INSTITUTO POLITÉCNICO DE LISBOA

Instituto Superior de Engenharia de Lisboa
Escola Superior de Tecnologia da Saúde de Lisboa

Previsão automática da mortalidade em UCI de doentes com síndrome da dificuldade respiratória aguda associada à COVID-19 utilizando radiografias de tórax e dados clínicos

Tiago Alexandre dos Santos Galvão

Trabalho Final de Mestrado para obtenção do grau de
Mestre em Engenharia Biomédica

Orientadores

Nuno Alexandre Soares Domingues (ISEL)
Pedro Miguel Torres Mendes Jorge (ISEL)
Luís Bento (CHULC)

Júri

Presidente:	Cecília da Cruz Calado (ISEL)
Vogais:	João Ferreira (ISCTE)
	Pedro Miguel Mendes Jorge (ISEL)

Novembro de 2023

INSTITUTO POLITÉCNICO DE LISBOA

Instituto Superior de Engenharia de Lisboa
Escola Superior de Tecnologia da Saúde de Lisboa

Previsão automática da mortalidade em UCI de doentes com síndrome da dificuldade respiratória aguda associada à COVID-19 utilizando radiografias de tórax e dados clínicos

Tiago Alexandre dos Santos Galvão

Trabalho Final de Mestrado para obtenção do grau de
Mestre em Engenharia Biomédica

Orientadores

Nuno Alexandre Soares Domingues (ISEL)
Pedro Miguel Torres Mendes Jorge (ISEL)
Luís Bento (CHULC)

Júri

Presidente:	Cecília da Cruz Calado
Vogais:	João Ferreira (ISCTE)
	Pedro Miguel Mendes Jorge (ISEL)

Novembro de 2023

Agradecimentos

É com grande gratidão que reconheço o apoio e orientação recebidos durante a elaboração desta dissertação.

Em primeiro lugar, gostaria de agradecer à Drª. Taia Cysneiros e ao Dr. Luís Bento, por toda a colaboração prestada e trabalho incansável desenvolvido na recolha de dados e definição de objetivos para a realização desta dissertação.

Aos meus orientadores, Professor Nuno Domingues e Professor Pedro Jorge por puxarem por mim na carreira académica e encontrarem sempre soluções para os problemas imprevistos.

Um agradecimento à minha família, em especial, à minha mãe Anabela, ao meu pai Joaquim, e ao meu irmão Tomás, por poder sempre contar com eles.

Aos meus amigos João, Belchior, Sara, Catarina e Raquel por me acompanharem nesta jornada do princípio ao fim.

À minha Ana, por ter sido a minha âncora como sempre, nunca me ter feito desistir e apoiar-me nos momentos mais difíceis.

Uma dedicatória especial ao Sr. Manuel, que nos deixou demasiado cedo, mas sempre acreditou em mim e fez sempre tudo pela família.

Um obrigado a todos.

Este trabalho foi realizado no âmbito do projeto DSAIPA/DS/0117/2020, financiado pela Fundação a Ciência e a Tecnologia, Portugal.

Resumo

A síndrome da dificuldade respiratória aguda associada à COVID-19 (ARDS-COV19), é uma síndrome pulmonar grave que resulta em insuficiência respiratória aguda. A ARDS é complexa e heterogénea, exigindo frequentemente ventilação mecânica invasiva (VMI) em unidades de cuidados intensivos (UCI). A identificação de grupos de risco é crucial para a medicina de precisão, embora a falta de métodos de diagnóstico seja limitativo. A radiografia torácica é um exame imagiológico, qualitativo e acessível, utilizado na rotina das UCIs. É essencial o desenvolvimento de um classificador multivariado e quantitativo, baseado em *radiomics*, para a previsão da mortalidade destes doentes sob VMI. Para este efeito foram incluídos 110 doentes ARDS-COV19 de uma UCI, com uma idade média de 63,2 ± 11,92 anos, sendo 61,2% do sexo masculino. A mortalidade foi de 47,3%. Radiografias do 1º e 3º dia de VMI foram recolhidas, pré-processadas e concatenadas. Características de *deep learning* foram então extraídas, utilizando uma rede neuronal convolucional pré-treinada (CheXnet). Estas características foram acopladas a variáveis clínicas (VC), para a construção de dois modelos de aprendizagem automática, um de regressão logística (LogReg) e um perceptrão multicamada (MLP). A idade, a razão PaO_2/FiO_2 do 3º dia de VMI e uma característica de imagem (DLF_258) foram utilizadas nos modelos finais. Os modelos que incluíram a DLF_258, apresentaram 89% (LogReg) e 82% (MLP) de probabilidade de terem melhor exatidão, do que os modelos de VC. No grupo de teste interno (23 doentes), o modelo de LogReg obteve os melhores resultados e menor *overfitting*, com uma *área under the ROC curve* (AUC) de 0,862 95%CI [0.654, 0.969], uma exatidão de 0,783 95%CI [0.563, 0.926] e um *score* de F1 de 0,783 95%CI [0.563, 0.926]. Apesar dos resultados promissores, o número de amostras foi reduzido, não existindo um teste externo. A recolha de dados e posterior validação são assim essenciais.

Palavras-Chave: ARDS; UCI; COVID-19; Radiografia; Mortalidade

Abstract

Acute respiratory distress syndrome associated with COVID-19 (ARDS-COV19) is a severe pulmonary syndrome leading to acute respiratory failure. ARDS is complex and heterogeneous, with patients frequently needing invasive mechanical ventilation (IMV) in intensive care units (ICUs). The identification of risk groups is crucial for precision medicine, although the lack of diagnostic methods can be limiting. Chest radiography is a qualitative and accessible imaging examination routinely used in ICU settings. The development of a multivariate and quantitative classifier based on radiomics is essential for predicting the mortality of patients under IMV. For this purpose, 110 ARDS-COV19 patients from an ICU, with an average age of 63.2 ± 11.92 years, of whom 61.2% were male, were included. The mortality rate was 47.3%. Chest X-rays from the 1st and 3rd days of IMV were collected, pre-processed, and concatenated. Deep learning features were then extracted using a pre-trained convolutional neural network (CheXnet). These features were combined with clinical variables (CV) to build two machine learning models: a logistic regression model (LogReg) and a multilayer perceptron (MLP). Age, the PaO_2/FiO_2 ratio on the 3rd day of IMV, and an image feature (DLF_258) were used in the final models. The models that included DLF_258 showed 89% (LogReg) and 82% (MLP) probability of having better accuracy than CV models. In the internal test group (23 patients), the LogReg model achieved the best results with lower overfitting, providing an area under the ROC curve (AUC) of 0.862, 95% CI [0.654, 0.969], an accuracy of 0.783, 95% CI [0.563, 0.926], and an F1 score of 0.783, 95% CI [0.563, 0.926]. Despite promising results, the sample size was limited, and external testing is lacking. Therefore, data collection and subsequent validation are essential.

Keywords: ARDS, ICU, COVID-19, X-ray, Mortality

1. Índice Geral

Índice de Tabelas

Índice de Figuras

Lista de Abreviaturas, Siglas e Símbolos

ARDS – Síndrome da Dificuldade Respiratória Aguda

ARDS-COV19 – ARDS associada a SARDS-CoV-2

AUC - *Area under the receiver operating characteristic curve*

CA – *Classification accuracy* (exatidão)

CLAHE - Equalização do histograma adaptativa com limite de contraste

CNN – Rede Neuronal Convolucional

CV –*Clinical variables* (variáveis clínicas)

CV2 - OpenCV

DL – *Deep Learning*

DLF – *Deep learning features*

dp – Pressão de distenção

FCL – Camada totalmente conectada

GB – *Gradient boosting*

IA – Inteligência Artificial

LogReg – Regressão logística

LSTM - *long short-term memory*

ML – *Machine Learning*

MLP – perceptrão multicamada

peep - pressão expiratória positiva final

PET - Tomografia por Emissão de Positrões

PF - Razão PaO_2/FiO_2

PS – Suporte de pressão

RC – Radiografia Convencional

RGB – Red, green, blue

RM – Ressonância Magnética

ROC - *receiver operating characteristic curve*

ROI – *region of interest*

R-TRX – Radiografia do tórax

SVM – *support vector machines*

TC – Tomografia Computorizada

UCI – Unidade de cuidados intensivos

VMI – Ventilação mecânica invasiva

1. Introdução

W. Rogers e colegas referem que, desde 1990, diversos investigadores têm feito previsões sobre os possíveis benefícios do ramo das *genomics*. O mesmo poderia apresentar o potencial de transformar a medicina terapêutica através do estudo das sequências genéticas e epigenéticas dos doentes, permitindo assim a medicina de precisão. A filosofia desta vertente terapêutica refere que a inclusão de informação genómica dos doentes na clínica diária, poderia trazer melhorias significativas no diagnóstico, previsão e tratamento de doenças heterogéneas. Recentemente esta definição tem tendencionalmente evoluído para além da genómica. A inclusão na prática clínica de fatores metabólicos, informação sobre proteínas e dados extraídos de imagens médicas, impulsionou uma mudança em como a medicina é praticada e estudada. O foco reside agora em tratamentos personalizados aos indivíduos e às possíveis variações da doença, não considerando apenas o desenvolvimento de protocolos *standard* baseados em estudos clínicos generalizados.

A síndrome do desconforto respiratório agudo, mais conhecida por ARDS (*acute respiratory distress syndrome)* é um exemplo de uma complicação heterogénea que poderá beneficiar da medicina de precisão, sendo inicialmente descrita em 1967[1]. Neste caso clínico um conjunto de doentes apresentava insuficiência respiratória aguda com infiltrados pulmonares bilaterais de substâncias mais densas que o ar [1]. Desde então, a definição da ARDS foi alvo de evolução ao longo dos anos. Atualmente existe um consenso médico na chamada definição de Berlim, que depende da avaliação de exames de imagem médica, como a radiografia do tórax (CRX), e da análise da gasometria do sangue arterial. A avaliação destes gases permite o cálculo da razão entre a pressão parcial do oxigénio e o oxigénio inspirado (PaO_2/FiO_2, ou razão PF) que por sua vez depende do valor da pressão expiratória positiva final (peep) de cada ciclo respiratório. Estes valores são assim essenciais na classificação atual da síndrome descrita pela definição de Berlim [2]. Neste contexto a ARDS é então definida por uma síndrome pulmonar aguda (instalação ou novo agravamento em uma semana) que envolve infiltrados pulmonares bilaterais e hipoxemia, não podendo ser justificada apenas por disfunção ventricular esquerda, sobrecarga de fluídos ou doença pulmonar crónica [3]. Apesar da definição aparentemente simples, a ARDS é uma síndrome de complexidade considerável, visto que a sua expressão e severidade está coassociada a diversas patologias, distúrbios clínicos, anormalidades pulmonares, anomalias radiográficas do tórax, diferentes expressões genéticas, variabilidades de lesões nas vias biológicas e variabilidades de evolução temporal, desde o desenvolvimento precoce de lesões pulmonares agudas até à necessidade de ventilação mecânica invasiva (VMI)[4–6].

Neste último fator reside a principal preocupação clínica. A ARDS é uma síndrome comum, heterogénea e grave, com a maioria dos doentes a necessitar de VMI nas unidades de cuidados intensivos. Esta encontra-se também associada a taxas de incidência, mortalidade e morbilidade consideráveis principalmente em situações de maior severidade da doença [7]. Visto que a definição da ARDS e a avaliação da sua severidade dependem da gasometria arterial, existe uma limitação no seu diagnóstico em países e centros clínicos, que na sua rotina, não dispõem destes métodos de diagnóstico. O referido cria também uma limitação na inclusão de prováveis doentes com ARDS em ensaios clínicos, visto que os mesmos não preenchem os requisitos de inclusão necessários. Outra problemática a considerar é o facto das opções terapêuticas para esta síndrome serem algo limitativas, apesar da mesma ser uma causa comum de insuficiência respiratória aguda [8]. Poucas opões terapêuticas têm demonstrado benefícios no prognóstico dos doentes para além da ventilação protetora dos pulmões, existindo a possibilidade desta limitação derivar da heterogeneidade subjacente apresentada pela ARDS [7,8]. A criação de novos métodos para identificação de subfenótipos

da doença e para classificação/avaliação da severidade de ARDS, é assim fundamental e representa em grande parte o estado de arte atual para atingir uma medicina de precisão.

Um subfenótipo é definido como um subgrupo dentro de uma entidade de doença que, apresenta maior risco de resultados desfavoráveis (enriquecimento prognóstico), ou, partilha fatores biológicos proximais que implicam reações semelhantes às medidas médicas tomadas (enriquecimento preditivo) [5,9]. Estas estratégias de enriquecimento oferecem o potencial de reduzir a heterogeneidade, permitindo assim, uma abordagem de medicina personalizada ao selecionar o subgrupo com maior probabilidade de obter benefícios das diversas terapêuticas [5,9]. Ensaios clínicos aleatorizados e estudos de corte recentes identificaram dois subfenótipos com características clínicas e biológicas diferentes utilizando *latent class analysis* (LCA), o subfenótipo hipoinflamatório e hiperinflamatório. Estes subfenótipos apresentam resposta distinta a valores de peep altos e baixos [6].

A ARDS pode também ser diferenciada entre primária e secundária de acordo com a etiologia da mesma. A primeira tipologia considera uma precedente lesão ou patologia pulmonar, enquanto a segunda considera etiologias provenientes de outras patologias. Recentemente uma nova tipologia tem sido considerada, sendo esta a ARDS associada à SARS-CoV-2 [10]. A pandemia da Covid-19 causou aumentos de mortalidade e morbilidade em todo o mundo, sobrecarregando os sistemas de saúde e UCI. Os doentes com Covid-19 de maior risco requerem VMI e continuam a ser uma preocupação presente estando fortemente associados à ARDS [7]. A investigação sobre a SARS-CoV-2, levou ao ensaio RECOVERY, que foi o primeiro a demonstrar um benefício forte e inequívoco dos corticosteroides em pacientes com esta infeção, que necessitam de suplementação de oxigénio e ainda mais benéfico naqueles que necessitam de ventilação mecânica invasiva (VMI) [11]. Isto demonstra a possibilidade e subfenótipos associados à ARDS associada à SARS-CoV-2 (ARDS-COV19), pois até à data presente, muitos estudos não conseguiram comprovar os benefícios desta terapêutica em doentes com ARDS dita “comum”. Vários estudos sobre biomarcadores proteicos têm demonstrado uma melhor compreensão da fisiopatologia da ARDS e podem oferecer uma possível abordagem baseada em personalização e atribuição de subfenótipos a esta síndrome. Estes biomarcadores proteicos da ARDS são frequentemente estudados no sangue (geralmente plasma) ou nos espaços aéreos do pulmão (geralmente usando fluido de lavagem bronco-alveolar) [12]. Este último tem a vantagem de estudar diretamente pulmões lesionados, no entanto, a padronização na diluição e aquisição destes fluídos são difíceis, não sendo isenta de complicações em doentes submetidos a VMI [12,13].

É necessário assim um método alternativo que possa despender estes métodos de diagnóstico invasivos. A solução poderá estar no potencial de diagnóstico quantitativo não aproveitado das imagens médicas, como no caso da radiografia torácica (R-TRX). A R-TRX portátil é um método de diagnóstico económico, disponível e utilizado diariamente nas UCI’s para diagnóstico e caracterização da ARDS e para averiguação da severidade da SARS-CoV-2. A R-TRX está fortemente associado à ARDS, dependendo a sua definição da avaliação qualitativa da mesma [14,15]. Alguns trabalhos têm analisado padrões de imagiologia torácica e a gravidade da ARDS. O score de avaliação radiográfica de edema pulmonar (RALE) foi desenvolvido como uma medida semi-quantitativa de edema pulmonar em pacientes com ARDS, com scores mais altos de RALE associados de forma independente a uma menor razão PaO2/FiO2 e menor sobrevida [16]. Uma alteração no RALE durante os primeiros dias após o início da ARDS está também associada de forma independente à sobrevivência, tornando-o um potencial método de diagnóstico para avaliação de novas terapêuticas, no entanto o mesmo não está associado à mortalidade em 28 ou 90 dias pós VMI [16]. Uma análise de doentes em UCI sob ventilação invasiva verificou que o score RALE tem uma excelente precisão diagnóstica para a ARDS, com uma *area under the receiver operating characteristic*

curve (AUC) de 0,91 [16,17]. Apesar das claras vantagens, o RALE contínua a ser um método semi-quantitativo, dependendo assim da avaliação subjetiva das radiografias de tórax. Este facto faz com que o método possa ser dispendioso em termos temporais e torna-o sujeito a variabilidades inter e intra-operador. Existe assim interesse num método automático de previsão do risco de mortalidade quantitativo utilizando as radiografias torácicas para consequente identificação de grupos de risco e possíveis subfenótipos da síndrome.

A aprendizagem automática (*machine learning)* em radiologia é um tópico de investigação em crescimento e tem sido utilizado com sucesso no diagnóstico da ARDS através da utilização de métodos computacionais quantitativos replicáveis. As metodologias de *radiomics p*ermitem a extração de características das imagens médicas quantificáveis, quer por métodos de estatística clássica de textura de imagem, como por métodos de aprendizagem profunda (*deep learning,* DL*)*, para a execução de várias tarefas de classificação, segmentação ou previsão de prognóstico dos doentes [18–20].

Os estudos referenciados neste capítulo focam-se na categorização ou previsão de severidade e mortalidade da ARDS através da investigação por diferentes ângulos, desde biomarcadores, até fisiologia pulmonar e imagiologia torácica, no entanto poucos utilizam dados clínicos e R-TRX em conjunto. É necessário manter uma abordagem completa, pois a SARS-CoV-2, como outras formas primárias da ARDS, reflete uma interação complexa de fatores biológicos que provavelmente não são passíveis de uma análise simples de variáveis isoladas [21].

Os que utilizam o conjunto destas técnicas demonstram metodologias semelhantes de *transfer learning*, extraindo características das radiografias de redes neuronais convulocionais (*convolutional neural network*, CNN's) pré-treinadas e procedendo à concatenação de dados clínicos/laboratorias para o treino de um classificador [22–24].

Abrem-se a assim portas para a presente investigação. Em colaboração com a equipa médica pneumologista do departamento do tórax do Hospital de São José (Centro Hospitalar de Lisboa Norte), foi possível recolher informação clínica e laboratorial de 110 doentes internados em UCI com ARDS-COV19. Essa mesma informação foi complementada por R-RTX portáteis do 1º e 3º dia de VMI nas UCI's de cada doente, permitindo assim estudar novas metodologias de treino de modelos de classificação da severidade da ARDS, utilizando as radiografias de tórax e as informações clínicas disponíveis.

1.1 Objetivo

O objetivo principal da presente investigação incide na criação de um modelo binário de classificação categórica da probabilidade de mortalidade de doentes com ARDS-COV19 internados em UCI, utilizando radiografias de tórax (do 1º e 3º dia de VMI em UCI) e a informação clínica/laboratorial disponível. É pretendido atingir valores de métricas de performance entre os 70-80% [25] (os mesmos são considerados aceitáveis em técnicas de *machine learning*), principalmente no que diz respeito à exatidão e AUC do modelo treinado, identificando assim possíveis grupos de risco e auxiliando a decisão clínica personalizada. Secundariamente, é pretendido avaliar se a adição da informação de imagens médicas aos dados clínicos no modelo promove melhorias de *performance* dos classificadores.

1.2 Estrutura da dissertação

A dissertação apresentada encontra-se descrita em seis capítulos distintos. O primeiro capítulo (Introdução) representa a introdução do trabalho, onde é feita uma reflexão sobre a temática e os seus conceitos básicos. O problema a considerar é também definido, assim como os principais objetivos e a estrutura da investigação. O segundo capítulo (Estado de Arte) desenvolve os conceitos chave, desde a estatística descritiva da ARDS, até possíveis fluxos de trabalho para os *radiomics,* considerando o tratamento de dados, a seleção das variáveis e os possíveis modelos de classificação. O terceiro capítulo (Metodologia) descreve os métodos aplicados através da análise dos dados disponíveis, por métodos de estatística descritiva, e através da descrição detalhada dos passos para a criação do modelo de classificação, previamente justificados no capítulo 2. No quarto capítulo (Resultados) são apresentados os principais resultados da investigação, comparando os diversos modelos de classificação utilizados. No quinto capítulo (Discussão), irá ser discutido os resultados e os achados, comparando com estudos relevantes e descrevendo as possíveis limitações do estudo. Por fim no sexto capítulo (Conclusão) são descritas as principais conclusões da dissertação, assim como o trabalho futuro a ser efetuado.

1.3 Contributos da dissertação

Os principais contributos desta dissertação residem na criação de um modelo de previsão de mortalidade específico para doentes de UCI com ARDS-COV19, considerando uma população homogénea da doença. Existem diversos estudos semelhantes publicados para doentes COVID-19, no entanto, nenhum referenciado e estudado teve em conta esta população de alto risco em específico ou identificou a patologia pela definição de Berlim. O modelo criado depende de variáveis e exames imagiológicos acessíveis à grande maioria das UCI's, sendo globalmente viável a sua aplicação. O estudo também demonstrou, através de métodos bayesianos o potencial de melhoria que as características quantitativas de radiografias poderiam trazer a modelos de previsão de mortalidade em UCI. Este facto salienta a importância da imagem médica na decisão terapêutica e identificação de grupos de risco na ARDS.

2. Revisão do estado de arte

No presente capítulo é apresentada uma introdução aos conceitos e etapas essenciais para aplicações baseadas em Inteligência Artificial (IA) e *machine learning* em imagem médica, assim como uma revisão do estado de arte de modelos e algoritmos de classificação da tipologia e prognóstico de doentes com ARDS e/ou Covid-19 presentes na literatura.

2.1 ARDS

Conforme referido na introdução, a ARDS é uma síndrome clínica de complexidade considerável com diversos níveis de severidade e etiologias distintas. A mesma, de acordo com a definição atual, é caracterizada pela diminuição súbita do oxigénio do sangue e pela presença de opacidades pulmonares bilaterais difusas, não podendo as mesmas serem justificadas por insuficiência cardíaca ou sobrecarga de fluidos no volume pulmonar[26]. A ARDS é assim um processo inflamatório pulmonar de origem aguda, que consequentemente promove o aumento da permeabilidade dos vasos sanguíneos pulmonares, com redução associada do funcionamento fisiológico do tecido pulmonar ventilado. Estes fatores traduzem-se na dispneia (dificuldade respiratória) aguda, com necessidade de VMI em doentes considerados críticos[2].

Esta síndrome passou por diversas definições clínicas desde a sua descrição inicial em 1967, onde *David G. Ashbaugh et al.* observaram um caso clínico de um grupo de doentes com hipoxemia grave sem lesões torácicas associadas e que não respondiam à terapêutica[1]. Posteriormente em 1994 foi criado o Comitê Americano-Europeu de Consenso sobre a ARDS para identificar a fisiopatologia da doença, o prognóstico dos doentes e as possíveis abordagens terapêuticas[27]. Ainda assim, o desenvolvimento do conhecimento desta síndrome heterogénia, desencadeou uma nova análise em 2011 em colaboração com a Sociedade Europeia de Cuidados Intensivos e a Sociedade Americana Torácica. Desta colaboração surgiu a definição de Berlim que ainda é amplamente aceite e utilizada globalmente [2].

2.1.1 Definição de Berlim

A definição de Berlim da ARDS inclui diversos critérios. Em primeiro lugar deverá ser uma síndrome aguda com expressão sintomática inicial até a uma semana após exposição ao fator de risco ou agressão desencadeante. Outro critério é a observação de padrões radiográficos tradutores de opacidades bilaterais difusas alveolares, através de radiografias torácicas ou de imagens por tomografia computorizada[2]. Este edema pulmonar observado não pode ser justificado por insuficiência cardíaca ou sobrecarga pulmonar de fluídos. A hipoxemia deve estar também presente, com uma razão PF inferior ou igual a 300 mmHg. A definição de Berlim também permite a classificação da ARDS por severidade da doença. Considerando um valor de peep superior ou igual 5 cmH_2O, a ARDS poderá ser ligeira se a razão de PF for menor ou igual a 300 e maior que 200 mmHg, moderada se for menor ou igual a 200 mmHg e maior que 100 mmHg e grave se for igual ou inferior a 100 mmHg [2].

A ARDS pode ter diversas etiologias, apresentando assim sintomatologia diferenciada e não específica. No entanto a hipoxemia e dispneia são elementos comuns, surgindo horas após o a causa inicial e sendo necessário VMI na maioria destes doentes em UCI's[28].

2.1.2 Fisiopatologia

O pulmão é estruturado para facilitar a eliminação de dióxido de carbono e transferência de oxigênio através das unidades alvéolo-capilares. Em condições normais, uma camada de células endoteliais estabelece uma barreira seletiva para fluidos e solutos, permitindo a reabsorção de fluido em casos de edema alveolar [3]. Na ARDS, a permeabilidade aumentada no endotélio pulmonar resulta em edema no interstício pulmonar, muitas vezes devido a lesões endoteliais. A lesão endotelial na ARDS pode ser causada por vários fatores, como inflamação, microrganismos, lesões por aspiração, isquemia, re-perfusão e transfusões. Estes fatores levam a um aumento da permeabilidade vascular, causando hipoxemia arterial e dificuldade na eliminação de dióxido de carbono devido a um desequilíbrio na ventilação-perfusão [3]. No processo de resolução da ARDS, semelhante ao edema cardiogénico, ocorre um transporte ativo de iões através do epitélio alveolar, criando um movimento osmótico do fluido. No entanto, na ARDS, a lesão direta da barreira endotelial e epitelial afeta a capacidade de reabsorção do fluido e a remoção de células inflamatórias e citocinas no pulmão [3].

A ARDS apresenta diferentes fases. A fase exsudativa está presente, tipicamente, durante a primeira semana da doença. Esta fase é marcada por edema intersticial e intra-alveolar, podendo existir hemorragia e necroses associadas. A fase proliferativa acontece entre o dia 7 e 21 da doença onde é possível observar inflamação crónica, endarterite e necrose do parênquima. A fase fibrótica desenvolve-se a partir do dia 21 e apresenta bronquiectasia de tração e fibrose colagenizada[29,30].

2.1.3 Etiologia e fatores de risco

De acordo com a sua etiologia, a ARDS pode ser considerada primária, caso provenha de uma doença pulmonar aguda (pneumonia, contusão pulmonar, lesão por inalação, etc..) ou secundária se tiver causas indiretas (trauma, sepsies, pancreatite, etc..). A sepsies devido à pneumonia é uma causa de incidência major. O alcoolismo, tabagismo e idade associada a comorbilidade também são fatores de risco a considerar, associados com maior incidência de ARDS e taxas de mortalidade. Já foi indicado na introdução que a ARDS é uma doença complexa com subfenótipos identificados que podem ser considerados também grupos de risco [30].

2.1.4 Epidemiologia

A incidência da ARDS varia globalmente em mais de 400%, sendo bastante dependente do contexto clínico e da fisiologia dos doentes, existindo também uma sob-identificação da mesma devido à definição e metodologias utilizadas [31]. *K. Hendickson et al* refere que num estudo de dimensão considerável apenas 60% dos casos de ARDS foram identificados apropriadamente por clínicos[31]. Um estudo que teve em conta 50 países e em 459 UCI's (referido com maior qualidade de evidência da incidência da ARDS), foi realizado pela LUNG-SAFE. Neste estudo verificou-se que 10% de todos os doentes de UCI e 23% de todos os

doentes submetidos a VMI foram identificados como doentes com ARDS, para um total de 5.5 casos por cama e por ano de UCI[31].

No que toca à mortalidade a ARDS, esta é uma síndrome de risco considerável. A razão PF correlaciona-se com o prognóstico da ARDS, o que leva a comunidade médica a manter a definição de Berlim. Os autores responsáveis por esta definição verificaram que a mortalidade em ARDS ligeira foi de 34,9%, em ARDS moderada foi de 40,3% e em ARDS severa foi de 46,1% (considerando diversos estudos). Apesar de ter existido uma descida de mortalidade nos últimos anos, quando consideradas todas as severidades da ARDS a mortalidade intra-hospitalar continua alta, na ordem dos 40% (**Figura 1**) [31].

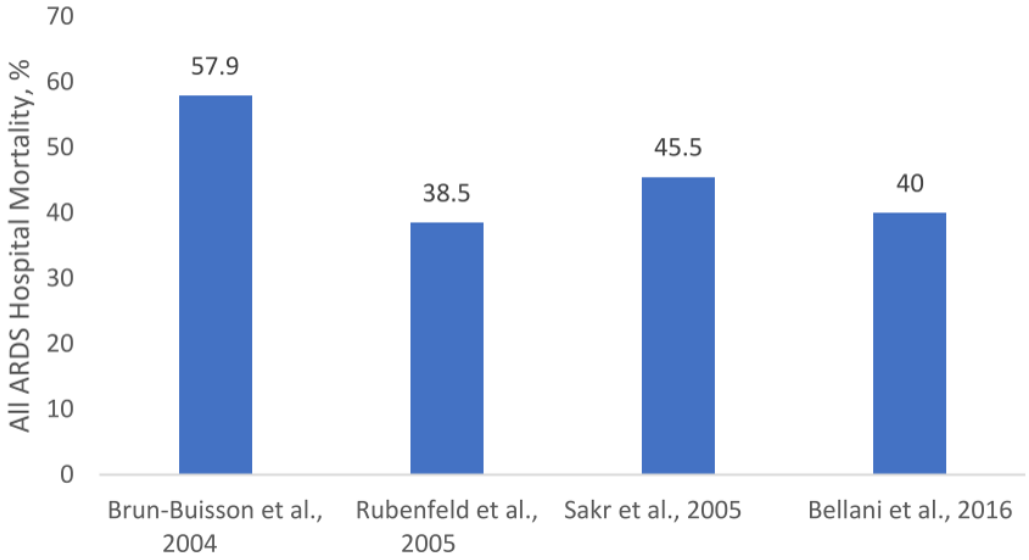


Figura 1 *- Mortalidade da ARDS ao longo dos anos, comparando vários estudos. Adaptado de K. Hendrickson (2021)*

2.1.5 ARDS-COV19

A pandemia COVID-19 foi declarada pela primeira vez em 11 de março de 2020 pela *World Health Organization*. Desde novembro desse ano o número de mortes registadas supera o valor de 1.368.000. A COVID-19 é uma doença multisistémica que pode desenvolver pneumonia viral e consequentemente ARDS. A mortalidade registada no mundo varia consideravelmente, chegando até aos 88,3% em alguns países e cidades. Deve ser considerado que os valor elevados possam ser devidos ao contexto de pandemia de uma doença até à data não bem conhecida, sem terapêuticas definidas e com limitações nos recursos humanos e hospitalares[31]. Ainda assim, a evidência atual sugere que os doentes com ARDS-COV19 apresentam pior prognóstico e maior risco de internamento em UCI [7].

A questão da ARDS-COV19 diferir das outras formas ARDS comum tem sido controversa. As preocupações iniciais sobre a rápida deterioração específica desta doença, o risco de aerossolização e transmissão da mesma aos prestadores de cuidados, a falta de terapêuticas eficazes e a perceção de alguns clínicos levaram à verificação de que alguns doentes revelaram complacência pulmonar mais alta do que o esperado para o grau de hipoxemia. A opinião de que a ARDS resultante da COVID-19 poderia ser excecional, e a frustração dos clínicos pela falta de tratamentos comprovadamente eficazes, foram por vezes associadas a

apelos para a aplicação de terapias previamente mostradas como ineficazes na ARDS em geral, e até mesmo para o uso de ventilação com volumes elevados [31].

Em 2020 houve um debate vigoroso sobre essas questões, com clínicos a defender cuidados de apoio inovadores enquanto outros acreditavam que os cuidados de apoio padrão da ARDS representavam a melhor abordagem. Embora alguns fatores de prognósticos difiram e o tempo desde o início dos sintomas da COVID-19, até a ARDS completa, seja por vezes mais longo do que a ARDS comum, a maioria das evidências até à data sugere que a ARDS-COV19 não apresenta diferenças importantes em relação à síndrome geral. O espectro de complacência pulmonar na ARDS-COVID-19 parece semelhante ao observado em estudos anteriores sobre ARDS [7,31].

A análise patológica também mostra achados semelhantes à ARDS. Portanto, a ARDS-COV19 é gerida com o conjunto de cuidados, baseados em evidências, que são à ARDS comum, incluindo estrita adesão à ventilação com baixo volume corrente, consideração da posição de decúbito ventral, peep elevada para ARDS mais grave, gestão conservadora de fluidos, avaliação protocolizada e mobilização precoce [7,31]. No entanto esta perceção poderá estar novamente a mudar.

É possível que devido à sua causa homogénea e a um fenótipo inflamatório potencialmente mais homogéneo, a ARDS-COVID-19 possa responder a terapias que falharam em ensaios clínicos com pacientes com uma gama heterogénea de causas e etiologias [7,31]. A eficácia aparente da terapia com esteroides em vários ensaios clínicos, um tratamento para o qual os ensaios na população geral de ARDS repetidamente forneceram evidências divergentes, pode ser um exemplo inicial desse fenómeno como referido no capítulo 1 (Introdução) [7,31].

O estudo e identificação de grupos de risco e/ou subfenótipos nestes casos de ARDS é assim essencial, para melhor compreensão da mesma num contexto de doença homogénea. Este fator pode ser difícil de estudar na ARDS comum com etiologias diversificadas, dificultando a terapêutica personalizada e a medicina de precisão. Os algoritmos de classificação automática, através da sua interpretação das variáveis, poderão auxiliar na descoberta deste fenómeno atual em investigação.

2.2 Radiologia convencional: radiografia do tórax

A radiografia do tórax é um exame de imagem médica amplamente utilizado no quotidiano da prática clínica. Tipicamente é o exame de imagem de primeira linha para o diagnóstico e avaliação de doenças respiratórias, cardiovasculares e sistémicas [32,33].

A radiologia convencional é a metodologia de exames que engloba todas as radiografias comuns bi-dimensionais (2D). Este método de imagem utiliza o raio-x, um tipo de radiação ionizante, com comprimento de onda entre os 0,01 e 10 nm para produzir imagens dos tecidos do corpo humano devido ao seu excelente efeito de penetração[34]. A produção destas imagens é possível devido ao efeito de atenuação do raio-x pela matéria, através da atenuação por efeito fotoelétrico, dispersão por efeito de *Compton* e de *Rayleigh* ou simples transmissão dos fotões. O processo de atenuação é governado pela lei *Lambert-Beer em que*[34]*:*

$$I = I_0 e^{-\mu d}$$

Equação 1 - Lei de Lambert-Beer

I = Intensidade dos fotões de raio-x transmitidos.

I_0 = Intensidade inicial dos fotões de raio-x.

μ = Coeficiente de atenuação linear.

d = Espessura do tecido atenuador.

Os efeitos físicos de interação com a matéria dominantes irão depender da energia do fotão incidente e do número atómico da matéria atenuante (μ). No caso de fotões de baixa energia, o efeito predominante encontra-se na absorção por efeito fotoelétrico, enquanto altas energias de fotões em materiais com número atómico baixo, o efeito predominante é a dispersão de Compton[34]. Quanto maior o número atómico do material mais atenuante o mesmo é, necessitando de menos espessura para absorver um feixe de radiação de uma determinada energia. Todos estes efeitos traduzem-se numa imagem radiográfica digital em níveis de cinzento com valores de pixel entre 0 e 255, onde (dependendo sempre da energia do feixe de raio-x e das características dos tecidos) regiões mais atenuantes e com maior densidade, como o osso, tendem a ser representadas a branco (valor de pixel = 255), enquanto regiões pouco atenuantes e de baixa densidade, como o ar, tendem a ser representadas a preto (valor de pixel = 0) [32,34]. Uma imagem digital bidimensional é uma representação matricial de um objeto do tipo f(x,y,z), referindo-se assim às coordenadas do valor do seu componente primário, o pixel. O valor de x identifica a coluna, enquanto o y identifica a linha. O valor de z é igual a 1 quando abordamos uma imagem em níveis de cinzento, ou varia entre 1 a 3 quando referimos uma imagem a cores, representadas tipicamente pelo formato RGB (*red, green, blue)*.

Considerando estes fatores, é possível verificar uma típica R-TRX na **Figura 2** [35], com a sua principal anatomia descrita e a sua representação matricial.

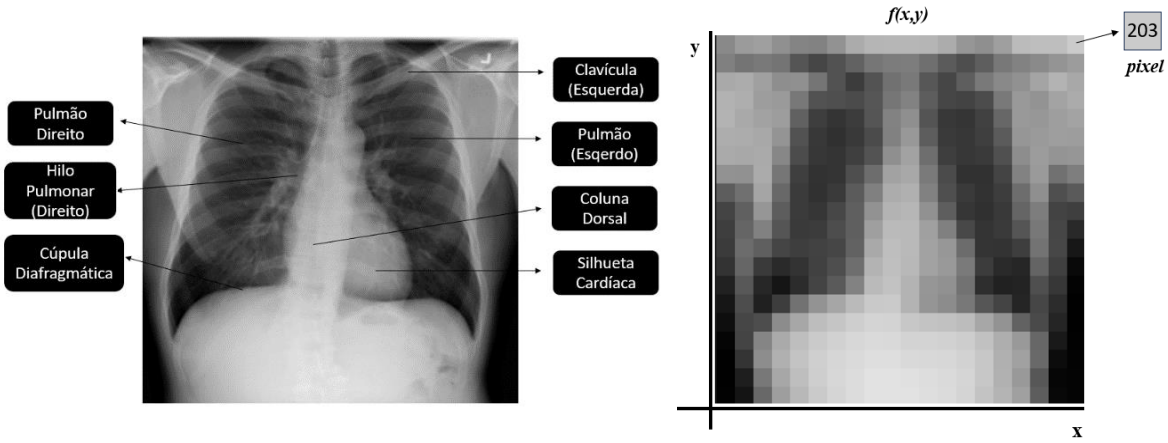


Figura 2 - *Radiografia Torácica e a sua representação matricial. Fonte: A.Tahir et al. (2021)*

A aquisição de uma radiografia depende de um gerador de raio-x e de um detetor da radiação para produção de imagens (**Figura 3**). O gerador é composto por 2 elétrodos (ânodo e o cátodo) numa câmara de vácuo. Nessa câmara, o cátodo produz eletrões altamente energéticos por efeito termiónico dos seus filamentos de tungsténio, criando corrente elétrica[34]. Os eletrões, por via de tensão elétrica aplicada, são então acelerados em direção ao ânodo onde são formados os fotões de raio-x por efeito de *bremsstrahlung* e de radiação característica do material do ânodo. A corrente aplicada no cátodo, a tensão de aceleração dos eletrões e o material do ânodo, são assim fatores determinísticos das características espetrais dos fotões de raio-x emitidos[34]. A manipulação destes parâmetros, por parte operador, irá afetar a qualidade da imagem produzida, sendo o mesmo discutido na secção 2.2.1.

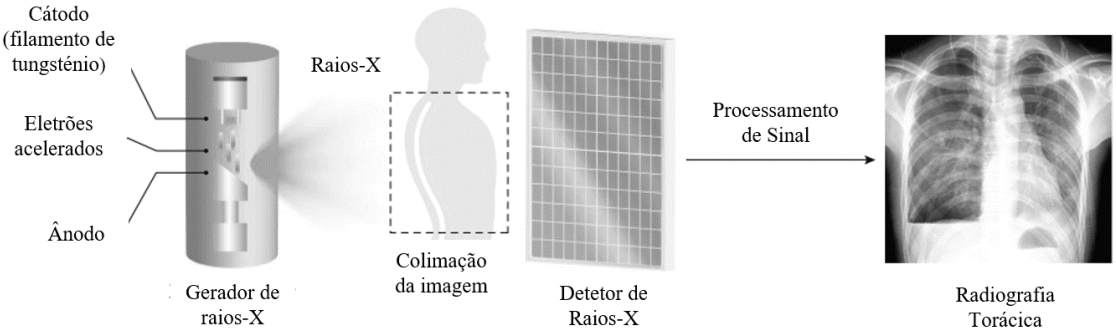


Figura 3 - *Descrição de um sistema de radiologia convencional. Adaptado de X. Ou* et al. *(2021)*

Após a emissão dos fotões de raio-x é necessário um detetor apropriado diretamente posterior ao objeto a radiografar convertendo o mesmo numa imagem através de diferentes métodos de processamento de sinal e tipologia de detetores[34]. A radiologia convencional apresentou diversas metodologias de conversão do sinal energético para luz visível ao longo dos anos, desde a radiografia analógica de filme, a radiografia computorizada e por fim a radiografia digital. Esta última é a mais comum nos tempos correntes e permite a conversão direta ou indireta dos fotões de raio-x em cargas elétricas, utilizando matrizes de componentes fotoelétricos, para consequente leitura da imagem. Este método é altamente eficiente na sua

taxa de conversão de fotões de raio-x, permitindo assim imagens com baixas doses de absorção de radiação e em tempo real[34]. Os detetores de conversão digital indireta apresentam uma camada inicial de filme cintilante (tipicamente de CsI) adjacente a uma matriz de fotodiodos de silício amorfo que produzem as cargas elétricas lidas pela matriz de transístores de filme fino (TFT) [34]. A metodologia de conversão direta, por outro lado, não utiliza uma camada de filme cintilante para criação de luminescência. O detetor da mesma é composto por uma camada fotocondutora de selénio amorfo, que converte os fotões de raio-x em corrente elétrica proporcional[34]. Os pares de buraco-eletrão criados movimentam-se de forma paralela e linear com difusão lateral limitada, permitindo assim separação eficiente de cada pixel correspondente na imagem final. Elétrodos positivos recolhem estes pares e as cargas armazenadas em condensadores são então lidas por uma matriz de TFT. Este método permite imagens com maior resolução espacial e taxas de conversão energética do que os detetores digitais indiretos[34].

A radiologia digital veio também facilitar a criação de equipamentos radiográficos portáteis de alta qualidade, utilizados diariamente na aquisição de radiografias em doentes intransportáveis, como por exemplo os doentes internados em UCI.

2.2.1 Técnicas e parâmetros de aquisição

Apesar da aparente simplicidade da técnica de radiologia convencional, a seleção dos métodos e parâmetros de aquisição da imagem podem ter um forte impacto na variabilidade qualitativa da R-TRX. Esta pode assim depender da direção e qualidade do feixe do raio-x, da distância da fonte-doente/fonte-detetor, da posição do doente e da exposição do raio-x[33,36].

No que toca à direção e distância da emissão do feixe de raio-x, este pode incidir de forma póstero-anterior (PA) ou ântero-posterior (AP), esquematizado na **Figura 4**. O primeiro método (PA) é o mais utilizado na clínica diária, onde o doente encontra-se em ortostatismo e apneia inspiratória, com a fonte de raio-x (gerador) a 1 metro e 80 centímetros do detetor[33,36].

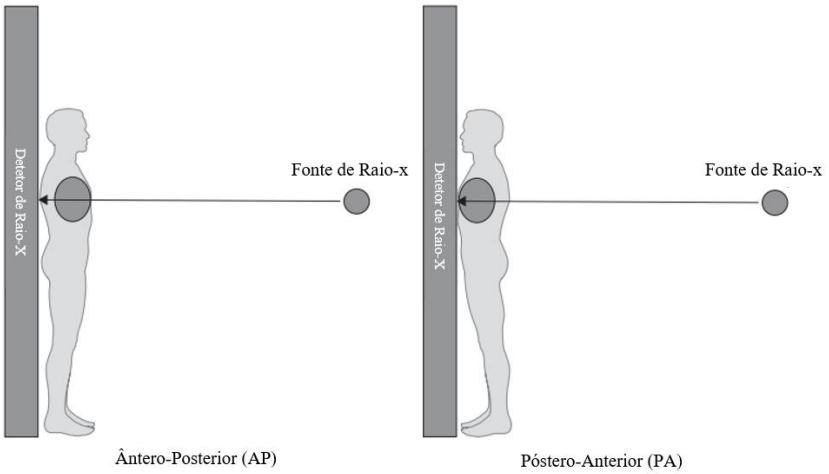


Figura 4 - *Esquematização das possíveis direções do feixe de raio -x (AP e PA)* – Adaptado de *J.Border (2011)*

Neste caso, a superfície anterior do tórax está em contacto com o detetor de raio-x, com incidência posterior do feixe. O segundo método (AP) é utilizado, tipicamente, na aquisição de R-TRX a doentes intransportáveis com equipamento portáteis. Neste caso os doentes encontram-se em decúbito dorsal e em respiração livre (caso não consigam efetuar a apneia inspiratória), com a superfície posterior do tórax em contacto com o detetor, incidindo o feixe anteriormente. A distância fonte-detetor, nesta metodologia, é tipicamente inferior à recomendada de 1 metro e 80 centímetros por questões de logística do espaço dos internamentos. A presença de artefactos metálicos altamente atenuantes é também mais comum, devido aos dispositivos médicos de avaliação e terapêutica presentes neste contexto clínico[33,36].

A direção do feixe e a distância da fonte-detetor / objeto-detetor importam devido ao efeito de magnificação. Considerando uma distância fonte-detetor fixa, quanto mais longe o objeto encontra-se do detetor, mais magnificado o mesmo estará na imagem final perdendo nitidez. Consequentemente ao aumentar a distância do da fonte-detetor, o objeto recupera nitidez e aproxima-se da sua dimensão real. Desta forma e para uma correta avaliação, é recomendado que o objeto a estudar esteja o mais próximo possível do detetor. No caso de uma R-TRX AP, o coração e o mediastino estão a uma maior distância do detetor, o que irá provocar magnificação dos mesmos. Este facto poderá interferir na avaliação médica das dimensões do coração, sobrevalorizando as mesmas para um possível diagnóstico de cardiomegália como poder ser observado na **Figura 5** [33,36].

