INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Diagnóstico cardíaco a partir de dados acústicos e clínicos

ELISETE MARIA SILVA DIAS ALVES DA CRUZ
Outubro de 2015

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Diagnóstico cardíaco a partir de dados acústicos e clínicos

Elisete Maria Silva Dias Alves da Cruz

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Arquiteturas, Sistemas e Redes

Orientador: Elsa Maria de Carvalho Ferreira Gomes

Júri:

Presidente:

[Nome do Presidente, Categoria, Escola]

Vogais:

[Nome do Vogal1, Categoria, Escola]

[Nome do Vogal2, Categoria, Escola] (até 4 vogais)

Porto, Outubro 2015

Resumo

Este documento foi redigido no âmbito da dissertação do Mestrado em Engenharia Informática na área de Arquiteturas, Sistemas e Redes, do Departamento de Engenharia Informática, do ISEP, cujo tema é diagnóstico cardíaco a partir de dados acústicos e clínicos.

O objetivo deste trabalho é produzir um método que permita diagnosticar automaticamente patologias cardíacas utilizando técnicas de classificação de *data mining*. Foram utilizados dois tipos de dados: sons cardíacos gravados em ambiente hospitalar e dados clínicos. Numa primeira fase, exploraram-se os sons cardíacos usando uma abordagem baseada em *motifs*. Numa segunda fase, utilizamos os dados clínicos anotados dos pacientes. Numa terceira fase, avaliamos a combinação das duas abordagens. Na avaliação experimental os modelos baseados em *motifs* obtiveram melhores resultados do que os construídos a partir dos dados clínicos. A combinação das abordagens mostrou poder ser vantajosa em situações pontuais.

Palavras-chave: Sons cardíacos, *motifs*, dados clínicos, classificação, diagnóstico

Abstract

This document was written as part of the Thesis of the MSc in computer science in the area of Architecture, System and Network, Department of Computer Engineering in ISEP. The main theme of this Thesis is to diagnose cardiac diseases, through acoustic and clinical data.

The goal of this work is to produce a process for automatically diagnosing heart problems using *data mining* classification techniques. Two types of data were used: heart sounds recorded in hospitals and clinical data. Initially, we explored the heart sounds using an approach based on *motifs*. In a second stage, we used the clinical data of the patients. In a third phase, we evaluated the combination of both approaches. Experimental evaluation showed that models based on *motifs* performed better than those built from clinical data. The combination of approaches has shown to be advantageous in specific situations.

Keywords: Heart Sounds, *motifs*, clinic data, classification, diagnostic

Agradecimentos

Agradeço a todas as pessoas que estiveram, de alguma forma, envolvidas neste trabalho.

Agradeço à professora Elsa Gomes todo o trabalho, dedicação e motivação que deu a este trabalho e que ajudou a este ser concretizado. Ela foi incansável e mostrou-se sempre disponível e com vontade de levar este trabalho a bom porto.

Agradeço à minha família pela paciência que tiveram e por me terem sempre apoiado durante todo este processo. Sem eles nada disto era possível, daí a minha gratidão.

Por fim gostaria de agradecer a todos os que permitiram e ajudaram-me a completar mais um ciclo de estudos, obrigada.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

CSV	*Comma Separated Values*
SAX	*Symbolic Aggregate ApproXimation*
TVP	*Taxa de verdadeiro positivo*
TFP	T*axa de falso positivo*
TP	*Verdadeiro positivo*
TN	*Verdadeiro negativo*
FP	*Falso positivo*
FN	*Falso negativo*
ISEP	*Instituto Superior de Engenharia do Porto*
MLP	*Multilayer Perceptron*
ANN	*Artificial Neural Network*
ECG	*Eletrocardiograma*
AUC	*Area Under ROC Curve*
IMC	*Índice de massa corporal*
JVM	*Java virtual machine*

1 Introdução

Hoje em dia a saúde é uma das grandes preocupações da sociedade atual. Existem diversos grupos de cientistas que se dedicam à investigação de diversas patologias.

Uma das áreas onde a investigação está a ser desenvolvida é na deteção de patologias cardíacas. As doenças cardíacas são a principal causa de mortalidade nos países desenvolvidos.

Neste trabalho iremos utilizar dados recolhidos num hospital pediátrico. Esses dados são de natureza áudio (gravações de sons cardíacos) e anotações feitas pelos clínicos no hospital. Esses dados serão utilizados para o desenvolvimento de uma metodologia de classificação automática de patologias.

O grande desafio encontra-se na análise dos dados recolhidos pois foram gravados em ambientes com muito ruido e nem sempre com os aparelhos mais adequados para esse efeito.

A ideia dos *motifs* é descobrir, nos sons cardíacos, padrões relevantes e frequentes que sirvam de atributos para classificação.

1.1 Motivação

A principal motivação deste trabalho é produzir ferramentas de auxílio aos profissionais de saúde na área das doenças cardíacas. Numa área em que a deteção precoce de problemas é muito importante, um método de auxílio rápido assume uma grande importância.

A nível pessoal, a principal motivação que me levou a escolher este tema foi estar relacionado com a área da saúde. É uma área pela qual eu tenho imenso prazer em trabalhar. Esse gosto aliado a tentar ajudar os médicos a descobrir patologias que podem ajudar as pessoas foi a alavanca de escolher este projeto como tese.

O facto de esta tese ser noutra área que não a minha foi outro dos motivos. Tive hipótese de aprender novos algoritmos e novas tecnologias. Gosto de aprender novas coisas e ter conhecimento em diversas áreas.

Outra motivação foi também os dados recolhidos serem de crianças. Se for possível encontrar padrões de patologias em crianças, que são mais frágeis que os adultos, é provável que possa ser aplicados os mesmos padrões para outras faixas etárias.

1.2 Objetivos e contribuições

Este trabalho tem como principal objetivo explorar a combinação de dados clínicos com dados áudio para produzir um método capaz de diagnosticar automaticamente patologias cardíacas. Em particular, pretende-se comparar três abordagens: usando *motifs*, os dados clínicos e a combinação de ambos os atributos.

Assim, contribuição principal deste trabalho é a avaliação da importância de combinar a informação clinica do doente com os sons cardíacos na deteção de patologia.

1.3 Estrutura da dissertação

No primeiro capítulo encontra-se a Introdução. Este capítulo permite apresentar o tema que será abordado e de que forma será estruturada esta dissertação.

No segundo capitulo, Ferramentas e Tecnologias utilizadas, é realizada uma breve apresentação das ferramentas e tecnologias utilizadas no decorrer deste trabalho.

No terceiro capítulo, *Data mining*, é descrito com mais detalhe a base das análises que serão posteriormente realizadas e as principais características dos algoritmos que foram utilizados.

No quarto capitulo, Classificação de sons cardíacos, é descrito de que forma o problema da classificação dos sons cardíacos é abordada e quais as abordagens que foram previamente realizadas.

O quinto capítulo, Modelação, contem duas partes. Inicialmente é realizada uma descrição dos procedimentos realizados e descrição dos conceitos teóricos que serão aplicados nas análises. Após a introdução dos conceitos e procedimentos são apresentados os resultados obtidos para cada análise e para as diferentes abordagens.

No sexto capítulo está uma breve conclusão deste trabalho e também algumas sugestões para um trabalho futuro de forma a poder melhorar ainda todo este processo.

Por último encontra-se o capítulo com as referências utilizadas como suporte para a realização deste trabalho.

2 Ferramentas e Tecnologias utilizadas

Para realizar este trabalho foram utilizadas diversas tecnologias. Neste capítulo pretende-se descrever essas tecnologias e ferramentas. Será abordada a questão de tratamento de ficheiros de som, conceito de *data mining* (ou mineração de dados), assim como será efetuada uma contextualização das técnicas de Aprendizagem Automática aplicadas ao problema de deteção de patologia nos sons cardíacos. Serão também descritas algumas medidas de avaliação dos métodos dos modelos de classificação.

2.1 Audacity no pré-processamento de dados

A Audacity é uma ferramenta gratuita que permite gravar e editar sons. [audacity]

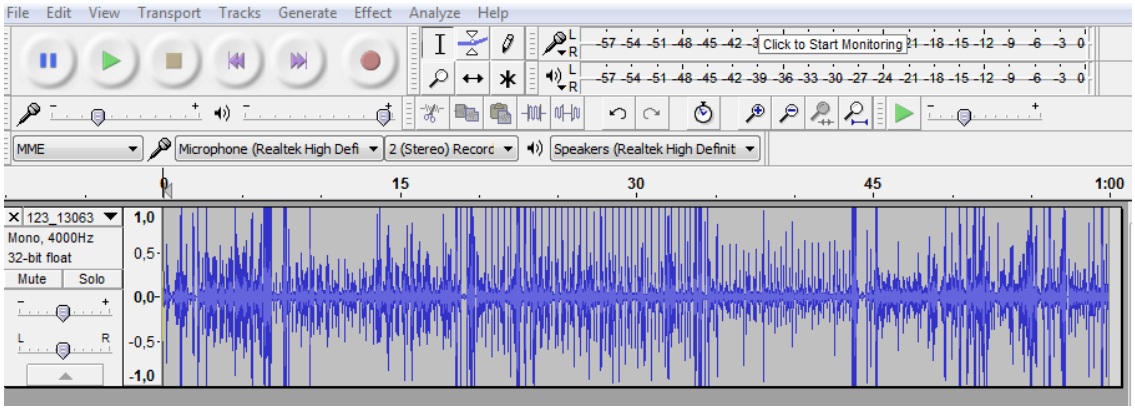

Figura 1 – Exemplo de um ficheiro por analisar no Audacity

Como é possível observar na Figura 1, esta ferramenta permite cortar, reduzir ruído e ver a amplitude de som do ficheiro. No contexto desta tese, foi utilizada para cortar os sons cardíacos originais fornecidos, com demasiado ruído e muito longos. Assim, o Audacity foi utilizado neste trabalho para cortar as partes relevantes nos ficheiros de som fornecidos. A técnica de cortar os ficheiros de som é referida na literatura (ver secção 4.2) pois, em ambiente real, é comum os ficheiros áudio captarem outros ruídos para além dos batimentos cardíacos tais como, ruído pulmonar, ruído do auscultador a roçar na roupa do auscultado, ruído ambiente, entre outros. É também comum os sons terem comprimentos muito díspares.

2.1.1 Java

Java é uma linguagem de programação orientada a objetos desenvolvida pela Sun Microsystems que permite desenvolver aplicações É uma linguagem que permite as que as suas aplicações sejam portáveis pois o código compilado é independente da máquina que o executa.

Este código é independente pois é executado numa máquina virtual (JVM) não necessitando assim de compilação para o código nativo da máquina.

Atualmente é das linguagens de programação mais utilizada. Esta linguagem foi escolhida para correr os algoritmos de classificação [Java].

2.1.2 NetBeans IDE

O NetBeans IDE permite o desenvolvimento rápido e fácil de aplicações de diversas linguagens. É uma ferramenta gratuita e tem uma grande comunidade de utilizadores em todo o mundo. A escolha deste IDE deveu-se a ser a ferramenta utilizada no ISEP para desenvolver aplicações de Java e Android [Netbeans].

2.2 Data mining

Data mining consiste na procura de padrões em grandes quantidades de dados. *Data mining* utiliza algoritmos matemáticos sofisticados para segmentar os dados e avaliar a probabilidade de eventos futuros. As principais propriedades são o descobrimento automático de padrões, a previsão de outputs, a criação de informação sobre a qual podem ser criadas ações e o foco em grandes quantidades de dados e base de dados [*Data mining*].

2.2.1 Weka

O Weka é um produto desenvolvido pela Universidade de Waikato na Nova Zelândia [Witten, 2005]. Este software é desenvolvido em Java e permite analisar *datasets* com algoritmos de aprendizagem.

O Weka contem ferramentas que permitem o pré-processamento, classificação, regressão, *clustering*, associação de regras, visualização e criação de novos esquemas de aprendizagem.

Figura 2 – Menu inicial do Weka

O Weka pode ser usado diretamente no conjunto de dados, por meio de uma interface gráfica (ver Figura 2), ou por uma API que pode ser chamada de programas em Java.

Nas experiências realizadas no âmbito deste trabalho, usou-se num programa em Java as classes da API do Weka.

Normalmente o tipo de ficheiros interpretado pelo Weka é ARFF. Contudo o Weka também suporta outros tipos de ficheiros como *input*. Neste caso, foram utilizados ficheiros do tipo CSV [Witten, 2005].

3 Data mining

Data mining, ou mineração de dados, consiste na descoberta de padrões em grandes conjuntos de dados com o objetivo de extrair informação útil.

A ferramenta Weka contém vários algoritmos cujo objetivo é encontrar um padrão nos dados recolhidos. Conforme será possível verificar nos capítulos seguintes, a classificação de sons cardíacos tentando definir alguns padrões é o grande objetivo deste trabalho. Para que tal seja possível foram utilizados diferentes algoritmos de classificação disponibilizados pelo Weka.

3.1 Técnicas de classificação

Nesta secção iremos abordar alguns dos algoritmos de classificação utilizados nas experiências realizadas no âmbito do trabalho que se descreve nesta dissertação.

3.1.1 Árvores de Decisão

As árvores de decisão permitem representar conhecimento e são um dos modelos mais usados em inferência indutiva. Este tipo de algoritmos permitem aprendizagem, ou seja através da construção da árvore é possível criar novos casos a fim de os classificar baseados nas análises anteriormente realizadas [Árvores de decisão].

3.1.1.1 J48

J48 é um dos algoritmos baseados em árvores de decisão. Para classificar um novo item, é necessário criar uma árvore de decisão baseada nos dados dos valores de amostra já disponíveis. Com os valores de amostra é possível identificar os atributos que permitem diferenciar as instâncias. Este algoritmo possibilita obter um maior ganho de informação. Se existir algum valor que não seja ambíguo, ou seja, tem bem definido a categoria a que pertencem e possui o mesmo valor que a variável desejada, termina a análise deste ramo da árvore e é atribuído ao ramo o valor obtido. Para os restantes casos são considerados outros atributos. No decorrer do processo ou obtemos uma decisão clara com a combinação dos atributos dados ou esgotamos os mesmos. Quer os atributos acabem ou não seja obtido um resultado ambíguo, é atribuído àquele ramo o valor que a maioria dos resultados que aquele ramo tem [*J48*].

Através do processo descrito anteriormente é possível obter uma árvore de decisão. Com os valores definidos na mesma é possível prever ou determinar o valor final de uma nova instância [Árvores de decisão2].

3.1.2 Random Forest

Random Forest é uma técnica eficiente que opera rapidamente sobre base de dados. É um conjunto de árvores de classificação ou regressão criadas através de amostras de dados de treino e seleção de recursos aleatória na indução de árvores.

Este algoritmo inicialmente desenha n número de árvores exemplo derivadas dos dados originais. Para cada exemplo, tenta escolher a melhor precisão possível de todos os dados. O range do erro pode ser obtido baseado nos dados já adquiridos pelo sistema.

Este algoritmo produz duas informações adicionais: a importância da medida das variáveis previsíveis e a medida da estrutura interna dos dados. Este algoritmo estima a importância das variáveis observando a previsão do crescimento do erro quando existe troca dos dados [*Random Forest*].

3.1.3 Rotation Forest

O *Rotation Forest* é um método que para cada árvore de decisão já existente ou considerada de treino, usa um diferente extrato de atributos para cada árvore. Este algoritmo baseia-se na construção de classificadores diversos e precisos. A principal heurística é aplicar métodos de extração e construir uma funcionalidade para cada classificador do conjunto. Como podemos verificar, este algoritmo utiliza também as árvores de decisão para obter resultados [*Rotation Forest*].

3.1.4 Support Vector Machines

Support Vector Machines são métodos supervisionados de aprendizagem utilizados quer para classificação quer para regressão. Este método permite aplicar técnicas lineares de classificação em dados não lineares. Permite também a existência de equações de Kernel que podem ser lineares, quadráticas, Gaussian ou qualquer outra técnica que permita atingir um determinado objetivo. Uma vez que os dados estejam divididos em duas categorias, o objetivo é obter um super plano que permite separar os dois tipos de instâncias. O super plano permite decidir o objetivo para futuras análises. Todos os dados utilizados neste método devem de ser binários. Caso os dados não sejam binários, este método interpreta-os como tal e completa a serie com dados binários [SVM].

3.1.4.1 SMO

SMO ou sequência mínima de otimização é um algoritmo para resolver o problema da programação quadrática derivada de otimizações matemáticas. Este problema surge durante as análises de *support vector machines*. Este algoritmo foi publicado em 1998.

3.1.5 *Multilayer Perceptron*

O *Multilayer Perceptron* é um algoritmo utilizado para a construção de redes neuronais. Esta rede neuronal pode ter uma ou diversas camadas entre a camada de introdução de dados e a camada de apresentação de resultados. A análise é efetuada sempre da introdução para a saída de resultados. Este tipo de rede é obtida e exercitada através do algoritmo de propagação de aprendizagem. O *Multilayer Perceptron* permite resolver problemas que não são linearmente separáveis [MLP].

3.2 *Descoberta de motifs em séries temporais*

Os *motifs* são padrões frequentes que se verificam nas séries temporais. Assim, um *motif* numa série temporal é uma subsequência repetida frequentemente.

Neste trabalho, foi utilizado o algoritmo *Multiresolution Motif Discovery* (*MrMotif*) para séries temporais [Castro, 2010b] para encontrar os padrões mais frequentes.

