na segunda e 5 neurónios na camada de saída, com taxa de aprendizagem igual a 0.9 e erro menor que 0.01.

O acerto total obtido utilizando 135 ROIs e pesos fixos foi de 89% e  $Az$  igual a 0.98, falso-negativo igual a 5% e falso-positivo igual a 7%.

A importância destes esquemas é realçado por Chan[22], que em programas de rastreio mostrou que com o uso de esquemas CAD melhoraram o desempenho dos radiologistas.

Assim, estes esquemas computacionais visam melhorar a visualização de imagens radiológicas, indicando e realçando a existência de estruturas que possam estar associadas à presença de tumores, auxiliando o radiologista, na interpretação da imagem e na elaboração de um diagnóstico mais preciso, fornecendo uma “segunda opinião”[23].

De realçar que estes esquemas, dada a sua importância em diversas aplicações radiológicas, com especial relevo na mamografia, já foi aprovado pelo *Food and Drug Administration* (FDA) como recurso ao diagnóstico do cancro da mama.

## 3.2 Imagem Digital

Uma imagem digital é definida como sendo uma função de duas variáveis  $f(x, y)$  de valores discretos, onde  $x$  e  $y$  são as coordenadas espaciais (planas).

O ponto  $(x, y)$  é conhecido como Pixel (Picture Element) e a amplitude de  $f$  é a intensidade do nível de cinza (Graylevel) nesse par de coordenadas  $(x, y)$ .

A escala de cinza é o intervalo de variação definido pela equação 3.1:

$$0 \leq f(x, y) \leq G \quad (3.1)$$

$G$  é o máximo valor da escala de cinza.

Se  $G = 255$ , significa que cada pixel tem resolução de 256 níveis de cinza, ou seja, 8 bits/pixel.

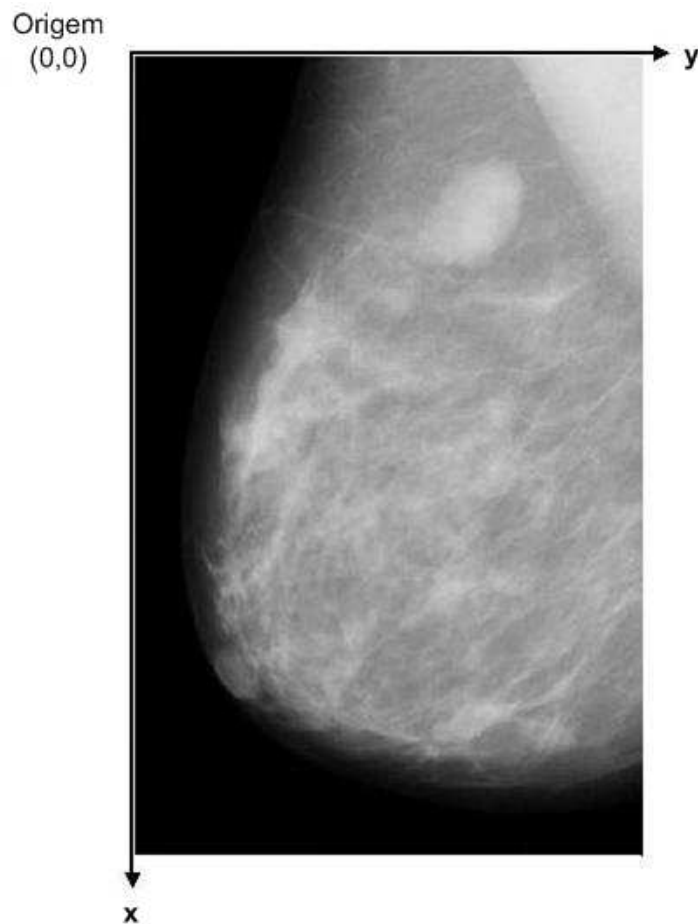


Figura 3.1: Convenção utilizada para os eixos em Processamento de Imagens

De salientar, que em processamento de imagens (figura 3.1) num arranjo matricial de  $M$  linhas por  $N$  colunas temos a seguinte convenção:

$$(0, 0) \leq (x, y) \leq (M - 1, N - 1) \quad (3.2)$$

ao contrário, já a toolbox de processamento de imagens do Matlab[11] apresenta a seguinte:

$$(1, 1) \leq (r, c) \leq (M, N) \quad (3.3)$$

### 3.3 Extracção de Características

O processamento das imagens visa desenvolver procedimentos para extrair informações de uma imagem.

Para se poder classificar automaticamente estruturas numa imagem, é obrigatório extrair várias medidas de um dado padrão e armazenar essas medida num vector que servirá de entrada para os classificadores automáticos.

De referir, que todo este processo de extracção e selecção das características que melhor permitem distinguir as diversas classes tem grande importância no desempenho dos classificadores.

Note-se, por exemplo, que os vários tipos de características seleccionadas devem ser relacionadas directamente com o problema.

Por isso, a selecção antecipada das melhores características que permitem caracterizar o problema, até de modo a evitar informações redundantes, são importantes para se obter uma classificação eficiente.

As características seleccionadas de modo a extrair informações das classes (benigno e maligno) em que se dividem as massas, pertencem aos grupos de textura e de forma.

#### 3.3.1 Características de Forma

O formato, tamanho e distribuição são factores importantes na análise e caracterização da massa. Diversos conjuntos de características morfológicas foram avaliadas e testadas[21, 24], sendo seleccionadas um conjunto de cinco características relacionadas com a forma e o tamanho: perímetro, área, rugosidade, brilho e irregularidade.

##### Perímetro ( $P$ )

É dado pelo número total de pixéis que pertencem ao contorno da massa.

##### Área ( $A$ )

É dado pelo número total de pixéis no interior da lesão.

A maneira mais simples de calcular a área de um objecto é contar o número de pixéis do objecto.

##### Rugosidade ( $R$ )

É uma medida derivada do perímetro e área de um objecto, conforme equação 3.4:

$$R = \frac{P^2}{4.\pi.A} \quad (3.4)$$

##### Brilho ( $B$ )

Corresponde à média da intensidade dos níveis de cinza presentes na lesão patológica seleccionada.

### Irregularidade ( $I$ )

Tal como a rugosidade, trata-se de uma medida derivada do perímetro e área de um objecto, conforme equação 3.5:

$$I = \frac{A}{P^2} \quad (3.5)$$

### 3.3.2 Características de Textura

Textura é um termo muito usado na visão computacional, sendo uma das principais características usadas para identificar objectos numa imagem.

Todos somos capazes de reconhecer a textura, mas é muito difícil definir textura com precisão. De facto, na própria literatura não existe um consenso geral capaz de a definir, encontrando-se várias definições. Pode-se referenciá-la como um conjunto de padrões existentes nas superfícies, perceptíveis ao olho humano e que possuem uma grande quantidade de informações sobre a sua natureza (ex: rugosidade, suavidade,...).

As três abordagens normalmente utilizadas para descrever textura em imagens são: abordagem estrutural, espectral e estatística[25].

#### Abordagem Estrutural

Esta abordagem considera as primitivas que formam padrões repetitivos e descreve estes, obedecendo às regras que os originam.

#### Abordagem Espectral

Esta abordagem é baseada no espectro de Fourier, sendo utilizada para detectar padrões periódicos de semi-periódicos de uma imagem, através da identificação de picos de alta frequência no espectro.

#### Abordagem Estatística

Esta abordagem propicia uma descrição da textura através de regras estatísticas que coordenam a distribuição e a relação entre os níveis de cinza.

As estatísticas de 1ª ordem, são calculadas a partir da probabilidade de observar um determinado valor de um pixel escolhido aleatoriamente na imagem.

Estas dependem somente do valor dos pixéis e não na interacção com os valores dos pixéis vizinhos. A média da intensidade de cinza da imagem é um exemplo de uma estatística de primeira ordem.

As estatísticas de 2ª ordem são calculadas a partir da probabilidade de observar os valores de um par de pixéis da imagem.

#### Estatística 1ª ordem

As estatísticas de 1ª ordem são muito simples, sendo calculadas a partir da função de probabilidade de níveis de cinza numa imagem. Esta função é também conhecida como histograma. Em imagens de escala de cinza o histograma é definido por:

$$H(g) = \frac{n_g}{N} \quad \text{com } g = 0, \dots, G \quad (3.6)$$

Onde  $N$  é o número total de pixels na imagem,  $n_g$  é o número de pixels com o valor  $g$  de intensidade na imagem.

A partir da função de probabilidade  $H$  do valor dos pixels é possível caracterizá-los com um conjunto de parâmetros estatísticos (chamados de 1ª ordem).

Abaixo, fica a lista de alguns parâmetros, onde  $P_n$  representa o parâmetro número  $n$ .

### Média

$$P_1 = \sum_{g=0}^G g.H(g) \quad (3.7)$$

### Desvio padrão

Mede a função de dispersão sobre a média.

$$P_2 = \sqrt{\sum_{g=0}^G (g - P_1)^2.H(g)} \quad (3.8)$$

### “Skewness”

$$P_3 = \frac{\sum_{g=0}^G (g - P_1)^3.H(g)}{(P_2)^3} \quad (3.9)$$

### Energia

$$P_4 = \sum_{g=0}^G H(g)^2 \quad (3.10)$$

### Entropia

$$P_5 = - \sum_{g=0}^G H(g) \log_2(H(g)) \quad (3.11)$$

### 1º Momento Central

É o mesmo parâmetro que a média.

### 2º Momento Central

$$P_7 = \sum_{g=0}^G (g - P_1)^2.H(g) \quad (3.12)$$

### 3º Momento Central

$$P_8 = \sum_{g=0}^G (g - P_1)^3.H(g) \quad (3.13)$$

### 4º Momento Central

$$P_9 = \sum_{g=0}^G (g - P_1)^4.H(g) \quad (3.14)$$

De referir contudo, que o cálculo de medidas de textura, a partir exclusivamente do histograma sofrem a limitação de possuir somente informações sobre a distribuição da intensidade, não possuindo informações sobre as posições dos pixels.

Assim sendo, será dada uma maior importância às estatísticas de 2ª ordem.



## Estatística 2ª ordem

A matriz de co-ocorrência entre níveis de cinza (*Grey Level Co-occurrence Matrix* (GLCM)) é o método mais utilizado sendo por isso dada uma maior ênfase.

Este método é também conhecido como método da dependência espacial de níveis de cinza (*Spatial Grey Level Dependence* (SGLD)).

A matriz SGLD calcula a probabilidade de ocorrer uma transição de nível de cinza entre dois pixels com uma relação espacial definida numa imagem.

O relacionamento espacial é definido em termos da distância  $d$  e do ângulo  $\theta$ .

Em outras palavras,  $P(i, j)$  indica a frequência de ocorrência entre pares de níveis de cinza, separados por uma distância  $d$  e com uma orientação  $\theta$  (que pode ser  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  e  $135^\circ$ ), conforme ilustrado na figura 3.2.

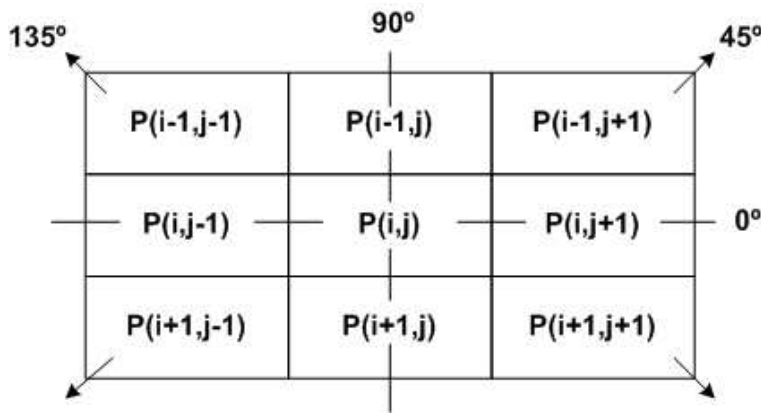


Figura 3.2: Variações angulares utilizadas na obtenção da matriz SGLD para uma distância  $d=1$

Na figura 3.3 está representado um exemplo de obtenção de uma matriz de co-ocorrência extraído de *Haralick*[26].

A partir de cada uma das matrizes SGLD, uma variedade de medidas estatísticas podem ser extraídas.

*Haralick et al.*[26, 27] definiram um conjunto de 14 características a serem extraídas a partir da nomenclatura a seguir apresentada:

$$p(i, j) = \frac{P(i, j)}{r} \quad (3.15)$$

Entrada  $(i, j)$  da matriz SGLD normalizada.

$$p_x(i) = \sum_{j=1}^{N_g} p(i, j) \quad (3.16)$$

corresponde à entrada  $i$  do vector de probabilidade marginal obtida pelo somatório das linhas de  $p(i, j)$ .

