Maria João Seiça Sousa

Modelo Ativo e Explicável de Aprendizagem Computacional para Deteção Automática da Doença de *Alzheimer*

1 2 9 0

UNIVERSIDADE D COIMBRA

1 2 9 0

UNIVERSIDADE D
COIMBRA

Maria João Seiça Sousa

Modelo Ativo e Explicável de Aprendizagem Computacional para Deteção Automática da Doença de *Alzheimer*

Dissertação no âmbito do Mestrado em Engenharia Biomédica, com especialização no ramo de Informática Clínica e Bioinformática, orientada pelo Professor Doutor Luís Miguel Machado Lopes Macedo, apresentada ao Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Julho de 2024

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

MODELO ATIVO E EXPLICÁVEL DE APRENDIZAGEM COMPUTACIONAL PARA DETEÇÃO AUTOMÁTICA DA DOENÇA DE *ALZHEIMER*

Maria João Seiça Sousa

Dissertação apresentada à Faculdade de Ciências e Tecnologia da Universidade de Coimbra para obtenção do grau de Mestre em Engenharia Biomédica com especialização em Informática Clínica e Bioinformática

Supervisor: Prof. Dr. Luís Macedo

Coimbra, Julho de 2024

Este trabalho foi desenvolvido em colaboração com:

Departmento de Engenharia Informática

Departamento de Física

Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Centre for Informatics and Systems of the University of Coimbra **(CISUC)**

Center For Responsible AI **(NextGenAI)**

Agradecimentos

Gostaria de expressar os meus agradecimentos a todos aqueles que contribuíram, não só para o desenvolvimento deste projeto mas também para o meu crescimento, tanto a nível pessoal como a nível académico.

Ao meu orientador, Professor Doutor Luís Macedo, pela disponibilidade e orientação constantes, fundamentais para o sucesso deste trabalho.

Aos meus pais e irmã, por terem tornado possível este meu caminho, estando sempre presentes em todas as etapas da minha vida. O amor, apoio e motivação dados foram bastante importantes para mim e para s conlusão deste percurso. Obrigada por estarem sempre ao meu lado e serem os meus maiores pilares!

À restante família, nomeadamente avós, padrinhos, tios e primos que me acompanham em todas as etapas da minha vida e motivam a alcançar os meus sonhos e a nunca desistir.

À minha madrinha Beatriz, que ao longo destes 5 anos foi apoio contante, fonte de inspiração e parceira em todas as ocasiões. A tua amizade incondicional foi muito importante para a minha vida e sei que te levo para sempre!

Aos meus amigos "Cheesecakes de Oreo", por terem caminhado ao meu lado nos últimos 2 anos. Foram casa na cidade que já era a minha casa e tornaram a palavra "Saudade" ainda mais marcante. Obrigada por todos os momentos inesquecíveis que vivemos juntos, levo-vos para a vida toda!

Ao meu amigo Rodrigo, que me acompanhou desde o início do meu percurso no ISEC. Obrigada por seres o melhor colega de licenciatura, estágio e mestrado que podia ter tido. Foram horas intermináveis a teu lado a fazer trabalhos, e centenas de momentos inesquecíveis de pura diversão e alegria.

Ao meu grupo de amigos de mestrado, os "Desintegrados", por me acolherem e partilharem comigo tantos momentos memoráveis. A vossa amizade é das melhores coisas que levo deste mestrado.

Por fim, mas não menos importante, agradeço à ISECOTUNA, a tuna que me acolheu desde o início do meu percurso académico. Foram muitos e bons momentos vividos a vosso lado, sempre acompanhados de amizade, companheirismo, música e claro, bafo do dragão. Obrigada por terem dado outro sabor à minha vida académica. Foram uma casa à qual sempre voltarei.

A todos os acima referidos e a todos os outros que também fizeram parte desta grande caminhada, o meu muito obrigada por tudo. Ser-vos-ei eternamente grata.

Resumo

A doença de *Alzheimer* é uma preocupação global na área da saúde, associada à neurodegeneração do cérebro humano e, consequentemente, à perda de memória e capacidade de comunicação. Uma vez que, até ao momento, esta doença não tem cura, é fundamental que seja detetada atempadamente para que se possam delinear estratégias para retardar o seu progresso, como por exemplo a estimulação cognitiva.

Este estudo teve como objetivo a inclusão da Inteligência Artificial (IA) no processo de diagnóstico da doença de *Alzheimer*, como ferramenta de auxílio para detetar, através de imagens Ressonância Magnética (MRI) do cérebro, pequenos sinais desta doença que possam passar despercebidos ao olho humano.

Para esse efeito, foi utilizado um *dataset* retirado da plataforma *Kaggle* constituído por 6400 imagens por MRI do cérebro de diversos pacientes, agrupadas em quatro classes, correspondentes às fases iniciais da doença: *Non Demented*, *Very Mild Demented*, *Mild Demented* e *Moderate Demented*. Numa primeira fase, foram desenvolvidos modelos *Machine Learning* (ML), utilizando as arquiteturas *Visual Geometry Group* de 16 camadas (VGG-16), *Visual Geometry Group* de 19 camadas (VGG-19) e *Residual Network* de 50 camadas (ResNet-50), todas elas *Convolutional Neural Network* (CNN) bastante reconhecidas. Os resultados demonstram que estes modelos apresentam bons desempenhos quando treinados e testados, obtendo valores acima de 86% para as métricas de *precision*, *recall* e *f1-score* e acima de 75% para as métricas de *Cohen's Kappa* e *Chance Adjusted Accuracy* (CAA).

Numa fase seguinte, procedeu-se à implementação de *Active Machine Learning* (AL), de modo a tentar mitigar as limitações do *dataset* e melhorar a performance do modelo, utilizando também para esse efeito a arquitetura ResNet-50, por ser a que apresentou o melhor desempenho na fase anterior. Os resultados deste modelo mostram que esta técnica não foi tão bem sucedida, obtendo valores de próximos de 68% para as métricas de *precision*, *recall* e *f1-score*, 48% para *Cohen´s Kappa* e 58% para CAA.

Numa fase final, após a avaliação do desempenho dos modelos propostos, e de modo a aumentar a confiança nas suas precisões, foi aplicada uma técnica de *Explainable AI* (XAI), o *Local Interpretable Model-agnostic Explanations* (LIME), que permite perceber quais foram as zonas do cérebro que o modelo considerou para a sua tomada de decisão.

Estes resultados contribuem para o avanço no controlo e intervenção na saúde, direcionando

futuros esforços para uma abordagem mais personalizada e eficaz. Com o projeto atual, que testa a colaboração entre humanos e a IA, pode ser possível alargar novos caminhos para o tratamento de dados e fazer previsões sobre a adesão a tratamentos/intervenções em doenças neurodegenerativas.

Palavras-chave – *Alzheimer*, *Explainable AI*, *Active Machine Learning*, Redes Neurais Convolucionais, VGG-16, VGG-19, ResNet-50, Classificação de Imagens, Deteção Automática, Diagnóstico Médico, *F1-Score*, *Precision*, *Recall*, *Cohen's Kappa*, *Chance Adjusted Accuracy*.

Abstract

Alzheimer's disease is a global health concern, associated with neurodegeneration of the human brain and, consequently, loss of memory and ability to communicate. Since this disease has no cure to date, it is essential that it is detected early so that strategies can be devised to slow down its progress, such as cognitive stimulation.

The aim of this study was to include Artificial Intelligence in the process of diagnosing *Alzheimer's* disease, as a tool to help detect small signs of this disease that may go unnoticed by the human eye, through MRI images of the brain.

For this purpose, a *dataset* taken from the *Kaggle* platform was used, consisting of 6400 Ressonância Magnética (MRI) images of the brains of various patients, grouped into four classes, corresponding to the early stages of the disease: *Non Demented*, *Very Mild Demented*, *Mild Demented* and *Moderate Demented*. In a first phase, *Machine Learning* (ML) models were developed using the *Visual Geometry Group* de 16 camadas (VGG-16), *Visual Geometry Group* de 19 camadas (VGG-19) and *Residual Network* de 50 camadas (ResNet-50) architectures, all of which are well-recognized *Convolutional Neural Network* (CNN). The results show that these models perform well when trained and tested, obtaining values above 86% for the *precision*, *recall* and *f1-score* metrics and above 75% for the *Cohen's Kappa* and *Chance Adjusted Accuracy* (CAA) metrics.

In the next phase, we implemented *Active Machine Learning* (AL) in order to try to mitigate the limitations of *dataset* and improve the model's performance, also using the ResNet-50 architecture for this purpose, as it was the one that showed the best performance in the previous phase. The results of this model show that this technique was not so successful, obtaining values close to 68% for the *precision*, *recall* and *f1-score* metrics, 48% for *Cohen's Kappa* and 58% for CAA.

In a final phase, after evaluating the performance of the proposed models, and in order to increase confidence in their accuracy, a *Explainable AI* (XAI) technique was applied, the *Local Interpretable Model-agnostic Explanations* (LIME), which allows us to see which areas of the brain the model considered for its decision-making.

These results contribute to advances in health monitoring and intervention, directing future efforts towards a more personalized and effective approach. With the current project, which tests the collaboration between humans and Inteligência Artificial (IA), it may be pos-

sible to broaden new avenues for processing data and making predictions about adherence to treatments/interventions in neurodegenerative diseases.

Keywords – *Alzheimer*, Explainable AI, Active Machine Learning, Convolutional Neural Network, VGG-16, VGG-19, ResNet-50, Image Classification, Automatic Detection, Medical Diagnosis, *F1-Score*, *Precision*, *Recall*, *Cohen's Kappa*, *Chance Adjusted Accuracy*.

Lista de Acrónimos

AL *Active Machine Learning*. ix, xi, xv, xvi, xvii, xxi, 2, 3, 10, 11, 23, 31, 52, 53, 55, 56, 57, 58, 61, 62

CAA *Chance Adjusted Accuracy*. ix, xi, 21, 37, 40, 42, 44, 50, 55, 58, 61
CISUC *Centre for Informatics and Systems of the University of Coimbra*. iii
CNN *Convolutional Neural Network*. ix, xi, 3, 11, 27

IA Inteligência Artificial. ix, x, xi, xx, 1, 2, 10, 15, 16, 18, 25, 26, 29, 31, 58, 62

LIME *Local Interpretable Model-agnostic Explanations*. ix, xi, xvi, 3, 17, 27, 28, 30, 39, 40, 42, 43, 47, 51, 52, 56, 57, 61, 62

ML *Machine Learning*. ix, xi, 2, 10, 17, 21, 22, 25, 27, 30, 34, 35, 52, 53, 58
MRI Ressonância Magnética. ix, xi, 2, 8, 9, 26, 27, 28, 29, 31, 32, 58, 61

NextGenAI *Center For Responsible AI*. iii

PET Tomografia por Emissão de Positrões. 8

ResNet *Residual Network*. xx, 14
ResNet-50 *Residual Network* de 50 camadas. ix, x, xi, xii, xv, xvi, xvii, xx, xxi, 3, 14, 16, 23, 27, 35, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 57, 61

SHAP *SHapley Additive exPlanations*. 17

TC Tomografia Computarizada. 8, 26

VGG *Visual Geometry Group*. xix, 11, 14, 43
VGG-19 *Visual Geometry Group* de 19 camadas. ix, x, xi, xii, xv, xvi, xvii, xx, xxi, 3, 12, 13, 14, 23, 35, 39, 40, 41, 42, 43, 44, 57, 61
VGG-16 *Visual Geometry Group* de 16 camadas. ix, x, xi, xii, xv, xvi, xvii, xix, xxi, 2, 12, 13, 23, 27, 35, 36, 37, 38, 39, 40, 44, 57, 61

XAI *Explainable AI*. ix, xi, xv, xx, 3, 15, 16, 17, 18, 26, 27, 28, 29, 31, 61

Lista de Figuras

Lista de Tabelas

Conteúdo

1

Introdução

O presente capítulo apresenta uma descrição do contexto e motivação do projeto desenvolvido (Secção 1.1), bem como os seus objetivos e metodologias (Secção 1.2). Além disso, são também apresentadas as contribuições científicas deste estudo (Secção 1.3) e a estrutura do documento (Secção 1.4).

1.1 Contexto do Projeto e Motivação

A presente dissertação foi desenvolvida no âmbito da conclusão do mestrado em Engenharia Biomédica, Ramo de Informática Clínica e Bioinformática, lecionado na Universidade de Coimbra.

A doença de *Alzheimer* é o tipo mais comum de demência, tratando-se de uma doença neurodegenerativa caracterizada pelo declínio cognitivo e perda de memória progressiva, que podem afetar as tarefas diárias e a capacidade de comunicação [1]. Não existe um diagnóstico precoce desta doença, no entanto, um diagnóstico atempado e exato da doença de *Alzheimer* é crucial para uma gestão eficaz da sua progressão, uma vez que permite intervenções atempadas e estratégias para abrandar o declínio cognitivo. Atualmente, não existe cura para esta doença, o que torna o seu diagnóstico precoce e exato ainda mais importante.

Nos últimos anos, a Inteligência Artificial (IA) [1] surgiu como uma ferramenta poderosa nos cuidados de saúde, incluindo no domínio do diagnóstico da doença de *Alzheimer*. Os algoritmos de IA podem analisar grandes quantidades de dados, como registos médicos, exames de imagiologia cerebral e informação genética, com uma rapidez e exatidão notáveis. Esta capacidade permite à IA detetar padrões e marcadores subtis associados à doença que podem passar despercebidos apenas pela observação humana, contribuindo assim para um diagnóstico atempado [2]. No entanto, apesar de estarem a ser testadas várias técnicas de IA para a doença de *Alzheimer*, ainda não existe nenhuma solução ativa nas clínica.

Com os avanços significativos na capacidade de análise e precisão da IA, nomeadamente

[1] Neste trabalho optou-se por utilizar certos termos em inglês, como *Active Machine Learning*, *Machine Learning* e *Explainable AI*, devido à falta de traduções amplamente reconhecidas no português técnico. Estes termos são amplamente utilizados na literatura científica e tecnológica, facilitando a compreensão e alinhamento com as fontes de referência. Por outro lado, o termo *Inteligência Artificial* foi traduzido para português, uma vez que este possui uma tradução bem estabelecida e compreendida na comunidade académica e profissional portuguesa.

no que diz respeito à área da saúde, é natural que se explorem formas de maximizar o seu potencial. Neste sentido, surge o conceito de IA colaborativa (*Colaborative-AI*), que é uma forma de interação entre os sistemas de IA e a inteligência humana, combinando os seus pontos fortes para obter um desempenho superior. A IA colaborativa reconhece que, embora a IA seja excelente no processamento de dados, na realização de cálculos e na identificação de padrões, podem faltar-lhe certas qualidades humanas, como o raciocínio de senso comum, a criatividade e o juízo ético. Por outro lado, os seres humanos possuem conhecimentos contextuais, intuição e inteligência social, mas podem ser limitados na sua capacidade de lidar com grandes conjuntos de dados. Ao integrar a IA e os agentes humanos num quadro de colaboração, o objetivo é obter resultados mais precisos, robustos e eticamente alinhados [3]. Um exemplo de IA colaborativa é o *Human-in-the-loop*, em que o contributo e a supervisão humanos são integrados no sistema, especialmente no processo de tomada de decisões, de modo a garantir a fiabilidade, a ética e o alinhamento com as expectativas humanas.

A escassez de dados e a sua dificuldade de etiquetagem pode constituir um desafio para a criação de *datasets* com dados suficientes para treinar com sucesso modelos de IA, especialmente na área da medicina [4]. Neste contexto, torna-se relevante a utilização de *Active Machine Learning* (AL) [5], um tipo *Machine Learning* (ML) mais apropriado para conjuntos de dados em que as etiquetas são mais difíceis e que se pode enquadrar no paradigma *Human-in-the-loop* e permite abordar o desafio da escassez de dados e da obtenção de anotações, permitindo que os escolham os caminhos de aprendizagem, neste caso, selecionando instâncias que, depois de anotadas, se espera que sejam as mais informativas, permitindo assim otimizar o modelo preditivo. Estes algoritmos iniciam-se com um pequeno conjunto de dados e rotulados, sendo iterativamente anotadas as amostras mais informativas, para serem integradas para atualizar o modelo. Desta forma, é possível reduzir a quantidade de dados necessária para treinar um modelo, tornando assim o processo de aprendizagem mais eficiente e económico.

Apesar de muitos modelos de ML terem mostrado uma elevada *accuracy*, estes geralmente apresentam uma falta de transparência e explicabilidade, elementos cruciais para a confiança nos sistemas IA, uma vez que os modelos podem não fornecer explicações claras sobre os seus resultados e o processo de raciocínio. Esta questão é particularmente importante tendo em conta a complexidade dos modelos e a forma como as decisões tomadas podem afetar a vida dos utilizadores, especialmente em contextos médicos em que é necessário decidir que procedimentos adotar para cuidar de um paciente [6, 7, 8, 9].

1.2 Objetivos e Metodologia

O presente estudo tem como principal objetivo detetar a doença de *Alzheimer*, nas suas diversas fases, através de imagens por Ressonância Magnética (MRI) do cérebro de pacientes, de modo a facitar o diagnóstico atempado desta patologia.

Para atingir este objetivo, foram desenvolvidos para este efeito, numa fase inicial, três modelos ML, baseados em diferentes arquiteturas: *Visual Geometry Group* de 16 camadas (VGG-16),

Visual Geometry Group de 19 camadas (VGG-19) e *Residual Network* de 50 camadas (ResNet-50), todas elas bastante reconhecidas no âmbito do reconhecimento de imagens. Para treinar estes modelos recorreu-se a um *dataset* retirado da plataforma *Kaggle*, constituído por 6400 imagens de quatro classes diferentes, correspondentes a quatro fases iniciais da doença, que foi dividido em sub-conjuntos para treino, teste e validação.

Posteriormente, uma vez que o número de amostras em cada classe não se encontra igualmente distribuído, recorreram-se a técnicas de *Data Augmentation* para equilibrar o conjunto de dados. Estas técnicas foram aplicadas ao conjunto de treino, tendo este sido utilizado para treinar um novo modelo, com a arquitetura ResNet-50.

Após se verificar que este modelo apresentava resultados inferiores aos anteriores, foi então construído um modelo de AL, de modo a tentar melhorar os resultados obtidos e a poupar recursos no processo de treino.

Numa fase final, foi aplicada a técnica de *Local Interpretable Model-agnostic Explanations* (LIME) a todos os modelos desenvolvidos, com o objetivos de aumentar a explicabilidade dos mesmos, permitindo assim aos utilizadores perceber como as decisões foram tomadas e consequentemente, o aumento na confiança nestes.

1.3 Contribuições Científicas

As contribuições científicas fundamentais originadas neste estudo compreendem:

- Vários modelos de IA explicáveis, capazes de identificar a doença de *Alzheimer* através de imagens MRI do cérebro, nas suas diferentes fases;
- Um modelo de AL capaz de identificar a doença, nas suas diversas fases, através de imagens MRI do cérebro;
- A singularidade do modelo proposto, uma vez que apesar de existirem vários modelos propostos para a deteção da doença de *Alzheimer*, nenhum deles combina técnicas de AL com *Explainable AI* (XAI);
- O potencial prático deste estudo, que demonstra resultados satisfatórios e mais confiáveis, devido ao uso de XAI que permite perceber como as decisões do modelo foram tomadas.

1.4 Organização do Documento

A restante parte do presente documento encontra-se estruturada da seguinte forma: o Capítulo 2 fornece uma visão geral sobre os conteúdos base necessários para a compreensão do estudo descrito nesta tese; no Capítulo 3 são apresentados os principais trabalhos relacionados com o presente trabalho, particularmente no que se refere à deteção de *Alzheimer* e ao uso de *Convolutional Neural Network* (CNN) e de XAI; o Capítulo 4 retrata o *dataset* utilizado neste estudo; o Capítulo 5 tem como principal objetivo a descrição das diferentes etapas deste trabalho, bem como a exposição e discussão dos resultados obtidos; por fim, no Capítulo 6, são

apresentadas algumas conclusões e limitação a este trabalho, bem como possíveis orientações para trabalhos futuros.

2

Fundamentos Teóricos

2.1 Doença de *Alzheimer*

A doença de *Alzheimer* é a forma mais comum de demência. Trata-se de uma doença neurodegenerativa progressiva que causa uma perda de memória progressiva, o que afeta a independência e o quotidiano dos indivíduos por ela afetados. Em 2020, estima-se que existiam no mundo cerca 50 milhões de indivíduos com a doença de Alzheimer, sendo que se previa a duplicação deste número a cada 5 anos, atingindo 152 milhões em 2050 [10].

2.1.1 Causas

Os neurónios (ou células nervosas) são a unidade funcional do sistema nervoso. Estas células são responsáveis por transmitir informação por todo o corpo humano, através de sinais elétricos e químicos, ajudando assim a coordenar todas as funções necessárias à vida. Para isso, existem três grandes categorias de neurónios: Sensoriais (que transportam informações dos órgãos dos órgãos como os olhos e o nariz para o cérebro), motores (que controlam a atividade muscular voluntária, como andar e falar, e transportam mensagens do cérebro para os músculos) e interneurónios (todos os neurónios que não são sensoriais ou motores) [11].

A doença de *Alzheimer* é caracterizada pela acumulação de proteínas anormais no cérebro (Amiloide-β e *tau*), o que perturba o funcionamento das células cerebrais, levando à sua morte. A proteína *Amiloide-*β, agrega-se formando placas que se depositam em diferentes regiões do cérebro. Estas placas são reconhecidas como um material estranho ao cérebro, desencadeando uma resposta inflamatória imunitária que conduz à morte celular e neurodegeneração. No que diz respeito à proteína *tau*, esta acumula-se dentro dos neurónios, formando emaranhados que os destabilizam, nomeadamente no transporte pelo axónio, levando à neurodegeneração [12, 13].

Estima-se que existam cerca de 86 bilhões de neurónios no cérebro, sendo que, com a doença de *Alzheimer*, estes vão sendo destruídos (morte celular), afetando assim, ao longo do tempo, a capacidade de pensar e a memória. Inicialmente, esta doença afeta o hipocampo (parte do cérebro essencial para formar memórias), sendo que, à medida que os neurónios vão morrendo, mais partes do cérebro vão sendo afetadas, levando assim a um encolhimento do córtex cerebral e à dilatação dos ventrículos do cérebro, tal como se pode verificar nas Figuras 2.2 e 2.1 [14].

Existem vários fatores que podem tornar um indivíduo mais suscetível a algum tipo de

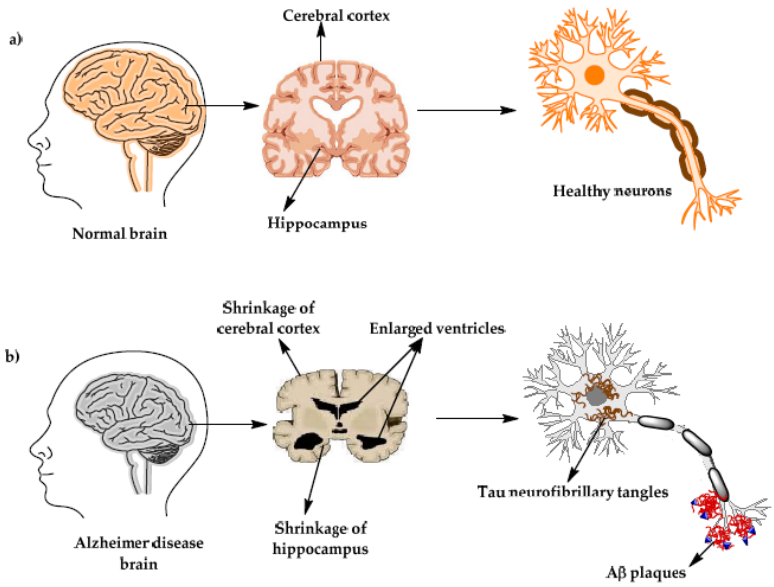


Figure 2.1: Comparação entre um cérebro normal e um cérebro com doença de *Alzheimer*: (a) Cérebro normal- O córtex cerebral e o hipocampo estão intactos e os neurônios são saudáveis; (b) Cérebro com doença de *Alzheimer*- Há atrofia do córtex cerebral, encolhimento do hipocampo e dilatação dos ventrículos. Observa-se também a presença de emaranhados neurofibrilares de tau e placas de beta-amiloide (A) (Imagem retirada de: [10]).

