UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS SOBRAL

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO (PPGEEC)

ANDRESSA GOMES MOREIRA

ANÁLISE DO DESEMPENHO DE REDES NEURAIS PROFUNDAS NA CLASSIFICAÇÃO E SEGMENTAÇÃO DE TUMORES CEREBRAIS EM IMAGENS DE RESSONÂNCIA MAGNÉTICA

SOBRAL

2024

ANDRESSA GOMES MOREIRA

ANÁLISE DO DESEMPENHO DE REDES NEURAIS PROFUNDAS NA CLASSIFICAÇÃO E SEGMENTAÇÃO DE TUMORES CEREBRAIS EM IMAGENS DE RESSONÂNCIA MAGNÉTICA

Dissertação apresentada Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC) da Universidade Federal do Ceará, Campus de Sobral, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Orientador: Prof. Dr. Iális Cavalcante de Paula Júnior

Coorientador: Prof. Dr. Fischer Jônatas Ferreira

SOBRAL

2024

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

M836a Moreira, Andressa Gomes.
Análise do Desempenho de Redes Neurais Profundas na Classificação e Segmentação de Tumores Cerebrais em Imagens de Ressonância Magnética / Andressa Gomes Moreira. – 2024.
120 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Campus de Sobral, Programa de Pós-Graduação em Engenharia Elétrica e de Computação, Sobral, 2024.
Orientação: Prof. Dr. Iális Cavalcante de Paula Júnior.
Coorientação: Prof. Dr. Fischer Jônatas Ferreira.

1. Deep Learning. 2. Ressonância Magnética. 3. Tumor Cerebral. 4. Classificação. 5. Segmentação. I. Título.

CDD 621.3

ANDRESSA GOMES MOREIRA

ANÁLISE DO DESEMPENHO DE REDES NEURAIS PROFUNDAS NA CLASSIFICAÇÃO E SEGMENTAÇÃO DE TUMORES CEREBRAIS EM IMAGENS DE RESSONÂNCIA MAGNÉTICA

Dissertação apresentada Programa de Pós-Graduação em Engenharia Elétrica e de Computação (PPGEEC) da Universidade Federal do Ceará, Campus de Sobral, como requisito parcial à obtenção do título de mestre em Engenharia Elétrica e de Computação. Área de Concentração: Sistemas de Informação

Aprovada em: 11 de Outubro de 2024

BANCA EXAMINADORA

Prof. Dr. Iális Cavalcante de Paula Júnior (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Fischer Jônatas Ferreira (Coorientador)
Universidade Federal di Ceará (UFC)

Prof. Dr. Carlos Alexandre Rolim Fernandes
Universidade Federal do Ceará (UFC)

Prof. Dr. Romuere Rodrigues Veloso e Silva
Universidade Federal do Piauí (UFPI)

Dedico este trabalho às minhas avós, Zuleide e Gonçala (In Memoriam).

AGRADECIMENTOS

Agradeço a Deus e à Nossa Senhora pelo dom da vida e por me sustentarem até aqui.

Aos meus pais pelos cuidados e por acreditarem que daria tudo certo. Um agradecimento especial à minha mãe, Marinete, que em todos os momentos me deu força para continuar.

Ao Prof. Dr. Iális Cavalcante de Paula Júnior por me orientar desde a graduação. Agradeço os ensinamentos e pela paciência durante todos esses anos.

Ao Prof. Dr. Fischer Jônatas Ferreira. Agradeço todo o apoio e os ensinamentos durante o mestrado.

Ao meu namorado, Carlos Augusto, pela força e por sempre me lembrar que sou capaz.

À minha amiga Stefane, por estar comigo nos momentos bons e ruins nos últimos sete anos. Ao meu amigo, Lucas Gabriel, por partilhar a experiência do mestrado comigo.

Aos meus amigos, Bruno, Débora e Pedro. Agradeço pela amizade e por toda a ajuda durante o desenvolvimento deste trabalho, sem vocês não teria conseguido.

Aos meus familiares e todos os que contribuíram de alguma forma para o desenvolvimento deste projeto.

Aos professores da Universidade Federal do Ceará (UFC) pela dedicação e pelo aprendizado.

À Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Funcap).

"Direi do Senhor: Ele é o meu Deus, o meu refúgio, a minha fortaleza, e nele confiarei."

(Salmos 91:2)

RESUMO

O desenvolvimento de tumores cerebrais malignos caracteriza o câncer cerebral. De acordo com o Instituto Nacional de Câncer (INCA), no Brasil são registrados cerca de 11 mil novos casos todos os anos, com um índice de mortalidade em aproximadamente 84% dos casos. Nesse contexto, a detecção precoce de tumores cerebrais pode elevar consideravelmente a taxa de sobrevivência dos pacientes. Os avanços em Inteligência Artificial (IA) têm aprimorado a análise de imagens médicas. Entretanto, a classificação e a segmentação de tumores cerebrais, por meio de imagens de ressonância magnética (RM), ainda são tarefas desafiadoras por diversos fatores, como, alterações de contraste, tamanho, formato, posição e variação da região tumoral, dependendo do paciente. Portanto, o principal objetivo deste trabalho é desenvolver um fluxo de processamento do treinamento e a análise do desempenho de arquiteturas de redes neurais profundas na classificação de tumores cerebrais em Meningioma, Glioma, Hipofisário, além de casos sem tumor e a detecção e segmentação da região tumoral. Para isso, foram utilizados três conjuntos de dados disponíveis publicamente da literatura de imagens de ressonância magnética. Na etapa de classificação, realizou-se o treinamento de quinze modelos de aprendizado profundo pré-treinados. Com a validação, assumiu-se métricas quantitativas amplamente aplicadas em problemas de classificação. Em seguida, foi conduzida uma análise minuciosa, por meio de testes estatísticos para verificar diferenças significativas entre as arquiteturas. Além disso, para realizar a segmentação dos tumores cerebrais, foram aplicadas as redes UNet, UNet++ e FPN combinadas com as codificadoras de melhor desempenho. Logo, os resultados indicaram que a EfficientNetB7 apresentou o melhor desempenho na classificação, com uma acurácia de 97,68%, precisão de 97,63%, *recall* de 97,69%, *F1-Score* de 97,64% e especificidade de 99,21%. Na segmentação, a combinação de EfficientNetB7 com FPN foi a mais eficaz, alcançando 99,52% de acurácia, 85,23% em *F1-Score*, 74,29% para a métrica Interseção sobre União e 4,56 na Distância de *Hausdorff*. O estudo apresentou resultados significativos frente ao que é explorado na literatura, evidenciando a eficácia na detecção e classificação de tumores cerebrais.

Palavras-chave: deep learning; ressonância magnética; tumor cerebral; classificação; segmentação.