O *MrMotif* tem como principais características:

- Baseia-se na metodologia *iSAX* para discretizar os sinais contínuos. O *iSAX* é uma generalização do *SAX* (*Symbolic Aggregate Approximation*) que permite a indexação e mineração de grandes conjuntos de dados [Shieh and Keogh, 2008].
- Procura padrões nas sequências discretas resultantes
- Encontra os *motifs* mais frequentes (Top-*K*)
- Trata-se de código *open-source* em Java

O *SAX* divide a série temporal em *frames*, em que cada *frame* corresponde a um símbolo que corresponde ao seu valor médio. Ou seja, transforma uma série temporal numa sequência discreta de símbolos. Representa-se simbolicamente como *SAX*(T, w, a), onde T representa o tempo real de comprimento n numa sequência e o W representa o tamanho da palavra que irá ser transformada em padrões. O número de símbolos a é o tamanho do alfabeto ou resolução. Na Figura 3, temos um exemplo de uma palavra de tamanho 8 com resolução 4 (alínea a) e com resolução 16 (alínea b).

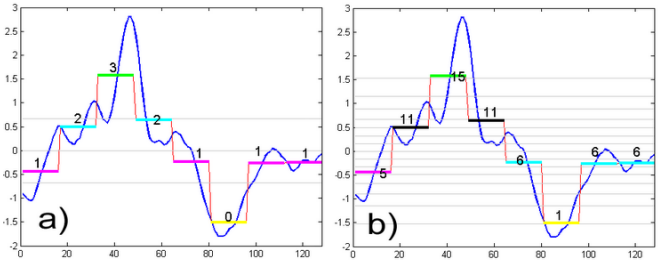

Figure 3: iSAX conversion process for time series X using w=8 and a) a=4; b) a=16 (code provided by SAX authors).

Figura 3 – Exemplo de discretização *SAX*, sendo a) a=4 e b) a=8 [Castro, 2010b]

Na Figura 4 é possível observar como é realizada uma discretização utilizando a ferramenta *iMotifs* [Castro, 2010b]. O tamanho do alfabeto é denominado de resolução.
O *iMotifs* (*Interactive Time Series Motif Discovery and Visualization Tool*) é uma versão visual e interativa do *MrMotif*.

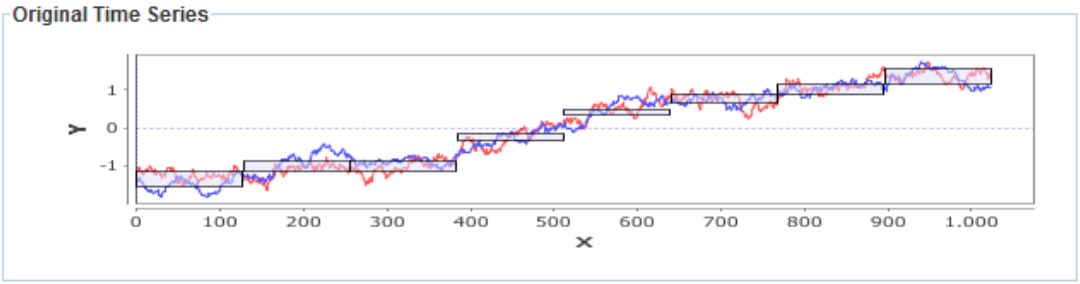

Figura 4 – Discretização Sax de segmentos temporais (*iMotifs* [Castro, 2010b])

Em suma, dada uma base de dados de séries temporais *D*, para um tamanho de *motif m* e um parâmetro *K*, por cada resolução (g_{min},g_{min} x 2,…, g_{max}), o *MrMotif* encontra os top-*K motifs*. Para além de devolver as top-*K motifs* mais frequentes, o *MrMotif* utilizado devolve também um *dataset* em que cada linha corresponde a uma série temporal e em a frequência que cada top-*K motif* é um atributo série temporal correspondente. Este *dataset* gerado pelo *MrMotif* é usado para a classificação [Gomes, 2013b].

Na Tabela 1, apresenta-se um pequeno excerto de um *dataset* gerado pelo *MrMotif*, neste caso para resolução 4 e para os 40 *motifs* mais frequentes, para as classes Normal (N) e Anormal (A).

Tabela 1 – Exemplo de *dataset* gerado pelo *MrMotif*

P1R4	P2R4	P3R4	…	P40R4	Classe
11	5	4	…	1	N
0	0	5	…	0	A
7	4	1	…	1	N
14	6	3	…	0	N
6	7	5	…	0	N
6	5	4	…	0	N
13	10	9	…	0	N

4 Classificação de sons cardíacos

Neste capítulo, precede-se à definição do problema e dos seus objetivos (secção 4.1).

Nas secções seguintes, faz-se uma revisão bibliográfica dos trabalhos que têm sido desenvolvidos por vários autores e os métodos e abordagens seguidas a nível do pré-processamento, da segmentação e da classificação de sons cardíacos gravados em ambiente real.

Faz-se também uma revisão do trabalho já desenvolvido no âmbito do projeto em que este trabalho se insere.

4.1 Definição do problema

Neste trabalho, pretende-se detetar patologias analisando os sons dos batimentos cardíacos em conjunto com informação clínica disponível. Temos um conjunto de dados acústicos obtidos em hospital complementados com dados clínicos dos pacientes.

O objetivo deste trabalho é produzir uma metodologia capaz de classificar sons cardíacos, para rastreio de patologias cardíacas de primeiro nível, em ambiente hospitalar. Pretende-se também comparar esta abordagem hibrida com abordagens anteriores baseadas apenas em dados acústicos ou apenas em dados clínicos.

Para ajudar a compreender o problema faremos uma breve descrição dos sons cardíacos. Na Figura 5 apresenta-se um esquema do coração humano, com os principais sons (lub, dub) do coração e os correspondentes S1 e S2. O S1 (ou lub) dá-se no início da sístole ventricular e o S2 (ou dub), no início da diástole. O coração produz outros sons (S3 e S4), frequentemente relacionados com um coração com patologia [Karnath, 2010].

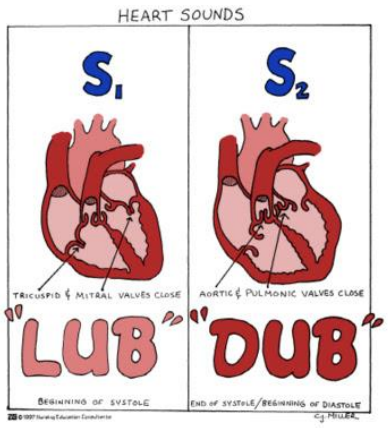

Figura 5 – Os principais sons cardíacos [HeartSounds]

Um som cardíaco normal é definido pelo claro som do tipo “lub dub lub dub”, em que o tempo “lub dub” é menor do que “dub lub” e o ritmo cardíaco varia entre os 60 e os 100 batimentos por minuto, em adultos. Neste caso, trata-se de sons de batimentos cardíacos obtidos em crianças, os batimentos variam entre os 40 e os 140 batimentos por minuto.

4.2 Revisão bibliográfica

O trabalho descrito nesta dissertação insere-se num projeto que tem vindo a ser desenvolvido. Nesta secção, descrevem-se as diferentes abordagens e resultados obtidos.

4.2.1 Recolha e Organização dos dados

A gravação de sons em ambientes reais acarreta o problema de incluir, junto com os batimentos cardíacos, ruídos de várias origens. Para além disso, é comum acontecer as gravações terem durações diferentes, muitas delas muito longas. Estas questões originam, entre outros problemas, um grande esforço computacional. Assim, uma técnica utilizada, relatada na literatura, é proceder ao corte dos sons, isolando em gravações mais curtas os períodos de interesse.

O projeto em que este trabalho se insere, teve início na participação no concurso “PASCAL Classifying Heart Sounds Challenge” [Bentley, 2011] com o objetivo de identificar patologias nos sons cardíacos. Estes sons, divididos em dois *dataset*s, foram recolhidos utilizando *smartphones* e um estetoscópio digital respetivamente [Gomes, 2012]. A página do concurso contém a informação de que os sons são de comprimentos variados (entre 1 segundo e 30 segundos) e que alguns foram cortados para reduzir o ruído excessivo e fornecer o fragmento saliente do som. Não referem, porém a ferramenta utilizada [Bentley, 2011].

Uma das ferramentas possíveis para esta tarefa é a Audacity, de acesso gratuito [Audacity]. Esta ferramenta encontra-se descrita na secção 2.1.

Encontra-se ainda na literatura, a utilização do Audacity para gravar os sons como um conjunto de dados. Este software permite gravar o som do coração e exibi-lo quando necessário [Arathy, 2013].

Posteriormente será descrito como foi efetuado o processo de tratamento de som com esta ferramenta no âmbito desta dissertação.

4.2.2 Pré-processamento

O processo de análise e classificações inicia-se, usualmente, pelo processo de pré-processamento dos sons.

No pré-processamento, é comum proceder-se à redução do ruído existente no sinal recorrendo a aplicação de filtros ao sinal e, posteriormente, encurtar o comprimento dos sinais mantendo, no entanto, a sua morfologia geral (decimação).

A redução do ruído assume grande importância no caso de sons gravados em ambiente clinico pois é comum encontrar, juntamente com a gravação dos sons cardíacos, e para além de ruídos pulmonares, o ruído do microfone a roçar na roupa do paciente e, inclusive, vozes.

A necessidade de encurtar o comprimento dos sinais prende-se com o facto de usualmente se tratar de registos longos, obrigando a elevados recursos computacionais.

No trabalho que aqui se descreve, tal como nos trabalhos desenvolvidos anteriormente, os ficheiros de áudio foram pré-processados.

No pré-processamento utilizaram-se as técnicas de decimação e um filtro passa-banda Chebyshev tipo I de 5ª ordem com frequência de corte inferior de 100Hz e frequência de corte superior de 882 Hz. O sinal foi normalizado relativamente ao valor máximo absoluto.

Após aplicação das técnicas descritas, foi calculado o envelope Shannon, seguindo o cálculo da média da energia Shannon realizado ao longo de uma janela contínua de 0.02 segundos com 0.01 segundo de sobreposição [Liang et al., 1997], [Gomes, 2012], [Gomes 2014], [Oliveira 2014], [Oliveira 2014b]. Esta técnica foi proposta por Liang [Liang et al., 1997], com o objetivo de atenuar o ruído e detetar mais facilmente os sons de baixa intensidade, os batimentos cardíacos.

Esta técnica de pré-processamento descrita previamente têm vindo a ser utilizada em diversos trabalhos anteriores inseridos neste projeto [Gomes, 2012], [Gomes 2013a], [Gomes 2013b], [Gomes 2014], [Oliveira 2014], [Oliveira 2014b] e, como foi referido, também foi utilizada no trabalho que aqui se descreve.

4.2.3 Segmentação e deteção de picos

Diversas abordagens de segmentação de sinais têm sido relatadas na literatura. A maioria explora sinais de eletrocardiograma (ECG) ou dados de pulso da carótida. Por exemplo, Groch apresenta uma solução, onde a segmentação se baseia nas características de sinal no domínio do tempo [Groch et al., 1992].

Por sua vez, Strunic utilizou estratégias de extração de sinal para reduzir anomalias e definir um *threshold* na amplitude de modo a selecionar os pontos e efetuar a segmentação [Strunic et al., 2007].

Marques et al. utilizaram Wavelets na segmentação [Marques, 2013] [Babaei, 2009]. Marques compara diversas técnicas com a energia de Shannon, aplicadas ao sinal. Entre essas técnicas constam, entre outras, as Wavetlets e Transformadas de Fourier. Lijuan, também usa

Wavelets para, juntamente com a energia normalizada de Shannon, extrair atributos para a classificação [Lijuan, 2012].

Em 2012, Gomes et.al aplicaram, após o pré-processamento, a segmentação aos sons cardíacos (S1 e S2) antes da fase de classificação [Gomes, 2012]. Na segmentação, o sinal do som cardíaco é dividido em segmentos diferentes: a sístole (S1) e a diástole (S2). Esta segmentação, seguida de um algoritmo de deteção de picos no envelope (S1 e S2), permitiu vencer o concurso "PASCAL Classifying Heart Sounds Challenge" [Bentley, 2011]. O principal desafio do referido concurso, consistia em identificar patologias cardíacas através de sons captados através de um estetoscópio digital e através de *smartphones* (fonocardiograma) [Gomes, 2012]. A abordagem seguida baseava-se nas distâncias entre S1 e S2, considerando S2 maior que S1, para batimentos cardíacos normais [Kumar, 2006], [Gupta, 2007].

4.2.4 Classificação

Os trabalhos descritos na literatura utilizam diversos algoritmos de classificação.

Strunic usou *Artificial Neural Network* (ANN) para classificar os batimentos cardíacos em Normal, Murmur sistólico causado pela regurgitação mitral (MR), sopro sistólico causada por aórtica Estenose (AS) e diástole Murmur causada por aórtica a regurgitação (AR) [Strunic et al., 2007]. A precisão obtida nos seus resultados caiu quando foram utilizados os dados recolhidos por um estetoscópio eletrônico com uma duração de cerca de 5 segundos. Encontrou problemas ao lidar com arquivos de áudio em grande parte, de diferentes comprimentos. Com arquivos de áudio selecionados de comprimento semelhante, o sistema ainda se mostrava mais incapaz de diferenciar os batimentos cardíacos do ruido de fundo, na maioria dos casos.

Também Karraz utilizou ANN. Extraiu o QRS (ondas Q, R e S do ECG) dos sinais e utilizou-os com atributos para construir a rede neuronal. Posteriormente classificou os sinais usando ferramenta Bayesiana [Karraz, 2006]. Babaei, também usou ANN [Babaei, 2009] assim como Ölmez [Ölmez, 2013].

Mandeep usou o classificador Naïve Bayes para classificar sons cardíacos (PCG) gravados de um estereoscópio [Mandeep, 2013].

Kampouraki e Kao utilizaram *Support Vector Machines* (SVMs). Kampouraki, para classificar ecocardiogramas (ECG) [Kampouraki, 2009] e Kao, para classificar sinais PCG, após segmentação [Kao, 2011]. Tanto ECG como a simulação de sons são muito diferentes de dados captados em ambiente real, em que os sons apresentam uma duração variável e contêm ruído.

Wang et.al. propõem um novo método para identificação automática de sons normais e com patologia. Após o pré-processamento dos sons, aplicam o *optimum multi-scale wavelet packet decomposition* (OMS-WPD) para extração de atributos e utilizam SVMs [Wang, 2014].

4.3 Abordagens anteriores

Nesta secção, faz-se uma revisão dos trabalhos publicados no âmbito do projeto onde o presente trabalho se enquadra.

Na secção 4.2.3, referimos alguns desses trabalhos. Num primeiro trabalho utilizou-se uma abordagem de segmentação dos sons, usando energia de Shannon e deteção de picos (S1 e S2) correspondentes à diástole e sístole do coração. Os atributos foram extraídos a partir de medidas de distâncias e estatísticas relacionadas com S1 e S2 [Gomes, 2012].

Numa fase posterior, partindo do mesmo pré-processamento dos sons e do envelope de Shannon, passou a fazer-se a extração de atributos usando o programa *Multiresolution Motif Discovery* (*MrMotif*) [Gomes, 2013b] para detetar os *motifs* mais frequentes. Como já referimos na secção 3.2, um *motif* (numa série temporal) é um padrão repetido frequentemente. O *MrMotif* permite encontrar estes padrões mais comuns, em séries temporais [Castro, 2010b].

Se pensarmos num registo áudio do som cardíaco como uma série temporal, diferentes tipos de som devem corresponder a diferentes *motifs*. Assim, aplicamos o *MrMotif* aos sons cardíacos para extrair os *motifs* que são utilizados como os atributos para classificação [Gomes, 2013], [Oliveira, 2014].

Na Figura 6 apresenta-se um exemplo de um som cardíaco (primeira imagem), do correspondente envelope de Shannon (segunda imagem) e de deteção de padrões no sinal, a azul e vermelho (terceira imagem).

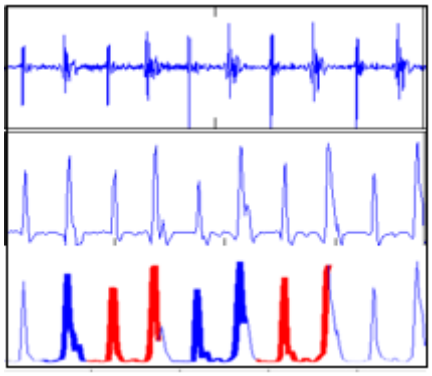

Figura 6 – Imagem 1: Sinal original; Imagem 2: Envelope do sinal; Imagem 3: Exemplo de padrões alternados a azul e vermelho ao longo da serie temporal (X -Tempo / Y-Amplitude).

Esta abordagem de tratar os sons cardíacos como séries temporais, tanto quanto temos conhecimento, não é usual. A frequência dos *motifs* nas séries temporais são os atributos usados na classificação.

Utilizando estes atributos, em Gomes et.al, faz-se uma comparação dos algoritmos de classificação. Utilizando a ferramenta Weka (secção 2.2.1) comparam-se os resultados obtidos, em termos de precisão média, com os algoritmos *Decision Trees*, *Logistic Regression*, *Rotation*

Forest e *Random Forest* [Gomes, 2013a]. O *Random Forest* [Breiman, 2001] obtém os melhores resultados neste tipo de abordagem baseada em *motifs*.

Em Oliveira et.al. é apresentado um algoritmo que combina a informação sobre os *motifs* mais frequentes com as características conhecidas dos sons cardíacos, em particular, do S1 e S2 [Oliveira, 2014].