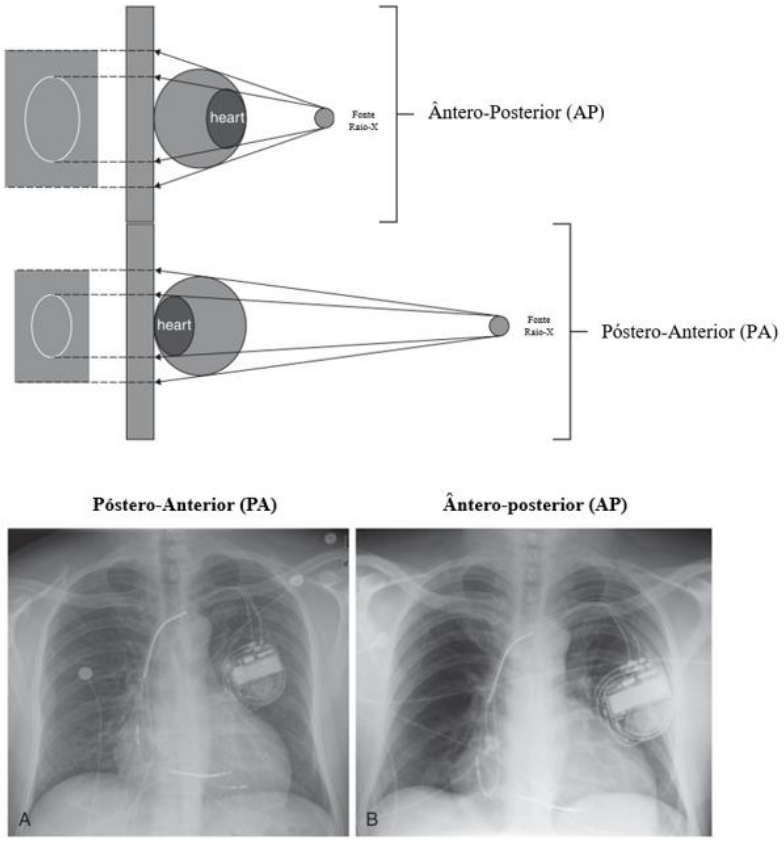


Figura 5 - *Exemplo do efeito de magnificação do coração na incidência AP. Podemos verificar que a R-RTX B apresenta maior dimensão da silhueta cardíaca, cobrindo também parte do parênquima pulmonar esquerdo. Adaptado de J. Border (2011)*

Por sua vez a aquisição em respiração livre pode também simular vasculatura pulmonar proeminente e edema intersticial. No que diz respeito à posição efetiva do doente durante a aquisição, o ortostatismo perpendicular ao detetor é preferível. Neste caso asseguramos igual magnificação das estruturas torácicas, ao contrário de uma aquisição em leito (decúbito dorsal) onde os tórax superior e inferior poderão estar a diferentes distâncias do detetor. Outra preocupação deve-se à avaliação dos derrames pulmonares. Quando existe fluídos nas cavidas torácicas, o efeito gravitacional de um doente em ortostatismo, irá acumular os mesmos na região pulmonar inferior, criando assim níveis (linhas) hidroaéreos claramente definidos. No entanto quando observamos o mesmo infiltrado pulmonar num doente em decúbito dorsal, o derrame poderá cobrir o restante parênquima pulmonar, criando um aparente efeito de aumento de densidade difusa prevenindo o correto diagnóstico (**Figura 6**)[33].

Figura 6 - *Esquema da representação de um derrame pleural num R-TRX em ortostatismo. Adaptado de: J. Broder (2011)*

A exposição da radiação é outro fator importante a considerar, dependendo a mesma do biótipo do doente e dos parâmetros técnicos de aquisição. Estes parâmetros são a tensão, definida em kV (Kilo-Volts), e a corrente, definida em mAs (miliamperes por segundo). A tensão irá controlar a quantidade e o potencial energético dos fotões de raio-x emitidos, afetando assim a transmissão, ou capacidade de "penetração" dos mesmos, nos tecidos do doente[37]. A corrente irá afetar a quantidade de fotões emitidos e que potencialmente chegarão ao detetor. Desta forma a tensão é essencial na definição do contraste final da imagem. O contraste de uma estrutura ou lesão, pode ser definida pela diferencial de intensidade entre a mesma e as estruturas envolventes. Quanto maior o diferencial maior o contraste[37]. Este fator é independente da quantidade de fotões emitidos e dependente do potencial de penetração dos mesmos a uma dada estrutura. Isto significa que para valores de tensão superiores, mais fotões são transmitidos pelos tecidos e chegam ao detetor de radiação, resultando numa imagem de tonalidade escura reduzindo o contraste final da imagem[37]. Quando o inverso acontece, mais fotões são absorvidos pelos tecidos e menos chegam ao detetor, resultando numa imagem de tonalidade mais clara para as estruturas absorventes, aumentando o contraste final da imagem. Considerando, no entanto, a corrente, esta apenas vai influenciar a potencial quantidade de fotões processados pelo detetor, afetando apenas a razão sinal/ruído da imagem final. Quanto mais fotões são detetados mais sinal é obtido, reduzindo consequentemente o ruído[37]. O R-TRX é uma técnica de contraste de longa escala onde é

utilizada alta tensão para uma maior transmissão. Neste caso temos uma aparente redução de contraste com menor diferenciação entre os tecidos *major*, no entanto obtemos uma maior *range* de níveis de cinzentos com presença de diferenças de densidade subtis e uniformes na região pulmonar. Esta técnica é essencial na avaliação deste tecido. Para contrabalançar o referido, esta técnica utiliza valores de corrente baixas, reduzindo os fotões emitidos e a tonalidade escura típica de uma imagem resultante da sobre-exposição à radiação. O aumento de 15% da tensão, corresponde ao aumento de 50% da corrente em termos de qualidade e exposição da imagem final[37].

Podemos então concluir que devido, a todas estas características da R-TRX, a uniformização e identificação da base de dados de acordo com as técnica e parâmetros de aquisição é essencial. Em caso de exemplo, se um algoritmo for treinado para a deteção de cardiomegalia apenas utilizando radiografias de rotina PA em ortostatismo, pode sobrestimar a dimensão da silhueta cardíaca classificando imagens adquiridas em AP e decúbito dorsal. É assim importante a criação de algoritmos altamente generalizados para os diferentes contextos clínicos ou, em caso de poucas amostras de dados, criar modelos de classificação específicos para os diferentes contextos clínicos com a sua população claramente descrita. No caso das radiografias portáteis em UCI, estas podem apresentar uma grande variabilidade de parâmeros de aquisição e de distância da fonte-detetor, sendo assim é essencial o pré-processamento das mesmas para otimização do contraste, redução do ruído e segmentação da região de interesse para potencial remoção dos artefactos metálicos.

2.2.2 Radiografias torácicas no contexto da ARDS

A R-RTX está integralmente ligada à ARDS. A mesma encontra-se presente na definição de Berlim clinicamente aceite. A ARDS depende atualmente da presença bilateral de infiltrados pulmonares envolvendo dois ou mais quadrantes na R-RTX frontal. Quando utilizada a tomografia computorizada como referência, autores revelam que a R-RTX apresenta uma sensibilidade de 68%-74% e uma especificidade de 73-74% na deteção destas anormalidades pulmonares resultantes da ARDS. Estas limitações do diagnóstico podem dever-se à baixa qualidade das imagens, com pouca capacidade de diagnóstico diferencial, e à própria avaliação qualitativa e subjetiva das mesmas. A própria definição de Berlim assume estas limitações e recomenda a criação de novos métodos para o esclarecimento de critérios de diagnóstico para as R-TRX[14,38].

No entanto as vantagens da utilização desta técnica de imagem tornaram-na assim indispensável na rotina clínica global das UCI. A R-TRX apresenta baixos custos económicos, é portátil, é sensível à deteção de patologia pulmonar (como o derrame pleural ou pneumotórax) e é fundamental na monotorização e colocação de diferentes dispositivos médicos[14,38]. Na ARDS a interpretação radiológica varia conforme a evolução patológica e fisiológica da doença. Nos primeiros dois dias da doença é normal não existirem alterações nas R-TRX, no entanto durante a primeira semana os infiltrados pulmonares progridem para consolidações bilaterais (tipicamente em mais de 3 lobos), podendo até apresentar uma densificação total do pulmão em casos de mais severidade da doença[14,38]. Sinais de fibrose pulmonar podem ser observados na fase final da doença acompanhadas de severa disfunção respiratória. A avaliação das R-RTX, para verificação da severidade e progressão da ARDS, não é assim passível de uma avaliação única e isolada temporalmente. Uma avaliação

longitudinal durante o internamento é assim necessária, visto que, é estimado que em cerca de 65% dos exames executados diariamente nas UCI's, são encontrados novos achados imagiológicos que afetam as decisões clínicas sobre o presente doente. Podemos observar a típica representação da ARDS, numa R-TRX, e a sua evolução ao longo dos dias de internamento na **Figura 7**[14,38].

Nos últimos anos tem existido assim um crescente interesse na avaliação quantitativa das R-RTX para avaliação da ARDS e do seu prognóstico. Uma das métricas com maior sucesso é a classificação do *score* de RALE desenvolvido por *M.Warren* e colegas em 2018 [16]. Este score divide cada pulmão em dois quadrantes, para um total de 4 quadrantes. Cada um destes

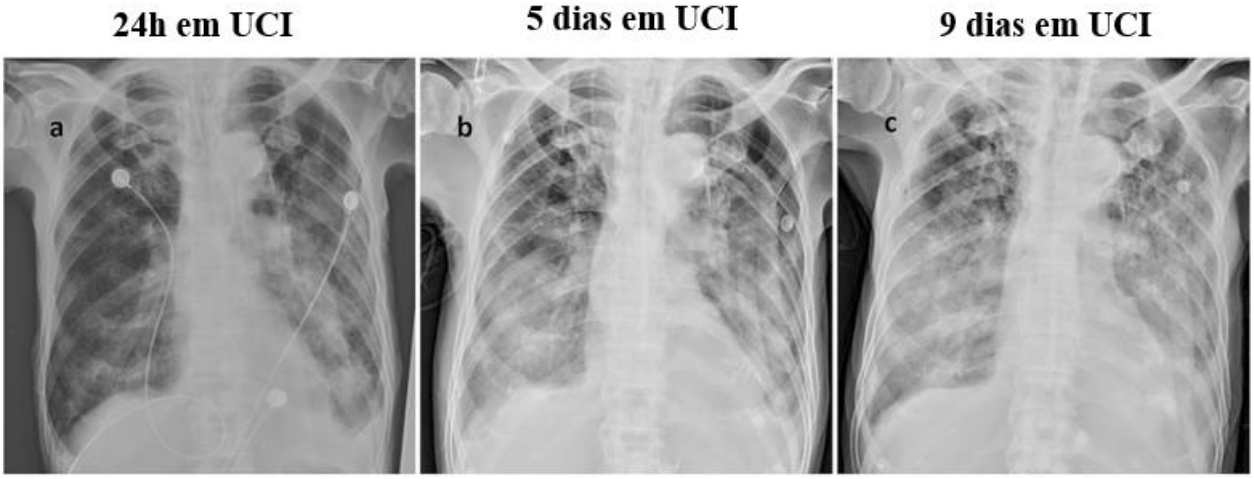


Figura 7 - *Progressão radiográfica da ARDS ao longo de 9 dias. Podemos verificar na imagem a) a presença de infiltrados bilaterais com apagamento das bases pulmonares (nos seios costofrénicos) e progressão da infiltração ao longo dos dias na imagem b) e c). Adaptado de: S. Huang et al. (2022)*

quadrantes é avaliado em termos de consolidação, de 0 a 4, (ocupação total do quadrante pelos infiltrados pulmonares) e em termos da densidade (atenuação dos infiltrados), de 0 a 3. Os scores de consolidação e densidade são então multiplicados em cada quadrante e o somatório de todos os quadrantes é realizado para obtenção de um score final. Valores de RALE superiores estão relacionados, de forma independente, com razões PF mais baixas (maior severidade de ARDS) e com piores taxas de sobrevivência. A variação do RALE, ao longo dos primeiros dias de internamento, está também associado à mortalidade ao fim de 90 dias após internamento[16]. Apesar do score de RALE ser promissor, o mesmo é um método semi-quantitativo e pode sofrer de variabilidades inter e intra-operador. A necessidade de uma avaliação por parte de um profissional treinado, como um radiologista, poderá também ser dispendioso em termos temporais e de logística de trabalho visto que é um exame de rotina realizado em média 0,7 vezes por dia em cada doente de UCI. Os algoritmos de *radiomics* e *deep learning* tem apresentado uma alternativa para obtenção de um método quantitativo de avaliação de RX-TRX na identificação de lesões pulmonares, ou na categorização do prognóstico do doente. Alguns destes estudos no contexto da ARDS e da COVID-19 irão ser descritos nos capítulos posteriores.

2.3 Machine learning em imagem médica: *Radiomics*

O conceito de *machine learning* (ML) refere-se a uma área de saber com foco no desenvolvimento e interpretação de modelos e algoritmos de aprendizagem automática no ramo da IA. Estes modelos têm como objetivo a realização de um determinado processo através de previsões e decisões que não são diretamente instruídas por lógica computacional tradicional. O algoritmo aprende padrões estatísticos e melhora o seu desempenho tendo em conta amostras de eventos anteriores. Estes eventos são tipicamente designados de "dados de treino". O algoritmo utiliza estas amostras para prever um *output* Y (*Target)* considerando novos *inputs* X não analisados pelo modelo anteriormente. Numa tarefa de classificação típica de imagem médica estes *outputs* são definidos como classes patológicas, biológicas ou de prognóstico de doente. O "treino" é referido como aprendizagem automática, e pode apresentar diferentes tipologias que irão ser exploradas na secção seguinte (2.4). Apesar do conceito ser utilizado há mais de 60 anos, apenas recentemente existiu um grande impulso na área. O aparecimento de *Deep Learning,* permitiu desenvolvimento de novo *hardware* e a facilidade de acesso à informação e a grande quantidades de dados permitiu que a implementação desta técnica para resolver diversos problemas em várias áreas[39].

No contexto da medicina, a radiologia obteve destaque no que toca a aplicações de IA e ML. A necessidade de classificar e segmentar imagens de exames complementares de diagnóstico (como a tomografia computorizada, a ressonância magnética e a radiografia convencional) em termos patológicos e de prognóstico do doente tornou-se uma realidade atual. Estas aplicações poderão apoiar a decisão médica trazendo ferramentas que aumentam a clareza, objetividade e quantificação de resultados a um ramo onde a avaliação qualitativa e semi-quantitativa é dominante. Apesar da alta capacidade visual de reconhecimento de padrões do ser humano, existe dificuldade na análise quantitativa complexa. Consequentemente, estas aplicações aumentam a possibilidade de reduzir a variabilidade inter-operador e intra-operador e evitar técnicas invasivas e mais dispendiosas como biópsias e cirurgias para o diagnóstico completo do doente[40]. No final do século XX ocorreu a transição da imagem médica analógica para digital, o que permitiu a análise computacional quantitativa de dados médicos através de métodos estatísticos clássicos (no início dos anos 60) para auxílio do diagnóstico. Surge assim o termo de *Computer Aided Diagnosis* (CAD), que à data era limitado, devido aos métodos probabilísticos utilizados e à pouca disponibilidade de dados. Só no final nos anos 80, através do desenvolvimento de algoritmos de ML e de reconhecimento de padrões, foi possível obter sistemas de CAD clinicamente viáveis[40]. O facto dos dados de imagem serem facilmente acessíveis e organizados, através dos servidores de armazenamento de PACS, (*Picture Archiving and Communication System*), também facilita a implementação e treino dos diferentes modelos de ML[41]. Nas últimas décadas, têm sido adotados com sucesso métodos quantitativos simples de análise de imagem baseados em observações qualitativas de radiografias, TC e RM como é o exemplo do score de RALE referido na introdução (1.1). Estes métodos têm ajudado os clínicos a reconhecer anormalidades, principalmente na deteção e classificação de lesões pulmonares e mamárias, abrindo a possibilidade de modelos puramente quantitativos serem vantajosos e clinicamente viáveis[40].

Uma das principais dificuldades na interpretação quantitativa, funcional e do prognóstico das imagens de RC e TC advém do facto de que o valor do pixel (2D no caso da RC) e do voxel

(3D no caso da TC) não representarem diretamente as propriedades biológicas, bioquímicas e fisiológicas das anormalidades que estão a ser observadas. Apesar de alguns métodos por ressonância magnética e TC espetral semi-paramétrica permitirem a obtenção de informação fisiológica e bioquímica das lesões, as técnicas de imagem mais comuns apenas permitem a dedução das mesmas através das propriedades morfológicas da lesão e do seu comportamento de atenuação aos raio-x (dependendo sempre dos coeficientes de atenuação e das propriedades do feixe de radiação emitido). Deste forma o défice de informação quantitativa (sem ser utilizado todo potencial das imagens clínicas), pode resultar na maior utilização de métodos de diagnóstico invasivos ou de exames de seguimento desnecessários [40]. Apesar das aparentes vantagens, o desenvolvimento de técnicas de quantificação e classificação automática das imagens médicas é desafiante devido à ausência de um *standard* universal para aquisição e reconstrução das mesmas, o que resulta em valores de pixel/voxel não reprodutíveis por população e patologia. Valores estes que seriam essenciais para o processamento, avaliação e estatística quantitativa. O desenvolvimento computacional permitiu a criação de algoritmos mais complexos e facilidade de análise de grandes quantidades de dados. Desta forma a barreira imposta pela não reprodutibilidade começou a ser ultrapassada, dependendo não só da análise dos valores de pixel/voxel da matriz, mas também das suas relações estatísticas com os pixéis/voxeis vizinhos, o que irá traduzir a corelação patológica [40].

Surge assim em 2012 o termo *radiomics*, que pode ser descrito como um processo que envolve extração de características de imagens médicas quantificáveis com o objetivo de obter classificações ou categorizações fenotípicas correlacionadas com a biologia, patologia, ou prognóstico de doentes, utilizando métodos avançados de *machine learning*[40,42]. Os *Radiomics* podem ainda ser divididos em pelo menos 2 metodologias distintas. O *radiomics hand-crafted* e o *radiomics* baseado em *Deep Learning*. O primeiro depende da extração de características referidas como *hand-crafted*. Estas características são provenientes de fórmulas matemáticas pré-concebidas que obtém informação da região de interesse da imagem através do seu histograma, da sua dimensão/forma e da sua textura de imagem. Posteriormente estas características são selecionadas e sofrem pré-processamento e normalização para servirem de *inputs* vetoriais em algoritmos de classificação de *machine learning* clássicos, como por exemplo a regressão logística. A segunda metodologia, que servirá de base para esta investigação, obtém características de imagem abstratas de forma automática, utilizando redes neuronais complexas e “afinadas” para a tarefa de classificação presente. [40,42]

Os métodos e os fluxos de trabalho das presentes metodologias irão ser descritos com maior detalhe nos capítulos seguintes, assim como as possíveis metodologias híbridas que poderão existir.

2.4 *Machine learning* supervisionado e não supervisionado

Os sistemas de classificação médicos baseados em *machine learning* podem ser definidos de acordo com a sua metodologia na fase de aprendizagem.

O mais comum destes métodos na literatura para aplicações da saúde e imagem médica é o treino supervisionado e será esta a metodologia utilizada na presente investigação [39,42]. No contexto das radiografias, este treino implica a existência de dados e imagens previamente categorizados com a sua classe de interesse para o *output* do algoritmo. Este processo denomina-se de "*data labelling"* e é essencial para o sistema de aprendizagem automática mapear a informação extraída dos pares vetoriais (*input-target),* a uma das classes de interesse pré-definidas, podendo estas serem representadas por um valor numérico ou vetorial. Após esta fase de aprendizagem, o algoritmo conseguirá classificar novos dados de *input* previamente não identificados. Podemos então aferir que estas categorizações servem também para a avaliação do próprio modelo de classificação, servindo de *grount-truth labels* (classe verdadeira previamente conhecida para comparação estatística com a classe prevista pelo modelo). Consequentemente, esta metodologia de classificação requer a experiência de radiologistas e outros especialistas médicos treinados para a correta identificação do prognóstico e/ou patologias presentes nas imagens de diagnóstico[39,43,44].

A aprendizagem não-supervisionada diverge da primeira metodologia por não necessitar de categorização prévia dos dados. Este método extrai padrões das características dos dados e deduz classes/subgrupos de acordo com as suas similaridades. O modelo mais utilizado para esta metodologia de aprendizagem é a análise por *k-means clustering* que agrupa os elementos dos *inputs* de acordo as distâncias euclidianas das suas características. Na classificação de novos dados o algoritmo identifica a presença das similaridades inferidas durante a aprendizagem. Posteriormente podemos analisar estatisticamente os grupos e obter possíveis correlações patológicas ou de prognóstico do doente em aplicações da saúde. As técnicas não supervisionada podem também ser úteis na redução da dimensionalidade dos dados. Exemplos do referido encontram-se em modelos como a *principal component analysis*, que utiliza transformações linear e não-lineares para redução do espaço dimensional das características e evitar viés (*bias)* do modelo de classificação para os dados de treino que não generalizam bem. [39,41–43,45]

Metodologias híbridas entre os dois métodos de aprendizagem podem também ser utilizadas em situações de poucos dados categorizados e muitos dados não categorizados. Nesta abordagem, denominada de aprendizagem semi-supervisionada, é realizado um mapeamento das características sem categorização às classes existentes dos dados que previamente categorizados. Posteriormente estes são utilizados num modelo de classificação supervisionado anteriormente descrito. Esta metodologia tem tido sucesso comprovado em diversos modelos de classificação. [39,41–43,45]

Apesar do grande potencial da aprendizagem não-supervisionada, a sua aplicação tem sido limitada quando comparada aos resultados obtidos através da aprendizagem supervisionada. Este facto deve-se, em parte, à grande quantidade de dados necessários à sua aplicação na saúde. O universo do *Deep Learning* tem vindo a reformular esta perspetiva. A utilização de *Autoencoders* é fundamental em arquiteturas não-supervisionadas de DL. Estas redes neuronais usam camadas escondidas progressivamente de menores dimensões, que em

conjunto com a regularização e redução das disparidades evita a aprendizagem de soluções triviais ao problema. O método de aprendizagem do ser humano é por si mesmo, em grande parte, não supervisionado. Assimilamos a semelhança de objetos ou conceitos sem categorizações previamente estabelecidas[39,42]. Esta técnica poderá ser assim o futuro da IA e, no caso em concreto da presente investigação, da identificação de sub-fénotipos de uma determinada patologia ou condição biológica como a RDSA.1

2.5 Escolha, seleção e amostragem dos dados

2.5.1 Seleção do *dataset*

A escolha da base de dados é a primeira fase de qualquer fluxo de trabalho para uma tarefa de classificação de *machine learning* em imagem médica. Num estudo de coorte típico é necessário identificar dentro de uma população, indivíduos/doentes com características em comum. Estas poderão ser, por exemplo, biométricas (peso, idade, género, etc.…), fenotípicas (expressão genética), patológicas ou *standards* de referência (resultados de exames/procedimentos de diagnóstico considerados *gold standard*.). A seleção de doentes poderá ser feita *a priori*, através de um conjunto de critérios de inclusão pré-definidos, ou à posteriori, onde os critérios de inclusão são definidos de acordo com a amostra disponível para recolha de dados. [43]

A seleção dos doentes por análise de informação clínica poderá resultar numa maior amostra, no entanto deve existir precaução no que toca ao viés (*bias)* de confirmação e/ou de verificação. No primeiro tipo de viés, existe a hipótese de seleção de doentes por pressupostos de características típicas que confirmam ou não o diagnóstico, sem existir um standard de referência para os mesmos. No segundo viés existe o risco de incluir doentes considerados patologicamente positivos ou negativos através de diversos métodos de diagnóstico, sem confirmação por métodos *gold standard*. Outra metodologia viável é a seleção de doentes por resultados imagiológicos prévios. Este método promove a seleção, pois garante que a maioria da população selecionada apresenta exames imagiológicos a serem utilizados para o treino do modelo de classificação. No entanto deve-se garantir que os doentes com imagens de baixa qualidade ou com artefactos são removidos, assim como doentes com exames prévios à instituição integrante da base de dados original. Finalmente existe a possibilidade de selecionar doentes através diagnósticos por interpretação histopatológica tecidular de biópsias ou outros procedimentos clínicos considerados *gold standard*. Este método apresenta a vantagem de obtenção de dados com um *ground truth label* seguro e confiável, no entanto, resulta em amostras de doentes menores com suscetibilidade adicional a variações interoperador da própria análise médica [43,46].

Tendo em consideração estes fatores, o melhor método a utilizar irá sempre depender da população disponível e de tipo de tarefa que queremos aplicar. No caso concreto desta investigação, para previsão da mortalidade em doentes COVID-19 com ARDS, o último método referido será o mais pertinente, visto que é necessário garantir a presença da doença na população com os *standards* de referência referidos na secção 2.1.

2.5.2 Desidentificação, exploração e limpeza dos dados

Um ponto importante a considerar durante os processos de investigação dependentes de imagens médicas, reside na proteção de dados dos doentes. Garantir a privacidade da informação dos doentes é essencial, existindo na literatura diversas metodologias usualmente utilizadas para este efeito. O ato de "esconder" ou eliminar as informações médicas revelantes denomina-se de desidentificação e ajuda a minimizar os riscos de identificação do doente durante o processo de *radiomics*.[43] A anonimização é um processo de remoção total e irreversível das informações pessoais identificáveis de uma base de dados, enquanto a pseudo-anonimização procura esconder essa informação, encriptando a mesma com valores numéricos aleatórios que podem ser acedidos por uma chave de encriptação. O processo de pseudo-anonimização também pode realizado através da utilização de um valor numérico (*index)* representativo do doente substituindo o número de processo hospitalar ou o seu nome [47]. No que toca às imagens médicas estas são tipicamente armazenadas em servidores locais. Estes servidores são denominados de *Picture and Archiving Communication System* e as imagens são armazenadas em formato DICOM (*Digital Imaging and Communications in medicine)*. Este formato apresenta meta-dados de informação adicional à imagem que necessitam de anonimização. Os dados são codificados em *headers* que apresentam informação pessoal direta (nome do doente, a sua data de nascimento e o seu número de processo hospitalar) e indireta (dia da aquisição da imagem, nome do operador, etc.), podendo os mesmos ser representados na imagem sobre valores de pixel. Existem *softwares open-source* de anonimização de dados DICOM, no entanto durante a exportação dos mesmos, a maioria dos PACS permite a desidentificação direta dos mesmos [43,46]. As imagens são tipicamente exportadas em outros formatos de imagem mais comuns que removem os meta-dados. São exemplos destes formatos o PNG e TIFF, que usam compressão *lossless* para garantir a qualidade da imagem através do armazenamento da informação original da mesma [46]. Este facto simplifica o pré-processamento das imagens sem perda da definição original da imagem e mantendo os seus contornos.

A correta limpeza de dados é outro processo inicial importante uma tarefa de *radiomics* e *machine learning* no geral. A qualidade das imagens extraídas é essencial para obter resultados de confiança, sendo assim necessária uma verificação rigorosa das mesmas. Este problema é ainda mais relevante se forem considerados grandes volumes de dados retirados de diferentes centros clínicos, visto que, irá existir uma maior heterogeneidade das imagens devido a diferentes equipamentos e protocolos de aquisição. É necessário recolher assim as imagens relevantes para o problema e excluir aquelas que poderão servir de *outliers* (amostras de dados que podem introduzir ruído na base de dados e treino do algoritmo, divergindo significativamente das restantes observações). No caso desta investigação, R-RTX com artefactos metálicos sobre a região de interesse, imagens ruidosas, imagens sub e sobre-expostas à radiação ou imagens com anatomia relevante, fora da área de colimação ("cortadas"), devem ser removidas **Figura 8**.

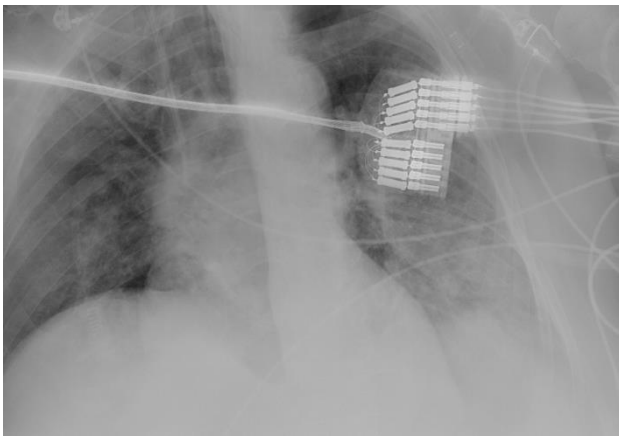


***Figura 8**- Exemplo de uma R-TRX recolhida de baixa qualidade com o parênquima pulmonar esquerdo fora da área de colimação e com dispostivos médicos a cobrir o parêquima pulmonar direito.*

A exploração dos dados iniciais e das suas variáveis (*features*) deve ser também efetuada através da sua visualização ou quantificação (estatística descritiva univariada e multivariada). Este fator irá permitir verificar as tendências globais, verificar possíveis variáveis de interesse para o problema presente e descobrir *outliers*. Desta forma é possível perceber se temos uma base de dados generalizável ou pouco equilibrada, em termos de classes e característica fenotípicas, o que poderá causar viés no modelo de classificação final. Na avaliação das variáveis discretas, contínuas ou categóricas podemos também descobrir possíveis *outliers* ou amostras de dados em falta.

2.5.3 Técnicas de Amostragem da base de dados

A amostragem de uma base de dados em *machine learning* depende da divisão da mesma em diversos sub-grupos para o propósito de treino e validação do modelo de classificação. É recomendada que esta divisão seja aleatória, mas estratificada, mantendo assim as distribuições de classes da base de dados inicial para maior generalidade (principalmente se existir um desequilíbrio entre as classes).[43]

A técnica de amostragem mais comum passa pela criação a criação de um grupo de treino, utilizado para a aprendizagem do algoritmo, um grupo de validação, usado para definição das metodologias de seleção de *features*, afinação e calibração dos hiperparâmetros do modelo, e um grupo de teste, não processado anteriormente pelo modelo, para avaliação de generalidade final (**Figura 9**) [41,43]. A proporção recomendada para a divisão destes grupos é de 80:10:10 respetivamente.

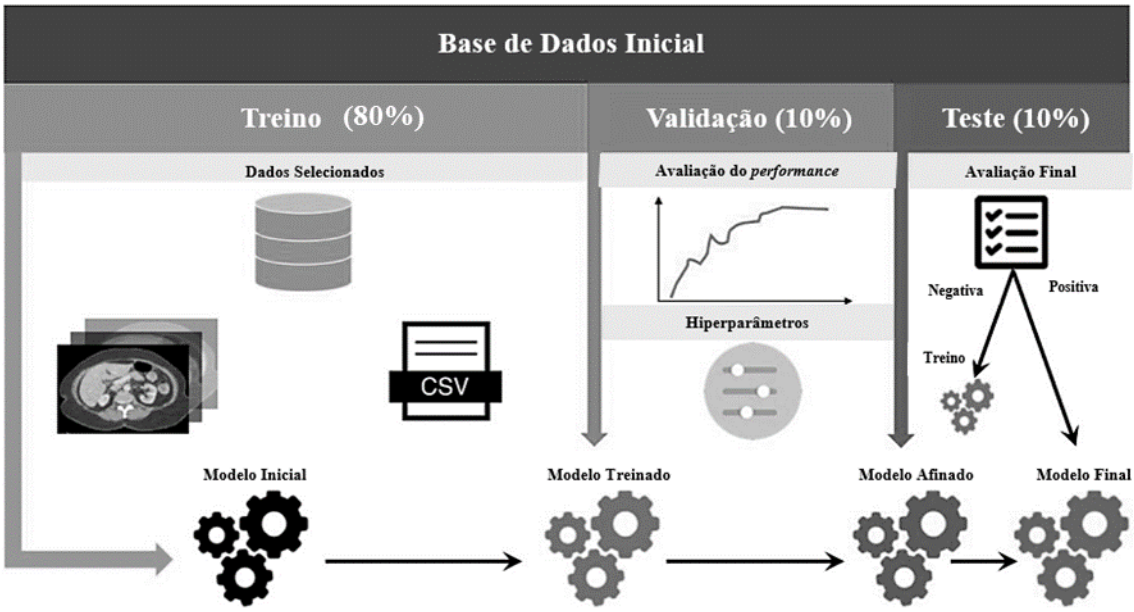


Figura 9 *- Esquematização da divisão da base de dados e as funções de cada subgrupo. Adaptado de:E. Montagnon (2020)*

No entanto ao utilizar esta metodologia, principalmente em bases de dados com poucas amostras, podemos correr o risco de não ter um grupo suficientemente representativo da população nos dados de validação. O risco de *overfitting* e *underfitting* está então presente. O primeiro risco refere-se ao facto de existir alta variância do modelo no grupo de treino/validação, com métricas de performance muito superiores no mesmo em relação ao grupo de teste, indicando não-generalidade. No segundo risco, podemos não ter informação suficientemente representativa das classes para o treino do modelo, resultando em performances relativamente baixas no grupo de treino e de teste devido a assunções demasiado simplistas das variáveis, indicando assim um alto viés (*bias*) [41,43].

Dado estes fatores existem métodos de reamostragem que podem ser aplicados se forma a utilizar a totalidade do grupo de treino para a validação. Estes métodos servem para determinar os processos de treino de modelo que melhor se adaptam à tarefa específica de forma generalizável, através da definição de *features* importantes, escolha de classificadores e calibração dos mesmos. São exemplos destes métodos mais utilizados o *k-fold cross validation*, *bootstrap* e *leave-one-out* (**Figura 10**).

K-fold cross validation **(CV)** é uma técnica que que divide os dados de treino em k subgrupos de treino e validação (nunca sobrepostos), com subsequente treino do algoritmo e classificação em cada um desses grupos. Quanto maior o valor de k mais subgrupos temos, mas menores as amostras de validação[48]. Esta técnica é conhecida por ser mais robusta e ter menos variância que uma simples divisão da base de dados, otimizando o uso dos dados disponíveis e permitindo realizar médias de métricas de *performance* entre os subgrupos para avaliação do erro preditivo dos modelos. Este facto permite que o mesmo seja utilizado para estimar o potencial de generalidade do modelo através do auxílio da escolha do processo de seleção de *feature* e dos hiperparâmetros do classificador mais indicados para a tarefa. Desta forma o CV é sensível à distribuição dos dados com classes desequilibradas, podendo resultar em grupos de teste e treino não representativas da amostra. A estratificação dos subgrupos é assim recomendada[48,49]. A decisão no valor de k não é consensual entre a literatura e é um

assunto debatido na área. Altos valores de k poderão aumentar o viés e reduzir a variância da estimativa de performance do modelo de cada subgrupo, pois existem mais amostras de treino nos mesmos. No entanto, isto torna a CV mais sensível a dados ruidosos ou *outliers* em bases de dados pequenas, resultando em grupos de teste não representativos e alta variabilidade de *performance* entre subgrupos. É importante referir que esta tendência só é válida para os subgrupos da própria CV, podendo a mesma inverter comparando modelos treinados com novos dados, mas com a mesma tipologia de aprendizagem (otimizada por CV). Isto deve-se ao facto de que quanto maior o valor de k, maior será a sobreposição das amostras de treino nos diferentes grupos criando modelos correlacionados e dependentes (o que gera variância e *overfitting* ao grupo de treino utilizado durante a otimização)[50–53]. Apesar deste fator, altos valores de k podem ser recomendados com pequenas bases de dados pois maximizamos o treino do classificador com maior quantidade de amostras, reduzindo o viés e *underfitting* do classificador[54]. Para um bom *trade-off* entre viés e variância o valor de k mais utilizado e recomendado na literatura é o 10, no entanto um valor de cinco também pode ser considerado, representando a proporção típica de 80:20 (treino:validação).

Leave-one-out pode ser considerado um método de validação cruzada, onde um treino é efetuado com todas as amostras dos dados menos uma (utilizada para a validação). O processo é repetido até todas as amostras terem sido classificadas[48]. Este processo é recomendado para bases de dados reduzidos, maximizando a fase de aprendizagem com o máximo de dados possível, enquanto oferece um processo de otimização generalizável. Está técnica pode ser considerada uma CV onde k = n (número de *samples*), aplicando-se os fatores referidos no parágrafo anterior. Autores referem que esta técnica pode apresentar baixo risco de viés nas avaliações das métricas de performance de diferentes modelos (erro preditivo), no entanto existe risco de variância dependendo da estabilidade do modelo, pois os subgrupos de treino têm mais amostras sobreponíveis entre si [55].

Boostraping é outro possível método de reamostragem onde é selecionada um grupo dos dados de treino para criação de um novo grupo que pode ser usado para validação. O grupo selecionado é então substituído por dados sintéticos que se aproximam da população original. Este método ajuda a prevenir *overfitting* introduzindo aleatoriedade e robustez ao algoritmo, sem assumir distribuições dos dados (não-paramétrico).[41] Este método pode correr o risco de introduzir viés em bases de dados pequenas pois as amostras sintéticas podem não representar a população. A variância pode ser também um problema, caso os classificadores e modelos complexos aprendam apenas a classificar corretamente as amostras sintéticas, não generalizando para as restantes populações[56]. No entanto estas desvantagens irão sempre depender da base de dados usada e da tarefa a aplicar, sendo a avaliação com um grupo de teste essencial.

V. Singh et al. estudaram os efeitos da escolha da técnica de amostragem estratificadas na performance de um modelo de classificação binária de *radiomics* de prognóstico de doentes, utilizando como classificadores o algoritmo de *gaussian naïve bayes*, regressão logística, análise linear discriminante e *random forest*. Os autores comparam 100 iterações diferentes de divisão de dados simples (proporções de 5:5 e 7:3), *tenfold cross validation* (10-CV), dez repetições de 10-CV e 500 iterações de *bootstraping*. A base de dados teve uma amostra de 715 doentes com *features* clínicas e de imagem (cintigrafia de perfusão do miocárdio). Os autores verificaram que a utilização de uma simples divisão poderia levar a uma diferença

significativa de 15% nas métricas de *performance* (AUC) entre os dados de validação e de treino nas diferentes iterações, recomendado os restantes métodos [55].

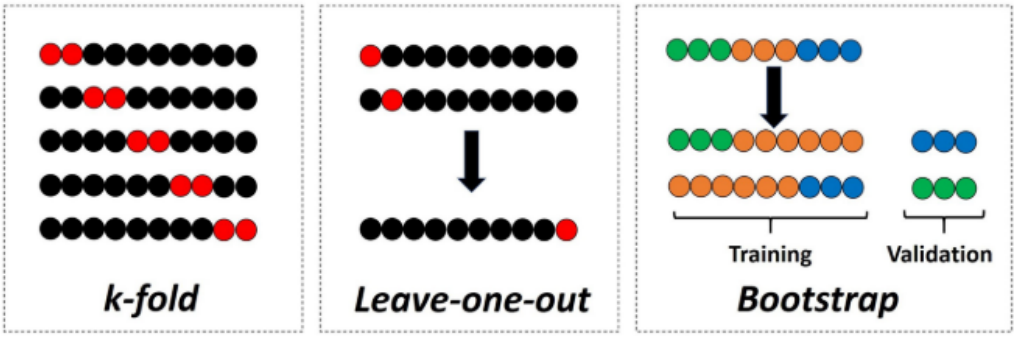


Figura 10 *– Esquematização de K-fold-CV, Leave-One-Out e Boostrap respetivamente. Os círculos vermelhos representam o grupo de teste nos dois primeiros métodos, enquanto as cores no último representam as diferentes classes. Fonte: Koçak et al (2021)*

C. An et al. realizaram um estudo semelhante com uma classificação binária tumoral de imagens de ressonância magnética utilizando *features* extraídas da região de interesse. Os mesmos testaram duas tarefas distintas, uma "fácil" (com 177 doentes) e uma "difícil" (com 258 doentes). Para testar a influência em bases de dados pequenas, os autores criaram dados com *undersampling* da base de dados total e utilizaram um modelo de baixa complexidade (regressão logística com regularização LASSO). Os autores consideraram 1000 iterações da simples divisão de treino e teste (7:3), testando a diferença de AUC entre diferentes técnicas de validação (*fivefold CV, 10 repetições de fivefold CV, nested fivefold CV, boostrap)* e o grupo de teste. A conclusão final residiu no facto de que a utilização de uma simples divisão entre dados de treino e teste pode não ser suficiente para averiguar o erro médio preditivo do modelo e consequentemente a generalidade do mesmo. Em tarefas complexas com bases de dados pequenas, a média da diferença de *performance* de AUC entre iterações pode variar até 0.092 (±0.071) onde num dos pares de treino/dados foi de 0.21. No uso da CV a variabilidade entre as iterações foi maior em bases de dados pequenas com tarefas complexas e maior quantidade de *features* (indicando variância), podendo ser justificado pela "maldição da dimensionalidade" que irá ser discutido em capítulos posteriores. Na comparação direta dos métodos de amostragem (utilizando as iterações com maior erro preditivo) não existiram diferenças significativas entre os mesmos, no entanto quando o erro preditivo é menor, o uso de *boostrap* foi o que permitiu a melhor generalidade e correção deste erro entre o grupo de validação e de teste.[57]

2.6 Pré-processamento dos dados numéricos e das radiografias

O pré-processamento de dados e imagens é essencial na transformação, limpeza, redução e *scaling* de uma base de dados para tarefas de *machine learning* em *radiomics*. Este processo promove a confiança na análise de dados, detetando ruído, introduzindo dados em falta e garantido que os mesmos cumprem os pré-requisitos dos diferentes classificadores e algoritmos de análise estatística [48,58].

2.6.1 Pré-processamento de variáveis numéricas e categóricas

A Imputação de dados em falta permite salvaguardar amostras com certas variáveis (*features*) não existentes através da criação sintética das mesmas. Este processo pode ser crítico quando se trabalha com bases de dados pequenas, através da manutenção do número de amostras. Existem diversos métodos de imputação de dados, podendo ser considerados univariados ou multivariados [58]. O primeiro usa técnicas como por exemplo, a imputação por média ou mediana (dos valores de uma determinada variável da base de dados) e a imputação *backward* ou *foward* (replica o valor da variável da amostra posterior ou anterior respetivamente). Estes métodos são simples e computacionalmente eficazes com baixa probabilidade de introdução de ruído, no entanto podem ser insensíveis a dados temporais [58]. Os métodos multivariados utilizam algoritmos de *k-nearest neighbour* e regressão linear (utilizando algorítmos de *machine learning* de regressão como *support vector machines*, SVM) para a criação de valores mais exatos. Estes métodos são mais complexos, apresentando boa *performance* quando existem grandes quantidades de medições em falta, no entanto é necessária uma base de dados significativa para a estimativa das substituições [58]. É importante referir que a imputação de dados deve ser realizada anteriormente à divisão dos dados (incluindo em cada subgrupo de uma CV), para evitar *data leakage* do grupo de teste para o grupo de treino, causando potencial *overfitting*.

A transformação de dados é outro passo necessário. Esta pode ser utilizada para tornar discretas variáveis contínuas, através do *binning* de igual espaçamento das mesmas (amostras são dividas em diversos intervalos de valores da variável definidos pelo utilizador), ou do *binning* de frequências (os intervalos são definidos pela quantidade equitativa de amostras entre os mesmos). O *one-hot-encoding* é outro processo que transforma variáveis categóricas em variáveis numéricas (discretas ou ordinais). Este processo é feito através da criação de uma matriz de colunas binárias onde o número de colunas é definido por L-1, representado L o número de categorias. Assim sendo uma variável categórica binária iria apresentar apenas uma coluna com valor de 0 e 1 codificando as categorias[58].

Por fim o *scaling* dos dados é também essencial, facilitando a visualização e análise estatística e preparando as mesmas para algoritmos de *machine learning* (por exemplo uma SVM), *deep learning* (redes neuronais) e regularização sensíveis à escala das variáveis. O *scaling* pode ser efetuado através da normalização por *max-min* ou por *z-score*. No primeiro método os valores são transformados para englobarem uma determina variação numérica. O caso mais típico é a normalização por *max-min*[58]*:*

$$f(x) = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Equação 2 – Normalização max-min

Em que: x, representa o valor da variável e x_{max} x_{min} , representam o valor máximo e mínimo dessa variável, respetivamente, com um intervalo final de valores de [0,1]

Este método pode ser utilizado em distribuições normais de dados, no entanto é sensível a *outliers* de valores extremos. A normalização por *z-score* é assim o método de *scaling* mais utilizado, transformando a variável para a mesma apresentar uma distribuição normal onde a média é zero e o desvio padrão é um[58]. Esta técnica pode ser definida por:

$$f(x) = \frac{x - \sigma}{\mu}$$

Exquação 3 – Normalização por z-score

Em que: x, representa o valor da variável, μ, representa a média da variável e σ representa o desvio padrão da variável.

2.6.2 Pré-processamento das radiografias

Os *radiomics* são dependentes dos parâmetros de imagem. Pode existir variação intramodalidade de imagem ou intermodalidade de imagem, no que toca à dimensão do pixel (tamanho da matriz), valor de pixel, (níveis de cinzento), e escala dos próprios níveis de cinzento[48]. Estas variações devem ser consideradas e minimizadas com técnicas de pré-processamento específicas para a imagem médica. Os processos mais típicos incidem na segmentação, normalização, redimensionamento, redução de ruído, otimização de contraste.

A segmentação é um dos passos mais críticos do pré-processamento de imagem quando são discutidos métodos clássicos de *radiomics*. A segmentação irá servir neste contexto para definir a região de interesse onde irão ser extraídas as *features* manuais do histograma e de textura da imagem. A segmentação manual das estruturas ou lesões continua a ser o *gold standard* quando realizada por indivíduos experientes na área. No entanto este método é dispendioso em termos temporais e poderá levar a problemas de reprodução e replicação devido à variabilidade inter e intra-operador.[48] Existem diversos métodos automáticos e semi-automáticos de segmentação com base em CNN's como a *U-Net* ou por métodos baseados em região, *tresholds* e contornos ativos (métodos de *snake*). No entanto estas técnicas podem não apresentar resultados homogéneos em toda a base de dados, devido à variação dos ruídos da imagem, protocolos de aquisição e artefactos clínicos.[48] Em relação a *radiomics* baseados em *deep learning*, a necessidade de segmentação pode ser verificada na literatura por *A. Tahir et al* [59]*.* Os autores verificaram numa tarefa de classificação de COVID-19, a segmentação providenciava resultados semelhantes das métricas de *performance*, enquanto melhorava a interpretação das mesmas, com foco da CNN na região do parênquima pulmonar (região de interesse) e não fora da mesma (**Figura 11***). M. Heidari et al*, também verifica que a remoção da região diafragmática via segmentação, melhorou a *performance* de uma CNN treinada para classificação de pneumonia [60].

