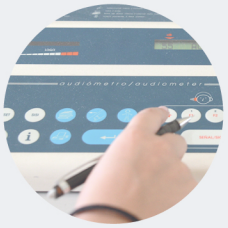
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA BIOMÉDICA






Criação de Bases de Dados de Imagens Histológicas Anotadas e Desenvolvimento de um Modelo de Classificação Automática de Patologias Mamárias

CRISTINA ISABEL DA SILVA MOREIRA
novembro de 2023

Instituto Superior de Engenharia do Porto

CRIAÇÃO DE BASE DE DADOS DE IMAGENS HISTOLÓGICAS ANOTADAS E DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO AUTOMÁTICA DE PATOLOGIAS MAMÁRIAS

Cristina Isabel da Silva Moreira

2023

Instituto Superior de Engenharia do Porto

Departamento de Física

P.PORTO


P.PORTO

Instituto Superior de Engenharia do Porto

CRIAÇÃO DE BASE DE DADOS DE IMAGENS HISTOLÓGICAS ANOTADAS E DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO AUTOMÁTICA DE PATOLOGIAS MAMÁRIAS

Cristina Isabel da Silva Moreira

Estudante n.º 1210159

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Biomédica, realizada sob a orientação do Doutor Luís Filipe Martins Pinto Coelho.

2023

Instituto Superior de Engenharia do Porto

Departamento de Física

P.PORTO


P.PORTO

AGRADECIMENTOS

Gostaria de expressar o meu profundo agradecimento ao meu orientador, Doutor Luís Filipe Coelho, pela orientação imensurável, pela paciência e por todo o apoio que foram cruciais ao longo deste percurso académico. A sua dedicação foi um pilar fundamental para a concretização deste trabalho.

Estendo os meus agradecimentos também aos meus colegas de trabalho pela colaboração para o desenvolvimento deste projeto e para o meu desenvolvimento pessoal e profissional. Um reconhecimento especial ao Doutor Eduardo Silva Ferreira pela sua disponibilidade e colaboração imprescindível neste projeto. A generosidade em fornecer o material permitiram que este trabalho não só começasse, mas também alcançasse os objetivos com sucesso.

Um agradecimento particular aos professores da Escola Superior de Saúde, em especial à Doutora Regina Silva que permitiram e facilitaram o acesso a equipamentos essenciais para a obtenção de material para este projeto.

Não posso deixar de agradecer a todos os amigos e familiares que estiveram ao meu lado, dando-me sempre suporte emocional e incentivo. Um agradecimento especial aos meus pais e ao meu irmão que foram um suporte incondicional ao longo desta jornada e claro em último, mas com ênfase ao meu namorado por todo o apoio e por toda a ajuda que foi essencial para que todos os desafios fossem superados e que tornou assim possível que esta etapa fosse concluída com sucesso.

Esta caminhada foi enriquecida pela presença e contribuição de cada um de vocês, e as palavras não podem expressar inteiramente a minha gratidão. Um GRANDE OBRIGADA a cada um de vocês em especial por fazerem parte desta enorme jornada.

"Há verdadeiramente duas coisas diferentes: saber e crer que se sabe. A ciência consiste em saber; em crer que se sabe reside a ignorância."

Hipócrates

RESUMO

O Cancro da Mama é uma das patologias mais prevalentes mundialmente e uma das principais causas de mortalidade por cancro associado ao sexo feminino em Portugal. O diagnóstico de patologias mamárias que é feito através de biópsias é uma tarefa complexa e detalhada para os Patologistas e como tal suscetível a erros. É exigido uma análise meticulosa e especializada de múltiplos campos microscópicos, onde a precisão é vital e os atrasos podem ser críticos. Neste sentido, o desenvolvimento de plataformas que auxiliem a um diagnóstico rápido e preciso é cada vez mais essencial.

Neste projeto foi desenvolvido um *dataset* de imagens histológicas de biópsias mamárias para o diagnóstico de patologias mamárias, com o objetivo de aplicar e avaliar a eficácia das redes neuronais na classificação e análise destes tecidos. O processo de criação do *dataset* resume-se à recolha do tecido, seguindo o seu processamento laboratorial onde as lâminas obtidas foram digitalizadas e submetidas a um processo de conversão e segmentação para formatos compatíveis com a posterior análise. A organização e categorização das imagens foi efetuada em código *Python* para a classificação automatizada, garantindo a integridade e precisão dos dados.

A fase de pré-processamento e organização do *dataset* foram essenciais para assegurar a qualidade e representatividade dos dados. A precisão das categorizações e a distribuição equilibrada das imagens nas respetivas categorias para treino e validação dos modelos foram cruciais. A normalização das imagens e a extração adequadas dos segmentos de interesse foram etapas fundamentais para preparar os dados para a análise das redes neuronais. Esta preparação dos dados assegurou que os modelos fossem treinados com as informações corretas e essenciais para garantir a eficácia da aprendizagem.

Para a aplicação das redes neuronais, foram selecionados os modelos SqueezeNet e InceptionV3, onde foram testados quatros cenários de classificação em ambas as arquiteturas e utilizadas quatro classes patológicas diferentes (Doença Fibrocística, Fibroadenoma, Carcinoma Lobular Invasivo e Carcinoma Ductal Invasivo). Estes modelos foram adaptados com camadas de entrada e saída personalizadas. A eficácia destes modelos foi avaliada com métricas estatísticas e gráficas incluindo a matriz de confusão, exatidão (*accuracy*), precisão (*precision*), sensibilidade (*recall*), *F1-score* e ainda as curvas de ROC e *Precision-Recall*. Ambos os modelos demonstraram uma boa performance com uma *accuracy* que variou entre os 88% e os 98% para todos os cenários testados. Porém, foi observado que o modelo InceptionV3 é o mais bem-sucedido, obtendo na maioria dos casos os valores mais altos de *accuracy* apesar de se ter observado alguma variação devido a fenómenos como *overfitting*.

Os resultados obtidos indicam que as redes neuronais podem ser ferramentas eficazes no diagnóstico de patologias mamárias a partir de imagens histológicas. A *accuracy* elevada dos modelos utilizados para desenvolvimento deste projeto, refletem a capacidade de reconhecer e classificar de forma precisa as características morfológicas relevantes nas imagens, demonstrando que a Inteligência Artificial tem um potencial significativo para melhorar a precisão e eficácia dos diagnósticos em Anatomia Patológica.

PALAVRAS-CHAVE

Cancro de Mama; Redes Neuronais; *Machine Learning*; SqueezeNet; InceptionV3.

ABSTRACT

Breast cancer is one of the most prevalent pathologies worldwide and a leading cause of cancer-related mortality among women in Portugal. Diagnosing breast pathologies through biopsies is a complex and detailed task for Pathologists and thus prone to errors. It requires meticulous and specialized analysis of multiple microscopic fields, where accuracy is vital, and delays can be critical. In this sense, the development of platforms that aid in rapid and precise diagnosis is increasingly essential.

In this project, a dataset of histological images from breast biopsies for the diagnosis of breast pathologies was developed, aiming to apply and evaluate the effectiveness of neural networks in classifying and analyzing these tissues. The dataset creation process involved tissue collection, followed by laboratory processing where the obtained slides were digitized and subjected to a conversion and segmentation process into formats compatible with subsequent analysis. The organization and categorization of the images were carried out using Python code for automated classification, ensuring the integrity and accuracy of the data.

The pre-processing and organization phase of the dataset were essential to ensure the quality and representativeness of the data. The accuracy of the categorizations and the balanced distribution of images in their respective categories for training and validating the models were crucial. Normalization of the images and proper extraction of the segments of interest were fundamental steps in preparing the data for neural network analysis. This data preparation ensured that the models were trained with the correct and essential information to guarantee the effectiveness of the learning.

For the application of neural networks, the SqueezeNet and InceptionV3 models were selected. Four classification scenarios were tested in both architectures using four different pathological classes (Fibrocystic Disease, Fibroadenoma, Invasive Lobular Carcinoma, and Invasive Ductal Carcinoma). These models were adapted with custom input and output layers. The efficacy of these models was evaluated using statistical and graphical metrics including the confusion matrix, accuracy, precision, recall, F1-score, and the ROC and Precision-Recall curves. Both models demonstrated good performance, with accuracy ranging between 88% and 98% for all tested scenarios. However, it was observed that the InceptionV3 model is the most successful, achieving the highest accuracy values in most cases, despite some variation due to phenomena such as overfitting.

The results indicate that neural networks can be effective tools in diagnosing breast pathologies from histological images. The high accuracy of the models used for this project's development reflects their ability to accurately recognize and classify relevant morphological characteristics in the images, demonstrating that Artificial Intelligence has significant potential to improve the precision and efficiency of diagnoses in Anatomical Pathology.

KEYWORDS

Breast Cancer; Neural Networks; Machine Learning; SqueezeNet; InceptionV3.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

SIGLA	***Descrição***
AUC	*Area Under the Curve*
CIS	Carcinoma Invasivos de Mama
CM	Cancro de Mama
CNNs	*Convolutional Neural Networks*
CSV	*Comma-separated values*
FN	Falsos Negativos
FP	Falsos Positivos
HE	Hematoxilina-Eosina
IA	Inteligência Artificial
ISEP	Instituto Superior de Engenharia do Porto
LSTMs	*Long Short-Term Memory networks*
MSMV-PFENet	*Multi-Scale Motion Vectores Pyramid Feature Extraction*
P.Porto	Instituto Politécnico do Porto
PR	*Precision-Recall*
RAM	*Random Access Memory*
RE	Recetores de Estrogénio
RNAs	Redes Neuronais Artificiais
RNNs	*Recurrent Neural Networks*
ROC	*Receiver Operating Characteristic*
RP	Recetores de Progesterona
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

1. INTRODUÇÃO

O Cancro da Mama (CM) é um dos cancros mais prevalentes do mundo essencialmente associado a pessoas do sexo feminino, embora também possa ocorrer em pessoas do sexo masculino e com tipologias muito agressivas. Atualmente o CM representa uma das maiores causas de morte por cancro em mulheres em Portugal. O risco de desenvolvimento do cancro da mama aumenta com a idade, a maioria dos casos são diagnosticado em mulheres entre os 50 e os 69 anos, no entanto, esta patologia tem surgido cada vez mais em mulheres em idades mais jovens[1].

O CM é uma patologia que engloba diversos subtipos e que se diferenciam essencialmente pelo seu perfil histológico, as suas respostas aos tratamentos e o possível prognóstico. Os subtipos de CM que existem assentam sobretudo nas diferentes caracteristicas morfológicas que podem ser observadas nas células que compõem estes tumores bem como, na resposta aos estudos imunológicos que são feitos para caracterização do tumor[2].

Na maioria dos casos é durante os exames imagiológicos de rastreio, como a mamografia ou a ecografia mamária que são detetados os nódulos e/ou tumores. Nestas situações é normalmente possível distingir de imediato através da imagem se estamos perante um tumor de origem benigna ou maligna. Os tumores benignos caracterizam-se por terem um crescimento lento, estarem bem delimitados, muitas vezes por uma cápsula de fibrose envolvente e não são tumores com capacidade infiltrativa nem metastática. Já os tumores malignos são detentores de todas as características opostas às mencionadas anteriormente, apresentam um crescimento muito acelerado, não se encontram circunscritos num local específico e têm capacidade de invadir e metastisar em vários locais do corpo[3], [4].

Enquanto os tumores benignos não exigem muita atenção, só são removidos quando atingem grandes dimensões uma vez que não apresentam risco direto para os pacientes. Estes normalmente só carecem da necessidade da realização de um exame imagiológico a cada seis meses para acompanhar a sua evolução e o crescimento. Os tumores malignos exigem uma rápida intervenção, protocolarmente as pacientes são encaminhadas para realizarem uma biópsia desse tecido para que seja realizada a caracterização do tumor histologicamente e imunologicamente. Estes dados são estudados pelos serviços de Anatomia Patológica, onde os tecidos recolhidos passam de forma sequencial por um conjunto de processos fisicos e químicos que permite aos médicos Patologistas ver microscopicamente que tipo de células se encontram naquele tumor. Posteriormente solicitam os estudos Imunohistoquímicos daquele tumor de modo a conseguirem orientar os médicos Oncologistas sobre qual poderá ser o melhor tratamento a realizar e a quais tratamentos aquele tumor irá responder de forma mais eficaz [1], [4], [5].

O cancro de mama invasivo é o tipo mais comum, distinguindo-se principalmente pela sua capacidade de invasão e metástase. O carcinoma ductal invasivo tem origem nas células dos ductos mamários enquanto o carcinoma lobular invasivo tem origem nas células lobulares. Já os cancros de mama *In Situ* como o carcinoma ductal *In Situ* e o carcinoma lobular *In Situ* caracterizam-se por um estadio menos agressivo da patologia[3].

A classificação dos tumores mamários também é feita tendo em conta a sensibilidade das células presentes para os recetores de estrogénio (RE), recetores de progesterona (RP) e a fator de crescimento HER2. Assim, podem ser classificados como Triplo-Negativo que são os que apresentam uma maior agressividade visto que não respondem aos tratamentos realizados. Os carcinomas podem ser também hormono-dependentes, sendo nestes casos tumores que respondem positivamente aos tratamentos anti-hormonais[6], [7].

Para além das patologias malignas, também existem patologias benignas no tecido mamário, entre as quais se destacam a Doença Fibrocística e os Fibroadenomas por serem as que apresentam uma maior incidência em idades jovens[8], [9].

Atualmente, a confirmação da presença de células neoplásicas em massas tumorais identificadas por exames de imagem é efetuada mediante biópsia da referida massa e subsequente análise microscópica das alterações celulares. Esta análise é efetuada por um médico Patologista, que realiza uma avaliação exaustiva de cada campo microscópico para assegurar a identificação de vários achados histológicos relevantes. A observação realizada por este profissional de saúde, exige rigor e perícia, dado que as características celulares responsáveis pelo diagnóstico final são muitas vezes subtis e susceptíveis de serem confundidas. A dificuldade de deteção destes detalhes é frequentemente aumentada por fatores intrínsecos ao erro humano, que podem decorrer de várias etapas, desde a recolha do tecido, até ao diagnóstico final, incluindo a fixação, o processamento, o corte e a coloração das amostras. Considerando que cada lâmina pode conter mais de 500 campos para análise, esta fase do processo revela-se morosa, representando um desafio significativo numa área da saúde onde a celeridade é cada vez mais crucial e onde os atrasos podem implicar consequências fatais.

Com a evolução da tecnologia têm surgido possibilidades de automatizar a análise do tecido, através de técnicas de análise da imagem utilizando *machine learning* para o desenvolvimento dos algoritmos. As redes neuronais convolucionais e a Inteligência Artificial (IA) têm tido um papel revolucionário no diagnóstico do cancro de mama. Os sistemas de IA que são desenvolvidos para auxiliar no diagnóstico do cancro de mama são treinados com amplos conjuntos de imagens médicas que normalmente são pré-anotadas por radiologistas ou patologistas. Aprendem a reconhecer características específicas de tumores malignos, como as margens irregulares, densidades anormais ou até a presença de microcalcificações[10].

A IA poderá futuramente servir de apoio aos médicos na tomada de decisões ao terem a capacidade de fornecer uma segunda opinião baseada na aprendizagem que fizeram ou até mesmo serem capazes de realçar algum pormenor que pode ter escapado ao olho humano. Além disso, conseguem processar e analisar um volume substancial de imagens rapidamente[11].

No panorama português e até mundial, dada a escassez de especialistas médicos em determinas áreas da saúde, a IA pode desempenhar um papel fundamental na garantia de diagnósticos mais rápidos e precisos, permitindo assim um alívio da carga de trabalho dos profissionais de saúde.

Contudo, para que a integração destas tecnologias seja bem-sucedida, é fundamental que exista uma infraestrutura robusta, uma formação específica para os profissionais de saúde na utilização destas novas ferramentas e um enquadramento regulamentar que assegure a qualidade e segurança dos diagnósticos realizados com o auxílio da IA[11], [12].

1.1. Enquadramento e pertinência

O CM mais comum é designado de Cancro de Mama Invasivo, que se caracteriza essencialmente pela sua capacidade de invasão dos tecidos que se encontram ao redor bem como a capacidade de se mestatizar em outras partes do corpo. Dentro deste tipo de CM existem dois subtipos, o Carcinoma Ductal Invasivo, que tal como o nome indica, este carcinoma tem origem nos ductos mamários e é o tipo de CM mais frequente, representando cerca de 70-80% de todos os casos. Existe ainda o Carcinoma Lobular Invasivo que tem origem nas células dos lóbulos (que são as glândulas produtoras de leite) e este tipo de CM representa cerca de 10% dos CM em Portugal[1], [4].

Existe ainda o Cancro de Mama *In Situ* (ou não invasivo) que tal como o nome indica é um tumor num estadio não invasor dos tecidos adjacentes e portanto são consideradas formas de cancro

menos agressivas quando comparadas com os subtipos de Carcinoma de Mama Invasivos. Também este tipo de CM se subdivide em dois subtipos, nomeadamente o Carcinoma Ductal *In Situ* e o Carcinoma Lobular *In Situ*, tendo estes tumores origem nas células dos ductos e nas células dos lóbulos respetivamente. Estes dois subtipos representam cerca de 10-15% dos CM em Portugal[1], [4], [5].

Após a classificação histológica dos tumores, no tipo e subtipo a que estes pertencem, é feita uma avaliação do comportamento biológico do tumor relativamente aos RE, RP e ao recetor do factor de crescimento HER2. Os tumores podem então ser classificados como Triplo-Negativo, que é o tipo de carcinoma mais agressivo, uma vez que não expressa RE, RP nem HER2, sendo assim um carcinoma que não responde aos tratamentos mais comuns que têm como alvo estes recetores. Pode ser classificado como um carcinoma HER2-positivo, onde existe uma superexpressão do gene HER2 e onde é feita uma terapia dirigida para este gene, permitindo assim que o desenvolvimento do tumor seja controlado. E existe ainda o subtipo de CM que é considerado Hormono-Dependente, onde se incluem todos os CM que expressam os RE e os RP onde podem ser realizados tratamentos relacionados com o bloqueio da ação destas hormonas e controlar desta forma o crescimento do tumor[13], [14].

Para além de todas as patologias malignas descritas até agora, também existem patologias benignas que surgem em mulheres de todas as faixas etárias e são normalmente detetadas nos exames de rotina ou até mesmo durante a auto-palpação. Atualmente as imagens do tecido mamário obtidas através da imagiologia já costumam ter uma boa resolução o que nos permite evitar que os pacientes tenham de fazer exames mais invasivos como é o caso das biópsias quando estamos perante uma patologia do foro benigno. No entanto, nem sempre isto acontece e há situações em que as imagens obtidas não são completamente esclarecedoras e é feita biópsia mamária para confirmar a benignidade do tecido anómalo que se encontra na imagem[8], [15], [16].

De entre as várias patologias benignas que existem, as mais comuns de causar algumas dúvidas e consequentemente as que têm uma maior percentagem de incidências são a Doença Fibrocística, que se caracteriza pela presença de cistos e fibrose no tecido mamário, acontecendo muitas vezes alterações do tamanho dos cistos relacionadas com o ciclo menstrual da mulher. O fibroadenoma é um dos tumores benignos mais comum em mulheres jovens, este é composto por um grande aglomerado de tecido conjuntivo fibroso que dá ao tumor uma consistência firme e que se move com facilidade quando é palpado, característica muito comum em tumores benignos[15], [16].

Na Figura 1 encontra-se um resumo esquemático das várias etapas que fazem parte do protocolo a que os pacientes são sujeitos, desde os exames de rotina, passando pela deteção do tumor, realização

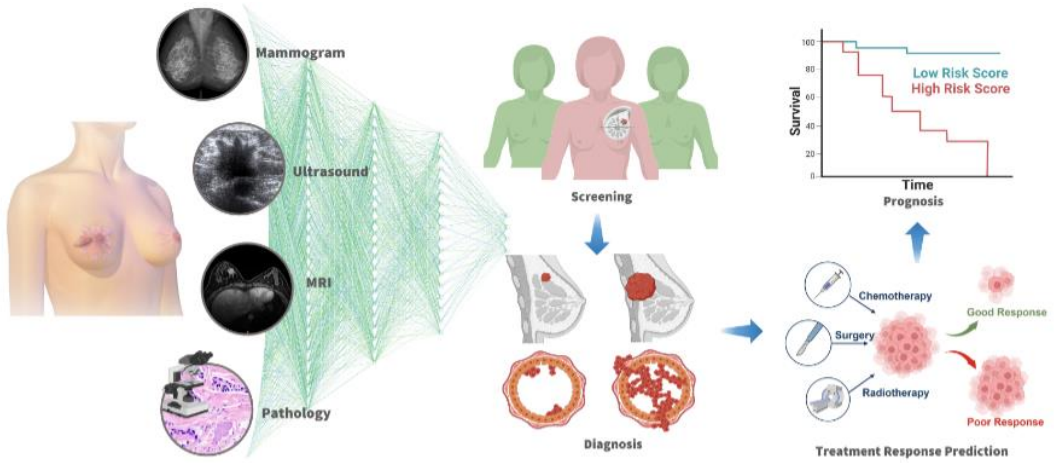


Figura 1 – Fluxograma das etapas de rastreio, diagnóstico e tratamento do cancro de mama, evidenciando os exames de imagiologia, a caracterização patológica do tecido bem como a análise de sobrevivência e resposta celular a terapias oncológicas[17]

de biópsia e posterior caracterização histológica do mesmo e em suma a escolha do tratamento [17].

Apesar das evoluções que têm sido feitas ao longos de décadas, ainda não existe um tratamento específico que nos permite garantir que todos os carcinomas possam ser tratáveis. Assim como a ciência do tratamento tem evoluído, também as células têm sofrido alterações e os carcinomas que têm surgido são cada vez mais agressivos para os pacientes e têm uma velocidade de crescimento e invasão cada vez maior, daí vivermos cada vez mais numa luta contra o tempo no que diz respeito ao diagnóstico e tratamento dos tumores.

Existem cada vez menos Patologistas em Portugal e o número de novos casos de cancro tem aumentado e as estatísticas apontam para que o aumento seja exponencial dentro de alguns anos. Surge assim a necessidade de recorrer a novas ferramentas com o objetivo de apoiar todos os profissionais de saúde e garantir respostas corretas e céleres a todos os pacientes.

A IA tem sido um incentivo para avanços significativos nas mais variadas áreas de saúde não sendo a área de Anatomia Patológica uma exceção. A IA é uma área tecnológica em expansão que se aprimora continuamente através de métodos de aprendizagem automática (*machine learning*). Esta abordagem permite que os sistemas computacionais aprendam e evoluam a partir dos dados e que melhorem a sua capacidade de reconhecimento de padrões e realização de tarefas complexas sem a intervenção humana direta [10].

Esta capacidade de aprendizagem contínua torna possível a aplicação da IA em diversas áreas como na saúde, onde pode auxiliar na deteção precoce de doenças e até mesmo ser um auxílio constante no trabalho de todos os profissionais de saúde. A IA abrange também áreas como processamento de linguagem natural e visão computacional, melhorando a interação entre humanos e máquinas. Contudo, a IA traz diversos desafios, nomeadamente em questões éticas e de privacidade relacionada com a colheita e uso de dados pessoais, problemas estes que requerem ainda alguma regulamentação do governo [11], [12], [18].

As tecnologias de *machine learning* são um subcampo da IA que se dedicam ao desenvolvimento de algoritmos e modelos estatísticos que permitem dotar os sistemas computacionais com capacidades de aprendizagem ou de melhoria do seu desempenho na realização de tarefas específicas através da aprendizagem. Não requerem uma programação explícita para cada nova adaptação, ao invés disso, este sistema utiliza padrões e interferências a partir dos dados fornecidos. Cientificamente, *machine learning* pode definir-se como o estudo e construção de modelos computacionais que podem identificar padrões complexos e fazer decisões ou previsões baseadas em dados, com pouca ou nenhuma intervenção humana após vários processos de treino. Estes modelos são treinados utilizando um conjunto de dados e algoritmos específicos, permitindo-lhes adaptar-se e generalizar a partir dos exemplos aprendidos para lidar com novos dados ou situações que não foram ainda observadas[11], [12], [18].

Na Anatomia Patológica, a implementação de sistemas baseados em IA tem revolucionado a forma em como o tecido é avaliado, nomeadamente a rapidez com que pode ser feita, bem como a possibilidade de destacar determinados pormenores menos comuns que até mesmo os Patologistas mais treinados não têm a capacidade de ver. Os sistemas são treinados utilizando um vasto reportório de imagens digitais de lâminas onde o diagnóstico é conhecido e onde já estão anotados os campos histológicos relevantes para determinado diagnóstico.

Dentro do campo de *machine learning*, podemos encontrar o *deep learning* – aprendizagem profunda, que se caracteriza pela utilização de redes neuronais artificiais com múltiplas camadas, denominadas de camadas ocultas. Estas redes são compostas por nós, que são análogos aos neurónios biológicos, que se encontram interligados e ajustam os "pesos sinápticos" através de um processo de treino interativo, permitindo assim que o sistema "aprenda" através dos dados fornecidos[19]–[21].

A arquitetura das redes neuronais profundas é projetada para que de forma automática e progressiva sejam extraídos vários dados importantes, iniciando-se por características mais gerais nas primeiras camadas ocultas e avança para características mais complexas nas camadas subsequentes. Este processo é descrito como uma hierarquização da aprendizagem[19], [20].

A *deep learning* tem mostrado ser fundamental para os avanços significativos que têm sido observados em tarefas como a percepção e cognição computacional como o reconhecimento de voz e de imagem, onde os modelos profundos superam outros métodos de aprendizagem. Um exemplo são as redes neuronais convolucionais (*Convolutional Neural Networks* - CNNs), que se especializam no processamento e análise de imagens digitais, recorrendo às propriedades consistentes das imagens para reduzir a complexidade computacional[22], [23].

As CNNs constituem uma categoria de redes neuronais profundas otimizadas para analisar dados estruturados em forma de matriz, como é o caso das imagens digitais, que são organizadas em linhas e colunas de píxeis. A arquitetura das CNNs é adaptada para capturar a natureza espacial e temporal dos dados, através do uso de operações matemáticas conhecidas como convoluções. Estas redes têm emergido como uma ferramenta muito poderosa na área da saúde, particularmente na identificação e distinção de células em aplicações de patologia digital. Esta classe de algoritmos de *deep learning*, destaca-se pela sua capacidade de processar imagens histológicas de alta resolução, sendo capaz de extrair as características essenciais para análise e classificação das células ou tecidos[24]–[28].

No campo da citologia e histologia as CNNs são utilizadas para realizar tarefas complexas de reconhecimento de padrões que normalmente exigem a interpretação de um patologista ou de um citotécnico. Através da análise de lâminas digitalizadas, estas redes aprendem a identificar as diferentes morfologias dos tecidos e a distinguir as células normais das células patológicas. A aplicação de filtros convolucionais permite às CNNs a deteção de características minuciosas das células tal como o formato do núcleo, a textura do citoplasma e até a presença de estruturas anormais[10], [29].

Adicionalmente, existem ainda variações como as Redes Neuronais Recorrentes (*Recurrent Neural Networks* - RNNs) e as Redes de Memória de Curto e Longo Prazo (*Long Short Term Memory networks* - LSTMs), que se têm revelado eficazes na resolução de problemas que envolvem sequências de dados, tais como a tradução automática e a modelagem linguística[30].

1.2. Questões e objetivos de investigação

Tendo em vista a busca de soluções para os desafios que atrás se apresentaram, foram formulados os seguintes objetivos para esta dissertação:

- O1. Preparação da Dados Histológicos: Escolha de pelos menos 100 casos diferentes de biópsias mamárias que incluam patologias benignas (Doença Fibrocística e Fibroadenoma) e malignas (Carcinoma Ductal Invasivo e Carcinoma Lobular Invasivo), para a realização de cortes histológicos e posterior coloração com HE (Hematoxilina-Eosina);
- O2. Elaboração de um Inventário de Imagens Histológicas: Contagem e divisão do número de imagens que foi possível obter de cada patologia;
- O3. Organização e preparação de imagens histológicas digitais: Executar o pré-processamento, conversão, segmentação, revisão e organização das imagens digitalizadas para formação do *dataset*;

- O4. Análise e Categorização: Realizar as devidas anotações e categorização de todas as imagens para assegurar o treino eficaz da rede neuronal;
- O5. Construir um modelo de classificação de imagens histológicos: Desenvolver um sistema de *machine learning* que permita classificar imagens histológicas como sendo pertencentes a tecidos saudáveis ou tecidos patológicos;
- O6. Avaliação de Desempenho: Analisar através das métricas de *accuracy*, *F1-score*, curvas de PR (*Precision-recall*) e curvas ROC (*Receiver Operator Characteristic*) o desempenho da rede neuronal;
- O7. Identificar Limitações e Proposta de Soluções: Investigar as limitações e desafios inerentes ao trabalho e ao treino das redes com dados não padronizados e propor soluções viáveis para estes desafios.

Foram ainda formuladas as seguintes questões de investigação:

- QI1. A partir de um conjunto de imagens histológicas de tecidos mamários, é possível distinguir as saudáveis das patológicas utilizando um sistema de inteligência artificial obtendo um elevado grau de confiança na estimativa?
- QI2. Impacto da Variabilidade de Dados nas Redes Neuronais: De que forma a inclusão de imagens com variabilidade de coloração, artefactos e processamento afeta a capacidade de aprendizagem e de generalização de uma rede neuronal?
- QI3. Comparação com Dados Padronizados: Quais serão as semelhanças ou diferenças dos resultados de redes neuronais treinadas com bases de dados de imagens de rotina em contraste com bases de dados de imagens padronizadas em termos de métricas de análise?
- QI4. Análise de Artefactos Histológicos: Qual é o papel dos artefactos histológicos, como os de fixação, coloração, entre outros na performance da classificação automática de patologias mamárias por redes neuronais?
- QI5. Impacto da aplicação de um sistema automático de apoio à decisão numa rotina laboratorial: Qual será o impacto que uma aplicação que automaticamente traz suporte a decisões num serviço de rotina pode ter para os Patologistas e para os pacientes?

1.3. Opções metodológicas

Para a elaboração deste projeto procedeu-se a uma extensa revisão bibliográfica, realizada através da consulta de bases de dados científicas reconhecidas, como a PubMed, a SpringerLink e a Google Scholar. Esta revisão foi complementada ainda por pesquisas em *websites* especializados e fóruns de discussão, relacionados com *machine learning* e *deep learning* e com a análise de imagens histológicas relacionada com estes tipos de aprendizagem. O propósito desta pesquisa profunda foi de alcançar e compreender quais as mais recentes evoluções práticas no campo da patologia digital e da IA aplicada à saúde.

1.4. Estrutura do trabalho

Neste trabalho foi feita uma descrição acerca dos temas inerentes ao desenvolvimento do mesmo de uma forma generalizada, seguindo-se de um enquadramento e pertinência atual dos temas que justificam o desenvolvimento deste projeto. Fora ainda feita uma descrição dos objetivos principais deste projeto, bem como das questões de investigação que se encontram subjacentes neste desenvolvimento.

O desenvolvimento deste projeto seguiu-se com a realização de um estado da arte sobre as principais características histológicas do Cancro da Mama, as principais questões para a obtenção de uma base de dados robusta, os fundamentos das Redes Neuronais bem como as suas principais aplicações na saúde e terminámos o capítulo com a análises de vários artigos científicos com dados relevantes que podem servir de bases de comparação para o desenvolvimento deste projeto.

Seguidamente foi feita uma descrição exaustiva de todas as metodologias utilizadas desde a aquisição de imagens através da sua digitalização, passando pelo seu pré-processamento e conversão, segmentação, revisão e organização para formação do *dataset*. Descreveu-se ainda sobre a escolha da arquitetura neuronal onde foi realizada a avaliação do desempenho das arquiteturas escolhidas.

Em suma, obtiveram-se gráficos com as principais métricas, como as curvas de *Precision-Recall* (PR) e *Receiver Operating Chatacteristic* (ROC) e tabelas que apresentam a matriz de confusão obtida no final de cada treino. Foi feita uma avaliação dos resultados obtidos tendo em consideração vários fatores, como o número de *epochs*, a arquitetura em causa e a matriz de confusão e foram retiradas as principais conclusões daqueles valores, seguindo-se de uma explicação breve de quais poderão ser as causas inerentes à obtenção dos resultados apresentados.

2. REVISÃO BIBLIOGRÁFICA

As patologias mamárias apresentam uma ampla variedade de morfologias, variando essencialmente entre patologias benignas e patologias malignas. A coloração de HE (Hematoxilina-Eosina) é crucial para distinguir estes padrões teciduais. Neste contexto, as bases de dados histológicas são essenciais para o avanço da patologia digital e da IA, fornecendo um vasto conjunto de imagens histológicas que posteriormente alimentarão os treinos de algoritmos para que estes sejam capazes de identificar todos os padrões celulares. Estes sistemas, apoiados por uma infraestrutura robusta de dados e métodos de pré-processamento avançados, entre outros, estão a revolucionar o diagnóstico médico, oferecendo assim ferramentas cada vez mais precisas para a classificação patológica, melhorando assim a precisão e rapidez do diagnóstico.