Numa outra fase, utilizaram-se técnicas de *text mining*. Estendendo a analogia do *text mining,* foi utilizado o modelo TFIDF (*term frequency, inverse document frequency*). Dado um *motif m* e um som *s* de uma coleção de sons, TFIDF(*m*; *s*), é calculada multiplicando TF (termo relativo de frequência, onde a frequência de *m* é dividida pelo numero de sons, *N*) por IDF(*m*;*s*).

$$TFIDF(m,s) = TF(m,s).log\frac{\#\{s \in Sounds\}}{\#\{s \in Sounds : m \in s\}}$$

Outra analogia realizada com *text mining* foi a ideia de explorar ocorrências sequenciais de termos em documentos (*n-grams*). Os *n-grams* mais frequentes correspondem a potenciais composições de termos relevantes. No trabalho [Gomes, 2014] foram utilizados *bigrams* (onde *n* é igual a 2). Por exemplo, “New York", é um *bigram* relevante que pode ser automaticamente identificado dos textos onde ocorre com frequência. Neste trabalho, avalia-se a capacidade discriminante destas técnicas na separação de classes (Normal e com Murmúrio cardíaco). É também feito um estudo da possibilidade de redução de custos. O classificador utilizado foi o *Random Forest* [Gomes, 2014].

5 Modelação

5.1 Preparação dos dados

Os dados analisados foram obtidos através de gravação de auscultações efetuadas com o *DigiScope Collector* (Figura 7), em ambiente hospitalar [Pereira,2011].

Figura 7 – Sistema *DigiScope Collector*

Os ficheiros de som obtidos têm a duração média de um minuto e com muito ruído. Para realizar este tipo de análises, o ideal é que a duração média dos ficheiros seja entre os 6 a 10 segundos. Esta duração foi utilizada em trabalhos anteriores ao projeto em que este trabalho se insere, daí a escolha destes valores. Foi decidido manter os mesmos padrões de forma a ser mais fácil a comparação de resultados entre estudos e análises anteriores e esta [Gomes, 2013b]. Assim, foi necessário recorrer a uma ferramenta de áudio (Audacity ver secção 2.1) para cortar os sons. Como os ficheiros originais eram de tamanhos bastante diversos, com muito ruído e de diferentes naturezas (vozes, por exemplo), a primeira parte do trabalho foi criar ficheiros mais pequenos com excertos de cada um dos ficheiros de som, onde os batimentos cardíacos eram mais percetíveis. Com este processo foi possível agrupar para cada ficheiro de som os seus melhores excertos.

Para aumentar a qualidade dos dados, escolheu-se um dos excertos de cada ficheiro de som. O critério de escolha usado nesta filtragem foi utilizar o ficheiro que continha mais dados relativamente aos batimentos e com menos ruido. A média de duração destes excertos foi de 8 segundos.

Este procedimento de cortar os sons permite-nos focar no problema de classificação de batimentos cardíacos. Esta técnica já foi utilizada por outros autores, como foi referido na secção 4.2.1.

5.2 Dataset

O *dataset* utilizado foi disponibilizado pela unidade de Cardiologia Unidade Materno-Fetal do Hospital Real Português no Recife, Brasil. Todos os dados recolhidos pertencem a crianças de diferentes faixas etárias sendo que a média de idades é de 7,4 anos. Este *dataset* contém informações cardíacas e clínicas sobre 102 rapazes e 56 raparigas e é composto por 158 gravações de sons cardíacos, divididos em duas classes: Normal e Anormal.

Este *dataset*, com 123 ficheiros da classe Normal (78%) e 35 da classe Anormal (22%) é bastante desequilibrado e requer algum cuidado na análise dos resultados. Um conjunto de dados é considerado desequilibrado se as categorias de classificação não são igualmente representadas.

5.3 Avaliação

Como foi referido anteriormente, o *dataset* utilizado só possuí duas classes para avaliação: classe normal (N) e classe anormal (A). Nesta secção irão ser abordadas as medidas de desempenho e análise do espaço ROC para duas classes.

A matriz confusão, que pode ser aplicada sempre que estamos a analisar um problema com duas classes, é muito útil visto que apresenta o número de predições corretas e incorretas para cada classe (Figura 8).

Na análise do problema dos sons cardíacos consideramos como classe positiva a classe anormal (A) e a negativa a classe normal (N). Esta decisão foi tomada pois pretendia-se verificar se os doentes possuem algum tipo de patologia cardíaca e não se são saudáveis. De seguida é apresentada a matriz confusão utilizada.

		Classe negativa (N)	
		+	-
Classe positiva (A)	+	VP	FN
	-	FP	VN

Figura 8 – Matriz confusão

Na matriz encontram-se quatro designações importantes que serão utilizadas em algumas análises seguintes. O VP corresponde ao número de verdadeiros positivos, ou seja, corresponde ao número de exemplos classificados como classe anormal, neste caso, e que estão bem classificados. O FP corresponde ao número de falsos positivos e indica todos os elementos que foram classificados como sendo elementos da classe anormal mas que pertencem à classe normal. O VN é o número de verdadeiros negativos, que corresponde a todos os valores que foram classificados como pertencentes à classe normal e essa classificação verificou-se correta. O FN é o número de falsos positivos, este valor corresponde a todos os

valores catalogados como pertencentes à classe normal mas que pertencem à classe anormal. Com base na matriz confusão é possível calcular uma série de medidas de desempenho.

Na secção seguinte iremos referir algumas medidas de desempenho que utilizaremos para avaliar os resultados obtidos. Abordaremos também a análise de curvas ROC.

5.3.1 Medidas de desempenho

A taxa de falsos negativos (TFN) indica a proporção de casos que foram classificados como anormais e são normais.

$$TFN = \frac{FN}{VP+FN}$$

A taxa de falsos positivos (TFP) representa a proporção de casos classificados como normais e neste caso deveriam ser anormais.

$$TFP = \frac{FP}{FP + VN}$$

A taxa de erro total é obtida pela seguinte fórmula:

$$Erro\ total = \frac{FP + FN}{FP + VP + FN + VN}$$

A *accuracy* (taxa de acerto) resulta da soma dos elementos da diagonal principal (da matriz de confusão) dividida pela soma de todos os elementos da matriz.

$$Accuracy = \frac{VP + VN}{VP + FN + VN + FP}$$

A precisão indica os valores calculados corretamente para a classe verdadeira, neste caso, a classe anormal.

$$Precisão = \frac{VP}{VP + FP}$$

O *Recall* indica a taxa de acerto na classe positiva.

$$Recall = \frac{VP}{VP + FN}$$

No contexto da média ponderada é obtido através da seguinte fórmula:

$$Recall = \sum_{i=1}^{K} w_i \frac{Aii}{\sum_{j=1}^{K} Aij}$$

Nesta fórmula o *wi*, representa a proporção dos exemplos da respetiva classe, o *A* representa a matriz de contingência e o *Aij* é o número de instâncias da classe *i* que estão

classificadas como *j*. Isto significa que *i* e *j* representam classes de classificação (anormal e normal).

Como, na avaliação de resultados, vamos ter em conta a média ponderada por classe, neste caso, o valor do *Recall* corresponde à *Accuracy*.

A taxa de verdadeiros positivos (TVP) pode ser também denominada por sensibilidade. Esta análise permite saber a taxa de acerto existente na classe positiva.

$$TVP = \frac{VP}{VP + FN}$$

A taxa de falsos positivos (TFP) corresponde à taxa de acerto na classe negativa.

$$TFP = \frac{VN}{VN + FP}$$

Das análises de valores referidos, a precisão e a taxa de verdadeiros positivos permitem retirar conclusões mais reais do modelo. Se por um lado a precisão é uma medida de exatidão do modelo, a taxa de verdadeiros positivos é uma medida de sua completude.

Dos valores de análise descritos, alguns são valores complementares e que não podem ser analisados isoladamente. Um desses exemplos é a precisão. A precisão só nos permite verificar uma das classes e não as duas. Por isso, temos a medida F (*F-Measure*) que é a média harmónica ponderada da precisão com a taxa de verdadeiros positivos [Gama,2012].

5.3.2 Análise ROC

A análise das curvas ROC é uma alternativa de análise a classificadores em problemas binários, como é o caso do trabalho que se descreve nesta dissertação. O gráfico ROC é um gráfico bidimensional num espaço denominado ROC, com dois eixos (X e Y). Cada um dos eixos representa as medidas de TFP e TVP.

A área abaixo da curva ROC designa-se por AUC (*Area Under ROC Curve*) e varia entre 0 e 1. Valores mais próximos de 1 são considerados melhores [Gama, 2015].

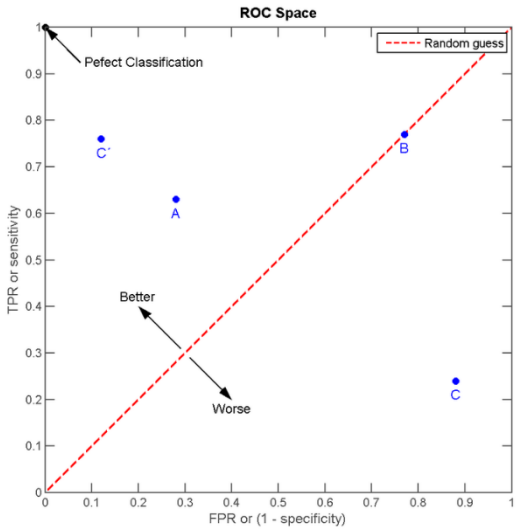

Figura 9 – Espaço ROC com três classificadores [Roc curve image]

Na Figura 9 encontram-se definidos os principais indicadores e aspetos que são relevantes nas curvas ROC. Os classificadores que realizam predições aleatórias estão representados pela linha diagonal. Qualquer classificador que se encontre abaixo dessa linha é considerado pior que o aleatório logo irrelevante. Neste gráfico, foram usadas as taxas de verdadeiros positivos e de falsos positivos (TPR e FPR respetivamente). Se a classificação realizada fosse perfeita, isto significava que todos os elementos seriam classificados corretamente, independentemente de serem positivos ou negativos, dizendo-se então que a avaliação atingiu o céu ROC. Caso todas as classificações estejam erradas é atingido o inferno ROC. Caso os classificadores se aproximem do ponto (0,0) as classificações são consideradas negativas e no caso dos classificadores se aproximarem do ponto (1,1) são consideradas classificações positivas.

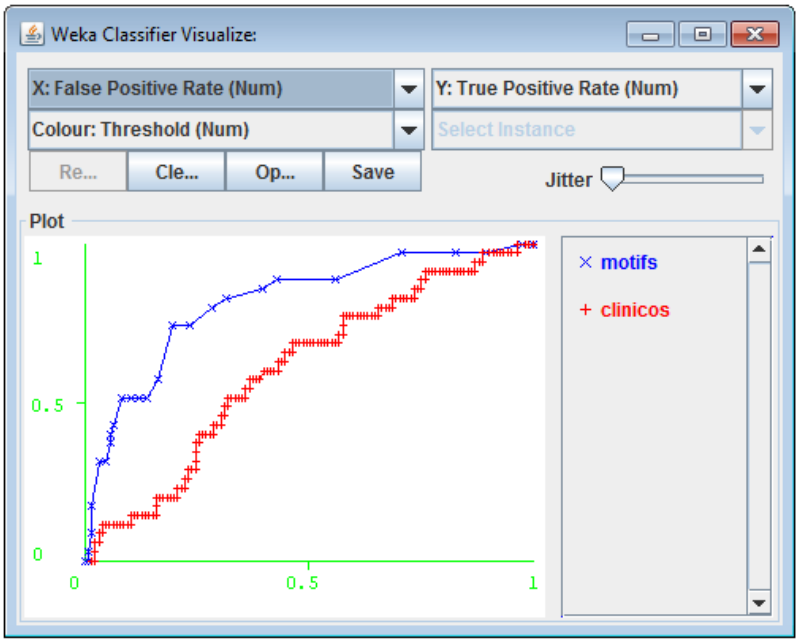

Figura 10 – Exemplo de curvas ROC

Na Figura 10 é possível observar duas curvas obtidas por dois modelos. Nos casos sem interceção, a curva que estiver mais próxima do ponto (0,1) corresponde ao melhor desempenho. Quando existe intersecção entre as curvas, significa que a escolha do melhor modelo será diferente consoante a região da curva em análise. Neste caso, pode comparar-se o desempenho de cada modelo pela área abaixo da curva (AUC).

5.3.3 Cross-validation

Dado um *dataset* e um algoritmo de classificação, as medidas acima descritas podem ser estimadas utilizando a metodologia de *cross-validation* (validação cruzada). Esta avaliação normalmente é realizada pela análise de desempenho em novos exemplos, ou seja, não são utilizados exemplos do conjunto de treino.

No método de avaliação cruzada *r-fold cross-validation*, o conjunto de exemplos é dividido em *r* subconjuntos do mesmo tamanho. As r-1 partições são utilizadas no treino de um algoritmo que é depois testado na partição restante. O processo é repetido *r* vezes utilizando, em cada ciclo, uma partição diferente para teste. O desempenho é avaliado pela média dos desempenhos observada em cada conjunto de teste [Gama, 2015].

No caso do trabalho que aqui se descreve, foi feita uma avaliação 10-*fold cross-validation*.

5.4 Classificação utilizando motifs

Nesta secção descrevemos a metodologia utilizada nos modelos de classificação, usando a abordagem dos *motifs*. Faz-se também uma análise dos resultados obtidos utilizando as diferentes medidas de avaliação.

5.4.1 Aplicação dos algoritmos de classificação

A metodologia seguida nas experiências realizadas segue os seguintes passos [Gomes et al., 2014]

1. Aplicar ao *dataset* original o pré-processamento. No pré-processamento, os sons cardíacos (referidos na secção 5.1 e 5.2) são filtrados, decimados e é calculado o envelope de energia de Shannon.

2. Aplicar o *MrMotif* ao *dataset* resultante do passo 1. Neste passo, é necessário escolher os valores para os parâmetros do *MrMotif*. Nas experiências realizadas, experimentamos diversas combinações dos parâmetros.
   O *MrMotif* recebe como parâmetros a o top-*K* (número *K* de *motifs* selecionados, de 4 a 64, que podem ser considerados), a resolução (número de *motifs* relevantes) *R*, o tamanho dos *motifs*, *W*, e o *overlap* das janelas de discretização. Tem-se ainda o *w*, tamanho da palavra, que é o número de símbolos utilizados na palavra resultante do *iSAX*.
   Ao longo das nossas experiências fizemos variar os parâmetros top-*K* (10, 20, 30 e 40), a resolução *R* e o tamanho dos *motifs* (4,20,30,40,50 e 64). O *overlap* foi mantido a 10 e o tamanho da palavra a 8.

3. O *MrMotif* gera um novo *dataset* em que cada linha corresponde à linha do *dataset* original e que contém, nas colunas, a frequência dos *motifs* mais relevantes na série temporal (os atributos). Tem-se, assim, como resultado do *MrMotif* uma tabela tabela***xRxKxW***.csv, que contém o *dataset* dos atributos e cujo nome fornece a informação sobre os parâmetros utilizados (resolução, top-*K* e tamanho dos *motifs*).

4. Aplicar um algoritmo de classificação ao *dataset* resultante do passo 3 e estimar a capacidade preditiva do modelo.

Foram aplicados os algoritmos de classificação *J48*, *MultiLayer Perceptron*, *Random Forest*, *Rotation Forest* e *SMO* com diferentes parâmetros. Foi utilizada validação cruzada com 10 *folds*.

5.4.2 Análise de resultados

Como referimos na secção anterior, foram utilizados diversos algoritmos de classificação com diferentes parâmetros.

Na secção 5.3 estão descritas as medidas de avaliação utilizadas para comparar os classificadores: *Recall*, *F-Measure* e area ROC (AUC).

5.4.2.1 Recall

Nesta subsecção apresentam-se os resultados obtidos nos diferentes modelos utilizando como referência a medida *Recall* (média ponderada).

De forma a resumir os resultados obtidos, na Tabela 2 encontram-se os dez melhores resultados obtidos (média ponderada) utilizando o algoritmo *J48*. Na tabela consta também uma coluna que contém os valores utilizados para o parâmetro *leaf* deste classificador.

Tabela 2 – Melhores resultados do *Recall* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K4W40	*leaf*=1	**0.848**	0.830	0.612
tabela4K4W40	*leaf*=2	**0.848**	0.830	0.612
tabela8K4W10	*leaf*=10	0.829	0.828	0.660
tabela8K4W10	*leaf*=15	0.829	0.828	0.660
tabela8K20W10	*leaf*=2	0.823	0.809	0.609
tabela4K20W30	*leaf*=10	0.823	0.805	0.620
tabela4K20W30	*leaf*=15	0.823	0.805	0.620
tabela16K64W10	*leaf*=2	0.816	0.813	0.641
tabela4K4W20	*leaf*=15	0.816	0.797	0.626

Como podemos observar o melhor resultado corresponde a um *Recall* de 0.848, para *K*=4 e *W*=40. Também podemos verificar que são obtidos os mesmos valores caso a *leaf* utilizada seja 1 ou 2.

Na Tabela 3, apresentam-se os melhores resultados obtidos utilizando o algoritmo *Random Forest* (média ponderada). Neste algoritmo de classificação fizemos variar os parâmetros *trees* e *features*.