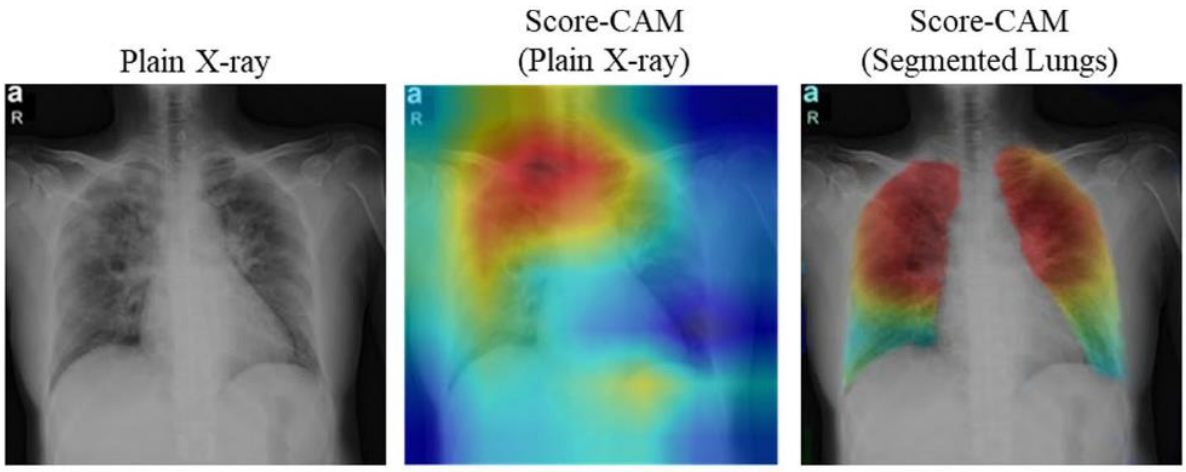


Figura 11 *– Visualização de mapas de ativação de classe (Score-CAM) de uma CNN treinada com (Segmented Lungs) e sem segmentação pulmonar (Plain X-ray) . É possível verificar que na ausência de segmentação a CNN foca a sua atenção na região da clavícula direita e não no pulmão.*

A normalização em tarefas de *radiomics* de DL é essencial devido às diferenças de intensidades do pixel que possa existir dentro de um grupo de dados. A diferença das técnicas de aquisição aplicadas ou a modificação da janela e nível da escala de cinzentos podem modificar o contraste e aparente sinal das mesmas, simulando, por exemplo, a identificação de patologias na CNN[42,48,61]. Este processo é essencial para manutenção da estabilidade do treino das CNN, proporcionando as mesmas numa escala comum de análise de valores de dimensão razoável, permitindo convergir facilmente, com uma eficiência computacional acrescida. As técnicas a aplicar podem variar, no entanto é típica a normalização por *max-min* dos valores de pixel, ou por métodos de Z-Score. Este último caso pode ser um pré-requisito essencial em CNN's pré-treinadas, que requerem uma normalização com médias e desvios padrão do grupo de dados utilizado no treino original[42,48,61,62].

O redimensionamento é também essencial no que diz respeito a manter a eficiência computacional do treino das redes neuronais, mantendo uma matriz comum entre as imagens. Tipicamente as CNN recebem como *input* imagens de dimensão inferior a 300x300 pixel, que é uma dimensão inferior à da típica R-RTX. O redimensionamento é assim essencial e pode ser realizado via técnicas de interpolação[46].

A redução do ruído é essencial no início de qualquer tarefa de pré-processamento, facilitando a aplicação de diferentes técnicas. O ruído em imagens médicas pode ser comum, dependendo do biótipo do doente e do método de aquisição das imagens (secção 2.2.1). No contexto das CNN's, o ruído pode esconder possíveis características de imagens interessantes para a presente classificação, impossibilitando a captura de certos padrões ou contornos da região de interesse. Devido a este fator, a aplicação de filtros de *blur* pode ser essencial, permitindo a redução de altas frequências e variações abruptas de pixel das imagens (através de técnicas de convolução) o que tem melhorado a *performance* das CNN's em diferentes tarefas[63–65]. *A. Gielczyk et al.* testam dois filtros de *blur* numa tarefa de classificação de COVID-19 utilizando R-RTX e uma CNN[66]. Os filtros utilizados foram o filtro gaussiano (cuja forma da filtragem apresenta um aspeto de uma curva gaussiana) e o filtro bilateral (que permite redução de ruído com manutenção de contornos, tendo em conta o domínio espacial e a intensidade dos pixéis). O autor verificou que a utilização do filtro gaussiano com uma equalização de histograma simples, levou a uma melhor exatidão, precisão e sensibilidade do modelo treinado. Este filtro foi assim selecionado para a presente dissertação.

A otimização do contraste é outro elemento que pode beneficiar a classificação de imagens médicas de forma qualitativa ou quantitativa, particularmente em R-RTX cujas variações de contraste são significativas no contexto de imagens de aparelhos portáteis em UCI, por razões já referidas. Para redução do impacto destas variáveis, a equalização do histograma poderá salientar padrões do parênquima pulmonar [60], associados com a ARDS-COV19. Esta técnica também tem demonstrado beneficiar estudos de *radiomics* baseados em DL[59]. A equalização do histograma é uma técnica que tem a capacidade de propagar os valores de intensidade mais frequentes pelo histograma, podendo ser definida pela seguinte equação:

$$\boldsymbol{y = T(x) = (L - 1)\sum_{I=0}^{X} p_x(X = i)}$$

Equação 4 – Equalização do Histograma

Onde $\boldsymbol{x}$ representa a intensidade do pixel original, $\boldsymbol{p_x(X = i)}$ representa a probabilidade do pixel ter essa intensidade, $\boldsymbol{T(x)}$ representa a transformação da função, $\boldsymbol{y}$ representa o novo valor de intensidade do pixel após a transformação e $\boldsymbol{L - 1}$ representa o valor máximo de intensidade para uma imagem de níveis de cinzento de 8 bits (255) [59].

Apesar da aplicação deste método apresentar benefícios em tarefas de classificação [66], existe outro método mais robusto chamado de equalização do histograma adaptativa com limite de contraste (CLAHE). Esta técnica realiza a equalização do histograma em pequenas regiões da imagem, melhorando os contornos e contrastes locais. Para prevenir o aumento do ruído, a mesma aplica um *threshold* limitante do aumento do contraste (*CLIP limit*) em determinada região[59,67]. Esta melhoria promove o aspeto natural das imagens quando. No caso das R-RTX previne a sobressaturação pulmonar encontrada nas técnicas de equalização de histograma comum (**Figura 12**). Este método foi aplicado para teste na presente dissertação, utilizando um CLIP *limit* conservador de 2.0 como recomendado por *S. Nefoussi et al.*, que verificou melhorias de sensibilidade e de AUC na deteção de pneumonia com CNN's em R-RTX, utilizando CLIP *limits* de 2.0 e 4.0 (em situações de classes não equilibradas)[67].

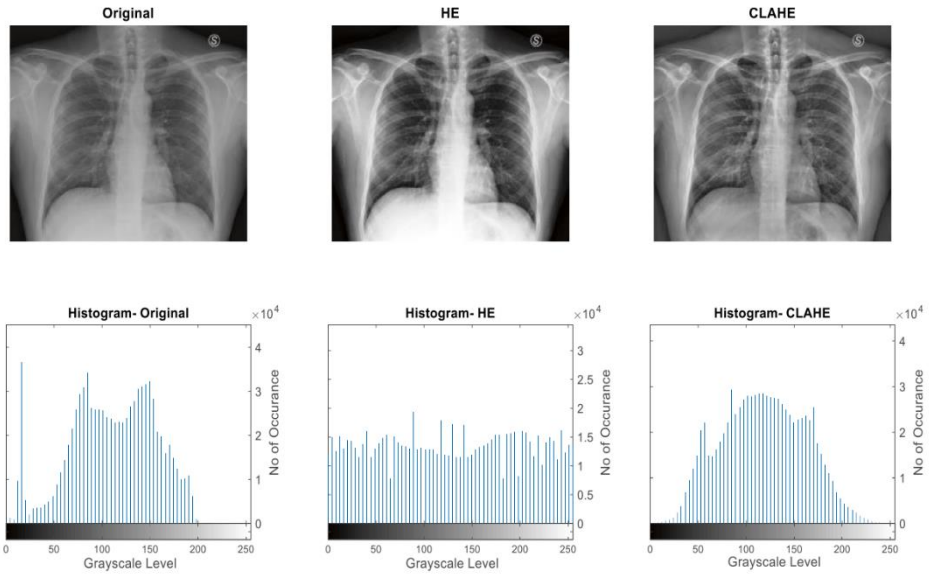


Figura 12 *- Efeitos e histogramas da aplicação da equalização de histograma comum (HE) e CLAHE. Adaptado de Tahir et al. (2022)*

2.7 Extração de *features* imagiológicas – *Deep Learning Radiomics*

DL é um subtipo de ML que possibilita a descrição complexa de representação de dados, através de hierarquias de estruturas de determinadas *features*. Este método utiliza multi-camadas de redes neuronais com um número finito de unidades não-lineares (neurónios). É possível assim a aprendizagem de representações digitais de dados complexos como imagens, através de cálculos de convolução de CNN's [44]. Esta afirmação torna apelativa a utilização de DL em imagem médica, principalmente na radiologia, onde é necessário a extração de *features* de imagem representativas dos fenótipos da mesma, permitindo a correta caracterização patológica e previsão do prognóstico do doente [44]. O processo de *Radiomics* clássico baseia-se na extração manual de *features* de textura, morfologia e distribuição de intensidades de pixel, através da segmentação da imagem por ROIs. Este processo pode ser demorado, dispendioso e dependente da experiência do operador. DL oferece a vantagem de processar, calcular e extrair *features* de imagem de forma automática. Estas *features* são geralmente mais robustas, ricas e apresentam diversos níveis de abstração não observáveis em métodos clássicos ou na avaliação qualitativa [42,44,68,69].

Uma CNN é um tipo de rede neuronal preparada para o processamento e análise de dados visuais como imagens. Existem diversas arquiteturas utilizadas na literatura para classificação de radiografias de tórax. São exemplos dessas redes a *AlexNet, VGG, ResNet, DenseNet, Inception (GoogLeNet)* e U-Net [70,71]. No entanto, é possível verificar alguns elementos em comum entre as mesmas, sendo estes as camadas que a compõem que irão ser descritas de seguida.

A camada de *input* é início da CNN. Esta camada pode representar imagens em formato de tensores. Um tensor é definido por uma estrutura de dados multidimensional que permite expandir as matrizes para mais de duas dimensões [72]. Este acréscimo de dimensionalidade pode representar os canais de uma imagem RGB, um conjunto com mais que uma imagem, ou uma sequência temporal das mesmas [72]. No caso em concreto e para o treino de uma CNN, o tensor de 100 R-RTX em níveis de cinzento de dimensão 224x224 píxel, seria representado por um tensor de (100,224,224,1).

A camada de convolução (CL) é o "coração" da CNN e apresenta o papel de obter representações reduzidas dos dados do input, de forma hierárquica, através de operações de convolução (**Figura 13**) [71]. Esta camada utiliza filtros (*kernels*) com ponderações adaptadas à tarefa de classificação em questão, aprendidas durante a fase de treino do modelo. Estes filtros realizam operações de convolução locais ao longo da imagem, permitindo obter uma nova matriz correspondente aos produtos escalares das mesmas. Estas matrizes são os chamados mapas de *features* e permitem obter informação importante sobre padrões, textura e contornos locais ao longo de todo o tensor inicial. O tamanho destes mapas é dependente do tamanho dos filtros e do tamanho do movimento que realiza ao longo da imagem (*stride*). Quanto maior o *stride* mais são os elementos da matriz que são ignorados, reduzindo a dimensão dos mapas. Para obter informações dos contornos, usualmente é necessário aplicar *padding*, ou seja, aplicar linhas e colunas extras nas bordas das imagens com um valor de pixel igual a zero (*zero-padding*), permitindo assim operações de convolução nessas regiões. Um dos grandes benefícios desta camada e que também ajuda na computação, é o facto das ponderações dos filtros serem partilhadas ao longo de todo o *input*, sendo só necessário a aprendizagem de um grupo de ponderações [70,71,73].

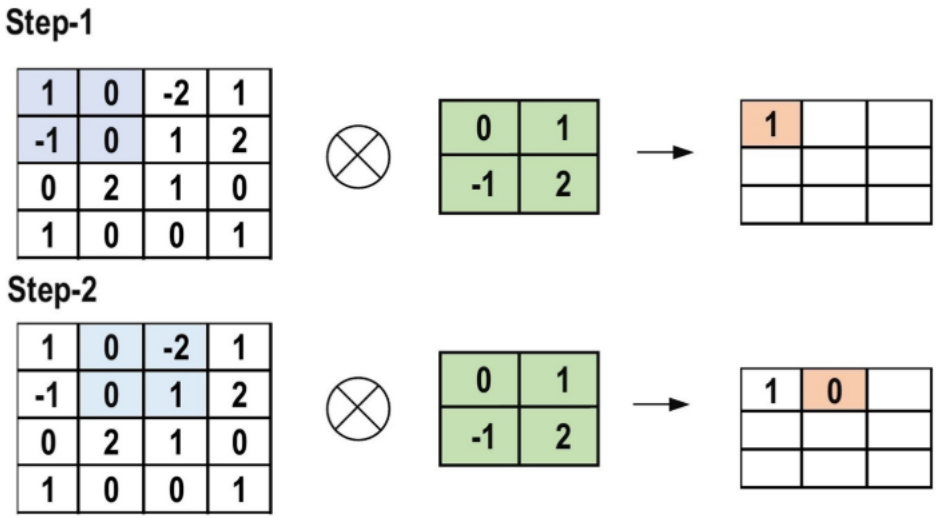


Figura 13 *- Exemplo de uma operação de convolução na camada convolucional. O filtro está representado a verde e o produto escalar a laranja. Adaptado de L. Alzubaidi et al. (2021)*

Posteriormente, existe, tipicamente, uma camada de ativação. Esta camada tem a funcionalidade de efetivamente ativar ou não os neurónios seguintes (mapeando o *input* ao *output*) e está presente em todo o tipo de redes neuronais. Este mapeamento funciona para todas as camadas de uma CNN com ponderações, quer sejam uma CL ou uma camada densa. O objetivo principal deste elemento da CNN é introduzir a não-linearidade, permitindo a aprendizagem de padrões e representações complexas [70,71,73]. Diversas funções podem ser utilizadas para este efeito, no entanto a função de ReLU é a mais comum. Esta função pode ser representada por:

$$f(x) = \max(0, x)$$

Equação 5 – Função de ReLU

Onde se o valor do *input* (x) for positivo, o *output* é igual a x, no entanto se for negativo é igual a 0. Estes fatores tornam a função computacionalmente eficiente e ajuda a mitigar o problema do *vanishing gradient* que irá ser explicado posteriormente nesta secção.

As camadas de *pooling* têm o objetivo de reduzir as dimensões espaciais do *input* e dos mapas de *features* produzidos pelas CL's, promovendo menor complexidade computacional e mantendo as informações dominantes das *features*. Os tipos de *pooling* mais utilizados são o *pooling* por média e por valor máximo (**Figura 14**). Na primeira tipologia é extraída a média de um grupo de valores dos mapas de *features*, enquanto na segunda é extraída o valor máximo de um grupo de valores. O *poolling* da média global também pode ser utilizado para efeitos de *flattening* das dimensões dos mapas de *features*, reduzindo o mesmo a um valor unidimensional (1D) [70,71,73].

As camadas de regularização têm o objetivo de prevenir o *overfitting* da CNN e podem incluir várias metodologias. As mais utilizadas são a camada de *dropout,* que remove aleatoriamente uma fração dos neurónios durante o treino, permitindo melhor generalidade da mesma (obriga a CNN a aprender *features* independentes) e a camada de normalização dos *batchs.* Esta camada normaliza os mapas de *features* por métodos de *z-score* entre camadas, permitindo mais estabilidade de treino, maior regularização e menores efeitos de *vanishing gradient*[70,71,73]*.*

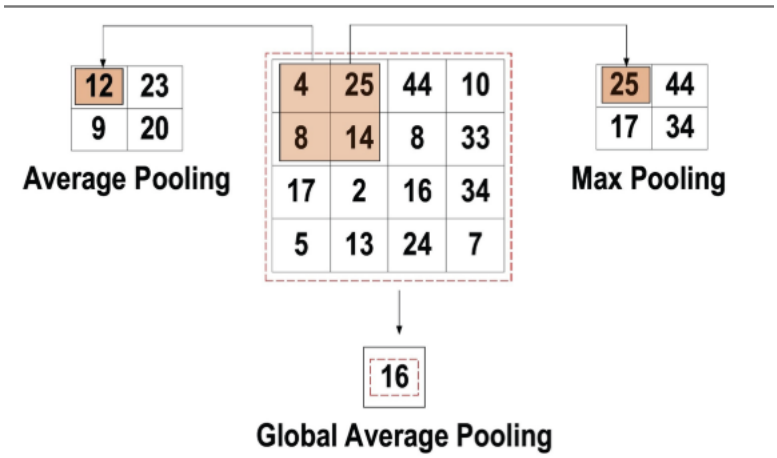


***Figura 14**- Exemplo de pooling por médias (Average Pooling), pooling por máximos (Max Pooling) e pooling da média global (Global Average Pooling)*

Por fim, existe a camada totalmente conectada (FCL): Esta é a última camada da CNN e vai ser responsável pela classificação final dada ao *input*. Nesta camada todos os neurónios estão interconectados aos neurónios da camada prévia (camada densa) e segue a logística de um peceptrão multicamada (rede neuronal *feedfoward*). Esta camada recebe como *input* os mapas de *feature* (após *flattening*) da última CL. Na **Figura 15** é possível verificar o exemplo de uma CNN completa [70,71,73].

Durante a fase de treino de uma CNN, existe uma iniciação aleatória das ponderações da rede através do *foward pass*. A CNN gera assim as primeiras previsões para os dados de treino, para efeitos de comparação com os *ground truth labels* e análise de uma função de *loss*. Esta função tem a capacidade de avaliar a discrepância entre estas duas variáveis e irá ser utilizada no processo de *backpropagation*. Nesta etapa, os gradientes da função de perda são calculados em relação às ponderações iniciais da CNN onde, através das regras de cadeias dos cálculos, os mesmos se propagam da camada de *output* para as camadas inicias. É aqui que aparece o termo *gradient descent*. Este algoritmo ajusta as ponderações na direção oposta do gradiente, minimizando a função de *loss,* de acordo com uma determinada taxa de aprendizagem que afeta a dimensão dos passos dados para a redução da função de *loss*. O ciclo é assim repetido durante múltiplas fases, ou épocas, e diferentes agrupamentos de dados permitindo a aprendizagem progressiva [70,71,73].

Contudo, um problema que pode afetar este procedimento é o efeito de *vanishing gradient,* onde os gradientes, ao longo da *backpropagation* e de vários ciclos, diminuem significativamente o seu tamanho em redes neuronais profundas, dificultando o treino das camadas iniciais onde a CNN para de aprender. Para além dos métodos já referidos, diversos arquiteturas de rede procuram solucionar esta problemática, como é o caso da *DenseNet-121*[71].

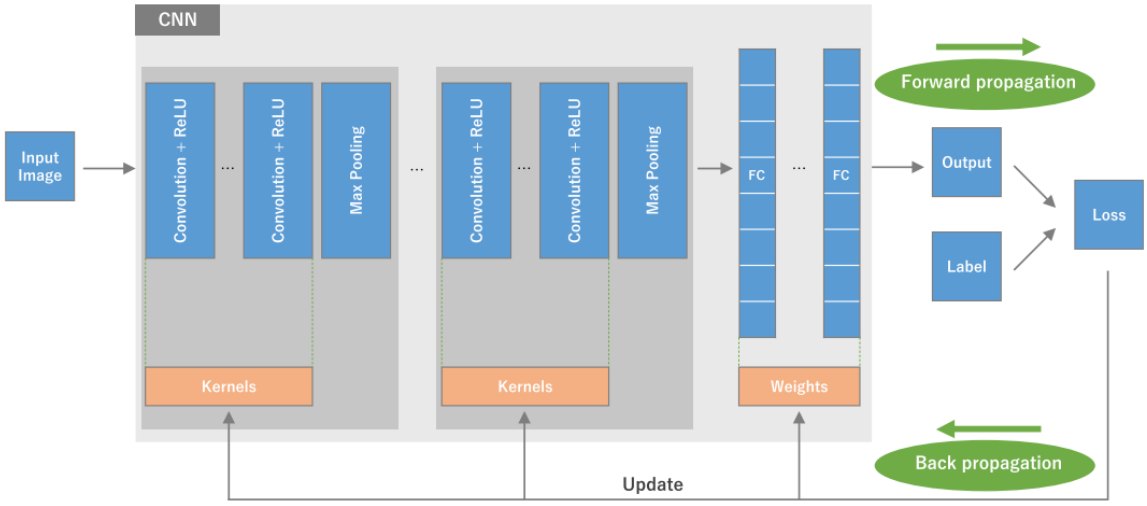


Figura 15 *- Representação de um CNN e das suas épocas de treino*

2.7.1 DenseNet-121

A *DenseNet* (**Figura 16**) foi criada com o objetivo de resolver o problema de *vanishing gradient*, através de da utilização de camadas altamente interconectadas e reutilização de *features*[74]. Esta arquitetura foi utilizada em diversas tarefas de classificação no âmbito dos *radiomics* com R-RTX[23,75–78]. Em particular, a arquitetura da DenseNet-121 é caracterizada por 121 camadas, sendo constituída por blocos densos e camadas de transição. Estes blocos são os componentes principais desta arquitetura e contribuem para a eficiência de otimização dos parâmetros do modelo, fluxo de gradientes aprimorado e representação eficaz de *features*[74].

A rede inicia-se com uma CL inicial, onde são extraídas as *features* iniciais das imagens de *inputs*. De seguida a rede apresenta quatro blocos densos, cada um contendo várias camadas densamente conectadas. A conectividade densa é uma característica distintiva dos *DenseNets*, onde cada camada recebe como *input* direto todas as camadas precedentes dentro do mesmo bloco, através da concatenação de mapas de *features*. Esta conectividade densa promove uma *backpropagation* eficiente, abordando os desafios associados à perda de informações em redes profundas anteriormente discutidas[74].

Dentro de cada bloco denso, camadas de *bottleneck* são integradas, utilizando convoluções 1x1 para reduzir o número de canais de entrada antes de convoluções subsequentes com *kernels* de 3x3. Esta escolha ajuda a controlar o custo computacional mantendo a capacidade representacional da rede, apesar da grande interconetividade. Existem também camadas de transição, intercaladas entre blocos densos, que desempenham um papel crucial na redução de dimensões espaciais através de uma combinação de convoluções 1x1 e *pooling* de médias. Um *pooling* da média global é aplicado após o último bloco denso (última camada de CL), resultando em dimensões espaciais unidimensionais de 1024 *features* usados para a classificação. A última camada é uma FCL, cujo número de neurônios corresponde ao número de classes na tarefa de classificação[74].

A *DenseNet*-121 apresenta características como eficiência de parâmetros, o que é especialmente vantajoso em comparação com arquiteturas tradicionais. A conectividade densa da arquitetura promove a reutilização eficiente de características entre as camadas, fomentando a representação de padrões complexos nos dados[71,74].

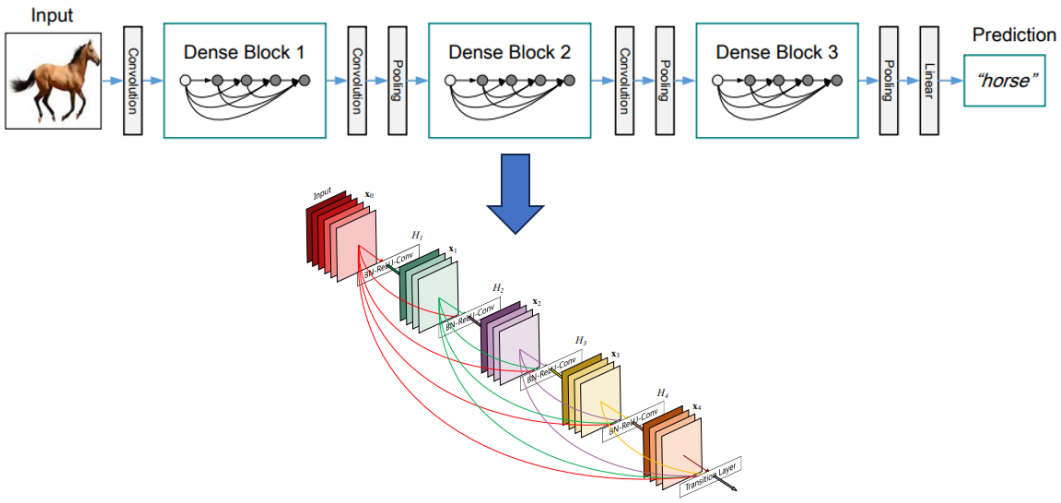


Figura 16 - *Representação da arquitetura de uma Densenet, onde é possível verificar a interconetividade de um bloco denso (seta azul)*. Adaptado de *G. Huang (2017)*

No entanto, o fator mais interessante para esta dissertação na utilização desta rede, reside no trabalho realizado por *P. Rajpurkar et al.* em 2017[79]. Estes autores utilizaram uma *DenseNet*-121*,* previamente treinada com a base de dados da *Imagenet,* e re-treinaram a mesma do princípio ao fim com uma taxa de aprendizagem baixa (*fine-tuning* total) utilizando a base de dados "*ChestX-ray14"*[80]*,* com o objetivo de classificar 14 patologias torácicas em R-RTX (pneumonia, cardiomegália ***etc***…). Os autores denominaram esta rede de CheXNet e na classificação de pneumonia, os autores obtiverem métricas de classificação superiores às de radiologistas treinados, provando a eficácia da rede. Esta CNN é *open-source*, o que levou à popularidade da sua utilização por outros autores em distintas tarefas de classificação[23,24,35,77]. As ponderações desta rede podem ser facilmente adquiridas e acopladas a qualquer arquitetura da *DenseNet-121.* No presente trabalho, esta rede pré-treinada para R-RTX foi utilizada para extração de *features* de qualidade sem necessidade de treino prévio para a previsão de mortalidade em doentes ARDS-COV19[79].

2.7.2 Técnicas de fusão de imagem

Devido ao facto do presente estudo dispor de duas R-RTX adquiridas em tempos de internamento distintos (d1 e d3), técnicas de fusão de dados multimodais foram considerados para resolver a problemática da *CheXNet* permitir apenas um *input*. A fusão de dados em DL tem sido uma realidade crescente no ramo da biomédica, permitindo a captura de relações complexas entre diversas modalidade clínicas na deteção de doença ou previsão de prognóstico dos doentes[81].

As técnicas de fusão de imagem podem ser divididas em três categorias. Em primeiro lugar, os métodos de *early fusion*, onde os dados são concatenados no *input* de uma CNN para uma

representação vetorial unimodal. Esta metodologia obriga os modelos de DL a não diferenciar as modalidades, originando *features* de representatividade combinada. Este método é computacionalmente eficiente e tem a vantagem do estudo das relações intermodais, ser efetuado desde os mapas de *features* iniciais, oferecendo características que estudam a relação dos *inputs* (não considerando os mesmos individualmente)[78,81]. As técnicas de *intermediate fusion* apresentam primeiro um mapa de *features* representativos de cada modalidade, antes da posterior junção para extração de *features* combinados e da classificação. Esta metodologia oferece a vantagem de flexibilidade, permitindo encontrar a profundidade certa para realizar a fusão numa CNN o que pode representar melhor as relações intermodais[78,81]. Os modelos de DL são particularmente úteis nesta metodologia, permitindo extração de *features* e combinação de camadas com relativa facilidade. Finalmente, existe a *late fusion,* onde as previsões de dois modelos (para cada modalidade) separados, são combinadas para uma classificação final. Este método permite a adaptação de cada modelo de DL para a modalidade específica, no entanto não oferece captura de *features* de representatividade intermodal, não representado as suas possíveis relações complexas e não-lineares[78,81].

Devido a estes fatores, foi selecionada a técnica de *early fusion* para a extração de *features* das R-RTX nesta dissertação, com o objetivo de obter uma possível relação temporal de melhoria ou agravamento dos padrões radiológicos. Tipicamente esta fusão em imagem é utilizada com recurso a *autoencoders*, que aprendem representatividades vetoriais comprimidas das imagens (métodos não supervisionados) antes da posterior concatenação[45,81]. No entanto, este método requer o treino destes modelos, sendo o mesmo limitado com poucas amostras. A concatenação pode ser realizada diretamente nas matrizes da imagem verticalmente (ao longo das linhas) e horizontalmente (ao longo das colunas). Este método foi utilizado com sucesso por *K. Lopez et al.,* onde os autores realizaram concatenação direta de R-RTX com imagens representativas do texto de relatórios médicos no *input* de uma DenseNet-121[78]. Os autores testaram vários tipos de fusão e obtiveram os melhores resultados com esta metodologia, sendo a mesma selecionada para esta dissertação experimentalmente.

2.8 *Transfer Learning* e as suas metodologias em *radiomics*

Uma grande vantagem da utilização das CNN's reside no facto de ser possível utilizar uma arquitetura pré-treinada, para um certo tipo de imagens e tarefa, para dados ou classificações distintas. Chama-se a este procedimento de *transfer learning*, sendo o mesmo útil na extração de *features* e em tarefas de classificação onde a quantidade de dados disponíveis é relativamente baixa. Na literatura existem quatro tipos de metodologias principais para o *transfer learning* em imagem médica*,* oferecendo cada uma vantagens e desvantagens (**Figura 17**) [82].

A metodologia híbrida de extração de *features* utiliza uma CNN pré-treinada apenas para extração de *features*, sem treino ou *fine-tuning* da mesma. Isto é conseguido descartando a último FCL de uma CNN, permitindo acesso aos mapas de *features* da última CL após *flattening*. Estas *features* são utilizadas para o treino de classificadores clássicos de ML como a regressão logística. Este método oferece a vantagem de ser computacionalmente eficiente e permitir o uso de modelos de ML, que podem obter melhores resultados com bases de

dados pequenas e oferecer melhor interpretação dos modelos finais. No entanto não existe treino da CNN para os novos *inputs* que poderão ser diferentes do treino original da mesma, comprometendo o verdadeiro potencial da qualidade de *features* extraídas[42,63,82].

A metodologia de extração de features é realizada através do "congelamento" do treino das camadas de convolução, apenas sendo treinada FCL final. Este método é semelhante ao anterior em termos computacionais e de vantagem em bases de dados pequenas. No entanto, nestas vertentes, a FCL pode induzir *overfitting* devido à complexidade do modelo quando comparado a algoritmos de ML tradicional. Este método oferece a vantagem de poder obter mapas de ativação de classe para visualização de mapas de atenção da CNN, visto que a totalidade da CNN é utilizada mesmo que não treinada e as ponderações em relação à classe podem ser analisadas retrospetivamente[82,83].

A metodologia de fine-tuning "congela" apenas as camadas de convolução iniciais, permitindo a re-aprendizagem dos *kernels* das CL finais e da camada de classificação final. Este método permite melhor adaptação da CNN aos *inputs* e tarefas atuais, no entanto requer uma base de dados com amostras suficientes para o treino[82].

A metodologia de fine-tuning total re-treina toda a CNN desde o início com um fator da taxa de aprendizagem pequena. Este método é frequentemente utilizado, no entanto, numa comparação de diversos estudos, esta metodologia não ofereceu vantagens de *performance* em relação às anteriores. Não existe também um *standard* dos parâmetros de *fine-tuning* a utilizar, recomendando-se assim a utilização das outras técnicas antes de tentar esta metodologia, visto que a mesma é computacionalmente exigente e requer maiores bases de dados[82,84].

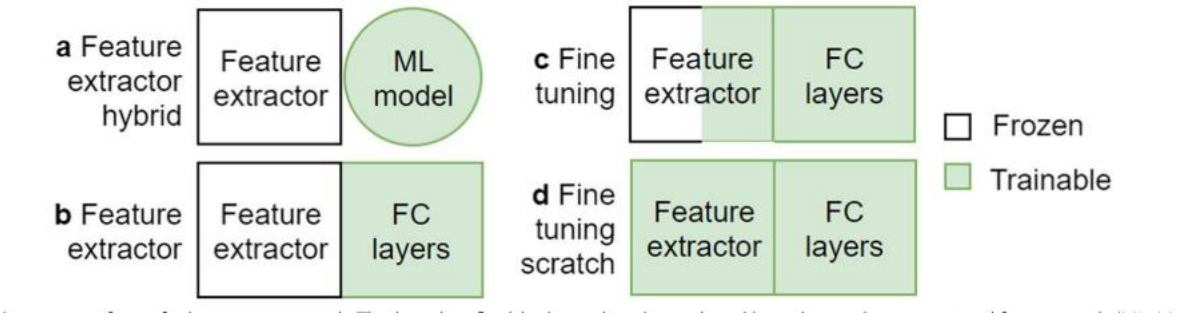


Figura 17 - *Esquematização dos quatro tipos de transfer learning encontrados na literatura. a) Método de extração de features híbrido (Feature extracor hybrid), b) método de extração de features (feature extracto), c) método de fine tuning e d) método de fine tuning total (fine tuning scratch). Adaptado de H. Kim et al. (2022)*

2.9 Redução da dimensionalidade – *Feature Selection*

A "maldição da dimensionalidade" é um termo utilizado em ML pra referir o risco de *overfitting* em grupos de dados de grande dimensionalidade de *features* em comparação com o numero de amostras[85]. Este fenómeno leva ao aumento da variância do modelo, onde o mesmo aprende com ruído e flutuações dos dados do grupo de treino, através da introdução de variáveis correlacionados e irrelevante. Este fenómeno faz com que o modelo sofra uma adaptação excedente ao grupo de treino, com *performance* significativamente reduzida no grupo de teste. Para mitigar esta problemática e reduzir o tempo computacional do treino, a seleção de *features* é um processo essencial nos treinos dos modelos de classificação. Este método permite a remoção de *features* possivelmente irrelevantes para as classes de interesse e que são redundantes entre si. No entanto, é preciso precaução na introdução desta metodologia, pois *features* aparentemente insignificantes na discriminação de classes, podem tornar-se relevantes na presença de outras variáveis de interesse. Existe a necessidade de atingir um equilíbrio entre aparente redundância, correlação e as possíveis interações de interesse[85,86]. O objetivo da seleção reside assim na seleção do subgrupo ótimo de variáveis que permitam a correta discriminação das classes de interesse. Tendo em conta estes fatores, existem diferentes metodologia de seriação das variáveis, importantes para tarefas de classificação, cada uma oferecendo vantagens e desvantagens (**Figura 18**)[85,86]. Para prevenir *overfitting* estes métodos devem ser aplicados durante a CV (em cada subgrupo) e não antes da mesma (*data leakage)*[87]*.*

Os métodos de filtragem realizam pontuação de *features* por *ranking* com base em determinados critérios e testes estatísticos, sendo independentes do algoritmo de classificação. As *features* são selecionadas ou excluídas antes da fase de treino/aprendizagem do modelo, promovendo eficiência do ponto de vista computacional. Este método é fácil de implementar e oferecem uma metodologia rápida de identificação de *features* potencialmente relevantes. Como estes métodos não consideram as relações com o classificador, são menos propícios a *overfitting* e podem ser particularmente úteis para conjuntos de dados de alta dimensionalidade. No entanto, os métodos de filtragem podem negligenciar dependências e interações entre *features*, pois avaliam cada uma de forma independente. Consequentemente, o desempenho pode ser reduzido quando a relação entre as características e a variável alvo é complexa ou não linear[85,88]. Os métodos de *threshold* ideal para a seleção de *features* é discutido na literatura, não existindo um consenso[86,89]. *R. Figueroa et al (2012)* recomenda a criação de curvas de *performance* para identificação do número ótimo de *features* [90]*.* No entanto, alguns métodos estatísticos utilizados para avaliação da relação entre *features* contínuas e variáveis categóricas são os testes de:

- ANOVA**:** Análise de variância das médias de dois ou mais grupos (classes), indicando se as mesmas diferem significativamente uma da outra. O *ranking* é realizado pelo valor de *score* de F obtido pelos testes. Este teste assume linearidade entre as *features* e o *target* assim como uma distribuição normal das *features*[88,91,92].
- *Gini Decrease*: Método utilizado habitualmente em algoritmos de árvores de decisão e *random forest*. Este teste mede a impuridade de *Gini* avaliando quantas vezes uma *feature* seria incorretamente classificada numa classificação aleatória de acordo com a distribuição das classes no grupo de dados. O *Gini Decrease* mede quanto esta impuridade desce de acordo com divisões dos dados baseadas em cada *feature*, determinando este valor a importância da mesma[91].

- *Fast Correlation Based Filter*: método multivariado que calcula a incerteza simétrica entre a variável contínua e o *target*, verificando quanta informação a mesma providencia em relação ao *target*. Subgrupos de *features* são testadas entre si para verificação de redundância e importância. Este método pode prevenir as desvantagens da filtragem[85].

Os métodos de *wrapper* determinam subgrupos de *features* com base na *performance* preditiva de um classificador treinado com essas *features*. Os mesmos envolvem treino e avaliação iterativa do modelo com diferentes subgrupos de *features*, selecionando o subconjunto que otimiza o desempenho do modelo. O método RFE (*Recursive Feature Elimination*) é um exemplo desta metodologia, utilizando um estimador (por exemplo os coeficientes de importância do classificador) para avaliar a importância das *features*, eliminado iterativamente as menos importantes até atingir um número pré-definido. Estes métodos consideram interações e dependências entre as *features*, proporcionando uma avaliação precisa da relevância das mesmas num modelo específico. Consequentemente, este método pode lidar com relações complexas e não lineares dos dados. No entanto estes métodos são computacionalmente dispendiosos e podem estar sujeitos a *overfitting*, especialmente com bases de dados pequenas. Esta metodologia é também sensível à escolha da métrica de avaliação e do algoritmo específico usado para o treino do modelo[85,88].

Os métodos embutidos incorporam a seleção de características como parte integral do processo de treino do modelo. Algoritmos como árvores de decisão, regressão logística e *random forests* realizam naturalmente a seleção de *features* durante a fase de treino por métodos de regularização que irão ser discutidos na secção seguinte. Esta metodologia explora a capacidade do modelo de avaliar a importância das *features* no contexto da tarefa atual. São eficientes do ponto de vista computacional em comparação com os métodos de invólucro e podem lidar com conjuntos de dados grandes. O desempenho dos métodos embutidos depende da adequação do modelo escolhido para a tarefa. Podem não ser tão flexíveis quanto os métodos de invólucro na captura de interações complexas entre características. As *features* selecionadas pode variar dependendo do algoritmo utilizado[85,88].

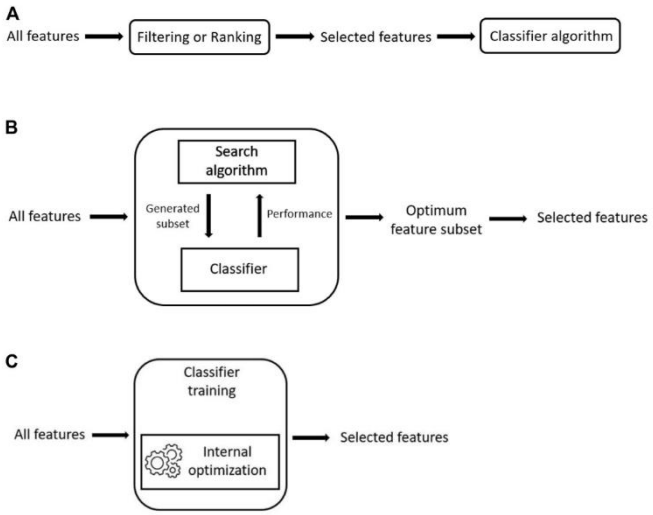


Figura 18 *- Esquematização das metodologias de seleção de features. a) técnicas de filtragem, b) técnicas de wrapper, c) técnicas embutidas. Adaptado de: N. Pudjihartono et al. (2022)*

2.10 Modelos de *machine learning* para classificação binária

Com o intuito de aplicar o método híbrido de extração de *features*, o presente capítulo apresenta os modelos de ML mais comuns para a aplicação de classificações binárias. Os hiperparâmetros de cada um irão também ser discutidos, sendo essenciais no *fine-tuning* final do modelo (modificação dos mesmos para obter os melhores resultados para a nossa tarefa e base de dados).

A regressão logística (LogReg) é um algoritmo popular no que toca a classificações binárias categóricas. Este classificador oferece a vantagem de proporcionar uma probabilidade de uma amostra pertencer a uma classe, utilizando uma função sigmoide para a previsão[88]. Esta função transforma uma combinação linear das *features* selecionados num valor probabilístico de zero a um. Caso a probabilidade seja superior a um *threshold* pré-definido (por exemplo 0,5) a amostra é classificada como a classe positiva[93,94]. A função pode ser definida por:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Equação 6 – Função de Regressão Logística

Onde σ(z) representa a probabilidade de uma amostra pertencer à classe positiva e z representa a combinação linear das *features* com ponderações previamente treinadas. A aprendizagem é realizada otimizando as ponderações de forma a que maximizem a probabilidade de classificar o evento como correto, através de técnicas como *gradient descend*[93,94]. Este classificador é bastante interpretável (modelo *white box*) pois permite verificar a contribuição de cada *feature* pelas ponderações aprendidas. A LogReg funciona melhor em casos onde é possível assumir linearidade entre as *features* (variável independente) e o *target* (variável dependente), ou seja quando as mesmas são linearmente separáveis[88]. Este classificador apresenta também a vantagem de sofrer menor variância tendo uma complexidade reduzida e apresentando sucesso nos resultados com grupos de dados de poucas amostras[88]. Em relação aos hiperparâmetros, a regressão logística tipicamente dispõe de 2 tipos de regularização que podem ser aplicadas. A regularização procura prevenir *overfitting* através da introdução de um termo de penalização à função de *loss* durante o treino no modelo[88]. Na regularização L1(Lasso), o termo de penalização é o valor absoluto das ponderações multiplicado por um parâmetro, definido pela letra “C”. Menores valores de C provocam regularizações mais fortes. A regularização L1 faz com que os pesos das *features* não importantes tendam para zero, existindo assim seleção de *features* integrada. A regularização L2 (*Ridge*), não resulta tipicamente em ponderações de valor zero resultando o termo de penalização do quadrado das mesmas multiplicadas por C. Este facto obriga o modelo a distribuir de forma equilibrada a importância das *features*[88].

O classificador *support vector machine* (SVM) é também tipicamente utilizado em tarefas de classificação categóricas ou de regressão. As SMV’s procura o melhor hiperplano capaz de distinguir as classes de interesse, servindo o mesmo como um limite de decisão[41,88,95]. Este hiperplano separa os diversos pontos dos dados entre classes e a sua dimensionalidade depende do número de *features* (2 *features* é representado por uma linha, 3 *features* por um plano, ect.). A SVM procura maximizar a margem, ou seja, a distância entre o hiperplano e os pontos de dados mais próximos para cada classe, efetivamente separando as mesmas (**Figura 19**)[41]. Estes pontos são então conhecidos como vetores de suporte, e quanto maior a

sua distância do hiperplano menor o erro de classificação e generalização[41]. O treino é realizado reduzindo a função de *loss,* penalizando classificações erradas e encorajando assim a descoberta de margens maiores. Devido a estes fatores, as SVM funcionam melhor em espaços de alta dimensionalidade, acomodando relações lineares ou não lineares dependendo dos *kernels* selecionados. Estes *kernels* são os principais hiperparâmetros a definir, podendo ser lineares, sigmoides, polinomiais, *etc.*. A SVM apresenta, no entanto, a desvantagem de ser altamente sensível ao ruído estatístico e apresenta piores resultados em classes bastante sobreponíveis. A mesma é também extremamente sensível à escolha de *kernels* e de hiperparâmetros[41].

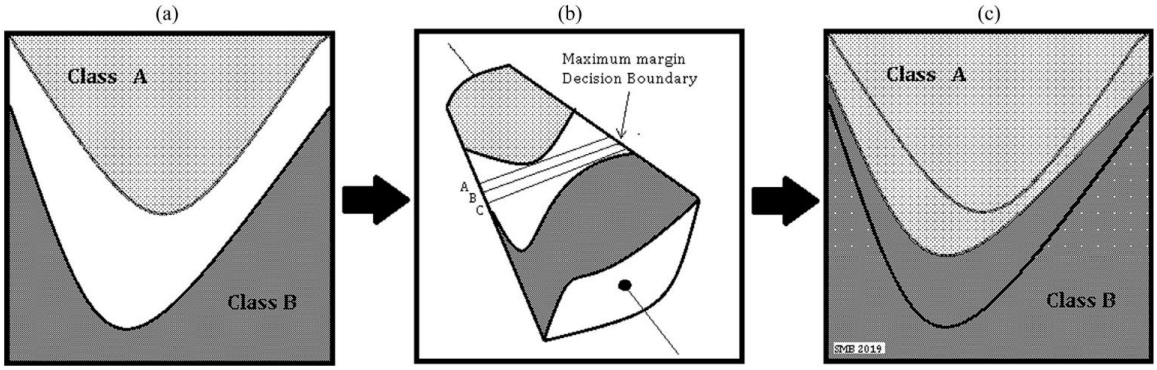


Figura 19 *– Representação do funcionamento de uma SVM. (a) representa duas classes numa superfície bidimensional que não podem ser separadas por uma linha reta. (b) representa a projeção SVM dos dados em num hiperespaço de dimensão superior onde existe uma fronteira de decisão linear com uma margem definida. As linhas A e C representam os vetores de suporte, e B a linha de decisão máxima entre eles. (c) a SVM devolve a região de decisão de volta à superfície original bidimensional. Adaptado de S. Borstelmann (2020)*

O perceptrão multicamada (MLP) é outro classificador que pode ser utilizado em tarefas binárias. O mesmo é uma versão simplificada de uma FCL no final das CNN's. O MLP é uma rede neuronal capaz de resolver problemas complexos, devido às suas múltiplas camadas de neurónios (*nodes*) interconectados com o *input* num *output* de previsão. A sua arquitetura apresenta uma camada de *input*, uma ou mais camadas ocultas e uma camada de *output*. Os neurónios de cada camada estão conectados aos da próxima e estas conexões apresentam ponderações previamente aprendidas[40,88,95]. Cada neurónio aplica uma função de ativação ao somatório do produto das ponderações com os *inputs*, o que introduz não-linearidade ao sistema, apresentando assim o MLP capacidade de aprender relações complexas entre as variáveis (**Figura 20**) [40,88,95]. A fase de aprendizagem da MLP reside em técnicas de *foward propagation* e *backwards propagation*. Na primeira, é realizada uma previsão de um dado *input* e na segunda as ponderações são alteradas através da descida dos gradientes de forma a reduzir a função de *loss*. Os principais hiperâmetros deste classificador residem assim na regularização, número de camadas ocultas, número de neurónios, tipo de função de ativação e taxa de aprendizagem. Os MLP's são especialmente eficazes em tarefas complexas não-lineares, no entanto com poucas amostras de dados, podem não obter bons resultados e causar *overfitting* (devido à grande complexidade do modelo) [40,88,95].