ABSTRACT

The development of malignant brain tumors characterizes brain cancer. According to the National Cancer Institute (INCA), approximately 11,000 new cases are registered in Brazil every year, with a mortality rate of approximately 84% of cases. In this context, early detection of brain tumors can considerably increase the survival rate of patients. Advances in Artificial Intelligence (AI) have improved the analysis of medical images. However, the classification and segmentation of brain tumors, using magnetic resonance imaging (MRI), are still challenging tasks due to several factors, such as changes in contrast, size, shape, position and variation of the tumor region, depending on the patient. Therefore, the main objective of this work is to develop a training processing flow and analyze the performance of deep neural network architectures in the classification of brain tumors in Meningioma, Glioma, Pituitary, as well as cases without tumor and the detection and segmentation of the tumor region. For this, three publicly available datasets from the literature of magnetic resonance images were used. In the classification stage, fifteen pre-trained deep learning models were trained. For validation, quantitative metrics widely applied in classification problems were assumed. Then, a thorough analysis was conducted using statistical tests to verify significant differences between the architectures. In addition, to segment the brain tumors, the UNet, UNet++, and FPN networks were applied in combination with the best-performing encoders. Therefore, the results indicated that EfficientNetB7 presented the best classification performance, with an accuracy of 97.68%, precision of 97.63%, *recall* of 97.69%, *F1-Score* of 97.64%, and specificity of 99.21%. In segmentation, the combination of EfficientNetB7 with FPN was the most effective, achieving 99.52% accuracy, 85.23% in *F1-Score*, 74.29% for the Intersection over Union metric and 4.56 in *Hausdorff Distance*. The study presented significant results compared to what is explored in the literature, evidencing the effectiveness in the detection and classification of brain tumors.

Keywords: deep learning; magnetic resonance imaging; brain tumor; classification; segmentation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

RM	Ressonância Magnética
INCA	Instituto Nacional de Câncer
TC	Tomografia Computadorizada
IA	Inteligência Artificial
DL	*Deep Learning*
CNN	*Rede Neural Convolucional*
DH	Distância de *Hausdorff*
IoU	Interseção sobre União
OMS	Organização Mundial da Saúde
ML	*Machine Learning*
FC	Camadas Totalmente Conectadas
Tanh	Tangente Hiperbólica
ReLU	Unidade Linear Retificada
VGG	Visual Geometry Group
CLAHE	Equalização de Histograma Adaptável Limitado por Contraste
AD	Aumento de Dados
VN	Verdadeiro Negativo
FP	Falso Positivo
FN	Falso Negativo
VP	Verdadeiro Positivo
DSC	Coeficiente de Similaridade de Dados
H_0	Hipótese Nula
H_A	Hipótese Alternativa

SUMÁRIO

1 INTRODUÇÃO

O cérebro é um dos órgãos mais complexos e sensíveis do corpo humano, responsável por controlar funções reguladoras como emoção, visão e reação. O crescimento de tumores cerebrais pode afetar de forma significativa essas funções. Os tumores cerebrais são considerados uma das piores doenças entre outros tipos de tumores devido à sua baixa taxa de sobrevivência entre os pacientes (DEANGELIS, 2001; ASIF *et al.*, 2022). Nesse contexto, existem dois tipos de tumores cerebrais, os malignos (cancerosos) e os benignos (não cancerosos). Os tumores benignos não possuem células cancerígenas e crescem lentamente, enquanto, os tumores malignos são caracterizados pela rápida disseminação para outros tecidos cerebrais (ASIF *et al.*, 2022).

Segundo o Instituto Nacional de Câncer (INCA), no Brasil são registrados cerca de 11 mil novos casos de pacientes com doenças tumorais cerebrais a cada ano, com uma taxa de mortalidade de aproximadamente de 84% (Instituto Nacional de Câncer, 2022). Para realizar a investigação e o diagnóstico de tumores cerebrais, os exames de imagens como Tomografia Computadorizada (TC) e Ressonância Magnética (RM) com contraste são as principais tecnologias utilizadas. Vale ressaltar que a ressonância magnética é a técnica de imagem médica mais comumente usada por se tratar de um processo não invasivo e oferecer imagens de alta resolução do tecido cerebral (AHMAD; CHOUDHURY, 2022; RAJ *et al.*, 2024).

O diagnóstico precoce e a localização precisa de tumores cerebrais podem aumentar as taxas de sobrevivência dos pacientes e fornecer opções oportunas para os médicos selecionarem planos de tratamento adequados na fase inicial da doença (KUMAR *et al.*, 2017). Entretanto, a detecção precoce e a classificação de tumores cerebrais é uma tarefa desafiadora por diversos fatores, por exemplo, alterações no tamanho, formato e posição do tumor no cérebro, dependendo do paciente (PEDDINTI *et al.*, 2021; OTTOM *et al.*, 2022). O processo de análise de imagens de RM e de segmentação manual do tumor requer tempo e esforço, devido ao processamento de grandes volumes de dados complexos, o que pode produzir resultados de segmentação imprecisos. Além disso, a classificação de características tumorais e normais é subjetiva e pode incluir uma margem de erro significativa (OTTOM *et al.*, 2022).

Nesse contexto, abordagens baseadas em Inteligência Artificial (IA) e *Deep Learning* (*DL*) têm aumentado o sucesso em pesquisas com imagens médicas e auxiliado a identificação e classificação precoce de tumores cerebrais. Técnicas computacionais auxiliam no desenvolvimento de modelos interpretáveis, capazes de reconhecer padrões automaticamente e diagnosticar

novos casos. Em geral, os sistemas automáticos de diagnóstico são destinados a complementar e auxiliar o diagnóstico médico (PADMAPRIYA; DEVI, 2024). Todavia, esses sistemas exigem uma grande quantidade de dados para obter um diagnóstico preciso. No entanto, a obtenção de tais dados não é uma atividade trivial (ASIF *et al.*, 2022). Logo, é necessário um esforço para desenvolver um sistema automatizado eficiente e que forneça resultados precisos.

Portanto, o objetivo desse projeto é desenvolver um fluxo de trabalho eficaz para auxílio ao diagnóstico e para classificar e detectar tumores cerebrais, por meio de imagens de ressonância magnética, com maior precisão e performance. É proposta a utilização de arquiteturas de redes neurais que garantam a assertividade. Vale ressaltar ainda que, apesar dos inúmeros benefícios, a proposta não substitui o diagnóstico feito por um médico especialista. O fluxo de trabalho serve como uma ferramenta de apoio para auxiliar os profissionais de saúde a realizarem diagnósticos precoce com maior precisão.

