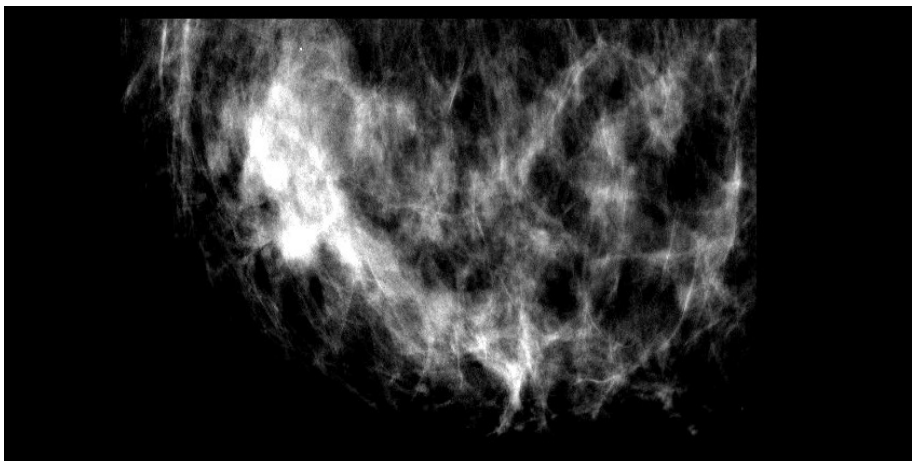
INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Eletrónica e Telecomunicações e de Computadores



Classificação automática de cancro da mama em imagiologia por micro-ondas

JOSÉ PAULO DOS SANTOS NUNES

(Bacharel)

Dissertação para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : Doutora Matilde Pós-de-Mina Pato
Doutor Pedro Renato Tavares Pinho

Júri:

Presidente: Doutor Nuno Miguel Machado Cruz

Vogais: Doutor Artur Jorge Ferreira
Doutora Matilde Pós-de-Mina Pato

FEVEREIRO, 2019

A ti Isabel, que deste sentido à minha vida

Agradecimentos

Em primeiro lugar, gostaria de agradecer aos meus orientadores, Dra. Matilde Pós-de-Mina Pato e Dr. Pedro Pinho, pelo apoio, orientação, ensinamentos e revisão do meu trabalho. Agradeço também ao Instituto de Telecomunicações pelo suporte à realização desta dissertação. Agradeço a todos os professores do Mestrado pelos conhecimentos que partilham que, sem exceção, me fizeram crescer quer na minha vertente profissional quer pessoal. Tarde, mas ainda em bom tempo, descobri a falta que faz continuar os estudos académicos, alargando o horizonte de competências, ao longo da carreira profissional. Agradeço de forma especial aos professores das cadeiras relacionadas com a Inteligência Artificial pelo mundo novo que me deram a descobrir, e pelas novas formas de raciocinar. Não poderei também esquecer-me de agradecer aos meus colegas de mestrado que se mostraram verdadeiros companheiros de caminhada e me ajudaram a vencer obstáculos e a chegar até aqui. A última palavra vai para a minha família e amigos que me deram o apoio necessário para concretizar esta aventura. De um modo muito particular os meus filhos, Inês e António, que muitas vezes foram obrigados a prescindir do meu apoio e dedicação, conseguindo sempre superar as minhas faltas e recebendo-me sempre com um sorriso.

Resumo

Ao longo das últimas duas décadas, a MWI (imagiologia por micro-ondas) tem atraído um crescente interesse para aplicações em diversas áreas da medicina, nomeadamente no diagnóstico do cancro da mama. A razão prende-se com o contraste dielétrico significativo existente entre os tecidos normais e os tecidos tumorais da mama (os tecidos tumorais apresentam um teor de água mais elevado) que quando expostos a frequências da gama das micro-ondas provocam uma maior dispersão do sinal.

Paralelamente a essas investigações decorrem estudos sobre a classificação desses sinais de micro-ondas utilizando técnicas de ML (aprendizagem automática). Enquadrada nesse âmbito, a presente dissertação pretende criar modelos de classificação e validar a viabilidade prática da mesma, com o objetivo de auxiliar os técnicos de saúde a tomarem as melhores decisões sobre os tratamentos e intervenções a efetuar em cada caso clínico.

Neste estudo, após a preparação e separação dos dados, é realizada a extração de caraterísticas através de DWT (Transformada discreta de *wavelet*), *Wavelets* interpolatórias e PCA (Análise de componentes principais), seguida da classificação usando SVM (Máquinas de vetores de suporte), LDA (Análise discriminante linear) e Random Forests. O melhor resultado de classificação do tamanho do tumor foi obtido com SVM e PCA, tendo sido obtido consistentemente melhor desempenho sempre que foi utilizada a média dos sinais das antenas. Também na classificação do tamanho do tumor se verificaram melhores resultados do que na classificação da sua malignidade.

Palavras-chave: Imagiologia por micro-ondas; Cancro da mama; Aprendizagem automática; Modelos de classificação

Abstract

Over the last two decades, the MWI (Microwave Imaging) has spiked interest in applications on several branches of medicine, namely in breast cancer diagnosis. The significantly dielectric contrast between normal and tumor tissues (tumor tissues having a higher water content) produces a noticeably different response of those tissues to the microwave radiation.
Investigations about the classification of those signals using Machine Learning, are also running along with these developments. This study will attempt to create classification models for microwave signals taken from the MWI, with the goal of optimizing the best choices of health specialists when it comes to performing treatments and interventions in each different clinic episode.
In this study, after data pre-processing and splitting, feature extraction was done using DWT (Discrete Wavelet Transform), Interpolatory Wavelets and PCA (Principal Components Analysis), followed by the classification itself with SVM (Support Vector Machines), LDA (Linear Discriminant Analysis) and Random Forests. The best result for tumour size was achieved with SVM and PCA, being observed better classification performance when the average of the antennas' signals was used. Classification of tumor size also obtained better results than the classification of its malignity.

Keywords: Microwave Imaging; Breast Cancer; Machine Learning; Classification Models.

Lista de acrónimos

AI *Artificial Inteligence*

CNN *Convolutional Neural Networks*

DNN *Deep Learning Neural Network*

DWT *Discret Wavelet Transform*

EEG Eletroencefalograma

EMD *Empirical mode decomposition*

FN *False Negative*

FP *False Positive*

ICA *Independent Component Analysis*

IMF *Intrinsic mode functions*

KNN *K-nearest neighbours*

LDA *Linear Discriminant Analysis*

LMNN *Largest margin nearest neighbour*

ML *Machine Learning*

MWI *Microwave Imaging*

PCA *Principal Component Analysis*

QDA *Quadratic Discriminant Analysis*

RBF *Radial Basis Function*

RF *Random Forests*

SVM *Support Vector Machines*

TN *True Negative*

TNR *True Negative Rate*

TP *True Positive*

TPR *True Positive Rate*

Índice

Lista de Figuras

Lista de Tabelas

Listagens

1

Introdução

Estudos para o desenvolvimento de meios e métodos que permitem a eficiência da gestão do conhecimento têm-se tornado imprescindíveis desde a informatização dos sistemas. O suporte à tomada de decisão está frequentemente associado a aplicações que envolvem a análise e a exploração de dados e que providenciam mecanismos de alto nível para auxílio a esses processos. Face à enorme quantidade de dados a manipular tornou-se imperativo que esses mecanismos sejam implementados em sistemas computacionais. As recentes evoluções no que se convencionou designar por Inteligência Artificial, em inglês AI (*Artificial Inteligence*), popularizaram ferramentas e técnicas que permitem simular comportamentos inteligentes e extrapolar informação e conhecimento a partir de enormes quantidades de dados. Uma das vertentes da AI encontra-se na área denominada de Aprendizagem Automática, em inglês ML (*Machine Learning*), a qual tem como objetivo agrupar enormes quantidades de informação complexa em classes com significado relevante para as mais diversas aplicações.
Nomeadamente na biomedicina técnicas de ML têm sido estudadas e utilizadas para a classificação de resultados de exames de rastreio de diversas doenças, por exemplo o cancro da mama.

Paralelamente a utilização de frequências na gama das micro-ondas tem merecido a atenção de vários investigadores nos últimos anos, resultando numa miríade

de estudos, protótipos e textos científicos. As investigações abrangem nomeadamente as seguintes vertentes: a tecnologia relacionada com as antenas e circuitos, a emissão e captação dos sinais, as frequências e tipos de radiação mais adequadas por tipo de aplicação, até ao processamento e análise dos sinais obtidos [6]. São exemplos nomeadamente os projetos apresentados pelo IEEE Microwave Theory & Techniques Society em [7].

As potenciais aplicações desta gama de frequências distribuem-se por variadas vertentes e áreas, tais como sistemas de RADAR (aplicações militares, polícia, segurança, desportos), comunicações, transportes, processamento industrial, medicina, ciência, exploração e computação.

Ao longo das últimas duas décadas, a MWI (Imagiologia por micro-ondas, do inglês *Microwave Imaging*) tem atraído um crescente interesse para aplicações em diversas áreas da medicina, nomeadamente na monitorização terapêutica e em exames de rastreio.
Especificamente na área da medicina os estudos sobre utilização de frequências na gama das micro-ondas distribuem-se por diferentes especialidades e com objetivos diferentes, nomeadamente em:

- **Deteção de doenças**, como por exemplo cancro da mama e AVC [8];
- **Tratamentos e cirurgias**, como por exemplo tratamento hipotérmico de tumores, tratamento da dor, doença do refluxo gástrico, cirurgias de redução da próstata, remoção de cicatrizes cirúrgicas, ablação endometrial [7].

Os exames atualmente disponíveis para o diagnóstico de cancro da mama apresentam algumas limitações e contraindicações, tais como:

1. **Mamografia** - utilização de raios-X, radiação ionizante, desconforto devido à compressão da mama [9], e assinalável percentagem de diagnósticos errados [10];
2. **Ecografia mamária** - depende da qualidade do aparelho e do técnico, difícil deteção de tumores de pequenas dimensões devido ao ruído [11].

Atendendo às desvantagens existentes nos atuais métodos para diagnóstico do cancro da mama, tem merecido maior atenção a utilização de frequências na gama das micro-ondas em exames da região mamária para diagnóstico precoce do cancro da mama [12]. As principais razões que motivam este elevado interesse prendem-se com as suas caraterísticas não invasivas e não ionizantes [6], e por ser potencialmente de baixo custo [13].

1.1 Enquadramento

Em todo o mundo existem milhões de pessoas que sofrem de cancro. A taxa de incidência do cancro continua a aumentar à escala mundial, devido ao envelhecimento e crescimento da população, mas também, pela adoção de comportamentos associados (por exemplo, tabagismo e inatividade física) ao desenvolvimento deste tipo de doença. Dados estatísticos comprovam que o cancro é a principal causa de morte nos países economicamente desenvolvidos e a segunda principal causa de morte nos países em desenvolvimento.
Em particular o cancro da mama é o tumor maligno com maior incidência nas mulheres na Europa e aparece como segunda causa de morte de cancro na mulher. Estima-se que na população europeia (estados membros da EU), uma em cada oito mulheres irá desenvolver cancro da mama até aos 85 anos, sendo que 20% dos casos acontecem em mulheres antes dos 50 anos de idade e 37% entre os 50 e os 64 anos [14].
Em Portugal, em 2012, houve cerca de 6.000 novos casos e cerca de 1.500 pessoas perderam a vida devido a este cancro. A taxa de mortalidade mundial foi, nesse ano, de 18,4/100.000 para uma incidência estimada em 85,6/100.000 [15].
É sabido que a mortalidade por cancro da mama tem vindo a diminuir na última década em resultado da aposta no diagnóstico cada vez mais precoce da doença e também em razão da otimização das abordagens terapêuticas para esta doença [16].

1.2 Objetivos da dissertação

Com esta dissertação pretende-se apresentar uma metodologia para a classificação em ML de sinais de exames de micro-ondas para diagnóstico de cancro da mama, que agrega a preparação e divisão dos dados com as técnicas de extração de caraterísticas e com os algoritmos de classificação. São assim construídas e validadas diversas arquiteturas e processos de classificação de sinais micro-ondas obtidos por técnicas inovadoras de diagnóstico precoce a partir de diversos modelos de tumores benignos, malignos, de diferentes tipos e dimensões. O presente trabalho surge na sequência e como complemento ao trabalho realizado por Conceição [17] na sua tese de doutoramento onde são desenvolvidas técnicas de MWI para deteção precoce de cancro da mama. Por se tratar de uma técnica ainda em fase de estudo, existem algumas preocupações com a qualidade dos resultados

obtidos e a sua eficiente aplicabilidade na medicina prática. Uma das sugestões para trabalho futuro indicadas nesse trabalho de Conceição consiste na aplicação de DWT (Transformada discreta de *wavelet*, na sigla inglesa) para extração de caraterísticas e a aplicação de SVM (Máquinas de vetores de suporte, na sigla inglesa) para classificação [17].

1.3 Organização da dissertação

Esta dissertação encontra-se dividida em cinco capítulos, correspondendo o presente capítulo à introdução. Na sequência desta introdução, o capítulo 2 apresenta o estado da arte da classificação em ML de sinais de micro-ondas em vários domínios de aplicação, estabelecendo-se um enquadramento teórico das técnicas e algoritmos usados nesta dissertação, assim como uma motivação para o uso de extração de caraterísticas como forma de reduzir a dimensionalidade dos dados a classificar. São apresentados os diversos métodos para preparação e divisão dos dados a classificar, assim como as técnicas de extração de caraterísticas e os algoritmos de classificação utilizados.
O capítulo 3 descreve o trabalho realizado, sendo realizada uma breve análise aos resultados obtidos e respetivas melhorias implementadas no capítulo 4.
Por fim, no capítulo 5 são apresentadas a análise crítica e as conclusões do trabalho prático realizado, e também são apontadas sugestões para trabalhos futuros.

1.4 Contribuição original

Na presente dissertação irá ser experimentada uma nova técnica de extração de caraterísticas baseada em *Wavelets* interpolatórias (tendo por base a tese de doutoramento de Pinho [1]) por forma a avaliar a possibilidade da sua utilização no processo de classificação de sinais de micro-ondas.
É também objetivo original desta dissertação analisar detalhadamente qual o grau de influência da etapa de preparação e divisão dos dados no desempenho final de todo o processo de classificação. Diversos estudos anteriores têm-se debruçado sobre a influência das técnicas de extração e dos algoritmos de classificação, e respetivas parametrizações, na qualidade do resultado final do processo de classificação, mas nenhum dos estudos consultados durante a elaboração desta dissertação aborda com o devido detalhe a fase de preparação e divisão dos dados.

2

Estado da arte

Neste capítulo é feita uma breve apresentação de alguns trabalhos realizados acerca da classificação de sinais de micro-ondas, com utilização de ML, com o objetivo de caraterizar o estado da arte nesse domínio de investigação. Aproveita-se a oportunidade para referir os diversos métodos, técnicas e algoritmos associados à classificação automática em ML, apresentando de forma clara e não exaustiva os seus respetivos conceitos essenciais, assim como as respetivas vantagens e desvantagens.
Começa-se por referir a constituição típica de um processo completo de classificação, os diversos módulos que o constituem, descrevem-se os métodos de preparação e divisão dos dados, as técnicas de extração de caraterísticas e os algoritmos de classificação.
São depois apresentados exemplos de investigações já realizadas relacionados com o tema desta dissertação e onde são utilizados esses conceitos.

2.1 Processos de classificação

Nomeadamente através da análise ao estado da arte da classificação automática foi possível concluir que todo o processo de classificação pode integrar vários por módulos encadeados, os quais incluem (Figura 2.1):

1. A preparação dos dados;

2. A extração e a redução de caraterísticas;

3. Os algoritmos de classificação automática.

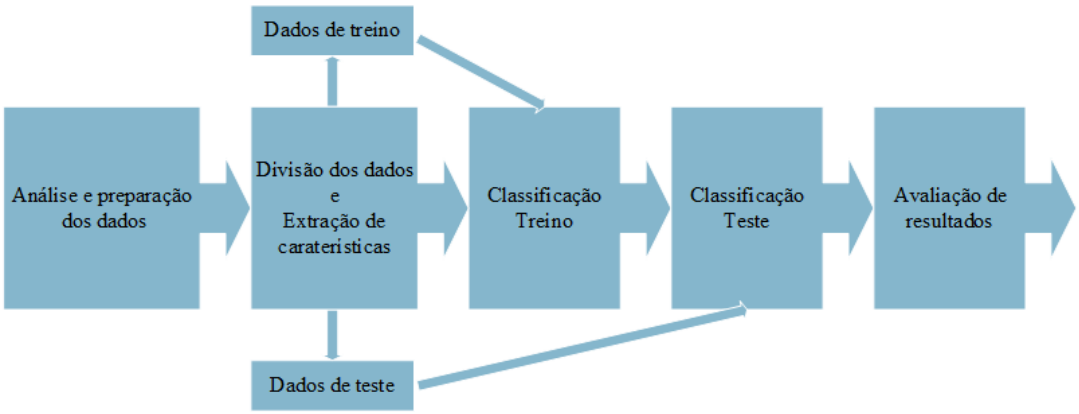


Figura 2.1: Processo de classificação

Todos estes módulos são determinantes no desempenho global da classificação, visando os dois primeiros simplificar e tornar mais eficaz a aprendizagem do algoritmo de classificação.