$N_g$  corresponde ao número total de níveis de cinza presentes na imagem.

$$p_y(j) = \sum_{i=1}^{N_g} p(i, j) \quad (3.17)$$

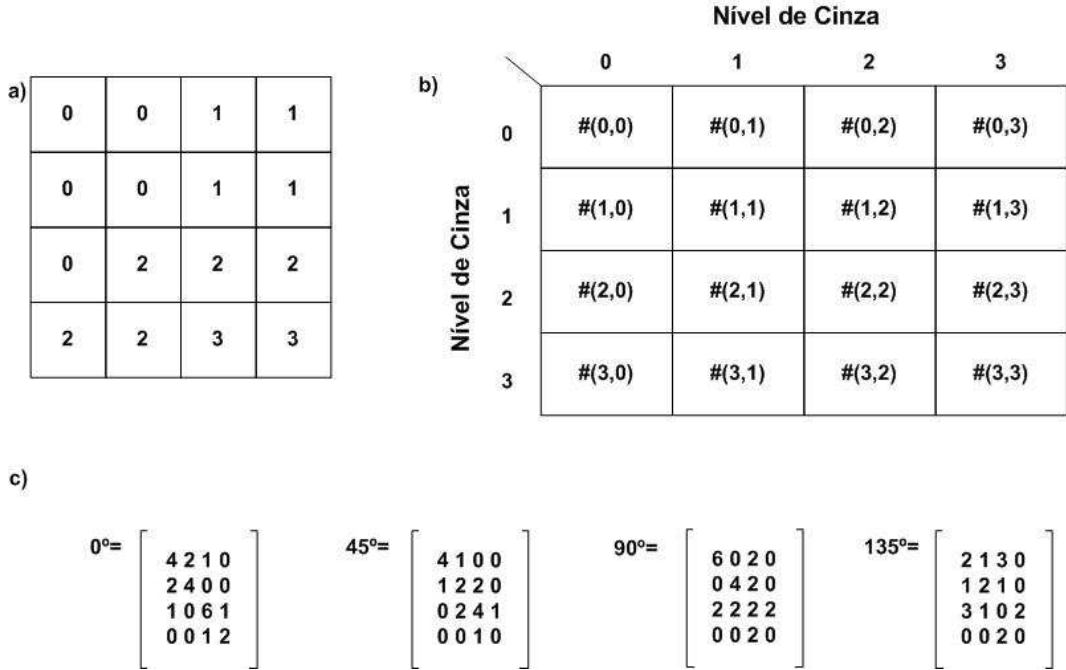


Figura 3.3: Cálculo matriz SGLD onde a) Imagem 4x4 pixel com 4 valores de nível de cinza, b) Construção da matriz SGLD, onde  $(i, j)$  indica os pares de níveis de cinza possíveis da imagem em questão e c) Cálculo das 4 matrizes SGLD quando  $d=1$

corresponde à entrada  $j$  do vector de probabilidade marginal obtida através do somatório das colunas de  $p(i, j)$ .

$$p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p(i, j) \quad \text{com } k = 2, \dots, 2N_g \quad (3.18)$$

O resultado da adição das entradas  $i$  e  $j$  calculadas a partir da matriz de probabilidade marginal.

$$p_{x-y}(k) = \sum_{\substack{i=1 \\ |i-j|=k}}^{N_g} \sum_{j=1}^{N_g} p(i, j) \quad \text{com } k = 0, \dots, N_g - 1 \quad (3.19)$$

O resultado da subtração das entradas  $i$  e  $j$  calculadas através da matriz de probabilidade marginal.

As 14 características são a seguir apresentadas.

### Segundo momento Angular (SMA)

$$F_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j)^2 \quad (3.20)$$

Esta é uma medida da uniformidade ou suavidade.

Texturas homogéneas terão um alto valor de energia em comparação com texturas não homogéneas, devido às texturas suaves possuírem densidades mais concentradas que as texturas rugosas.

### Contraste

$$F_2 = \sum_{\substack{n=0 \\ |i-j|=n}}^{N_g-1} n^2 \left( \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right) \quad (3.21)$$

Esta característica reflecte a quantidade de variações locais de níveis de cinza presentes numa imagem.

### Correlação

$$F_3 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (ij)p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad (3.22)$$

onde  $\mu_x$ ,  $\mu_y$ , correspondem às médias nas direcções x e y, enquanto que  $\sigma_x$ ,  $\sigma_y$ , representam os desvios padrão.

Correlação mede a dependência linear dos níveis de cinza dentro da imagem, ou seja, mede o grau em que as linhas e colunas da matriz SGLD se assemelham mutuamente.

Valores elevados de correlação são obtidos quando os elementos da matriz são uniformemente iguais, ao passo que, valores baixos são obtidos quando os valores dos elementos na matriz possuem grandes diferenças.

### Variância

$$F_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu_x)^2 p(i, j) \quad (3.23)$$

A variância diz respeito à heterogeneidade da matriz de co-ocorrência, ou seja, trata-se de um indicador da variação de tonalidade de fundo da imagem.

### Momento da diferença inversa (MDI) ou homogeneidade

$$F_5 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1 + (i - j)^2} \cdot p(i, j) \quad (3.24)$$

Este denota a homogeneidade da matriz de co-ocorrência, sendo maior quanto menor for o valor do contraste.

### Entropia

$$F_6 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \cdot \log_2(p(i, j)) \quad (3.25)$$

A entropia da imagem é uma medida da aleatoriedade do valor do nível de cinza de uma imagem.

Valores elevados de entropia indicam grande aleatoriedade na imagem.

### Média da soma

$$F_7 = \sum_{i=2}^{2N_g} i p_{x+y}(i) \quad (3.26)$$

**Variância da soma**

$$F_8 = \sum_{i=2}^{2N_g} (i - F_7)^2 p_{x+y}(i) \quad (3.27)$$

**Entropia da soma**

$$F_9 = - \sum_{i=2}^{2N_g} p_{x+y}(i) \cdot \log_2(p_{x+y}(i)) \quad (3.28)$$

**Variância da diferença**

$$F_{10} = \text{variancia de } p_{x-y} \quad (3.29)$$

**Entropia da diferença**

$$F_{11} = - \sum_{i=0}^{N_g-1} p_{x-y}(i) \cdot \log_2(p_{x-y}(i)) \quad (3.30)$$

**Medida de informação de correlação 1**

$$F_{12} = \frac{HXY - HXY1}{\max(HX, HY)} \quad (3.31)$$

onde,

$$HX = - \sum_{j=1}^{N_g} p_x(j) \cdot \log_2(p_x(j)) \quad (3.32)$$

que corresponde à entropia vertical, e

$$HY = - \sum_{i=1}^{N_g} p_y(i) \cdot \log_2(p_y(i)) \quad (3.33)$$

que corresponde à entropia horizontal.

$$HXY = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \cdot \log_2(p(i, j)) \quad (3.34)$$

$$HXY1 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \cdot \log_2(p_x(i) \cdot p_y(j)) \quad (3.35)$$

$$HXY2 = - \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i) \cdot p_y(j) \cdot \log_2(p_x(i) \cdot p_y(j)) \quad (3.36)$$

**Medida de informação de correlação 2**

$$F_{13} = \sqrt{1 - \exp[-2.0(HXY2 - HXY)]} \quad (3.37)$$

**Máximo coeficiente de correlação**

$$F_{14} = \sqrt{\text{(segundo maior valor proprio de Q)}} \quad (3.38)$$

onde,

$$Q(i, j) = \sum_k \frac{p(i, k) \cdot p(j, k)}{p_x(i) \cdot p_y(k)} \quad (3.39)$$

## 3.4 Redes Neuronais Artificiais

As RNAs são técnicas computacionais que apresentam um modelo inspirado no funcionamento do cérebro e foram propostas com o intuito de resolver problemas em que os seres humanos possuem boa performance, porém de difícil tratamento pelos computadores tradicionais.

### 3.4.1 Breve Cronologia

A história das redes neuronais (citado em [9, 28]), remonta ao início dos anos 40, mais precisamente no ano 1943, quando um dos primeiros modelos lógico-matemático de um neurónio artificial foi proposto pelo neurofisiologista *Warren McCulloch* e pelo matemático *Walter Pitts*, apresentado na figura 3.4.

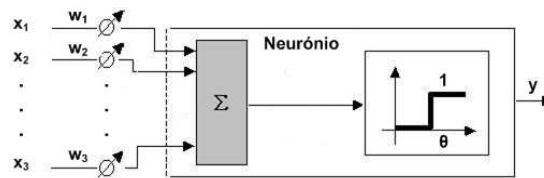


Figura 3.4: Neurónio proposto por *McCulloch e Pitts*

O estudo baseava-se no facto, de num dado instante de tempo, o neurónio, ou está a disparar pulsos eléctricos ou está parado, correspondendo assim ao “verdadeiro e falso” do cálculo lógico, ou ao “zero e um” da álgebra linear.

No entanto, o primeiro neuro-computador a obter sucesso, o *perceptron*, foi criado somente em 1958, por *Frank Rosenblatt*.

Após *Rosenblatt, Widrow and Hoff*, em 1963, desenvolveu um novo tipo de elemento de processamento chamado *Adaline*, equipado com uma poderosa lei de aprendizagem.

Em 1969, *Minsky e Papert*, no seu livro *Perceptrons* apontaram as limitações teóricas dos modelos de redes neuronais com uma camada. Devido a esta visão pessimista que desacreditou a área, a investigação sobre as redes neuronais ficou parada durante praticamente duas décadas.

No início dos anos 80, *Hopfield* (1982) introduziu o conceito de minimização de energia da física nas redes neuronais o que deu origem ao termo *momentum*, resgatando o interesse nas redes neuronais.

Em 1986, *David Rumelhart*, com a publicação do livro “*Parallel Distributed Processing*” (Processamento Distribuído Paralelo), apresentou uma solução poderosa para o treino de uma rede de várias camadas que destronou os perceptrões e que ficou conhecida como o algoritmo de aprendizagem *Backpropagation*, cujas bases teóricas para um método de cálculo de derivadas parciais ordenadas já havia sido desenvolvido por *Werbos* em 1974.

Desde 1987, muitas universidades anunciaram a formação de institutos de pesquisa para a neuro computação.

### 3.4.2 Estrutura do Neurónio Biológico

O córtex cerebral é formado por um conjunto extremamente complexo de células, os neurónios. Estes têm um papel essencial na determinação do funcionamento e comportamento do corpo

humano e do raciocínio.

Um neurónio biológico é formado pelo corpo celular (ou soma), que contém o núcleo da célula, pelos dendritos, que recebem os impulsos eléctricos e pelo axónio, através do qual os impulsos são enviados.

A figura 3.5 mostra a estrutura de uma célula neuronal.

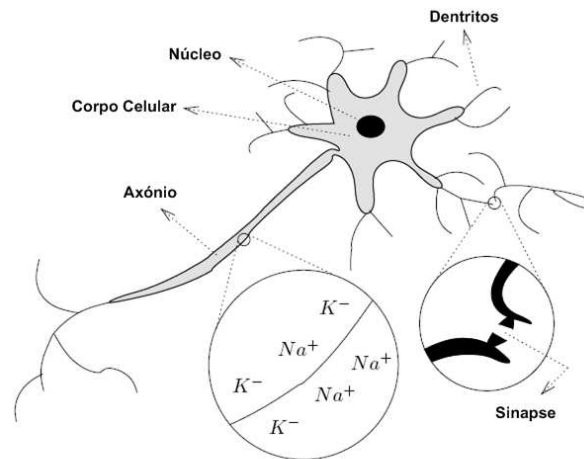


Figura 3.5: Constituição da Célula Neuronal

Os neurónios interligam-se através das ligações sinápticas ou sinapses. A comunicação é realizada unidirecionalmente por sinais nervosos eléctricos e químicos (nas regiões terminais das ligações sinápticas) de uma forma descontínua, isto é, por impulsos. Estes últimos são desencadeados em cada neurónio sempre que um certo potencial de activação é ultrapassado, em resultado da recepção de um ou mais sinais nervosos nas sinapses.

A propagação de um impulso eléctrico ao longo de um dendrito ou de um axónio dá-se através da alteração da concentração dos iões  $K^-$  e  $Na^+$  em ambos os lados da membrana.

### 3.4.3 Funcionamento do Neurónio Artificial

O neurónio artificial, cuja estrutura está representada na figura 3.4, tenta modelizar algumas das características biológicas. Neste modelo geral, o neurónio é composto por uma ou mais entradas ( $x_1, x_2, \dots, x_n$ ) onde, para cada entrada está associado um peso ( $w_1, w_2, \dots, w_n$ ).

É então calculada a soma ponderada entre os sinais de entrada e os pesos das conexões, peso esse indicativo do grau de importância que determinada entrada possui em relação a esse mesmo neurónio.

O resultado final ( $s$ ) é aplicado à função de activação, também designada de função de transferência.

Esta função transmite o sinal para a saída, de modo a produzir um estado de activação do neurónio que corresponde à frequência de descarga do neurónio biológico.

Considerando a função de activação do *Perceptron* (função degrau), o modelo matemático

apresentado terá as seguintes equações:

$$s = \sum_{i=1}^n x_i w_i \quad (3.40)$$

$$f(s) = \begin{cases} 1 & \text{if } s \geq L \\ -1 & \text{if } s < L \end{cases} \quad (3.41)$$

onde,

- $n$  = número de entradas
- $w_i$  = peso associado à entrada  $x_i$
- $L$  = valor do limiar do neurónio

O bias é uma entrada adicional que pode ser acrescentada e que possui o valor de entrada fixo de +1.

O seu peso de conexão ( $w_0$ ) é ajustável durante o processo de aprendizagem como qualquer outro peso de conexão.

O resultado final da soma ponderada ( $s$ ) será agora dado pela equação 3.42.

$$s = \sum_{i=1}^n x_i w_i + w_0 \quad (3.42)$$

As funções de activação mais conhecidas na construção das redes neuronais artificiais, têm sido, entre outras a função degrau, que possui decisões ríspidas, principalmente para valores extremos, ou seja, ultrapassando um limite determinado, a função dispara o valor 0 ou 1 (ou -1), a função semi-linear e a função sigmoideal.