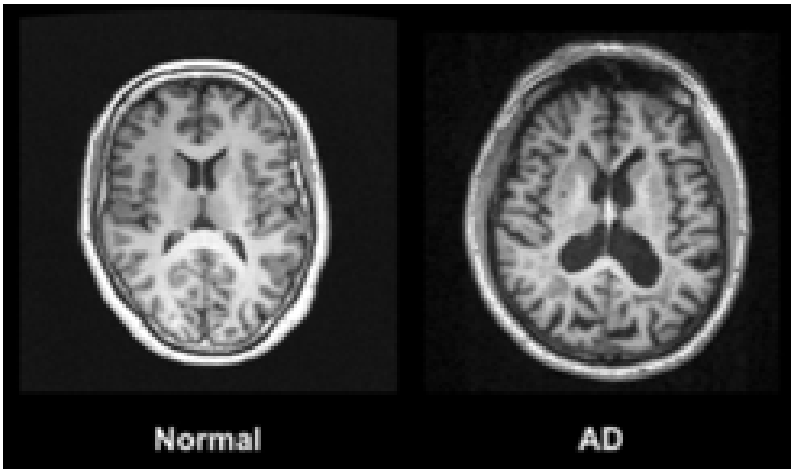


Figure 2.2: Imagens de ressonância magnética de um cérebro normal versus um cérebro com doença de Alzheimer: À esquerda-Cérebro normal mostrando ventrículos e córtex cerebral em condições normais; À direita: Cérebro com doença de Alzheimer (AD) mostrando atrofia cortical significativa e dilatação dos ventrículos(Imagem adptada de:[15]).

demência, incluindo *Alzheimer*. Entre estes fatores, destacam-se a idade e a genética. Para além destes, também fatores como a hipertensão, a obesidade, o tabagismo, a depressão, a diabetes, a falta de atividade física, uma menor educação e o uso excessivo de álcool podem aumentar o risco da doença. Estudos demonstram também a importância da estimulação cognitiva, uma vez que, frequentemente, os indivíduos apresentam um declínio cognitivo mais acentuado em situações de menor estimulação, como, por exemplo, ao se aposentarem [16].

2.1.2 Fases do *Alzheimer*

A doença de *Alzheimer* apresenta três fases principais: Ligeira (*early stage* ou *mild*), moderada (*middle-stage* ou *moderate*) e avançada/severa (*late stage*,*severe stage* ou *advanced stage*) [17].

2.1.2.1 *Early Stage*

Na fase inicial de *Alzheimer*, o indivíduo pode viver de forma independente. No entanto, começa a ter algumas dificuldades como, por exemplo, esquecimento, dificuldade de lembrar-se de algumas palavras, dificuldade de concentração ou de aprender coisas novas, dificuldade acrescida de planear ou organizar, etc. Estes sintomas podem não ser fáceis de detetar, uma vez que, nesta fase, o indivíduo mantém a grande maioria das suas habilidades e pode não necessitar de qualquer auxílio. No entanto, nesta fase os indivíduos podem começar a sentir dificuldades em completar tarefas que costumam realizar sem qualquer problema [17].

2.1.2.2 *Middle Stage*

A fase moderada da doença de *Alzheimer* é, geralmente, a mais longa, podendo durar vários anos. Ao longo desta fase, o doente começa a mostrar mais sintomas de demência e a necessitar de mais cuidados, à medida que as suas habilidades cognitivas vão continuando a deteriorar. Alguns dos principais sintomas desta fase são, por exemplo, não conseguir dizer qual o dia em que está, ser incapaz de se lembrar de algumas informações de si próprio (como número de telemóvel, morada, escola onde andou, etc.), entre outros [17].

2.1.2.3 *Late Stage*

Nesta fase da doença, o paciente com *Alzheimer* demonstra sinais avançados de demência, perdendo a capacidade de cuidar de si próprio e de se comunicar. Alguns do sintomas comuns desta fase passam por uma grave perturbação de memória, de processamento de novas informações e de reconhecimento de tempo ou local, perda da capacidade de reconhecer a fala, assim como uma perda de capacidade de comer, andar e ir à casa de banho sem ajuda [17].

2.1.3 Diagnóstico e Tratamento

Tipicamente, o processo de diagnóstico da doença de *Alzheimer* passa por diversas fases, que se completam entre si. Destaca-se a importância das primeiras fases, pois é nestas que é feita a deteção dos sinais iniciais da doença (tais como perda de memória ou sintomas de depressão),

o que permite uma melhor gestão da doença e do tratamento. Além disso, é também feita uma avaliação que tem como objetivo detetar as possíveis causas destes sintomas e compreender os fatores de risco do paciente (tais como a idade, casos confirmados em familiares diretos, etc.), de modo a garantir que os sintomas efetivamente estão relacionados com esta patologia, e não advêm de outras causas como, por exemplo efeitos colaterais de medicamentos [13].

Na Figura 2.3 encontra-se esquematizado o processo de diagnóstico desta patologia.

Atualmente, ainda não há uma cura para o *Alzheimer*, pelo que é importante que o diagnóstico seja feito o mais cedo possível, de modo a permitir uma melhor gestão da doença. Existem alguns medicamentos que podem ser utilizados para amenizar os sintomas da doença, mas não para parar a sua progressão [18]. Além disso, devem ser aplicadas terapias não farmacológicas (como dietas, exercício e atividades de estimulação cognitiva), de forma a tentar manter, ou melhorar, a função cognitiva e a capacidade de realizar as atividades do quotidiano[13].

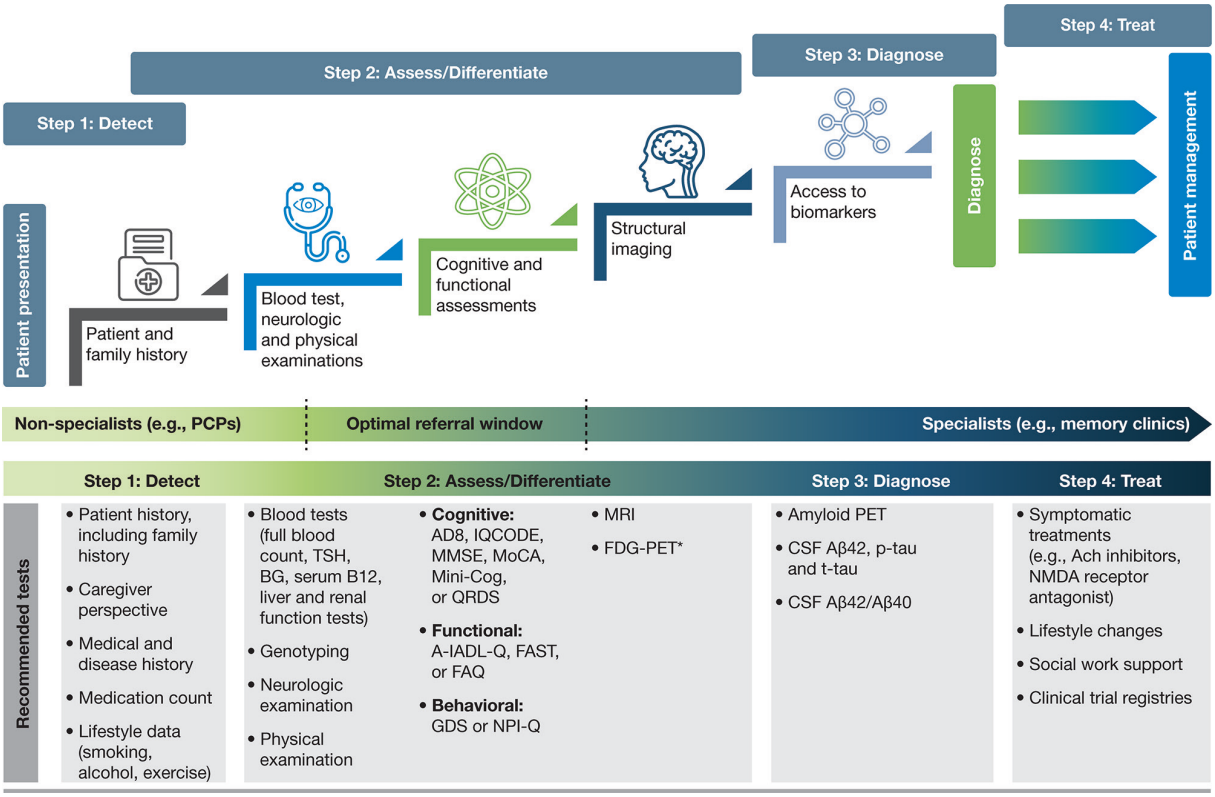


Figure 2.3: Diagnóstico de Alzheimer (Imagem retirada de: [13]).

2.2 Técnicas de Imagiologia

Ao longo dos anos, diversas técnicas de imagiologia têm sido estudadas para o diagnóstico da doença de *Alzheimer*. Na década de 1970, a Tomografia Computarizada (TC) foi pioneira neste meio, fornecendo imagens detalhadas do cérebro em corte transversal. No entanto, ao longo do tempo esta técnica foi sendo substituída, no contexto da doença de *Alzheimer*, por técnicas mais recentes, tais como a Ressonância Magnética (MRI) ou a Tomografia por Emissão de

Positrões (PET), que oferecem uma maior sensibilidade e especificidade na deteção de alterações relacionadas com esta doença [19].

Para a realização do presente trabalho, foram utilizadas imagens obtidas através de MRI. Neste sentido, esta secção irá abordar, de forma superficial, alguns dos conceitos básicos da ressonância magnética.

A MRI é uma técnica de imagem médica, não invasiva, que utiliza campos magnéticos e ondas de rádio para produzir imagens detalhadas do interior do corpo humano. Esta técnica é bastante utilizada para diagnóstico, uma vez que permite a visualização de estruturas internas do corpo, com grande precisão. Desta forma, a MRI torna-se particularmente útil para deteção de lesões cerebrais, tumores, doenças cardíacas e lesões musculoesqueléticas [20].

Esta técnica de imagem utiliza o fenómeno de ressonância magnética, que se baseia na interação entre um campo magnético aplicado e um núcleo que possui *spin*. O *spin* nuclear, mais precisamente denominado de momento angular do *spin* nuclear, é uma propriedade intrínseca de um átomo. O núcleo pode ser considerado como estando constantemente a rodar em torno de um eixo, a uma taxa ou velocidade constante, sendo que este eixo de auto-rotação é perpendicular à direção de rotação. Esta rotação do núcleo, sob a influência do campo magnético, é a base do fenómeno de ressonância magnética, oferecendo conhecimentos valiosos em aplicações de diagnóstico e de pesquisa médica [21].

Para a obtenção da imagem MRI, o paciente é colocado dentro do equipamento, onde é lhe aplicado um forte campo magnético externo (em redor da região de interesse). O campo magnético aplicado cria uma interação com os átomos de hidrogénio do corpo do paciente, levando-os a alinharem-se temporariamente, seja paralelamente ou antiparalelamente, com o campo magnético. De seguida, é aplicado um pulso de radiofrequências, de curta duração, com o objetivo de alterar a orientação dos *spins* dos átomos de hidrogénio presentes no corpo do paciente, resultando numa absorção de energia por parte dos átomos de hidrogénio. Quando o pulso de radiofrequências é interrompido, os *spins* dos átomos começam a regressar ao seu estado original, libertando energia sob a forma de sinais radiofrequência. Estes sinais são recebidos pelo equipamento MRI, sendo posteriormente processados para obter a imagem final [22].

No contexto da doença de *Alzheimer*, a MRI apresenta-se particularmente útil, pois permite detetar alterações estruturais no cérebro, características da doença, como a atrofia cerebral. Além disso, a sua capacidade de fornecer imagens detalhadas e precisas torna-a uma ferramenta valiosa no acompanhamento da progressão da doença ao longo do tempo, o que é fundamental para entender a evolução da doença e avaliar a eficácia das intervenções terapêuticas utilizadas. No entanto, é importante ressaltar que a imagem por MRI, por si só, não é capaz de oferecer um diagnóstico definitivo, pelo que deve ser utilizada em conjunto com outras informações clínicas, como o historial médico e um exame neuropsicológico.

2.3 *Active Machine Learning*

Machine Learning (ML) é um campo da Inteligência Artificial (IA), que se foca no desenvolvimento de algoritmos e modelos que permitam que os sistemas aprendam com os dados e utilizem esses conhecimentos para tomar decisões ou efetuar previsões [23].

Active Machine Learning (AL) é uma área do ML que permite que o algoritmo escolha os dados com que deve aprender, de modo a melhorar a sua *accuracy*. Desta forma, o modelo não precisa de ser alimentado com grandes quantidades de dados rotulados (anotados), pois permite que o algoritmo selecione os dados mais úteis para a sua aprendizagem. Assim sendo, o AL visa maximizar o desempenho do modelo, com o menor número de dados de treino rotulados possível, o que é bastante útil quando o processo de classificação de dados é demasiado caro ou demorado, uma vez que permite que o modelo seja treinado com eficiência, concentrando-se apenas nos dados mais pertinentes[5].

Ns sistemas de AL podem ser distinguidas duas componentes principais: oráculo e classificador. O oráculo refere-se à entidade responsável por fornecer etiquetas (*labels*) às amostras que lhe são apresentadas pelo algoritmo de AL. O classificador trata-se do modelo de aprendizagem automática que está a ser treinado iterativamente com os dados rotulados, obtidos através da aprendizagem ativa (AL). Além destas componentes, existe também o *dataset* utilizado, que pode ou não conter algumas anotações (rótulos) e uma *query strategy*, que se trata de um algoritmo para decidir quais as amostras devem ser encaminhadas para o oráculo. O desempenho do classificador melhora à medida que recebe etiquetas mais informativas do oráculo, permitindo-lhe fazer melhores previsões sobre dados não vistos.

A interação entre o oráculo e o classificador é crucial no processo de aprendizagem ativa, uma vez que o algoritmo visa selecionar as instâncias mais informativas para a rotulagem, de modo a melhorar o desempenho do classificador (maximizando a aprendizagem) e a minimizar o custo da rotulagem.[5, 24].

Em AL, o processo de treino inicia-se com um pequeno conjunto de dados rotulados. O sistema de aprendizagem ativa seleciona amostras não rotuladas, que envia ao oráculo para rotulação. Os rótulos fornecidos pelo oráculo são então incorporados no conjunto de treino do classificador, que é reajustado com estes novos dados. Após as pontuações da avaliação inicial, este processo repete-se até se atingir uma pontuação aceitável. Na Figura 2.4 encontra-se esquematizado este processo de treino.

Para selecionar as amostras que irão ser rotuladas, existem várias estratégias, tais como:

1. **Incerteza do modelo:** seleciona amostras para as quais o modelo atual tem maior incerteza de previsão. Esta estratégia tem como objetivo reduzir a incerteza, consultando as amostras que estão na fronteira de decisão ou onde as previsões do modelo são inconsistentes.
2. **Diversidade de amostras:** seleciona amostras que representam diferentes regiões do espaço de características, de forma a garantir que o modelo é exposto a uma variedade de

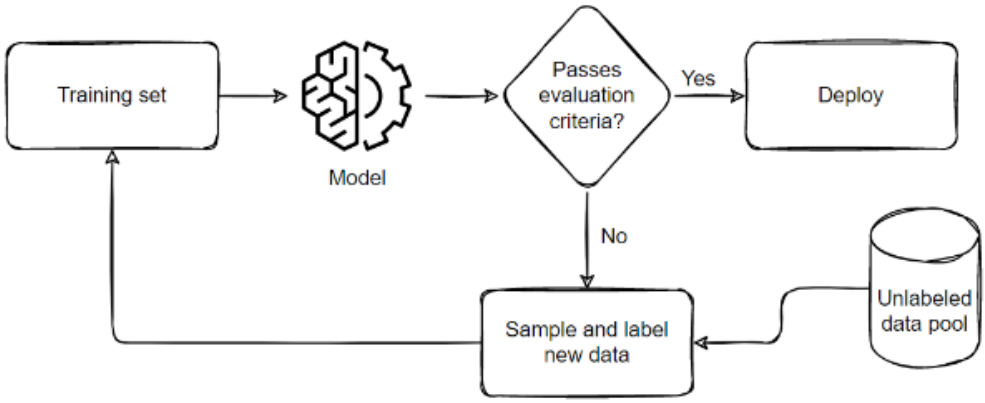


Figure 2.4: Esquema representativo do processo de treino em AL (Imagem retirada de: [25]).

exemplos durante o treino, o que melhora a sua generalização e robustez.

3. **Mudança esperada do modelo:** seleciona amostras que se espera que causem uma maior mudança no modelo quando anotadas. O objetivo é priorizar as instâncias que provavelmente terão um impacto significativo na melhoria do desempenho do modelo.

Uma métrica bastante utilizada por estas estratégias é a margem de classificação, que seleciona amostras próximas da fronteira de decisão do modelo, onde a diferença entre as probabilidades das classes é maior, ou seja, onde são mais difíceis de classificar com segurança. Esta abordagem torna-se praticamente útil, uma vez que permite que o modelo aprenda com exemplos mais desafiadores e que podem causar uma maior mudança no modelo quando rotulados e incorporados no conjunto de treino. Desta forma, o modelo consegue aumentar a sua capacidade de generalização, o que se traduz numa melhoria mais eficaz do desempenho, com um menor número de amostras rotuladas [5].

De modo a otimizar o desempenho dos modelos, é também importante considerar o ajuste adequado dos hiperparâmetros do modelo. Os hiperparâmetros são parâmetros externos ao modelo, como por exemplo a taxa de aprendizagem) que afetam a sua configuração e seu comportamento ao longo do processo de treino. Estes parâmetros são definidos antes do treino e têm impacto significativo na generalização do modelo para novas amostras, bem como na prevenção de *overfitting* ou *underfitting*, pelo que a sua escolha adequada se torna particularmente importante em cenários onde a rotulagem de dados é custosa ou demorada, uma vez que permite que os recursos sejam alocados de forma mais inteligente, concentrando-se nos aspetos mais relevantes do modelo.

2.4 Modelo *Visual Geometry Group* (VGG)

O modelo VGG trata-ta se uma *Convolutional Neural Network* (CNN) desenvolvida, em 2014, pelo *Visual Geometry Group* da Universidade de *Oxford*. Esta arquitetura demonstrou-se revolucionária, servindo de base para vários modelos inovadores de reconhecimento de objetos. É uma das arquiteturas de reconhecimento de imagem mais populares, tendo variantes como

Visual Geometry Group de 16 camadas (VGG-16) ou *Visual Geometry Group* de 19 camadas (VGG-19) [26].

2.4.1 Arquitetura do modelo VGG-16

Esta arquitetura foi proposta por Simonyan e Zisserman [27] no âmbito do concurso ImageNet (ILSVR) em 2014, e tornou-se um dos primeiros modelos a obter um bom desempenho no âmbito do processamento de imagens.

Este algoritmo de deteção e classificação de imagens foi construído considerando como entrada imagens RGB de 224x244 (neste trabalho, o tamanho das imagens de entrada foi de 128x128), sendo capaz de classificar 1000 imagens de 1000 categorias diferentes com uma *accuracy* de 92,7%.

O modelo VGG-16 é composto por 21 camadas, sendo que apenas 16 destas têm pesos. É constituído por 13 camadas de convoloção (responsáveis por extrair características das imagens de entrada), 5 camadas de *max-pooling* (responsáveis para reduzir o tamanho das características extraídas) e 3 camadas totalmente conectadas (responsáveis pela classificação final da rede, utilizando os vetores de características obtidos pelas camadas anteriores para atribuir as imagens a uma das categorias de classificação) [28, 26]. Esta arquitetura utiliza camadas de convolução com filtros de tamanho contante de 3x3 e *stride* 1, bem como camadas de *pooling* com filtro *max pooling* 2x2 e *stride* 2, sendo que estas camadas estão dispostas de forma consistente em toda a arquitetura [29, 30]. Na Figura 2.5 encontra-se a representação gráfica desta arquitetura, sendo que cada bloco representa uma camada específica, contendo informações acerca do tipo de camada, tamanho do filtro, número de filtros e função de ativação utlizada.

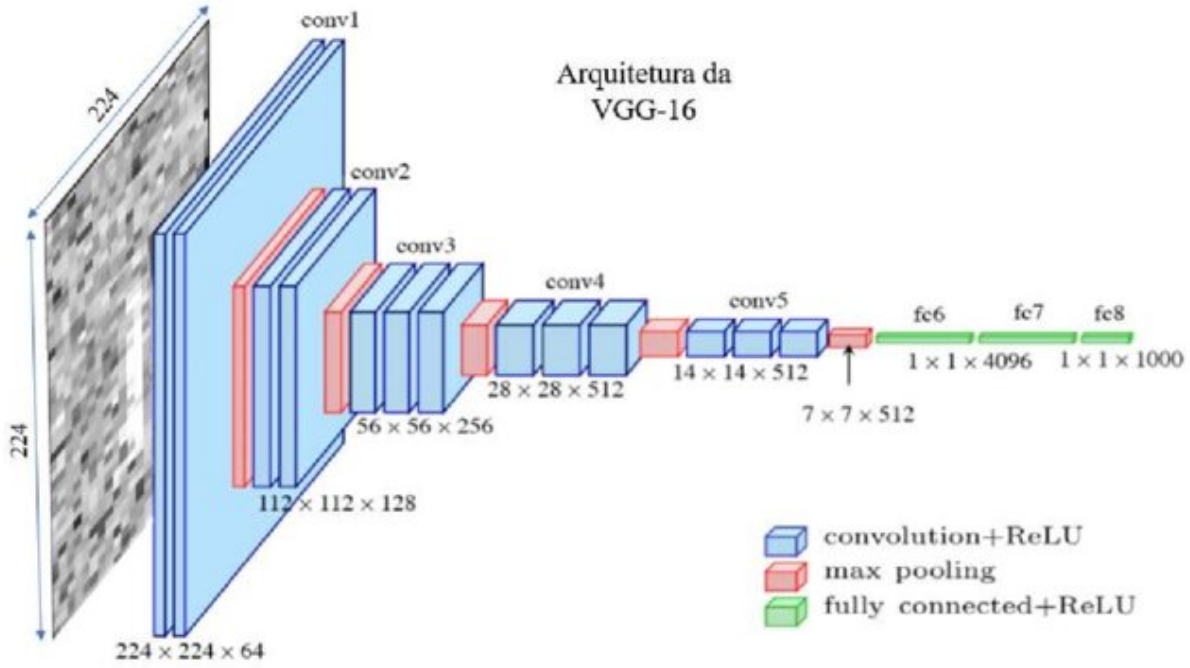


Figure 2.5: Diagrama da arquitetura VGG-16 (Imagem retirada de: [28]).

De forma mais detalhada, a estrutura desta rede é:

- **Camadas de convolução (*conv*):** A rede inicia-se com duas camadas de convolução (*conv1* e *conv2*). Nesta arquitetura, cada camada de convolução é seguida por uma função ativação ReLU (*Rectified Linear Unit)*), para introduzir não linearidade. Tal como demonstrado na Figura 2.5, o número de filtros em cada uma destas camadas é: *conv1*-64,*conv2*-128,*conv3*-256 e *conv4* e *conv5*-512.

- **Camadas de *pooling*:** Após cada camada de convolução, existe uma camada de *pooling*, com o objetivo de reduzir a dimensão espacial, tornando assim a rede mais eficiente a nível computacional e ajudando a prevenir o *overfitting*. Nesta arquitetura, é utilizada a técnica de *max-pooling*.

- **Camadas totalmente conectadas:** Após as camadas convolucionais, o VGG-16 possui 3 camadas totalmente conectadas (*fc6, fc7 e fc8*), seguidas por uma função de ativação ReLU. Cada neurónio em cada uma destas camadas está conectado a todos os neurónios da anterior, permitindo assim aprender representações mais complexas.

- **Camada de saída:** Geralmente, a última camada da arquitetura é uma camada *softmax*, que fornece a distribuição de probabilidade sobre as classes de saída, pelo que é utilizada para classificação [29, 30].

Embora o VGG-16 ser uma arquitetura poderosa e amplamente utilizada para reconhecimento de imagens, esta apresenta algumas desvantagens. Devido à sua profundidade (composta por 16 camadas), a rede torna-se computacionalmente intensiva. Além disso, em conjuntos de dados mais pequenos, a sua complexidade pode levar a problemas de *overfitting* , pois a rede pode memorizar os dados de treino, ao invés de aprender padrões gerais entre estes. No entanto, apesar destas limitações, o VGG-16 continua a ser uma das arquiteturas mais populares, tendo alcançado desempenhos excecionais em concursos de reconhecimento de imagens.

2.4.2 Arquitetura do modelo VGG-19

A arquitetura VGG-19 é semelhante à VGG-16, apresentando apenas mais três camadas de convolução, o que aumenta a sua profundidade, número de parâmetros e capacidade de aprendizagem de relações mais complexas[29, 31].

De uma forma geral, esta arquitetura é uma extensão da VGG-16 com uma maior profundidade e capacidade de aprendizagem. No entanto, esta apresenta alguns desafios computacionais e de possível *overfitting*. Uma vez que utiliza mais parâmetros, o seu treino é mais lento e exige mais recursos computacionais, além de que se pode tornar mais propensa ao *overfitting*, especialmente para conjuntos de dados menores. Apesar disto, a sua profundidade adicional pode permitir que a rede aprenda representações mais abstratas e complexas das imagens, o que pode ser benéfico em tarefas de reconhecimento de objetos.