2.1. Histologia e Coloração dos tecidos

As morfologias teciduais das patologias mamárias apresentam uma heterogeneidade significativa que pode ser meticulosamente discriminada através da análise histológica. Nas condições benignas das patologias, como na Doença Fibrocística, observam-se alterações proliferativas e regenerativas, evidenciadas por cistos, fibrose e hiperplasia epitelial, mantendo, contudo, a arquitetura lobular. Por outro lado, o Fibroadenoma caracteriza-se pela proliferação benigna de componentes estromais e glandulares, formando nódulos bem delimitados, sem evidências de invasão celular para os tecidos circundantes. Na Figura 2 e Figura 3 podemos ver uma imagem histológica onde estão presentes as principais características destas patologias benignas[15], [16].

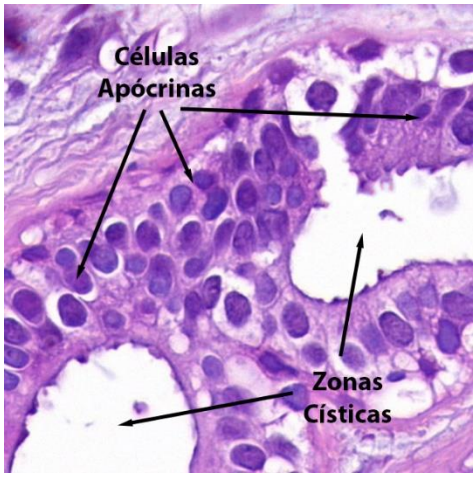


Figura 2 – Imagem histológica de tecido mamário representativo da Doença Fibrocística, onde se pode ver as células apócrinas e as zonas císticas comuns desta patologia

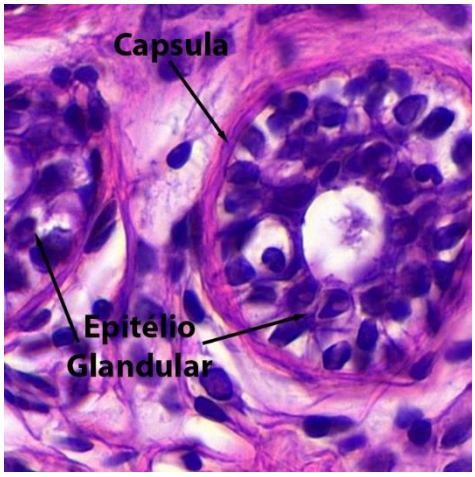


Figura 3 – Imagem histológica de tecido mamário representativo de um Fibroadenoma, onde podem ser vistos os nódulos bem delimitados pela cápsula bem como o epitélio glandular em camada dupla.

Por contraste, as patologias malignas, nomeadamente o Carcinoma Ductal Invasivo e o Carcinoma Lobular Invasivo, são marcadas pela presença de células neoplásicas com atipias nucleares e citoplasmáticas, que rompem a barreira basal e infiltram todo o estroma mamário. No Carcinoma Ductal Invasivo, o padrão de crescimento é tipicamente idiossincrático, expondo as estruturas ductais anormais e uma desorganização arquitetural. O Carcinoma Lobular Invasivo distingue-se pela apresentação de um padrão que pode ser facilmente confundido devido à tendência de apresentar células neoplásicas que se disseminam em filas lineares e normalmente estas células não apresentam tamanhos exagerados como é

típico dos padrões de neoplasia[31], [32]. Na Figura 4 e Figura 5 podemos ver imagens histológicas onde é possível observar as principais características das patologias malignas que foram referidas.

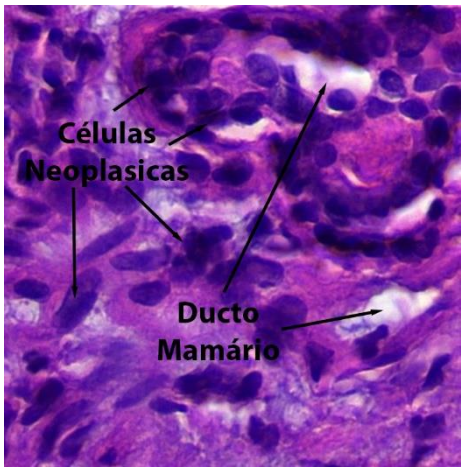


Figura 4 – Imagem histológica de tecido mamário representativo de Carcinoma Ductal Invasivo onde é possível observar as células neoplásicas, uma desorganização arquitetural típica desta patologia e ainda os ductos mamários preenchidos por células neoplásicas

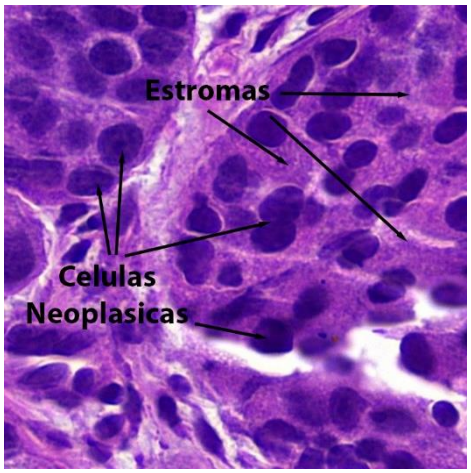


Figura 5 – Imagem histológica de tecido mamário, representativo de Carcinoma Lobular Invasivo onde é possível ver as células neoplásicas numa disposição mais linear e ainda o estroma bem evidente

A coloração de HE desempenha um papel fundamental na distinção destes padrões morfológicos. A Hematoxilina é um corante básico que cora os núcleos das células com uma tonalidade azul-violeta, permitindo assim a identificação de várias atipias nucleares como a hipercromasia, pleomorfismo bem como o aumento da relação núcleo/citoplasma. A Eosina é um corante com ácido responsável pela coloração do citoplasma com uma cor rosada, possibilitando assim a avaliação da coesão celular, da presença de invasão e da fibrose intersticial que possa existir. A combinação destes dois corantes, não só viabiliza a caracterização das alterações patológicas mamárias, como também acentua os contrastes entre o tecido normal e o tecido patológico, fatores essenciais para a delimitação das fronteiras das lesões de forma precisa[33], [34].

2.2. Importância das Bases de Dados na Patologia Digital

As bases de dados histológicas têm desempenhado um papel fundamental na evolução da patologia digital, promovendo os avanços da aprendizagem automática e da IA no âmbito do diagnóstico médico. Estes repositórios digitais, compreendem uma vasta panóplia de imagens histológicas selecionadas especificamente e anotadas, que fornecem a essência para o desenvolvimento e aprimoramento de algoritmos computacionais capazes de realizar análises complexas e precisas, mimetizando a acuidade humana[35], [36].

Os sistemas que estão a ser desenvolvidos na área da identificação celular, quando cupulados à patologia digital, tem permitido a identificação de padrões subtis que frequentemente podem escapar à análise do olho humano. A aplicação da aprendizagem das máquinas com estas bases de dados, não só potenciam a deteção de características morfológicas chave para o diagnóstico, como também contribuem para a descoberta de biomarcadores de prognóstico e terapêuticos, viabilizando respostas mais rápidas para a realização de tratamentos personalizados que os pacientes cada vez mais necessitam[37], [38].

2.2.1. Robustez das Bases de Dados

No contexto da patologia computacional, a criação e manutenção de bases de dados robustas e meticulosamente anotadas é uma necessidade incontornável. A qualidade e a diversidade dos dados histológicos, aliadas à precisão das anotações fornecidas por especialistas, constituem a base fundamental do treino eficaz de algoritmos de aprendizagem automática. É imperativo que tais bases de dados não só abranjam um amplo espectro de variações patológicas, mas também reflitam as nuances inter e intra observador no que toca à interpretação das imagens histológicas[37].

As anotações detalhadas, que incluem, mas não se limitam a diagnósticos, graus de diferenciação, margens e outras características patológicas relevantes, enriquecem o processo de aprendizagem dos algoritmos. Este enriquecimento traduz-se numa capacidade ampliada de reconhecer e aprender a variabilidade inerente às amostras patológicas, o que é determinante quer para o desenvolvimento da precisão no diagnóstico quer para a minimização de falsos positivos ou falsos negativos[35].

São vários os desafios que têm de ser enfrentados para a construção das bases de dados sólidas e robustas nomeadamente, os desafios relacionados com a variação da técnica nas etapas de preparação e análise das lâminas. A heterogeneidade da coloração de HE é essencial para os diagnósticos diferenciais, mas também se caracteriza como um dos obstáculos mais proeminentes. As variações nos protocolos de coloração, relacionadas com a marca dos reagentes que são usados, nos tempos em que o tecido fica mergulhado em cada reagente, bem como, na saturação dos reagentes resultam em diferenças significativas nas tonalidades finais das várias estruturas que compõem o tecido[33], [34].

De realçar ainda, que a preparação das lâminas também pode ser um entrave ao padrão esperado. Aquando do corte do bloco de parafina, apesar do aparelho que o realiza deva estar sempre calibrado, uma vez que a espessura do corte é tão fina, existem vários fatores que podem influenciar nesta espessura como a temperatura a que o bloco se encontra no momento da realização do corte, o estado da faca com que o corte é realizado e até mesmo a experiência e precisão do técnico que executa a tarefa. Estas alterações são sempre muito mínimas, mas como as células também apresentam tamanhos na ordem dos μm qualquer que seja a alteração pode impactar diretamente na morfologia celular e tecidual que é visualizada nas imagens. Após estes passos serem concluídos, também a digitalização das lâminas se torna um desafio, uma vez que existem várias complexidades técnicas ligadas aos equipamentos, nomeadamente a variação da iluminação, a resolução e até mesmo a calibração dos equipamentos visto que são equipamentos muito sensíveis e podem estes fatores introduzir novos artefactos e inconsistência nas imagens obtidas[39].

A necessidade da construção de bases histológicas é uma tarefa complexa e exigente que requer uma equipa multidisciplinar de Técnicos de Anatomia Patológica, Médicos, Engenheiros e outros profissionais qualificados para garantir a robustez destas bases de dados que sejam capazes de proporcionar avanços significativos na automatização e na precisão do diagnóstico na área da Anatomia Patológica.

2.2.2. Pré-processamento dos dados

A garantia da consistência das imagens histológicas bem como a sua harmonização é um fator fundamental para a eficácia dos algoritmos de aprendizagem profunda. A variabilidade intrínseca no processo de aquisição de imagens pode decorrer de vários fatores, todos eles já mencionados anteriormente.

O processo de normalização da coloração das imagens assume uma importância crítica, sendo essencial para reduzir as variações cromáticas entre as imagens obtidas das diversas lâminas. Algoritmos

de normalização de cor, como o método de Reinhard ou *stain normalization* devem ser aplicados para assegurar que as características morfológicas sejam o foco para a aprendizagem do modelo, e não as variações de artefactos de cor. A normalização da coloração procura alinhar as propriedades cromáticas das imagens para uma representação padrão e consistente. Estes métodos permitem então ajustar as cores das imagens de origem para cores que correspondam estatisticamente à coloração de uma imagem de referência predefinida[40].

Antes da introdução das imagens na rede neuronal, deve ser realizado um pré-processamento cuidado, isto inclui o ajuste do contraste e brilho, a correção de artefactos e o recorte das regiões de interesse. O aumento de dados, através de técnicas como rotação, zoom, e *flip* horizontal ou vertical, é também essencial para ampliar o conjunto de dados, proporcionando ao modelo uma aprendizagem mais robusta e generalizada[41], [42].

A verificação e anotação das imagens por patologistas experientes é indispensável para a formação e avaliação dos algoritmos. Este processo deve ser interativo, onde o *feedback* dos especialistas serve para aferir a qualidade das imagens e das anotações, garantindo uma boa representatividade das mesmas.

Após o pré processamento, é imprescindível aplicar técnicas de validação robustas para testar a eficácia do modelo. A utilização de conjuntos de dados de validação e teste independentes permite avaliar a capacidade de generalização do modelo. Na Figura 6 está representado um esquema resumo de todos os parâmetros que devem ser realizados no pré-processamento para garantir que a informação que será utilizada para a aprendizagem é informação de qualidade capaz de refletir bons resultados.

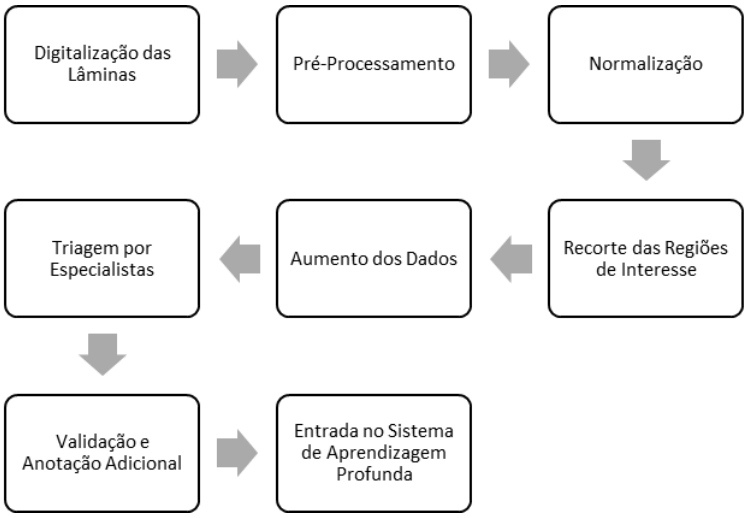


Figura 6 – Representação esquemática dos principais parâmetros do pré-processamento

A implementação destas estratégias é essencial para assegurar a qualidade e a fiabilidade dos sistemas de diagnóstico auxiliados por IA. É imprescindível reconhecer que, sem um pré-processamento adequado, o risco de má interpretação dos resultados é significativamente elevado. Por isso, estas medidas são essenciais para desenvolver modelos preditivos que sejam não só tecnicamente exatos, mas também clinicamente relevantes[20].

2.3. Redes Neuronais

2.3.1. Principais fundamentos

As redes neuronais artificiais (RNAs) constituem uma esfera da IA inspirada nos mecanismos biológicos do cérebro humano. Emergiram como um campo de estudo nas décadas de 1940 e 1950 com o objetivo de simular o processamento neuronal para realizar tarefas de aprendizagem automática. Desde então, o desenvolvimento de algoritmos e o aumento exponencial da capacidade computacional têm contribuído para o aprimoramento e diversificação das RNAs[43].

Uma rede neuronal é uma estrutura computacional que consiste em unidades de processamento, denominados neurónios artificiais, organizadas em camadas. Estes neurónios interligam-se e transmitem sinais, de forma análoga ao sistema nervoso. A origem das RNAs remonta ao modelo de neurónio de McCulloch-Pitts, proposto em 1943, que estabeleceu a base teórica para neurónios binários e a sua capacidade de realizar cálculos lógicos simples[43].

Existem diversas tipologias de RNAs, classificáveis de acordo com a sua arquitetura e o tipo de aprendizagem. Incluem redes de *feedforward*, onde a informação se move estritamente num sentido progressivo; redes recorrentes, que têm conexões cíclicas; redes convolucionais, especializadas em processar dados com uma topologia de grade, como imagens[44], [45].

O funcionamento de uma RNA inicia-se com a introdução de dados no sistema, seguida pelo seu processamento sucessivo através de várias camadas ocultas. Em cada neurónio, realiza-se o cálculo de uma soma ponderada das entradas, e subsequente aplicação de uma função de ativação que introduz uma componente não linear ao processo. A fase de treino da rede implica a afinação dos pesos sinápticos, uma tarefa efetuada com base na minimização do erro da saída gerada. Esta otimização é frequentemente realizada através de algoritmos de retro propagação do erro ou *backpropagation*, que ajustam os pesos de forma interativa com o objetivo de melhorar a precisão da rede. Na Figura 7 podemos ver uma representação e comparação entre a estrutura e funcionamento de um neurónio biológico e a sua versão artificial numa rede neuronal.

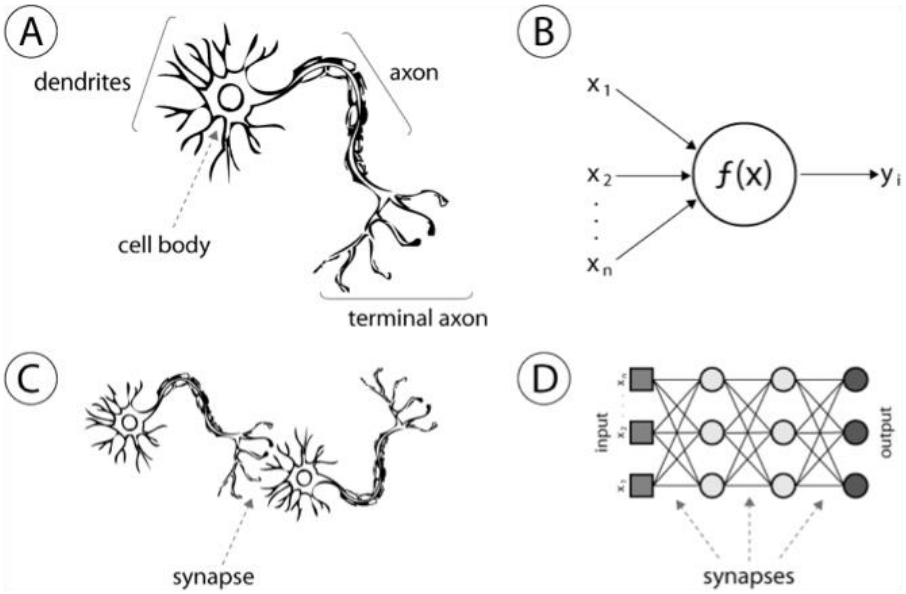


Figura 7 – A imagem A representa a estrutura básica de um neurónio biológico; A imagem B ilustra um modelo abstrato de um neurónio artificial capaz de receber várias informações; A imagem C demonstra a sinapse que é o

ponto de comunicação entre dois neurónios biológicos e a imagem D apresenta uma RNA que são uma coleção de neurónios artificiais interligados, "comunicando" entre si.

O processo de treino de uma RNA é complexo e exige um grande número de dados. Aspetos como o *overfitting*, que ocorrem quando a rede aprende de forma excessivamente específica os dados de treino em detrimento da capacidade de generalização. Para atenuar essas adversidades, recorre-se a estratégias como a validação cruzada, e técnica de *dropout* (método de regularização) e a normalização por lotes[46].

Ainda de realçar que no âmbito da saúde a capacidade de interpretar os modelos de RNA é de extrema importância. Sabendo que o funcionamento interno das RNAs pode ser opaco, ou seja, os processos através dos quais elas chegam a uma determinada conclusão ou previsão não são facilmente compreensíveis, mesmo para os criadores do modelo, o que origina por vezes várias questões éticas.

2.3.2. Aplicações na Saúde

As RNAs e as CNNs constituem dois paradigmas distintos dentro do domínio da aprendizagem profunda, cada uma com características e aplicabilidades específicas. As RNAs são sistemas de *deep learning* que imitam a arquitetura do cérebro humano, permitindo que a máquina aprenda a partir de dados de entrada. As CNNs são uma especialização das RNAs projetadas especificamente para processar dados com uma clara organização espacial e hierárquica de características, como é o caso de imagens. Utilizam filtros convolucionais para percorrer os dados de entrada e gerar mapas de características que encapsulam a presença de padrões específicos em distintas áreas dos dados. Acrescentam-se, ainda, camadas de *pooling* que diminuem a dimensão dos dados enquanto preservam as feições mais marcantes, dotando as CNNs da capacidade de reconhecer padrões independentemente da sua posição na imagem[27], [37].

No âmbito da saúde, as CNN sobrepõem-se às RNAs por múltiplas razões. Primeiramente, a natureza das imagens médicas que podem englobar tomografias, ressonâncias magnéticas, imagens histológicas, entre outras, requer uma análise e interpretação de padrões complexos e detalhados, muitas vezes de natureza espacial e hierárquica. As CNNs têm a capacidade de deter tais padrões, assimilando características em diversos níveis de abstração, o que é essencial para um diagnóstico preciso[22], [24].

Por contrapartida, a competência das CNNs para manipular grandes quantidades de dados e a sua eficiência computacional ao processarem imagens amplas e complexas tornam-nas particularmente aptas para a prática médica, onde o volume de dados é substancial e a precisão é essencial. A resiliência das CNNs face a variações e distorções nos dados também representa um papel crucial, uma vez que as imagens médicas podem diferir significativamente devido a disparidades das imagens nas variadas áreas médicas devido a fatores externos[24], [26].

No contexto da área da saúde, a capacidade de extrair padrões complexos de imagens assentam essencialmente nas seguintes arquiteturas: SqueezeNet; InceptionV3; AlexNet; DenseNet; EfficientNet; MobileNet V3 e ResNet.

A arquitetura SqueezeNet caracteriza-se pelo seu design inovador que visa minimizar o número de parâmetros enquanto preserva a precisão analítica. Uma característica fulcral é o módulo "fire" que constitui o núcleo desta arquitetura. A eficiência é alcançada através da implementação de camadas de "squeeze", que realizam convulsões 1x1 para reduzir a profundidade dos dados de entrada, seguidas por camadas "expand", compostas por convulsões 1x1 e 3x3, para reconstruir a profundidade da representação. Esta estratégia resulta numa arquitetura compacta que demonstra a capacidade de manter

a precisão, enquanto reduz significativamente a quantidade de parâmetros e a complexidade computacional[47]–[49].

A Figura 8 [50] apresenta uma representação esquemática da arquitetura SqueezeNet. O processo inicia-se com uma imagem de entrada que segue através de uma série de camadas convolucionais (C1 e C2), módulos "fire" (Fire1 a Fire4), e operações de *mac pooling* (P1, P2 e P3). Conclui-se com um vetor de saída que corresponde às classes de previsão. Este fluxo de operações exemplifica a capacidade do SqueezeNet em condensar um modelo de aprendizagem profunda sem comprometer o seu desempenho analítico.

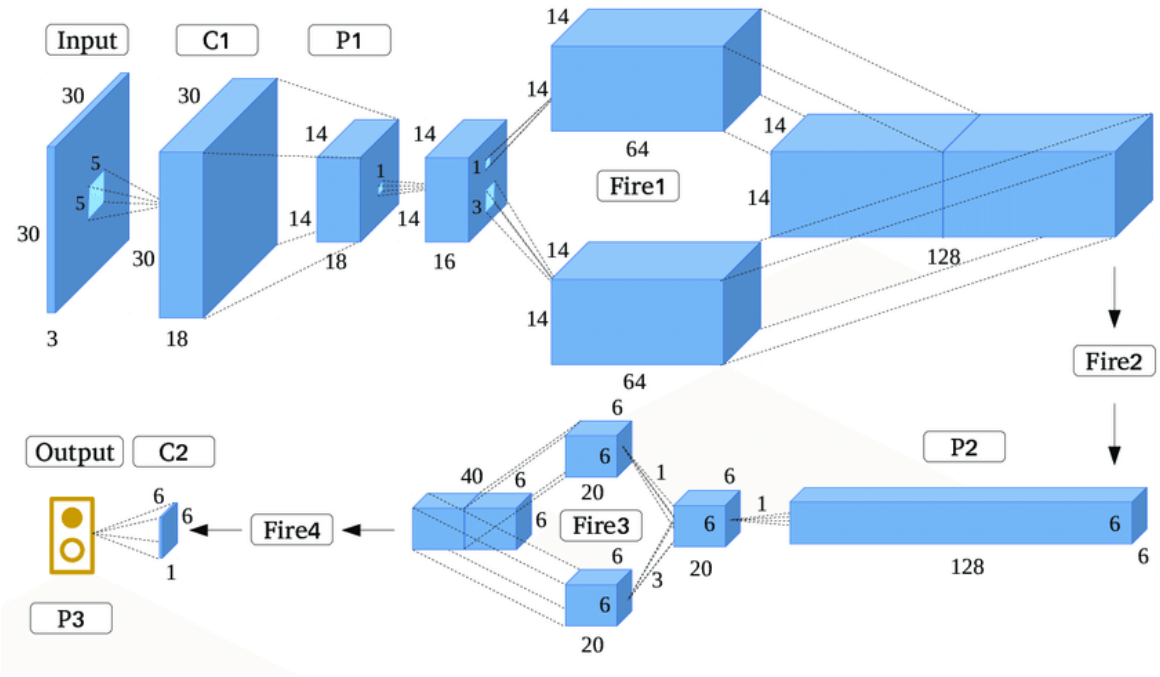


Figura 8 – Representação gráfica da arquitetura SqueezeNet[50]

A InceptionV3, representa um avanço significativo na arquitetura de redes neuronais convolucionais. Esta estrutura é delineada para otimizar o desempenho computacional e a acuidade classificativa, através da disposição e paralelização inteligente dos módulos "Inception". Estes módulos combinam operações de convolução e *pooling* que operam em diferentes escalas e são fundidos por encadeamento, favorecendo a captura de informações em múltiplas resoluções espaciais[51], [52].

A Figura 9 [53] ilustra a orquestração das operações na InceptionV3. Observa-se que, subsequente à entrada de dados, a imagem é submetida a uma série de transformações convolucionais, seguidas de processos de *pooling* que visam a compactação espacial das características. O encadeamento das características é uma etapa crucial, integrando dados de múltiplas fontes de convulsão e *pooling*. Posteriormente, a rede emprega camadas totalmente conectadas para interpretar o conjunto de características extraídas. A inclusão de camadas de *dropout* serve para mitigar o sobreajustamento, e finalmente, a camada *softmax* converte as saídas da rede numa distribuição de probabilidade, fundamental para a tarefa de classificação[54].

Desta forma, a arquitetura de InceptionV3 demonstra a eficácia de um design modular e paralelo, que não só promove a diversidade de características extraídas, mas também contribui para a redução do número de parâmetros sem comprometer a profundidade e complexidade da rede.

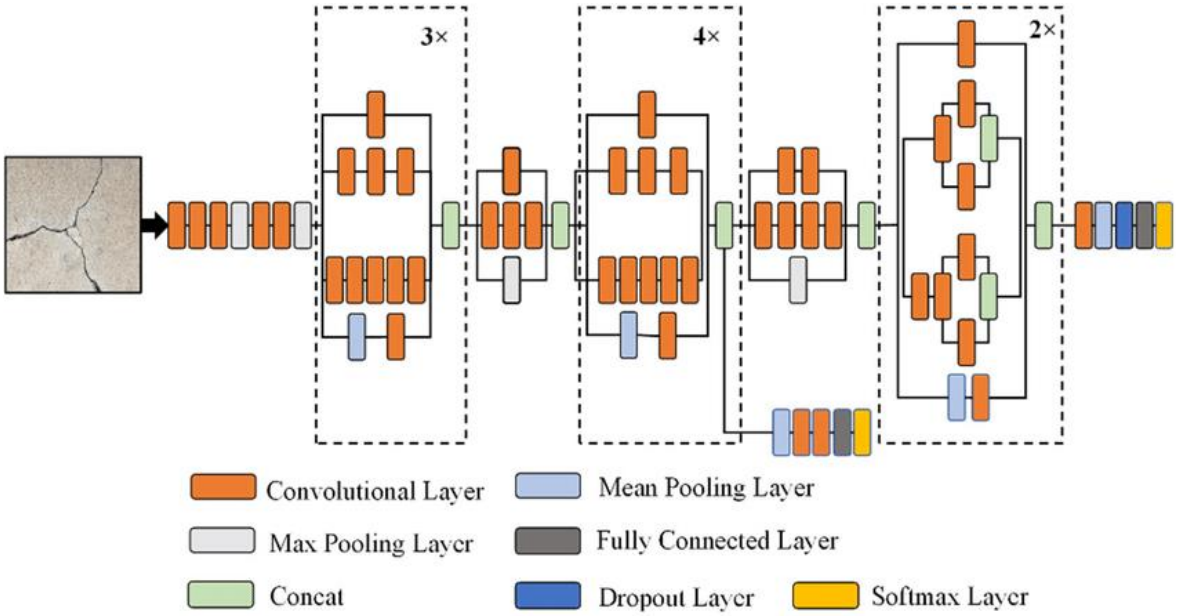


Figura 9 – Representação gráfica da arquitetura Inception V3 [53]

AlexNet, é uma das primeiras arquiteturas de aprendizagem profunda e que ainda é relevante devido à sua simplicidade e eficácia. A técnica de *dropout* que introduziu é um componente essencial para evitar o sobre ajuste, uma apreciação importante no treino de redes com conjuntos de dados médicos, que muitas vezes são limitados[55]–[57].

A DenseNet fornece uma abordagem única de conexão entre camadas, o que promove a reutilização de características e melhora a eficiência da rede, aspetos valiosos na análise detalhada de imagens histológicas, por exemplo[58], [59].

A EfficientNet, por meio da sua abordagem de escalamento sistemático, redefine os parâmetros de eficiência e acuidade. No contexto da saúde, esta capacidade traduz-se na viabilização de modelos computacionais de elevada potência mesmo em ambientes restritos em termos de recursos, como é o caso de hospitais cujas infraestruturas tecnológicas são menos desenvolvidas[60].

MobileNetV3 é projetada especificamente para dispositivos móveis e portanto, adequada para telessaúde e diagnósticos em locais remotos. A sua eficiência computacional é de particular importância para o processamento de imagens em aplicações móveis de saúde[61].

Por último, a ResNet permite treinar redes extremamente profundas sem perder eficácia na aprendizagem. Isso é essencial para situações de problemas de saúde complexos, onde a capacidade de aprender padrões profundamente enraizados pode ser a chave para identificar corretamente situações raras[55], [58].

2.4. Estudos no âmbito da Inteligência Artificial e o Cancro da Mama

Na Tabela 1 estão citados vários artigos científicos onde foram utilizadas diferentes redes neuronais para diferentes tipos de tecido mamário (normal, benignos e malignos) e de onde foram retiradas várias métricas de avaliação dos resultados obtidos.

Podemos observar que as melhores percentagens avaliativas encontram-se essencialmente nos estudos em que foi feita uma classificação binária quando comparados com outros estudos que fazem uma classificação mais complexa. Também é possível perceber uma evolução ao longo dos anos no que diz respeito às redes neuronais utilizadas, nomeadamente nos anos mais recentes o uso de novas versões

de algumas redes neuronais e até mesmo o uso de redes neuronais mais específicas como é o caso da MSMV-PFENet (*Multi-Scale Motion Vectores Pyramid Feature Extraction*).

Tabela 1 – Tabela comparativa dos principais dados recolhidos de alguns artigos científicos onde foram utilizadas diversas arquiteturas neuronais, com várias classes de patologias mamárias e de onde foram recolhidos os dados da avaliação de desempenho

Ano	Pre-Processamento	Redes Neuronais	Número de Classes:	Métricas avaliadas	Ref.
2019	Macenko	CNN;	4 classes: 2 patologias benignas; 2 patologias benignas;	*Accuracy*:87%	[62]
2019	Macenko	GoogleNet, Visual Geometry Group Network (VGGNet) e ResNet;	2 classes: Maligno; Benigno	*Accuracy*: GoogleNet – 93,5%; VGGNet – 94,15%; ResNet – 94,35%;	[63]
2020	Color enhancement	ResNet50, DenseNet, Inception V3 e VGG16;	4 classes: Benigno; Carcinoma in situ; Carcinoma Invasivo; Tecido Normal;	*Accuracy*: ResNet – 87%; DenseNet-87%; InceptionV3-83%; VGG16-78,1%;	[64]
2020	Não aplicado	MobileNet e Inception;	6 classes: Tecido Normal; Fibroadenoma; Doença Fibrocística; Carcinoma invasivo de tipo não específico; Carcinoma Lobular Invasivo; Carcinoma Ductal In Situ;	Sensibilidade: MobileNet-80,5%; Inception-73,8%; Especificidade: MobileNet-96,1%; Inception-94,7%;	[65]
2022	Não aplicado	MSMV-PFENet;	4 classes: Tecido Normal; Benigno; Carcinoma In Situ; Carcinoma Invasivo;	*Accuracy*: Imagens-94,8%; Patch- 93%;	[66]

Como forma de conclusão da revisão de vários artigos, saliento o artigo "*Artificial Inteligence Improves the Accuracy in Histologic Classification of Breast Lesions*", uma vez que faz uma avaliação de como os sistemas de IA podem ser um auxílio essencial para melhorar a precisão dos Patologistas nos diagnósticos.

- ***Artificial Inteligence Improves the Accuracy in Histologic Classification of Breast Lesions*** **(**[67]**)**

Neste estudo, investigou-se o impacto do uso de algoritmos de aprendizagem na precisão da classificação de Carcinoma Invasivos de Mama (CIS) por patologistas, com particular ênfase na distinção entre CIS microinvasivos e CIS não microinvasivos em fotografias microscópicas de regiões de interesse. O estudo foi desenvolvido em três fases diferentes, com patologistas com diferentes níveis de experiência, e utilizou-se um algoritmo de aprendizagem automática denominado de Algoritmo A, o qual foi posteriormente comparado com a classificação padrão sem auxílio da ferramenta, denominado de Algoritmo B.