2.2 Preparação e divisão dos dados

A tarefa de preparação e divisão de dados é uma das mais sensíveis dentro da análise de dados, pois a qualidade dos dados reflete-se diretamente na qualidade do conhecimento gerado a partir dos mesmos [18]. Os métodos a considerar nesta etapa do processo de classificação são os seguintes:

Preparação e transformação dos conjuntos de dados
Adaptação dos dados à plataforma aplicacional onde irão ser processados e adição de atributos calculados tendo já em vista o objetivo final da classificação.
Neste módulo o conjunto de dados original é importado do seu formato original e convertido num formato adequado ao seu processamento na plataforma aplicacional escolhida. Podem posteriormente ser adicionados atributos adicionais, calculados com base nos seus atributos originais tendo em vista a simplificação e otimização da classificação dos dados.

Seleção e divisão dos conjuntos de dados (treino, teste e validação)
Na disciplina de aprendizagem automática é comum a divisão dos dados disponíveis em três conjuntos: um usado na aprendizagem para treinar o algoritmo de classificação, um para o validar (e otimizar), e um terceiro de teste para verificar o desempenho da classificação. É de referir que os dados do conjunto de teste não devem ter sido ser usados na fase de aprendizagem, pois iria falsear os resultados.
A utilização dos conjuntos de validação e de teste no treino iria também potenciar a sobre aprendizagem do algoritmo, ou seja, o algoritmo poderia ficar demasiado especializado nas amostras disponíveis para a criação do modelo e seria totalmente inútil na classificação de novas amostras [18].
Uma regra de bom senso na divisão nestes três conjuntos de dados é usar 50% dos dados disponíveis para treino, 25% para validação e 25% para teste [19].
Também é frequente dispensar o conjunto de validação quando a quantidade de amostras disponíveis é insuficiente para tal. Nesses casos a divisão aconselhada é 2/3 para treino e 1/3 para teste [18].

Balanceamento dos dados
O equilíbrio dos dados a analisar relativamente às classes a classificar constitui outro problema, pois interfere na qualidade da aprendizagem e influencia o desempenho final da classificação. Os algoritmos tendem a especializar-se na classe maioritária e ignoram as amostras da classe minoritária [20, 21].

2.2.1 *Holdout*

Holdout (preservar) é o método de divisão dos dados que precisamente descreve o acto de preservar uma parte dos dados (normalmente 1/3) para teste, ficando com os restantes (2/3) para treino. Esta divisão deve ser aleatória e tentar garantir a existência de proporcionalidade na representatividade de cada classe em ambos os conjuntos [18].

2.2.2 k-Fold Cross Validation

A necessidade de medir o desempenho do algoritmo de classificação e a frequente escassez do conjunto de dados disponível está na origem deste método de teste de validação cruzada. Com k-Fold os dados disponíveis são divididos (*folded*)

em k porções, todas de igual dimensão, mantendo a proporcionalidade entre as amostras de cada uma das classes. À vez, guarda-se uma das porções para teste usando-se as outras *k-1* porções para treino. O processo repete-se k vezes e no final mede-se o desempenho juntando os resultados dos k testes. São frequentemente usados os valores 5 e 10 para valor de k, passando nesses casos a técnica a ser designada por 5-Fold ou 10-Fold [19]. Cada uma destas iterações treino-teste é equivalente ao *Holdout*, mas difere no modo de divisão do conjunto de dados uma vez que neste caso o conjunto de treino é constituído por 50% ou 90% (5-Fold ou 10-Fold) das amostras e o conjunto de teste fica com os restantes 50% ou 10%.

2.2.3 Bootstrap

Este método é utilizado principalmente quando o conjunto de dados disponível é reduzido. E resolve esse problema através da repetição (reamostragem) de amostras.
Suponhamos que temos um conjunto com 10 amostras. Escolhemos aleatoriamente por dez vezes uma amostra e adicionamo-la ao conjunto de treino. Este conjunto de treino ficará com 10 amostras incluindo certamente algumas repetições. As amostras que não forem escolhidos formarão o conjunto de teste.
Probabilisticamente demonstra-se que existe uma probabilidade de $0,368$ de uma amostra não ser escolhida e, portanto, figurar no conjunto de teste, ou seja, o conjunto de teste ficará com $36,8\%$ das amostras [18].

2.3 Técnicas para extração de caraterísticas

A finalidade da utilização de técnicas de extração de caraterísticas é obter uma redução na dimensionalidade dos dados, tendo como consequências a redução do tempo computacional da classificação e a otimização dessa classificação.

O processo de extração de caraterísticas pode ser definido como:

- Por um lado, a escolha das caraterísticas distintivas de um sinal, as que o definem concretamente no âmbito do problema a resolver;
- Por outro lado, trata-se de descartar caraterísticas redundantes ou de pouca importância na definição do sinal.

Este processo além de simplificar, reduzir a dimensionalidade dos sinais a tratar, diminuindo a necessidade de capacidade de processamento dos algoritmos de classificação, também melhora o seu desempenho ao colocar em evidência as caraterísticas que melhor distinguem os sinais uns dos outros.

2.3.1 Transformada Discreta de *Wavelet*

As transformadas discretas de *Wavelet* são um modo de extração de caraterísticas muito popular em aprendizagem automática para sinais no domínio do tempo e frequência. Fazem a análise do sinal no domínio do tempo-frequência projetando os vetores de entrada num novo espaço de caraterísticas. As caraterísticas resultantes da aplicação da DWT são os coeficientes da *wavelet*. São esses coeficientes que serão usados na classificação em vez do sinal original.
Existem diversos métodos (filtros) para concretizar essa transformação, sendo os mais conhecidos Haar [22], Daubechies [23], Least Asymetric, Best Localized e Coiflet. No entanto existe a noção de que não é tanto o tipo de filtro que influencia o resultado mas sim o comprimento da *wavelet*, sendo que quanto maior for esse comprimento maior será a sensibilidade às diferenças entre os grupos de classificação [24].
Outro factor com importante influência nos resultados é a escolha da quantidade de níveis de decomposição. Essa escolha deve ter como base os componentes de frequência dominantes dos sinais a classificar [25].

2.3.2 *Wavelets* interpolatórias

Foi incluído no âmbito deste estudo a utilização de *Wavelets* interpolatórias para extração de caraterísticas dos sinais de micro-ondas.
Embora não tivesse sido encontrada qualquer referência à sua utilização para extração de caraterísticas em estudos anteriores, as *Wavelets* interpolatórias endereçam adequadamente o objetivo da redução da dimensionalidade dos sinais, dado que simplificam enormemente uma função ortogonal permitindo a sua representação usando apenas os pontos que não são possíveis de calcular através da interpolação de outros pontos.
A motivação para o desenvolvimento e estudo das *Wavelets* interpolatórias surgiu no âmbito da investigação de novas técnicas (numéricas adaptativas) para a resolução de equações diferenciais (particularmente as equações de Maxwell),

aproveitando o recente progresso dos recursos computacionais realizada por Pinho na sua tese de doutoramento. O fundamento das *Wavelets* interpolatórias é o seguinte: uma função f representada por N pontos numa grelha uniforme pode ser representada também numa base de *wavelets*, com um erro ϵ , por N_s pontos (representação esparsa) em que N_s é muito menor que N. Verifica-se que para calcular os coeficientes de interpolação não é necessário realizar nenhuma operação complexa (tal como por exemplo o cálculo de integrais), o que representa uma grande vantagem em relação à representação dessas funcões noutras bases [1].
A interpolação de um determinado ponto é realizada a partir dos seus $2p$ pontos vizinhos mais próximos. A interpolação pode ser linear ($p = 1$), em que se usam os dois pontos vizinhos mais próximos, ou cúbica ($p = 2$) onde são utilizados os 4 pontos vizinhos mais próximos [1].
Uma vez que o cálculo está dependente dos pontos vizinhos poderá haver problemas nas fronteiras da função. Isso acontece na interpolação cúbica, pois vamos necessitar de pontos que caem fora da fronteira. O modo utilizado nesta implementação para resolver esse problema foi a utilização de simetria. Neste modo considera-se a existência de pontos simétricos no exterior da função (ver Figura 2.2) [1].

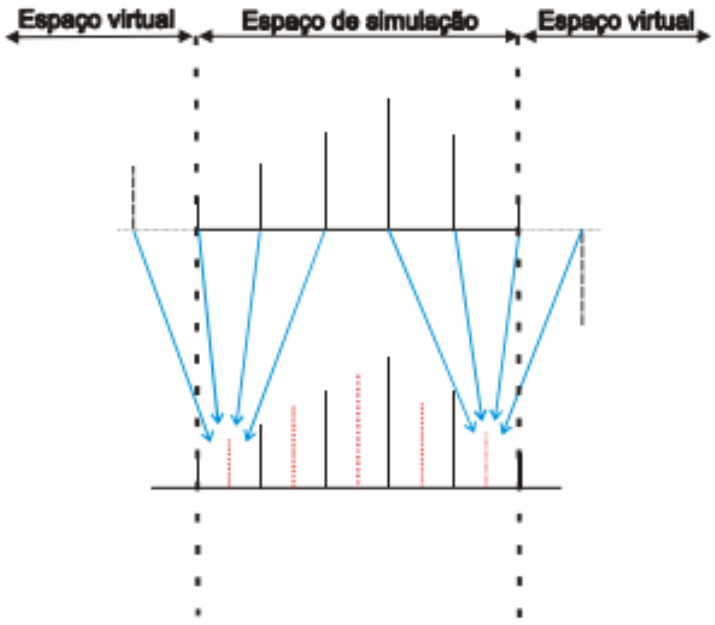


Figura 2.2: Condições fronteira para a interpolação cúbica (simetria) [1]

A interpolação é realizada por vários níveis e recursivamente, pelo que um ponto pode ser representado pela interpolação de 2 pontos (ou 4), e estes por sua vez representados pela interpolação de 2 pontos (ou 4), e assim sucessivamente.

Quando o valor da diferença entre o valor da função num ponto e o respetivo valor interpolado é considerado abaixo de um dado limiar (erro de interpolação ϵ) esse detalhe assume-se nulo. Deste modo, obtém-se uma representação esparsa da função. Esta é formada por um número de pontos N_s inferior ao número de pontos N da função no nível mais fino [1].
Este método permite uma grande simplificação dos dados, tendo como resultado a representação dos sinais através de um vetor esparso que contém apenas os pontos absolutamente necessários para definir o sinal, que são os pontos fundamentais da função (coeficientes significativos) e que a caraterizam univocamente. Dessa forma adequa-se a ser utilizado como técnica de extração de caraterísticas dos sinais a classificar.

2.3.3 Análise de Componentes

A PCA (Análise de componentes principais, na sigla inglesa), a ICA (Análise de componente independente, na sigla inglesa) e a LDA (Análise de discriminante linear, na sigla inglesa) são técnicas de extração de caraterísticas.

PCA
A PCA procura encontrar correlações entre as várias caraterísticas (componentes) dos dados resultando na sua representação através dos seus componentes principais. A redução de caraterísticas através da representação dos dados pelos seus componentes principais permite expressar de uma forma mais simplificada a variabilidade dos sinais. O primeiro componente principal é a combinação linear das caraterísticas dos dados que apresenta a máxima variância de todas as combinações, por isso reflete a máxima variabilidade possível dos dados. São definidos sucessivos componentes principais, não correlacionados uns com os outros, como sendo as combinações lineares com sucessivamente menor variância das anteriores. Dependendo da análise que se pretende fazer decide-se o número de componentes principais a considerar. Cada componente principal reflete a variação dos dados numa determinada percentagem, cada vez mais pequena. Assim, se for razoável trabalhar com apenas 80% da variação dos dados, utilizam-se apenas os primeiros n componentes principais que cumulativamente refletem essa percentagem de variação.
De notar que não existe maneira de determinar o número de componentes principais a utilizar em cada caso, sendo essa uma decisão intuitiva e baseada na experiência [26].
Na implementação da PCA em R a função *prcomp* permite definir (entre outros

parâmetros) uma tolerância, percentagem da variância de cada sucessivo componente principal relativa ao primeiro componente principal, e *rank*, o número máximo de componentes principais a utilizar [27]. Ambos os parâmetros limitam a quantidade de componentes principais retornados pela função.

ICA

A ICA baseia-se na independência probabilística de cada característica (componente) dos dados, ou seja que não podem ser inferidas umas a partir das outras. Partindo do princípio que cada um dos dados é resultante de uma mistura a ICA tem como objetivo obter as características fonte cuja mistura deu origem a esses dados [25]. A ICA foi desenvolvida para resolver o problema da separação cega de fontes: separar os sinais originais de um sinal obtido pela mistura de vários sinais, sem se ter conhecimento prévio da forma como cada sinal original contribuiu para o sinal final. A ICA pressupõe que os vários sinais foram misturados linearmente.

O funcionamento da ICA é intuitivamente percebido através do exemplo do som de, por exemplo, uma festa. O som oriundo dessa festa é composto pelos vários sons de música, vozes, dança, tilintar de copos, etc. A ICA pretende obter os sons elementares que deram origem a esse som, como se existisse um microfone em cada fonte sonora [28].

LDA

O objetivo da LDA é obter uma única variável composta, o discriminante, a partir da combinação linear das características (componentes) originais do sinal. Resulta daí que o discriminante não é mais que uma uma combinação linear ponderada das características do sinal original. O valor do discriminante **D** é calculado de acordo com a equação (2.1) onde, **w** são os pesos a aplicar aos valores das várias dimensões originais do sinal **Z**. Os valores dos pesos **w** são escolhidos de modo a maximizar a diferença entre a média dos discriminantes da cada classe [25].

$$D = w_1 Z_1 + w_2 Z_2 + w_3 Z_3 + ... + w_p Z_p \tag{2.1}$$

Considerando um determinado número de componentes independentes que melhor descrevem os dados, a LDA seleciona a combinação linear que apresenta a maior diferença, em média, entre as classes.

Tanto a PCA como a LDA tentam encontrar a combinação linear das características que melhor representam os dados. Mas a LDA tenta objetivamente encontrar as diferenças entre as classes existentes nos dados, daí também ser utilizada diretamente como um classificador, embora seja mais frequentemente utilizado como

técnica de extração de caraterísticas [29]. No estudo [25] no domínio da classificação de sinais de eletroencefalogramas (EEG) foram encontradas vantagens na aplicação de PCA, ICA e LDA aos coeficientes de *wavelet* obtidos da aplicação de DWT a esses sinais (para a classificação foi utilizado o algoritmo SVM).
Os resultados desse estudo demonstram um melhor desempenho do algoritmo de classificação SVM quando aplicado PCA, ICA e LDA aos coeficientes *wavelet* obtidos pela aplicação de DWT ao sinal de EEG. De entre os três métodos o melhor desempenho foi obtido com LDA [25].

2.4 Algoritmos de classificação

Classificadores, também designados por algoritmos de aprendizagem automática (ML), são modelos baseados em algoritmos matemáticos que, através de um processo de aprendizagem, podem classificar automaticamente ocorrências (amostras) de um determinado domínio. Por exemplo, podem-se ensinar classificadores a classificar clientes em categorias, o risco de um pedido de crédito, se um tumor é maligno ou benigno, por exemplo.
O processo de aprendizagem é supervisionado, ou seja, são usadas amostras de teste cuja classificação é conhecida à partida, e a aprendizagem é feita em lote. Seguindo o exemplo dos pedidos de crédito, o treino do algoritmo de classificação é realizado com base em pedidos de crédito anteriores que foram analisados e classificados por pessoas especialistas na matéria. Após a aprendizagem o sistema deverá poder classificar automaticamente novos pedidos de crédito, com um bom desempenho (elevada percentagem de acertos).
Contribuem para o bom desempenho dos algoritmos de classificação, a qualidade dos dados disponíveis para treino, a preparação desses dados, o processo de extração de caraterísticas, e a própria adequação do algoritmo de classificação às caraterísticas dos dados a classificar.

Existem técnicas que permitem melhorar os resultados dos algoritmos de classificação, mesmo que estes já sejam bons, pois pretende-se atingir os melhores resultados possíveis. Uma das técnicas é reunir vários algoritmos de classificação (mesmo que estes isoladamente apresentem resultados fracos) de modo a obter um resultado final de qualidade superior ao das partes. Dessa técnica fazem parte:

Bagging
Consiste na utilização do mesmo algoritmo (nalguns casos podem mesmo ser

considerados outros algoritmos) treinado várias vezes com diferentes partições dos dados disponíveis para treino. A classificação final é obtida por votação dos diferentes modelos assim obtidos.

Boosting
Vários algoritmos vão sendo iterativamente treinados com o mesmo conjunto de dados. Mas em que em cada iteração são selecionados com maior peso as amostras mal classificadas na iteração anterior. Isto consegue-se aumentando o peso da probabilidade de seleção para iteração seguinte às amostras mal classificados na iteração anterior, e mesmo diminuindo esse peso às bem classificadas. A decisão da classificação final é baseada na classificação de cada algoritmo, tendo menos importância (menos votos) os algoritmos com menor precisão, e mais importância (mais votos) os com maior precisão.