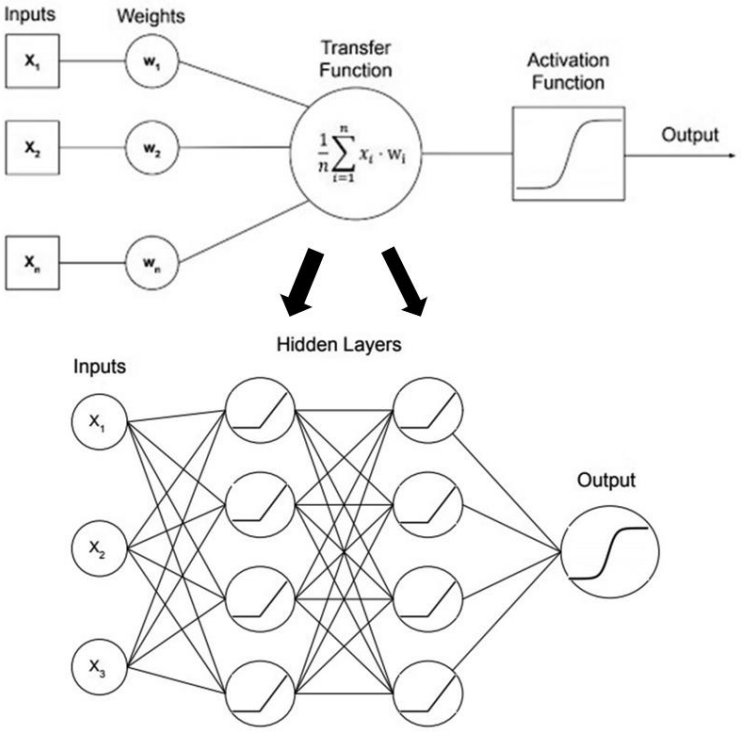


Figura 20 – *Representação de um MLP (em baixo), onde é possível verificar a arquitetura de apenas um neurónio (em cima) com a função de transferência (transfer function) a demonstrar o sumatório do produto das ponderações (W) com os inputs (x). Adaptado de W. Rogers et al. (2020)*

Finalmente existem os classificadores de *gradient boosting* (GB). Este é um método de aprendizagem combinada que utiliza sequencialmente as previsões de múltiplos classificadores "fracos", para um *boost* do desempenho do modelo. Esta melhoria é conseguida através da aplicação de ajustes, em diversas iterações, dos erros residuais de uma árvore para a seguinte[40,88,96]. Este ajuste é realizado através da aplicação do gradiente negativo da função de *loss* da previsão combinada das árvores de decisão. Os hiperparâmetros incluem assim a contribuição de cada árvore para a classificação final, o número de árvores de decisão e a sua profundidade. O método mais popular de GB está presente no XGBoost (*Extreme Gradient Boosting*) que aplica técnicas de regularização, remove ramos das árvores pouco úteis e é altamente eficiente em grandes quantidades de dados[41,88]. Apesar das aparentes vantagens deste método em tarefas complexas e não lineares, o mesmo apresenta resultados menos positivos em grupos de dados com poucas amostras e com *outliers* (existindo um foco excessivo no erro das árvores de decisão produzidas pelos mesmos)[97].

2.10.1 Calibração dos modelos probabilísticos

Um passo essencial na criação de modelos probabilísticos, por vezes ignorado, reside na calibração dos modelos de classificação probabilísticos, como a LogReg[98]. Este processo refere-se ao ajuste das probabilidades obtidas pela classificação do modelo, de forma a refletirem as verdadeiras probabilidades do evento de interesse ocorrer. Um modelo bem calibrado quantifica corretamente a incerteza associada às suas previsões, enquanto que, um modelo mal calibrado apresenta sub-confiança e/ou sobre-confiança nos seus resultados[98,99]. O método usualmente utilizado para a avaliação da calibração de modelos de ML, depende da criação de curvas de calibração. Para a criação destes gráficos agrupam-se grupos de probabilidades produzidos pelos modelos e calcula-se a proporção de verdadeiros positivos nesses grupos de amostras. Posteriormente, a média das probabilidades é mapeada em relação à proporção de verdadeiros positivos no grupo. Um modelo perfeitamente calibrado iria apresentar uma reta diagonal onde f(x) = y [98,99]. Modelos que apresentem a sua curva superior a essa reta mostram subconfiança na região probabilística, enquanto que a representação inferior à reta de referência, demonstra sobreconfiança (**Figura 21**)[98].

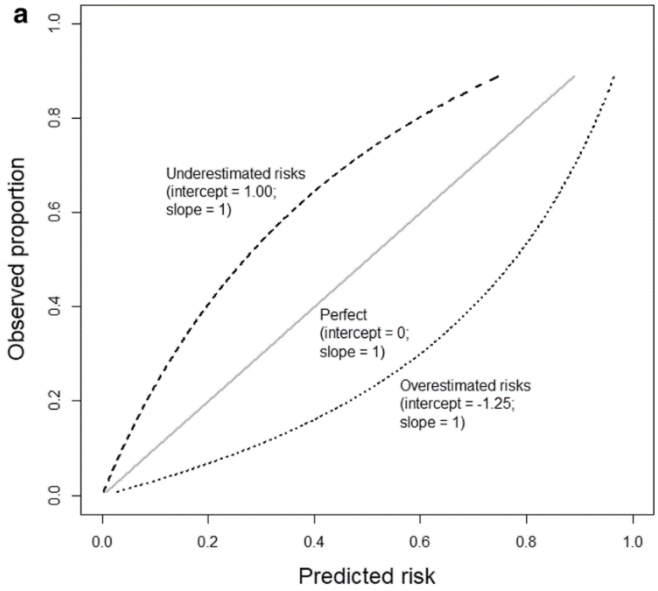


Figura 21 *– Curva de calibração que demonstra o modelo perfeito (Perfect), um modelo sobreconfiante (Overestimated risks) e um modelo subconfiante (Underestimated risks). Adaptado de B. Calster et al. (2019)*

Quanto aos possíveis métodos de calibração para classificações binárias, a regressão isotónica é uma possibilidade. Este método é não-paramétrico e ajusta as probabilidades para que as mesmas sejam "não decrescente", ao longo de diferentes intervalos de confiança ou de grupos, preservando a ordem relativa das probabilidades previstas. Este método é flexível e utilizada para calibrações complexas[99]. Modelos sigmoides (*Plat Scalling*) são, no entanto, os mais comuns para classificações binárias. Estes métodos utilizam a regressão logística aplicado aos *outputs* de um classificador, comprimindo as probabilidades entre valores de 0 a 1. Apesar de ser simples, computacionalmente eficiente e útil em SVM's, este método promove uma curva sigmoide que pode levar à sobre-confiança, sendo também menos útil em classificadores probabilísticos por natureza [99].

2.11 Métricas de *performance* e interpretação do modelo

A avaliação dos modelos treinados é um passo essencial em qualquer tarefa de ML e DL, tanto em termos de validação, como de teste. As diferentes métricas indicam os comportamentos dos modelos e averiguam também a existência de *overfitting* ou *underfitting*.[100] As métricas utilizadas nesta dissertação encontram-se resumidas nesta secção.

Exatidão (*Accuracy ou* CA) é definida como a razão entre o número de amostras corretamente classificadas e o total de amostras avaliadas. Esta proporção é muito usada no âmbito de medicina, onde um valor de CA de 1 representa uma classificação perfeita de todas as amostras de um grupo, enquanto uma CA de 0 representa uma totalidade de classificações incorretas por parte do modelo[100].

$$Accuracy = \frac{\#\ amostras\ corretamente\ classificadas}{\#\ todas\ as\ amostras} = \frac{VP + VN}{VP + FP + VN + FN}$$

Equação 7 – Fórmula de cálculo da Accuracy

Em que:

VP: Verdadeiro positivo – Número de amostras positivas corretamente identificadas.

VN: Verdadeiro positivo – Número de amostras negativas corretamente identificadas.

FP: Falso positivo – Número de amostras incorretamente classificadas como positivas.

FN: Falso negativo – Número de amostras incorretamente classificadas como negativas.

Recall (sensibilidade) é outra métrica utilizada para avaliar a taxa de sucesso do modelo em devolver amostras corretamente classificadas da classe positiva considerando todas as amostras positivas. Esta utiliza a razão entre o número de amostras corretamente classificadas como positivas e todas as amostras verdadeirametne associadas a esta classe. Esta métrica revela a capacidade do modelo detetar corretamente a classe positiva de interesse [100].

$$Recall = \frac{\#\ amostras\ classificadas\ corretamente\ como\ positivas}{\#\ amostras\ verdadeiramente\ positivas} = \frac{VP}{VP + FN}$$

Equação 8 – Fórmula de cálculo de Recall

Especificidade (Spec) avalia a capacidade do modelo de classificar a classe negativa corretamente, evitando falsos positivos. Esta métrica é dada pela proporção de VN sobre todas as amostras negativas em termos de *ground truth label*[100].

$$Specificity = \frac{\#\ amostras\ classificadas\ corretamente\ como\ negativas}{\#\ amostras\ verdadeiramente\ negativas} = \frac{VN}{VN + FP}$$

Equação 9 – Fórmula de cálculo da Especificidade

Precisão (*Prec*) avalia a exatidão das classificações positivas do modelo. O seu cálculo é efetuado a partir da razão de amostras corretamente classificadas como positivas e todas amostras que o modelo classificou como positivas [100].

$$Precision = \frac{\#\ amostras\ corretamente\ classificadas\ como\ positivas}{\#\ amostras\ classificadas\ como\ positivas} = \frac{VP}{VP + FP}$$

Equação 10 – Fórmula de cálculo da Precision

Por fim, existe um indicador para sistemas de classificação denominado de F1 *score* (F1). F1 relaciona a *Prec* e R*ecall* através da média harmónica das mesmas. Esta métrica é importante para identificação do equilíbrio entre as duas métricas e é especialmente importante na indicação do *performance* em casos de desequilíbrio entre as classes de interesse e em tarefas onde a classificação de FP e FN tem consequências sérias[100].

$$F1\ score\ (F1) = 2 \times \frac{Prec \times recall}{Prec + recall} = \frac{2 \times VP}{2 \times VP + FP + FN}$$

Equação 11 – Fórmula de cálculo de F1 score

Para além das métricas apresentadas a avaliação da curva ROC (*Receiver Operating Characteristic*) e da área sob a mesma (AUC) é essencial em modelos de classificação binários. Esta curva verifica o desempenho do modelo nos diversos *thresholds* de decisão possíveis. Esta curva reflete a *performance* do modelo ao longo de todo o seu eixo de operação. A mesma é mapeada através da *recall* e da taxa de falsos positivos (1-Spec) onde cada ponto da curva representa um possível *threshold* [25,101,102]. Um modelo perfeito passaria pelo canto superior esquerdo do gráfico, indicando 100% de *recall* e 0% de falsos positivos (**Figura 22**). Desta forma a AUC avalia numericamente a qualidade global do modelo, segundo a sua capacidade de distinguir as classes, sendo importante em casos de desequilibro populacional entre as mesmas e na comparação de modelos [25,101,102]. No caso de ser usada CV, é recomendado que a mesma seja calculada utilizando a concatenação de todas a previsões, segundo *T.Fawcett* [101].

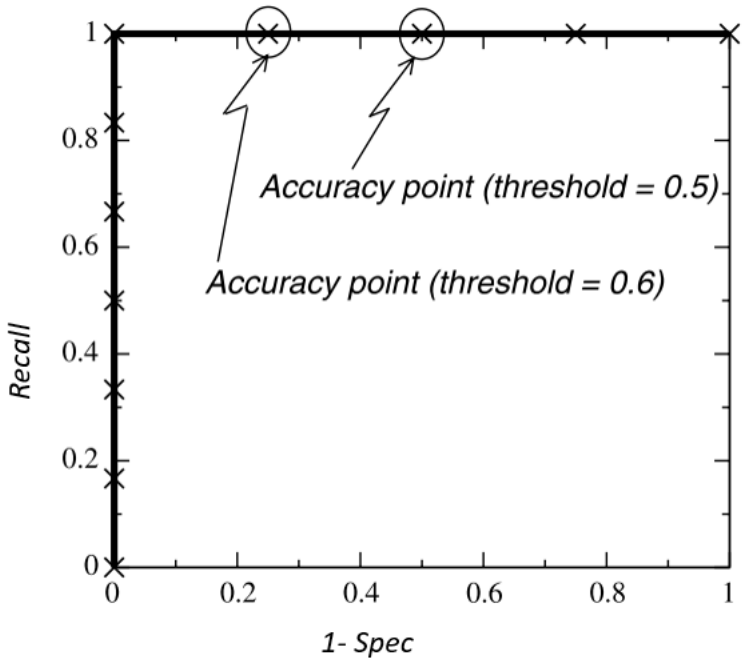


Figura 22 *- Exemplo da Curva ROC perfeita com uma AUC = 1. Adaptado de T. Fawcett (2006)*

2.11.1 Interpretação dos modelos de classificação

A interpretação clínica de modelos de *radiomics* (quer seja de ML ou DL) é essencial para garantir confiança no algoritmo e para fomentar aprendizagem médica com os mesmos. Os modelos de ML são tipicamente modelos de *white-box*, ou seja transparentes[103]. Estes modelos são facilmente interpretáveis matematicamente e a sua lógica interna pode ser compreendida, através das ponderações dadas a cada *feature* e dos parâmetros utilizados. É facilmente compreendido como o modelo de classificação chegou a tal decisão de classificação. Os modelos de DL, por outro lado, são considerados modelos de *black-box* visto que apresentam grande complexidade e níveis de abstração consideráveis[103]. Os mecanismos internos não facilmente interpretados fugindo à intuição humana (o que os torna úteis em tarefas complexas como reconhecimento de imagem). No entanto, existem diversas técnicas que poderão ajudar na interpretação de modelos.

Os valores de SHAP (*Shapley Additive exPlanations*), são frequentemente utilizados na explicação de modelos de ML tradicionais. Estes valores baseiam-se na teoria de jogo cooperativo e distribuem o grau de importância de uma determina *feature*, de forma justa, de acordo com a sua margem de contributo para uma previsão[104]. Esta metodologia é utilizada por diversos autores para obter explicações das lógicas em modelos de classificação de imagem, através do relação dos valores de *features* com o seu contributo para uma determinada classe[105–107]. As bibliotecas de *Python* dedicadas à construção destes valores, possuem gráficos que resumem o comportamento expectável dos modelos e o valor de cada *feature*.

Em modelos de DL, principalmente em CNN's, a tarefa de interpretação é desafiante. Os mapas de ativação de classe por ponderação dos gradientes (Grad-CAM) são os métodos mais utilizados para compreensão visual da decisão das CNN's. Esta técnica produz *heat-maps* na imagem original que focam as zonas de maior relevância para a decisão do modelo numa determinada classe[83]. Como o nome indica, estes métodos dependem da existência de uma classe ativa (classificada) e da sua probabilidade classificação. Seguidamente os gradientes da probabildiade em relação aos mapas de *features* da última CL são calculados, sofrendo os mesmos *pooling* da média global para obter as ponderações de importância de cada mapa de acordo com a classe classificada e gerar a assim um mapa de atenção[83]. Apesar da utilidade desta técnica, a mesma requer uma classe ativada através de uma FCL diretamente conectada à última LC, sendo difícil a sua utilização em situações de *transfer learning* híbrido por extração de *features*.

2.12 Estudos relacionados e ferramentas a utilizar

Apesar do tópico de interesse, na revisão bibliográfica não foram encontrados estudos de previsão de mortalidade ou de grupos de risco especificamente para doentes ARDS-COV19, sendo esta uma vantagem da dissertação atual em termos de homogeneidade populacional. No entanto foi possível encontrar estudos semelhantes considerando previsão de mortalidade em doentes COVID-19 internados em UCI, sobre VMI e em departamentos de urgência. Nesta secção também serão descritas as ferramentas de *software* a utilizar na metodologia.

2.12.1 Estudos relacionados

D. Gourdeau et al utiliza metodologias híbridas de extração de *features* (*transfer learning*) de radiografias para previsão da mortalidade em doentes COVID-19 sob VMI nas UCI[23]. Os autores utilizaram uma CheXNet pré-treinada para a extração de *features* de radiografias centradas na região pulmonar (técnica de *cropping*) para consequente treino de uma SVM. As variáveis clínicas utilizadas basearam-se na aglomeração de comorbilidades e informações demográficas dos doentes num único *score* de risco de mortalidade. Apesar de apenas ser utilizada uma base de dados de 160 doentes, os autores consideraram como *input* dos modelos as diversas radiografias disponíveis dos mesmos, para um total de 278 R-RTX para treino e 184 R-RTX para teste. Isto significa que o modelo foi treinado, utilizando múltiplas radiografias dos mesmos doentes como amostras independentes, existindo um risco de *viés* em validação externa por pouco potencial de generalidade. As 1024 *features* de imagens extraídas sofreram filtragem utilizando uma metodologia de *ranking* por o valor F do teste de ANOVA, resultando em duas *features* de imagem de *deep learning* de interesse. Os autores verificaram melhor AUC, CA, e F1 no modelo que combinou as *features* de imagem com os dados clínicos. Apesar de existir grande probabilidade dos doentes submetidos a VMI apresentarem ARDS-COV19, esta condição não foi referida pelos autores, não existindo dados suficientes para comprovar o mesmo, segundo a definição de Berlim

J. Cheng et al. desenvolveram um método de previsão de mortalidade em doentes COVID-19 internados em UCI, utilizando 422 doentes para treino (5645 R-RTX) e 108 doentes de teste externo[108]. Os autores utilizam uma arquitetura complexa de *deep learning* treinada de raiz

para extração de *features* longitudinais das radiografias previamente segmentadas. Em primeiro lugar, uma sequência de radiografias é utilizada como *input* de uma CNN *Resnet*-50 para extração de *features* e posterior codificação por *vision transformers* (dividem os mapas de *features* em secções para obter representações vetoriais que consideram relações locais e globais). Estas *features* foram então combinadas por *pooling* de valores médios e máximos com posterior utilização de duas FCL para uma *output* binário. As variáveis clínicas utilizadas correspondem a comorbilidades, informações demográficas e análises clínicas/gasometria, não referenciando a razão pf nem parâmetros de ventilação. Estas variáveis são usadas para treinar uma MLP com três camadas densas para *output* binário. O modelo combinado tem em conta a soma ponderada dos *outputs* de ambos os modelos, para uma classificação final. Os autores criaram diversos modelos para teste e verificaram que a introdução de informação imagiológica melhorou a *performance* do modelo no grupo de teste externo, principalmente com maior componente longitudinal.

Jiao et al. por outro lado pretenderam avaliar a severidade da COVID-19 no momento da admissão hospitalar no departamento de urgência utilizando metodologia de *deep learning*, diferenciando do objetivo da atual dissertação [109]. Os autores realizaram uma classificação binária entre doentes críticos e não-críticos, onde o parâmetro da classe positiva (críticos) ficou definido pela morte ou admissão na UCI. Os autores utilizaram um total de 1285 doentes para treino, 183 para validação e 366 para teste interno. Os mesmos também apresentaram um grupo de teste externo de 475 doentes. Para cada doente, foram extraídas R-RTX adquiridas em contexto de urgência hospitalar e variáveis clínicas baseadas em informações demográficas, análise clínicas e comorbilidades. A razão pf não foi mencionada neste estudo. Os autores utilizam uma CNN *EfficientNet* pré-treinada com a base de dados *ImageNet* (*transfer learning por fine tuning* total) e uma FCL para a classificação binária. Para as variáveis clínicas foi utilizada uma MLP com três camadas densas para o mesmo efeito. O modelo combinado baseou-se na soma ponderada de ambas as classificações para obter uma classificação final. Estes modelos não são assim treinados utilizando ambas as *features* de imagem e as variáveis clínicas na sua "raiz", não considerado as possíveis interações entre as mesmas. Com semelhança aos outros estudos mencionados, os autores verificarem uma melhoria significativa na aplicação de *features* das radiografias na *performance* dos modelos.

Os resultados obtidos na presente dissertação são comparados com estes estudos na secção 5.2 (Discussão).

2.12.2 Ferramentas utilizadas

Para a criação dos modelos de previsão de mortalidade foram utilizados uma diversidade de bibliotecas de *python* (para processamento de imagem e criação de CNN's) e *softwares* estatísticos, de *data mining* e de construção de modelos.

O OpenCV foi utilizada para o processamento de imagem e destaca-se como uma biblioteca notável no campo da visão computacional, oferecendo uma ampla gama de funcionalidades[110]. Esta disponibiliza recursos como filtragem, transformações geométricas e deteção de contornos em imagens. Adicionalmente, a sua capacidade de integração com módulos de ML e DL torna-a uma escolha vantajosa para implementação de algoritmos nesse

domínio. O *software* de processamento de imagem *ImageJ* foi também utilizada para efeitos de criação de máscaras e segmentação pulmonar[111].

A plataforma *Orange Data Mining* surge como uma solução abrangente para análise de dados[112]. O seu conjunto de ferramentas abrange desde visualização de dados, com utilização de gráficos interativos, e até técnicas avançadas de modelagem preditiva. Este efeito é conseguido através da conexão de *widgets* que representam diferentes funções. O mesmo oferece também recursos para pré-processamento de dados, incluindo procedimentos de limpeza e transformação de conjuntos de dados, sendo uma escolha determinante na etapa preparatória de análise estatística. A capacidade de permitir análise de *data mining*, revelando padrões e *insights*, adiciona um componente significativo à sua utilidade.

A API *Keras* foi utilizada para a construção das CNN's e extração de *features* das R-RTX[113]. Esta é uma biblioteca comum e simples para a programação de redes neurais na linguagem *Python*, simplificando consideravelmente o processo de construção e experimentação com modelos de aprendizagem profunda. A sua abstração facilita a criação de redes neurais complexas, utilizando conceitos como camadas, funções de ativação e otimizadores. A flexibilidade e a integração com *backends* populares, como TensorFlow e Theano, ampliam a sua aplicabilidade em diversas tarefas, desde classificação de imagem até processamento de linguagem natural.

O *MedCalc* foi utilizado no cálculo de intervalos de confiança das métricas de *performance*. Este destaca-se como um software estatístico especializado em dados médicos, oferecendo um conjunto diversificado de ferramentas para análise estatística e a sua interpretação. As suas funcionalidades incluem testes t, ANOVA, regressão, análise de sobrevivência, bem como cálculos específicos para avaliação diagnóstica, como análise de proporções e comparações de curvas ROC.

3. Materiais e métodos

3.1 Base de dados e estatística descritiva das variáveis clínicas

3.1.1 Descrição da base de dados

Em colaboração com a equipa médica pneumologista do departamento do tórax do Hospital de São José (Centro Hospitalar de Lisboa Norte), um estudo de coorte retrospetivo foi realizado considerando todos os doentes com ARDS-COV19 que foram submetidos VMI e admitidos nas UCI's do Hospital São José e do Hospital Curry Cabral entre abril de 2020 e janeiro de 2021.

Os critérios de inclusão para o estudo foram a confirmação da infeção por SARSCoV-2 através de teste de PCR de nasofaringe ou PCR de lavagem bronco-alveolar, o diagnóstico de ARDS de acordo com a definição de Berlim e a necessidade de ventilação mecânica invasiva. Os critérios de exclusão incluíram idade inferior a 18 anos, gravidez, outras causas contribuintes para o ARDS, como traumatismo, infeção respiratória sincrónica com outros agentes, morte nas primeiras 48 horas de internamento na unidade de cuidados intensivos e falta de dados sobre as configurações de ventilação ou R-RTX de qualidade limitada. É importante referir que os pacientes em oxigenação por membrana extracorpórea (ECMO) também foram excluídos desta análise, devido à preocupação de que as estratégias de ventilação de repouso pulmonar, poderiam fornecer dados de ventilação consideravelmente distintos, não sendo comparáveis aos pacientes que não estavam sobre ECMO. Os dados dos doentes referem-se às primeiras 72 horas de internamento na unidade de cuidados intensivos, período em que o SARSCoV-2 foi o único agente infecioso isolado, ao contrário do que ocorreu mais tarde durante o internamento prolongado na unidade de cuidados intensivos, quando outros agentes foram isolados no contexto de pneumonia associada à ventilação. Existe também um potencial viés ao excluir pacientes em ECMO, uma vez que provavelmente seriam mais jovens, no entanto, devido ao grande número de pacientes durante o pico da pandemia, a idade média da nossa coorte foi de 63,2±11,92 anos.

No início da investigação foram adquiridos 89 doentes, com exclusão de dois por apresentarem R-RTX com artefactos metálicos e/ou parênquima pulmonar "cortado", como referido na secção 2.5.2, para um total de 87 doentes. Estas amostras serviram para o treino e validação do algoritmo. Na fase terminal da investigação, 23 doentes foram recolhidos para o teste do modelo, servindo assim de grupo de teste interno. Esta divisão permite uma proporção entre grupo de treino e grupo teste, de 8:2, o que corresponde ao recomendado pela literatura (capítulo 2.5.3).

No total, foram assim considerados 110 doentes, com uma idade média de 63,2±11,92 anos (mínimo: 26, máximo: 83). Existem mais doentes do sexo masculino do que do feminino, correspondendo a 61,2% (68) do total de doentes, onde 25,37% (17) dos mesmos apresentavam ARDS grave (razão PaO2/FiO2 inferior a 100) *vs.* 12,20% (5) das mulheres (**Figura 23**). A mortalidade após 72 horas de internamento em UCI (incluindo após transferência hospitalar) foi de 47,3% (52) no total, servindo esta variável binária de *target* para a classificação do modelo de *machine learning* a treinar (sobrevivente/não-sobrevivente).

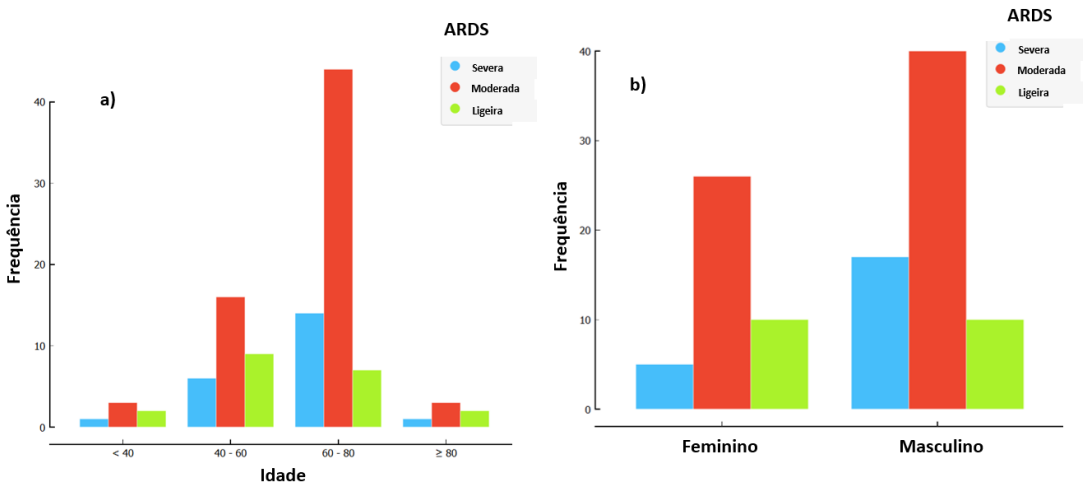
A classe não-sobrevivente foi considerada a classe positiva com codificação binária de 1, enquanto os sobreviventes foram considerados a classe negativa com codificação binária de 0. Foram recolhidos dados sobre idade, sexo, gases arteriais, como a razão pf e $PaCO_2$ (co2), e configurações da ventilação, como a peep, suporte de pressão (ps) e a subtração da peep à ps, denominado de dp (como substituto da *driving pressure*[114]). Estas variáveis foram obtidas durante o primeiro dia de ventilação mecânica invasiva (d1) e entre as 48 a 72 horas da mesma (d3), para um total de 12 variáveis/*features* (onde 10 são variáveis clínicas de d1 e d3) e 1 *target* (mortalidade).

R-TRX portáteis de incidência AP de d1 e d3 para cada doente foram também recolhidas. As mesmas foram descarregadas em formato PNG e com os *headers* de informação pessoal de DICOM removidas, para um total de 220 R-RTX. As imagens foram pseudo-anonimizadas com um valor de 0 a 110 para identificação da amostragem correspondente na base de dados em formato de CSV (110x13).

3.1.2 Análise descritiva univariada da base de dados

Nas **Tabela 1**,**Tabela 2** e **Tabela 3**, é possível verificar a estatística descritiva da base de dados total, de treino e teste (apenas considerando as informações clínicas fornecidas). Para averiguar a existência de diferenças significativas entre as variáveis do grupo de treino e de teste, o *students t-test* foi utilizado. A hipótese nula (H_0), ou seja, ambas as médias são estatisticamente iguais, foi rejeitada considerando um valor $p \leq 0,05$. A análise foi realizada utilizando o *software Orange Data Mining*, após importação do ficheiro CSV com os seguintes *widgets*:

- ***Distributions,*** que permite obter gráficos de frequência de variáveis discretas ou contínuas (através de *binning* das mesmas) e fazer adaptação dos dados a diferentes tipos de distribuição. Foi selecionada a distribuição normal, obtendo a média e desvio padrão de cada variável.
- ***Feature Statistics,*** onde é possível observar a média, mediana, moda, mínimos, máximos e gráfico de distribuição de frequências por classe de *target.*



(pf_d1). a) Gráfico de frequência da severidade da ARDS em cada grupo etário. b) Gráfico de frequências da severidade da ARDS em cada sexo

Tabela 1 – *Estatística descritiva da base de dados total.* ***n=****número de amostras;* ***NS=****Não-Sobrevivente;* ***Méd.±d.p*** *= Média ± Desvio Padrão;* ***Min:Max*** *= Valores mínimos e máximos da variável;* ***C_v****=coeficiente de variação;* ***Falta(%)*** *= Percentagem de amostras em falta. A negrito encontra-se a classe target*

Variáveis	**Categóricas (n=110)**					
Mortalidade (*target: NS)*	**Não-Sobrevivente**		**n=52**		**47,3%**	
	Sobrevivente		n=58		52,7%	
Sexo	Masculino		n=68		61,2%	
	Feminino		n=42		38,18%	
Contínuas						
	Méd.±d.p	**Mediana**	**Moda**	**Min:Max**	**C_v**	**Falta (%)**
Idade	63.2 ±11,87	64	64	26:83	0,19	0
pf_d1	148,50± 52,92	147,5	150	51:297	0,36	2 (2%)
pf_d3	163,73±53.07	160	118	52:365	0,32	1(1%)
ps_d1	16,22±3,19	16	16	8:26	0,20	3(3%)
ps_d3	15,49±3,50	18	16	0:26	0,23	7(6%)
peep_d1	12,76±3,29	14	14	6:20	0,26	0(0%)
peep_d3	11,61±3,49	12	12	0:22	0,30	5(5%)
dp_d1	3,48 ±5,14	4	0	-6:18	1,48	3(3%)
dp_d3	3,81± 4,23	4	6	-6:14	1,11	7(6%)
co2_d1	45,84± 11,36	44,6	46	18,1:89	0,25	1(1%)
co2_d3	47,32± 9,61	46,2	40,3	26,4:90,6	0,20	2(2%)

Tabela 2 – *Estatística descritiva da base de dados de treino.* ***n=****número de amostras;* ***NS=****Não-Sobrevivente;* ***Méd.±d.p*** *= Média ± Desvio Padrão;* ***Min:Max*** *= Valores mínimos e máximos da variável;* ***C_v****=coeficiente de variação;* ***Falta(%)*** *= Percentagem de amostras em falta.*

Variáveis	**Categóricas (n=87)**					
Mortalidade (*target: NS)*	**Não-Sobrevivente**		**n=42**		**48,28%**	
	Sobrevivente		n=45		51,72%	
Sexo	Masculino		n=54		62,07%	
	Feminino		n=33		37,93%	
Contínuas						
	Méd.±d.p	**Mediana**	**Moda**	**Min:Max**	**C_v**	**Falta (%)**
Idade	64 ±11,6	64	64	26:83	0.18	0
pf_d1	150,41± 50,66	150	150	51:297	0.34	2 (2%)
pf_d3	163,13±50.71	159	118	52:365	0.31	1(1%)
ps_d1	15,99±3,03	16	16	10:26	0.19	3(3%)
ps_d3	15,40±3,28	16	16	5:26	0.21	5(6%)
peep_d1	12,76±3,30	14	14	6:20	0.26	0(0%)
peep_d3	11,63±3,33	12	12	3:22	0.29	4(5%)
dp_d1	3,25 ±5,15	3	6	-6:17	1.59	3(3%)
dp_d3	3,76± 4,31	4	6	-6:14	1.15	5(6%)

co2_d1	46,32± 11,53	44.9	46	26:89	0.25	1(1%)
co2_d3	46,53± 9,053	45.8	40.3	26,4:78,0	0.19	2(2%)

Tabela 3 - *Estatística descritiva da base de dados de teste.* ***n=****número de amostras;* ***NS=****Não-Sobrevivente;* ***Méd.±d.p*** *= Média ± Desvio Padrão;* ***Min:Max*** *= Valores mínimos e máximos da variável;* ***C_v****=coeficiente de variação;* ***Falta(%)*** *= Percentagem de amostras em falta.*

Variáveis	**Categóricas (n=23)**			
Mortalidade (*target: NS)*	**Não-Sobrevivente**	**n=10**	**43,48%**	
	Sobrevivente	n=13	56,52%	
Sexo	Masculino	n=14	60,87%	
	Feminino	n=9	39,13%	
Contínuas				
	Méd.±d.p	**Mediana**	**Moda**	**Min:Max**	**C_v**	**Falta (%)**
Idade	60.17 ±12,29	61	67	31:83	0.20	0
pf_d1	141,48± 60,01	128	92	63:297	0.42	0
pf_d3	165,96±61.05	162	182	68:286	0.37	0
ps_d1	17,09±3,60	17	18	8:24	0.21	0
ps_d3	15,81±4,24	17	18	0:20	0.27	2(9%)
peep_d1	12,78±3,23	14	14	6:18	0.25	0
peep_d3	11,55±4,04	12	12	0:18	0.35	1(4%)
dp_d1	4,30± 5,01	4	4	-318	1.16	0
dp_d3	4± 3,92	2	2	-2:12	0.98	2(9%)
co2_d1	44,06± 10,56	43.4	18.1	18,1:70	0.24	0
co2_d3	50,243± 10,93	48	36.1	36,1:90,6	0.22	0

3.1.3 Análise exploratória multivariada dos dados de treino

A análise estatística multivariada foi realizada apenas no grupo de dados de treino (n=87), pois este será o grupo que os diferentes algoritmos de classificação irão utilizar para realizar as suas previsões. Esta análise é essencial para garantir a integridade do processo de avaliação das variáveis e das suas relações com as classes, evitando o *leaking* dos dados de teste e a aplicação de conhecimentos globalizados nos modelos finais (*overfitting)*. A análise multivariada foi realizada utilizando o *software Orange Data Mining*, aplicando os seguintes *widgets*:

- ***Box Plot:*** Esta ferramenta permite a avaliação das distribuições das variáveis contínuas ou discretas (categóricas) por classe. O mesmo utiliza um diagrama de caixas representando a média, o desvio padrão, a mediana, o primeiro (25%) e o terceiro (75%) quartil. Este *widget* também verifica a existência de diferenças estatisticamente significativas variáveis dos *subgrupos* de interesse (classes). Para variáveis contínuas é utilizado o ***Student's t-test*** (*t-test*) de amostras independentes, assumindo a independência das varáveis, a sua distribuição normal e 85 graus de liberdade. Este teste verifica se a média de uma *feature* entre subgrupos é estatisticamente diferente, utilizando o valor de estatística e distribuição de *t*. O valor de *p* é também calculado para efeitos de rejeição da hipótese nula (H_0), ou seja, da não existência significativa de diferenças das médias entre os subgrupos. As variáveis categóricas são avaliadas utilizando o teste de *chi-squared* (x_2), com 1 grau de liberdade, que estuda a independência das distribuições de frequências das mesmas proporcionado também um valor de *p*. Foi considerado um valor de $p \leq 0.05$ para rejeição de H_0.
- ***Scatter Plot*****:** Este *widget* permite a criação de um gráfico bidimensional de correlação multivariável. Cada amostra é mapeada de acordo com os valores de duas variáveis (uma representada no eixo X e outro representada no eixo Y). As amostras são codificadas por cor dependendo da classe a que pertencem (tipicamente o *target*), formando assim regiões coloridas que descrevem a capacidade das variáveis selecionadas separarem as classes das amostras. Esta ferramenta permite encontrar as melhores projeções de variáveis discriminantes da classe, através de um sistema de *scoring e ranking*. Estes *scores* são calculados através do agrupamento dos dez vizinhos mais próximos de cada combinação de variáveis, verificando quantos grupos tem apenas amostras da mesma cor (classe). A avaliação final é realizada através do cálculo da média do número de vizinhos da mesma cor por par de variáveis.
- ***Correlations*****:** Esta ferramenta permite o cálculo dos coeficientes de correlação de *Pearson* entre as variáveis selecionadas, ordenando as mesmas. A interpretação usual da medicina dos mesmos foi utilizada [115]. Neste caso correlações com um coeficiente superior ou igual +0,8/-0,8 são consideradas "fortes", com um coeficiente superior ou igual a +0,6/-0,6 e inferior a +0,8/-0,8 são consideradas moderadas, com um coeficiente superior ou igual a +0,3/-0,3 e inferior a +0,6/-0,6 são consideradas ligeiras e com um coeficiente inferior a +0,3/-0,3 são consideradas fracas.

Tabela 4 – *Estatística bivariada do grupo de treino* em relação à mortalidade. A negrito encontram-se as variáveis com diferenças estatisticamente significativas (*p≤0,05), utilizando o student's t-test.* ***n****=número de amostras;* ***Méd.±d.p*** *= Média ± Desvio Padrão*

	Não sobrevivente (n=42)	**Sobrevivente (n=45)**		
Feature	**Méd.±d.p**	**Méd.±d.p**	**t-test *score***	**Valor *p***
Idade	**68,64±9,4**	**59,67 ±11,8**	**3,932**	**< 0,001**
pf_d1	**138,26 ±45,61**	**162,27 ±52,49**	**2,253**	**0.027**
pf_d3	**146,67 ±43,52**	**178,84 ±43,52**	**3,114**	**0,003**
ps_d1	16±3,2	15,98 ±2,8	0,034	0,973
ps_d3	14,88 ±3,1	15,90 ±3,3	1,443	0,153
peep_d1	12,40 ±3,2	13,09 ±3,4	0,972	0,334
peep_d3	11,95 ±4,0	11,29 ±2,5	0,912	0,365
dp_d1	3,62 ±5,2	2,91 ±5,1	0,636	0,526
dp_d3	3,95 ±4,7	3,55 ±3,9	0,425	0,672
co2_d1	46,74 ±11,46	45,918 ±11,57	0,331	0,741
co2_d3	47,49 ±9,89	45,584 ±8,04	0,974	0,333

Na **Tabela 4**, é possível verificar a comparação das médias das *features* contínuas entre as classes de interesse (Não-sobrevivente e Sobreviventes). São apresentados também os resultados dos testes estatísticos (*t-test*), e o seu valor *p*. A variável *sexo* não se encontra representada. As comparações da distribuição de frequências da mesma não demonstraram diferenças estatisticamente significativas na discriminação das classes (*p=0,129)* considerando um teste de x_2 com *score* de 2,30, considerando um grau de liberdade.

Na **Figura 24**, estão ilustrados os gráficos de frequência e de diagrama de caixas correspondentes às variáveis que melhor discriminam as classes, considerando as suas médias por classe.

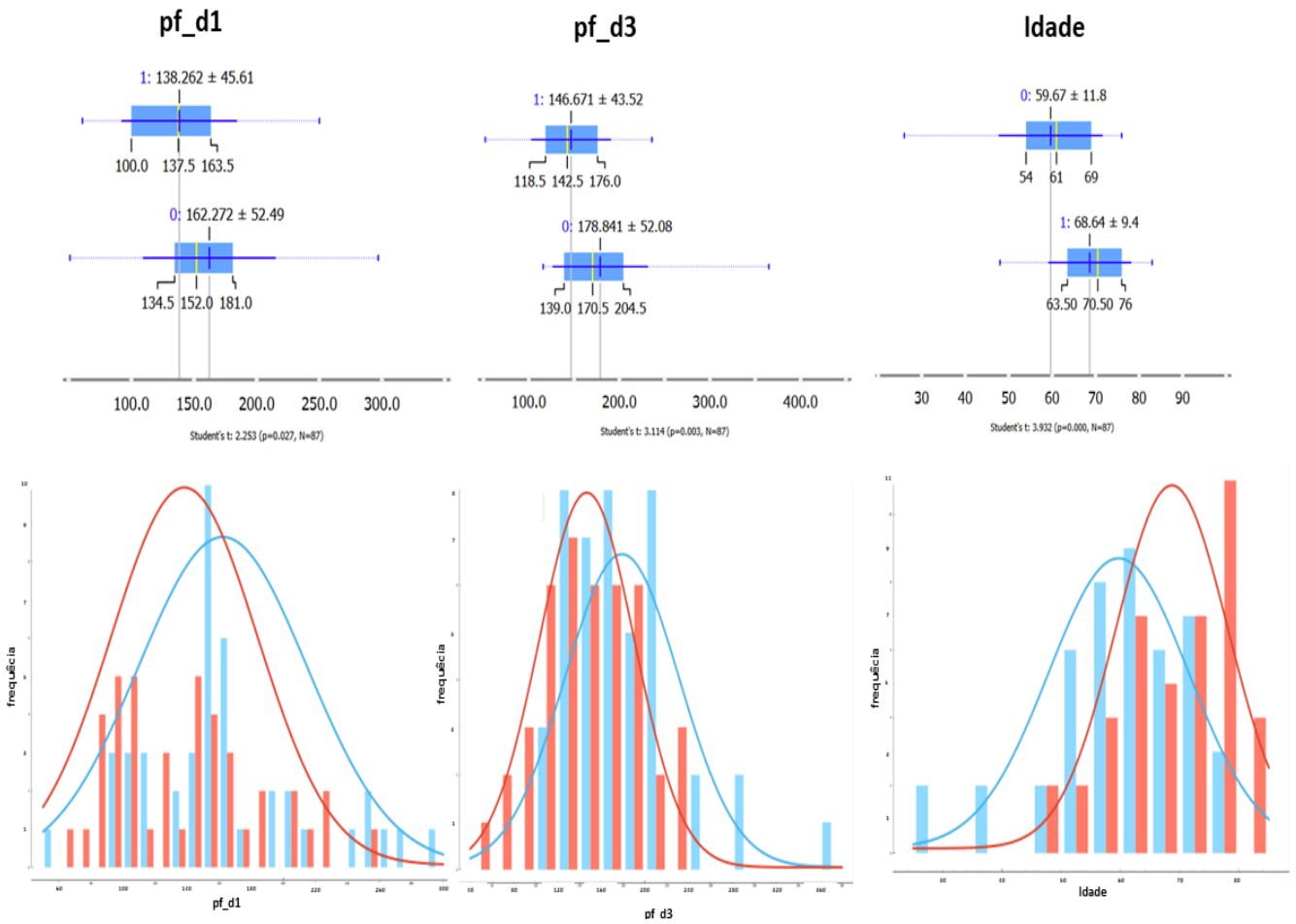


Figura 24 - *Gráficos de Diagrama (esquerda) e de frequência (direita) para cada feature cuja média discrimina significativamente a classe.* **Número 0**: Sobrevivente; **Número 1**: Não-sobrevivente; **Cor Azul:** Sobrevivente; **Cor Vermelha** Não-Sobrevivente

Na **Figura 25** é possível visualizar os *scatter plots* entre as variáveis de maior interesse aparente. É possível verificar que existe uma separação considerável entre as regiões de mortalidade das amostras. A mortalidade poderá estar associada a maiores valores de idade e menores valores de pf_d1 e pf_d3, no entanto é possível observar *outliers* desta tendência.