Para atingir os objetivos, foram utilizados os bancos de dados *Brain Tumor Classification (MRI)* (BHUVAJI *et al.*, 2020), *Figshare* (CHENG, 2017) e Br35H (CHAKRABARTY, 2017), disponíveis publicamente na literatura de imagens de ressonância magnética para realizar o treinamento dos modelos de classificação e segmentação de tumores cerebrais. A priori, para realizar a classificação dos tipos de tumores cerebrais em Meningioma, Glioma, Hipofisário e casos sem tumor, foi realizado treinamento e uma análise comparativa da precisão e assertividade de quinze modelos pré-treinados, sendo eles, AlexNet (KRIZHEVSKY *et al.*, 2012), DenseNet121 (HUANG *et al.*, 2018), DenseNet169 (HUANG *et al.*, 2018), DenseNet201 (HUANG *et al.*, 2018), EfficientNetB2 (TAN, 2019), EfficientNetB7 (TAN, 2019), GoogLeNet (SZEGEDY *et al.*, 2015), MobileNetV2 (SANDLER *et al.*, 2018), MobileNetV3 (HOWARD *et al.*, 2019), ResNet18 (HE *et al.*, 2015), ResNet50 (HE *et al.*, 2015), ResNet101 (HE *et al.*, 2015), SqueezeNet (IANDOLA *et al.*, 2016), VGG16 (SIMONYAN; ZISSERMAN, 2014), VGG19 (SIMONYAN; ZISSERMAN, 2014).

Ademais, para avaliar o desempenho das arquiteturas foram usadas as métricas de avaliação, como, acurácia, precisão, *recall*, *F1-Score* e especificidade. Em seguida, foi conduzida uma extensa análise, por meio da aplicação de testes estatísticos, para avaliar a significância dos resultados e para identificar possíveis diferenças entre as médias dos resultados obtidos. Por fim, foi realizado o procedimento de validação externa para avaliar a capacidade de generalização das redes em em contextos diferentes daqueles em que os dados de treinamento foram obtidos.

Em seguida, para realizar a detecção e segmentação da área tumoral foi utilizado o

conceito de codificadores e decodificadores. Nessa etapa, foi realizado o treinamento e análise do desempenho de arquiteturas. Para isso, foram usados os codificadores DenseNet201, EfficientNetB7 e ResNet50. Já para as redes decodificadoras utilizou-se as arquiteturas UNet (RONNEBERGER *et al.*, 2015), Unet++ (ZHOU *et al.*, 2018) e FPN (LIN *et al.*, 2017), amplamente reconhecidas em problemas de segmentação de imagens médicas. Logo, para avaliar os modelos de segmentação foram aplicadas as métricas acurácia, *F1-score*, Interseção sobre União (IoU), Distância de *Hausdorff* e tempo de treinamento. Além disso, o processo de aplicação de testes estatísticos foi realizado para avaliar a significância dos resultados. Finalmente, realizou-se a análise visual dos resultados obtidos pelas arquiteturas, por meio da geração de máscara de segmentação da área real do tumor.

Logo, entre os modelos de classificação analisados, a arquitetura EfficientNetB7 destacou-se com as melhores taxas nas métricas quantitativas, alcançando uma acurácia de 97,68%, precisão de 97,63%, *recall* de 97,69%, *F1-Score* de 97,64% e especificidade de 99,21%, embora tenha exigido o maior tempo de treinamento. Ademais, no contexto de segmentação de tumores cerebrais, as redes FPN e EfficientNetB7 resultados significativos frente ao que é explorado na literatura, com 99,52% de acurácia, 85,23% em *F1-Score*, 74,29% em IoU e 4,56 na Distância de *Hausdorff*.

Portanto, as principais contribuições deste trabalho são listadas a seguir:

- Foi realizada uma análise comparativa das principais redes de *DL* aplicadas à análise de imagens médicas, focada na detecção, segmentação e classificação de tumores cerebrais, incluindo Meningioma, Glioma, Hipofisário e casos sem tumor;
- Uma análise estatística detalhada foi conduzida em várias arquiteturas de redes neurais para atividades de classificação e segmentação, com o objetivo de avaliar a confiabilidade das conclusões e verificar se as diferenças observadas são estatisticamente relevantes ou apenas resultantes de variações aleatórias;
- Experimentos foram conduzidos utilizando uma combinação diversificada de conjuntos de dados, incluindo a adição de um conjunto de validação externa, para garantir uma generalização mais robusta do modelo.

1.1 Objetivos

1.1.1 Objetivo geral

Este trabalho possui como objetivo desenvolver um fluxo de trabalho eficaz para classificar e segmentar tumores cerebrais em imagens de RM com maior precisão e desempenho, por meio da utilização de modelos robustos de *Rede Neural Convolucional* (*CNN*).

1.1.2 Objetivos Específicos

- Elaborar um sistema capaz de classificar tumores cerebrais em Meningioma, Glioma, Hipofisário e casos Sem Tumor, por meio de uma extensa análise de modelos de *CNN* pré-treinados;
- Implementar o conceito de codificadores e decodificadores para desenvolver um sistema de segmentação de tumores cerebrais em imagens de RM;
- Realizar uma análise estatística para validar o desempenho dos modelos nos cenários de classificação e segmentação de imagens de tumores cerebrais;
- Validar a generalização e a robustez dos modelos através da abordagem de validação externa com dados independentes;

1.2 Contribuições Científicas

MOREIRA, ANDRESSA G.; SANTOS, STEFANE A. dos; OLIVEIRA, MICHELE F. de; PAULA JÚNIOR, IÁLIS C. DE; ASSIS, DÉBORA F. Classificação de Tumores Cerebrais em Imagens de Ressonância Magnética. In: **Anais do XXIV Simpósio Brasileiro de Computação Aplicada à Saúde**. Porto Alegre, RS, Brasil: SBC, 2024. p. 424–435. ISSN 2763-8952. Disponível em: https://sol.sbc.org.br/index.php/sbcas/article/view/28837.

1.3 Trabalhos Relacionados

Devido à necessidade de um diagnóstico preciso e precoce e aos fatores que dificultam a atividade de detecção e classificação de tumores cerebrais, diversos estudos no âmbito da IA foram realizados para auxiliar o diagnóstico automático de novos casos (EL-DAHSHAN *et al.*, 2014; TANDEL *et al.*, 2020; ALI *et al.*, 2022; RAGHAVENDRA *et al.*, 2023). Nesta seção são compilados trabalhos da literatura que empregam o conceito de IA para a classificação de

tumores cerebrais.

Em Khan *et al.* (2022) os autores propõem dois modelos de aprendizado profundo para classificar tumores cerebrais em normais e anormais, como também de acordo com os tipos de tumores meningioma, glioma e hipofisário. Para a realização dos experimentos foram utilizados dois conjuntos de dados disponíveis publicamente na literatura: *Harvard Medical* e *Figshare*. A arquitetura proposta é uma *CNN* de 23 camadas combinada com a rede VGG16. Como resultados, a rede alcançou uma acurácia de 97,8%, precisão de 96,5%, *recall* de 97,4%, *F1-Score* de 96,4%. Todavia, o estudo não apresenta a aplicação de validação externa ou o uso de testes estatísticos para validar o cenário existente de resultados.

