Ana Sofia Besteiro Lebreiro

DEEP LEARNING PARA SEGMENTAÇÃO E CLASSIFICAÇÃO AUTOMÁTICA DE SONS RESPIRATÓRIOS ADVENTÍCIOS

UNIVERSIDADE D COIMBRA

1 2 9 0

UNIVERSIDADE D COIMBRA

Ana Sofia Besteiro Lebreiro

DEEP LEARNING PARA SEGMENTAÇÃO E CLASSIFICAÇÃO AUTOMÁTICA DE SONS RESPIRATÓRIOS ADVENTÍCIOS

Dissertação no âmbito do Mestrado me Engenharia Informática, especialização em Sistemas Inteligentes, orientada pelo Professor Doutor Rui Pedro Pinto de Carvalho e Paiva e pelo Professor Doutor Paulo Fernando Pereira de Carvalho apresentada à Faculdade de Ciências e Tecnologia / Departamento de Engenharia Informática.

Outubro de 2021

Departamento de Engenharia Informática

Deep Learning para Segmentação e Classificação Automática de Sons Respiratórios Adventícios

Ana Sofia Besteiro Lebreiro

Dissertação no âmbito do Mestrado em Engenharia Informática, especialização em Sistemas Inteligentes, orientada pelo Professor Doutor Rui Pedro Pinto de Carvalho e Paiva e pelo Professor Doutor Paulo Fernando Pereira de Carvalho apresentada à Faculdade de Ciências e Tecnologia / Departamento de Engenharia Informática.

Outubro de 2021

UNIVERSIDADE D
COIMBRA

Resumo

As patologias do foro respiratório são das mais mortíferas em todo o mundo. A maioria dos pacientes com doenças respiratórias possuem uma respiração caraterizada pela existência de sons adventícios, como sibilâncias e fervores, ao longo do ciclo respiratório.

Neste trabalho propomos estudar a aplicabilidade de arquiteturas de aprendizagem profunda para classificar e segmentar automaticamente sons adventícios (sibilâncias e fervores) e outros sons presentes nos ciclos respiratórios dos pacientes. Todos os ficheiros de áudio utilizados nesta tese provêm da base de dados *Respiratory Sound Database* [1]. Definiram-se arquiteturas de aprendizagem profundas como CNN, LSTM e outras variantes, assim como modelos clássicos SVM e k-NN, em conjunto com técnicas de processamento de som, de forma a criar modelos otimizados, tanto em ambientes controlados como em ambientes reais de operação.
Na fase de classificação de eventos pré segmentados manualmente, são comparadas abordagens clássicas de engenharia de atributos e aprendizagem computacional, e modelos de classificação de aprendizagem profunda com diferentes tipos de atributos de entrada. A metodologia adotada passou por um pré-processamento dos ficheiros de áudio através da filtragem do som de forma a concentrar a análise na gama de frequências relevantes para o estudo. A partir do som filtrado são extraídos atributos para construir diferentes conjuntos de dados a aplicar nos classificadores. De forma a conhecer melhor os dados e a proceder-se a uma seleção de atributos, foi realizada uma análise exploratória. A seleção é realizada a partir de três algoritmos distintos para compreender o desempenho dos classificadores com diferentes atributos evitando redundância de dados. Os classificadores são avaliados a partir do cálculo de várias métricas estatísticas. O classificador CNN alcançou o melhor resultado alcançando 81% de exatidão e 81% de sensibilidade superando os classificadores clássicos com resultados máximos de 76,17% de exatidão.

Na fase de segmentação automática e classificação, são explorados diversos mecanismos para aplicar a segmentação sonora, prosseguindo para uma classificação *frame* a *frame* com vista a identificar a localização temporal dos eventos adventícios. A avaliação desta fase é realizada a diferentes níveis: *frame* a *frame*, detetando 51,2% das *frames* de sibilâncias e 50,1% de fervores, por evento e por paciente.

Palavras-Chave: Engenharia de atributos, Sons Respiratórios Adventícios, Classificação, Segmentação, Aprendizagem profunda

Abstract

Respiratory pathologies are among the deadliest in the world. Most patients with respiratory diseases have a breathing characterized by adventitious sounds, such as wheezes and crackles, throughout the respiratory cycle.

In this work we propose to study the applicability of deep learning architectures to automatically classify and segment adventitious sounds (wheezes and crackles) and normal sounds existing in patients' breathing cycles. All audio files used in this thesis belong to the Respiratory Sound Database [1]. Several deep learning architectures such as CNN, LSTM and other variants are evaluated, as well as some classical models, i.e., SVM and k-NN, together with sound processing techniques, in order to create optimized models, both in controlled environments and in real-life operation scenarios.

In the phase of event classification of manually segmented events, classical feature engineering and machine learning approaches are compared against deep learning classification models with different types of input features. The methodology adopted involves the pre-processing of audio files through sound filtering, in order to focus the analysis on the range of frequencies relevant to the study. From some filters, resources are extracted to build different datasets to be applied to the classifiers. In order to better understand the data and proceed with a selection of resources, an exploratory analysis is performed. The selection is carried out using three different algorithms to understand the performance of the classifiers with different features, avoiding data redundancy. The classifiers are evaluated by calculating various statistical metrics. The CNN classifier obtained the best results reaching 81% accuracy and 81% recall, beating the classic classifiers which obtained a maximum result of 76,17% accuracy.

In the automatic segmentation and classification phase, several mechanisms are explored to perform sound segmentation, proceeding to a frame-by-frame classification in order to identify the temporal location of adventitious events. The evaluation of this phase is carried out at different levels: frame by frame, which detected 51,2% of the wheezes frames and 50,1% of the crackles, by event and by patient.

Keywords: Features Engineering, Adventitious Respiratory Sounds, Classification, Segmentation, Deep Learning

Agradecimentos

Em primeiro lugar, aos meus orientadores Professor Doutor Rui Pedro Paiva e Professor Doutor Paulo Carvalho pelo apoio continuo ao longo deste ano, pela colaboração, opiniões e críticas construtivas muito importantes para a minha aprendizagem.

Aos CISUC pela oportunidade e confiança que depositaram em mim ao concederem-me uma bolsa de investigação.

Aos meus parceiros de projeto Bruno Rocha e Diogo Pessoa que me integraram muito bem na equipa. Agradeço o facto de estarem sempre disponíveis para me auxiliar e esclarecer qualquer dúvida ou problema.

A todos os meus amigos e colegas de curso, com quem pude aprender muito tanto a nível académico como pessoal. Alexandre, Caio, Martinho e Sara agradeço por todas as noitadas a estudar, pelos momentos parvos e de risada e, principalmente, pela amizade ao longo dos anos. Sou grata por tê-los conhecido e são pessoas que levo para a vida.

À Vera Estanqueiro, a pessoa que esteve sempre ao meu lado, nos momentos de felicidade, de stress e tristeza. Que teve muita paciência comigo e que nunca me deixou desistir, ajudando-me e motivando-me para ser melhor todos os dias.

A toda a minha família, em particular aos meus tios(as) e avós pelo carinho e pelas palavras de incentivo e confiança que muitas vezes precisava de ouvir.

E por último, um agradecimento muito especial aos meus pais e irmão que sempre acreditaram em mim. Pelo apoio incondicional nesta e em todas as minhas etapas, por serem os melhores modelos que alguma vez poderia ter e porque sem eles este caminho seria impossível de ser concretizado. Obrigado por tudo o que sou hoje!

Conteúdo

Acrónimos

AED Análise Exploratória de Dados

ANN *Artificial Neural Network*

CAS *Continuous Adventitious Sounds*

CNN *Convolutional Neural Network*

DAS *Discontinuous Adventitious Sounds*

DNN *Deep Neural Network*

DPOC Doença Pulmonar Obstrutiva Crónica

DWT *Discrete Wavelet Transform*

EMD *Empirical Mode Decomposition*

FFT *Fast Fourier Transform*

GMM *Gaussian Mixture Modelling*

GRU *Gated Recurrent Units*

k-NN *k-nearest neighbors algorithm*

LR *Logistic Regression*

LSTM *Long Short-Term Memory*

MFCC *Mel-frequency cepstral coefficients*

MRMR *Minimun Redundancy Feature Selection*

NLP *Natural Language Processing*

OST *Optimized S-transform*

ReLU *Rectified Linear Unit*

RNN *Recurrent Neural Network*

SA Sons Adventícios

SR Sons Respiratórios

STFT *Short-Time Fourier Transform*

SVM *Support Vector Machine*

Lista de Figuras

Lista de Tabelas

Introdução

Este capítulo pretende enquadrar o leitor no tema do trabalho proposto. É realizada uma contextualização, são discutidas as motivações subjacentes ao trabalho e apresentados os objetivos principais da dissertação e a estrutura do presente documento.

1.1 Motivação e Contexto

As patologias do foro respiratório são uma grande preocupação médica, sendo estas as principais causadoras de milhões de mortes todos os anos. A Doença Pulmonar Obstrutiva Crónica (DPOC) e as infeções do trato respiratório inferior são, há várias décadas, das doenças mais mortíferas a nível mundial [2], como é ilustrado na Figura 1. Em 2018, as doenças do aparelho respiratório apresentavam uma taxa de mortalidade de 11,7% em Portugal [3], o que causa uma enorme pressão sob os sistemas saúde [4] e, consequentemente, a nível económico e social. Esta situação impõe uma necessidade de investir e procurar novas abordagens que permitam contornar esta situação num futuro próximo.

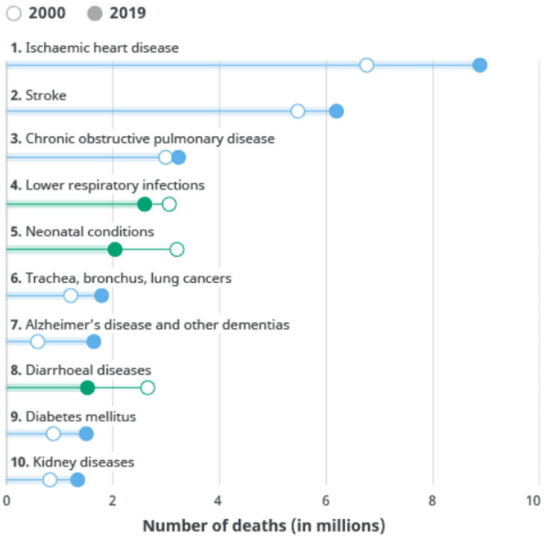


Figura 1 – As 10 principais causas de morte mundiais [2]

As doenças respiratórias são muitas vezes caraterizadas por sons anormais /sons adventícios. Atualmente, os médicos ainda dependem do estetoscópio, por ser pouco intrusivo e economicamente acessível, para analisar primeiramente a respiração do paciente e identificar alguma possível patologia detetando sons adventícios. Contudo, o estetoscópio é um instrumento pouco preciso sendo a sua eficácia dependente de inúmeros fatores como: nível de acuidade auditiva do utilizador, ruído ambiental, movimentações do paciente e concentração. Devido a estas limitações uma das opções (não ideal) mais utilizadas é a realização de exames ao tórax através de raio-X, sendo este mais intrusivo para o paciente

devido à radiação a que está sujeito, caro e pouco objetivo, podendo novamente ocorrerem erros humanos a nível da análise visual do raio-X [5].
Para ultrapassar as limitações das soluções apresentadas, a área de aprendizagem computacional surgiu como um possível suporte a este problema, principalmente devido ao investimento e evolução do grande domínio de Inteligência Artificial ao longo dos últimos anos. Devido ao aperfeiçoamento de vários dispositivos de extração de sons pulmonares é possível analisar computacionalmente o áudio extraído tendo em vista a aplicabilidade nos algoritmos de aprendizagem computacional. Aquando da recolha de áudio, o paciente respira durante um determinado período e, de seguida, os especialistas analisam o áudio e anotam o tipo de sons presentes: som normal, adventícios ou apenas ruído. Toda esta análise é um trabalho árduo, moroso e apresenta uma elevada variabilidade inter e intra anotador. Neste contexto, algoritmos de aprendizagem profunda (*deep learning*), juntamente com técnicas de extração de informação e análise dos sinais de som, têm sido explorados para classificar e identificar automaticamente eventos ao nível da saúde.

Em particular, salienta-se o grande poder das redes neuronais profundas e a falta de exploração desta potencialidade na segmentação automática de sons respiratórios (SR), sendo esta uma técnica com elevado potencial para melhorar os resultados atuais do estado da arte. No âmbito desta proposta estamos interessados em explorar novas combinações de mecanismos de exploração de sons respiratórios (pré-processamento e extração de atributos) juntamente com diversas arquiteturas de aprendizagem profunda para classificação e segmentação de eventos respiratórios, com vista a alcançar resultados futuramente aplicáveis em centros médicos, facilitando o trabalho dos profissionais de saúde.

Existem duas motivações principais por detrás deste trabalho. Primeiramente, a extrema necessidade do mercado face à prevenção de doenças respiratórias. Quanto mais rápido e fácil for detetar patologias respiratórias, mais rapidamente os pacientes podem iniciar o tratamento, o que leva a uma tentativa da diminuição da taxa de mortalidade em todo o mundo. Igualmente importante é a relevância dos resultados obtidos para trabalhos futuros, quer referente a doenças respiratórias quer ao estudo de classificação e segmentação de som. Se forem obtidos bons resultados nos algoritmos de classificação desenvolvidos em comparação com os do estado da arte, esta dissertação pode ser um bom ponto de partida para a criação de modelos de aprendizagem profunda para aplicar a segmentação e na classificação dos sons, tendo uma previsão dos resultados que estas irão alcançar.

Estas motivações estão diretamente correlacionadas com os nossos objetivos. Temos o intuito de responder a diversas questões:

- Qual o melhor método de pré-processamento a aplicar aos sinais dos sons respiratórios?
- De entre os classificadores apresentados nesta dissertação, quais as arquiteturas de aprendizagem profunda que obtêm melhores resultados a classificar e segmentar eventos a partir de ciclos respiratórios?
- Será que as arquiteturas conseguem ultrapassar a barreira de ruído existente nos ficheiros?
- Existem algumas arquiteturas que funcionam melhor com apenas um determinado tipo de atributos?

1.2 Objetivos e Abordagens

Esta dissertação tem como objetivo principal a utilização de abordagens de aprendizagem profunda para realizar a segmentação e classificação de sons respiratórios adventícios, tanto em ambientes controlados como em contextos sujeitos a perturbações acústicas ambientais. A primeira fase da tese vai de encontro a uma fase de adaptação ao tema e ao trabalho que tem sido realizado ao longo dos anos pelos restantes elementos da equipa de investigação, tendo de seguida como foco uma revisão da literatura existente na área. Tendo esta base, serão desenvolvidas técnicas clássicas e de aprendizagem profunda. Em particular, os objetivos a atingir nesta dissertação são:

- Análise exploratória dos atributos recolhidos;
- Otimização do pipeline de pré-processamento de sinais respiratórios;
- Definição de uma arquitetura de aprendizagem profunda para a segmentação e classificação de eventos respiratórios usando sons respiratórios e a sua avaliação comparativa com metodologias clássicas;
- Avaliação dos modelos em contextos controlados e em contextos de operação reais.

Com o foco de cumprir os objetivos acima definidos, o trabalho foi organizado em várias fases. Cada uma das etapas do trabalho resultou numa conclusão que é extremamente útil para o desenvolvimento da fase seguinte.

Primeiramente, foi necessária uma pesquisa profunda e o estudo de conceitos relacionados com sons respiratórios, sons adventícios e as suas características. Também foi realizada uma aprendizagem das estruturas dos sinais sonoros, frequências e amplitudes mais comuns nesta fonte de informação.

O objetivo principal foca-se na classificação e segmentação usando técnicas de aprendizagem profunda. Para o alcançar, é importante comparar os resultados de aprendizagem profunda com abordagens clássicas. Como tal, essa comparação é um objetivo secundário bastante importante que complementará o principal.

Posto isto, para realizar uma classificação de sons respiratórios é necessário obter um conjunto de dados consistente, diversificado, com uma grande quantidade de dados e o mais realista possível. Estes fatores destacam-se uma vez que para criar um bom classificador é necessário treiná-lo com uma grande quantidade de dados de forma que exista a possibilidade de dividir o conjunto de dados em treino e teste, contendo assim em cada uma das fases uma grande quantidade de dados. Estes, por sua vez, devem ser diversificados para que os classificadores aprendam a reconhecer os diferentes padrões de características dos sinais dos sons respiratórios. Como temos o objetivo de aplicar o resultado desta tese à vida real, os dados utilizados devem ser recolhidos com base em pacientes reais.

Neste trabalho, o problema de segmentação e classificação de sons adventícios foi dividido em dois sub-problemas. O primeiro focou-se na classificação de eventos previamente segmentados manualmente. No segundo, todo o processo, tanto de segmentação como de classificação, é automático. Em ambos os problemas foi utilizada a *Respiratory Sound Database* (RSD) [1], a qual contém anotações realizadas por especialistas. Esta é a base de dados pública de maior dimensão e mais realista conhecida.

Utilizaram-se diversas formas de entrada para as redes sendo uma delas as características dos segmentos de som (eventos). Para isso, existiu uma extração de atributos

utilizando a *MIRtoolbox 1.7.2* do MATLAB que contém métodos especializados na extração de atributos de áudio. Esta fase está descrita detalhadamente no capítulo 4.

A análise exploratória de dados é necessária de forma a ter um conhecimento mais aprofundado do conjunto de dados. Nesta fase, foram aplicados métodos estatísticos para averiguar a distribuição dos dados e como estes estão relacionados uns com os outros. O método escolhido para verificar a relação dos atributos entre eles e em relação ao alvo final foi a matriz de correlação de Pearson. Para verificar se os dados seguem uma distribuição normal foi utilizado o teste de Shapiro. As conclusões finais desta fase serão aplicadas na fase seguinte de seleção de atributos e, deste modo, escolher os métodos mais adequados para realizar a sua seleção.

A seleção de atributos é uma fase bastante importante visto que irá ter um grande peso nos resultados dos classificadores que irão ser implementados. O conjunto de dados a aplicar nos classificadores terá de possuir apenas os melhores atributos que caraterizam os eventos sonoros, uma vez que o poder computacional das máquinas é limitado. A seleção dos melhores atributos dependem do critério de seleção dos selecionadores aplicados, descritos no capítulo 4.

Os classificadores de aprendizagem profunda escolhidos para implementar e testar, tendo em conta o estado da arte, foram: Rede Neural Convolucional (*Convolutional Neural Network* - CNN) [1], Redes de memória de curto e longo prazo (*Long Short-Term Memory* - LSTM) e GRU, variantes de Redes Neurais Recorrentes (*Recurrent Neural Network* **-** RNN). Os dados fornecidos aos classificadores de aprendizagem profunda serão os atributos referidas anteriormente, imagens de diversos espectrogramas, espectrogramas de mel, o sinal sonoro e onduletas de sons adventícios e outros sons existentes nos ficheiros. Os classificadores clássicos escolhidos foram: SVM e k-NN. Neste caso, os dados de entrada das redes basearam-se nos mesmos atributos referidas anteriormente.

A fase de treino foi desenvolvida de forma bastante pormenorizada, visto que o conjunto de dados foi bastante variado e além disso teve de existir um balanceamento de classes antes de a iniciar. O balanceamento foi realizado a partir de técnicas de subamostragem, aumento de dados, sobreamostragem, manipulando algumas características das ondas sonoras dos eventos como altura, intensidade e frequência. Este processo irá ser descrito mais detalhadamente no capítulo 4.

O segundo problema baseava-se na segmentação e classificação de eventos. A criação do conjunto de dados de treino e teste foi trabalhada de maneira a esta ser constituída por *frames* de espectrogramas de mel, usando diferentes parâmetros para efetuar a segmentação. Após a realização dos classificadores de ambas as abordagens é necessário avaliá-los de forma a perceber se estes correspondem às expectativas, se alcançam bons resultados ao classificar uma grande percentagem dos sons ou se, pelo contrário, possuem uma grande taxa de erro. Deste modo, serão aplicadas diferentes métricas de desempenho como: precisão, sensibilidade, especificidade, F1-Score e exatidão. Também será desenvolvida uma análise gráfica aplicando a matriz de confusão aos vários modelos criados. Relativamente à análise da segmentação foram criadas métricas para avaliar o modelo ao nível da classificação das *frames*, eventos e pacientes.

[1] Neste documento, optou-se por utilizar algumas expressões ou acrónimos em inglês tanto para este como para outros termos (e.g., *frame*, etc.), em virtude de esta ser uma prática generalizada.

1.3 Estrutura do documento

Este relatório encontra-se estruturado em 6 capítulos:

- **Capítulo 1 Introdução** – É introduzido o tema da dissertação, assim como seus os objetivos, contexto, metodologia e objetivos a alcançar.
- **Capítulo 2 Conceitos e Métodos** – São apresentados alguns conceitos e definições necessárias para a interpretação do trabalho realizado;
- **Capítulo 3 Estado de Arte** – O estado da arte é realizado, contendo uma análise critica e comparativa de diversos algoritmos para a classificação de sons.
- **Capítulo 4 Classificação de eventos adventícios** – É neste capítulo que é descrito todo o trabalho realizado de forma a cumprir o primeiro grande objetivo, a classificação de eventos adventícios previamente segmentados manualmente;
- **Capítulo 5 Segmentação Automática de sons adventícios** – Continuação do trabalho realizado referente ao objetivo da segmentação automática e classificação de sons adventícios;
- **Capítulo 6 Conclusões e Trabalho Futuro** – São apresentadas as principais conclusões sobre as metas alcançadas ao longo do trabalho e algumas direções de trabalhos futuros.

2 Conceitos e métodos

Este capítulo é dedicado à apresentação de diversos conceitos fundamentais para a compreensão desta dissertação. Serão explicados alguns conceitos médicos referentes a sons respiratórios, assim como a descrição de metodologias aplicadas em aprendizagem computacional, mais concretamente sobre aprendizagem profunda, mencionadas ao longo do documento.

2.1 Sons Respiratórios

São considerados sons respiratórios quaisquer sons produzidos a partir dos pulmões, boca ou traqueia (sons vocais não estão incluídos). Este conceito abrange sons produzidos por uma respiração normal, sons adventícios (e.g., sibilâncias e fervores) e outros sons (e.g., tosse, espirros, roncos e sons provocados por músculos respiratórios), tal como se ilustra na Figura 2.

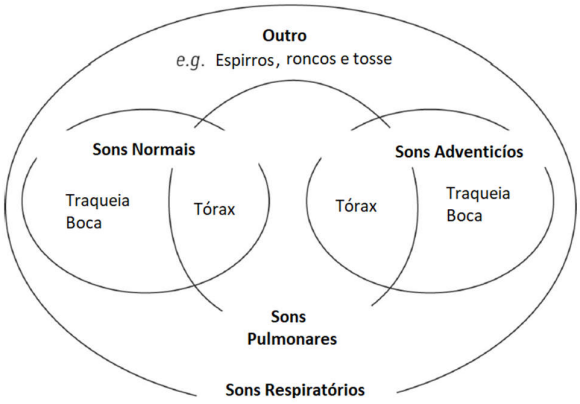


Figura 2 - Relação entre os termos sons respiratórios, sons adventícios, sons pulmonares e sons fisiológicos [4]

Um som respiratório normal escutado a partir do tórax do paciente apresenta uma frequência baixa durante a inspiração e praticamente inaudível ao expirar. Já se estes forem escutados na traqueia, o ato de respirar provoca sons com uma frequência maior e, consequentemente, mais fáceis de detetar [6]. Certas patologias do foro respiratório alteram o modo de respirar das pessoas, tendo como consequência a produção de sons distorcidos ao inalar e/ou exalar em relação ao som de uma respiração de um paciente são. Estes sons denominam-se como sons adventícios.

Os sons adventícios manifestam-se como distorções ou uma sobreposição adicional ao som respiratório saudável [7]. Podem identificar-se como roncos, sibilâncias, grasnidos, fervores entre outros dependendo das suas características sonoras.

A *International Lung Sound Association* em 1976, [8] separou em duas categorias os diferentes sons adventícios (SA) devido às suas diferentes caraterísticas: sons adventícios contínuos (CAS) e sons adventícios descontínuos (DAS).

2.1.1. Sons Adventícios Contínuos

Os sons adventícios contínuos são carateristicamente sons agudos possuindo uma natureza musical, ou seja, são sons uniformes que resultam de vibrações regulares, que permanecem continuamente ao longo da respiração. Normalmente, os CAS possuem uma duração de 250ms ou superior e apresentam uma forma de onda sinusoidal. Estes sons advêm de uma obstrução das vias aéreas provocadas por espasmos do músculo liso brônquico, presença de secreções ou inflamação [9]. As sibilâncias e os roncos são dois tipos de CAS.

As sibilâncias são os CAS mais comuns, apresentando uma frequência bastante diversificada, entre 100Hz e 1000Hz. Podem ser detetadas mais facilmente, uma vez que são geralmente mais intensos que os roncos [8], sendo detetadas por vezes a partir da respiração realizada de boca aberta ou da auscultação na traqueia.

2.1.2. Sons Adventícios Descontínuos

Os sons adventícios descontínuos, como os fervores, são sons breves e explosivos que podem ser resultado de diferentes tipos de patologias, dependendo da sua deteção a nível torácico quando efetuada uma auscultação [9]. A sua reprodução ocorre quando existe uma abertura/fecho abrupta das vias aéreas ou pela passagem de ar pelas secreções das vias respiratórias.

Os fervores são sons não musicais (produzem uma vibração irregular) e podem ser classificados dependendo da sua duração e frequência. Estes normalmente atingem frequências entre 60 Hz e 2000 Hz [10] e têm uma duração inferior a 25ms. Na Figura 3 são ilustradas amostras de ondas sonoras e espectrogramas de sons respiratórios, nomeadamente som saudável, sibilância e fervor. Os espectrogramas são gráficos que permitem analisar a densidade espectral de energia de um sinal sonoro em função do tempo. Como se pode verificar na Figura 3, os espectrogramas são constituídos com cores bastante diferentes que indicam a intensidade espectral de energia.

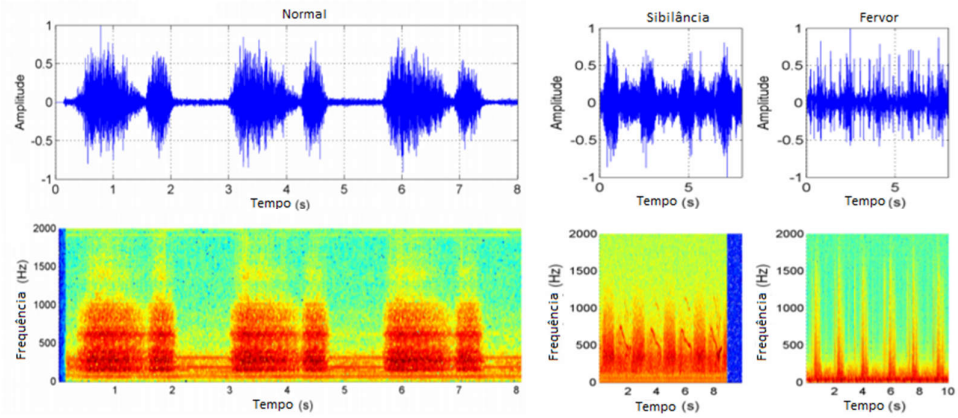


Figura 3 – Ondas sonoras (em cima) de uma respiração normal, sons adventícios e os seus respetivos espectrogramas (em baixo) [11]

Dependendo da sua duração, intensidade, permanência no ciclo respiratório e se estão relacionados a sintomas de tosse, os fervores podem ser designados como finos ou grossos [12], sendo estes distinguidos maioritariamente pela sua frequência e duração: baixa frequência e maior duração (grossos) ou alta frequência e menor duração (finos). Os fervores finos ocorrem frequentemente em pacientes com doenças pulmonares intersticiais, sendo produzidos sons agudos durante a inspiração. Em comparação, os fervores grossos são mais audíveis, com ondas sonoras com menor frequência e uma maior duração. A sua existência também está relacionada com determinadas patologias, como por exemplo a bronquite crónica ou o edema pulmonar grave, que é caraterizada por sons mais graves que os fervores finos, tanto na inspiração como expiração [13].