A função de activação utilizada na construção da rede neuronal multicamada foi contudo, a função tangente hiperbólica, que é descrita matematicamente, pela equação 3.43.

$$f(x) = f_{max} \tanh(sx) = f_{max} \frac{\sinh(sx)}{\cosh(sx)} = f_{max} \frac{e^{2sx} - 1}{e^{2sx} + 1} \quad (3.43)$$

### 3.4.4 Caracterização das RNAs

Uma vez definido o princípio de funcionamento do neurónio artificial é possível estudar as propriedades das redes de neurónios interconectados, de forma a que as saídas de uns estão ligadas às entradas de outros.

As *Redes Neuronais Artificiais* (RNA) possuem assim processamento altamente paralelo apresentando as seguintes propriedades neurobiológicas:

- O conhecimento é adquirido pela rede, a partir do seu ambiente, através de um processo de aprendizagem.
- Forças de conexão entre neurónios, conhecidos como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

- Os pesos sinápticos são ajustados (durante o treino) através de algoritmos, para que possam representar o conhecimento adquirido e com isso melhorar o seu desempenho. Estas alterações caracterizam o processo de aprendizagem de uma rede neuronal.
- As aplicações das redes neuronais artificiais estão relacionadas com a sua arquitectura, processo de aprendizagem e comportamento dinâmico.

### Classificação quanto à Arquitectura

A arquitectura da RNA define a forma como os neurónios estão organizados, diferenciando-se umas das outras em função do número de camadas e dos arranjos das conexões.

Conforme pode ser visualizado na figura 3.6, geralmente as RNA possuem:

- Camada de entrada, que é utilizada para captar estímulos externos e repassá-los para a camada escondida;
- Camadas escondidas, cujas saídas dos neurónios não são acessadas por procedimentos externos à RNA;
- Camada de saída, que retrata o resultado produzido pela RNA em resposta a uma dada entrada.

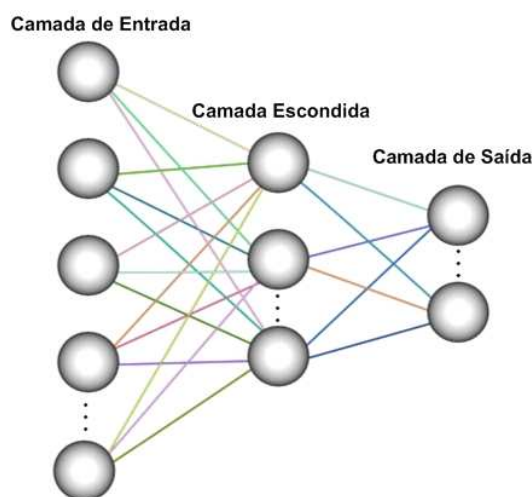


Figura 3.6: Camadas de uma RNA genérica

Em relação ao tipo de conexões, as RNAs podem ser classificadas em *feedforward* e *feedback*. As Feedforwards, são redes de fluxo directo, onde todos os pontos de conexão seguem numa direcção, propagando-se para a frente.

A rede Feedback ou cíclica (realimentação), é aquela em que o sinal de saída de um neurónio serve como entrada para a mesma camada, uma camada anterior, entre camadas alternadas ou para partes específicas de uma camada qualquer.

### Classificação quanto ao Processo de Aprendizagem

A característica mais importante das RNAs é a sua capacidade para adquirir conhecimento e melhorar o seu desempenho ao longo do tempo.



O processo de treino da rede sucede-se basicamente de duas formas: supervisionado ou não supervisionado.

A aprendizagem supervisionada exige a disponibilidade de um conjunto de treino formado por pares de vectores de entrada e de saída, ou seja, os pares de entrada são apresentados à rede para ser verificado se a saída obtida foi a desejada.

Caso esta seja diferente, a rede irá calcular um erro, representativo da distância à resposta desejada.

Com isso, a rede ajusta os pesos até que o erro seja mínimo.

Na aprendizagem não supervisionada, o conjunto de treino é composto somente por vectores de entrada.

### 3.4.5 O algoritmo *Backpropagation*

O conceito de retropropagação (do inglês, *backpropagation*) do erro introduzido por *Rumelhart*(1986) possibilitou a implementação de redes multicamada, sendo o principal e mais conhecido algoritmo de treino para redes MLP.

Este tem como objectivo o cálculo do erro para a camada de saída e propagar este no sentido saída-entrada, ajustando os pesos de todas as camadas, através da retropropagação do erro.

Este algoritmo de aprendizagem supervisionada (também conhecido por *regra delta generalizada*), é derivado do método da regra delta desenvolvida por *Widrow and Hoff* e consiste numa optimização não linear baseada na técnica de gradiente descente sobre uma superfície de erro formada pelo quadrado da diferença entre a saída desejada e a saída gerada pela rede. Durante o treino com o algoritmo *backpropagation*, a rede opera numa sequência de duas etapas, como demonstra a figura 3.7.

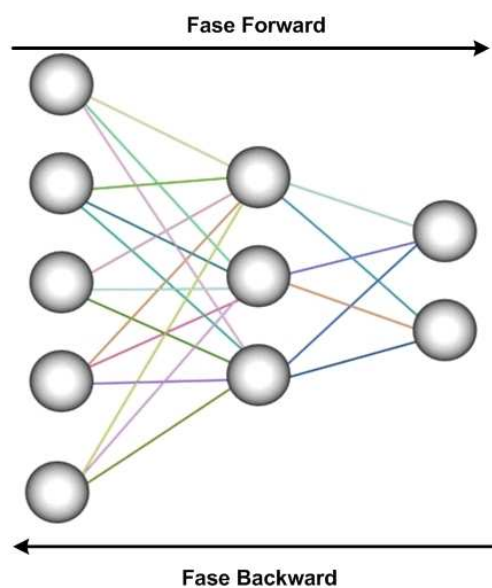


Figura 3.7: Esquema de funcionamento do algoritmo *backpropagation*

Na etapa *forward*, são apresentados os vectores de entrada à rede e calculada a resposta produzida pela camada de saída.

Na etapa *backward*, a saída obtida é comparada com a saída desejada. Se estas não forem iguais é calculado o erro, sendo este retropropagado desde a camada de saída até à camada de entrada, para o cálculo da actualização dos pesos.

## Capítulo 4

# Material e Métodos

Os esquemas CAD tal como descrito no capítulo anterior visam auxiliar os radiologistas na tarefa de analisar achados radiológicos suspeitos e posteriormente contribuir no diagnóstico dando uma “segunda opinião”, cabendo sempre a decisão final ao radiologista. De modo a estruturar o problema para mais fácil interpretação dos passos necessários elaborou-se um diagrama de blocos apresentado na figura 4.1.

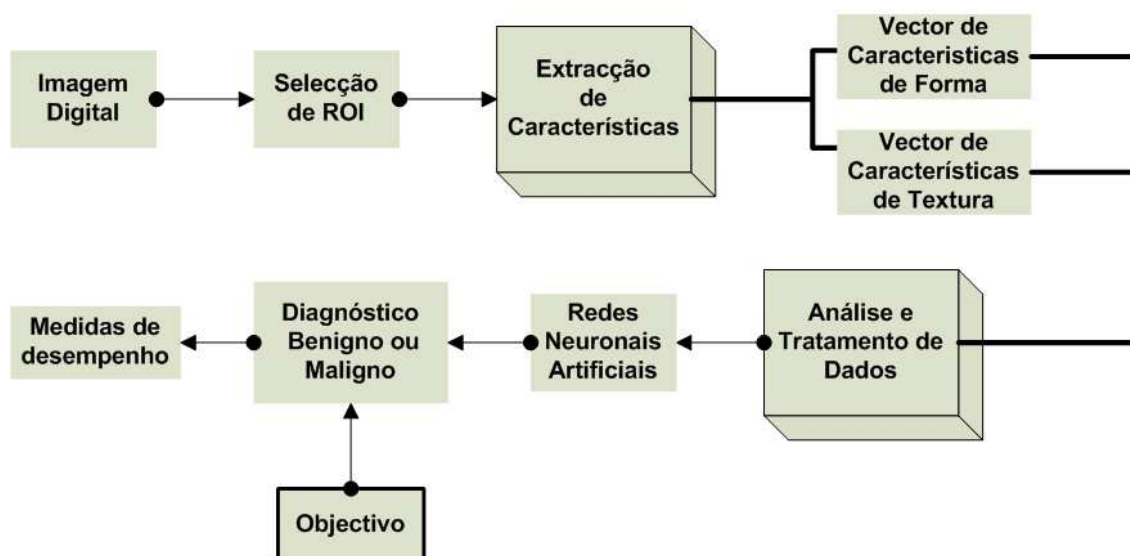


Figura 4.1: Diagrama de blocos da estrutura do esquema CAD implementado

### 4.1 Base de dados Mamográfica

As imagens mamográficas digitais utilizadas foram obtidas da base de dados de domínio público, miniMIAS (do inglês, *Mammographic Image Analysis Society*) [29]. Esta base de dados é constituída por um total de 322 mamogramas (de 161 pacientes), com uma resolução de  $200\mu m$  por pixel (de 8 bits), tendo assim cada imagem, um tamanho de  $1024 \times 1024$  pixéis.

Destes 322 mamogramas foram seleccionadas 73 imagens que continham massas de diversas classes de lesões patológicas, tal como referenciado na tabela 4.1.

Classes	Tipo			Total	
	Fatty	Fatty-Glandular	Dense-Glandular		
Circumscribed Masses	B: 11	B: 6	B: 3	B: 20	24
	M: 2	M: 2	M: 0	M: 4	
Spiculated Masses	B: 2	B: 4	B: 5	B: 11	19
	M: 3	M: 3	M: 2	M: 8	
Miscellaneous (Ill-defined Masses)	B: 4	B: 2	B: 1	B: 7	15
	M: 5	M: 2	M: 1	M: 8	
Asymmetry Masses	B: 1	B: 2	B: 3	B: 6	15
	M: 3	M: 2	M: 4	M: 9	
<b>Total</b>	<b>31</b>	<b>23</b>	<b>19</b>	<b>73</b>	

Tabela 4.1: Composição da base de dados (B: Benigno, M: Maligno).

## 4.2 Selecção de ROI

Esta é uma etapa fundamental, uma vez que a região seleccionada representa a parte da imagem mamográfica que se pretende analisar, onde se encontra o laúdo mamográfico.

Assim sendo, neste processo selecciona-se somente as regiões que contenham lesões patológicas suspeitas.

A extracção da lesão que se pretende classificar foi obtida com o recurso à listagem fornecida pelo MIAS.

Nesta lista é fornecida as coordenadas  $(x, y)$  do centro da anormalidade bem como, o raio aproximado (em pixéis) da circunferência que engloba a anormalidade.

Note-se que, quando na lista que incorpora a informação de todas as imagens, nos informa das coordenadas  $(x, y)$  e do raio aproximado da lesão, temos de ter em conta a convenção dos eixos para estas coordenadas. Assim sendo, na matriz o valor de  $x$  corresponde ao número da coluna e o valor de  $1024 - y$  corresponde ao número da linha.

Na imagem apresentada como exemplo na figura 4.2, o MIAS informa que as coordenadas do centro da lesão é  $(595, 864)$ , logo na matriz de intensidade de cinza dessa mesma imagem corresponde ao pixel com linha igual a 160  $(1024-864)$  e coluna igual a 595.

## 4.3 Extracção de Características

O processo de extracção e selecção das características que melhor permite distinguir diversas classes tem uma grande importância no desempenho dos classificadores.

As características seleccionadas pertencem ao grupo de textura e de forma, tal como representado no diagrama da figura 4.3.