Tabela 3 – Melhores resultados do *Recall* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K64W30	*trees*=25, *features*=1, *seed*=1	**0.861**	0.838	0.726
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.854	0.839	0.743
tabela4K40W40	*trees*=50, *features*=2, *seed*=1	0.854	0.833	0.732
tabela4K64W30	*trees*=30, *features*=1, *seed*=1	0.854	0.829	0.728
tabela4K64W30	*trees*=50, *features*=1, *seed*=1	0.854	0.829	0.700
tabela16K20W10	*trees*=25, *features*=2, *seed*=1	0.848	0.827	0.719
tabela4K30W30	*trees*=30, *features*=1, *seed*=1	0.848	0.833	0.764
tabela4K30W30	*trees*=40, *features*=2, *seed*=1	0.848	0.830	0.749
tabela4K30W30	*trees*=50, *features*=2, *seed*=1	0.848	0.830	0.750
tabela4K40W40	*trees*=30, *features*=2, *seed*=1	0.848	0.823	0.728
tabela4K40W30	*trees*=30, *features*=1, *seed*=1	0.848	0.830	0.748

O melhor resultado obtido foi um valor de *Recall* de 0.861. Este valor foi obtido utilizando um *K*=64, *W*=30 no *MrMotif*, 25 *trees*, o número de *features* e de *seed* igual a 1 no algoritmo *Random Forest*.

Na Tabela 4, apresentam-se os melhores resultados obtidos utilizando o algoritmo *Rotation Forest* (média ponderada).

Tabela 4 – Melhores resultados do *Recall* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=30	**0.861**	0.841	0.737
tabela4K40W30	*iterations*=40	0.854	0.836	0.723
tabela4K50W30	*iterations*=40	0.854	0.836	0.729
tabela4K50W30	*iterations*=30	0.854	0.836	0.739
tabela4K50W30	*iterations*=20	0.854	0.833	0.706
tabela4K20W30	*iterations*=30	0.848	0.827	0.725
tabela4K20W30	*iterations*=25	0.848	0.830	0.735
tabela4K30W30	*iterations*=25	0.848	0.830	0.711
tabela4K40W30	*iterations*=50	0.848	0.827	0.705
tabela4K40W40	*iterations*=30	0.848	0.827	0.699
tabela4K40W40	*iterations*=25	0.848	0.827	0.645

O melhor resultado obtido foi um valor de *Recall* de 0.861. Este valor foi obtido utilizando K=40, W=30 e 30 *iterations* como parâmetro.

Também se efetuaram experiências com o algoritmo *SMO* mas os resultados obtidos foram irrelevantes.

Na Tabela 5, apresentam-se os melhores resultados obtidos utilizando o algoritmo *MultiLayer Perceptron* (média ponderada).

Tabela 5 – Melhores resultados do *Recall* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K20W30	*learning rate*=0.3	**0.829**	0.799	0.688
tabela8K4W20	*learning rate*=0.3	0.810	0.757	0.734
tabela4K4W10	*learning rate*=0.3	0.804	0.745	0.730
tabela4K4W30	*learning rate*=0.3	0.804	0.774	0.736
tabela4K4W40	*learning rate*=0.3	0.804	0.769	0.709
tabela4K30W30	*learning rate*=0.3	0.797	0.765	0.688
tabela8K4W10	*learning rate*=0.3	0.791	0.728	0.710
tabela4K20W40	*learning rate*=0.3	0.791	0.782	0.712
tabela4K64W30	*learning rate*=0.3	0.791	0.782	0.711

O melhor resultado obtido foi um valor de *Recall* de 0.829. Este valor foi obtido utilizando K=20, W=30 no *MrMotif* e um *learning rate* de 0.3 no MLP.

Comparando os melhores resultados de cada técnica de classificação obtemos os resultados apresentados na Tabela 6.

Tabela 6 – Resumo os melhores resultados de *Recall*

Algoritmo	***Dataset***	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
J48	tabela4K4W40	*leaf*=1	0.848	0.830	0.612
Random Forest	tabela4K64W30	*trees*=25, *features*=1, *seed*=1	**0.861**	0.838	0.726
Rotation Forest	tabela4K40W30	*iterations*=30	**0.861**	0.841	0.737
SMO	tabela32K50W40	*seed*=1	0.785	0.696	0.514
MultiLayer Perceptron	tabela4K20W30	*learning rate*=0.3	0.829	0.799	0.688

Na Tabela 6, é possível observar que, para medida de avaliação *Recall*, os melhores algoritmos de classificação são a *Random Forest* e a *Rotation Forest*, com um valor de 0.861. Ambos os resultados foram obtidos para o mesmo número de *motifs* relevantes mas o parâmetro *motifs* selecionados varia. No caso do *Random Forest* é igual a 64 e no *Rotation Forest* é igual a 40.

5.4.2.2 F-Measure

O *F-Measure* é uma medida combinada da *Precision* com a TVP (taxa de valores positivos), como vimos na secção 5.3.1 relativa a medidas de avaliação de modelos.

De forma a resumir os resultados obtidos, a informação foi agrupada por resultados de cada algoritmo por tabela. Na Tabela 7, encontram-se os dez melhores resultados obtidos pelo algoritmo *J48* referente à média ponderada.

Tabela 7 – Melhores resultados do *F-Measure* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K4W40	*leaf*=1	0.848	**0.830**	0.612
tabela4K4W40	*leaf*=2	0.848	**0.830**	0.612
tabela8K4W10	*leaf*=10	0.829	0.828	0.660
tabela8K4W10	*leaf*=15	0.829	0.828	0.660
tabela16K64W10	*leaf*=2	0.816	0.813	0.641
tabela16K4W10	*leaf*=1	0.810	0.810	0.634
tabela16K4W10	*leaf*=2	0.810	0.810	0.634
tabela16K4W10	*leaf*=10	0.810	0.810	0.634
tabela16K4W10	*leaf*=15	0.810	0.810	0.634
tabela16K20W10	*leaf*=10	0.810	0.810	0.634
tabela16K20W10	*leaf*=15	0.810	0.810	0.634

Como podemos observar o melhor resultado obtido segundo a medida F (ou *F-Measure*) foi 0.830. Como aconteceu no caso dos valores do *Recall*, os melhores resultados foram obtidos com K=4 e W=40 e caso a *leaf* utilizada seja 1 ou 2.

Na Tabela 8, apresentam-se os melhores resultados obtidos utilizando o algoritmo *Random Forest* (média ponderada).

Tabela 8 – Melhores resultados do *F-Measure* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.854	**0.839**	0.743
tabela4K64W30	*trees*=25, *features*=1, *seed*=1	0.861	0.838	0.726
tabela4K40W40	*trees*=50, *features*=2, *seed*=1	0.854	0.833	0.732
tabela4K30W30	*trees*=30, *features*=1, *seed*=1	0.848	0.833	0.764
tabela8K50W40	*trees*=50, *features*=1, *seed*=1	0.842	0.833	0.738
tabela4K30W30	*trees*=40, *features*=2, *seed*=1	0.848	0.830	0.749
tabela4K30W30	*trees*=50, *features*=2, *seed*=1	0.848	0.830	0.750
tabela4K40W30	*trees*=30, *features*=1, *seed*=1	0.848	0.830	0.748
tabela4K50W20	*trees*=40, *features*=2, *seed*=1	0.848	0.830	0.717
tabela4K50W20	*trees*=50, *features*=2, *seed*=1	0.848	0.830	0.741
tabela4K64W30	*trees*=30, *features*=1, *seed*=1	0.854	0.829	0.728

O melhor resultado obtido foi um valor de *F-Measure* de 0.839. Este valor foi obtido utilizando K=40, W=30, 20 *trees* e tanto o valor da *feature* como o valor de *seed* igualado a 1.

Na Tabela 9 apresentam-se os melhores resultados obtidos utilizando o algoritmo *Rotation Forest* (média ponderada).

Tabela 9 – Melhores resultados do *F-Measure* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=30	0.861	**0.841**	0.737
tabela4K40W30	*iterations*=40	0.854	0.836	0.723
tabela4K50W30	*iterations*=40	0.854	0.836	0.729
tabela4K50W30	*iterations*=30	0.854	0.836	0.739
tabela4K50W30	*iterations*=20	0.854	0.833	0.706
tabela4K20W30	*iterations*=25	0.848	0.830	0.735
tabela4K30W30	*iterations*=25	0.848	0.830	0.711
tabela4K50W30	*iterations*=50	0.848	0.830	0.692
tabela8K20W10	*iterations*=50	0.842	0.828	0.764
tabela4K20W30	*iterations*=30	0.848	0.827	0.725
tabela4K40W30	*iterations*=50	0.848	0.827	0.705

O melhor resultado obtido foi um valor de *F-Measure* de 0.841. Este valor foi obtido utilizando K=40, W=30 e 30 *iterations* como parâmetro.

Na Tabela 10 apresentam-se os melhores resultados obtidos utilizando o algoritmo *MultiLayer Perceptron* (média ponderada).

Tabela 10 – Melhores resultados do *F-Measure* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K20W30	*learning rate*=0.3	0.829	**0.799**	0.688
tabela4K20W40	*learning rate*=0.3	0.791	0.782	0.712
tabela4K64W30	*learning rate*=0.3	0.791	0.782	0.711
tabela4K4W30	*learning rate*=0.3	0.804	0.774	0.736
tabela4K4W40	*learning rate*=0.3	0.804	0.769	0.709
tabela4K30W30	*learning rate*=0.3	0.797	0.765	0.688
tabela8K4W20	*learning rate*=0.3	0.810	0.757	0.734
tabela4K4W10	*learning rate*=0.3	0.804	0.745	0.730
tabela8K4W10	*learning rate*=0.3	0.791	0.728	0.710

O melhor resultado obtido foi um valor de *F-Measure* de 0.799. Este valor foi obtido utilizando K=20, W=30 e um *learning rate* de 0.3.

Comparando os melhores resultados de cada modelo de classificação obtemos a seguinte tabela:

Tabela 11 – Resumo os melhores resultados de *F-Measure*

Algoritmo	*Dataset*	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	tabela4K4W40	*leaf*=1	0.848	0.830	0.612
Random Forest	tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.854	0.839	0.743
Rotation Forest	tabela4K40W30	*iterations*=30	0.861	**0.841**	0.737
MultiLayer Perceptron	tabela4K20W30	*learning rate*=0.3	0.829	0.799	0.688

Na Tabela 11 é possível verificar que relativamente à medida *F* o melhor resultado foi 0.841 obtido pelo algoritmo de classificação *Rotation Forest*. Este valor foi obtido utilizando como parâmetros do *MrMotif*: 40 *motifs* selecionados e 30 relevantes e 30 *iterations*.

5.4.2.3 Área ROC

Nesta secção, apresentamos os melhores resultados de área ROC obtidos nos modelos testados. De forma a resumir os resultados obtidos, apresentam-se os dez melhores resultados obtidos por técnica de classificação referente à média ponderada.

Na Tabela 12 apresentam-se os melhores resultados obtidos utilizando o algoritmo *J48*

Tabela 12 – Melhores resultados da área ROC com *J48*

Dataset	Parâmetros	*Recall*	*F-Measure*	Área ROC
tabela4K50W30	*leaf*=1	0.797	0.785	**0.706**
tabela8K30W10	*leaf*=10	0.801	0.806	0.677
tabela8K40W10	*leaf*=10	0.810	0.806	0.677
tabela8K50W10	*leaf*=10	0.810	0.806	0.677
tabela8K64W10	*leaf*=10	0.810	0.806	0.677
tabela8K20W10	*leaf*=10	0.791	0.782	0.671
tabela8K20W10	*leaf*=15	0.804	0.798	0.663
tabela8K30W10	*leaf*=15	0.804	0.798	0.663
tabela8K40W10	*leaf*=15	0.804	0.798	0.663
tabela8K50W10	*leaf*=15	0.804	0.798	0.663
tabela8K64W10	*leaf*=15	0.804	0.798	0.663

Como podemos observar o melhor resultado corresponde a uma área *ROC* de 0.706. Os melhores resultados foram obtidos, no algoritmo *J48*, para os parâmetros K=50 e W=30 e *leaf*=1.

Na Tabela 13, apresentam-se os melhores resultados obtidos utilizando o algoritmo *Random Forest*.

Tabela 13 – Melhores resultados da área ROC com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K40W10	*trees*=40, *features*=2, *seed*=1	0.823	0.794	**0.816**
tabela8K40W10	*trees*=20, *features*=2, *seed*=1	0.797	0.754	0.813
tabela8K40W10	*trees*=50, *features*=2, *seed*=1	0.797	0.759	0.811
tabela8K40W10	*trees*=30, *features*=2, *seed*=1	0.791	0.749	0.807
tabela8K64W10	*trees*=40, *features*=2, *seed*=1	0.810	0.763	0.806
tabela8K64W10	*trees*=50, *features*=2, *seed*=1	0.810	0.763	0.806
tabela8K40W10	*trees*=25, *features*=2, *seed*=1	0.804	0.764	0.801
tabela4K20W30	*trees*=30, *features*=1, *seed*=1	0.829	0.807	0.797
tabela4K20W30	*trees*=25, *features*=1, seed=1	0.829	0.807	0.794
tabela8K64W10	*trees*=30, *features*=2, *seed*=1	0.804	0.752	0.793
tabela4K20W30	*trees*=40, *features*=1, *seed*=1	0.823	0.802	0.792

O melhor resultado obtido é de uma área *ROC* de 0.816 com os parâmetros K=40 e W=10 no *MrMotif*, *trees*=40, *features*=2 e *seed*=1, no *Random Forest*.

Na Tabela 14 apresentam-se os melhores resultados obtidos utilizando o algoritmo *Rotation Forest*.

Tabela 14 – Melhores resultados da área ROC com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K40W10	*iterations*=20	0.816	0.793	**0.797**
tabela8K20W10	*iterations*=25	0.823	0.809	0.791
tabela8K30W10	*iterations*=30	0.829	0.803	0.781
tabela8K50W10	*iterations*=20	0.823	0.794	0.781
tabela8K40W10	*iterations*=25	0.829	0.807	0.772
tabela8K20W30	*iterations*=30	0.791	0.764	0.770
tabela8K20W10	*iterations*=50	0.842	0.828	0.764
tabela8K30W10	*iterations*=50	0.823	0.798	0.763
tabela8K30W10	*iterations*=25	0.835	0.813	0.761
tabela8K50W30	*iterations*=30	0.791	0.776	0.760
tabela8K20W10	*iterations*=20	0.797	0.774	0.760

O melhor resultado obtido é de uma área ROC de 0.797 com o K=40, W=10 e *iterations*=20.

Na Tabela 15 apresentam-se os melhores resultados obtidos utilizando o algoritmo *MultiLayer Perceptron*

Tabela 15 – Melhores resultados da área ROC para *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K4W30	*learning rate*=0.3	0.804	0.774	**0.736**
tabela8K4W20	*learning rate*=0.3	0.810	0.757	0.734
tabela4K4W10	*learning rate*=0.3	0.804	0.745	0.730
tabela4K20W40	*learning rate*=0.3	0.791	0.782	0.712
tabela4K64W30	*learning rate*=0.3	0.791	0.782	0.711
tabela8K4W10	*learning rate*=0.3	0.791	0.728	0.710
tabela4K4W40	*learning rate*=0.3	0.804	0.769	0.709
tabela4K20W30	*learning rate*=0.3	0.829	0.799	0.688
tabela4K30W30	*learning rate*=0.3	0.797	0.765	0.688

O melhor resultado obtido foi de uma área ROC de 0.736. Este resultado foi obtido com K=4, W=30 e um *learning rate* de 0.3.

Comparando os melhores resultados de cada técnica de classificação obtemos a seguinte tabela:

Tabela 16 – Resumo os melhores resultados da área ROC

Algoritmo	***Dataset***	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
J48	tabela4K50W30	*leaf*=1	0.797	0.785	0.706
Random Forest	tabela8K40W10	*trees*=40, *features*=2, *seed*=1	0.823	0.794	**0.816**
Rotation Forest	tabela8K40W10	iterayions=20	0.816	0.793	0.797
SMO	tabela32K50W40	*seed*=1	0.785	0.696	0.514
MultiLayer Perceptron	tabela4K4W30	*learning rate*=0.3	0.804	0.774	0.736

Na Tabela 16 é possível verificar que, relativamente à área ROC, o modelo com melhores resultados de classificação é o *Random Forest*, para o *dataset* resultante do *MrMotif* com 40 *motifs* selecionados e 10 *motifs* relevantes como parâmetros.

5.5 Classificação usando atributos clínicos

Nesta secção descrevemos a metodologia utilizada nos modelos de classificação, utilizando dados clínicos como atributos. Faz-se também a análise dos resultados obtidos utilizando as diferentes medidas de avaliação.

O conjunto de dados originais, com os atributos clínicos dos pacientes, é composto por 18 atributos. Estes atributos podem ser visualizados na Tabela 17. Na segunda coluna dessa tabela apresenta-se também uma breve descrição dos atributos.

Tabela 17 – Atributos dos dados recolhidos dos pacientes

Atributos	Descrição
Weight	Peso em Quilogramas
Height	Altura em centímetros
Height percentile	Percentil da altura
BMI	Índice de massa muscular (IMC)
Age_Years	Idade em anos
Age_months	Idade em meses
Sex	Sexo
AuscultationPosition	Posição em como foi efetuada a auscultação
SystemicPressureMethod	Método de medição da tensão
SystolicSystemicPressure_mmHg	Valores obtidos da medição da pressão sistólica
DiastolicSystemicPressure_mmHg	Valores obtidos da medição da pressão diastólica
Hypertension	Indicação se a criança tem ou não hipertensão
S1Status	Classificação do período sistólico do movimento cardíaco (lub)
S2Status	Classificação do período diastólico do movimento cardíaco (dub)
PulmonaryComponent	Informações sobre a componente pulmonar
CardiacPathology	Informações sobre a existência de patologias cardíacas
CardiacPathologyType	Tipo de patologia cardíaca
Murmur	Indica tipo de murmur verificado

Destes atributos, numa primeira fase, selecionamos apenas os apresentados na Tabela 18.