2.4.1 Análise Discriminante Linear e Quadrática

LDA (Análise discriminante linear, na sigla inglesa) e QDA (Análise discriminante quadrática, na sigla inglesa) são algoritmos de classificação bayesianos. Nestes dois algoritmos, na fase de treino são calculados o vetor média e a matriz de covariância em relação à distribuição das várias classes, que depois são usadas nas fórmulas para calcular o designado valor discriminante para cada classe. A diferença entre os dois está no facto de o LDA simplificar a fórmula presumindo uma matriz de covariância global a todas as classes enquanto que o QDA usa uma matriz de covariância para cada classe, o que torna a fórmula do LDA linear e a do QDA quadrática.
Ambos os métodos apresentam melhor desempenho quando as densidades de cada classe são aproximadamente normais e se conseguem boas estimativas com os dados disponíveis para treino. O QDA necessita geralmente de um maior número de amostras para treino do que o LDA [30].

2.4.2 Support Vector Machines

Neste método, proposto por Cortes e Vapnick [31], as diversas caraterísticas de todas as amostras de treino são analisadas num hiperespaço (espaço n-dimensional) e tenta-se encontrar um hiperplano que separe as amostras de cada uma das classes. A pesquisa da dimensão do espaço e do hiperplano que separa as classes constitui um problema de otimização sendo resolvido à custa de uma função *kernel* (na linguagem das SVM) cujo tipo deve ser selecionado de acordo com as

caraterísticas dos dados a classificar.
No treino da SVM poderão ser encontradas várias possibilidades de hiperplanos. O algoritmo seleciona a solução que proporciona maior margem de separação entre as classes. Essa margem é definida pelos designados vetores de suporte, que são as amostras de cada classe mais próximas (no hiperespaço) da outra classe [32].

A função *kernel* é a raiz das SVM, sendo a sua escolha, e a respetiva parametrização, um grande desafio. É crucial para o sucesso da classificação a correta definição do parâmetro de regularização C, que penaliza a violação da margem de separação entre as classes. Esta escolha influencia grandemente o desempenho da SVM, sendo que uma escolha desadequada pode levar à sobre ou sub especialização do algoritmo de classificação.
Quando se utiliza o *kernel* radial (RBF) também o valor do coeficiente γ influencia o desempenho da classificação [33].

As SVM apenas permitem classificação binária (entre duas classes) por definição. Para conseguir a classificação de múltiplas classes o *package* R libsvm utilizado neste trabalho ([34]) recorre à técnica um contra um, na qual são executados diversos sub classificadores binários e, sendo a classe correta encontrada por votação [34].

No estudo [25] foi utilizado o kernel RBF (radial basis function, em inglês), com a otimização dos parâmetros realizada com recurso à validação cruzada 10-Fold (referida na secção 2.2.2). Depois de escolhidos os parâmetros ótimos foi treinado de novo o algoritmo de classificação para construir o modelo final.

As SVM apresentam normalmente melhor desempenho com menor esforço computacional quando comparadas com outros métodos, mas apresentam a desvantagem de aumento exponencial da complexidade do problema com o aumento da quantidade de amostras para treino [35].

2.4.3 Random Forests

Este algoritmo de classificação, RF, é baseado em árvores e foi criado por Leo Breimen [36]. As amostras a classificar são divididas por várias árvores. Em cada árvore é escolhido de forma aleatória um determinado número de caraterísticas a partir das quais se seleciona a melhor partição para esse nó. Cada amostra é colocada nessa floresta de árvores sendo a sua classificação obtida pela "votação" da classificação e cada árvore.

É muito importante para o desempenho que as árvores possuam baixa correlação entre elas e que cada árvore em si seja "resistente" a erros de classificação.

2.5 Arquiteturas de classificação

As arquiteturas de classificação são constituídas pela composição das diversas etapas sucessivas de classificação, cada uma delas classificando diferentes classes cada vez com maior detalhe, com o objetivo de apurar a classificação final. Ou seja, cada etapa de classificação subsequente recebe os resultados da etapa anterior e processa-os para lhes atribuir uma classificação mais detalhada [17].
No caso presente de classificação de sinais de micro-ondas de tumores mamários poderão ser usadas as diferentes etapas de classificação esquematizadas na Figura 2.3, classificando sucessivamente com maior granularidade, as seguintes classes dos tumores:

- Tumor maligno ou benigno (classificação grosseira) e depois se é microlobulado ou espiculado ou se é macrolobulado ou liso (classificação fina)
- Tumor grande ou pequeno (classificação grosseira) e depois se é de raio 10 ou 7,5 mm ou se é de raio 2,5 ou 5 mm (classificação fina)

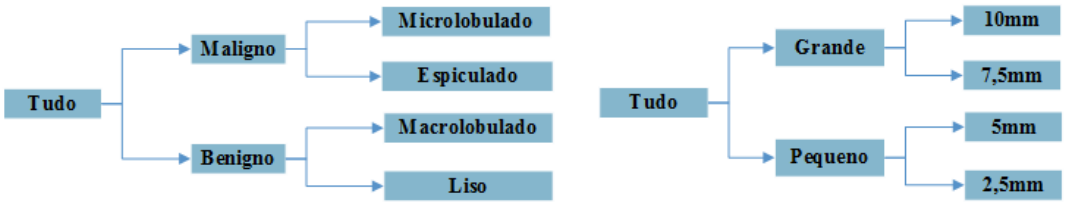


Figura 2.3: Arquiteturas de classificação (adaptado de [17])

2.6 Métricas de classificação

A avaliação dos diversos processos de classificação é realizada através de métricas retiradas da matriz de confusão dos resultados obtidos.
Na matriz de confusão irão constar a quantidade de amostras positivas bem classificadas TP (Verdadeiros positivos, na sigla inglesa), quantidade de amostras negativas bem classificadas TN (Verdadeiros negativos, na sigla inglesa), quantidade de amostras positivas mal classificadas FP (Falsos positivos, na sigla inglesa) e quantidade de amostras negativas mal classificadas FN (Falsos negativos,

na sigla inglesa).
Desses valores podem ser obtidos a quantidade de amostras bem classificadas (rácio de acertos, *Accuracy*, do inglês)(2.4), a Sensibilidade (rácio de verdadeiros positivos, TPR, na sigla inglesa)(2.2) e a Especificidade (rácio de verdadeiros negativos, TNR, na sigla inglesa)(2.3).

$$TPR = \frac{TP}{(TP + FN)} \times 100\% \tag{2.2}$$

$$TNR = \frac{TN}{(TN + FP)} \times 100\% \tag{2.3}$$

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)} \times 100\% \tag{2.4}$$

A fórmula de cálculo do desempenho do algoritmo de classificação depende de qual o erro que é considerado mais grave: atribuir um valor positivo a uma amostra negativa (FP) ou atribuir um valor negativo a uma amostra positiva (FN).
No caso do rastreio do cancro da mama é crítica a classificação de uma amostra positiva (Maligno) em negativa (FN), pelo que deve ser considerado na avaliação da classificação também o rácio de verdadeiros positivos (TPR).

Nos casos em que existem múltiplas classes as fórmulas anteriores podem simplesmente ser aplicadas classe a classe obtendo-se um valor de TPR para cada classe. Para determinação de valores de desempenho globais basta dividir o somatório dos valores de cada classe pelo número de classes.

A curva ROC (*Receiver Operating Characteristic*) permite representar graficamente a variação dos valores de TPR e TNR com diferentes valores de parametrização dos algoritmos, sendo utilizada quando essa variação é linear como forma de encontrar o ponto de equilíbrio entre TPR e TNR.

2.7 Estudos sobre deteção de cancro da mama

A utilização de micro-ondas em aplicações biomédicas está a ser estudada desde há 40 anos, tendo recentemente sido desenvolvidos sistemas de rastreio de cancro da mama para testes de avaliação clínica com recurso a pacientes [37]. Paralelamente têm sido desenvolvidos estudos para classificação usando ML dos sinais

obtidos quer em experimentações envolvendo voluntárias quer obtidos de simulações matemáticas. Nas secções seguintes serão descritos alguns destes estudos envolvendo a classificação de sinais através de ML.

2.7.1 Investigação de classificadores para cancro da mama

Num desses estudos [2], de 2014, foi usado um sistema de radar de micro-ondas no domínio do tempo para avaliar o desempenho de algoritmos de aprendizagem automática (ML) na deteção de existência de cancro da mama. Foram testados os algoritmos LDA e SVM (ver secção 2.4.2), usando PCA para extração das suas caraterísticas principais (e consequente redução da dimensionalidade dos sinais). Foram usados dados experimentais obtidos de artefactos de mama dieletricamente realistas. Os artefactos foram colocados numa cavidade em forma de taça, com 16 antenas de micro-ondas distribuídas pela sua superfície (Figura 2.4). A aquisição dos sinais é realizada ativando uma antena de cada vez, a qual faz propagar uma radiação de micro-ondas através do artefacto, obtendo-se as leituras dos sinais dispersos através da matéria dos artefactos nas restantes 15 antenas. Obteve-se dessa forma um total de 240 sinais biestáticos em cada ensaio.

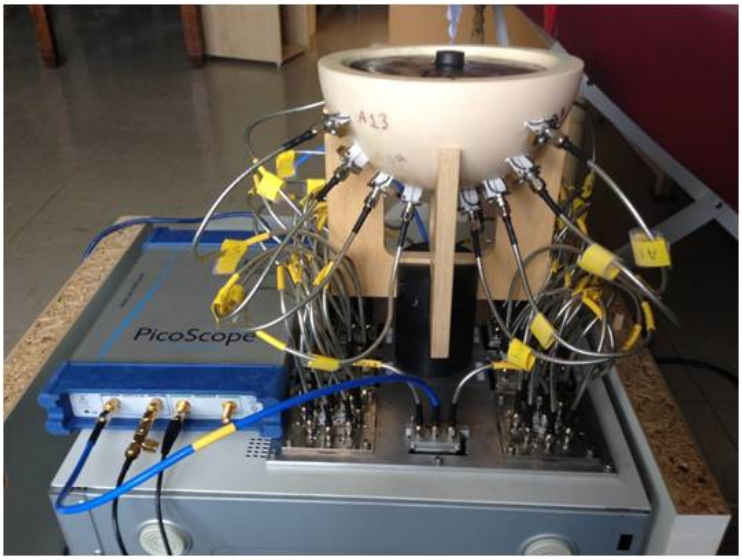


Figura 2.4: Sistema experimental utilizado [2]

Os algoritmos de classificação foram treinados com uma quantidade igual (100) de artefactos com e sem tumor. Os tumores foram colocados em duas localizações diferentes nos artefactos.
Os dados para teste foram obtidos de 30 artefactos totalmente novos, fabricados para o efeito, tentando reproduzir as situações clínicas reais em que o treino será

realizado com os dados de pacientes conhecidos, sendo os resultados obtidos posteriormente aplicados na classificação dos dados de novos pacientes.
A aplicação de PCA reduziu a dimensionalidade do sinal, otimizando o tempo de processamento dos algoritmos de classificação. Para o SVM foi selecionado o *kernel* RBF tendo os valores ótimos para os seus parâmetros C (parâmetro de regularização) e γ (coeficiente do *kernel*) sido obtidos por *cross-validation* (ver secção 2.2.2).
Este estudo concluiu ter sido demonstrada a viabilidade de utilização destes algoritmos de ML para a deteção de cancro da mama, tendo sido observado melhor desempenho do SVM em relação ao LDA.
O desempenho total do SVM foi $73,64\%$ (com $76,71\%$ de sucesso na deteção de existência de tumor e $67,48\%$ de sucesso na deteção de ausência de tumor), contra $70,30\%$ (com $70,45\%$ de sucesso na deteção de existência de tumor e $70,00\%$ de sucesso na deteção de ausência de tumor) do LDA.
Foram também efetuados testes de classificação usando apenas os sinais das antenas mais relevantes, que conforme apurado em trabalho anterior dos mesmos autores, são os grupos de antenas alinhadas com os quatro quadrantes da cavidade onde eram colocados os artefactos. Nesses testes o desempenho de ambos os algoritmos melhorou, mantendo-se o melhor desempenho do SVM, atingindo este $87,63\%$ de sucesso na deteção de existência de tumor (contra $77,40\%$ do LDA) e de $57,00\%$ de sucesso na deteção de ausência de tumor (contra $54,79\%$ do LDA).

2.7.2 Classificação de tumores usando *Deep learning*

No estudo [38], de 2017, foi analisada a utilização de métodos *Deep Learning* na classificação de sinais de radar de micro-ondas obtidos de modelos numéricos de artefactos de tumores imersos em tecido adiposo homogéneo.
A utilização de PCA permitiu reduzir a dimensionalidade dos sinais antes da sua utilização para treino com DNN e com CNN.
Os valores para os parâmetros dos algoritmos de classificação foram otimizados usando *10-Fold Cross Validation* (ver secção 2.2.2), tendo-se utilizados grupos de amostras de treino estratificados de acordo com a sua classe, preservando a percentagem de cada classe (por forma a compensar o desequilibrio do conjunto de dados relativamente à quantidade de amostras de cada classe). Os resultados eram validados no final de cada série de treino, tendo-se verificado que deixou de haver melhorias de desempenho ao final de 300 séries.

Os resultados revelaram um desempenho de $92,81\%$ usando uma DNN de 3 camadas com 300 neurónios nas duas primeiras camadas. Já com CNN foi obtido um desempenho de apenas $89,58\%$ usando 4 camadas convolucionais com 20 filtros cada, um *kernel* de (3×3) e 9 camadas de 300 neurónios completamente ligados.
A utilização de SVM (com $\gamma = 3 \times 10^{-3}$ e $C = 3 \times 10^{-5}$) a seguir a DNN mostrou uma melhoria significativa de desempenho para $93,44\%$.

2.7.3 Classificação de sinais de micro-ondas de modelos de mama

Noutro estudo [39], de 2016, foram usados dois modelos de mama para avaliar o desempenho de dois algoritmos de classificação: SVM (ver secção 2.4.2) e KNN (K vizinhos mais próximos, na sigla inglesa). Um dos modelos (com 1.534 amostras) foi construído com pele e tecido adiposo. No outro modelo (com 2.395 amostras), além de pele e tecido adiposo foi colocado também tecido fibroglandular. Cada amostra podia conter um tumor com o raio variando de 2 a 10 mm.
Com SVM os resultados foram superiores tendo tido um sucesso 80% na deteção de tecido tumoral com o conjunto de dados sem tecido fibroglandular e apenas 62% no outro.
Os parâmetros SVM utilizados foram o par: $(C = 65,79; \gamma = 4,32 \times 10^3)$, com um *kernel* RBF escolhido de forma totalmente empírica. Aqueles valores foram encontrados após vários testes usando o método *10-Fold Cross Validation* (ver secção 2.2.2), fazendo simultaneamente variar os valores dos parâmetros entre $(1 < C < 10^4; 10^{-4} < \gamma < 0,1)$.

2.7.4 Deteção de cancro da mama usando EMD

No estudo [40], de 2017, também foi usado o algoritmo SVM (ver secção 2.4.2), mas numa versão sensível ao custo (*cost-sensitive SVM* - 2v-SVM) para classificar a existência de cancro da mama. O principal objetivo do estudo era avaliar qual o melhor método para extração de caraterísticas entre EMD (Modo de decomposição empírico, na sigla inglesa) e PCA.
EMD é uma técnica simples que decompõe o sinal (estacionário) num pequeno número de IMF (Funções de modo intrínseco, na sigla inglesa). A motivação para utilizar EMD adveio do facto de este método de extração ser potencialmente mais

robusto a vibrações do sistema de aquisição devidas, por exemplo, ao ato de respirar das voluntárias.
O sinais foram obtidos através um sistema multi-estático com 16 antenas, onde uma antena de cada vez emite um sinal UWB enquanto as restantes antenas captavam os sinais dispersos através dos tecidos do seio, perfazendo um total de 240 valores registados por ensaio. Dado que foram usadas voluntárias saudáveis, num total de 12, foi posteriormente simulada a existência de tumores em diversas localizações dos dois seios de cada voluntária. Cada uma das 12 voluntárias foi submetida a vários testes ao longo de 8 meses, tendo sido colecionados 96 amostras no total.
Para avaliar o desempenho dos diferentes modos de extração de caraterísticas foram organizados 50 conjuntos de dados, cada um dividido entre treino e teste, garantido-se o equilíbrio entre as amostras com e sem tumor. Para o algoritmo $2v$-SVM foram encontrados os parâmetros ótimos através de *k-Fold Cross Validation* (ver secção 2.2.2) com os valores a variar entre $10^{-5} < v < 1$.
O estudo constatou que os resultados obtidos quando se usou EMD foram superiores aos quando se usou PCA, tendo-se também observado uma pequena melhoria com a conjunção dos dois métodos.

2.8 Estudos sobre deteção de AVC (acidente vascular cerebral)

Existem vários trabalhos realizados sobre este tema que demonstram o sucesso da classificação de sinais de micro-ondas, neste caso em exames ao cérebro.