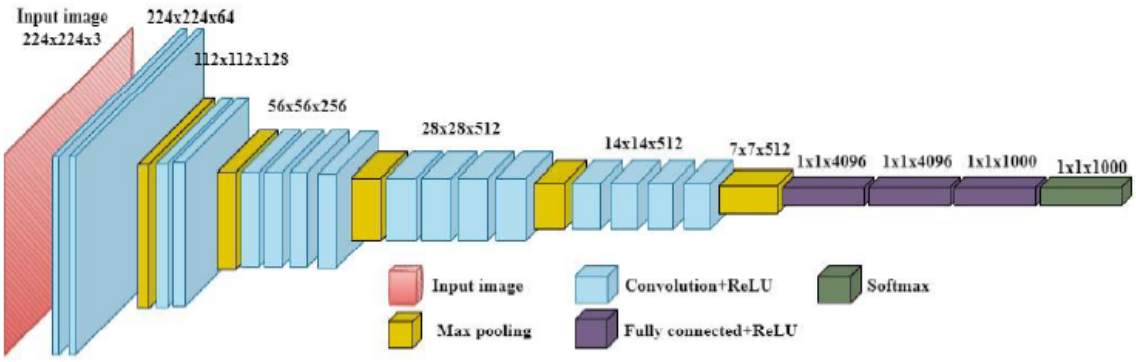


Figure 2.6: Diagrama da arquitetura VGG-19 (Imagem retirada de: [31]).

2.5 Modelo *Residual Network* (ResNet)

O modelo ResNet é, à semelhança do VGG, uma *Convolutional Neural Network* que revolucionou o campo da visão computacional, desde a sua introdução, em 2015. Desenvolvida pela *Microsoft Research* [32], a ideia central por detrás desta arquitetura é a utilização de conexões residuais, que permitem a construção de redes muito mais profundas, sem sacrificar o seu desempenho. Para tal, o ResNet conta com a introdução de blocos residuais que permitem que a rede aprenda as diferenças entre os dados de entrada e a saída esperada, o que facilita o treino de redes profundas. Existem várias variantes do ResNet, cada uma adaptada para diferentes necessidades e complexidades de tarefas. Um exemplo notável é o *Residual Network* de 50 camadas (ResNet-50), que se destaca como uma escolha popular numa ampla gama de aplicações de reconhecimento de imagens.

2.5.1 Arquitetura do modelo ResNet-50

O ResNet-50 é uma variação da rede que possui 50 camadas: 48 de convolução, 1 *MaxPool* e 1 *Average Pool*. Esta arquitetura é bastante conhecida pela sua eficiência e desempenho em trabalhos de reconhecimento de imagens.

Esta arquitetura é capaz de lidar com o problema do gradiente desaparecido, que ocorre em redes neurais profundas devido à diminuição do gradiente à medida que é propagado para camadas iniciais, o que dificulta o ajuste eficaz dos pesos e a aprendizagem de representações úteis. O ResNet-50 utiliza blocos residuais para aprender a diferença entre a saída desejada e a entrada, facilitando assim o treino de redes mais profundas[33, 34].

Na arquitetura da *ResNet-50*, as camadas encontram-se organizadas em diferentes estágios (*stages*). De forma mais detalhada, as camadas que constituem cada um deles são:

- ***Stage 1:*** A primeira camada desta rede consiste numa convolução 7x7, com filtros de 64 e um *stride* de 2, seguida de uma camada de *max-pooling* para diminuir a resolução da imagem de entrada e extrair características iniciais. Nesta arquitetura, é também bastante comum existir, neste estágio, entre as camadas de convolução e de *max-pooling* , uma camada de *Batch Normalization* (utilizada para normalizar a entrada de cada camada,

o que ajuda a acelerar o treino da rede e a melhorar a sua convergência) e a função de ativação ReLU (para introduzir não linearidades, permitindo assim que a rede aprenda relações mais complexas dos dados). Além disso, antes deste estágio, é bastante comum utilizar-se uma camada de *Zero Padding*, de modo a garantir que a saída da convolução mantenha a mesma altura e largura que a entrada.

- ***Stage 2:*** O segundo estágio nesta arquitetura é geralmente constituído por um bloco de convolução (*conv block*), seguido por um "*ID Block*". No *conv block* encontram-se várias camadas de convolução 3x3, para extraír características complexas da imagem, geralmente seguidas de *batch normalization* e da função de ativação ReLU. Após este bloco, encontra-se o "*ID Block*", um tipo específico de bloco residual, que inclui uma conexão direta (*skip connection*) para adicionar a entrada original à saída do bloco, permitindo assim que a rede aprenda a residual (diferença) entre a entrada e a saída do bloco, o que facilita o treino de redes profundas e ajuda a evitar problemas de degradação do desempenho à medida que a rede se torna mais profunda. A combinação destes dois blocos desempenha um papel crucial na extração de características complexas da imagem e no treino eficaz da rede.
- ***Stage 3:*** Este estágio segue a mesma estrutura do anterior, mas com convoluções 1x1, utilizadas para reduzir a dimensionalidade e o número de parâmetros.
- ***Stage 4:*** Semelhante ao anterior, mas com mais blocos residuais para aprender representações ainda mais complexas da imagem.
- ***Stage 5:*** O último estágio desta arquitetura é semelhante aos anteriores, coma a adição de camadas totalmente conectadas, após a sequência de blocos residuais, para realizar a classificação final da rede. Estas camadas são, geralmente, uma camada de *Average Pooling* (utilizada para reduzir a dimensionalidade dos dados e agregar informações espaciais das características extraídas pelas camadas convolucionais anteriores. Esta camada calcula a média dos valores em cada mapa de características, resultando numa representação compacta e global das características) seguida de uma camada de *Flattening* (responsável por converter a saída da camada de *Average Pooling* num vetor unidimensional, preparando assim os dados para serem inseridos nas camadas totalmente conectadas subsequentes) e por uma camada final, totalmente conectada, que recebe o vetor unidimensional resultante do *Flattening* e realiza a classificação das características extraídas em relação às classes de interesse [33].

Na Figura 2.7 encontra-se um diagrama que sumariza o acima descrito.

2.6 *Explainable AI* (XAI)

Os algoritmos de IA são utilizados para auxiliar o utilizador na tomada de decisões. No entanto, o utilizador não tem qualquer conhecimento sobre como um modelo IA tomou as suas decisões ou realizou as tarefas. Assim, muitos dos modelos são difíceis de compreender e confiar, uma vez que funcionam como "caixas negras". Além disso, estes modelos acarretam também

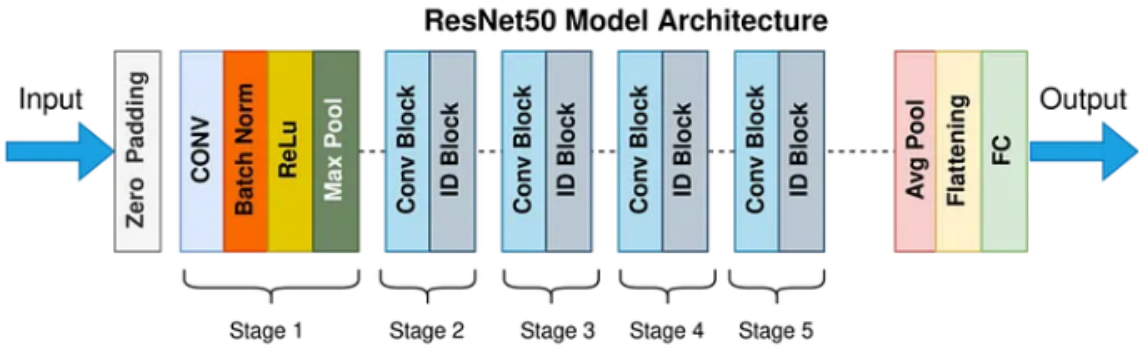


Figure 2.7: Diagrama da arquitetura ResNet-50 (Imagem retirada de: [35]).

várias implicações legais e éticas, pelo que surgiu a necessidade de modelos IA explicáveis (*Explainable AI*-XAI) [36, 37].

A explicabilidade refere-se à compreensão do processo interno de um modelo, fornecendo uma explicação, legível por humanos, dos métodos, procedimentos e resultados do modelo. Na XAI, os dados de treino são fornecidos ao modelo como entrada e, com base nos requisitos ou domínio de aplicação, são selecionadas as metodologias para a previsão e as técnicas de XAI a ser utilizadas para explicar o funcionamento interno do modelo e os seus resultados. Com base neste conhecimento, os utilizadores podem melhorar o modelo e expor falhas que nele detetem, contribuindo assim para que possam tomar decisões acertadas para melhorar o modelo, o que se traduz numa maior confiança neste, uma vez que desta forma é possível garantir a equidade (isto é, que as previsões são imparciais e não discriminam implícita ou explicitamente os grupos sub-representados), privacidade e robustez [38, 7]. Na Figura 2.8 encontra-se esquematizado o processo de XAI.

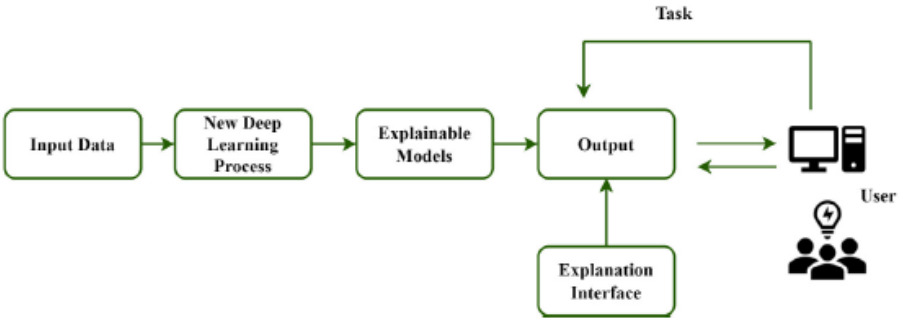


Figure 2.8: Processo de XAI (Imagem retirada de: [38]).

Existem dois tipos de abordagens de XAI: orientada para o conhecimento (*knowledge-driven*) e orientada para os dados (*data-driven*). *Knowledge-driven* trata-se de uma abordagem que incorpora o conhecimento humano (externo ao modelo) para tornar as decisões do modelo mais compreensíveis, através de regras pré-definidas ou estruturas específicas de modelagem. *Data-driven* consiste numa abordagem em que a explicabilidade é derivada diretamente dos dados utilizados para o treinar do modelo, pelo que se foca em entender como as características es-

pecíficas dos dados de treino influenciam as saídas do modelo [38]. Alguns exemplos de técnicas de *Explainable AI data-driven* são o *SHapley Additive exPlanations* (SHAP) e o *Local Interpretable Model-agnostic Explanations* (LIME).

Um das técnicas bastante utilizadas em XAI é o LIME. Este trata-se de um "modelo agnóstico" (isto é, um modelo que pode ser aplicado a um vasto conjunto de modelos de ML, sem depender das suas características internas), que tem como objetivo explicar a previsão do modelo para instâncias individuais [39, 40]. A ideia central do LIME é aproximar, localmente, um modelo complexo a um modelo mais simples e interpretável, que seja suficientemente representativo da vizinhança da instância de dados que se pretende explicar, permitindo assim que o modelo deixe de funcionar como uma "caixa negra".

A abordagem do LIME consiste numa perturbação das entradas do modelo, observando como este se comporta nas novas previsões, de modo a aprender seu o funcionamento, através de um modelo linear com a ponderação das perturbações. Para explicar um classificador de imagens, o LIME destaca os superpíxeis (isto é, a coleção de píxeis que cobrem uma área ligada de uma imagem) que mais justificam a eleição de uma determinada classe. Os superpíxeis devem corresponder a padrões específicos de uma imagem, sendo que o utilizador apenas pode especificar a resolução das áreas selecionadas. Isto coloca uma dificuldade adicional, uma vez que características significativas podem estar em diferentes superpíxeis. Contudo, é possível ter algum tipo de controlo sobre este fator, sabendo o tamanho relativo das áreas afetadas, o que permite que se possa ajustar o número de superpíxeis (através do parâmetro *ratio*, sendo que quanto menor este valor, maior o número de superpíxeis) e, consequentemente, o seu tamanho.[41]

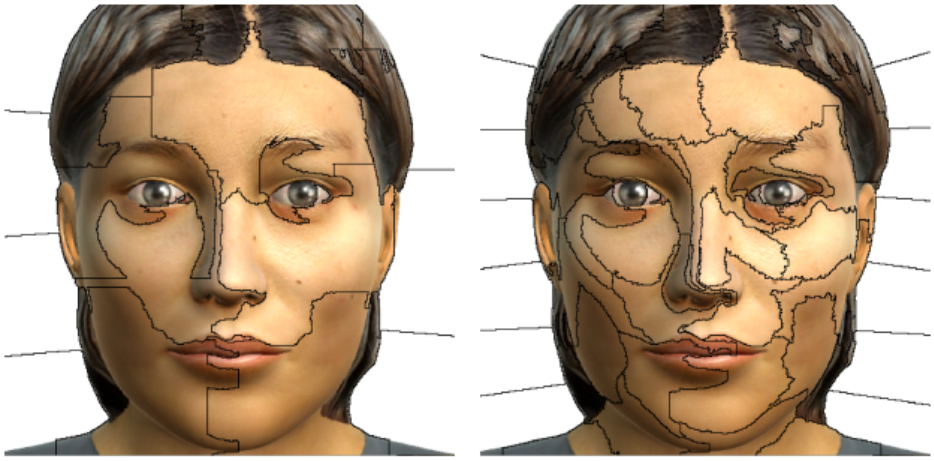


Figure 2.9: Exemplo de uma imagem dividida com diferentes números de superpíxeis (Imagem retirada de: [41]).

Esta é uma abordagem visa oferecer transparência e interpretabilidade aos modelos de ML, o que ajuda os utilizadores a entender como as decisões foram tomadas. Em *Python*, esta técnica pode ser implementada através da biblioteca LIME.

A utilização de XAI torna-se especialmente relevante em setores críticos, como na área da saúde, uma vez que existe uma necessidade acrescida de compreender as decisões tomadas

pelos modelos de IA. Estas explicações permitem que os profissionais compreendam as razões por trás de uma previsão ou diagnóstico, aumentando assim a confiança nas ferramentas de IA e facilitando a integração destas tecnologias no processo de tomada de decisões. Além disso, a XAI ajuda a garantir a transparência e ética dos modelos, promovendo assim uma prática médica mais justa e precisa.

2.7 Ferramentas de Avaliação

Para avaliar a performance num problema multiclasse, é comum usar um conjunto de métricas como ***accuracy***, ***precision***,***recall***, ***f1-score***, ***Cohen's Kappa*** e ***Chance Adjusted Accuracy***.

2.7.1 *Accuracy*

A métrica ***accuracy*** (Equação 2.1) é frequentemente utilizada para avaliar o desempenho de um modelo ao longo das épocas de treino (*epochs*). Esta métrica mede a proporção entre o número de amostras corretamente classificadas e o número total de amostras do *dataset*, resultando num valor entre 0 e 1, em que o valor 1 representa uma *accuracy* perfeita, em que todas as amostras foram corretamente classificadas, e 0 representa uma *accuracy* nula, o que significa que nenhuma amostra foi classificada corretamente.

$$accuracy = \frac{\text{Amostras corretamente classificadas}}{\text{Total de Amostras do } dataset} \tag{2.1}$$

Apesar da *accuracy* ser bastante popular para a avaliação do desempenho de um modelo, esta métrica pode causar uma falsa impressão de bom desempenho quando os dados não são balanceados, isto é, quando há uma grande disparidade na distribuição das classes. Nestes casos, a *accuracy* pode ser influenciada pela classe dominante, pelo que não fornece uma compreensão completa do desempenho do modelo. Por exemplo, num conjunto de dados em que 80% dos dados pertençam a uma só classe, basta classificar esses exemplos corretamente para se obter uma *accuracy* de 80%, mesmo que todos os elementos de outra classe estejam classificados incorretamente.

2.7.2 *Precision*

A ***precision*** (Equação 2.2) representa a relação entre o número de amostras classificadas como pertencentes a uma classe, que realmente lhe pertencem, e o número total de amostras classificadas nessa classe.

$$precision = \frac{\text{Amostras corretamente classificadas numa determinada classe}}{\text{Total de Amostras classificadas nessa classe}} \tag{2.2}$$

Uma *precision* elevada indica que o classificador tem uma baixa taxa de falsos positivos, ou seja, que a maioria das amostras classificadas como pertencentes a uma classe realmente lhe

pertencem.

2.7.3 *Recall*

A ***recall*** (Equação 2.3) representa a relação entre o número total de amostras corretamente classificadas numa determinada classe e o número total de amostras pertencentes a essa classe, mesmo que classificadas numa outra classe. Esta métrica é também frequentemente chamada de taxa de *sensitivity*.

$$recall = \frac{\text{Amostras corretamente classificadas numa determinada classe}}{\text{Total de Amostras pertencentes a essa classe}} \tag{2.3}$$

Para problemas de classificação multiclasse, o valor final desta métrica é obtido usando a média ponderada (*weighted average*). Nesta abordagem, o *recall* é calculado individualmente para cada classe e, em seguida, cada valor obtido é ponderado pelo número de amostras presentes nessa classe. Desta forma, é possível assegurar que o impacto de cada classe no valor final desta métrica é proporcional à sua frequência no conjunto de dados, o que é especialmente importante em conjuntos de dados não balanceados, onde algumas classes podem ter significativamente mais amostras do que outras.

2.7.4 *F1-Score*

A ***F1-score*** (Equação 2.4) é uma métrica que combina a *precision* e a *recall*, sendo geralmente descrita como a média harmónica entre as duas.

$$f1\text{-}score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{2.4}$$

Esta métrica fornece uma avaliação mais ampla acerca do desempenho do modelo, pelo que é bastante útil em casos em que as classes não estão equilibradas. De um modo geral, quanto maior o valor de *f1-score*, melhor o desempenho do modelo.

À semelhança da métrica *recall*, o resultado final desta métrica é obtido através de uma média ponderada dos valores obtidos para as diferentes classes. Isso significa que o *f1-score* é calculado individualmente para cada classe, e em seguida, cada valor de *f1-score* é ponderado pelo número de amostras presentes nessa classe.

2.7.5 *Cohen's Kappa*

Em modelos com *datasets* não balanceados (isto é, modelos em que as classes estão representadas de forma desigual nos dados), a avaliação do seu desempenho pode ser um desafio, uma vez que a métrica *accuracy* não é suficiente para refletir a eficácia do modelo. Nestes casos, é bastante comum recorrer-se ao coeficiente de *Cohen's Kappa*, que é utilizado para medir a concordância entre as classificações observadas e esperadas, tendo em consideração a possibilidade

de concordâncias aleatórias. Desta forma, esta métrica fornece uma medida mais robusta da concordância entre as classificações observadas e as esperadas.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.5}$$

A Equação 2.5 representa o cálculo deste coeficiente, em que:

- p_o representa a *accuracy* observada, calculada como a proporção de dados que o modelo classificou corretamente, através da Equação 2.6, em que n_{ii} representa número de vezes que o modelo previu corretamente a classe i e N o número total de amostras.

$$p_o = \frac{\sum_{i=1}^{k} n_{ii}}{N} \tag{2.6}$$

- p_e diz respeito à *accuracy* esperada ao acaso, calculada através da Equação 2.7 , em que $n_{i\cdot}$ diz respeito ao número total de casos na classe i nas previsões do modelo e $n_{\cdot i}$ ao número total de casos na classe i nas *labels* verdadeiras.

$$p_e = \sum_{i=1}^{k} \left(\frac{n_{i\cdot} \cdot n_{\cdot i}}{N^2} \right) \tag{2.7}$$

O valor deste coeficiente varia entre -1 e 1, podendo os valores de κ serem interpretados da seguinte forma:

- $\kappa < 0$: Concordância menor do que a esperada pelo acaso; Indica que as classificações são piores do que as esperadas por puro acaso.
- $\kappa = 0$: Concordância equivalente ao acaso; Não há concordância significativa além do acaso.
- $0 < \kappa \leq 0.20$: Concordância mínima (ligeira); Indica uma concordância muito baixa além do acaso.
- $0.21 \leq \kappa \leq 0.40$: Concordância fraca; Há alguma concordância entre as classificações, mas é considerada baixa.
- $0.41 \leq \kappa \leq 0.60$: Concordância moderada; Este intervalo indica uma concordância razoável entre as classificações.
- $0.61 \leq \kappa \leq 0.80$: Concordância substancial; Este intervalo representa um bom nível de concordância, indicando que as classificações são bastante consistentes.
- $0.81 \leq \kappa \leq 1.00$: Concordância quase perfeita; Indica um nível muito alto de concordância, quase idêntica entre as classificações [42].

2.7.6 *Chance Adjusted Accuracy*

Chance Adjusted Accuracy (CAA) trata-se de uma métrica de avaliação de desempenho em modelos de ML, bastante semelhante à *Cohen's Kappa*, que ajusta a *accuracy* de um modelo tendo em consideração a *accuracy* esperada pelo acaso. Esta métrica é bastante utilizada para avaliar problemas de classificação em ML, principalmente quando há uma disparidade entre o número de amostras em cada classe.

A CAA representa um valor entre -1 e 1, calculado a partir da Equação 2.8, onde *Accuracy* representa a *accuracy* observada (proporção de classificações corretas) e $Accuracy_{\text{acaso}}$ a *accuracy* esperada pelo acaso, que pode ser calculada como a soma das probabilidades de classificação correta para cada classe, se as classificações fossem feitas aleatoriamente [24]. O valor desta métrica pode ser interpretado como:

- CAA < 0: A *accuracy* do modelo é pior do que o esperado pelo acaso.
- CAA $= 0$: A *accuracy* do modelo é igual à esperada pelo acaso. Isto é, o modelo não está melhor que a classificação aleatória.
- $0 <$ CAA ≤ 0.20: O modelo tem uma leve melhoria em relação ao acaso; Concordância mínima (ligeira).
- $0.21 \leq$ CAA ≤ 0.40: O modelo tem uma pequena melhoria; Concordância fraca.
- $0.41 \leq$ CAA ≤ 0.60: O modelo apresenta uma melhoria moderada; Concordância moderada.
- $0.61 \leq$ CAA ≤ 0.80: O modelo tem uma boa melhoria; Concordância substancial.
- $0.81 \leq$ CAA ≤ 1.00: O modelo é significativamente melhor que o acaso; Concordância quase perfeita.

$$\text{CAA} = \frac{Accuracy - Accuracy_{\text{acaso}}}{1 - Accuracy_{\text{acaso}}} \tag{2.8}$$

2.7.7 Matriz de Confusão

A **matriz de confusão** é uma ferramenta crucial na avaliação do desempenho de modelos de classificação. Esta matriz organiza os resultados das previsões do modelo em relação aos verdadeiros rótulos das amostras, fornecendo uma visão detalhada das diferentes formas como o modelo está a classificar as amostras. Numa matriz de confusão, os rótulos verdadeiros e as previsões são organizados numa matriz, onde cada linha representa as amostras pertencentes a uma classe real e cada coluna representa as amostras previstas como pertencentes a uma classe. Os elementos desta matriz representam a contagem de amostras para cada combinação de classe verdadeira e classe prevista.

A interpretação desta matriz pode fornecer *insights* valiosos sobre o desempenho do modelo, uma vez que permite identificar facilmente e de forma visual quais as classes que o modelo

confunde com mais frequência, o que pode indicar onde o modelo precisa de ser ajustado ou áreas onde o conjunto de dados pode ser desbalanceado [43].

2.8 Dados não balanceados

Na área do ML, a qualidade dos dados é fundamental para o sucesso dos modelos desenvolvidos. No entanto, é comum depararmo-nos com um desafio significativo: o não balanceamento dos dados. Este problema é de extrema relevância, pois pode distorcer a capacidade do modelo aprender de forma precisa e eficaz [44, 45].

O não balanceamento dos dados refere-se a situações em que as classes de um conjunto de dados não estão representadas de forma equilibrada, ou seja, uma ou mais classes têm significativamente mais ou menos amostras do que outras classes. Estas situações podem levar a desafios na construção de modelos ML, principalmente quando se trata de classes minoritárias, uma vez que os algoritmos podem ter dificuldade em aprender padrões dessas classes, devido à falta de amostras para treino. Assim, a sua capacidade de generalização diminui, pois o modelo tende a favorecer as classes maioritárias, deixando de parte as minoritárias, pelo que a sua capacidade de representação diminui. Em contextos médicos, por exemplo, o não balanceamento das classes pode ser problemático quando se trata de prever eventos raros, como doenças graves, que podem ser a classe minoritária [44, 46].

Este problema pode afetar o modelo de várias maneiras. Este pode tornar-se tendencioso em relação à classe maioritária, resultando assim num baixo desempenho na classificação da classe minoritária e, consequentemente, numa baixa capacidade de generalização para novos dados não balanceados, o que se traduz numa baixa capacidade de lidar com a distribuição real dos dados no mundo real. Além disso, as métricas de avaliação tradicionais, como a *accuracy* podem ser enganosas para este tipo de conjuntos, uma vez que um modelo pode obter uma *accuracy* elevada apenas por prever corretamente amostras da classe maioritária, sem realmente aprender de forma eficaz a classe minoritária. Estes modelos tornam-se também mais sensíveis a ruídos e exemplos atípicos, o que pode levar a decisões erradas ou instáveis [47].