Tabela 18 – Atributos utilizados para a primeira análise

Age_Years
BMI
Sex
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg
PulmonaryComponent

Estes atributos foram escolhidos por indicação de uma médica de família. Em sua opinião, estes são os atributos mais importantes no estudo de diagnósticos cardíacos.

Numa segunda análise selecionámos os atributos apresentados na Tabela 19.

Tabela 19 – Atributos utilizados para a segunda análise

Weight
Height
Age_months
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg

Numa terceira análise selecionamos os atributos apresentados na Tabela 20.

Tabela 20 – Atributos utilizados para a terceira análise

Height_percentile
BMI
Age_years
Sex
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg
PulmonaryComponent

Numa quarta análise selecionamos os atributos apresentados na Tabela 21.

Tabela 21 – Atributos utilizados na quarta análise

Height_percentile
BMI
Age_years
Sex
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg
PulmonaryComponent
S1Status
S2Status

5.5.1 Aplicação dos algoritmos de classificação

Nesta secção serão apresentados os resultados obtidos pelos diferentes algoritmos de classificação aplicados a quatro diferentes *datasets* de atributos clínicos.

5.5.1.1 Primeira análise

Na Tabela 22 são apresentados os melhores resultados obtidos para cada algoritmo (baseada na Tabela 18)

Tabela 22 – Melhores resultados referente ao *Recall*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=10	0.774	0.675	0.459
Random Forest	*trees*=40, *features*=2, *seed*=1	**0.780**	0.717	0.559
Rotation Forest	*iterations*=50	0.767	0.672	0.554
SMO	*seed*=1	0.774	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.717	0.689	0.591

Como podemos verificar, o melhor valor de *Recall* obtido é de 0.780 com o algoritmo *Random Forest*. Este resultado foi obtido utilizando como parâmetros 40 *trees*, 2 *features* e *seed*=1.

Na Tabela 23 apresentam-se os melhores resultados observados para a *F-Measure*.

Tabela 23 – Melhores resultados referente ao *F-Measure*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=10	0.774	0.675	0.459
Random Forest	*trees*=40, *features*=3, *seed*=1	0.780	**0.731**	0.560
Rotation Forest	*iterations*=50	0.767	0.672	0.554
SMO	*seed*=1	0.774	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.717	0.689	0.591

Podemos observar que relativamente ao *F-Measure*, o melhor valor obtido é de 0.731 com o algoritmo *Random Forest*. Os parâmetros utilizados para a obtenção deste resultado foram 40 *trees*, 3 *features* e o *seed*=1.

Tabela 24 – Melhores resultados referente à área ROC

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=2	0.767	0.672	0.474
Random Forest	*trees*=25, *features*=1, *seed*=1	0.767	0.701	**0.602**
Rotation Forest	*iterations*=50	0.767	0.672	0.554
SMO	*seed*=1	0.774	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.717	0.689	0.591

Relativamente à área ROC o melhor resultado obtido foi de 0.602 também com algoritmo *Random Forest*. Neste caso os parâmetros usados foram 25 *trees*, 1 *feature* e *seed*=1 como é possível verificar na Tabela 24.

5.5.1.2 Segunda análise

Na Tabela 25 podem observar-se os melhores resultados obtidos para cada algoritmo para o *dataset* composto pelos atributos apresentados na Tabela 19.

Tabela 25 – Melhores resultados referente ao *Recall*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=15	**0.774**	0.675	0.459
Random Forest	*trees*=25, *features*=1, *seed*=1	0.761	0.705	0.552
Rotation Forest	*iterations*=20	**0.774**	0.686	0.535
SMO	*seed*=1	**0.774**	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.736	0.696	0.595

Neste caso, o melhor valor obtido de *Recall* foi de 0.774. Este valor foi conseguido através de três algoritmos: *J48*, *Rotation Forest* e *SMO*. No caso do *J48* a *leaf* era 15, no caso do Rotation foram usadas 20 *iterations* e no caso do *SMO seed*=1.

Tabela 26 – Melhores resultados referente ao *F-Measure*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=15	0.774	0.675	0.459
Random Forest	*trees*=25, *features*=1, *seed*=1	0.761	**0.705**	0.552
Rotation Forest	*iterations*=20	0.774	0.686	0.535
SMO	*seed*=1	0.774	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.736	0.696	0.595

Para o *F-Measure* o melhor valor obtido foi 0.705 referente ao algoritmo *Random Forest*. Os parâmetros utilizados foram 25 *trees*, *features*=1 e *seed*=1, conforme é possível observar na Tabela 26.

Tabela 27 – Melhores resultados referente à área ROC

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=15	0.774	0.675	0.459
Random Forest	*trees*=30, *features*=2, *seed*=1	0.761	0.705	0.563
Rotation Forest	*iterations*=25	0.767	0.672	0.594
SMO	*seed*=1	0.774	0.675	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.736	0.696	**0.595**

No caso da área ROC, o melhor valor obtido foi 0.595 com o algoritmo *MultiLayer Perceptron* com um *learning rate* de 0.3, conforme é constatado na Tabela 27.

5.5.1.3 Terceira análise

De seguida é apresentada a tabela com o melhor resultado obtido para cada algoritmo utilizando o *dataset* composto pelos atributos definidos na Tabela 20.

Tabela 28 – Melhores resultados referente ao *Recall*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=2	0.741	0.662	0.443
Random Forest	*trees*=40, *features*=1, *seed*=1	0.772	0.699	0.518
Rotation Forest	*iterations*=20	0.759	0.672	0.464
SMO	*seed*=1	**0.778**	0.682	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.671	0.650	0.504

No caso de *Recall* o melhor valor obtido foi 0.778 através o algoritmo *SMO* com o parâmetro *seed*=1 (Tabela 28)

Tabela 29 – Melhores resultados referente ao *F-Measure*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=2	0.741	0.662	0.443
Random Forest	*trees*=50, *features*=3, *seed*=1	0.759	**0.700**	0.580
Rotation Forest	*iterations*=20	0.759	0.672	0.464
SMO	*seed*=1	0.778	0.682	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.671	0.650	0.504

Relativamente ao *F-Measure* o melhor valor obtido foi 0.700. O algoritmo que permitiu atingir este valor foi o *Random Forest* com 50 *trees*, 3 *features* e 1 *seed*.

Tabela 30 – Melhores resultados referente à área ROC

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=2	0.741	0.662	0.443
Random Forest	*trees*=20, *features*=3, *seed*=1	0.741	0.672	**0.588**
Rotation Forest	*trees*=40	0.753	0.669	0.515
SMO	*seed*=1	0.778	0.682	0.500
MultiLayer Perceptron	*learning rate*=0.3	0.671	0.650	0.504

No caso da área ROC , o melhor resultado obtido foi 0.588 com o algoritmo de *Random Forest*. Para a obtenção deste resultado foram utilizados como parâmetros 20 *trees*, 3 *features* e *seed*=1, conforme é possível observar na Tabela 30.

5.5.1.4 Quarta análise

De seguida é apresentada a tabela com o melhor resultado obtido para cada algoritmo para o *dataset* composto pelos atributos definidos na Tabela 21.

Tabela 31 – Melhores resultados referente ao *Recall*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=1	0.772	0.708	0.486
Random Forest	*trees*=20, *features*=2, *seed*=1	**0.791**	0.743	0.573
Rotation Forest	*iterations*=20	0.766	0.675	0.557
SMO	*seed*=1	0.785	0.696	0.514
MultiLayer Perceptron	*learning rate*=0.3	0.709	0.675	0.561

No caso de *Recall* o melhor valor obtido foi 0.791 através do algoritmo *Random Forest*. Os parâmetros utilizados foram 20 *trees*, 2 *features* e *seed*=1 (Tabela 31).

Tabela 32 – Melhores resultados referente ao *F-Measure*

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=1	0.772	0.708	0.486
Random Forest	*trees*=20, *features*=2, *seed*=1	0.791	**0.743**	0.573
Rotation Forest	*iterations*=20	0.766	0.675	0.557
SMO	*seed*=1	0.785	0.696	0.514
MultiLayer Perceptron	*learning rate*=0.3	0.709	0.675	0.561

Relativamente ao *F-Measure* o melhor valor obtido foi 0.743. O algoritmo que permitiu atingir este valor foi o *Random Forest* com 20 *trees*, 2 *features* e *seed*=1.

Tabela 33 – Melhores resultados referente à área ROC

Algoritmo	Parâmetros	*Recall*	*F-Measure*	Área ROC
J48	*leaf*=1	0.772	0.708	0.486
Random Forest	*trees*=30, *features*=2, *seed*=1	0.785	0.716	**0.581**
Rotation Forest	*iterations*=40	0.766	0.675	0.566
SMO	*seed*=1	0.785	0.696	0.514
MultiLayer Perceptron	*learning rate*=0.3	0.709	0.675	0.561

No caso da área ROC, o melhor resultado obtido foi 0.581 com o algoritmo *Random Forest*. Para a obtenção deste resultado foram utilizados como parâmetros 30 *trees*, 2 *features* e *seed*=1 (Tabela 33).

5.5.2 Análise de resultados

De acordo com a primeira análise podemos concluir que o algoritmo de classificação que obtém melhores resultados para conjunto de atributos enumerados na Tabela 18 é o *Random Forest*. Os valores obtidos foram *Recall* de 0.780, *F-Measure* de 0.731 e uma área ROC de 0.602.

Relativamente à segunda análise já verificamos que dependendo da medida, os melhores resultados são conseguidos por diferentes modelos de classificação. Os atributos utilizados encontram-se referidos na Tabela 19. Se a medida de avaliação considerada for o *Recall* existem três algoritmos que permitem obter melhores resultados: *J48*, *Random Forest* e *SMO* com um valor de 0.774. No caso da medida F (*F-Measure*) é o algotimo *Random Forest* com o valor de 0.705 que proporciona melhor valor e para a área ROC com o valor de 0.595 é o algoritmo Multilayer Perceptron.

Relativamente à terceira análise (Tabela 20), podemos verificar que para a medida de *Recall*, o melhor algoritmo é o *SMO* com o valor de 0. 778. Relativamente ao *F-Measure* e área ROC , os melhores valores obtidos são proporcionados pelo algoritmo *Random Forest*. A medida F tem o valor de 0.700 e a área ROC 0.588.

Na quarta análise, o algoritmo *Random Forest* permite obter os melhores resultados. Este algoritmo permite obter valores de *Recall* de 0.791, *F-Measure* de 0.743 e área ROC de 0.581. Os dados utilizados nesta análise encontram-se na Tabela 21.

É importante ter em conta a inexistência de balanceamento dos dados, problema comum em dados reais. Neste *dataset*, com 123 ficheiros de classe normal (78%) e 35 da classe anormal (22%), as classes são bastante desequilibradas e este cenário requer algum cuidado ao analisar os resultados obtidos. Nestes casos, para que os resultados sejam aceitáveis, a taxa de acerto preditiva deve ser maior do que a taxa de acerto obtida quando se atribui a classe maioritária a todo e qualquer objeto [Gama, 2015]. Ou seja, para um resultado ser considerado interessante deve ser superior a 0.780. Nas experiências efetuadas consegue-se uma *Recall* de 0.790 para o 4º conjunto de dados clínicos usando o *Random Forest*. Apesar de este valor ser superior a 0.78 é bastante próximo do mesmo.

Por outro lado, os resultados obtidos são estatisticamente muito próximos dos reportados na literatura [Ferreira, 2012]. Neste trabalho, onde é apresentado um estudo com o objetivo de detetar patologia cardíaca usando um *dataset* similar, a *Recall* é também aproximadamente 0.790.

Pudemos assim observar que quanto à avaliação dos modelos de classificação utilizados nas nossas experiências, estes apresentaram uma baixa capacidade discriminante quando comparados com abordagem usando os *motifs* (0.860).

5.6 Classificação usando atributos combinados

Nesta secção descrevemos a metodologia utilizada nos modelos de classificação, utilizando os atributos obtidos na abordagem dos *motifs* combinados com os atributos clínicos. Faz-se também uma análise dos resultados obtidos nas diferentes medidas de avaliação.

Esta técnica de combinar atributos foi tinha sido utilizada em trabalhos anteriores e com obtenção de bons resultados [Gomes, 2013b]. A ideia desta abordagem consiste numa tentativa de melhorar o classificador combinando dois tipos de informação: os *motifs* que resultam dos sons cardíacos recolhidos e os dados clínicos com informação dos pacientes, que foram fornecidos. De salientar que este tipo de análise requer o cuidado de fazer corresponder os dois tipos de atributos à mesma linha no *dataset*, ou seja, que os dados pertençam ao mesmo paciente.

5.6.1 Aplicação dos algoritmos de classificação

Para realizar esta análise, foram combinados os atributos que anteriormente foram analisados separadamente. Foram realizadas duas análises com os dados combinados.

5.6.1.1 Primeira análise

Numa primeira análise foram utilizados todos os *datasets* obtidos e analisados na secção 5.4.2 juntamente com os seguintes dados clínicos:

Tabela 34 – Atributos utilizados para a primeira análise

Age_Years
BMI
Sex
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg
PulmonaryComponent

Para melhor compreensão dos resultados serão analisados inicialmente os 5 melhores resultados de cada modelo e posteriormente a comparação entre os melhores resultados conseguidos.

Na tabela 35 apresentam-se os melhores resultados obtidos de *Recall*, para o algoritmo *J48*.

Tabela 35 – Melhores resultados do *Recall* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K4W10	*leaf*=1	**0.842**	0.822	0.619
tabela8K4W10	*leaf*=2	0.829	0.803	0.570
tabela8K4W10	*leaf*=15	0.829	0.828	0.660
tabela4K4W40	*leaf*=1	0.823	0.805	0.688
tabela8K30W10	*leaf*=1	0.823	0.814	0.612

O melhor resultado obtido foi para um valor de *Recall* de 0.842 com um *leaf*=1 e 4 *motifs* selecionados e 10 *motifs* relevantes no *MrMotif*.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos encontram-se na Tabela 36.

Tabela 36 – Melhores resultados do *Recall* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K50W30	*trees*=50, *features*=2, *seed*=1	**0.854**	0.829	0.777
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.848	0.823	0.786
tabela4K64W40	*trees*=25, *features*=1, *seed*=1	0.848	0.823	0.709
tabela4K40W30	*trees*=30, *features*=1, *seed*=1	0.848	0.820	0.776
tabela4K50W30	*trees*=30, *features*=3, *seed*=1	0.848	0.827	0.775

O melhor resultado obtido foi um valor de *Recall* de 0.854 com os parâmetros de 50 *trees*, 2 *features*, *seed*=1, 50 *motifs* selecionados e 30 *motifs* relevantes no *MrMotif*.

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 37 – Melhores resultados do *Recall* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=30	**0.861**	0.844	0.738
tabela4K40W40	*iterations*=20	0.854	0.839	0.717
tabela4K50W30	*iterations*=30	0.854	0.833	0.755
tabela4K50W30	*iterations*=20	0.848	0.823	0.750
tabela4K64W30	*iterations*=50	0.848	0.830	0.741

O melhor resultado obtido foi um valor de *Recall* de 0.861 com um número de *trees* de 30, 40 *motifs* selecionados e 30 *motifs* relevantes.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos foram:

Tabela 38 – Melhores resultados do *Recall* com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela64K30W40	*seed*=2	**0.803**	0.743	0.576
tabela64K30W40	*seed*=10	**0.803**	0.743	0.576
tabela64K30W40	*seed*=1	0.790	0.718	0.548
tabela32K50W40	*seed*=1	0.785	0.696	0.514
tabela64K64W20	*seed*=1	0.785	0.696	0.514

O melhor resultado obtido foi um valor de *Recall* de 0.803 obtido para o mesmo número de *motifs* relevantes (40) e número de *motifs* selecionados (30). A variante destes resultados é o número de *seeds* utilizados, neste caso, 2 e 10.

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos encontram-se descritos na Tabela 39.

Tabela 39 – Melhores resultados do *Recall* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K64W30	*learning rate*=0.3	**0.772**	0.758	0.722
tabela32K4W10	*learning rate*=0.3	0.766	0.731	0.599
tabela64K50W40	*learning rate*=0.3	0.759	0.672	0.566
tabela64K64W40	*learning rate*=0.3	0.759	0.672	0.561
tabela64K4W20	*learning rate*=0.3	0.753	0.679	0.517

O melhor resultado obtido foi um valor de *Recall* de 0.772 com 64 *motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3.

Tabela 40 – Resumo os melhores resultados de *Recall*

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela8K4W10	0.842	0.822	0.619
Random Forest	tabela4K50W30	0.854	0.829	0.777
Rotation Forest	tabela4K40W30	**0.861**	0.844	0.738
SMO	tabela64K30W40	0.803	0.743	0.576
MultiLayer Perceptron	tabela16K64W30	0.772	0.758	0.722

Da análise da Tabela 40 é possível concluir que o melhor valor de *Recall* obtido foi 0.861 através do algoritmo *Rotation Forest*. Os parâmetros utilizados para a obtenção deste valor foram de 40 *motifs* selecionados e 30 *motifs* relevantes no *MrMotif* e o número de *iterations*=30.

Relativamente ao *F-Measure*, para o algoritmo *J48* foram obtidos os resultados apresentados na Tabela 41.