2.8.1 Aprendizagem automática em diagnóstico de AVC

Em [41], de 2014, são referenciadas várias técnicas de ML para classificação de sinais de micro-ondas com o objetivo de distinguir o tipo de AVC ocorrido (hemorrágico ou isquémico). Neste estudo são alvo de análise os algoritmos SVM (ver secção 2.4.2) e LMNN (Vizinho mais próximo com maior margem, na sigla inglesa), uma derivação do algoritmo KNN. Sendo também reconhecidas as dificuldades encontradas com a reduzida dimensão do conjunto de dados face à quantidade de caraterísticas dos sinais.
No caso da SVM foram utilizados dois *kernel*, o linear e o radial (RBF). Os parâmetros ideais para estes dois kernel foram encontrados usando o algoritmo de

afinação interno ao *package* utilizado (LIBSVM para MatLab). Os resultados deste estudo mostraram que com LMNN não se obtiveram resultados convincentes, e mesmo com SVM o desempenho em todos os testes rondou apenas os 50% de acertos. Na análise crítica a estes resultados julgou-se que o modo de determinar dos parâmetros SVM não foi o mais adequado devido ao desequilíbrio do conjunto de dados em relação às duas classes que se pretendia encontrar.

2.8.2 Localização de AVC usando classificação de sinais de micro-ondas

Em [3], de 2016, os autores utilizaram SVM (ver secção 2.4.2) para o componente principal de um mecanismo de localização da zona afetada do cérebro. Nesse mecanismo são utilizados conjuntos de vários canais (cada canal é formado por uma antena emissora e uma antena recetora) colocados ortogonalmente. Dessa forma a localização do AVC é determinada pelo local da interseção dos feixes dos canais onde foi detetado o AVC (2.5). O algoritmo SVM foi usado para determinar os canais que detetaram AVC.

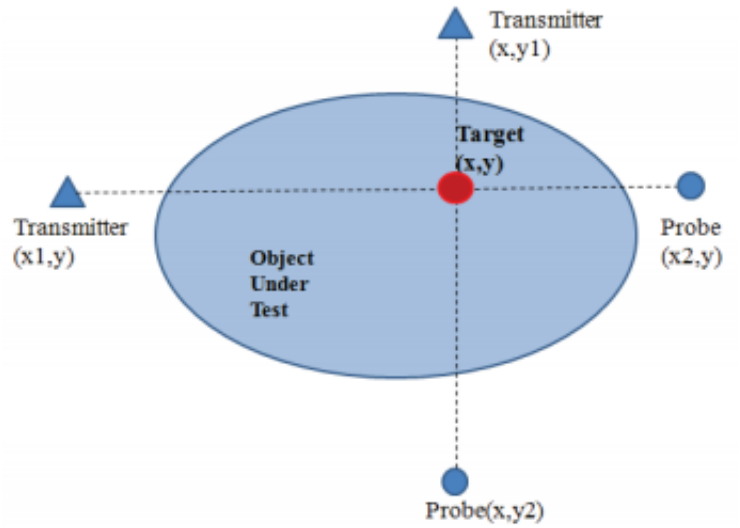


Figura 2.5: Localização do AVC usando interseção de dois canais [3]

Este método foi selecionado pelo seu bom desempenho e por não necessitar de um número elevado de amostras para treino. Os parâmetros ótimos para usar no modelo SVM foram encontrados através do método *k-Fold Cross Validation* (com k = 5) (ver secção 2.2.2). Para treinar e testar o algoritmo de classificação foram

usados cinco conjuntos de dados com diferentes tipos de interferências no sinal e diferentes quantidades de amostras positivas e negativas. Os resultados obtidos atingiram os 80% de sucesso na classificação garantido assim o sucesso na localização do AVC. Também permitiram demonstrar que o método é robusto relativamente à existência de ruído branco nos sinais.

2.8.3 Técnicas de processamento e imagem para rastreio baseado em micro-ondas

Também o estudo [42], de 2017, incide sobre a classificação do tipo de AVC, sendo neste caso todas as amostras do conjunto de dados oriundas de doentes com AVC. Para isso foi usado o algoritmo SVM (ver secção 2.4.2) aplicado à vetorização da imagem do cérebro, obtida a partir de modelos numéricos. Os parâmetros ótimos para o kernel RBF foram determinados através de *k-Fold Cross Validation* (ver secção 2.2.2).
Os resultados deste estudo apontam para um elevado nível de sucesso na classificação (92%), embora na presença de ruído a percentagem baixe para 81%.

3

Processo de classificação

Neste capítulo descreve-se o trabalho prático desenvolvido para a concretização dos objetivos citados na secção 1.2. São descritos detalhadamente todos os ensaios realizados, incluindo todo o processamento de preparação e divisão dos dados, extração de caraterísticas e classificação.
Durante este trabalho foram avaliadas as diversas combinações de métodos preparação e divisão de dados, com técnicas de extração de caraterísticas e com algoritmos de classificação. Foi desenvolvida uma plataforma aplicacional de ensaios que possibilitou essas avaliações, e que poderá servir para outras avaliações futuramente se pretendam realizar no mesmo âmbito. O código foi desenvolvido na linguagem R [1] utilizando a ferramenta RStudio [2] (quer a linguagem, quer a ferramenta são *open source*, ou seja, de utilização gratuita).
No decurso do trabalho efetuado foram surgindo algumas dúvidas e tomadas opções com base nos resultados que se iam encontrando, tendo nalguns casos sido encontradas soluções alternativas às consideradas inicialmente. Foi também necessário limitar a quantidade de ensaios de modo a não tornar demasiado longa esta dissertação e a análise dos resultados obtidos.

[1] https://www.r-project.org/
[2] https://www.rstudio.com/

3.1 Resumo dos ensaios efetuados

Durante a realização deste trabalho prático foram efetuados vários testes com várias variantes de processos de classificação dos sinais de micro-ondas que serviram de base de para este estudo. Com base nos resultados foram selecionados os ensaios de classificação esquematizados no quadro da Figura 3.1. De forma idêntica foram determinadas as parametrizações a utilizar em cada uma das técnicas de extração e algoritmos de classificação.

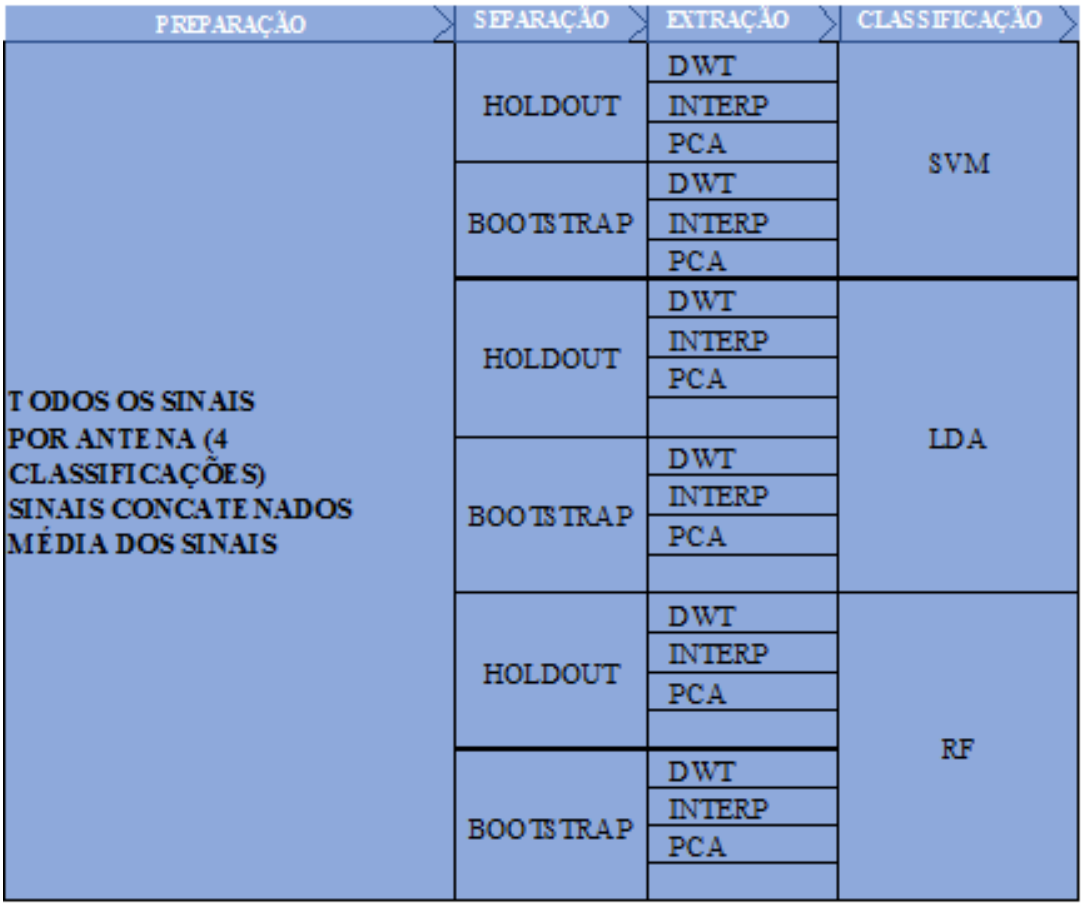


Figura 3.1: Quadro de ensaios efetuados

Conforme é possível inferir desse quadro foram utilizados:

- Quatro modos de preparação dos dados;
- Duas técnicas de divisão dos dados;
- Três métodos de extração de caraterísticas, com diversas parametrizações (13 variantes);

- Três algoritmos de classificação (dois deles foram experimentados sem extração prévia de caraterísticas).

Todos os ensaios incidiram na classificação de duas classes: Maligno e Grande. São decerto essas as classes mais importantes no rastreio do cancro da mama, pois são as que fornecem aos clínicos informação sobre a malignidade e tamanho do tumor. Seria decerto também importante para esses clínicos obter informação mais detalhada, como por exemplo a quantidade de espículos. Mas, conforme referido atrás, foi necessário limitar a quantidade de ensaios. Pela mesma razão optou-se por não se utilizar arquiteturas de classificação (referidas na secção 2.5), também porque os resultados não seriam significantes atendendo à quantidade reduzida de amostras disponíveis no conjunto de dados.
Ainda assim obtiveram-se um total de 656 resultados de desempenho da classificação (resultantes de 328 ensaios para cada uma das duas classes) que irão ser analisados nos capítulos seguintes, procurando-se determinar quais as melhores soluções de classificação para este tipo de sinais e quais os fatores que podem influenciar o sucesso dessa classificação.

3.2 Descrição dos dados

Os dados a utilizar na presente dissertação foram obtidos de simulações (modelos matemáticos, *numeric phantoms*, no inglês) de 10 diferentes modelos de tumores que foram desenvolvidos especificamente para este tipo de estudos [17]. Cada modelo inclui 4 tamanhos e 4 formas diferentes de tumores, tendo uma das formas (espiculada) 3 variantes. Perfazendo, portanto, um total de 240 simulações de tumores.
Os raios dos tumores são de $2,5$ mm, 5 mm, $7,5$ mm e 10 mm. Os tipos de tumor são liso, macrolobulado, microlobulado e espiculado (Figura 3.2), podendo neste último caso apresentarem 3, 5 ou 10 espículos.
Os tumores microlobulados e espiculados são malignos e os macrolobulados e lisos são benignos.

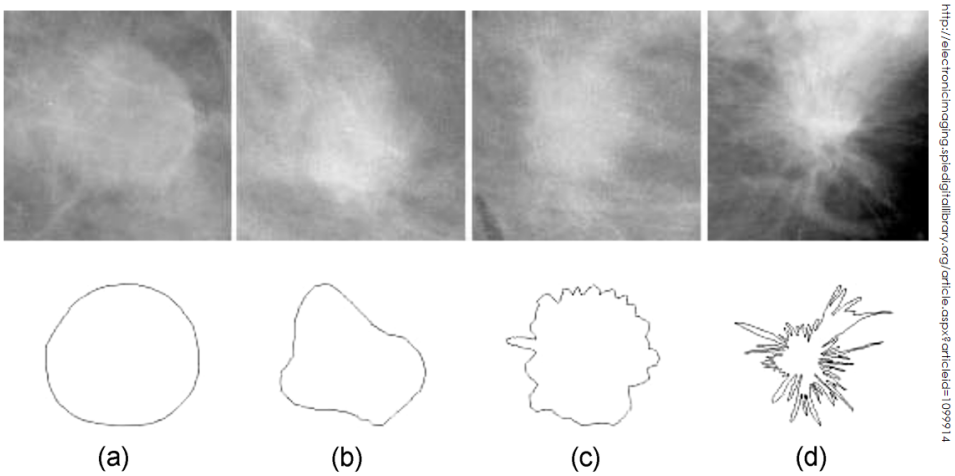


Figura 3.2: Tipos de tumores [4]
(a) liso; (b)macrolobulado; (c)microlobulado; (d)espiculado

Os sinais foram recolhidos em quatro antenas de micro-ondas colocadas em quatros ângulos ao redor do modelo (com antenas colocadas de 90 em 90 graus: 0^o, 90^o, 180^o e 270^o), obtendo-se assim 4 sinais em cada simulação (96 sinais por cada um dos 10 modelos).
A Figura 3.4, que se refere a um protótipo com 16 antenas colocadas à volta de um tumor (*phantom*), é apresentada como exemplo visual da simulação matemática efetuada para geração destes sinais. É claro que, no caso presente, só existiriam 4 antenas colocadas nos ângulos atrás referidos.
Os sinais estão no domínio temporal e apresentam 4.000 valores de amplitude por antena.

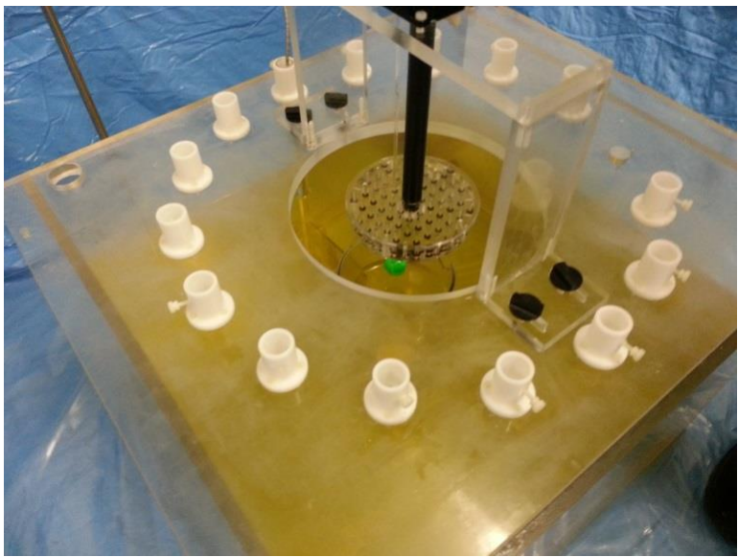


Figura 3.4: Protótipo com várias antenas [5]

Caraterísticas do conjunto de dados

O conjunto de dados para esta dissertação (cujas caraterísticas se apresentam na tabela 3.1) foi fornecido em 240 ficheiros MAT[3], contendo cada ficheiro uma variável *signalOutMinusNoTumorMonostatic* com os sinais calibrados recolhidos nas quatro antenas.

Tabela 3.1: Caraterísticas do conjunto de dados

Quantidade de instancias	240	**Caraterística**	Multivariado
Quantidade de dimensões	4.000	**Tipo de dimensões**	Real
Quantidade de atributos	4	**Tipo de atributos**	Discreto

As dimensões a usar na classificação são os 4.000 valores do sinal captado pelas antenas ao longo do tempo.
Os 4 atributos que caraterizam o tumor figuram no nome do respetivo ficheiro MAT e são os seguintes:

- O número do modelo: valor varia entre 1 a 10
- O tipo do tumor: valor varia entre 1 e 4

  1 - tipo espiculado

  2 - tipo microlobulado

  3 - tipo macrolobulado

  4 - tipo liso
- O raio (tamanho) do tumor: com os valores possíveis de 25, 50, 75 e 100

  - correspondentes a 2, 5 mm, 5 mm, 7, 5 mm e 10 mm de raio)
- O número de espículos

  - No tipo de tumor 1 toma os valores de 3, 5 ou 10 consoante o número de espículos

  - Sendo igual a 0 (sem espículos) para tipos de tumor 2, 3 e 4.

Como exemplo apresenta-se no gráfico da Figura 3.5 o sinal captado por uma antena, onde se pode observar a evolução da amplitude ao longo dos 4.000 instantes de tempo.