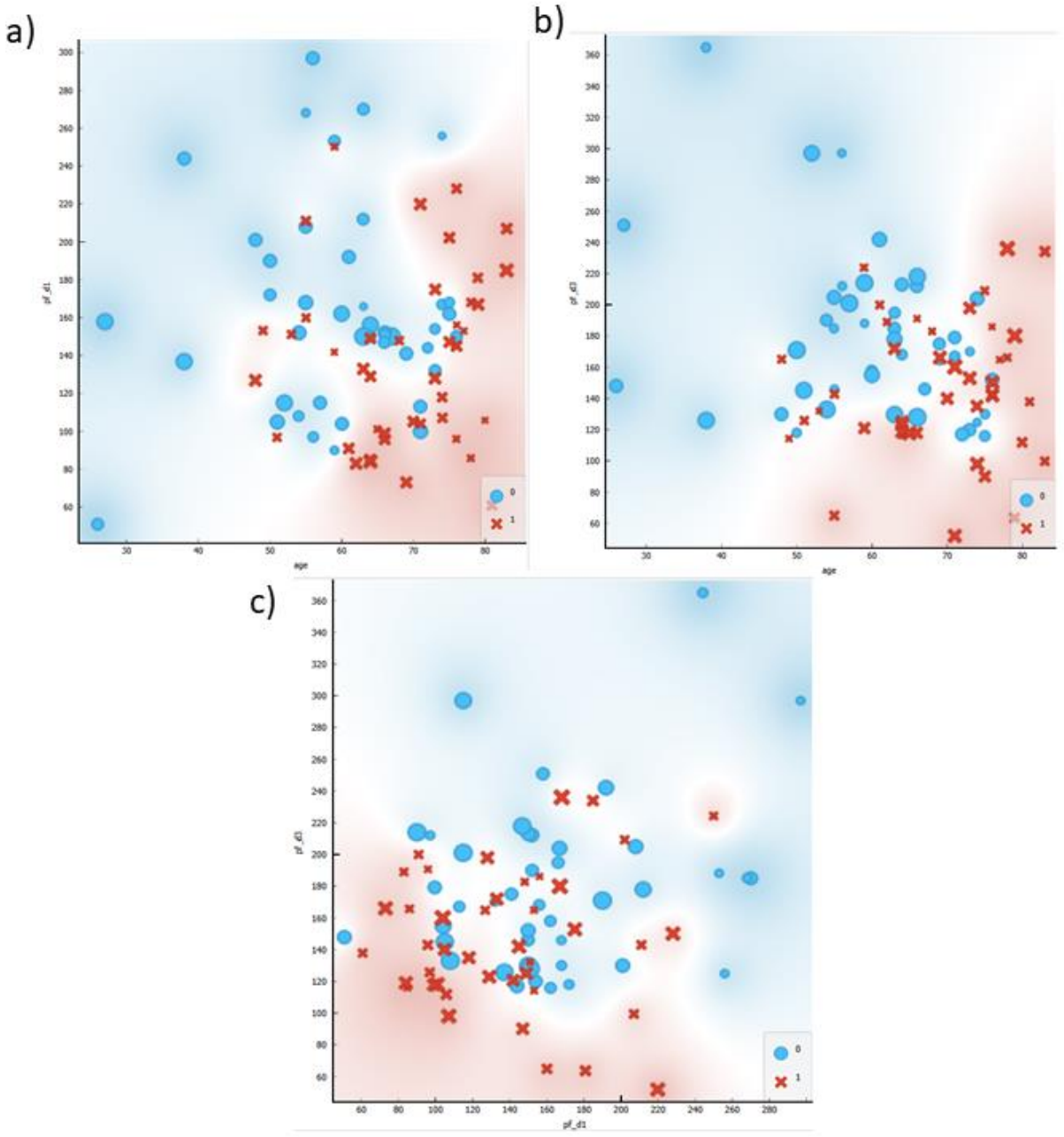
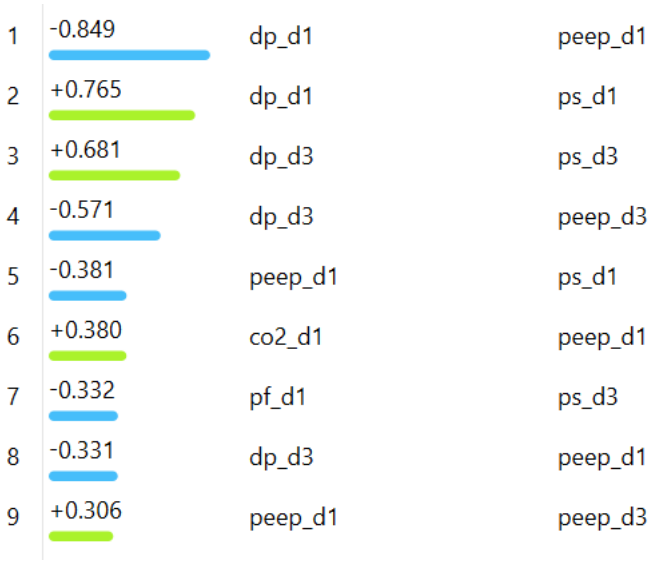


Figura 25 ***-*** *Scatter plots das amostras entre duas possíveis features de interesse.* ***a)*** *Gráfico entre Idade(age) no eixo x e pf_d1 no eixo y.* ***b)*** *Gráfico entre idade (age) no eixo x e pf_d3 no eixo y.* ***c)*** *Gráfico entre pf_d1 no eixo x e pf_d3 no eixo y. Nos três gráficos a cor vermelha, o valor um e as cruzes representam as amostras dos não-sobreviventes, enquanto a cor azul, o valor zero e os círculos representam as amostras dos sobreviventes.*

Por fim, na **Figura 26**, estão representados os coeficientes de correlação de *Pearson* entre as variáveis clínicas. São demonstradas apenas as *features* que apresentam uma correlação ligeira, moderada ou forte. É importante referir que os autores do *Orange Data Mining* não providenciam valor *p* nas correlações devido ao problema do *texas sharp shooter,* que desenha um alvo depois de ter realizado os "disparos". Em estatística a formulação da hipótese deve ser sempre o 1º passo, sendo assim necessário a recolha de novos dados para comprovar a mesma, ou neste caso, comprovar as correlações.



1	-0.849	dp_d1	peep_d1
2	+0.765	dp_d1	ps_d1
3	+0.681	dp_d3	ps_d3
4	-0.571	dp_d3	peep_d3
5	-0.381	peep_d1	ps_d1
6	+0.380	co2_d1	peep_d1
7	-0.332	pf_d1	ps_d3
8	-0.331	dp_d3	peep_d1
9	+0.306	peep_d1	peep_d3

Figura 26 – *Cálculo do Orange Data Mining dos coeficientes de correlação de Pearson entre as variáveis clínicas.* São apenas visualizados os pares que apresentam pele menos a hipótese de uma correlação ligeira

3.2 Pré-processamento das R-RTX

O processamento das radiografias foi realizado de acordo com a revisão da literatura e as metodologias descritas na secção 2.6.2.

Em primeiro lugar foi realizada a filtragem para redução do ruído e otimizado o contraste da imagem, utilizando um filtro gaussiano (*gaussian blur)* e a técnica de equalização do histograma, CLAHE. Estas técnicas foram efetuadas em *python* através da biblioteca *OpenCV* (*cv2)*. A função de pré-processamento foi definida através da aplicação da função *cv2.GaussianBlur()* e *cv2.createCLAHE()* a um *input* de uma R-RTX lida pela função *cv2.imread(),* que por sua vez permite a representação matricial de imagens PNG, em formato de *arrays* 2D de níveis de cinzentos. Na função de filtragem gaussiana, um *kernel* conservador de (3,3) foi selecionado para efeitos de convolação. Este *kernel* assume o compromisso de redução de ruído, enquanto preserva a definição e contornos da imagem original. Por sua vez, a técnica de CLAHE foi aplicada utilizando um *clip-limit* de 2, como recomendado na literatura (secção 2.6.2) em tarefas de *radiomics* de R-RTX,e um *tileGridSize* pequeno de (8,8) para captura de detalhes finos e variações locais. A função de pré-processamento foi aplicada iterativamente, utilizando um *for loop* e a função *glob(),* a todas as 220 R-RTX . O processo encontra-se exemplificado na **Figura 27**, onde é também possível observar os histogramas resultantes do pré-processamento.

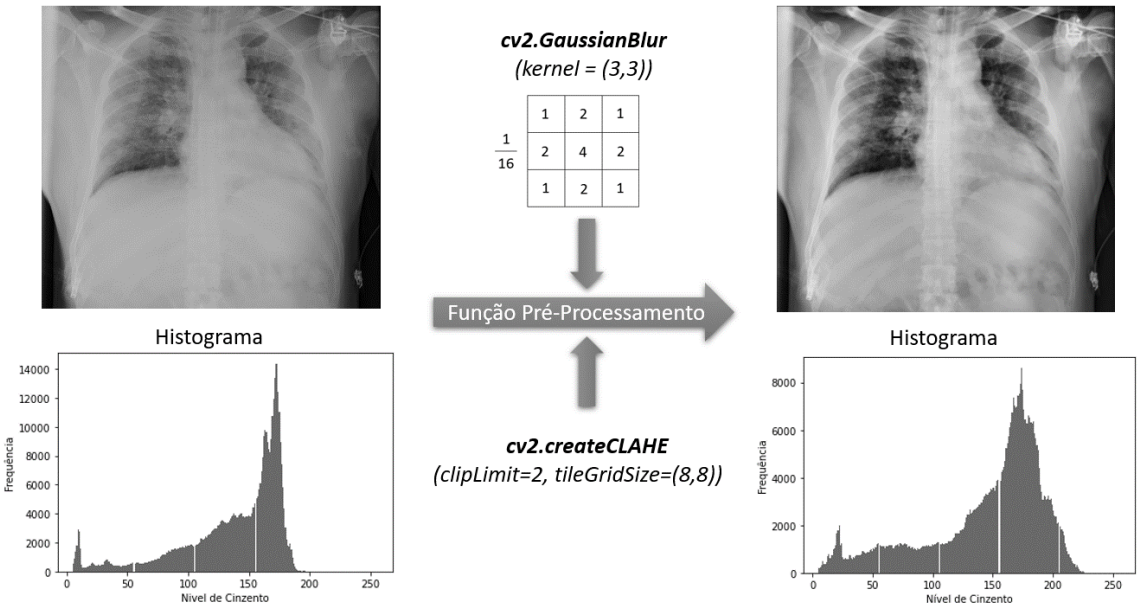


Figura 27 - *Esquematização da metodologia de pré-processamento com as R-RTX do doente nº20 do 1º dia de VMI. Os histogramas apresentados correspondem à imagem diretamente acima dos mesmos*

O segundo passo, passou pela segmentação da região de interesse (parênquima pulmonar), ou ROI, garantido a extração de *features* de imagem de interesse clínico relacionadas com a ARDS. As imagens foram segmentadas manualmente utilizando a ferramenta *open-source ImageJ*. Utilizando o método *polygonal selection*, foram desenhados os contornos do pulmão esquerdo e pulmão direito o que permite utilizar a função *create mask*. Esta função cria uma máscara onde a área da ROI é definida pelo valor de pixel 255, enquanto a restante área da

imagem é definida pelo valor de pixel 0. Durante a segmentação a silhueta cardíaca não foi considerada, com expectativa de reduzir o impacto da variabilidade de distância fonte-detetor entre as radiografias. Existe a possibilidade da mesma ser confundida com maiores ou menores infiltrados pulmonares alterando a classificação do modelo (secção 2.2.1). Sempre que possível, os artefactos metálicos foram removidos de forma razoável, existindo a potencialidade da não inclusão de parte do parênquima pulmonar. As segmentações foram validadas por profissionais de saúde (técnico de radiologia e médica pneumologista).

Após criação das máscaras, as mesmas foram aplicadas às imagens pré-processadas utilizando o operador de lógica *bitwise* AND. Foi também efetuado *cropping* à ROI correspondente da máscara. Esta experiência foi realizada por dois motivos teóricos. O primeiro: reduzir o tamanho da matriz da imagem, o que poderá ajudar na manutenção das relações inter-pixel e do *aspect ratio* da ROI quando for efetuado o redimensionamento (*downsizing*) da mesma por interpolação. Segundo: eliminar o máximo de pixéis de valor 0 fora da ROI que não têm interesse para a extração de *features* por parte da CNN. O *cropping* foi conseguido utilizando a seguinte lógica computacional em python: Foram detetadas na máscara todas as localizações de pixel (coluna_x,linha_y) que correspondem à condição (x,y) = 255 utilizando a função de *numpy np.where*(). A função retorna um *tuple* de 2 *arrays*, a primeira (0) corresponde aos valores de x da matriz, enquanto a segunda (1) contém os valores de y da mesma, correspondente a cada ponto de valor 255. Funções de min/max foram utilizadas para descobrir as localizações extremas de x e y, permitindo a construção de uma máscara que representa a mínima dimensão da matriz que inclui o parênquima pulmonar, representada por (min_x:max_x, min_y:max_y). Este processo foi executado iterativamente para todas as 220 R-RTX. A **Figura 28** esquematiza o processo da segmentação.

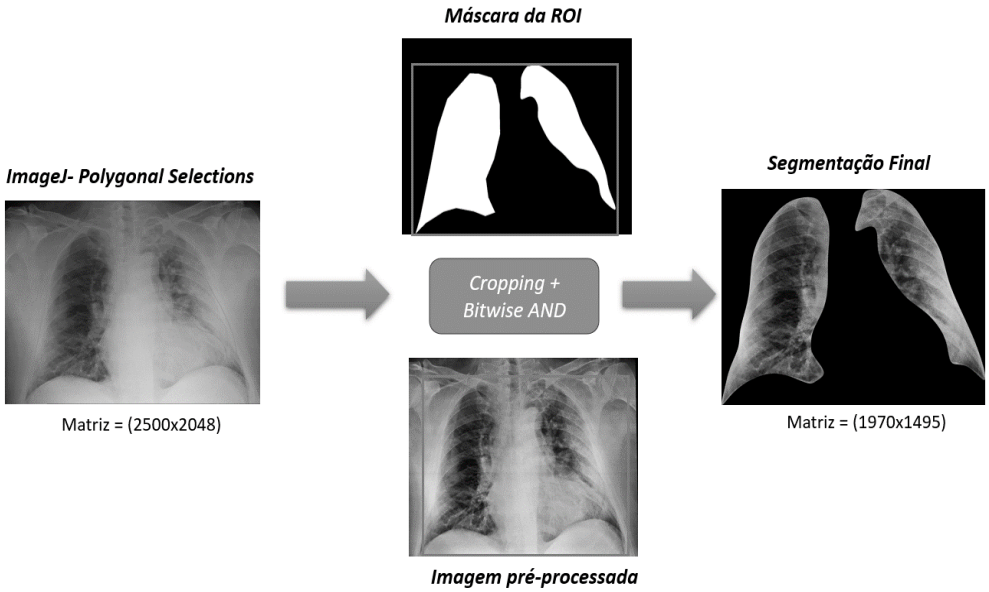


Figura 28- *Esquematização da metodologia de segmentação utilizada*

3.3 Extração de *features* de *deep learning*

Pelos motivos mencionados na secção 2.7, a extração de *features* das imagens foi realizada utilizando uma metodologia de *radiomics* baseada em *deep learning*. Esta metodologia permite uma extração automática de *features* de alto nível de abstração que podem demonstrar relações complexas em relação à previsão de mortalidade na ARDS-COV19. Devido à dimensão de dados limitada (87 amostras nos dados de treino), a metodologia de *transfer learning* pareceu ser a mais adequada. O *fine-tuning* de uma CNN pré-treinada foi considerada, no entanto, o tamanho da base dados utilizada foi substancialmente inferior à de autores que selecionaram esta técnica para previsão de mortalidade em UCI's e para a deteção de ARDS utilizando R-TRX [20,109]. O método híbrido de extração de *features* (descrito no capítulo 2.8) foi também utilizado. Esta metodologia necessita de uma CNN pré-treinada para extração de *features*, sem *fine-tuning*, para posterior classificação com algoritmos de *machine learning* clássicos.

A CNN utilizada foi a CheXNet [79]. Como referido na secção 2.7.1 esta rede foi treinada com R-RTX para a classificação de 14 patologias torácicas, apresentado sucesso nesta tarefa, principalmente na identificação de pneumonia. Para além do claro possível benefício da utilização da mesma rede para a tarefa atual, a CheXNet oferece a vantagem de ser *open-source* com múltiplas replicações na literatura de *transfer learning,* por exemplo*,* em concursos de *machine leaning* da plataforma *Kaggle.* B.Chou e M.Lee criaram em 2020 um repositório no *website Github* que permite a implementação acessível da CheXNet através da biblioteca *Keras* [116]. Os mesmos autores também disponibilizaram os pesos (*weights*) de uma *DenseNet*-121 pré treinada com a mesma base de dados e parâmetros dos autores originais. Estes pesos foram assim utilizados, evitando o treino computacionalmente intensivo de uma nova réplica da CNN.

Visto que existe um total de duas R-RTX (d1 e d3) por doente e por previsão da classe, técnicas de fusão de dados tiverem de ser consideradas. O uso de técnicas de *early fusion* foram selecionadas na tentativa de capturar padrões de ambas as radiografias. Este facto poderá permitir obter *features* de imagem que considerem a evolução temporal do estado clínico do doente, não incidindo apenas num tempo clínico fixo, que pode ser limitativo no contexto de cuidados intensivos e da medicina personalizada. Devido à variabilidade e ausência de relação espacial entre as radiografias de d1 e d3, a concatenação das mesmas pela dimensão do canal poderá não ser benéfica. Experimentalmente, foi realizada a concatenação vertical das radiografias no *input* da CNN como exemplificado por K. Lopez *et al*. [78] numa tarefa de classificação multimodal. Considerando duas radiografias com uma matriz de dimensão de(500x500) a concatenação vertical das matrizes resultaria numa *array* de dimensão de (1000x500).

A construção da CNN foi realizada utilizando a biblioteca *Keras*. Em primeiro lugar, as R-RTX concatenadas foram pré-processadas de acordo com os requisitos da aplicação da *DenseNet*-121 do *Keras*. As mesmas foram convertidas em formato RGB (copiando o canal único nas 3 dimensões) e redimensionadas por interpolação, utilizando a função do *OpenCV*, *cv2.resize(),* para obter uma matriz final de (224,224,3), requerida pela CNN. Após o referido, as imagens dos pacientes foram transformadas em *arrays* de *numpy* através da função *np.array()* permitindo a criação de um tensor de dimensão (110,224,244,3), ou seja 110 matrizes concatenadas de (224,244,3). O *rescaling* foi concebido utilizando a função *keras.applications.densenet.preprocess_input ()*, que normaliza as imagens, para uma *range* de [0-1] com a metodologia de *z-score*, utilizando as médias e desvios padrões dos 3 canais da base de dados ImageNet, onde a CNN foi originalmente treinada.

Para a criação da CNN, foi aplicada a função *keras.applications.DenseNet121()* com os argumentos que permitissem a utilização da mesma sem pesos treinados prévios, sem a camada de classificação final e com *average pooling,* introduzido posteriormente à ultima função de ativação ReLu da última camada de convolução. As camadas da rede foram congeladas e foi adicionada uma nova camada densa de classificação para replicar a *CheXNet,* com 14 neurónios de *output* para as diferentes classes, ativadas por uma função sigmoidal (*keras.layers.Dense* (14,*activation = sigmoid*). A criação desta camada permitiu introduzir os pesos pré-treinados da *CheXNet,* utilizando a função *model.load_weights*(), sendo posteriormente removida utilizando a função *model.layers (-1).output* para criação de um modelo de extração de *features*. A última camada de convolução apresenta 1024 mapas de *features* de (7x7) que sofreram redimensionamento através da camada de *average pooling*, criando assim 1024 *deep learning features (DLF)* 1D. Estas *features* foram concatenadas aos dados clínicos, através do índice chave representativo das imagens, utilizando a função da biblioteca *pandas*, *pd.DataFrame ()* e *pd.merge()*. As *DataFrames* foram então exportadas em formato CSV para o grupo de treino (1036,87) e para o grupo de teste (1036,23) O processo descrito nesta secção encontra-se esquematizado na **Figura 29**.

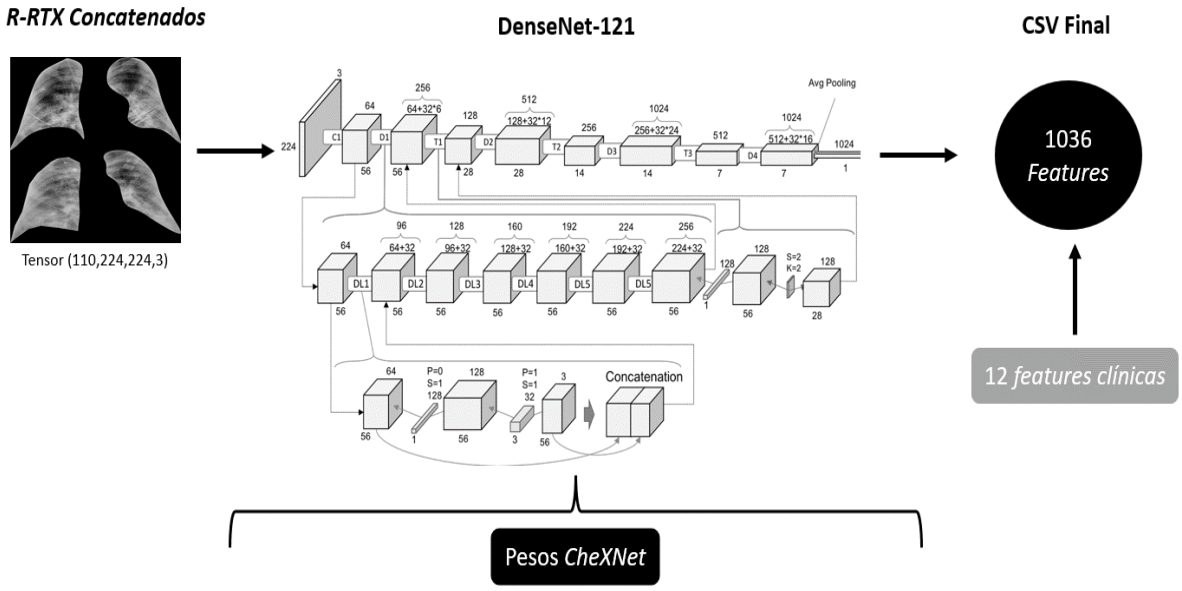


Figura 29 - *Extração de deep features utilizando a CheXNet*

3.4 Análise exploratória multivariada das DLF

Após a extração das DLF, a mesma metodologia da secção 3.1.3 foi aplicada. Devido à grande dimensionalidade da base de dados da DLF (1024 *features*), foi utilizado o *widget box plot,* com o objetivo de filtrar as *features* de potencial interesse para a tarefa atual através do uso do *t-test*. Apenas as DLF com valores de $p \leq 0,01$, foram consideradas para análise, maximizando a filtragem. Na **Tabela 5**, é possível verificar a comparação das médias das DLF entre as classes de interesse (Não-sobrevivente e Sobreviventes). São apresentados também os resultados dos testes estatísticos (*t-test*), e o seu valor *p*.

Tabela 5 - *Estatística bivariada das DLF do grupo de treino em relação à mortalidade com diferenças estatisticamente significativas (p≤0,05), utilizando o student's t-test. n=número de amostras; Méd.±d.p = Média ± Desvio Padrão*

	Não-Sobrevivente (n=42)	**Sobrevivente (n=45)**		
DLF	**Méd.±d.p**	**Méd.±d.p**	**t-test *score***	**Valor *p***
DLF_258	$2,06x10^{-3} \pm 6,77x10^{-4}$	$2,49x10^{-3} \pm 6,98x10^{-4}$	2.869	0,005
DLF_858	$0,169 \pm 5,48x10^{-3}$	$0,172 \pm 5,48x10^{-3}$	2,795	0.006
DLF_411	$9,58x10^{-4} \pm 6,61x10^{-5}$	$1,00x10^{-3} \pm 8,27x10^{-5}$	2,798	0,006
DLF_963	0,62±0,01	0,63 ±0,02	2,731	0,008
DLF_843	$0,215 \pm 9,02x10^{-3}$	0,22 ±0,01	2,668	0,009

É possível visualizar, na **Figura 30**, os coeficientes de correlação de P*earson* entre as DLF e apenas entre as DLF da **Tabela 5** (onde *p≤0,01*). Nesta análise é verificado o facto destas *features* estarem altamente correlacionados entre si, sendo a seleção de *features* essencial para o objetivo de -não introdução de ruído durante a aprendizagem dos classificadores. Também podemos verificar que a DLF_258 aparenta uma fraca correlação com as restantes.

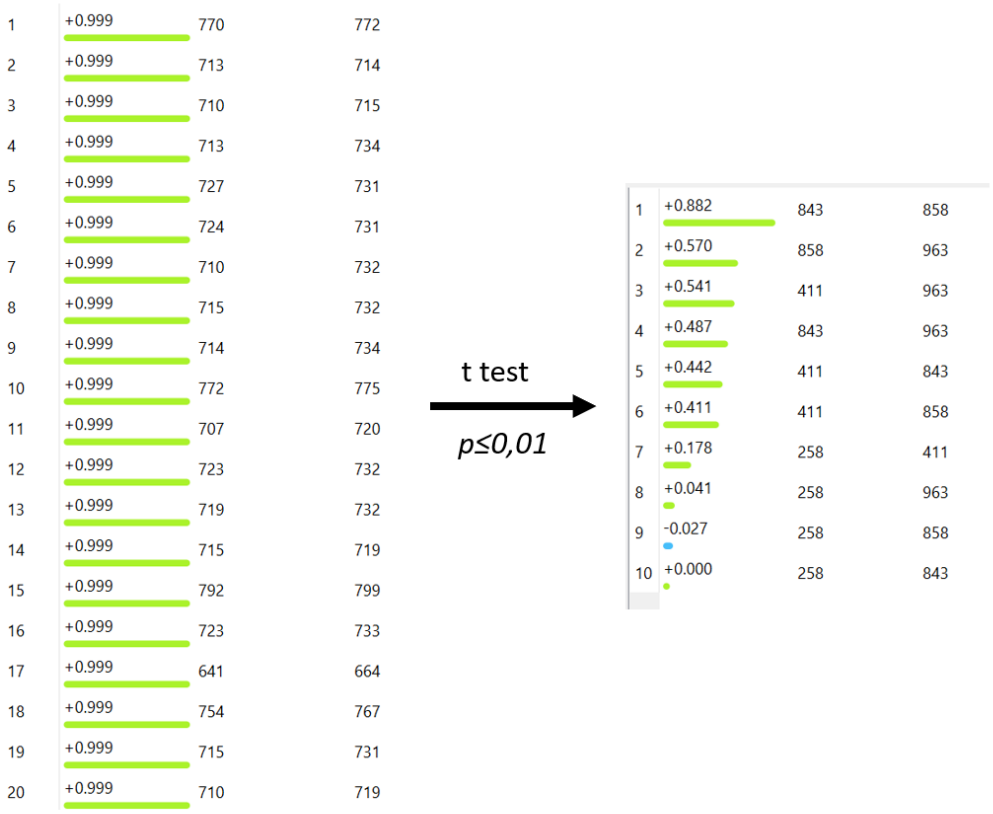


Figura 30 - *Cálculo do Orange Data Mining dos coeficientes de correlação de Pearson entre* todas as DLF e após filtragem das que obtiveram um valor de *p≤0,01* no *t-test* da tabela 5.

Na **Figura 31**, é possível verificar os gráficos de frequência e de diagrama de caixas correspondentes às três DLF que melhor discriminam as classes (de acordo com o *t-test* da **Tabela 5**), considerando as suas médias por classe. Por sua vez, na **Figura 32**, é possível visualizar os *scatter plots* entre as DLF da **Tabela 5** que possam ter maior interesse para a presente tarefa. As três melhores projeções são apresentadas e calculadas automaticamente pelo *widget*. Podemos verificar que DLF_258 encontra-se apresentada em todos os pares e que o risco de mortalidade poderá estar associados valores inferiores das DLF apresentadas

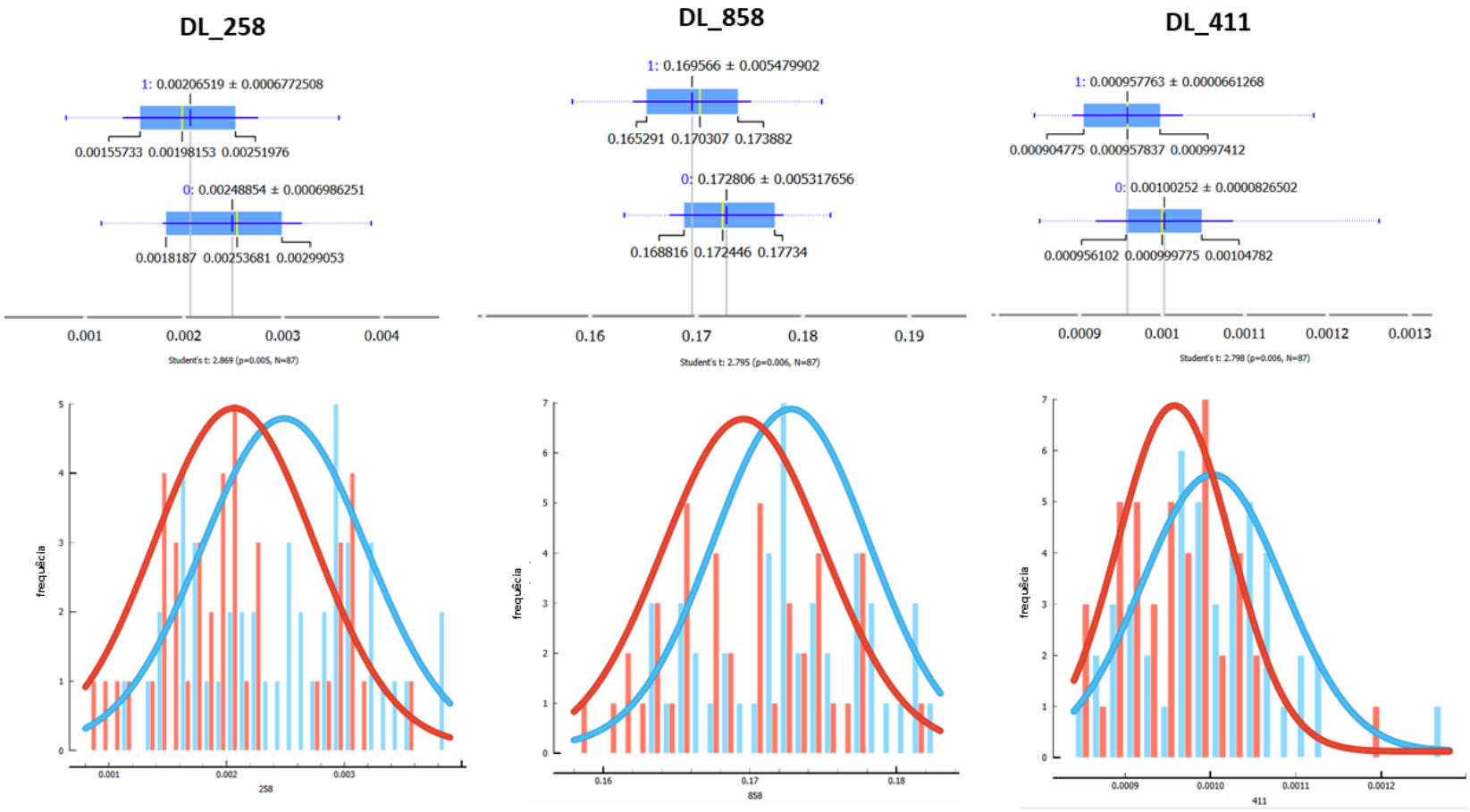


Figura 31 *- Gráficos de Diagrama (topo) e de frequência (fundo) das três DLF cuja média melhor discrimina a classe. Número 0: Sobrevivente; Número 1: Não-sobrevivente; Cor Azul: Sobrevivente; Cor Vermelha Não-Sobrevivente*

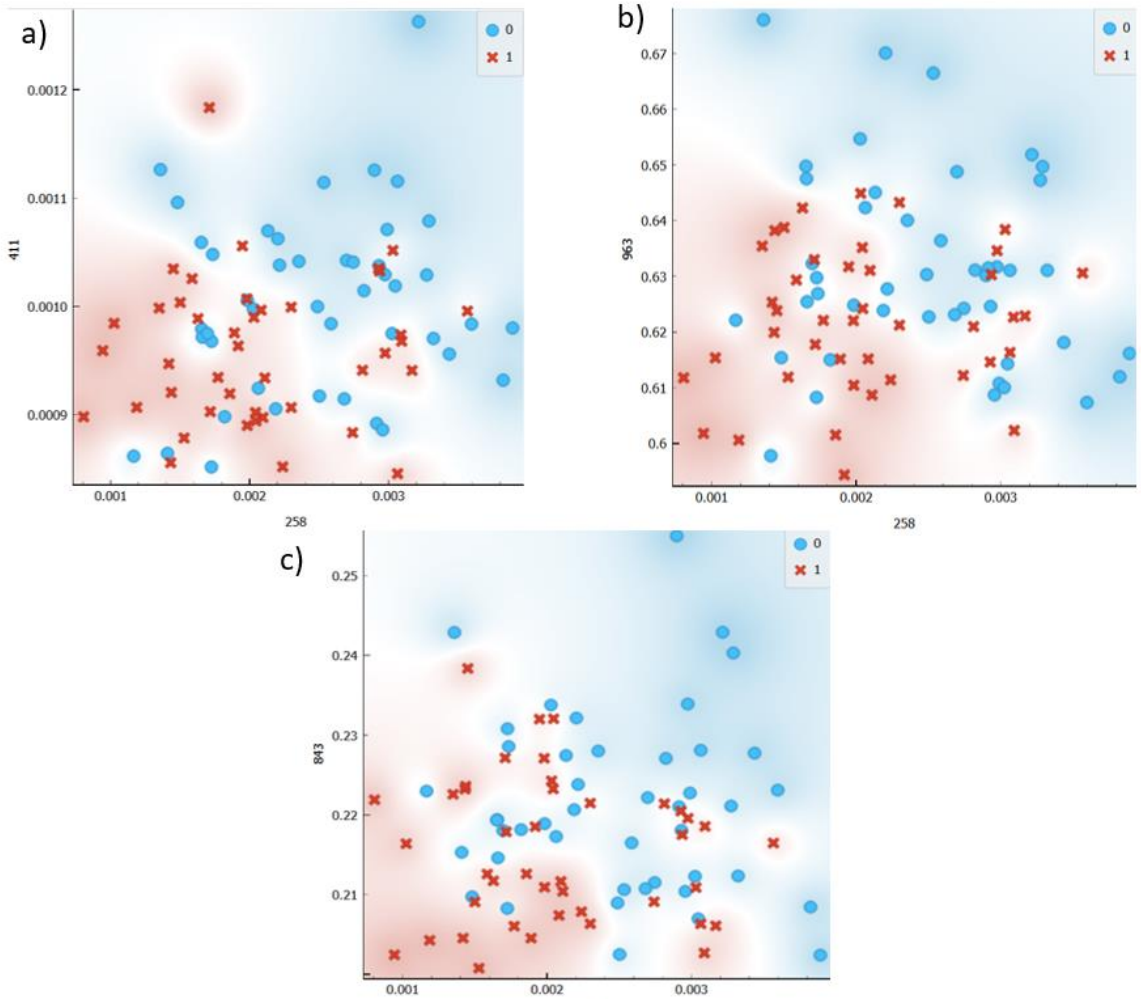


Figura 32 - *Scatter plots das amostras entre duas possíveis DLF's de interesse. a) Gráfico entre DLF_258 no eixo x e DLF_411 no eixo y. b) Gráfico DLF_258 no eixo x e DLF_963 no eixo y. c) Gráfico entre DLF_258 no eixo x e DLF_843 no eixo y. Nos três gráficos a cor vermelha, o valor um e as cruzes representam as amostras dos não-sobreviventes, enquanto a cor azul, o valor zero e os círculos representam as amostras dos sobreviventes*

3.4.1 Análise multivariada entre as DLF e variáveis clínicas

Em última análise, foi verificada a correlação entre as variáveis clínicas e as DLF, procedendo-se também à criação dos melhores *scatter plots* discriminativos da classe de interesse.

É possível visualizar, na **Figura 33**, os coeficientes de correlação de *Pearson* entre cada DLF da **Tabela 3** (onde *p≤0,01*) e as variáveis clínicas da **Tabela 4** (onde *p≤0,05*). Nesta análise é verificado que as *features* estão fracamente correlacionados entre si, potencialmente beneficiando a *performance* dos classificadores.

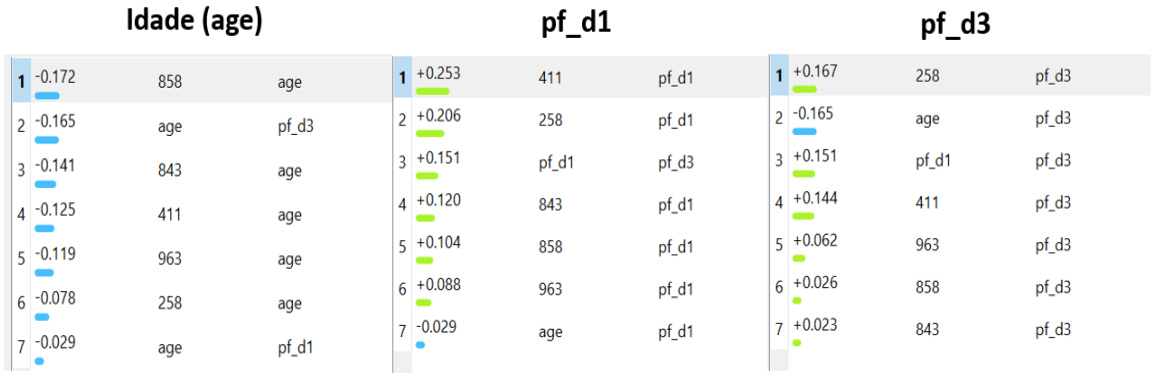


Figura 33 - *Cálculo do Orange Data Mining dos coeficientes de correlação de Pearson entre* as DLF selecionadas e as possíveis variáveis clinicas de interesse.

Na **Figura 34**, é possível visualizar os *scatter plots* entre as DLF da **Tabela 5**, que possam ter maior interesse para a presente tarefa, e as variáveis clínicas da **Tabela 4**. As três melhores projeções são apresentadas e calculadas automaticamente pelo *widget*. É verificado que DLF_258 encontra-se apresentada em 2 dos pares e que o risco de mortalidade poderá estar associado a valores inferiores das DLF e da pf_d3 e superiores de idade.

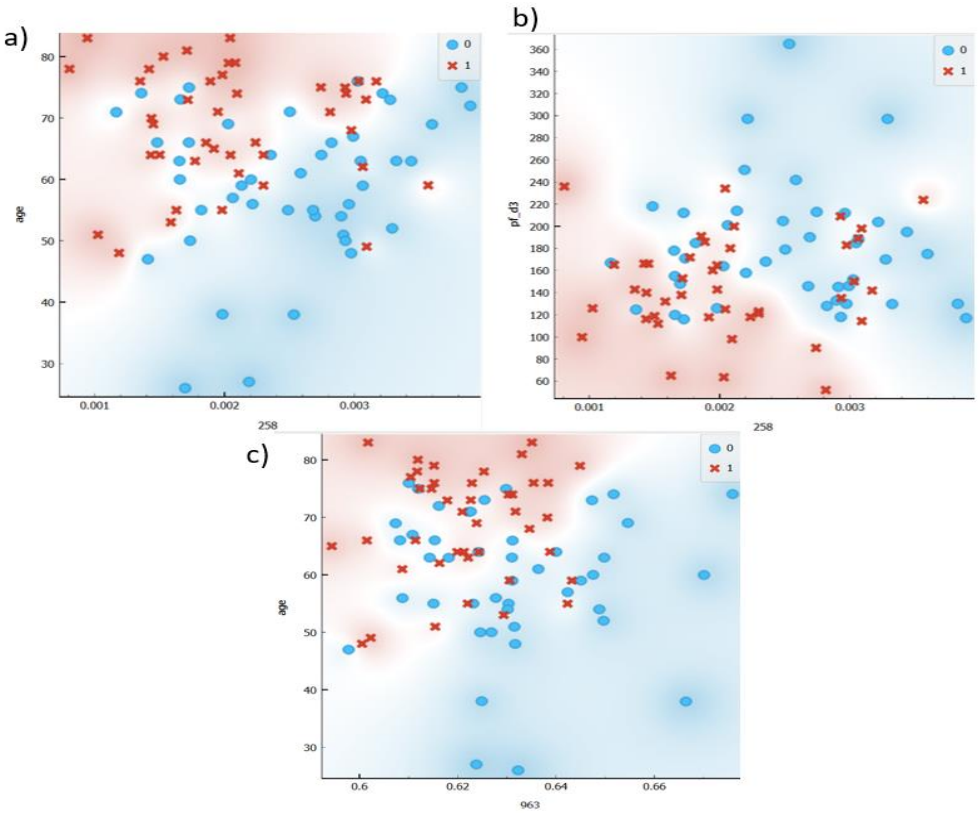


Figura 34 *- Scatter plots das amostras entre possíveis DLF's e variáveis clínicas de interesse. a) Gráfico entre DLF_258 no eixo x e idade (age) no eixo y. b) Gráfico DLF_258 no eixo x pf_d3 no eixo y. c) Gráfico entre DLF_963 no eixo x e idade (age) no eixo y. Nos três gráficos a cor vermelha, o valor 1 e as cruzes representam as amostras dos não-sobreviventes, enquanto a cor azul, o valor zero e os círculos representam as amostras dos sobreviventes.*

3.5 Construção dos modelos de classificação

Os modelos de classificação binária foram testados e treinados utilizando a ferramenta do *Orange Data Mining*, *Test and Score*. Este *widget* permite o treino, validação e testagem de diferentes classificadores, através de métodos de *cross validation* (CV), *leave-one-out* e de amostragem aleatória. As métricas selecionadas para avaliação foram a AUC_p (AUC considerando a Não-Sobrevivência como classe positiva) $AUC_{méd}$ (AUC média entre as 2 classes tendo em conta cada uma como positiva e considerando a quantidade de amostras das mesmas atribuindo pesos), a exatidão (*CA*), precisão (*Prec),* sensibilidade (*Recall)*, *F1-Score (F1)* especificidade (*Spec)*, calculando o *widget* a média das mesmas em CV. Este *widget* também permite a comparação das métricas de *performance* da CV de diferentes classificadores utilizando a interpretação Bayesiana do *t-test* descrita em 2015 por *G. Corani* e *A. Benavoli* [117]. Este método tem em consideração a possível correlação que possa existir entre os grupos de treino da CV (o que pode causar erros de tipo I e tipo II em testes paramétricos) e apresenta resultados mais fidedignos quando comparado a outros testes não-paramétricos, como o *Wilcoxon signed rank test* [117]. Quando utilizado em apenas um grupo de dados, este teste calcula analiticamente a distribuição dos resultados *a priori,* considerando que os dois classificadores apresentam *performance* semelhante. De seguida, múltiplas validações, utilizando a CV, são realizadas e a distribuição das amostras é adaptada

posteriormente. Este fator permite o cálculo simulado de quantas vezes o modelo X iria ganhar (métricas superiores) ao modelo Y em múltiplos testes de classificação. Desta forma, em vez de proporcionar um valor de *p* para rejeitar a hipótese nula, que pode ser limitado na avaliação dos modelos [118], o teste proporciona valores de probabilidade do modelo X ser superior ao modelo B, baseados em distribuições. Para além deste fator, este teste permite definir regiões da distribuição assumindo diferenças negligenciáveis (o *software* permite a definição destas diferenças). Devido ao grupo de dados pequeno de apenas 87 amostras, uma diferença negligenciável de apenas 0,05 foi utilizada (5%) pois em saúde qualquer melhoria pode ser significativa.

Os classificadores escolhidos são os tipicamente utilizados para tarefas de classificação binária com a explicação presente na secção 2.10. Inicialmente foi selecionado um perceptrão multicamada (MLP), um algoritmo de regressão logística (LogReg), uma *support vector machine* (SVM), um algoritmo de gradiente boosting (GB) com os hiperparâmetros pré-definidos do *Orange Data Mining* (**Figura 35**). Para validação dos modelos, foi utilizada a técnica de 10-*fold* CV estratificada, permitindo mais dados de treino em cada subgrupo de validação para um melhor *trade-off* entre viés e variância (secção 2.5.3). No total foram criados 10 subgrupos das 87 amostras de treino, onde sete dos grupos de validação tinham nove amostras e três tinham 8 amostras.

Com o objetivo de validar a possível mais-valia da adição das DLF das R-TRX para a classificação, foram criados dois grupos de treino distintos resultando em 2 tipologias de modelos:

- **Modelo A:** 87 amostras com variáveis de imagem (DLF) e clínicas, onde o *target* foi definido como a Não-Sobrevivência dos doentes (valor binário:1)
- **Modelo B:** 87 amostras em que apenas foram consideradas variáveis clínicas, onde o *target* foi definido como a Não-Sobrevivência dos doentes (valor binário:1)

Após comparação de ambas as tipologias de modelo, os que obtiveram melhores métricas de *performance* foram selecionados para posterior *fine-tuning* (afinação) e calibração com consequente testagem no grupo de teste (23 doentes).

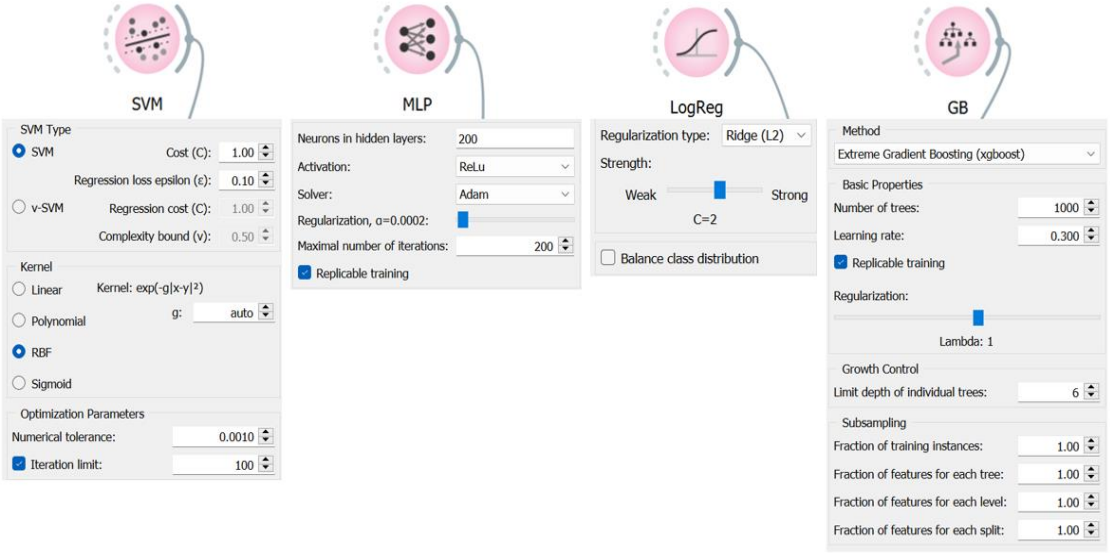


Figura 35 *- Hiperparâmetros pré-definidos do Orange Data Mining. SVM (support vector machine), MLP (perceptrão multicamada), LogReg (Regressão Logística), GB (Gradiente Boosting)*

3.5.1 Pré-processamento dos dados

O pré-processamento dos dados foi realizado utilizando o *widget preprocess*, onde vários métodos de processamento estão presentes. Foram utilizados os seguintes:

- One-Hot-Encoding: Permite transformar as variáveis categóricas em variáveis numéricas binárias (como no caso da variável sexo).
- Imputação por média: Permite preencher as variáveis em falta utilizando a média do total das amostras, permitindo a manutenção da totalidade do grupo de treino.
- Estandardização por *z-score*: permite o *rescaling* das variáveis fazendo com que a média seja zero e o desvio padrão seja um. Este passo é essencial para a realização dos treinos dos classificadores, permitindo também menor interferência dos *outliers* (secção 2.6.1).