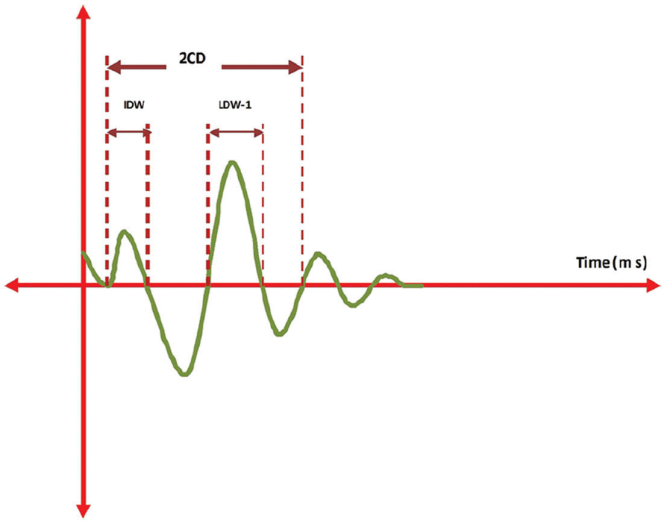


Figura 4 - Onda sonora de um fervor [32]

Através da análise do formato das ondas produzidas pelos fervores é possível distinguir os fervores finos dos grossos de acordo com três parâmetros: Largura da Deflexão Inicial (IDW), que representa a duração entre o início do som e a primeira deflexão, o Ciclo de Duas Durações (2CD), que simboliza a duração entre o início do fervor e os dois primeiros ciclos do fervor, e a Maior Largura de Deflexão (LDW) que, como representada na Figura 4, possui fervores grossos com uma duração média de IDW e 2CD de 1,5ms e 10ms respetivamente, enquanto os fervores finos possuem 0,7ms e 5ms [14].

2.2 Atributos

A partir dos dados crus do conjunto de dados inicial, e através de técnicas e algoritmos de análise de dados, são extraídas características que caraterizam os objetos que se pretendem analisar. Os objetos sob análise podem possuir uma quantidade gigantesca de atributos, podendo estas ser redundantes para caracterizar o objetivo a analisar. Um bom conjunto de dados com finalidade de ser aplicado a técnicas de aprendizagem computacional deve possuir atributos que apresentem uma fraca correlação entre elas, mas uma forte correlação, isoladamente ou em combinação com outras, com as classes a classificar.

Alguns exemplos de atributos utilizadas na classificação de sons respiratórios adventícios são as seguintes [10]:

Atributos Espectrais

Os atributos espectrais são extraídos a partir do espectro sonoro que é formado pelas ondas que compõem o sinal. Estes atributos são conhecidos maioritariamente por serem baseadas no conteúdo de frequência do sinal: centroide espectral (representa o centro de gravidade do espectrograma), propagação (descreve o desvio padrão do espectrograma relativamente ao centroide), brilho (reflete a quantidade de informação de alta frequência e a relação da energia de uma frequência em relação à total), entre outras. A partir destas podem ser analisadas: a energia, variância da frequência na onda sonora, qual o centro geométrico e a assimetria da distribuição, entre outros. Podem ser obtidas através da conversão de sinais temporais para o domínio da frequência utilizando a transformada de Fourier. Estes atributos são utilizados para identificar timbres sonoros diferentes, alturas, ritmos e melodias [15].

Atributos MFCC

Os atributos obtidos através dos *Mel-Frequency Cepstral Coefficients* (MFCC) são características de um sinal e descrevem a forma geral do envelope espectral. Estão fortemente relacionadas com o timbre do som. O seu método de extração pode ser observado a partir da Figura 6, podendo-se verificar que inicialmente é aplicada uma Transformada de Fourier (F) ao sinal sonoro, obtendo-se um espectro composto pela relação entre a potência e as diversas frequências do sinal. A este, é aplicado um logaritmo (*log*) para obter o espectro em relação à amplitude. As frequências do sinal são traduzidas para uma escala logarítmica, a escala de Mel (Figura 5). Por último, é aplicada uma transformação com uma forte propriedade de compactação de energia, a transformada discreta de cosseno, que é utilizada para processamento de sinais e compressão de dados. Este processo pode ser visto como uma função inversa da Transformada de Fourier (F^{-1}), a partir da qual se retiram os coeficientes do sinal (C). A maioria das informações dos sinais tendem a concentrar-se em componentes de baixa frequência, pelo que, tipicamente, são apenas extraídos os treze primeiros componentes [16]. Este processo pode ser traduzido pela equação: $C\big(x(t)\big) = F^{-1}[\log(F[x(t)])]$.

Figura 6 - Processo MFCC

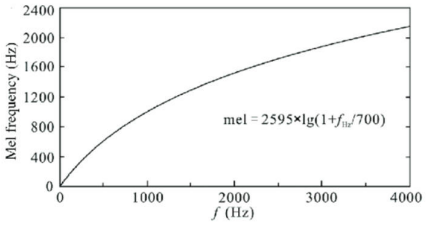


Figura 5 - Escala de Mel

Atributos Melódicos

A melodia é a interseção de altura e ritmo, ou seja, é criada através da subida e descida de notas. Esta é criada quando uma sequência de tons é produzida durante um período de tempo. Os atributos melódicos descrevem a melodia ou o timbre de um determinado som. Estas são computadas a partir da curva da altura que carateriza as frequências do sinal sonoro. Algumas destas características correspondem ao comprimento, altura e suavidade da curva da altura [10].

2.3 Aprendizagem computacional

A aprendizagem computacional é um ramo da área de Inteligência Artificial baseado no conceito de implementação de mecanismos de aprendizagem em diversos sistemas. Este domínio utiliza diferentes tipos de dados como entrada para algoritmos matemáticos de aprendizagem com vista a que este aprenda e possa tornar-se autónomo e automático aquando confrontado com situações semelhantes.

Os sistemas são treinados utilizando dados históricos, para obter uma determinada resposta. Existem três formas de aprendizagem principais: supervisionada, não supervisionada ou por reforço. Na aprendizagem supervisionada o algoritmo irá ser treinado utilizando os dados de entrada e a saída desejada como referência, tendo como objetivo a criação de um modelo que consegue prever a saída desejada apenas a partir dos dados de entrada. A aprendizagem não supervisionada não conhece previamente a classe da respetiva entrada quando se pretende treinar o algoritmo. Este centra-se na deteção automática de padrões nos dados para previsões futuras. A aprendizagem por reforço também tem como objetivo a criação de um modelo que seja capaz de tomar decisões, contudo neste caso o algoritmo recebe recompensas ou penalizações a cada iteração que possa ocorrer ao longo do treino até que o modelo aprenda o que se pretenda.

A aprendizagem computacional recorre a diversas técnicas para dar resposta a um problema que se pretende resolver. Estes problemas podem ter como natureza: a regressão, classificação, aglomerativa, redução de dimensionalidade, entre outros.

O foco deste trabalho é a aprendizagem supervisionada. Neste sentido, podem-se identificar três abordagens principais: i) abordagens clássicas, baseadas na extração de atributos do domínio e na utilização de algoritmos clássicos de aprendizagem computacional, e.g., árvores de decisão, SVM, etc.; ii) abordagens baseadas em aprendizagem profunda com aprendizagem automática de atributos pela rede

(aprendizagem de atributos), as quais se baeiam em arquiteturas de redes neuronais profundas; iii) abordagens híbridas, baseadas na extração de atributos, as quais são usadas como entradas para redes neuronais profundas. Neste trabalho, o foco incidirá na segunda abordagem, i.e., baseada em aprendizagem profunda com aprendizagem de atributos.

Antes disso, importa introduzir alguns conceitos-chave de redes neuronais artificiais (ANN). A origem dos algoritmos de redes neuronais tem como principal inspiração o funcionamento do cérebro humano. Os neurónios são as unidades básicas da estrutura do cérebro humano que permitem a transmissão de informação de umas células para as outras. Estes comunicam entre si, recebendo informações e transmitindo-as ao próximo neurónio. Os modelos criados têm como premissa redes neuronais artificiais, constituídas por neurónios que se conectam entre si através de pesos.

A unidade mais básica das ANN é o perceptrão, que é constituída apenas por um neurónio. Como se pode observar na Figura 7, esta recebe vários valores de entrada que são multiplicados pelos pesos correspondentes, sendo de seguida o valor da soma destas multiplicações enviado para a função de ativação que irá decidir o valor de saída desta rede.

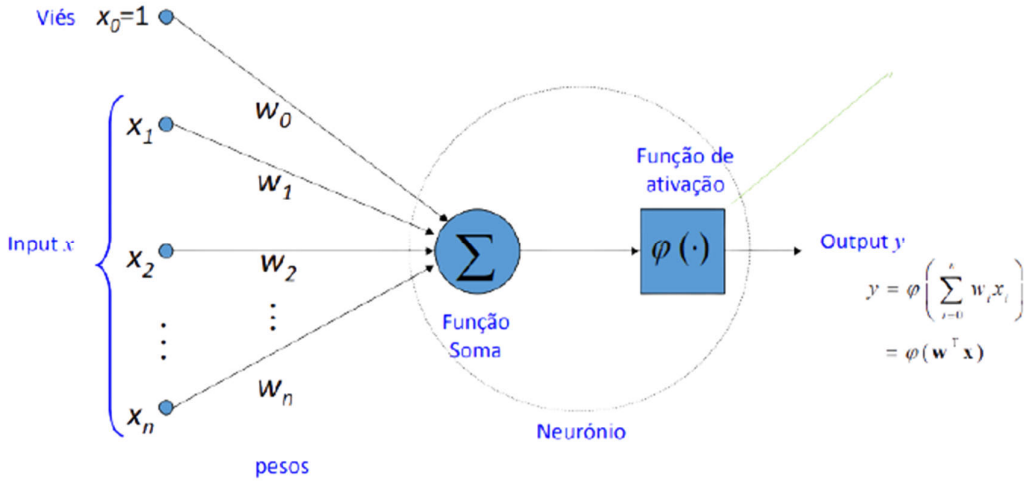


Figura 7 - Arquitetura de um Perceptrão [30]

Normalmente, um neurónio, mesmo tendo muitos elementos de entrada, não consegue resolver os problemas identificados acima. Deste modo, as ANN são compostas por uma camada de entrada, onde são introduzidos os dados, uma camada de saída com o resultado da toma de decisão do sistema e por uma ou mais camadas escondidas (redes multicamadas) com n neurónios. Assim como no perceptrão, cada neurónio possui um viés e uma função de ativação. O viés pertencente a cada um dos neurónios e os pesos das entradas vão sendo ajustados de acordo com o resultado pretendido, com o objetivo de aperfeiçoar o modelo computacional.

Existem várias variantes de arquiteturas de redes neuronais tais como:

Rede neuronal *Feedforward* - Esta rede neuronal foi a primeira e mais simples rede neuronal desenvolvida, contendo apenas uma camada de entrada, uma camada escondida de neurónios e a camada de saída.

Redes neuronais multicamada – Difere das redes descritas acima por possuir duas ou mais camadas escondidas.

Em termos de algoritmos de treino, o algoritmo de retropropagação do erro é, provavelmente, o mais utilizado. É um algoritmo multicamada em que o resultado obtido no final de cada iteração é utilizado para calcular o erro de classificação e através deste os pesos das camadas anteriores à camada de saída serão atualizados, havendo uma passagem de informação do final para o início, ao contrário da *feedforward* em que a informação apenas circula da camada de entrada para a de saída.

2.3.1. Aprendizagem Profunda

A aprendizagem profunda é uma subárea das redes neuronais que diverge das abordagens clássicas. Os problemas que os algoritmos de aprendizagem computacional tentam resolver têm vindo a tornar-se cada vez mais desafiantes. Deste modo, as ANN tendem a possuir um número maior de camadas ocultas, ou seja, as redes tendem a ser mais profundas. Os algoritmos *de* aprendizagem profunda são autodidatas, tendo a capacidade de aprender à medida que vão filtrando informações ao longo das camadas ocultas.

Técnicas de aprendizagem profunda são bastante eficientes ao necessitar de processar uma grande quantidade de dados devido às suas numerosas camadas escondidas. Algumas arquiteturas profundas referidas ao longo da dissertação são os seguintes:

- ***Convolutional Neural Network* (CNN)**

  As CNN são redes neuronais frequentemente aplicadas a reconhecimento/classificação de imagens e processamento na área de *Natural Language Processing* (NLP). Estas redes possuem diversos tipos de entrada, por exemplo uma imagem a ser classificada em forma de matriz.

  Esta rede neuronal artificial, para além das camadas de entrada e saída, contém camadas escondidas designadas por camadas convolucionais. Nestas camadas cada neurónio é um filtro (matriz de pesos) que é aplicado à matriz de entrada da rede, computando uma nova matriz, dando como entrada na camada seguinte. Esta nova matriz pode ter uma dimensionalidade menor que a inicial ao usar um passo superior a um, ou seja, em vez de percorrer todos os valores da matriz com o filtro este salta o número de valores igual ao passo definido. As camadas também possuem funções de ativação como *softmax* e *Rectified Linear Unit (ReLU)*. As CNN também podem ter uma camada de *pooling* que serve para simplificar a camada anterior, normalmente utilizando *maxpooling* no qual o maior elemento selecionado pela área é passado para a matriz seguinte, mantendo a informação mais importante. A última camada das CNN é a *fully connected*, em que sua saída é constituída por *N* neurónios,

sendo *N* a quantidade de classes a classificar. A Figura 8 permite uma visualização mais clara de uma arquitetura das CNN e como as camadas se relacionam.

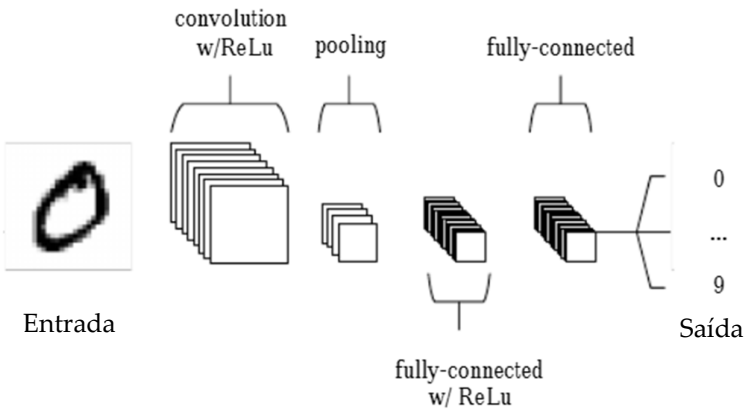


Figura 8 - Arquitetura CNN simples [28]

- **Autoencoders**

Os *autoencoders* são redes neuronais de aprendizagem não supervisionada, com um objetivo bastante definido: comprimir dados e aprender a reconstruí-los tendo como resultado uma representação semelhante à original. A sua arquitetura varia consoante o objetivo que o programador desejar alcançar. Os *autoencoders* possuem três componentes: *encoder*, *code* e *decoder* (Figura 9). O *encoder* é a secção da rede que comprime os dados da camada de entrada, dando deste modo origem ao *code* que representa os dados comprimidos. O *decoder* faz o oposto do *encoder* e, reconstrói os dados comprimidos inicialmente. Os *autoencoders* profundos, são compostos por duas redes neuronais *belief*.

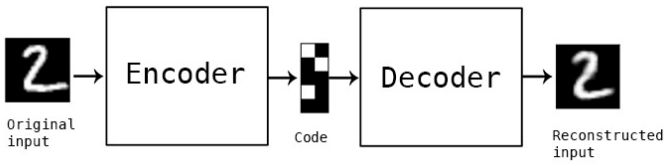


Figura 9 - Arquitetura de um autoencoder

- **Recurrent Neural Network (RNN)**

As redes RNN são redes neuronais com memória. Estas conseguem que a informação em cada neurónio ou camada persista ao longo do tempo, sendo transmitida essa informação de uma etapa para a seguinte. Uma RNN, ao contrário de outros classificadores, partilha a informação do viés com todas as camadas da rede reduzindo, deste modo, o número de parâmetros que precisa de atualizar e aprender ao longo do treino [17]. Pode-se assim, pensar numa RNN como múltiplas copias da

mesma, uma vez que todas possuem a mesma informação. Utiliza os dados de saída da rede anteriormente obtidos como dados de entrada na próxima iteração do processo de classificação. A Figura 10 apresenta a arquitetura de uma RNN. A amarelo é apresentada a camada de entrada, a azul as camadas com neurónios recorrentes e a laranja a camada de saída. Os neurónios a azul são neurónios recorrentes e apresentam recursividade.

Figura 10 - Arquitetura RNN [38]

- **Long Short-Term Memory (LSTM)**
  É uma variante da arquitetura RNN que apresenta normalmente melhores resultados uma vez que estes estão elaborados de forma a possuir "memória" retendo a informação por longos períodos. Estas redes neuronais são bastante utilizadas para prever séries temporais. As LSTM possuem uma estrutura composta por três redes neuronais e diferentes blocos de memória: as células. A finalidade destas células é reter informação e a dos *gates* é manipular a memória. O Forget *gate*, tendo duas entradas (entrada atual e saída anterior) filtra a informação que é útil. O *input gate* adiciona as informações úteis ao estado da célula. A Figura 11, apresenta uma arquitetura LSTM. A diferença entre esta arquitetura e a das RNN é que as camadas escondidas possuem neurónios com memória.

Figura 11 - Arquitetura LSTM [29]

2.4 Métricas de desempenho

A matriz de confusão é uma tabela que permite, de uma forma simplificada, visualizar o desempenho do modelo ao nível das classes a serem classificadas. Esta indica quantas

amostras de cada classe foram classificadas corretamente ou classificadas como outra. A Tabela I representa o modelo de uma matriz de confusão para um problema de duas classes (Positiva e Negativa). Para problemas de múltiplas classes têm que se considerar a classe positiva como aquela que se pretende analisar e a negativa como todas as outras existentes no problema.

		Classe Prevista	
		Positivo	Negativo
Classe Real	Positivo	TP	FP
	Negativo	FN	TN

Tabela I – Matriz de confusão modelo

As amostras, depois de passarem pela fase de teste do modelo, podem ser consideradas como:

- Verdadeiro positivo (TP), quando a amostra foi prevista pertencente à classe real da amostra, ou seja, classificada corretamente;
- Falso positivo (FP), quando a amostra que estamos a prever é classificada incorretamente como outra classe;
- Verdadeiro Negativo (TN), quando a amostra pertence a uma classe diferente daquela que estamos a analisar e a previsão classifica-a corretamente;
- Falso negativo (FN), quando a amostra pertence a uma classe diferente daquela que estamos a analisar e foi classificada incorretamente.

A partir da matriz de confusão é possível extrair métricas de desempenho para auxiliar a avaliação dos modelos. Estas têm um papel fundamental para se medir a qualidade dos modelos criados. São essenciais para quantificar o quão bom e confiável é o modelo, principalmente na fase de teste. É necessário ter atenção às métricas escolhidas para avaliar os modelos, uma vez que algumas são mais adequadas que outras dependendo dos dados que se pretende classificar. Para medir o desempenho dos modelos que iremos desenvolver nesta tese utilizámos as seguintes métricas:

- **Exatidão**
  Esta métrica relaciona todas as amostras corretamente classificadas, tanto negativas como positivas, considerando o número total de amostras, independentemente da classe. Através do complemento desta métrica pode-se obter a percentagem de erro dos modelos. [18]

$$\text{Exatidão} = \frac{TP + TN}{TP + TN + FP + FN}$$

É preciso analisar esta métrica com bastante cuidado, uma vez que o seu resultado pode induzir em erro. Se um modelo for testado com um conjunto de dados com classes muito desequilibradas como 990 sons normais e 10 sons adventícios e o modelo classificar

todas as amostras como sons normais, o resultado da exatidão irá ser de 99%, mas não simboliza que este possua um bom desempenho. Desta forma é necessário também analisar mais métricas como as que irão ser descritas já de seguida.

- **Precisão**
  Fornece uma percentagem de quantas amostras classificadas como positivas são realmente positivas. Como se pode ver na fórmula abaixo, esta tem em conta o número de amostras que foram classificadas corretamente de forma positiva (TP) e divide-as pelo total de amostras classificadas como positivas (TP e FP).

$$Precisão = \frac{TP}{TP + FP}$$

- **Sensibilidade**
  Avalia o quão bom é o modelo a classificar uma determinada classe, dando uma maior ênfase às amostras classificadas como falsas negativas.

$$Sensibilidade = \frac{TP}{TP + FN}$$

- **Especificidade**
  Estima a probabilidade de as amostras da classe negativa serem corretamente classificadas, ou seja, a percentagem de amostras negativas que foram identificadas corretamente.

$$Especificidade = \frac{TN}{TN + FP}$$

- **F1-Score**

  Esta métrica é considerada a média harmónica entre a precisão e a sensibilidade. Quando o F1-Score tem um resultado elevado (tanto a precisão como a sensibilidade têm que possuir uma percentagem elevada) pode-se afirmar que o modelo é capaz de classificar corretamente grande parte das suas amostras tendo em conta o acerto por classe. Sendo assim, esta métrica é conhecida por resumir de forma bastante completa o desempenho do modelo a analisar.

$$F1\ Score = \frac{2 * (Precisão * Sensibilidade)}{Precisão + Sensibilidade}$$

3 Estado de Arte

Existem inúmeros estudos sobre classificação sonora, tanto na área médica como noutras áreas, com resultados bastante positivos. O estado da arte presente neste capítulo foca a apresentação de artigos, tendo em vista uma análise crítica das metodologias aplicadas. Os artigos apresentados foram escolhidos tendo em conta os seus bons resultados de classificação e/ou segmentação e a sua aplicabilidade tendo em conta os objetivos da tese.

Também tínhamos como objetivo apresentar uma variedade de classificadores, juntamente com as diferentes formas de processar os sinais, aumentando assim o conhecimento das capacidades de algoritmos de diferentes abordagens. Estas dividem-se em abordagens clássicas e de aprendizagem profunda criando uma analogia entre os vários processos de classificação. Salientamos, ainda, o facto de que todos os artigos presentes nesta revisão foram apresentados em conferências e/ou publicados em revistas de cariz científico. Este capítulo encontra-se dividido em estudos sobre segmentação e/ou classificação de sons adventícios (SA) cujos dados provêm da mesma base de dados (*Respiratory Sound Database*) [1] que serve de apoio para esta tese, e estudos relativos à utilização de técnicas de classificação e segmentação de som em geral.

Base de dados *Respiratory Sound Database*

A base de dados *Respiratory Sound Dataset* utilizada é constituída por ficheiros de áudio com amostras de 126 pacientes. As equipas da Escola de Ciências de Saúde da Universidade de Aveiro (ESSUA), Universidade de Aristóteles de Salónica, Grécia (AUTH) e Universidade de Coimbra (laboratório de informática clínica pertencente ao grupo de computação adaptativa) foram as responsáveis por adquirir as amostras presentes nesta fonte de dados, que originalmente serviu de apoio ao desafio ICBHI 2017 [1].

Os ficheiros de som que constituem a base de dados foram recolhidos em dois locais diferentes e extraídos usando equipamentos heterogéneos (AKGC417L[2], Meditron[3], Litt3200[4] e LittC2SE[5]) aplicados a diferentes zonas do tórax e traqueia do paciente. A base de dados contém gravações com durações entre 10s e 90s, sendo 1864 fervores, 886 sibilâncias e 506 áudio com ambos os sons adventícios. Estes ficheiros encontram-se devidamente anotados por especialistas, contendo 920 ficheiros de texto com a informação do início e fim dos eventos existentes no ciclo respiratório/áudio, podendo estes ser sons adventícios (fervor ou sibilância) ou som normal. Os especialistas realizaram as anotações usando o programa Audacity através de uma inspeção visual e acústica. Os ciclos respiratórios possuem um comprimento muito variável, tendo de 0.2 a 16.2 segundos de duração. Tabela II apresenta resumidamente as principais características da RSD.

[2] Microfone AKG C417L

[3] Estetoscópio eletrónico WelchAllyn Meditron Master Elite

[4] Estetoscópio eletrónico 3M Litmmann 3200

[5] Estetoscópio 3M Littmann Classic II SE

Nº de ficheiros de áudio	920
Nº de Participantes	126
Informação dos Participantes	Faixa etária: 49 Crianças e 77 Adultos Sexo: 46 Feminino e 79 Masculino Idade (Média ±desvio padrão): 43 ± 32,2 anos
Zonas de recolha de dados	Tc: Traqueia Al: Anterior Esquerdo (1) Ar: Anterior Direito (2) Pl: Posterior esquerdo (4) Pr: Posterior direito (3) Ll: Lateral esquerdo (6) Lr: Laterial direito (5) (Ver locais na Figura 12)

Tabela II – Demografia da base de dados *Respiratory Sound Database*

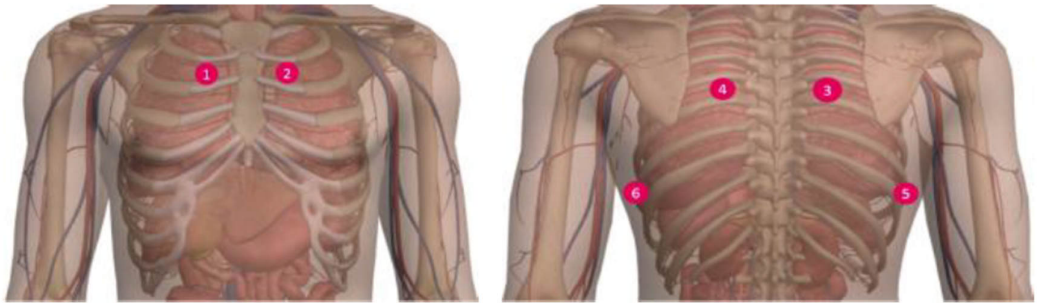


Figura 12 - Potenciais localizações de aquisição de sons (vermelho) [42]

A utilização de pacientes com caraterísticas fisionómicas (idade, sexo, massa corporal e altura) bastante distintas, tanto saudáveis como patologicamente identificados (infeções do trato respiratório superior, asma, doença pulmonar obstrutiva crónica, infeções do trato respiratório inferior, pneumonia e bronquiectasia), e utilizando diversos mecanismos de recolha em zonas corporais diferentes permitiu adquirir uma boa base para criar um conjunto consistente de dados com o qual se pudesse trabalhar futuramente, assemelhando-se a um ambiente real devido à diversidade das variáveis enumeradas.

3.1 Segmentação e classificação de SA provenientes da base de dados *Respiratory Sound Database*

Na abordagem proposta em [19] foi desenvolvida uma combinação do classificador *random forest classifier* e o método empírico de decomposição (EMD), tendo como objetivo a classificação de diversos elementos dos sons para distinguir diferentes patologias respiratórias (multi-classificação). Os ciclos respiratórios foram segmentados em 4 fases: fase inspiratória (FI), fase expiratória (FE), fase inspiratória inicial (FII) e fase inspiratória tardia (FIT), sendo os dois primeiros a primeira e segunda metade de um ciclo respiratório e os dois

últimos a primeira e segunda metade do FI, respetivamente. Com o objetivo de isolar os fervores, aplicaram um filtro aos SR ficando apenas com o sinal de frequência entre 120Hz e 1800Hz, removendo assim ruído ambiental, corporal ou de batimentos cardíacos. A técnica janela de Hamming foi aplicada para prevenir *frequency leakage*. Através da técnica EMD, o sinal foi decomposto em funções de modo intrínseco, sendo apenas as três primeiras utilizadas para a extração de atributos. Foram extraídos atributos espectrais, assim como variações estatísticas incluindo o desvio padrão, média, mediana, valor máximo e mínimo e por último extraída a média dos primeiro 13 MFCC das quatro fases respiratórias. Foram aplicados 15 classificadores com conjunto de dados formados pelas diferentes fases do som. Pode-se concluir qua a melhor classificação em relação à identificação das patologias é alcançada quando se utilizam as fases FII e FI.