[3]a extensão MAT está associada a MatLab MAT-File

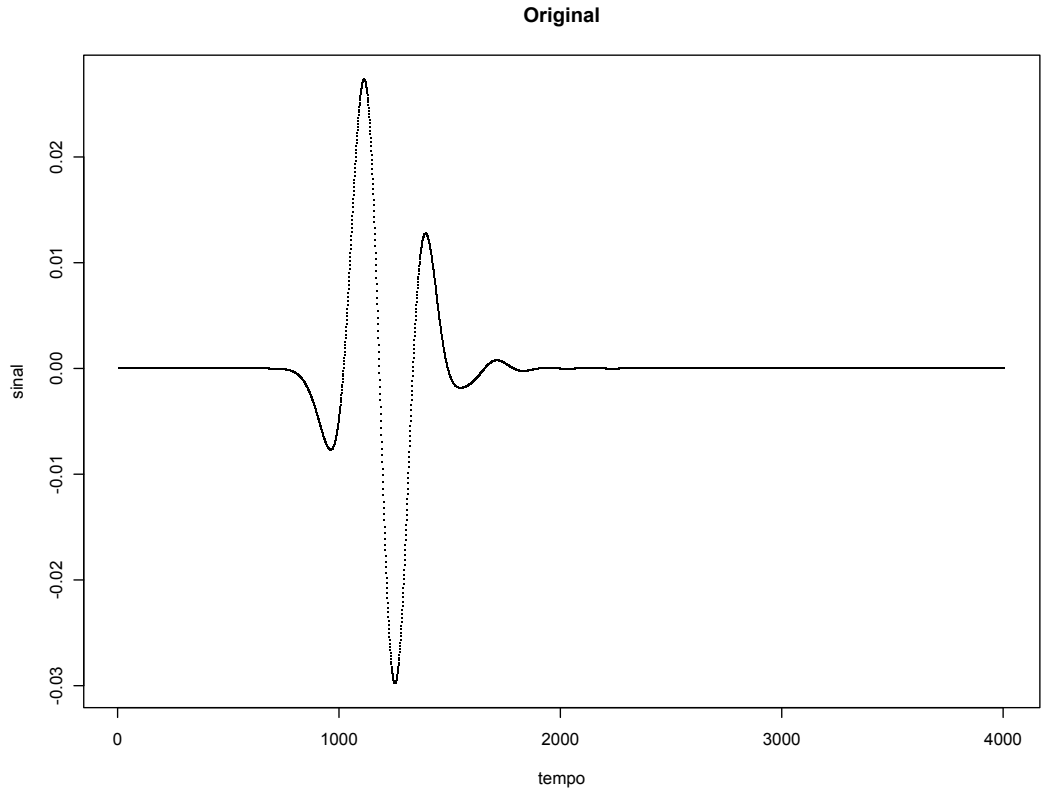


Figura 3.5: Sinal de uma antena

3.3 Preparação dos dados

Nas secções seguintes são descritas as transformações realizadas aos sinais de micro-ondas previamente à utilização das técnicas de extração de caraterísticas e algoritmos de classificação.
As transformações visaram não só a adição de caraterísticas, que irão facilitar e tornar mais direto o processo de classificação, mas também testar as várias possibilidades de conjugação e agrupamento dos sinais das quatro antenas para verificar quais apresentam melhores resultados na classificação.

Importação do conjunto de dados

Os ficheiros Matlab foram importados para uma estrutura de dados do tipo matriz (*data frame* do R), contendo tanto os atributos referidos na secção 3.2 como o respetivo sinal nas suas linhas. Foram além disso adicionados outros atributos, de tipo booleano, calculados a partir dos primeiros, tendo em vista a classificação binária dos sinais:

- **Maligno** - Se o tumor é do tipo microlobulado ou espiculado, ou não (tipo liso ou macrolobulado);
- **Grande** - Se o tumor é de raio $7,5$ ou 10 mm, ou não (raio $2,5$ ou 5 mm);
- De assinalar que a classificação binária (1 ou 0, equivalente a Verdadeiro ou Falso) apresenta grandes vantagens para os algoritmos de classificação em geral, dada a simplificação da decisão (ou pertence a essa classe ou não pertence), face à decisão entre várias classes.

De seguida essa matriz de dados foi separada em duas: uma apenas com os atributos e outra com os sinais de micro-ondas, estando ambas logicamente relacionadas através do número da amostra. Dessa forma todo o processamento de classificação incidirá apenas sobre a matriz com os sinais, adicionando a informação dos atributos a partir da segunda matriz sempre que necessário.

Simplificação dos sinais

Pela análise visual dos gráficos dos sinais (Figura 3.5) é possível constatar que praticamente não existe variação entre eles nos seus extremos inicial e final.
Perante esta constatação achou-se por bem realizar a simplificação dos sinais, omitindo as partes iniciais e finais dos mesmos, com o objetivo de otimizar os tempos de processamento e o desempenho da classificação.
Tendo em vista o objetivo da classificação dos sinais, que se baseia na distinção das caraterísticas dos vários sinais, optou-se pela sua simplificação através da remoção das suas posições iniciais e finais consecutivas com valores iguais (ou bastante próximos) entre todos os sinais, ou seja, cuja variância no conjunto de todas as amostras se aproximasse de zero. Dessa forma não são perdidas as características distintivas dos sinais das diferentes amostras, mantendo-se assim a sua integridade na perspetiva da classificação.
Na Figura 3.6 apresenta-se o sinal obtido após a remoção dos seus pontos iniciais e finais cuja variância é inferior a uma centésima de milésima (10^{-5}). Este valor foi o que revelou o melhor compromisso entre a redução do número de pontos do sinal e o desempenho da classificação.

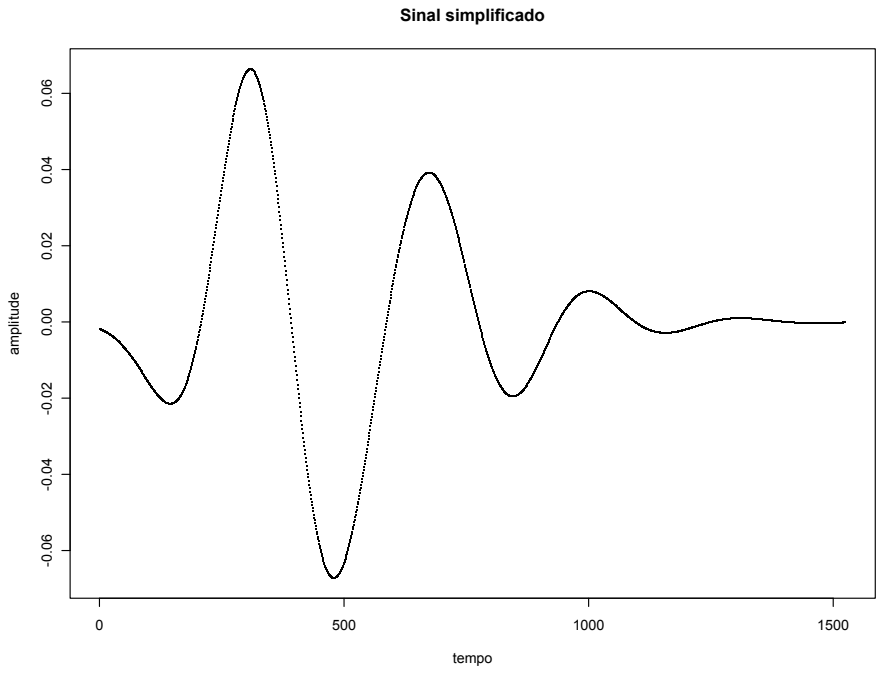


Figura 3.6: Sinal truncado nos pontos iniciais e finais com variância inferior a 10^{-5}

Foram estes sinais simplificados, sem os pontos iniciais até à abcissa 796 e sem os pontos finais a partir da abcissa 2.322, os utilizados em todos os ensaios descritos neste relatório. Ou seja, dos 4.000 pontos iniciais foram considerados apenas os

1.525 pontos da parte central do sinal, significando uma redução para menos de 40% das caraterísticas que servirão de base ao processo de classificação.

Conjugação dos sinais

O processo de classificação poderá ser realizado considerando todos os sinais das quatro antenas como um todo ou classificando-os em separado. Podem ainda ser utilizadas várias opções de conjugação dos sinais das quatro antenas de cada amostra.

Considerar as quatro antenas como um todo apresenta como vantagem a maior quantidade de amostras disponíveis para treino e teste dos processos de classificação. O que já não acontece quando conjugamos os sinais das 4 antenas correspondentes a uma amostra, situação em que ficamos com 240 amostras, em vez de 960.

Na classificação dos sinais das quatro antenas em separado (identificando a posição da antena) são realizados quatro processos de classificação, sendo o resultado final da classificação obtido por votação. Esta forma de classificação permite apresentar também um grau de confiança na classificação: por exemplo, caso os sinais das quatro antenas da mesma amostra sejam classificados na mesma classe poderemos afirmar que a confiança na classificação é de 100%.

- A dificuldade coloca-se quando existe um empate na classificação, já que, no nosso caso as antenas são em número par. A opção foi assumir a classificação com valorização mais negativa (maligno, por exemplo) de modo a não aumentar o número de falsos negativos.

Resumo da preparação dos dados

Assim, após esta fase de preparação dos dados, nas condições referidas anteriormente, obtiveram-se:

1. Os sinais simplificados, omitindo-lhes a parte inicial e final;
2. Os sinais com mais dois atributos: Maligno (ou não) e Grande (ou não);
3. Quatro conjuntos de sinais a classificar:
    - O conjunto de todos os sinais das 4 antenas (960 amostras);
    - Quatro conjuntos com os sinais de cada uma das 4 antenas (4×240 amostras);
    - O conjunto com os sinais das 4 antenas concatenados (240 amostras);
    - O conjunto com a média dos sinais das 4 antenas (240 amostras).

3.4 Divisão dos dados

Nesta secção são apresentadas as técnicas que foram usadas na divisão do conjunto de dados em duas partes (ver secção 2.2):

- Uma parte para ser utilizada no treino da classificação;
- E outra parte para ser usada em testes, como forma de validação do modelo de classificação.

Holdout

Seguindo a técnica de *Holdout* (ver secção 2.2.1) a divisão dos dados contempla dois conjuntos disjuntos:

- Um com 2/3 das amostras para treino;
- E outro com o restante 1/3 para teste.
- Nota: não foi considerado um conjunto para validação atendendo à quantidade reduzida de dados disponíveis (tal como referido na secção 2.2 o conjunto de validação é frequentemente dispensado).

Constatando a existência de uniformidade entre as amostras dos 10 modelos existentes no conjunto de dados disponibilizado para este trabalho (cada modelo abrange todo o universo de possibilidades de tumores), considerou-se uma boa prática fazer a divisão por modelo de tumor, utilizando 7 modelos para treino (cerca de 2/3 de 10) e os restantes 3 para teste. A divisão dos modelos foi feita de forma aleatória.
Desta forma manteve-se a proporcionalidade entre os diferentes tipos de tumores, já que cada modelo possui a mesma quantidade de tipos e dimensões de tumores, e também a disjunção dos dois conjuntos.
Este modo de divisão foi utilizado quer quando se considerou indiferentemente o conjunto de todos os sinais (960 amostras), quer quando se conjugaram ou se consideraram separadamente os sinais das quatro antenas (240 amostras).

Bootstrap

Na técnica de *Bootstrap* (ver secção 2.2.3) foi repetida por 10 vezes a seleção aleatória de um dos 10 modelos de simulação (permitindo-se a seleção de modelos repetidos). Foram assim obtidos os 10 modelos que foram usados para treino dos

algoritmos de classificação. Dos restantes modelos não selecionados foram selecionados 3 para testes, correspondentes a 30% do conjunto de dados. A opção por fixar em 3 a quantidade de modelos para teste prende-se com a necessidade de comparar os resultados com a técnica de *Holdout*, onde são utilizados também 3 modelos para testes.

3.5 Extração de caraterísticas

Cada um dos diferentes métodos de extração de caraterísticas transforma os sinais em diferentes formatos de resultados, tendo os mesmos influência direta no desempenho do algoritmo de classificação automática. Foram assim experimentados os seguintes métodos de extração conjugados com os vários algoritmos de classificação na tentativa de encontrar a conjugação com melhores desempenhos.

Extração de caraterísticas com DWT

A extração de caraterísticas através de DWT (ver secção 2.3.1) foi concretizada usando o *package* R Wavelets ([43]), sendo selecionada a sua parametrização a partir da parametrização que apresentou melhores resultados na tese de Conceição [17]), ou seja, com o filtro Coiflet 6.
Foram utilizadas quatro diferentes parametrizações onde se variaram apenas os filtros, Coiflet e Daubechies, com diferentes comprimentos 6, 8 e 12, mantendo-se os restantes parâmetros constantes: com 5 níveis de decomposição e o tipo de sinal periódico.

- Esta seleção de quatro parametrizações foi obtida de forma empírica e por experimentação prévia de diversas outras parametrizações.

Extração de caraterísticas com PCA

Conforme referido na secção 2.3.3 a análise de componentes principais, PCA, procura encontrar correlações entre as várias caraterísticas (componentes) do sinal original como forma de redução da quantidade de caraterísticas a utilizar na classificação. Isso é conseguido através da análise prévia das caraterísticas de todo o conjunto de dados (de treino e, posteriormente de teste), com o objetivo de conservar apenas os componentes com maior influência na variância das suas caraterísticas [26].
Neste caso foram experimentadas seis diferentes parametrizações obtidas de forma empírica e por experimentação de diversas outras parametrizações. Neste caso os valores de tolerância variaram entre 10^{-6} $(0,000001\%)$ e 10^{-1} $(0,1\%)$ e o número

máximo de componentes principais (*rank*) entre 20 e 50.
Com estas parametrizações a quantidade de caraterísticas resultantes do processo de extração foi reduzida para valores entre as 6 e as 51.

- Relembra-se que os sinais depois de removidos os valores iniciais e finais com variância próxima do zero ficaram com 1.525 caraterísticas, ou 6.100 quando se concatenaram os sinais das 4 antenas).

Extração de caraterísticas com *Wavelets* interpolatórias

Para a obtenção de *Wavelets* interpolatórias dos sinais de micro-ondas, foi realizada a adaptação do código em linguagem C desenvolvido no âmbito da tese de doutoramento de Pinho [1] para sua utilização no ambiente R. O código foi adaptado e integrado no R através da função dyn.load (*Foreign Function Interface*).
Foi utilizada a função *spr_data_to_sparse* (3.1) que executa a interpolação, por níveis, de um dado vetor *u*.

```
void spr_data_to_sparse (double* u, int u_size, int* ell_p, int* n0_p, int* interp_
    points_p, double* eps_p)
{
 double *u_new;
 int len = u_size;
 len--;
 len |= len >> 1;
 len |= len >> 2;
 len |= len >> 4;
 len |= len >> 8;
 len |= len >> 16;
 len++;
 u\_new = (double *) malloc(len*sizeof(double));
 memcpy(u\_new, u, u\_size*sizeof(double));
 int ii;
 for(ii=u\_size; ii<len; ii++) {
   u\_new[ii] = 0.0;
 }
 int ell; int n0; int interp\_points; double eps;
 ell = *ell\_p; n0 = *n0\_p; interp\_points = *interp\_points_p; eps = *eps\_p;
 int n = (n0-1) * (1<<ell) + 1;
 int spc = 1;
 int i, j;

 for (j=ell; j >= 1; j--)
 {
  spc <<= 1;
  for (i=spc/2; i < n; i+=spc)
  {
   if (!\_isnan(u_new[i]))
   {
    if (fabs(u\_new[i] - interp2(u\_new, ell, n0, interp\_points, i, spc)) < eps)
    {
```

```
33          u\_new[i] = NAN;
          }
35      }
      }
37  }
  memcpy(u, u\_new, u\_size*sizeof(double));
39 }
```

Listagem 3.1: Função *spr_data_to_sparse*

A função coloca a NAN os pontos que não são necessários para reconstruir a função, ou seja, os pontos em que o valor calculado por interpolação dos pontos vizinhos é muito próximo (ou igual) do valor verdadeiro (a menos de um erro *eps*). *ell* representa o número de níveis de decomposição, $n0$ a quantidade de amostras no nível grosseiro e *interp_points* o tipo de interpolação (2 para interpolação binária, ou 4 para interpolação cúbica).
As três parametrizações utilizadas nos testes efetuados foram obtidas de forma empírica após a experimentação de várias possibilidades de conjunção de valores. Assim, foram utilizados para o número de níveis de decomposição os valores 4 e 5 e para valor do erro 10^{-4} e 2×10^{-4}. O valor para a quantidade de amostras no nível grosseiro foi calculado através da fórmula (3.1).

$$n0 = (sizeof(u) - 1)/(2^{ell}) + 1 \tag{3.1}$$

Os resultados obtidos com as parametrizações utilizadas mostram que se obtém melhores resultados com menor valor de ϵ e com 5 níveis de decomposição, quer para a classificação de Maligno quer para a classificação de Grande (tabela 3.2).

Tabela 3.2: Resultados das *Wavelets* interpolatórias com diferentes parametrizações

PREPARAÇÃO		ENSAIO			Wavelet Interpolatória			RESULT. %	
dataset	separa.	extração	classif.	classe	nível	interp.	epsilon	Acc	TPR
Média	Holdout	Interp	RF	Maligno	5	4	0,0001	86,11	93,75
Média	Holdout	Interp	RF	Maligno	4	4	0,0002	86,11	93,75
Média	Holdout	Interp	RF	Maligno	5	4	0,0002	85,42	92,19
Média	Holdout	Interp	RF	Maligno	4	4	0,0001	84,72	97,92
Média	Holdout	Interp	SVM	Grande	5	4	0,0001	94,44	94,44
Média	Holdout	Interp	SVM	Grande	4	4	0,0002	94,44	91,67
Média	Bootstrap	Interp	SVM	Grande	5	4	0,0002	94,44	97,22
Média	Bootstrap	Interp	SVM	Grande	4	4	0,0001	93,06	94,44

A aplicação de *Wavelets* interpolatórias aos sinais revelou-se bastante eficaz ao nível da simplificação do sinal, conforme se pode observar comparando as duas figuras, Figura 3.7 e Figura 3.8, que representam o mesmo sinal respetivamente antes da interpolação e após a interpolação.