Na primeira fase, observou-se que a utilização do Algoritmo A aumentou significativamente a precisão dos patologistas na identificação de CIS, sugerindo que ferramentas de diagnóstico auxiliadas por IA podem ser valiosas para reduzir resultados falsos-negativos e melhorar a saúde do paciente. Na segunda fase, mesmo quando a precisão do algoritmo A era menor que a dos patologistas mais experientes, estes mantiveram a sua classificação inicial, sugerindo uma confiança na própria avaliação acima do algoritmo. Contudo, os patologistas menos experientes, mostraram-se mais propensos a alterar as suas classificações em favor das sugestões do algoritmo.

A terceira fase não foi conduzida devido à perceção de que a precisão dos patologistas seria superior à do próprio algoritmo. A pesquisa destacou também a consistência dos patologistas sobre os erros potenciais do Algoritmo A, o que levou a um aumento do número de diagnósticos alternativos e à melhoria na precisão, especialmente entre os patologistas com maior experiência.

Um ponto relevante neste estudo foi que, em certos casos, dois patologistas tinham uma consistência de classificação inferior entre si comparados com as classificações do algoritmo, o que sugere que diferentes patologistas podem ter interpretações variáveis mesmo com a ajuda de algoritmos de diagnóstico auxiliado por computador.

As conclusões do estudo indicam que a implementação de sistemas de diagnósticos auxiliados por computador, em diagnósticos patológicos, pode melhorar a precisão, especialmente com uma supervisão adequada, e que os patologistas podem beneficiar desta tecnologia como um complemento à sua análise e não de um substituto. Este estudo foi pioneiro na comparação do impacto dos algoritmos de aprendizagem na precisão dos diagnósticos na área da Anatomia Patológica.

Nos estudos realizados entre 2019 e 2020, destacou-se a eficácia de vários modelos de redes neuronais convolucionais como várias de CNN, GoogleNet, VGGNet, ResNet50, DenseNet, InceptionV3, VGG16 e MobileNet na classificação de patologias mamárias. Estes modelos demonstraram uma alta precisão, especialmente na distinção entre patologias benignas e malignas, com resultados promissores que evidenciam a capacidade destas tecnologias em melhorar o diagnóstico. Em 2022, a introdução do MSMV-PFENet marcou um avanço significativo, alcançando elevadas taxas de accuracy na classificação de patologias mamárias, refletindo uma evolução contínua e o aprimoramento das técnicas de IA na área da saúde.

O estudo "*Artificial Intelligence Improves the Accuracy in Histologic Classification of Breast Lesions*", mostra o impacto que a IA tem no diagnóstico de CIS. Este estudo demonstra que a precisão dos diagnósticos pode aumentar com o auxílio de ferramentas de diagnóstico, reduzindo falsos-negativos especialmente em patologistas com menos experiência. Assim, é possível perceber que a IA deve ser usada como um complemento à avaliação humana e que deve ser feita sempre uma abordagem integrada que combine a experiência humana com as novas ferramentas.

3. MÉTODOS E APLICAÇÃO

Para o desenvolvimento deste projeto foi efetuada uma recolha, no Laboratório Dr. Eduardo da Silva Ferreira, de 100 diferentes cortes histológicos efetuados num bloco de parafina com uma dimensão de 3μm (micrómetros) de espessura de tecido de biópsia de mama previamente recolhido e processado. Seguidamente os cortes foram corados na coloração de HE juntamente com o serviço de rotina diária do laboratório. Depois de coradas as lâminas foram montadas utilizando *Entellan*, um meio de montagem que nos permite proteger o corte histológico e uma lamela.

Posteriormente seguiu-se com a digitalização das lâminas no digitalizador *Pannoramic Midi ll*, com uma ampliação de 20x. Esta digitalização permitiu-nos nos a visualização da imagem no programa *SlideViewer*, onde era possível ampliar até 60x a imagem e ver as estruturas dos tecidos com a nitidez necessária.

Após a obtenção das digitalizações foi realizada uma conversão do formato das imagens no software *QuPath,* seguindo com segmentação das imagens com uma dimensão de 512x512 pixéis e ainda foi efetuado uma revisão e organização de todas as imagens de forma a criar o *dataset*. Posteriormente foi efetuada uma preparação dos dados obtidos, seguindo da distribuição da mesma, finalizando com o processamento dos dados nas arquiteturas neuronais.

Numa perspetiva geral, este capítulo irá focar-se na descrição do procedimento do estudo em causa. Inicialmente, descreverá o procedimento laboratorial desde a manipulação das biópsias até à digitalização final das lâminas. De seguida, irá ser descrito todo o processo de manipulação de imagens, criação do *dataset* e preparação dos dados para o treino do modelo. Antes de prosseguir com o treino, são apresentados alguns dos modelos disponíveis para treino, uma breve explicação do porquê das escolhas dos modelos SqueezeNet e InceptionV3, e ainda uma apresentação dos cenários de classificação que vão ser abordados. No final é apresentado como é que o treino se processa e como é feita a sua validação.

3.1. Aquisição de imagens

A criação de um *dataset* de imagens histológicas a partir de biópsias de mama é um processo meticuloso e detalhado, que envolve várias etapas desde a colheita do tecido até à aquisição de imagens de alta qualidade.

Assim, a primeira etapa é a realização da biópsia mamária, que pode ser feita por diferentes métodos, biópsia por agulha fina, biopsia por agulha grossa (denomina-se de core biópsia) ou uma biópsia excisional. Para a criação deste *dataset* foram recolhidas imagens somente de core biópsia. Este procedimento é realizado por um especialista que procura garantir a integridade do tecido e minimizar da melhor forma os dados da arquitetura tecidular.

Após a recolha dos tecidos, os fragmentos devem ser imediatamente submersos numa solução fixadora, a mais utilizada é o formol a 10%, onde se deve utilizar uma quantidade de formol a 10% numa proporção 10x superior ao tamanho de qualquer tecido que seja recolhido. A fixação é crucial para preservar a estrutura celular e a morfologia do tecido e evitar a degradação enzimática que acontece após a excisão do tecido.

Os tecidos seguiram então para o processamento que consiste numa desidratação feita por uma série crescente de álcoois e segue para o processo de diafanização realizado pelo xilol. Este reagente é responsável pela preparação final do tecido para que este possa ser embebido em parafina. Este processo tem uma duração aproximada de 12H e após o processamento, o tecido segue para o passo da inclusão

onde este é emoldurado num bloco de parafina que nos possibilita a obtenção de um bloco sólido essencial na fase seguinte.

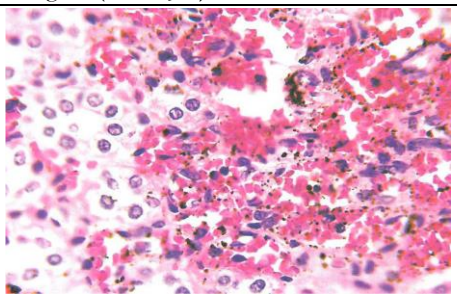
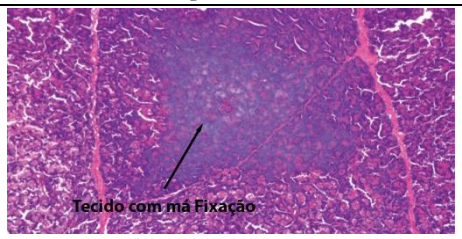
O processo seguiu-se com o corte do bloco que foi realizado no micrótomo para obter secções finas de aproximadamente 3 μm de espessura. Esta espessura permite que seja possível uma visualização clara da estrutura do tecido e da sua configuração celular. O corte obtido foi então colocado numa lâmina de vidro que seguiu posteriormente para a etapa de coloração por HE.

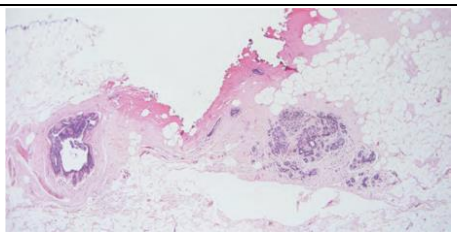
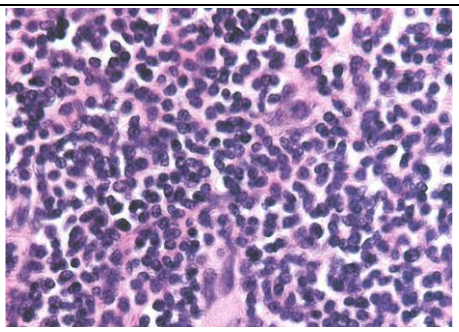
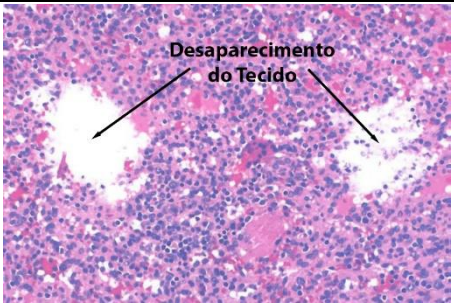
Assim que todos estes passos estão concluídos, as lâminas são montadas. Esta é a fase final da preparação dos tecidos, que consiste na colocação de uma lamela, ou de outro material, que permita a observação microscópica do tecido e simultaneamente a sua preservação.

Por último as lâminas foram observadas por um patologista responsável pelo diagnóstico final da paciente e uma vez que o objetivo era a criação de uma base de dados com estes tecidos, as lâminas foram digitalizadas, com auxílio do digitalizador *Pannoramic Midi ll*, a uma ampliação de 20x.

No decorrer de todos os procedimentos acima descritos, a intervenção humana é preponderante na execução da maioria das tarefas. Esta dependência aumenta, consequentemente, a probabilidade da ocorrência de erros. Nem todas as lâminas preparadas resultam em imagens de qualidade ótima, essencial para um diagnóstico assertivo. Vários fatores podem conduzir à aquisição de imagens inapropriadas para diagnóstico, comumente referidas como "imagens de lixo". Na Tabela 2, é apresentado um resumo de algumas problemáticas que podem emergir no contexto laboratorial. Esta tabela inclui exemplos ilustrativos destas questões, acompanhados de uma explicação sucinta e fundamentada acerca das causas subjacentes a estas situações.

Tabela 2 – Resumo dos problemas mais comuns observados na rotina laboratorial, acompanhados das causas e de imagens representativas

Tipo de Problema	Imagem (descrição)	Causa
Fixação	 Tecido marcado com pigmentos acastanhados resultantes do reagente de fixação.	Presença de pigmentos de formalina num corte histológico, devido a uma possível reação da formalina com a hemoglobina ou devido ao tempo prolongado de fixação. Os pormenores celulares acabam por ficar invisíveis;
Fixação	 Tecido que apresenta uma zona central pouco específica devido a situações de má fixação	Má fixação do tecido na região central da imagem pode dever-se a peças cirúrgicas de grande dimensão onde o formol não foi capaz de penetrar e fixar aqueles grupos celulares. Não é possível perceber as características celulares naquela zona

Coloração	 Coloração de HE ténue quando comparada com as outras imagens	Coloração de HE muito ténue onde as estruturas celulares são de difícil observação. A coloração de HE normalmente é ajustada de acordo com os gostos dos patologistas, a intensidade da marcação dos corantes varia de local para local, no entanto nesta situação é possível que este tecido tenha sido corando em corantes muito saturados e o tempo que o tecido esteve em contacto com o corante foi insuficiente para a sua correta coloração.
Descalcificação	 Tecido mamário com microcalcificações- presença de tecido muito fragmentado, na zona central, onde não é possível destingir algumas das características celulares daquela zona.	Quando o fragmento seccionado é de tecido ósseo ou tecidos com microcalcificações é comum este tipo de imagens histológicas onde não é possível observar as estruturas celulares nem tecidulares, uma vez que, o tecido ósseo apresenta uma grande rigidez e quando não é feita uma correta descalcificação, após o corte deste tecido no micrótomo, o aspeto é o que pode ser observado na imagem.
Microtomia	 Representação de tecido com uma espessura superior àquela que é considerada ótima. Existem várias camadas de células sobrepostas.	Imagem histológica onde se observa uma sobreposição grande das células, impossibilitando ver os detalhes celulares devido a ter sido feito um corte no micrótomo com uma espessura superior ao protocolado.
Microtomia ou Processamento	 Representação de um tecido onde existem áreas em que este está ausente.	Imagem histológica onde se observam áreas de tecido ausente, que podem ser áreas importantes para o diagnóstico final. Estas situações podem dever-se a causas como um mau processamento ou erros na microtomia.

3.2. Pré-processamento

Após o processo laboratorial dos tecidos em estudo e da digitalização das lâminas, foi necessário fazer um pré-processamento das imagens obtidas das digitalizações das lâminas. Para isso foi necessário recorrer a conversão do formato obtido na digitalização, e a segmentação das imagens para que seja usado imagens com tamanho 512x512 pixéis. Os próximos subcapítulos vão se focar nestas duas fases importantes para a preparação do *dataset* que servirá para o treino dos modelos.

3.2.1. Conversão

O processo de conversão constitui uma etapa crucial para a manipulação eficiente e a extração de informação relevante.

Numa fase inicial, o processo de conversão foi realizado com a utilização do software *SlideViewer* (Figura 10), um programa especializado na leitura de ficheiros no formato .mrxs. Este formato, comum em imagens de patologia digital, contém dados de alta resolução que são essenciais para um diagnóstico preciso e para pesquisas de qualidade. Através do *SlideViewer*, conseguimos aceder não apenas a imagens das lâminas, mas também às etiquetas a elas associadas. Estas etiquetas são fundamentais, pois nelas consta o número do caso, um dado crucial que assegura a correta correspondência entre a imagem digitalizada e a sua origem clínica, permitindo um seguimento exato e uma organização rigorosa dos dados.

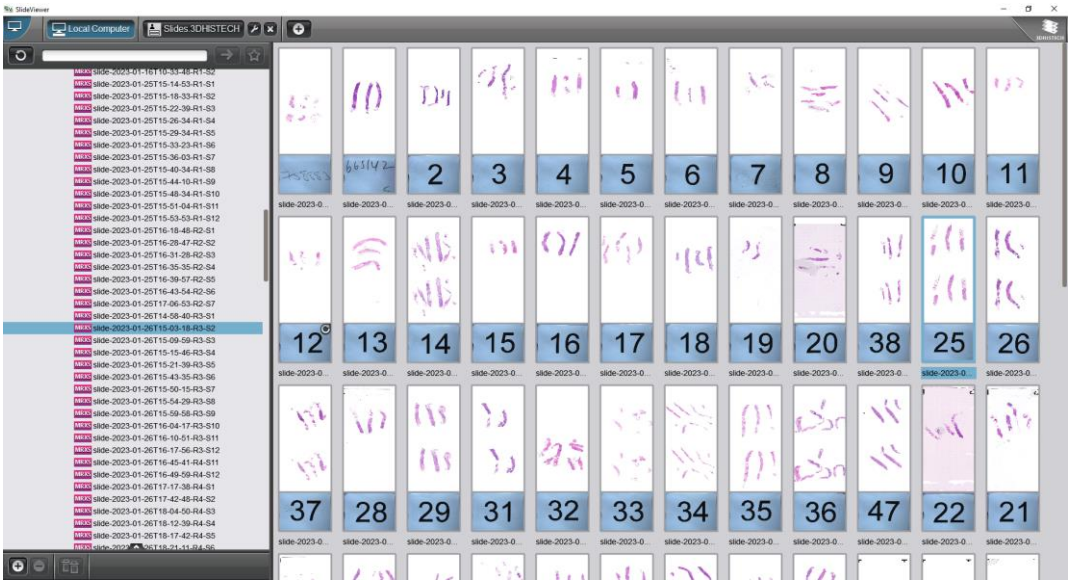


Figura 10 – Ambiente de utilizador do *SlideViewer*

Porém, a extensão é proprietária, pelo que é escasso ou quase inexistente ferramentas ou bibliotecas que facilitem o uso deste formato para treino de modelo de redes neuronais. Para isso é necessário converter a extensão .mrxs numa imagem com extensão JPG. Para isso, procedemos com o uso do *QuPath* (Figura 11), um software de código aberto concebido para a análise de imagens patológicas. No contexto deste caso, aplicamos esta ferramenta com o intuito de converter imagens do formato original .mrxs para .ome.tif. Este formato, derivado do *Open Microscopy Environment*, é amplamente reconhecido pela sua interoperabilidade e adequação à análise computacional em larga escala. O *QuPath* não só facilita

esta conversão, mas também preserva as características essenciais das imagens, como a resolução e a profundidade de cor, assegurando que a informação diagnóstica não seja comprometida durante o processo.

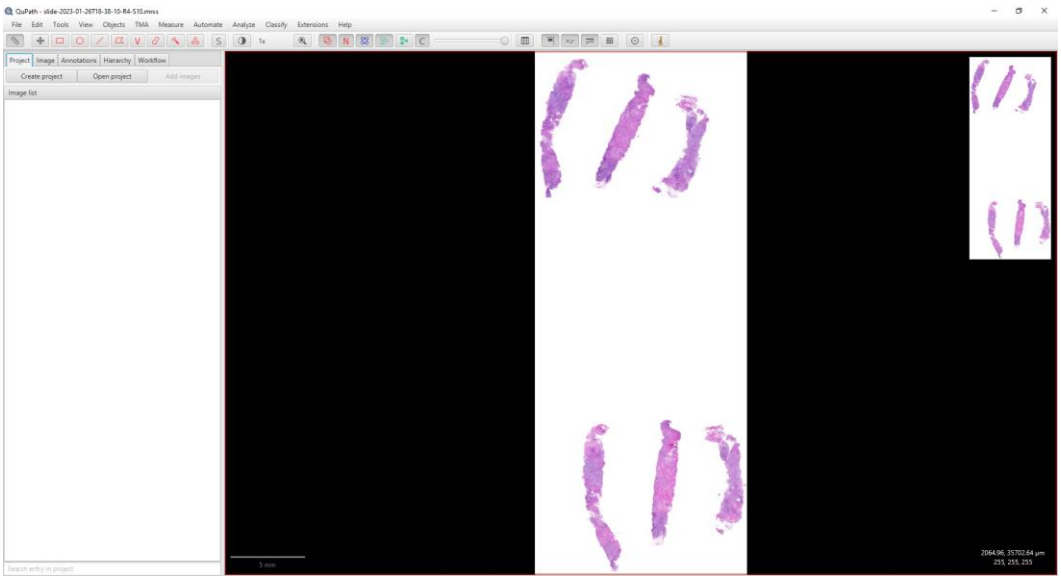


Figura 11 – Ambiente de utilizador do *QuPath*

3.2.2. Segmentação de imagens

Apesar das imagens do tipo 'ome.tif' poderem ser facilmente convertidas em imagens do formato .jpg, estas iriam resultar em imagens com resoluções muito elevadas e com tamanho grande de memória, pois cada imagem poderia ter dezenas de *Gigabytes*. Para além dessa situação, haveria dois problemas na altura de proceder com o treino do modelo. O primeiro problema seria a quantidade de imagens que seria muito pequena para o *dataset*. O segundo problema seria que, na altura do treino, as imagens ao serem grandes, seria necessário um processamento computacional muito grande para o modelo ser treinado com sucesso.

Para resolver ambos os problemas, foi necessário implementar um programa automatizado que opera sobre as imagens 'ome.tif'. Este script foi projetado para recortar as imagens em segmentos de dimensões 512x512 pixéis (Figura 12), um tamanho adequado para análise padrão, facilitando o processamento e a análise subsequente por algoritmos de redes neuronais.

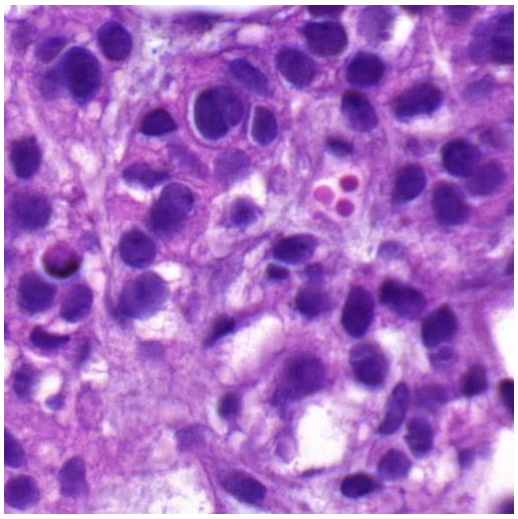


Figura 12 – Exemplo de uma imagem recortada em 512x512 pixéis (imagem não está à escala)

Adicionalmente, o script possui uma função de seleção que descarta qualquer segmento de imagem que não contenha conteúdo relevante, nomeadamente segmentos que se apresentem completamente brancos. Estas imagens são descartadas pois não adicionam valor analítico e economizam espaço de armazenamento. As imagens que contêm dados relevantes são então gravadas no formato .jpg.

Assim, para além de serem resolvidos os problemas da extensão e do tamanho das imagens, ao segmentá-las foi possível aumentar em grande número a quantidade de entradas para o modelo de treino da rede neuronal.

3.2.3. Organização das imagens convertidas

A organização dos dados convertidos foi estruturada de maneira que a construção do *dataset* fosse o mais simplificada possível. As imagens no formato .jpg são armazenadas dentro de pastas nomeadas segundo o número do caso correspondente ao paciente, conforme indicado nas etiquetas iniciais. Esta organização permite uma identificação rápida e eficiente das imagens, essencial para a organização dos dados. A correta organização e catalogação das imagens é importante, visto que assegura a integridade, veracidade e qualidade do estudo e dos seus resultados obtidos. Uma incorreta submissão e organização dos casos poderia invalidar, impossibilitar a obtenção de resultados, ou até mesmo falsear os resultados e conclusões deste estudo.

Para além da organização das imagens por caso, foi necessário confirmar as imagens de cada caso para apagar imagens não relevantes que, por variadas razões, como impurezas na digitalização ou erro de código, possam ter escapado ao programa de segmentação anterior.

3.2.4. Criação do *dataset*

Para a criação de um *dataset* é necessário guardar toda a informação referente a uma imagem num ficheiro que qualquer código consiga ler e interpretar a informação gravada. A melhor maneira e a mais comum em treinos de redes neuronais é a gravação das informações em ficheiros *Comma-separated Values* (CSV), que como o nome indica, são ficheiros com valores que estão separados por vírgulas.

A construção de um arquivo .csv serve como um meio eficiente para compilar e estruturar informações de forma que facilite análises e processamentos computacionais posteriores.

Para isso foi construído um *script* em *Python* que serve para automatizar o processo de classificação de acordo com os valores de entrada gerados pelo utilizador, de acordo de que caso se esteja a tratar.

O procedimento inicia com a definição de uma função "atribuir_categoria", responsável por recolher a classificação das imagens médicas, baseada na interação do utilizador, que compara o número do caso com o caso clínico. Esta função é projetada para garantir que cada diretório de imagens seja associado a uma categoria específica: benigna ou maligna. De seguida o utilizador define o tipo de imagem dependendo se a categoria anterior é benigna ou maligna. Se for benigna, o utilizador define se a imagem é de um caso de Fibroadenoma (representado por 1) ou doença Fibrocística (representado por 2). Se o caso for maligno, então escolhe entre um caso de Carcinoma ductal (representado por 1) ou Carcinoma Lobular (representado por 2).

O ficheiro CSV é então criado através da função "criar_csv". Com o uso da biblioteca "csv" do *Python*, um objeto "DictWriter" é inicializado com os nomes dos campos correspondentes às categorias de classificação. Um cabeçalho é escrito no início do arquivo para facilitar a compreensão e o acesso aos dados.

Através da função "os.walk", o código percorre todos os subdiretórios contidos no diretório raiz fornecido. Para cada conjunto de arquivos num subdiretório, a função "atribuir_categoria" é invocada para classificar o diretório atual. Esta classificação é aplicada a cada imagem dentro do diretório onde estas se encontram.

Para cada imagem válida (com a extensão .jpg), é escrito um novo registo no arquivo .csv. Este registo contém o caminho completo da imagem e as categorias atribuídas, criando assim um mapeamento direto entre a imagem e suas características de classificação do caso em que a imagem se insere.

3.3. Distribuição e Amostragem para Treino e Validação de Modelos de Redes Neuronais

Para que o treino da rede neuronal seja bem sucedido, é importante uma preparação adequada dos dados. Para isso é necessário garantir a aleatoriedade dos dados, bem como um equilíbrio dos dados entre as várias características que queremos que o modelo preveja. Durante este subcapítulo vai ser abordado como foi realizado a distribuição e amostragem dos dados para o treino e validação dos modelos.

3.3.1. Preparação dos dados

A preparação dos dados inicia-se com a definição de uma semente aleatória (*seed*), garantindo a reprodutibilidade das amostras aleatoriamente selecionadas durante a execução do código. A importação do dataset é feita através da função "read_csv" da biblioteca "pandas", que permite uma manipulação eficaz dos dados anteriormente definidos para cada caso.

3.3.2. Distribuição de dados

Antes de proceder à amostragem, é essencial entender a distribuição das imagens pelas diferentes categorias: malignas ou benignas, e as subcategorias correspondentes a carcinoma (ductal e lobular) e a benigno (fibroadenoma e doença fibrocística). Através da função "value_counts", é possível obter a contagem exata de imagens para cada categoria no dataset, o que é crucial para a definir a alocação equilibrada de imagens nos conjuntos de treino e validação, onde é possível ver um exemplo na Figura 13.

```
Distribuição de mal_ou_beg: 1    131301
0     55327
Name: mal_ou_beg, dtype: int64
Distribuição de carcinoma: 1    75975
2    55326
Name: carcinoma, dtype: int64
Distribuição de benigno: 1    30320
2    25007
Name: benigno, dtype: int64
```

Figura 13 – Exemplo da distribuição dos dados pelas categorias

3.3.3. Alocação de Amostras

Após a distribuição dos dados é necessário fazer a alocação das amostras para o treino e validação do modelo. Teoricamente, o valor base correto a utilizar para calcular a quantidade máxima necessária de entradas sem desequilibrar os dados de treino e de validação seria utilizar a subcategoria cuja quantidade de imagens do *dataset* seja menor. De acordo com a imagem Figura 13 , podemos constatar que a subcategoria com menor quantidade é a subcategoria 2 da categoria benigno, que corresponde ás imagens dos casos de Doença Fibrocistica com 25007 imagens.

Ao calcular através das 25007, podemos então definir que para equilibrar a alocação total, seriam usadas 25007 imagens de cada subcategoria. Esta alocação iria garantir que cada categoria "carcinoma" e "benigno" teriam o mesmo número de imagens, que seriam 50014. Estas imagens seriam usadas na categoria "mal_ou_beg", que é o campo que define se cada imagem é de um caso benigno ou maligno, pelo que estaria assim garantido o equilíbrio nesta categoria. No final, a alocação teria um total de 100028 imagens. Em seguida, é boa prática dividir estas imagens numa proporção de 80-20% para que 80% destas imagens sejam usadas para treino do modelo e os restantes 20% sejam usadas para validação deste.

Porém, devido ao processamento limitado da máquina que corre o treino do modelo, foi impossível utilizar a totalidade do *dataset* com os valores indicados anteriormente. Por isso, depois de algumas tentativas para definir a quantidade de imagens a usar nos modelos de treino e validação, foi definido que uma quantidade de aproximadamente 18000 imagens era a quantidade suportada pela máquina de 32 Gigabytes de memória RAM (*Random Access Memory*).

Para ter uma melhor noção da quantidade das imagens, foram definidas variáveis com o cálculo da alocação das imagens no código de maneira a fazer a alocação correta das imagens de treino e de validação.

3.3.4. Amostragem Estratificada

A seguinte etapa debruçou-se no planeamento e estratificação da amostragem tanto para treino como validação. É esta parte do código que é responsável por garantir a quantidade e a não sobreposição entre os conjuntos.

Esta amostragem foi definida ao utilizar a função "sample" da biblioteca "pandas". Esta função seleciona aleatoriamente, com base na "seed" previamente definida, um número específico de imagens de treino ("n_train_per_type") e de validação ("n_val_per_type") para cada subcategoria. A seleção aleatória e a subsequente exclusão das imagens selecionadas para treino antes da amostragem para validação garantem que não haja sobreposição entre os conjuntos.

3.3.5. Formação dos conjuntos de dados

Após a realização da amostragem para todas as categorias, os conjuntos de dados de treino e validação são formados pela concatenação das amostras individuais. A função "sample" é novamente aplicada para baralhar os dados, seguida de "reset_index" para assegurar a correta indexação das imagens. Esta etapa é crucial para evitar qualquer ordem intrínseca que possa influenciar a aprendizagem do modelo.

3.3.6. Validação da amostragem

Numa fase final da amostragem, foi necessário recorrer à função "assert" que confirma se o número de imagens nos conjuntos de treino e validação corresponde ao esperado (14.400 e 3.600, respetivamente). Este passo é de extrema importância para verificar a integridade do processo de amostragem e garantir que o modelo de rede neuronal será treinado e validado com a quantidade adequada de dados que foram definidos anteriormente.

3.4. Processamento de dados

Depois de definida a quantidade de imagens e de se ter realizado a alocação destas, é necessário fazer o processamento dos dados escolhidos pelo código apresentado anteriormente. A etapa de preparação dos dados é vital para assegurar que o modelo da rede neuronal receba a informação de forma correta e formatada de acordo com as suas necessidades. Ao adotar práticas como normalização e monitorização do progresso, assegura-se não só a qualidade do treino e validação, mas também a eficiência e transparência do processo.

3.4.1. Carregamento e transformação de imagens

O procedimento de processamento dos dados inicia com o carregamento das imagens a partir dos caminhos fornecidos no "DataFrame" "df". Cada imagem é carregada com as dimensões específicas – "image_height" e "image_width" - adequadas para a arquitetura da rede que se pretende treinar. As dimensões destas imagens são 512x512 pixéis. Como as imagens já têm essa dimensão, a especificação da dimensão pelo código garante que são estas as dimensões utilizadas na preparação como no treino dos dados.

Durante o carregamento, as imagens são convertidas em vetores e normalizadas dividindo-se os valores dos pixéis por 255. Esta normalização é uma prática comum no processamento de imagens, pois converte a escala de cor de cada pixel para um valor entre 0 e 1, facilitando a convergência do modelo durante o treino.

Ao longo do carregamento e processamento das imagens, é realizado uma monitorização do processo de carregamento (Figura 14). É de grande importância monitorizar o processo de carregamento, especialmente quando se processa um grande número de imagens. Assim, implementa-se um sistema de relatório de progresso que informa o utilizador a cada 10% das imagens processadas. Esta funcionalidade não só ajuda a verificar que o processo está a progredir, mas também estima o tempo restante até á conclusão da fase em curso.

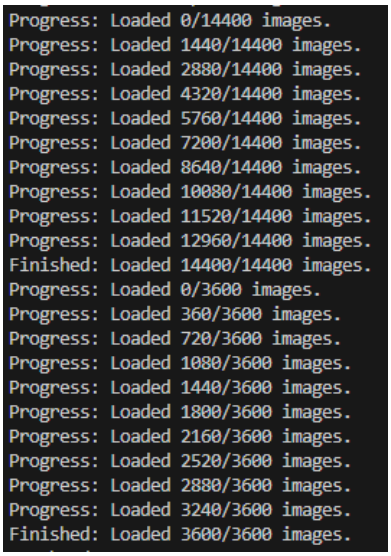


Figura 14 – Monitorização do progresso do carregamento das imagens

3.4.2. Extração de Etiquetas

A extração correta de etiquetas a partir do conjunto de dados é um elemento fulcral na configuração do processo de aprendizagem supervisionada em modelos de redes neuronais. Para cada imagem carregada, as etiquetas correspondentes às categorias "mal_ou_beg" (maligno ou benigno), "carcinoma" e "benigno" são extraídas e armazenadas em listas separadas. Estas etiquetas serão utilizadas como o verdadeiro valor (*ground truth*) durante o treino e a validação do modelo, permitindo que a rede aprenda a classificar as imagens nas categorias corretas.

Estas listas de etiquetas são essenciais pois definem os "alvos" que o modelo tentará prever corretamente durante o treino. A precisão da extração destas etiquetas é diretamente proporcional à qualidade da aprendizagem do modelo.

A representação vetorial resultante destas será utilizada para treinar o modelo, através de uma função de perda que mede a discrepância entre as previsões do modelo e as verdadeiras etiquetas. Além disso, na fase de validação, a comparação entre as etiquetas previstas e as verdadeiras etiquetas permitirá avaliar a capacidade do modelo em generalizar o que foi aprendido para os novos dados.

3.4.3. Conversão e Construção dos Conjuntos de Dados

Após o carregamento de todas as imagens e a extração das etiquetas, as listas são convertidas em vetores da biblioteca "NumPy". Esta conversão é crucial porque estruturas de dados do "NumPy" são as mais apropriadas para alimentar os modelos de redes neuronais construídos com bibliotecas *Python* como "TensorFlow" ou "Keras", devido à sua eficiência e compatibilidade.

Finalmente, a função "prepare_data" retorna os conjuntos de imagens ("X") e as respetivas etiquetas ("y") para as categorias "mal_ou_beg", "carcinoma" e "benigno". Estes conjuntos são separados em dados de treino ("X_train", "y_train_mal_ou_beg", "y_train_carcinoma", "y_train_benigno") e de validação ("X_val", "y_val_mal_ou_beg", "y_val_carcinoma", "y_val_benigno"). Depois deste passo, as imagens e os dados a eles associados estão prontos para serem utilizados no treino e validação do modelo de rede neuronal.