É importante referir que este *widget* foi conectado diretamente ao *test and score*, aplicando o pré-processamento a cada subgrupo de treino da 10-*fold* CV individualmente. Esta metodologia permite evitar *data leakage* dos subgrupos de validação para os subgrupos de treino.

3.5.2 Seleção de *features*

A seleção de *features* é um dos passos mais importantes desta investigação, devido à grande dimensionalidade de *features* da base de dados, as poucas amostras e a forte correlação entre as variáveis presentes. Desta forma, foi utilizada a CV para determinar qual o melhor processo de seleção de *features*, para aplicação no modelo final, em termos de métodos utilizados e do número total das mesmas. O *Orange Data Mining* apenas permite a seleção *features* por métodos de filtragem, com diversos testes disponíveis para *ranking* da possível importância das variáveis na discriminação das classes do *target*. Os métodos selecionados foram a *ANOVA* e *Gini Decrease*, sendo apropriados para variáveis numéricas e *targets* categóricos (secção 2.9). Para selecionar o melhor método, os mesmos foram aplicados em CV utilizando o possível número ótimo de *features*.

J. Hua et al.[86] referem que à medida que a correlação entre variáveis aumenta o número ideal de *features* corresponde a $\sqrt{n}$, onde n é o número de amostras. Por outro lado, para classificação categóricas, *V. Lakshmanan et al.* recomendam que $m \geq 10\ n \times C$, onde m é o número de amostras, n é o número de *features* e C é o número de classes [119].

Aplicando esta fórmula, cada subgrupo de treino da CV irá ter no mínimo 78 amostras para uma classificação binária, sendo o número recomendado de *features* aproximadamente quatro. Podemos verificar os resultados da comparação entre os testes estatísticos na **Figura 36** onde o *ranking* por *ANOVA* mostrou resultados mais promissores na maioria dos classificadores, tanto no modelo A como no modelo B. Para seleção do número ideal de *features* utilizando a *ANOVA* foram aplicadas curvas de *performance* considerando a exatidão, com adição iterativa das *features*, como recomendado por *R. Figueroa et al (2012)* [90]. É possível verificar na **Figura 37** (gráfico de *performance*) que, para o modelo A, o valor ótimo de número de *features* foi 3 (de 1036), enquanto para o modelo B, foi 2, para os classificadores que apresentavam melhor exatidão (atingindo um decréscimo e *plateau* a partir desse valor). Para verificar quais as

features que iriam ser selecionadas foi conectado o *widget* de *ranking* à base de dados total onde, utilizando o teste de *ANOVA*, foi verificado que para o modelo A foi selecionado a DLF_258, idade e pd_d3, enquanto que para o modelo B, apenas foi selecionada a idade e pf_d3 (**Figura 38**). A seleção de *features* foi aplicada no *widget preprocess* conectado diretamente ao *test and score*, realizando a seleção de *features* após CV e para cada subgrupo, não existindo *data leakage* e evitando *overfitting* [120].

Gini Decrease

	Model	AUC	CA	F1	Prec	Recall	Spec
A	MLP_A	0.495	0.506	0.394	0.483	0.333	0.667
	LogReg_A	0.428	0.402	0.316	0.353	0.286	0.511
	SVM_A	0.456	0.494	0.371	0.464	0.310	0.667
	GB_A	0.438	0.448	0.415	0.425	0.405	0.489
B	LogReg_B	0.709	0.586	0.581	0.568	0.595	0.578
	MLP_B	0.628	0.563	0.578	0.542	0.619	0.511
	SVM_B	0.590	0.552	0.552	0.533	0.571	0.533
	GB_B	0.556	0.540	0.512	0.525	0.500	0.578

Anova

	Model	AUC	CA	F1	Prec	Recall	Spec
A	MLP_A	0.733	0.724	0.700	0.737	0.667	0.778
	LogReg_A	0.759	0.690	0.667	0.692	0.643	0.733
	SVM_A	0.754	0.713	0.675	0.743	0.619	0.800
	GB_A	0.676	0.609	0.585	0.600	0.571	0.644
B	LogReg_B	0.721	0.621	0.602	0.610	0.595	0.644
	MLP_B	0.709	0.667	0.667	0.644	0.690	0.644
	SVM_B	0.653	0.575	0.584	0.553	0.619	0.533
	GB_B	0.599	0.529	0.518	0.512	0.524	0.533

Figura 36 *–Comparação das métricas de perfomance médias, utilizando 10-fold cross validation, dos diferentes métodos de seleção de features utilizados para o modelo A (A) e para o modelo B (B). MLP = perceptrão multicamada, LogReg = regressão logística, SVM = support vector machine, GB = gradient boosting.*

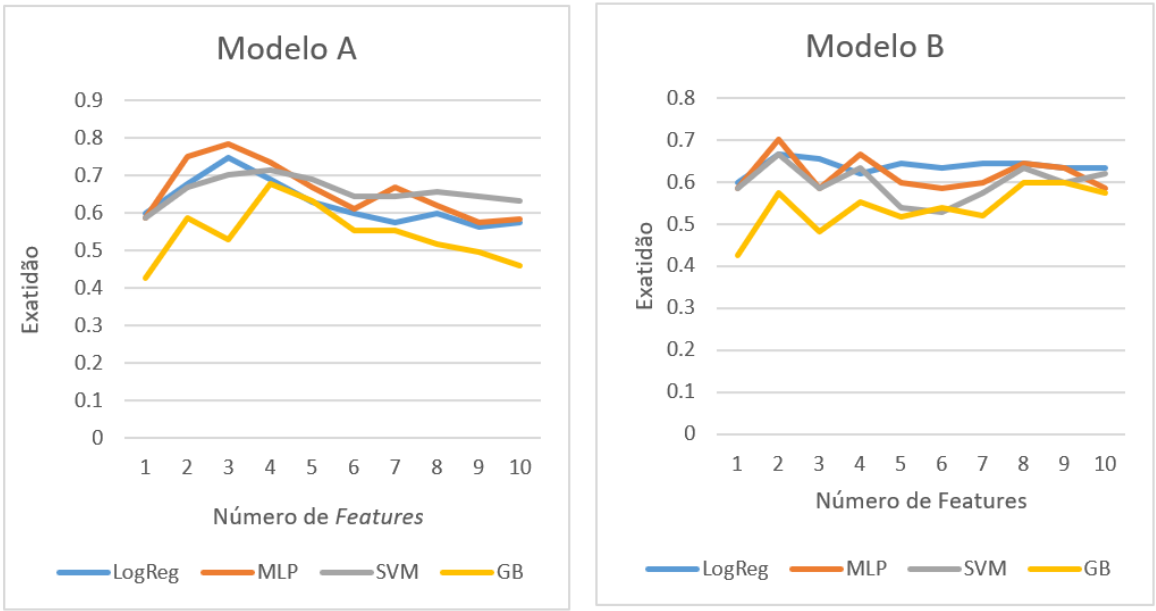


Figura 37 – Curvas de *performance* de exatidão em função do número de features para o modelo A (esquerda) e o modelo B(direita). MLP = *perceptrão multicamada, LogReg = regressão logística, SVM = support vector machine, GB = gradient boosting.*

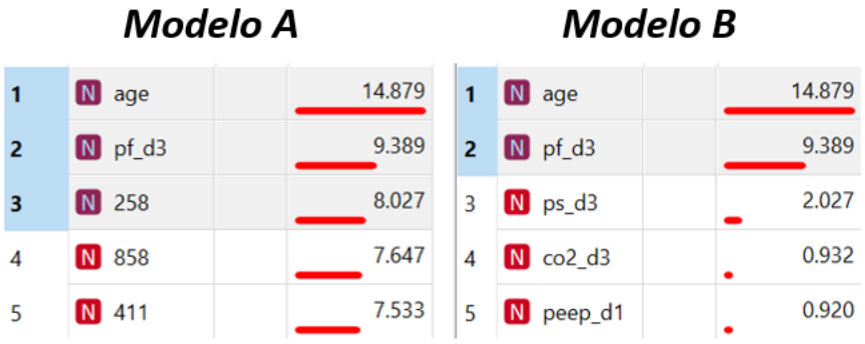


Figura 38 – *Ranking dos valores da estatística F do teste de ANOVA utilizado para seleção de features. No modelo A foram selecionadas três features, enquanto no modelo B foram selecionadas duas.*

3.5.3 Seleção dos classificadores finais

Após a escolha do melhor método de seleção de *features*, o mesmo foi aplicado e foi realizada mais uma ronda de CV. Para seleção dos classificadores a utilizar, foi avaliada os resultados probabilísticos da interpretação Bayesiana do *t-test* considerando uma diferença negligenciável de 0,05. Um classificador foi considerado melhor que outro, caso as suas métricas de *performance* apresentem um valor probabilístico de superioridade maior que o de inferioridade ou equivalência. A seleção dos classificadores foi considerada apenas para o Modelo A.

Na **Figura 39**, é possível averiguar a média das métricas da ronda de CV considerando os mesmos subgrupos (através da função *cross-validate by feature*) e as probabilidades referidas (para o modelo A e o modelo B), onde foi averiguado que a LogReg e a MLP, seriam os classificadores a utilizar para o treino final. No modelo A em termos de AUC, não existem diferenças significativas entre a MLP, LogReg e SVM (com maior probabilidade de serem iguais), no entanto em termos de CA, a SVM apresenta uma probabilidade inferior a 10% de ser superior ao classificador MLP e LogReg (com uma probabilidade inferior a 50% de ser idêntico). No modelo B não existem diferenças significativas entre o classificador de MLP, LogReg e SVM. No entanto para obter uma comparação direta da influência das DLF no potencial de classificação dos modelos, apenas foram selecionados os modelos de MLP e LogReg para ambos os casos. Na **Figura 40**, podem-se observar as curvas ROC dos diferentes classificadores do modelo A.

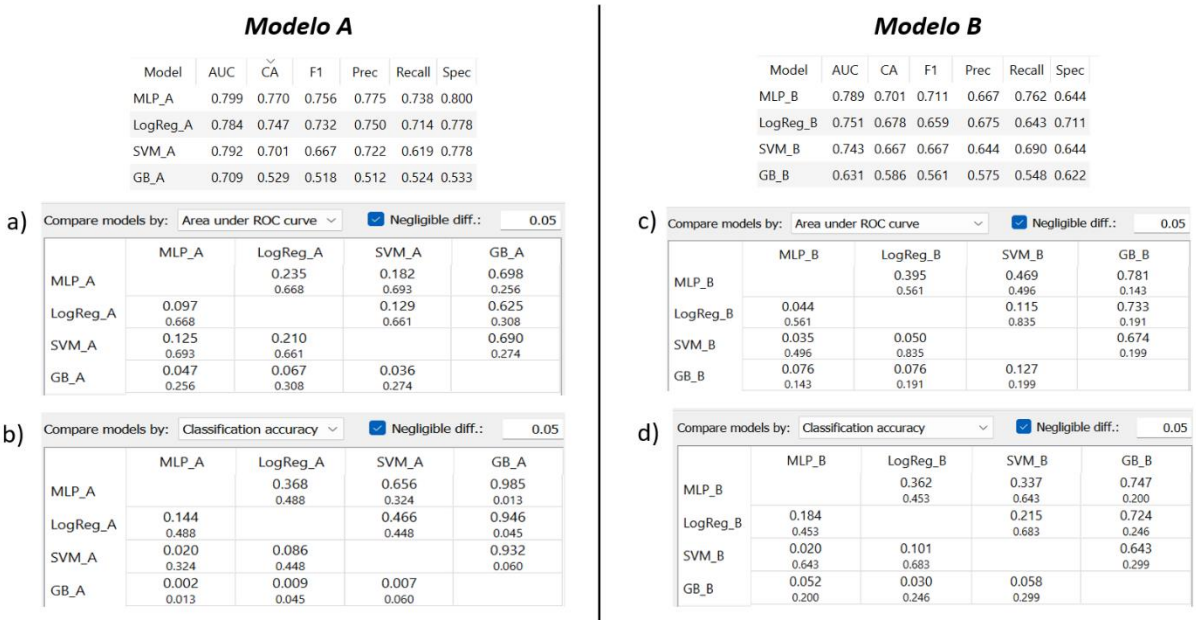


Figura 39 - *Métricas de performance para cada modelo (A e B) nos respetivos classificadores. As tabelas a), b), c) e d) apresentam as probabilidades da interpretação bayesiana do t-test, onde o número maior (superior) representa a probabilidade do classificador da linha ser superior ao da coluna e o número pequeno (inferior) representa a probabilidade de serem idênticos, considerando uma diferença negligenciável de 0,05. Estes testes foram realizados para a métrica de AUC e CA no modelo A (a) e b) respetivamente) e para o modelo B (c) e d) respetivamente).*

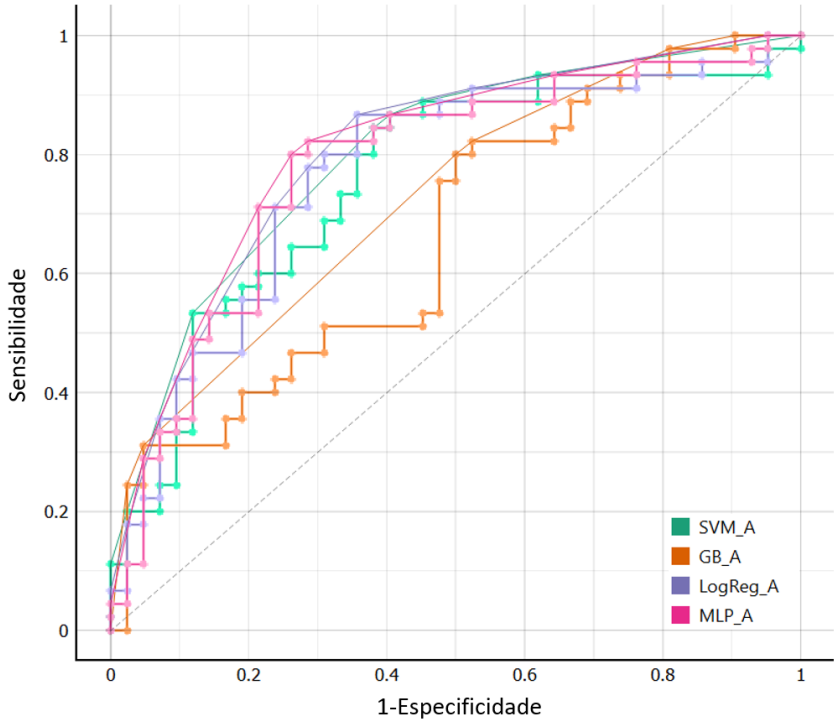


Figura 40 *- Curva ROC (Orange Data Mining) dos classificadores do modelo A. SVM_A = Support vector machine do modelo A, GB =gradient boosting do modelo A; LogReg_A = Regrssão Logística do Modelo A, MLP_A = perceptrão multicamada do modelo A*

3.5.4 Comparação e seleção dos modelos

A comparação e seleção entre os modelos A e B foi efetuada utilizando a mesma metodologia da secção 3.5.3. A LogReg do modelo A foi comparada diretamente com a LogReg do modelo B e MLP do modelo A foi comparada diretamente com a MLP do modelo B. Este método foi utilizado para averiguar o potencial benefício da utilização das DLF, com manutenção dos restantes processos de construção do modelo previamente otimizadas para cada tipologia (A e B). As matrizes de confusão para cada subgrupo da CV e para as previsões na totalidade dos subgrupos da CV (**Figura 41**) foram extraídas para efeitos de cálculo de intervalos de confiança 95% (95%CI) para uma perspetiva estatística de estimativa [121]. Os mesmos foram calculados automaticamente utilizando o *software* estatístico *MedCalc® Statistical Software version 22.013*. Esta ferramenta utiliza o cálculo binomial exato de *Clopper-Pearson* [56,122] para métricas proporcionais de testes binários. Esta utiliza a totalidade das previsões da CV (por exemplo para a *recall* foi considerado o valor de VP+FN da matriz de confusão total da CV no denominador da proporção), enquanto que para a AUC, são utilizados métodos de comparação não-paramétricos de *E. Delong (1988)* [123] através do cálculo das diferenças entre AUC's, do erro padrão e posterior aplicação de métodos binomiais exatos.

A

Tabela 6 e a

Tabela 7 resumem os resultados obtidos. As **Tabela 8** representa os 95%CI obtidos para cada modelo. No caso do classificador MLP, existe maior probabilidade do modelo A ser superior ao modelo B, em 4 métricas de *performance*, do que ser inferior ou idêntico ao mesmo (considerando uma diferença negligenciável das métricas de 5% entre os modelos).

É possível verificar que para a comparação entre a MLP_A e MLP_B, a contabilização das DLF no processo de seleção de *features* (3 *features* por subgrupo), apresentou probabilidade superior, correspondendo a um aumento significativo (superior a 5%) de *performance* em 4 métricas, quando comparado à probabilidade de ser inferior ou idêntica (CA, *recall,* Prec, Spec, F1). É possível verificar a curva ROC dos modelos referidos na **Figura 42**.

Tabela 6 – *Média das métricas de performance dos modelos MLP_A e MLP_B com respetivos resultados de teste-t bayesiano. P_AB = Probabilidade do modelo A ser melhor que o B; P_BA = Probabilidade do modelo B ser melhor que o modelo A; $P_{5\%}$_AB = Probabilidade do modelo A ser melhor que o B considerando uma diferença de performance de 5% negligenciável; $P_{5\%}$_BA = Probabilidade do modelo B ser melhor que o A considerando uma diferença de performance de 5% negligenciável; $P_{neg5\%}$ = Probabilidade dos modelos serem idênticos considerando uma diferença negligenciável de performance de 5%.* Maiores probabilidades a negrito

Modelo	$AUC_{méd}$	AUC_p	CA	*Recall*	Prec	Spec	F1
MLP_A	0,778	0,799	0,770	0,738	0,775	0,738	0,756
MLP_B	0,738	0,789	0,701	0,762	0,667	0,762	0,711
Bayesian t-test	**P_AB**	0,549	**0,820**	0,427	**0,873**	**0,883**	**0,767**
	P_BA	0,451	0,180	0,573	0,127	0,117	0,233
	$P_{5\%}$_AB	0,321	**0,613**	0,216	**0,729**	**0,791**	**0,516**
	$P_{5\%}$_BA	0,233	0,069	0,109	0,053	0,062	0,086
	$P_{neg5\%}$	0,455	0,318	0,557	0,218	0,147	0,398

Tabela 7 – *Média das métricas de performance dos modelos LogReg_A e LogReg_B com respetivos resultados de teste-t bayesiano. P_AB = Probabilidade do modelo A ser melhor que o B; P_BA = Probabilidade do modelo B ser melhor que o modelo A; $P_{5\%}$_AB = Probabilidade do modelo A ser melhor que o B considerando uma diferença de performance de 5% negligenciável; $P_{5\%}$_BA = Probabilidade do modelo B ser melhor que o A considerando uma diferença de performance de 5% negligenciável; $P_{neg5\%}$ = Probabilidade dos modelos serem idênticos considerando uma diferença negligenciável de performance de 5%.* Maiores probabilidades a negrito

Modelo	$AUC_{méd}$	AUC_p	CA	*Recall*	Prec	Spec	F1
LogReg_A	0,770	0,784	0,747	0,714	0,750	0,778	0,732
LogReg_B	0,701	0,751	0,678	0,643	0,675	0,711	0,659
Bayesian t-test	**P_AB**	**0,739**	**0,886**	**0,769**	**0,875**	**0,758**	**0,864**
	P_BA	0,261	0,114	0,231	0,122	0,242	0,136
	$P_{5\%}$_AB	0,364	**0,630**	**0,568**	**0,734**	**0,547**	**0,713**
	$P_{5\%}$_BA	0,062	0,026	0,104	0,204	0,107	0,056
	$P_{neg5\%}$	**0,574**	0,453	0,328	0,230	0,346	0,230

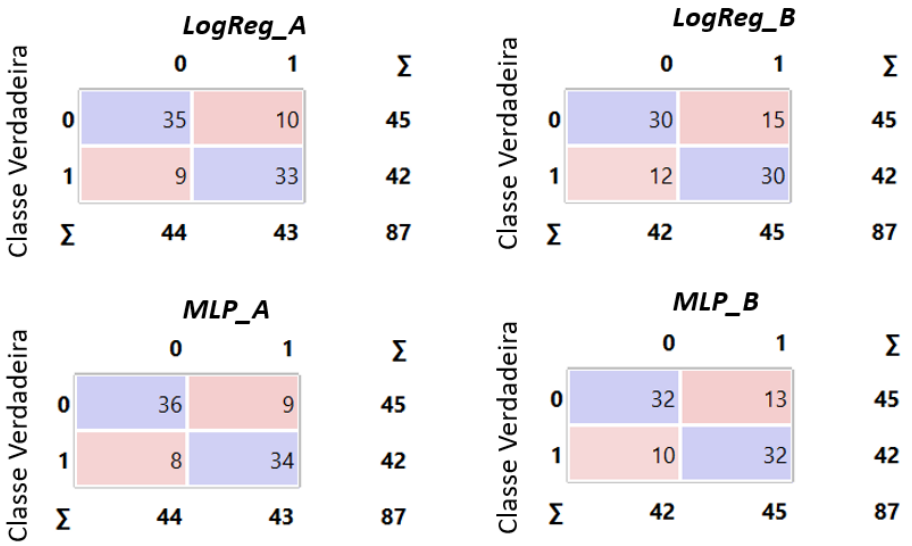


Figura 41 *- Matriz de confusão do validação cruzada do classificador MLP_A, MLP_B, LogReg_A, LogReg_B. LogReg_A = regressão logística do modelo A, MLP = perceptrão multicamada do modelo A LogReg_B = Regressão logística do modelo B, MLP_B = perceptrão multicamada do modelo B , 0 = Sobrevivente, 1- Não Sobrevivente.*

Tabela 8 – *Média das métricas de classificação para os modelos LogReg_A, LogReg_B, MLP_A e MLP_B (utilizando 10-fold CV), considerando os seus intervalos de confiança a 95% (95%CI) calculados pelos métodos descritos no capítulo 3.5.4.*

Métrica	LogReg_A	MLP_A	LogReg_B	MLP_B
$AUC_{méd}$	0,770 95%CI [0.667 ,0.853]	0,778 95%CI [0.677 ,0.860]	0,701 95%CI [0.593 ,0.794]	0,738 95%CI [0.633 ,0.827]
CA	0,747 95%CI [0.643 , 0.834]	0,770 95%CI [0.667 ,0.853]	0,678 95%CI [0.569 ,0.774]	0,701 95%CI [0.593 ,0.794]
Recall	0,714 95%CI [0.554 , 0.843]	0,738 95%CI [0.580 ,0.861]	0,643 95%CI [0.480 ,0.784]	0,762 95%CI [0.606 ,0.880]
Prec	0,750 95%CI [0.588 , 0.873]	0,775 95%CI [0.615 ,0.891]	0,675 95%CI [0.509 ,0.814]	0,667 95%CI [0.516 ,0.796]
Spec	0,778 95%CI [0.629 , 0.888]	0,738 95%CI [0.586 ,0.858]	0,711 95%CI [0.560 ,0.834]	0,711 95%CI [0.557 ,0.836]
F1	0,732 95%CI [0.623 ,0.824]	0,756 95%CI [0.669 ,0.844]	0,659 95%CI [0.546 ,0.760]	0,711 95%CI [0.606 ,0.802]

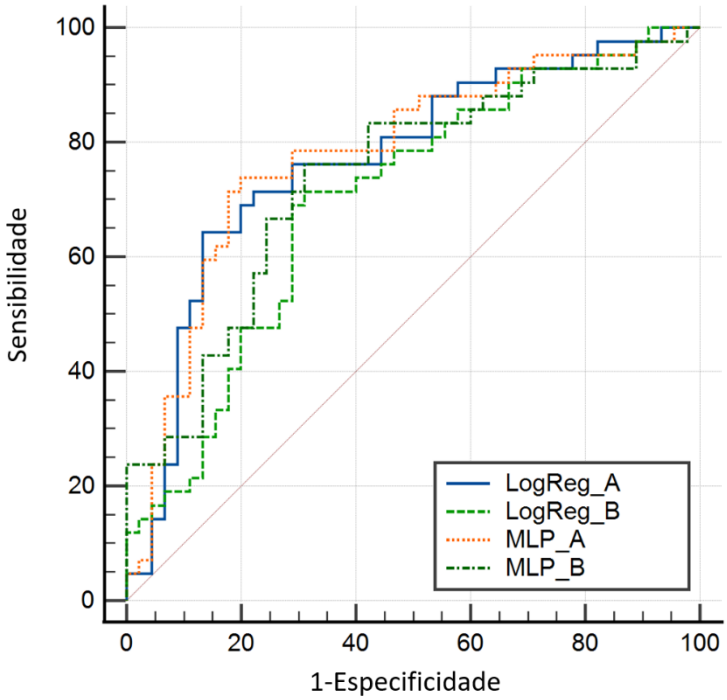


Figura 42 - *Curva ROC (MedCalc) dos classificadores do modelo A e do modelo B. LogReg_A = regressão logística do Modelo A,. LogReg_B = regressão logística do Modelo B; MLP_A = perceptrão multicamada do modelo A; MLP_B = perceptrão multicamada do modelo B*

3.5.5 Fine-Tuning e Calibração

O *fine tuning* foi realizado alterando iterativamente os hiperparâmetros de LogReg e MLP do modelo A utilizando a CV. No modelo de LogReg não foi verificado melhorias nas métricas de classificação alterando os diferentes tipos de regularização e o valor de C, com manutenção da pré-definição do *software*. Na MLP foi verificada uma melhoria nas métricas de *performance* em CV apenas através da manutenção da função de ativação pré-definida (ReLU) e aumentando os neurónios utilizados de 200 para 300 na *hidden layer*. Na **Tabela 9** é possível averiguar a influência da modificação deste hiperparâmetro da MLP nas métricas de *performance*. Podemos verificar que existe uma alta probabilidade (>70%) do classificador de MLP *fine-tuned (MLP_F)* ser superior ao MLP sem *fine-tuning (MLP_NF)* nas diferentes métricas, no entanto, podemos verificar que existe alta probabilidade (>80) desse aumento de *performance* resultar em apenas 5% de diferença (alta probabilidade de serem idênticos considerando essa diferença negligenciável). Para efeitos de *fine-tuning* este aumento foi aceite e considerado significativo, procurando pequenos *boosts* de *performance*. Curvas ROC e de *Precision-Recall,* nos diferentes *thresholds* foram criadas para melhor observação das melhorias de *performance*, utilizando o *widget performance curve e ROC analysis (***Figura 43***)*. É também possível verificar que foi possível atingir uma AUC de 0,820. O modelo de MLP afinado (MLP_F) foi assim selecionado para o treino final.

Tabela 9 - *Média das métricas de performance dos modelos MLP_F e MLP_NF com respetivos resultados de teste-t bayesiano. P_F-NF = Probabilidade do modelo F ser melhor que o NF; P_NF_F = Probabilidade do modelo NF ser melhor que o modelo F;* $P_{5\%}$*_F_NF = Probabilidade do modelo F ser melhor que o NF considerando uma diferença de performance de 5% negligenciável;* $P_{5\%}$*_NF_F = Probabilidade do modelo NF ser melhor que o F considerando uma diferença de performance de 5% negligenciável;* $P_{neg5\%}$ *= Probabilidade dos modelos serem idênticos considerando uma diferença negligenciável de performance de 5%.* Maiores probabilidades a negrito

Modelo	$AUC_{méd}$	AUC_p	CA	*Recall*	Prec	Spec	F1
MLP_F	0,787	0,820	0,782	0,762	0,780	0,800	0,771
MLP_NF	0,778	0,799	0,770	0,738	0,775	0,800	0,756
Bayesian t-test	**P_F-NF**	**0,771**	**0,746**	**0,746**	**0,873**	0,460	**0,825**
	P_NF_F	0,229	0,254	0,254	0,127	0,540	0,233
	$P_{5\%}$_F_NF	0,160	0,020	0,164	0,211	0,145	0,041
	$P_{5\%}$_NF_F	0,014	0,002	0,020	0,063	0,191	0,002
	$P_{neg5\%}$	**0,825**	**0,978**	**0,816**	**0,726**	0,664	**0,957**

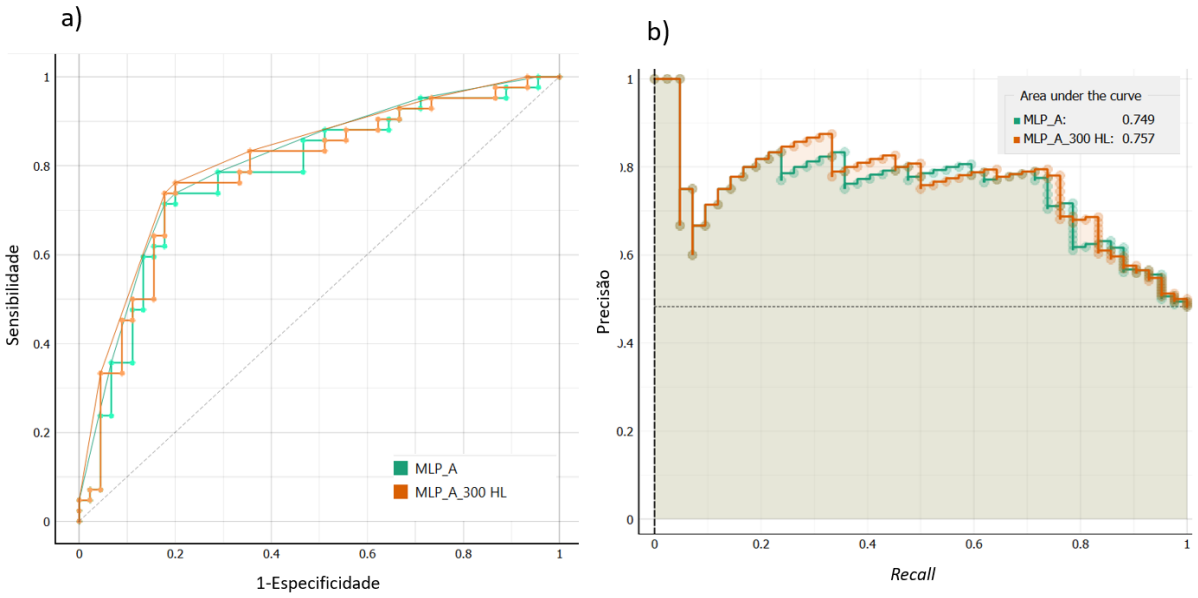


Figura 43 – *Curvas de performance do classificador MLP e MLP com fine-tuning*. a) Curva ROC do classificador MLP com e sem *fine tuning*; b) Curva de *Precision-Recall* (cada ponto representa um *treshold* de classificação) do classificador MLP com e sem *fine tuning*. MLP_A = perceptrão multicamada do modelo A; MLP_A_300_HL: perceptrão multicamada do modelo A *fine tuned* com 300 *hidden layers.*

3.5.6 Calibração dos modelos finais

Conforme descrito nos modelos anteriores, foi selecionado o classificador de *MLP fine-tuned* e LogReg do modelo A para efeitos de calibração. Este processo foi realizado utilizando o *widget calibrated learner*, diretamente conectado aos classificadores em questão e ao *test and score*, permitindo avaliação por CV. Esta ferramenta dispõe de métodos de calibração isotónica e sigmoide. Ambos os métodos foram testados, tanto no classificador de LogReg, como no classificador *fine-tuned* MLP. A curva de calibração foi desenhada utilizando os resultados médios da CV, através do *widget calibration curve*. Na **Figura 44** podem-se averiguar os resultados da construção desta curva para cada um dos classificadores, onde o eixo yy representa a fração de verdadeiros positivos num certo grupo de classificados e o eixo xx representa a probabilidade de classificação da classe positiva, nesses mesmos grupos, fornecida pelo classificador. A reta diagonal onde f(x) = x, representa um classificador perfeitamente calibrado. É possível averiguar que em ambos os classificadores, a metodologia de calibração isotónica resultou numa curva de calibração mais retificada e próxima da reta de referência (principalmente nos valores de probabilidade centrais). Quanto à otimização do *threshold* probabilístico de classificação entre classes, o mesmo foi definido através da criação de curvas de *performance* de CA e F1 para cada *threshold* utilizando o mesmo *widget* (**Figura 45**). Foi procurado um valor de *threshold* que permitisse a melhor CA possível, sem perda significativa de F1. A **Figura 45** reflete o *threshold* escolhido de 0,5 para a classificação entre sobreviventes e não-sobreviventes. A calibração isotónica, implicaria a redução da sensibilidade dos algoritmos em CV. Existe uma probabilidade superior a 60% dos algoritmos sem calibração serem melhores que os algoritmos com calibração, tendo em conta esta métrica (**Figura 46**). Devido a tal fator, não foi aplicada calibração a ambos os classificadores

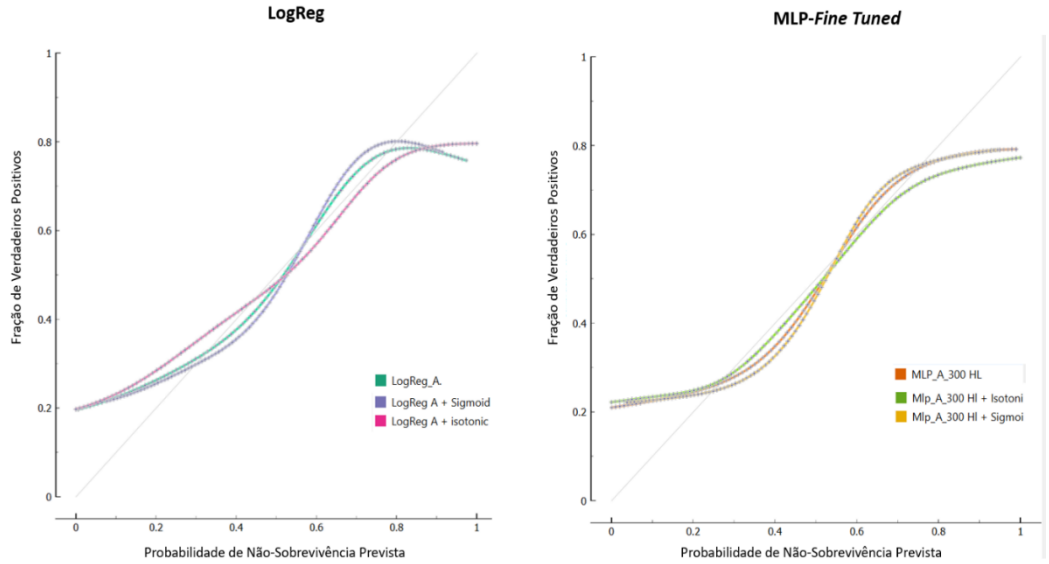


Figura 44 – *Curvas de calibração da 10-fold cross validation para a regressão logística do modelo A (LogReg_A) e o perceptrão multicamada fine-tuned do modelo A (MLP_A_300HL), utilizando calibração sigmóidea e isotónica*

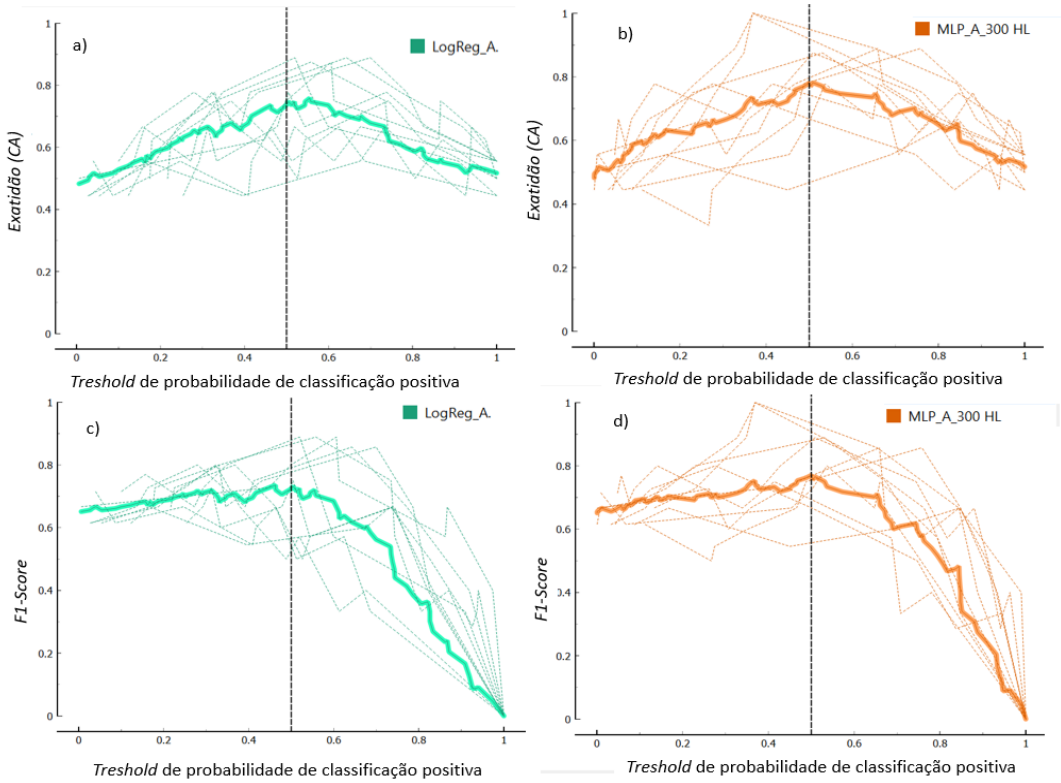


Figura 45 *- Curvas de performance para exatidão (CA) e score de F1 (F1-Score) para o classificador de regressão logística (LogReg_A), a) e c) respetivamente, e para o classificador de MLP Fine-tuned, b) e d) respetivamente. Em cada gráfico as linhas a tracejadas representam o valor da métrica representada em cada threshold para cada subgrupo da validação cruzada. A linha preenchida representa a média da validação cruzada. Em cada gráfico a linha preta vertical, representa o threshold de classificação selecionado (0,5).*

Model	AUC	CA	F1	Prec	Recall	Spec
Mlp_A_300 Hl + Isotonic	0.813	0.747	0.718	0.778	0.667	0.822
MLP_A_300 HL	0.820	0.782	0.771	0.780	0.762	0.800
LogReg_A.	0.784	0.747	0.732	0.750	0.714	0.778
LogReg A + isotonic	0.778	0.724	0.692	0.750	0.643	0.800

Compare models by: Recall — Negligible diff.: 0.05

	Mlp_A_300 Hl + Isotonic	MLP_A_300 HL	LogReg_A.	LogReg A + isotonic
Mlp_A_300 Hl + Isotonic		0.014 0.224	0.166 0.355	0.399 0.375
MLP_A_300 HL	0.761 0.224		0.474 0.406	0.808 0.169
LogReg_A.	0.479 0.355	0.120 0.406		0.645 0.332
LogReg A + isotonic	0.225 0.375	0.023 0.169	0.023 0.332	

Figura 46 *- Métricas de performance para cada modelo (LogReg e MLP-Fine tuned) e a sua respetiva calibração isotónica. A tabela em inferior apresenta a interpretação bayesiana do t-test, onde o número maior (superior) representa a probabilidade do classificador da linha ser superior ao da coluna e o número pequeno (inferior) representa a probabilidade de serem idênticos, considerando uma diferença negligenciável de 0,05. Este teste foi realizado para a métrica de recall.* MLP_A_300_HL: perceptrão multicamada do modelo A *fine tuned* com 300 *hidden layers; LogReg_A: = Regressão logística do modelo A; +isotonic: Modelo + Calibração isotónica.*

3.5.7 Interpretação dos modelos finais

Após seleção dos melhores processos de treino dos modelos, melhores *features* e melhores classificador por CV, foi realizado o treino final do modelo A contabilizando a totalidade das 87 amostras do grupo de treino. Foi assim treinado um MLP e um classificador de LogReg *fine tuned* com os processos descritos para avaliação da *performance* no grupo de teste interno (23 doentes). Com o objetivo de tornar os modelos de *white box* ainda mais interpretáveis clinicamente, foi realizado uma avaliação por valores de SHAP dos modelos finais treinados, utilizando o *widget Explain Model*. Esta ferramenta produz um gráfico que representa, por ordem de importância, as *features* mais impactantes para as classificações no grupo de treino, assim como a dimensionalidade dos seus valores e o impacto dos mesmos na decisão final do classificador (valor de SHAP mais distantes de zero). É possível verificar na **Figura 47** que a variável mais importante na totalidade das decisões foi a idade, seguida da DLF_258 e da pf_d3, considerando o grupo de treino (87 amostras).

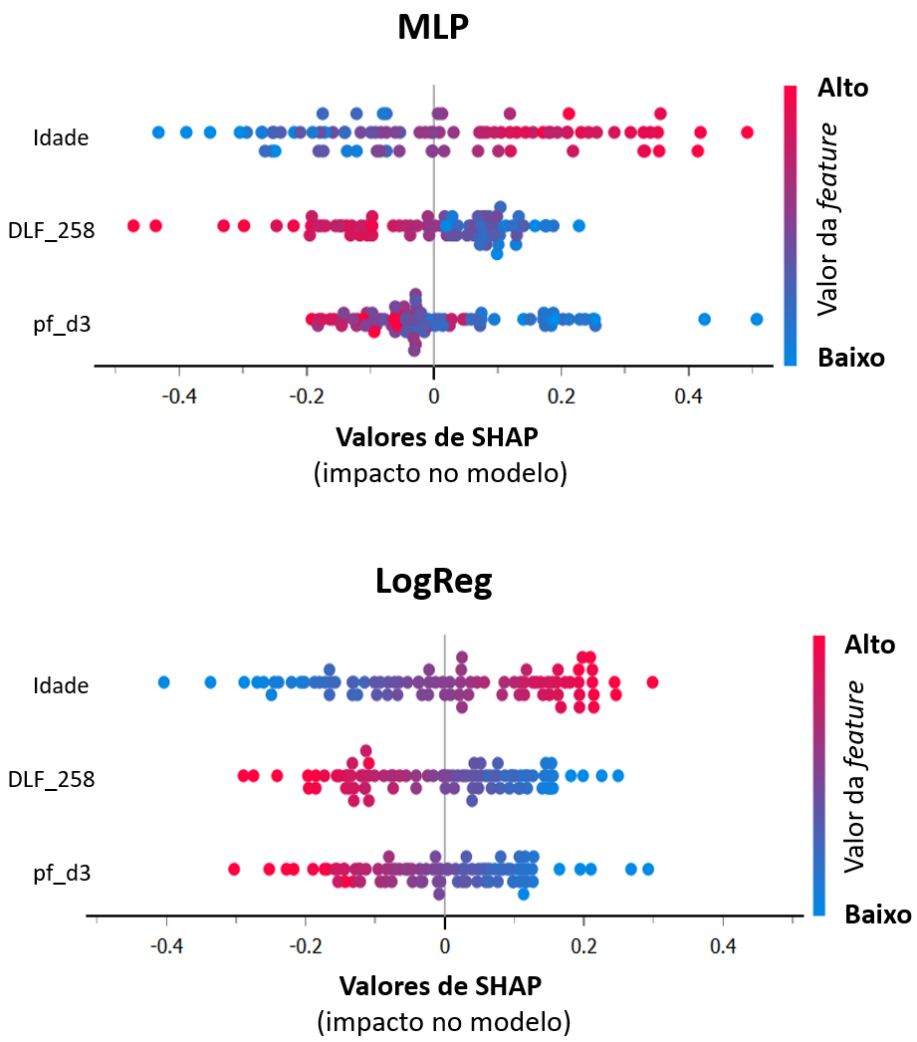


Figura 47 *- Análise dos valores de SHAP nas classificações do grupo de treino (87) amostras* com o classificador MLP (em cima) e LogReg (em baixo)*. Cada círculo representa uma amostra diferente, com maiores valores das features associados à tonalidade vermelha. O eixo x apresenta os valores de SHAP e o eixo y demonstra, por ordem decrescente de importância (de cima para baixo), o nome das features*

4. Resultados

O presente capítulo apresenta os resultados de classificação obtidos no grupo de teste (23 amostras) com os modelos treinados descritos no capítulo 3, assim como a sua comparação com os resultados obtidos na fase de validação por *10-fold CV* e no grupo de treino.

4.1.1 Métricas de *performance*

As

Tabela 10 e **Tabela 11** demonstram as métricas de *performance* dos modelos de LogReg e MLP respetivamente, tanto no grupo de treino, de teste e na CV. As diferenças entre as mesmas podem também ser observadas para averiguar o erro preditivo e *overfitting*. É possível visualizar as curvas ROC da CV e do grupo de teste, para cada modelo, na **Figura 48**, assim como curvas de *precision-recall* para o grupo de CV e treino (**Figura 49**) e para o grupo de CV e teste (**Figura 50**). Na **Figura 51** verificam-se as matrizes de confusão, para cada modelo, do grupo de teste.