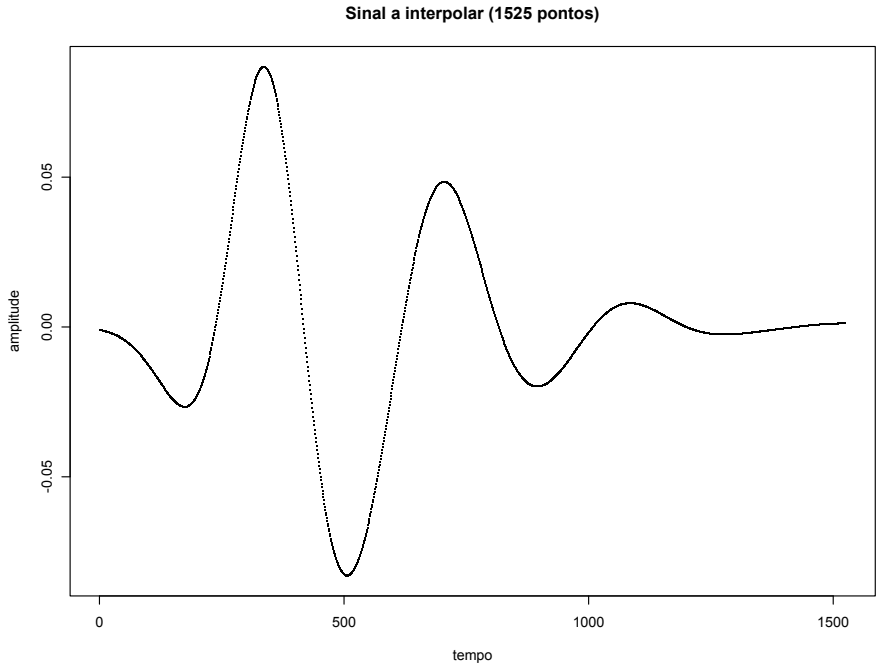


Figura 3.7: Sinal original - a interpolar

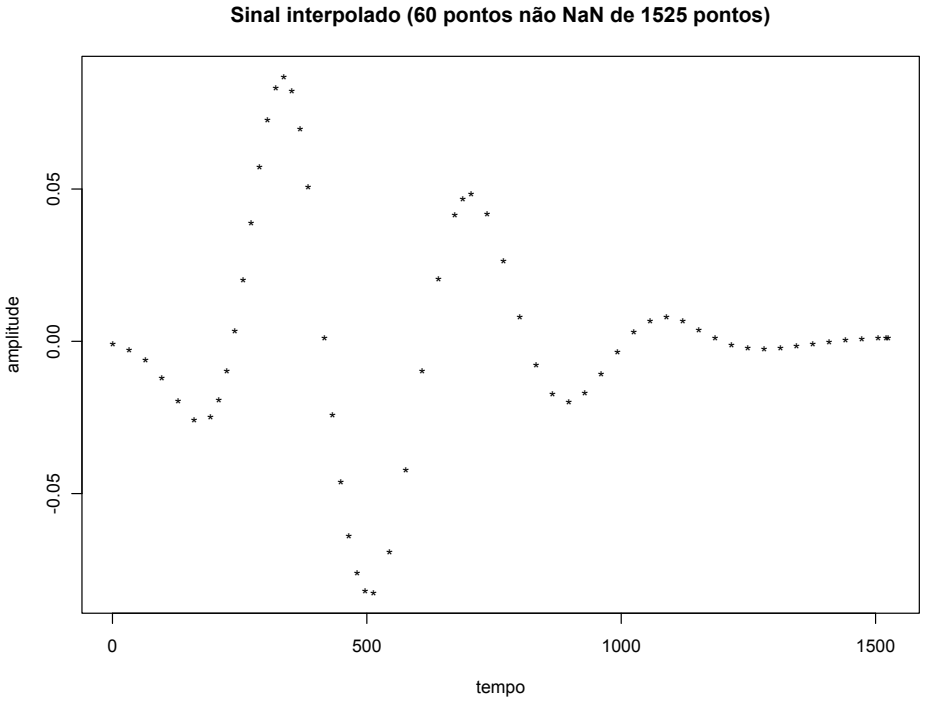


Figura 3.8: Sinal interpolado

Verifica-se neste exemplo uma simplificação de 96%, ou seja, apenas 60 pontos (de 1.525) são necessários para representar este sinal. Todos os restantes pontos

podem ser obtidos por interpolação destes 60.
Efetivamente este método permite uma grande simplificação dos sinais, mas o seu resultado não pode ser diretamente aplicado aos algoritmos de classificação. Atendendo a que essa simplificação é conseguida pela "eliminação" de pontos do sinal (os que conseguem ser obtidos pela interpolação dos seus vizinhos), nem todos os sinais, na sua versão interpolada, irão ter valores significativos nas mesmas abcissas. O que impede a utilização dos sinais assim representados no treino dos algoritmos de classificação.
Foi, portanto, necessário encontrar uma forma de uniformizar todos os sinais por forma a que todos apresentem valores nas mesmas abcissas (ou seja, todos tenham valores nas mesmas caraterísticas). Com esse objetivo determinou-se o conjunto de abcissas que tinha valores significativos em pelo menos um dos sinais (de ambos os conjuntos de treino e de teste). "Completaram-se" depois todos os sinais interpolados de modo a todos ficarem com valores significativos nesse conjunto de abcissas, colocando-lhe o valor original do sinal nas abcissas que tinham sido "eliminadas". Foram depois removidas as abcissas que não tinham valor significativo em nenhum dos sinais obtendo-se um sinal interpolado uniformizado conforme representado na Figura 3.9 (no caso deste conjunto de dados foram apenas 79 abcissas tinham valores significativos em pelo menos um dos sinais).

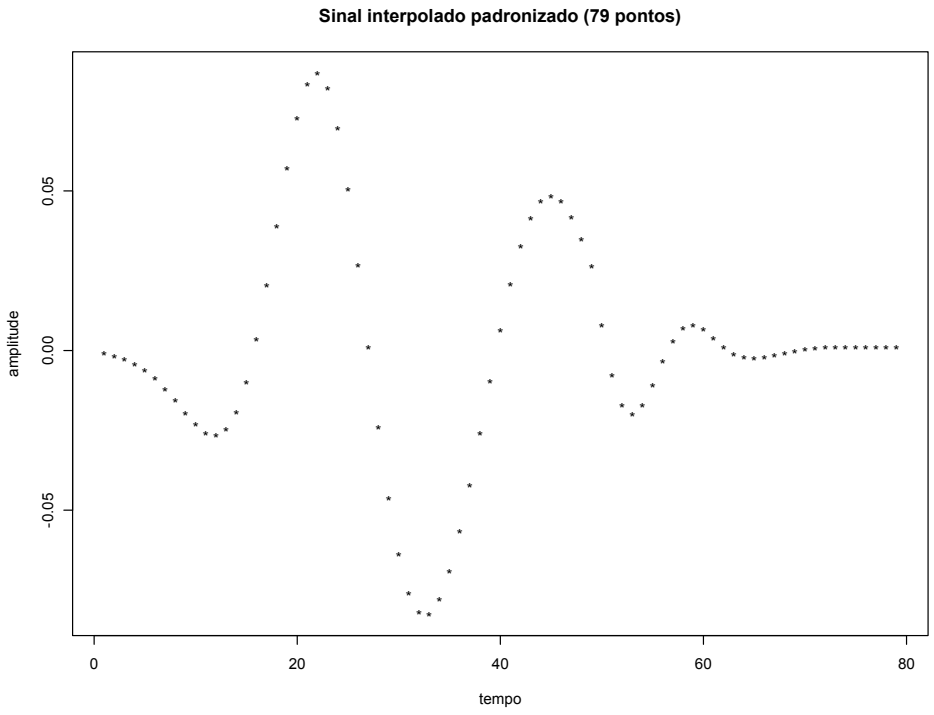


Figura 3.9: Sinal interpolado - uniformizado

3.6 Classificação

Neste parágrafo descrevem-se os algoritmos de classificação que foram utilizados neste trabalho e a forma como foram utilizados, apresentando-se também os resultados obtidos com esses algoritmos.

SVM

O algoritmo de classificação *Support Vector Machines* (SVM) referido na secção 2.4.2 foi utilizado após a aplicação das diferentes técnicas de extração de características mencionadas na secção 3.5 aos conjuntos de dados obtidos pelos diferentes métodos de preparação e divisão de dados referidos nas secções 3.3 e 3.4.
O algoritmo SVM foi usado para classificar as amostras nas classes Maligno e Grande, tendo sido otimizados os respetivos parâmetros utilizando o algoritmo de afinação interno ao *package* utilizado [34]. Os valores iniciais de parametrização foram escolhidos de forma aleatória, após várias tentativas, e foram os seguintes:

- γ entre 10^{-4} e 10 e custo entre 10 e 100.000, para classificação na classe Maligno;
- γ igual a 10^{-6} e custo entre 10 e 100.000, para classificação na classe Grande;
- A seleção do kernel foi fixada em RBF (radial).

LDA

O algoritmo *Linear Discriminant Analisys* (LDA) referido na secção 2.3.3 foi utilizado apenas para classificação, embora se possa aplicar também para extração de caraterísticas. Tal como os restantes foi utilizado para classificar os sinais nas classes Maligno e Grande após a extração de caraterísticas e também diretamente (sem extração prévia de caraterísticas).
Para a aplicação da função *lda* no R [44] foi necessário encontrar o valor para o parâmetro tolerância que permitisse a distinção entre as classes a classificar.

Random Forests

O algoritmo de classificação RF referido na secção 2.4.3 foi utilizado após a aplicação das diferentes técnicas de extração de caraterísticas mencionadas na secção 3.5 aos dados obtidos pelos diferentes métodos de preparação e divisão de dados referidos nas secções 3.3 e 3.4.
O método RF foi usado para classificar as amostras nas classes Maligno e Grande,

com a parametrização de omissão (número de árvores igual a 500) do *package* R utilizado ([45]).

4

Análise de resultados

Neste capítulo são apresentados os resultados obtidos no trabalho prático desenvolvido descrito no capítulo 3. Conforme descrito nesse capítulo foram realizados os ensaios de classificação (de duas classes: Maligno e Grande) esquematizados no quadro da Figura 3.1 tendo sido obtidos 656 resultados de desempenho da classificação (resultantes de 328 ensaios para cada uma das duas classes) que irão ser analisados detalhadamente no presente capítulo, procurando-se avaliar quais as melhores soluções de classificação para este tipo de sinais e quais os fatores que podem influenciar o sucesso dessa classificação. Essa avaliação é feita com base na percentagem de acertos (*accuracy*)(2.4), sendo apresentado também o rácio de verdadeiros positivos (TPR)(2.2) já que, conforme referido em 2.6, é considerado um erro grave a classificação de uma amostra positiva (tumor maligno) em negativa (tumor benigno).

4.1 Preparação dos dados

Tendo sido realizadas as transformações aos sinais de micro-ondas descritas em 3.3, perante os resultados obtidos apresentados nas tabelas com os melhores resultados de classificação obtidos com cada uma das quatro formas de preparação dos sinais (4.1, 4.2, 4.3 e 4.4), constatou-se que quando foi usada a média dos sinais das 4 antenas se obteve o melhor desempenho: para a classe Maligno

obteve-se um máximo de 90, 28% de acertos e um máximo de 97, 22% para a classe Grande.

Tabela 4.1: Melhores resultados para a média dos sinais das 4 antenas

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
Média	Holdout	DWT	RF	Maligno	88,89	95,83
Média	Holdout	DWT	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	95,83
Média	Holdout	PCA	RF	Maligno	87,50	100,00
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67

Tabela 4.2: Melhores resultados para o conjunto dos sinais das 4 antenas

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
TodasAsAntenas	Holdout	PCA	SVM	Maligno	87,15	86,98
TodasAsAntenas	Holdout	DWT	RF	Maligno	86,46	94,27
TodasAsAntenas	Holdout	PCA	SVM	Maligno	86,11	84,90
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	86,11	90,63
TodasAsAntenas	Bootstrap	Interp	SVM	Grande	92,36	93,75
TodasAsAntenas	Bootstrap	DWT	RF	Grande	91,67	97,22
TodasAsAntenas	Bootstrap	DWT	SVM	Grande	91,32	96,53
TodasAsAntenas	Bootstrap	DWT	SVM	Grande	91,32	95,14
TodasAsAntenas	Bootstrap		RF	Grande	91,32	96,53

Tabela 4.3: Melhores resultados para os sinais das 4 antenas separados

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
PorAntena	Holdout	Interp	RF	Maligno	87,50	93,75
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	95,83
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	93,75
PorAntena	Bootstrap	PCA	SVM	Maligno	84,72	93,75
PorAntena	Holdout	PCA	SVM	Maligno	83,33	81,25
PorAntena	Bootstrap	PCA	SVM	Grande	91,67	94,44
PorAntena	Bootstrap	DWT	SVM	Grande	91,67	97,22
PorAntena	Bootstrap	DWT	SVM	Grande	91,67	97,22
PorAntena	Holdout	PCA	SVM	Grande	90,28	86,11
PorAntena	Holdout	PCA	LDA	Grande	90,28	86,11

Tabela 4.4: Melhores resultados para os sinais das 4 antenas concatenados

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Concatenação	Bootstrap	PCA	SVM	Maligno	85,07	89,58
Concatenação	Holdout	Interp	RF	Maligno	84,72	97,92
Concatenação	Holdout	Interp	RF	Maligno	84,72	97,92
Concatenação	Bootstrap	PCA	SVM	Maligno	84,72	89,06
Concatenação	Holdout	Interp	RF	Maligno	83,33	97,92
Concatenação	Holdout	Interp	SVM	Grande	91,67	88,89
Concatenação	Bootstrap	DWT	SVM	Grande	91,67	97,22
Concatenação	Bootstrap	DWT	SVM	Grande	91,67	97,22
Concatenação	Bootstrap	Interp	RF	Grande	91,67	95,83
Concatenação	Bootstrap	DWT	RF	Grande	91,32	95,14

Por contraste o pior modo de preparação dos dados foi a concatenação dos sinais das 4 antenas tanto na classificação da classe Maligno (máximo de 85, 07%) como na classificação da classe Grande (máximo de 91, 67%).

4.2 Divisão dos dados

Nesta secção são apresentados os resultados obtidos com as duas técnicas de divisão do conjunto de dados em duas partes: *Holdout* e *Bootstrap*.

Holdout

Com a divisão por *Holdout* os valores máximos de desempenho na classificação da classe Maligno foram atingidos com o algoritmo RF (90, 28%) e da classe Grande com o algoritmo SVM (94, 44%), conforme se pode ver na tabela 4.5. Das técnicas de extração de caraterísticas destacam-se a DWT e as *Wavelets* interpolatórias. Do modo de preparação de dados destaca-se a média das 4 antenas.

Tabela 4.5: Melhores resultados com *Holdout*

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
Média	Holdout	DWT	RF	Maligno	88,89	95,83
Média	Holdout	DWT	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	95,83
Média	Holdout	PCA	RF	Maligno	87,50	100,00
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67
Média	Holdout	DWT	SVM	Grande	93,06	91,67
Média	Holdout	DWT	SVM	Grande	91,67	91,67
Média	Holdout	DWT	SVM	Grande	91,67	91,67

Bootstrap

Como se pode constatar da tabela de melhores resultados com *Bootstrap* (tabela 4.6) os melhores desempenhos na classificação na classe Maligno foram obtidos com os algoritmos SVM e RF e com as técnicas de extração PCA e DWT. Quanto ao modo de preparação de dados apenas não aparece nestes 5 melhores resultados a concatenação dos sinais das 4 antenas.

Já na classificação da classe Grande o modo de preparação de dados com melhores resultados foi a média das 4 antenas, constando nestes melhores resultados todas a técnicas de extração e algoritmos de classificação, sendo que a combinação com o melhor resultado de todos foi a utilização de PCA com SVM.

Tabela 4.6: Melhores resultados com Boostrap

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
Média	Bootstrap	DWT	RF	Maligno	86,11	97,92
Média	Bootstrap	DWT	RF	Maligno	86,11	95,83
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	95,83
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	93,75
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Bootstrap	DWT	RF	Grande	94,44	100,00
Média	Bootstrap	Interp	SVM	Grande	94,44	97,22

4.3 Extração de caraterísticas

Nesta secção são apresentados os resultados obtidos com as diversas técnicas de extração de caraterísticas utilizadas: DWT, PCA e *Wavelets* interpolatórias.

DWT

Na tabela 4.7 são apresentados os 5 melhores valores de desempenho obtidos com esta técnica, quer para a classe Maligno quer para a classe Grande.
Os valores máximos de desempenho na classificação da classe Maligno com esta técnica de extração de caraterísticas foram obtidos com o algoritmo de classificação RF, com o método de divisão *Holdout* e utilizando a média dos sinais das 4 antenas (90, 28%). Para a classificação na classe Grande observa-se que os melhores valores também foram obtidos com RF (95, 83%), mas também se obtiveram bons valores com SVM, tendo em comum a utilização da média dos sinais das 4 antenas. Nos melhores resultados de classificação desta classe Grande destaca-se também o método de divisão de dados *Bootstrap*.