Os autores Zhu *et al.* (2024) definem uma abordagem para diagnosticar tumores cerebrais por meio de aprendizado profundo e um algoritmo meta-heurístico. O método consiste em extrair características de imagens de ressonância magnética cerebral com a rede AlexNet, reduzir a complexidade da arquitetura empregando uma rede *Extreme Learning Machine* (ELM) como uma camada de classificação e ajustar os parâmetros da rede ELM usando um Algoritmo de Otimização de *Grasshopper Amended* (AGOA). Para a análise do desempenho foi utilizado o conjunto de dados publicamente disponível, obtido do *The Cancer Imaging Archive* (TCIA) de pacientes com glioblastoma. Como resultados, o método atingiu para a métrica Coeficiente de correlação de Matthew (MCC) de 90%, uma acurácia de 96%, precisão de 94%, *recall* de 94%, F1-Score de 96% e especificidade 96%. Em comparação, neste trabalho foram encontradas pontuações superiores para as métricas avaliadas, com uma acurácia de 97,68%, precisão de 97,63%, *recall* de 97,69%, *F1-Score* de 97,64% e especificidade 99,21%.

Os autores em El-Assiouti *et al.* (2023) apresentam novas técnicas de aumento de dados, definidas como RegionInpaint, aumento de corte, e aumento do RegionMix, a fim de melhorar o desempenho da identificação de tumores cerebrais. Foram utilizados os conjuntos de dados SPMRI e Br35H, compostos por imagens ressonância magnética cerebral distribuídas nas classes “Tumor” e “Não tumoral”. Para a tarefa de classificação binária foram realizados testes com diversas arquiteturas pré-treinadas. O melhor resultado para a precisão de teste alcançado foi de 96,88% para o modelo VGG19 treinado com imagens do conjunto de dados Br35H com técnicas de aumento. Além disso, para a tarefa de segmentação a rede VGGUNET obteve uma acurácia de validação de 98,67%. Apesar dos resultados promissores obtidos por esses estudos, algumas limitações ainda podem ser abordadas. Por exemplo, os estudos empregaram conjuntos de dados relativamente pequenos, o que pode limitar sua generalização. Além disso,

as pontuações para as medidas de avaliação obtidas foram inferiores às do presente trabalho.

Em Rehman *et al.* (2020) os autores propuseram um *framework* para classificação de tumores cerebrais utilizando o conjunto de dados de tumor cerebral Figshare. No estudo foram utilizadas três arquiteturas de redes neurais convolucionais: AlexNet, GoogLeNet e VGGNet, para classificar tumores cerebrais, como meningioma, glioma e hipófisário. No estudo são aplicados modelos de técnicas de aumento de dados para garantir a generalização dos resultados, aumentar as amostras do conjunto de dados e reduzir a chance de overfitting. Como resultado, a arquitetura VGG16 obteve a maior acurácia com 98,96% em termos de classificação, após o ajuste fino. Embora o modelo tenha obtido alta taxa para a métrica de acurácia após o ajuste fino, o trabalho não apresenta os resultados para as demais métricas. Todavia, são apresentados os valores obtidos sem o ajuste fino, na qual, o modelo atingiu 89,76% de acurácia, 87,81% para a métrica recall, 94,64% para especificidade e 88,67% para a precisão.

No trabalho Yan *et al.* (2022) é proposto um modelo melhorado da rede U-Net, denominado, SEResU-Net, que combina a rede residual profunda e a Rede Squeeze-and-Excitation. Os experimentos foram realizados utilizando os conjuntos de dados BraTS2018 e BraTS2019. Para avaliar os resultados foram utilizadas as métricas: coeficiente de similaridade de dados, sensibilidade, especificidade e distância de Hausdorff (HD). Por fim, a rede SEResU-Net obteve para o coeficiente de similaridade de dados 93,73%, 91,08% e 87,58% para todo o tumor, o núcleo do tumor e o tumor realçado, respectivamente.

Em Bindu e Devi (2024), o objetivo principal reside na classificação binária de tumores cerebrais. Para este propósito, o estudo empregou o método *de transfer learning* para extrair os recursos de imagens de ressonância magnética, utilizando quatro Redes Neurais Convolucionais (CNNs) pré-treinadas. Adicionalmente, os autores consideraram imagens segmentadas do conjunto de dados original para treinamento. Para a segmentação, os autores utilizam métodos clássicos, como, K-means e Fuzzy C-means, mas não apresentam resultados quantitativos para os métodos aplicados. Já os resultados experimentais demonstram uma classificação com uma acurácia de 96,98%.

Portanto, em diversas aplicações médicas, modelos de aprendizagem profunda são utilizados em tarefas de auxílio ao diagnóstico. Enquanto alguns estudos focam no desenvolvimento de modelos para a classificação de tumores cerebrais (SRIVASTAVA *et al.*, 2023; ISLAM *et al.*, 2023; KUSHWAHA; MAIDAMWAR, 2022), outros se dedicam à segmentação da área tumoral (LIU *et al.*, 2023; WULANDARI *et al.*, 2018; HARSHAVARDHAN *et al.*, 2017).

Neste trabalho, é implementado um fluxo de trabalho que realiza tanto a segmentação quanto a classificação do tumor, visando apoiar os profissionais de saúde na detecção e diagnóstico da doença. É importante destacar que, além dessas tarefas, o presente estudo conduz uma análise abrangente dos resultados, por meio de testes estatísticos e validação externa. Essa abordagem, que se diferencia de outros trabalhos na área, proporciona maior rigor analítico e confiabilidade nos resultados obtidos.

A Tabela 1 apresenta a síntese dos trabalhos relacionados, comparando este estudo com as pesquisas existentes na literatura. A tabela destaca os principais aspectos analisados,como, a base de dados utilizada, o pré-processamento aplicado, as arquiteturas de classificação e segmentação empregadas, a realização de validação externa, as principais métricas de avaliação, a aplicação de testes estatísticos, e os resultados mais relevantes alcançados.

1.4 Organização do Documento

O Capítulo 2 apresenta os principais conceitos teóricos abordados no desenvolvimento deste projeto. Em seguida, o Capítulo 3 descreve as etapas metodológicas realizadas na elaboração deste trabalho. O Capítulo 4 apresenta uma análise e discussão dos resultados experimentais. Por fim,o Capítulo 5 apresenta as considerações finais deste projeto e determina os trabalhos futuros.

Tabela 1 – Sumarização dos Trabalhos Relacionados.