Tabela 10 – *Métricas de performance do modelo final LogReg_A no grupo de treino (LogReg_train), na validação cruzada (LogReg_CV) e no grupo de teste (LogReg_Test).* Δ_{CV_Treino} *= Diferença entre a métrica do grupo de treino e a validação cruzada.* Δ_{Teste_Treino} *= Diferença entre as métricas do grupo de treino e do grupo de teste. 95%CI calculado com os métodos descritos no capítulo 3.5.4*

Métricas	**LogReg_Train**	**LogReg_CV**	**LogReg_Test**	**Δ_{CV_Treino}**	**Δ_{Teste_Treino}**
$AUC_{méd}$	0,820 95%CI [0.723 ,0.894]	0,770 95%CI [0.667 ,0.853]	**0,862** 95%CI [0.654 ,0.969]	-0,092	0,042
AUC_p	0,820	0,784	**0,862**	-0,036	0,042
CA	0,782 95%CI [0.681 ,0.863]	0,747 95%CI [0.643 ,0.834]	**0,783** 95%CI [0.563 ,0.926]	-0,035	-0,0
F1	0,776 95%CI [0.672, 0.859]	0,732 95%CI [0.623 ,0.824]	**0,783** 95%CI [0.563, 0.926]	-0,044	0,007
Prec	0,776 95%CI [0.623 ,0889]	0,750 95%CI [0.588 ,0.873]	**0,692** 95%CI [0.385 ,0.909]	-0,026	0,084
Recall	0,787 95%CI [0.633 ,0.898]	0,714 95%CI [0.554 ,0.843]	**0,900** 95%CI [0.555 ,0.997]	-0,073	0,113
Spec	0,778 95%CI [0.629 ,0.888]	0,778 95%CI [0.629 ,0.888]	**0,692** 95%CI [0.385 ,0.909]	0,000	-0,086

Tabela 11 - *Métricas de performance do modelo final MLP_A no grupo de treino (LogReg_train), na validação cruzada (LogReg_CV) e no grupo de teste (LogReg_Test). Δ_{CV_Treino} = Diferença entre a métrica do grupo de treino e a validação cruzada. Δ_{Teste_Treino} = Diferença entre as métricas do grupo de treino e do grupo de teste. 95%CI calculado com os métodos descritos no capítulo 3.5.4*

Métricas	MLP_Train	MLP_CV	MLP_T	Δ_{CV_Treino}	Δ_{Teste_Treino}
$AUC_{méd}$	0,895 95%CI [0.810 ,0.950]	0,787 95%CI [0.686 ,0.867]	**0,815** 95%CI [0.600 ,0.945]	-0,108	-0,080
AUC_p	0,895	0,820	**0,815**	-0,075	-0,080
CA	0,816 95%CI [0.718 ,0.891]	0,782 95%CI [0.681 ,0.863]	**0,783** 95%CI [0.563 ,0.926]	-0,034	-0,033
F1	0,810* 95%CI [0.709 ,0.887]	0,771 95%CI [0.666 ,0.856]	**0,762** 95%CI [0.528, 0.918]	-0,039	-0,048
Prec	0,810 95%CI [0.659 ,0.914]	0,780 95%CI [0.718 ,0.891]	**0,727** 95%CI [0.390 ,0.965]	-0,03	-0,030
Recall	0,810 95%CI [0.659 ,0.914]	0,762 95%CI [0.606 ,0.879]	**0,800** 95%CI [0.444 ,0.975]	-0,048	-0,010
Spec	0,822 95%CI [0.679 ,0.920]	0,800 95%CI [0.654 ,0.904]	**0,769** 95%CI [0.462 ,0.945]	-0,022	-0,022

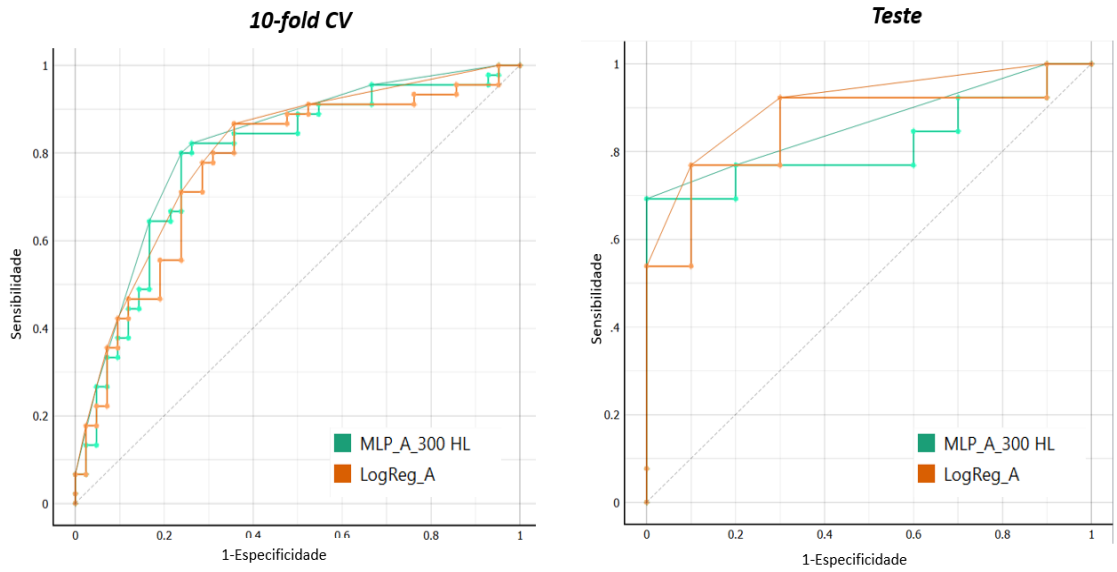


Figura 48 *– Curva ROC do classificador MLP e LogReg finais nos dados de CV (esquerda) e de teste (direita); LogReg_A = regressão logística do modelo A, MLP_A_300_HL* = perceptrão multicamada *fine-tuned* com 300 neurónios na *hidden layer.*

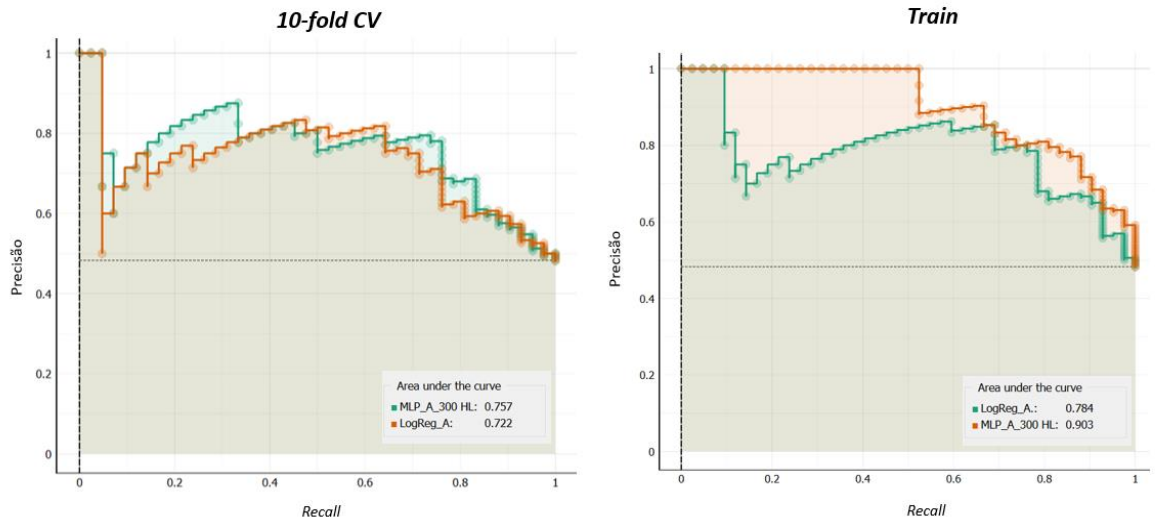


Figura 49 *– Curvas Precision-Recall do classificador MLP e LogReg finais nos dados de CV (esquerda) e treino (direita). É possível verificar a area under the curve em cada gráfico. LogReg_A = regressão logística do modelo A, MLP_A_300_HL =* perceptrão multicamada *fine-tuned* com 300 neurónios na *hidden layer.*

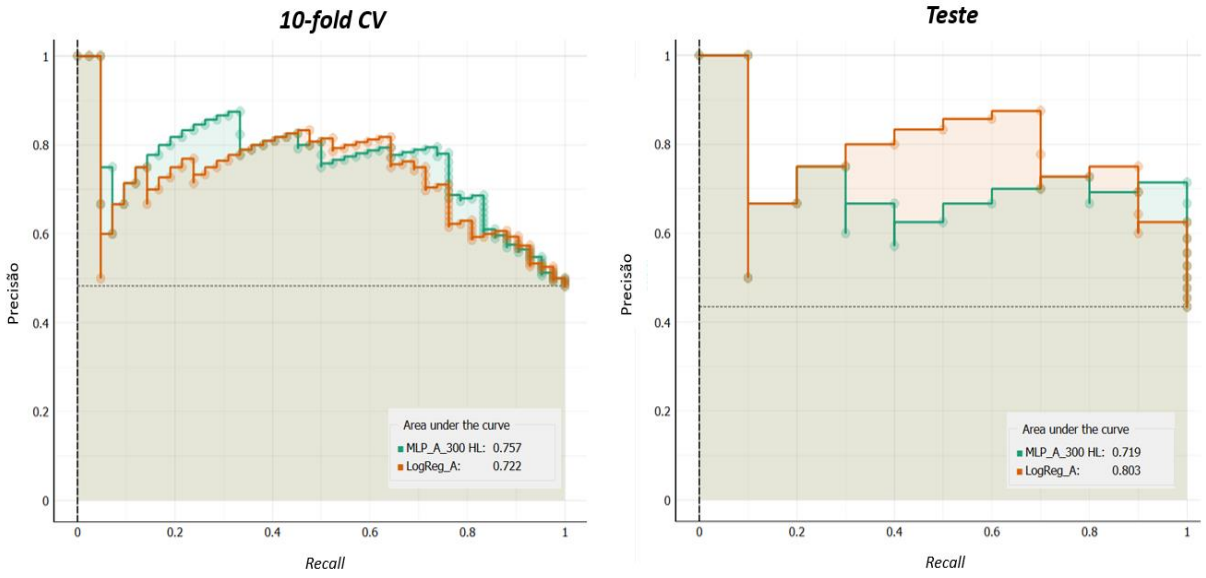


Figura 50 *– Curva Precision-Recall do classificador MLP e LogReg finais nos dados de CV (esquerda) e teste (direita). É possível verificar a area under the curve em cada gráfico. LogReg_A = regressão logística do modelo A, MLP_A_300_HL =* perceptrão multicamada *fine-tuned* com 300 neurónios na *hidden layer.*

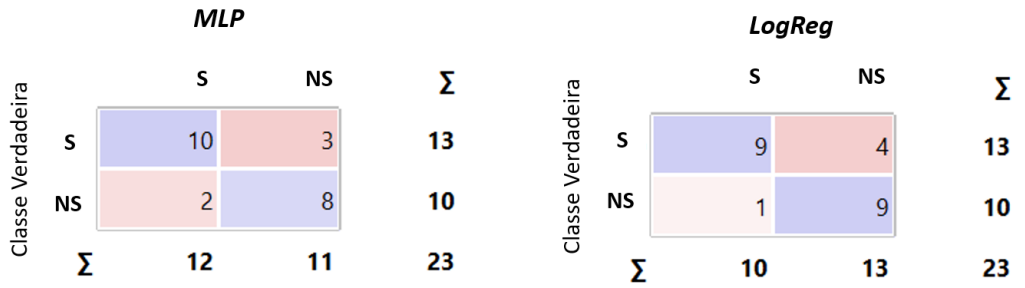


Figura 51 *– Matriz de confusão do grupo teste do classificador MLP (esquerda) e LogReg (direita) finais. LogReg = regressãoa logística do modelo A, MLP = perceptrão multicamada fine-tuned, S = Sobrevivente, NS = Não sobrevivente*

4.1.2 Avaliação da calibração dos modelos finais

Para efeitos de avaliação dos modelos finais, foram criadas curvas de calibração dos modelos finais do grupo de teste, comparando com os criados durante fase de validação (CV). Podemos verificar os resultados na **Figura 52**.

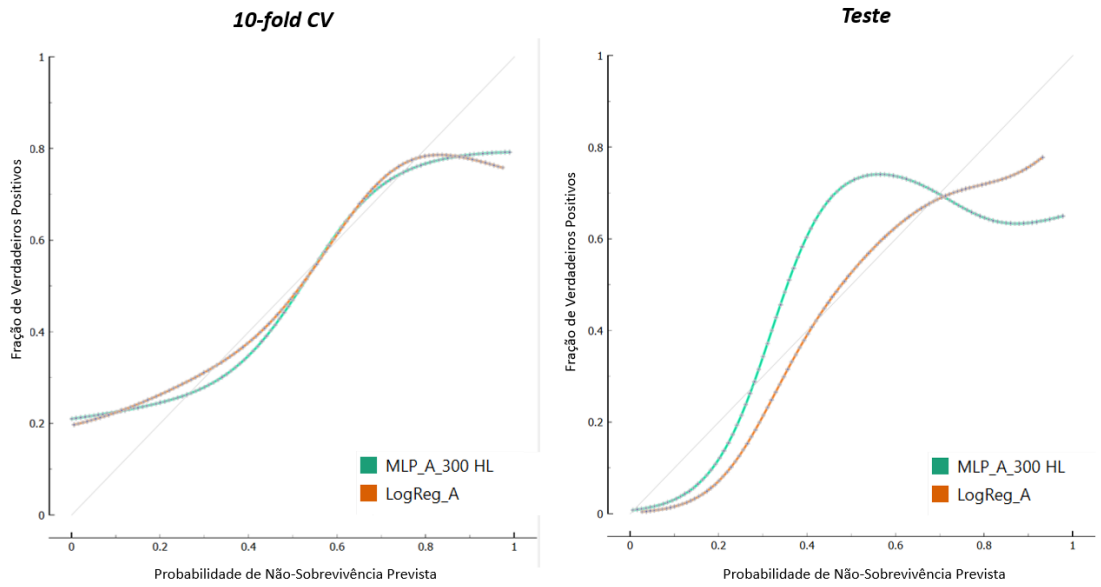


Figura 52 *– Curvas de calibração do classificador MLP e LogReg finais nos dados de CV (esquerda) e teste (direita). A reta onde f(x) = y desenhada, representa uma calibração perfeita. LogReg_A = regressão logística do modelo A, MLP_A_300_HL* = perceptrão multicamada *fine-tuned* com 300 neurónios na *hidden layer.*

4.1.3 Interpretação das previsões

Com o intuído de obter uma melhor interpretação do racional utilizado pelos modelos nas classificações do grupo de teste, foi criada uma tabela onde é possível verificar as previsões de cada modelo (com o seu erro preditivo), as *features* correspondentes e o *label* real dos 23 doentes. É possível observar na **Tabela 12** que existiu três casos comuns de *missclassification* em ambos os modelos (Doente nº37, nº107 e nº112). Também existiram casos de doentes com ARDS-COV19 moderada em que os modelos conseguiram prever a sobrevivência. Após esta observação, foi utilizado o *widget explain predictitons* do *Orange Data Mining*, para compreender as possíveis razões destes acontecimentos, analisando também as radiografias. Os resultados da aplicação desta ferramenta consideram o valor de SHAP para avaliar a contribuição de cada *feature* na previsão selecionada.

Tabela 12 - *Tabela de classificações dos modelos finais do grupo de teste. As features com barras vermelhas são pertencentes à classe de não-sobreviventes, enquanto as que têm barras azuis pertencem à classe de sobreviventes. A dimensão desta barra representa a dimensão relativa do valor da feature. No label de Morte, o valor 1 representa os não-sobreviventes, enquanto o valor 0 representa os sobreviventes. LogReg_A = regressão logística do modelo A, MLP_A_300_HL* = perceptrão multicamada *fine-tuned* com 300 neurónios na *hidden layer, error = erro de classificação, 258 = DLF_258, age = idade*

MLP_A_300_HL	error	LogReg_A	error	Morte	indice	258	pf_d3	age
0.27 : 0.73 → 1	0.268	0.40 : 0.60 → 1	0.403	1	26	0.00230307	175	73
0.51 : 0.49 → 0	0.487	0.41 : 0.59 → 1	0.593	0	36	0.00167328	176	66
0.07 : 0.93 → 1	0.068	0.07 : 0.93 → 1	0.068	1	37	0.00216406	68	81
0.04 : 0.96 → 1	0.956	0.42 : 0.58 → 1	0.575	0	38	0.00158623	284	83
0.88 : 0.12 → 0	0.120	0.92 : 0.08 → 0	0.081	0	48	0.00294691	127	36
0.98 : 0.02 → 0	0.019	0.80 : 0.20 → 0	0.204	0	66	0.000987966	277	54
0.99 : 0.01 → 0	0.007	0.97 : 0.03 → 0	0.028	0	77	0.00305114	266	47
0.45 : 0.55 → 1	0.446	0.38 : 0.62 → 1	0.383	1	100	0.00185536	163	67
0.86 : 0.14 → 0	0.138	0.72 : 0.28 → 0	0.275	0	101	0.00287451	146	57
0.34 : 0.66 → 1	0.664	0.54 : 0.46 → 0	0.461	0	102	0.00227764	125	57
0.45 : 0.55 → 1	0.446	0.46 : 0.54 → 1	0.458	1	103	0.00206961	143	62
0.05 : 0.95 → 1	0.052	0.35 : 0.65 → 1	0.353	1	104	0.00109952	110	51
0.53 : 0.47 → 0	0.533	0.44 : 0.56 → 1	0.442	1	105	0.00322634	88	66
0.81 : 0.19 → 0	0.186	0.48 : 0.52 → 1	0.519	0	106	0.00120913	193	60
0.63 : 0.37 → 0	0.631	0.59 : 0.41 → 0	0.591	1	107	0.00347142	162	74
0.76 : 0.24 → 0	0.237	0.54 : 0.46 → 0	0.455	0	108	0.00176524	182	61
0.34 : 0.66 → 1	0.344	0.31 : 0.69 → 1	0.306	1	109	0.00149206	144	64
0.99 : 0.01 → 0	0.007	0.94 : 0.06 → 0	0.057	0	110	0.00195855	286	48
0.91 : 0.09 → 0	0.093	0.75 : 0.25 → 0	0.249	0	111	0.00257605	169	56
0.15 : 0.85 → 1	0.853	0.24 : 0.76 → 1	0.760	0	112	0.00185603	114	67
0.43 : 0.57 → 1	0.430	0.30 : 0.70 → 1	0.300	1	113	0.0015248	157	67
0.99 : 0.01 → 0	0.007	0.97 : 0.03 → 0	0.034	0	114	0.00268344	182	31
0.02 : 0.98 → 1	0.022	0.33 : 0.67 → 1	0.325	1	115	0.001883	80	56

4.1.4 Estudo de casos de classificações incorretas do grupo de teste

Nas figuras desta secção, podemos observar os gráficos explicativos das previsões por valores de SHAP e as radiografias correspondentes aos doentes classificados incorretamente por ambos os algoritmos.

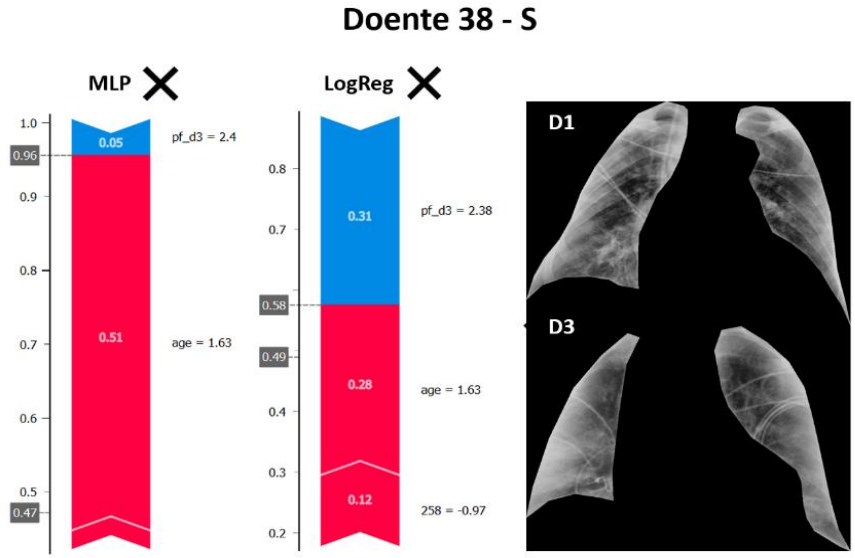


Figura 53 *– Gráficos de valores de SHAP para a classificação do doente 38 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. S = Ground-truth label de sobrevivência, age = Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta*

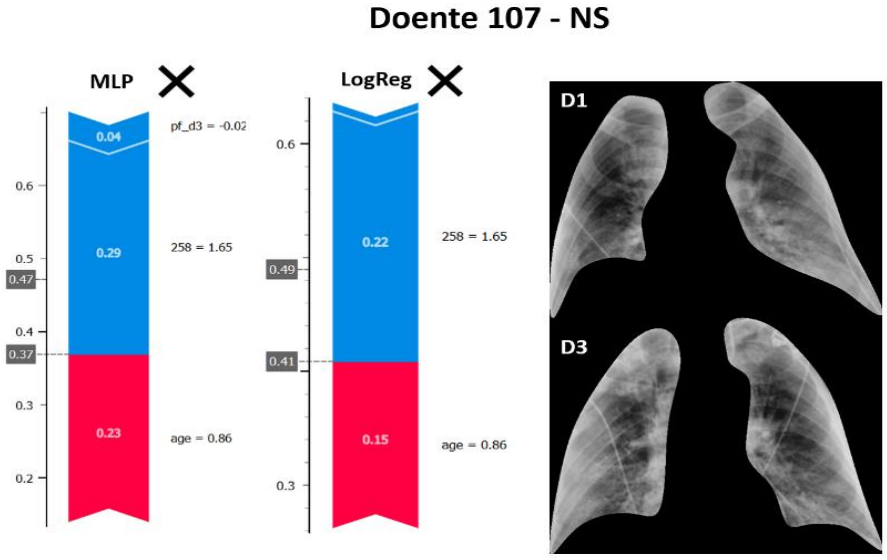


Figura 54 *- Gráficos de valores de SHAP para a classificação do doente 107 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. NS = Ground-truth label de não-sobrevivência, age Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta*

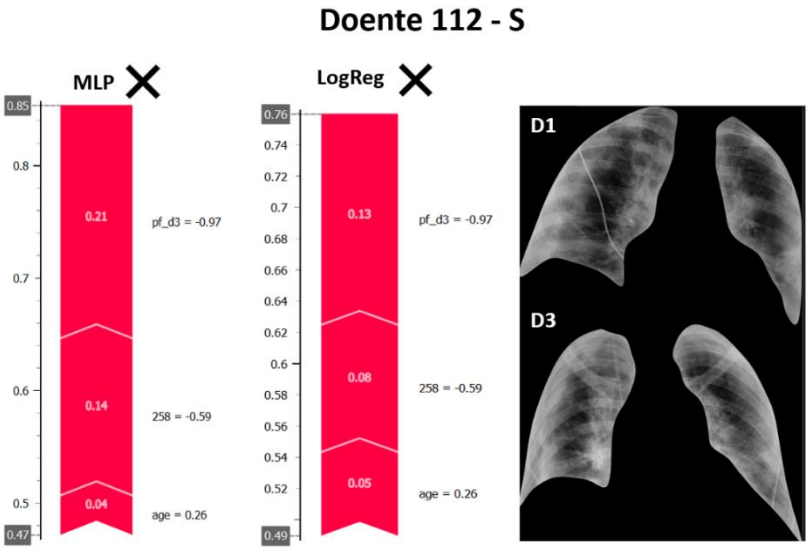


Figura 55 - *Gráficos de valores de SHAP para a classificação do doente 112 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. S = Ground-truth label de sobrevivência, age = Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta*

4.1.5 Estudo de outras classificações de interesse

Nesta secção estão apresentados os resultados de classificações de interesse, onde existiu *miss-classification* em apenas um dos modelos. É possível observar os casos em que os modelos conseguiram prever com sucesso a sobrevivência dos doentes, apesar dos mesmos apresentarem ARDS moderada, que está associado a piores prognósticos.

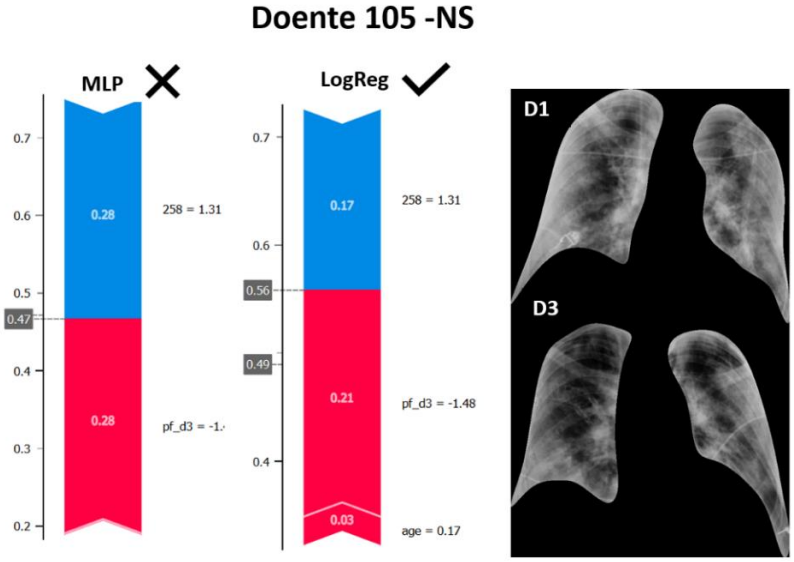


Figura 56 *- Gráficos de valores de SHAP para a classificação do doente 105 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. NS = Ground-truth label de não-sobrevivência, age = Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta e os certos correta*

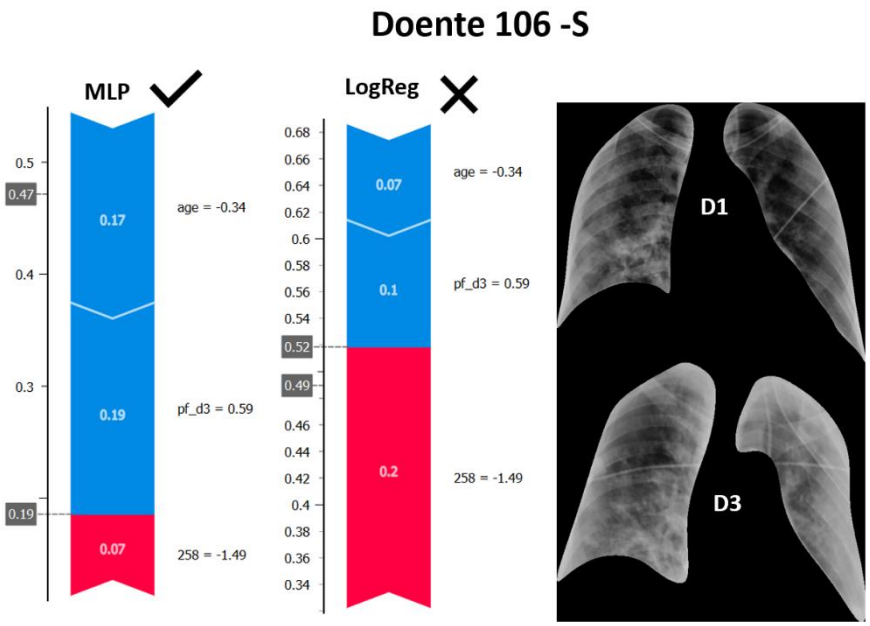


Figura 57 - *Gráficos de valores de SHAP para a classificação do doente 106 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. NS = Ground-truth label de sobrevivência, age = Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta e os certos correta*

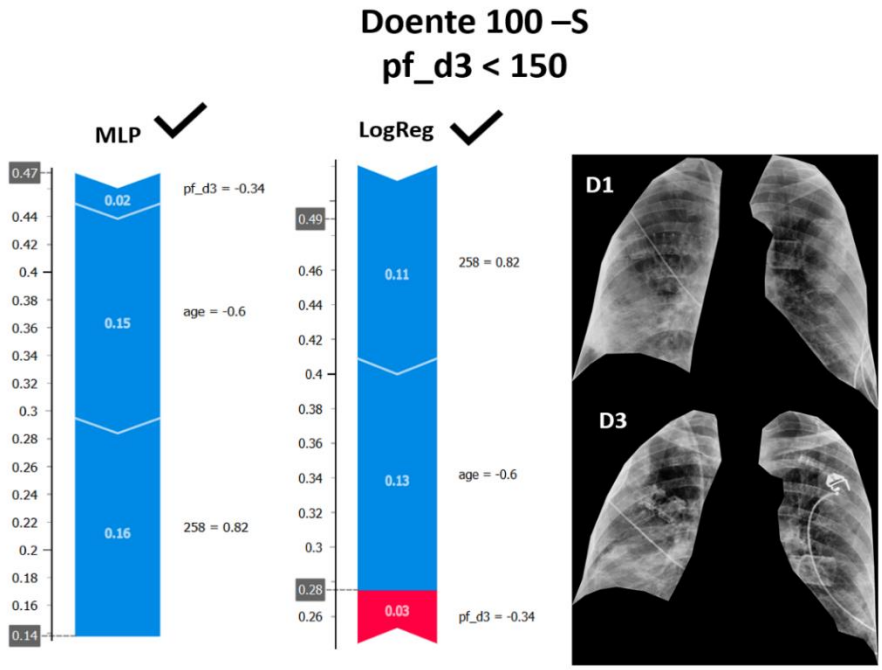


Figura 58 - *Gráficos de valores de SHAP para a classificação do doente 100 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado. S = Ground-truth label de não-sobrevivência, age= Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta e os certos correta*

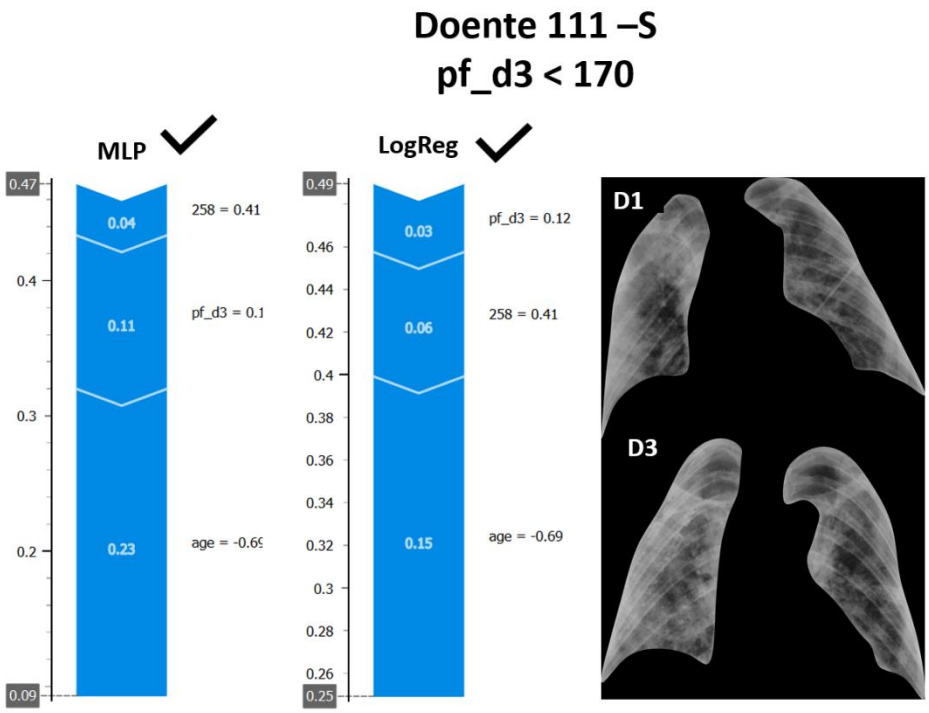


Figura 59 - *Gráficos de valores de SHAP para a classificação do doente 111 e R-RTX correspondente. Setas maiores indicam maior impacto da feature na previsão (vermelho para a classe positiva e azul para a classe negativa). As caixas cinzentas com tracejado indicam a probabilidade da classe positiva, enquanto a sem tracejado indica a probabilidade média do grupo classificado NS = ground-truth label de sobrevivência, age= Idade, 258 = DLF_258, D1 = 1ª dia de VMI, D3 = 3º dia de VMI. As cruzes indicam classificação incorreta e os certos correta*

5. Discussão

5.1 Análise dos Resultados

A presente metodologia de criação de modelos de previsão de mortalidade em doentes com ARDS-COV19 após o 1º dia de VMI, mostraram resultados satisfatórios e com um risco de *overfitting* reduzido, tendo em conta os objetivos propostos e a base de dados reduzida considerando a restante literatura do tópico. Os algoritmos propostos realizam uma avaliação multivariável da doença, o que é fundamental na avaliação e severidade da ARDS, permitindo também o automatismo e o potencial de utilização na identificação de grupos de risco em UCI. Estes fatores podem permitir a utilização destes classificadores em diversas unidades de saúde, pois utilizam parâmetros e métodos de diagnostico acessíveis na maioria das UCI's.

Os presentes resultados indicam o potencial benefício da utilização dos *radiomics* para a extração de *features* de *deep learning* DLF) de radiografias torácicas, importantes para a previsão da mortalidade da ARDS-COV19. Mesmo com poucas amostras (87) é possível utilizar estas características de imagem para o benefício de tarefas de classificação, através da utilização de metodologias de *transfer learning* com CNN's pré-treinadas e seleção de *features*, quando tipicamente são métodos reservados a bases de dados com muitas amostras (*big data)*.

Observando a **Tabela 4** é possível verificar que nas variáveis clínicas e do doente, apenas a idade e a razão pf de d1 e d3, apresentaram médias significativamente diferentes entre os grupos de sobreviventes e não-sobreviventes ($p \leq 0,05$). Estes resultados foram expectáveis, visto que maiores faixas etárias estão associadas a risco de mortalidade superiores em doentes com ARDS comum e ARDS-COV19, principalmente a partir dos 80 anos[124,125]. Por sua vez a ARDS moderada e grave, definida pela razão pf na definição de Berlim, também está associada a maior taxa de mortalidade (40,3% e 46,1%), de acordo com um estudo realizado em UCI's de 50 países[126]. Apesar da possível boa capacidade de discriminação entre as duas classes destas variáveis, os *scatter plots* da **Figura 25** demonstram a existência de *outliers* no gráfico a) e b) da tendência apresentada. Esta observação demonstra que estas variáveis poderão não ser suficientes para uma correta classificação do risco de mortalidade.

Curiosamente, o sexo não demonstrou diferenças significativas na distribuição de frequências entre as duas classes de interesse. No presente estudo existiram mais doentes do sexo masculino do que feminino. A justificação reside em estudos epidemiológicos efetuados que revelam maior probabilidade de desenvolvimento de ARDS-COV19 em doentes do sexo masculino, o que por sua vez, leva a internamento dos mesmos em UCI[127,128]. Os mesmos estudos revelam também maior taxa de mortalidade em doentes do sexo masculino com comorbilidades associadas, quando comparado ao sexo feminino[127,128]. O mesmo não foi verificado no presente estudo (estatisticamente) podendo ser justificado através do número reduzido de amostras não equilibradas entre os sexos, o que não permitiu a inclusão desta variável nos treinos dos modelos finais devido à grande dimensionalidade dos dados. Na **Figura 26** é possível verificar que existe uma correlação forte entre as variáveis dp, peep e ps, o que corresponde às expetativas, pois o cálculo da dp dependeu da peep e da ps[114].

A metodologia aplicada para a extração de *features* das imagens permitiu adquirir com sucesso variáveis que discriminavam bem as classes de interesse, principalmente a DLF_258 (($p \leq 0,005$) (**Tabela 5**). Na **Figura 30** e na **Figura 33** pode-se verificar que esta *feature* é a que apresenta correlação mais fraca com as restantes variáveis (clínicas e de imagem), apresentando bom potencial de aplicação na construção dos modelos de ML. A redução do valor desta variável parece estar associada à mortalidade e em conjunto com a idade,

promove uma boa separação dos grupos de amostras de cada classe. Este comportamento é verificado no gráfico a) da **Figura 34**, onde também existem menos amostras de não-sobreviventes na região azul de sobrevivência, indicando boa identificação desta classe. Apesar destes fatores a metodologia utilizada da concatenação prévia das imagens foi experimental e baseada no trabalho multimodal de *K. Lopez et al*.[78]. Esta poderá não ser a solução mais otimizada, pois a CheXNet original foi treinada para tarefas distintas e com um *input* apenas uma radiografia sem segmentação.

Esta problemática poderá levar a *features* de menor qualidade que seria solucionada com *fine-tuning* prévio ou re-treino total da CNN, no entanto não existem amostras suficientes para esse processo (87 doentes). Outra solução seria a utilização de técnicas de fusão tardia, onde a radiografia de d1 e d3 seriam processadas pela CNN separadamente para efeitos de concatenação final. Este método seria menos "crude" e mais otimizado, no entanto iria aumentar a dimensionalidade dos dados para o dobro e as características obtidas não iriam refletir o conteúdo de ambas as radiografias, removendo a caracterização evolutiva das mesmas.

A concatenação direta das matrizes de duas imagens poderia ter sido evitada com as técnicas de fusão intermédia, onde duas camadas convolucionais para cada radiografia iriam produzir mapas de *features* que seriam então concatenados para as dimensões de *input* da CheXNet, no entanto este método iria requerer o treino prévio da CNN, que não é aconselhável em números de amostras reduzidos. Certos autores utilizam *autoenconders* pré-treinados em radiografias, para primeiro serem extraídos os componentes representativos das imagens antes da concatenação, podendo beneficiar a metodologia[81,129]. Esta técnica utiliza métodos não-supervisionados para codificação dos *inputs* em representações vetoriais compactas, para depois descodificarem as mesmas para obter imagens semelhantes às originais. O método que explora de uma maneira mais interessante para o presente objetivo, seria o uso de LSTM's (redes neuronais recorrentes de *long short-term memory*) para uma avaliação longitudinal das radiografias[76,130,131]. Estas redes recebem *inputs* de quatro dimensões (4D), considerando a componente temporal de uma série de imagens em tarefas de classificação. As LSTM's capturam as dependências dentro de uma série temporal, através do esquecimento ou memorização de informações relevantes para a tarefa[76,130,131]. O uso destas redes depende da existência de várias referências temporais, que não estão presentes neste estudo devido à existência de apenas duas (d1 e d3). Os autores que utilizam estas técnicas com sucesso, tipicamente dispõem de 5 a 8 radiografias em diferentes dias de internamento[130,131].

Durante o processo de seleção de *features* foi possível verificar aspetos interessantes. Em primeiro lugar verificou-se através da **Figura 37** que o número ótimo de *features* a utilizar para as amostras disponíveis, seria de três (considerando todas as variáveis disponíveis). Este valor é próximo do recomendado por *V. Lakshmanan et al.*, confirmando-se a relação[119]. A partir deste valor, é possível verificar que iria existir introdução de ruído no treino do modelo, com valores de exatidão inferiores. O teste de ANOVA efetuado, também permitiu ordenar as *features* por importância de diferenciação de classes com maior sucesso.

É possível verificar na **Figura 38** que a DLF_258 aparenta maior importância que todas as restantes variáveis clínicas, com a exceção da idade da pf_d3, salientando assim a possível necessidade da introdução de dados de imagem médica em avaliações de risco dos doentes de UCI. A pf_d3 revelou ser mais importante que a pf_d1, possivelmente devido ao facto de ser uma medição mais próxima temporalmente do desfecho a classificar (mortalidade), revelando o estado da severidade da doença.

Na avaliação dos diferentes classificadores, verificou-se na **Figura 39** que a LogReg e a MLP ofereciam melhor exatidão e AUC do que os restantes classificadores, tanto no modelo A como no modelo B. Esta revelou maior probabilidade de serem superiores aos restantes, do que serem inferiores ou idênticos (considerando um aumento de 5% da performance insignificante). Esta observação deve-se ao facto da regressão logística ser um classificador principalmente desenhado para classificação binárias, funcionando especialmente bem em circunstâncias onde o limite de decisão é claramente definido por uma linha de separação de classes, apresentando variáveis com relações aproximadamente lineares com o *target* (o que pode ser observado nos *scatter plots* anteriormente referidos). Este método também tem tendência a apresentar bons resultados com poucas amostras, sendo menos sensível ao ruído e *outliers* quando comparado a classificadores de GB e SVM (modelos que necessitam de mais amostras). Os classificadores de SVM podem resultar em circunstâncias semelhantes de linearidade, no entanto são altamente sensíveis aos *kernels* selecionados e aos seus hiperparâmetros, os quais não foram modificados neste estudo para efeitos de simplificação.

Refletindo sobre o MLP, os melhores resultados podem resultar da sua capacidade de capturar interações complexas entre variáveis, independentemente da relação de linearidade das mesmas. No entanto estes algoritmos são extremamente sensíveis aos seus hiperparâmetros e apresentam uma grande complexidade, existindo a possibilidade dos seus resultados promissores deverem-se a *overfitting* durante o processo de treino nos subgrupos de validação cruzada.

Através dos resultados obtidos da

Tabela 6 e

Tabela 7, foi possível comprovar a diferença e probabilidade de melhoria entre os modelos A (com *features* de imagem) e os modelos B (sem *features* de imagem). Pode-se verificar que para a LogReg, o modelo A apresentou melhorias em todas as métricas de *performance*. O Modelo A apresenta uma probabilidade de ser superior ao modelo B de 73,9% para a AUC, de 88,6% para a CA, de 86,4% para o *score* de F1 e de 75,8% para a especificidade. Existe também uma probabilidade significativa dessa melhoria ser superior a 5% em duas destas métricas (63% para a CA e 71,3% para o *score* de F1). A mesma relação está presente para o modelo de MLP, excluindo melhorias da AUC e *recall*, mostrando maior probabilidade de serem idênticas, considerando uma diferença negligenciável de 5% (45,5% para a AUC e 55,7% para a *recall*). O benefício da utilização de dados de imagem juntamente com informações clínicas foi similarmente demonstrado por autores dos estudos relacionados (descritos na secção 2.12). No entanto nenhum destes autores oferece uma perspetiva bayesiana probabilística para identificar o grau e a generalidade dessas melhorias para outros subgrupos de treino, baseando-se apenas em estatística de estimativa com intervalos de confiança, ou demonstrando a quantidade de subgrupos de CV que ofereceram melhores resultados[23,24,108,109]. Na literatura referenciada e estudada, este é o primeiro estudo que oferece esta perspetiva.

Avaliando os modelos finais (A) no grupo de teste (

Tabela 10 e **Tabela 11**) pode-se verificar que a LogReg obteve uma $AUC_{méd}$ de 0,862 95%CI [0.654, 0.969], uma CA de 0,783 95%CI [0.563, 0.926] e um *score* de F1 de 0,783 95%CI [0.563, 0.926], enquanto o modelo de MLP obteve $AUC_{méd}$ de 0,815 95%CI [0.600, 0.945], uma CA de 0,783 95%CI [0.563, 0.926] e um *score* de F1 de 0,727 95%CI [0.528, 0.918]. Apesar de resultados semelhantes no grupo de teste, a $AUC_{méd}$ do modelo LogReg foi superior, com possibilidade de indicar melhor capacidade de discriminação das classes e

generalidade. No entanto, é importante considerar que a quantidade de amostras presentes no grupo de teste interno (23 doentes), não é suficiente para validar os modelos, existindo a possibilidade de variabilidade estatística comprovada pela amplitude dos intervalos de confiança. É expectável que os valores obtidos através da validação cruzada, sejam mais próximos da realidade considerando a LogReg. No caso da MLP, este pode não ser o caso, devido a vários indícios de *overfitting*. Em primeiro lugar existe uma maior diferença da MLP entre o grupo de treino e o grupo de teste, referente à CA e AUC, do que na LogReg.

Na diferença da validação cruzada para o grupo de treino verifica-se a mesma tendência na AUC, *recall*, Spec e Prec. Apesar de uma maior diferença de *recall* entre o grupo de teste e treino na LogReg (0,113) do que na MLP (-0,010), esta foi positiva, podendo indicar boa generalidade. No entanto, é mais provável esta diferença derivar da variabilidade do grupo de teste pequeno. Observando as matrizes de confusão do grupo de teste (**Figura 51**), apesar de existir apenas mais um verdadeiro positivo e menos um verdadeiro negativo na LogReg quando comparado à MLP, foi o suficiente para alterar o *recall* (MLP) de 0,800 para 0,900 (LogReg). Devido a este facto, considerando as métricas de *recall* e Prec, é sempre necessário uma avaliação das suas curvas nos diversos *thresholds* de classificação para avaliação do *overfitting* [132]. Observando assim a **Figura 49** e a **Figura 50**, é possível verificar que existiu uma diferença significativa nas áreas das curvas de precisão-*recall* de treino, validação cruzada e teste no modelo de MLP quando comparado ao modelo de LogReg, sendo um indicador claro de *overfitting* com superioridade da curva do grupo de treino.

Outro indicador do *overfitting* ocorrido, passa pela avaliação das curvas de calibração da validação cruzada e do grupo de teste na **Figura 52**. *M. Pavlou et al.* refere que "modelos *overfitted* têm a tendência a subestimar a probabilidade de um evento em doentes de baixo risco e sobrestima-los em doentes de alto risco [133]. É observado que ambos os modelos sobrestimam, no grupo de teste, a probabilidade de mortalidade nos doentes de baixo risco. No entanto, a partir deste ponto a curva de LogReg é melhor calibrada, mais robusta e semelhante à validação cruzada. A MLP subestima os doentes de médio risco e sobrestima consideravelmente os doentes de alto risco, indicando *overfitting*. Na **Figura 46** também é possível verificar que existiu maior variabilidade de CA e *score* de F1, em cada subgrupo de validação cruzada no modelo de MLP quando comparado ao de LogReg, indicando assim maior variância e resultados de validação cruzada menos fidedignos.

A robustez da LogReg quando comparada ao MLP, pode também ser verificada nas avaliações dos valores de SHAP de cada *feature* nas classificações do grupo de treino. Podemos verificar na **Figura 47** que a LogReg dá uma importância semelhante a cada uma das 3 *features*, para uma verdadeira classificação multivariada, no entanto o MLP poderá dar uma importância excessiva à idade e pouca importância à pf_d3. Na prática clínica, tal importância poderá não ser aceite, dado que umaimportância excessiva em apenas uma variável cria assim um modelo menos robusto, cujos resultados possam ser duvidosos na aplicação clínica diária. Curiosamente ambos os modelos distinguem maior importância da DLF_258 em relação à pf_d3 (método *standard* para a avaliação da severidade da ARDS), indicando mais uma vez o potencial dos fatores imagiológicos na análise de risco dos doentes com ARDS-COV19.

Devido a estas condicionantes o modelo de LogReg apresenta os melhores resultados, sendo assim o modelo a considerar nas análises futuras da presente discussão. Para melhor compreensão dos modelos e do significado da DLF_258, procedeu-se à análise e interpretação de classificações individuais de interesse no grupo de teste, através de valores de SHAP e da avaliação qualitativa das radiografias pelos profissionais de saúde envolvidos.