Tabela 4.7: Melhores resultados com DWT

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
Média	Holdout	DWT	RF	Maligno	88,89	95,83
Média	Holdout	DWT	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	95,83
TodasAsAntenas	Holdout	DWT	RF	Maligno	86,46	94,27
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Bootstrap	DWT	RF	Grande	94,44	100,00
Média	Holdout	DWT	SVM	Grande	93,06	91,67
Média	Bootstrap	DWT	SVM	Grande	93,06	97,22
Média	Bootstrap	DWT	SVM	Grande	93,06	94,44

PCA

Pelos resultados de desempenho obtidos com PCA (apresentados na tabela 4.8 de valores máximos) verifica-se que para a classe Maligno obtiveram-se bons desempenhos quer com o algoritmo SVM, quer com o RF (variando os métodos de divisão entre *Holdout* e *Bootstrap*) e, quer usando os sinais de todas as antenas quer a média dos sinais das 4 antenas. Sendo que o melhor resultado foi obtido com a média das 4 antenas, com a divisão por *Holdout* e com a classificação por RF ($87,50\%$).

Quanto à classificação na classe Grande o melhor resultado foi obtido com a média das 4 antenas, com a divisão por *Bootstrap* e com a classificação por SVM ($97,22\%$). Obtendo-se também bons resultados com a classificação por LDA e usando os sinais das 4 antenas em separado (com votação para obter a classificação final), mas sempre com a divisão de dados por *Bootstrap*.

Tabela 4.8: Melhores resultados com PCA

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	PCA	RF	Maligno	87,50	100,00
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
TodasAsAntenas	Holdout	PCA	SVM	Maligno	87,15	86,98
Média	Holdout	PCA	RF	Maligno	86,11	91,67
Média	Holdout	PCA	RF	Maligno	86,11	100,00
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Bootstrap	PCA	LDA	Grande	93,06	91,67
PorAntena	Bootstrap	PCA	SVM	Grande	91,67	94,44
Média	Bootstrap	PCA	SVM	Grande	91,67	91,67

Wavelets interpolatórias

O desempenho obtido com este método de extração de caraterísticas, conforme se pode verificar na tabela 4.9, os melhores resultados na classe Grande foram obtidos com a classificação através de SVM (94, 44%) enquanto que para Maligno foram com a classificação através de RF (87, 50%). Destacando-se também o método de divisão *Holdout* e a utilização da média dos sinais das 4 antenas.

Tabela 4.9: Melhores resultados com *Wavelets* interpolatórias

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
PorAntena	Holdout	Interp	RF	Maligno	87,50	93,75
Média	Holdout	Interp	RF	Maligno	86,11	93,75
Média	Holdout	Interp	RF	Maligno	86,11	93,75
Média	Holdout	Interp	RF	Maligno	86,11	91,67
TodasAsAntenas	Holdout	Interp	RF	Maligno	85,42	92,19
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67
Média	Bootstrap	Interp	SVM	Grande	94,44	97,22
Média	Bootstrap	Interp	SVM	Grande	93,06	94,44
PorAntena	Bootstrap	Interp	SVM	Grande	92,36	93,75

4.4 Classificação

Nesta secção são apresentados os resultados obtidos colocando em evidência os algoritmos de classificação utilizados: SVM, LDA e Random Forests. Através

desses resultados podemos constatar que o algoritmo que produziu melhores resultados foi o SVM com 97, 22% de acertos na classe Grande.

SVM

Analisando os valores apresentados na tabela dos melhores resultados com SVM (tabela 4.10), verifica-se que os melhores desempenhos se obtiveram para a classe Grande, com extração através de PCA (97, 22%), mas também com *Wavelets* interpolatórias e DWT, mas sempre com a média dos sinais das antenas. Os melhores desempenhos na classificação de Maligno foram inferiores aos obtidos para a classe Grande, tendo sido obtidos sempre com a extração de caraterísticas através de PCA, usando quer os sinais de todas as antenas (87, 50%) quer a classificação das 4 antenas em separado (86, 11%).

Tabela 4.10: Melhores resultados com SVM

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
TodasAsAntenas	Holdout	PCA	SVM	Maligno	87,15	86,98
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	95,83
PorAntena	Bootstrap	PCA	SVM	Maligno	86,11	93,75
TodasAsAntenas	Holdout	PCA	SVM	Maligno	86,11	84,90
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67
Média	Bootstrap	Interp	SVM	Grande	94,44	97,22
Média	Holdout	DWT	SVM	Grande	93,06	91,67

LDA

Os resultados obtidos sem extração de prévia de caraterísticas foram bastante pobres, como se pode verificar na tabela 4.11, o que revela que o referido algoritmo não tem bom desempenho como classificador de sinais com elevada quantidade de caraterísticas.

Tabela 4.11: Classificação com LDA sem extração de caraterísticas

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
PorAntena	Bootstrap		LDA	Maligno	65,28	68,75
TodasAsAntenas	Bootstrap		LDA	Maligno	63,19	63,54
Concatenação	Bootstrap		LDA	Maligno	63,19	63,54
Média	Bootstrap		LDA	Maligno	61,11	66,67
PorAntena	Holdout		LDA	Maligno	56,94	58,33
Média	Bootstrap		LDA	Grande	77,78	72,22
TodasAsAntenas	Bootstrap		LDA	Grande	73,96	69,44
Concatenação	Bootstrap		LDA	Grande	73,96	69,44
PorAntena	Bootstrap		LDA	Grande	73,61	69,44
Média	Holdout		LDA	Grande	63,89	69,44

O desempenho do algoritmo LDA, quer na classificação da classe Maligno quer na classe Grande, pode ser verificado na tabela 4.12. Dela se pode concluir que os melhores resultados para a classe Maligno são inferiores aos obtidos para a classe Grande. Sendo que na classe Maligno os melhores resultados foram obtidos com o método de divisão *Holdout* e com as técnicas de extração PCA (82, 29%) e DWT (81, 94%). Para a classe Grande foi a técnica PCA que se destacou com melhores resultados (95, 83%), mas também se obtiveram bons resultados com as *Wavelets* interpolatórias (91, 67%) e com DWT (90, 28%). O método de preparação de dados que mais se destacou com melhores resultados foi quando se utilizou a média dos sinais das 4 antenas.

Tabela 4.12: Melhores resultados com LDA

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
TodasAsAntenas	Holdout	PCA	LDA	Maligno	82,29	81,25
Média	Holdout	DWT	LDA	Maligno	81,94	81,25
Média	Holdout	PCA	LDA	Maligno	81,94	79,17
TodasAsAntenas	Holdout	PCA	LDA	Maligno	81,25	80,21
TodasAsAntenas	Holdout	PCA	LDA	Maligno	80,90	80,21
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Bootstrap	PCA	LDA	Grande	93,06	91,67
Média	Holdout	Interp	LDA	Grande	91,67	88,89
PorAntena	Holdout	PCA	LDA	Grande	90,28	86,11
TodasAsAntenas	Holdout	DWT	LDA	Grande	90,28	86,81

Random Forests

Também neste caso foram analisados comparativamente os resultados obtidos sem extração de prévia de caraterísticas. Ao contrário do que aconteceu com o LDA estes resultados foram considerados bons, em linha com os obtidos quando se utilizou extração de caraterísticas, como se pode verificar na tabela 4.13.

Tabela 4.13: Classificação com RF sem extração de caraterísticas

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout		RF	Maligno	84,72	89,58
TodasAsAntenas	Holdout		RF	Maligno	83,33	90,63
Concatenação	Holdout		RF	Maligno	83,33	97,92
Média	Bootstrap		RF	Maligno	81,94	95,83
Concatenação	Bootstrap		RF	Maligno	79,17	94,27
Média	Bootstrap		RF	Grande	93,06	97,22
TodasAsAntenas	Bootstrap		RF	Grande	91,32	96,53
Concatenação	Bootstrap		RF	Grande	91,32	96,53
TodasAsAntenas	Holdout		RF	Grande	88,54	82,64
Média	Holdout		RF	Grande	87,50	83,33

Analisando o desempenho geral com RF na tabela 4.14 verifica-se que os melhores desempenhos foram obtidos para a classe Grande (95, 83%) e sempre com o método de divisão *Bootstrap*. Curiosamente os melhores desempenhos para a classe Maligno foram obtidos com o método de divisão *Holdout* (90, 28%). A técnica de extração de caraterísticas com sucesso mais constante com este algoritmo

foi nitidamente a DWT mas também se obteve um resultado muito bom com as *Wavelets* interpolatórias.
O método de preparação de dados que mais se destacou com melhores resultados foi quando se utilizou a média dos sinais das 4 antenas.

Tabela 4.14: Melhores resultados com RF

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
Média	Holdout	DWT	RF	Maligno	88,89	95,83
PorAntena	Holdout	Interp	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	95,83
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Bootstrap	DWT	RF	Grande	94,44	100,00
Média	Bootstrap	DWT	RF	Grande	93,06	97,22
Média	Bootstrap	DWT	RF	Grande	93,06	100,00
Média	Bootstrap		RF	Grande	93,06	97,22

4.4.1 Classes de classificação

Como tem vindo a ser constatado nas análises de resultados nas diversas vertentes referidas nos parágrafos anteriores a classificação na classe Grande tem apresentado melhores desempenhos que a classificação na classe Maligno. Esta é uma constatação que intuitivamente já poderíamos esperar pois é para nós humanos mais fácil distinguir um tamanho de que uma forma. O curioso é isso também acontecer com os processos de ML.
Os melhores resultados para cada classe são apresentados na tabela 4.15 de onde pode extrair as seguintes particularidades:

- Todos os melhores desempenhos para a classe Grande foram obtidos utilizando a média dos sinais das 4 antenas. Obtiveram-se bons resultados com qualquer dos algoritmos de classificação, destacando-se o SVM (97, 22%) entre eles, e com qualquer das técnicas de extração de caraterísticas.

- Para a classe Maligno os melhores resultados foram obtido sempre com o algoritmo de classificação RF e com o método de divisão *Holdout*, maioritariamente sobre a média dos sinais das 4 antenas. O melhor desempenho obteve-se usando DWT para extração (90, 28%), mas também se obtiveram bons resultados as *Wavelets* interpolatórias (87, 50%).

Tabela 4.15: Melhores resultados para cada classe

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
Média	Holdout	DWT	RF	Maligno	88,89	95,83
PorAntena	Holdout	Interp	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	93,75
Média	Holdout	DWT	RF	Maligno	87,50	95,83
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67

4.4.2 Impacto da simplificação dos sinais

De modo a confirmar as vantagens com a simplificação dos sinais (reduzindo-os de 4.000 pontos para 1.525 por eliminação dos seus extremos), procedeu-se a alguns ensaios de classificação com SVM e RF sem essa simplificação. Verificou-se que com os sinais simplificados se obtém ganhos significativos no tempo de processamento (de extração e classificação), observando-se ainda ligeiras melhorias no desempenho na classificação. Os testes foram feitos usando os sinais simplificados de todas as antenas e com a divisão entre treino e teste através de *Holdout*. Os resultados foram resumidos na tabela 4.16.
Por exemplo, usando DWT para extração e SVM para classificação demorou menos 65% no tempo de processamento (de 5 minutos para $1,784$ minutos), e uma ligeira melhoria no desempenho médio (de $84,02\%$ para $84,29\%$). Verifica-se que os ganhos no tempo de processamento são ainda maiores quando se usam as *Wavelets* interpolatórias, de $13,315$ para $3,232$ minutos (83% menos tempo), mas nesse caso verifica-se que o desempenho da classificação (com SVM) é melhor quando a interpolação é feita a partir dos sinais originais (de $80,87\%$ para $83,01\%$).

Tabela 4.16: Desempenho sem simplificação de sinais

EXTRAÇÃO	CLASSIFICAÇÃO	DESEMPENHO MÉDIO (COM SIMPLIFICAÇÃO)	DEGRADAÇÃO DO TEMPO DE PROCESSAMENTO
DWT	SVM	84,02% (84,29%)	65%
INTERP	SVM	83,21% (82,97%)	83%
PCA	SVM	80,21% (83,19%)	
DWT	RF	87,5% (88,59%)	
PCA	RF	80,78% (83,51%)	

4.4.3 Melhores resultados globais

Analisando os melhores resultados globais de desempenho da classificação na tabela 4.17 com os 5 melhores desempenhos na classificação de cada classe Maligno e Grande podemos constatar que o melhor resultado para a classe Maligno foi de $90,28\%$ de acertos quando se usou o algoritmo de classificação RF, a técnica de extração DWT, o modo de divisão de dados *Holdout* aplicados sobre a média dos sinais das 4 antenas. O melhor resultado para a classe Grande foi de $97,22\%$ de acertos quando se usou o algoritmo de classificação SVM, a técnica de extração PCA, o modo de divisão de dados *Bootstrap* aplicados sobre a média dos sinais das 4 antenas.

Tabela 4.17: Melhores resultados globais

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Bootstrap	DWT	RF	Grande	95,83	100,00
Média	Bootstrap	PCA	LDA	Grande	95,83	94,44
Média	Holdout	Interp	SVM	Grande	94,44	94,44
Média	Holdout	Interp	SVM	Grande	94,44	91,67

5

Conclusões e trabalho futuro

Neste capítulo são apresentados um resumo de todo o estudo e trabalho efetuados no âmbito desta dissertação e as conclusões pertinentes resultantes desse trabalho. Apresenta-se também uma avaliação crítica dos resultados de classificação obtidos com os vários ensaios efetuados, os quais foram descritos detalhadamente no capítulo 4, e algumas sugestões para trabalhos futuros.

5.1 Resumo da dissertação

O desenvolvimento desta dissertação iniciou-se com a pesquisa de informação já existente sobre a temática do cancro da mama, quer na vertente da problemática social, quer ao nível da prevenção, diagnóstico, tratamentos e probabilidades de cura perante o estágio e caraterísticas da doença.
Seguiu-se a leitura de estudos e publicações sobre tudo o que diz respeito às novas aplicações das micro-ondas, a qual acabou por revelar uma miríade de vertentes relacionadas com essa tecnologia atualmente em desenvolvimento e investigação. Foram encontrados estudos sobre a eletrónica utilizada na construção das antenas e dos seus circuitos, sobre o comportamento da emissão e receção dessas radiações face ao tipo e configurações de antenas utilizados, sobre quais as frequências mais adequadas para cada meio ambiente e tipo de aplicação, e também sobre as interferências e ruídos produzidos nos diversos tipos de materiais que essas radiações atravessam.

Na vertente de diagnóstico de doenças e tratamentos médicos a pesquisa de informação já publicada desvendou a diversidade de aplicações já concretizadas e em investigação. Concretamente no rastreio e diagnóstico do cancro da mama constatou-se a existência de desenvolvimentos em diferentes áreas:

- Criação de protótipos de aparelhos de rastreio por forma a testar a sua ergonomia e a acoplação das antenas aos tecidos humanos;
- Obtenção de imagens médicas a partir dos sinais de micro-ondas obtidos em exames de rastreio;
- Finalmente, e afinal a maior parte de toda esta investigação, a classificação destes sinais com recurso a técnicas de aprendizagem automática (ML), a qual é a motivação e o tema da presente dissertação.

Face à informação consultada sobre a classificação de sinais de micro-ondas com recurso a técnicas de aprendizagem automática (ML) e, na sequência de estudos anteriores semelhantes a este, foram então selecionadas as ferramentas, métodos de preparação de dados, técnicas de extração de caraterísticas e algoritmos de classificação a utilizar no trabalho prático desta dissertação. Na preparação desse trabalho prático, começou-se pela aprendizagem das ferramentas a utilizar, seguindo-se a análise do conjunto de dados fornecido para o trabalho, e o estudo dos *packages* R que implementam as técnicas e algoritmos selecionados, incluindo as respetivas configurações e parametrizações. No final dessa preparação, e com o conhecimento obtido com a mesma, juntamente com o obtido no estudo anteriormente realizado, foram definidas as etapas do processo de classificação e escolhidos os vários ensaios a realizar.
Para avaliação do desempenho dos processos de classificação foi usada a percentagem de acertos (*accuracy*)(2.4). É apresentado também o rácio de verdadeiros positivos (TPR), já que é um indicador importante na classificação de doenças, pois reflete a quantidade de erros graves na classificação (classificar uma amostra positiva como negativa). Não foram apresentadas curvas ROC de desempenho em virtude de se ter constatado que a variação da parametrização dos algoritmos não produz um impacto linear nos valores de TPR e TNR, não sendo desse modo possível identificar o ponto de equilíbrio para esses dois indicadores (ver secção 2.6).

5.2 Conclusões principais

Como resultado de todos os ensaios realizados no decurso da parte prática desta dissertação, e face aos valores de desempenho obtidos nesses ensaios, é apresentado abaixo um resumo das conclusões mais pertinentes.