2.8.1 *Oversampling*

O ***oversampling*** é uma técnica de balanceamento de dados, utilizada em conjuntos de dados não balanceados em que uma ou mais classes apresentam um número de amostras significativamente menor, em relação a outras classes. Esta técnica tem como objetivo aumentar o número de amostras das classes minoritárias, sem necessidade de se coletar novas amostras. Para tal, as amostras da classe minoritária são replicadas ou geradas sinteticamente, de forma aleatória, aumentando assim a sua representação no conjunto de dados, o que permite equilibrar a distribuição das classes e fornecer mais informações ao modelo sobre a classe minoritária.

O método de *oversampling* deve ser aplicado apenas ao conjunto de treino, sendo particularmente útil quando há poucas amostras na classe minoritária e é importante aumentar a sua representação para melhorar a capacidade do modelo aprender com esta classe e, consequente-

mente, melhorar o desempenho do modelo. No entanto, se o número de amostras gerado for excessivo, o risco de *overfitting* pode aumentar. Neste cenário, verifica-se que o modelo tende a se ajustar muito bem aos dados de treino, mas pode falhar em generalizar de forma adequada para novos dados.

2.8.2 *Undersampling*

O ***undersampling*** é, contrariamente ao *oversampling*, uma técnica que procura resolver o problema do não balanceamento dos dados diminuindo a quantidade de amostras presentas na classe maioritária, de modo a igualá-la à classe minoritária. Nesta técnica, amostras da classe maioritária são removidas do conjunto de dados, de forma aleatória, para equilibrar a distribuição entre as classes, criando assim um conjunto de dados balanceado.

A técnica de *udersampling* é eficaz para lidar com conjuntos de dados muito grandes, onde o *oversampling* pode ter elevados custos computacionais. No entanto, apesar de o *undersampling* poder ajudar a equilibrar a distribuição de classes, este pode também pode levar à perda de informações importantes contidas nas amostras removidas, afetando assim a capacidade do modelo aprender de forma eficaz [46, 47].

2.9 Resumo

No presente capítulo, é abordada, de forma abrangente, uma ampla gama de fundamentos teóricos relacionados com a temática da pesquisa. Inicialmente, são explorados os diversos aspetos da doença de *Alzheimer*, incluindo as suas causas, fases, diagnóstico e tratamento, bem como as técnicas de imagiologia utilizadas para o diagnóstico desta doença. Numa fase seguinte, é abordado o campo do AL e são descritas, em detalhe, as arquiteturas VGG-16, VGG-19 e ResNet-50, utilizadas para o treino dos modelos. Posteriormente, são também abordadas as ferramentas de avaliação e algumas estratégias de processamento de dados para enfrentar desafios como o não balanceamento dos dados. Os conceitos abordados ao longo deste capítulo servem de base teórica para estudos e análises subsequentes neste trabalho, proporcionando uma compreensão abrangente e fundamentada para as investigações futuras.

3

Trabalhos relacionados

Nos últimos anos, com o crescimento de doenças neurodegenerativas, como o *Alzheimer*, houve uma maior necessidade de estudos e pesquisas para as compreender e perceber como estas afetam o cérebro humano. Neste sentido, uma vez que os danos causados são irreversíveis, não é surpresa que tenham surgido imensas pesquisas no campo da Inteligência Artificial (IA), como *Machine Learning* (ML) e *Deep Learning*, com o objetivo de proporcionar diagnósticos mais precisos e precoces, permitindo assim uma melhor gestão deste tipo de doenças.

3.1 IA na doença de *Alzheimer*

Kantayeva [48] elaborou um artigo que realiza uma revisão de diversos trabalhos que exploram o uso da IA na saúde, particularmente em casos de *Alzheimer* e outros tipos de demência, oferecendo uma discussão acerca da relevância e impacto destas técnicas neste contexto. Para tal, foram utilizados 25 trabalhos, publicados na base de dados *Scopus* num período de 20 anos (2001 a 2021), sendo que as principais questões a que se pretendia responder eram: "Qual é a relevância e o impacto do ML na área da saúde, em particular na demência?", "Qual é o objetivo da utilização dos métodos de ML?","Quais os métodos de ML mais utilizados na área da demência?" e "Que características são utilizadas para classificar ou prever a demência?". A estratégia de pesquisa foi composta por três fases. Em primeiro lugar foram selecionados os artigos a rever, através da leitura dos títulos dos artigos e da análise das *keywords* (de modo a verificar se continham palavras de interesse, como *Machine Learning* ou *Health*)e dos *abstracts* (para confirmar se o artigo realmente ia de encontro ao pretendido). De seguida, foi feita uma análise bibliométrica e uma discussão dos estudos considerados. Os autores concluíram que os algoritmos de ML apresentam resultados promissores para auxílio ao diagnóstico de demência, sendo que a maioria dos trabalhos referidos utilizam a imagem por ressonância magnética como principal característica para a previsão, acompanhada de dados demográficos tais como a idade e o género do paciente. No entanto, salientaram algumas limitações a estes algoritmos, tais como a disponibilidade e o pré-processamento dos dados.

Yao [49] elaborou um artigo de revisão que aborda os últimos avanços a nível da IA e da técnica de imagem por ressonância magnética para diagnóstico da doença de *Alzheimer*, com o objetivo de fornecer uma visão global dos vários algoritmos e modelos que foram desenvolvidos para analisar imagens MRI e diagnosticar esta doença. Para tal, os autores oferecem uma

discussão acerca do potencial de vários modelos de *Deep Learning* e *Transfer Learning* utilizados, refererindo a importância da explicabilidade (*Explainable AI* (XAI)) para promover a transparência e a confiança nos sistemas de IA, o que pode constituir uma ferramenta importante para melhorar a eficácia e a qualidade dos cuidados prestados aos doentes. Este artigo salienta ainda a importância da deteção precoce e das intervenções atempadas para a doença de *Alzheimer* e a forma como os diagnósticos auxiliados pela IA podem ajudar a atingir esse objetivo. No entanto, são também destacados alguns desafios a este diagnóstico, tal como a falta de dados universais de ressonância magnética dada a variação nas imagens devido aos diferentes equipamentos utilizados, pelo que os autores referem uma necessidade de mais investigação, de modo a minimizar estas limitações e a aumentar a eficiência e precisão dos modelos.

Abd El-Latif [50] propôs um modelo de *Lightweight Deep Learning* para deteção precoce da doença de *Alzheimer*, utilizando imagens Ressonância Magnética (MRI). O modelo proposto foi concebido para aprender e extrair *features* de grandes conjuntos de dados (o que o torna adequado para analisar imagens médicas), distinguindo-se por ser "leve", uma vez que só usa sete camadas, tornando assim o sistema menos complexo e mais rápido. Este modelo foi testado com recurso a um conjunto de dados disponível na plataforma *Kaggle*, tendo-se obtido uma *accuracy* de 99,22% para classificação binária e de 95,23% para classificação em quatro categorias, mostrando-se, assim, eficiente e eficaz na deteção da doença de *Alzheimer*, possuindo um elevado desempenho sem a necessidade de camadas mais profundas.

Uma vez que as fases iniciais da doença de *Alzheimer* apresentam características bastante semelhantes e pode ser difícil distingui-las, Khalid [51] elaborou um estudo que teve como objetivo desenvolver sistemas eficazes para ajudar os médicos e os radiologistas a diagnosticar precocemente a doença de *Alzheimer* e a prever a sua progressão. O autor propôs a utilização de abordagem híbridas para extração de *features* de imagens MRI, combinando as *features* extraídas para serem introduzidas num algoritmo de rede neural *feedforward* (FFNN), de modo a diagnosticar a doença e prever a sua progressão. Deste estudo, resultaram em três metodologias, cada uma delas com dois sistemas. Para a primeira metodologia foram utilizados os modelos *GoogLeNet* e *DenseNet-121*, separadamente, para a extração de *features* e uma rede FFNN. A segunda metodologia consiste numa rede FFNN com *features* combinadas dos modelos *GoogLeNet* e *Dense-121*, antes e após a redução de alta dimensionalidade, através do algoritmo de Análise de Componentes Principais (PCA). Por fim, a terceira metodologia consiste na utilização de uma rede FFNN com *features* combinadas da aplicação separada dos modelos *GoogLeNet* e *Dense-121* e das *features* extraídas por *Discrete Wavelet Tranfor*(DWT), *Local Binary Pattern* (LBP) e *Gray Level Co-occurrence Matrix* (GLCM), denominados de métodos tradicionais ou artesanais. Todos estes sistemas produziram resultados excelentes na deteção da doença de *Alzheimer* e na previsão das fases da sua progressão, sendo que, com a terceira metodologia, a FFNN alcançou uma *accuracy* de 99,7%.

Alshmrany [52] também propôs uma abordagem híbrida, combinando técnicas de *Transfer Learning* e com *Deep Learning* para deteção da doença de *Alzheimer* por meio de de exames de Tomografia Computarizada (Tomografia Computarizada (TC)) ou Ressonância Magnética

(MRI). O modelo centra-se na classificação dos doentes nas diversas fases da doença, através da aprendizagem por transferência (*Tranfer Learning* com modelos pré-treinados, tais como *Residual Network* de 50 camadas (ResNet-50), *Visual Geometry Group* de 16 camadas (VGG-16) e DenseNet121, juntamente com redes *Convolutional Neural Network* (CNN) num grande conjunto de dados. Adicionalmente, os autores utilizaram técnicas de *downsampling* para previnir o *overfitting*. Este modelo obteve resultados promissores, com uma *accuracy* final de 96,6%, representando uma melhoria significativa em comparação com estudos anteriores, conforme destacado pelos próprios autores.

AlSaeed [53] propôs um algoritmo de ML para deteção da doença de *Alzheimer*, constituído por duas fases: extração/seleção de *features* e classificação binária (com ou sem *Alzheimer*) de imagens MRI. Para a primeira fase, os autores realizaram uma segmentação dos tecidos através de diversas técnicas, tais como *fuzzy possibilistic*, *speeded-up robust features* (SURF) e redes neurais convolucionais ResNet-50 pré-treinadas, de modo a extrair características importantes das imagens MRI. Na segunda fase pretendia-se classificar as imagens, pelo que os autores recorreram a vários classificadores, tais como *support Vector Machines* (SVM) e *Random forest* (RF). O algoritmo proposto foi testado com recurso a dois *datasets* e através de várias métricas, tendo sido obtida uma *accuracy* de 96.875% para o primeiro *dataset* e 95.83% para o segundo.

Nithya [34] desenvolveu um modelo de *Deep Learning* para deteção precoce e classificação na doença de *Alzheimer*. O modelo proposto visa superar métodos tradicionais, oferecendo maior precisão na classificação das fases iniciais da doença, possibilitando assim diagnósticos mais confiáveis e oportunos. Além disso, o modelo é projetado para ser acessível em diversos ambientes de saúde, exigindo menos recursos computacionais, devido à eficiente arquitetura ResNet-50. O modelo proposto utiliza técnicas de pré-processamento inovadoras, como "*Contrast Limited Adaptive Histogram Equalization* (CLAHE)" e "*Boosted Anisotropic Diffusion Filters* (BADF)", para aprimorar a qualidade das imagens de entrada. Esta abordagem de pré-processamento garante uma representação ótima dos dados, o que permite que o modelo capture, com elevada precisão, padrões subtis relacionados com doença. Além disso, o modelo emprega o algoritmo de *clustering K-means* para a segmentação precisa de regiões cerebrais em exames de MRI, o que facilita a análise e a extração eficiente de características. Os resultados obtidos com este estudo foram altamente promissores, tendo o modelo alcançado uma *accuracy* de 95% e uma perda mínima de 0,12.

3.2 XAI na saúde

Magesh [54] elaborou um estudo com o objetivo de desenvolver um modelo de ML para classificar com precisão imagens como tendo ou não doença de *Parkinson* que fornecesse uma razão plausível para a previsão. Para tal, recorreu ao uso de XAI através do *Local Interpretable Model-agnostic Explanations* (LIME), de modo a que a solução fosse interpretável, permitindo assim que os médicos especialistas pudessem compreender por que razão a máquina pensa de uma determinada forma, fornecendo informações cruciais para o processo de tomada de decisões. O modelo proposto utiliza a arquitetura VGG-16 para deteção precoce da doença de *Parkinson*

através imagens SPECT DaTSCAN e foi treinado utilizando *Transfer Learning*, obtendo uma *accuracy* de 95,2%, uma sensibilidade de 97,5% e uma especificidade de 90,9%. Os autores concluíram que o modelo proposto é uma ferramenta promissora para o diagnóstico e tratamento precoce da doença de *Parkinson*, uma vez que fornece resultados de classificação exatos. Além disso, consideram também que a utilização de LIME proporcionou uma visão significativa das decisões geradas pelo modelo, o que se traduz numa melhoria da interpretabilidade dos resultados da classificação. Desta forma, o modelo proposto tem a capacidade de fornecer aos médicos um apoio valioso na tomada de decisões.

Também Davagdorj [55] elaborou um estudo com XAI na área da saúde, nomeadamente com doenças não transmissíveis (isto é, doenças que não são causadas por agentes infeciosos e não são passíveis de serem transmitidas de uma pessoa para outra, como é o caso de doenças cardiovasculares, diabetes ou doenças neuropsiquiátricas como o *Alzheimer*). Os autores tinham como objetivo propor um algoritmo de *Deep Learning* para previsão precoce destas doenças, pelo que propuseram um algoritmo composto por três partes: seleção de *features*, treino do modelo e explicação do modelo. Para o treino do modelo, aplicaram um classificador de rede neural profunda (DNN), ajustado com hiper-parâmetros, ao subconjunto de *features* selecionado. De modo a conseguir oferecer uma explicação do modelo, tanto a nível global como a nível local, os autores recorreram ao DeepSHAP. A eficácia do modelo foi avaliada num conjunto de dados reais, tendo este estudo demonstrado resultados promissores em termos de precisão de previsão e explicabilidade

3.2.1 *Explainable AI* no *Alzheimer*

Mulyadi [56] propôs um algoritmo de aprendizagem profunda (*Deep Learning*) para modelação da progressão da doença de *Alzheimer* em imagens MRI 3D, em conjunto com *eXplainable AD Likelihood Map Estimation (XADLiME)* para identificação de biomarcadores da doença, através de imagens funcionais do cérebro, tendo por objetivo fornecer um método explicável e preciso para identificação e diagnóstico de biomarcadores da doença de *Alzheimer*. O modelo proposto utiliza a aprendizagem de protótipos orientada clinicamente, de modo a estabelecer protótipos que capturam a gama de amostras que refletem possíveis vias de progressão da doença. Assim, o modelo estima mapas de probabilidade de *Alzheimer* a partir de exames de MRI 3D e promove a explicabilidade a partir de perspetivas clínicas e morfológicas. O mapa de probabilidade estimado é utilizado para várias tarefas clínicas, como o diagnóstico e o prognóstico, mantendo o desempenho do diagnóstico e a interpretabilidade. Para a avaliação do modelo, os autores compararam-no com cinco modelos de classificação base através de várias métricas, o que lhes permitiu concluir que o modelo proposto apresenta resultados comparáveis ou até mesmo superiores em várias tarefas clínicas, ao mesmo tempo que fornece uma melhor compreensão mais abrangente das vias de progressão da doença de *Alzheimer*, promovendo assim a explicabilidade.

Qiu [57] propôs também uma estratégia de *Deep Learning* para diagnosticar com precisão a doença de *Alzheimer* através de dados MRI. Os autores apresentam um modelo que combina uma

Fully Convolutional Network (FCN), responsável pela construção de mapas de alta resolução da probabilidade da doença a partir da estrutura cerebral, com um perceptrão multicamada (*Multilayer Perceptron*- MLP), que processa os mapas da FCN para gerar visualizações presisas do risco da doença. O modelo proposto é capaz de fornecer probabilidades locais intuitivas que são facilmente interpretáveis, contribuindo para o movimento crescente da XAI na medicina. Para o desenvolvimento e validação do modelo, foram utilizados quatro *datasets* diferentes, tendo-se obtido bons desempenhos de precisão em todos eles, o que sugere um forte grau de especialização.

Kwanseok Oh[58] realizou um estudo com objetivo de abordar o desafio da interpretabilidade em modelos de *Deep Learning* para previsão da doença de *Alzheimer*. O modelo proposto combina o raciocínio contrafactual com os mapas explicativos visuais, de modo a transformá-los em características quantitativas, o que permite uma avaliação mais objetiva e mensurável da decisão do modelo. Para o desenvolvimento deste estudo, os autores utilizaram um *dataset* de imagens MRI de 1540 indivíduos para elaborar um mapa do efeito da doença nas diversas regiões por ela afetadas, de modo a investigar quantitativamente as variações anatómicas na densidade da matéria cinzenta. Desta forma os investigadores puderam concluir que, apesar de existirem alguma zonas que se destacam em todos os casos, existem também áreas específicas para cada nível de gravidade da doença, o que se traduz numa variação do número de regiões de interesse em cada fase do *Alzheimer*, pelo que ficaram convencidos que o seu mapa refletia as áreas mais prováveis para classificar as diferentes fases. O modelo proposto obteve uma melhoria no desempenho de previsão face a modelos anteriores, ao mesmo tempo que promove a interpretabilidade, uma vez que apresenta potencial para um previsão mais eficaz da doença de *Alzheimer*.

3.3 Resumo

Neste capítulo, foram apresentados alguns estudos relativos à IA na saúde, mais concretamente na doença de *Alzheimer*. Através destes, é possível perceber que existem várias estratégias a ser estudadas, que apresentam resultados promissores no auxílio ao diagnóstico da doença de *Alzheimer*. Além disso, estes estudos apresentam também a XAI como um fator crucial para promover a transparência e confiança dos sistemas de IA na área da saúde. A capacidade de explicar as decisões tomadas pelos modelos de IA é de extrema importância, especialmente em contextos médicos, um vez que é fundamental compreender as razões por trás de uma decisão, de modo a aumentar a confiança dos profissionais de saúde nestes sistemas, contribuindo assim para a aceitação da IA na área da saúde. Além disso, a utilização de XAI também abre caminho para avanços mais significativos no diagnóstico e tratamento de doenças como o *Alzheimer*. Na Tabela 3.1 encontra-se uma breve comparação dos modelos propostos nos artigos acima referidos.

* *Accuracy* não especificada. No entanto, os autores destes trabalhos consideram os seus resultados promissores quando comparados com os de outros estudos analisados para a realização dos seus.

Autores	Técnicas Utilizadas	Desempenho do modelo
Abd El-Latif [50]	*Lightweight DL*	*Accuracy*: 99.22% (binária), 95.23% (quatro categorias)
Khalid [51]	Abordagens híbridas (ML + DL)	*Accuracy*: 99.7%
Alshmrany [52]	Transfer Learning, DL	*Accuracy*: 96.6%
AlSaeed [53]	ML, Seleção de Features	*Accuracy*: 96.875% (Dataset 1), 95.83% (Dataset 2)
Nithya [34]	DL	*Accuracy*: 95%
Magesh [54]	DL, Explainable AI (LIME)	*Accuracy* 95.2%
Davagdorj [55]	DL, Explainable AI (DeepSHAP)	*
Mulyadi [56]	DL, XADLiME	*
Qiu [57]	DL	*
Kwanseok Oh [58]	DL, Raciocínio Contrafactual	*

Table 3.1: Comparação dos Trabalhos *Related Work*.

4

Materiais, Métodos, Resultados e Discussão

A doença de *Alzheimer* é uma condição neurodegenerativa progressiva que afeta principalmente idosos, caracterizada pela deterioração cognitiva e pela perda de memória, impactando significativamente as tarefas do dia a dia. A sua deteção nas fases iniciais é crucial, uma vez permite intervenções mais atempadas e eficazes para retardar a progressão dos sintomas e melhorar a qualidade de vida dos pacientes. Além disso, o seu diagnóstico precoce oferece também a oportunidade de planear o tratamento e o suporte necessário, ajudando tanto os pacientes quanto as suas famílias a lidar melhor com os desafios futuros.

O presente estudo propôs-se a explorar o uso da Inteligência Artificial (IA) na deteção precoce da doença de *Alzheimer*, através de imagens Ressonância Magnética (MRI), de modo a oferecer aos profissionais de saúde uma ferramenta de auxílio eficaz para identificar alterações cerebrais relacionadas com a doença. De modo a tornar os resultados mais claros e interpretáveis, recorreu-se ao uso de *Explainable AI* (XAI), facilitando assim a compreensão e a confiança dos profissionais de saúde nas decisões geradas pelos modelos.

Através da pesquisa efetuada para a realização do capítulo anterior, foi possível adquirir um melhor conhecimento acerca dos estudos já realizados nesta área, bem como das técnicas que já estão a ser exploradas. No entanto, foi também possível verificar que ainda não existe nenhum trabalho para diagnóstico da doença de *Alzheimer* que inclua técnicas de *Active Machine Learning* (AL). A implementação destas técnicas pode ser bastante vantajosa na área da medicina, uma vez que permite uma abordagem mais eficiente e adaptativa na construção dos modelos. Na área da medicina, a obtenção de dados rotulados de grande qualidade pode ser bastante complicada, pelo que se torna importante o estudo destas técnicas para melhorar os resultados dos modelos de IA, de modo a maximizar a eficiência dos dados disponíveis, reduzindo a necessidade de rotulagem extensiva, mesmo quando o conjunto de dados é limitado.

Para a realização deste trabalho, escolheu-se a linguagem de programação *Python*. Esta linguagem é amplamente utilizada em IA, devido à sua sintaxe intuitiva, à vasta gama de bibliotecas especializada e à sua popularidade. Assim, a sua utilização é uma tendência atual em IA, pelo que a sua escolha como linguagem de programação central neste trabalho facilita a comunicação com outros pesquisadores e o acesso aos recursos disponíveis na comunidade, o

que contribui para um desenvolvimento mais eficiente e colaborativo do projeto.

O presente capítulo apresenta, o *dataset* utilizado para neste trabalho, bem como os métodos e processos utilizados no desenvolvimento dos modelos e os resultados obtidos através da sua aplicação.

4.1 *Dataset*

Para a elaboração do presente trabalho, recorreu-se a um *dataset* retirado da plataforma *Kaggle* [59]. Este conjunto de dados é constituído por imagens MRI do cérebro de vários indíduos, de tamanho 128x128 e pré-processadas, provenientes de várias plataformas de hospitais e repositórios públicos.

As imagens no *dataset* representam secções transversais do plano axial, extraídas das digitalizações 3D originais em localizações não normalizadas. Esta abordagem permite capturar diferentes cortes do cérebro, fornecendo assim uma visão abrangente para a análise.

O *dataset* é composto por 6400 imagens, estando estas divididas em quatro classes: *Non Demented*,*Very Mild Demented*, *Mild Demented* e *Moderate Demented*. A Tabela 4.1 apresenta a distribuição das imagens pelas classes, enquanto que a Figura 4.1 ilustra, de forma gráfica, esta distribuição.

Fase da doença	Número de Imagens
Non Demented	3200
Very Mild Demented	2240
Mild Demented	896
Moderate Demented	64

Table 4.1: Distribuição do *dataset*.

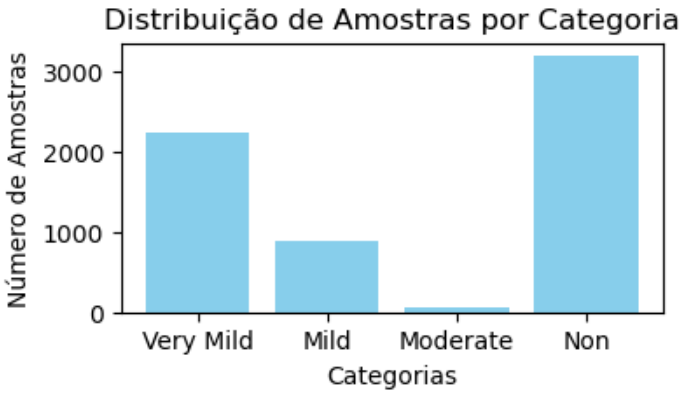


Figure 4.1: Distribuição do *dataset*.

Na Figura 4.2 encontram-se algumas imagens exemplificativas das classes presentes neste *dataset*.