O artigo [10] teve como objetivo estudar a influência da duração dos sons adventícios respiratórios na classificação automática. Deste modo, utilizaram-se os sons adventícios com durações variáveis e fixas. Os sons adventícios com durações variáveis foram obtidos através da RSD, construída com a participação dos autores deste artigo. Os segundos sons foram gerados através de um script criando eventos de 50ms e 150ms, num total de 963 sibilâncias e 4158 fervores. Todos os sinais sonoros sofreram uma reamostragem de 4000Hz. A experiência realizada passou pela criação de diversos conjuntos de dados, a partir da extração de atributos das *frames* dos espectrogramas dos sons adventícios. Num total, foram extraídos 81 atributos sendo: 25 espectrais, 26 MFCC e 30 melódicos, sendo a posteriori calculadas cinco estatísticas destes atributos: média, desvio padrão, mediana e o valor mínimo e máximo. Para cada um dos conjuntos de dados criados foram selecionados os melhores atributos utilizando o algoritmo *Minimun Redundancy Feature Selection* (MRMR). Cada um destes conjuntos de dados serviram para alimentar o treino de três classificadores diferentes: *Linear Discriminant Analysis* (LDA), SVM com uma função de base radial (SVMrbf), árvores aumentadas com subamostragem aleatória *(*RUSBoost*)*. Também foram construídas CNNs e treinadas com imagens de espectrogramas e espectrogramas melódicas. Todos os algoritmos foram treinados dez vezes com *seeds* diferentes. Os conjuntos de dados constituídos com três classes alcançaram valores de precisão de 96,9% usando sons com durações fixas e 81.8% com durações variáveis.

Em [20] é proposto um modelo LSTM para classificar ficheiros de áudio relativamente à saúde respiratória dos pacientes. Estes podem ser classificados de oitos formas diferentes: saudáveis ou com uma das sete doenças referidas na base de dados descrita acima. De cada um dos ficheiros de áudio foi extraída uma amostra de um segundo. A partir destas, foram extraídos diversos atributos (MFCCs, *Zero-Crossing Rate* (ZCR), coeficientes harmónicos, codificadores preditivos lineares (LPC) e predição linear percetual (PLP)) que tinham como objetivo serem utilizadas como entrada do modelo desenvolvido, sendo, por fim, aplicado um método de normalização de dados. As amostras foram divididas em 70% para treino, 10% para validação e 20% para teste. A rede neuronal é constituída por uma camada LSTM e uma camada *fully connected* composta por 8 neurónios. O modelo alcançou uma exatidão de 46% na fase de teste.

O artigo [21] foca-se na implementação de uma rede neuronal profunda CNN simples para classificar sons adventícios. O destaque deste estudo é o desenvolvimento de um

conjunto de novas técnicas: *device specific-tuning* (divisão do conjunto de teste em subconjuntos tendo em conta o dispositivo de recolha do som), *concatenation-based augmentation (*que consiste em criar uma nova amostra de uma classe a partir da concatenação de duas amostras aleatórias da mesma classe), *blank region clipping* (técnica para eliminar dos espectrogramas de mel regiões vazias que se localizam entre os 1500-2000Hz) *e smart padding* (técnica de transformar os ficheiros de áudio com um tamanho fixo a partir da cópia de ciclos já existentes do mesmo paciente ou amostras de ciclos vizinhos). Estas tinham o propósito de aumentar o número de amostras e o potencial das mesmas. Os dados sofreram um pré-processamento que consistiu na normalização padrão do sinal entre [-1, 1], reamostragem para 4000Hz e na aplicação de um filtro passa-banda Butterworth de 5ª ordem, eliminando possíveis ruídos como batimento cardíaco ou pessoas a falar. Por fim, após todas as transformações, os sinais foram convertidos em espectrogramas de Mel de forma a alimentar o modelo de aprendizagem profunda CNN. Este modelo foi baseado na ResNet-34 (rede neural residual), através da qual se inicializaram os pesos da rede. É constituída por duas camadas lineares com funções de ativação ReLU e a última possui uma função de ativação *softmax*, sendo aplicado um *dropout* a todas as camadas *fully connected*. A rede é treinada com uma *loss cross-entropy* tendo o interesse de minimizar a perda da classificação. Foram criadas duas experiências para testar o classificador em que a primeira se baseava na classificação dos ciclos respiratórios como: normal, com fervores, com sibilâncias, ou com fervores e sibilâncias. Já a segunda pretendia apenas distinguir ciclos respiratórios entre normais (respiração sem sons adventícios) e anormais (respiração com sons adventícios). Para a experiência com quatro classes, o conjunto de dados foi dividido de duas formas diferentes: 60% de treino, 40% de teste e 80% de treino e 20% de teste, respetivamente. Para a classificação de quatro classes, a melhor pontuação (média da sensibilidade e especificidade) foi de 56.2%, sensibilidade de 40.1% e especificidade de 72.3% quando aplicado utilizado o classificador CNN em conjunto com as técnicas *concatenation-based augmentation*, *blank region clipping* e *device specific fine-tuning*. Para a experiência com divisão de 80/20 do conjunto de dados o melhor método foi o mesmo, mas com uma pontuação superior de 68.5%, sensibilidade de 53.7% e especificidade 83.3%. O melhor resultado obtido foi na classificação de duas classes com uma divisão 80/20 e o mesmo classificador anteriormente descrito tendo uma pontuação de 77%, especificidade de 80.9% e 73.1% de sensibilidade.

A abordagem proposta em [22] baseia-se no desenvolvimento de métodos de Redes neuronais profundas (*Deep Neural Networks* (DNN)) e autoencoders para classificar ciclos respiratórios em quatro categorias diferentes: fervores, sibilâncias, ambos ou normal. À semelhança dos artigos anteriores, foi realizado um pré-processamento no qual se aplicou uma reamostragem de todos os ciclos respiratórios de 4000Hz e aplicou-se a duplicação de amostras para que todos os áudios tivessem 10s de duração. De seguida, para a eliminação de ruído, foi aplicado um filtro passa-banda de 100-2000Hz. Por último, os sinais são convertidos em espectrogramas através do uso de um filtro gama com 1024 FFT. Visto que existe uma falta de balanceamento ao nível das quatro classes, implementaram-se duas técnicas de aumento de dados: sob amostragem das imagens dos espectogramas e de seguida aplicado o critério de Fisher. As imagens geradas são inseridas como entrada no classificador DNN. A rede neuronal profunda criada foi uma CNN com cinco camadas, sendo a primeira a de entrada, e as outras referentes a *batch normalization, convolutional* com função de ativação *ReLU, max pooling, global max pooling e dropout* que sofre um incremento de 5% ao longo das camadas inicializado a 10%. A última camada é composta por duas camadas *fully-connected*,

uma com função de ativação ReLU e outra com *softmax*. A rede *autoencoder* é constituída pela fase *encoder* e a fase *decoder*. A fase encoder é formada pelas primeiras cinco camadas da rede CNN, já a fase de *decoder* (descodificador) possui quatro blocos para inverter a convolução e usa ReLU para reconstruir a imagem. O conjunto de dados foi dividido em treino e teste, numa proporção de 60% e 40%, respetivamente. A rede foi treinada com uma taxa de aprendizagem de 0,0001, um *batch size* de 50, 100 *epochs* e para otimização foi utilizado o método Adam. O desempenho obtido pelo Autoencoder foi superior ao da rede CNN apenas por 2%, alcançando uma pontuação média de 43%, especificidade de 62% e sensibilidade de 33%. A fusão dos métodos de classificação CNN e Autoencoder alcançaram um F-score de 49%, uma pontuação média de 42%, especificidade de 69% e sensibilidade de 30%.

No estudo [23] é proposto um novo método capaz de distinguir fervores, sibilâncias e sons normais: as redes residuais profundas. Antes de proceder à classificação são extraídos diversos segmentos de som que contenham uma das classes que se desejam classificar. De seguida é aplicada uma otimização da Transformada de S (S-*transform*) da qual são extraídos os coeficientes de *optimized S-transform* (OST), à posteriori criadas imagens RGB, espectrogramas de OST, sofrendo um reescalonamento com um tamanho fixo de 224x224xRGB, através do método bilinear. Depois dos segmentos sonoros passarem por todas as fases de pré-processamento os mesmos são divididos em 70% para treino e 30% para teste. As imagens do subconjunto de dados de treino são inseridas na rede neural residual profunda para esta ser treinada. A rede neural residual consiste em três fases: pré-processamento, processamento e classificação. Na primeira, as imagens RGB correspondentes a um mapa de atributos são pré-processadas e normalizadas através das camadas de convolução, *batch-normalization*, ReLU e *MaxPooling*. Na segunda, o conjunto de dados é processado por blocos de identidade e blocos convolucionais. A última fase é constituída por três camadas, a *average-pooled, flatten e full connection*, com função de ativação *softmax*. A rede obteve o melhor valor de exatidão num valor de 98.79%, sensibilidade de 96.27% e especificidade de 100%, quando treinada com um tamanho 16 de *batch* e com 20000 iterações.

Uma nova forma de prever anomalias respiratórias foi proposta em [24]. O trabalho consiste em gerar modelos de RNN capazes de classificar patologias e sons adventícios, ou seja, duas resoluções diferentes, uma com duas classes (sibilâncias e fervores) e outra com classes múltiplas (sibilâncias e fervores, dois ou normal). Foram desenvolvidas duas arquiteturas avançadas de RNN's, uma LSTM e uma GRU, e duas derivações BiLSTM e BiGRU, tendo em vista a resolução do problema. Para todas as arquiteturas foi implementada a mesma configuração: 2 camadas com 256 células cada com a função de ativação de *tanh, dropout* que variaria entre 30% a 60%, *batch normalization, batch size* de 32, sendo o algoritmo de otimização utilizado o Adam. Foi desenhado um pré-processamento com várias fases: cada ciclo respiratório foi segmentado com base numa janela deslizante (25ms a 500ms) usando uma sobreposição de 50%, e para cada janela foram extraídos os MFCCs, sendo concatenados por último (*frame*). De seguida os dados são normalizados utilizando a técnica Min-Max e Z-score. O conjunto de dados foi dividido em subconjunto de dados de treino e testes com um rácio de 80% e 20%, respetivamente, sendo à posteriori o modelo treinado com 100 *epochs*. Analisando o impacto da normalização nos resultados de teste deparamos nos com uma *acurracy* mais elevada quando utilizada normalização com Z-score para os dois problemas, alcançando 81% para classificação de duas classes e 74% para

classes múltiplas. Em média, as redes LSTM e BiLSTM foram as que tiveram melhor desempenho para a classificação de 4 classes. Em comparação a modelos de artigos publicados anteriormente os resultados alcançados mostram-se competitivos.

O artigo [25] apresenta uma proposta de um modelo CNN capaz de distinguir ciclos respiratórios com sibilâncias ou fervores, os dois ou normais, como na proposta descrita anteriormente, mas realizam um tratamento dos ciclos e geração de modelos de forma bastante diferentes. Antes de extrair dados interessantes dos ciclos respiratórios para alimentar o modelo, foi aplicado uma reamostragem para 4000Hz e um filtro de passa-banda Butterworth de 12ª ordem com um corte entre 120-1800Hz de modo a eliminar ruído dos sinais. Estes foram convertidos em espectrogramas e onduletas, que serviram como dados de entrada para a CNN. Numa primeira abordagem foram utilizados apenas espectrogramas, na segunda experiência apenas onduletas e por fim a utilização de ambas. O conjunto dos dados foi dividido em treino (60%) e teste (40%) e de seguida treinado o modelo composto por uma camada *Bach normalization*, função de ativação ReLU, e camadas f*ully connected*, sendo primeiramente utilizada uma aprendizagem de transferência baseada na rede VGG-16. Depois de treinado e testado o modelo através da primeira abordagem (espectrogramas) alcançou-se um *F1-score* de 46%, sensibilidade de 36% e especificação de 69%. Já o modelo relativo à segunda (onduletas) atingiu um *F1-score* de 46%, sensibilidade de 37% e especificidade de 62%. A terceira experiência obteve um *F1-score* ligeiramente abaixo de 42%, sensibilidade de 28% e especificidade de 81%.

3.2 Segmentação e classificação de som

O artigo em [26] apresenta o desenvolvimento de um modelo de aprendizagem profunda, com aprendizagem semi-supervisionada, para classificar sons pulmonares automaticamente. Os dados utilizados foram gravados em quatro centros clínicos da Índia, a partir de 284 pacientes com doenças respiratórias, utilizando como ferramentas de recolha um telemóvel (aplicação *android*) e um estetoscópio eletrónico, resultando em 11627 ficheiros de áudio extraídos de 11 diferentes zonas de auscultação. O conjunto de dados escolhido focou-se na presença de sibilâncias ou fervores, uma vez que estes são os sons adventícios mais comuns em SR. Os dados sofreram um pré-processamento, sendo filtrados os sinais com uma janela de Hamming na gama de 200Hz a 1600Hz e aplicada reamostragem para 4000Hz. Cada sinal foi convertido em espectrogramas utilizando a *Short-Time Fourier Transform* (STFT). De forma a evitar que ocorressem diferenças entre os sons recolhidos nas mesmas zonas do corpo, os dados foram normalizados por zona de extração. Foi construído um *denoising autoencoder (DA)* com três camadas de 50 neurónios. Esta última técnica ajudou a remover ruído existente nos sons. Em termos de métodos de classificação foram utilizadas duas *Suport Vector Machines* (SVM), uma para a classificar sibilâncias e a outra para classificar fervores, conseguindo obter resultados de exatidão de 86% e 74%, respetivamente. O método proposto requer apenas que 5% dos ficheiros de som sejam anotados manualmente antes da extração de atributos significativos, poupando assim grandes esforços e recursos às equipas de investigação clínica.

No estudo [27] foi desenvolvida uma abordagem baseada em *Gaussian Mixture Model* (GMM) para detetar sibilâncias. Os sons respiratórios foram adquiridos utilizando um

estetoscópio de alta frequência no diafragma e no pescoço. Foram aplicados aos sinais um filtro com uma frequência de 4000Hz e eliminada as frequências <20Hz (zona de sons musculares e alguns sons cardíacos). Os sons respiratórios foram processados usando uma *Fast Fourier Transform* (FFT) discreta e aplicada uma janela de Hamming com um comprimento de 58ms com uma sobreposição de 50%. Foram extraídas para treino do classificador 2100 atributos de sibilâncias e 3700 atributos de sons normais a partir de MFCCs. De modo a que este não detetasse ruído espectral, foi aplicado o algoritmo Shabtai-Musih. Para que os sons respiratórios fossem classificados, foi treinado um modelo GMM. Este método foi dividido em duas fases: treino e reconhecimento. Na primeira fase foi usado o algoritmo *Expectation-Maximization* (EM) para construir modelos acústicos de sons respiratórios normais e sibilâncias. Na segunda fase foi utilizada a verosimilhança do classificador GMM para estimar qual o melhor modelo que classifica os SR. Para a validação do modelo foram usados 36 ficheiros de áudio. O modelo demonstrou ter um bom desempenho, nomeadamente uma especificidade de 99% e uma sensibilidade de 88% na deteção de sibilâncias principalmente quando eram aplicados poucos atributos como entrada.

Para além dos dois trabalhos anteriores, não foram encontrados outros artigos abordando a segmentação e classificação de sons respiratórios adventícios. Como tal, nos parágrafos seguintes, são revistas estratégias de segmentação e classificação aplicadas a outros tipos de sons (e.g., sons cardíacos), as quais podem ser potencialmente aplicadas a sons respiratórios.

Em [28] é proposto um método baseado em Bi-LSTM com técnicas de atenção para segmentar fonocardiogramas (PCG) mesmo com ruído e sinais irregulares, sendo este o método mais promissor para segmentar sons cardíacos. Indo de encontro aos autores, esta metodologia pode ser aplicável para deteção e segmentação de outros sinais biomédicos, como por exemplo sons adventícios, razão pela qual descrevo esta abordagem. Foram utilizados três conjuntos de dados diferentes: *PhysioNet/ CinC Challenge 2016* (PCC), *M3dicine Human Heart Sound Database* (M3-Hu) e *M3dicine Animal Heart Sound Database* (M3-An), compostos por sons cardíacos de seres humanos e animais. O PCC contem arquivos obtidos em âmbito clínico e configurações não clínicas, tendo uma enorme variedade de dados em termos de pacientes e qualidade de dados, sendo constituído por 3126 gravações. O M3-Hu consiste num arquivo de 170 gravações digitais de sons cardíacos saudáveis. Já o M3-Na contém 105 sons digitais extraídos de uma vasta gama de espécies animais. Foram avaliados diferentes atributos desde o método clássico de onduletas até MFCCs e deltas. O método proposto foi testado e comparado com outras abordagens (DNN, CNN, General recursive neural network (GRNN), LSTM e Bidirecional LSTM (Bi-LSTM)). Os resultados para os três conjuntos de dados foram medidos através da exatidão tendo sido obtidos os seguintes valores: PCC com 96.9% ±0.13, M3-Hu com 97.1% ±0.32 e M3-An com 96% ±0.17.

No estudo em [29] propôs-se uma abordagem para deteção de eventos sem utilizar explicitamente informação à priori sobre a duração do estado. O seu estudo foi direcionado para a análise de sons do coração, tendo sido definido um problema de dois focos: segmentação e classificação. Para atingirem o objetivo de segmentação do som utilizaram uma abordagem baseada em redes neuronais profundas recorrentes. As RNNs são apropriadas para este estudo uma vez que aprendem dependências temporais contidas nos

dados por elas próprias. O som do coração é considerado um evento monofónico, apesar de este poder ser contaminado com sons referentes a outras partes do corpo, o que constitui um cenário polifónico. As LSTMs são definidas como DNNs capazes de modelar dependências temporais, tendo já sido aplicadas para deteção de eventos polifónicos. GRNNs são uma simplificação das LSTMs que contêm menos parâmetros, mas demonstram um desempenho comparável. Assim sendo, foram as redes escolhidas para fazer a segmentação do som pretendida. O conjunto de dados utilizado foi retirado de uma base de dados de sons do coração, *PhysioNet/CinChallenge* 2016 [30] (utilizado no artigo indicado anteriormente). Este contém sons de pacientes normais e pacientes patológicos, que se dividiam em 9 grupos. Porém o seu diagnostico é binário (normal ou não normal). Além desta informação, o conjunto de dados ainda fornece anotações sobre o estado de cada uma das amostras de sons do coração (S1, sístole, S2, diástole). Primeiramente foram comparadas diferentes arquiteturas básicas de RNNs, LSTMs, GRNNs, e implementações bidirecionais BiGRNNs que superaram as outras arquiteturas anteriores. Extraíram atributos a partir de espectrogramas e juntamente com os atributos do invólucro construíram o conjunto de dados para análise. Estes dados foram utilizados para treinar modelos BiGRNN. Esta rede consiste em duas camadas ocultas com 200 neurónios cada uma, funções de ativação retificadoras e técnicas de *dropout* para regularização. O método proposto (BiGRNN) alcança níveis de desempenho comparativos ao método LR-HSMM. A proposta deste artigo alcança excelentes resultados tendo F-score 95.6%.

Na abordagem proposta em [31], é apresentada uma primeira construção de um modelo baseado em CNN. Este modelo foi desenvolvido para segmentar sons cardíacos ao contraio da aplicabilidade que outros autores deram às CNNs, nomeadamente para classificar sons cardíacos normais de anormais. A *PhysioNet/CinC challenge* 2016 foi a base de dados utilizada, considerando 427 gravações de sons cardíacos de 130 pacientes. Destes, 181 foram recolhidos a partir de pacientes com patologias cardíacas e 246 de pacientes saudáveis. Foi realizado um pré-processamento aos sinais PCG e extraídos atributos de envelogramas. O pré-processamento passou por filtrar a frequência dos sinais, realizando um corte de 25Hz e 400Hz. Os atributos foram extraídos de 4 envelopes diferentes, sendo estes normalizados de forma a minimizar a variância dos atributos recolhidas. Para cada som cardíaco, são extraídas 4 fases do sinal a partir dos envelogramas normalizados, em que t indica o instante temporal e $s(t)$ é definido como a sequência que contem o estado associado a cada instante, ou seja, o estado 0 corresponde a um som S1 (primeiro som cardíaco), o estado 1 a um intervalo sistólico, o estado 2 corresponde ao som S2 (segundo som cardíaco), o estado 3 a um intervalo diastólico. De seguida é treinada a CNN com funções de ativação *ReLU*, a qual possui uma camada Max *pooling* responsável pela subamostragem das saídas das camadas escondidas. E por último, usa os pesos resultantes do treino, de forma a aplicá-los numa CNN de teste. A proposta apresentada obteve uma sensibilidade média de 93.4%, identificando quase na perfeição todos estados presentes na sequência de sinais.

O artigo [32] tem como base o desafio *Acoustic Scenes and Events* (DCASE) 2016 que tem como objetivo a classificação de 15 sons interiores e exteriores, tais como autocarro, café, carro, centro da cidade, floresta, biblioteca, comboio, entre outros, contendo um dos maiores e variados conjuntos de dados desta área, possuindo 13 horas de gravação. Destes, foram extraídos atributos a partir de 6 técnicas diferentes de processamento de sinal (MFCC, Binaural MFCC, log espectro de Mel e duas escalas temporais diferentes, extraídas usando OpenSMILE [33]). Desenvolveram-se vários, classificadores, tanto baseados em abordagens

clássicas como (GMM e *i-vector*), como baseados em arquiteturas de redes neuronais e aprendizagem profunda (DNN, RNN, CNN). A arquitetura destes foi alterando ao longo da experiência de comparação obtendo-se um total de 30 modelos para o problema de classificação e cinco arquiteturas diferentes. Depois de treinados e testados, o classificador com o desempenho mais baixo foi o GMM, independentemente dos atributos utilizadas. Concluiu-se que os modelos de redes neuronais apresentavam um melhor desempenho quando o número de atributos de entrada aumentavam.

A deteção de eventos raros em áudio a partir da segmentação de sinais é o objetivo da proposta [34]. Apresenta-se um sistema híbrido que combina uma rede LSTM com uma rede convolucional com entrada de uma dimensão (1D ConvNet). Os ficheiros sonoros proveem da base de dados *DCASE 2017 Challenge Task 2* que contém eventos sonoros com sons de bebés a chorar, vidro a partir e tiros. Como ruído de fundo existem 15 cenas de áudio diferentes. Foram criados 4 conjuntos de misturas de som, contendo cada um 15000 ficheiros de som, tendo em conta o balanceamento das classes (5000 por classe), contendo todas 30s, sendo estes divididos em treino e teste numa proporção de 80-20%, respetivamente. Os sinais foram convertidos em espectrogramas de Mel a partir de uma janela de 46ms e sobreposição de 50%, extraindo 128 *mel-filter banks* do espectro de cada *frame*. Por último, cada espectrograma de mel é dividido usando um intervalo de tempo em segmentos. Os segmentos são usados para alimentar a rede (1D ConvNet) que converte as *frames* do espectro em características espectrais. Esta é composta por uma camada 1D convolucional, executado o processo de *Batch Normalization*, aplicada a função de ativação ReLU e no final uma camada *pooling* e aplicado um *dropout* de 0.3. Os atributos que saem desta rede irão alimentar a rede LSTM. Esta foi estruturada para ser unidirecional *backward*, sendo utilizada como função de ativação a função *tanh* (tangente hiperbólica) e aplicado um *dropout* de 0.3 nas duas camadas internas LSTM. O modelo é treinado usando o algoritmo de otimização Adam, um *mini-batch* de 256 e um rate de aprendizagem inicial de 0.001 e reduzindo 0.01 a cada época. Depois do modelo ser treinado foi necessário definir um limiar para determinar a ausência ou presença de um evento sonoro a partir da probabilidade da sequência. O sistema no final conseguiu resultados com um F-score para cada classe superior a 90% (choro de bebé 97.6%, vidro a partir 99.6%, tiro 91.6%).

3.3 Abordagens e limitações do estado de arte

Existem muitos estudos sobre a classificação de sons adventícios utilizando técnicas de aprendizagem profunda e abordagens clássicas. Há 10 anos, as técnicas clássicas eram as mais utilizadas, contrastando com a aplicação de redes neuronais profundas que só se iniciou mais tarde como comprovam os estudos em [35] [17]. Para além disso, as aplicações de técnicas de aprendizagem profunda nas áreas da informática na saúde também eram bastante escassas, como se pode observar a partir da Figura 13. Os estudos apresentados no decorrer deste capítulo, possuem limitações tais como: falta de variedade em termos de conjunto de dados utilizados, utilização de amostras de áudio muito curtas, pouca variedade de técnicas de extração de atributos baseando-se quase sempre em MFCCs e escassa existência de estudos referentes a segmentação de SR. Esta última limitação tem um maior peso que as outras uma vez que os estudos apresentados apenas se referem a sons não respiratórios (muitas vezes

referentes a sons cardíacos). As técnicas apresentadas devem poder ser aplicadas a SA, mas ao contrário dos sons cardíacos, que são bastante regulares e que podem ser recolhidos a partir de 2 ou 3 zonas do corpo, os SA são altamente variáveis entre os pacientes e podem ser recolhidos a partir de diversas zonas do corpo, diferentes partes constituintes do pulmão ou vias respiratórias [26], podendo assim dificultar a sua classificação.

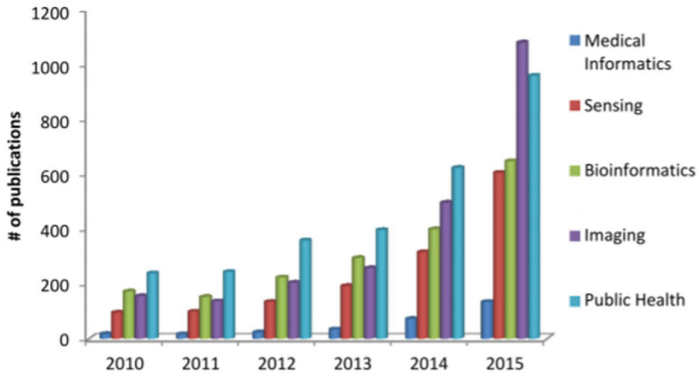


Figura 13 - Distribuição de artigos publicados que usam aprendizagem profunda em subáreas da informática em saúde. As estatísticas de publicação são obtidas no Google Scholar [17]

Os resultados e as conclusões retiradas pelos autores dos artigos analisados nesta revisão são, por vezes, sobrevalorizados devido à falta de informação e métricas mais específicas. Em concreto, os artigos referem apenas valores globais, inflacionados pela falta de balanceamento de classes muitas vezes omitida pelos autores. Desta forma a taxa de sucesso é influenciada pela existência de muitos falsos positivos.

Outro problema encontrado na análise dos artigos está relacionado com a comparação dos resultados dos estudos com abordagens que alcançaram valores inferiores, ou seja, a pesquisa de trabalhos existentes na área foi limitada.

Através das Tabela III e IV pode-se resumir as características principais dos estudos anteriormente referidos, observando-se que as abordagens clássicas demonstraram ter bons resultados, mas as técnicas baseadas em aprendizagem profunda apresentam resultados superiores.