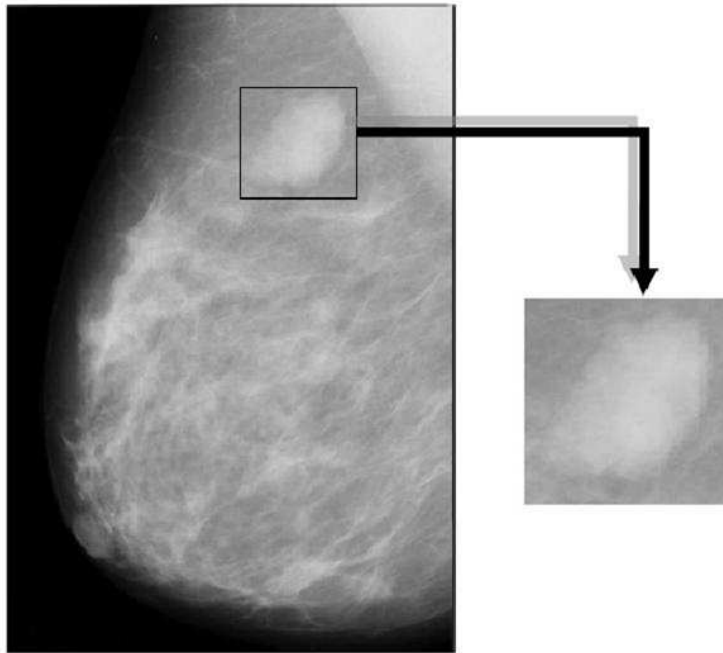


Figura 4.2: Mamograma original e extracção da ROI de uma massa benigna circunscrita

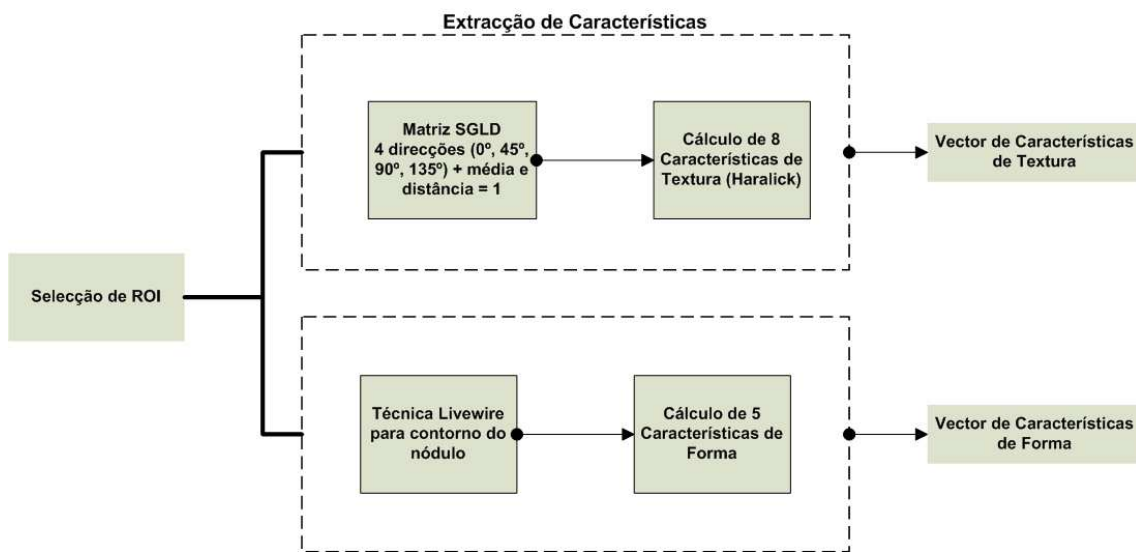


Figura 4.3: Diagrama de blocos da Extracção de Características

### 4.3.1 Características de Forma

Para a extracção destas características foi utilizada uma técnica interactiva[30, 31] que permitiu com o mínimo de interacção do usuário traçar o contorno da região suspeita nas ROIs seleccionadas(ver figura 4.4a)).

A técnica utilizada é comumente conhecida como *Live-wire* (ou *intelligent scissors*) e permite a segmentação da lesão patológica de uma forma rápida e precisa.

Esta técnica tem duas componentes fundamentais:

Uma função de custo local que permite um menor custo para obter o contorno e um processo de expansão de forma otimizada de modo a obter um contorno óptimo do objecto de interesse, com base na função de custo e dos pontos fornecidos pelo usuário.

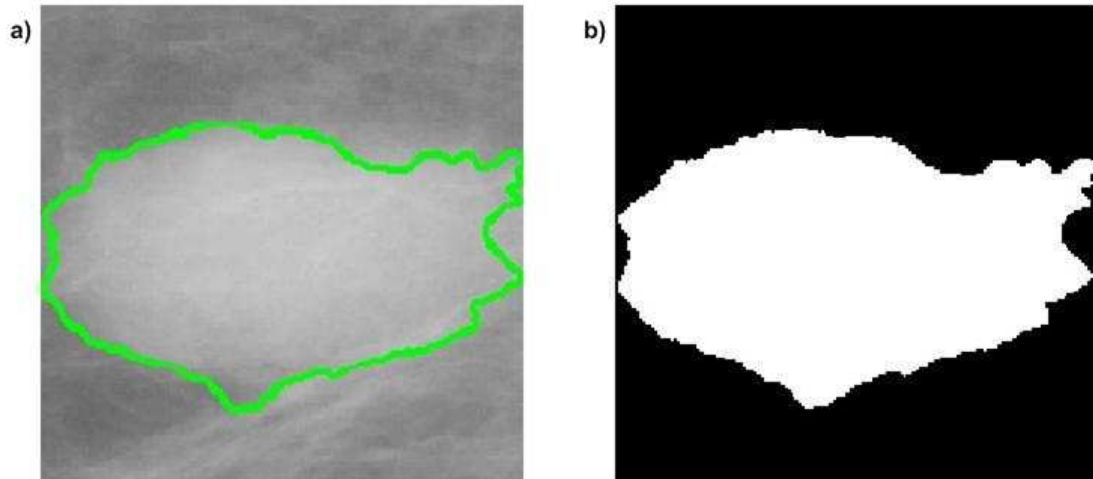


Figura 4.4: a) Segmentação de uma ROI com a técnica *Live-wire* b) Aplicação da função *poly2mask*

Esta função devolve um vector com as coordenadas (x,y) de todos os pixéis pertencentes à fronteira.

Com esta informação, é possível calcular o perímetro, tal como está descrito no capítulo 3.3.1. Para o cálculo da área recorreu-se primeiramente à função *poly2mask*, de modo a converter a ROI numa imagem binária, em que na parte interior da massa segmentada com a técnica *Live-wire* fica a um e da parte de fora fica a zero (ver figura 4.4b)), sendo assim possível obter a área do modo descrito no capítulo 3.3.1.

O brilho foi obtido recorrendo à função *Mean* do Matlab, que calcula a média da intensidade de níveis de cinza presentes no objecto segmentado, antes de ter utilizado a função *poly2mask*.

A rugosidade ( $R$ ) e a irregularidade ( $I$ ) são medidas derivadas do perímetro ( $P$ ) e da área ( $A$ ) do objecto, sendo respectivamente calculadas de acordo com as fórmulas 3.4 e 3.5.

### 4.3.2 Características de Textura

Tal como representado no diagrama de blocos da figura 4.3, numa primeira fase são calculadas as matrizes simétricas de co-ocorrência entre pares de níveis de cinza (matriz SGLD) para cada região de interesse.

Estas matrizes foram obtidas com recurso à função *graycomatrix* da Toolbox de Processamento de Imagens do Matlab.

Foram assim obtidas quatro matrizes de co-ocorrência para uma distância  $d=1$  e as variações angulares mostradas na figura 3.2, para cada ROI.

Com base nas matrizes de SGLD, Haralick et al.[26, 27] determinaram 14 medidas, previamente apresentadas no capítulo 3.3.2. Entretanto, *Ohanian e Dubes*[32] mostraram que apenas um pequeno grupo de características é suficiente para caracterizar a textura.

Assim, foram extraídas oito características para cada uma das 4 matrizes de co-ocorrência, sendo elas: Segundo Momento Angular, Contraste, Correlação, Variância, Momento da Diferença Inversa, Média da Soma, Entropia da Soma e Entropia.

De realçar, que esta função antes do cálculo das características, normaliza as matrizes SGLD de acordo com a equação 3.15, de maneira que a soma de todos os seus elementos (de cada matriz) seja um.

Haralick [26, 27] demonstra que os valores obtidos para estas quatro matrizes, devem ser combinados através da média entre eles, de modo a formar o vector final de probabilidade de co-ocorrência.

Foi assim implementada a função *ObterTextura* que recebe como parâmetro de entrada a ROI de uma dada imagem, devolvendo 5 vectores de 8 linhas por uma coluna, onde os valores representam cada uma das características pela ordem acima apresentada.

Assim sendo, o primeiro vector corresponde às características obtidas através da matriz SGLD para uma orientação de  $0^\circ$  (e  $180^\circ$ ) e assim sucessivamente, sendo que o último vector e que foi utilizado na parte experimental deste trabalho corresponde ao vector final de probabilidade de co-ocorrência.

## 4.4 Análise e Tratamento dos dados

Antes de se proceder à configuração da arquitectura da rede neuronal procedeu-se à análise e tratamento dos dados obtidos.

Esta análise consistiu no tratamento dos chamados *outliers*, na normalização e na separação dos conjuntos de treino e teste, tal como é evidenciado no diagrama da figura 4.5.

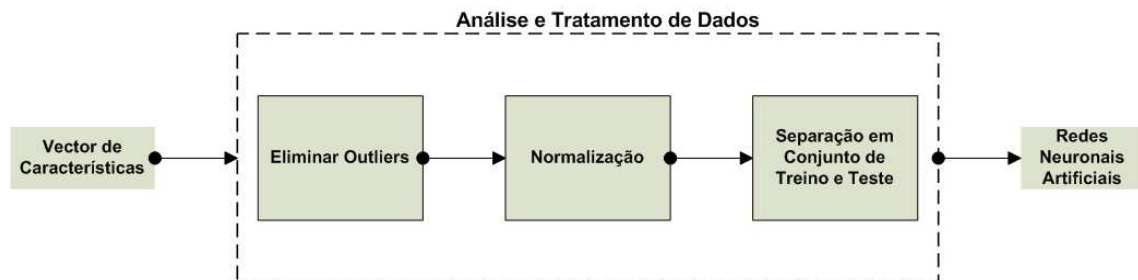


Figura 4.5: Diagrama de blocos dos procedimentos a efectuar no vector de características antes da entrada nas RNAs

### 4.4.1 Tratamento dos *Outliers*

Os outliers são considerados os pontos fora da curva, ou seja, os dados que possuem valores que não estão coerentes com os restantes.

Torna-se assim necessário tratar os outliers, ajustando-os ou eliminando-os (nem sempre é possível), pois estes dificultam a convergência da rede, podendo até impedir a aprendizagem. Neste processo foram eliminadas seis ROI's que apresentavam valores díspares dos restantes.

Uma ROI pertencente à classe “*Circumscribed Masses*”, três pertencentes à “*Miscellaneous (ill-defined masses)*” e duas pertencentes à classe “*Asymmetry Masses*”.

#### 4.4.2 Normalização

A normalização dos dados consiste em colocar os valores numa mesma faixa, de modo a evitar que uma dimensão se sobreponha às outras e que o aprendizado fique estagnado. Existem vários métodos de normalização, sendo utilizado o da equação 4.1, que permite que os dados sejam normalizados em torno da sua média e do seu desvio padrão.

$$X_{new} = \frac{x - \bar{x}}{\sigma_x} \quad (4.1)$$

A função implementada denominada *normalizar* tem como parâmetro de entrada o vector contendo os dados que se pretende normalizar e retorna na saída o vector normalizado.

$$dataNorm.X = normalizar(data.X)$$

Os dados ficam assim normalizados com média igual a zero e variância igual a um.

#### 4.4.3 Separação dos conjuntos de treino e teste

O método utilizado é comumente conhecido como “*Leave-one-out*” (por vezes chamado de “*round-robin*” ou “*jack-knife*”), uma vez que a quantidade de dados disponível era reduzida (67 ROIs consideradas).

Nesta abordagem, um conjunto de dados de tamanho N é dividido em N maneiras diferentes num conjunto de treino de tamanho N-1 e num conjunto de teste de tamanho 1.

Para cada uma das N divisões, o treino é executado com N-1 dados e em seguida, o teste é feito com o dado que ficou de fora.

A performance é depois estimada com a média dos N testes.

Para uma mais fácil interpretação, a figura 4.6 retrata o funcionamento deste método.

### 4.5 Redes Neurais Artificiais

Depois de ter os vectores dos dados para treino e teste devidamente separados e normalizados procede-se numa primeira fase ao treino do classificador, seguindo uma fase de classificação dos dados de teste e por fim avalia-se o erro.

Para o treino foram utilizados 3 tipos de RNAs, o *Perceptron*, o *Pocket* e uma MLP com algoritmo de aprendizagem *backpropagation*.

#### 4.5.1 Estrutura Linear

As primeiras 2 RNAs referenciadas (*Perceptron* e *Pocket*)[33] utilizam a aprendizagem *perceptron* para encontrar um hiperplano de separação para os dados de treino passados como parâmetros de entrada juntamente com a classe a que pertencem.



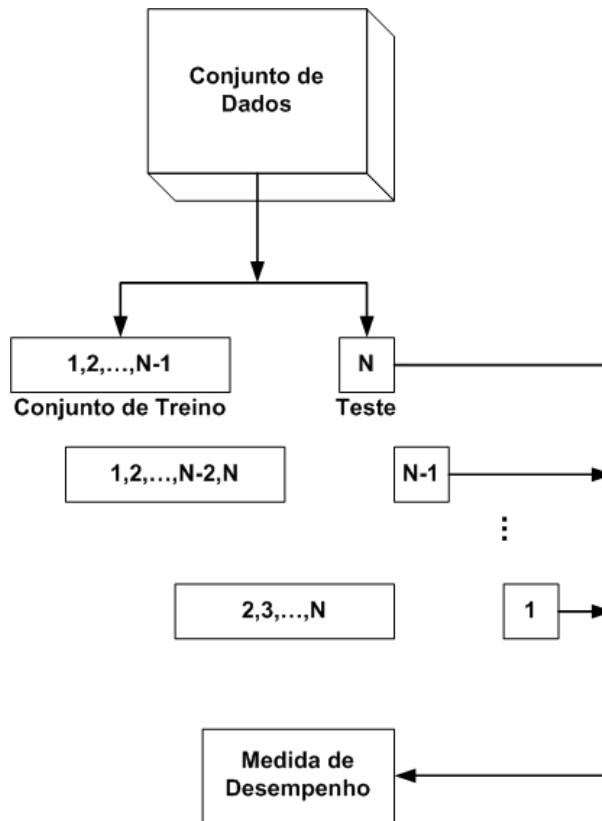


Figura 4.6: Método *Leave-One-Out*. Adaptado de [2]

O treino de um classificador linear binário para um erro nulo é equivalente a encontrar o hiperplano:

$$H = \{x. < w.x > + b = 0\} \quad (4.2)$$

que separa os vectores de treino para a classe  $y = 1$  da classe  $y = 2$ .