Ref.	*Dataset*	Pré-proces.	Classificação	Segmentação	Valid. Ext	Métricas	Testes	Resultados
(KHAN *et al.*, 2022)	Harvard Medical, *Figshare*.	Redimensionamento.	CNN proposta.	✗	✗	Acurácia Precisão, *Recall*, FPR, TRN, *F1-Score*, ROC.	✗	Acc: 97,8%, Prec: 96,5%, *Recall*: 97,4%, FPR: 0,016, TRN: 0,983, F1: 96,4%, ROC: 98,9%.
(ZHU *et al.*, 2024)	TCIA	✗	AlexNet, ELM, AGOA.	✗	✗	MCC Acurácia Precisão, *Recall*, *F1-Score*, Especificidade.	✗	MCC: 90%, Acc: 96%, Prec: 94%, *Recall*: 94%, F1: 96%, Esp: 96%.
(EL-ASSIOUTI *et al.*, 2023)	SPMRI, Br35H.	Filtro Gaussiano, Tons de Cinza, Binarização, Erosão, Dilatação, Corte, Normalização.	VGG19.	VGGUNET.	✗	Acurácia Geral, Precisão, *Recall*, *F1-Score*, AUC, Dice.	✗	Classificação: Acc: 96,8%, Prec: 96,06%, *Recall*: 96,73%, F1: 96,37%, AUC: 99,34%. Segmentação: Acc: 98,67%, Prec: 85,82%, *Recall*: 88,72%, F1: 96,4%, DICE: 85,82%.
(REHMAN *et al.*, 2020)	*Figshare*.	Alongamento de contraste.	AlexNet, GoogLeNet, VGG16.	✗	✗	Acurácia Precisão, *Recall*, Especificidade.	✗	ACC (ajuste fino): 98,69%, Acc: 89,76%, Prec: 88,67%, *Recall*: 87,81%, Esp: 94,64%.
(YAN *et al.*, 2022)	BraTS2018, BraTS2019.	Normalização, Corte.	✗	SEResU-Net	✗	Dice, *Recall*, Especificidade, DH.	✗	Dice: 94,13%, *Recall*: 93,84%, Esp: 95,79%, DH: 2,13.
(BINDU; DEVI, 2024)	MRI Kaggle.	Equalização histograma, OTSU.	AlexNet, GoogleNet, VGG-16, VGG-19, ResNet-18, ResNet-50.	K-means, Fuzzy C-Means.	✗	Acurácia.	✗	Classificação: Acc: 96,98%.
Este Trabalho	Brain Tumor Classification, *Figshare*, Br35H.	Redimensionamento, Tons de Cinza, Binarização, Filtragem Bilateral, CLAHE.	AlexNet, DenseNet121, DenseNet169, DenseNet201, EfficientNetB2, EfficientNetB7, GoogLeNet, MobileNetV2, MobileNetV3, ResNet18, ResNet50, ResNet101, SqueezeNet, VGG16, VGG19.	Codificadores: DenseNet201, EfficientNetB7, ResNet50. Decodificadores: UNet, UNet++, FPN.	Classificação: Acc: 99,01%, Prec: 98,79%, *Recall*: 98,91%, F1: 98,85%, Esp: 99,66%.	Acurácia Precisão, *Recall*, *F1-Score*, Especificidade, IoU, DH.	*Shapiro-Wilk*, *Levene*, *Friedman*, *Nemenyi*.	Classificação: Acc: 97,68%, Prec: 97,63%, *Recall*: 97,69%, F1: 97,64%, Esp: 99,21%. Segmentação: Acc: 99,52%, F1: 85,23%, IoU: 74,29%, DH: 4,56.

Fonte: Elaborado pela autora.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, será apresentado um embasamento teórico sobre os principais conceitos abordados no desenvolvimento deste projeto. As seções conduzem uma revisão da literatura sobre tumores cerebrais (Seção 2.1), IA para imagens médicas (Seção 2.2), *CNN* (Seção 2.3) e segmentação semântica (Seção 2.4). Também abordam técnicas de pré-processamento de imagens (Seção 2.5), aumento de dados (Seção 2.6), *grid search* (Seção 2.7), validação cruzada (Seção 2.8), métricas de avaliação (Seção 2.9), testes estatísticos (Seção 2.10) e validação externa (Seção 2.11). Com isso, serão levantados aspectos relevantes relacionados ao tema e identificadas lacunas para serem exploradas em relação à questão central de pesquisa.

2.1 Tumor Cerebral

O cérebro humano é considerado um dos principais órgãos, responsável por diversos processos e funções reguladoras do corpo, como, emoções, habilidades motoras, memória, visão, respostas e respiração. Tais funções podem ser significativamente interrompidas pelo desenvolvimento de células cerebrais disformes, que podem tornam-se cancerígenas, definidas como tumor cerebral. Os tumores cerebrais podem ser classificados de duas formas, sendo tumor cerebral primário ou um tumor cerebral metastático. Os tumores cerebrais primários se desenvolvem no interior do cérebro e representam o desenvolvimento dos tecidos cerebrais. Em contrapartida, os tumores cerebrais metastáticos são definidos como aqueles tumores que se desenvolvem em uma parte do corpo e eventualmente se espalham para outros locais, incluindo o cérebro humano (ANANTHARAJAN *et al.*, 2024).

O cérebro possui múltiplas camadas de proteção e defesa que afastam substâncias estranhas e mantêm um sistema interno estável. O crânio rígido é uma defesa estrutural contra traumas físicos. Além do crânio, existem outras barreiras estruturais e funcionais, como a barreira hematoencefálica (BHE) e mecanismos de autorregulação, responsáveis por garantir a homeostase do ambiente cerebral e por criar um ambiente inóspito para o crescimento de células tumorais (CHA, 2006a). Dessa forma, a barreira hematoencefálica regula o fluxo de íons, oxigênio e nutrientes entre o sangue e o parênquima cerebral e impede a entrada de substâncias no parênquima e de compostos sanguíneos no sistema nervoso central (SNC). Em suma, a BHE regula seletivamente a passagem de substâncias entre o sangue e o cérebro, permitindo a entrada de nutrientes essenciais enquanto bloqueia a passagem de substâncias potencialmente

tóxicas. Logo, independente das características genética ou histológica, os tumores cerebrais estão confinados pelas barreiras inerentes do cérebro (CHA, 2006a; COUREUIL *et al.*, 2017).

Portanto, a qualidade e expectativa de vida dos pacientes são melhoradas significativamente pela identificação precoce da doença e pelos planos de tratamento (ANANTHARAJAN *et al.*, 2024). Nesse contexto, a apresentação clínica das metástases cerebrais é variável. A cefaléia é um sintoma comum, mas não exclusiva para o diagnóstico. Exames de neuroimagem são recomendados para avaliar características de cefaléia atípica. Em geral, as características atípicas incluem aumento rápido da frequência da dor de cabeça, falta de coordenação, sinais neurológicos localizados e dor de cabeça causando o despertar do sono. As convulsões são outro sintoma comum em pacientes. A neuroimagem deve ser considerada para todos os pacientes que apresentam crise. Além disso, os pacientes podem apresentar sintomas neurológicos focais, como fraqueza, perda sensorial, distúrbios visuais e comprometimento cognitivo ou comportamental (MCFALINE-FIGUEROA; LEE, 2018).