3.5. Seleção e Adaptação das Arquiteturas de Redes Neuronais

3.5.1. Avaliação Preliminar de Modelos de *Deep Learning*

A convolução é uma operação linear que envolve a multiplicação de uma matriz de pesos, chamada de filtro ou *kernel*, com uma matriz de entrada, para produzir uma matriz de saída denominada de mapa de características. Este processo permite que a rede identifique padrões como arestas, cantos e outras características visuais, que são essenciais para o reconhecimento de imagens e vídeos.

O processo de convolução das CNNs é particulamente vantajoso no âmbito da Anatomia Patológica, uma vez que mantém a relação espacial entre os píxeis, aspeto essencial durante a análise de células que, por vezes, apresentam diferenças estruturais muito subtis. Esta propriedade é determinante para a identificação de células em distintas etapas do ciclo celular ou para o reconhecimentode modificações morfológicas iniciais que podem surgerir um diagnóstico diferencial.

A diversidade de modelos CNN oferece assim um leque de opções para o treino do estudo em questão, incluindo AlexNet, DenseNet, EfficientNet, InceptionV3, MobileNetV3, ResNet e SqueezeNet. Na Tabela 3 é possível ver as características destas, as suas principais diferenças e a sua performance.

Tabela 3 – Comparação das Principais Arquiteturas de Redes Neuronais Convolucionais Aplicadas em Patologia Digital

Arquitetura	Características Principais	Principais Diferenças	Performance Benchmarks
SqueezeNet [49]	Utiliza módulos *fire* para reduzir parâmetros e operações; Estratégia de *squeeze* e *expand*;	Muito compacta, com menos parâmetros, no entanto mantém a precisão;	Treino tradicional com foco na economia de recursos computacionais;
InceptionV3 [54]	Módulos *Inception* com fatores de convolução para eficiência computacional;	Uso de convoluções assimétricas para reduzir a complexidade;	Treino com uso intensivo de dados e técnicas de regularização;
AlexNet [57]	Uma das primeiras arquiteturas profundas; Usa *rectified linear unit*, normalização local e *dropout*;	Introduziu técnicas inovadoras, mas é superada em precisão;	Uso de grandes volumes de dados e técnicas de regularização como *dropout*;
DenseNet [59]	Conectividade densa onde cada camada está conectada a todas as outras;	Reduz o problema do desvanecimento de gradiente; eficiente no uso de parâmetros;	Treino com eficiência de parâmetros, requer menos *epochs*.
EfficientNet [60]	Balanço escalável entre largura, profundidade e resolução das camadas;	Abordagem de escalamento sistemático composto para aumento da eficiência;	Treino escalável com aumento proporcional das dimensões da rede;
MobileNet V3 [61]	Otimizada para dispositivos móveis; Usa módulos *bottleneck* com atenção baseada em *squeeze;*	Com foco no desempenho em plataformas com recursos limitados;	Treino com técnicas de economia computacional, como cortes e quantização;
ResNet [52]	Uso de conexões residuais para permitir treinos de redes mais profundas;	Melhora a propagação de gradiente e permite treino de redes com centenas de camadas;	Treino com inicialização cuidadosa e lotes de normalização;

No contexto do presente estudo, estes modelos CNN foram testados, utilizando um *dataset* de dimensão reduzida para a realização de testes exploratórios, o que permitiu a obtenção de uma visão preliminar sobre o desempenho de cada um no contexto específico da classificação de imagens médicas.

Através da análise dos resultados (Tabela 4), os modelos SqueezeNet e InceptionV3 destacaram-se pelo seu desempenho promissor, evidenciando uma rápida taxa de aprendizagem aliada a um tempo de treino razoavelmente curto – aproximadamente 20 minutos por *epoch* para o SqueezeNet e aproximadamente 2 horas por *epoch* para o InceptionV3. Esta eficiência tornou-os nos candidatos ideais para a aplicação no âmbito deste estudo.

Tabela 4 – Tabela da análise de eficiência, precisão e exatidão dos vários modelos

Arquitetura	Teste	Tempo/*epoch*	Conclusões Finais
SqueezeNet	Confusion Matrix [[91 9] [3 97]] Classification Report precision recall f1-score support Class 0 0.97 0.91 0.94 100 Class 1 0.92 0.97 0.94 100 accuracy 0.94 200 macro avg 0.94 0.94 0.94 200 weighted avg 0.94 0.94 0.94 200	20 minutos por *epoch* num *dataset* de 18000 imagens	Modelo muito eficiente e com uma precisão e exatidão alta
InceptionV3	Confusion Matrix [[100 0] [1 99]] Classification Report precision recall f1-score support Class 0 0.99 1.00 1.00 100 Class 1 1.00 0.99 0.99 100 accuracy 0.99 200 macro avg 1.00 0.99 0.99 200 weighted avg 1.00 0.99 0.99 200	2 horas por *epoch* num *dataset* de 18000 imagens	Modelo muito preciso e exato e com eficiência aceitável
AlexNet	Confusion Matrix [[63 37] [39 61]] Classification Report precision recall f1-score support Class 0 0.62 0.63 0.62 100 Class 1 0.62 0.61 0.62 100 accuracy 0.62 200 macro avg 0.62 0.62 0.62 200 weighted avg 0.62 0.62 0.62 200	4 horas por *epoch* num *dataset* de 1000 imagens	Modelo de eficiência média e precisão e exatidão aceitáveis
DenseNet	Confusion Matrix [[89 11] [1 99]] Classification Report precision recall f1-score support Class 0 0.99 0.89 0.94 100 Class 1 0.90 0.99 0.94 100 accuracy 0.94 200 macro avg 0.94 0.94 0.94 200 weighted avg 0.94 0.94 0.94 200	7-8 horas por *epoch* num *dataset* de 1000 imagens	Modelo preciso e exato, mas pouco eficiente
EfficientNet	----	>10 horas por *epoch* num *dataset* de 1000 imagens	Modelo pouco eficiente
MobileNet V3	----	>10 horas por *epoch* num *dataset* de 1000 imagens	Modelo pouco eficiente
ResNet	----	>10 horas por *epoch* num *dataset* de 1000 imagens	Modelo pouco eficiente

3.5.2. Cenários de Classificação

Com a seleção dos modelos estabelecida, procedeu-se à configuração dos mesmos para responder a quatro cenários de classificação distintos. Vão ser estes quatro cenários que vão ser estudados para a análise deste estudo de classificação de imagens histológicas.

Os quatro cenários de classificação são:

- **Classificação binária do campo "mal_ou_beg":** cenário em que os dois modelos são treinados para identificar se uma imagem faz parte de um caso benigno (0) ou maligno (1)
- **Classificação binária do campo "carcinoma":** cenário onde é utilizado apenas imagens de casos malignos para classificar as imagens como sendo Carcinomas Ductais (0) ou Carcinomas Lobulares (1).
- **Classificação binária do campo "benigno":** cenário onde é utilizado apenas imagens de casos malignos para classificar as imagens como sendo Fibroadenoma (0) ou Doença Fibrocística (1).
- **Classificação multiclasse combinada:** cenário mais complexo onde o modelo é proposto a prever simultaneamente a categoria "mal_ou_beg (binário), a categoria ternária "carcinoma" como sendo Não maligno (0), Carcinomas Ductais (1) ou Carcinomas Lobulares (2), e a categoria ternária "benigno" como sendo Não benigno (0), Fibroadenoma (1) ou Doença Fibrocística (2).

3.5.3. Personalização e Implementação dos Modelos

Para implementar os modelos escolhidos anteriormente, foi utilizado os modelos pré-treinados onde foram adicionadas camadas de entrada e de saída personalizadas. A implementação começou com a importação do modelo base, utilizando pesos pré-treinados do 'imagenet' e excluindo as camadas superiores para permitir a personalização. A inclusão de uma "GlobalAveragePooling2D" visou a redução da dimensionalidade, seguida por uma camada densa com 1024 unidades e ativação "relu", promovendo a aprendizagem de características de alto nível.

As camadas de saída foram criteriosamente desenhadas para corresponder às necessidades específicas de cada tarefa de classificação:

- **"mal_ou_beg_out":** Camada densa com ativação 'sigmoid' para a previsão binária.
- **"carcinoma_out" e "benigno_out":** Camadas densas com ativação 'softmax' para as classificações ternárias, tratando-se de um problema de classificação multi-classe.

Estas camadas personalizadas são essenciais, pois permitem que o modelo realize previsões específicas para cada uma das categorias de diagnóstico, otimizando o desempenho nas tarefas atribuídas.

3.5.4. Compilação e Treino dos Modelos

Após a fase de seleção e adaptação das arquiteturas, é necessário proceder à compilação dos modelos. Esta fase constitui a preparação dos modelos para o treino, definindo-se o optimizador, as funções de perda e as métricas de avaliação.

Um optimizador é um algoritmo usado para alterar os atributos do modelo, como os pesos das ligações da rede neuronal, a fim de minimizar a função de perda. Para ambos os modelos, optou-se pelo optimizador Adam, que é um dos mais utilizados em *Deep Learning* devido à sua eficiência. O Adam é uma combinação de dois optimizadores - *AdaGrad* e *RMSProp* - destacando-se por adaptar a taxa de aprendizagem durante o treino para cada parâmetro do modelo. Esta característica torna-o

particularmente robusto para conjuntos de dados grandes e complexos, onde o espaço de parâmetros é vasto e as relações entre os dados são complexas e pouco claras.

A taxa de aprendizagem no optimizador foi estabelecida em 0.0001, uma escolha conservadora que visa uma convergência estável durante o treino, evitando oscilações abruptas que podem decorrer de taxas mais elevadas. Este valor foi escolhido pelo método de tentativa-erro, onde o valor escolhido foi o que teve resultados mais promissores e mais estáveis.

Para analisar as perdas e as métricas utilizaram-se as funções "binary_crossentropy" para as saídas binárias e "sparse_categorical_crossentropy" para as saídas ternárias. A exatidão (*accuracy*) foi a métrica escolhida para acompanhar a performance dos modelos, oferecendo uma leitura direta e compreensível do sucesso das previsões em relação às etiquetas verdadeiras.

Definido assim todos os parêmetros, foi precedido o treino do modelo através de um processo iterativo de 15 *epochs*, permitindo aos modelos ajustar progressivamente os seus pesos internos. Ao longo das *epochs*, os modelos aprenderam a minimizar as funções de perda e, por conseguinte, a aumentar a exatidão (*accuracy*) e a precisão (*precision*) das previsões.

3.6. Validação

A fase de validação é importante no desenvolvimento de modelos de *machine learning*, pois permite avaliar a performance do modelo em dados que não foram utilizados durante o treino, simulando uma aplicação prática e fornecendo uma estimativa da capacidade de generalização do modelo. No código é utilizado a função "validate_model", que executa essa tarefa crucial ao aplicar o modelo treinado ao conjunto de validação e separando as probabilidades previstas para cada classe.

Esta fase irá permitir tirar valores para de seguida ser possível, através de várias métricas estatísticas e gráficas, analisar a precisão e exatidão de cada um dos modelos.

4. RESULTADOS E DISCUSSÃO

Neste capítulo irá ser explicado cada uma das métricas utilizadas para a avaliação do desempenho dos modelos e todos os resultados obtidos dos treinos dos modelos para os quatro cenários descritos no capítulo passado. A apresentação dos resultados será dividida pelos cenários, ao apresentar as métricas obtidas para os dois modelos utilizados – SqueezeNet e InceptionV3. Em cada cenário é feita uma comparação entre os resultados obtidos nos dois modelos, analisando e discutindo quais são os prós e os contras no uso de cada modelo no cenário em específico.

As métricas utilizadas para apresentar os resultados foram descritos no capítulo anterior, que são: matriz confusão, exatidão (*accuracy*), precisão (*precision*), sensibilidade (*recall*), *f1-score*, curvas PR, curvas ROC e AUC, e curva de exatidão.

No final é discutido os resultados na globalidade com conclusões finais sobre qual modelo mais adequado para o caso em questão.

4.1. Avaliação de Desempenho dos Modelos de Classificação

O passo final no treino de um modelo de *machine learning* é a recolha de dados e avaliação de desempenho desses modelos perante os resultados da fase de validação, pois é desta forma que é possível compreender a eficácia e a aplicabilidade dos mesmos em cenários reais. Nos próximos subcapítulos irá ser explicado quais são as métricas utlizadas, nomeadamente a matriz de confusão, exatidão (*accuracy*), precisão (*precision*), *f1-score*, *recall* (sensibilidade), perdas, além das curvas de PR e de ROC, incluindo sua versão multinível.

4.1.1. Matriz confusão

A matriz de confusão é uma tabela que permite a visualização do desempenho do modelo de classificação. Esta matriz compara as etiquetas verdadeiras com as previsões do modelo, fornecendo uma visão clara dos valores falsos (falsos positivos e falsos negativos) e dos valores verdadeiros (verdadeiros positivos e verdadeiros negativos). As métricas fornecem uma representação visual detalhada do desempenho do modelo, revelando não apenas o número de previsões corretas, mas também os tipos específicos de erros que o modelo cometeu.

A sua forma mais básica reflete os resultados de um modelo de classificação binária, e com isso contém 4 elementos principais, como demonstra a Figura 15:

1) **Verdadeiros Positivos (VP):** Casos em que o modelo previu corretamente a classe positiva (1).
2) **Falsos Positivos (FP):** Casos em que o modelo previu incorretamente a classe positiva (1).
3) **Verdadeiros Negativos (VN):** Casos em que o modelo previu corretamente a classe negativa (0).
4) **Falsos Negativos (FN):** Casos em que o modelo previu incorretamente a classe negativa (0).

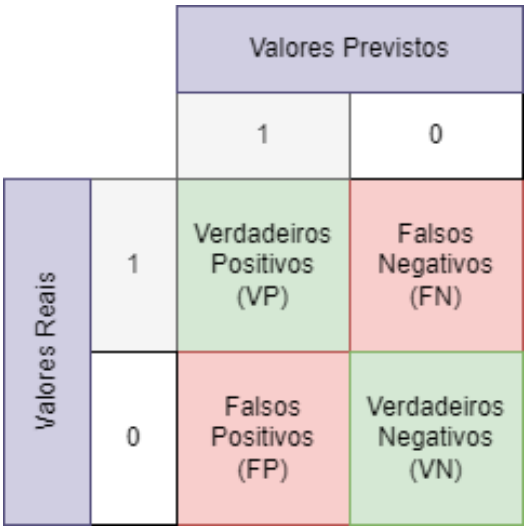


Figura 15 – Organização de uma matriz confusão binária

Porém, a matriz confusão não é apenas usada para problemas binários, pelo que pode ser adaptada para problemas multiclasse (Figura 16), seguindo sempre o mesmo critério dos positivos, onde o modelo acerta, e dos negativos, onde o modelo classifica incorretamente.

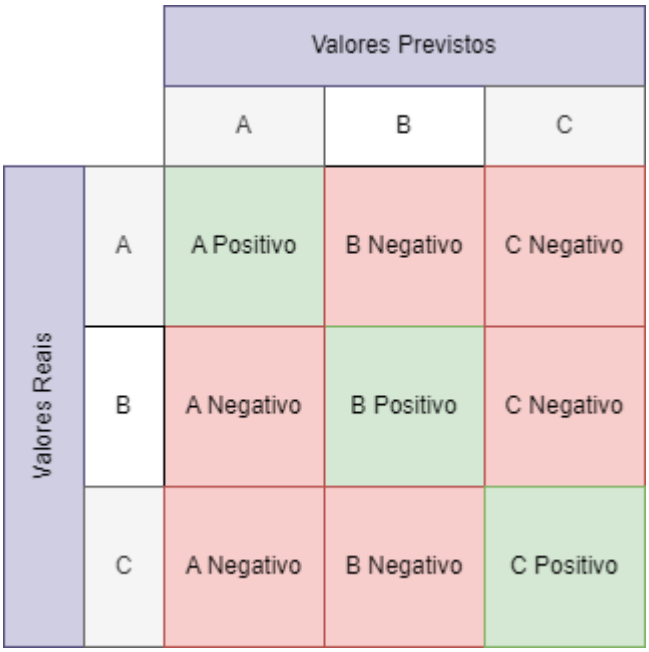


Figura 16 – Organização de uma matriz confusão multiclasse

4.1.2. Exatidão (*accuracy*)

A exatidão (ou *accuracy*) é uma das principais métricas para avaliar o desempenho de modelos de classificação. Matematicamente, a exatidão é definida como a proporção de previsões corretas (tanto positivas quanto negativas) em relação ao total de previsões feitas pelo modelo. É expressa pela fórmula:

$$accuracy = \frac{Verdadeiros\ Positivos\ (VP) + Verdadeiros\ Negativos\ (VN)}{Total\ das\ entrada}$$

Um valor alto de exatidão pode indicar que o modelo tem um bom desempenho geral em prever as etiquetas corretamente para as amostras. Porém, se o valor for baixo, pode sugerir que o modelo tem dificuldades em classificar as amostras corretamente.

Embora a exatidão seja uma métrica importante, ela sozinha pode induzir em erro, especialmente em casos onde o conjunto de dados seja desequilibrado. Nestas situações, um modelo pode alcançar uma alta exatidão simplesmente prevendo sempre a classe mais comum, ignorando as classes menos frequentes.

4.1.3. Precisão (*Precision*)

A precisão é uma métrica importante na classificação e consiste na proporção de identificações corretas realizadas pelo modelo, ou seja, mede a qualidade das previsões corretas do modelo. A precisão é particularmente importante em contextos onde a previsão errada é muito significativa para a natureza do estudo.

O cálculo da precisão tem a seguinte equação, onde x é a etiqueta que o modelo tenta prever:

$$precision = \frac{Verdadeiro\ x}{Verdadeiro\ x + Falso\ x}$$

Se o valor da precisão for alto, quer dizer que a maioria das previsões de ser x é correta, caso contrário pode sugerir que maior parte das previsões são incorretas.

Porém, a precisão não tem em conta os falsos y, ou seja, situações em que o modelo previsão que um dado era y, mas na realidade era x. Portanto, um modelo pode ter uma alta precisão, mas ainda assim ser ineficaz se muitos casos x reais não forem detetados.

4.1.4. Sensibilidade (*Recall*)

A sensibilidade (*recall*) é uma métrica usada para analisar todos os casos reais de uma classe específica. Esta métrica mede a capacidade de o modelo identificar corretamente uma classe em relação ao número total de casos dessa classe. O *recall* é calculado da seguinte forma:

$$recall = \frac{Verdadeiros\ x}{Verdadeiros\ x + Falsos\ y}$$

Um valor elevado de *recall* indica que um modelo é eficaz a identificar a maioria dos casos de uma classe específica. Pelo sentido contrário, um baixo valor de *recall* indica que um modelo não está a identificar os casos de uma classe específica. Embora um alto *recall* seja importante, um modelo com *recall* de 100% pode ser ineficaz se também tiver muitos falsos positivos (baixa precisão). Por exemplo, um teste que identifica todos como "doentes" terá um *recall* perfeito, mas seria inútil na prática.

4.1.5. *F1-score*

O *f1-score* é uma métrica utilizada para combinar a *precison* e a *recall* numa única medida. É usado principalmente quando se tenta encontrar um equilíbrio máximo entre as duas métricas. O *f1-score* é uma média harmónica da *precision* e do *recall*, calculado pela seguinte fórmula:

$$f1score = 2 * \frac{precision * recall}{precision + recall}$$

Um valor alto do *f1-score* indica que o modelo realiza previsões com precisão alta e com elevada sensibilidade. Porém, se o *f1-score* for baixo, pode ser indício de que o modelo tem problemas de precisão, de *recall* ou ambos. Apesar do *f1-score* ser uma métrica muito importante, pode não ser a ideal quando o objetivo do modelo é priorizar uma métrica em relação à outra.

4.1.6. Curvas de *Precision-Recall*

As curvas PR apresentam-se em forma de gráficos de avaliação usados para entender o desempenho de um modelo de classificação em diferentes patamares de decisão. A curva tem como eixo horizontal os valores de *recall*, enquanto o eixo vertical contém os valores de precisão. Num modelo ideal (Figura 17 [68]), a precisão e o *recall* terão taxas altas simultaneamente, o que originaria um gráfico onde a curva se aproximaria do canto superior direito. A área abaixo da curva também é uma boa indicadora, sendo que quanto maior a área, maior é a precisão e o *recall* do modelo.

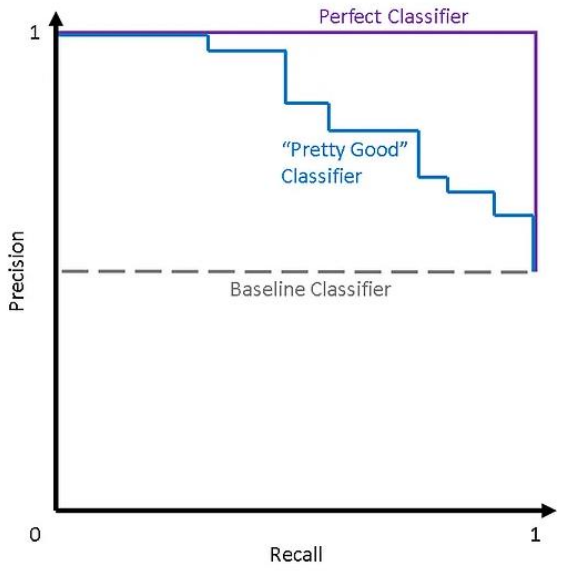


Figura 17 – Dois exemplos de potenciais curvas PR. A roxo podemos ver um exemplo de um modelo com curva perfeita, e a azul uma curva razoável [68]

4.1.7. Curva de ROC e AUC

A curva de ROC é um gráfico utilizado para avaliar a performance de modelos de classificação binária. A curva representa a relação entre a taxa de previsões corretas de um caso x e a taxa previsões incorretas desse mesmo caso para diferentes patamares de decisão. A *Area Under the Curve* (AUC) é a área sob a curva ROC e serve como um resumo numérico da capacidade do modelo de distinguir entre as classes.

O gráfico da curva ROC como exemplificado na Figura 18 [69], consiste no eixo horizontal que representa a proporção de um caso y que foi identificado como x, e no eixo vertical que representa os casos x que foram corretamente identificados. Num modelo ideal, o valor do eixo vertical deverá ser máximo (1) e o valor no eixo horizontal ser mínimo (0), em que a AUC será 1. Caso a curva coincidir com a função identidade (do tipo $f(x) = x$, então o modelo representa um desempenho aleatório e a AUC será de 0,5. Por isso, um AUC elevado sugere um modelo que consegue distinguir bem as classes.

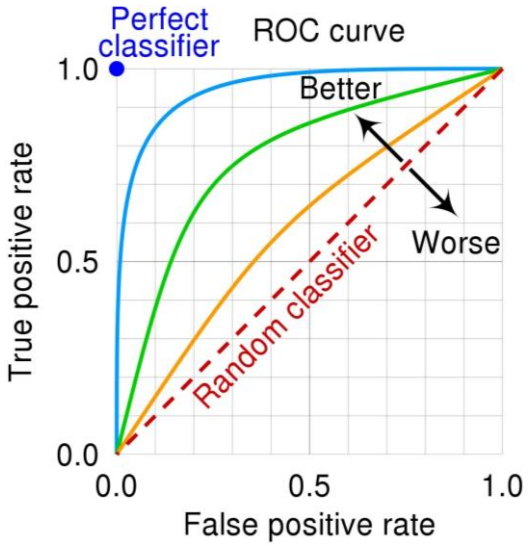


Figura 18 – Curva ROC com várias situações possíveis: o ponto azul indica o modelo ideal enquanto a linha tracejada a vermelho indica que o modelo é aleatório [69]

Em contextos de classificação multiclasse, a curva ROC pode ser adaptada para avaliar o desempenho do modelo de cada classe em relação todas as outras, resultando em múltiplas curvas ROC e AUC's. Esta abordagem fornece uma visão mais detalhada do desempenho do modelo em cada classe individual.

4.1.8. Curva de exatidão (*accuracy*)

A curva de exatidão é um gráfico que mostra a variação da exatidão (*accuracy*) de um modelo de classificação ao longo do tempo, geralmente ao longo das *epochs* de treinamento e validação. Esta curva é uma ferramenta visual usada para avaliar como a exatidão do modelo evolui à medida que ele aprende com os dados durante o processo de treino.

Geralmente, este gráfico tem como eixo horizontal o tempo em *epochs*, enquanto o eixo vertical indica a exatidão do modelo, calculada como a proporção das previsões corretas em relação ao número total de previsões. Se a curva é crescente, então indica que o modelo está a aprender ao longo das *epoch* dos dados que lhe são fornecidos. Porém, se a curva estagna ou decresce, pode ser um sinal de problemas como *overfitting*, onde o modelo aprende demasiados detalhes, ou *underfitting*, onde o modelo não consegue aprender suficiente dos dados de treino.

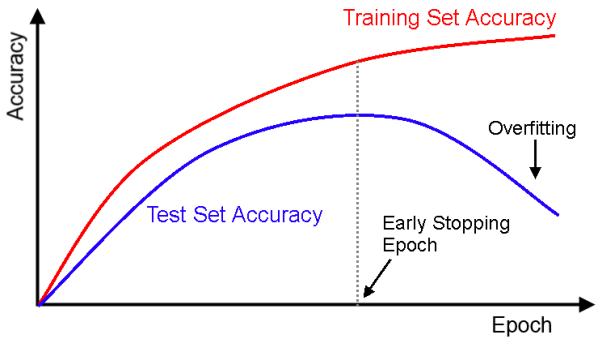


Figura 19 – Exemplo do progresso das curvas de exatidão. A azul é a curva de teste acompanhada com a curva de treino, mas que devido a *overfitting*, a exatidão da validação baixa[70]

Tal como é possível ver na Figura *19* [70], as curvas de exatidão são colocadas em comparação entre os dados de treino e os de validação. A comparação entre ambas as curvas dão informações importante sobre o progresso do modelo:

- **Exatidão Semelhante entre Treino e Validação:** Indica que o modelo é estável e está a evoluir.
- **Alta Exatidão no Treino, mas Baixa na Validação:** Sugere que o modelo se encontra em *overfitting*, onde o modelo não está a aprender mais com os dados, interferindo com a capacidade de classificação do modelo.
- **Baixa Exatidão em Ambos Treino e Validação:** Pode indicar *underfitting*, onde o modelo é muito simples para capturar a complexidade dos dados.

A análise da curva de exatidão auxilia no ajuste de parâmetros de treino, como a taxa de aprendizagem, o número de *epochs* ou a arquitetura do modelo, para melhorar a capacidade de aprendizagem e classificação do modelo.

4.2. Apresentação de resultados

4.2.1. Cenário 1: Classificação binária do campo "mal_ou_beg"

O treino para a classificação binária do campo "mal_o_beg" iniciou-se com o treino do modelo em SqueezeNet. A execução do modelo teve uma duração média de 20 a 25 minutos por *epoch* o que perfaz uma duração total de 5 horas a 6 horas e 15 minutos ao longo das 15 *epochs* que foram executadas.

Do treino deste modelo foi obtida a matriz confusão da Figura 20. Desta matriz é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence à classe benigna, representado por zero (0), ou então à classe maligna, representado por um (1).

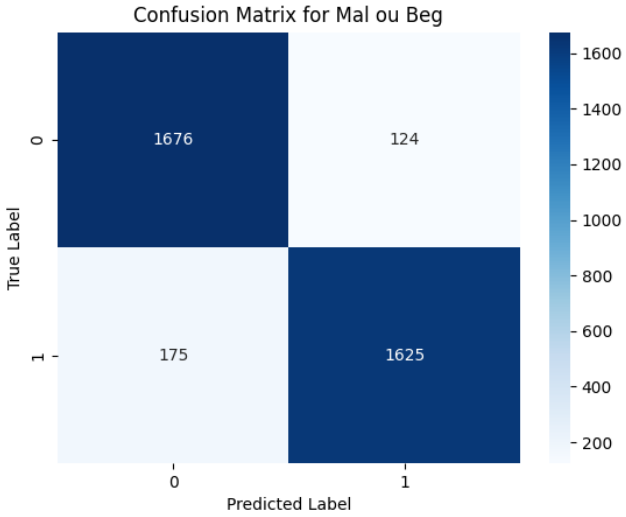


Figura 20 – Matriz confusão do modelo executado em SqueezeNet para a classificação binária do campo "mal_ou_beg". O Valor zero (0) corresponde à classe benigna e um (1) à classe maligna

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1625 casos foram corretamente identificados como malignos.
- **Falsos Negativos (FN):** 175 casos foram erradamente classificados como benignos, quando na verdade eram malignos.
- **Verdadeiros Negativos (VN):** 1676 casos foram corretamente identificados como benignos.

- **Falsos Positivos (FP):** 124 casos foram erradamente classificados como malignos, quando na verdade eram benignos.

Podemos reparar que, apesar de na grande maioria das vezes o modelo prever corretamente as classes de uma imagem, este confunde em maior quantidade casos que são malignos com os casos benignos, do que propriamente casos benignos com malignos que são erradamente identificados como tal.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 21. Aqui é representado as métricas simples de desempenho do modelo como as precisões (*precision*), exatidões (*accuracy*), sensibilidade (*recall*) e *f1-score*.

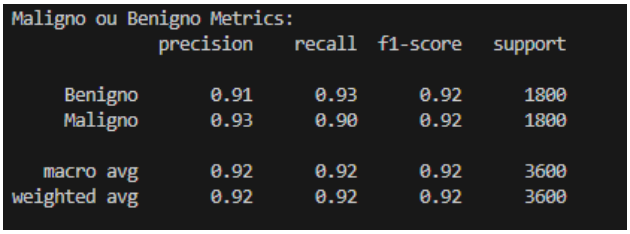


Figura 21 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo "mal_ou_beg"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe Benigna (0):**
  - A **precisão**, que indica a proporção de identificações corretas de benignidade entre todas as identificações de benignidade feitas pelo modelo, foi de 91%;
  - O ***recall***, que mede a proporção de casos benignos corretamente identificados pelo modelo em relação a todos os casos reais de benignidade, foi de 93%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 92%.
- **Para a classe Maligna (1):**
  - A **precisão** foi de 93%;
  - O ***recall*** foi de 90%;
  - O ***f1-score*** foi de 92%.
- **Exatidão** geral do modelo foi de 92%.

Com estas métricas, podemos reparar que o modelo apresenta valores de precisão e exatidão relativamente altos e com um bom equilíbrio entre a precisão e a sensibilidade. Porém, é de notar que o modelo é ligeiramente menos eficiente na identificação de todos os casos reais de malignidade (90% de *recall*), do que na identificação de casos reais de benignidade (93% de *recall*).

A métrica obtida de seguida foi o gráfico da curva PR. Como já explicado anteriormente, ela demonstra a relação entre a precisão e a sensibilidade para a classificação binária em questão.

A curva PR ilustrada na Figura 22 demonstra que à medida que o *recall* aumenta, a precisão permanece alta de maneira consistente, onde apenas diminuí para valores mais elevados do *recall*. A AUC da curva PR é alta (0,98), o que indica um bom desempenho do modelo. O valor próximo de 1 revela uma alta capacidade do modelo distinguir entre classes com precisão mesmo em diferentes patamares da classificação.

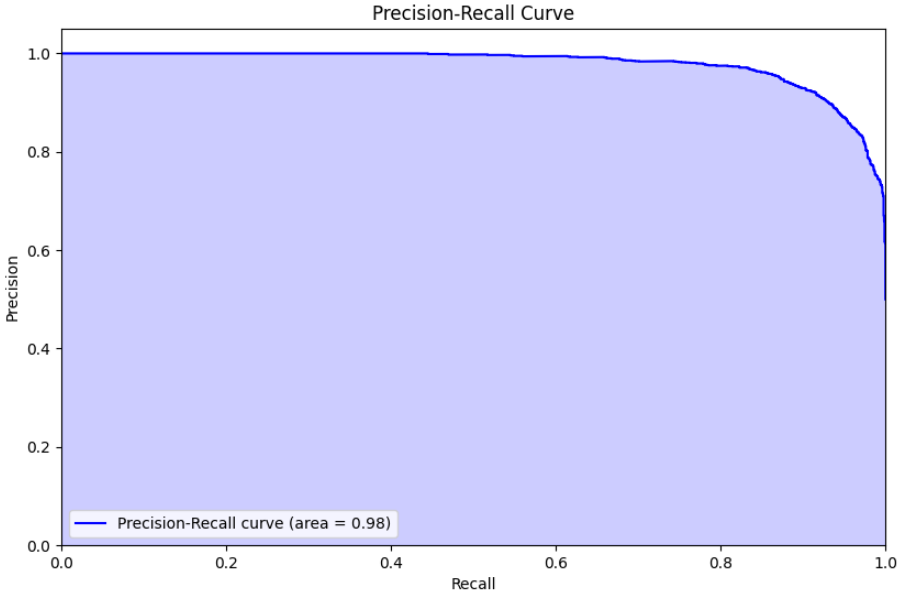


Figura 22 – Gráfico da Curva PR e respetiva AUC do modelo executado em SqueezeNet para a classificação binária do campo "mal_ou_beg"

A curva ROC é igualmente importante na avaliação da performance do modelo de classificação binária. Como descrito no capítulo anterior, o gráfico traça a taxa de verdadeiros positivos (sensibilidade, contra a taxa de falsos positivos.