O problema é formalmente definido resolvendo o seguinte conjunto de inequações lineares:

$$< w.x_i > + b \geq 0, \quad y_i = 1 \quad (4.3)$$

$$< w.x_i > + b < 0, \quad y_i = 2 \quad (4.4)$$

Onde  $x_i$  representa os dados de treino passados nos parâmetros de entrada com a correspondente classe  $y_i$ .

Se estas inequações tiverem solução então o conjunto de dados é linearmente separável.

Pode ser passado um parâmetro opcional que especifica uma condição de paragem, sendo neste caso, o número máximo de iterações (*options.tmax*).

É possível ainda especificar o modelo inicial que inclui o vector normal (*model.w*), o bias do hiperplano (*model.b*) e o número da iteração (*model.t*).

Caso estes não sejam especificados, serão inicializados a zero e o número de iterações máximo é colocado a infinito, sendo que, o algoritmo só termina quando o *perceptron* convergir.

Os parâmetros de saída são o model que irá ser utilizado na fase de classificação dos dados de teste. Trata-se de uma estrutura que contém os parâmetros do classificador linear, sendo de destacar o vector normal ( $.w$ ) e o bias ( $.b$ ) do hiperplano.

A diferença entre estes dois classificadores prende-se com o modo como é feita a actualização dos pesos ao longo das iterações.

Enquanto que no *perceptron*, em cada iteração, os pesos são actualizados, no *Pocket* os pesos só são actualizados se necessário. Em cada iteração é executada uma rotina do algoritmo *perceptron*, escolhendo aleatoriamente um padrão de entrada do conjunto de treino.

Se o neurónio corrente não proporcionar a saída desejada, então o vector peso é actualizado. Desta forma, o vector peso que satisfaça um número óptimo de padrões de entrada tem uma maior probabilidade de se manter inalterável durante as iterações.

Assim, durante a execução, o algoritmo *Pocket* mantém o vector peso que não tenha sido modificado durante o maior número de iterações.

Na segunda fase é utilizada a função *linclass* de modo a classificar o vector dos dados de teste usando para isso uma função discriminante linear.

Neste caso particular binário um que  $y = \{1, 2\}$ , o classificador linear, dado o parâmetro  $w \in \mathbb{R}^n$  e o bias  $b \in \mathbb{R}$ , é representado por uma única função discriminante.

$$f(x) = \langle w.x \rangle + b \quad (4.5)$$

O vector de entrada  $x \in \mathbb{R}^n$  é atribuído à classe  $y \in \{1, 2\}$  da seguinte forma:

$$q(x) = \begin{cases} 1 & \text{if } f(x) \geq 0 \\ 2 & \text{if } f(x) < 0 \end{cases} \quad (4.6)$$

Esta função tem como parâmetros de entrada, o vector contendo os dados de teste e a estrutura model obtida no processo de treino com os parâmetros do classificador linear ( $.w$  e  $.b$ ). Esta função devolve dois vectores, um contendo a classificação dos dados de teste e outro com os valores da função discriminante.

Na terceira fase é utilizada a função *error*, em que se passa a classificação obtida para os dados de teste e a classificação real para esses mesmos dados, com o objectivo de avaliar o erro obtido na classificação.

#### 4.5.2 Estrutura Multicamada

Na segunda fase do processo de classificação foi utilizado uma *Multi-Layer Perceptron* (MLP) com algoritmo de treino *backpropagation*[9].

Trata-se portanto de um algoritmo de treino supervisionado, que implementa o método do gradiente descente na soma dos quadrados dos erros.

A topologia da rede MLP é constituída por uma camada de entrada (*input layer*), uma camada escondida (*hidden layer*) e uma camada de saída (*output layer*).

A função *BackpropagationCGD* apresenta um conjunto de parâmetros a serem informados, tais como: o número de neurónios a serem utilizados na camada escondida, um critério de paragem com erro menor que um valor estipulado, além dos padrões de entrada para treino

(*datatreino.X*), com a classe associada (*datatreino.y*).

Nesta função é ainda passado o dado que havia ficado de fora aquando do uso da técnica *Leave-one-out* para teste (*datatest.X*). Somente depois da informação de todos esses parâmetros, o treino pode ser iniciado.

A saída desta função devolve a classe do dado passado para teste, bem como os pesos para a camada escondida e de saída e ainda o erro associado ao treino.

Um dos problemas encontrados no treino da rede neuronal do tipo MLP diz respeito ao número de neurónios na camada escondida.

Infelizmente não existe qualquer método ou procedimento sistemático e determinístico para o projecto de arquitectura da rede a implementar, assim como para a especificação dos parâmetros envolvidos.

De facto, não há nenhuma concordância entre os diversos trabalhos que vêm sendo desenvolvidos com o objectivo de obter a arquitectura que permite a obtenção da melhor classificação. Para determinar o número de neurónios na camada escondida Freisleben[34] relata que o número de neurónios nessa camada deve ser múltiplo do número de padrões de entrada menos um.

Por outro lado, *Baum e Haussler*[35] sugerem que o número de neurónios para a camada escondida deve ser calculado através da equação 4.7.

$$j = \frac{m \cdot e}{n + z} \quad (4.7)$$

Onde,

- $j$  = número de neurónios na camada escondida
- $m$  = número de dados do conjunto de treino
- $e$  = erro
- $n$  = número de entradas
- $z$  = número de saídas

No entanto, este método é muito dependente da taxa de erro  $e$ .

*Clarence Tan e Gerhard Wittig*[36] consideram este método demasiado restritivo e demonstram que o número de neurónios na camada escondida deve ser superior a um, mas inferior ao número de entradas, formando assim uma estrutura piramidal.

## 4.6 Medidas de Desempenho

As métricas de erro apropriadas para os problemas de classificação envolvem a frequência relativa das decisões correctas e incorrectas.

Adicionalmente, as métricas de nível superior são obtidas a partir do desempenho destas métricas mais simples.

### 4.6.1 Verdadeiro e Falso, Positivos ou Negativos

No contexto de detectar a presença ou ausência de uma anormalidade numa imagem, os termos verdadeiro-positivo, falso-positivo, verdadeiro-negativo e falso-negativo são comumente utilizados.

Nestes termos, o positivo ou negativo refere-se à decisão tomada pelo algoritmo de classificação.

O verdadeiro ou falso relaciona-se de como a decisão do algoritmo está de acordo com o estado clínico actual.

Assim, o verdadeiro-positivo (TP, do inglês *true-positive*) indica que o algoritmo decidiu correctamente e que a anormalidade existe.

O falso-positivo (FP, inglês *false-positive*) indica que o algoritmo decidiu erradamente e que a anormalidade não existe.

O verdadeiro-negativo (TN, do inglês *true-negative*) indica que o algoritmo decidiu correctamente e que a anormalidade não existe.

O falso-negativo (FN, do inglês *false-negative*) indica que o algoritmo decidiu erradamente e que a anormalidade existe.

Estas definições são de fácil interpretação através de figura 4.7.

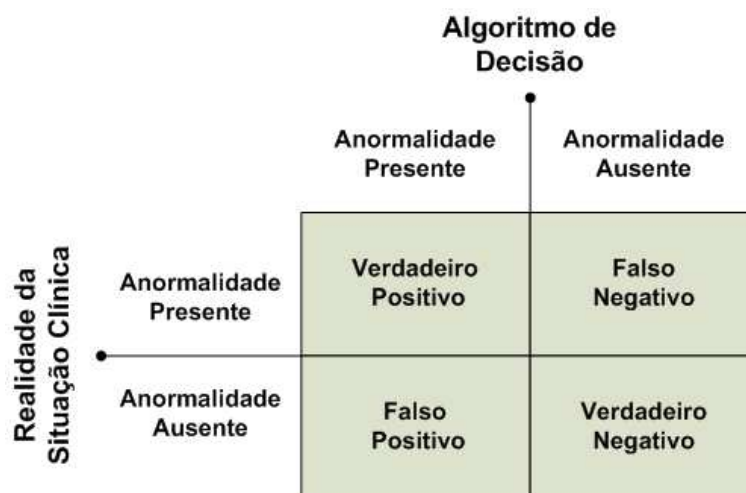


Figura 4.7: Definição de verdadeiro e falso positivos e negativos. Adaptado de [2]

### 4.6.2 Sensibilidade, Especificidade e Precisão

A sensibilidade ( $S$ ) é obtida pelo número de decisões TP a dividir pelo número de casos realmente positivos, tal como demonstra a equação 4.8.

$$S = \frac{TP_s}{TP_s + FN_s} \quad (4.8)$$

A sensibilidade varia assim, entre 0 que indica que nenhuma anormalidade foi detectada e 1 (ou 100%), indicativo de que todas as anormalidades foram classificadas correctamente.

A fracção verdadeiro positivo (TPF) corresponde precisamente à sensibilidade, ou seja, à fracção das anomalias existentes que são correctamente classificadas como tal.

A especificidade de um algoritmo de classificação corresponde ao número de decisões TN a dividir pelo número de casos realmente negativos.

Tal como a sensibilidade, a especificidade varia entre 0 e 1 (ou entre 0 e 100, em termos de percentagem).

A especificidade (E) é dada pela equação 4.9.

$$E = \frac{TN_s}{TN_s + FP_s} \quad (4.9)$$

A fracção falso positivo (FPF) corresponde a  $1 - \textit{Especificidade}$ , ou seja, a fracção dos casos normais que são falsamente classificados como anormais.

Mencionar qualquer um destes dois termos isoladamente é geralmente insignificante.

Uma sensibilidade de 1 é facilmente alcançada para um algoritmo que decide que a anormalidade existe.

Uma especificidade de 1 é facilmente alcançada para um algoritmo que decide que a anormalidade não existe.

Numa situação de ter grande sensibilidade, passa pela aceitação de uma menor especificidade e vice-versa.

A precisão do diagnóstico representa a percentagem de resultados correctamente classificados, podendo ser descrita pela equação 4.10.

$$P = \frac{TP_s + TN_s}{TP_s + TN_s + FP_s + FN_s} \quad (4.10)$$

### 4.6.3 Curvas ROC

Neste trabalho vamos focar na análise da curva característica de resposta do observador (curva ROC).

A curva ROC é tipicamente, um método de avaliação directo e a sua maior vantagem é que se apresenta de forma gráfica, com o TPF (ou sensibilidade) sobre o eixo y e com o FPF (ou  $1 - \textit{especificidade}$ ) sobre o eixo x.

Neste gráfico, o ponto ideal de funcionamento é o canto superior esquerdo, com a fracção TP=1 e a fracção FP=0.

Um gráfico ROC típico é representado na figura 4.8.

A diagonal sobre o gráfico, a partir de  $(\textit{TPF}, \textit{FPF}) = (0,0)$  a  $(\textit{TPF}, \textit{FPF}) = (1,1)$ , representa o nível de desempenho em que a classificação é feita aleatoriamente.

Uma curva ROC trata-se então de uma demonstração bidimensional da performance de um classificador. Para comparar curvas ROC é preciso reduzir estas curvas a um valor escalar.

Um método comum para se proceder à comparação de desempenho entre duas curvas ROC é a área sob a curva ROC (Az).

Esta área pode ser vista como a capacidade do sistema ser, simultaneamente sensível e específico, sendo que quanto maior for, melhor será a precisão diagnóstica da metodologia em análise.

Como a Az é uma porção da área do quadrado unitário (do espaço ROC), o seu valor varia

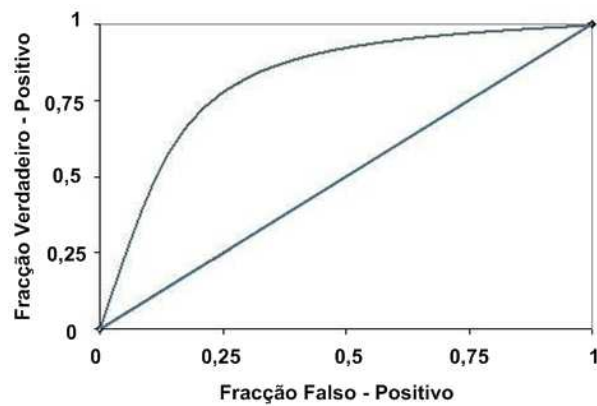


Figura 4.8: Curva ROC (*Receiver Operating Characteristics Curve*)

entre 0.0 e 1.0.

Recorreu-se ao simulador acessível na página web <http://www.rad.jhmi.edu/jeng/javarad/roc/JROCFITi.html>, que de um modo simples e rápido, gera a curva ROC e calcula o valor da área  $A_z$ . Este simulador aceita vários formatos de entrada, sendo escolhido o formato 3 para duas classes, que visa a passagem dos valores obtidos de TP, TN, FP, FN.

## Capítulo 5

# Resultados e discussão

Neste capítulo são apresentados os resultados da aplicação dos classificadores estudados, passando como parâmetros de entrada as diferentes características previamente apresentadas. Os resultados apresentados correspondem aos testes feitos em duas fases distintas, com o objectivo de classificação de nódulos mamários.