Tabela III – Resumo dos artigos sobre diversos sons

Referência	Amostra	Local de gravação	Sons	Extração de atributos	Classificadores	Segmentação	Resultados
[26]	4 centros clínicos 284 pacientes 11627 arquivos de som respiratório	11 zonas de auscultação	Sibilâncias e fervores	---	*denoising autoencoder* + duas SVM	---	Exatidão = sibilâncias 86% e fervores 74%
[27]	2100 atributos de sibilâncias e 3700 atributos de sons normais validação = 36 ficheiros de áudio	Diafragma e pescoço	Sibilâncias e sons respiratórios normais	MFCC	GMM	---	Especificidade = 99% Sensibilidade = 88%
[28]	(1) 3126 gravações de anomalias cardíacas (2) 170 sons normais (3) 105 sons de animais	---	Sons cardíacos de seres humanos e animais	Onduletas MFCC Envelope	Bi-LSTM + Attention	Bi-LSTM + Attention	(1) Exatidão = 96.9% ±0.13 (2) Exatidão = 97.1% ±0.32 (3) Exatidão = 96% ±0.17
[29]	3126 gravações de anomalias cardíacas	---	Sons cardíacos humanos	Melódicas Onduletas Envelope	BiGRNN (camadas ocultas com 200 neurónios cada uma Funções de ativação retificadoras)	-	Sensibilidade méd. = 96.1% F méd. = 95.6% Erro méd. 0.09%
[31]	427 gravações 130 pacientes 181 sons de patologias cardíacas 246 de pacientes saudáveis	---	Sons cardíacos humanos patológicos e saudáveis	---	CNN	CNN	Sensibilidade média = 93.4% ! Falta de informação sobre resultados mais

							específicos de classificação e segmentação
[32]	13 horas de áudio	---	15 sons indoor e outdoor	MFCC Binaural MFCC Log espectro de Mel Duas diferentes escalas temporais extraídas usando OpenSMILE	(1) GMM (2) i-vector (3) DNN (4) RNN (5) CNN	---	Melhores resultados: (1) Exatidão Média = 72.5% (2) Exatidão Média = 81.7% (3) Exatidão Média = 84.2% (4) Exatidão Média = 80.2% (5) Exatidão Média = 82.2%
[34]	15000 gravações	---	Bebés a chorar, vidro a partir e tiros	MFCC Log espectro de Mel	Sistema híbrido (1D ConvNet + LSTM)	Sistema híbrido (1D ConvNet + LSTM) através de um Limiar	F-score: choro de bebé = 97.6% vidro a partir = 99.6% Tiros = 91.6%.

Tabela IV – Resumo dos artigos baseados em RSD

Referência	Amostra	Classes	Extração de atributos	Classificadores	Segmentação	Resultados
[19]	6898 ciclos respiratórios reais 126 participantes 963 sibilâncias e 4158 fervores computados	Sibilâncias e fervores	Atributos espectrais MFCC	*Random Forest Classifier*	EMD	2 segmentos do sinal: Exatidão = 88%, Especificidade = 97%, Precisão = 91% 4 segmentos do sinal: Exatidão = 69%, Especificidade = 94% Precisão = 66%
[10]	6898 ciclos respiratórios 126 participantes 963 sibilâncias e 4158 fervores computados	Sibilâncias e fervores	Atributos espectrais Atributos MFCC Atributos Melódicos	LDA SVMrbf RUSBoost CNN	---	De entre todos os conjunto de dados os melhores resultados formam obtidos quando classificavam 3 classes Exatidão = 96,9% Exatidão = 81.8%
[20]	Segmentos de 1s cada ficheiro de áudio	Tipos de doenças	MFCCs, taxa de cruzamento zero, coeficientes harmónicos, codificadores preditivos lineares e predição linear percetual	LSTM	---	Exatidão 46%
[21]	6898 ciclos respiratórios reais 126 participantes + Aumento de dados	4 classes: normal, fervores, sibilância, dois e 2 classes: Normal, anormal	Log espectro de Mel	CNN	---	2 classes: Exatidão = 77%, especificidade=80.9% sensibilidade =73.1% 4 classes: Exatidão =56.2% Sensibilidade=40.1% Especificidade = 72.3%

[22]	6898 ciclos respiratórios reais 126 participantes + Aumento de dados	normal, fervores, sibilância, dois	Espectrogramas	CNN e Autoencoder	---	Autoencoder: pontuação média = 43% especificidade = 62% sensibilidade = 33% CNN + Autoencoder: F-score = 49% pontuação média = 42% especificidade = 69% sensibilidade = 30%
[23]	segmentos de som que contenham uma das classes	fervores, sibilâncias e sons normais	Espectrogramas de OST	Rede neural residual profunda	---	Exatidão = 98.79% sensibilidade = 96.27% especificidade = 100%
[24]	6898 ciclos respiratórios segmentado com base numa janela deslizante (25ms a 500ms) usando uma sobreposição de 50%	2 classes: fervores, sibilância 4 classes: normal, fervores, sibilância, dois	MFCCs	LSTM GRU BiLSTM	---	4 classes: Exatidão = 81% 2 classes: Exatidão = 74%
[25]	6898 ciclos respiratórios reais 126 participantes	sibilâncias ou fervores, os dois ou normais	espectrogramas e onduletas	CNN	---	espectrogramas: *F1-score* = 46% sensibilidade = 36% especificidade = 69% onduletas: *F1-score* = 46%, sensibilidade = 37% e especificidade = 62%

4 Classificação de eventos adventícios

A primeira fase do trabalho consistiu no desenvolvimento de modelos capazes de distinguir eventos (pré-segmentados manualmente) com sibilâncias, fervores (realçando o facto de estes também possuírem ruído, uma vez que foram captados em ambiente hospitalar) ou outro som ausente de sons adventícios podendo este ser som respiratório normal ou qualquer tipo ruído. Os áudios sofreram um pré-processamento antes de segmentarmos manualmente os eventos. Através destes realizou-se a extração de atributos, a sua análise e posteriormente a seleção das que melhor se adaptavam ao problema. Os conjuntos de dados necessitaram de sofrer um balanceamento de classes uma vez que as suas amostras se encontravam bastante desequilibradas. Foram elaboradas duas abordagens: com algoritmos clássicos (SVM e k-NN) e com algoritmos de aprendizagem profunda (CNN, LSTM, BiLSTM e GRU). Por fim, comparámos o desempenho alcançado a partir de modelos das duas abordagens. Estes modelos foram testados com diversos atributos de entrada de forma a comparar o desempenho dos diferentes modelos relativamente ao nível do tipo de dados que lhes é fornecido.

4.1 Pré-Processamento

Começámos por analisar os ficheiros de texto referentes aos eventos existentes nos ciclos respiratórios para averiguar de que forma é que estes estavam anotados. De seguida, desenvolvemos um *script* para cortar os 920 ficheiros de áudio existentes na base de dados (ciclos respiratórios), de forma a obter um ficheiro de áudio por evento detetado na respiração. Cada ficheiro de eventos possuía o início e fim da ocorrência, em segundos, de diversos sons adventícios ao longo de um ciclo respiratório. Foram extraídos 10 766 ficheiros de áudio relativos a eventos, sendo estes: 1898 sibilâncias e 8877 fervores e 5024 de outros sons existentes nos ficheiros, obtendo um total de 15790 ficheiros de áudio. Como se pode observar na Figura 14, o conjunto de dados possui um grande desequilíbrio em relação ao número de amostras de casa classe.

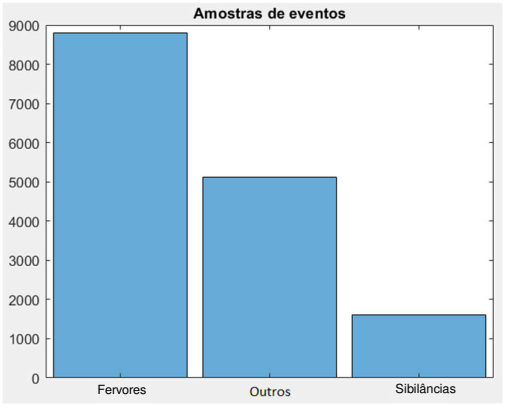


Figura 14 - Número de amostras por classe

De acordo com o trabalho de investigação desenvolvido em [10] sobre a base de dados RSD, concluiu-se que a informação do sinal mais interessante se encontrava abaixo de 2000Hz. Deste modo, aplicou-se uma reamostragem dos dados para uma taxa de amostragem de *4000Hz* a todos os arquivos existentes, uniformizando-os e respeitando o Teorema de *Nyquist*.

4.2 Abordagem com algoritmos clássicos

Esta abordagem vai ao encontro ao objetivo de classificar eventos adventícios a partir de algoritmos de aprendizagem computacional clássicos. A secção foca-se na extração de atributos para estes classificadores, assim como também na análise de dados tendo em conta o impacto dos atributos para cumprir o objetivo, como por exemplo a correlação que existe entre os atributos e a classe dos eventos adventícios a classificar. Os algoritmos SVMrbf e k-NN foram alimentados com diferentes conjuntos de dados, sendo também realizadas diversas experiências de forma a encontrar o melhor classificador possível. Estes pontos irão ser abordados de forma detalhada nas subsecções seguintes.

4.2.1. Extração de atributos

Com o objetivo de classificar os diferentes sons a partir de abordagens clássicas, foi necessário extrair dos mesmos algumas características, sendo esta a primeira fase da criação de um dos conjuntos de dados. Como os ficheiros de áudio dos eventos têm uma duração extremamente reduzida foi fundamental definir um *hopsize* e tamanho da janela em ms, de maneira a segmentar o sinal do evento em múltiplas *frames*, mais concretamente de 16ms em 16ms (tamanho da janela de Hamming) e um *hopsize* de 4%. A representação deste processo encontra-se na Figura 15.

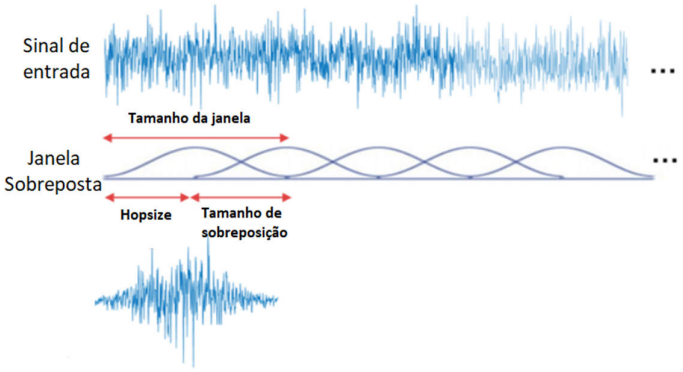


Figura 15 - Demonstração do processo inicial de extração de atributos [43]

Também foi utilizada a técnica de *zero padding* a alguns sinais, uma vez que os métodos da *MIRtoolbox* 1.7.2 definidos para extrair os atributos (Tabela V) estavam com dificuldades em processar áudios com uma duração tão pequena. Deste modo, foram acrescentados 0 ao final de cada som de forma que os todos ficheiros fossem processados com os mesmos parâmetros (*Zero Padding*), ficando com uma duração de 1s. A base de dados continha 87 eventos anotados com mais de 2s. Nestes casos foi considerado apenas o primeiro segundo, visto que a anotação desses eventos era menos precisa.

A partir dos espectrogramas criados computacionalmente através da *toolbox*, extraímos 81 atributos por cada *frame*, sendo estas: 25 referentes ao espectro, 26 MFCC e 30 melódicas. A Tabela V apresenta os diferentes tipos de atributos extraídos de acordo com a sua especificidade. Como os atributos foram extraídas em relação a cada *frame* do espectrograma, foram calculadas 5 estatísticas: média, mediana, desvio padrão, valor máximo e valor mínimo para cada um dos 81 atributos, obtendo no final 405 atributos por evento respiratório. Todos estes atributos e a respetiva classe (tipo de evento) estão armazenados em ficheiros *.mat* (ficheiros de dados binários do Matlab) para serem à posteriori analisadas, criando assim um primeiro conjunto de dados de eventos respiratórios composto por atributos.

Tipo de atributos	Nome dos atributos
Melódicos	Pitch Pitch Smoothing Inharmonycity Inharmonycity Smoothing Voicing Voicing Smoothing
MFCC	MFCC Delta- MFCC
Espectrais	Centroid Spread Skewness Kurtosis Zero-crossing Rate Flatness Roughness Irregularity Flux Flux Inc Flux Halfwave Flux Median Brightness Brightness 400 Ratio Brightness 800 Ratio Roll-off Outlier Ratio Interquartile Ratio

Tabela V – Atributos extraídos[6]

4.2.2. Análise Exploratória de dados (AED)

Ao longo do trabalho desenvolvido foram realizadas análises distintas com diferentes finalidades.

Análise #1

O conjunto de dados é constituído por múltiplas *frames* (segmentos de uma janela temporal que compõem um evento) de diversos sons com sibilâncias, extraídas a partir da

[6] Por uma questão de preservar o sentido dos atributos extraídos, estas ser nomeadas em inglês

auscultação de vários pacientes utilizando estetoscópios diferentes em localizações diferentes do corpo do paciente. A este respeito, colocam-se as seguintes questões:

- Será que os atributos extraídos poderão estar a discriminar sons/*frames* com base nas características intrínsecas dos aparelhos de gravação (estetoscópios) e/ou com base na localização dos mesmos?
- Será que existem tipos de atributos (espectrais ou melódicas) que são mais sensíveis em termos de variação de estetoscópio?

Esta experiência teve como base a implementação de uma *One-Class* SVM para classificar *frames* e, através desta, identificar possíveis *frames* consideradas como anomalia. Primeiramente foram construídos pequenos conjuntos de dados com base na localização do estetoscópio que conteriam apenas sibilâncias de forma a uniformizar ao máximo as variáveis de cada som presente no conjunto de dados. Utilizámos apenas os dois atributos com mais peso/mais relevantes na classificação dos sons tendo em conta os estudos previamente realizados [10], partindo do mesmo conjunto de dados e dos mesmos atributos extraídas anteriormente. Para os atributos espectrais classificaram-se *frames* das "*Centroid*" e "*Brightness* 4" ratio e para os atributos melódicos as "*MelpitchHF*" e "*Melinharmonicity*".

A *One-Class* SVM é um algoritmo de aprendizagem não supervisionada que irá tratar os elementos do conjunto de dados, como pertencentes a uma só classe. Nesta experiência, o *One-Class* SVM irá encontrar a fronteira não linear de acordo com os dados fornecidos que permite identificar todas as *frames* que não se enquadram dentro da fronteira como anomalias. Este caso observa-se na Figura 16. O limite que separa as anomalias do resto dos dados ocorre onde o valor do contorno é 0.

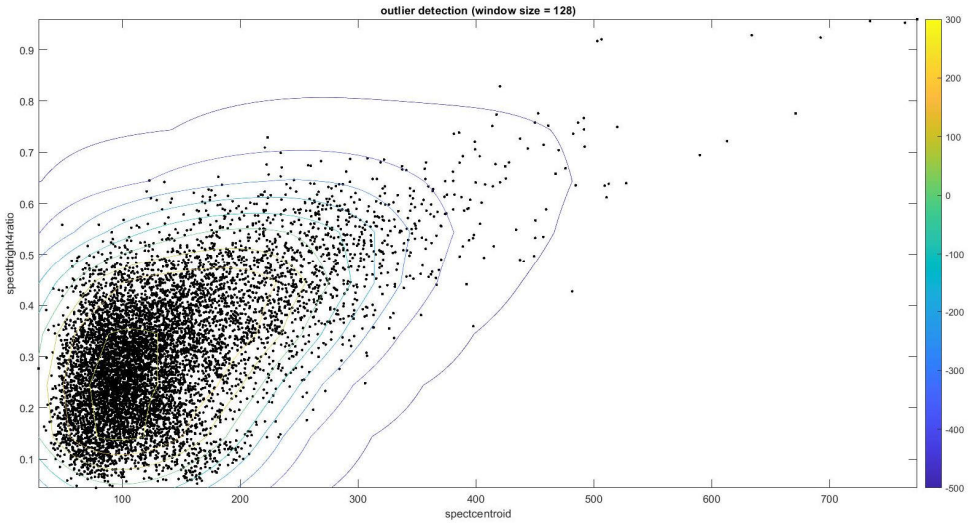


Figura 16 - Fronteiras de decisão da One Class SVM com atributos espectrais

Posto isto, são consideradas anomalias todas as *frames* com a pontuação inferior a 0. Nos gráficos das Figuras 17 e 18, são apresentadas as pontuações que cada *frame* obteve. Usando os atributos "*centroid*" e "*brightness* 4 ratio" obtiveram-se:

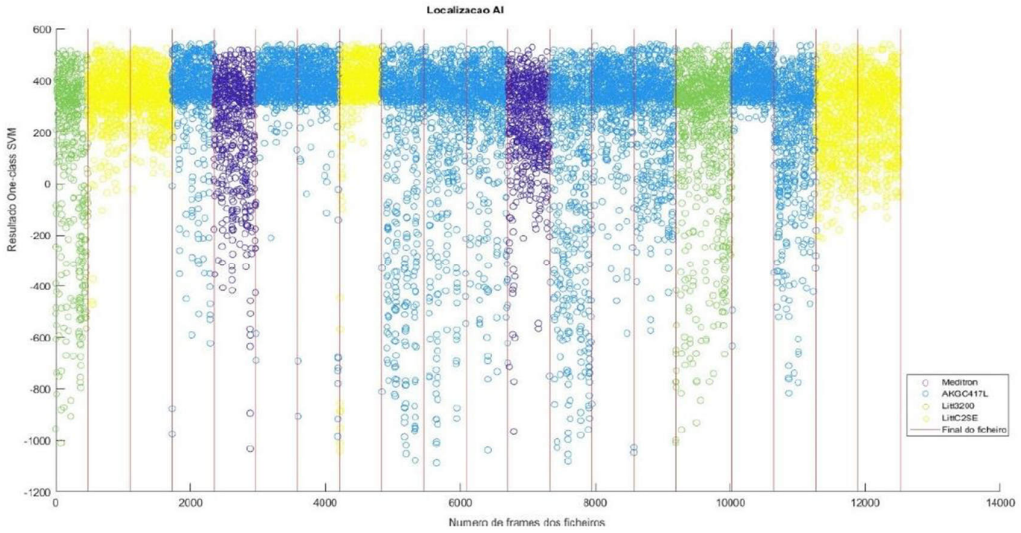


Figura 17 - Resultado One class SVM anomalias Al

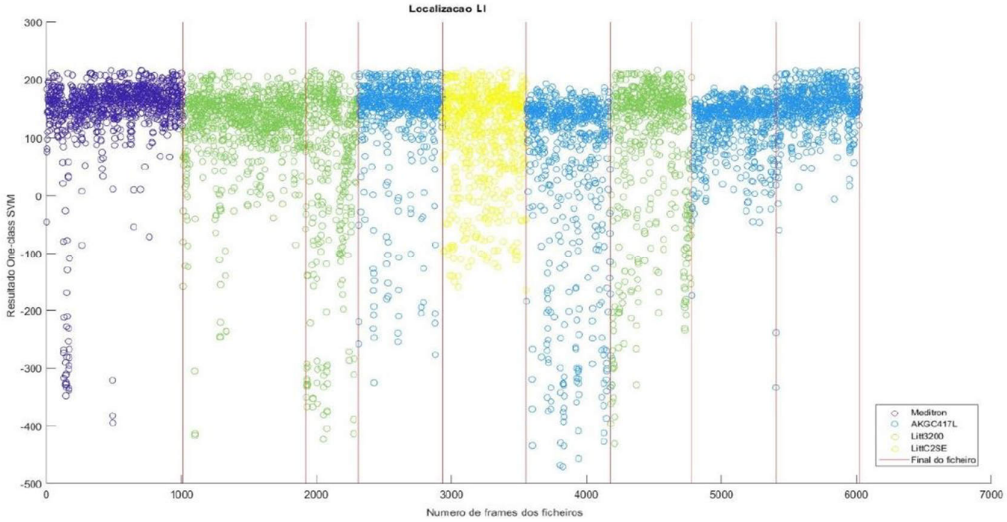


Figura 18 - Resultado One class SVM anomalias Ll

Usando os atributos melódicas *"MelpitchHF"* e *"Melinharmonicity"*:

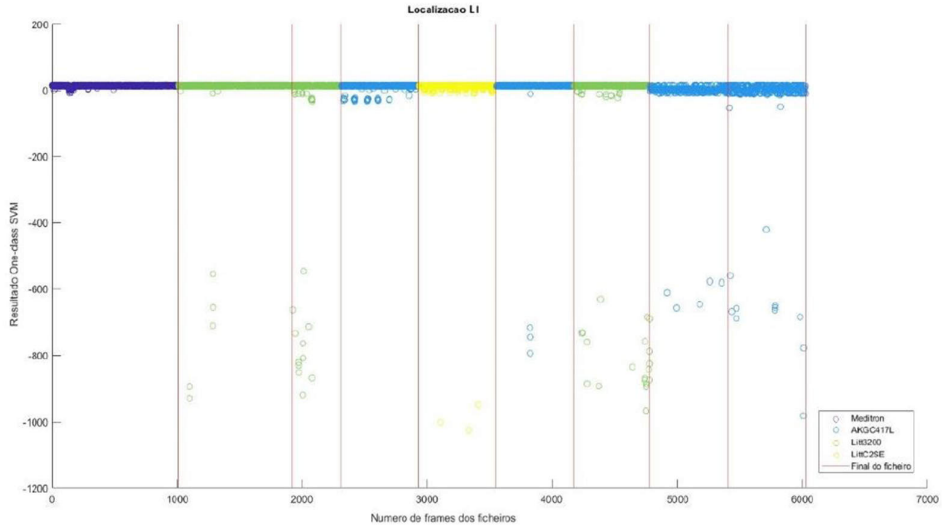


Figura 20- Resultado dos atributos melódicos na localização LI

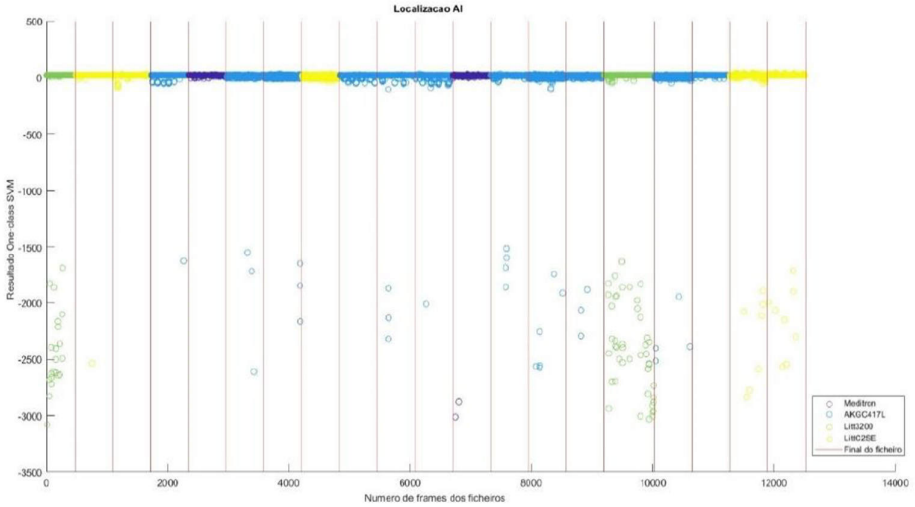


Figura 19 – Resultado dos atributos melódicos na localização AI

Na segunda experiência, filtrámos ainda mais este conjunto de dados, apenas contendo sons de sibilâncias com espectrogramas semelhantes (analisados e comparados visualmente). A partir da Figura 21, é possível compreender a dificuldade a olho nu de distinguir espectrogramas de ciclos respiratórios mesmo utilizando mecanismos de recolha diferentes. Estes espectrogramas são uma boa representação de como irão ser as imagens que alimentarão os classificadores de aprendizagem profunda.

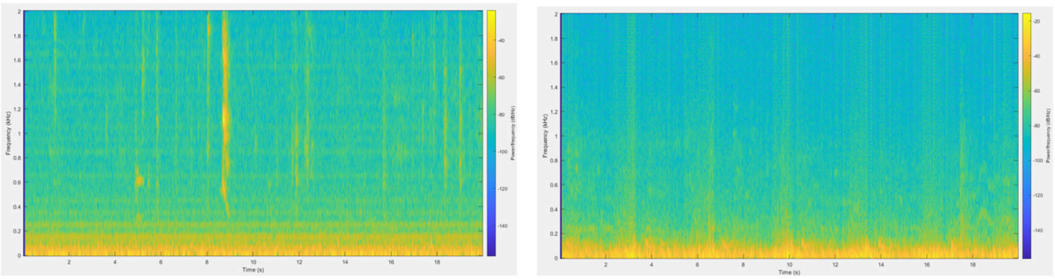


Figura 21 - Espectrogramas de sibilâncias da localização LI

Apesar da dificuldade, através das suas características conseguimos realizar uma boa filtragem por inspeção visual, sendo criados gráficos cada vez mais semelhantes. A figura 22 representa uma amostra dos resultados desta experiência. Todos os gráficos podem ser observados no Apêndice A.

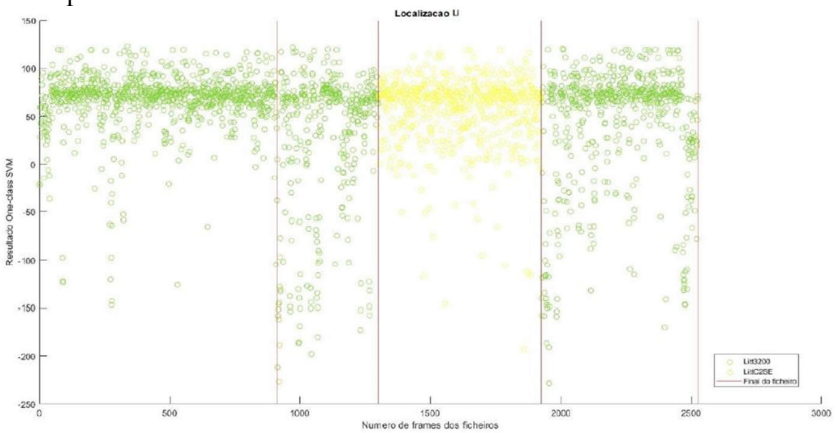


Figura 22 – Resultado da classificação das *frames* na localização LI

Através do primeiro conjunto de gráficos, apercebemo-nos que a variação dos estetoscópios não é um problema, visto que existem bastantes anomalias independentemente do estetoscópio utilizado. Não sabemos até que ponto a classificação das *frames* melódicas versus espectrais é relevante uma vez que estas tomam valores bastante diferentes. Os atributos melódicos possuem bastantes valores nulos nas suas *frames* e isso deve ter interferido na classificação do SVM.

No segundo conjunto de gráficos os sons foram filtrados de uma forma mais rigorosa utilizando os espetrogramas para uniformizar o máximo possível os sons dos conjuntos de dados. Neste caso pode-se concluir que as *frames* extraídas usando o LittC2SE são as que são menos classificadas como anomalias. Estas, quando colocadas em classificação com *frames* de todos os outros estetoscópios, são as possuem menos anomalias (caso Lr e Al).

Pode-se concluir que o classificador não está a aprender a classificar os diferentes estetoscópios relativamente às *frames*. Importa também referir que, certamente, existiram alguns fatores a influenciar estos atributos para além do estetoscópio (espaço de recolha, duração do som…).

Análise #2

A seleção de atributos é um fator extremamente importante para enriquecer o conjunto de dados fornecido como entrada a qualquer classificador. Desta forma, é essencial compreender quais das estatísticas dos atributos extraídas na Secção 4.2.1 são as mais adequadas para identificar e distinguir os eventos respiratórios. Colocam-se então as seguintes questões:

- Qual/quais o(s) método(s) estatístico(s) mais indicado(s) para realizar a seleção de atributos?
- Que relação existe entre os atributos e o resultado esperado?