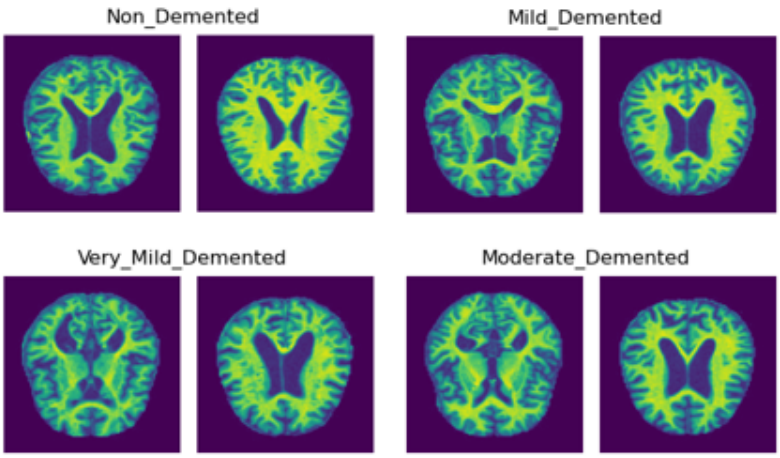


Figure 4.2: Imagens de cada classe presente no *dataset*

4.2 Divisão dos dados

Numa fase inicial, o *dataset* foi dividido em três subconjuntos, para treino, teste e validação, em proporções de 80%, 10% e 10%, respetivamente. O conjunto de treino é utilizado para que o modelo aprenda os padrões entre os dados. O conjunto de validação é utilizado para avaliar o desempenho do modelo ao longo do treino, permitindo assim o ajuste dos hiperparâmetros e a prevenção de *overfitting*. Por fim, o conjunto de teste é utilizado para avaliar o desempenho final do modelo, fornecendo uma avaliação imparcial do quão bem o modelo generaliza para dados não vistos. Esta divisão estratégica dos dados permite que o treino seja adequado e com ajustes precisos, garantindo assim um modelo robusto e confiável.

4.3 Implementação dos Modelos

4.3.1 *Early Stopping*

Durante o treino do modelo, foi utilizada a técnica de *Early Stopping*, de modo a evitar o *overfitting* e melhorar o desempenho geral do modelo, aperfeiçoando assim a sua capacidade preditiva. Esta técnica consiste na interrupção do treino de uma rede neural, quando há indícios de que o modelo está a ajustar-se em demasia aos dados de treino, o que poderia levar a uma redução da sua capacidade de generalização. Assim sendo, esta ferramenta monitoriza o desempenho do modelo, num conjunto de validação, ao longo da fase de treino, interrompendo o processo quando o este começa a piorar e guardando a melhor performance obtida [60]. O uso desta técnica foi fundamental para a obtenção de resultados mais confiáveis e de um modelo com boa capacidade de generalização para novos dados, bem como para a economização do tempo de treino e recursos computacionais.

4.3.2 Parâmetros de Treino

A seleção dos parâmetros de treino num modelo *Machine Learning* (ML) é um dos passos mais críticos para o seu sucesso e eficácia. Estes parâmetros determinam como o algoritmo aprende através dos dados, influenciando diretamente o seu desempenho e a precisão final. Os ajustes inadequados destes parâmetros podem levar a problemas como *overfitting* ou *underfitting*, onde o modelo não captura bem os padrões existentes entre os dados. Desta forma, a escolha dos parâmetros de treino torna-se fundamental para desenvolver modelos robustos que generalizem bem para novos dados, garantindo assim previsões precisas e confiáveis, além de poderem otimizar o tempo de processamento e os recursos computacionais necessários, tornando assim o desenvolvimento do modelo mais eficiente. De modo a otimizar o desempenho e a precisão dos modelos desenvolvidos, foi realizada uma busca dos melhores hiperparâmetros antes do início do treino, tendo esses sido usados, em conjunto com outros parâmetros específicos, durante a fase de treino.

4.3.2.1 Busca dos melhores hiperparâmetros

A busca por hiperparâmetros ideais é uma componente fundamental no desenvolvimento e implementação de modelos de ML. Hiperparâmetros são parâmetros externos ao modelo que não são atualizados durante o processo de treino, mas que têm um impacto significativo no desempenho do modelo, como, por exemplo, o número de unidades em camadas densas de redes neurais, o número de camadas, a taxa de aprendizagem do otimizador, entre outros. A escolha adequada desses parâmetros é fundamental para obter um modelo com elevado desempenho [61]. No presente trabalho, esta busca foi efetuada de modo a maximizar a *accuracy* de validação, através dos hiperparâmetros *hp_units* e *hp_learning_rate*, que definem, respetivamente, o número de unidades em cada camada densa e a taxa de aprendizagem para o otimizador *Adam*. Além disso, foi também considerada a *patience* do *Early Stopping*, configurada para 5 épocas, com o objetivo de evitar o *overfitting*, interrompendo o treino quando a melhoria da *accuracy* de validação estabiliza.

No fim desta pesquisa, obteve-se um valor de 160 *hp_units*, uma taxa de aprendizagem de 0.001 (*hp_learning_rate*) e uma *patience* de 5 para o *Early Stopping*.

4.3.2.2 Otimizador *Adam*

A escolha do otimizador *Adam* prende-se com o facto de este apresentar várias vantagens que o tornam uma escolha popular em problemas de otimização, como no treino de modelos de ML. Este otimizador tem a capacidade de ajustar automaticamente a taxa de atualização dos parâmetros durante o treino, combinando métodos como o do gradiente estocástico e o método de momento, o que ajuda a acelerar a convergência do modelo durante o treino. Além disso, é também capaz de lidar com conjuntos de dados grandes e complexos, pelo que se torna adequado para lidar com o problema de deteção de *Alzheimer* através de imagens *MRI* [62]. As fórmulas utilizadas por este otimizador baseadas no trabalho de *Kingma* e *Ba* [63].

No presente trabalho, os parâmetros do Adam foram configurados com os valores padrão,

beta_1 = 0.9 e *beta_2* = 0.999, além da taxa de aprendizagem (*learning rate*) ajustada para 0.001.

4.4 Fase Experimental 1

Numa fase inicial, foram desenvolvidos, para o efeito pretendido, três modelos de ML, utilizando as arquiteturas *Visual Geometry Group* de 16 camadas (VGG-16), *Visual Geometry Group* de 19 camadas (VGG-19) e *Residual Network* de 50 camadas (ResNet-50). A escolha destas arquiteturas recaiu sobre o facto de estas serem bastante populares e já terem resultados comprovados na área de reconhecimento de imagens.

4.4.1 Modelo VGG-16

Primeiramente, começou-se por implementar a arquitetura VGG-16 (descrita no capítulo 2). Como referido anteriormente, foi utilizada a técnica de *Early Stopping* de modo a previnir o *overfitting* e melhorar a capacidade preditiva do modelo. Assim, o treino deste modelo foi interrompido na iteração 27.

Com o objetivo de analisar o comportamento do modelo ao longo das épocas de treino (*epochs*) e identificar possíveis casos de *overfitting*, foram gerados gráficos de *accuracy* e *loss* (calculada através da função "*sparse categorical crossentropy*") nos conjuntos de treino e de validação. Esta escolha gráfica fornece uma visão abrangente do processo de aprendizagem do modelo e permite uma análise aprofundada do seu comportamento ao longo do tempo.

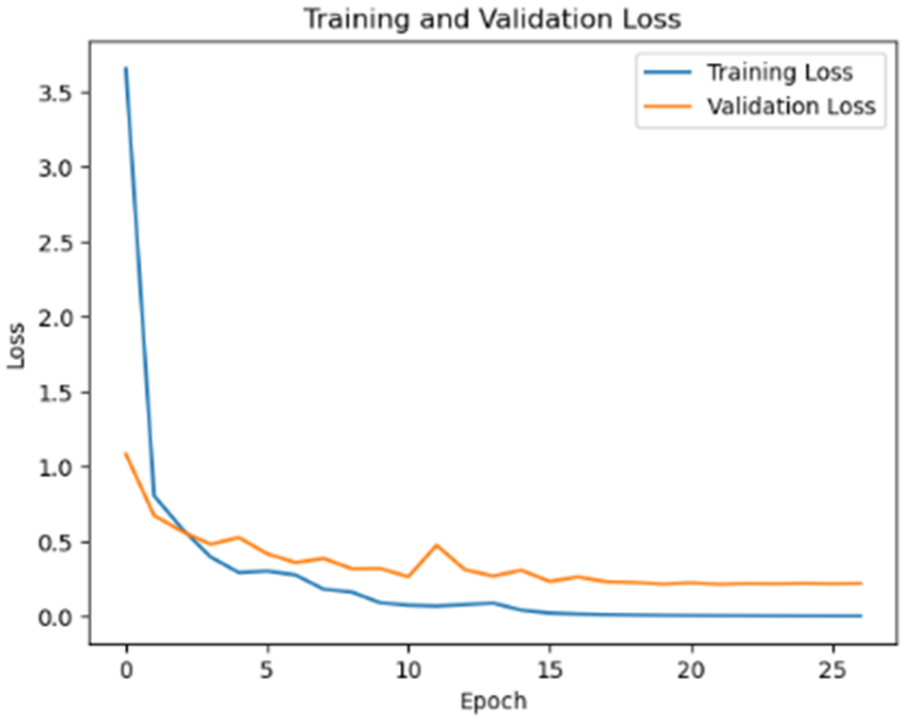
No conjunto de Figuras 4.3 encontram-se representadas, de forma gráfica, as métricas *accuracy* e *loss* de treino e validação, ao longo das iterações do modelo.

Geralmente, uma vez que o modelo está a ser otimizado para os dados de treino, a *accuracy* obtida ao longo do treino neste conjunto tende a ser maior do que a obtida no conjunto de validação, pelo que a curva de validação, que serve como um indicador do quão bem o modelo generaliza para novos dados, se deve encontrar abaixo da curva de treino. Na Figura 4.3a podemos concluir que tal não se verifica nas primeiras épocas do treino. No entanto, a situação regulariza-se antes da *epoch* 5, pelo que isso não constitui um fator de preocupação, uma vez que se trata apenas de uma flutuação que pode ser devida à aleatoriedade nos dados de validação. Além disso, verificamos também que a *accuracy* em ambos os conjuntos aumentou ao longo das *epoch*, o que indica que o modelo está a conseguir aprender com os dados, e que a *accuracy* começou a estabilizar nas últimas épocas de treino, indicando convergência.

No que diz respeito às curvas de *loss*, estas devem ser decrescentes ao longo das *epochs*, sendo isso indicativo de que o modelo está a aprender a ajudar-se aos dados. Na Figura 4.3b podemos verificar que as curvas apresentam esse comportamento decrescente ao longo das épocas de treino, começando a estabilizar a partir da *epoch* 20, o que indica que o modelo se está a aproximar de um mínimo local da função de perda.

Pela análise de ambas as curvas, verificamos que não existem sinais evidentes de *overfitting*,

(a) *Accuracy* de treino e validação ao longo das *epochs*.



(b) *Loss* de treino e validação ao longo das *epochs*.

Figure 4.3: Resultados de treino e validação do modelo VGG-16.

Modelo	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Cohen's Kappa*	CAA
VGG-16	90.19 %	90.00%	90.00%	90.00%	85.00%	85.00%

Table 4.2: Resultados das métricas obtidas com o modelo VGG-16.

o que é um bom sinal e indica que o modelo não necessita de técnicas mais complexas para o evitar.

Para avaliar o desempenho geral do modelo e a sua capacidade de generalização para os dados de teste, foram calculadas algumas métricas adicionais, como *precision*, *recall*, *f1-score*, *Cohen's Kappa* e *Chance Adjusted Accuracy* (CAA), explicadas no Capítulo 2. Os resultados obtidos para estas métricas encontram-se representados na Tabela 4.2.

Analisando estes resultados, verificamos que os valores das métricas *Cohen´s Kappa* e CAA se apresentam ligeiramente mais baixos que os restantes, o que seria de esperar, uma vez que estas têm em consideração o não balanceamento das classes. No entanto, atendendo aos intervalos de interpretação referidos no Capítulo 2, ambas as métricas indicam uma concordância quase perfeita entre as previsões do modelo e as classes reais, sugerindo que o desempenho do modelo é significativamente melhor do que o esperado pelo acaso, o que é especialmente relevante em conjuntos de dados não balanceados, onde a *accuracy* por si só pode não ser uma métrica confiável. No que diz respeito às métricas de *precision*, *recall* e *f1-score*, estas apresentam valores consistentes de 90%, o que indica uma boa capacidade do modelo em identificar corretamente as classes e minimizar os erros de classificação. Desta forma, os resultados obtidos sugerem que o modelo VGG-16 está bem ajustado para a tarefa de classificação, com um desempenho consistente e confiável, mesmo face ao não balanceamento das classes.

Também para avaliar o desempenho do modelo, foi elaborada uma matriz de confusão, após as previsões do modelo. Esta matriz oferece uma representação visual da performance do modelo, pois organiza as suas previsões em relação aos rótulos reais. Na Figura 4.4, encontra-se representada a matriz de confusão obtida.

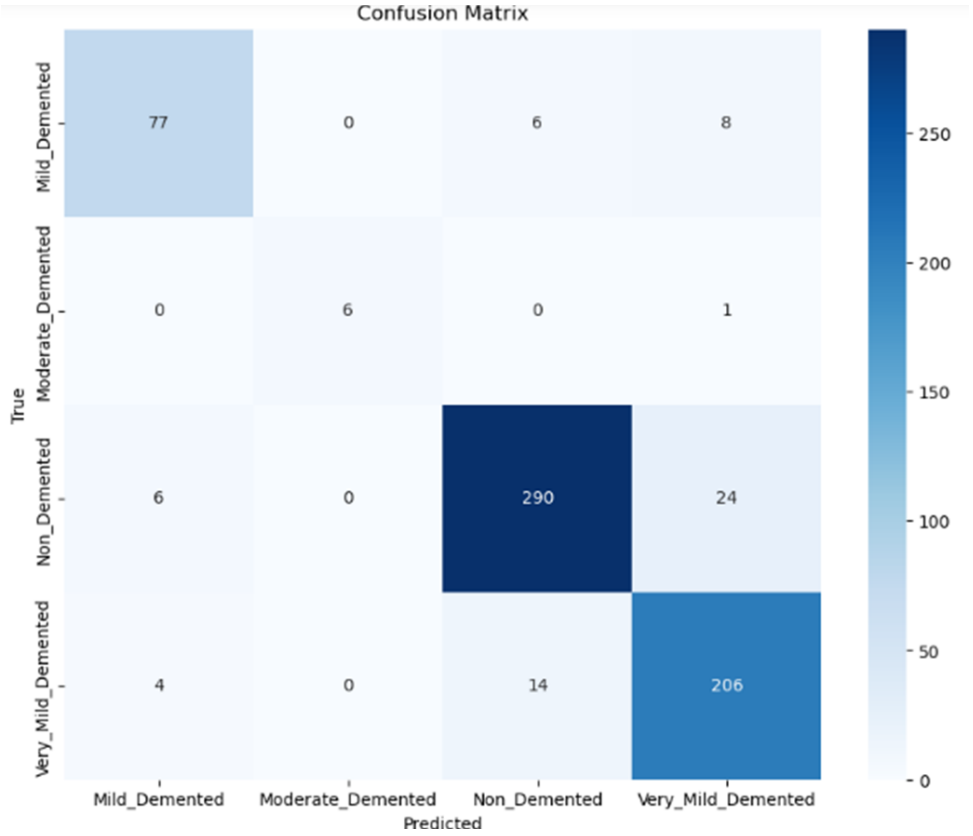


Figure 4.4: Matriz de confusão obtida para as previsões do modelo VGG-16.

Como referido no Capítulo 2, a diagononal principal da matriz representa o número de elementos de uma determinada classe que foram classificados como pertencendo a essa mesma classe, ou seja, através dela podemos verificar o número de dados corretamente classificados. Através da análise da matriz obtida (Figura 4.4), podemos verificar que modelo identificou corretamente 77 imagens da classe *Mild Demented*, 6 imagens da classe *Moderate Demented*, 290 imagens da classe *Non Demented* e 206 imagens da classe *Very Mild Demented*. Ao analisar as restantes células da matriz, verificamos que exitiram 10 casos em que modelo classificou as imagens, erradamente, como *Mild Demented*, 20 casos como *Non Demented* e 33 casos como *Very Mild Demented*. Das imagens que realmente pertenciam a *Mild Demented*, 6 foram classificadas de forma errada como *Non Demented* e 8 como *Very Mild Demented*. Em contrapartida, das imagens que pertenciam à classe *Moderate Demented*, apenas uma foi mal classificada, como pertencendo a *Very Mild Demented*. No que diz respeito à classe *Non Demented*, foram classificadas, erradamente, 6 imagens como pertencentes a *Mild Demented* e 24 como *Very Mild Demented*. Por fim, da classe *Very Mild Demented*, 4 imagens foram mal classificadas como *Mild Demented* e 14 como *Non Demented*.

Através desta análise, verificamos que as maiores taxas de erro se encontram nas classes minioritárias (*Mild Demented* e *Moderate Demented*), em que o modelo as classifica como pertencentes a uma das classes maioritárias. Isto pode ser um fator indicativo de que o modelo apresenta uma certa tendência para as classes maioritárias, o que era expectável uma vez que as classes se apresentam bastante desbalanceadas.

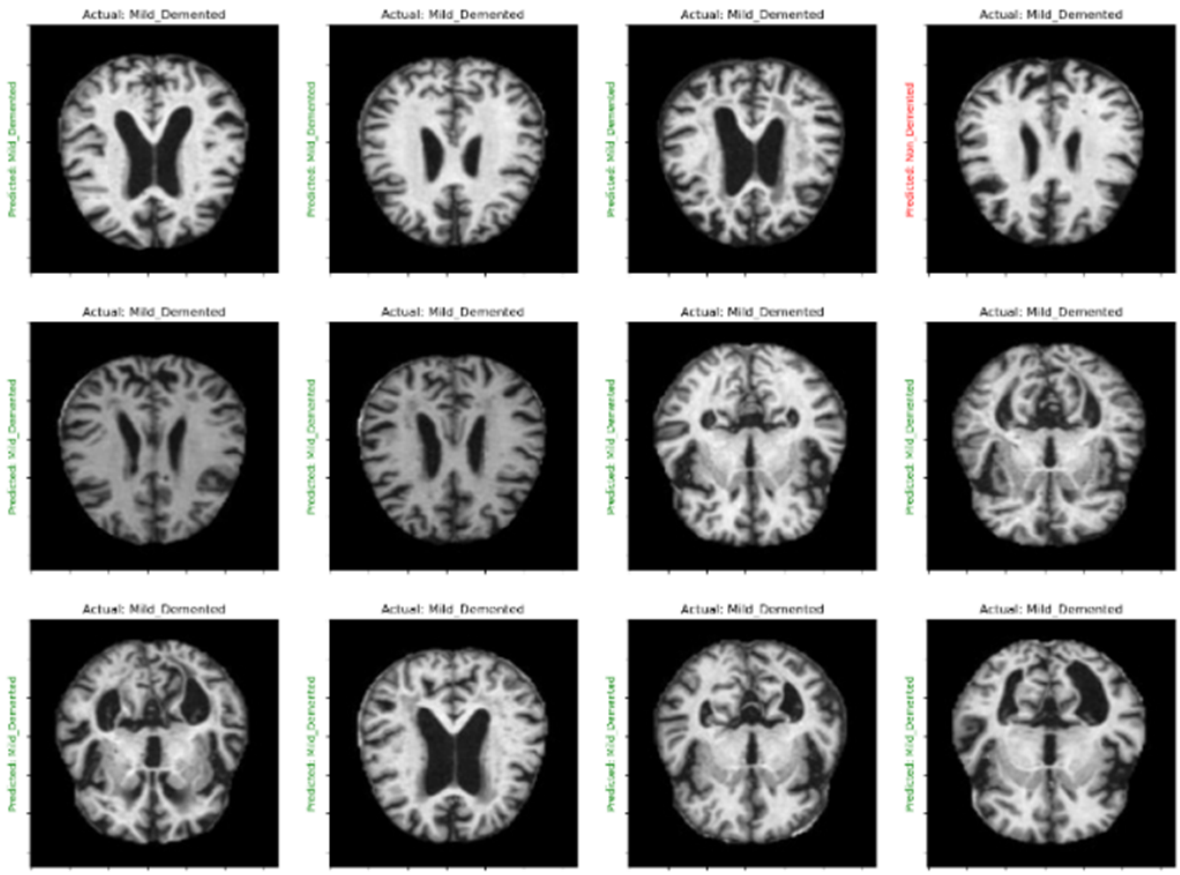


Figure 4.5: Exemplos de previsões efetuadas com o modelo VGG-16.

Após a avaliação do desempenho do modelo, efetuaram-se as previsões, estando alguns exemplos ilustrados na Figura 4.5.

De modo a aumentar a confiança nas previsões do modelo, para que os especialistas o possam aplicar para auxílio do diagnóstico da doença de *Alzheimer*, recorreu-se ao *Local Interpretable Model-agnostic Explanations* (LIME), referido no Capítulo 2, para tonar o modelo explicável. Esta ferramenta permite marcar os contornos das regiões consideradas mais importantes para a previsão do modelo. Na Figura 4.6 ilustra-se a aplicação desta técnica a uma das imagens previstas pelo modelo.

4.4.2 Modelo VGG-19

Após a aplicação da arquitetura VGG-16, foi aplicada, para efeitos de comparação, a arquitetura VGG-19. Como referido no 2, esta arquitetura é bastante semelhante à anterior, apresentando apensas um maior número de camadas de convolução, sendo também bastante popular e tendo resultados comprovados na área do reconhecimento de imagens.

O processo de treino foi semelhante ao anterior, tendo sido aplicada a técnica de *Early Stopping* para previnir o *overfitting*, pelo que o treino do modelo foi interrompido na *epoch* 23.

Para avaliar o desempenho do modelo, foram geradas as curvas de aprendizagem (curva de

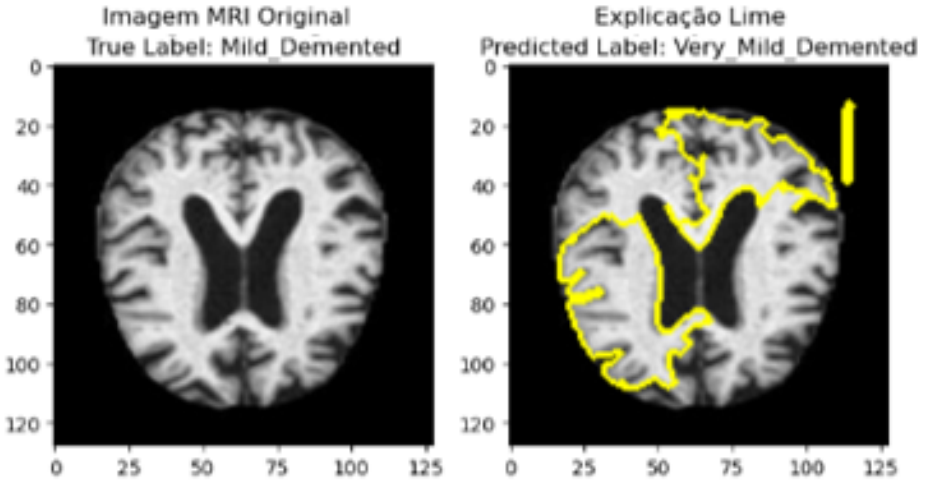


Figure 4.6: Aplicação de LIME a uma previsão do modelo VGG-16.

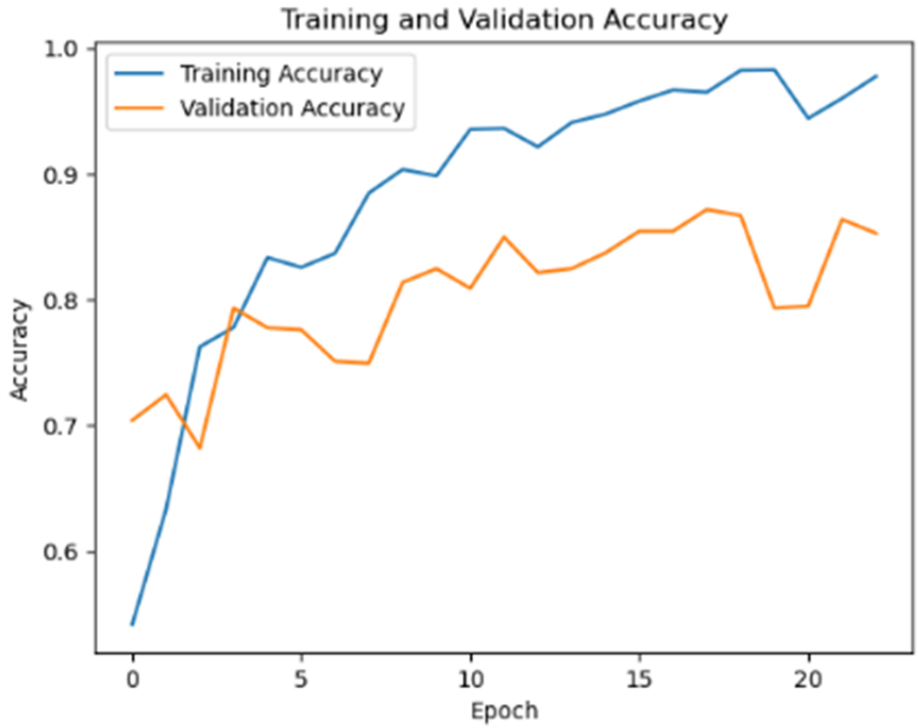
accuracy e curva de *loss*), representadas no conjunto de Figuras 4.7.