Numa primeira abordagem são realizados testes com os classificadores lineares *Perceptron* e *Pocket*, sendo que, para a segunda abordagem de testes fica reservado o uso do classificador multicamada com algoritmo de aprendizagem *Backpropagation*.

### 5.1 Descrição dos Dados de Entrada

A partir das imagens e posterior selecção das ROIs, foram extraídas as diferentes características seguindo a metodologia apresentada no capítulo 4.

Assim sendo, foram formados três conjuntos de características para o processo de treino dos classificadores:

- Conjunto de características  $F_1$ : cinco características de forma extraídas de 67 ROIs
- Conjunto de características  $F_2$ : oito características de textura extraídas de 67 ROIs
- Conjunto de características  $F_3$ : cinco características de forma mais oito características de textura extraídas de 67 ROIs

### 5.2 Classificadores Lineares

Para a comparação dos algoritmos lineares, *Perceptron* e *Pocket*, é importante tentar perceber se a actualização dos pesos no *Pocket* permite uma melhor generalização e consequentemente uma melhor classificação do que a tradicional actualização dos pesos no *Perceptron*.

Teoricamente para os classificadores lineares não existe nenhum critério de paragem. Para um número infinito de iterações e se os dados forem linearmente separáveis, o algoritmo converge, quando todos os elementos do conjunto de treino forem correctamente classificados.

Assim como critério de paragem, procede-se à variação do parâmetro que define o número máximo de iterações, com o intuito de visualizar o número de padrões correctamente aprendidos ao longo das iterações.

A tabela 5.1 mostra os resultados alcançados com o algoritmo *Perceptron*, variando os conjuntos de treino e o número de iterações máximo, ao passo que a tabela 5.2 mostra os resultados com o algoritmo *Pocket*.

Perceptrão									
Conjunto	Topologia	iteração	Classificação no conjunto de treino						
			TN	TP	FN	FP	S (%)	E (%)	P (%)
$F_1$	5/1	100	34	6	18	9	25.0	79.1	59.7
		500	38	4	20	5	16.7	88.4	62.7
		1000	38	2	22	5	8.3	88.4	59.7
		2000	38	2	22	5	8.3	88.4	59.7
		50000	38	4	20	5	16.7	88.4	62.7
$F_2$	8/1	100	32	4	20	11	16.7	74.4	53.7
		500	32	9	15	11	37.5	74.4	61.2
		1000	34	11	13	9	45.8	79.1	67.2
		2000	36	9	15	7	37.5	83.7	67.2
		50000	29	13	11	14	54.2	67.4	62.7
$F_3$	13/1	100	35	5	19	8	20.8	81.4	59.7
		500	38	4	20	5	16.7	88.4	62.7
		1000	38	7	17	5	29.2	88.4	67.2
		2000	37	7	17	6	29.2	86.0	65.7
		50000	30	12	12	13	50.0	69.8	62.7

Tabela 5.1: Treinos realizados com o algoritmo *Perceptron*

Nesta primeira fase foi possível verificar que os dados não são linearmente separáveis, ou seja, não é possível encontrar um hiperplano de separação entre as duas classes, uma vez que, para todas as experiências efectuadas, quer com o *Perceptron*, quer com o *Pocket*, o algoritmo não converge, cumprindo o critério de paragem do número de iterações máximo.

Relativamente às classificações é notória uma diminuição de falsos-negativos (algoritmo decide que a anormalidade não existe quando na realidade ela existe) com o algoritmo *Pocket*, porém um aumento de falsos-positivos (algoritmo decide que a anormalidade existe quando na realidade ela não existe). Assim sendo é possível afirmar que apesar da precisão do diagnóstico ser equivalente nos dois casos, o algoritmo *Pocket* possui geralmente uma sensibilidade superior ao algoritmo *Perceptron*, perdendo para este na especificidade.

### 5.3 Classificadores Não Lineares

Para o processo de treino e teste dos classificadores não lineares, utilizaram-se os mesmos três conjuntos de treino apresentados no capítulo 5.1, com o objectivo de avaliar o processo de classificação de uma rede neuronal artificial multicamada, com algoritmo de aprendizagem *backpropagation*.

Estes conjuntos de características serviram de entrada para a rede MLP, onde se procedeu a vários treinos mudando a topologia da rede e o critério de paragem, conforme tabela 5.3.

Os treinos foram validados através da técnica *leave-one-out* apresentado no capítulo 4.4.3. Foram retiradas as percentagens de verdadeiro-positivo (TP), falso-negativo (FN), falso-



Pocket									
Conjunto	Topologia	iteração	Classificação no conjunto de treino						
			TN	TP	FN	FP	S (%)	E (%)	P (%)
$F_1$	5/1	100	24	16	8	19	66.7	55.8	59.7
		500	30	11	13	13	45.8	69.8	61.2
		1000	26	14	10	17	58.3	60.5	59.7
		2000	25	13	11	18	54.2	58.1	56.7
		50000	26	15	9	17	62.5	60.5	61.2
$F_2$	8/1	100	25	14	10	18	58.3	58.1	58.2
		500	36	5	19	7	20.8	83.7	61.2
		1000	31	11	13	12	45.8	72.1	62.7
		2000	31	10	14	12	41.7	72.1	61.2
		50000	25	15	9	18	62.5	58.1	59.7
$F_3$	13/1	100	31	11	13	12	45.8	72.1	62.7
		500	26	13	11	17	54.2	60.5	58.2
		1000	26	17	7	17	70.8	60.5	64.2
		2000	37	11	13	6	45.8	86.0	71.6
		50000	34	15	9	9	62.5	79.1	73.1

Tabela 5.2: Treinos realizados com o algoritmo *Pocket*

MLP com algoritmo de aprendizagem <i>Backpropagation</i>				
Camada de Entrada		Camada Escondida	Camada de Saída	Critério de Paragem
				Erro <
5	8	13	1	0.1
				0.01
				0.005
				0.001
				0.1
				0.01
				0.005
				0.1
				0.01
				0.005
				0.1
				0.01
				0.005
				0.01
				0.01

Tabela 5.3: Topologias utilizadas para a rede neuronal MLP

positivo (FP), verdadeiro-negativo (TN) e, com base nestes valores, calculada a precisão do diagnóstico, bem como a sensibilidade e especificidade, de acordo com as fórmulas apresentadas no capítulo 4.6.

Procedeu-se ainda ao cálculo da área sob a curva ROC ( $A_z$ ). É importante recordar que quanto maior for o valor da área (mais próxima da unidade), melhor será o comportamento do sistema em análise.

De todos os treinos realizados, foram seleccionados para cada conjunto de dados o melhor resultado conforme tabela 5.4 e geradas as respectivas curvas ROC conforme figura 5.1.

MLP com algoritmo de aprendizagem <i>Backpropagation</i>										
Conjunto	Topologia	Erro <	Classificação							
			TN	TP	FN	FP	S (%)	E (%)	P (%)	$A_z$
$F_1$	5/2/1	0.05	41	3	21	2	12.5	95.3	65.7	0.539
$F_2$	8/3/1	0.01	34	9	15	9	37.5	79.1	64.2	0.583
$F_3$	13/3/1	0.05	36	6	18	7	25.0	83.7	62.7	0.544

Tabela 5.4: Resultados obtidos para cada conjunto de dados com a rede MLP

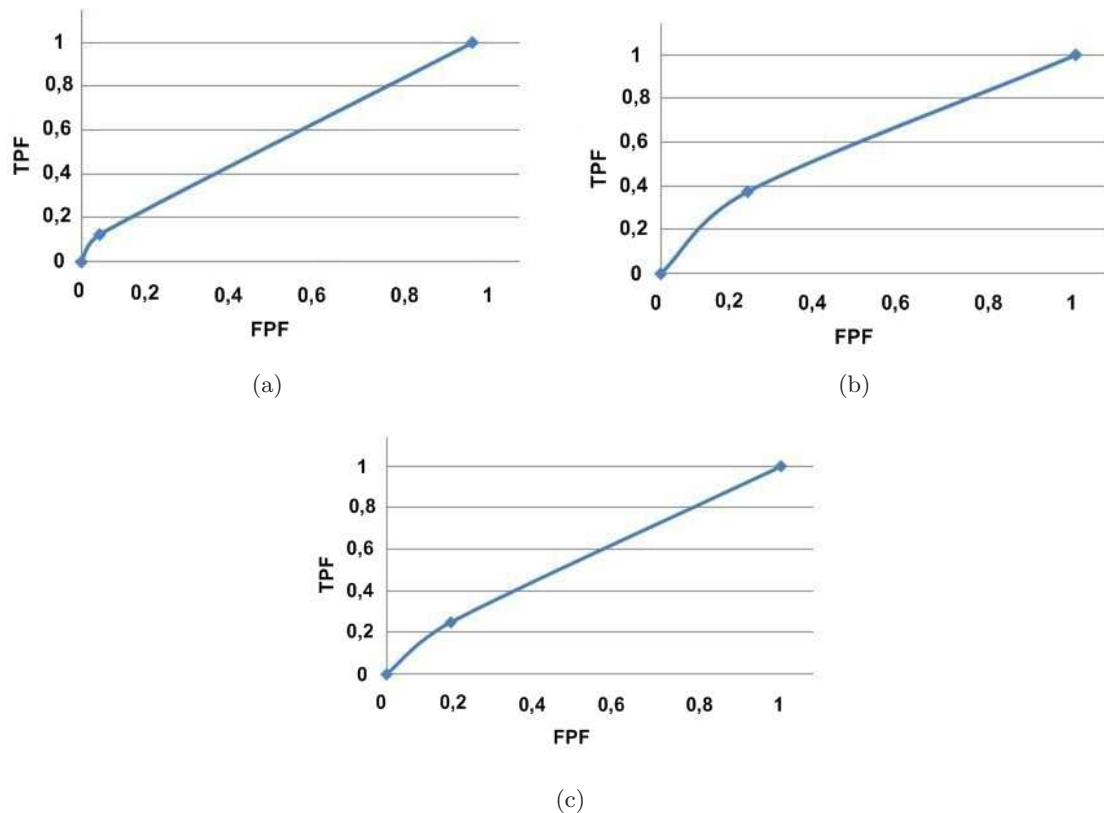


Figura 5.1: Curvas ROC dos resultados obtidos com a rede MLP para a topologia a) 5/2/1, b) 8/3/1 e c) 13/3/1

Foram obtidos valores de  $A_z=0.583$ , com uma precisão diagnóstica de 64.2% para o conjunto de características de textura e valores de  $A_z=0.539$  com uma precisão de 65.7% para o conjunto de características de forma.

Se para o conjunto de textura os valores obtidos estão de acordo com a diversa literatura, o

mesmo não se verifica para as características de forma.

Este facto pode ser em parte explicado por uma importante limitação da metodologia proposta para o contorno dos nódulos e sobretudo pelo processo de validação dos mesmos, não ter sido realizado por um especialista.

De realçar que o tamanho dos dados (67 massas, sendo 43 benignas e 24 malignas) e a sua desproporção (maior número de massas benignas) podem prejudicar uma análise mais precisa dos métodos em estudo.

Assim sendo, efectuou-se uma análise com um conjunto de dados mais equilibrado (igual número de massas benignas e malignas), o que foi conseguido com uma adaptação da técnica *leave-one-out*. Esta adaptação consistiu em manter o mesmo princípio de funcionamento da técnica *leave-one-out* apresentada no capítulo 4.4.3, mas com uma ligeira alteração. Assim para cada iteração, são seleccionados aleatoriamente 23 dados dos 43 classificados como benignos e 23 dos classificados como malignos, deixando em cada iteração um dado maligno de fora.

Nesta situação a técnica *leave-one-out* é executada somente sobre os dados malignos, sendo realizadas para cada conjunto de treino 24 iterações.

A tabela 5.5 mostra os resultados obtidos, sendo de notar, a ausência de casos TN e FP, uma vez que o dado passado para teste é sempre maligno.

MLP com algoritmo de aprendizagem <i>Backpropagation</i>					
Conjunto	Topologia	Erro <	Classificação		
			TP	FN	P (%)
$F_1$	5/2/1	0.05	14	10	58.3
$F_2$	8/2/1	0.05	15	9	62.5
$F_3$	13/3/1	0.05	14	10	58.3

Tabela 5.5: Resultados obtidos com a técnica *Leave-One-Out* modificada

Os resultados alcançados com esta técnica permitiram obter uma precisão semelhante aquando do uso da técnica *leave-one-out* tradicional, não trazendo por isso grandes vantagens.

De facto, se por um lado a técnica *leave-one-out* adaptada permitiu um maior equilíbrio nos dados (igual número de massas benignas e malignas), por outro lado, foi necessária uma redução do número de dados (46 massas, sendo 23 benignas e 23 malignas, contra as 67) para o treino do classificador.

## 5.4 Avaliação das Características de Forma

Os resultados verificados com o conjunto de dados  $F_2$ , que correspondem às características de textura, estão de acordo com a literatura estudada[8, 20, 21] e previamente apresentada no capítulo 3.1. Relativamente aos dados do conjunto  $F_1$  que correspondem às características de forma o mesmo não se verifica.

De facto, em Túlio André e Rangaraj Rangayyan[20] os valores de  $A_z$  estão compreendidos entre 0.48 e 0.63 para as características de textura e valores de 0.75 até 0.99 para as características de forma.