Melhor algoritmo de classificação

O melhor desempenho de classificação foi obtido com o algoritmo SVM, com $97,22\%$ de acertos na classificação da classe Grande. Este algoritmo também já apresentou o melhor resultado noutros estudos de classificação de sinais de micro-ondas em diagnóstico de cancro da mama (por exemplo em [17] e [2]), pelo que se conclui ser este o algoritmo com maiores potencialidades de aplicação em sistemas reais de diagnóstico de cancro da mama baseados em micro-ondas. Este resultado foi obtido quando se utilizou a média dos sinais das quatro antenas, a divisão de dados por *Bootstrap* e a extração de caraterísticas com PCA. Para se obter esse resultado utilizaram-se um *kernel* RBF (radial), 10^{-3} para valor de γ e 10 para valor do custo.

Utilização de *Wavelets* interpolatórias

Num dos objetivos originais desta dissertação pretendeu-se avaliar a utilização das *Wavelets* interpolatórias na extração de caraterísticas destes sinais de micro-ondas. Os resultados obtidos nesses ensaios permitem concluir que as *Wavelets* interpolatórias comprovaram ser adequadas para o efeito, tendo o seu melhor desempenho ($94,44\%$ de acertos) sido obtido com o algoritmo SVM para a classificação da classe Grande.

Esta técnica é assim merecedora de futuros estudos, necessitando particularmente de estudos centrados em encontrar outras formas de representação dos sinais interpolados. Esta necessidade de transformar os resultados da interpolação para um formato que possibilite a sua utilização pelos algoritmos de classificação, abre portas à investigação de outras formas de transformação dos sinais interpolados que tragam maiores vantagens ao processo de classificação.

Influência dos métodos de preparação dos dados

Outro objetivo original desta dissertação foi a demonstração da influência dos

métodos de preparação dos dados no resultado final da classificação. Face aos resultados obtidos nos diversos ensaios realizados considera-se que ficou demonstrada tal influência.
Como é facilmente observável nas tabelas apresentadas na secção onde são analisados os resultados obtidos com os métodos de preparação dos dados (ver secção 4.1), existem diferenças de desempenho consideráveis entre os diferentes métodos. Essas diferenças estão resumidas na tabela 5.1, e permitem concluir que o método de preparação de dados com melhor desempenho é a média dos sinais das 4 antenas, quer na classificação da classe Maligno quer na da Grande.

Tabela 5.1: Melhores resultados com os diferentes métodos de preparação dos dados

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
PorAntena	Holdout	Interp	RF	Maligno	87,50	93,75
Concatenação	Bootstrap	PCA	SVM	Maligno	85,07	89,58
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
TodasAsAntenas	Bootstrap	Interp	SVM	Grande	92,36	93,75
PorAntena	Bootstrap	DWT	SVM	Grande	91,67	97,22
Concatenação	Holdout	Interp	SVM	Grande	91,67	88,89

Essa constatação vem contrariar a hipótese colocada inicialmente, de que a classificação do sinal de cada antena isoladamente (sendo a classificação final obtida por votação) seria a mais adequada. Dado que a base de raciocínio dessa hipótese reside na maior ou menor proximidade de cada antena ao tumor (é intuitivo que as antenas mais próximas do tumor o detetam com maior facilidade), proporciona-se a realização de trabalhos futuros usando esta forma de classificação, mas com um maior número de antenas e com diversos posicionamentos do tumor na mama.
É também necessário desenvolver estudos que abordem as influências da preparação dos dados em todo o processo de classificação, por forma a encontrar uma forma de determinar qual a melhor escolha perante a necessidade concreta de realizar determinada classificação de sinais (de micro-ondas ou outros).

Influência dos métodos de divisão dos dados

Relativamente aos métodos de divisão dos dados, face aos resultados obtidos e que se encontram resumidos na tabela com os melhores resultados obtidos com

os diferentes métodos de separação dos dados 5.2, ficou também demonstrada a sua influência no desempenho da classificação.

Tabela 5.2: Melhores resultados com os diferentes métodos de separação dos dados

PREPARAÇÃO		CLASSIFICAÇÃO			RESULT. %	
dataset	separa.	extração	classif.	classe	Acc	TPR
Média	Holdout	DWT	RF	Maligno	90,28	95,83
TodasAsAntenas	Bootstrap	PCA	SVM	Maligno	87,50	94,27
Média	Bootstrap	PCA	SVM	Grande	97,22	97,22
Média	Holdout	Interp	SVM	Grande	94,44	94,44

O modo de divisão dos dados que apresentou melhores desempenhos na classificação da classe Maligno foi o *Holdout* (com 90, 28% de acertos) enquanto que para a classificação na classe Grande foi o *Bootstrap* (com 97, 22% de acertos). Não sendo possível vislumbrar uma justificação para este facto, será um tema a analisar em futuros estudos que incidam na forma como a divisão dos dados entre treino e teste afeta os resultados da classificação face à classe que se pretende classificar.

Viabilidade da classificação dos sinais de micro-ondas de rastreio do cancro da mama

Dos desempenhos de classificação obtidos nos diversos ensaios, quer observando os valores de *Accuracy*, que os de TPR, pode-se concluir positivamente sobre a viabilidade da aplicação de processos de classificação de ML em sinais de micro-ondas de rastreio do cancro da mama e, consequentemente, da sua potencial aplicabilidade em exames de rastreio reais. É possível constatar que os valores de desempenho obtidos foram bastante satisfatórios, tendo-se atingido bastante frequentemente valores acima dos 90% de acertos (com um máximo de 97, 22%), e valores de TPR acima de 95%).

O passo seguinte para confirmar essa aplicabilidade prática será repetir estes ensaios de classificação com base em sinais obtidos em rastreios reais em voluntários e utilizando protótipos de dispositivos de rastreio adequados à realização de exames com seres humanos. Apenas nessa altura, e perante os sinais e condições reais, será possível determinar que método de preparação e divisão de dados, que técnica de extração de caraterísticas e que algoritmo de classificação (e respetivas parametrizações) serão os mais adequados.

Diferença de desempenho entre as duas classes de classificação

Relativamente à classificação de cada uma das duas classes torna-se facilmente percetível, através das várias tabelas de melhores resultados apresentadas no capítulo 4, que a classificação na classe Grande apresenta consistentemente melhor desempenho que a classificação na classe Maligno. Talvez porque, à semelhança da visão humana, os algoritmos matemáticos utilizados distingam melhor o tamanho do que a forma.

- Relembra-se que, no caso dos tumores da mama, o que diferencia um tumor maligno é o facto de o seu perfil ser mais ou menos rendilhado.

5.3 Trabalho futuro

Com base no trabalho realizado nesta dissertação, nos resultados obtidos e respetivas conclusões, são apresentadas as seguintes sugestões para a realização de trabalhos futuros:

- **Novas formas de utilização de *Wavelets* interpolatórias**
  Sugere-se realização de estudos para encontrar diferentes formas de representação para os sinais interpolados que eventualmente potenciem o desempenho dos algoritmos de classificação.

- **Influência do método de preparação dos dados na classificação**
  Os resultados obtidos nesta dissertação permitiram perceber a necessidade de estud-os futuros centrados no objetivo de determinar de que forma o método de preparação dos dados, no caso de sinais oriundos de várias antenas, influencia o desempenho de todo o processo de classificação.
  É igualmente pertinente a realização de estudos para avaliar o desempenho da classificação dos sinais de cada antena isoladamente, mas em rastreios usando uma maior quantidade de antenas e com diversas localizações dos tumores.

- **Influência do modo de divisão de dados na classificação de diferentes classes**
  Atendendo à particularidade dos resultados obtidos nesta dissertação (diferenças evidentes no desempenho da classificação de diferentes classes), serão necessários estudos acerca da influência do método de divisão do

conjunto de dados, em treino e teste, no desempenho na classificação de diferentes classes.

- **Utilização de *Deep Learning***
  Este tipo de classificação está atualmente no topo do estado da arte da classificação em ML, mas atendendo à escassez dos dados de teste utilizados e às ferramentas de desenvolvimento utilizadas, não foi possível endereçar este método de classificação na presente dissertação.

Referências

[1] Pinho, Pedro Renato Tavares: *Resolução das equações de Maxwell por análise multiresolução usando wavelets interpolatória*. Tese de Doutoramento, Universidade de Aveiro, 2004. (pp. xvii, 4, 10, 11, and 36)

[2] Santorelli, Adam, Emily Porter, Evgeny Kirshin, Yi Jun Liu e Milica Popovic: *Investigation of classifiers for tumor detection with an experimental time-domain breast screening system*. Progress In Electromagnetics Research, 144:45–57, 2014. (pp. xvii, 18, and 59)

[3] Wu, Yizhi, Mingda Zhu, Danmei Li, Youtao Zhang e Yifan Wang: *Brain stroke localization by using microwave-based signal classification*. Em *Electromagnetics in Advanced Applications (ICEAA), 2016 International Conference on*, páginas 828–831. IEEE, 2016. (pp. xvii and 22)

[4] *Spie digital library*. https://www.spiedigitallibrary.org/. (pp. xvii and 28)

[5] Medeiros, Hugo Filipe de Carvalho da *et al.*: *Classificação de tumores de cancro na mama através de radar de banda ultra-larga de microondas*. Tese de Doutoramento, Faculdade de Ciências e Tecnologia, 2013. (pp. xvii and 28)

[6] Nikolova, Natalia K: *Microwave near-field imaging of human tissue: Hopes, Challenges, Outlook*. 2012. (p. 2)

[7] IEEE: *Microwave Theory and Techniques Society (MTT–S) @ONLINE*. December 2016. http://www.mtt.org/fellowship-projects. (p. 2)

[8] Persson, Mikael, Andreas Fhager, Hana Dobšíček Trefná, Yinan Yu, Tomas

McKelvey, Göran Pegenius, Jan Erik Karlsson e Mikael Elam: *Microwave-based stroke diagnosis making global prehospital thrombolytic treatment possible*. IEEE Transactions on Biomedical Engineering, 61(11):2806–2817, 2014. (p. 2)

[9] Patlak, Margie, S Nass, I Henderson e J Lashof: *Mammography and beyond: developing technologies for the early detection of breast cancer*. Atlanta, GA, USA: Nat. Acad. Press, 2001. (p. 2)

[10] Bird, Richard E, Terry W Wallace e Bonnie C Yankaskas: *Analysis of cancers missed at screening mammography.* Radiology, 184(3):613–617, 1992. (p. 2)

[11] Chen, S C, Y C Cheung, C H Su, M F Chen, T L Hwang e S Hsueh: *Analysis of sonographic features for the differentiation of benign and malignant breast tumors of different sizes*. Ultrasound in obstetrics & gynecology, 23(2):188–193, 2004. (p. 2)

[12] Stuchly, Maria A: *Applications of microwaves in medicine*. IEEE AP-S Lecture, 2006. (p. 2)

[13] Nilavalan, R., J. Leendertz, I. J. Craddock, A. Preece e R. Benjamin: *Numerical analysis of microwave detection of breast tumours using synthetic focussing techniques*. 2004. (p. 2)

[14] DONNA, EUROPA: *Guide to breast health*. February 2013. https://www.europadonna.org/wp-content/uploads/2013/02/EDGuideToBreastHealth.pdf. (p. 3)

[15] CANCER, IARC INTERNATIONAL AGENCY FOR RESEARCH ON: *Breast cancer incidence, mortality, and prevalence worldwide*. October 2016. https://www.iarc.fr/en/media-centre/iarcnews/2016/BreastCancer_Graphics_GCO_Oct2016.pdf. (p. 3)

[16] Saúde, Direção Geral de: *Estatísticas da saúde*. https://www.dgs.pt/publicacoes/estatisticas-da-saude.aspx. (p. 3)

[17] Conceição, Raquel: *The Development of Ultra Wideband Scanning Techniques for Detection and Classification of Breast Cancer (Cap. 5)*. Tese de Doutoramento, National University of Ireland Galway, September 2010. http://hdl.handle.net/10379/1821. (pp. xvii, 3, 4, 16, 27, 35, and 59)

[18] Witten, Ian H, Eibe Frank, Mark A Hall e Christopher J Pal: *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016. (pp. 6, 7, and 8)

[19] Friedman, Jerome, Trevor Hastie e Robert Tibshirani: *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001. (pp. 7 and 8)

[20] Chawla, Nitesh V, Nathalie Japkowicz e Aleksander Kotcz: *Special issue on learning from imbalanced data sets*. ACM Sigkdd Explorations Newsletter, 6(1):1–6, 2004. (p. 7)

[21] Chawla, Nitesh V: *Data mining for imbalanced datasets: An overview*. Em *Data mining and knowledge discovery handbook*, páginas 875–886. Springer, 2009. (p. 7)

[22] Haar, Alfred: *On the theory of orthogonal function systems*. https://pdfs.semanticscholar.org/3b08/b61ba914626db518b6add5b73ac21d62f0c1.pdf. (p. 9)

[23] Daubechies, Ingrid: *Ten lectures on wavelets*. https://doi.org/10.1137/1.9781611970104. (p. 9)

[24] Zhang, Zitong, Qawi K Telesford, Chad Giusti, Kelvin O Lim e Danielle S Bassett: *Choosing wavelet methods, filters, and lengths for functional brain network construction*. PLoS One, 11(6):e0157243, 2016. (p. 9)

[25] Subasi, Abdulhamit e M Ismail Gursoy: *EEG signal classification using PCA, ICA, LDA and support vector machines*. Expert Systems with Applications, 37(12):8659–8666, 2010. (pp. 9, 12, 13, and 15)

[26] University, The Pennsylvania State: *Principal components analysis (PCA) lesson*. STAT 505 - Applied Multivariate Statistical Analysis online course. (pp. 11 and 35)

[27] Team, R Core e contributors worldwide: *The r stats package - principal components analysis*. 2018. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html. (p. 12)

[28] Shlens, Jonathon: *A tutorial on independent component analysis*. arXiv preprint arXiv:1404.2986, 2014. https://arxiv.org/pdf/1404.2986.pdf. (p. 12)

[29] Pushpa Rathi, V.P.Gladis e Dr.S. Palani: *Brain tumor mri image classification with feature selection and extraction using linear discriminant analysis*. arXiv preprint arXiv:1208.2128, 2012. http://arxiv.org/abs/1208.2128. (p. 13)

[30] Friedman, Jerome H: *Regularized discriminant analysis*. Journal of the American statistical association, 84(405):165–175, 1989. (p. 14)

[31] Cortes, Corinna e Vladimir Vapnik: *Support-vector networks*. Machine learning, 20(3):273–297, 1995. (p. 14)

[32] Osuna, Edgar, Robert Freund e Federico Girosit: *Training support vector machines: an application to face detection*. Em *Computer vision and pattern recognition, 1997. Proceedings, 1997 IEEE computer society conference on*, páginas 130–136. IEEE, 1997. (p. 15)

[33] Huang, Cheng Lung, Hung Chang Liao e Mu Chen Chen: *Prediction model building and feature selection with support vector machines in breast cancer diagnosis*. Expert Systems with Applications, 34(1):578–587, 2008. (p. 15)

[34] Meyer, David: *Support vector machines: The interface to `libsvm` in package `e1071`*. 2004. (pp. 15 and 40)

[35] Schohn, Greg e David Cohn: *Less is more: Active learning with support vector machines*. Em *ICML*, páginas 839–846. Citeseer, 2000. (p. 15)

[36] Breiman, Leo: *Random forests*. Machine learning, 45(1):5–32, 2001. (p. 15)

[37] O'Loughlin, Declan, Martin James O'Halloran, Brian M Moloney, Martin Glavin, Edward Jones e Muhammad Adnan Elahi: *Microwave breast imaging: Clinical advances and remaining challenges*. IEEE Transactions on Biomedical Engineering, 2018. (p. 17)

[38] Gerazov, Branislav e Raquel C Conceicao: *Deep learning for tumour classification in homogeneous breast tissue in medical microwave imaging*. Em *Smart Technologies, IEEE EUROCON 2017-17th International Conference on*, páginas 564–569. IEEE, 2017. (p. 19)

[39] Sacristán, Jorge, Bárbara Luz Oliveira e Stephen Pistorius: *Classification of electromagnetic signals obtained from microwave scattering over healthy and tumorous breast models*. Em *Electrical and Computer Engineering (CCECE), 2016 IEEE Canadian Conference on*, páginas 1–5. IEEE, 2016. (p. 20)

[40] Song, Hongchao, Yunpeng Li, Mark Coates e Aidong Men: *Microwave breast cancer detection using empirical mode decomposition features*. arXiv preprint arXiv:1702.07608, 2017. (p. 20)

[41] Sundström, Christoffer: *Machine learning algorithms for stroke diagnosis*. Tese de Mestrado, Chalmers University of Technology, Gothenburg, Sweden, 2014. http://publications.lib.chalmers.se/records/fulltext/200455/200455.pdf. (p. 21)

[42] Guo, Lei: *Processing and imaging techniques for microwave-based head imaging*. 2017. (p. 23)

[43] Aldrich, Eric: *A package of functions for computing wavelet filters, wavelet transforms and multiresolution analyses*. 2013. https://cran.r-project.org/web/packages/wavelets/wavelets.pdf. (p. 35)

[44] Brian Ripley, Bill Venables *et al.*: *Support functions and datasets for venables and ripley's mass*. 2018. https://cran.r-project.org/web/packages/MASS/MASS.pdf. (p. 40)

[45] Liaw, Andy e Matthew Wiener: *Breiman and cutler's random forests for classification and regression*. 2018. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. (p. 41)