Tabela 41 – Melhores resultados do *F-Measure* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K4W10	*leaf*=15	0.829	**0.828**	0.660
tabela8K4W10	*leaf*=1	0.842	0.822	0.619
tabela8K30W10	*leaf*=1	0.823	0.814	0.612
tabela8K30W10	*leaf*=2	0.823	0.811	0.612
tabela4K4W30	*leaf*=10	0.823	0.811	0.614

O melhor resultado obtido foi um valor de *F-Measure* de 0.828 com um *leaf*=15, 4 *motifs* selecionados e 10 *motifs* relevantes.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos encontram-se descritos na Tabela 42.

Tabela 42 – Melhores resultados do *F-Measure* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K50W30	*trees*=50, *features*=2, *seed*=1	0.854	**0.829**	0.777
tabela4K50W30	*trees*=30, *features*=3, *seed*=1	0.848	0.827	0.775
tabela4K50W30	*trees*=50, *features*=3, *seed*=1	0.848	0.827	0.762
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.848	0.823	0.786
tabela4K64W40	*trees*=25, *features*=1, *seed*=1	0.848	0.823	0.709

O melhor resultado obtido foi um valor de *F-Measure* de 0.829 com os parâmetros de 50 *trees*, 2 *features*, *seed*=1, 50 *motifs* selecionados e 30 *motifs* relevantes no *MrMotif*.

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 43 – Melhores resultados do *F-Measure* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=30	0.861	**0.844**	0.738
tabela4K40W40	*iterations*=20	0.854	0.839	0.717
tabela4K50W30	*iterations*=30	0.854	0.833	0.755
tabela4K50W30	*iterations*=20	0.848	0.823	0.750
tabela4K64W30	*iterations*=50	0.848	0.830	0.741

O melhor resultado obtido foi um valor de *F-Measure* de 0.844 com um número de *iterations* de 30, 40 *motifs* selecionados e 30 *motifs* relevantes.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos foram:

Tabela 44 – Melhores resultados do *F-Measure* com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela64K30W40	*seed*=2	0.803	**0.743**	0.576
tabela64K30W40	*seed*=10	0.803	**0.743**	0.576
tabela64K30W40	*seed*=1	0.790	0.718	0.548
tabela32K50W40	*seed*=1	0.785	0.696	0.514
tabela64K64W20	*seed*=1	0.785	0.696	0.514

O melhor resultado obtido foi um valor de *F-Measure* de 0.743 obtido para o mesmo número de *motifs* relevantes (40) e número de *motifs* selecionados (30). A variante destes resultados é o número de *seeds* utilizados, neste caso, 2 e 10.

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos são apresentados na Tabela 45:

Tabela 45 – Melhores resultados do *F-Measure* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K64W30	*learning rate*=0.3	0.772	**0.758**	0.722
tabela32K4W10	*learning rate*=0.3	0.766	0.731	0.599
tabela4K40W20	*learning rate*=0.3	0.734	0.731	0.667
tabela4K50W30	*learning rate*=0.3	0.741	0.730	0.681
tabela32K50W10	*learning rate*=0.3	0.728	0.729	0.690

O melhor resultado obtido foi um valor de *F-Measure* de 0.758 com 64 *motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3.

De seguida, na Tabela 46, é apresentado o resumo dos melhores resultados obtidos pelos algoritmos.

Tabela 46 – Resumo os melhores resultados de *F-Measure*

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela8K4W10	0.829	0.828	0.660
Random Forest	tabela4K50W30	0.854	0.829	0.777
Rotation Forest	tabela4K40W30	0.861	**0.844**	0.738
SMO	tabela64K30W40	0.803	0.743	0.576
MultiLayer Perceptron	tabela16K64W30	0.772	0.758	0.722

Da análise da Tabela 46 é possível concluir que o melhor valor de *F-Measure* obtido foi 0.844 através do algoritmo *Rotation Forest*. Este resultado foi obtido através de 40 *motifs* selecionados e 30 *motifs* relevantes obtidos do *MrMotif* e 30 *iterations*.

Relativamente à área ROC , para o algoritmo *J48* foram obtidos os resultados apresentados na Tabela 47.

Tabela 47 – Melhores resultados da área ROC com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K4W40	*leaf*=1	0.823	0.805	**0.688**
tabela8K4W10	*leaf*=10	0.823	0.805	0.681
tabela8K30W10	*leaf*=10	0.804	0.793	0.678
tabela8K40W10	*leaf*=10	0.804	0.793	0.678
tabela8K50W10	*leaf*=10	0.804	0.793	0.678

O melhor resultado obtido foi um valor de área ROC de 0.688 com um *leaf*=1, 4 *motifs* selecionados e 40 *motifs* relevantes.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos foram:

Tabela 48 – Melhores resultados da área ROC com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K20W10	*trees*=20, *features*=2, *seed*=1	0.804	0.779	**0.808**
tabela8K40W10	*trees*=50, *features*=3, *seed*=1	0.823	0.784	0.802
tabela8K20W10	*trees*=40, *features*=2, *seed*=1	0.804	0.774	0.802
tabela8K20W10	*trees*=50, *features*=2, *seed*=1	0.797	0.765	0.802
tabela8K20W10	*trees*=30, *features*=2, *seed*=1	0.804	0.779	0.798

Conforme observado na Tabela 48, o melhor resultado obtido foi um valor de 0.808 para a área ROC com os parâmetros de 20 *trees*, 2 *features*, *seed*=1, 20 *motifs* selecionados e 10 *motifs* relevantes.

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 49 – Melhores resultados da área ROC com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=20	0.842	0.822	**0.795**
tabela4K20W30	*iterations*=20	0.835	0.813	0.794
tabela8K30W10	*iterations*=50	0.810	0.769	0.794
tabela4K50W30	*iterations*=50	0.842	0.822	0.792
tabela4K20W30	*iterations*=40	0.835	0.813	0.790

O melhor resultado obtido foi um valor de área ROC de 0.795 com um número de *iterations* de 20, 40 *motifs* selecionados e 30 *motifs* relevantes.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos encontram-se na Tabela 50.

Tabela 50 – Melhores resultados da área ROC com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela64K30W40	*seed*=2	0.803	0.743	**0.576**
tabela64K30W40	*seed*=10	0.803	0.743	**0.576**
tabela64K30W40	*seed*=1	0.790	0.718	0.548
tabela32K50W40	*seed*=1	0.785	0.696	0.514
tabela64K64W20	*seed*=1	0.785	0.696	0.514

O melhor resultado obtido foi um valor da área ROC de 0.576 obtido para o mesmo número de *motifs* relevantes (40) e número de *motifs* selecionados (30). A variante destes resultados é o número de *seeds* utilizados, neste caso, 2 e 10.

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos foram:

Tabela 51 – Melhores resultados da área ROC com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K64W30	*learning rate*=0.3	0.772	0.758	**0.722**
tabela32K20W10	*learning rate*=0.3	0.728	0.720	0.707
tabela32K50W10	*learning rate*=0.3	0.728	0.729	0.690
tabela4K50W30	*learning rate*=0.3	0.741	0.730	0.681
tabela4K64W30	*learning rate*=0.3	0.715	0.714	0.680

Conforme é possível verificar na Tabela 51, o melhor resultado obtido foi um valor da área ROC de 0.722 com 64 *motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3.

Da análise da Tabela 52, pode observar-se que o melhor valor da área ROC obtido foi 0.808 com o algoritmo *Random Forest*. Os parâmetros utilizados neste algoritmo foram20 *trees*, 2 *features*, *seed*=1 e 20 *motifs* selecionados e 10 *motifs* relevantes no *MrMotif*.

Tabela 52 – Resumo os melhores resultados da área ROC

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela4K4W40	0.823	0.805	0.688
Random Forest	tabela8K20W10	0.804	0.779	**0.808**
Rotation Forest	tabela4K40W30	0.842	0.822	0.795
SMO	tabela64K30W40	0.803	0.743	0.576
MultiLayer Perceptron	tabela16K64W30	0.772	0.758	0.722

5.6.1.2 Segunda análise

Numa segunda análise foram utilizados todos os atributos obtidos e analisados na secção 5.4.2 juntamente com os seguintes dados clínicos:

Tabela 53 – Atributos utilizados para a segunda análise

Height_percentile
BMI
Age_years
Sex
SystolicSystemicPressure_mmHg
DiastolicSystemicPressure_mmHg
PulmonaryComponent
S1Status
S2Status

Na Tabela 53 encontram-se os atributos relativos aos dados clínicos utilizados para combinar com os atributos baseados em *motifs* e assim fazer a análise combinada.

Para melhor compreensão dos resultados serão inicialmente analisados os 5 melhores resultados de cada algoritmo e posteriormente a comparação entre o melhor resultado de cada algoritmo.

Relativamente ao *Recall*, para o algoritmo *J48* foram obtidos os seguintes resultados

Tabela 54 – Melhores resultados do *Recall* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K30W10	*leaf*=2	**0.829**	0.817	0.680
tabela8K4W10	*leaf*=15	**0.829**	0.828	0.660
tabela4K20W30	*leaf*=10	0.823	0.805	0.620
tabela4K4W30	*leaf*=10	0.823	0.811	0.614
tabela8K4W10	*leaf*=10	0.823	0.805	0.681

O melhor resultado obtido foi um valor de *Recall* de 0.829. Este valor é partilhado por duas gamas de parâmetros diferentes. Um dos resultados é obtido com 30 *motifs* selecionados, 10 *motifs* relevantes e *leaf* =2, o outro é obtido por 4 *motifs* selecionados, 10 *motifs* relevantes e *leaf*=15.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos encontram-se na Tabela 55.

Tabela 55 – Melhores resultados do *Recall* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*trees*=25, *features*=1, *seed*=1	**0.854**	0.833	0.749
tabela4K40W30	*trees*=40, *features*=1, *seed*=1	**0.854**	0.829	0.745
tabela4K40W30	*trees*=50, *features*=1, *seed*=1	**0.854**	0.833	0.745
tabela4K64W30	*trees*=25, *features*=1, *seed*=1	0.848	0.823	0.778
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.842	0.822	0.754

O melhor resultado obtido foi um valor de *Recall* de 0.854 com três gamas de parâmetros. Em comum tem o número de *motifs* selecionados (40) e relevantes (30), *features* e *seed*=1. Só o número de *trees* é que é variável sendo utilizados os valores 25,40 e 50.

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 56 – Melhores resultados do *Recall* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=25	**0.848**	0.830	0.744
tabela4K40W30	*iterations*=30	**0.848**	0.830	0.765
tabela4K30W40	*iterations*=40	**0.848**	0.830	0.715
tabela4K50W30	*iterations*=40	**0.848**	0.830	0.724
tabela4K64W30	*iterations*=40	**0.848**	0.827	0.706

O melhor resultado obtido é partilhado com diversas gamas de parâmetros sendo que não existe um padrão que possa ser retirado dos mesmos. O melhor resultado de *Recall* obtido foi 0.848.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos estão definidos na Tabela 57.

Tabela 57 – Melhores resultados do *Recall* com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K50W40	*seed*=1	**0.797**	0.724	0.543
tabela32K50W40	*seed*=1	**0.797**	0.724	0.543
tabela64K64W20	*seed*=1	**0.797**	0.724	0.543
tabela8K30W30	*seed*=1	**0.797**	0.724	0.543
tabela32K50W40	*seed*=2	**0.797**	0.724	0.543

À semelhança do algoritmo anterior, para o algoritmo *SMO* não é possível observar um padrão de parâmetros comum. O melhor valor de *Recall* obtido é 0.797.

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos foram:

Tabela 58 – Melhores resultados do *Recall* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K50W30	*learning rate*=0.3	**0.804**	0.798	0.720
tabela4K30W20	*learning rate*=0.3	0.778	0.763	0.691
tabela64K20W20	*learning rate*=0.3	0.778	0.755	0.596
tabela64K50W40	*learning rate*=0.3	0.778	0.712	0.590
tabela4K4W10	*learning rate*=0.3	0.772	0.758	0.679

O melhor resultado obtido foi um valor de *Recall* de 0.804 com 50 *motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3, conforme é possível verificar na Tabela 58.

De seguida é apresentada a tabela com os melhores resultados obtidos.

Tabela 59 – Resumo os melhores resultados de *Recall*

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela8K30W10	0.829	0.817	0.680
Random Forest	tabela4K40W30	**0.854**	0.833	0.749
Rotation Forest	tabela4K40W30	0.848	0.830	0.744
SMO	tabela16K50W40	0.797	0.724	0.543
MultiLayer Perceptron	tabela4K50W30	0.804	0.798	0.720

Da análise da Tabela 59 é possível concluir que o melhor valor de *Recall* obtido foi 0.854 através do algoritmo *Random Forest*. Como foi referido anteriormente, o resultado obtido tem o número de *motifs* selecionados (40) e relevantes (30), *features* e *seed*=1. O parâmetro que é variável é o número de *trees* (25,40 e 50).

Relativamente ao *F-Measure*, para o algoritmo *J48* foram obtidos os seguintes resultado apresentados na Tabela 60.

Tabela 60 – Melhores resultados do *F-Measure* com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K4W10	*leaf*=15	0.829	**0.828**	0.660
tabela8K30W10	*leaf*=2	0.829	0.817	0.680
tabela4K4W30	*leaf*=10	0.823	0.811	0.614
tabela4K4W30	*leaf*=15	0.823	0.811	0.614
tabela16K20W10	*leaf*=15	0.810	0.810	0.634

O melhor resultado obtido foi um valor de *F-Measure* de 0.828 com um *leaf*=15, 4 *motifs* selecionados e 10 *motifs* relevantes.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos foram:

Tabela 61 – Melhores resultados do *F-Measure* com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*trees*=25, *features*=1, *seed*=1	0.854	**0.833**	0.749
tabela4K40W30	*trees*=50, *features*=1, *seed*=1	0.854	**0.833**	0.745
tabela4K40W30	*trees*=40, *features*=1, *seed*=1	0.854	0.829	0.745
tabela4K64W30	*trees*=25, *features*=1, *seed*=1	0.848	0.823	0.778
tabela4K40W30	*trees*=20, *features*=1, *seed*=1	0.842	0.822	0.754

Conforme é possível verificar na Tabela 61, o melhor resultado obtido foi um valor de *F-Measure* de 0.833. Estes valores foram obtidos utilizando quase todos os parâmetros idênticos exceto o número de *trees* utilizadas (25 e 50).

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 62 – Melhores resultados do *F-Measure* com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*iterations*=25	0.848	**0.830**	0.744
tabela4K40W30	*iterations*=30	0.848	**0.830**	0.765
tabela4K30W40	*iterations*=40	0.848	**0.830**	0.715
tabela4K50W30	*iterations*=40	0.848	**0.830**	0.724
tabela4K40W30	*iterations*=50	0.848	**0.830**	0.787

O melhor resultado obtido foi um valor de *F-Measure* de 0.830. Este valor foi obtido por diferentes gamas de parâmetros logo não foi possível determinar um padrão.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos estão representados na Tabela 63:

Tabela 63 – Melhores resultados do *F-Measure* com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K50W40	*seed*=1	0.797	**0.724**	0.543
tabela32K50W40	*seed*=1	0.797	**0.724**	0.543
tabela64K64W20	*seed*=1	0.797	**0.724**	0.543
tabela8K30W30	*seed*=1	0.797	**0.724**	0.543
tabela32K50W40	*seed*=2	0.797	**0.724**	0.543

À semelhança do algoritmo anterior, o melhor valor de *F-Measure* foi obtido em diferentes gamas de valores sem evidência de um parâmetro comum. O melhor resultado de *F-Measure* obtido foi 0.724

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos foram:

Tabela 64 – Melhores resultados do *F-Measure* com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K50W30	*learning rate*=0.3	0.804	**0.798**	0.720
tabela4K30W20	*learning rate*=0.3	0.778	0.763	0.691
tabela4K4W10	*learning rate*=0.3	0.772	0.758	0.679
tabela8K4W10	*learning rate*=0.3	0.772	0.758	0.627
tabela64K20W20	*learning rate*=0.3	0.778	0.755	0.596

O melhor resultado obtido foi um valor de *F-Measure* de 0.798 com 50 *motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3.

Na Tabela 65 encontra-se um resumo dos melhores resultados obtidos para todos os algoritmos analisados relativamente à medida F.

Tabela 65 – Resumo os melhores resultados de *F-Measure*

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela8K4W10	0.829	0.828	0.660
Random Forest	tabela4K40W30	0.854	**0.833**	0.749
Rotation Forest	tabela4K40W30	0.848	0.830	0.744
SMO	tabela16K50W40	0.797	0.724	0.543
MultiLayer Perceptron	tabela4K50W30	0.804	0.798	0.720

Da análise da Tabela 65 é possível concluir que o melhor valor de *F-Measure* obtido foi 0.833 através do algoritmo *Random Forest*. Foram utilizados como parâmetros 40 *motifs* selecionados, 30 *motifs* relevantes e a variação do número de *trees* utilizadas (25 e 50).

Relativamente à área ROC , para o algoritmo *J48* foram obtidos os seguintes resultados:

Tabela 66 – Melhores resultados da área ROC com *J48*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K40W30	*leaf*=1	0.797	0.793	**0.709**
tabela4K30W30	*leaf*=1	0.785	0.783	0.696
tabela4K50W30	*leaf*=1	0.791	0.785	0.683
tabela8K4W10	*leaf*=10	0.823	0.805	0.681
tabela8K30W10	*leaf*=2	0.829	0.817	0.680

O melhor resultado obtido foi um valor de área ROC de 0.709 com um *leaf*=1, 40 *motifs* selecionados e 30 *motifs* relevantes.