Assim sendo, efectuou-se um estudo para avaliar estas últimas características. Na figura 5.2 são apresentados os valores da área, perímetro e brilho que foram previamente normalizados tal como foi referido no capítulo 4.4.2 para as 67 ROIs estudadas. A rugosidade e a irregularidade não foram tidas em conta uma vez que, dependem exclusivamente do perímetro e da área.

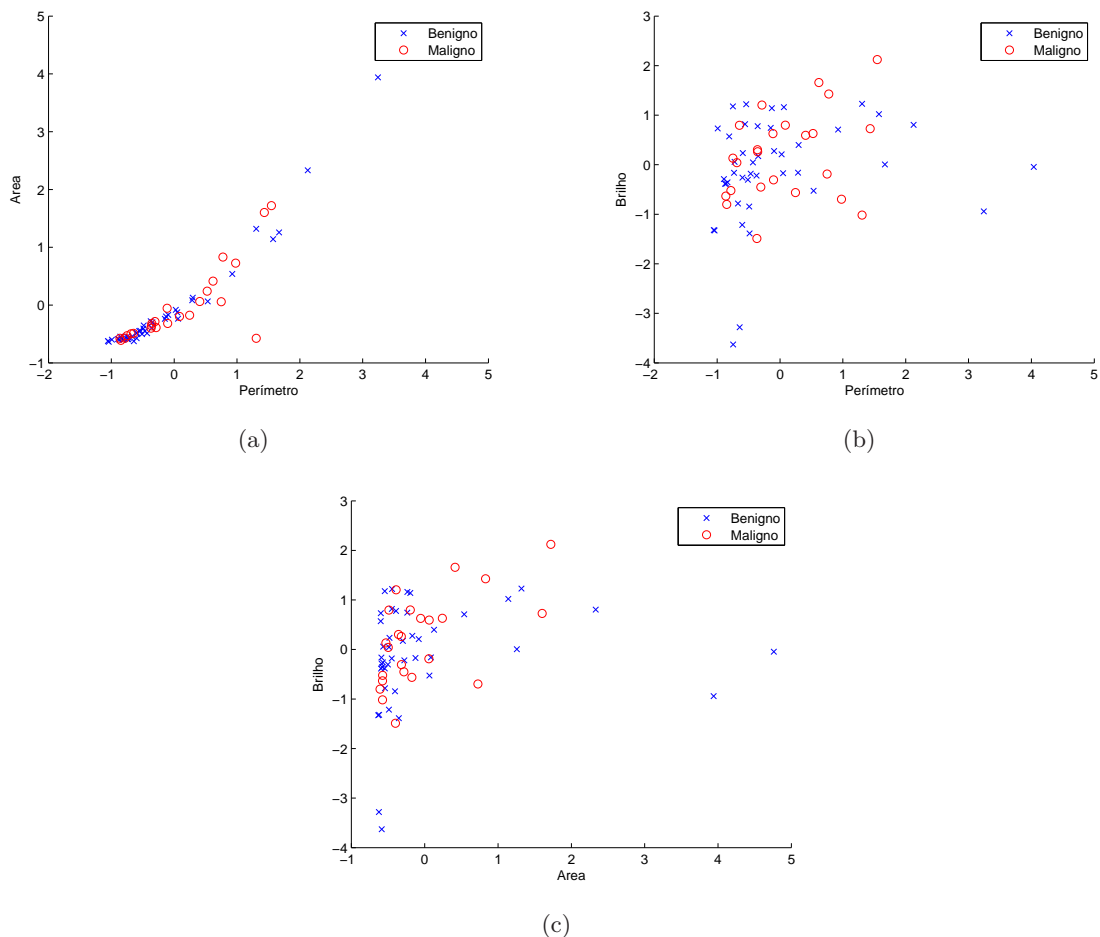


Figura 5.2: Avaliação das características

A análise conjunta de cada par de características mostra que os valores para ambas as classes se encontram misturados, não sendo possível identificar regiões onde uma das classes predomine. A utilização da segmentação semi-automática faz com que a extracção do contorno do nódulo seja demorada e que seja necessária uma boa precisão visual.

Este último factor contribui com informação errada, uma vez que a validação do contorno da massa não foi realizada por um especialista. Mesmo quando realizada por um especialista, muitas das vezes este tem dificuldade em contornar correctamente a massa devido sobretudo, à sobreposição e composição dos tecidos mamários.

Se para alguns nódulos o contorno é bastante perceptível (no caso da base de dados MIAS, os nódulos da classe *Circumscribed Masses*), para outros torna-se extremamente difícil distinguir e delimitar correctamente.

## Capítulo 6

# Conclusões

Neste trabalho foram propostos dois grupos distintos de características, sendo que um abordou as texturas dos nódulos e o outro a sua forma.

Para o processo de classificação dos nódulos mamários foram estudadas três técnicas de classificação baseadas em redes neuronais artificiais de modo a determinar a sua benignidade ou malignidade.

Duas técnicas foram baseadas em redes neuronais artificiais lineares e a terceira baseada em redes multicamada com algoritmo de aprendizagem *Backpropagation*.

A utilização das duas técnicas lineares permitiram concluir que não é possível traçar um hiperplano de separação entre as classes, logo os dados não são linearmente separáveis.

É de realçar contudo, que o processo de aprendizagem com o algoritmo *Pocket*, em que a actualização dos pesos ao longo das iterações difere do algoritmo utilizado no *Perceptron*, permitiu obter uma sensibilidade superior a este, apesar da precisão de diagnóstico ser semelhante em ambos os algoritmos. Assim, o algoritmo *Pocket* revelou ser mais equilibrado no processo de classificação.

Para o processo de classificação utilizando a rede neuronal multicamada com algoritmo de aprendizagem *backpropagation*, os valores obtidos de  $Az$  para o conjunto de treino de textura encontram-se de acordo com a diversa literatura, não sendo contudo características suficientemente discriminatórias de benignidade ou malignidade.

Em relação ao conjunto de características de forma, da qual se esperava (até pela diversa literatura apresentada) melhores resultados, estes contudo, não se verificaram.

Este facto, comprovado pelos valores obtidos de  $Az = 0.539$  com uma precisão de diagnóstico de somente 65,7% é explicado sobretudo pela validação do contorno da massa não ter sido realizado por um especialista.

A análise conjunta de cada par de características de forma permitiu concluir que não é possível identificar regiões onde uma das classes predomine.

O conjunto de características  $F_3$  que advém da junção dos outros dois conjuntos obteve valores de  $Az$  intermédios destes, não trazendo vantagens no processo de classificação dos nódulos mamários.

## 6.1 Trabalhos Futuros

Com a experiência adquirida neste trabalho, são a seguir apresentadas algumas sugestões tidas como primordiais no sentido de melhorar o processo de classificação:

- É imprescindível a obtenção de um maior conjunto de imagens que tenham disponíveis a classificação das lesões através de biópsia;
- Este trabalho restringiu-se apenas a classificar massas, sendo interessante incluir as microcalcificações;
- Como o processo de contorno da massa é extremamente difícil e sujeito a erros é importante que este seja validado por um especialista, diminuindo assim os erros de precisão;
- Incluir parâmetros clínicos dos pacientes, como por exemplo a idade, como entrada para os classificadores;
- Aplicar outras técnicas de classificação, como por exemplo, o *Radial Basis Function* (RBF).

# Bibliografia

- [1] D. B. Kopans, *Breast Imaging*. 1989.
- [2] M. Sonka and J. M. Fitzpatrick, “Medical image processing and analysis,” vol. 2, pp. 567–1004, Bellingham: WA: SPIE Press, 2000.
- [3] W. H. Organization, “Mortality database,” 2004.
- [4] J. Ferlay, F. Bray, P. Pisani, and D. Parkin, “Cancer incidence, mortality and prevalence worldwide,” 2004.
- [5] <http://www.cancer.org/> (Acedido em 5/11/2008).
- [6] P. Borgen, G. Wong, V. Vlamis, and e. al., “Current management of male breast cancer. a review of 104 cases,” *Ann Surg.*, vol. 215, no. 5, pp. 451–459, 1992.
- [7] I. Fentiman, A. Fourquet, and G. Hortobagyi, “Male breast cancer,” *The Lancet*, vol. 367, no. 9510, pp. 595–604, 2006.
- [8] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels, “Gradient and texture analysis for the classification of mammographic masses,” *IEEE Transactions on Medical Imaging*, vol. 19, no. 10, pp. 1032–1043, 2000.
- [9] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley Interscience Publication, 1973.
- [10] R. Duda, D. Nitzan, and P. Barret, “Uses of range and reflectance data to find surface regions,” *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. PAMI.1, no. 3, pp. 259–271, 1979.
- [11] <http://www.mathworks.com/products/matlab/> (Último acesso a 4/11/2008).
- [12] C. J. D’Orsi, L. W. Bassett, S. A. Feig, V. P. Jackson, D. B. Kopans, M. N. Linver, E. A. Sckles, and C. B. Stelling, “Breast imaging reporting and data system (bi-rads),” *Reston, VA: American college of radiology*, 1998.
- [13] D. Getty, R. Pickett, C. D’Orsi, and J. Swets, “Enhanced interpretation of diagnostic images,” *Invest. Radiology*, vol. 23, pp. 240–252, 1988.
- [14] J. Swets, D. Getty, R. Pickett, C. D’Orsi, S. Seltzer, and B. McNeil, “Enhancing and evaluating diagnostic accuracy,” *Med Decis Making*, vol. 11, pp. 9–18, 1991.

- [15] C. E. Floyd Jr, J. Y. Lo, A. J. Yun, D. C. Sullivan, and P. J. Kornguth, "Prediction of breast cancer malignancy using an artificial neural network," *Cancer*, vol. 74, pp. 1944–1948, 1994.
- [16] J. A. Backer, P. J. Kornguth, J. Y. Lo, M. E. Williford, and C. E. Floyd Jr, "Breast cancer: prediction with artificial neural network based on bi-rads standardized lesions," *Radiology*, vol. 196, pp. 817–822, 1995.
- [17] J. Y. Lo, J. A. Backer, P. J. Kornguth, and C. E. Floyd Jr, "Computer-aided diagnosis of breast cancer: artificial neural network approach for optimized merging of mammographic features," *Academic Radiology*, vol. 2, pp. 841–850, 1995.
- [18] J. A. Backer, P. J. Kornguth, J. Y. Lo, and C. E. Floyd Jr, "Artificial neural network: improving the quality of breast biopsy recommendations," *Radiology*, vol. 198, pp. 131–135, 1996.
- [19] M. Giger, "Computer-aided diagnostic of breast lesions in medical images," *Computing in science engineering*, vol. 2, pp. 39–45, 2000.
- [20] T. C. Andre and R. M. Rangayyan, "Classification of tumors and masses in mammograms using neural networks with shape and texture features," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 3, (Cancun, Mexico), pp. 2261–2264, 2003.
- [21] P. B. Ribeiro, *Classificação por análise de contornos de nódulos mamários utilizando redes neurais artificiais*. PhD thesis, Universidade de São Paulo, 2006.
- [22] H. P. Chan, K. Doi, C. Vyborny, R. Schmidt, K. Metz, T. Ogura, Y. Wu, and H. MacMahon, "'improvement in radiologists" detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," *Invest. Radiology*, vol. 25, no. 10, pp. 1102–1110, 1990.
- [23] M. L. Giger, "Computerized analysis of images in the detection and diagnosis of breast cancer," *Seminars in Ultrasound, CT, and MRI*, vol. 25, no. 5, pp. 411 – 418, 2004. Computers in Radiology.
- [24] K. Rodenacker, "A feature set for cytometry on digitized microscopic images," *Analytic Cellular Pathology*, vol. 25, no. 1, pp. 1–36, 2003.
- [25] R. Gonzalez and R. Woods, *Digital Imaging Processing*. second ed., 2002.
- [26] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems Man, and Cybernetics*, vol. SMC 3, no. 6, pp. 610–621, 1973.
- [27] R. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, 1979.
- [28] A. K. Jain, J. Mao, and K. Mohiuddin, "Artificial neural networks: A tutorial," *IEEE Computer*, vol. 29, pp. 31–44, 1996.



- [29] J. Suckling, J. Parker, and et al., “The mammographic image analysis society digital mammogram database,” in *Proc. 2nd International Workshop Digital Mammography*, pp. 375–378, 1994.
- [30] G. Hamarneh and J. Hradsky, “DTMRI Segmentation using DT-Snakes and DT-Livewire,” *IEEE International Symposium on Signal Processing and Information Technology*, pp. 513 – 518, 2006.
- [31] A. Chodorowski, U. Mattsson, M. Langille, and G. Hamarneh, “Color lesion boundary detection using live wire,” *Proceedings of SPIE Medical Imaging: Image Processing*, vol. 5747, pp. 1589–1596, 2005.
- [32] P. P. Ohanian and R. C. Dubes, “Performance evaluation for four classes of textural features,” *Pattern Recognition*, vol. 25, no. 8, pp. 819–833, 1992.
- [33] <http://cmp.felk.cvut.cz/cmp/software/stprtool/> (Acedido em 5/11/2008).
- [34] B. Freisleben, “Stock market prediction with backpropagation networks,” in *Proceedings of the 5th international conference on Industrial and engineering applications of artificial intelligence and expert systems* (Springer-Verlag, ed.), vol. 604, pp. 451–460, 1992.
- [35] E. B. Baum and D. Haussler, “What size net gives valid generalization?,” *Neural Computation*, vol. 1, no. 1, pp. 151–160, 1989.
- [36] C. N. Tan and G. E. Wittig, “A study of the parameters of a backpropagation stock price prediction model,” *IEEE Computer Society Press*, pp. 288–291, 1993.