Para realizar o diagnóstico de tumores cerebrais um dos pré-requisitos é o conhecimento de sua incidência e prevalência, classificação histopatológica e evolução clínica. O diagnóstico e a classificação dos tumores cerebrais seguem a classificação da Organização Mundial da Saúde (OMS). Além disso, a neuroimagem desempenha um papel fundamental no diagnóstico de tumores cerebrais. Para realizar a investigação e o diagnóstico de tumores cerebrais, os exames de imagens como TC e RM com contraste são tecnologias comumente utilizadas. Fatores como calcificações, componentes císticos, realce de contraste e intensidade de sinal nas imagens ponderadas permitem a caracterização dos tumores. Ademais, o diagnóstico pode ser feito através de uma biópsia, que consiste na retirada de uma amostra de tecido do tumor cerebral (CHOURMOUZI *et al.*, 2014; AHMAD; CHOUDHURY, 2022).

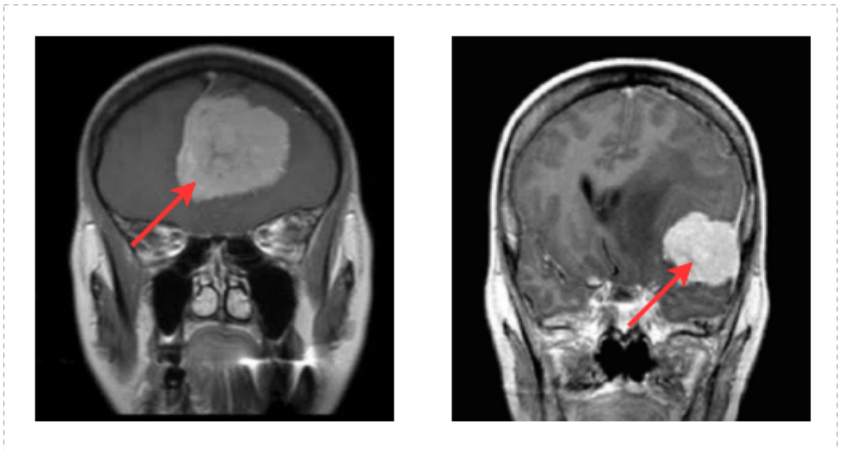
O tratamento de tumores cerebrais podem variar de acordo com o tipo de tumor e o prognóstico da doença. A cirurgia pode ser necessária dependendo do diagnóstico e da extensão da doença. A radioterapia cerebral consiste em outro tipo de tratamento e resulta na melhora da função neurológica. No entanto, está associada à neurotoxicidade, particularmente à fadiga e à disfunção neurocognitiva. Além desses tratamentos, existem as terapias sistêmicas, como imunoterapia e quimioterapia. Essas terapias são tratamentos médicos administrados para atingir o sistema circulatório do paciente e serem distribuídas por todo o corpo, alcançando células cancerígenas que podem estar presentes em várias partes do organismo (MCFALINE-FIGUEROA; LEE, 2018).

Logo, os tumores cerebrais são classificados de acordo com os agrupamentos histológicos principais. Dessa forma, os tumores de meninges incluem os tumores meningioma e hemangioblastoma. Os tumores de tecido neuroepitelial são denominados de glioma, astrocitoma (grau II), astrocitoma anaplásico (grau III), glioblastoma (grau IV), oligodendroglioma e ependimoma. Por fim, os tumores de células germinativas e tumores da região selar são compostos por tumores hipofisários e craniofaringioma (FISHER *et al.*, 2007). Neste trabalho, serão detalhados os tumores meningioma, glioma e hipofisário, cada um pertencente a um tipo de agrupamento histológico.

2.1.1 Meningioma

Os meningiomas são, em sua maioria, neoplasias biologicamente benignas e de crescimento lento, que surgem das células meningoteliais da aracnóide (MCFALINE-FIGUEROA; LEE, 2018). Os meningiomas podem ser detectados ao longo das superfícies externas do cérebro, bem como dentro do sistema ventricular, local onde surgem as células aracnóides. As localizações mais comuns incluem a face parassagital da convexidade cerebral, a convexidade do hemisfério lateral, a asa esfenoidal, a fossa craniana média e o sulco olfatório. A Figura 1 apresenta um exemplo da localização do tumor cerebral meningioma (CHOURMOUZI *et al.*, 2014).

Figura 1 – Exemplo de tumor cerebral Meningioma.



Fonte: Adaptado de (BHUVAJI *et al.*, 2020).

De acordo com o esquema de classificação da OMS, os meningiomas podem ser classificados como graus I a III, com base na sua histologia. Dessa forma, os meningiomas de grau I, identificados como meningiomas benignos, são os mais comuns e apresentam prognóstico

favorável. Entretanto, os meningiomas de graus II e III são mais agressivos e associados a uma sobrevida de 78% e 44% em 5 anos, respectivamente. O meningioma grau II, apresenta aumento do número de mitoses, celularidade e proporção nuclear/citoplasmática. O meningioma grau III é caracterizado pela presença de características malignas, como anaplasia celular evidente, alta taxa mitótica, áreas de necrose e invasão cerebral (CHA, 2006b; CHOURMOUZI *et al.*, 2014).

A apresentação clínica e os sintomas estão relacionados ao tamanho e à localização do tumor. Isto posto, menos de 10% dos meningiomas são sintomáticos, porém, em casos específicos podem apresentar sintomas como dores de cabeça, convulsões ou sintomas neurológicos focais. O diagnóstico é realizado muitas vezes incidentalmente em neuroimagem ou radiografias. Em casos de incerteza diagnóstica, pode-se realizar a biópsia ou ressecção para estabelecimento de um diagnóstico preciso (MCFALINE-FIGUEROA; LEE, 2018).

2.1.2 Glioma

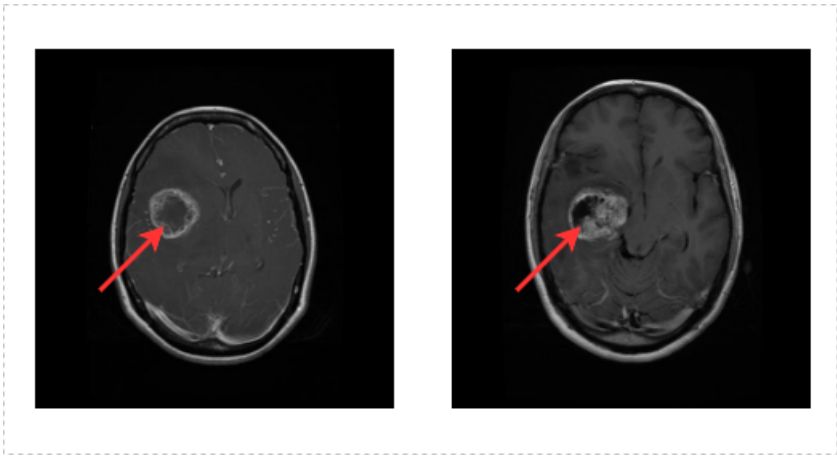
Os gliomas são os tumores cerebrais primários mais comuns do sistema nervoso central e incluem astrocitomas, oligodendrogliomas, ependimomas e uma variedade de histologias raras. Entre eles, o glioblastoma, um astrocitoma grau IV, destaca-se como o tipo mais comum e agressivo. A classificação desses tumores baseia-se, principalmente, em características histológicas, como celularidade, atipia nuclear, atividade mitótica, vascularização e necrose. Os gliomas estão localizados, principalmente, no cerebelo, na região hipotálamo-quiasmática, no nervo óptico, no tronco cerebral e nos hemisférios cerebrais (CHOURMOUZI *et al.*, 2014; MCFALINE-FIGUEROA; LEE, 2018).