Analisando estas curvas, verificamos que as curvas de *accuracy* apresentam um comportamento maioritariamente crescente e as de *loss* um comportamento descrescente, como era expectável. No entanto, verificamos também que este comportamento não é constante ao longo do tempo, principalmente nas curvas que dizem respeito aos dados de validação, uma vez que apresentam um comportamento irregular de subida e descida, o que pode ser um sinal de *overfitting*. Além disso, verificamos também que a curva *loss* dos dados de validação não estabilizou e apresentou uma subida no final das épocas de treino, o que indica a presença de *overfitting*. Assim sendo, conclui-se que este modelo necessitaria de outras técnicas de prevenção de *overfitting* para poder ser utilizado. No entanto, como já se tinha obtido um modelo, aparentemente eficaz com a arquitetura anterior e ainda estava planeada a implementação de uma arquitetura diferente, não foram aplicadas este tipo de técnicas.

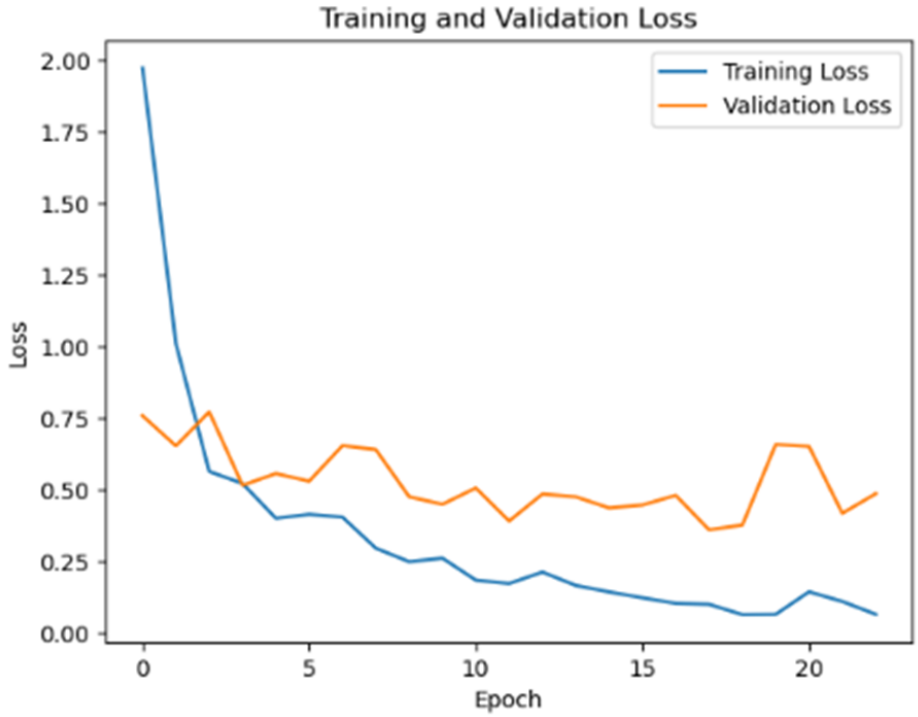
Apesar de o modelo apresentar *overfitting*, o seu desempenho foi também avaliado através das métricas *precision*, *recall* e *f1-score*,*Cohen's Kappa* e CAA, bem como pela matriz de confusão. Na Tabela 4.3 e na Figura 4.8, encontram-se, respetivamente, os valores destas métricas e a matriz de confusão resultantes.

Modelo	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Cohen's Kappa*	CAA
VGG-19	86.45 %	87.00%	86.00%	86.00%	76.00%	76.00%

Table 4.3: Resultados das métricas obtidas com o modelo VGG-19.



(a) *Accuracy* de treino e validação ao longo das *epochs*.



(b) *Loss* de treino e validação ao longo das *epochs*.

Figure 4.7: Resultados de treino e validação do modelo VGG-19.

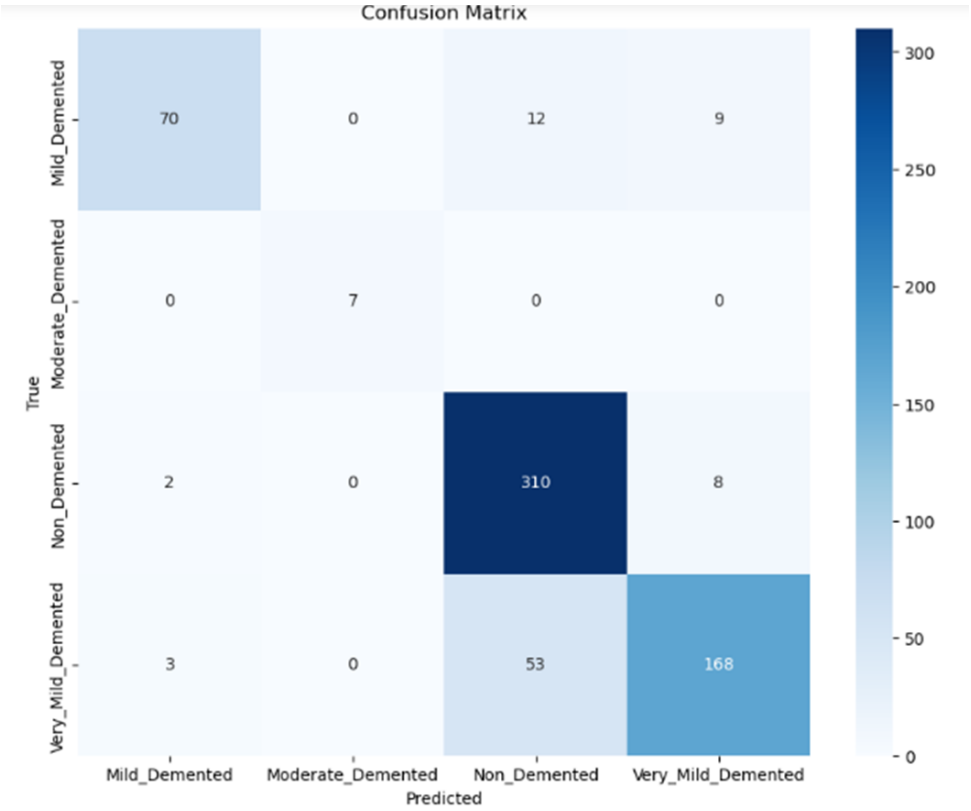


Figure 4.8: Matriz de confusão obtida para as previsões do Modelo VGG-19.

Através dos valores obtidos para as métricas *Cohen's Kappa* e CAA, verificamos que o modelo apresenta uma concordância substancial entre as previsões efetuadas e as classes reais, o que sugere que o modelo é eficaz na classificação. No que diz respeito às restantes métricas, estas indicam um desempenho satisfatório do modelo.

Analisando a matriz de confusão obtida, verificamos que 70 imagens foram corretamente identificadas como pertencentes à classe *Mild Demented*, 7 a *Moderate Demented*, 310 a *Non Demented* e 168 como pertencentes a *Very Mild Demented*. Efetuando uma análise para as restantes células semelhante à efetuada para a matriz de confusão obtida na arquitetura anterior, verificamos que, apesar de o modelo apresentar indícios de *overfitting*, as taxas de erro nas previsões de teste são, de um modo geral, menores do que as anteriores, mesmo nas classes minoritárias, o que indica uma melhor generalização das classes. No entanto, o maior equilíbrio na matriz de confusão pode ocorrer devido ao facto de modelo elaborado com a arquitetura *VGG-16* apresentar os erros mais distribuídos por todas as classes, o que resulta em melhores métricas globais. Em contrapartida, o modelo VGG-19 comete os erros em classes mais "fáceis" de prever (classes maioritárias), como podemos verificar pelo aumento significativo da taxa de erro na classe *Very Mild Demented*, o que é consistente com a presença de *overfitting*, uma vez que o modelo poderá estar sobreajustado a outras classes ou a padrões específicos nos dados de treino.

Após esta análise, foram também efetuadas as previsões e a aplicação do LIME, que se encontram nas Figuras 4.9 e 4.10, respetivamente.

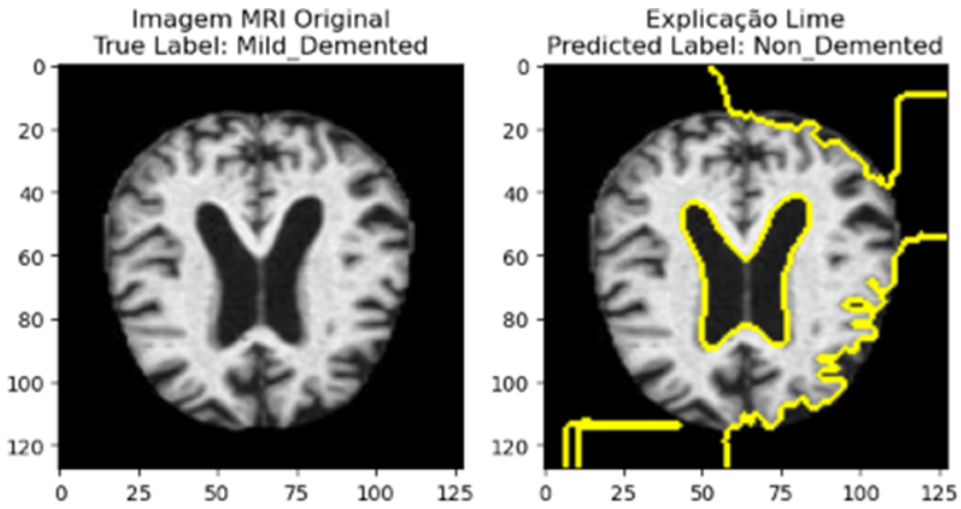


Figure 4.10: Aplicação de LIME a uma das previsões do modelo VGG-19.

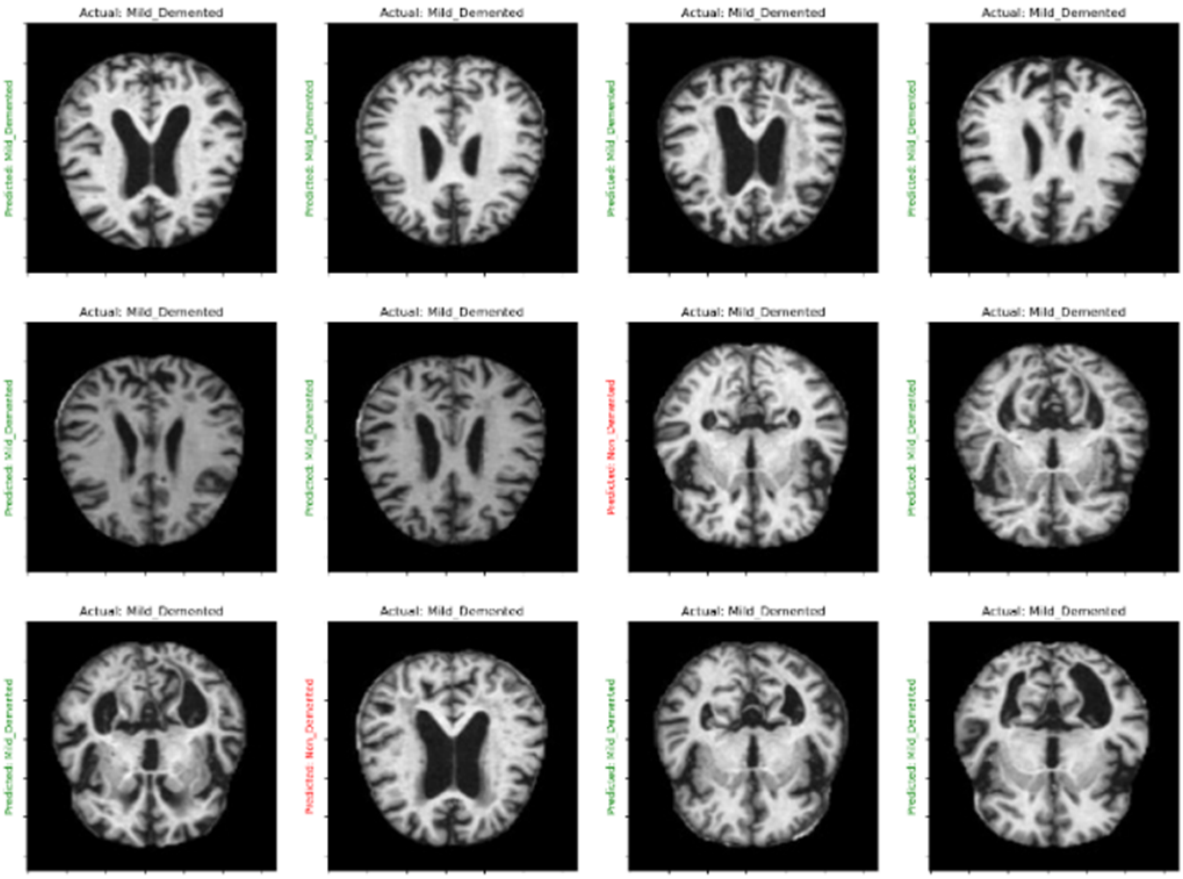


Figure 4.9: Exemplos de previsões efetuadas com o modelo VGG-19.

4.4.3 Modelo ResNet-50

Após a implementação das duas principais variações do modelo *Visual Geometry Group* (VGG), foi também implementada, nos mesmos moldes, a arquitetura ResNet-50. Durante a

fase de treino do modelo com esta arquitetura, ocorreu o *Early Stoppping* na *epoch* 35, tendo sido obtidas as curvas de aprendizagem presentes no conjunto de Figuras 4.11.

Analisando estes gráficos, verificamos novamente que as curvas de *accuracy* apresentam um comportamento crescente ao longo das *epochs*, mantendo-se a curva de validação abaixo da curva de treino. No que diz respeito às curvas de *loss*, estas apresentam um comportamento decrescente ao longo das épocas de treino e é possível observar que, nas últimas iterações e após a estabilização das curvas, a *loss* de validação começou a sofrer um ligeiro aumento, antes do *Early Stopping*, o que indica que, se o treino continuasse, o modelo iria apresentar sinais de *overfitting*, pelo que a sua interrupção permitiu evitar essa situação.

Após esta análise, foram calculadas as métricas adicionais para avaliar o desempenho do modelo e a sua capacidade de generalização para novos dados. Os resultados obtidos encontram-se na Tabela 4.4. Estes resultados sugerem um bom desempenho do modelo e uma concordância quase perfeita entre as previsões do modelo e as classes reais.

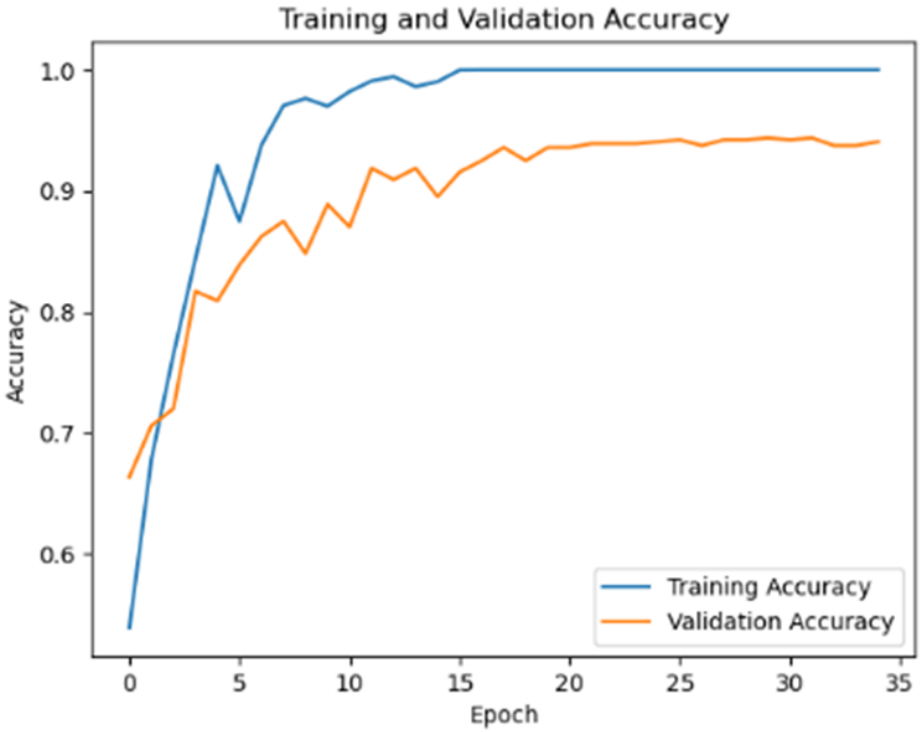
Modelo	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Cohen's Kappa*	CAA
ResNet-50	94.30 %	94.00%	94.00%	94.00%	91.00%	91.00%

Table 4.4: Resultados das métricas obtidas com o modelo ResNet-50.

Após as previsões e também para avaliar o desempenho geral do modelo e a sua capacidade de generalização para novos dados, foi elaborada a matriz de confusão, presente na Figura 4.12.

Analisando esta matriz, verificamos que o modelo acertou nas previsões de 77 imagens da classe *Mild Demented*, 6 imagens da classe *Moderate Demented0*, 312 imagens da classe *Non Demented* e 204 imagens da classe *Very Mild Demented*. No que diz respeito às restantes células, podemos concluir que as taxas de erro associadas a cada classe diminuíram face às obtidas através das arquiteturas VGG-16 e VGG-19. Desta forma, combinando a análise da matriz com os resultados das métricas de desempenho obtidas, podemos concluir que o modelo ResNet-50 se demonstra mais eficaz na deteção de *Alzheimer*, nas suas diversas fases.

À semelhança do realizado nas arquiteturas anteriores, após esta avaliação visualizaram-se algumas previsões (Figura 4.13) e aplicou-se *LIME* a uma delas (Figura 4.14), de modo a perceber como o modelo tomou a decisão de a classificar numa determinada classe.



(a) *Accuracy* de treino e validação ao longo das *epochs*.

(b) *Loss* de treino e validação ao longo das *epochs*.

Figure 4.11: Resultados de treino e validação do modelo ResNet-50.

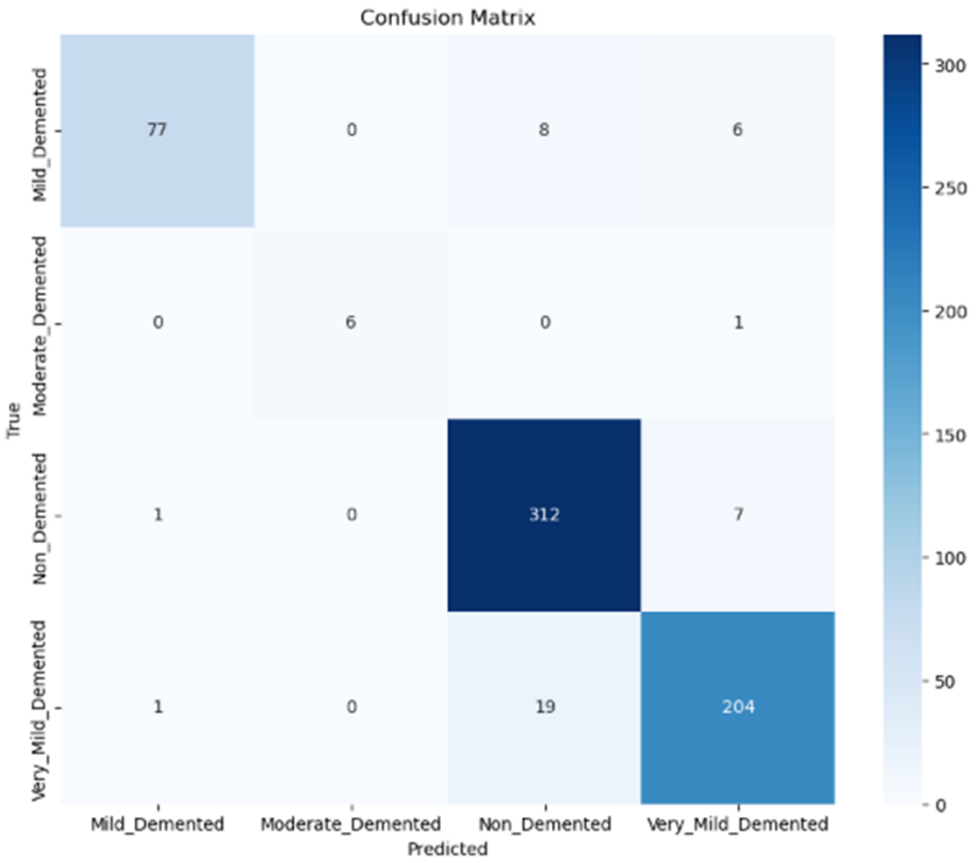


Figure 4.12: Matriz de confusão obtida para as previsões do Modelo ResNet-50.

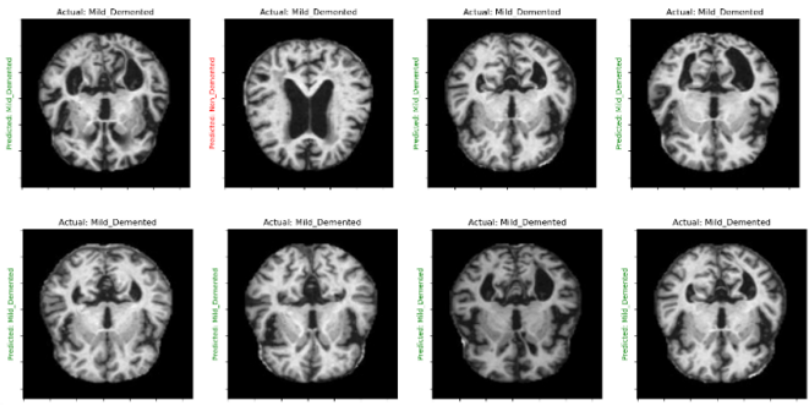


Figure 4.13: Exemplos de previsões efetuadas com o modelo ResNet-50.

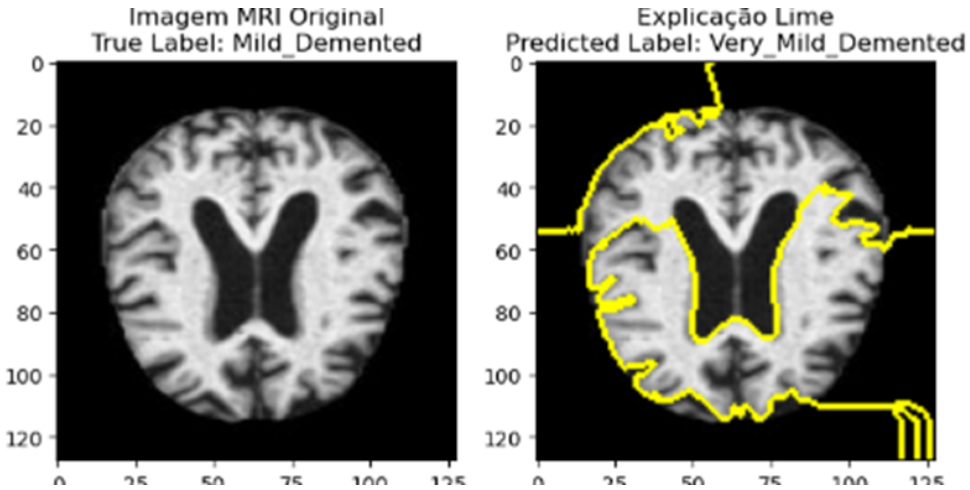


Figure 4.14: Aplicação de LIME a uma das previsões do modelo ResNet-50.

4.5 Fase Experimental 2

Apesar de os modelos acima apresentados apresentarem resultados satisfatórios, o *dataset* utilizado apresenta-se desbalanceado, o que pode afetar negativamente o desempenho destes, levando a métricas enganosamente boas em classes maioritárias, enquanto que as minoritárias apresentam taxas de erro mais elevadas. Para combater este possível problema, foi aplicada uma técnica de *data augmentation* nas classes minoritárias. Após este processo, o novo conjunto de dados foi utilizado para treinar o modelo de uma das arquiteturas.

4.5.1 *Oversampling*

Como referido no Capítulo 2, *data augmentation* por *oversampling* é uma técnica que permite aumentar, de forma artificial, o tamanho de um *dataset*, através da criação de novas amostras, derivadas das existentes. A implementação desta técnica permite fornecer ao modelo, durante o treino, uma maior variedade de dados da classe minoritária, de modo a que este possa aprender melhor os padrões entre estes dados e melhorar a sua capacidade de generalização para

novos dados, ao mesmo tempo que ajuda a mitigar problemas de *overfitting*.