Ao analisar o gráfico da Figura 23, podemos concluir que esta curva revela uma elevada capacidade de separação das duas classes. A AUC desta curva é de 0,98, o que demonstra um desempenho próximo do ideal. Este valor sugere que o modelo tem uma taxa elevada de verdadeiros positivos e uma baixa taxa de falsos positivos, o que mostra ser capaz de identificar corretamente a maioria dos casos malignos com um número relativamente baixo de casos falsos de malignidade.

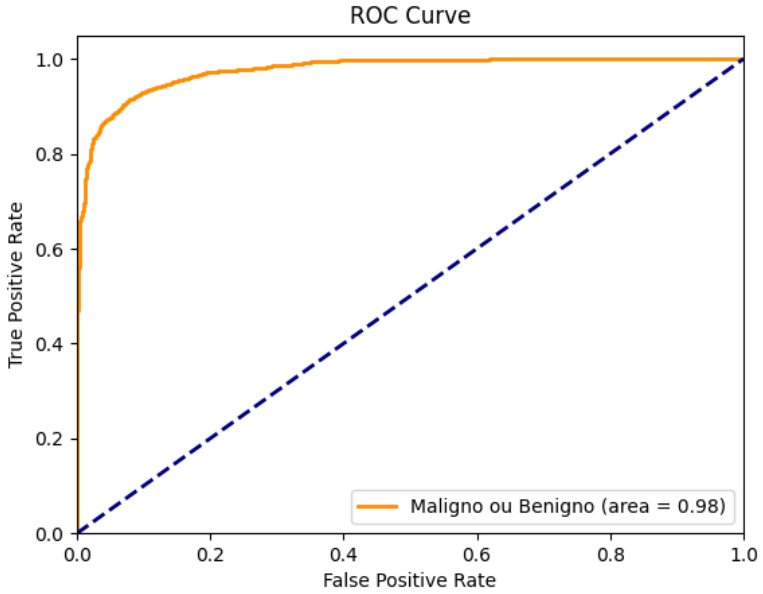


Figura 23 – Curva ROC e respetivo valor AUC do modelo SqueezeNet na classificação binária do campo "mal_ou_beg"

Por último, foi obtido o gráfico de evolução da exatidão do modelo ao longo dos 15 *epochs*. Este gráfico consiste em duas curvas de exatidão, em que uma pertence aos valores de exatidão na fase de treino do modelo, e a segunda curva representa os valores de exatidão na validação.

Ao analisar o gráfico da Figura 24, podemos observar que após a primeira *epoch*, os valores de exatidão na fase de treino são próximos de 75%, mas que aumenta gradualmente ao longo das *epochs* até décima quinta *epoch*, onde os valores chegam acima de 92,5%. Já a exatidão da validação começa num patamar ligeiramente abaixo de 80% e mantém-se acima do valor de exatidão de treino durante as primeiras *epoch*. De seguida, o crescimento do valor de exatidão da validação diminui para valores entre os 88% e os 92%, onde termina na última *epoch*. Apesar da tendência crescente da exatidão de validação, é de notar uma ligeira instabilidade dos valores, que pode ser causada por vários fatores tais como alguma inconsistência dos dados do *dataset* ou a necessidade de alterações no modelo para aumentar a estabilidade da aprendizagem.

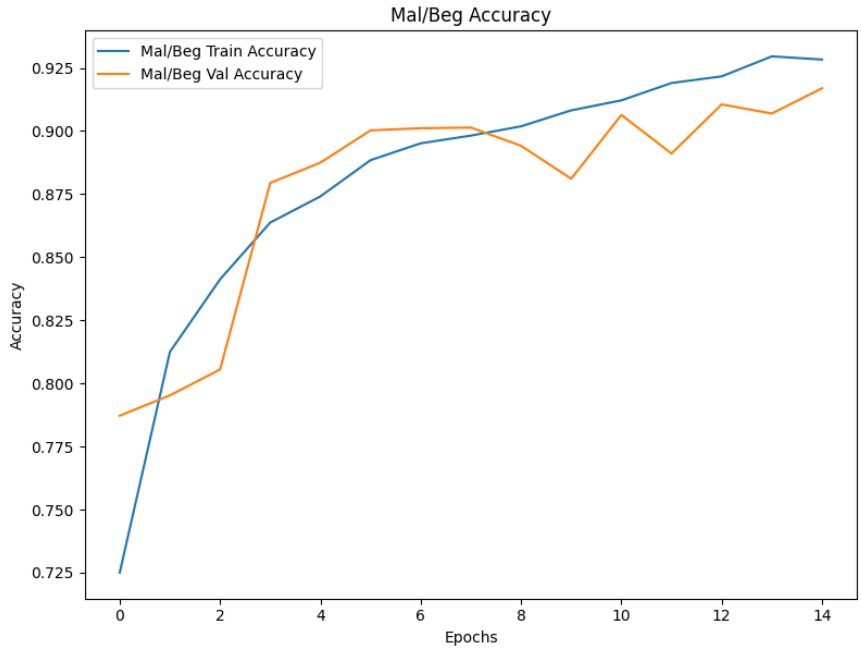


Figura 24 – Curva de exatidão do modelo SqueezeNet na classificação binária do campo "mal_ou_beg"

De seguida, foi executado o modelo InceptionV3, cujo tempo de execução varia entre 2 horas e 2 horas e 30 minuto por *epoch*, o que perfaz um total de 30 a 37 horas e meia ao longo das 15 *epochs* que foram executadas.

Do treino deste modelo foi obtida a matriz confusão da Figura 25. Desta matriz é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence à classe benigna, representado por zero (0), ou então à classe maligna, representado por um (1).

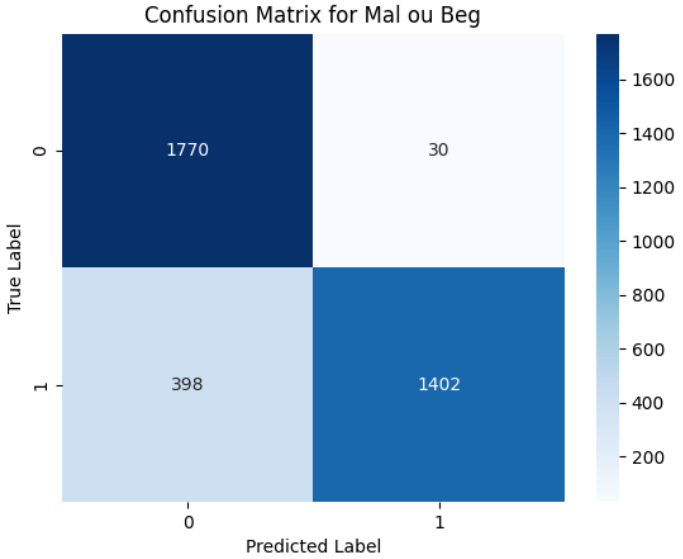


Figura 25 – Matriz confusão do modelo executado em InceptionV3 para a classificação binária do campo "mal_ou_beg". O Valor zero (0) corresponde à classe benigna e um (1) à classe maligna

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1402 casos foram corretamente identificados como malignos.
- **Falsos Negativos (FN):** 398 casos foram erradamente classificados como benignos, quando na verdade eram malignos.
- **Verdadeiros Negativos (VN):** 1770 casos foram corretamente identificados como benignos.
- **Falsos Positivos (FP):** 30 casos foram erradamente classificados como malignos, quando na verdade eram benignos.

Podemos reparar que, apesar de na grande maioria das vezes o modelo prever corretamente as classes de uma imagem, este confunde numa quantidade considerável de casos que são malignos com os casos benignos, do que propriamente casos benignos com malignos que são erradamente identificados como tal. Isto pode indiciar um problema ao longo da execução do modelo.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra Figura 26.

Maligno ou Benigno Metrics:

	precision	recall	f1-score	support
Benigno	0.82	0.98	0.89	1800
Maligno	0.98	0.78	0.87	1800
accuracy			0.88	3600
macro avg	0.90	0.88	0.88	3600
weighted avg	0.90	0.88	0.88	3600

Figura 26 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo "mal_ou_beg"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe Benigna (0):**
  - A **precisão**, que indica a proporção de identificações corretas de benignidade entre todas as identificações de benignidade feitas pelo modelo, foi de 82%;

  - O ***recall***, que mede a proporção de casos benignos corretamente identificados pelo modelo em relação a todos os casos reais de benignidade, foi de 98%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 89%.
- **Para a classe Maligna (1):**
  - A **precisão** foi de 98%;
  - O ***recall*** foi de 78%;
  - O ***f1-score*** foi de 87%.
- **Exatidão** geral do modelo foi de 88%.

Com estas métricas, podemos notar que apesar de o modelo ter um desempenho alto, também revelam que há uma diferença na sua capacidade de identificar corretamente casos benignos em comparação com casos malignos. Para a classe benigna, a precisão de 82% sugere que, quando o modelo prevê um caso benigno, é correto em 82% dos casos. O *recall* de 0.98 para esta classe indica que o modelo foi capaz de identificar 98% de todas as lesões benignas verdadeiras. Isto resulta num *f1-score* de 0.89, que sugere um desempenho equilibrado.

No que diz respeito à classe maligna, a precisão elevada de 98% indica que o modelo está certo grande parte das vezes que realiza a previsão de malignidade. No entanto, o *recall* de 78% revela que o modelo não identifica parte dos casos malignos, deixando um número significativo deles por diagnosticar, o que se traduz num *f1-score* de 87%. Esta diferença pode dever-se a diferentes casos, como por exemplo *overfitting*, ou até mesmo a presença de características histológicas similares entre casos benignos e malignos

A curva PR ilustrada na Figura 27 apresenta que à medida que o *recall* aumenta, a precisão permanece alta de maneira consistente, onde apenas diminuí para valores mais elevados do *recall*. A AUC da curva PR é alta (0,98), o que indica um bom desempenho do modelo. O valor próximo de 1 revela uma alta capacidade do modelo distinguir entre classes com precisão mesmo em diferentes patamares da classificação.

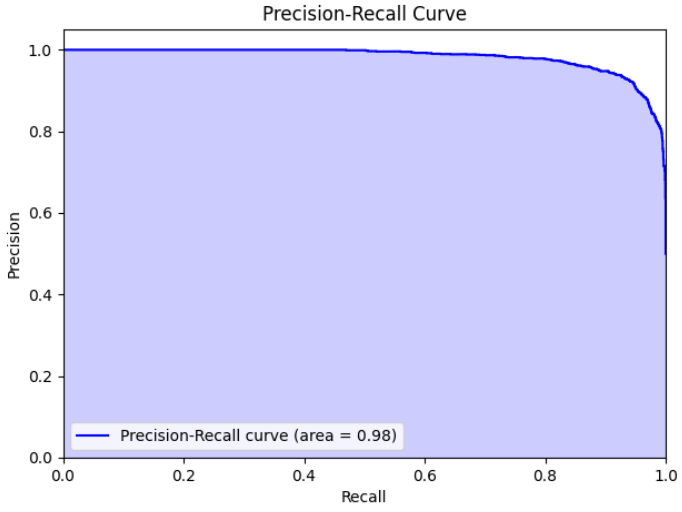


Figura 27 – Gráfico da Curva PR e respetiva AUC do modelo executado em InceptionV3 para a classificação binária do campo "mal_ou_beg"

Ao analisar o gráfico ROC da Figura 28, podemos concluir que esta curva revela uma elevada capacidade de separação das duas classes. A AUC desta curva é de 0,98, o que demonstra um desempenho próximo do ideal. Este valor sugere que o modelo tem uma taxa elevada de verdadeiros positivos e uma baixa taxa de falsos positivos, o que mostra ser capaz de identificar corretamente a maioria dos casos malignos com um número relativamente baixo de casos falsos de malignidade.

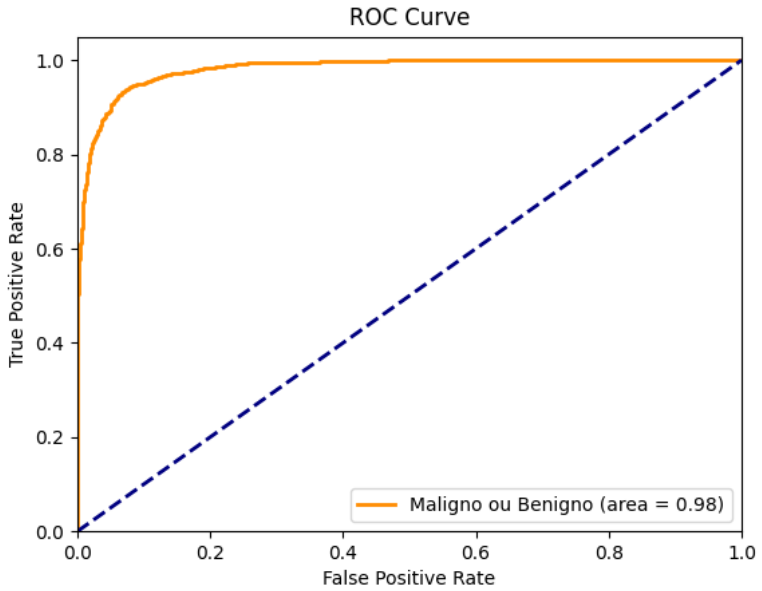


Figura 28 – Curva ROC e respetivo valor AUC do modelo InceptionV3 na classificação binária do campo "mal_ou_beg"

Ao analisar o gráfico da Figura 29, podemos observar que logo nas primeiras *epochs*, a exatidão de treino do modelo atinge valores superior a cerca de 95%, o que demonstra que este modelo tem uma grande capacidade de aprendizagem para o cenário estudado, mantendo-se estável ao longo das restantes *epochs*. A curva de exatidão da fase de validação acompanha a curva de treino, porém exibe uma maior variação ao longo das *epoch*, com alguma tendência decrescente em certas fases.

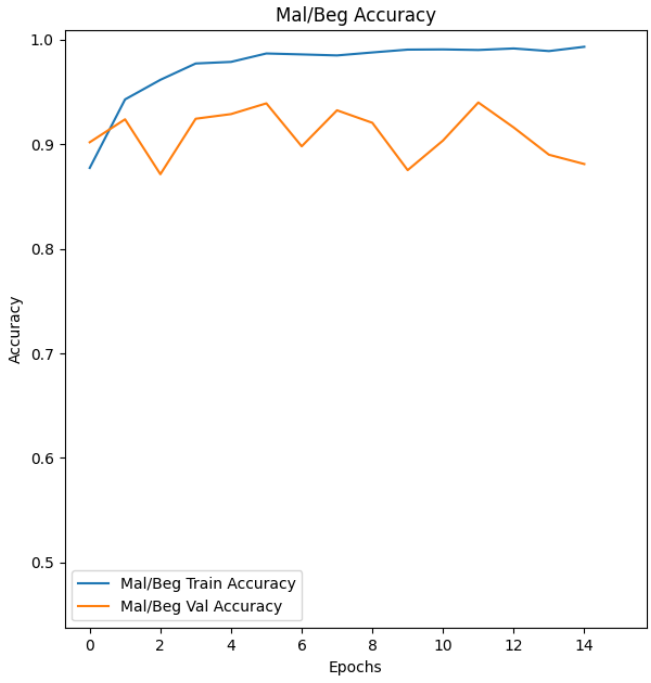


Figura 29 – Curva de exatidão do modelo InceptionV3 na classificação binária do campo "mal_ou_beg"

A discrepância entre as exatidões de treino e validação, com maior destaque nas oscilações nas curvas de validação, pode ser indicativa que o modelo possa estar a entrar em *overfitting*, em que o modelo começa a aprender padrões específicos que não ajudam a classificação final dos casos. A situação em causa poderá ter sido a causa do facto do modelo ter previsto erradamente casos malignos mais frequentemente quando estes eram benignos.

- ✓ COMPARAÇÃO/DISCUSSÃO

Numa análise comparativa entre o resultado dos dois modelos de rede neuronal, ambas mostraram uma alta eficiência. O modelo SqueezeNet mostrou uma performance mais estável nas métricas de precisão e *recall*, mantendo um bom equilíbrio na classificação dos dados.

Por outro lado, a InceptionV3 alcançou uma precisão mais alta na identificação de casos malignos, porém com um *recall* inferior, o que sugere uma sensibilidade menor na previsão da generalidade de todos os casos verdadeiramente malignos. Esta característica é muito importante na prática médica, pois falhar na identificação de um caso maligno tem consequências significativas para os pacientes em causa. Para além disso, o facto deste modelo apresentar uma maior volatilidade pode significar um *overfitting*, onde o modelo pode estar a captar informações erradas aos dados fornecidos.

Para este cenário, o modelo SqueezeNet seria a melhor escolha para o modelo não só pela eficiência computacional no treino, mas também pela estabilidade no seu desempenha e equilíbrio na classificação entre casos malignos e benignos.

Para além das questões técnicas, é ainda possível perceber que o modelo InceptionV3 apresenta uma grande quantidade de resultados falsos negativos, o que representa um problema grave no contexto do clínico. A existência de falsos negativos resulta num não diagnóstico de uma patologia de foro maligno e consequentemente em atrasos na administração de tratamentos essenciais para o paciente, diminuindo assim a probabilidade de sobrevivência. Preferencialmente, um modelo deverá errar num maior número de falsos positivos quando comparado com os falsos negativos, pois apesar de o primeiro cenário poder induzir as pacientes a situações de ansiedade, também exige a realização de exames adicionais e uma posterior não confirmação do diagnóstico inicial o que não representa um dano tão grande quanto omitir um tratamento necessário face a um diagnóstico incorreto de benignidade.

4.2.2. Cenário 2: Classificação binária do campo "benigno"

À semelhança do cenário 1, o treino para a classificação binária do campo "benigno" Iniciou-se com o treino do modelo em SqueezeNet. A execução do modelo teve uma duração média de 20 a 25 minutos por *epoch* o que perfaz uma duração total de 5 horas a 6 horas e 15 minutos ao longo das 15 *epochs* que foram executadas.

Do treino deste modelo foi obtida a matriz confusão da Figura 30. Desta matriz é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence à patologia Fibroadenoma, representado por zero (0), ou então a Doença Fibrocística, representado por um (1).

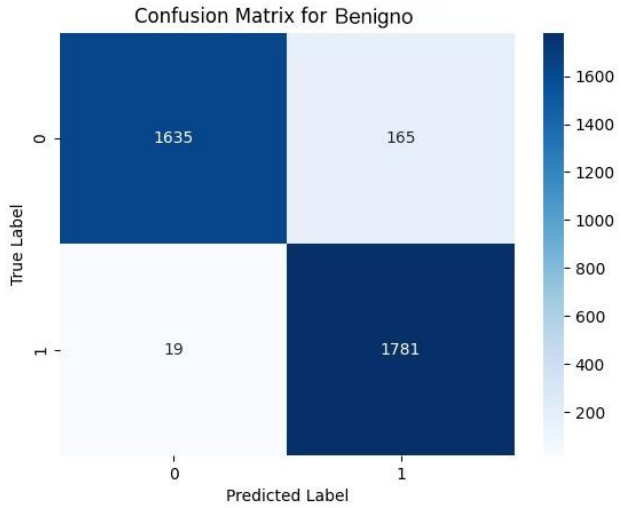


Figura 30 – Matriz confusão do modelo executado em SqueezeNet para a classificação binária do campo "benigno". O Valor zero (0) corresponde à patologia Fibroadenoma e um (1) a Doença Fibrocística

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1781 casos foram corretamente identificados como Doença Fibrocística.
- **Falsos Negativos (FN):** 19 casos foram erradamente classificados como Fibroadenoma, quando na verdade eram Doença Fibrocística.
- **Verdadeiros Negativos (VN):** 1635 casos foram corretamente identificados como Fibroadenoma.
- **Falsos Positivos (FP):** 165 casos foram erradamente classificados como Doença Fibrocística, quando na verdade eram Fibroadenoma.

Podemos reparar que a grande maioria das vezes o modelo prevê corretamente ambas as classes. Também podemos notar que muito dificilmente o modelo prevê erradamente como Fibroadenoma quando o caso era Doença Fibrocística, o que é muito positivo na validação deste modelo.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 31.

Benigno Metrics:

	precision	recall	f1-score	support
Fibroadenoma	0.99	0.91	0.95	1800
Doença Fibrocística	0.92	0.99	0.95	1800
accuracy			0.95	3600
macro avg	0.95	0.95	0.95	3600
weighted avg	0.95	0.95	0.95	3600

Figura 31 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo "benigno"

A partir desta imagem podemos obter as seguintes métricas:

- **Para Fibroadenoma (0):**
  - A **precisão**, que indica a proporção de identificações corretas de Fibroadenoma entre todas as identificações de Fibroadenoma feitas pelo modelo, foi de 99%;
  - O ***recall***, que mede a proporção de casos de Fibroadenoma corretamente identificados pelo modelo em relação a todos os casos reais de Fibroadenoma, foi de 91%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 95%.
- **Para Doença Fibrocística (1):**

- A **precisão** foi de 92%;
- O ***recall*** foi de 99%;
- O ***f1-score*** foi de 95%.

– **Exatidão** geral do modelo foi de 95%.

Para a patologia Fibroadenoma, o modelo alcançou uma precisão de 99%, indicando uma alta taxa de diagnósticos corretos para esta patologia em relação ao total de casos de Fibroadenoma identificados. O *recall* foi de 91%, refletindo a proporção de casos de Fibroadenoma corretamente identificados em relação ao total de casos reais. O *F1-score* foi de 95%, sugerindo um equilíbrio adequado entre as duas métricas.

Em relação à Doença Fibrocística, a precisão foi ligeiramente inferior, com um valor de 92%, enquanto o *recall* foi de 99%, indicando que quase todos os casos reais foram corretamente identificados pelo modelo. O F1-score manteve-se constante em 95%, demonstrando um desempenho robusto do modelo para ambas as patologias.

A exatidão global do modelo foi de 95%, o que implica que 95% das classificações foram corretas, o que reforça a consistência do desempenho do modelo.

A curva PR ilustrada na Figura 32 demonstra que à medida que o *recall* aumenta, a precisão permanece alta de maneira consistente. A AUC da curva PR é alta, o que indica um bom desempenho do modelo perto do ideal. Apesar do gráfico indicar um valor AUC de 1, este é um valor arredondado, mas que mesmo assim revela uma capacidade muito alta do modelo distinguir entre classes com precisão mesmo em diferentes patamares da classificação. É de notar que se a AUC fosse realmente 1, então a curva PR seria uma reta horizontal com precisão 1,0, mesmo quando o *recall* chegasse a 1

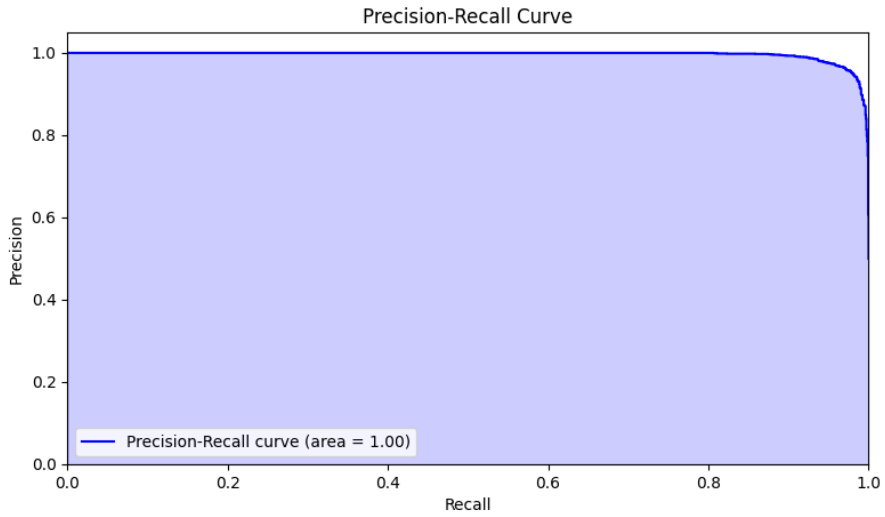


Figura 32 – Gráfico da Curva PR e respetiva AUC do modelo executado em SqueezeNet para a classificação binária do campo "benigno"

Ao analisar o gráfico da Figura 33, podemos concluir que esta curva apresenta uma capacidade quase perfeita na classificação das duas classes. A AUC desta curva, similar ao que aconteceu na curva PR, é de aproximadamente 1, o que demonstra um desempenho próximo do ideal. Este valor sugere que o modelo tem uma taxa elevada de verdadeiros positivos e uma baixa taxa de falsos positivos, o que mostra ser capaz de identificar corretamente a maioria dos casos de Doença Fibrocística com um número relativamente baixo de casos falsos desta patologia.

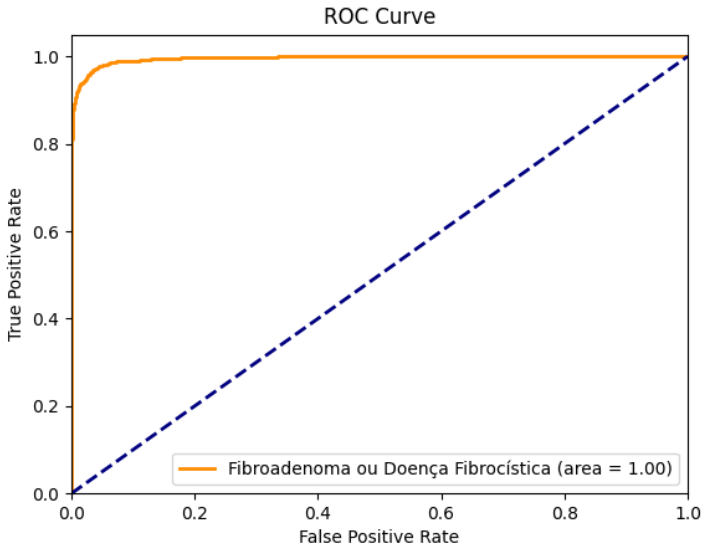


Figura 33 – Curva ROC e respetivo valor AUC do modelo SqueezeNet na classificação binária do campo "benigno"

A última métrica para este modelo é a curva de exatidão ao longo dos 15 *epochs* de treino. Ao analisar as curvas de exatidão da Figura 34, podemos notar que a curva de exatidão de validação apresenta um comportamento volátil nas primeiras *epoch*, o que é comum nas primeiras *epochs* de treino à medida que os pesos da rede estão a ser ajustados substancialmente. A curva de exatidão da fase de treino tem uma tendência crescente ao longo das *epochs*, com um maior declive nas primeiras *epochs* demonstrando, posteriormente, uma tendência de estabilização, aproximando-se dos 97,5%. Isto indica que o modelo aprendeu eficazmente as características dos dados de treino. Por outro lado, a exatidão da validação, apesar de apresentar algumas flutuações, mostra um aumento significativo, sugerindo que o modelo não está apenas a memorizar os dados de treino, mas também a generalizar bem para os dados de validação.

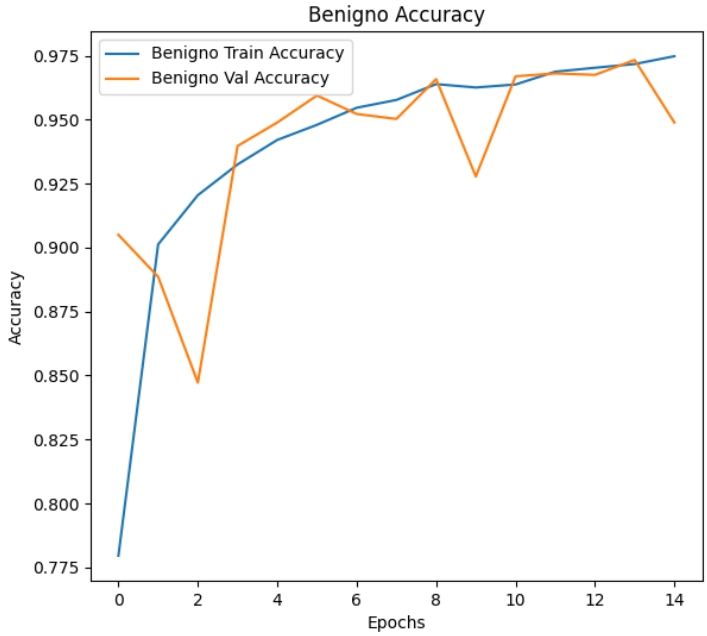


Figura 34 – Curva de exatidão do modelo SqueezeNet na classificação binária do campo "benigno"

Após completar o modelo SqueezeNet, foi executado o modelo InceptionV3 em que o tempo de execução varia entre 2 horas e 2 horas e 30 minuto por *epoch*, o que perfaz um total de 30 a 37 horas e meia ao longo das 15 *epochs* que foram executadas.

No treino deste modelo foi obtida a matriz confusão da Figura 35, é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence a um caso de Fibroadenoma, representado por zero (0), ou a um caso de Doença Fibrocística, representado por um (1).

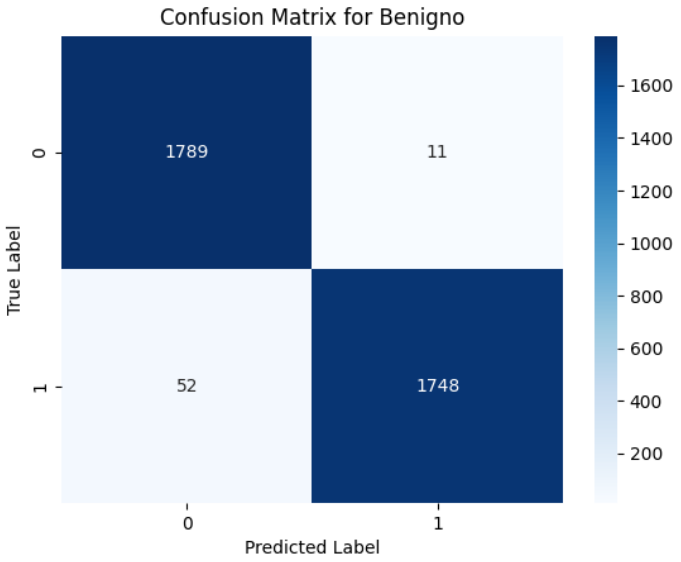


Figura 35 – Matriz confusão do modelo executado em InceptionV3 para a classificação binária do campo "benigno". O Valor zero (0) corresponde a Fibroadenoma e um (1) a Doença Fibrocística

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1748 casos foram corretamente identificados como Doença Fibrocística.
- **Falsos Negativos (FN):** 52 casos foram erradamente classificados como Fibroadenoma, quando na verdade eram Doença Fibrocística.
- **Verdadeiros Negativos (VN):** 1789 casos foram corretamente identificados como Fibroadenoma.
- **Falsos Positivos (FP):** 11 casos foram erradamente classificados como Doença Fibrocística, quando na verdade eram Fibroadenoma.

Nesta matriz confusão podemos concluir que o modelo tem uma quantidade relativamente baixa de erros, com maior tendência a confundir alguns casos de Fibroadenoma como Doença Fibrocística. Esta quantidade de erros é positiva na validação deste modelo.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 36.

Benigno Metrics:

	precision	recall	f1-score	support
Fibroadenoma	0.97	0.99	0.98	1800
Doença Fibrocística	0.99	0.97	0.98	1800
accuracy			0.98	3600
macro avg	0.98	0.98	0.98	3600
weighted avg	0.98	0.98	0.98	3600

Figura 36 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo “benigno”

A partir desta imagem podemos obter as seguintes métricas:

- **Para Fibroadenoma (0):**
  - A **precisão**, que indica a proporção de identificações corretas de Fibroadenoma entre todas as identificações de Fibroadenoma feitas pelo modelo, foi de 97%;
  - O ***recall***, que mede a proporção de casos de Fibroadenoma corretamente identificados pelo modelo em relação a todos os casos reais de Fibroadenoma, foi de 99%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 98%.
- **Para Doença Fibrocística (1):**
  - A **precisão** foi de 99%;
  - O ***recall*** foi de 97%;
  - O ***f1-score*** foi de 98%.
- **Exatidão** geral do modelo foi de 98%.

Podemos reparar que para a patologia Fibroadenoma, o modelo obteve uma precisão de 97%, o que demonstra uma taxa alta de diagnósticos corretos para Fibroadenoma. O *recall* foi de 99%, o que indica que quase todos os casos de Fibroadenoma foram identificados. Pelo *f1-score*, podemos dizer que o desempenho do modelo em identificar este tipo de patologia é muito eficiente.

Para a Doença Fibrocística, a precisão foi de 99%, que indica uma taxa de diagnósticos corretos de Doença Fibrocística mais alto que da patologia anterior, o que indica que 99% dos casos dados como Doença Fibrocística estão corretos. O *recall* foi de 97%, que apesar de ser inferior que para a patologia anterior, continua a ser um valor elevado, confirmado pelo f1-score de 98% que indica que o modelo teve um bom desempenho.

A exatidão global do modelo chegou a 98%, o que é um resultado bastante positivo, pois indica que cerca de 98% das classificações foram acertadas.