A OMS divide os tipos de tumores cerebrais em 4 graus distintos. O grau I, ou astrocitoma pilocítico, são tumores de caráter benigno e de crescimento lento. O grau II, ou astrocitoma difuso, são tumores de baixo grau de malignidade, mas com tendência a progressão para graus mais altos. O grau III, denominado de astrocitoma anaplásico, já possui um grau de agressividade maior, com alta atividade mitótica. Por fim, os tumores de grau IV, ou glioblastoma, possuem alto grau de agressividade e proliferação vascular (NEUROCIRURGIA, 2024).

Pacientes diagnosticados com tumores com maior grau de agressividade, podem apresentar dores de cabeça, convulsões ou sintomas neurológicos focais. Devido à sua natureza agressiva, os sintomas podem se desenvolver rapidamente. Em termos de diagnóstico, é necessário a confirmação patológica por biópsia ou ressecção cirúrgica. Todavia, a ressonância magnética do cérebro com e sem contraste é uma das principais modalidades de escolha para neuroimagem

para a detecção de tumores cerebrais. Dessa forma, a aparência do tumor pode variar, sendo caracterizado, principalmente, por uma lesão de massa supratentorial, heterogeneamente realçada, com necrose central e sinal de substância branca circundante (MCFALINE-FIGUEROA; LEE, 2018). A Figura 2 apresenta uma imagem de ressonância magnética do cérebro com o tumor cerebral glioma.

Figura 2 – Exemplo de tumor cerebral Glioma.



Fonte: Adaptado de (BHUVAJI *et al.*, 2020).

2.1.3 Hipofisário

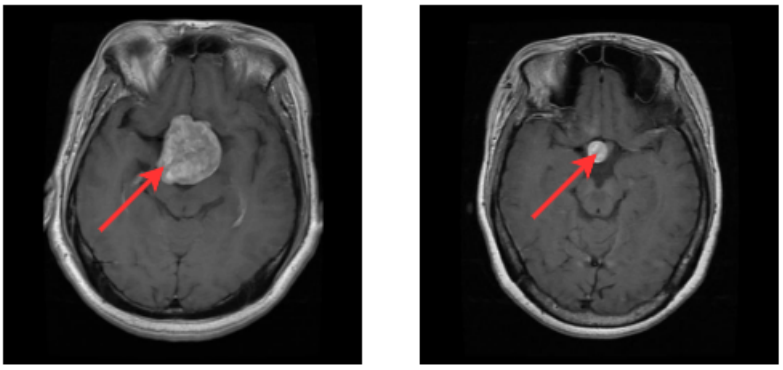
Tumores hipofisários são crescimentos anormais na glândula hipófise, localizada na base do cérebro. Esses tumores, conhecidos como adenomas hipofisários, podem ser assintomáticos e descobertos incidentalmente durante exames de imagem para outras condições e são assintomáticos. Embora tais tumores não causem sintomas e não precisem de tratamento, alguns podem crescer e interferir na produção hormonal, resultando em sintomas variados dependendo do tipo de hormônio afetado (DAI *et al.*, 2021).

Dessa forma, os tumores hipofisários são responsáveis por aproximadamente 10 a 15% das neoplasias intracranianas e são os segundos tumores cerebrais primários mais comuns. A maioria dos tumores são geralmente benignos e classificados como adenomas hipofisários. Estes tumores são classificados pela OMS com base em características histopatológicas, como o conteúdo hormonal das células tumorais e suas características ultraestruturais. Todavia, alguns tumores apresentam metástase cranioespinhal, e esses tumores são conhecidos como carcinomas hipofisários. Essas neoplasias são muito raras, compreendendo menos de 1% de todos os tumores hipofisários, sendo a maioria tumores hormonalmente ativos (METE; LOPES, 2017; DAI *et al.*,

2021).

Em casos raros, os adenomas hipofisários podem ser agressivos, crescendo rapidamente e sendo resistentes ao tratamento convencional. Esses tumores são desafiadores para o manejo clínico, e a identificação precoce é crucial para a aplicação de estratégias terapêuticas mais agressivas. A definição e classificação desses tumores ainda são temas de debate, com a necessidade de critérios mais objetivos para melhorar o tratamento e o prognóstico dos pacientes (DAI *et al.*, 2021). A Figura 3 apresenta um exemplo da localização do tumor cerebral hipofisário.

Figura 3 – Exemplo de tumor cerebral Hipofisário.



Fonte: Adaptado de (BHUVAJI *et al.*, 2020).

2.2 Inteligência Artificial para Imagens Médicas

A IA é uma ampla área da ciência da computação que visa construir métodos automáticos para resolver problemas que normalmente exigem inteligência humana (PANNU, 2015; SARKER, 2022). Como complemento, a área de Visão Computacional, lida com uma grande variedade de problemas, como segmentação de imagens, reconhecimento de objetos, detecção e reconstrução. Este campo visa modelar e entender o mundo visual extraindo informações úteis de imagens digitais, inspiradas por tarefas complexas da visão humana. O conceito de *Machine Learning* (*ML*), por sua vez, é uma subárea da IA que constrói sistemas capazes de aprender automaticamente a partir de dados e observações. Ademais, *DL* é um subcampo de *ML*, aplicado na resolução de problemas complexos de análise e interpretação de informações visuais. O termo profundo se refere a modelos de redes neurais de múltiplas camadas, como as *CNN*s. Dessa forma, existe o crescimento considerável do conceito de *DL* em problemas de análise de imagens

médicas (OLVERES *et al.*, 2021).

As modalidades de imagem mais populares usadas para análise cerebral são a TC, RM, ultrassom, tomografia computadorizada por emissão de fótons simples (SPECT), tomografia por emissão de pósitrons (PET) e raio-X. Todavia, as imagens de RM têm se tornado a principal tecnologia de imagem médica, pois oferece imagens de melhor contraste em comparação com outras técnicas de imagem médica, de forma não invasiva, para a realização do diagnóstico, avaliação e monitoramento. Recentemente, as abordagens baseadas *ML* e *DL* estão sendo comumente utilizadas para identificar o tumor cerebral a partir de imagens de RM, pois fornecem resultados de detecção bastante precisos. Vale ressaltar que, o diagnóstico em estágio inicial e identificação do tumor cerebral são tarefas desafiadoras, devido ao tamanho, formato, posição e forma do tumor no cérebro, além de existir uma falta de informações precisas sobre o tamanho do tumor resultante de imagens de baixa resolução (OLVERES *et al.*, 2021; AHMAD; CHOUDHURY, 2022).