Para aplicar a técnica de data augmentation, foram utilizados dois métodos principais: a função ***ImageDataGenerator*** do *Keras* e as camadas de pré-processamento de *data augmentation* do *TensorFlow*. Primeiramente, o ***ImageDataGenerator*** foi configurado com as seguintes transformações:

- ***width_shift_range***: Deslocamento horizontal aleatório das imagens em até 10% da largura total;
- ***height_shift_range***: Deslocamento vertical aleatório das imagens em até 10% da altura total;
- ***shear_range***: Aplicação de uma transformação de cisalhamento com um ângulo de até 20 graus;
- ***zoom_range***: Zoom aleatório nas imagens em até 20%;
- ***horizontal_flip***: Espelhamento horizontal aleatório das imagens;
- ***vertical_flip***: Espelhamento vertical aleatório das imagens;
- ***fill_mode***: Método de preenchimento dos pixels vazios após as transformações, definido como 'nearest'.

Além disso, foram utilizadas camadas de data augmentation de *TensorFlow*, implementadas através de um *tf.keras.Sequential*, com as seguintes transformações:

- **RandomFlip**: Espelhamento horizontal aleatório das imagens;
- **RandomZoom**: Zoom aleatório nas imagens em até 10%.

Estas transformações são aplicadas de forma aleatória a cada imagem durante o processo de *data augmentation*, aumentando assim a diversidade do conjunto de dados e ajudando a prevenir o *overfitting*.

4.5.2 Modelo ResNet-50 com *Data Augmentation*

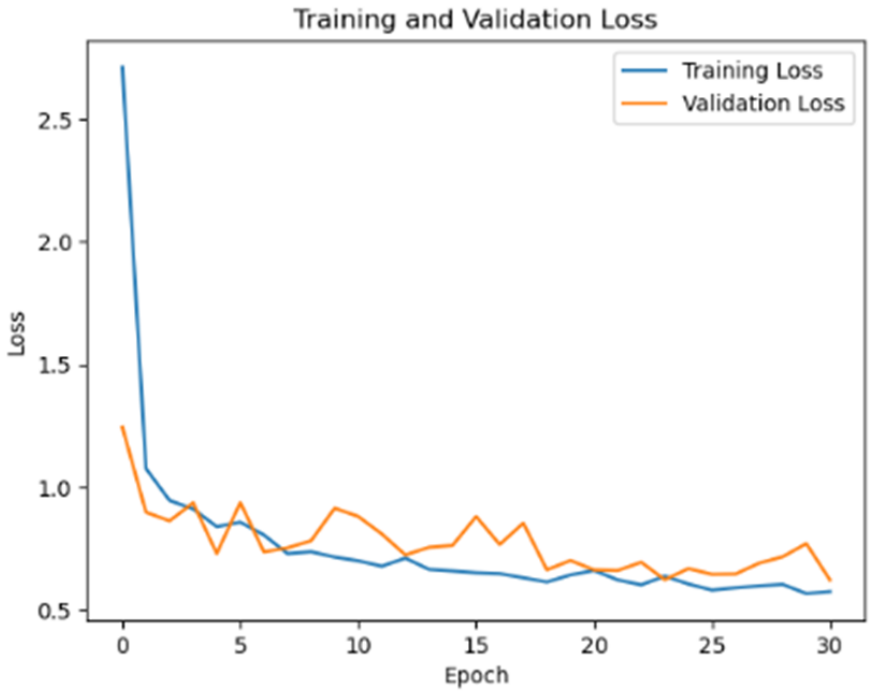
A arquitetura escolhida para efetuar esta comparação foi a ResNet-50, uma vez foi a que apresentou melhores resultados na fase anterior.

O processo de treino foi semelhante ao efetuado para os modelos anteriores, tendo também sido aplicado *Early Stopping* para prevenir o *overfitting*, pelo que o treino do modelo foi interrompido na *epoch* 31.

Para avaliar o comportamento do modelo ao longo das *epochs*, foram geradas as curvas de aprendizagem representadas no conjunto de Figuras 4.15.

Através da análise destas curvas, verificamos que estas apresentaram um comportamento bastante oscilatório ao longo das épocas de treino, o que pode ser devido à maior variedade nas

(a) *Accuracy* de treino e validação ao longo das *epochs*.



(b) *Loss* de treino e validação ao longo das *epochs*.

Figure 4.15: Resultados de treino e validação do modelo ResNet-50 com *Data Augmentation*.

amostras, causada pelo *data augmentation*. Apesar dessas oscilações, verificamos que o comportamento geral das curvas vai de encontro ao esperado, isto é, um comportamento crescente na curva de *accuracy* e decrescente na curva de *loss*.

Após esta análise, foram calculadas as métricas adicionais *precision*, *recall*, *f1-score*, *Cohen's Kappa* e CAA, de modo a avaliar a *performance* global do modelo e a sua capacidade de generalização para novos dados. Os resultados obtidos encontram-se representados na Tabela 4.5.

Modelo	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Cohen's Kappa*	CAA
ResNet-50 com *Data Augmentation*	71.96 %	72.00%	71.00%	71.00%	51.00%	53.00%

Table 4.5: Resultados das métricas obtidas com o modelo ResNet-50 com *Data Augmentation*.

Comparando estes resultados com os obtidos anteriormente para o modelo ResNet-50 sem *data augmentation*, verificamos que os valores das métricas de desempenho sofreram uma grande descida, com uma diminuição acentuada da concordância entre as previsões do modelo e as classes reais, que indica que este modelo é apenas ligeiramente melhor do que o esperado pelo acaso. Esta diminuição pode ser devida ao aumento da variabilidade dos dados, que pode ter causado ao modelo um aumento da dificuldade em aprender padrões existentes entre eles. Além disso, a natureza artificial das amostras geradas pode ter introduzido padrões irreais ao *dataset*, o que não contribui de forma positiva para a aprendizagem do modelo. Neste sentido, seria necessária a existência de um *dataset* real e balanceado para tornar possível uma mellhor aprendizagem.

Apesar de os resultados obtidos através deste método terem sido piores face aos anteriores, foram ainda realizadas as previsões com este modelo (Figura 4.16) e traçada a matriz de confusão, presente na Figura 4.17. Pela observação desta matriz, verificamos que o modelo acertou nas previsões de 50 imagens da classe *Mild Demented* (das 91 que tinha desta classe classificar), em 7 imagens da classe *Moderate Demented* (em 7 imagens que tinha para classificar), 273 imagens da classe *Non Demented* (das 320 que tinha para classificar) e em 132 imagens da classe *Very Mild Demented* (em 224 que tinha desta classe para classificar). Posto isto, é notório que o modelo teve mais dificuldades em aprender os padrões entre os dados, de modo a poder generalizar de forma eficaz para novos dados.

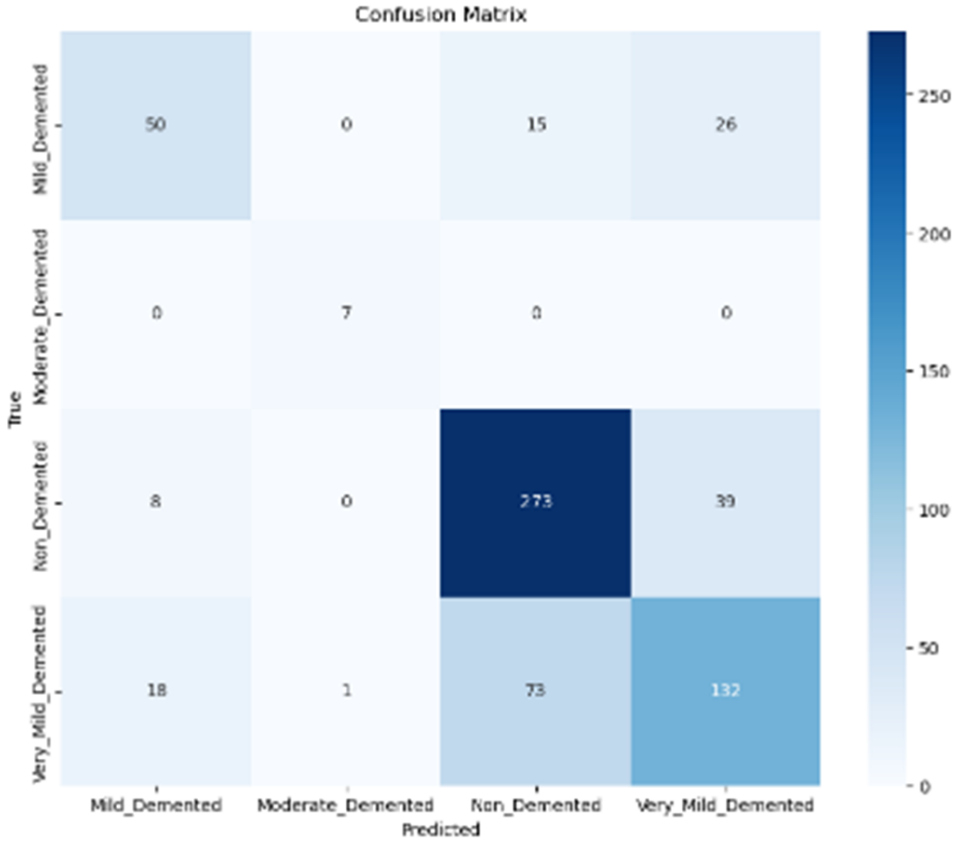


Figure 4.17: Matriz de confusão obtida para as previsões do Modelo ResNet-50 com *Data Augmentation*.

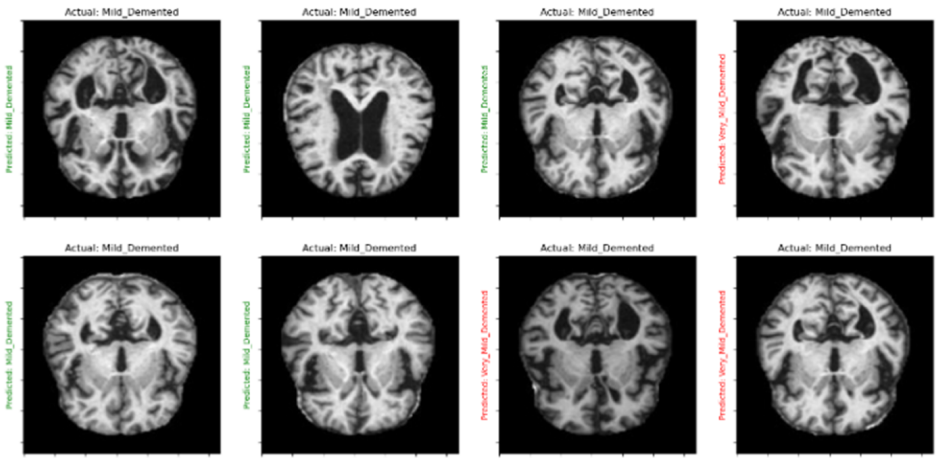


Figure 4.16: Exemplos de previsões efetuadas com o modelo ResNet-50 com *Data Augmentation*.

Para finalizar a análise efetuada para o modelo ResNet-50 com *data augmentation*, foi, à semelhança do realizado para os modelos anteriores, aplicado LIME a uma das previsões efetuadas, de modo a poder verificar com base em que regiões do cérebro o modelo tomou a decisão de classificar uma determinada imagem como pertencendo a uma determinada classe. O

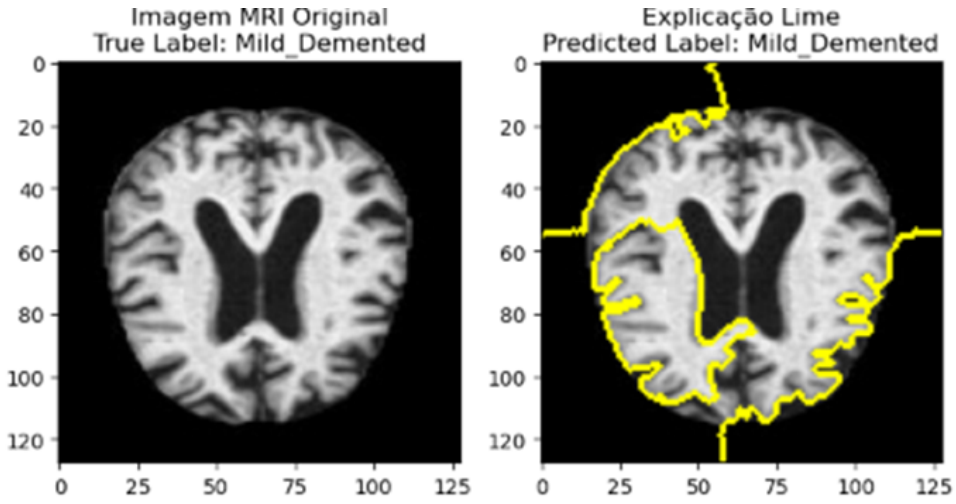


Figure 4.18: Aplicação de LIME a uma das previsões do modelo ResNet-50 com *Data Agmentation*.

resultado obtido encontra-se representado na Figura 4.18.

Embora os resultados obtidos através deste método tenham sido inferiores aos anteriores, neste trabalho, optou-se por não aplicar o conjunto de dados resultante do processo de *data augmentation* às restantes arquiteturas, devido a questões de tempo e de foco nos objetivos, uma vez que se pretendia testar o uso de AL no *dataset* original e para o caso em estudo, ao invés do impacto desta técnica.

4.6 Fase experimental 3

Após se ter verificado que a aplicação de *data augmentation* não foi favorável ao caso em estudo, por alterar a realidade dos dados, tornando assim mais difícil a aprendizagem de padrões específicos por parte do modelo, iniciou-se o estudo da aplicação de AL a uma das arquiteturas anteriormente referidas, neste caso à arquitetura ResNet-50. Como referido no Capítulo 2, AL é uma área do ML bastante utilizada quando não existe uma grande quantidade de dados disponível, uma vez que permite selecionar de maneira inteligente quais amostras devem ser anotadas e utilizadas para treinar o modelo, isto é, permite selecionar as amostras mais informativas e que apresentam um maior potencial para melhorar a performance do modelo, tornando-o mais eficiente. Esta abordagem, apesar de não melhorar diretamente a performance do modelo em comparação com modelos treinados de forma passiva (modelos ML), o AL possibilita alcançar resultados semelhantes com uma menor quantidade de dados. Esta abordagem pode mitigar as limitações do dataset original (não balanceamento das classes), melhorando assim a eficiência do processo de anotação e, consequentemente, facilitando a obtenção de resultados robustos nas métricas de desempenho e na matriz de confusão.

4.6.1 *Pytorch*

Para a realização desta fase, recorreu-se ao uso de *Pytorch*, uma biblioteca bastante conhecida e utilizada em ML pela sua flexibilidade e facilidade de utilização. Esta biblioteca apresenta uma grande capacidade de manipular os dados de forma dinâmica e eficiente, pelo que facilita a implementação de estratégias para seleção de amostras, reduzindo assim a necessidade de rotular manualmente grandes conjuntos de dados.

Pytorch suporta a criação de redes neurais complexas com uma sintaxe intuitiva, além de oferecer uma integração robusta com GPUs para acelerar o treino dos modelos. Além disso, esta biblioteca conta com uma ampla comunidade de desenvolvedores e pesquisadores, que contribuem com uma vasta gama de bibliotecas complementares e tutoriais, tornando assim a aplicação de técnicas avançadas de ML mais acessível. A biblioteca *Pytorch* possui também um forte suporte para a deteção de erros ou falhas no código, o que permite que os desenvolvedores identifiquem e resolvam os problemas de forma mais eficaz durante o desenvolvimento do modelo [64].

A utilização de *Pytorch* nesta fase do projeto facilita a implementação de AL e pode resultar em modelos mais precisos e eficientes, uma vez que permite que o modelo aprenda com menos dados rotulados, economizando assim tempo e recursos no processo de treino.

4.6.2 Modelo AL ResNet-50

O treino do modelo de AL iniciou-se com um conjunto inicial de 100 amostras rotuladas, selecionadas aleatoriamente a partir do conjunto de treino disponível, sendo que a escolha deste número visa equilibrar a representatividade inicial dos dados e aumentar a eficiência do processo de treino. Este processo foi conduzido iterativamente ao longo de várias fases onde, em cada iteração, foram selecionadas 10 amostras através de um processo conhecido como "*query strategy*", baseado na maior incerteza do modelo em relação a essas amostras. Estas amostras foram então rotuladas e utilizadas para atualizar o modelo, que passou por um novo ciclo de treino.

Durante cada fase do treino, o modelo foi sendo refinado com o conjunto de dados rotulados atualizado, permitindo-lhe assim focar-se nas áreas de maior incerteza e mais desafiantes no espaço de dados. Este ciclo de solicitação de novos dados, rotulagem e atualização contínua do modelo foi fundamental para melhorar o seu desempenho ao longo do tempo, bem como para a utilização eficiente dos recursos de rotulagem disponíveis.

Este processo foi repetido ao logo de 460 iterações, tendo o desempenho do modelo sido monitorizado, tal como se pode verificar no conjunto de Figuras 4.19.

No final do processo de AL, o modelo demonstrou uma melhoria substancial em relação ao seu desempenho inicial, destacando a eficácia do AL na otimização do uso dos recursos de rotulagem e na melhoria contínua do desempenho do modelo.

Com a utilização do AL, espera-se que o modelo obtenha, com uma menor quantidade de dados, resultados semelhantes aos obtidos para a mesma arquitetura de forma passiva. No

(a) *Accuracy* de treino e validação ao longo das iterações.

(b) *Loss* de treino e validação ao longo das *epochs*.

Figure 4.19: Resultados de treino e validação do modelo ResNet-50 com *Data Augmentation*.

entanto, apesar de o *dataset* ter sido quase completamente utilizado após as 460 iterações, verificamos que o desempenho do modelo AL ficou bastante aquém do modelo ResNet-50, que tinha sido treinado usando o mesmo dataset. Esta diferença de desempenhos pode dever-se ao facto de os modelos terem sido treinados a partir de *frameworks* diferentes (*Tensorflow* e *Pytorch*), o que pode impactar a otimização e treino do modelo. Além disso, o processo de seleção de amostras em AL pode ter levado o modelo a aprender padrões diferentes, influenciando o seu desempenho final.

Após a análise das curvas obtidas, foram calculadas, para o modelo obtido no final do processo de AL, as métricas *precision*, *recall*, *f1-score*, *Cohen's Kappa* e CAA para avaliar o desempenho do modelo no conjunto de teste e a sua capacidade de generalização para novos dados. Os resultados obtidos encontram-se na Tabela 4.6.

Modelo	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*Cohen's Kappa*	CAA
AL ResNet-50	68.38 %	68.48%	68.38%	68.36%	47.92%	57.84%

Table 4.6: Resultados das métricas obtidas com o modelo AL com ResNet-50

Através destes resultados, podemos verificar novamente que a *accuracy* do modelo AL foi ligeiramente inferior à obtida pelo modelo ResNet-50 com *data augmentation*. No que diz respeito às métricas *precision*, *recall* e *f1-score*, o valor destas também apresenta uma diminuição face aos obtidos pelos modelos anteriores, com a mesma arquitetura, o que sugere que a capacidade do modelo identificar corretamente as classes específicas pode estar comprometida, quando comparada com o modelo base (ResNet-50 sem *data augmentation*). Quanto ao valor da métrica *Cohen's Kappa*, que avalia a concordância entre as previsões do modelo e os rótulos reais, observamos uma concordância razoável, dentro dos intervalos apresentados no Capítulo 2. Por fim, o valor da métrica CAA também indica uma concordância moderada entre os rótulos reais e as previsões do modelo. No entanto, é relevante notar que, apesar dos resultados inferiores observados com o modelo AL, houve um leve aumento na CAA em comparação com o modelo ResNet-50 com data augmentation, o que pode sugerir que, embora os resultados gerais sejam menos favoráveis, as previsões do modelo AL podem ser mais robustas.

Após a obtenção dos valores das métricas de desempenho, foi também realizado um estudo detalhado da matriz de confusão. A análise desta matriz fornece *insights* adicionais sobre como o modelo quais as classes que podem estar mais propensas a erros de previsão. A matriz obtida encontra-se na Figura 4.20.

Pela diagonal principal desta matriz, podemos verificar que o modelo acertou nas previsões de 48 imagens da classe *Mild Demented* (das 91) que tinha para classificar, 4 imagens da classe *Moderate Demented* (em 7 que tinha para classificar), 245 imagens da classe *Non Demented* (das 320 que tinha para classificar) e 142 imagens da classe *Very Mild Demented* (das 142 que tinha para classificar). Através destes resultados e analisando as restantes células da matriz, onde verificamos um aumento nas taxas de erro das previsões, podemos confirmar a diminuição

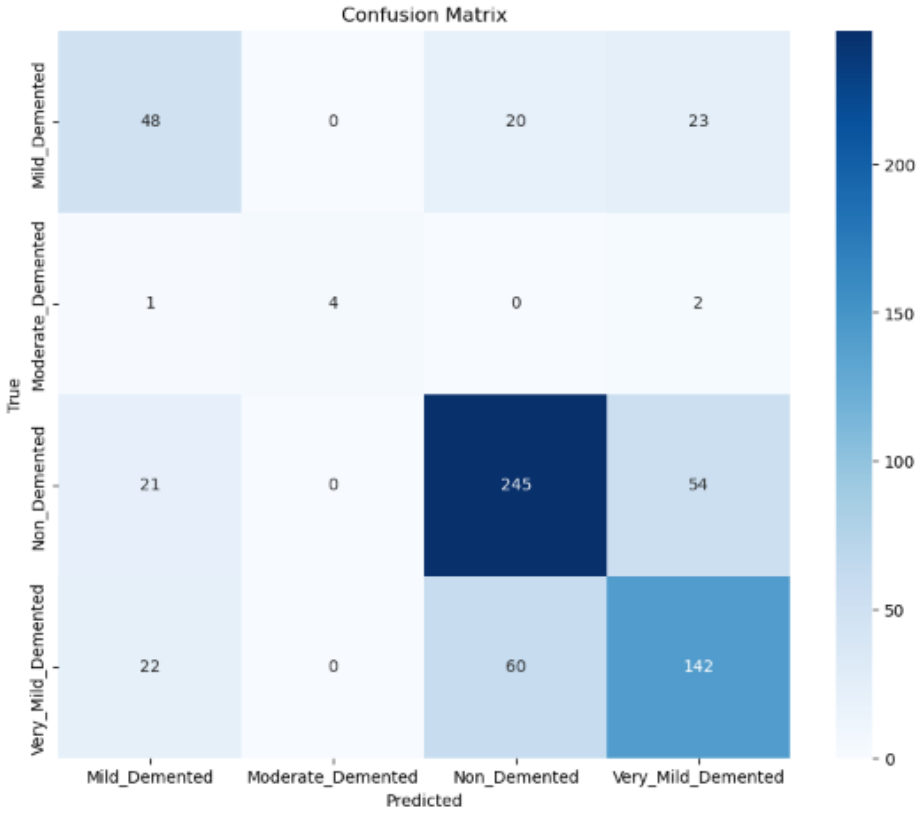


Figure 4.20: Matriz de confusão obtida para as previsões do Modelo AL

do desempenho geral do modelo e da sua capacidade de generalização para novas amostras, o que indica que o modelo AL necessitaria de ajustes futuros para melhorar o seu desempenho, de modo a permitir a sua aplicação no mundo real.

Para concluir a análise dos resultados obtidos com este modelo, foram observadas algumas previsões efetuadas pelo modelo e aplicada a técnica de LIME para as tornar explicáveis, tal como se pode verificar nas Figuras 4.21 e 4.22, respetivamente.