Devido ao facto do conjunto de dados criado possuir 15790 eventos distintos, para realizar a análise estatística necessária utilizámos uma amostra com cerca de 10% dos eventos totais (773 fervores, 227 sibilâncias e 500 normais), devido a limitações computacionais e com vista a podermos extrapolar os resultados obtidos com esta amostra para o conjunto de dados total.

Para identificar quais os testes estatísticos adequados à tarefa de seleção de atributos, começámos por aplicar o teste de Shapiro-Wilk para verificar se os dados seguem uma distribuição normal. De acordo com os resultados obtidos, apresentados na Figura 23, podemos aferir que os atributos possuem valores de *p-value* bastante diferentes ao longo da análise, quase nulos. Posto isto, a H0 do teste de Shapiro-Wilk, "a amostra provém de uma população normal", é refutada e podemos afirmar com um nível de significância de 5% que a amostra não provém de uma população normal.

```
ShapiroResult(statistic=0.9131295680999756, pvalue=1.8143986190840419e-28)
ShapiroResult(statistic=0.9483645558357239, pvalue=1.5744757424803478e-22)
ShapiroResult(statistic=0.8441915512084961, pvalue=5.12210757905643e-36)
ShapiroResult(statistic=0.9117304086685181, pvalue=1.1605119399317916e-28)
ShapiroResult(statistic=0.8509692549705505, pvalue=2.0746465101555637e-35)
ShapiroResult(statistic=0.8027223944664001, pvalue=2.4864513834117765e-39)
ShapiroResult(statistic=0.9200136661529541, pvalue=1.780184810659743e-27)
ShapiroResult(statistic=0.6537182927131653, pvalue=0.0)
```

Figura 23 - Resultado do teste de Shapiro-Wilk

A legenda da classe presente nos conjuntos de dados foi convertida de categórica para numérica, sendo fervores (1), outros sons (2) e sibilâncias (3). De seguida vamos verificar qual a relação entre os atributos extraídas e o resultado esperado a partir da correlação dos atributos.

4.2.2.1. Seleção de atributos

Uma vez que foram calculados e extraídos 405 atributos é necessário reduzir o seu número, devido tanto à limitação computacional como com vista a evitar redundâncias e *overfitting*. Tendo em conta a análise anterior, foram aplicados dois métodos de seleção de atributos, o teste Kruskal-Wallis, o *Minimun Redundancy Maximum Relevance* (MRMR) e o Relieff.

Estes três métodos qualificaram os 405 atributos, do melhor para o pior, em relação ao grupo-alvo a classificar (sibilâncias, fervores e outros). Desta forma conseguimos extrair dois bons conjuntos de atributos para aplicar em métodos de segmentação e classificação. O número de atributos a selecionar irá sendo variado para testar o comportamento de diversos classificadores, mas atualmente o conjunto de dados é constituído pelos 25 melhores atributos classificados pelo MRMR, Kruskal-Wallis e Relieff, respetivamente.

Foram construídas três matrizes de correlação, Figuras 24, 25 e 26, para conseguirmos visualizar quais os atributos com mais correlação relativamente à classe. Os atributos com mais correlação entre si, possuem uma grande quantidade de informação semelhante, podendo levar a redundâncias futuras caso não sejam identificados.

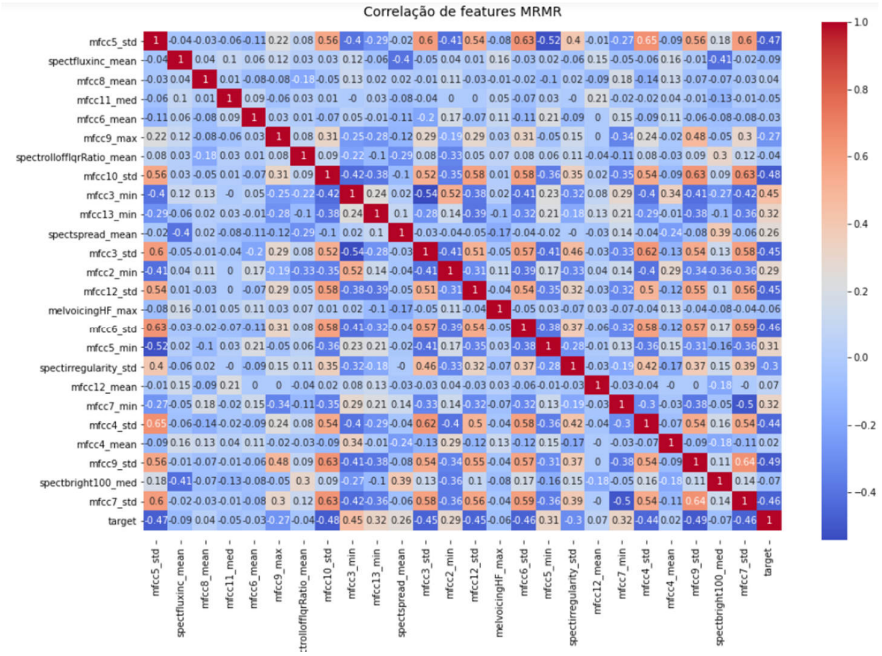


Figura 24 - Matriz de Correlação MRMR

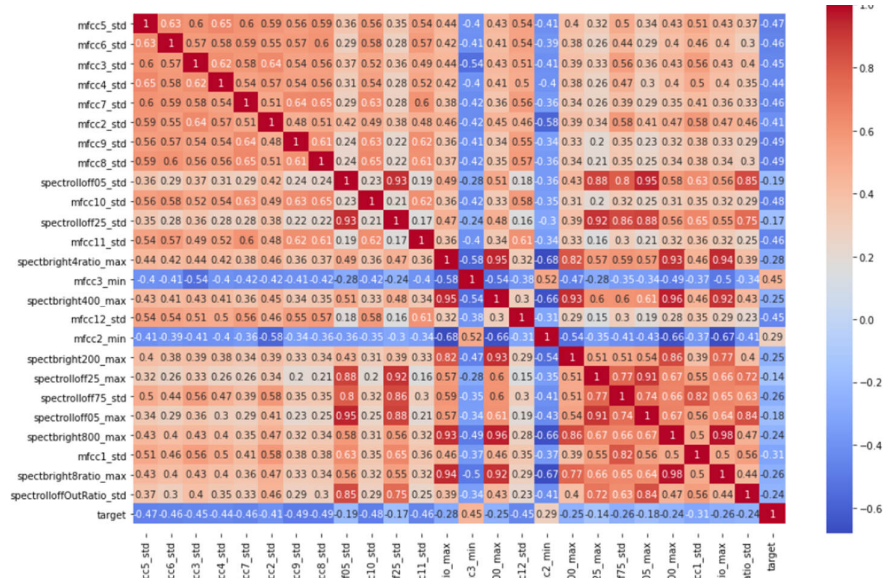


Figura 25 - Matriz de Correlação Kruskal-Wallis

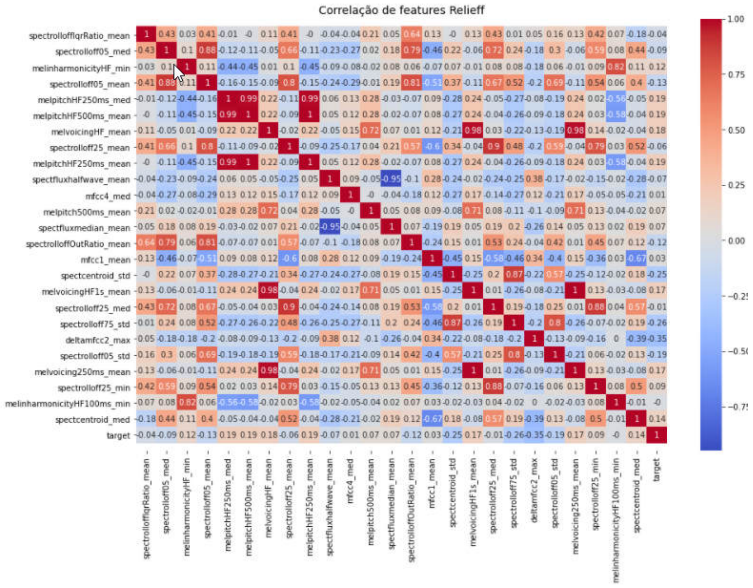


Figura 26 -Matriz de correlação de Relieff

O método MRMR realiza a seleção de atributos tendo em conta a similaridade dos mesmos, eliminando a redundância entre os atributos. O selecionador Kruskal-Wallis determina o poder discriminatório dos atributos através do qual se calculou o *ranking* usando o valor de *chi-square* para selecionar aquelas que possuem um maior valor, que apresenta valores de correlação bastante elevados. A correlação obtida não é muito elevada, e é principalmente negativa (-0,4). Este fator não é necessariamente desvantajoso uma vez que correlação não indica casualidade. Ainda assim o selecionador MRMR apresenta uma correlação mais acentuada entre os atributos e o resultado esperado. Estes três métodos foram aplicados para poder comparar os resultados dos classificadores com diferentes elementos de entrada.

4.2.3. Balanceamento de dados

À semelhança da maioria dos conjuntos de dados médicos, a base de dados com que estamos a trabalhar tem um desequilíbrio de classes bastante acentuado. Para prevenir a existência de *overfitting* na fase de treino dos modelos foram aplicadas técnicas de balanceamento de dados. A classe 'outros' sofreu subamostragem, uma vez que compõe mais de metade das amostras do conjunto de dados. Desta forma foram reduzidas amostras aleatórias desta classe. Para diversificar as outras classes aplicámos várias técnicas de aumento de dados evitando a duplicação dos eventos. As técnicas de aumento de dados aplicadas a áudio basearam-se em ligeiras modificações do sinal das amostras. Considerando os dados, foram excluídas técnicas de adição de ruído, uma vez que os áudios foram recolhidos em ambiente hospitalar, possuindo naturalmente ruído ambiental. Foram executadas as seguintes técnicas:

- *Pitch Shiftting* é uma técnica que altera o tom do som ou a frequência do sinal, tornando-os mais graves ou mais agudos. A probabilidade desta técnica ser aplicada no evento é de 60% e a variação do tom entre -2 e 2.

- O método *Time Shifting* consiste na deslocação do áudio com um determinado passo de segundos, tanto para a frente como para trás. Se o som sofrer uma deslocação para a esquerda, os primeiros milissegundos de som serão eliminados e acrescentado silencio no fim. Se for realizado para a direita, na primeira parte do som é acrescentado silêncio. A probabilidade desta técnica ser aplicada no evento é de 90% e a velocidade do áudio variou entre -1 e 1.
- O volume foi outro das características dos sinais do evento. O método de compressão foi utilizado para modificar este parâmetro alterando a intensidade do som, mais alto ou mais baixo. A probabilidade desta técnica ser aplicada no evento é de 70% e a sua variação oscila entre -0.5 e 0.5.

Os eventos adventícios que foram selecionados como amostras para gerar novas amostras da mesma classe foram sempre os mesmos escolhidos para a todos os conjuntos de dados selecionados e a partir destes extrair diferentes atributos. As técnicas por vezes foram aplicadas individualmente e para outras amostras juntámos a aplicação dos procedimentos descritos acima. Para definir os valores de oscilação dos parâmetros baseamos-mos em artigos, como por exemplo [36], que realizaram aumento de dados semelhantes e tivemos em conta o tipo de som com que estamos a trabalhar.

4.2.4. Modelos

Os dois algoritmos de aprendizagem não profunda propostos são: o SVM com a função base radial (SVMrbf) tendo em conta os bons resultados dos artigos [26] e [10] e o k-NN.

O modelo SVM é composto por um núcleo que segue uma função de aprendizagem gaussiana (SVMrbf). Uma vez que estamos frente a um problema não binário foi utilizado um modelo *error-correcting output codes* (ECOC) que realiza a classificação a partir de várias SVM através do método um contra um. O fator de probabilidade dado às classes foi uniforme, ou seja, as probabilidades das classes terão um fator igual.

O k-NN classifica as amostras a partir da distância entre os k vizinhos mais próximos. A métrica de distância foi variada (euclidiana ou correlação) ao longo das experiências com vista a alcançar os melhores resultados possíveis.

4.2.5. Avaliação

De maneira a conseguirmos analisar o desempenho dos algoritmos SVM e k-NN foram realizadas várias experiências que consistem na variação do tipo de atributos que compõem o conjunto de dados. Estes são compostos por atributos extraídos usando a *MIRtoolbox* 1.7.2 (referentes à Secção 4.2.1). A composição dos conjuntos de dados passou por um dos selecionadores de atributos: MRMR, Kruskal-Wallis ou Relieff, referidas na Secção 4.2.2.1. Na utilização do selecionador Relieff variou-se o valor de k (nº de observações vizinhas) entre 15 e 50. A divisão do conjunto de dados em treino e teste baseou-se numa pré divisão dos ficheiros de áudio existentes na base de dados que possuem um rácio de cerca de 60%/40%, respetivamente. Antes de realizar o balanceamento o conjunto de dados para treino era composto por: 1173 'sibilâncias', 4500 'outro', 3053 'fervores'. Após a tentativa de balanceamento este era composto por 4500 'sibilâncias', 4500 'outro', 4500 'fervores'. Analisámos o impacto do balanceamento usando as técnicas referidas na Secção 4.2.3. As

métricas de avaliação usadas foram: Exatidão, especificidade (E), precisão (P), sensibilidade (S) e F1-Score. Cada classificador foi treinado 10 vezes com o propósito de obter os melhores modelos com a capacidade de identificar eventos. As métricas por classe ('sibilâncias', 'fervores' e 'outro') foram calculadas usando a técnica "um contra todos".

Na Tabela VI pode-se observar apenas os resultados obtidos pelos melhores modelos SVM dentro de todas as experiências realizadas, passando estas por alteração da função de decisão da SVM, quantidade de atributos a selecionar pelos selecionadores de atributos, e o tamanho da vizinhança (K).

Tipo de Selecionador de atributos	**Balanceamento**	**S**	**P**	**E**	**F1-Score**	**Exatidão**
Relieff k=15 (20 atributos)	Não	82,82%	81,20%	81,20%	82%	74,97%
MRMR (20 atributos)	Não	83,96%	82,17%	81,2%	83,05%	76,17%
Relieff k=15 (100 atributos)	Sim	42,69%	76,92%	86,79%	54,91%	50,45%
MRMR (100 atributos)	Sim	46,62%	79,85%	87,86%	58,87%	52,8%
Kruskal-Wallis (20 atributos)	Sim	61,92%	67,34%	69,03%	64,52%	57,9%
Kruskal-Wallis (100 atributos)	Não	86,32%	79,81%	77,48%	82,94%	75,39%

Tabela VI – Resultados obtidos a partir de SVM

O SVMrbf obteve o melhor valor de exatidão quando alimentado pelos vinte atributos que o selecionador MRMR indicava como melhores. As técnicas de balanceamento não surtiram efeito uma vez que os valores das métricas são bastante melhores quando o conjunto de eventos fornecido ao classificador não é balanceado.

O algoritmo k-NN também foi treinado com os mesmos conjuntos de dados que o SVM a fim de se realizar uma análise comparativa do desempenho o mais justa possível. A Tabela VII contém os melhores resultados alcançados pelo modelo considerando os diferentes selecionadores de atributos e o balanceamento do conjunto de dados de treino. Ao utilizar o Relieff como selecionador obtivemos uma grande diferença da quantidade de atributos a selecionar dependendo se existiu ou não balanceamento. Com balanceamento de dados obtivemos 66,39% de exatidão usando 100 atributos, já sem balanceamento os melhores resultados ocorreram quando foram apenas selecionadas 45, obtendo 71% de exatidão. Através do selecionador MRMR obtivemos entre o conjunto de dados balanceado e não balanceado uma diferença de exatidão de cerca de 4%. O melhor resultado com este

selecionador foi de 68,31% de exatidão. O valor superior a 85% de sensibilidade é bastante positivo uma vez que este é referente ao acerto dos eventos pelo classificador.

Tipo de Selecionador de atributos	Balanceamento	S	P	E	F1Score	Exatidão
Relieff k=25 (100 atributos) correlação	Sim	72,51%	78,95%	80%	75,59%	66,39%
MRMR k=15 (100 atributos) correlação	Sim	68,21%	79,83%	82,21%	73,55%	64,10%
MRMR k=25 (45 atributos) euclidiana	Não	87,50%	70,52%	62,26%	78,10%	68,31%
Relieff k=25 (45 atributos)	Não	88,13%	72,30%	65,16%	79,43%	71%
Kruskal-Wallis k= 15 (100 atributos) Correlação	Sim	70,94%	78,95%	80,49%	74,74%	66,21%
Kruskal-Wallis k=15 (20 atributos) euclidiana	Não	92,5%	70%	59,25%	79,74%	69,96%

Tabela VII – Resultados obtidos a partir de k-NN

Podemos observar que quando o conjunto de dados não é balanceado o classificador k-NN alcança melhores valores quando são selecionados menos atributos, ao contrário de quando é balanceado usando 100 atributos. O classificador SVM utiliza menos atributos que o k-NN para obter resultados superiores.

De acordo com os resultados apresentados nas Tabelas VI e VII o selecionador de atributos Relieff é o que apresenta melhores resultados para ambos os classificadores, sendo necessárias apenas entre 20 e 45 atributos (dependendo do classificador). O que é um número bastante reduzido em comparação com todos os atributos extraídas. Estes atributos são as que melhor se relacionam com as classes a classificar.

4.3 Abordagem com algoritmos de aprendizagem profunda

Esta abordagem consiste na classificação de eventos adventícios a partir de classificadores de aprendizagem profunda, nomeadamente a partir de arquiteturas CNN, BiLSTM, LSTM e GRU. Os resultados destes algoritmos serão comparados com os obtidos pelos classificadores de aprendizagem computacional avaliados no capítulo anterior de forma a perceber quais são os melhores meios para conseguir identificar corretamente os eventos adventícios. Esta secção foca-se também na geração de novos atributos para alimentar os classificadores de aprendizagem profunda.

4.3.1. Pré-Processamento e geração de atributos

Os atributos selecionados na Secção anterior foram apenas um dos formatos de entrada dos modelos. Após o pré-processamento dos ficheiros de som da base de dados e obtido os eventos de fervores, sibilâncias e outros, queríamos testar diferentes géneros de entradas para avaliar o comportamento dos classificadores dependendo dos dados com os quais os classificadores são alimentados com vista a encontrar o melhor formato de dados para a identificação de eventos. Considerando este fator, gerámos diferentes conjuntos de dados de:

Sinal

Os dados provenientes do sinal dos eventos foram os mais fáceis de extrair. Através da leitura dos arquivos dos eventos conseguimos obter uma matriz de MxN dimensões em que M significa o número de amostras do áudio e o N o número de canais do ficheiro. Neste caso possuíamos apenas um canal ficando cada evento caraterizado por um *array*. Os dados foram normalizados em valores de [-1 1], devido à existência de valores negativos na matriz inicial.

Espectrogramas

Delineámos a geração dos espectrogramas (Figura 27) a partir da aplicação de janelas de Hamming (devido à obtenção de melhores resultados em pesquisas anteriores da equipa) com um tamanho de 64ms, uma sobreposição de 75% e 512 nfft (número de pontos na FFT). De seguida, transformámos a informação que foi retornada para decibéis (db). No fim, normalizámos os dados relativos à imagem espetral em [0 1].

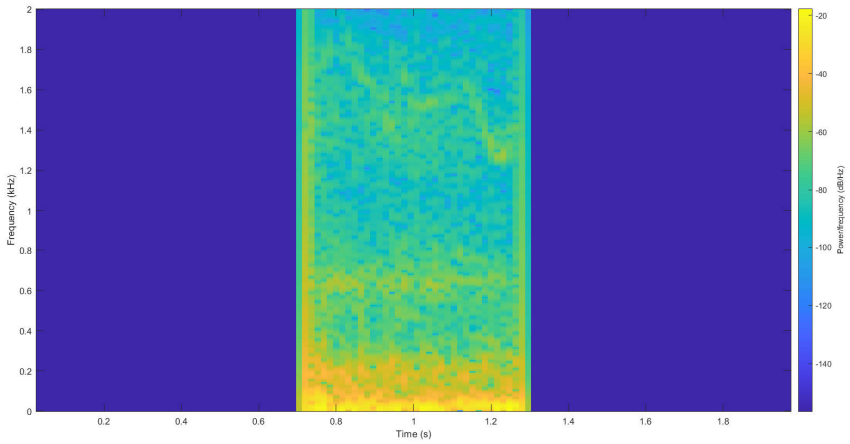


Figura 27 – Espectrograma de sibilância

Espectrogramas de Mel

Os espectrogramas de Mel derivam dos espectrogramas, uma vez que são criados a partir da conversão das frequências dos espectrogramas para uma escala de mel. Os atributos MFCC provêm destes espectros a partir do cálculo de DCT. A Figura 28 representa essa mesma transformação em relação ao espectrograma do evento da Figura 27. Desta forma, gerámos também espectrogramas de Mel com os mesmos parâmetros definidos para a extração dos espectrogramas.

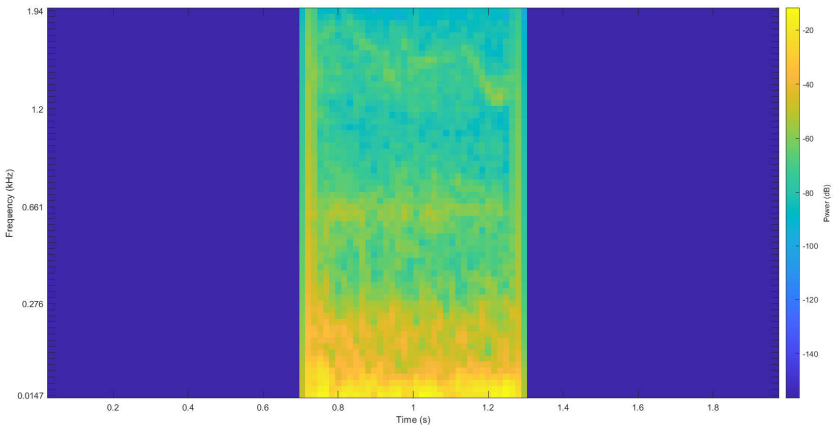


Figura 28 - Espectrograma de Mel de sibilância

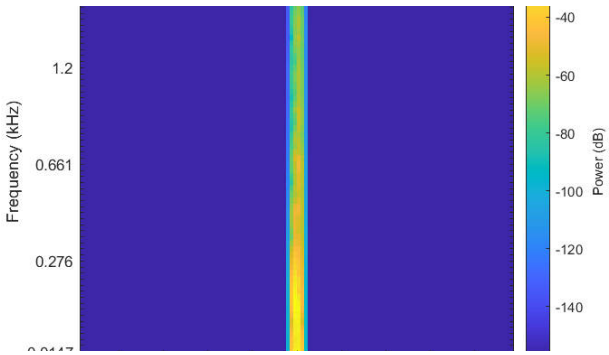


Figura 29 - Espectrograma de Mel de fervor

Onduletas

As onduletas provêm da *Discrite Wavelet Transform* (DWT), que nos fornece uma representação tempo-frequência do sinal. Esta transformada foi desenvolvida como alternativa à STFT utilizada pelos espectrogramas. Os atributos foram obtidos tendo em conta a duração do som e a frequência de amostras. As onduletas depois de geradas tiveram que sofrer uma redução de tamanho por motivos computacionais, passando de imagens [101x8001] a um tamanho de [101x800]. As Figuras 30 e 31 são dois exemplos de onduletas.

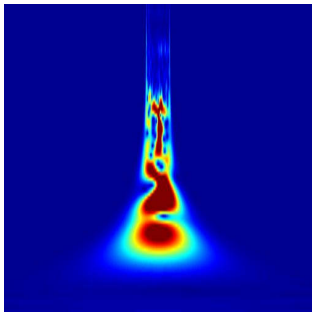


Figura 31 – DWT: Evento de fervor

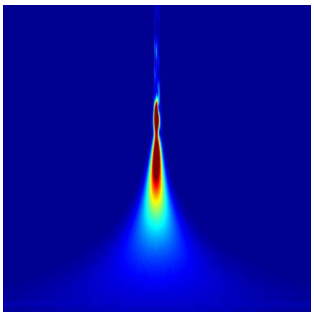


Figura 30 – DWT: Evento de sibilância

Cromagramas

Os cromagramas são gráficos que têm por base os diversos tons que um áudio pode ter. Estes representam as mudanças de tom ao longo do tempo, refletindo de forma simples a energia do som presente em todas as doze notas musicais. Os dados foram extraídos a partir da análise da decomposição da janela sonora em *frames*.

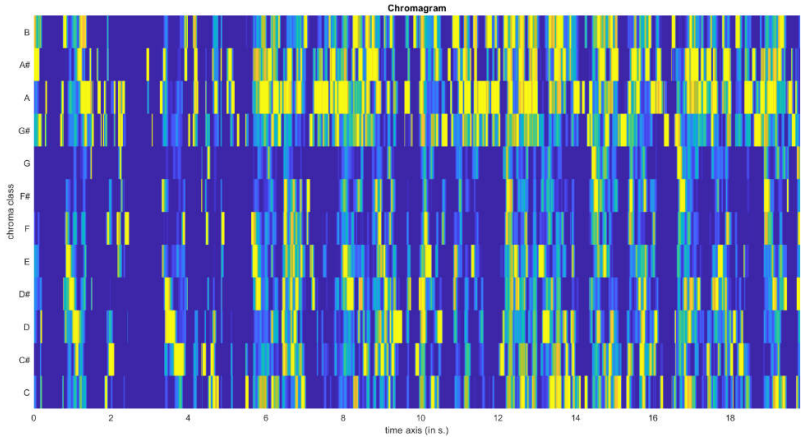


Figura 32 – Cromograma de um evento

4.3.2. Modelos

Os modelos são a peça principal para a classificar os eventos dos conjuntos de dados criados. Desta forma, desenvolvemos diversas arquiteturas com parâmetros distintos de treino de maneira a poder compará-los não só entre modelos de aprendizagem profunda e não profunda, mas também o impacto dos seus constituintes.

Os classificadores CNN, BiLSTM, LSTM e GRU foram desenvolvidos tendo em atenção o tipo de dados que as redes iriam receber na sua camada de entrada, fomos elaborando diversas arquiteturas e ajustando os parâmetros que compõem as camadas internas das redes. Como resultado, a camada de saída retorna sempre a classe com a qual classificou o evento de entrada.

Começando pelos modelos baseados em Redes Neuronais Recorrentes (RNN), como se pode observar na Figura 33, a rede BiLSTM é composta por uma camada de entrada que irá variar consoante o tipo de dados de entrada, uma camada BiLSTM com x neurónios, uma camada *fullyconnected* com tamanho de 3 (o número total de classes presentes no problema), uma camada *softmax* e por último uma camada de classificação que calcula a função a otimizar para a classificação: entropia cruzada. O número de neurónios foi uma variável testada nas experiências realizadas.

O modelo GRU, para além da camada de entrada, tem duas camadas GRU uma a seguir à outra com 100 células cada uma, uma camada *fully connected*, uma camada *softmax* e uma camada de classificação como a arquitetura anterior.

O classificador LSTM é bastante semelhante ao BiLSTM, apenas alterando a camada BiLSTM para LSTM. Alterámos os parâmetros de treino realizando algumas experiências, assim como o número de células das camadas internas das redes.