Relativamente ao algoritmo *Random Forest*, os melhores resultados obtidos estão representados na Tabela 67:

Tabela 67 – Melhores resultados da área ROC com *Random Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K20W30	*trees*=40, *features*=3, *seed*=1	0.823	0.802	**0.795**
tabela4K40W40	*trees*=20, *features*=1, *seed*=1	0.816	0.784	0.789
tabela8K40W10	*trees*=20, *features*=3, *seed*=1	0.797	0.747	0.789
tabela4K64W30	*trees*=40, *features*=1, *seed*=1	0.842	0.810	0.785
tabela4K20W30	*trees*=30, *features*=3, *seed*=1	0.816	0.800	0.784

O melhor resultado obtido foi um valor da área ROC de 0.795 com os parâmetros de 40 *trees*, 3 *features*, *seed*=1, 20 *motifs* selecionados e 30 *motifs* relevantes.

Relativamente ao algoritmo *Rotation Forest*, os melhores resultados obtidos foram:

Tabela 68 – Melhores resultados da Área ROC com *Rotation Forest*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela8K50W10	*iterations*=25	0.829	0.803	**0.792**
tabela4K40W30	*iterations*=50	0.848	0.830	0.787
tabela8K20W10	*iterations*=50	0.791	0.760	0.781
tabela8K20W10	*iterations*=20	0.823	0.802	0.772
tabela4K30W30	*iterations*=50	0.829	0.811	0.769

O melhor resultado obtido foi um valor da área ROC de 0.792 com um número de *iterations* de 25, 50 *motifs* selecionados e 10 *motifs* relevantes, como é possível observar na Tabela 68.

Relativamente ao algoritmo *SMO*, os melhores resultados obtidos foram:

Tabela 69 – Melhores resultados da área ROC com *SMO*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela16K50W40	*seed*=1	0.797	0.724	**0.543**
tabela32K50W40	*seed*=1	0.797	0.724	**0.543**
tabela64K64W20	*seed*=1	0.797	0.724	**0.543**
tabela8K30W30	*seed*=1	0.797	0.724	**0.543**
tabela32K50W40	*seed*=2	0.797	0.724	**0.543**

O melhor resultado obtido foi um valor da área ROC de 0.543. Este valor é partilhado por uma gama de parâmetros, à semelhança do que já aconteceu em algoritmos e valores anteriores.

Relativamente ao algoritmo *MultiLayer Perceptron*, os melhores resultados obtidos estão descritos na Tabela 70.

Tabela 70 – Melhores resultados da área ROC com *MultiLayer Perceptron*

Dataset	**Parâmetros**	***Recall***	***F-Measure***	**Área ROC**
tabela4K50W30	*learning rate*=0.3	0.804	0.798	**0.720**
tabela16K64W30	*learning rate*=0.3	0.766	0.745	0.704
tabela4K64W30	*learning rate*=0.3	0.709	0.714	0.704
tabela4K64W40	*learning rate*=0.3	0.753	0.739	0.693
tabela4K30W20	*learning rate*=0.3	0.778	0.763	0.691

O melhor resultado obtido foi um valor da área ROC de 0.720 com *50 motifs* selecionados, 30 *motifs* relevantes e um *learning rate* de 0.3.

Em suma, na Tabela 71, é possível observar os melhores resultados obtidos da aplicação dos algoritmos anteriormente apresentados relativamente à área ROC.

Tabela 71 – Resumo os melhores resultados da área ROC

Algoritmo	***Dataset***	***Recall***	***F-Measure***	**Área ROC**
J48	tabela4K40W30	0.797	0.793	0.709
Random Forest	tabela4K20W30	0.823	0.802	**0.795**
Rotation Forest	tabela8K50W10	0.829	0.803	0.792
SMO	tabela16K50W40	0.797	0.724	0.543
MultiLayer Perceptron	tabela4K50W30	0.804	0.798	0.720

Da análise da Tabela 71 é possível concluir que o melhor valor da área ROC obtido foi 0.795 através do algoritmo *Random Forest*. Os parâmetros utilizados foram de 40 *trees*, 3 *features*, *seed*=1, 20 *motifs* selecionados e 30 *motifs* relevantes relativos ao *MrMotif*.

5.6.2 Análise de resultados

Na secção 5.6.1 estão descritas duas análises que complementam as realizadas anteriormente. No caso das análises efetuadas nas secções 5.4 e 5.5, os *datasets* continham os atributos separados. Para esta análise foram utilizados os atributos dos *datasets* anteriormente referenciados combinados, ou seja, juntando as colunas que continham os dados de ambos os atributos para cada linha do *dataset*.

Na primeira análise os melhores resultados obtidos pelos modelos utilizados foram 0.861 para a *Recall*, 0.844 para a *F-Measure* e 0.808 para a área ROC. Estes resultados foram obtidos por dois algoritmos. No caso da *Recall* e *F-Measure* pelo do algoritmo *Rotation Forest* e no caso da área ROC pelo algoritmo *Random Forest*.

Na segunda análise, os melhores resultados obtidos foram uma *Recall* de 0.854, *F-Measure* de 0.833 e uma área ROC de 0.795. No caso destes dados, os melhores valores foram todos obtidos pelo algoritmo *Random Forest*.

Comparando as duas análises observamos que, relativamente a todos os valores considerados para análise, a primeira análise obteve melhor resultado. Também é possível concluir que a melhor análise não tem em conta os dados refentes ao S1 e S2 *status*.

5.7 Discussão de resultados

Nesta secção apresentamos uma discussão de resultados, comparando os resultados obtidos nas três experiências descritas nas secções anteriores.

5.7.1 Comparação dos resultados obtidos

Nesta secção comparam-se os melhores resultados obtidos pelos modelos de classificação para os três tipos de experiências: utilizando a abordagem dos *motifs* (áudio), utilizando os dados clínicos e combinado os atributos das duas abordagens anteriores.

Na Tabela 72 apresentam-se os melhores resultados obtidos segundo a medida *Recall*. No caso dos *motifs* (primeira e segunda linha da tabela) obteve-se um valor de 0.861 e no caso dos dados clínicos (terceira linha da tabela) obteve-se 0.791 para o *dataset* resultante da combinação de dados da 4ª análise (ver secção 5.5.1.4). Em ambos os casos, o *Random Forest* obteve a melhor *Recall* (ex aequo com o *Rotation Forest* no caso dos *motifs*).

Tabela 72 – Resumo dos melhores resultados para o *Recall*

Dados	Algoritmo	Parâmetros	*Recall*
***Motifs*(tabela8K40W10)**	*Random Forest*	*trees*=25, *features*=1, *seed*=1	**0.861**
***Motifs* (tabela8K40W10)**	*Rotation Forest*	*iterations*=30	**0.861**
Clinicos(4ªanálise)	*Random Forest*	*trees*=20, *features*=2, *seed*=1	**0.791**

Na Tabela 73 apresentam-se os melhores resultados obtidos para a *F-Measure*. Tal como no caso do *Recall*, obteve-se um melhor resultado no caso dos *motifs*, 0.841 com o algoritmo *Rotation Forest*. Para os dados clínicos, obteve-se 0.743 para a combinação de dados da 4ª análise (ver secção 5.5.1.4) e com o algoritmo *Random Forest*.

Tabela 73 – Resumo dos melhores resultados para a *F-Measure*

Dados	Algoritmo	Parâmetros	*F-Measure*
***Motifs*(tabela4K40W30)**	*Rotation Forest*	*iterations*=30	**0.841**
Clínicos(4ª análise)	*Random Forest*	*trees*=20, *features*=2, *seed*=1	**0.743**

Por último, apresentam-se na Tabela 74 os resultados obtidos para a área ROC. Novamente, os *motifs* conseguiram os melhores resultados com o *Random Forest* (0.816) contra os dados clínicos do primeiro conjunto de atributos (secção 5.5.1.1).

Tabela 74 – Resumo dos melhores resultados para a área ROC

Dados	Algoritmo	Parâmetros	Área ROC
***Motifs*(tabela8K40W10)**	*Random Forest*	*trees*=40, *features*=2, *seed*=1	**0.816**
Clinicos(1ª análise)	*Random Forest*	*trees*=25, *features*=1, *seed*=1	**0.602**

Para completar esta análise faremos uma análise das curvas ROC para estes pares de resultados para a classe Anormal. Esta análise, como já referimos anteriormente, é usada para avaliar classificadores em problemas de duas classes (secção 5.3).

Na Figura 11 apresenta-se a curva ROC para os melhores modelos segundo a medida *Recall*, obtidos na classificação, utilizando como atributos os *motifs* (curva 'x', a azul) e os dados clínicos (curva '+', a vermelho).

Como podemos observar, os resultados para os *motifs* atingem muito rapidamente uma TVP elevada, com grande destaque relativamente aos resultados correspondentes aos atributos clínicos. Esta tendência mantém-se ao longo das curvas.

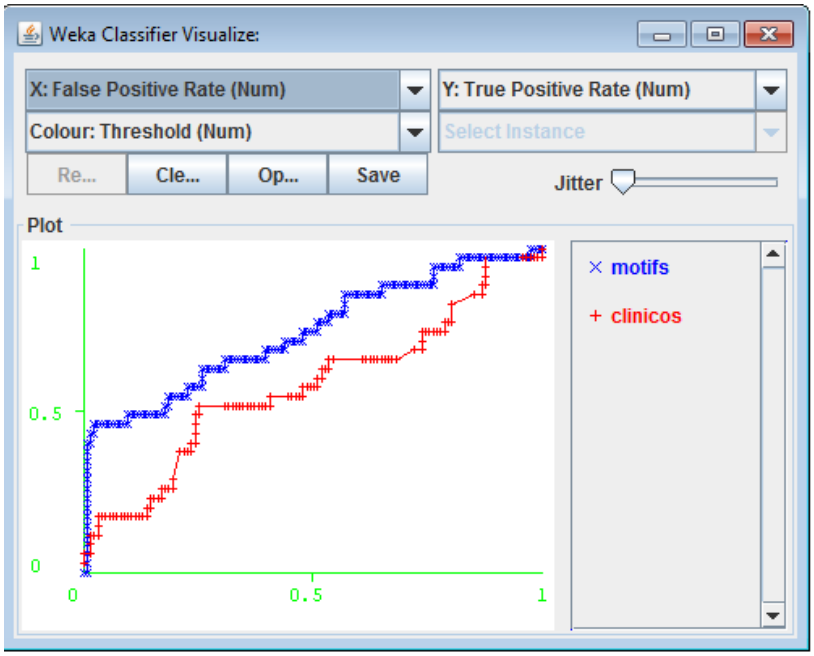

Figura 11 – Análise ROC para o *Recall*

Na Figura 12 apresenta-se a curva ROC para os melhores resultados de *Recall* obtidos na classificação, usando como atributos os *motifs* e os dados clínicos. Novamente, a curva a azul ('x') corresponde aos *motifs* e a vermelha ('+') os dados clínicos.

Tal como na figura anterior, observa-se também que os resultados para os atributos *motif* atingem muito rapidamente uma TVP elevada quando comparados com os resultados correspondentes aos atributos clínicos.

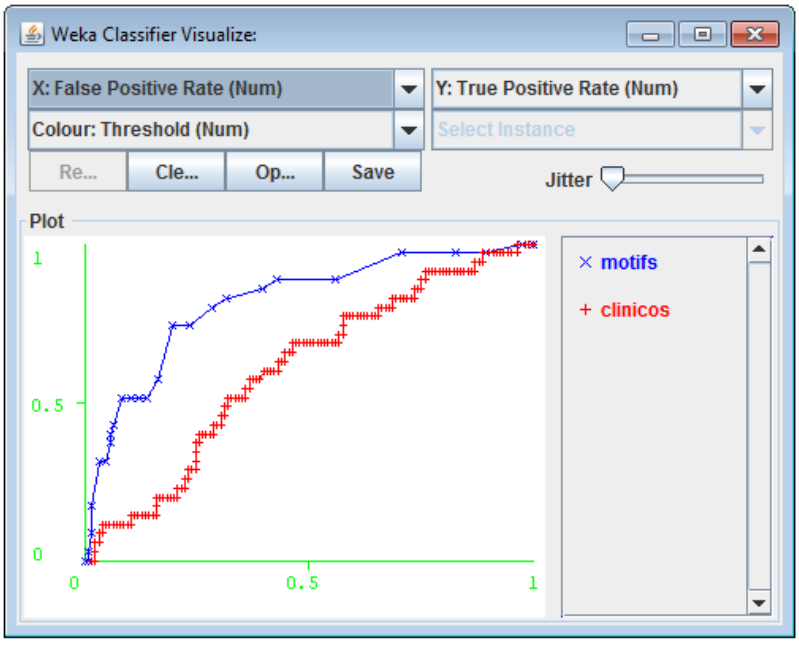

Figura 12 – Análise ROC para a ROC área (AUC)

No que diz respeito à combinação dos parâmetros dos *motifs* com os dados clínicos, pelo que nos foi possível observar, não vieram acrescentar melhorias aos resultados obtidos utilizando apenas os *motifs* como atributos.

Na Tabela 75 apresentam-se dos resultados de *Recall* obtidos usando os *motifs* (duas primeiras linhas da tabela), os dados clínicos (terceira linha) e os atributos combinados (última linha da tabela). Como podemos observar a combinação de atributos não melhorou o resultado do *Recall*.

Tabela 75 – Resumo dos melhores da *Recall* para atributos combinados

Dados	**Algoritmo**	**Parâmetros**	***Recall***
***Motifs*(tabela8K40W10)**	*Random Forest*	*trees*=25, *features*=1, *seed*=1	**0.861**
***Motifs*(tabela8K40W10)**	*Rotation Forest*	*iterations*=30	**0.861**
Clinicos(4ª análise)	*Rotation Forest*	*trees*=20, *features*=2, *seed*=1	**0.791**
Combinados(tabela4K40W30_comb)	*Rotation Forest*	*iterations*=20	**0.861**

Na Tabela 76 apresentam-se os resultados obtidos pelas diferentes abordagens, avaliando os resultados pela medida *F-Measure*. Neste caso, a combinação de atributos proporcionou uma ligeira melhoria de resultados (3ª casa decimal), como podemos observar na última linha da tabela.

Tabela 76 – Resumo dos melhores resultados *F-Measure* para atributos combinados

Dados	**Algoritmo**	**Parâmetros**	**F-meas**
***Motifs*(tabela4K40W30)**	*Rotation Forest*	*iterations*=30	**0. 841**
clínicos(4ª análise)	*Random Forest*	*trees*=20, feat=2, s=1	**0.743**
Combinados(tabela4K40W30_comb)	*Rotation Forest*	*iterations*=30	**0.844**

Por último, apresentamos os resultados para a medida área ROC para cada uma das abordagens. Podemos novamente observar, na Tabela 77, que a combinação de atributos não melhorou os resultados.

Tabela 77 – Resumo dos melhores resultados de área ROC para atributos combinados

Dados	Algoritmo	Parâmetros	Área ROC
***Motifs*(tabela8K40W10)**	*Random Forest*	*trees*=40, *features*=2, *seed*=1	**0.816**
Clínicos(1ª análise)	*Random Forest*	*trees*=25, *features*=1, *seed*=1	**0.602**
Combinados (tabela4K40W40_comb)	*Random Forest*	*trees*=20, *features*=2, *seed*=1	**0.808**

De forma análoga ao que fizemos para os atributos *motifs* e clínicos, fizemos também uma análise das curvas ROC para os melhores modelos, segundo estas medidas.

Na Figura 13, podemos observar que as curvas ROC para os modelos com melhor *Recall*, quando comparamos os atributos *motifs* com os atributos combinados, se sobrepõem. Assim, pelo que podemos observar, juntar aos *motifs* os dados clínicos não parece ter melhorado os resultados.

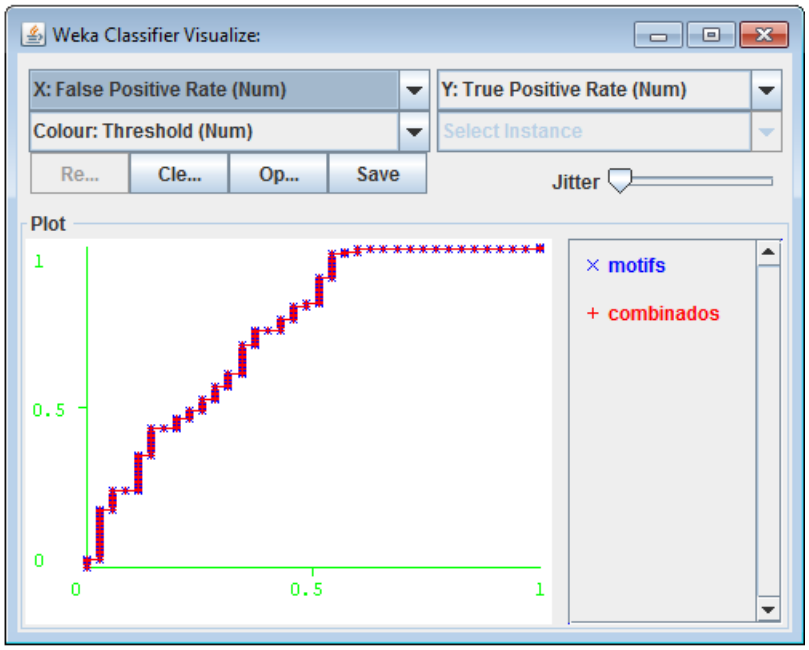

Figura 13 – Análise ROC para os melhores modelos segundo o *Recall* para atributos combinados

No caso da análise ROC para os melhores modelos segundo a área ROC, apesar dos *motifs* e dos atributos combinados obterem resultados médios idênticos, existe uma região onde a curva ROC, correspondente aos atributos combinados ('+', a vermelho), é claramente superior à outra curva (Figura 14). Nessa região, conseguimos ver que é possível, combinando os atributos *motifs* e clínicos, obter uma TVP elevada (cerca de 80%) para uma TFP relativamente baixa (abaixo de 20%).