O doente nº38 (**Figura 53**) trata-se de um falso positivo verificar que o algoritmo LogReg apresentou uma menor probabilidade deste ser positivo (0,58) em comparação com o MLP (0,96). Na LogReg a mesma classificou o doente como positivo (não-sobrevivente) devido ao maior valor de idade e menor valor da DLF_258. Podemos verificar que da radiografia de d1 para d3 existiu, de facto, um agravamento dos padrões radiológicos com maior densidade na base pulmonar esquerda (mais "branca").

O doente 107 (**Figura 54**) foi um falso negativo para ambos os modelos, devido ao elevado valor da DLF_258 adicionando ao facto de parecer ter ocorrido uma melhoria dos padrões da radiografia de d1 para d3, com redução da densidade das bases pulmonares esquerdas. Mais uma vez, a LogReg teve menor certeza nesta classificação do que a MLP.

No doente 105 (**Figura 56**), a LogReg classificou corretamente o mesmo como não sobrevivente apesar da melhoria das radiografias de d1 para d3, devido à reduzida pf_d3 e com algum impacto da idade. Mais uma vez a DLF_258 foi elevada neste caso.

No doente nº100 (**Figura 58**) a LogReg classificou o mesmo como sobrevivente devido a reduzida pf_d3, idade e DLF_259. Este caso é interessante pois não parece existir uma diferença significativa entre ambas as radiografias, no entanto, ambas apresentam densificações consideráveis.

A análise destes doentes leva à possibilidade da DLF_258 considerar ambas as radiografias em termos de melhoria ou agravamento e não apenas uma delas. No entanto, esta avaliação é puramente qualitativa e estudos quantitativos de correlação com o score de RALE devem ser ponderados.

5.2 Comparação dos resultados com os estudos relacionados

Considerando os estudos relacionados descritos na secção 2.12.1, é possível verificar que a metodologia aplicada apresenta resultados semelhantes com uma menor quantidade de amostras.

D. Gourdeau et al utilizaram metodologias híbridas de extração de *features* de radiografias para previsão da mortalidade em doentes COVID-19 sob VMI em UCI[23]. O melhor modelo criado pelos autores foi o que utilizou a combinação de um score de risco associado a variáveis clínicas e duas DLF. Este modelo demonstrou uma AUC de 0,743 95%CI [0.732, 0.746], uma CA de 0,755, uma Spec de 0,828 e uma *recall* de 0,487 (os autores não apresentaram 95%CI). É possível verificar que o modelo LogReg criado nesta dissertação apresentou melhor *performance* nestas métricas, principalmente em *recall* (0,900) e com exceção da Spec (0,692). Existe também um menor risco de viés devido à utilização de um doente por amostra. Os autores também verificaram um *overfitting* significativo entre o grupo de validação (AUC: 0,88) e o grupo de teste (AUC:0,78), que não foi observável nesta dissertação (LogReg). No entanto, a comparação direta dos dois modelos deve ser realizada com precaução, visto que o maior número de amostras no grupo de teste dos autores permitiu obter resultados estatisticamente robustos e com menores 95%CI na AUC. Apesar de existir grande probabilidade dos doentes submetidos a VMI apresentarem ARDS-COV19, esta condição não foi referida pelos autores, bem como não existem dados para comprovar o mesmo, segundo a definição de Berlim. Os autores também não oferecerem metodologias de interpertabilidade dos modelos, sofrendo o estudo de um efeito de *black-box.*

J. Cheng et al. desenvolveram um método de previsão de mortalidade em doentes COVID-19 internados em UCI[108]. Os autores utilizam uma arquitetura complexa de *deep learning* treinada

de raiz para extração de *features* longitudinais das radiografias previamente segmentadas. Dos modelos criados os que apresentam maior similaridade com o estudo atual, são o modelo que apenas utiliza radiografias durante o internamento de UCI e o que combina radiografias longitudinais da UCI e pré-UCI com variáveis clínicas. No primeiro modelo, em grupo de teste externo, os autores apresentam uma AUC de 0,697 95%CI [0.615, 0.776], uma CA 0,657 95%CI [0.583, 0.732], um *score de* F1 0,638 95%CI [0.547, 0.729] e uma Spec de 0,644 95%CI [0.546, 0.745]. No segundo modelo apresentaram uma AUC de 0,727 95%CI [0.645, 0.809], uma CA 0,732 95%CI [0.667, 0.806], um *score de* F1 0,707 95%CI [0.620, 0.786] e uma Spec de 0,746 95%CI [0.648, 0.833]. Podemos verificar que a dissertação com o modelo LogReg (

Tabela 10) atual obteve melhores resultados que ambos os modelos, tanto em validação cruzada como no grupo de teste, com exceção da Spec no grupo de teste. Os autores também não proporcionam métricas obtidas no grupo de validação, tornando a avaliação de possível *overfitting* difícil com um efeito de *blackbox* também presente. No entanto, a comparação direta dos dois modelos deve ser realizada com precaução, visto que o maior número de amostras no grupo de teste dos autores (neste caso externo) permitiu aos mesmos obter resultados estatisticamente robustos e com menores 95%CI nas métricas de *performance*. Apesar de existir uma probabilidade dos doentes submetidos a VMI na UCI apresentarem ARDS-COV19, esta condição não foi referida pelos autores, nem existem dados para comprovar o mesmo, segundo a definição de Berlim. O presente modelo defendido nesta dissertação, poderá oferecer melhor *performance* devido à população mais homogénea, sendo específico e adaptado a uma determinada população.

Jiao et al. pertenderam avaliar a severidade da COVID-19 no momento da admissão hospitalar no departamento de urgência utilizando metodologias de *fine tuning total* de *deep learning*, diferenciando o objetivo da atual dissertação [109]. O modelo combinado com técnicas de fusão tardias de dados de imagem e de variáveis clínicas obteve, no grupo de teste interno, uma AUC de 0,837 95%CI [0.820, 0.849], uma *score* de F1 de 0,806 95%CI [0.790, 0.817] e uma Spec de 0,820 95%CI [0.810, 0.830]. Os valores de CA não são apresentadas. É possivel verificar que estes valores do grupo de treino são superiores ao modelo de LogReg criado nesta dissertação (**Tabela 10**), tanto na validação cruzada como no grupo de teste. No entanto deve ser considerado que o grupo de treino destes autores esteve na ordem dos milhares, enquanto o grupo de treino utilizado no presente estudo encontrou-se na ordem das dezenas (87 doentes). As imagens obtidas em contexto de urgência durante a admissão hospitalar, apresentam também maior qualidade quando comparadas a R-RTX portáteis. Este fator beneficia algoritmos de classificação e a qualidade das *features* extraídas [23]. Na avaliação do grupo de teste as métricas reduziram substancialmente no estudo referido, indicando a presença de *overfitting* para o grupo de treino e teste (AUC de 0,736 95%CI [0.717, 0.754], um *score* de F1 de 0,690 95%CI [0.674, 0.703]. No entanto, a comparação direta dos dois modelos deve ser realizada com precaução, visto que o maior número de amostras no grupo de teste dos autores (neste caso externo) permitiu novamente obter resultados estatisticamente robustos e com menores 95%CI nas métricas de *performance*. O objetivo e a população do estudo também foi distinta, apesar de um objetivo semelhante de prever a grupos de risco. A qualidade e origem das imagens, assim como as variáveis clínicas utilizadas, são distintas das desta dissertação, sendo a comparação entre os mesmos difícil.

As comparações dos estudos relacionados indicam que a metodologia aplicada poderá obter previsões de mortalidade semelhantes ou melhores que modelos mais complexos, considerando uma população mais homogénea e interações entre *features* de imagem e variáveis clínicas. Foi também possível obter uma interpretação considerável da influência

das variáveis em comparação com os estudos referidos. No entanto, a dimensão dos dados de teste e a ausência de um grupo externo, limita a comparação e credibilidade estatística dos resultados desta dissertação.

5.3 Limitações do estudo

A principal limitação do presente trabalho incide nas reduzidas amostras da base de dados. Menores dados de treino e poucos dados de teste podem limitar a avaliação dos modelos, introduzindo viés e variância principalmente em tarefas de maior complexidade[55,57]. Apesar da verificação da importância da DLF_258 na previsão de mortalidade de doentes com ARDS-COV19, a impossibilidade de *fine-tuning* e calibração de uma CNN *end-to-end*, não permite obter uma interpretação visual do que a mesma representa. Esta limitação deve-se ao facto das técnicas dependerem da ativação de classe e dos pesos dados pela FCL (numa rede *fine-tuned)* ou treinada para uma determinada classificação (técnicas de GRAD-CAM), não contabilizando assim classificadores externos de ML clássicos [83]. Os mapas de *features* e filtros possivelmente obtidos por si só, são abstratos e podem não oferecer utilização na interpretação clínica. Foi realizada uma avaliação qualitativa, no entanto, esta não é suficiente. Foi possível observar *overfitting* em ambos os modelos, principalmente nas curvas de calibração. No entanto a MLP apresentou-se como o modelo com menos generalidade. Este facto pode-se dever à definição de hiperparâmetros complexos por metodologias menos robustas. Autores indicam a necessidade de realizar validação cruzada com um grupo extra de teste, não considerado no treino e validação de cada subgrupo, através de um *loop* externo e interno de validação cruzada (*nested cross validation)* [55,134]. Esta metodologia em combinação com a pesquisa iterativa de combinação hiperparâmetros (*GridSearch*) poderia resultar em menor *overfitting*, no entanto o *Orange Data Mining*, não possui esta funcionalidade. A metodologia de obtenção de intervalos de confiança também poderia ser melhorada através da aplicação de múltiplas iterações de *bootstrap*, oferecendo maior confiança nos resultados de validação através da criação e validação de múltiplos grupos de dados sintéticos, aleatórios e independentes [55,134]. No entanto, esta técnica não está disponível por iterações automáticas nos *softwares* utilizados e pode obter resultados menos fidedignos em bases de dados com poucas amostras devido à variabilidade estatística e maior impacto da presença de *outliers* [134].

6. Conclusões e Trabalho Futuro

A ARDS é uma síndrome heterogénea e grave com alta incidência e taxa de mortalidade nas UCI's de todo o mundo. Esta síndrome pode ter diferentes etiologias, estando coassociada a diferentes tipos de doença como é o caso da Sars-CoV-2, não sendo passível de uma análise uni-variada. Nestes contextos mais graves, os doentes tipicamente necessitam de VMI, sendo essencial a identificação de grupos de risco nos primeiros dias de internamento para a correta decisão clínica e terapêutica com base na medicina personalizada, o que se pode tornar difícil devido à heterogeneidade apontada. O ensaio RECOVERY, demonstrou pela primeira vez o benefício da aplicação de esteroides na terapêutica da ARDS, provavelmente devido à homogeneidade presente no estudo com doentes ARDS-COV19, apontando para a possível presença de subfenótipos e para a importância da investigação em amostras homogêneas desta doença. A presente dissertação incidiu nesta crescente necessidade, utilizando uma amostra de 110 doentes com ARDS-COV19 comprovada pela definição de Berlim e por métodos de PCR.

Com o objetivo da criação de um método automático de previsão de mortalidade nos primeiros dias de VMI em UCI'S, foram utilizados parâmetros de ventilação, valores de gasometria e radiografias de tórax do 1º e 3º dia de VMI para construção de modelos de *machine learning* generalizáveis e com níveis de *performance* promissores. Devido ao grande volume de imagens produzidas e à sua eficácia econômica, a radiografia de tórax é uma proposta atraente para a investigação de modelos de previsão de mortalidade em doentes com ARDS. Ao contrário da tomografia computorizada, a radiografia é utilizada diariamente em doentes internados e intransportáveis, permitindo averiguar a evolução da doença. Os padrões radiológicos da mesma, são utilizados na própria definição de Berlim e na avaliação do prognóstico do doente nas UCI's.

Consequentemente foi possível extrair 1024 características numéricas para a previsão de mortalidade de doentes com ARDS-COV19, através do pré-processamento da imagem (segmentação e otimização do contraste/ruído) e da extração automática de *deep learning features*, utilizando uma CNN pré-treinada com R-RTX (CheXnet) para tarefas de classificação de diversas patologias torácicas. Na tentativa de obter informação evolutiva de ambas as radiografias, técnicas de *early fusion* foram aplicadas. Para este efeito procedeu-se concatenação das radiografias do 1º e 3º dia de VMI no input da CNN. Estas *features* foram então concatenadas às variáveis clínicas para treino de dois classificadores de *machine learning* tradicionais, um MLP e um algoritmo de LogReg (técnica híbrida de extração de *features*). Após rigorosa análise das variáveis, foram selecionadas as mais importantes por técnicas de filtragem e de validação cruzada, evitando *overfitting* e a "maldição da dimensionalidade". Foi verificado que a utilização da idade do doente, da sua razão PF do 3º dia de VMI e da DLF_258, permitiu a melhor exatidão nos classificadores. Foi possível concluir também que a introdução da DLF_258 no treino dos classificadores teve uma probabilidade superior a 75% de melhorar a exatidão e score de F1 dos mesmos, ou superior a 60% se uma melhoria de 5% for considerada negligenciável. Através da análise dos valores de *SHAP* dos modelos treinados, verificou-se que a DLF_258 foi mais importante que a pf_d3 na previsão da mortalidade dos doentes durante o treino, indicando o potencial da importância dos *radiomics* na ARDS.

Após *fine-tuning* e treino final dos modelos com 87 amostras, foi possível obter uma AUC de 0,86, exatidão de 0,78 e sensibilidade de 0,90 na classificação do grupo de teste interno (23

amostras) com o algoritmo de LogReg, mostrando este melhor calibração, robustez e menor probabilidade de *overfitting* quando comparado ao MLP. Apresentou também boa capacidade de discriminação de classes e de identificação de doentes com maior risco de mortalidade, apesar de uma baixa especificidade e variabilidade nos subgrupos de validação cruzada. As métricas foram semelhantes ao grupo de treino, mas superiores à validação cruzada. Este facto pode indicar uma ótima generalidade do modelo, no entanto é mais provável que os resultados sejam demasiado otimistas e tenham sucedido devido a aleatoriedade estatística e variabilidade das poucas amostras de teste, como indicado pelas dimensões dos intervalos de confiança das métricas de *performance* do grupo de teste. A quantidade de dados foi assim a principal limitação desta dissertação introduzindo maior risco de viés e variância. O facto de não ter existido um treino ou *fine-tune* da CNN devido à reduzida quantidade de dados, também impossibilita a localização/identificação do que representa a DL_258 através de técnicas comuns como *Gradient-weighted Class Activation Mapping* produzindo um algoritmo de *blackbox.* No entanto a análise da contribuição desta *feature* na classificação demonstrou possibilidade da mesma estar associada a modificações dos padrões radiológicos do 1º e 3º dia de VMI. O *input* das imagens concatenadas na CNN, é também significativamente diferente das imagens originais com que a mesma foi treinada, com possibilidade de não produzindo DLF's otimizadas.

Apesar do referido, estas limitações devem servir de impulsão para trabalhos futuros, visto que foi demonstrado por métodos estatísticos bayesianos que a introdução de informação radiológica tem uma alta probabilidade de promover a identificação de grupos de risco na ARDS-COV19 em algoritmos de classificação de *machine learning*, considerando bases de dados pequenas. O mesmo ainda não tinha sido demonstrado concretamente na literatura mencionada, apenas por estimativa, intervalos de confiança e testes de hipóteses, que podem ser limitativos. Os modelos construídos conseguiram alcançar métricas de *performance* promissoras em validação cruzada e no grupo de teste interno (considerando a reduzida a dimensão dos dados), usando informações disponíveis em qualquer UCI, independentemente dos recursos económicos e profissionais. É esperado que os modelos sejam em breve validados prospectivamente e que sejam úteis na seleção de doentes com maior risco de mortalidade.

O trabalho reuniu assim interesse da equipa clínica, no qual se inseriu e novo(a)s imagens/dados clínicos estão a ser ativamente recolhidos, tanto em Portugal como em centros clínicos de outros países. No futuro, novas metodologias deverão ser aplicadas e otimizadas. O *fine-tune* de uma CNN *end-to-end* deverá ser explorado, assim como a utilização de características de *radiomics hand-crafted* que poderão produzir resultados mais interpretáveis e com menores dados de treino necessários. A diferenciação dessas características entre doentes vacinados e não-vacinados para a COVID-19 necessita também de analise para identificação de possíveis subfenótipos de ARDS neste contexto. A exploração de diferentes metodologias de pré-processamento das imagens deve ser também considerada, verificando quais os métodos que promovem melhores resultados para a tarefa em questão e não apenas considerando aqueles que funcionaram anteriormente na literatura científica.

A combinação de métodos quantitativos de *radiomics* com a avaliação clínica e bioquímica diária em UCI, poderá ser assim o futuro da medicina personalizada no contexto da ARDS-COV19, servindo de mapa ao que a heterogeneidade da mesma quer esconder.

7. Bibliografia

1. Ashbaugh D, Boyd Bigelow D, Petty T, Levine B. ACUTE RESPIRATORY DISTRESS IN ADULTS. Lancet [Internet]. 1967 Aug;290(7511):319–23. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673667901687

2. ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin Definition. JAMA [Internet]. 2012 Jun 20;307(23):2526–33. Available from: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2012.5669

3. Matthay MA, Zemans RL, Zimmerman GA, Arabi YM, Beitler JR, Mercat A, et al. Acute respiratory distress syndrome. Nat Rev Dis Prim [Internet]. 2019 Mar 14;5(1):18. Available from: http://dx.doi.org/10.1038/s41572-019-0069-0

4. Ware LB, Matthay MA. The acute respiratory distress syndrome. N Engl J Med [Internet]. 2000 May 4;342(18):1334–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10793167

5. Wildi K, Livingstone S, Palmieri C, LiBassi G, Suen J, Fraser J. Correction to: The discovery of biological subphenotypes in ARDS: a novel approach to targeted medicine? J Intensive Care [Internet]. 2021 Dec 25;9(1):22. Available from: https://jintensivecare.biomedcentral.com/articles/10.1186/s40560-021-00534-y

6. Maddali M V, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, et al. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. Lancet Respir Med [Internet]. 2022 Apr;10(4):367–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/35026177

7. Krynytska I, Marushchak M, Birchenko I, Dovgalyuk A, Tokarskyy O. COVID-19-associated acute respiratory distress syndrome versus classical acute respiratory distress syndrome (a narrative review). Iran J Microbiol [Internet]. 2021 Dec 22;13(6):737–47. Available from: https://publish.kne-publishing.com/index.php/IJM/article/view/8072

8. Acute Respiratory Distress Syndrome Network, Brower RG, Matthay MA, Morris A, Schoenfeld D, Thompson BT, et al. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. N Engl J Med [Internet]. 2000 May 4;342(18):1301–8. Available from: http://journals.lww.com/00132586-200102000-00017

9. Shankar-Hari M, Rubenfeld GD. Population enrichment for critical care trials: phenotypes and differential outcomes. Curr Opin Crit Care [Internet]. 2019 Oct;25(5):489–97. Available from: http://journals.lww.com/00075198-201910000-00010

10. Meyer NJ, Gattinoni L, Calfee CS. Acute respiratory distress syndrome. Lancet [Internet]. 2021 Aug;398(10300):622–37. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0140673621004396

11. RECOVERY Collaborative Group, Horby P, Lim WS, Emberson JR, Mafham M, Bell JL, et al. Dexamethasone in Hospitalized Patients with Covid-19. N Engl J Med [Internet]. 2021 Feb 25;384(8):693–704. Available from: http://www.nejm.org/doi/10.1056/NEJMoa2021436

12. Villar J, Ferrando C, Martínez D, Ambrós A, Muñoz T, Soler JA, et al. Dexamethasone treatment for the acute respiratory distress syndrome: a multicentre, randomised controlled trial. Lancet Respir Med [Internet]. 2020 Mar;8(3):267–76. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32043986

13. Matthay MA, Arabi YM, Siegel ER, Ware LB, Bos LDJ, Sinha P, et al. Phenotypes and personalized medicine in the acute respiratory distress syndrome. Intensive Care Med [Internet]. 2020 Dec;46(12):2136–52. Available from: https://doi.org/10.1007/s00134-020-06296-9

14. Zompatori M, Ciccarese F, Fasano L. Overview of current lung imaging in acute respiratory distress syndrome. Eur Respir Rev [Internet]. 2014 Dec 1;23(134):519–30. Available from: http://err.ersjournals.com/lookup/doi/10.1183/09059180.00001314

15. Hui TCH, Khoo HW, Young BE, Mohamed S, Mohideen H, Lee YS, et al. Clinical utility of chest radiography for severe COVID-19. 2020;10(7):1540–50.

16. Warren MA, Zhao Z, Koyama T, Bastarache JA, Shaver CM, Semler MW, et al. Severity scoring of lung e dema on the chest radiograph is associated with clinical outcomes in ARDS. Thorax [Internet]. 2018 Sep;73(9):840–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29903755

17. Jabaudon M, Audard J, Pereira B, Jaber S, Lefrant JY, Blondonnet R, et al. Early Changes Over Time in the Radiographic Assessment of Lung Edema Score Are Associated With Survival in ARDS. Chest [Internet]. 2020 Dec;158(6):2394–403. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32659235

18. Le S, Pellegrini E, Green-Saxena A, Summers C, Hoffman J, Calvert J, et al. Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). J Crit Care [Internet]. 2020 Dec;60:96–102. Available from: https://doi.org/10.1016/j.jcrc.2020.07.019

19. Reamaroon N, Sjoding MW, Gryak J, Athey BD, Najarian K, Derksen H. Automated detection of acute respiratory distress syndrome from chest X-Rays using Directionality Measure and deep learning features. Comput Biol Med [Internet]. 2021 Jul;134:104463. Available from: http://www.ncbi.nlm.nih.gov/pubmed/33993014

20. Sjoding MW, Taylor D, Motyka J, Lee E, Co I, Claar D, et al. Deep learning to detect acute respiratory distress syndrome on chest radiographs: a retrospective study with external validation. Lancet Digit Heal [Internet]. 2021 Jun;3(6):e340–8. Available from: http://dx.doi.org/10.1016/S2589-7500(21)00056-X

21. Forel JM, Guervilly C, Hraiech S, Voillet F, Thomas G, Somma C, et al. Type III procollagen is a reliable marker of ARDS-associated lung fibroproliferation. Intensive Care Med [Internet]. 2015 Jan;41(1):1–11. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25354475

22. Johnson A. Identifying and Assessing the Severity of Acute Respiratory Distress Syndrome with Machine Learning Methods. 2020;

23. Gourdeau D, Potvin O, Biem JH, Cloutier F, Abrougui L, Archambault P, et al. Deep learning of chest X-rays can predict mechanical ventilation outcome in ICU-admitted COVID-19 patients. Sci Rep [Internet]. 2022 Dec 13;12(1):6193. Available from: https://doi.org/10.1038/s41598-022-10136-9

24. Stubblefield J, Hervert M, Causey JL, Qualls JA, Dong W, Cai L, et al. Transfer learning with chest X-rays for ER patient classification. Sci Rep [Internet]. 2020 Dec 1;10(1):20900. Available from: https://doi.org/10.1038/s41598-020-78060-4

25. Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. J Thorac Oncol [Internet]. 2010 Sep;5(9):1315–6. Available from: http://dx.doi.org/10.1097/JTO.0b013e3181ec173d

26. Wilson JG, Calfee CS. ARDS Subphenotypes: Understanding a Heterogeneous Syndrome. Crit Care [Internet]. 2020 Dec 24;24(1):102. Available from:

https://ccforum.biomedcentral.com/articles/10.1186/s13054-020-2778-x

27. Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, et al. Report of the American-European consensus conference on ARDS: Definitions, mechanisms, relevant outcomes and clinical trial coordination. Intensive Care Med [Internet]. 1994 Mar;20(3):225–32. Available from: http://link.springer.com/10.1007/BF01704707

28. Hudson LD, Milberg JA, Anardi D, Maunder RJ. Clinical risks for development of the acute respiratory distress syndrome. Am J Respir Crit Care Med. 1995 Feb;151(2 Pt 1):293–301.

29. Goldstone J. The pulmonary physician in critical care * 10: Difficult weaning. Thorax [Internet]. 2002 Nov 1;57(11):986–91. Available from: https://thorax.bmj.com/lookup/doi/10.1136/thorax.57.11.986

30. Tomashefski JFJ. Pulmonary pathology of acute respiratory distress syndrome. Clin Chest Med. 2000 Sep;21(3):435–66.

31. Hendrickson KW, Peltan ID, Brown SM. The Epidemiology of Acute Respiratory Distress Syndrome Before and After Coronavirus Disease 2019. Crit Care Clin [Internet]. 2021;37(4):703–16. Available from: https://doi.org/10.1016/j.ccc.2021.05.001

32. Kelly B. The chest radiograph. Ulster Med J [Internet]. 2012 Sep;81(3):143–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780323399524000044

33. Broder J. Imaging the Chest. In: Diagnostic Imaging for the Emergency Physician [Internet]. Elsevier; 2011. p. 185–296. Available from: https://www.sciencedirect.com/science/article/pii/B9781416061137100055

34. Ou X, Chen X, Xu X, Xie L, Chen X, Hong Z, et al. Recent Development in X-Ray Imaging Technology: Future and Challenges. Research [Internet]. 2021 Jan;2021. Available from: https://spj.science.org/doi/10.34133/2021/9892152

35. Tahir AM, Chowdhury MEH, Khandakar A, Rahman T, Qiblawey Y, Khurshid U, et al. COVID-19 infection localization and severity grading from chest X-ray images. Comput Biol Med [Internet]. 2021;139(August):105002. Available from: https://doi.org/10.1016/j.compbiomed.2021.105002

36. Huda W, Abrahams RB. X-Ray-Based Medical Imaging and Resolution. 2015;(April):393–7.

37. Huda W, Abrahams RB. Radiographic Techniques, Contrast, and Noise in X-Ray Imaging. Am J Roentgenol [Internet]. 2015 Feb;204(2):W126–31. Available from: https://www.ajronline.org/doi/10.2214/AJR.14.13116

38. Huang S, Wang YC, Ju S. Advances in medical imaging to evaluate acute respiratory distress syndrome. Chinese J Acad Radiol [Internet]. 2022;5(1):1–9. Available from: https://doi.org/10.1007/s42058-021-00078-y

39. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature [Internet]. 2015 May 28;521(7553):436–44. Available from: http://www.nature.com/articles/nature14539

40. Rogers W, Thulasi Seetha S, Refaee TAG, Lieverse RIY, Granzier RWY, Ibrahim A, et al. Radiomics: from qualitative to quantitative imaging. Br J Radiol [Internet]. 2020 Apr;93(1108):20190948. Available from: https://www.birpublications.org/doi/10.1259/bjr.20190948

41. Borstelmann SM. Machine Learning Principles for Radiology Investigators. Acad Radiol [Internet]. 2020 Jan;27(1):13–25. Available from: https://doi.org/10.1016/j.acra.2019.07.030

42. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: From machine learning to deep learning. Phys Medica [Internet]. 2021 Mar;83(February):9–24. Available from: https://doi.org/10.1016/j.ejmp.2021.02.006

43. Montagnon E, Cerny M, Cadrin-Chênevert A, Hamilton V, Derennes T, Ilinca A, et al. Deep learning workflow in radiology: a primer. Insights Imaging [Internet]. 2020 Dec 10;11(1):22. Available from: https://insightsimaging.springeropen.com/articles/10.1186/s13244-019-0832-5

44. Zhang X, Zhang Y, Zhang G, Qiu X, Tan W, Yin X, et al. Deep Learning With Radiomics for Disease Diagnosis and Treatment: Challenges and Potential. Front Oncol [Internet]. 2022 Feb 17;12(February):1–25. Available from: https://www.frontiersin.org/articles/10.3389/fonc.2022.773840/full

45. Raza K, Singh NK. A Tour of Unsupervised Deep Learning for Medical Image Analysis. Curr Med Imaging Former Curr Med Imaging Rev [Internet]. 2021 Sep;17(9):1059–77. Available from: https://pubmed.ncbi.nlm.nih.gov/33504314/

46. Willemink MJ, Koszek WA, Hardell MSC, Wu MSJ, Rubin DL, P MSM. Preparing Medical Imaging Data for Machine Learning. 2020;(21).

47. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell [Internet]. 2020 Jun 8;2(6):305–11. Available from: http://dx.doi.org/10.1038/s42256-020-0186-1

48. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: A practical guide for beginners. Diagnostic Interv Radiol. 2019;25(6):485–95.

49. Montesinos López OA, Montesinos López A, Crossa J. Overfitting, Model Tuning, and Evaluation of Prediction Performance. In: Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet]. Cham: Springer International Publishing; 2022. p. 109–39. Available from: https://doi.org/10.1007/978-3-030-89010-0_4

50. Nti IK, Nyarko-Boateng O, Aning J. Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. Int J Inf Technol Comput Sci [Internet]. 2021 Dec 8;13(6):61–71. Available from: https://www.mecs-press.org/ijitcs/ijitcs-v13-n6/v13n6-5.html

51. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res. 2010;11:2079–107.

52. BURMAN P. A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. Biometrika [Internet]. 1989;76(3):503–14. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/76.3.503

53. Tian L, Cai T, Goetghebeur E, Wei LJ. Model evaluation based on the sampling distribution of estimated absolute prediction error. Biometrika [Internet]. 2007 Feb 28;94(2):297–311. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/asm036

54. Nti IK, Nyarko-Boateng O, Aning J. Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. Int J Inf Technol Comput Sci. 2021 Dec;13(6):61–71.

55. Singh V, Pencina M, Einstein AJ, Liang JX, Berman DS, Slomka P. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. Sci Rep [Internet]. 2021 Jul 14;11(1):14490. Available from: https://doi.org/10.1038/s41598-021-93651-5

56. Isaksson A, Wallman M, Göransson H, Gustafsson MG. Cross-validation and bootstrapping are unreliable in small sample classification. Pattern Recognit Lett. 2008;29(14):1960–5.

57. An C, Park YW, Ahn SS, Han K, Kim H, Lee SK. Radiomics machine learning study with a small sample size: Single random training-test set split may lead to unreliable results. PLoS One [Internet]. 2021;16(8 August):1–13. Available from: http://dx.doi.org/10.1371/journal.pone.0256152

58. Fan C, Chen M, Wang X, Wang J, Huang B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. Front Energy Res [Internet]. 2021 Mar 29;9(March):1–17. Available from: https://www.frontiersin.org/articles/10.3389/fenrg.2021.652801/full

59. Tahir AM, Qiblawey Y, Khandakar A, Rahman T, Khurshid U, Musharavati F, et al. Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images. Cognit Comput. 2022;2019(December 2019).

60. Heidari M, Mirniaharikandehei S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. Int J Med Inform [Internet]. 2020;144(June):104284. Available from: https://doi.org/10.1016/j.ijmedinf.2020.104284

61. Wagner MW, Namdar K, Biswas A, Monah S, Khalvati F, Ertl-Wagner BB. Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. Neuroradiology [Internet]. 2021;63(12):1957–67. Available from: https://doi.org/10.1007/s00234-021-02813-9

62. Caseneuve G, Valova I, LeBlanc N, Thibodeau M. Chest X-Ray image preprocessing for disease classification. Procedia Comput Sci [Internet]. 2021;192:658–65. Available from: https://doi.org/10.1016/j.procs.2021.08.068

63. Demircioğlu A. Predictive performance of radiomic models based on features extracted from pretrained deep networks. Insights Imaging [Internet]. 2022 Dec 9;13(1):187. Available from: https://doi.org/10.1186/s13244-022-01328-y

64. Salvi M, Acharya UR, Molinari F, Meiburger KM. The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. Comput Biol Med [Internet]. 2021 Jan;128:104129. Available from: https://doi.org/10.1016/j.compbiomed.2020.104129

65. Olisah CC, Smith L. Understanding unconventional preprocessors in deep convolutional neural networks for face identification. SN Appl Sci [Internet]. 2019 Nov 30;1(11):1511. Available from: https://doi.org/10.1007/s42452-019-1538-5

66. Giełczyk A, Marciniak A, Tarczewska M, Lutowski Z. Pre-processing methods in chest X-ray image classification. Mehmood Z, editor. PLoS One [Internet]. 2022 Apr 5;17(4):e0265949. Available from: http://dx.doi.org/10.1371/journal.pone.0265949

67. Nefoussi S, Amamra A, Amarouche IA. A Comparative Study of Chest X-Ray Image Enhancement Techniques for Pneumonia Recognition. In 2021. p. 276–88. Available from: http://link.springer.com/10.1007/978-3-030-69418-0_25

68. Lin W, Hasenstab K, Moura Cunha G, Schwartzman A. Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. Sci Rep [Internet]. 2020 Dec 23;10(1):20336. Available from: https://doi.org/10.1038/s41598-020-77264-y

69. Truhn D, Schrading S, Haarburger C, Schneider H, Merhof D, Kuhl C. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing

Lesions at Multiparametric Breast MRI. Radiology [Internet]. 2019 Feb;290(2):290–7. Available from: http://pubs.rsna.org/doi/10.1148/radiol.2018181352

70. Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. Neurocomputing [Internet]. 2021;444:92–110. Available from: https://doi.org/10.1016/j.neucom.2020.04.157

71. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data [Internet]. 2021 Mar 31;8(1):53. Available from: https://doi.org/10.1186/s40537-021-00444-8

72. Hackenberger BK. Tensors all around us. Croat Med J [Internet]. 2019 Aug;60(4):369–74. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6734579/

73. Patil A, Rane M. Convolutional Neural Networks: An Overview and Its Applications in Pattern Recognition. In: Smart Innovation, Systems and Technologies [Internet]. Insights into Imaging; 2021. p. 21–30. Available from: https://insightsimaging.springeropen.com/track/pdf/10.1007/s13244-018-0639-9.pdf

74. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [Internet]. IEEE; 2017. p. 2261–9. Available from: https://ieeexplore.ieee.org/document/8099726/

75. Le Dinh T, Lee SH, Kwon SG, Kwon KR. COVID-19 Chest X-ray Classification and Severity Assessment Using Convolutional and Transformer Neural Networks. Appl Sci [Internet]. 2022 May 11;12(10):4861. Available from: https://www.mdpi.com/2076-3417/12/10/4861

76. Wu Y, Rocha BM, Kaimakamis E, Cheimariotis GA, Petmezas G, Chatzis E, et al. A deep learning method for predicting the COVID-19 ICU patient outcome fusing X-rays, respiratory sounds, and ICU parameters. Expert Syst Appl [Internet]. 2024;235(June 2023):121089. Available from: https://doi.org/10.1016/j.eswa.2023.121089

77. Rahman T, Chowdhury MEH, Khandakar A, Mahbub Z Bin, Hossain MSA, Alhatou A, et al. BIO-CXRNET: a robust multimodal stacking machine learning technique for mortality risk prediction of COVID-19 patients using chest X-ray images and clinical data. Neural Comput Appl [Internet]. 2023;35(24):17461–83. Available from: https://doi.org/10.1007/s00521-023-08606-w

78. Lopez K, Fodeh SJ, Allam A, Brandt CA, Krauthammer M. Reducing Annotation Burden Through Multimodal Learning. Front Big Data. 2020;3(June).

79. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017 Nov 14;3–9. Available from: https://arxiv.org/abs/1711.05225

80. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proc - 30th IEEE Conf Comput Vis Pattern Recognition, CVPR 2017. 2017;2017-Janua:3462–71.

81. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: A review. Brief Bioinform. 2022;23(2):1–15.

82. Kim HE, Cosa-Linan A, Santhanam N, Jannesari M, Maros ME, Ganslandt T. Transfer learning for medical image classification: a literature review. BMC Med Imaging [Internet]. 2022 Dec 13;22(1):69. Available from: https://doi.org/10.1186/s12880-022-

00793-7

83. Salahuddin Z, Woodruff HC, Chatterjee A, Lambin P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Comput Biol Med [Internet]. 2022;140(October 2021):105111. Available from: https://doi.org/10.1016/j.compbiomed.2021.105111

84. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Adv Neural Inf Process Syst. 2019;32(NeurIPS).

85. Pudjihartono N, Fadason T, Kempa-liehr AW. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. 2022;2(June):1–17.

86. Hua J, Xiong Z, Lowey J, Suh E, Dougherty ER. Optimal number of features as a function of sample size for various classification rules. Bioinformatics [Internet]. 2005 Apr 15;21(8):1509–15. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bti171

87. Demircioğlu A. Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights Imaging [Internet]. 2021;12(1):1–10. Available from: https://doi.org/10.1186/s13244-021-01115-1

88. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci [Internet]. 2021 May 22;2(3):160. Available from: https://doi.org/10.1007/s42979-021-00592-x

89. Benhar H, Idri A, Hosni M. Impact of threshold values for filter-based univariate feature selection in heart disease classification. Heal 2020 - 13th Int Conf Heal Informatics, Proceedings; Part 13th Int Jt Conf Biomed Eng Syst Technol BIOSTEC 2020. 2020;5(Biostec 2020):391–8.

90. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak [Internet]. 2012 Dec 15;12(1):8. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-12-8

91. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal [Internet]. 2020;143:106839. Available from: https://doi.org/10.1016/j.csda.2019.106839

92. Refaeilzadeh P, Tang L, Liu H. On comparison of feature selection algorithms. AAAI Work - Tech Rep. 2007;WS-07-05(January 2007):34–9.

93. Sperandei S. Understanding logistic regression analysis. Biochem Medica [Internet]. 2014;24(1):12–8. Available from: http://www.biochemia-medica.com/en/journal/24/1/10.11613/BM.2014.003

94. Stoltzfus JC. Logistic regression: A brief primer. Acad Emerg Med. 2011;18(10):1099–104.

95. Zanaty EA. Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification. Egypt Informatics J [Internet]. 2012 Nov;13(3):177–83. Available from: http://dx.doi.org/10.1016/j.eij.2012.08.002

96. Tan H. Machine Learning Algorithm for Classification. J Phys Conf Ser [Internet]. 2021 Aug 1;1994(1):012016. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/1994/1/012016

97. Li AH, Bradic J. Boosting in the Presence of Outliers: Adaptive Classification With

Nonconvex Loss Functions. J Am Stat Assoc [Internet]. 2018 Apr 3;113(522):660–74. Available from: https://doi.org/10.1080/01621459.2016.1273116

98. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, et al. Calibration: The Achilles heel of predictive analytics. BMC Med. 2019;17(1):1–7.

99. Silva Filho T, Song H, Perello-Nieto M, Santos-Rodriguez R, Kull M, Flach P. Classifier calibration: a survey on how to assess and improve predicted class probabilities. Mach Learn [Internet]. 2023;112(9):3211–60. Available from: https://doi.org/10.1007/s10994-023-06336-7

100. Hicks SA, Strümke I, Thambawita V, Hammou M, Riegler MA, Halvorsen P, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep [Internet]. 2022 Apr 8;12(1):5979. Available from: https://doi.org/10.1038/s41598-022-09954-8

101. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006;27(8):861–74.

102. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Casp J Intern Med [Internet]. 2013;4(2):627–35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24009950

103. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognit Comput [Internet]. 2023; Available from: https://doi.org/10.1007/s12559-023-10179-8

104. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Nips [Internet]. 2017 May 22;16(3):426–30. Available from: http://arxiv.org/abs/1705.07874

105. Bang M, Eom J, An C, Kim S, Park YW, Ahn SS, et al. An interpretable multiparametric radiomics model for the diagnosis of schizophrenia using magnetic resonance imaging of the corpus callosum. Transl Psychiatry [Internet]. 2021;11(1):1–8. Available from: http://dx.doi.org/10.1038/s41398-021-01586-2

106. Moura LV de, Mattjie C, Dartora CM, Barros RC, Marques da Silva AM. Explainable Machine Learning for COVID-19 Pneumonia Classification With Texture-Based Features Extraction in Chest Radiography. Front Digit Heal [Internet]. 2022 Jan 17;3(January):1–13. Available from: https://www.frontiersin.org/articles/10.3389/fdgth.2021.662343/full

107. Severn C, Suresh K, Görg C, Choi YS, Jain R, Ghosh D. A Pipeline for the Implementation and Visualization of Explainable Machine Learning for Medical Imaging Using Radiomics Features. Sensors. 2022;22(14).

108. Cheng J, Sollee J, Hsieh C, Yue H, Vandal N, Shanahan J, et al. Correction to: COVID-19 mortality prediction in the intensive care unit with deep learning based on longitudinal chest X-rays and clinical data. Eur Radiol [Internet]. 2022 Jul 23;32(7):5034–5034. Available from: https://link.springer.com/10.1007/s00330-022-08680-z

109. Jiao Z, Choi JW, Halsey K, Tran TML, Hsieh B, Wang D, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. Lancet Digit Heal [Internet]. 2021 May;3(5):e286–94. Available from: http://dx.doi.org/10.1016/S2589-7500(21)00039-X

110. Bradski G. The OpenCV Library. Dr Dobb's J Softw Tools. 2000;

111. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods [Internet]. 2012 Jul 28;9(7):671–5. Available from: https://www.nature.com/articles/nmeth.2089

112. Janez D, Curk T, Erjavec A. Orange: Data Mining Toolbox in Python Janez. J Mach

Learn Res. 14(SpecialIssue1):295–300.

113. van Merriënboer B, Bahdanau D, Dumoulin V, Serdyuk D, Warde-Farley D, Chorowski J, et al. Blocks and Fuel: Frameworks for deep learning. 2015 Jun 1;1–5. Available from: http://arxiv.org/abs/1506.00619

114. Williams EC, Motta-Ribeiro GC, Vidal Melo MF. Driving Pressure and Transpulmonary Pressure. Anesthesiology [Internet]. 2019 Jul 1;131(1):155–63. Available from: https://pubs.asahq.org/anesthesiology/article/131/1/155/18072/Driving-Pressure-and-Transpulmonary-PressureHow-Do

115. Akoglu H. User's guide to correlation coefficients. Turkish J Emerg Med. 2018 Sep;18(3):91–3.

116. Chou B, Lee M. CheXNet-Keras [Internet]. GitHub repository. 2020 [cited 2023 Jun 5]. Available from: https://github.com/brucechou1983/CheXNet-Keras

117. Corani G, Benavoli A. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. Mach Learn [Internet]. 2015 Sep 24;100(2–3):285–304. Available from: http://dx.doi.org/10.1007/s10994-015-5486-z

118. Andrade C. The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives. Indian J Psychol Med [Internet]. 2019;41(3):210–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/31142921

119. Lakshmanan V, Robinson S, Munn M. Machine Learning Design Patterns: Solutions to Common Challenges in Data Preparation, Model Building, and MLOps Title. O'Reilly; 2020. 390 p.

120. Demšar J, Zupan B. Hands-on training about overfitting. Palagi PM, editor. PLOS Comput Biol [Internet]. 2021 Mar 4;17(3):e1008671. Available from: https://dx.plos.org/10.1371/journal.pcbi.1008671

121. Claridge-Chang A, Assam PN. Estimation statistics should replace significance testing. Nat Methods [Internet]. 2016 Feb 28;13(2):108–9. Available from: https://doi.org/10.1038/nmeth.3729

122. Tsai WY, Chi Y, Chen CM. Interval estimation of binomial proportion in clinical trials with a two-stage design. Stat Med. 2008;27(1):15–35.

123. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988 Sep;44(3):837–45.

124. Killien EY, Mills B, Vavilala MS, Watson RS, O'Keefe GE, Rivara FP. Association between age and acute respiratory distress syndrome development and mortality following trauma. J Trauma Acute Care Surg. 2019;86(5):844–52.

125. Chiumello D, Modafferi L, Fratti I. Risk Factors and Mortality in Elderly ARDS COVID-19 Compared to Patients without COVID-19. J Clin Med [Internet]. 2022 Sep 1;11(17):5180. Available from: https://www.mdpi.com/2077-0383/11/17/5180

126. Bellani G, Laffey JG, Pham T, Fan E, Brochard L, Esteban A, et al. Epidemiology, Patterns of Care, and Mortality for Patients With Acute Respiratory Distress Syndrome in Intensive Care Units in 50 Countries. JAMA. 2016 Feb;315(8):788–800.

127. Kelada M, Anto A, Dave K, Saleh SN. The Role of Sex in the Risk of Mortality From COVID-19 Amongst Adult Patients: A Systematic Review. Cureus [Internet]. 2020 Aug 29;12(8). Available from: https://www.cureus.com/articles/36138-the-role-of-sex-in-the-risk-of-mortality-from-covid-19-amongst-adult-patients-a-systematic-review

128. Gujski M, Jankowski M, Rabczenko D, Goryński P, Juszczyk G. The Prevalence of Acute Respiratory Distress Syndrome (ARDS) and Outcomes in Hospitalized Patients with COVID-19—A Study Based on Data from the Polish National Hospital Register. Viruses [Internet]. 2022 Jan 1;14(1):76. Available from: https://www.mdpi.com/1999-4915/14/1/76

129. Khan E, Rehman MZU, Ahmed F, Alfouzan FA, Alzahrani NM, Ahmad J. Chest X-ray Classification for the Detection of COVID-19 Using Deep Learning Techniques. Sensors [Internet]. 2022 Feb 5;22(3):1211. Available from: https://www.mdpi.com/1424-8220/22/3/1211

130. Duanmu H, Ren T, Li H, Mehta N, Singer AJ, Levsky JM, et al. Deep learning of longitudinal chest X-ray and clinical variables predicts duration on ventilator and mortality in COVID-19 patients. Biomed Eng Online [Internet]. 2022;21(1):1–15. Available from: https://doi.org/10.1186/s12938-022-01045-z

131. Santeramo R, Withey S, Montana G. Longitudinal detection of radiological abnormalities with time-modulated LSTM. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. 2018 Jul 16;11045 LNCS:326–33. Available from: http://arxiv.org/abs/1807.06144

132. McWhite CD, Papoulas O, Drew K, Dang V, Leggere JC, Sae-Lee W, et al. Co-fractionation/mass spectrometry to identify protein complexes. STAR Protoc [Internet]. 2021 Mar;2(1):100370. Available from: https://www.sciencedirect.com/science/article/pii/S2666166721000770

133. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ [Internet]. 2015 Aug 11;351:h3868. Available from: https://www.bmj.com/lookup/doi/10.1136/bmj.h3868

134. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learn [Internet]. 2018;107(12):1895–922. Available from: https://doi.org/10.1007/s10994-018-5714-4