Para a fase de treino das redes optámos por usar a mesma composição de parâmetros para todos os modelos para realizar uma comparação o mais justa possível. Aplicámos a função 'Adam' para treino, um *minibatchsize* de 128, uma taxa de aprendizagem inicial de 0,01, sendo esta atualizada de forma '*piecewise*', ou seja, a cada época de treino. A fase de treino será mais especificada na subsecção seguinte.

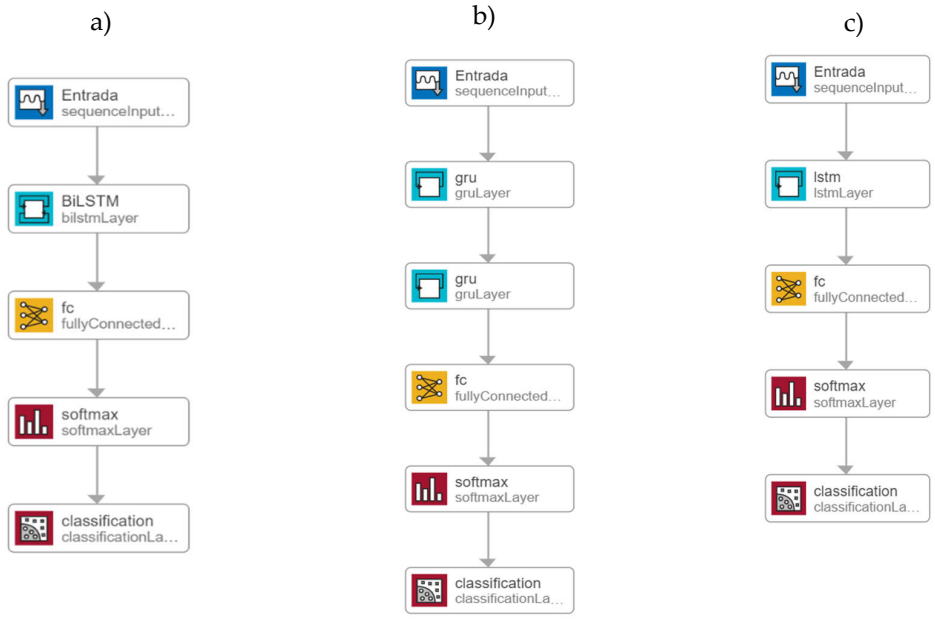


Figura 33 - Arquiteturas dos classificadores baseados em RNN: a) BiLSTM, b) GRU e c) LSTM

Para além dos modelos baseados em RNNs, estruturámos diferentes modelos CNN. Como se pode verificar na Figura 34, um dos modelos elaborados possui dupla entrada de conjunto

de dados de treino. Estes passam por uma camada convolucional, função ReLU, uma camada maxpool2d e uma camada de *dropout*. Estas camadas voltam a ser utilizadas a seguir ao primeiro *dropout*, sendo este processo realizado de forma semelhante para o outro tipo de entrada. De seguida, o resultado dos dois é concatenado nesta mesma fase, mas não sem antes passar pela camada de *flatten*. Depois de concatenados, os dados vão para uma camada *dense* com função de ativação ReLU e um *dropout*. As últimas camadas são *dense* e *softmax*.

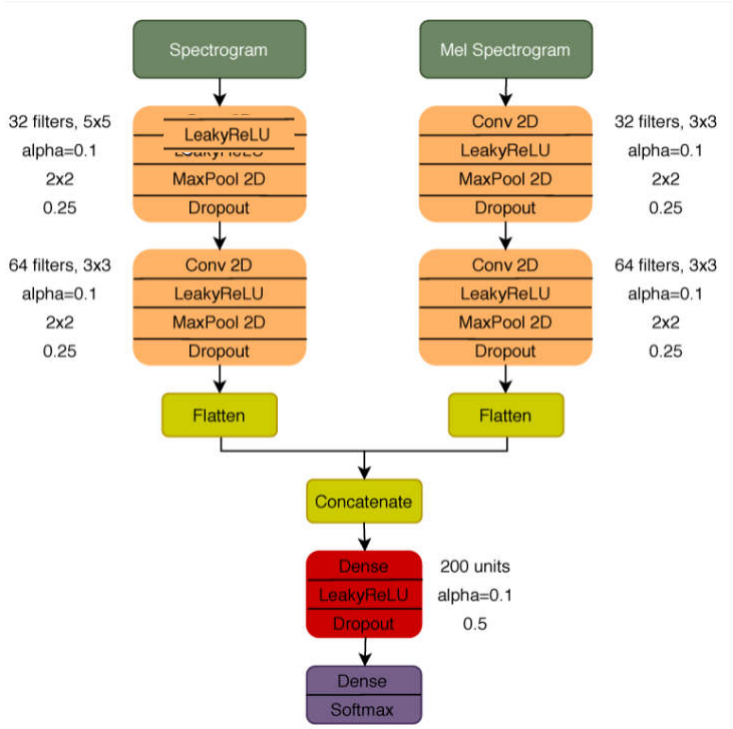


Figura 34 - Arquitetura CNN com duas entradas

Também desenvolvemos um modelo CNN com apenas uma entrada de dados (Figura 35). A camada de entrada da rede irá adaptar-se consoante o tipo de atributos que se irão usar para a classificação. Todas as camadas de convolução possuem uma função de ativação Relu seguidas de uma camada *Max Pooling* e uma camada de *dropout* de 25%. O resultado da última camada de convolução passa por uma camada *Flatten* e de seguida por uma camada dense com uma função de ativação Relu e aplicado um *dropout* de 10%. Por último, os dados passam por uma camada *Dense* com a função objetivo *softmax*.

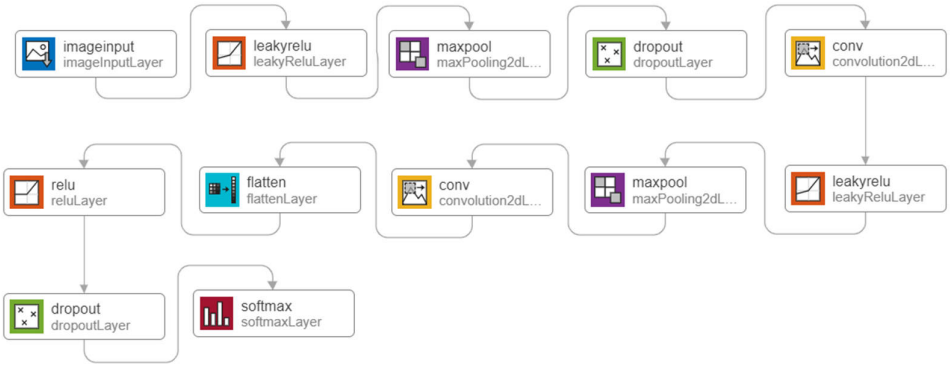


Figura 35 - Arquitetura CNN com uma entrada

4.3.3. Avaliação

Os conjuntos de dados construídos para este problema foram divididos sempre com as mesmas amostras de treino e teste para que a avaliação e comparação dos classificadores seja a mais justa possível. Os eventos pertencentes ao conjunto de dados de treino e teste encontravam-se previamente selecionados, assim como aconteceu nos conjuntos de dados para os algoritmos clássicos. Quisemos testar o comportamento dos algoritmos CNN, BiLSTM, LSTM e GRU relativamente ao tipo de dados que recebem nas suas camadas de entrada da rede. As métricas utilizadas para avaliar este comportamento foram: exatidão, precisão (P), sensibilidade (S) e *F1-score*.

Começámos por construir um classificador CNN com uma camada de entrada (Figura 33 da Secção anterior). Este foi treinado com 50 épocas e 5 vezes, sendo testado apenas o modelo com melhores valores de exatidão. A função de treino aplicada foi a Adam, com uma percentagem de aprendizagem de 0,01%. A partir da Tabela VIII podemos observar os resultados obtidos relativamente ao melhor modelo para cada um dos tipos de entrada da rede e dependendo se o conjunto de dados de treino foi ou não balanceado. Esta análise deve-se ao facto de realizar uma verificação da eficácia do balanceamento de dados de treino.

Tipo de entrada	**Balanceamento**	**S**	**P**	**F1-Score**	**Exatidão**
Espectro Mel	Não	81%	80%	80%	81%
Espectro	Não	79%	79%	79%	79%
Onduleta	Não	70%	69%	69%	71%
Sinal	Não	77%	77%	76%	80%
Cromagrama	Não	38%	35%	35%	50%
Espectro Mel	tempo + volume + altura	76%	77%	76%	80%
Espectro	tempo + volume + altura	76%	77%	76%	80%
Onduleta	tempo + volume + altura	69%	68%	69%	70%

Tabela VIII – Resultados obtidos a partir da arquitetura CNN de uma entrada

Neste caso, o único tipo de entrada capaz de se destacar pela negativa são os cromagramas alcançando apenas 50% de exatidão e um F1-Score de 35%. Através das matrizes de confusão presentes nas Figuras 36 e 37, é possível retirarmos informação mais detalhada sobre o desempenho do classificador relativamente a cada uma das classes. A utilização do sinal do evento e de espectrogramas de Mel apresentam os melhores resultados do classificador, alcançando 81% e 80% de exatidão respetivamente. A classe onde existe uma maior dificuldade a ser identificada é a 'sibilância' tendo uma percentagem de sensibilidade mais baixa que as outras (71% para sibilâncias e 92% para fervores). O número reduzido de amostras de 'sibilâncias' no conjunto de dados de teste pode estar a prejudicar esta classe. As restantes matrizes de confusão não foram apresentadas por motivos de organização do documento.

Figura 36 – Matriz de confusão do classificador CNN com entrada de sinal

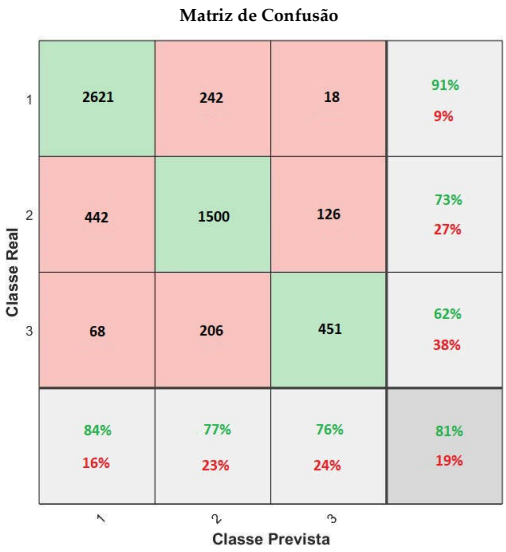


Figura 37 – Matriz de confusão do classificador CNN com entrada de espectrograma de mel

Realizámos outra experiência que consistiu em combinar os diferentes tipos de atributos utilizadas na primeira experiência. Adaptámos a rede dualCNN presente no artigo [10], publicado por outros membros da equipa, para receber todas as combinações de dois tipos de atributos diferentes. Esta experiência consistiu na possibilidade de dois tipos de atributos acrescentarem valor à rede em vez de esta obter um tipo de informação. Mais uma vez o classificador, foi executado 5 vezes com 50 épocas. Os resultados obtidos encontram-se na tabela IX.

Tipo de entrada	Balanceamento	S	P	F1Score	Exatidão
mel + spectro	Não balanceado	76%	81%	81%	82%
sinal + mel	Não balanceado	78%	81%	81%	82%
mel + spectro	tempo + volume + altura	81%	78%	79%	82%
sinal + spectro	Não balanceado	77%	79%	78%	81%

spectro + onduletas	Não balanceado	77%	78%	77%	81%
mel + onduletas	Não balanceado	78%	80%	78%	82%

Tabela IX – Resultados obtidos a partir da arquitetura CNN de duas entradas

Os resultados de exatidão são extremamente parecidos aos obtidos com a CNN com apenas um formato de dados de entrada. Novamente a classe com mais dificuldade a ser classificada é a sibilâncias com 62% de sensibilidade (matriz de confusão na Figura 38), contudo também temos que ter em atenção que a quantidade de amostras de teste de eventos de sibilâncias é bastante reduzida, o que pode afetar a classificação.

Figura 38 - Matriz de confusão do classificador CNN com entrada de sinal e espectrogramas de mel

O classificador BiLSTM foi treinado com diferentes atributos com informação extraída do sinal dos arquivos de som dos eventos. Foram testados diferentes parâmetros na sua arquitetura, como pode ser verificado na Tabela X. Em relação ao sinal obtivemos um máximo de 60% de exatidão aquando aplicado um conjunto de treino balanceado e uma camada escondida com um tamanho de 100 neurónios. Em relação aos outros atributos alcançámos 78,32% de exatidão quando introduzidos 45 atributos e 100 de neurónios.

Tipo de entrada	Balanceamento	S	P	F1Score	Exatidão
30 atributos	Não balanceado	87,50%	76,33%	81,54%	74,48%
45 atributos 10 neurónios	Não balanceado	89,58%	78,63%	83,75%	76,77%
45 atributos 100 neurónios	Não balanceado	89,52%	80,58%	84,82%	78,32%
45 atributos 400 neurónios	Não balanceado	91,73%	75,80%	83%	77,50%
45 atributos 100 neurónios	tempo + volume	76,35%	88,48%	81,97%	75,83%

45 atributos 100 neurónios	tempo + altura	74,52%	87,26%	80,39%	56,26%
45 atributos 100 neurónios	tempo + volume + altura	57,84%	93,25%	71,40%	51,23%
Sinal 400 neurónios	Não balanceado	95,41%	34,50%	50,68%	57,75%
sinal 400 neurónios	tempo + volume + altura	69,28%	71,69%	70,46%	59,90%
sinal 100 neurónios	balanceado	64,90%	73,61%	68,98%	60,03%
sinal 100 neurónios	Não balanceado	99,93%	40%	57,13%	50,94%

Tabela X – Resultados obtidos a partir da arquitetura CNN de duas entradas

A Tabela XI contém os resultados do classificador GRU. O sinal desta vez apresenta valores semelhantes à tabela anterior. Por outro lado, a seleção de 45 atributos permite alcançar uma exatidão de 79,78% com um tamanho de 400 neurónios na camada interior. Os valores da sensibilidade de todas as experiências são bastante satisfatórios alcançando quase sempre um resultado superior a 70%.

Tipo de entrada	**Balanceamento**	**S**	**P**	**F1Score**	**Exatidão**
45 atributos 10 camadas	Não balanceado	99,80%	1,30%	2,57%	51,24%
45 atributos 100 camadas	Não balanceado	99,70%	1,30%	2,57%	51,30%
45 atributos 400 camadas	Não balanceado	88,14%	82,80%	85,39%	79,78%
sinal 400	balanceado	65,29%	75,73%	70,13%	61,16%
sinal 400 400 camadas	Não balanceado	94,80%	36,54%	52,75%	58,36%
sinal 400 400 camadas	Balanceado	73,36%	71,08%	72,20%	60,93%

Tabela XI – Resultados obtidos a partir da arquitetura GRU

O classificador LSTM foi o que obteve piores resultados de exatidão alcançando no máximo 60%.

Tipo de entrada	**Balanceamento**	**S**	**P**	**F1Score**	**Exatidão**
sinal 400	não balanceada	96,42%	31,57%	47,57%	56,99%
sinal 100	não balanceada	85,64%	49,34%	62,61%	59,72%

sinal 400	tempo + volume + altura	69,66%	71,18%	70,41%	59,86%
sinal 100	tempo + volume + altura	68,96%	72,48%	70,67%	60,22%

Tabela XII – Resultados obtidos a partir da arquitetura LSTM

Para concluir esta avaliação, classificar eventos adventícios é um grande desafio uma vez que estamos a tratar de sons extremamente curtos. O balanceamento das classes na fase de treino também foi um desafio, pois tínhamos de ter em atenção as propriedades dos eventos com que estamos a trabalhar. Apesar destes fatores, conseguimos obter no melhor caso, entre todas as experiências elaboradas, uma exatidão de 82% com uma sensibilidade de 81% a partir de espectrogramas de mel. Os conjuntos de dados com que tivemos mais dificuldade a obter resultados foram os constituídos por cromagramas. Os modelos clássicos em comparação com os modelos de aprendizagem profunda alcançam resultados inferiores, visto que o melhor modelo clássico desenvolvido tem uma exatidão 8% inferior ao melhor modelo de aprendizagem profunda.

5 Segmentação e classificação de sons adventícios

A segmentação e classificação de sons adventícios exige tanto a identificação como a localização deste tipo de sons nos ficheiros de áudio. Deste modo, criámos diversos conjuntos de dados aplicando técnicas de pré-processamento diferentes das utilizadas no problema exposto no capítulo 4. Devido aos bons resultados apresentados pelos modelos de aprendizagem profunda (expostos na Secção 4.3.3) construímos redes CNN e LSTM. Foram elaboradas diversas experiências de forma a verificar qual a melhor abordagem para resolver o problema associado à segmentação e classificação de sons adventícios. Também definimos várias métricas com vista a avaliar os classificadores. Todos os passos envolvidos nestes processos serão explicados mais detalhadamente nos parágrafos seguintes.

5.1 Pré-Processamento

A criação dos conjuntos de dados foi realizada com base nos eventos utilizados no problema anterior. Como a existência de trabalhos sobre segmentação de sons respiratórios é escassa, apoiámo-nos na técnica de segmentação do artigo [34]. Os arquivos de som da base de dados já tinham sofrido uma reamostragem para 4000Hz (pré-processamento da Secção 4.1), não foi necessário executar novamente este processo.

Considerando os resultados do problema anterior, decidimos transformar cada evento em espectrogramas de Mel, sendo estes à posteriori segmentados. Este processo encontra-se representado na Figura 39. Para a geração das imagens dos espectros, recorremos à aplicação de uma janela de Hamming com um tamanho de (32ms ou 64ms)x taxa de amostragem do sinal do evento (dependendo da experiência); uma sobreposição de 50% e aplicámos 128 filtros passa-banda. Após o processamento do espectro, a partir da localização dos centros das janelas geradas e do cálculo do *hopsize*, segmentámos o espectro em várias *frames*, atribuindo a cada uma delas uma classe: '1' fervor, '2' outro e '3' sibilância. Cada sequência de *frames* compõe um evento.

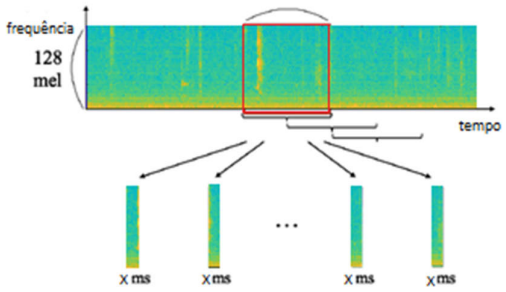


Figura 39 – Processo de segmentação

A elaboração de várias experiências e análises levou a que o processo precedente, tivesse sido executado para a composição de vários conjuntos de dados, sendo necessário alterar alguns parâmetros na geração dos espectros como por exemplo o tamanho das janelas calculadas (em ms).

5.1.1. Modelos

Este problema de classificação foi abordado de 3 formas: abordagem com três classes, abordagem binária e abordagem por paciente. Começámos por construir modelos que recebem conjuntos de dados com três classes como no problema anterior. Modelámos assim o classificador CNN com as arquiteturas apresentadas nas Figura 40. A camada de entrada irá receber uma imagem referente a cada *frame*, enquanto na do LSTM irá receber um conjunto de imagens, uma sequência de *frames* que compõem cada um dos arquivos sonoros.

A camada convolucional possui vinte filtros de 60x1 e as camadas *fully connected* têm um tamanho de 200, 50 e 2 respetivamente.

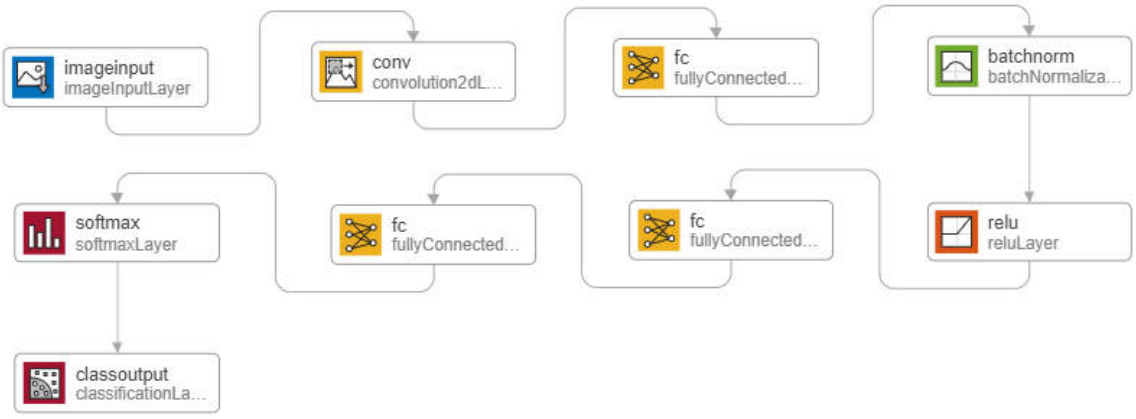


Figura 40 – Arquiteturas dos modelos CNN

Também elaborámos experiências com apenas duas classes que irão ser explicadas no subcapítulo da avaliação dos modelos. Posto isto, a única alteração que os modelos sofreram foi o tamanho da última camada *fully conneted* relativamente aos modelos CNN, que passou de três para dois, dependendo das classes existentes no problema.

Para a fase de treino das redes decidimos utilizar os mesmos hiperparâmetros da experiência anterior, usando 50 épocas.

Devido ao forte desequilíbrio dos dados, a favor da classe 'outro', a identificação dos eventos adventícios na sequência de *frames* que compõem os ficheiros torna-se um grande desafio. Por este motivo, ao resultado do classificador foram aplicados **filtros/máscaras** para a decomposição de *frames*. Um dos filtros resulta na observação da classe de um conjunto de *frames* e realizar uma média das observações para trás e para a frente da observação atual. Definimos, depois de várias tentativas, um *timestep* igual a um, uma vez que este não poderia ser maior devido ao tamanho dos eventos adventícios serem bastante curtos e, por conseguinte, o número de *frames* que os constituem ser bastante reduzido. Também criámos um filtro semelhante ao anterior, mas aplica a mediana ao conjunto de observações. Ainda

outro modo de aplicar o filtro foi desenvolvido com o objetivo de facilitar a identificação das *frames* adventícias. Este consistiu na verificação do elemento anterior da sequência. Após aplicar a média dos segmentos como é executado no filtro anterior, e a *frame* é considerada da classe 'sibilância' ou 'fervor' e a *frame* procedente a esta é classificada como da classe 'outro', o filtro transforma a *frame* procedente igual à anterior permitindo que a deteção das *frames* que compõem os eventos adventícios sejam mais fáceis de detetar ao aumentar em uma unidade o seu tamanho. Este último filtro só tem lógica ser aplicado uma vez que temos como objetivo detetar eventos adventícios que são bastante raros em relação à outra classe.

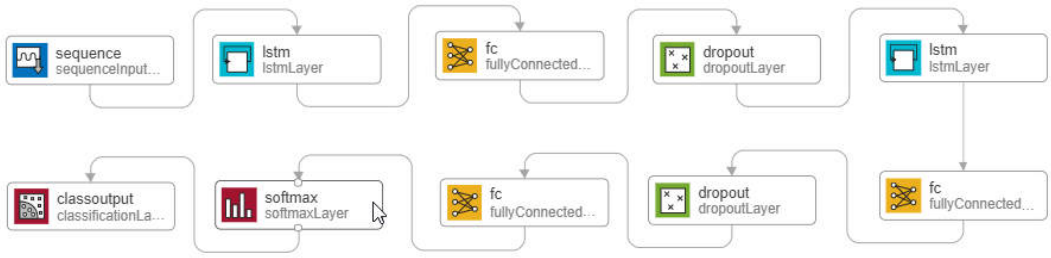


Figura 41 – Arquiteturas dos modelos CNN

A Figura 41 representa uma rede LSTM, constituída por duas camadas LSTM com 100 e 50 unidades escondidas, com um modo de saída sequencial, ou seja, são retornados valores para cada *frame* que compõe a sequência de entrada. As camadas *fully connected* são constituídas por 150, 50 e 3 células respetivamente, seguidas por uma camada de *dropout* com um valor de 0.3 de probabilidade.

5.1.2. Avaliação dos modelos

Foram calculadas duas formas de avaliar o desempenho dos modelos tanto a nível geral do classificador como em relação a cada uma das classes:

i) Por *frames*; o desempenho '*frame* a *frame*' consiste no número total de *frames* corretamente classificadas em relação ao número de *frames* que compõem as gravações de áudio.

ii) Por eventos, os quais constituem sequências de *frames* da mesma classe.
O desempenho por evento é calculado tendo em conta o número total de eventos existentes no ficheiro de áudio e se o classificador identifica a existência de *frames* dentro do período de duração dos eventos existentes. Neste caso foram definidos dois limiares: a) um limiar mínimo de acerto de 20% das *frames* dentro da duração de um determinado evento para que este seja identificado como corretamente classificado; b) um limiar igual a uma *frame* bem identificada dentro do período os eventos.

Os resultados obtidos ao nível dos eventos são os mais importantes, uma vez que queremos verificar se as redes conseguem identificar quando ocorre uma *frame* que carateriza a existência de um evento adventício. Em relação à classe, foi calculada a sensibilidade e a precisão considerando as duas formas de avaliação.

Inicialmente foi aplicada a divisão pré-definida ao conjunto de arquivos sonoros existentes na base de dados que se baseia numa divisão de 60% para treino e 40% para teste.

O pré-processamento permitiu a criação do conjunto de dados de treino e teste compostos por *frames* de espectrogramas de Mel.

5.1.2.1. Abordagem com três classes

O grande problema de identificar sons adventícios na respiração é o facto da sua existência ao longo do ciclo respiratório ter uma duração extremamente curta (as sibilâncias com uma duração de 250ms ou superior e os fervores com uma duração inferior a 25ms). Ao segmentarmos o ficheiro sonoro em pequenos fragmentos de som, aqueles que irão possuir informação de sons adventícios serão muito menores em termos de quantidade do que os restantes segmentos do sinal. Esses eventos podem ser denominados de eventos raros. Desta forma, é necessário utilizar técnicas de **balanceamento das amostras**, uma vez que as sequências de *frames* de cada um dos ficheiros serão compostas por classes muito desequilibradas. Os ficheiros de pacientes saudáveis não foram incluídos nos conjuntos de dados utilizados. Deste modo, dos 920 arquivos, 589 pertencem a pacientes não saudáveis. Aqui, o conjunto de treino de *frames* é composto por: 27585 'sibilâncias', 13829 'fervores' e 105303 'outro', denotando-se que a classe 'outros' é cerca de 282% superior à classe 'sibilâncias' e 661% superior à classe 'fervores'. Para diminuir o *overfitting* dos modelos foi aplicado subamostragem para obter 13829 amostras da classe 'outros' e de ´sibilâncias´.

Esta abordagem baseia-se na classificação de *frames* (sibilâncias, fervores e outros) utilizando um classificador CNN com a arquitetura descrita anteriormente. O conjunto de dados composto por *frames* foi extraído dos espectrogramas de Mel, com uma janela de 64ms. Os resultados apresentados na Tabela XIII são relativamente ao melhor modelo obtido entre as 10 execuções, relativamente à análise do conjunto de dados de treino balanceado.