A curva PR ilustrada na Figura 37 demonstra que à medida que o *recall* aumenta, a precisão permanece alta de maneira consistente. A AUC da curva PR é alta, o que indica um bom desempenho do modelo perto do ideal. Apesar do gráfico indicar um valor AUC de 1, este é um valor arredondado, mas que está muito perto de ter uma capacidade perfeita do modelo distinguir entre classes.

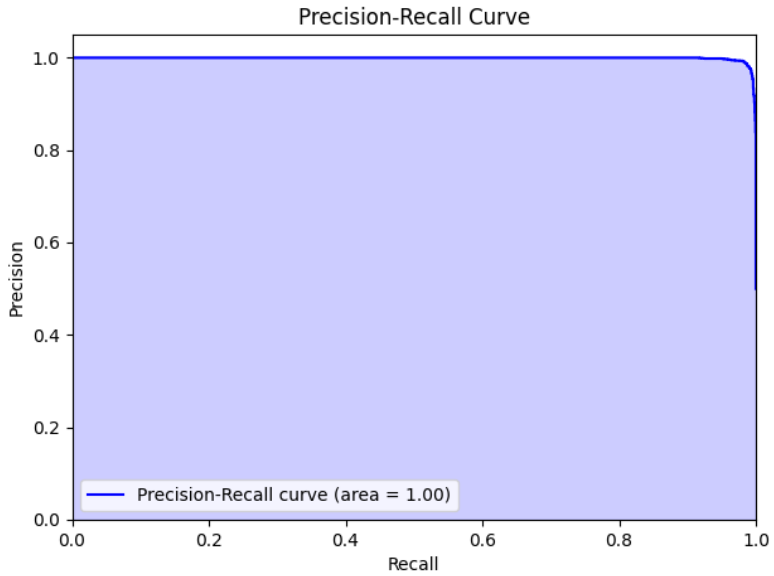


Figura 37 – Gráfico da Curva PR e respetiva AUC do modelo executado em InceptionV3 para a classificação binária do campo "benigno"

Ao analisar o gráfico ROC da Figura 38, podemos concluir que esta curva apresenta uma capacidade quase perfeita na classificação das duas classes. Podemos reparar que a curva é muito próxima de ser reta em 1. A AUC desta curva, similar ao que aconteceu na curva PR, é de aproximadamente 1, o que demonstra um desempenho próximo do ideal. Este valor sugere que o modelo tem uma taxa elevada de verdadeiros positivos e uma baixa taxa de falsos positivos, o que mostra ser capaz de identificar corretamente a maioria dos casos de Doença Fibrocística com um número relativamente baixo de casos falsos desta patologia.

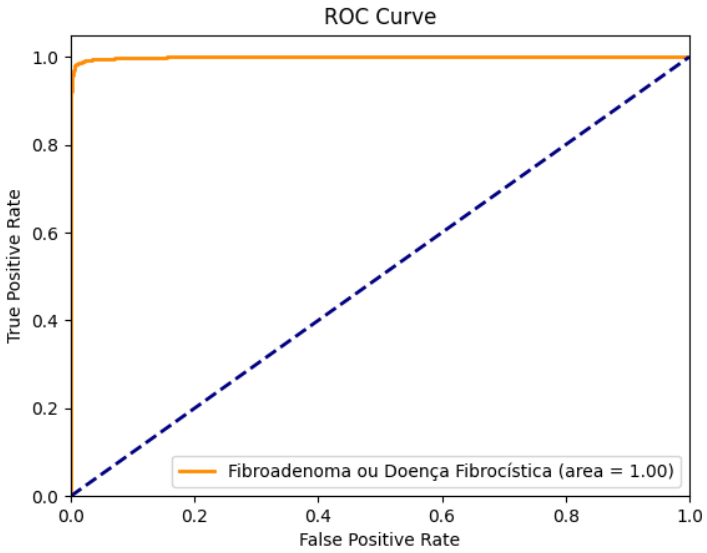


Figura 38 - Curva ROC e respetivo valor AUC do modelo InceptionV3 na classificação binária do campo "benigno"

O gráfico das curvas de exatidão obtidas deste modelo (Figura 39), demonstra uma curva de exatidão de treino mantém-se consistentemente acima de 98%, o que é indicativo de um elevado desempenho por parte do modelo. Por outro lado, a curva laranja, referente à exatidão de validação apresenta uma variação maior, com valores a oscilar entre aproximadamente 92% e 98%. É notável uma queda significativa por volta da décima *epoch*, onde a exatidão de validação desce abruptamente abaixo dos 94%, antes de retornar aos valores anteriores. Esta variação na precisão de validação sugere um possível *overfitting* do modelo aos dados de treino, o que é indicado pela discrepância entre as exatidões de treino e de validação. No entanto, a recuperação após a queda indica que o modelo foi capaz de generalizar a partir do ajuste realizado em resposta ao *overfitting*.

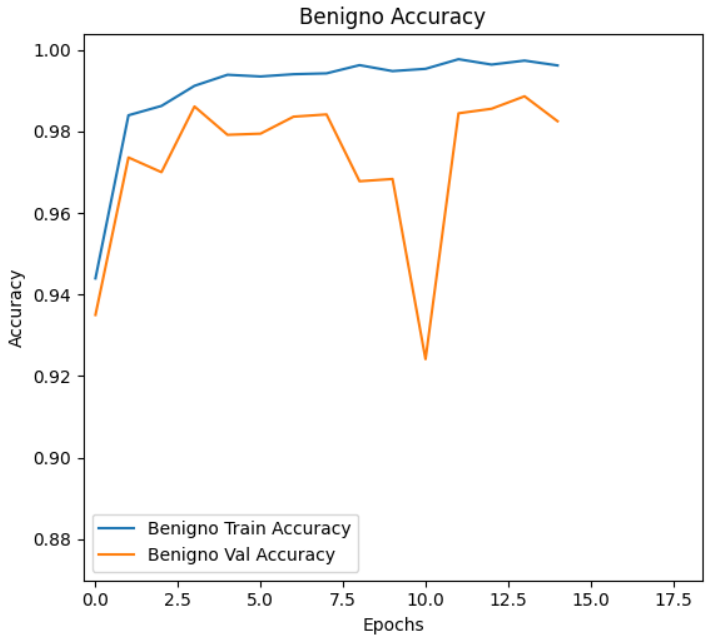


Figura 39 – Curva de exatidão do modelo InceptionV3 na classificação binária do campo "benigno"

✓ COMPARAÇÃO/DISCUSSÃO

Ao comparar os resultados dos dois modelos, podemos concluir que para este cenário ambos mostraram métricas de desempenho elevadas. A matriz de confusão da SqueezeNet revelou um elevado número de casos corretos para ambas as patologias com poucos erros, o que indica uma distinção precisa entre as classes. A InceptionV3 apresentou resultados semelhantes, com uma ligeira melhoria na identificação de Fibroadenoma.

Esta tendência manteve-se nas métricas de precisão, *recall* e *F1-score*, onde ambos os modelos apresentaram bons desempenhos, porém a InceptionV3 apresentou uma ligeira vantagem, evidenciando um equilíbrio entre a exatidão e a capacidade de identificação de casos de Doença Fibrocística.

Outra métrica em que os dois modelos se diferenciaram ligeiramente foi nas curvas de exatidão. Enquanto a SqueezeNet demonstrou uma ligeira volatilidade na exatidão de validação, a InceptionV3 exibiu uma exatidão de validação mais estável, sugerindo uma melhor generalização para os dados não vistos, apesar do pico negativo numa das *epochs* que pode ser indicativo de *overfitting*.

Em conclusão, ambos os modelos parecem capazes de serem eficientes na classificação de patologias benignas. Porém, o InceptionV3, apesar de alguma volatilidade na exatidão, apresenta resultados mais equilibrados e promissores, pelo que seria o modelo mais eficaz neste cenário. Considerando que a Doença Fibrocística, possui um potencial de evoluir para uma possível componente maligna, a minimização de falsos negativos assume uma relevância clínica prioritária, visto que cada caso não detetado pode retardar intervenções cruciais para evitar a progressão para estados neoplásicos malignos. Ainda que o modelo SqueezeNet apresente uma menor taxa de falsos positivos, o impacto clínico de um falso negativo na Doença Fibrocística é mais crítico, justificando a preferência pelo modelo InceptionV3 na presente análise.

4.2.3. Cenário 3: Classificação binária do campo “carcinoma”

Tal como nos cenários anteriores, o treino para a classificação binária do campo “carcinoma” Iniciou-se com o treino do modelo em SqueezeNet. A execução do modelo teve uma duração média de

20 a 25 minutos por *epoch* o que perfaz uma duração total de 5 horas a 6 horas e 15 minutos ao longo das 15 *epochs* que foram executadas.

Do treino deste modelo foi obtida a matriz confusão da Figura 40. Desta matriz é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence à patologia Carcinoma Ductal, representado por zero (0), ou então a Carcinoma Lobular, representado por um (1).

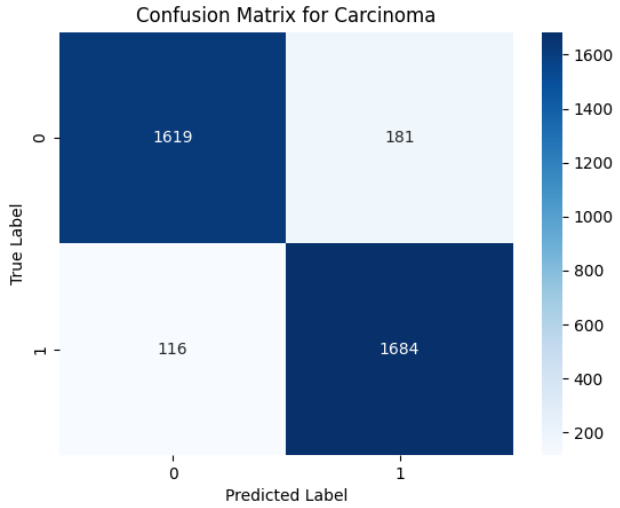


Figura 40 – Matriz confusão do modelo executado em SqueezeNet para a classificação binária do campo "carcinoma". O Valor zero (0) corresponde a Carcinoma Ductal e um (1) a Carcinoma Lobular

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1684 casos foram corretamente identificados como Carcinoma Lobular.
- **Falsos Negativos (FN):** 116 casos foram erradamente classificados como Carcinoma Ductal, quando na verdade eram Carcinoma Lobular.
- **Verdadeiros Negativos (VN):** 1619 casos foram corretamente identificados como Carcinoma Ductal.
- **Falsos Positivos (FP):** 181 casos foram erradamente classificados como Carcinoma Lobular, quando na verdade eram Carcinoma Ductal.

Podemos observar que a grande maioria das vezes o modelo prevê corretamente ambas as classes. Porém, é de notar que ao errar, o modelo confunde mais facilmente Carcinomas Ductais como sendo Carcinomas Lobulares do que o inverso.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 31.

```
Carcinoma Metrics:
                precision    recall  f1-score   support

Carcinoma Ductal       0.93      0.90      0.92      1800
Carcinoma Lobular      0.90      0.94      0.92      1800

        accuracy                           0.92      3600
       macro avg       0.92      0.92      0.92      3600
    weighted avg       0.92      0.92      0.92      3600
```

Figura 41 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo "carcinoma"

A partir desta imagem podemos obter as seguintes métricas:

- **Para Carcinoma Ductal (0):**
  - A **precisão**, que indica a proporção de identificações corretas de Carcinoma Ductal entre todas as identificações de Carcinoma Ductal feitas pelo modelo, foi de 93%;
  - O ***recall***, que mede a proporção de casos de Carcinoma Ductal corretamente identificados pelo modelo em relação a todos os casos reais de Carcinoma Ductal foi de 90%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 92%.
- **Para Carcinoma lobular (1):**
  - A **precisão** foi de 90%;
  - O ***recall*** foi de 94%;
  - O ***f1-score*** foi de 92%.
- **Exatidão** geral do modelo foi de 92%.

Para Carcinoma Ductal, o modelo alcançou uma precisão de 93%, que apesar de ser um resultado alto, é um resultado que fica aquém de alguns dos resultados obtidos em cenários anteriores. O *recall* foi de 92%, refletindo a proporção de casos de Carcinoma Ductal corretamente identificados em relação ao total de casos reais. O *F1-score* foi de 92%, sugerindo um equilíbrio adequado entre as duas métricas, mesmo que ligeiramente mais baixo que em alguns modelos de cenários anteriores.

Em relação à Carcinoma Lobular, a precisão foi ligeiramente inferior, com um valor de 90%, enquanto o *recall* foi de 94%, o que indica que foram identificados 94% dos casos de Carcinoma Lobular. O *F1-score* manteve-se constante em 92%, o que demonstra um modelo equilibrado para ambas as patologias, apesar de menos eficiente que parte dos modelos treinados em cenários anteriores.

A exatidão global do modelo foi de 92%, o que implica que 92% das classificações foram corretas, o que indica bom desempenho do modelo, apesar de mais uma vez ser inferior ao já registado em cenários anteriores.

A curva PR ilustrada na Figura 42 demonstra que à medida que o *recall* aumenta, a precisão do modelo se mantém elevada, com o valor a decair apenas para valores elevados do recall. O AUC desta curva é 0.98, valor próximo de 1, o que indica que o modelo é equilibrado.

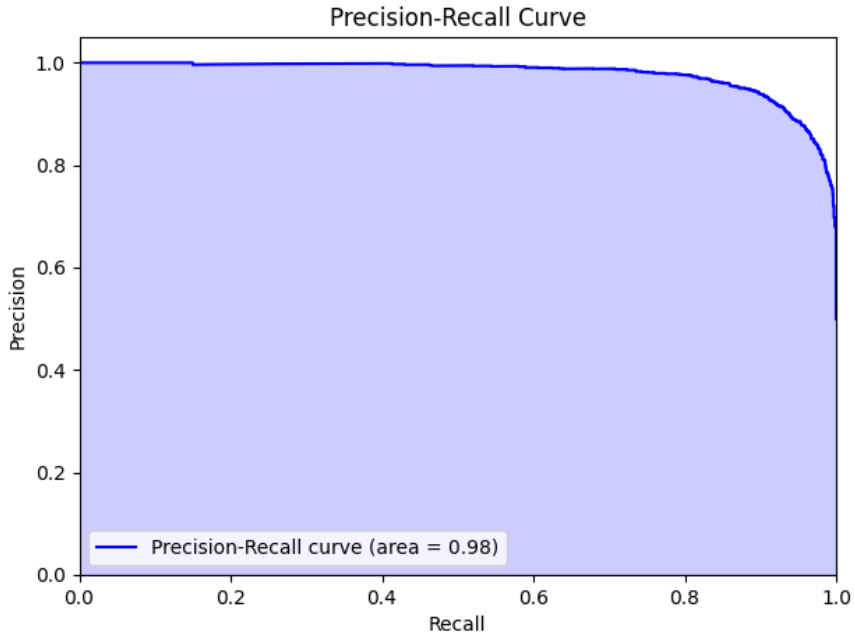


Figura 42 – Gráfico da Curva PR e respetiva AUC do modelo executado em SqueezeNet para a classificação binária do campo "carcinoma"

Ao analisar o gráfico da Figura 43, podemos concluir que esta curva apresenta um bom equilibro na distinção entre as duas patologias malignas. A AUC desta curva, tem o valor de 0,98 o que demonstra um bom desempenho do modelo e uma taxa relativamente alta de identificar Carcinomas Ductais em relação à taxa de casos erradamente identificados para esta patologia.

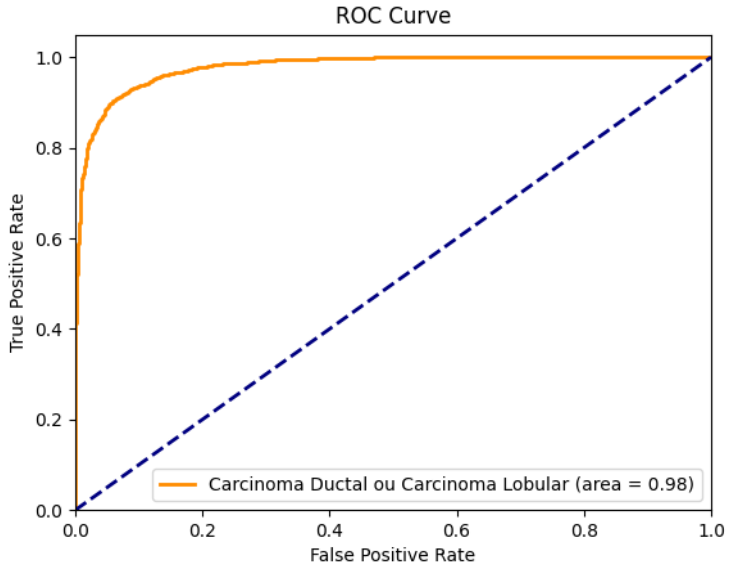


Figura 43 – Curva ROC e respetivo valor AUC do modelo SqueezeNet na classificação binária do campo "carcinoma"

A curva de exatidão ao longo dos 15 *epochs* de treino da Figura 44, permite analisar e concluir que ambas as curvas de exatidão (treino e validação) apresentam um comportamento crescente ao longo das 15 *epochs*. A primeira manteve-se crescente ao longo do tempo, não conseguindo estabilizar ao final das últimas *epochs*. A curva de exatidão da fase de validação, apesar de alguma volatilidade, manteve a tendência que a curva de treino, acompanhando de perto em termos de valor a curva de treino. Ambas as evoluções indicam que possivelmente o modelo não atingiu o potencial de desempenho e que seria necessário mais algumas *epochs* para que os resultados fossem melhores.

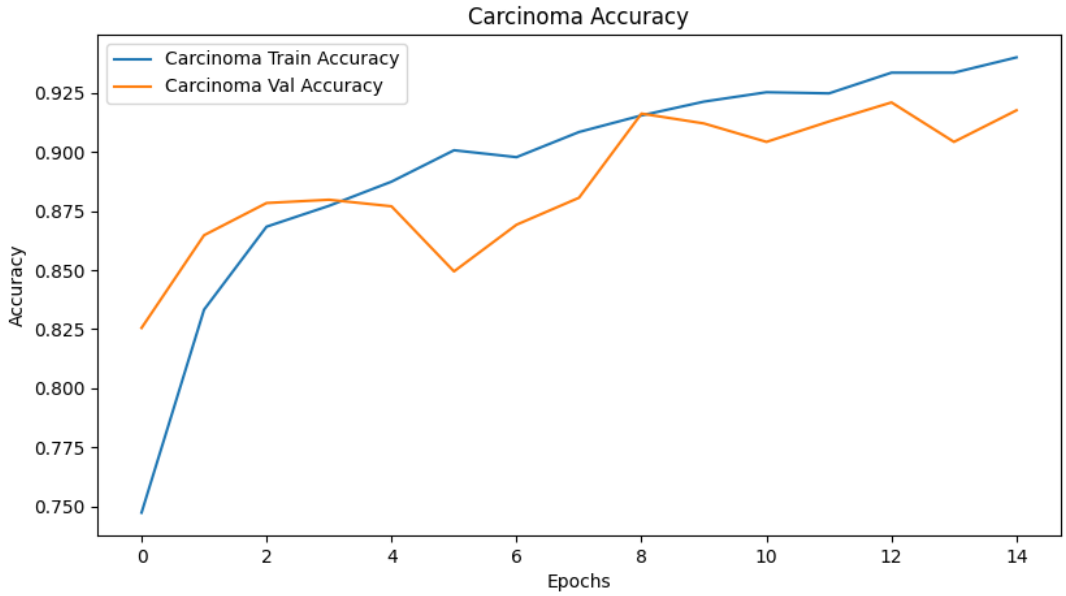


Figura 44 – Curva de exatidão do modelo SqueezeNet na classificação binária do campo "carcinoma"

Após completar o modelo SqueezeNet, foi executado o modelo InceptionV3 em que o tempo de execução varia entre 2 horas e 2 horas e 30 minuto por *epoch*, o que perfaz um total de 30 a 37 horas e meia ao longo das 15 *epochs* que foram executadas.

Do treino deste modelo foi obtida a matriz confusão da Figura 45. Desta matriz é possível avaliar em termos quantitativos a capacidade deste modelo distinguir se uma imagem pertence à patologia Carcinoma Ductal, representado por zero (0), ou então a Carcinoma Lobular, representado por um (1).

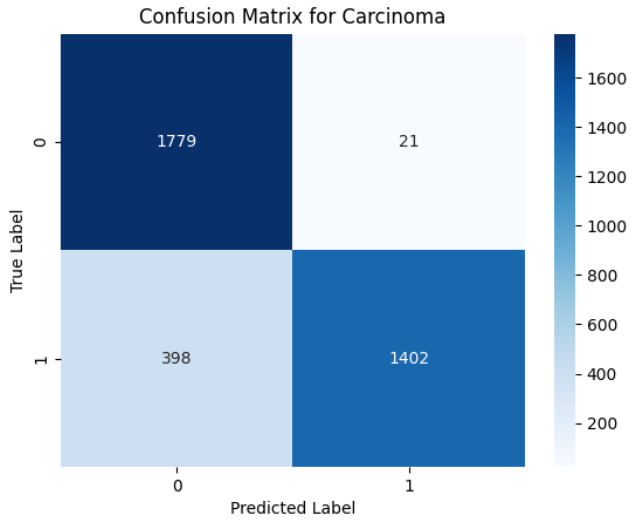


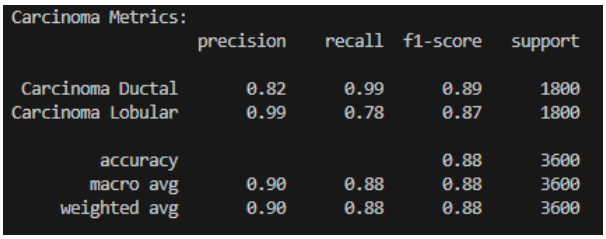
Figura 45 – Matriz confusão do modelo executado em InceptionV3 para a classificação binária do campo "carcinoma". O Valor zero (0) corresponde a Carcinoma Ductal e um (1) a Carcinoma Lobular

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1402 casos foram corretamente identificados como Carcinoma Lobular.
- **Falsos Negativos (FN):** 398 casos foram erradamente classificados como Carcinoma Ductal, quando na verdade eram Carcinoma Lobular.
- **Verdadeiros Negativos (VN):** 1779 casos foram corretamente identificados como Carcinoma Ductal.
- **Falsos Positivos (FP):** 21 casos foram erradamente classificados como Carcinoma Lobular, quando na verdade eram Carcinoma Ductal.

Podemos observar que a grande maioria das vezes o modelo prevê corretamente ambas as classes. Porém, é de notar que ao contrário do modelo anterior para este cenário, o modelo erra mais facilmente Carcinomas Lobulares como sendo Carcinomas Ductais do que o inverso.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 46.



Carcinoma Metrics:

	precision	recall	f1-score	support
Carcinoma Ductal	0.82	0.99	0.89	1800
Carcinoma Lobular	0.99	0.78	0.87	1800
accuracy			0.88	3600
macro avg	0.90	0.88	0.88	3600
weighted avg	0.90	0.88	0.88	3600

Figura 46 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo "carcinoma"

A partir desta imagem podemos obter as seguintes métricas:

- **Para Carcinoma Ductal (0):**
  - A **precisão**, que indica a proporção de identificações corretas de Carcinoma Ductal entre todas as identificações de Carcinoma Ductal feitas pelo modelo, foi de 82%;
  - O ***recall***, que mede a proporção de casos de Carcinoma Ductal corretamente identificados pelo modelo em relação a todos os casos reais de Carcinoma Ductal foi de 99%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 89%.
- **Para Carcinoma lobular (1):**
  - A **precisão** foi de 99%;
  - O ***recall*** foi de 78%;
  - O ***f1-score*** foi de 87%.
- **Exatidão** geral do modelo foi de 90%.

Para Carcinoma Ductal, o modelo alcançou uma precisão de 82%, o que é baixo de acordo com o cenário de diagnóstico de tipos de carcinomas. Apesar disso, o *recall* foi de 99%, refletindo a proporção de casos de Carcinoma Ductal corretamente identificados em relação ao total de casos reais. O *F1-score* foi de 89%, sugerindo um equilíbrio baixo em relação aos modelos já testados.

Em relação à Carcinoma Lobular, a precisão 99%, o que demonstra uma taxa alta de diagnósticos corretos para Carcinoma Lobular, enquanto o *recall* foi de 78%, o que revela que o modelo não identifica parte dos casos de Carcinoma Lobular, deixando um número significativo deles por diagnosticar, o que se traduz num f1-score de 87%.

A exatidão global do modelo foi de 90%, o que implica que 90% das classificações foram corretas, o que indica um razoável desempenho do modelo, apesar de mais uma vez ser inferior ao já registado em cenários anteriores.

A curva PR ilustrada na Figura 47 demonstra que à medida que o *recall* aumenta, a precisão do modelo se mantém elevada, com o valor a decair apenas para valores elevados do recall, como acontece na generalidade dos modelos. O AUC desta curva é 0.98, valor próximo de 1, o que indica que o modelo é equilibrado.

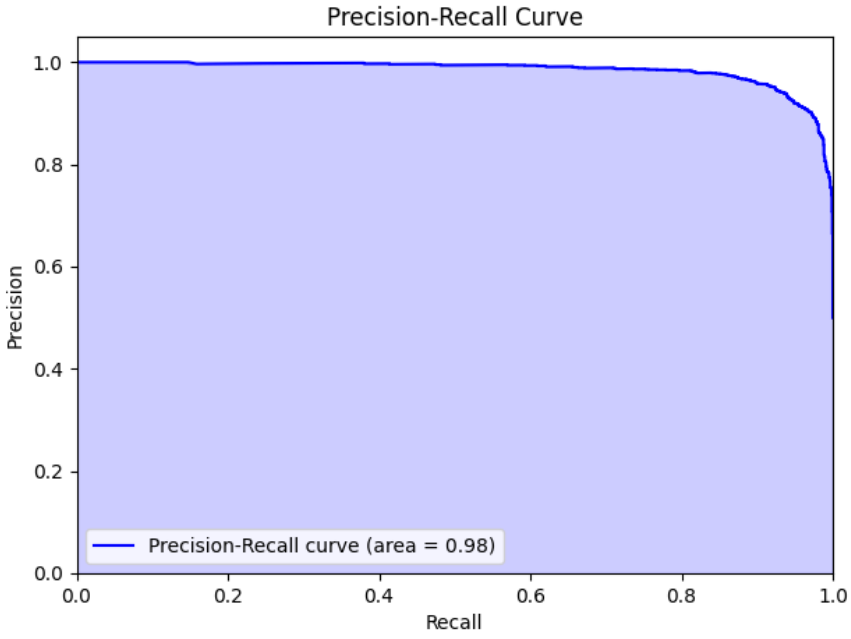


Figura 47 – Gráfico da Curva PR e respetiva AUC do modelo executado em InceptionV3 para a classificação binária do campo "carcinoma"

Ao analisar o gráfico da Figura 48, podemos concluir que esta curva apresenta um bom equilibro na distinção entre as duas patologias malignas. Tal como no modelo SqueezeNet, a AUC desta curva tem o valor de 0,99 o que demonstra um bom desempenho do modelo e uma taxa relativamente alta de identificar Carcinomas Ductais em relação à taxa de casos erradamente identificados para esta patologia.

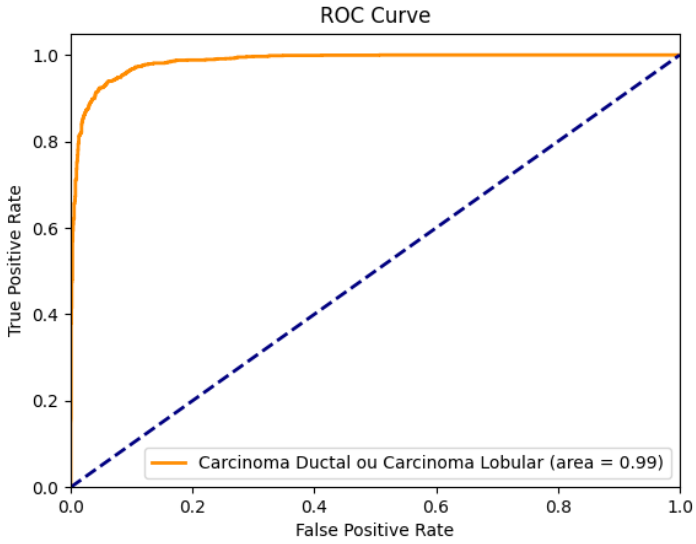


Figura 48 – Curva ROC e respetivo valor AUC do modelo InceptionV3 na classificação binária do campo "carcinoma"

A curva de exatidão ao longo dos 15 *epochs* de treino da Figura 49 demonstra uma rápida subida da curva de treino e que rapidamente estabiliza por volta dos 98%. Porém, a curva de validação evoluí dos 85% para 95% nos primeiros *epoch* e depois mantém-se muito volátil, até ao ponto que começa a decrescer. Este comportamento indicia que o modelo entrou em *overfitting* e não está a recuperar da aprendizagem de padrões que não facilitam a decisão.

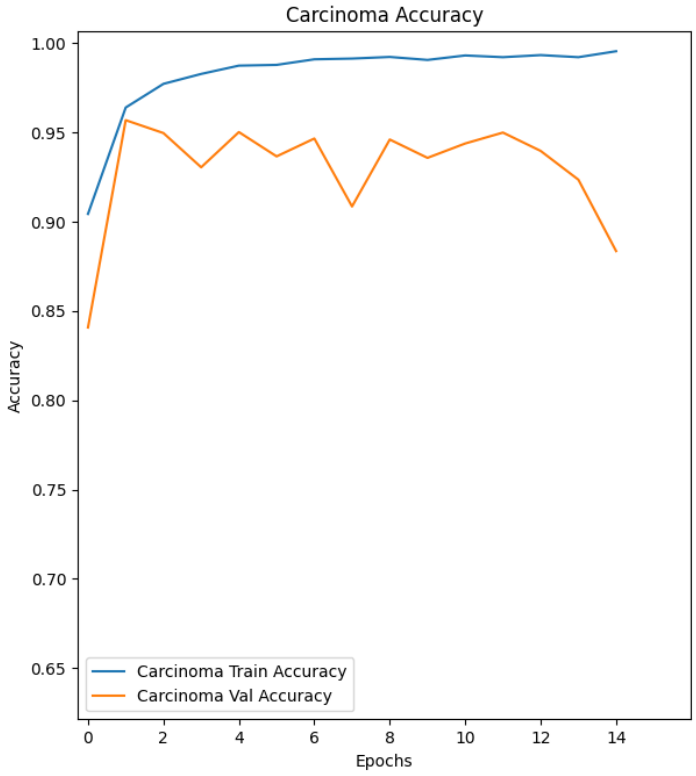


Figura 49 – Curva de exatidão do modelo InceptionV3 na classificação binária do campo "carcinoma"

✓ COMPARAÇÃO/DISCUSSÃO

Ao comparar todos os resultados dos dois modelos para este cenário, podemos concluir que apesar deste cenário apresentar valores ligeiramente mais baixos que em outros cenários, o SqueezeNet foi o que mostrou melhor resultados e com maior potencial. As matrizes confusão apresentaram melhor equilibro no SqueezeNet que no InceptionV3, bem como nas métricas como precisão, exatidão, *recall* e *f1-score*.

O facto de as curvas de exatidão serem muito instáveis no InceptionV3, demonstram que o modelo entra em *overfitting*, o que não é positivo para o cenário. Porém, a evolução das curvas respetivas para o SqueezeNet mostram que ainda é possível melhorias na aprendizagem da classificação de classes e com isso, potencialmente mais precisão que aquela que foi obtida apenas com 15 *epochs*.

Na seleção do modelo SqueezeNet em detrimento do modelo InceptionV3 para a classificação de carcinomas, a decisão passou também pela análise dos falsos negativos e falsos positivos apresentados pelas matrizes de confusão. A maior incidência de falsos negativos no modelo InceptionV3 na classificação do Carcinoma Ductal Invasivo é essencial uma vez que este apresenta uma maior agressividade, o que poderia resultar na falta de diagnóstico desta patologia. Assim, o modelo SqueezeNet, ao reduzir os falsos negativos, proporciona uma vantagem clínica significativa, assegurando uma maior probabilidade de diagnóstico do Carcinoma Ductal Invasivo, permitindo intervenções mais rápidas e adequadas, podendo assim aumentar a esperança de vida dos pacientes.

4.2.4. Cenário 4: Classificação Multiclasse Combinada

Para o último cenário em estudo, as métricas são ligeiramente diferentes sendo que temos 3 classes diferentes para previsão. O facto de duas destas classes serem multiclasse em vez de serem binárias as

métricas em estudo mudam ligeiramente, mas que utilizam as mesmas bases teóricas e matemáticas que foram utilizadas até agora.

Como nos cenários anteriores, foi testado o modelo SqueezeNet em que a duração média de 20 a 25 minutos por *epoch* o que perfaz uma duração total de 5 horas a 6 horas e 15 minutos ao longo das 15 *epochs* que foram executadas.

A primeira matriz conseguida foi a de categoria "mal_ou_beg" que é uma categoria binária onde zero (0) representa os casos benignos e um (1) os casos malignos (Figura 50).