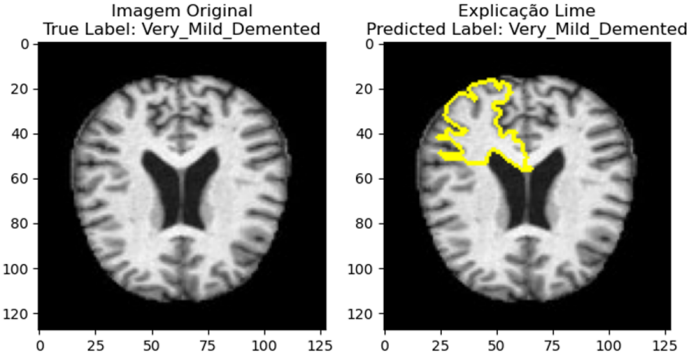


Figure 4.22: Aplicação de LIME a uma das previsões do modelo AL

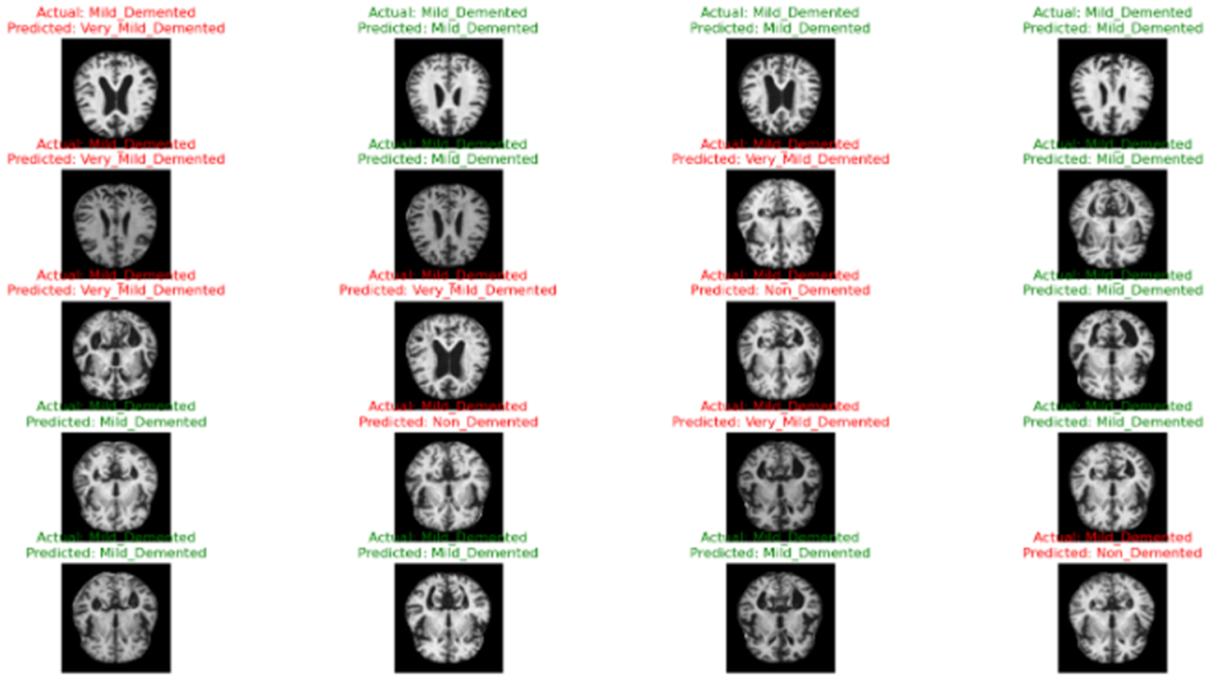


Figure 4.21: Exemplos de previsões efetuadas com o modelo AL.

4.7 Fase experimental 4

A fase final deste trabalho consistiu na escolha aleatória de uma imagem do conjunto de treino e na sua previsão e aplicação de LIME, através dos vários modelos. Esta abordagem tem como objetivo verificar como os diferentes modelos afetam a interpretabilidade das previsões para a mesma imagem.

Inicialmente, foi selecionada, de forma aleatória, uma imagem do conjunto de teste. Esta imagem foi então submetida aos diferentes modelos utilizados neste estudo (VGG-16, VGG-19, ResNet-50, ResNet-50 com *data augmentation* e AL), sendo que cada modelo gerou uma previsão para a imagem selecionada.

Posteriormente, para cada modelo, foi aplicado o método LIME para gerar explicações interpretáveis sobre como o modelo chegou à sua decisão para essa imagem específica. Essas explicações são fundamentais para entender não apenas o que o modelo previu, mas também quais as características da imagem que tiveram mais influência na decisão do modelo.

Essa abordagem permitiu uma comparação direta das explicações geradas pelos diferentes modelos, destacando como as diferentes arquiteturas afetam a interpretabilidade das previsões. Os resultados obtidos encontram-se representados na Figura 4.23.

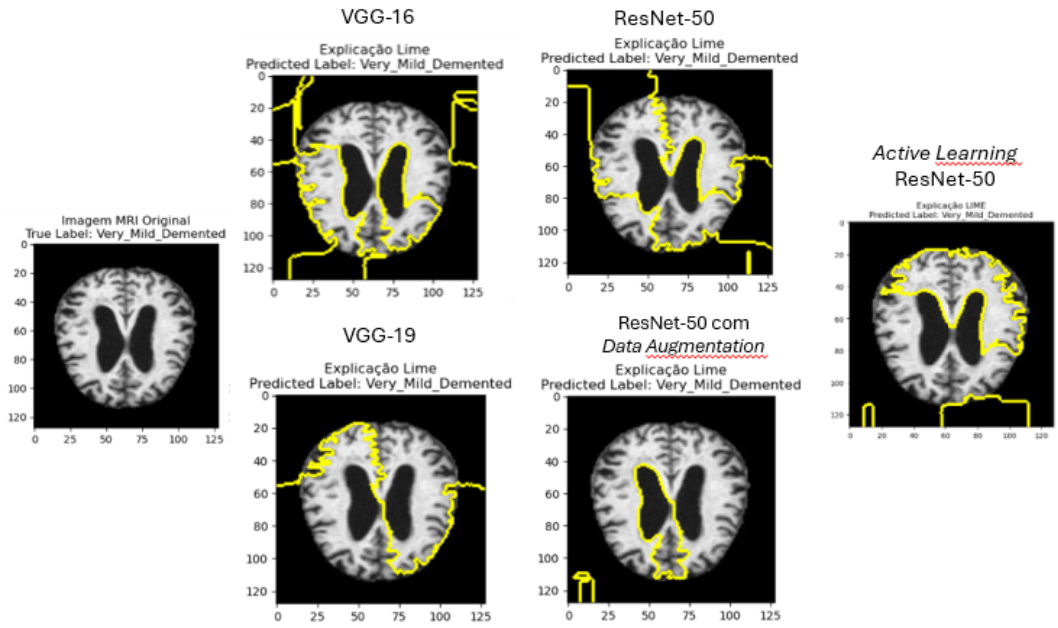


Figure 4.23: Explicações obtidas pelos diferentes modelos para a mesma imagem.

Ao analisar as explicações de cada modelo, foi possível identificar padrões de comportamento distintos, além de potenciais áreas de melhoria ou foco para futuras iterações do modelo. No entanto, é importante ressaltar que, para determinar a precisão das explicações fornecidas por cada modelo, seria necessário o envolvimento de um especialista médico para validar qual interpretação é mais apropriada ou a necessidade de ajustes nos modelos.

4.8 Resumo

Neste capítulo, foi abordada a implementação de modelos de IA para a deteção da doença de Alzheimer utilizando imagens de MRI do cérebro. Foram apresentados os materiais, métodos, resultados e discussões relacionados ao desenvolvimento desses modelos, com ênfase na importância da prevenção de *overfitting* por meio de estratégias como o *Early Stopping*. O *dataset* utilizado foi obtido da plataforma *Kaggle*, contendo 6400 imagens pré-processadas divididas em quatro classes: *Non Demented*, *Very Mild Demented*, *Mild Demented* e *Moderate Demented*.

O desempenho dos modelos foi avaliado através de métricas como *precision*, *recall*, *f1-score*, *Cohen's Kappa* , CAA e matriz de confusão, tendo estes demonstrado uma concordância substancial entre as previsões e as classes reais. A utilização de *Python* como linguagem central e a divisão estratégica dos dados contribuíram para o desenvolvimento de modelos robustos e confiáveis. Além disso, a aplicação de técnicas de AL destacou-se como uma abordagem promissora para aprimorar a eficiência dos modelos na área da medicina, apesar de os resultados obtidos, com menos dados, não terem alcançado um desempenho tão satisfatório quanto os obtidos com os modelos ML. A explicabilidade dos resultados foi enfatizada através da aplicação de LIME, evidenciando a influência dos diferentes modelos na interpretação das previsões.

Os resultados obtidos demonstram a eficácia dos modelos de IA na deteção precoce da

doença de Alzheimer, fornecendo *insights* valiosos para a prática clínica e o bem-estar dos pacientes.

5

Conclusão e Trabalhos Futuros

O presente estudo tem como objetivo primordial a deteção da doença de *Alzheimer*, nas suas fases iniciais, através da análise de imagens Ressonância Magnética (MRI) do cérebro. Esta deteção nas fases iniciais é bastante importante para que se possam delinear estratégias para tentar atrasar a progressão da doença.

Para atingir este propósito, numa primeira fase foram utilizados os modelos *Visual Geometry Group* de 16 camadas (VGG-16), *Visual Geometry Group* de 19 camadas (VGG-19) e *Residual Network* de 50 camadas (ResNet-50), sem a aplicação de *data augmentation*. De modo a tentar combater as limitações do *dataset* causadas pelo não balanceamento das classes, numa segunda fase recorreu-se à aplicação de *data augmentation* ao modelo ResNet-50, com o intuito de aumentar a variabilidade das amostras de treino e, potencialmente, melhorar a capacidade do modelo de generalizar para novos dados. Numa terceira fase, foi implementada a técnica de *Active Machine Learning* (AL) com o modelo ResNet-50, com o objetivo de selecionar de maneira inteligente as amostras mais informativas para anotação e treino do modelo, o que é especialmente útil em casos em que os dados são limitados. Por fim, numa quarta fase, realizou-se a comparação das explicações geradas pelos diferentes modelos utilizando técnicas de *Explainable AI* (XAI), especificamente *Local Interpretable Model-agnostic Explanations* (LIME).

Os resultados do primeiro teste experimental foram altamente positivos, com métricas de desempenho, como *precision*, *recall* e *f1-score* excedendo a marca de 86% e as métricas de *Cohen's Kappa* e *Chance Adjusted Accuracy* (CAA) a exceder os 75%. No entanto, os resultados da segunda fase não foram tão favoráveis, tendo sido obtidos valores de cerca de 71% para as primeiras métricas referidas, de 51% para a métrica *Cohen's Kappa* e de 53% para a métrica de CAA. No que diz respeito aos resultados da terceira fase, estes demonstraram ser ainda menos favoráveis que os da fase anterior, com as métricas de *precision*, *recall* e *f1-score* a rondar os 68%, *Cohen's Kappa* de cerca de 48% e CAA de cerca de 58%.

Os resultados menos favoráveis obtidos na segunda fase experimental podem ser atribuídos ao aumento na variabilidade dos dados, que podem ter causado uma dificuldade acrescida na aprendizagem de padrões existentes entre os dados, por parte do modelo. Além disso, as amostras geradas apresentam uma natureza artificial, o que pode ter introduzido padrões irreais aos dados, contribuindo assim para a maior dificuldade de aprendizagem do modelo.

No que diz respeito aos resultados obtidos na terceira fase experimental, o facto de estes

serem menos favoráveis pode ser atribuído às diferentes *frameworks* utilizadas para o desenvolvimento dos modelos (*Tensorflow* e *Pytorch*). As diferenças entre estas *frameworks* podem influenciar o processo de treino e a otimização dos modelos, resultando em desempenhos variados. Além disso, o desempenho do Active Learning pode ter sido impactado pela seleção das amostras ou pelo próprio método de amostragem utilizado, que pode não ter sido ideal para o conjunto de dados específico.

Na fase final deste trabalho, foi selecionada uma amostra aleatória do conjunto de teste, que foi submetida aos diferentes modelos para previsão e explicação por parte do LIME. Através desta abordagem foi possível verificar que as explicações geradas com os diferentes modelos são bastante distintas, pelo que se ressalta a importância da inclusão de um especialista, como trabalho futuro, de modo a avaliar qual das explicações geradas mais se aproxima da correta e perceber como se pode melhorar o modelo e o processo explicativo, para que este possa ser utilizado em casos práticos da vida real. Desta forma, destaca-se a importância da colaboração com especialistas, de modo a garantir que as explicações geradas sejam úteis e precisas, orientando possíveis ajustes e melhorias no modelo.

Adicionalmente, uma sugestão para futuras pesquisas é a utilização de um *dataset* real e balanceado, e o teste de diferentes técnicas de *data augmentation*, de modo a mitigar as limitações causadas pelo não balanceamento das classes. Além disso, destaca-se também a importância do treino destes modelos com diferentes arquiteturas. Outra possibilidade seria o treino do modelo AL utilizando uma *framework* diferente, ou ajustando os parâmetros de treino. Estes passos podem contribuir para a melhoria da performance e robustez do modelo, facilitando assim a sua aplicação em cenários reais.

Em síntese, esta pesquisa demonstrou o potencial da Inteligência Artificial (IA) no auxílio dos diagnósticos médicos, nomeadamente na doença de *Alzheimer*. Embora os resultados não tenham sido significativamente superiores aos métodos tradicionais, o estudo destacou a importância de abordar as limitações dos conjuntos de dados e de utilizar abordagens mais sofisticadas como *data augmentation* e AL. Além disso, sublinhou a necessidade de um envolvimento mais ativo de especialistas médicos para validar as explicações geradas pelos modelos, garantindo assim que as soluções desenvolvidas sejam não apenas tecnicamente robustas, mas também clinicamente relevantes.

A integração da IA na medicina tem o potencial de transformar significativamente a prática médica, oferecendo ferramentas avançadas para diagnósticos mais rápidos e precisos. A colaboração contínua entre cientistas de dados e profissionais de saúde será essencial para desenvolver soluções que possam ser implementadas de forma eficaz e segura, melhorando assim os resultados para os pacientes e contribuindo para avanços significativos na área da saúde.

Referências

[1] C. A. Lane, J. Hardy, and J. M. Schott, "Alzheimer's disease," *European Journal of Neurology*, vol. 25, no. 1, pp. 59–70, 2018.

[2] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *Am. J. Med.*, vol. 132, pp. 795–801, July 2019.

[3] N. Jahan, I. B. Rashid, O. Al Numan, A. S. M. T. Hasan, and N. Begum, "Collaborative ai in smart healthcare system," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, pp. 1–5, 2021.

[4] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, pp. 44–53, 08 2017.

[5] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, 07 2010.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-agnostic interpretability of machine learning," *ArXiv*, vol. abs/1606.05386, 2016.

[7] C. Molnar, *Interpretable Machine Learning*. 2 ed., 2022.

[8] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. D. Ser, W. Samek, I. Jurisica, and N. Díaz-Rodríguez, "Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence," *Information Fusion*, vol. 79, pp. 263–278, 2022.

[9] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[10] Z. Breijyeh and R. Karaman, "Comprehensive review on alzheimer's disease: Causes and treatment," *Molecules*, vol. 25, p. 5789, Dec. 2020.

[11] L. R.Squire, D. Berg, F. E. Bloom, S. du Lac, A. Ghosh, and N. C. Spitzer, *Fundamental Neuroscience*. Elsevier Science, 2013.

[12] S. Khan, K. H. Barve, and M. S. Kumar, "Recent advancements in pathogenesis, diagnostics and treatment of alzheimer's disease," *Curr. Neuropharmacol.*, vol. 18, no. 11, pp. 1106–1125, 2020.

[13] A. P. Porsteinsson, R. S. Isaacson, S. Knox, M. N. Sabbagh, and I. Rubino, "Diagnosis of early alzheimer's disease: Clinical practice in 2021," *J. Prev. Alzheimers Dis.*, pp. 1–16, 2021.

[14] "About Alzheimerx2019;s — Alzheimerapos;s Disease Research Center — adrc.pitt.edu." `https://www.adrc.pitt.edu/alzheimers-disease/about-alzheimers/`. [Accessed 16-10-2023].

[15] A. Chandra, G. Dervenoulas, M. Politis, and Alzheimer's Disease Neuroimaging Initiative, "Magnetic resonance imaging in alzheimer's disease and mild cognitive impairment," *J. Neurol.*, vol. 266, pp. 1293–1302, June 2019.

[16] G. Livingston, J. Huntley, A. Sommerlad, D. Ames, C. Ballard, S. Banerjee, C. Brayne, A. Burns, J. Cohen-Mansfield, C. Cooper, S. G. Costafreda, A. Dias, N. Fox, L. N. Gitlin, R. Howard, H. C. Kales, M. Kivimäki, E. B. Larson, A. Ogunniyi, V. Orgeta, K. Ritchie, K. Rockwood, E. L. Sampson, Q. Samus, L. S. Schneider, G. Selbæk, L. Teri, and N. Mukadam, "Dementia prevention, intervention, and care: 2020 report of the lancet commission," *Lancet*, vol. 396, pp. 413–446, Aug. 2020.

[17] J. Neugroschl and S. Wang, "Alzheimer's disease: Diagnosis and treatment across the spectrum of disease severity," *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, vol. 78, no. 4, pp. 596–612, 2011.

[18] E. Passeri, K. Elkhoury, M. Morsink, K. Broersen, M. Linder, A. Tamayol, C. Malaplate, F. T. Yen, and E. Arab-Tehrany, "Alzheimer's disease: Treatment strategies and their limitations," *Int. J. Mol. Sci.*, vol. 23, p. 13954, Nov. 2022.

[19] W. M. van Oostveen and E. C. M. de Lange, "Imaging techniques in alzheimer's disease: A review of applications in early diagnosis and longitudinal monitoring," *Int. J. Mol. Sci.*, vol. 22, p. 2110, Feb. 2021.

[20] M. Cristina, F. Nunes, S. Hage, I. Masao, and I. Ii, "Imagem por ressonância magnética: princípios básicos magnetic resonance imaging-basics resumo," pp. 1287–1295.

[21] *Production of Net Magnetization*, ch. 1, pp. 1–9. John Wiley Sons, Ltd, 2003.

[22] A. A. Mazzola, "Artigo de revisão introdução ressonância magnética: princípios de formação da imagem e aplicações em imagem funcional magnetic resonance: principles of image formation and applications in funcional imaging."

[23] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, p. 160, Mar 2021.

[24] R. Monarch, R. Munro, and C. Manning, *Human-in-the-Loop Machine Learning: Active Learning and Annotation for Human-centered AI.* Manning, 2021.

[25] K. Team, "Keras documentation: Review Classification using Active Learning — keras.io." https://keras.io/examples/nlp/active_learning_review_classification/. [Accessed 22-04-2024].

[26] D. A. Rocha, F. M. F. Ferreira, and Z. M. A. Peixoto, "Diabetic retinopathy classification using vgg16 neural network," *Research on Biomedical Engineering*, vol. 38, pp. 761 – 772, 2022.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[28] L. Falqueto, R. Paes, and A. Passaro, "Knn e rede neural convolucional para o reconhecimento de plataformas de petróleo em imagens sar do sentinel-1," *Aplicações Operacionais em Áreas de Defesa*, vol. 24, pp. 29–33, 09 2023.

[29] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-vgg16 cnn model for big data places image recognition," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 169–175, 2018.

[30] G. Learning, "Everything you need to know about VGG16 — mygreatlearning." https://medium.com/@mygreatlearning/everything-you-need-to-know-about-vgg16-7315defb5918. [Accessed 24-04-2024].

[31] T.-H. Nguyen, T.-N. Nguyen, and B.-V. Ngo, "A vgg-19 model with transfer learning and image segmentation for classification of tomato leaf disease," *AgriEngineering*, vol. 4, no. 4, pp. 871–887, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 12 2015.

[33] B. Mandal, A. Okeukwu, and Y. Theis, "Masked face recognition using resnet-50," 4 2021.

[34] V. P. Nithya, N. Mohanasundaram, and R. Santhosh, "An early detection and classification of alzheimer's disease framework based on ResNet-50," *Curr. Med. Imaging Rev.*, vol. 20, Oct. 2023.

[35] N. Kundu, "Exploring ResNet50: An In-Depth Look at the Model Architecture and Code Implementation — nitishkundu1993." https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d [Accessed 24-04-2024].

[36] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and Precise4Q consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Med. Inform. Decis. Mak.*, vol. 20, p. 310, Nov. 2020.

[37] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information Fusion*, vol. 99, p. 101805, 2023.

[38] S. A. and S. R., "A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends," *Decision Analytics Journal*, vol. 7, p. 100230, 2023.

[39] J. An, Y. Zhang, and I. Joe, "Specific-input lime explanations for tabular data based on deep learning models," *Applied Sciences*, vol. 13, no. 15, 2023.

[40] "Explainable AI(XAI) Using LIME - GeeksforGeeks — geeksforgeeks.org." `https://www.geeksforgeeks.org/introduction-to-explainable-aixai-using-lime/`. [Accessed 03-01-2024].

[41] G. del Castillo Torres, M. F. Roig-Maimó, M. Mascaró-Oliver, E. Amengual-Alcover, and R. Mas-Sansó, "Understanding how CNNs recognize facial expressions: A case study with LIME and CEM," *Sensors (Basel)*, vol. 23, p. 131, Dec. 2022.

[42] B. Wieckowska, K. B. Kubiak, P. Jóźwiak, W. Moryson, and B. Stawińska-Witoszyńska, "Cohen's kappa coefficient as a measure to assess classification improvement following the addition of a new marker to a regression model," *Int. J. Environ. Res. Public Health*, vol. 19, p. 10213, Aug. 2022.

[43] G. James, D. Witten, T. Hastie, and R. Tibshirani, *Springer Texts in Statistics*. 01 2013.

[44] M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, vol. 3, p. 224, 04 2013.

[45] X. Lai, Y. Lu, L. Zhang, Y. Feng, and G. Zhang, "Imbalanced-type incomplete data fuzzy modeling and missing value imputations," in *Proceedings of the 2021 5th International Conference on Machine Learning and Soft Computing*, ICMLSC '21, (New York, NY, USA), p. 33–37, Association for Computing Machinery, 2021.

[46] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[47] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[48] G. Kantayeva, J. Lima, and A. I. Pereira, "Application of machine learning in dementia diagnosis: A systematic literature review," *Heliyon*, vol. 9, no. 11, p. e21626, 2023.

[49] Z. Yao, H. Wang, W. Yan, Z. Wang, W. Zhang, Z. Wang, and G. Zhang, "Artificial intelligence-based diagnosis of alzheimer's disease with brain mri images," *European Journal of Radiology*, vol. 165, p. 110934, 2023.

[50] A. A. A. El-Latif, S. A. Chelloug, M. Alabdulhafith, and M. Hammad, "Accurate detection of alzheimer's disease using lightweight deep learning model on MRI data," *Diagnostics (Basel)*, vol. 13, p. 1216, Mar. 2023.

[51] A. Khalid, E. M. Senan, K. Al-Wagih, M. M. A. Al-Azzam, and Z. M. Alkhraisha, "Automatic analysis of MRI images for early prediction of alzheimer's disease stages based on hybrid features of CNN and handcrafted features," *Diagnostics (Basel)*, vol. 13, p. 1654, May 2023.

[52] S. Alshmrany, G. M. U. D. Dar, and S. I. Ansarullah, "Open peer review on qeios an improved hybrid transfer learning-based deep learning model for alzheimer's disease detection using ct and mri scans,"

[53] D. AlSaeed and S. F. Omar, "Brain MRI analysis for alzheimer's disease diagnosis using CNN-based feature extraction and machine learning," *Sensors (Basel)*, vol. 22, p. 2911, Apr. 2022.

[54] P. R. Magesh, R. D. Myloth, and R. J. Tom, "An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery," *Computers in Biology and Medicine*, vol. 126, p. 104041, 2020.

[55] K. Davagdorj, J.-W. Bae, V.-H. Pham, N. Theera-Umpon, and K. H. Ryu, "Explainable artificial intelligence based framework for non-communicable diseases prediction," *IEEE Access*, vol. 9, pp. 123672–123688, 2021.

[56] A. W. Mulyadi, W. Jung, K. Oh, J. S. Yoon, K. H. Lee, and H.-I. Suk, "Estimating explainable alzheimer's disease likelihood map via clinically-guided prototype learning," *NeuroImage*, vol. 273, p. 120073, 2023.

[57] S. Qiu, P. S. Joshi, M. I. Miller, C. Xue, X. Zhou, C. Karjadi, G. H. Chang, A. S. Joshi, B. Dwyer, S. Zhu, M. Kaku, Y. Zhou, Y. J. Alderazi, A. Swaminathan, S. Kedar, M.-H. Saint-Hilaire, S. H. Auerbach, J. Yuan, E. A. Sartor, R. Au, and V. B. Kolachalama, "Development and validation of an interpretable deep learning framework for Alzheimer's disease classification," *Brain*, vol. 143, pp. 1920–1933, 05 2020.

[58] K. Oh, D.-W. Heo, A. W. Mulyadi, W. Jung, E. Kang, and H.-I. Suk, "Quantifying explainability of counterfactual-guided mri feature for alzheimer's disease prediction," 12 2022.

[59] "Alzheimer MRI Preprocessed Dataset — kaggle.com." `https://www.kaggle.com/datasets/sachinkumar413/alzheimer-mri-dataset`. [Accessed 08-10-2023].

[60] M. Vilares Ferro, Y. Doval Mosquera, F. J. Ribadas Pena, and V. M. Darriba Bilbao, "Early stopping by correlating online indicators in neural networks," *Neural Netw.*, vol. 159, pp. 109–124, Feb. 2023.

[61] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *ArXiv*, vol. abs/2007.15745, 2020.

[62] J. Cao, D. Zhao, C. Tian, T. Jin, and F. Song, "Adopting improved adam optimizer to train dendritic neuron model for water quality prediction," *Math. Biosci. Eng.*, vol. 20, no. 5, pp. 9489–9510, 2023.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[64] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," *ArXiv*, vol. abs/1912.01703, 2019.