Experiência	Exatidão	Precisão	Sensibilidade	Especificidade	F1-Score
Balanceada	78,5%	5,3%	2%	98,3%	3%

Tabela XIII – Resultados obtidos a partir da arquitetura CNN com o conjunto de dados de *frames* balanceado

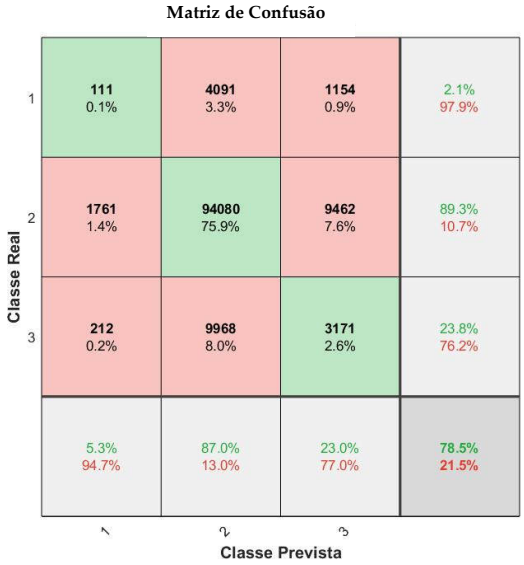


Figura 42 – Matriz de confusão do classificador CNN com 3 classes

Como se pode verificar através dos resultados presentes na Tabela XIII e na matriz de confusão presente na Figura 42, foi bastante difícil para o classificador distinguir os diferentes tipos de *frames*. Os elementos da classe 'sibilância' (3) e 'fervor' (1) foram bastante confundidos com os da classe 'outro'(2), obtendo uma exatidão de 78,5%. Este fator leva a que as restantes métricas possuam um valor bastante baixo, uma vez que têm como base os elementos corretamente identificados (TP).

Na Tabela XIV podemos observar a sensibilidade (S) e a precisão (P) do classificador ao nível das classes.

Classe	**S**	**P**
Outro	89,3%	87%
Sibilância	23,8%	23%
Fervores	2,1%	5,3%

Tabela XIV – Sensibilidade e precisão de cada classe obtidas pelo classificador CNN

As *frames* de fervores foram bastante difíceis de classificar possuindo uma sensibilidade inferior a 2,5%. Já a classe relativa às sibilâncias alcançou uma sensibilidade de 23,8%, uma vez que grande parte da sua amostra foi corretamente identificada. Por sua vez, a precisão não alcança grandes resultados, pois existem muitas amostras das outras duas classes identificadas de forma errada como sibilância. Estes valores também podem estar a ser influenciados devido a uma grande diferença de quantidade de amostras relativamente a cada classe no conjunto de teste, uma vez que este não pode ser balanceado porque não é possível retirar da sequência temporal dos ficheiros determinadas *frames*.

A construção de uma arquitetura LSTM tinha como objetivo a identificação dos eventos adventícios através de uma aprendizagem com séries temporais. Deste modo, esta

foi treinada e testada através de conjuntos de dados constituídos por ficheiros de áudio decompostos em sequências de *frames* (processo descrito em 5.1). Como todos os ficheiros são predominantemente compostos por *frames* 'outros', o classificador não conseguiu classificar nenhuma *frame* adventícia nas sequências que componham o conjunto de dados de teste.

5.1.2.2. Abordagem binária

Devido aos resultados pouco positivos decidimos criar um modelo com apenas duas classes, formando dois conjuntos de dados e dois classificadores binários: um conjunto com *frames* de ficheiros de áudio de apenas eventos 'outros' e 'fervores', e outro com a classe 'sibilâncias' e 'outros'. Esta experiência desenvolveu-se devido à necessidade de facilitar a aprendizagem dos classificadores, uma vez que existiam dificuldades na aprendizagem com três classes. Para a criação destes conjuntos de dados foram utilizados os ficheiros dos pacientes que continham apenas fervores e noutro com apenas sibilâncias. A divisão entre conjunto de dados de treino e teste utilizada foi a pré-definida inicialmente, 60/40. A partir dos arquivos com sibilâncias foram extraídas as *frames* considerando uma janela de 64ms, enquanto que para os arquivos que continham fervores foi aplicada uma janela de 32ms (uma vez que estes são bem mais curtos que os outros sons adventícios). Mais uma vez, para diminuir o *overfitting*, foi aplicado uma subamostragem às amostras 'outro' do conjunto de dados de treino limitando-as a um máximo de 14300 amostras de cada classe ('sibilâncias' e 'outro') e 10000 amostras de cada classe ('fervores' e 'outro'). O classificador utilizado foi o mesmo que na abordagem anterior, uma CNN. Os resultados relativos à análise do conjunto de dados de treino balanceado são apresentados nas Tabelas XVI, XVII, e XVIII referem-se aos melhores modelos obtidos entre 10 execuções.

Através da Tabela XV e das matrizes de confusão presentes nas Figuras 43 e 44 conseguimos observar que o valor de exatidão do classificador com um conjunto de dados binário ('outro - 2' e 'sibilância - 1') rondou os 71%, em relação aos de 78,5% alcançados na abordagem anterior. A exatidão geral apesar de diminuir, conseguimos obter uma sensibilidade bastante superior à experiência anterior. Este fator demonstra que a distinção entre as classes melhora quando dividimos o problema em binário. O balanceamento do conjunto de dados não elevou o valor da exatidão, mas fez com que o F1-score atinge-se o triplo da não balanceada.

Experiência	Exatidão	Precisão	Sensibilidade	Especificidade	F1-Score
Não balanceada	87%	92,6%	5,4%	99,9%	10,3%
Balanceada	71,3%	23,5%	48,4%	74,9%	31,6%

Tabela XV – Resultados obtidos a partir da arquitetura CNN com o conjunto de dados de *frames* binário 'Sibilâncias' e 'Outro'

Figura 44 - Matriz de confusão do classificador CNN Balanceado

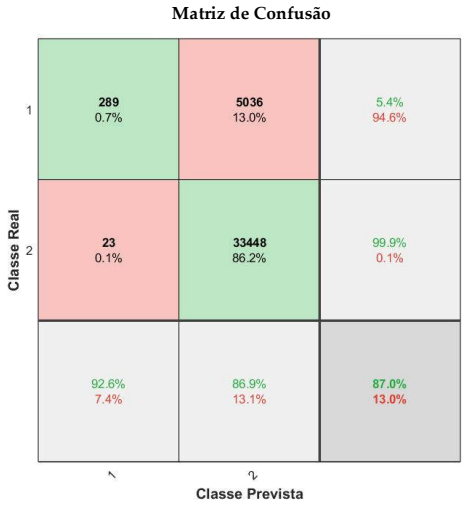


Figura 43 - Matriz de confusão do classificador CNN não balanceada

A Tabela XVI apresenta os resultados das métricas: sensibilidade (S) e precisão (P) relativamente às duas classes do problema. Novamente o classificador tem dificuldade um identificar as *frames* dos eventos adventícios sibilâncias, mas obteve uma sensibilidade de 48,4% em relação à alcançada do problema anterior (23,8%). Em relação à localização de eventos adventícios conseguimos identificar corretamente, com um limiar de 20% de *frames*, 49,1% de todos os eventos com sibilâncias existentes no conjunto de dados de teste e 67,3% quando aplicado um limiar de uma *frame*.

Eventos					*Frame* a *Frame*	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Outro	61,5%	58,5%	31%	52,6%	74,9%	90,1%
Sibilância	**49,1%**	52,2%	**67,3%**	45,6%	**48,4%**	23,5%

Tabela XVI – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento

As Tabelas XVII e XVIII contêm os resultados obtidos ao nível das classes quando é aplicado um filtro ao conjunto de *frames* resultantes da classificação na fase de teste com vista a corrigir algumas *frames* que tenham sido mal classificadas. Este filtro tem em consideração as *frames* vizinhas da sequência. Tanto a aplicação de um filtro por média e mediana apresentam resultados semelhantes com um limiar de 20% e inferiores com um limiar de 1 *frame*, em relação a sem filtro.

Filtro Média						
Eventos					***Frame* a *Frame***	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Outro	61,2%	58,4%	38,5%	52,6%	75,5%	90,1%
Sibilância	49,1%	52,1%	59,6%	45,3%	48%	23,8%

Tabela XVII – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de média

Filtro Mediana						
Eventos					***Frame* a *Frame***	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Outro	61,5%	58,5%	38,5%	52,6%	75,5%	90,1%
Sibilância	49,1%	52,2%	59,6%	45,3%	48%	23,8%

Tabela XVIII – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de mediana

Na Tabela XIX são apresentados os valores obtidos da classe sibilâncias a partir do filtro que tem como objetivo aumentar o tamanho dos eventos adventícios encontrados. Neste caso em relação à deteção de eventos ao aplicar um limiar de 20% conseguimos obter uma ligeira melhoria na precisão da classe e um aumento de 49,1% para 57,8% de sensibilidade. Do ponto de vista das *frames* também existiu uma melhoria de cerca de 3% em termos da sensibilidade obtida daquela classe e a precisão manteve-se.

Filtro Média + Transformação						
Eventos					***Frame* a *Frame***	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Sibilância	57,8%	53,5%	59,6%	45,3%	51,2%	23,2%

Tabela XIX – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de média + transformação

De acordo com os resultados apresentados na Tabela XIX, a aplicação de um filtro que permita aumentar o número de *frames* que compõem os eventos adventícios ajuda na localização dos mesmos.

As tabelas seguintes são referentes aos resultados alcançados pelo classificador binário treinado com um conjunto de *frames* categorizadas como 'outro' e 'fervores'. A partir da Tabela XX e das matrizes de confusão podemos observar que apesar de a exatidão ser ligeiramente pior do que o classificador binário de sibilâncias, ou seja, 57%, a precisão do classificador em comparação piorou bastante. Este resultado significa que o classificador tem dificuldade em distinguir as classes. O resultado de 50% de sensibilidade aquando usado um conjunto de dados balanceado é bastante positivo, simbolizando que existe uma grande percentagem de amostras de ambas as classes corretamente identificadas, mas também existe uma grande amostra que não é corretamente classificada.

Experiência	Exatidão	Precisão	Sensibilidade	Especificidade	F1-Score
Não balanceada	93,9%	59,8%	3,1%	99,8%	6%
Balanceada	57%	7%	50%	57,4%	12,5%

Tabela XX – Resultados obtidos a partir da arquitetura CNN com o conjunto de dados de *frames* binário 'Fervor' e 'Outro'

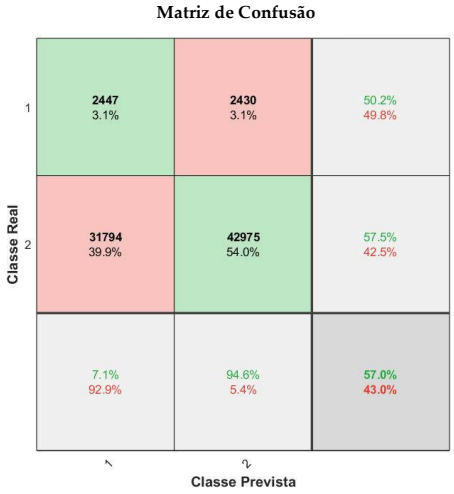


Figura 46 – Matriz de confusão do classificador CNN Balanceado com 'Outro' e 'Fervor'

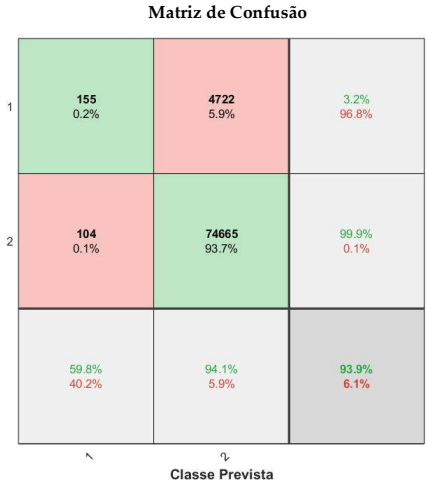


Figura 45 – Matriz de confusão do classificador CNN não Balanceado com 'Outro' e 'Fervor'

Os **eventos fervores** conseguem ser corretamente identificados na sua maioria, alcançando uma sensibilidade de 76% (considerando um limiar de 20%). Por sua vez, a classificação das ***frames* de fervores (1)** apresenta uma sensibilidade de 50,2% o que é bastante positivo em relação à alcançada de 2,1% na abordagem anterior. Isto pode confirmar que a aplicação de uma janela mais pequena (32ms) ajuda a detetar os eventos 'fervores', uma vez que estes são mais curtos que as outras duas classes e com uma janela de 64ms tornam-se impossíveis de detetar.

Eventos					*Frame* a *Frame*	
	Limiar 20		**Limiar 1**			
Classe	S	P	S	P	S	P
Outro	27%	54%	24%	51,5%	57,5%	94,6%
Fervores	76%	53,9%	76,5%	49,2%	50,1%	7%

Tabela XXI – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento

A aplicabilidade dos filtros não mostrou grande efeito uma vez que piorou todos os resultados obtidos. Neste caso a aplicação do filtro que aumenta da sequência de *frames* de fervores não ajuda a obter melhores resultados uma vez que o classificador está a classificar muitas *frames* da classe 'outro' como da classe 'fervores'.

Filtro Média						
Eventos					*Frame* a *Frame*	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Outro	32,8%	47,4%	32%	47%	59%	94,4%
Fervores	62,1%	51,3%	62,4%	46,8%	47,8%	7%

Tabela XXII – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de média

Filtro Mediana						
Eventos					*Frame* a *Frame*	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Outro	32,8%	47,37%	31,6%	46,6%	59,1%	94,6%
Fervores	62,1%	51,4%	62,4%	46,8%	47,8%	7%

Tabela XXIII – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de mediana

Filtro Média + Transformação						
Eventos					*Frame* a *Frame*	
	Limiar 20%		**Limiar 1 *frame***			
Classe	S	P	S	P	S	P
Fervores	64,7%	45,4%	64,8%	45,9%	52,3%	7%

Tabela XXIV – Resultados da sensibilidade e precisão de cada classe, *frame* a *frame* e por evento após filtro de média + transformação

Através desta abordagem concluímos que um classificador binário ajuda na identificação dos eventos adventícios que são bastante raros em comparação com a classe 'outros'. Conseguimos identificar 67,3% dos eventos de sibilância existentes no conjunto de dados de teste com um limiar de uma *frame* e 76,5% dos eventos de fervores com o mesmo limiar. O balanceamento de *frames* ajudou bastante a classificação. Quanto às *frames* o classificador confunde bastante as *frames* 'fervores' com as 'outro'.

5.1.2.3. Abordagem classificação por paciente

Esta análise teve como objetivo verificar se existem distinções na classificação por paciente. Por este motivo criámos um classificador adequado a cada paciente individualmente. Um dos problemas que enfrentámos foi o número extremamente reduzido de ficheiros por paciente. Para contornar este problema foi necessário recorrer ao aumento de dados para gerar ficheiros semelhantes ao do paciente, ficando no mínimo com seis ficheiros de som por paciente sendo divididos em 80% para treino e 20% para teste. Este número parece muito reduzido para classificadores de aprendizagem profunda, mas como vamos recorrer à segmentação um ficheiro decompor-se-á em múltiplas *frames*.

Os resultados apresentados na Tabela XXV são a média (M) e a variância (V) das métricas de todos os pacientes.

Exatidão		**Precisão**		**Sensibilidade**		**Especificidade**		**F1-Score**	
M	**V**	M	**V**	M	**V**	M	**V**	M	**V**
73,49%	4%	33,38%	12,5%	36,46%	13,9%	73,53%	12,8%	33,25%	10%

Tabela XXV – Média das métricas alcançadas pelo modelo

Tendo em conta os resultados presentes na Tabela XXV verificamos que para quase as métricas calculadas (exceto exatidão) existe uma variância de cerca de 12,5%. Este fator indica que para alguns pacientes o classificador apesar de apresentar uma exatidão elevada não consegue distinguir as classes corretamente. Os valores da exatidão média do classificador está a ser influenciada devido à presença nos conjuntos de testes a existência de muitas *frames* 'outro, como já acontecia nas outras experiências. Os melhores resultados foram obtidos pelos pacientes com os ids: 122, 170 e 192 com uma exatidão superior a 95% e os piores foram com os ids 145 e 218: com uma exatidão de 39% e 26%, respetivamente. O paciente 145 possui poucos ficheiros o que não ajuda na fase de treino do classificador. O paciente 218 também não tem muitos ficheiros, mas o que pode levar ainda mais a dificuldade do classificador é ter recolha de dois estetoscópios diferentes. Os ficheiros de som destes pacientes possuíam eventos pertencentes às três classes.

Tendo em conta as classes a classificar, a classe 'sibilância' possui uma sensibilidade inferior à de 'fervores', assim como aconteceu ao longo de todas as experiências anteriores. O melhor valor de sensibilidade ao nível de fervores foi obtido pelo paciente 176 e ao nível das sibilâncias pertence ao paciente 221. Existiram casos de pacientes em que por vezes a sensibilidade das classes referentes a sons adventícios era perto de 0%, afetando a média global. Este fator deve-se ao facto de o balanceamento de dados ter sido realizado a partir dos ficheiros dos pacientes os conjuntos de dados não ficaram bem balanceados ao nível do tipo de eventos adventícios. Mesmo tendo isto em conta, a variância do modelo não é muito

elevada alcançando no máximo 5,2% para a classe de 'fervores', como se pode visualizar na Tabela XXVI.

Classe	**Precisão**		**Sensibilidade**	
	Média	Variância	Média	Variância
Sibilância	33,26%	5,2%	15,24%	2,3%
Outro	89,31%	0,01%	89,36%	0,01%
Fervores	22,09%	4%	21,7%	2,1%

Tabela XXVI – Média das métricas alcançadas pelo modelo por classe

Mais uma vez, a aplicação do filtro não permite alcançar melhores resultados que os anteriores. A percentagem de *frames* com sons adventícios é pequena, alcançando no caso das sibilâncias 11,06% e dos fervores 16,18%. O filtro alterou a percentagem de melhor valor de sensibilidade ao nível das classes, sendo agora atribuídas aos pacientes 193 e 201.

Classe	**Precisão**		**Sensibilidade**	
	Média	Variância	Média	Variância
Sibilância	32,10%	5,4%	11,06%	1,8%
Outro	88,04%	0%	86,53%	0%
Fervores	17,83%	2,8%	16,18%	2,5%

Tabela XXVII – Média das métricas alcançadas pelo modelo por classe depois do filtro média

Existem pacientes bastante diversificados: uns possuem muitos ficheiros de áudio de um só estetoscópio; outros possuem muito poucos ficheiros, mas de vários meios de recolha; ainda existem pacientes com apenas um ou dois evento adventício de uma determinada classe e os restantes eventos são de outra aumentando a dificuldade de treino para uma determinada classe o que prejudica o cálculo das métricas gerais.

6 Conclusões e Trabalho Futuro

As patologias respiratórias são uma grande preocupação na comunidade médica. Esta dissertação teve como objetivo principal aplicar algoritmos de aprendizagem profunda com vista a segmentar e classificar automaticamente eventos adventícios. Tendo em atenção este foco, dividimos este propósito em dois grandes problemas: "Classificação de eventos adventícios" e "Segmentação e Classificação Automática de sons adventícios".

A classificação de eventos adventícios (previamente segmentados) é um problema que ainda persiste mesmo tendo sido trabalhado no meio da comunidade científica ao longo anos. Com este estudo contribuímos para a continuação da investigação da aplicabilidade de diferentes tipos de atributos em diversos algoritmos de classificação. Tratámos o problema através de duas abordagens distintas. Uma consistiu na classificação de eventos por meios de algoritmos clássicos (SVM e k-NN) e outra através de algoritmos de aprendizagem profunda (CNN, dualCNN, LSTM, BiLSTM e GRU). A *feature* que se destacou pela negativa foi a utilização de cromogramas visto que alguns eventos têm tonalidades bastante parecidas, alcançando apenas 50% de exatidão, por outro lado, atributos como espectrogramas de mel obtiveram resultados bastante positivos alcançando uma exatidão de 81% e uma sensibilidade de 81%. A combinação de atributos através do dualCNN também obteve resultados positivos, apresentando para quase todos os tipos de atributos resultados de exatidão superiores a 80%. Comparativamente, as duas abordagens alcançaram resultados interessantes, mas como pudemos analisar, os algoritmos de aprendizagem profunda demonstraram a sua superioridade a distinguir os eventos sonoros.

Depois de provar que os algoritmos conseguem classificar eventos adventícios definiu-se, através do segundo problema, a exploração da segmentação e classificação automática de sons adventícios. Neste caso, tivemos como objetivo a criação de modelos para identificar a localização e existência de eventos adventícios automaticamente a partir dos ficheiros de áudio, sem a necessidade de identificação manual dos eventos. Este problema ainda é um grande desafio para a comunidade científica, uma vez que foi pouco explorado. Abordámos o problema de três formas distintas, utilizando algoritmos CNN: abordagem com três classes, e devido aos resultados pouco satisfatórios definimos também uma abordagem binária. Para explorar a eventual personalização das abordagens, definimos ainda uma abordagem por paciente. Em nenhuma das abordagens conseguimos alcançar resultados relevantes do ponto de vista clínico; alcançando na primeira experiência 7,3% de sensibilidade relativamente às *frames* pertencentes à classe 'fervores'. Na segunda abordagem conseguimos identificar 48,4% das *frames* 'sibilâncias' e 50,1% das *frames* pertencentes a fervores. A abordagem por paciente levanta algumas questões pertinentes sobre o problema, uma vez que o algoritmo de classificação funciona muito bem para uns pacientes e para outros não consegue identificar eventos adventícios. Existem pacientes que contém ficheiros com muitos eventos adventícios de uma classe e muito poucos de outra. Este facto prejudica as métricas do classificador (de um modo geral e por classe) uma vez que este vai abordá-lo como um problema com três classes apesar de só ter um ou dois eventos adventícios de uma classe apenas.

Com esta dissertação ficou demonstrado que os algoritmos de aprendizagem profunda conseguem distinguir os diversos eventos quando pré-segmentados a partir de diferentes tipos de atributos. O problema da classificação automática continua a ser um desafio e, apesar dos resultados conseguimos levantar algumas questões sobre o tema e consideramos uma boa primeira abordagem.

6.1 Trabalho Futuro

Conforme mencionado, tanto no capítulo do Estado de Arte como nas Conclusões, este trabalho aborda um tema que pode ser explorado de diversas formas. Uma vez que o problema da segmentação e classificação automática ainda não foi devidamente resolvido pela comunidade científica, acreditamos que ainda existam muitas novas oportunidades de investigação e formas de aprimorar o trabalho já realizado. Os próximos parágrafos são referentes a algumas sugestões do autor para trabalhos futuros.

Um aspeto que iria ajudar a melhorar projetos em curso e futuras investigações sobre problemas respiratórios e sons adventícios seria aumentar a base de dados em termos de diversidade de pacientes e das doenças diagnosticadas, uma vez que existem doenças que são pouco representadas. Além deste fator, deveriam ser analisadas em maior detalhe as anotações dos eventos presentes na base de dados.

Sugerimos, também, o alargamento das anotações e classificação de sons adventícios como roncos, fervores finos e grossos, uma vez que estes também estão relacionados com as patologias existentes na base de dados.

Quanto à segmentação e classificação automática de sons adventícios, seria interessante continuar esta investigação através da aplicação dos conjuntos de dados de *frames* gerados a outros classificadores de aprendizagem profunda como LSTMs, assim como a variância de alguns parâmetros de modelação e de treino, como por exemplo testes centrados na função objetivo dos classificadores, nas diferentes arquiteturas e diferentes métodos de balanceamento de dados. A aplicação de janelas de Hamming mais pequenas que as utilizadas na segmentação neste trabalho (inferiores a 32ms) também seria uma aplicação com valor de análise.

Por fim, outra oportunidade de investigação estaria relacionada com a aplicação de novos filtros/máscaras ao resultado dos classificadores de forma a facilitar a identificação de eventos adventícios e eliminar "ruído" na classificação, tendo sempre em atenção que estamos a trabalhar com eventos de curta duração.

Referências

[1] B. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. Kahya, N. Jakovljevic, T. Turukalo, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, N. Maglaveras, R. Pedro Paiva, I. Chouvarda, and P. de Carvalho, "Respiratory Sound Database," ICBHI 2017 Challenge, 2017. [Online]. Available: https://bhichallenge.med.auth.gr/. [Acedido em Novembro 2020].

[2] W. H. Organization, "The top 10 causes of death," World Health Organization, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death. [Acedido em Dezembro 2020].

[3] Pordata, "Óbitos por causas de morte," Base de Dados Portugal Contemporâneo, 2018. [Online]. Available: https://www.pordata.pt/Portugal/%c3%93bitos+por+algumas+causas+de+morte+(percentagem)-758-235714. [Acedido em Dezembro 2020].

[4] United for Lung Health, Fórum Internacional de Sociedades Respiratórias, "O Impacto Global da Doença Respiratória - Segunda Edição," 2017.

[5] V. S. M. Dias, "Classificação de sons respiratórios adventícios em crianças," Fevereiro 2020.

[6] A.R.A. Sovijärvi, F. Dalmasso, J. Vanderschoot, L.P. Malmberg, G. Righini, S.A.T. Stoneman, "Definition of terms for applications of respiratory sounds," 2000.

[7] Nandini Sengupta, Md Sahidullah, Goutam Saha, "Computers in Biology and Medicine," *Lung sound classification using cepstral-based statistical features,* pp. Google Scholar, ScienceDirect, Agosto 2016.

[8] Malay Sarkar, Irappa Madabhavi, Narasimhalu Niranjan, Megha Dogra, "Auscultation of the respiratory system," *Ann Thorac Med,* p. Google Scholar , 2015.

[9] SPP - Sociedade Portuguesa de Pneumologia, J. Agostinho Marques, Isabel Gomes, Pedro Silveira, Ana Rosa Santos, "Semiologia do Aparelho Resporatório," em *Curso Interactivo de Pneumologia*, Permanyer Portugal, 1997.

[10] Bruno Machado Rocha, Diogo Pessoa, Alda Marques, Paulo Carvalho, Rui Pedro Paiva, "Automatic Classification of Adventitious Respiratory Sounds: A (Un)Solved Problem?," *Sensors,* vol. 21, p. Google Scholar, 2021.

[11] A. Padilla-Ortiza e D. Ibarra, "Lung and Heart Sounds Analysis: State-of-the-Art," *Critical Reviews in Biomedical Engineering*, 2018.

[12] Sandra Reichert, Raymond Gass, Christian Brandt and Emmanuel Andrès, "Analysis of Respiratory Sounds: State of the Art," *Circulatory, respiratory and pulmonary medicine,* vol 2, 2008.

[13] H. U. I. D. Y. O. Y. M. Mitsuru Munakata, "Thorax," *Spectral and waveform characteristics of fine and,* pp. 46:651-657, 1 Setembro 1991.

[14] Saudi Thoracic Society, "Annals of Thoracic Medicine," *Auscultation of the respiratory system,* pp. Volume 10, Issue 3, Julho-Setembro 2015.

[15] N. Caka, "ResearchGate," University of Prishtina, Março 2015. [Online]. Available: https://www.researchgate.net/post/What-are-the-Spectral-and-Temporal-Features-in-Speech-signal. [Acedido em Janeiro 2021].

[16] O. Lartillot, MIRtoolbox 1.7.2 User's Manual, 2019.

[17] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang, "Deep Learning for Health Informatics," *IEEE Journal of Biomedical and Health Informatics,* vol. 21, nº 1, 2017.

[18] H. Dalianis, Clinical Text Minig, Springer Open, 2018.

[19] Liqun Wu and Ling Li, "Investigating into segmentation methods for diagnosis of respiratory".

[20] Chelsea Villanueva, Joshua Vincent, Alexander Slowinski, Mohammad-Parsa Hosseini.*Respiratory Sound Classification Using Long-Short Term Memory.*

[21] Microsoft Research India, Microsoft, "RESPIRENET: A DEEP NEURAL NETWORK FOR ACCURATELY DETECTING," 31 Outubro 2020.