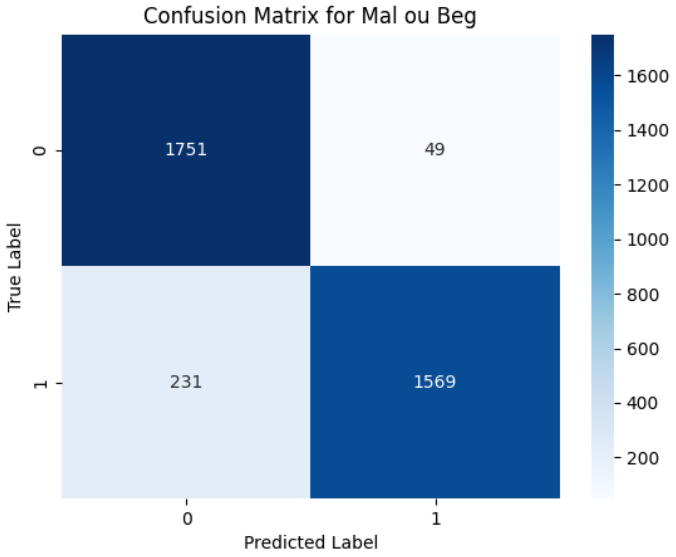


Figura 50 – Matriz confusão do modelo executado em SqueezeNet para a classificação binária do campo binário "mal_ou_beg". O Valor zero (0) corresponde a caso benigno e um (1) a caso maligno

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1569 casos foram corretamente identificados como benignos.
- **Falsos Negativos (FN):** 231 casos foram erradamente classificados como benignos, quando na verdade eram malignos.
- **Verdadeiros Negativos (VN):** 1751 casos foram corretamente identificados como benignos.
- **Falsos Positivos (FP):** 49 casos foram erradamente classificados como malignos, quando na verdade eram benignos.

Podemos reparar que, apesar a grande maioria das vezes o modelo prever corretamente as classes de uma imagem, este confunde numa quantidade considerável de casos que são malignos com os casos benignos, do que propriamente casos benignos com malignos que são erradamente identificados como tal.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 51.

Maligno ou Benigno Metrics:

	precision	recall	f1-score	support
Benigno	0.88	0.97	0.93	1800
Maligno	0.97	0.87	0.92	1800
accuracy			0.92	3600
macro avg	0.93	0.92	0.92	3600
weighted avg	0.93	0.92	0.92	3600

Figura 51 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo binário "mal_ou_beg"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe Benigna (0):**
  - A **precisão**, que indica a proporção de identificações corretas de benignidade entre todas as identificações de benignidade feitas pelo modelo, foi de 88%;
  - O ***recall***, que mede a proporção de casos benignos corretamente identificados pelo modelo em relação a todos os casos reais de benignidade, foi de 97%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 93%.
- **Para a classe Maligna (1):**
  - A **precisão** foi de 97%;
  - O ***recall*** foi de 87%;
  - O ***f1-score*** foi de 92%.
- **Exatidão** geral do modelo foi de 92%.

Com estas métricas, podemos concluir que o modelo teve um alto desempenho, mas que existe uma diferença significativa na capacidade de identificar corretamente casos benignos em comparação com os casos malignos. Para a classe benigna, a precisão de 88% sugere que o modelo está correto 88% das vezes que prevê benignidade, o que é um resultado razoável, mas que pode ser melhorado. O *recall* de 0.97 para esta classe indica que o modelo foi capaz de identificar 97% de todas as lesões benignas verdadeiras, o que comparando com outras métricas indicam mais uma vez que o modelo prevê com alguma regularidade casos malignos como sendo benignos. Isto resulta num *f1-score* de 0.93, que sugere um desempenho equilibrado e eficiente.

No que diz respeito à classe maligna, a precisão elevada de 97% indica que o modelo está certo grande parte das vezes que realiza a previsão de malignidade. Porém o *recall* de 0.87 para esta classe indica que o modelo foi capaz de identificar apenas 87% de todas as lesões malignas verdadeiras, o que se traduz num *f1-score* de 92%. Estes resultaram foram razoáveis, mas que necessitam de melhorias porque 13% dos casos malignos não serem diagnosticados é relativamente baixo para a importância do cenário.

A exatidão global é a métrica que valida todas as premissas anteriores, pois o seu valor de 92% é um bom valor que no indica que o modelo é aceitável para a distinção entre maligno e benigno.

A segunda matriz conseguida foi a de categoria "benigno" que é uma categoria ternária, que ao invés de conter dois valores possíveis, contém três, onde zero (0) representa os casos que não são benignos, um (1) os casos de Fibroadenoma e dois (2) como Doença Fibrocística (Figura 52).

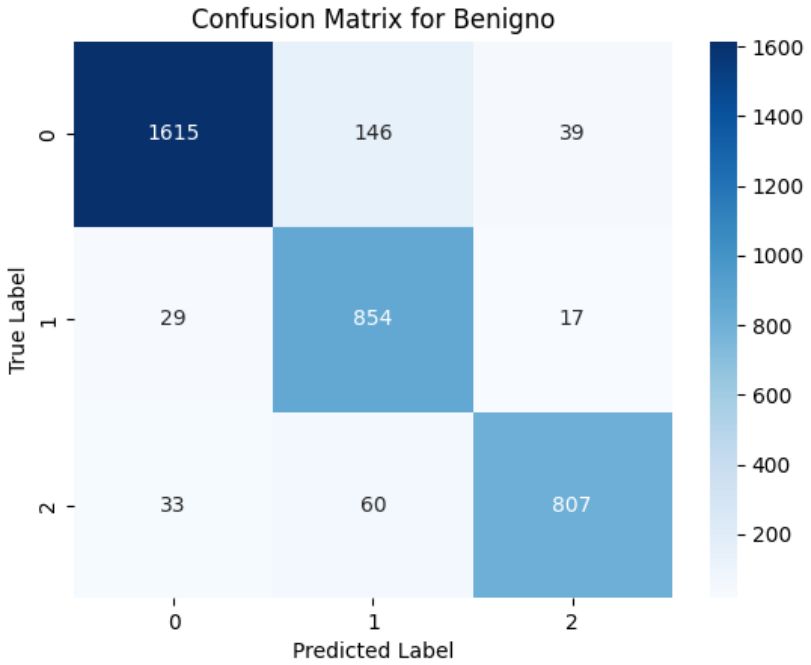


Figura 52 – Matriz confusão do modelo executado em SqueezeNet para a classificação ternária do campo "benigno". O Valor zero (0) representa os casos que não são benignos, um (1) os casos de Fibroadenoma e dois (2) como Doença Fibrocística

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Casos Corretos de "Não benigno":** 1615 casos foram corretamente identificados como não benignos.
- **Casos Corretos de Fibroadenoma:** 854 casos foram corretamente identificados como Fibroadenoma.
- **Casos Corretos de Doença Fibrocística:** 807 casos foram corretamente identificados como Doença Fibrocística.
- **Casos Incorretos de "Não benigno":** 62 casos foram incorretamente identificados como não benignos.
- **Casos Incorretos de Fibroadenoma:** 206 casos foram incorretamente identificados como Fibroadenoma.
- **Casos Incorretos de Doença Fibrocística:** 56 casos foram incorretamente identificados como Doença Fibrocística.

Desta matriz podemos reparar que o modelo acerta grande parte das vezes no diagnóstico das imagens na categoria de classificação benigna. Podemos também reparar que o modelo identifica mais vezes casos incorretos de Fibroadenoma, com maior quantidade se a imagem não for benigna.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 53.

Benigno Metrics:

	precision	recall	f1-score	support
Categoria 0	0.96	0.90	0.93	1800
Categoria 1	0.81	0.95	0.87	900
Categoria 2	0.94	0.90	0.92	900
accuracy			0.91	3600
macro avg	0.90	0.91	0.91	3600
weighted avg	0.92	0.91	0.91	3600

Figura 53 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo ternário "benigno"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe "Não Benigno" (0):**
  - A **precisão**, que indica a proporção de identificações corretas casos não benignos entre todas as identificações de benignidade feitas pelo modelo, foi de 96%;
  - O ***recall***, que mede a proporção de casos não benignos corretamente identificados pelo modelo em relação a todos os casos reais não benignos, foi de 90%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 93%.
- **Para a classe Fibroadenoma (1):**
  - A **precisão** foi de 81%;
  - O ***recall*** foi de 95%;
  - O ***f1-score*** foi de 87%.
- **Para a classe Doença Fibrocística (1):**
  - A **precisão** foi de 94%;
  - O ***recall*** foi de 90%;
  - O ***f1-score*** foi de 92%.
- **Exatidão** geral do modelo foi de 91%.

Pelos valores destas métricas, podemos reparar que na generalidade o modelo é equilibrado e eficiente a prever a classificação de imagens quanto à sua benignidade. Existe uma discrepância em Fibroadenoma que confirma através da baixa precisão da classe de Fibroadenoma a suspeita levantada na matriz anterior.

Sendo um classificador mais complexo por ter três valores possíveis, podemos dizer que a exatidão global desta fase de classificação é boa, pelo que acerta no diagnóstico 91% das vezes.

A última matriz conseguida foi a de categoria "carcinoma" que é uma categoria ternária, que ao invés de conter dois valores possíveis, contém três, onde zero (0) representa os casos que não são malignos, um (1) os casos de Carcinoma Ductal e dois (2) como Carcinoma Lobular (Figura 54).

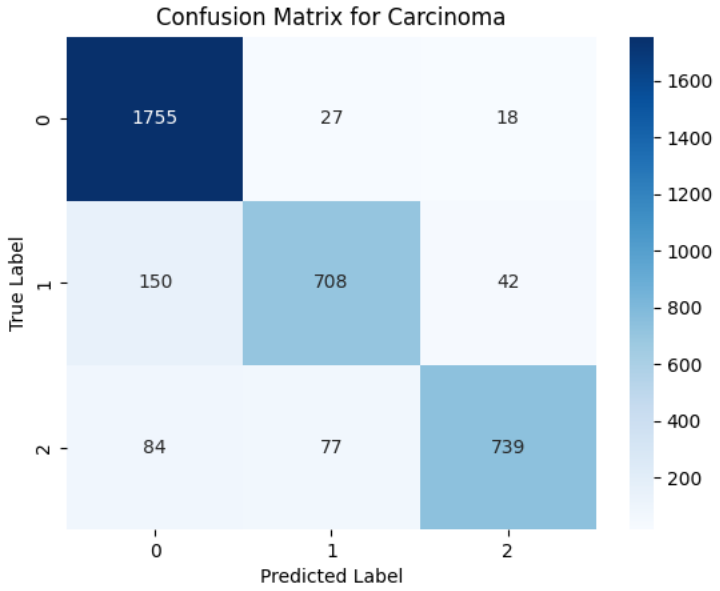


Figura 54 – Matriz confusão do modelo executado em SqueezeNet para a classificação ternária do campo "carcinoma". O Valor zero (0) representa os casos que não são malignos, um (1) os casos de Carcinoma Ductal e dois (2) como Carcinoma Lobular

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Casos Corretos de "Não maligno":** 1755 casos foram corretamente identificados como não malignos.
- **Casos Corretos de Carcinoma Ductal:** 708 casos foram corretamente identificados como Carcinoma Ductal.
- **Casos Corretos de Carcinoma Lobular:** 739 casos foram corretamente identificados como Carcinoma Lobular.
- **Casos Incorretos de "Não maligno":** 234 casos foram incorretamente identificados como não malignos.
- **Casos Incorretos de Carcinoma Ductal:** 104 casos foram incorretamente identificados como Carcinoma Ductal.
- **Casos Incorretos de Carcinoma Lobular:** 60 casos foram incorretamente identificados como Doença Carcinoma Lobular.

Desta matriz podemos reparar que o modelo acerta grande parte das vezes no diagnóstico das imagens na categoria de classificação benigna. É identificável que o modelo classifica mais vezes casos incorretos de não maligno, com maior quantidade se a imagem for de casos de Carcinoma Ductal.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 55.

Carcinoma Metrics:

	precision	recall	f1-score	support
Categoria 0	0.88	0.97	0.93	1800
Categoria 1	0.87	0.79	0.83	900
Categoria 2	0.92	0.82	0.87	900
accuracy			0.89	3600
macro avg	0.89	0.86	0.87	3600
weighted avg	0.89	0.89	0.89	3600

Figura 55 – Métricas de validação na execução do modelo SqueezeNet para a classificação do campo ternário "benigno"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe "Não Maligno" (0):**
  - A **precisão**, que indica a proporção de identificações corretas de não malignos entre todas as identificações de benignidade feitas pelo modelo, foi de 88%;
  - O ***recall***, que mede a proporção de casos não malignos corretamente identificados pelo modelo em relação a todos os casos reais não malignos, foi de 97%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 93%.
- **Para a classe Carcinoma Ductal (1):**
  - A **precisão** foi de 87%;
  - O ***recall*** foi de 79%;
  - O ***f1-score*** foi de 83%.
- **Para a classe Carcinoma Lobular (1):**
  - A **precisão** foi de 92%;
  - O ***recall*** foi de 82%;
  - O ***f1-score*** foi de 87%.
- **Exatidão** geral do modelo foi de 89%.

Através destas métricas, podemos concluir que este modelo teve valores razoáveis de eficiência e precisão, porém aquém do esperado. Estes valores devem-se ao facto de apenas 79% dos casos de Carcinoma Ductal foram identificados, o que juntando à conclusão que foi obtida da matriz anterior, podemos concluir que a maior parte dos Carcinomas Ductais que não foram corretamente classificados, foram classificados como não sendo malignos.

A exatidão geral do modelo demonstrou um valor razoável de 89%, mas que é ligeiramente mais baixo que esperado.

Ao analisar a curva ROC para a classificação de maligno e benigno (Figura 56), podemos concluir que esta apresenta uma um bom equilíbrio na distinção entre maligno e benigno. A AUC desta curva apresenta um valor de 0,99, o que é próximo do ideal.

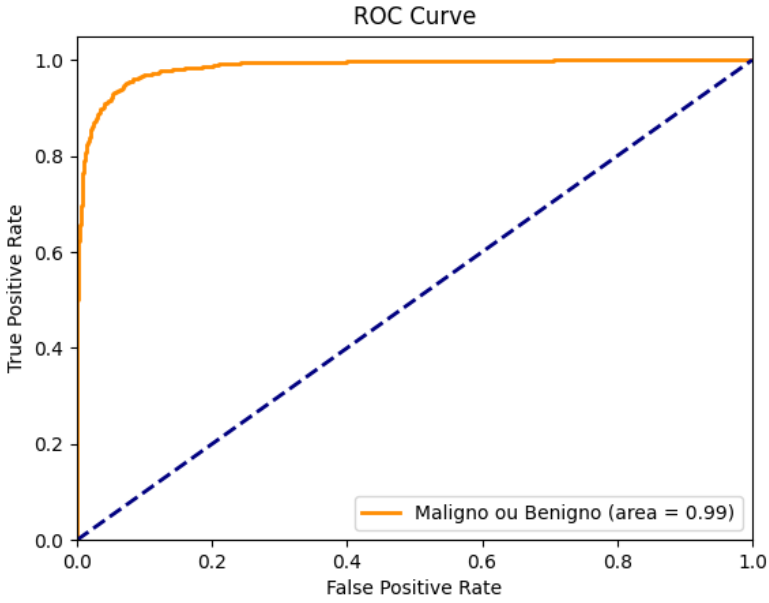


Figura 56 – Curva ROC e respetivo valor AUC do modelo SqueezeNet na classificação binária do campo "mal_ou_beg"

O mesmo acontece para as curvas ROC multiclasse que estão representadas na Figura 57 e na Figura 58. Podemos reparar que a área de todas as curvas onde maior parte tem um valor 0,99, o que demonstra uma boa capacidade de distinção de cada classe em relação às restantes classes. A única curva que teve valor inferior (0,97), foi a curva ROC da classe Carcinoma Ductal da categoria "carcinoma", o que reflete os valores obtidos nas métricas anteriores para este cenário.

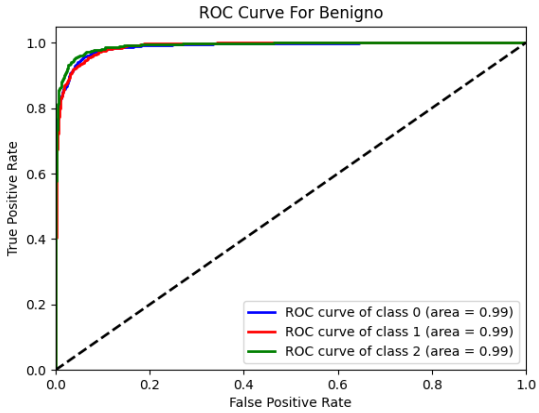


Figura 58 – Curvas ROC e respetivos valores AUC do modelo SqueezeNet na classificação ternária do campo "benigno"

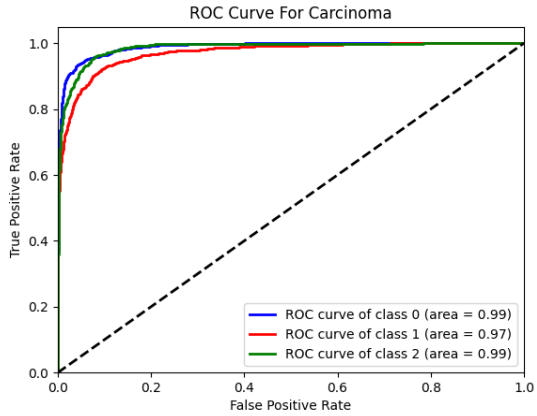


Figura 57 – Curvas ROC e respetivos valores AUC do modelo SqueezeNet na classificação ternária do campo "carcinoma"

No final, foi obtido os gráficos de exatidão ao longo das 15 *epochs* ocorridas para cada uma. NaFigura 59, podemos reparar que todas as curvas de exatidão de treino têm tendência crescente ao longo das *epochs*, o que demonstra uma aprendizagem contínua do modelo. As exatidões das curvas de treino ficaram entre os 92% e os 95%, o que indica que este modelo ainda tem capacidade que aprender com os dados de treino, com um potencial de exatidão maiores. O mesmo acontece às curvas de exatidão de

validação, que acompanham as curvas de treino ao logo das *epochs*, que apesar de terem alguma volatilidade, não demonstram sinais de *overfitting* nem *underfitting*. Os valores de exatidão das curvas de validação ficaram para valores entre os 89 e os 92% e que poderiam ser favorecidas com mais *epochs* de treino.

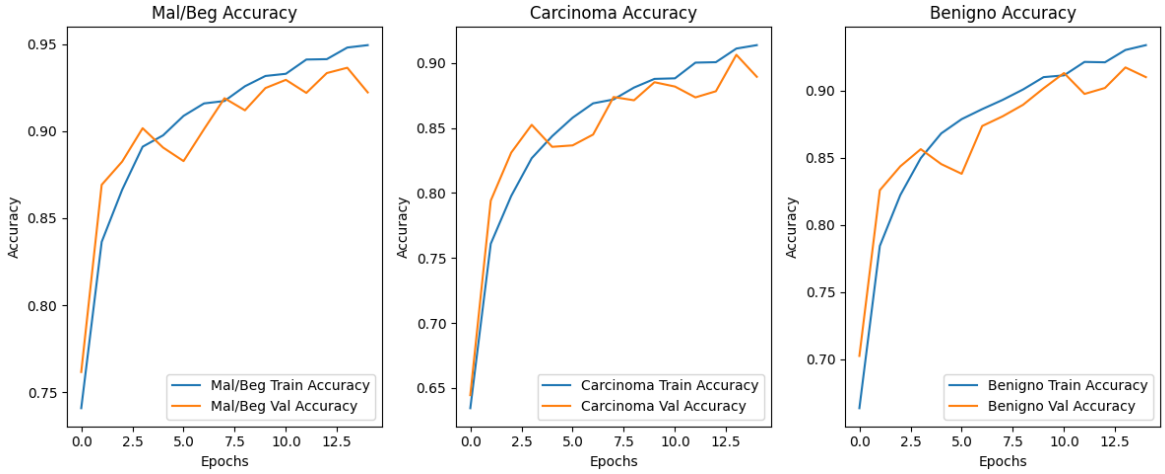


Figura 59 – Curvas de exatidão do modelo SqueezeNet no cenário de classificação multiclasse

Tal como em todos os cenários anteriores, após completar o modelo SqueezeNet, foi executado o modelo InceptionV3 em que o tempo de execução varia entre 2 horas e 2 horas e 30 minuto por *epoch*, o que perfaz um total de 30 a 37 horas e meia ao longo das 15 *epochs* que foram executadas.

A primeira matriz conseguida foi a de categoria "mal_ou_beg" que é uma categoria binária onde zero (0) representa os casos benignos e um (1) os casos malignos (Figura 60).

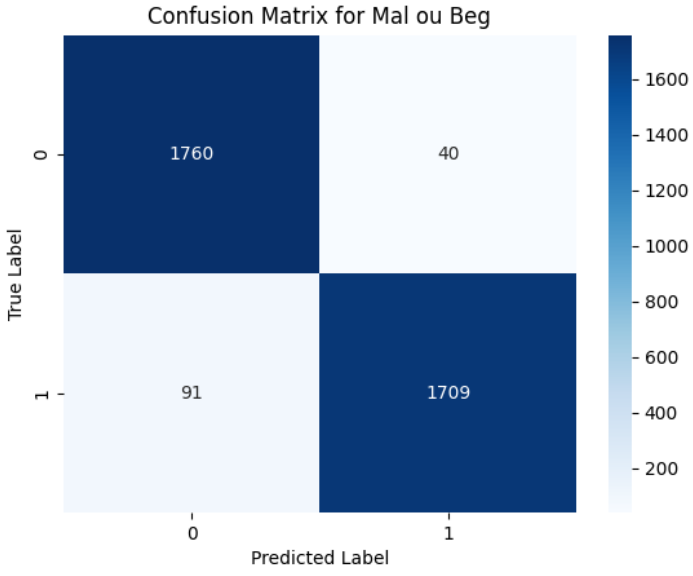


Figura 60 – Matriz confusão do modelo executado em InceptionV3 para a classificação binária do campo binário "mal_ou_beg". O Valor zero (0) corresponde a caso benigno e um (1) a caso maligno

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Verdadeiros Positivos (VP):** 1709 casos foram corretamente identificados como benignos.
- **Falsos Negativos (FN):** 91 casos foram erradamente classificados como benignos, quando na verdade eram malignos.

- **Verdadeiros Negativos (VN):** 1760 casos foram corretamente identificados como benignos.
- **Falsos Positivos (FP):** 40 casos foram erradamente classificados como malignos, quando na verdade eram benignos.

Podemos reparar que, apesar de na grande maioria das vezes o modelo prever corretamente as classes de uma imagem, este confunde numa quantidade considerável de casos que são malignos com os casos benignos, do que propriamente casos benignos com malignos que são erradamente identificados como tal.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 61.

Maligno ou Benigno Metrics:

	precision	recall	f1-score	support
Benigno	0.95	0.98	0.96	1800
Maligno	0.98	0.95	0.96	1800
accuracy			0.96	3600
macro avg	0.96	0.96	0.96	3600
weighted avg	0.96	0.96	0.96	3600

Figura 61 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo binário "mal_ou_beg"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe Benigna (0):**
  - A **precisão**, que indica a proporção de identificações corretas de benignidade entre todas as identificações de benignidade feitas pelo modelo, foi de 95%;
  - O ***recall***, que mede a proporção de casos benignos corretamente identificados pelo modelo em relação a todos os casos reais de benignidade, foi de 98%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 96%.
- **Para a classe Maligna (1):**
  - A **precisão** foi de 98%;
  - O ***recall*** foi de 95%;
  - O ***f1-score*** foi de 96%.
- **Exatidão** geral do modelo foi de 96%.

Podemos reparar que o modelo teve um alto desempenho, mas que existe alguma diferença na capacidade de identificar corretamente casos benignos em comparação com os casos malignos. Para a classe benigna, a precisão de 95% sugere que, quando o modelo prevê um caso benigno, é correto em 95% dos casos, o que é um resultado bastante positivo. O *recall* de 0.98 para esta classe indica que o modelo foi capaz de identificar 98% de todas as lesões benignas verdadeiras. Isto resulta num *f1-score* de 0.96, que sugere um desempenho equilibrado e eficiente.

No que diz respeito à classe maligna, a precisão elevada de 98% indica que o modelo está certo grande parte das vezes que realiza a previsão de malignidade. O *recall* de 0.98 para esta classe indica que o modelo foi capaz de identificar 98% de todas as lesões malignas verdadeiras, o que se traduz num *f1-score* de 96%. Estes resultaram foram bastante positivos e demonstram um desempenho eficaz, e com bastante precisão.

A exatidão global é a métrica que valida todas as premissas anteriores, pois o seu valor de 96% é elevado e garante assim que o modelo, na generalidade, acerte no seu diagnóstico de malignidade ou benignidade 96% das vezes.

A segunda matriz conseguida foi a de categoria "benigno" que é uma categoria ternária, que ao invés de conter dois valores possíveis, contém três, onde zero (0) representa os casos que não são benignos, um (1) os casos de Fibroadenoma e dois (2) como Doença Fibrocística (Figura 62).

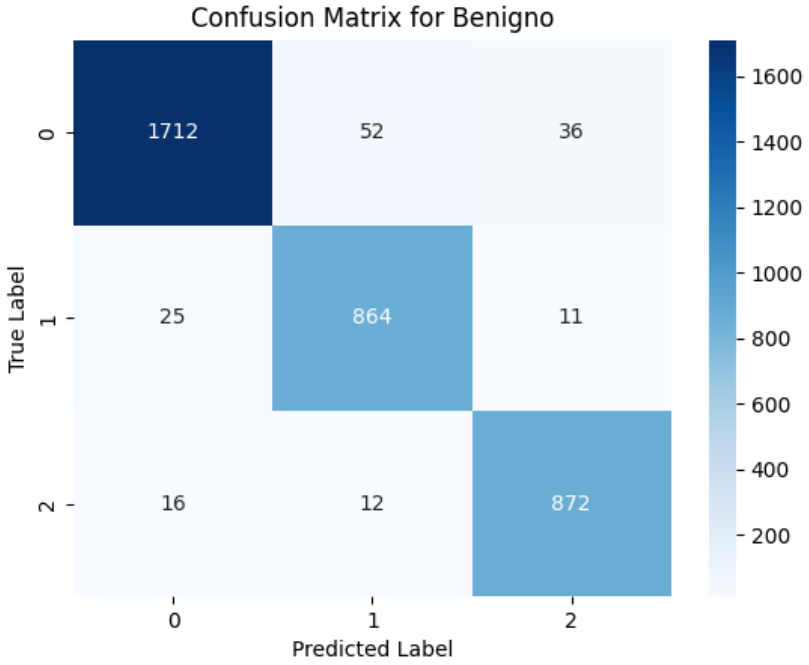


Figura 62 – Matriz confusão do modelo executado em InceptionV3 para a classificação ternária do campo "benigno". O Valor zero (0) representa os casos que não são benignos, um (1) os casos de Fibroadenoma e dois (2) como Doença Fibrocística

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Casos Corretos de "Não benigno":** 1712 casos foram corretamente identificados como não benignos.
- **Casos Corretos de Fibroadenoma:** 864 casos foram corretamente identificados como Fibroadenoma.
- **Casos Corretos de Doença Fibrocística:** 872 casos foram corretamente identificados como Doença Fibrocística.
- **Casos Incorretos de "Não benigno":** 41 casos foram incorretamente identificados como não benignos.
- **Casos Incorretos de Fibroadenoma:** 64 casos foram incorretamente identificados como Fibroadenoma.
- **Casos Incorretos de Doença Fibrocística:** 47 casos foram incorretamente identificados como Doença Fibrocística.

Desta matriz podemos reparar que o modelo acerta grande parte das vezes no diagnóstico das imagens na categoria de classificação benigna. Tal como aconteceu no SqueezeNet, podemos reparar que o modelo identifica mais vezes casos incorretos de Fibroadenoma, com maior quantidade se a imagem não for benigna. Isto poderá indicar que estes casos em específicos possam ter semelhanças, o que pode confundir o modelo de classificação.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 63.

Benigno Metrics:

	precision	recall	f1-score	support
Categoria 0	0.98	0.95	0.96	1800
Categoria 1	0.93	0.96	0.95	900
Categoria 2	0.95	0.97	0.96	900
accuracy			0.96	3600
macro avg	0.95	0.96	0.96	3600
weighted avg	0.96	0.96	0.96	3600

Figura 63 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo ternário "benigno"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe "Não Benigno" (0):**
  - A **precisão**, que indica a proporção de identificações corretas casos não benignos entre todas as identificações de benignidade feitas pelo modelo, foi de 98%;
  - O ***recall***, que mede a proporção de casos não benignos corretamente identificados pelo modelo em relação a todos os casos reais não benignos, foi de 95%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 96%.
- **Para a classe Fibroadenoma (1):**
  - A **precisão** foi de 93%;
  - O ***recall*** foi de 96%;
  - O ***f1-score*** foi de 95%.
- **Para a classe Doença Fibrocísitica (1):**
  - A **precisão** foi de 95%;
  - O ***recall*** foi de 97%;
  - O ***f1-score*** foi de 96%.
- **Exatidão** geral do modelo foi de 96%.

Pelos altos valores destas métricas, podemos reparar que na generalidade o modelo é eficaz e preciso a prever a classificação de imagens quanto à sua benignidade. Existe uma discrepância mais pequena em Fibroadenoma que confirma a suspeita levantada na matriz, em que concluímos que o modelo errava mais vezes em casos de Fibroadenoma.

Sendo um classificador mais complexo por ter três valores possíveis, podemos dizer que a exatidão global desta fase de classificação é muito boa, pelo que acerta no diagnóstico 96% das vezes.

A última matriz conseguida foi a de categoria "carcinoma" que é uma categoria ternária, que ao invés de conter dois valores possíveis, contém três, onde zero (0) representa os casos que não são malignos, um (1) os casos de Carcinoma Ductal e dois (2) como Carcinoma Lobular (Figura 64).

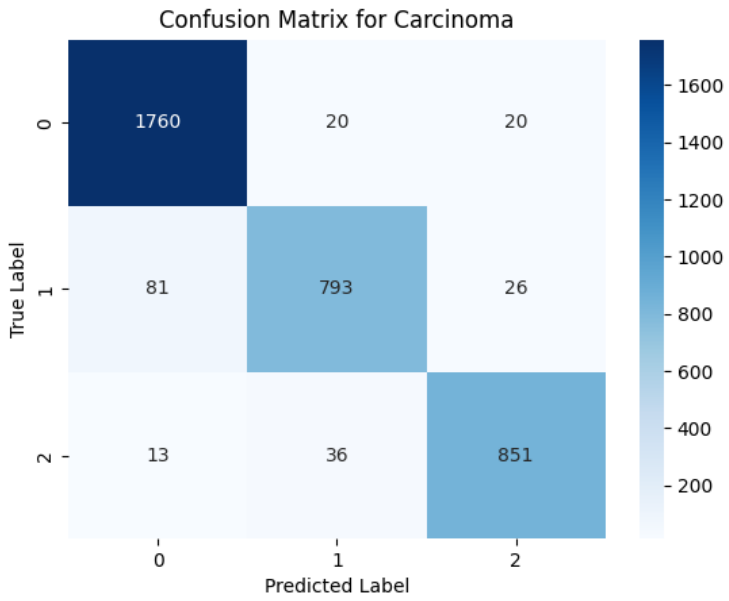


Figura 64 – Matriz confusão do modelo executado em InceptionV3 para a classificação ternária do campo "carcinoma". O Valor zero (0) representa os casos que não são malignos, um (1) os casos de Carcinoma Ductal e dois (2) como Carcinoma Lobular

Os resultados da matriz confusão podem ser interpretados da seguinte maneira:

- **Casos Corretos de "Não maligno":** 1760 casos foram corretamente identificados como não malignos.
- **Casos Corretos de Carcinoma Ductal:** 793 casos foram corretamente identificados como Carcinoma Ductal.
- **Casos Corretos de Carcinoma Lobular:** 851 casos foram corretamente identificados como Carcinoma Lobular.
- **Casos Incorretos de "Não maligno":** 94 casos foram incorretamente identificados como não malignos.
- **Casos Incorretos de Carcinoma Ductal:** 56 casos foram incorretamente identificados como Carcinoma Ductal.
- **Casos Incorretos de Carcinoma Lobular:** 46 casos foram incorretamente identificados como Doença Carcinoma Lobular.

Desta matriz podemos reparar que o modelo acerta grande parte das vezes no diagnóstico das imagens na categoria de classificação benigna. Da mesma forma que aconteceu no modelo SqueezeNet, podemos também reparar que o modelo identifica mais vezes casos incorretos de não maligno, com maior quantidade se a imagem for de casos de Carcinoma Ductal. Isto poderá indicar que estes casos em específicos possam ter semelhanças, o que pode confundir o modelo de classificação.

O modelo transforma estes resultados em métricas que são representados pelo terminal de execução como mostra a Figura 65.