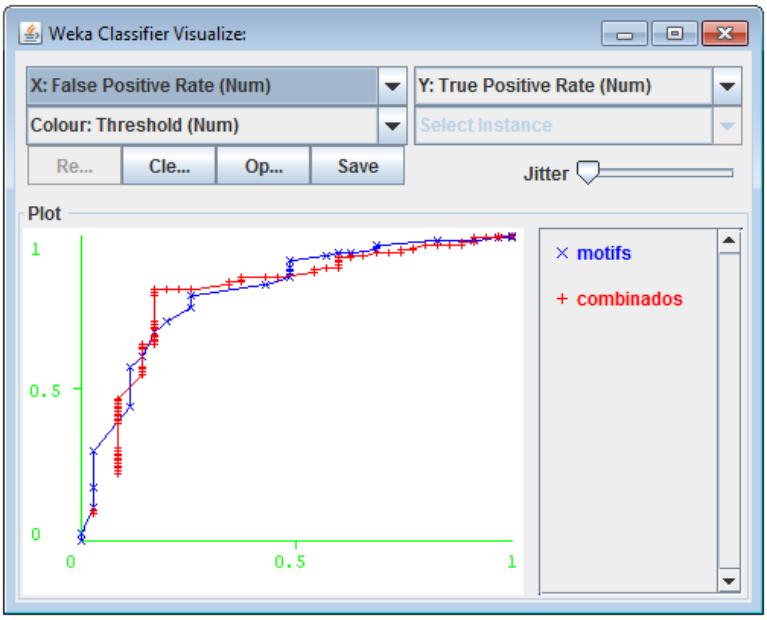

Figura 14 – Análise ROC para a área ROC para atributos combinados

Como para os atributos combinados o resultado da medida F assume um valor ligeiramente melhor (apenas na terceira casa decimal) do que o valor obtido com os *motifs*. Foi realizada, também, a análise ROC dos melhores resultados obtidos segundo esta medida. Apesar da pequena diferença na medida F para os atributos combinados (Tabela 76), como se pode observar na Figura 15, as curvas ROC estão praticamente sobrepostas.

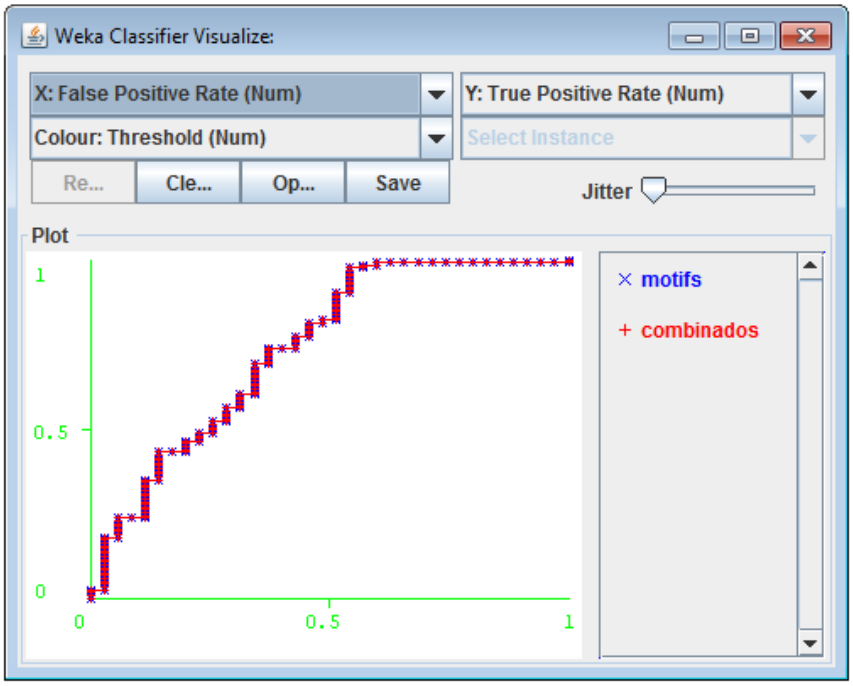

Figura 15 – Análise ROC para a *F-Measure* para atributos combinados

6 Conclusões e trabalho futuro

Nesta dissertação apresenta-se um trabalho onde foram exploradas diferentes abordagens com o objetivo de identificar automaticamente a existência de patologias cardíacas a partir de sons e de dados clínicos de pacientes, recolhidos num hospital.

Numa primeira fase, explorou-se a abordagem usando *motifs*. Nesta abordagem trataram-se os ficheiros áudio de sons cardíacos que se sabia pertencerem a uma de duas classes: Normal ou Anormal. Esta abordagem já tinha sido utilizada em trabalhos relacionados no projeto em que o trabalho descrito nesta dissertação se insere.

Numa segunda fase, estudámos os dados clínicos anotados dos pacientes Os nossos resultados indicaram que, para os atributos considerados, o classificador tem um poder discriminante bastante inferior aos atributos baseados em *motifs* retirados dos sons cardíacos.

Numa terceira fase, estudámos também os dados combinados, juntando os atributos dos *motifs* com os resultados clínicos. Apesar de os resultados se mostrarem semelhantes aos conseguidos com os dados somente obtidos pelos *motifs,* analisando as curvas ROC observamos que existe uma região onde a curva ROC, correspondente aos atributos combinados, é claramente superior à curva correspondente aos *motifs*, conseguindo obter uma TVP elevada (cerca de 80%) para uma TFP relativamente baixa (abaixo de 20%).

Apesar dos resultados obtidos com os atributos combinados indicarem que não compensa adicionar os dados clínicos aos *motifs*, o resultado da observação da curva ROC parece promissor e pensamos que valerá a pena adquirir mais dados para melhorar a tarefa de classificação.

6.1 Trabalho futuro

Em termos de continuação deste trabalho foram consideradas as seguintes tarefas:

- Tentar obter mais e melhores sons para análise utilizando diferentes equipamentos e em ambientes sem tanto ruído,
- Utilizar a ferramenta Audacity para melhorar os sons obtidos e retirar o ruído. Tentar usufruir melhor das funcionalidades que a ferramenta possuí,
- Alargar esta análise outro tipo de batimentos de forma a conseguir mais classes de análise,
- Fazer recolha de dados clínicos de adultos e verificar se a tendência de padrões definidos para as crianças pode ser aplicada em adultos

Referências

[Arathy, 2013] Arathy R, Gowriprabha V, Vysakh V, PC based Heart Sound Monitoring System, International Journal of Computer Applications, Vol 3 (16), 0975-8887, 2013.

[Babaei, 2009] S. Babaei and A. Geranmayeh. Heart sound reproduction based on neural network classi_cation of cardiac valve disorders using wavelet transforms of pcg signals. Comp. in Bio. and Med., 39(1):8-15, 2009

[Bentley, 2011] P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. www.peterjbentley.com/heartchallenge, 2011

[Breiman, 2001] L. Breiman. *Random Forest*s. Machine Learning, 45(1):5-32, 2001.

[Castro, 2010a] N. Castro and P. J. Azevedo. Multiresolution *Motif* Discovery in Time Series. In SDM, pages 665-676,2010.

[Castro, 2010b] N. Castro. Multiresolution *motif* discovery in time series website. http://www.di.uminho.pt/ castro/*MrMotif*. [último acesso: Out 2014]

[DEI, 2006] Departamento de Engenharia Informática, http://www.dei.isep.ipp.pt/ [último acesso: Jul 2015]

[Ferreira, 2006] P. G. Ferreira, P. J. Azevedo, C. G. Silva, and R. M. M. Brito. Mining approximate *motifs* in time series. In Discovery Science, pages 89-101, 2006.

[Ferreira, 2012] Ferreira, P.; Pereira, D.; Mourato, F.; Mattos, S.; Cruz-Correia, R.; Coimbra, M.; Dutra, I., "Detecting cardiac pathologies from annotated auscultations," in Computer-Based Medical Systems (CBMS), 2012 25th International Symposium on , vol., no., pp.1-6, 20-22 June 2012.

[Gama, 2015] Gama, João ; de Carvalho, A. C. P. L. F. ; Faceli, Katti ; Lorena, A. C. ; OLIVEIRA, M.. Extração de Conhecimento de Dados *Data mining*. 2 ed. Edições Sílabo, 2015.

[Gomes, 2012] http://www.peterjbentley.com/heartworkshop/challengepaper1.pdf [último acesso: Outubro 2015]

[Gomes, 2013a] E. F. Gomes, P. J. Bentley, E. Pereira, M. Coimbra, and Y. Deng. Classifying heart sounds - approaches to the pascal challenge. In D. Stacey, J. Sol_e-Casals, A. L. N. Fred, and H. Gamboa, editors, HEALTHINF, pages 337{340. SciTePress, 2013.

[Gomes, 2013b] E. F. Gomes, A. M. Jorge, and P. J. Azevedo. Classifying heart sounds using multiresolution time series *motifs*: an exploratory study. In B. C. Desai, A. M. de Almeida, and S. P. Mudur, editors, C3S2E, 23-30. ACM, 2013.

[Gomes, 2014] Gomes, E. F., Jorge, A. M. & Azevedo, P. J., 2014. Classifying heart sounds using SAX *motifs*, *Random Forest*s and text mining techniques. *In 18th International Database Engineering & Applications Symposium, IDEAS 2014*, 7-9 July, pp. 334-337.

[Groch, 1992] M. W. Groch, J. R. Domnanovich, and W. D. Erwin.
A new heart-sounds gating device for medical imaging. IEEE Transactions on Biomedical Engineering, 39:307-310, 1992.

[Gupta, 2005] C. N. Gupta, R. Palaniappan, S. Rajan, S. Swaminathan, and S. M. Krishnan. Segmentation and classi_cation of heart sounds. In Canadian Conference on
Electrical and Computer Engineering, pages 1674-1677, 2005.

[Gupta, 2007] C. N. Gupta, R. Palaniappan, S. Swaminathan, and S. M. Krishnan. Neural network classi_cation of homomorphic segmented heart sounds. Applied Soft Computing, 7:286-297, 2007.

[Heart, 2014] http://www.texasheart.org/HIC/Anatomy/systole.cfm acedido em 07/10/14.

[Karnath, 2002] B. Karnath and W. Thornton. Auscultation of the heart. Hospital Physician, 38(9):39-43, sep. 2002

[Karraz, 2006] G. Karraz and G. Magenes. Automatic Classification of Heartbeats using Neural Network Classifier based on a Bayesian Framework. In Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pages 4016-4019, 2006.

[Kampouraki, 2009] A. Kampouraki, G. Manis, and C. Nikou. Heartbeat Time Series Classification With Support Vector Machines. IEEE Transactions on Information Technology in Biomedicine, 13:512-518, 2009.

[Kao, 2011] W.-C. Kao, L.-W. Cheng, C.-Y. Chien, and W.-K. Lin. Robust brightness measurement and exposure control in real-time video recording. IEEE T. Instrumentation and Measurement, 60(4):1206-1216, 2011.

[Kumar, 2006] D. Kumar, R. Carvalho, M. Antunes, R. Gil, J. Henriques, and L. Eugenio. A New Algorithm for Detection of S1 and S2 Heart Sounds. In International Conference on Acoustics, Speech, and Signal
Processing, volume 2, 2006.

[Lee V., 2005] Lee, Valentino. Aplicações móveis: arquitectura, projecto e desenvolvimento, pages 23-36, 2005

[Liang, 1997] H. Liang, S. Lukkarinen, and I. Hartimo. Heart sound segmentation algorithm based on heart sound envelogram. In Computers in Cardiology 1997, pages 105-108, Sep 1997.

[Lijuan, 2012] Jia, Lijuan; Song, Dandan; Tao, Linmi; Lu, Yao.Heart Sounds Classification with a Fuzzy Neural Network Method with Structure Learning, Springer Berlin Heidelberg, 130-140, 2012.

[Lin, 2002] J. Lin, E. Keogh, S. Lonardi, and P. Patel. Finding *motifs* in time series. In Proceedings of the 2nd Workshop on Temporal *Data mining*, pages 53-68,
2002.

[Mandeep, 2013] Singh, Mandeep; Cheema, Amandeep. Heart Sounds Classification using *Feature* Extraction of Phonocardiography Signal. International Journal of computer Applications, vol 7(4),0975-8887, 2013.

[Marques, 2013] N. Marques, R. Almeida, A. Rocha, and M. Coimbra. Exploring the stationary wavelet transform detail coefficients for detection and identification of the S1 and S2 heart sounds. In Computing in Cardiology
Conference (CinC), 2013, pages 891{894, Sept 2013.

[Oliveira, 2014] Soraia Cruz Oliveira, Elsa Ferreira Gomes, Alípio Mário Jorge: Heart sounds classification using *motif* based segmentation. IDEAS 2014: 370-371

[Oliveira, 2014b] Soraia Cruz Oliveira, Dissertação de mestrado. Classificação de Sons Cardíacos usando *motifs*: Desenvolvimento de uma aplicação móvel. 2014

[Ölmez, 2013] Tamer Ölmez, Zümray Dokur. Classification of heart sounds using an artificial neural network, Pattern Recognition Letters, Volume 24, Issues 1–3, Pages 617-629, ISSN 0167-8655, January 2003.

[Palm, 2010] Palm, D., Burns, S., Pasupathy, T., Deip, E., Blair, B., Flynn, M., Drewek, A., Sjostrand, M., Stephenson, B., and Nordehn, G. (2010). Artificial Neural Network Analysis of Heart Sounds Captured From an Acoustic Stethoscope and Emailed Using iStethoscopePro. Journal of Medical Devices, 4(2):027531+.

[Pereira, 2011] D. Pereira, F. Hedayioglu, R. Correia, T. Silva, I. Dutra, F. Almeida, S. Mattos, and M. Coimbra. Digiscope - unobtrusive collection and annotating of auscultations in real hospital environments. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, pages 1193-1196, 30 2011-sept. 3 2011.

[Shieh and Keogh, 2008] Shieh, J. & Keogh, E., 2008. iSAX: Indexing and Mining Terabyte Sized Time Series.

[Strunic, 2007] S. L. Strunic, F. Rios-Gutierrez, R. Alba-Flores, G. Nordehn, and S. Burns. Detection and Classification of Cardiac Murmurs Using Segmentation Techniques and Artificial Neural Networks. In IEEE Symposium on Computational Intelligence and *Data mining*, pages 128-133, 2007.

[Witten, 2005] Witten, I. H. and Frank, E. (2005). *Data mining*: Practical Machine Learning Tools and Techniques.

[Yankov, 2007] D. Yankov, E. J. Keogh, J. Medina, B. Y. chi Chiu, and V. B. Zordan. Detecting time series *motifs* under uniform scaling. In P. Berkhin, R. Caruana, and X. Wu, editors, KDD, pages 844{853. ACM, 2007.

[Wang, 2014] Yan Wang, Wenzao Li, Jiliu Zhou, Xiaohua Li, Yifei Pu, Identification of the normal and abnormal heart sounds using wavelet-time entropy *features* based on OMS-WPD, Future Generation Computer Systems, Volume 37, July 2014,Pages 488-495, http://dx.doi.org/10.1016/j.future.2014.02.009.

Referências URL:

[árvores de decisão]	http://home.iscte-iul.pt/~dmt/publ/tx/Arvores_de_Decisao_INDEG_ISCTE.pdf último acesso: Outubro 2015
[árvores de decisão 2]	http://web.tecnico.ulisboa.pt/ana.freitas/bioinformatics.ath.cx/bioinformatics.ath.cx/indexf23d.html?id=199 último acesso: Outubro 2015
[árvores de decisão 3]	http://www.novaims.unl.pt/docentes/vlobo/escola_naval/SAD/SAD_EN_8_arvores.pdf último acesso: Outubro 2015
[audacity]	http://sourceforge.net/projects/audacity/ ultimo acesso: Outubro 2015
[Data mining]	http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm último acesso :Outubro 2015
[HeartSounds]	http://allfornursing.blogspot.pt/2012_10_01_archive.html último acesso :Outubro 2015
[Java]	https://www.java.com último acesso: Outubro 2015
[J48]	http://www.d.umn.edu/~padhy005/Chapter5.html último acesso: Outubro 2015
[MultiLayer Perceptron]	http://neuroph.sourceforge.net/tutorials/MultiLayer Perceptron .html último acesso: Outubro 2015
[netbeans]	https://netbeans.org/ ultimo acesso: Outubro 2015
[Random Forest]	https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm último acesso: Outubro 2015
[Rotation Forest]	http://link.springer.com/chapter/10.1007%2F978-3-540-72523-7_46#page-1 último acesso: Outubro 2015
[support vector machine]	http://research.microsoft.com/pubs/69644/tr-98-14.pdf último acesso: Outubro 2015
[roc area image	https://es.wikipedia.org/wiki/Curva_ROC#/media/File:ROC_space-2.png
[Weka]	http://www.cs.waikato.ac.nz/ml/weka/ último acesso: Outubro 2015
[Weka informação]	http://www.ibm.com/developerworks/br/opensource/library/os-weka1/ último acesso: Outubro 2015
[Weka data analysis]	http://www.cs.usfca.edu/~pfrancislyon/courses/640fall2015/WekaDataAnalysis.pdf