[22] Dat Ngo, Lam Pham, Anh Nguyen, Ben Phan, Khoa Tran, Truong Nguyen, "Deep Learning Framework Applied For Predicting Anomaly of," 26 Dezembro 2020.

[23] Hai Chen, Xiaochen Yuan, Zhiyuan Pei, Mianjie Li, Jianqing Li, "Triple-Classification of Respiratory Sounds Using," 26 Março 2019.

[24] A. T. Diego Perna, "Deep auscultation: Predicting respiratory anomalies".

[25] Koki Minami, Huimin Lu, Hyoungseop Kim, Shingo Mabu, Yasushi Hirano, Shoji Kido, "Automatic Classification of Large-Scale Respiratory Sound Dataset," Outubro 2019.

[26] D. Chamberlain, R. Kodgule, D. Ganelin e a. R. R. F. Vivek Miglani, "Application of Semi-Supervised Deep Learning to Lung Sound Analysis".

[27] Lin, Bor-Shing Lin, Bor-Shyh, "Automatic Wheezing Detection Using Speech Recognition Technique," *Taiwanese Society of Biomedical Engineering,* Agosto 2016.

[28] T. Fernando, S. Denman, S. Sridharan e C. Fookes, "Heart Sound Segmentation Using Bidirectional LSTMs With Attention," *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS,* vol. 24, nº 6, 2020.

[29] Elmar Messner, Matthias Zohrer, Franz Pernkopf, "Heart Sound Segmentation—An Event Detection Approach Using Deep Recurrent Neural Networks," *IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING,* vol. 65, nº 9, 2018.

[30] PhysioNet, "PhysioNet," [Online]. Available: https://physionet.org/about/database/. [Acedido em Janeiro 2021].

[31] F. Renna, J. Oliveira e M. T. Coimbra, "Convolutional Neural Networks for Heart Sound Segmentation," em *26th European Signal Processing Conference (EUSIPCO)*, 2018.

[32] Juncheng Li, Wei Dai, Florian Metze, Shuhui Qu, Samarjit Das.*A Comparison Of Deep Learning Methods For Environmental Sound*.

[33] audEERING, "openSMILE 3.0," audEERING, [Online]. Available: https://www.audeering.com/opensmile/. [Acedido em Janeiro 2021].

[34] Hyungui Lim, Jeongsoo Park, Kyogu Lee, Yoonchang Han, "RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL," 16 Novembro 2017.

[35] Renard Xaviero Adhi Pramono, Stuart Bowyer, Esther Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review".

[36] Zeenat Tariq, Sayed Khushal Shah, Yugyung Lee, "Lung Disease Classification using Deep Convolutional Neural Network," 2019.

[37] F. Lindsten, "A remark on zero-padding for increased frequency," 4 Novembro 2010.

[38] R. N. Keiron O'Shea, "An Introduction to Convolutional Neural Networks," Novvembro 2015.

[39] t. d. science, "The mostly complete chart of Neural Networks, explained," 4 Agosto 2017. [Online]. Available: https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464. [Acedido em Janeiro 2021].

[40] J. M. S. -. isep, "Redes Neuronais Conceitos".

[41] M. i. retrieval, "Mel Frequency Cepstral Coefficients (MFCCs)," [Online]. Available: https://musicinformationretrieval.com/mfcc.html. [Acedido em Janeiro 2021].

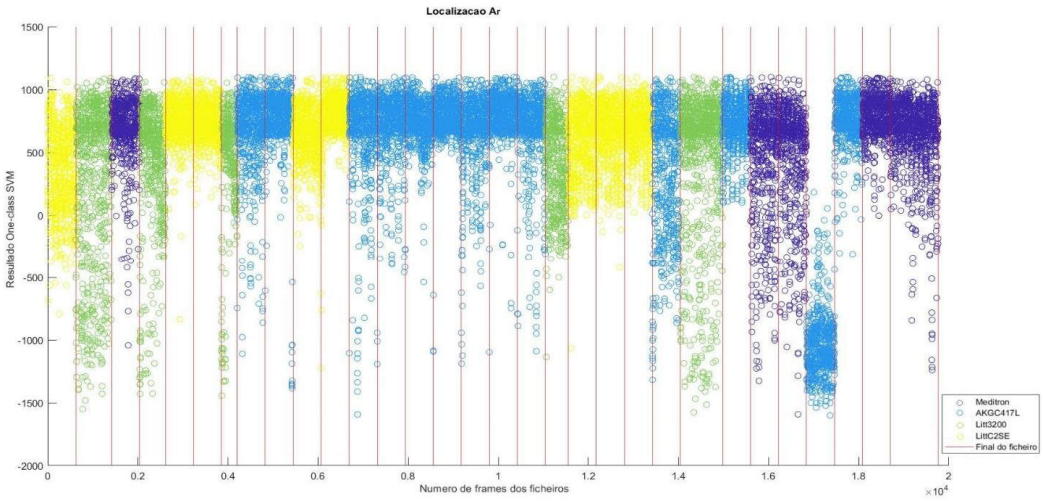
[42] L. Mendes, I. M. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, V. Tsara, C. Teixeira, P. Carvalho, J. Henriques, R. P. Paiva, "Detection of wheezes using their signature in the spectrogram space," 2020.

[43] I. The MathWorks. [Online]. Available: https://www.mathworks.com/help/dsp/ref/dsp.stft.html. [Acedido em 29 10 2021].

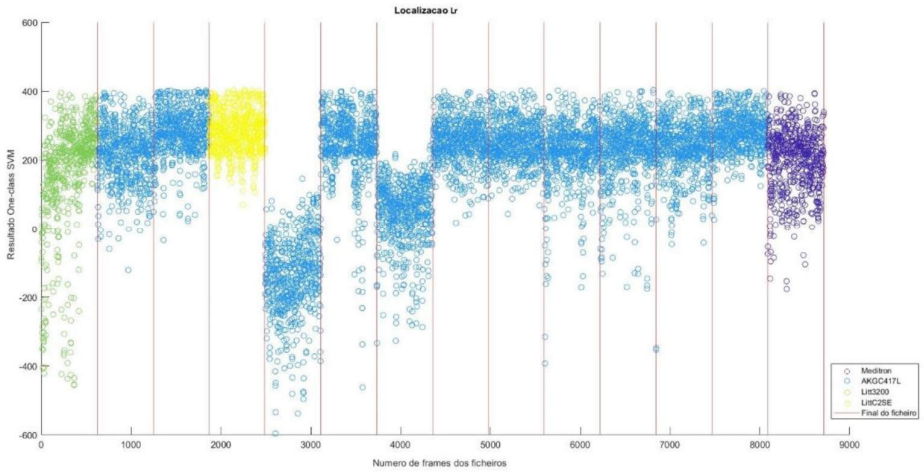
Apêndices

Apêndice A

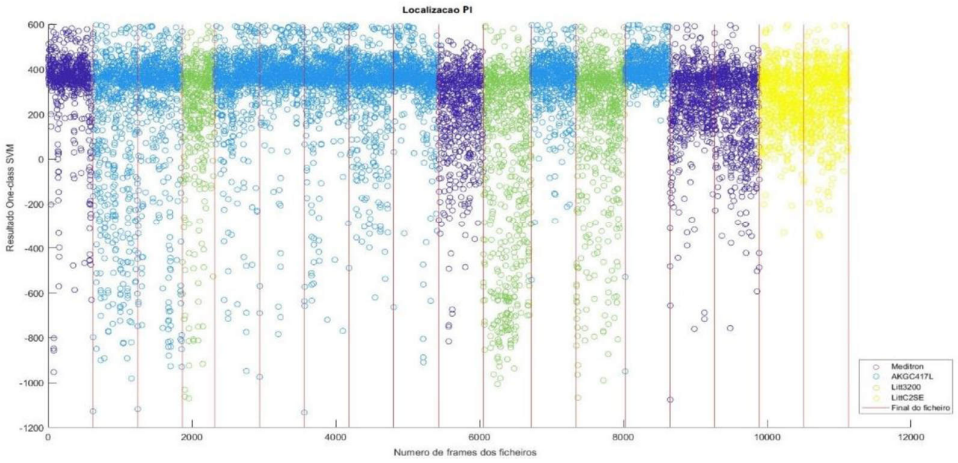
Neste apêndice estão presentes todos os gráficos que completam a análise 1, realizado no capítulo 4.
Usando os atributos *"centroid"* e "*brightness* 4" ratio obtiveram-se:



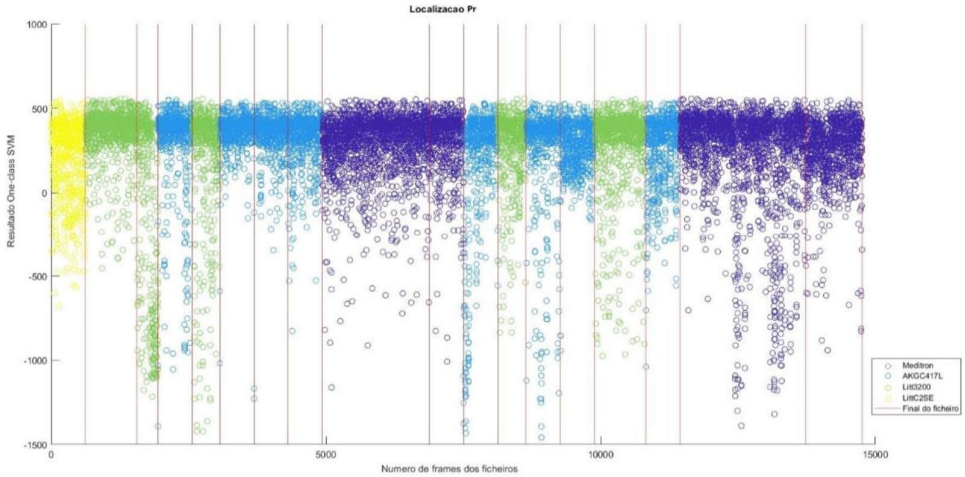
A 1 - Resultado One class SVM anomalias Ar



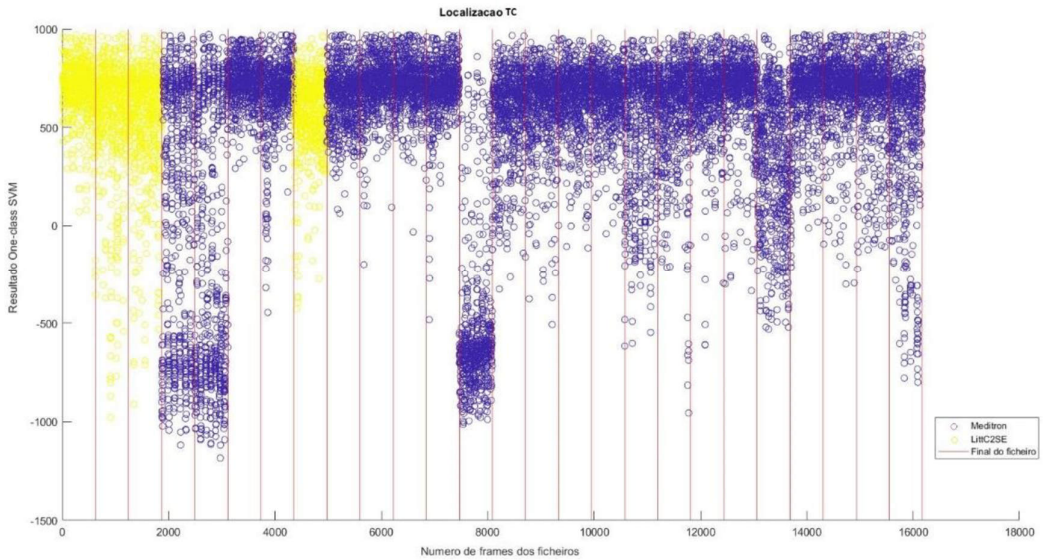
A 2 - Resultado One class SVM anomalia Lr



A 3 - Resultado One class SVM anomalia Pl

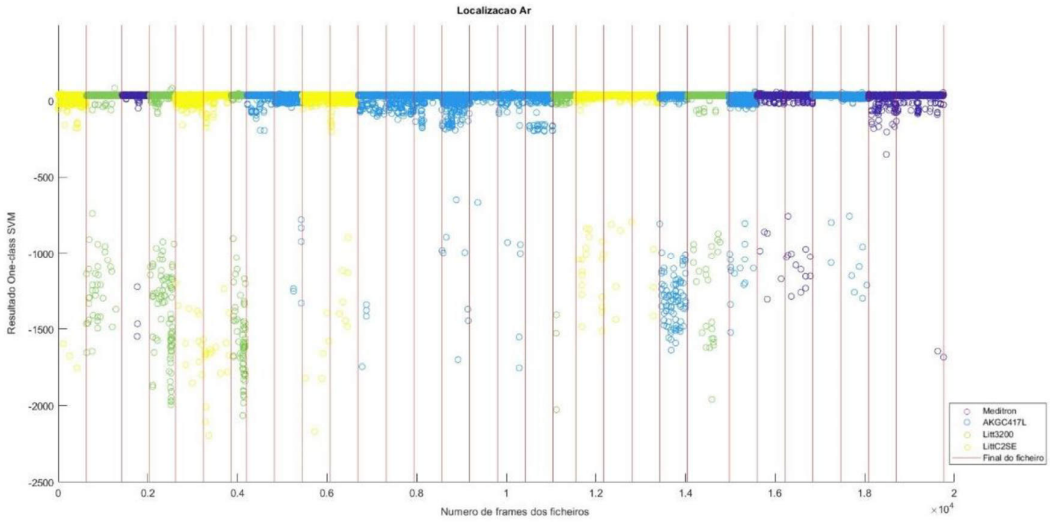


A 4 - Resultado One class SVM anomalia Pr

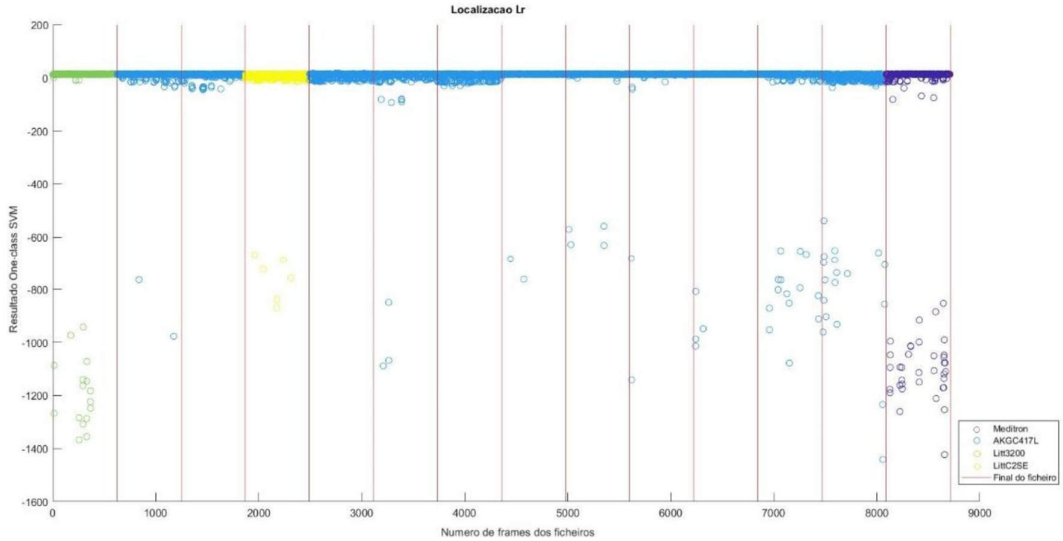


A 5 - Resultado One class SVM anomalia Tc

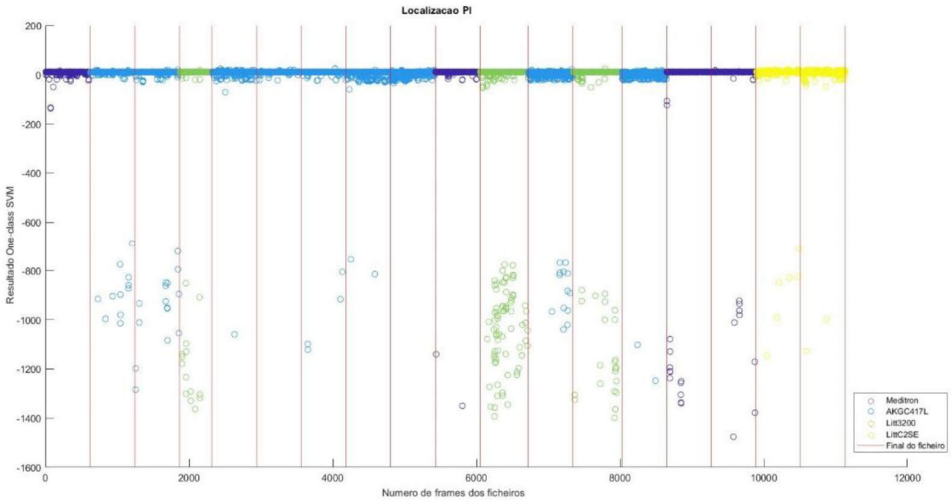
Usando-se os atributos melódicos*:*



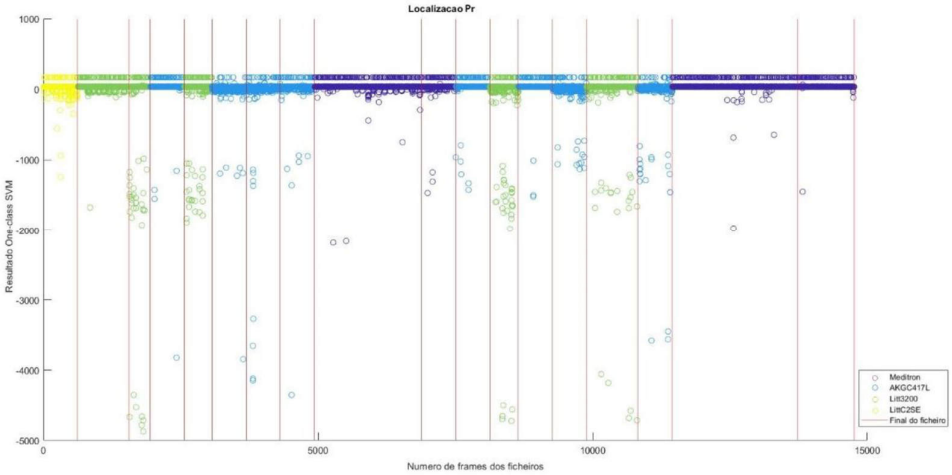
A 6 - Resultado One class SVM anomalia Ar



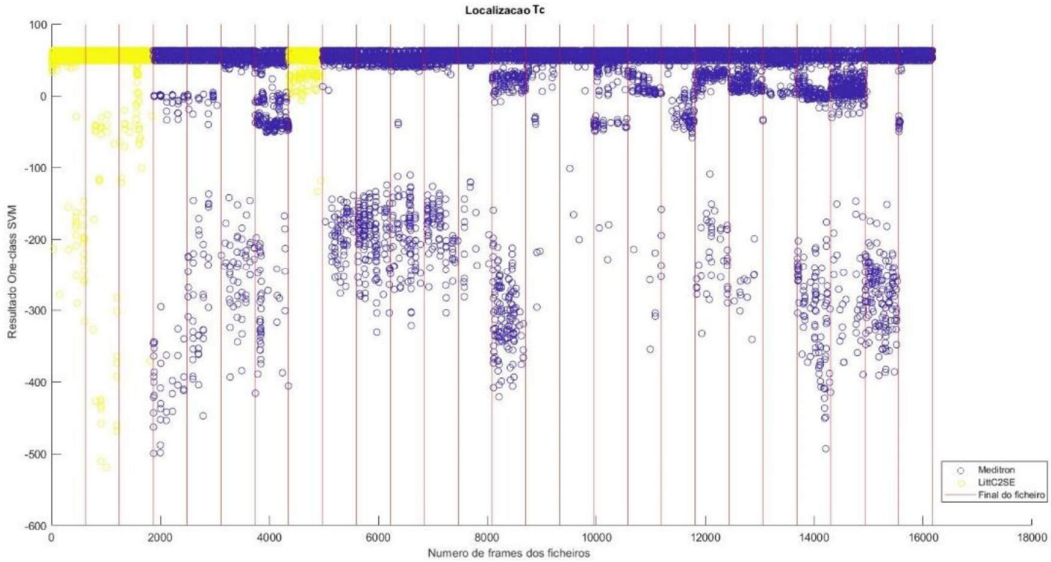
A 7 - Resultado One class SVM anomalias Lr



A 8 - Resultado One class SVM anomalias Pl

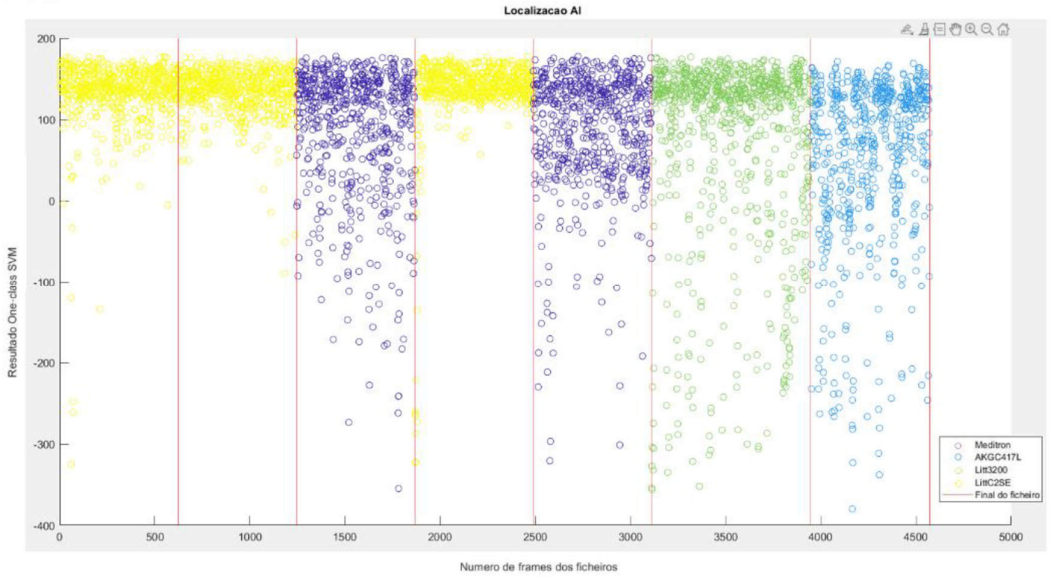


A 9 - Resultado One class SVM anomalias Pr

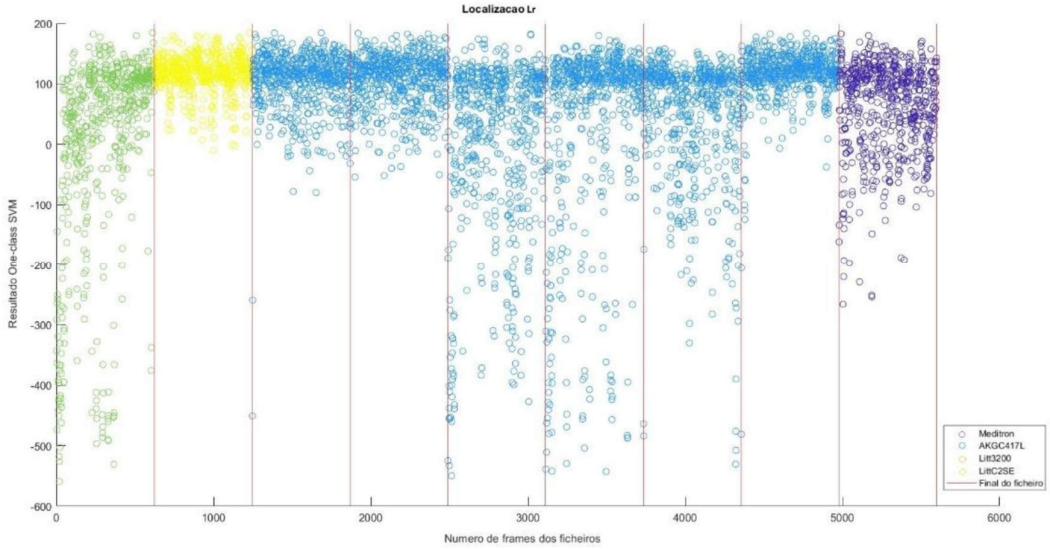


A 10 - Resultado One class SVM anomalias Tc

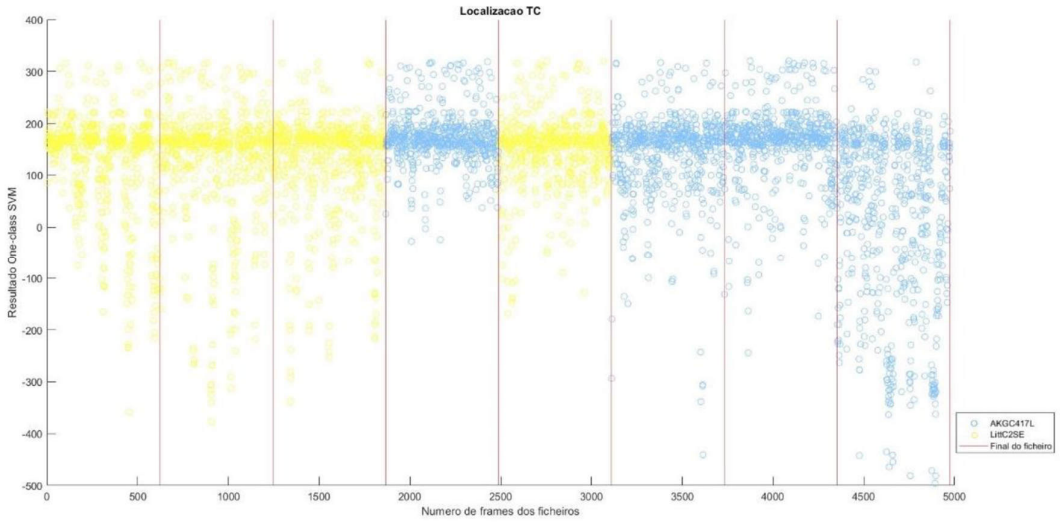
Na segunda experiência, utilizando os espectrogramas como referência obteve-se:



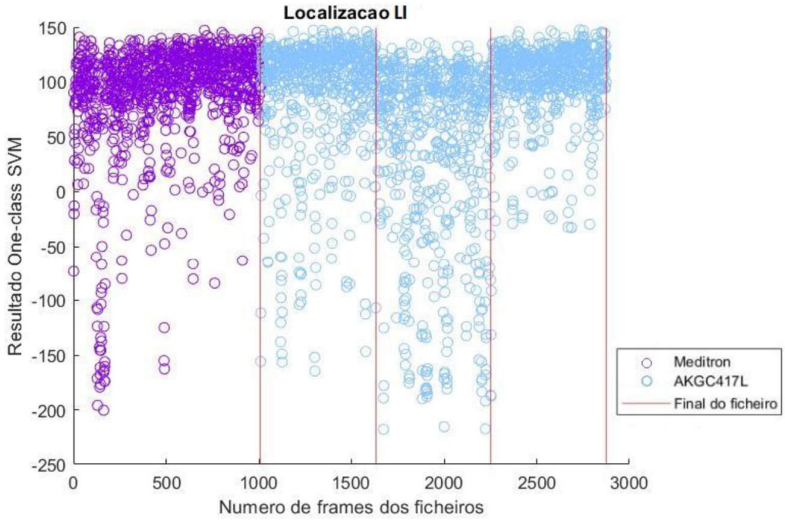
A 11 - Resultado One class SVM anomalias Al



A 12 - Resultado One class SVM anomalias Lr



A 13 - Resultado One class SVM anomalias Tc



A 14 - Resultado One class SVM anomalias LI