Carcinoma Metrics:

	precision	recall	f1-score	support
Categoria 0	0.95	0.98	0.96	1800
Categoria 1	0.93	0.88	0.91	900
Categoria 2	0.95	0.95	0.95	900
accuracy			0.95	3600
macro avg	0.94	0.93	0.94	3600
weighted avg	0.95	0.95	0.95	3600

Figura 65 – Métricas de validação na execução do modelo InceptionV3 para a classificação do campo ternário "benigno"

A partir desta imagem podemos obter as seguintes métricas:

- **Para a classe "Não Maligno" (0):**
  - A **precisão**, que indica a proporção de identificações corretas de não malignos entre todas as identificações de benignidade feitas pelo modelo, foi de 95%;
  - O ***recall***, que mede a proporção de casos não malignos corretamente identificados pelo modelo em relação a todos os casos reais não malignos, foi de 98%;
  - O ***f1-score***, que relaciona *precision* e *recall*, foi de 96%.
- **Para a classe Carcinoma Ductal (1):**
  - A **precisão** foi de 93%;
  - O ***recall*** foi de 88%;
  - O ***f1-score*** foi de 91%.
- **Para a classe Carcinoma Lobular (1):**
  - A **precisão** foi de 96%;
  - O ***recall*** foi de 91%;
  - O ***f1-score*** foi de 95%.
- **Exatidão** geral do modelo foi de 95%.

Através destas métricas, mais uma vez o modelo mostrou ser capaz de eficientemente e com elevada precisão classificar as imagens quanto à sua malignidade. Porém, através do *recall* da classe de Carcinoma Ductal, podemos reparar que apenas 88% dos casos desta patologia foram identificados como tal, o que juntando à conclusão que foi obtida da matriz anterior, podemos concluir que a maior parte dos Carcinomas Ductais que não foram corretamente classificados, foram classificados como não sendo malignos.

Embora ligeiramente mais baixo que a classificação de benignidade das imagens, a exatidão do modelo voltou a mostrar um elevado valor de 95%, o que é bastante positivo no contexto deste cenário.

Ao analisar a curva ROC para a classificação de maligno e benigno (Figura 66), podemos concluir que esta apresenta uma um bom equilíbrio na distinção entre maligno e benigno. A AUC desta curva apresenta um valor de 0,99, o que é próximo do ideal, como podemos ver pela forma da curva.

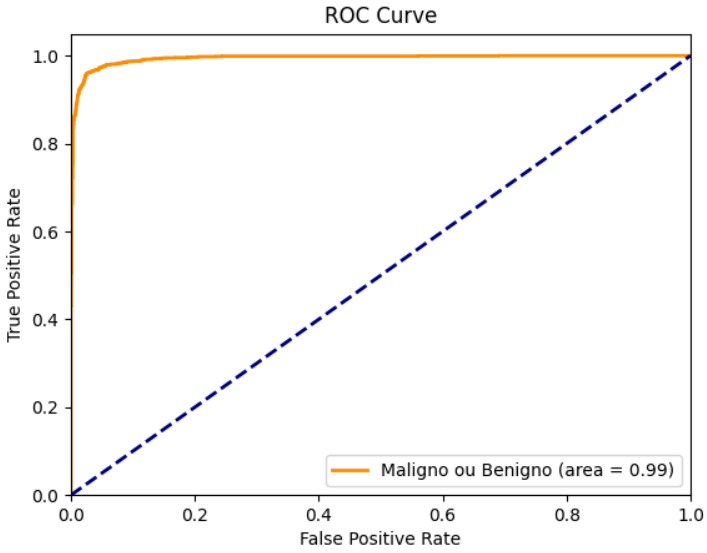


Figura 66 – Curva ROC e respetivo valor AUC do modelo InceptionV3 na classificação binária do campo "mal_ou_beg"

O mesmo acontece para as curvas ROC multiclasse que estão representadas na Figura 68 e na Figura 67. Podemos reparar que a área de todas as curvas são cerca de 0,99, o que demonstra uma boa capacidade de distinção de cada classe em relação às restantes classes.

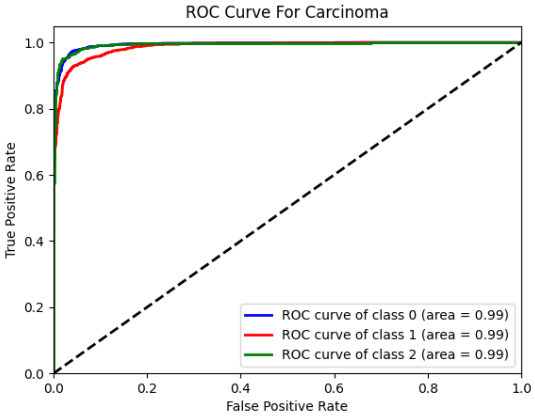


Figura 68 – Curvas ROC e respetivos valores AUC do modelo InceptionV3 na classificação ternária do campo "carcinoma"

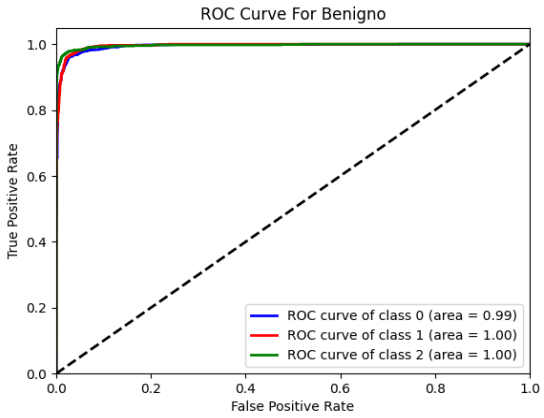


Figura 67 – Curvas ROC e respetivos valores AUC do modelo InceptionV3 na classificação ternária do campo "benigno"

No final, foi obtido os gráficos de exatidão ao longo das 15 *epochs* ocorridas para cada uma. Ao analisar esse gráfico apresentado na Figura 69, podemos reparar que todas as curvas de exatidão de treino crescem nos primeiros *epochs*, mas que rapidamente estabilizam acima dos 99%. O mesmo não acontece às curvas de exatidão de validação, que após acompanharem o crescimento das curvas de treino nos primeiros *epochs* deixam de crescer ou estabilizar, com alguma volatilidade e algum decréscimo na exatidão. Isto demonstra de maneira clara que modelo rapidamente entra em *overfitting*.

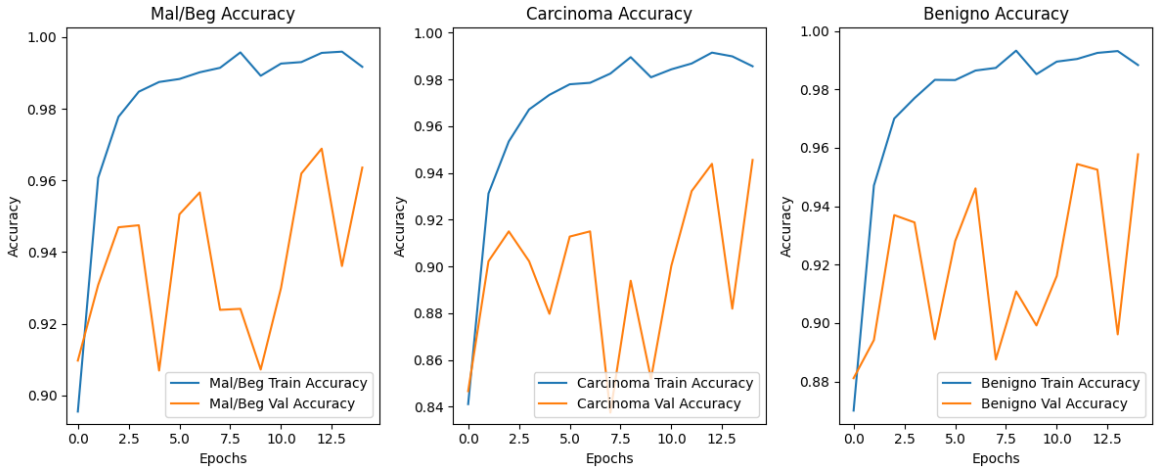


Figura 69 – Curvas de exatidão do modelo InceptionV3 no cenário de classificação multiclasse

✓ COMPARAÇÃO/DISCUSSÃO

No final da análise destes dois modelos para o cenário multiclasse, podemos concluir que apesar do modelo InceptionV3 ter exatidões mais voláteis ao longo das *epochs*, o que implica que o modelo possa estar em *overfitting*, teve exatidões mais elevadas que as obtidas no SqueezeNet. O facto de o InceptionV3 ter métricas melhores que o SqueezeNet faz com que este este modelo tenha sido o mais indicado para este cenário, mesmo com a volatilidade demonstrada.

Porém, podemos reparar que o modelo SqueezeNet poderia beneficiar do aumento de *epochs* no treino do modelo, já que as curvas de exatidão dos mesmos mostram ainda capacidade crescente, tanto no treino como na validação.

Com estes resultados, podemos dizer que provavelmente o modelo InceptionV3 é um bom modelo para cenários de decisão mais complexas como é o caso deste cenário multiclasse, onde o modelo teve de realizar várias previsões em diferentes categorias usando os mesmos dados de treino.

4.3. Discussão de resultados

Depois de executar e analisar os modelos SqueezeNet e InceptionV3 para todos os cenários, é possível tirar algumas conclusões através das métricas que foram analisadas. A métrica considerada mais importante para estes tipos de cenários é a exatidão (*accuracy*) que é possível retirar da validação do modelo. Na Tabela 5 podemos observar todos os valores de *accuracy* obtidos em cada modelo e em cada arquitetura.

Tabela 5 – Tabela de comparação de exatidão nos modelos para todos os cenários

Cenários (accuracy)\Rede Neuronal	Cenário 1 (Maligno/Benigno)	Cenário 2 (Benigno)	Cenário 3 (Carcinoma)	Cenário 4 (Multiclasse)
SqueezeNet	92%	95%	92%	Maligno/Benigno: 92% Benigno: 91% Carcinoma: 89%
InceptionV3	88%	98%	90%	Maligno/Benigno: 96% Benigno: 96% Carcinoma: 95%

Podemos concluir, a partir da tabela, que as percentagens mais altas pertencem ao modelo InceptionV3. Porém, tal como analisado noutras métricas, foi possível observar volatilidade nos resultados deste modelo, o que pode ser indício de *overfitting* na aprendizagem. Essa instabilidade é notada nos valores de exatidão desta arquitetura ao longo de todos os cenários.

Também podemos reparar que o modelo SqueezeNet teve menos eficácia no cenário multiclasse quando comparado com os restantes. Como o modelo é mais simplificado e eficiente em termos de tempo, este será o melhor modelo para treino de cenários simples como os três primeiros, onde o modelo apenas tinha de classificar categorias binárias.

No caso do InceptionV3, podemos observar que apesar de se mostrar instável nestas condições, aparenta ser mais eficiente em cenários de classificação mais complexos, como é o caso do último cenário abordado neste estudo. Para as restantes três categorias, este modelo foi eficaz e mostrou uma boa precisão e exatidão. Isso deve-se ao facto de a arquitetura do InceptionV3 ser mais complexa, o que faz com que a aprendizagem do modelo seja mais rápida e que consiga aproveitar os dados de maneira a classificar melhor as categorias de problemas mais complexos.

No caso do treino dos dois modelos, apesar do InceptionV3 demorar mais tempo a treinar, este necessita de menos *epochs*, o que foi possível de observar pela rápida estabilização das curvas de exatidão de treino deste modelo, o que por vezes levou a uma instabilidade e *overfitting* na curva de validação. No caso do SqueezeNet, podemos reparar pelas curvas que os modelos poderiam obter uma maior percentagem nas métricas de precisão e exatidão se o treino do modelo fosse maior, mas não podemos garantir tal situação uma vez que este também poderia entrar em *overfitting*, piorando assim o treino do modelo.

Ambos os modelos são promissores para este tipo de cenários de diagnóstico, porém, como é necessário um grau elevado de exatidão e precisão, o modelo InceptionV3 consegue colmatar a volatilidade do modelo, aumentando ainda mais a exatidão deste. Podemos então considerar o modelo InceptionV3 como o melhor modelo tendo em conta os valores de exatidão que conseguiu obter. Para tornar o modelo ainda mais consistente, seria necessário alterar pequenos parâmetros deste para conseguir valores ainda mais elevados e mais consistentes e evitar situações de *overfitting*, tal como aconteceu nos cenários estudados.

5. CONCLUSÃO

Este capítulo será uma conclusão da criação de uma base de dados de imagens histológicas anotadas e do desenvolvimento de um modelo de classificação de algumas patologias mamárias, onde será feita uma ponte entre a tecnologias da IA, *machine learning* e *deep learning*, e a área da saúde, mais concretamente na área da histologia e da área da prevenção e diagnóstico do cancro da mama.

5.1. Conclusões finais

Este projeto de mestrado mostra que é possível haver avanços significativos e com qualidade na aplicação da IA para o diagnóstico e análise de imagens histológicas no contexto de várias patologias mamárias. Foram cumpridos todos os objetivos estabelecidos, desde a recolha, preparação e organização de dados histológicos até à construção e avaliação de um modelo de classificação eficaz.

As redes neuronais, especificamente as arquiteturas SqueezeNet e InceptionV3, demonstraram uma elevada precisão na classificação de tecidos mamários, com destaque para a InceptionV3 que foi capaz de alcançar valores de 98% de *accuracy* em cenários específicos. Esta eficácia reflete não só a relevância das escolhas metodológicas, mas também a capacidade de lidar com a variabilidade e desafios inerentes aos dados histológicos. Esta constatação responde à primeira e segunda questões de investigação, uma vez que dados os valores atingidos, concluímos que é possível distinguir imagens de patologias mamárias com um elevado grau de confiança, e ainda que é possível observar que existe robustez nestas tecnologias face à diversidade dos dados clínicos.

Relativamente à terceira questão de investigação, a comparação entre as métricas obtidas com dados padronizados e não padronizados, revelou que as redes neuronais, apesar de apresentarem algumas variações nas métricas, mantêm-se eficientes e confiáveis mesmo sob condições de dados não padronizados. Esta conclusão é crucial, pois salienta a aplicabilidade prática destas tecnologias em ambientes clínicos reais, onde a padronização dos dados nem sempre é possível.

Assim, este projeto contribuiu para uma melhor compreensão das implicações da variabilidade dos dados e dos artefactos histológicos na performance das redes neuronais, concluindo que estas arquiteturas são capazes de realizarem uma boa aprendizagem, mesmo perante imagens com alguma distorção devido a fatores externos, respondendo assim de à quarta questão de investigação.

Os resultados obtidos comprovam a viabilidade da utilização de sistemas de IA no suporte à decisão clínica, sendo estes capazes de melhorar significativamente a eficácia e a precisão do diagnóstico e, consequentemente, a vida e a saúde dos pacientes, sendo esta conclusão uma resposta à última questão de investigação.

Em suma, este trabalho não só atingiu todos os objetivos inicialmente propostos com sucesso, mas também destacou áreas de investigação futura na interligação entre a inteligência artificial e a patologia, sendo indicativo de caminhos promissores para avanços na integração destes sistemas na medicina de diagnóstico.

5.2. Limitações e investigação futura

Durante a realização deste estudo, houve algumas limitações que influenciaram o seu desempenho e desenvolvimento. Nas questões laboratoriais, a não uniformização da coloração, devido a colorações com corantes de várias concentrações e aos vários artefactos que podem surgir, são situações que podem ter limitado e influenciado o treino dos modelos ao induzir informação muito diversa na sua

aprendizagem. A falta de mais recursos computacionais foi mais um dos fatores limitadores, uma vez que limitou a quantidade de dados que foram utilizados para treinar os modelos bem como adicionar mais complexidade aos cenários também não foi possível devido a esta situação.

Como passos futuros na melhoria deste estudo, passa por adicionar um maior número de casos clínicos, incluindo um maior número de patologias mamárias, aumentando assim a relevância clínica do treino destes modelos. Pode também ser testado alterações no processo de digitalização, bem como a influência da ampliação da digitalização na eficiência e exatidão dos modelos treinados.

Como investigação futura e que abrange o aumento de casos na base de dados do sistema, é a importância de treinar estes modelos em plataformas com capacidades computacionais elevadas de maneira a melhorar o estudo de classificação de patologias mamárias. Finalmente, como produto final poderia ser utilizar estes modelos treinados para criar um software que fizesse a classificação de patologias mamárias dependendo das imagens que eram fornecidas, de maneira a dar apoio à decisão clínica ao patologista e assim diminuir erros humanos e tornar mais eficiente o diagnóstico destas patologias.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] "Cancro da Mama : Liga Portuguesa Contra o Cancro." Accessed: Nov. 12, 2023. [Online]. Available: https://www.ligacontracancro.pt/cancro-da-mama/

[2] D. S. C. Vieira, R. M. Dufloth, F. C. L. Schmitt, and L. C. Zeferino, "Carcinoma de mama: novos conceitos na classificação," *Revista Brasileira de Ginecologia e Obstetrícia*, vol. 30, no. 1, pp. 42–47, Jan. 2008, doi: 10.1590/S0100-72032008000100008.

[3] "Câncer de mama in situ - Sintomas, diagnóstico e tratamento | BMJ Best Practice." Accessed: Nov. 12, 2023. [Online]. Available: https://bestpractice.bmj.com/topics/pt-br/717

[4] P. Thakur, K. Baraskar, and V. K. Shrivastava, "Histopathological Characteristics: Breast Cancer Subtypes Depending on Receptor Status, Clinical and Pathological Staging of Breast Cancer," *Breast Cancer: From Bench to Personalized Medicine*, pp. 31–46, Jan. 2022, doi: 10.1007/978-981-19-0197-3_2/COVER.

[5] H. Tsuda *et al.*, "Histological classification of breast tumors in the General Rules for Clinical and Pathological Recording of Breast Cancer (18th edition)," *Breast Cancer*, vol. 27, no. 3, pp. 309–321, May 2020, doi: 10.1007/S12282-020-01074-3/METRICS.

[6] K. P. Serra *et al.*, "[The new classification of breast cancers: finding the luminal A]," *Rev Bras Ginecol Obstet*, vol. 36, no. 12, pp. 575–580, 2014, doi: 10.1590/SO100-720320140005158.

[7] D. S. C. Vieira, R. M. Dufloth, F. C. L. Schmitt, and L. C. Zeferino, "Carcinoma de mama: novos conceitos na classificação," *Revista Brasileira de Ginecologia e Obstetrícia*, vol. 30, no. 1, pp. 42–47, Jan. 2008, doi: 10.1590/S0100-72032008000100008.

[8] "Patologias da Mama: tipos, sintomas e quem está em risco – Augusta Matos Clínicas." Accessed: Nov. 12, 2023. [Online]. Available: https://www.augustamatosclinicas.pt/patologias-da-mama-tipos-sintomas-e-quem-esta-em-risco/

[9] "Fibroadenoma da Mama - Sintomas e tratamento | MD.Saúde." Accessed: Nov. 12, 2023. [Online]. Available: https://www.mdsaude.com/ginecologia/fibroadenoma-da-mama/

[10] D. A. Naik, R. M. Mohana, G. Ramu, Y. S. Lalitha, M. SureshKumar, and K. V. Raghavender, "Analyzing histopathological images by using machine learning techniques," *Applied Nanoscience (Switzerland)*, vol. 13, no. 3, pp. 2507–2513, Mar. 2023, doi: 10.1007/S13204-021-02217-4/METRICS.

[11] "Artificial Intelligence (AI): What Is AI and How Does It Work? | Built In." Accessed: Nov. 12, 2023. [Online]. Available: https://builtin.com/artificial-intelligence

[12] "Machine Learning vs. AI: Differences, Uses, and Benefits | Coursera." Accessed: Nov. 12, 2023. [Online]. Available: https://www.coursera.org/articles/machine-learning-vs-ai

[13] J. W. Prichard, C. R. Mehr, D. G. Hicks, and E. Hammond, "Predictive Biomarkers in Breast Cancer: ER, PR, and HER-2/NEU," *Handbook of Practical Immunohistochemistry: Frequently Asked Questions*, pp. 293–312, Jan. 2022, doi: 10.1007/978-3-030-83328-2_15/COVER.

[14] H. Zhao and Y. Gong, "The Prognosis of Single Hormone Receptor-Positive Breast Cancer Stratified by HER2 Status," *Front Oncol*, vol. 11, p. 643956, May 2021, doi: 10.3389/FONC.2021.643956/BIBTEX.

[15] M. Guray and A. A. Sahin, "Benign Breast Diseases: Classification, Diagnosis, and Management," *Oncologist*, vol. 11, no. 5, pp. 435–449, May 2006, doi: 10.1634/THEONCOLOGIST.11-5-435.

[16] "Understanding Your Pathology Report: Benign Breast Conditions | American Cancer Society." Accessed: Nov. 12, 2023. [Online]. Available:

https://www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/breast-pathology/benign-breast-conditions-pathology.html

[17] L. Luo *et al.*, "Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions," Apr. 2023, Accessed: Nov. 09, 2023. [Online]. Available: http://arxiv.org/abs/2304.06662

[18] "What Is Machine Learning? Definition, Types, and Examples | Coursera." Accessed: Nov. 12, 2023. [Online]. Available: https://www.coursera.org/articles/what-is-machine-learning

[19] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, Sep. 2021, doi: 10.1007/S12525-021-00475-2/TABLES/2.

[20] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 6, pp. 1–20, Nov. 2021, doi: 10.1007/S42979-021-00815-1/FIGURES/6.

[21] "Train Deep Learning Network with Nested Layers - MATLAB & Simulink." Accessed: Nov. 12, 2023. [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/train-deep-learning-network-with-nested-layers.html

[22] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data 2021 8:1*, vol. 8, no. 1, pp. 1–74, Mar. 2021, doi: 10.1186/S40537-021-00444-8.

[23] M. Z. Alom *et al.*, "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics (Basel)*, vol. 8, no. 3, Mar. 2019, doi: 10.3390/ELECTRONICS8030292.

[24] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med Image Anal*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/J.MEDIA.2017.07.005.

[25] J. Gu *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognit*, vol. 77, pp. 354–377, May 2015, doi: 10.1016/J.PATCOG.2017.10.013.

[26] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, Jun. 2019, doi: 10.1007/S13748-019-00203-0.

[27] D. Shen, G. Wu, and H. Il Suk, "Deep Learning in Medical Image Analysis.," *Annu Rev Biomed Eng*, vol. 19, pp. 221–248, Jun. 2017, doi: 10.1146/ANNUREV-BIOENG-071516-044442.

[28] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *J Digit Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi: 10.1007/S10278-019-00227-X.

[29] K. Lee *et al.*, "Deep Learning of Histopathology Images at the Single Cell Level," *Front Artif Intell*, vol. 4, p. 754641, Sep. 2021, doi: 10.3389/FRAI.2021.754641/BIBTEX.

[30] B. Ghojogh and A. Ghodsi ALIGHODSI, "Recurrent Neural Networks and Long Short-Term Memory Networks: Tutorial and Survey," Apr. 2023, Accessed: Nov. 12, 2023. [Online]. Available: https://arxiv.org/abs/2304.11461v1

[31] C. Yang *et al.*, "Comparison of Overall Survival Between Invasive Lobular Breast Carcinoma and Invasive Ductal Breast Carcinoma: A Propensity Score Matching Study Based on SEER Database," *Front Oncol*, vol. 10, p. 590643, Dec. 2020, doi: 10.3389/FONC.2020.590643/BIBTEX.

[32] Z. Chen *et al.*, "Invasive lobular carcinoma of the breast: A special histological type compared with invasive ductal carcinoma," *PLoS One*, vol. 12, no. 9, p. e0182397, Sep. 2017, doi: 10.1371/JOURNAL.PONE.0182397.

[33] "An Intro to H&E Staining: Protocol, Best Practices, Steps & More." Accessed: Nov. 12, 2023. [Online]. Available: https://www.leicabiosystems.com/pt-pt/knowledge-pathway/he-staining-overview-a-guide-to-best-practices/

[34] C. Munien and S. Viriri, "Classification of Hematoxylin and Eosin-Stained Breast Cancer Histology Microscopy Images Using Transfer Learning with EfficientNets," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/5580914.

[35] M. S. Hosseini *et al.*, "Atlas of digital pathology: A generalized hierarchical histological tissue type-annotated database for deep learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 11739–11748, Jun. 2019, doi: 10.1109/CVPR.2019.01202.

[36] "Benefits of Digital Pathology." Accessed: Nov. 12, 2023. [Online]. Available: https://www.news-medical.net/life-sciences/Benefits-of-Digital-Pathology.aspx

[37] A. V. Parwani, "Next generation diagnostic pathology: Use of digital pathology and artificial intelligence tools to augment a pathological diagnosis," *Diagn Pathol*, vol. 14, no. 1, pp. 1–3, Dec. 2019, doi: 10.1186/S13000-019-0921-2/METRICS.

[38] J. Hung *et al.*, "Keras R-CNN: Library for cell detection in biological images using deep neural networks," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–7, Jul. 2020, doi: 10.1186/S12859-020-03635-X/FIGURES/1.

[39] "Steps to Tissue Processing for Histopathology." Accessed: Nov. 12, 2023. [Online]. Available: https://www.leicabiosystems.com/pt-pt/knowledge-pathway/an-introduction-to-specimen-processing/

[40] I. K. Park, I. D. Yun, and S. U. Lee, "A color normalization algorithm for image indexing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 1351, pp. 96–103, 1997, doi: 10.1007/3-540-63930-6_109/COVER.

[41] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/S40537-019-0197-0/FIGURES/33.

[42] "Top Data Augmentation Techniques: Ultimate Guide for 2023." Accessed: Nov. 12, 2023. [Online]. Available: https://research.aimultiple.com/data-augmentation-techniques/

[43] "Artificial Neural Network - Basic Concepts." Accessed: Nov. 12, 2023. [Online]. Available: https://www.tutorialspoint.com/artificial_neural_network/artificial_neural_network_basic_concepts.htm

[44] G. Auda and M. Kamel, "Modular neural networks: a survey.," *Int J Neural Syst*, vol. 9, no. 2, pp. 129–151, 1999, doi: 10.1142/S0129065799000125.

[45] "Convolutional Neural Networks (LeNet) — DeepLearning 0.1 documentation." Accessed: Nov. 12, 2023. [Online]. Available: https://web.archive.org/web/20171228091645/http://deeplearning.net/tutorial/lenet.html

[46] "Cross-Validation: A Technique To Prevent Overfitting In Neural Networks – Surfactants." Accessed: Nov. 12, 2023. [Online]. Available: https://www.surfactants.net/cross-validation-a-technique-to-prevent-overfitting-in-neural-networks/

[47] A. M. Alqudah, H. Alquraan, I. A. Qasmieh, A. Alqudah, and W. Al-Sharu, "Brain Tumor Classification Using Deep Learning Technique -- A Comparison between Cropped, Uncropped, and Segmented Lesion Images with Different Sizes," Jan. 2020, doi: 10.30534/ijatcse/2019/155862019.

[48] W. Setiawan, A. Ghofur, F. Hastarita Rachman, and R. Rulaningtyas, "Deep Convolutional Neural Network AlexNet and Squeezenet for Maize Leaf Diseases Image Classification," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Nov. 2021, doi: 10.22219/KINETIK.V6I4.1335.

[49] A. Maiti, A. Abarda, M. Hanini, and A. Oussous, "An Optimal Model Combining SqueezeNet and Machine Learning Methods for Lung Disease Diagnosis," *Current Medical Imaging Formerly Current Medical Imaging Reviews*, vol. 20, Oct. 2023, doi: 10.2174/0115734056258742230920062315.

[50] "SqueezeNet-like architecture. | Download Scientific Diagram." Accessed: Nov. 12, 2023. [Online]. Available: https://www.researchgate.net/figure/SqueezeNet-like-architecture_fig8_320723863

[51] P. Bedi, N. Ningshen, S. Rani, and P. Gole, "Explainable Predictions for Brain Tumor Diagnosis Using InceptionV3 CNN Architecture," pp. 125–134, 2024, doi: 10.1007/978-981-99-4071-4_11.

[52] D. Tiwari, M. Dixit, and K. Gupta, "PERFORMANCE COMPARISON OF THE RESNET50 AND INCEPTIONV3 DEEP TRANSFER LEARNING MODELS OVER THE BREAST CANCER THERMOS GRAM DATASET," *BSSS Journal of Computer*, vol. 13, no. 1, pp. 1–9, Jun. 2022, doi: 10.51767/JC1301.

[53] "The architecture of Inception-V3 model. | Download Scientific Diagram." Accessed: Nov. 12, 2023. [Online]. Available: https://www.researchgate.net/figure/The-architecture-of-Inception-V3-model_fig5_349717475

[54] B. Swathi, K. S. Kannan, S. Sreenivasa Chakravarthi, G. Ruthvik, J. Avanija, and C. Chandra Mohan Reddy, "Skin Cancer Detection using VGG16, InceptionV3 and ResUNet," *2023 4th International Conference on Electronics and Sustainable Communication Systems, ICESC 2023 - Proceedings*, pp. 812–818, 2023, doi: 10.1109/ICESC57686.2023.10193609.

[55] "Difference between AlexNet, VGGNet, ResNet, and Inception | by Aqeel Anwar | Towards Data Science." Accessed: Nov. 12, 2023. [Online]. Available: https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96

[56] Y. N. Fu'Adah, I. Wijayanto, N. K. C. Pratiwi, F. F. Taliningsih, S. Rizal, and M. A. Pramudito, "Automated Classification of Alzheimer's Disease Based on MRI Image Processing using Convolutional Neural Network (CNN) with AlexNet Architecture," *J Phys Conf Ser*, vol. 1844, no. 1, p. 012020, Mar. 2021, doi: 10.1088/1742-6596/1844/1/012020.

[57] "AlexNet: The Architecture that Challenged CNNs | by Jerry Wei | Towards Data Science." Accessed: Nov. 12, 2023. [Online]. Available: https://towardsdatascience.com/alexnet-the-architecture-that-challenged-cnns-e406d5297951

[58] "Architecture comparison of AlexNet, VGGNet, ResNet, Inception, DenseNet | by Khush Patel | Towards Data Science." Accessed: Nov. 12, 2023. [Online]. Available: https://towardsdatascience.com/architecture-comparison-of-alexnet-vggnet-resnet-inception-densenet-beb8b116866d

[59] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, Aug. 2016, doi: 10.1109/CVPR.2017.243.

[60] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *36th International Conference on Machine Learning, ICML 2019*, vol. 2019-June, pp. 10691–10700, May 2019, Accessed: Nov. 12, 2023. [Online]. Available: https://arxiv.org/abs/1905.11946v5

[61] A. Howard *et al.*, "Searching for MobileNetV3," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2019-October, pp. 1314–1324, May 2019, doi: 10.1109/ICCV.2019.00140.

[62] K. Roy, D. Banik, D. Bhattacharjee, and M. Nasipuri, "Patch-based system for Classification of Breast Histology images using deep learning," *Comput Med Imaging Graph*, vol. 71, pp. 90–103, Jan. 2019, doi: 10.1016/J.COMPMEDIMAG.2018.11.003.

[63] "A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters, 125, 1–6 | 10.1016/j.patrec.2019.03.022." Accessed: Nov. 09, 2023. [Online]. Available: https://sci-hub.se/https://www.sciencedirect.com/science/article/abs/pii/S0167865519301059?via%3Dihub

[64] Y. Wang *et al.*, "Breast Cancer Image Classification via Multi-Network Features and Dual-Network Orthogonal Low-Rank Learning," *IEEE Access*, vol. 8, pp. 27779–27792, 2020, doi: 10.1109/ACCESS.2020.2964276.

[65] V. A. Kuiava, E. L. Kuiava, R. Rodriguez, A. E. Beck, J. P. M. Rodriguez, and E. O. Chielle, "Método de diagnóstico histopatológico de nódulos mamários por meio do algoritmo de aprendizagem profunda," *J Bras Patol Med Lab*, vol. 55, no. 6, pp. 620–632, Mar. 2020, doi: 10.5935/1676-2444.20190055.

[66] L. Liu, W. Feng, C. Chen, M. Liu, Y. Qu, and J. Yang, "Classification of breast cancer histology images using MSMV-PFENet," *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-22358-y.

[67] A. Polónia *et al.*, "Artificial Intelligence Improves the Accuracy in Histologic Classification of Breast Lesions," *Am J Clin Pathol*, vol. 155, no. 4, pp. 527–536, Apr. 2021, doi: 10.1093/AJCP/AQAA151.

[68] "Precision-Recall Curves. Sometimes a curve is worth a thousand… | by Doug Steen | Medium." Accessed: Nov. 13, 2023. [Online]. Available: https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248

[69] "Roc Curve. A Receiver Operating Characteristic… | by Sahil Tikkal | May, 2023 | Medium | Medium." Accessed: Nov. 13, 2023. [Online]. Available: https://medium.com/@tikkalsahil/roc-curve-c04101df8a19

[70] "A Deep Dive Into Learning Curves in Machine Learning | ml-articles – Weights & Biases." Accessed: Nov. 13, 2023. [Online]. Available: https://wandb.ai/mostafaibrahim17/ml-articles/reports/A-Deep-Dive-Into-Learning-Curves-in-Machine-Learning--Vmlldzo0NjA1ODY0