# Apêndice A

## Programas

### A1. Algoritmo para o cálculo das características de Forma

```
function [ValForma] = ObterForma(ROI)
%Calcula 5 características de forma
%Inicialmente através da função lwcontour é gerado o contorno da massa,
%sendo retornado os pontos pertencentes ao contorno segmentado.
%Botão esquerdo do rato para criar o ponto inicial e traçar o contorno.
%Para terminar o programa e devolver as coordenadas, pressionar qualquer
%tecla.
%Em seguida é calculada cada uma das 5 características:
%Forma 1: Perimetro
%Forma 2: Area
%Forma 3: Brilho
%Forma 4: Rugosidade
%Forma 5: Irregularidade
%
%Inputs:
%ROI – Região de interesse.
%
%Outputs:
%ValForma – Valores relativos às características apresentadas, pela ordem
%acima referidas.
c = lwcontour(ROI,1);
C=poly2mask(c(1,:),c(2,:),length(ROI),length(ROI));
C=C';
d=find(C==0);
ROI(d)=[];
%Cálculo das características:
ValForma(1:5,1)=0;
%Perimetro
ValForma(1,1)=length(c);
%Area
ValForma(2,1)=sum(C(:));
%Brilho
ValForma(3,1)=mean(ROI);
```

```

%Rugosidade
ValForma(4,1)=(ValForma(1,1)^2)/(4*pi*ValForma(2,1));
%Irregularidade
ValForma(5,1)=ValForma(2,1)/(ValForma(1,1)^2);

```

## A2. Algoritmo para o cálculo das características de Textura

```

function [Par0,Par45,Par90,Par135,ParMedio]=ObterTextura(ROI);
% Calcula 8 parâmetros de Haralick.
% Inicialmente a função calcula através da função graycomatrix do Matlab
% as matrizes de co-ocorrência para as 4 orientações e normaliza-as.
% Em seguida calcula para cada uma delas as seguintes características:
%
% Haralick 1: Segundo Momento Angular
% Haralick 2: Contraste
% Haralick 3: Correlação
% Haralick 4: Soma dos Quadrados – Variância
% Haralick 5: Momento da Diferença Inversa
% Haralick 6: Entropia
% Haralick 7: Média da Soma
% Haralick 8: Entropia da Soma
%
% Inputs:
% ROI = região de interesse com 256 níveis de cinza
%
% outputs:
% Par0 – valores relativos a matriz com orientação de 0 graus
% Par45 – valores relativos a matriz com orientação de 45 graus
% Par90 – valores relativos a matriz com orientação de 90 graus
% Par135 – valores relativos a matriz com orientação de 135 graus
% ParMedio – Média dos valores anteriores

%Cálculo das 4 Matrizes de Co-ocorrência
offsets = [0 1;-1 1;-1 0;-1 -1];
GLCM=graycomatrix(ROI, 'offset', offsets, 'NumLevels',256, 'Symmetric', true);

% Cálculo das 4 Matrizes de probabilidade de co-ocorrência entre níveis de
% cinza
for i=1:4
    Matriz_CoOcorrencia_Pro(:, :, i)=GLCM(:, :, i)/sum(sum(GLCM(:, :, i)));
end

%Cálculo das características na direção 0°
Par0 = CalcValTextura(Matriz_CoOcorrencia_Pro(:, :, 1));
%Cálculo das características na direção 45°
Par45 = CalcValTextura(Matriz_CoOcorrencia_Pro(:, :, 2));
%Cálculo das características na direção 90°

```

```

Par90 = CalcValTextura(Matriz_CoOcorrencia_Pro(:, :, 3));
%Cálculo das características na direção 135°
Par135 = CalcValTextura(Matriz_CoOcorrencia_Pro(:, :, 4));
ParMedio = (Par0 + Par45 + Par90 + Par135)/4;

% Cálculo das 8 características de Haralick
function Par=CalcValTextura(glcm)
Par(1:8,1) = 0;
% Matrizes Auxiliares
is=0:255;
is=is(ones(1,256),:);
js=is';
% Calcula Haralick 1: Segundo Momento Angular
energy = glcm.^2;
Par(1,1) = sum(energy(:));
% Calcula Haralick 2: Contraste
s = size(glcm);
[c,r] = meshgrid(1:s(1),1:s(2));
r = r(:);
c = c(:);
term1 = abs(r - c).^2;
term2 = glcm.^1;
Par(2,1)=sum(term1 .* term2(:));
% Calcula Haralick 3: Correlação
mr = sum(r .* glcm(:));
Sr = sqrt(sum((r - mr).^2 .* glcm(:)));

mc = sum(c .* glcm(:));
Sc = sqrt(sum((c - mc).^2 .* glcm(:)));

term1 = sum((r - mr) .* (c - mc) .* glcm(:));
Par(3,1) = term1 / (Sr * Sc);
% Calcula Haralick 4: Soma dos Quadrados - Variância
Par(4,1)= sum(sum(((r - mr).^2).*glcm(:)));
% Calcula Haralick 5: Momento da Diferença Inversa
term1 = (1 + abs(r - c).^2);
Par(5,1) = sum(glcm(:) ./ term1);
% Calcula Haralick 6: Entropia
glcm1=glcm;
glcm1(glcm1==0) = [];
Par(6,1)= -1*sum(glcm1(:).*log2(glcm1(:)));
% Calcula Haralick 7: Média da Soma
for n=0:512
    Mask=abs(is+js)==n;
    Pxmaisy=sum(sum(Mask.*glcm));
    Pxmaisy2(n+1,1)=Pxmaisy;
    if (Pxmaisy ~= 0)

```

```

        Par(8,1)=Par(8,1)+Pxmaisy*log2(Pxmaisy);
    end;
end
Par(7,1)=sum(Pxmaisy2.*([0:5 1 2]'));
% Calcula Haralick 8: Entropia da Soma
Par(8,1) = -1 * Par(8,1);

```

### A3. Algoritmo para a Normalização

```

function ValXNorm = normalizar(ValX)
%Esta função normaliza os dados passados como parâmetros de entrada em
%torno da sua média e do seu desvio padrão.
%
%Inputs:
%ValX – Valores que se pretende normalizar
%
%Outputs:
%ValXNorm – Valores normalizados
%
[x,y]=size(ValX);
for i=1:x
    Media(i)=mean(ValX(i,:));
    Desvio(i)=std(ValX(i,:));
end
for i=1:x
    for j=1:y
        ValXNorm(i,j)=(ValX(i,j)-Media(i))/Desvio(i);
    end
end
end

```

### A4. Técnica *Leave-one-out*

```

function [datatreino, datateste]= LeaveOneOut(data,i)
%Esta função aplica a técnica Leave-One-Out, de modo a separar o conjunto
%de dados passado como parâmetro de entrada, num conjunto para treino e
%outro para teste.
%
%Inputs:
%data – conjunto de dados normalizados
%i – número da iteração
%
%Outputs:
%datatreino – conjunto de dados para treino
%datateste – elemento deixado de fora para teste
%
[linhas, colunas]=size(data.X);
if (i==1)
    datatreino.y=data.y(:,2:colunas);

```

```

    datatreino.X=data.X(:,2:colunas);
    datatestey=data.y(:,i);
    datatestex=data.X(:,i);
elseif (i==67)
    datatreino.y=data.y(:,1:colunas-1);
    datatreino.X=data.X(:,1:colunas-1);
    datatestey=data.y(:,i);
    datatestex=data.X(:,i);
else
    datatreino.y=[data.y(:,1:i-1),data.y(:,i+1:colunas)];
    datatreino.X=[data.X(:,1:i-1),data.X(:,i+1:colunas)];
    datatestey=data.y(:,i);
    datatestex=data.X(:,i);
end

```

#### A5. Algoritmo para treino e teste do classificador *Perceptron*

```

function [TN,TP,FN,FP]= CorrerPerceptron(data,tmax)
%Esta função serve para analisar o classificador linear Perceptron quanto à
% sua convergência.
%
%Inputs:
%data – conjunto de dados normalizados
%tmax – condição de paragem: número de iterações máximo
%
%Outputs:
%TN – verdadeiros-negativos
%TP – verdadeiros-positivos
%FN – falsos-negativos
%FP – falsos-positivos
%
options.tmax=tmax;
model=perceptron(data,options);
[ypred,dfce]=linclass(data.X,model);
TN=0;
TP=0;
FN=0;
FP=0;
for i=1:length(data.X)
if ypred(i)==data.y(i)
if ypred(i)==1
TN=TN+1;
else
TP=TP+1;
end
else
if ypred(i)==1

```

```

        FN=FN+1;
    else
        FP=FP+1;
    end
end
end
end

```

#### A6. Algoritmo para treino e teste do classificador *Pocket*

```

function [TN,TP,FN,FP]= CorrerPocket(data ,tmax)
%Esta função serve para analisar o classificador linear Pocket quanto à
%sua convergência.
%
%Inputs:
%data – conjunto de dados normalizados
%tmax – condição de paragem: número de iterações máximo
%
%Outputs:
%TN – verdadeiros–negativos
%TP – verdadeiros–positivos
%FN – falsos–negativos
%FP – falsos–positivos
%
options.tmax=tmax;
model=pocket(data ,options);
[ypred ,dfce]=linclass(data.X,model);
TN=0;
TP=0;
FN=0;
FP=0;
for i=1:length(data.X)
if ypred(i)==data.y(i)
    if ypred(i)==1
        TN=TN+1;
    else
        TP=TP+1;
    end
else
    if ypred(i)==1
        FN=FN+1;
    else
        FP=FP+1;
    end
end
end
end
end

```

#### A7. Algoritmo para treino e teste do classificador MLP



```

function [TN,TP,FN,FP]= CorrerBackpropagation(data,Nunit,error)
%Esta função utiliza a técnica Leave-One-Out, para treinar e testar o
%classificador multicamada, com algoritmo de aprendizagem Backpropagation.
%Retorna os valores de TN,TP,FN e FP de modo a avaliar o desempenho do
%classificador.
%
%Inputs:
%data – conjunto de dados normalizados
%Nunit – número de neurónios para a camada escondida
%error – condição de paragem: erro menor que valor definido
%
%Outputs:
%TN – verdadeiros-negativos
%TP – verdadeiros-positivos
%FN – falsos-negativos
%FP – falsos-positivos
%
TN=0;
TP=0;
FN=0;
FP=0;
i=0;
ans=find(data.y==2);
data.y(ans)=0;
[linhas,colunas]=size(data.X);
for i=1:colunas
    [datatreino, datateste]= LeaveOneOut(data,i);
    [ypred, Wh, Wo, errors] = BackpropagationCGD...
    (datatreino.X, datatreino.y, datateste.X, [Nunit, error]);
    if ypred==datateste.y
        if ypred==1
            TN=TN+1;
        else
            TP=TP+1;
        end
    else
        if ypred==1
            FN=FN+1;
        else
            FP=FP+1;
        end
    end
end
end

```

**A8. Algoritmo para treino e teste do classificador MLP utilizando conjunto de treino com igual número de casos benignos e malignos**

```

function [TN,TP,FN,FP]= CorrerBackpropagationAdaptado(data,Nunit,error)
%Esta função utiliza uma variante da técnica Leave-One-Out, de modo a
%equilibrar o conjunto de dados passados como parâmetros de entrada para o
%processo de treino do classificador (igual número de massas benignas e
% malignas).
%Retorna os valores de TN,TP,FN e FP de modo a avaliar o desempenho do
%classificador.
%De notar, que esta função só deve ser utilizada quando o número de massas
% malignas é inferior ao número de massas benignas.
%
%Inputs:
%data – conjunto de dados normalizados
%Nunit – número de neurónios para a camada escondida
%error – condição de paragem: erro menor que valor definido
%outputs:
%TN – verdadeiros-negativos
%TP – verdadeiros-positivos
%FN – falsos-negativos
%FP – falsos-positivos
TN=0;
TP=0;
FN=0;
FP=0;
i=0;
ans=find(data.y==2);
data.y(ans)=0;
n=length(ans);
for i=1:n
    h=find(data.y==1);
    number=randsample(h,n-1);
    datatreino.X=data.X(:,number);
    datatreino.y=data.y(:,number);
    j=find(data.y==0);
    datateste.y=data.y(j(i));
    datateste.X=data.X(:,j(i));
    if (i==1)
        datatreino.X=[datatreino.X,data.X(:,j(2:n))];
        datatreino.y=[datatreino.y,data.y(:,j(2:n))];
    elseif (i==24)
        datatreino.X=[datatreino.X,data.X(:,j(1:n-1))];
        datatreino.y=[datatreino.y,data.y(:,j(1:n-1))];
    else
        datatreino.X=[datatreino.X,data.X(:,j(1:i-1)),data.X(:,j(i+1:n))];
        datatreino.y=[datatreino.y,data.y(:,j(1:i-1)),data.y(:,j(i+1:n))];
    end
    [ypred,Wh,Wo,errors]=BackpropagationCGD...
    (datatreino.X,datatreino.y,datateste.X,[Nunit,error]);

```

```
if ypred==datatest.y
    if ypred==1
        TN=TN+1;
    else
        TP=TP+1;
    end
else
    if ypred==1
        FN=FN+1;
    else
        FP=FP+1;
    end
end
end
```