Portanto, a IA desempenha um papel crucial na análise de imagens médicas, o que melhora de forma significativa o diagnóstico, por meio da automatização da extração de características complexas e precisas. O desenvolvimento contínuo de algoritmos de IA está revolucionando a análise de imagens médicas, tornando-se uma ferramenta essencial para melhorar a especificidade dos diagnósticos, superando outros métodos quantitativos tradicionais. Desse modo, para garantir a eficácia da IA, é necessário utilizar grandes volumes de dados rotulados e padronizados. O processo envolve o pré-processamento dos dados, a inserção em arquiteturas de aprendizado profundo, e a otimização de hiperparâmetros. Além disso, técnicas de aumento de dados são empregadas para superar a limitação de dados rotulados, garantindo que os modelos de IA alcancem um desempenho plausível no diagnóstico de imagens médicas (YOON *et al.*, 2019).

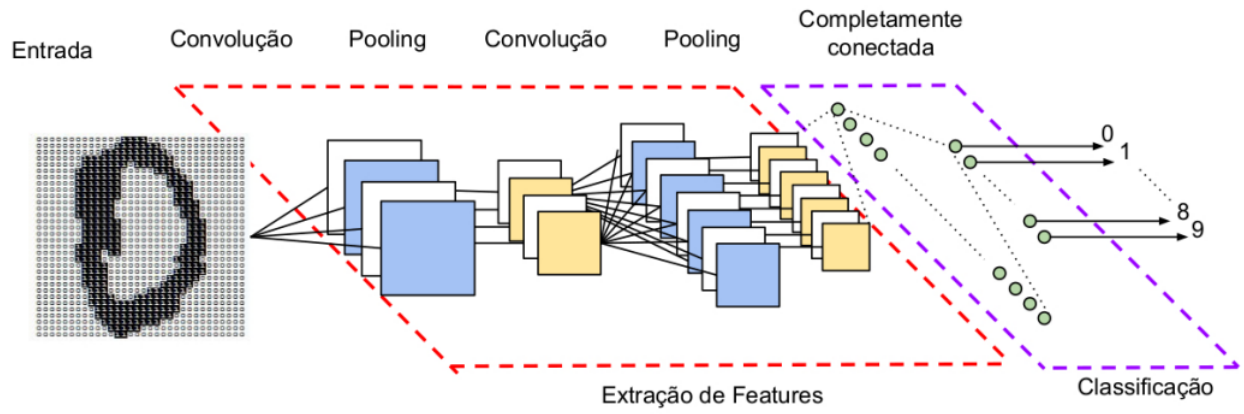
2.3 Rede Neural Convolucional (CNN)

As *CNN*s ou ConvNets, introduzidas por Lecun *et al.* (1998), representam um algoritmo de aprendizado profundo que aceitam imagens de entradas e são ideais em tarefas de aprendizado automático e classificação de imagens. Inspiradas no processamento visual biológico, as *CNN*s desempenham a função de aplicar filtros aos dados durante o processamento da rede, imitando assim o funcionamento do sistema visual humano (VARGAS *et al.*, 2016).

As *CNN*s são compostas por três tipos de camadas com diferentes funções, definidas

como, camadas convolucionais, camadas de *pooling* e Camadas Totalmente Conectadas (FC). Nesse sentido, as camadas convolucionais executam operações de convolução e aplicam funções de ativação sobre os dados de entrada. Em seguida, as camadas de *pooling* subsequentes reduzem as dimensões das camadas de anteriores. Por fim, as FC são responsáveis por combinar as saídas geradas pelas camadas convolucionais com um vetor unidimensional que representa as probabilidades de cada característica pertencer a um rótulo específico (O'SHEA; NASH, 2015; VAZ; BALAJI, 2021). A Figura 4 apresenta a estrutura básica da arquitetura *CNN*.

Figura 4 – Exemplo de arquitetura CNN para classificação de imagens.



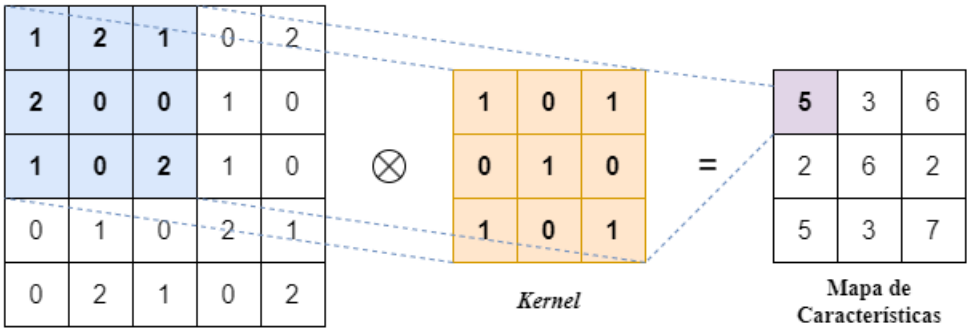
Fonte: (VARGAS *et al.*, 2016).

As camadas convolucionais desempenham um papel crucial no funcionamento das *CNN*s. São camadas responsáveis pela extração de características dos dados de entrada, por meio de operações de convolução e de ativação. Nesse contexto, na operação linear de convolução, ocorre a multiplicação elemento a elemento entre uma matriz de recursos denominada *kernel* e a matriz de entrada, chamada tensor. Este processo gera um mapa de características, obtido pela soma dos produtos de cada elemento das matrizes sobrepostas, que atua como uma entrada para a próxima camada na rede (O'SHEA; NASH, 2015; VARGAS *et al.*, 2016).

A Figura 5 exibe uma representação visual de uma camada convolucional. No exemplo, a operação de convolução é representada por uma matriz de entrada de tamanho 5 x 5 e um *kernel* de tamanho 3 x 3. O resultado é um mapa de características de tamanho 3 x 3. Logo, é realizado o produto entre os valores da matriz de entrada e os valores do *kernel*. Cada um dos valores são multiplicados e somados para obter a saída na posição correspondente do mapa de características. Dessa forma, na iteração em destaque, é realizada a operação de entre os valores

da matriz de entrada e do *kernel*, sendo: $(1 \cdot 1) + (2 \cdot 0) + (1 \cdot 1) + (2 \cdot 0) + (0 \cdot 1) + (0 \cdot 0) + (1 \cdot 1) + (0 \cdot 0) + (2 \cdot 1) = 5$. O resultado corresponde a uma posição no mapa de características. O processo é repetido até que toda a matriz de entrada seja percorrida.

Figura 5 – Representação visual de uma camada convolucional.



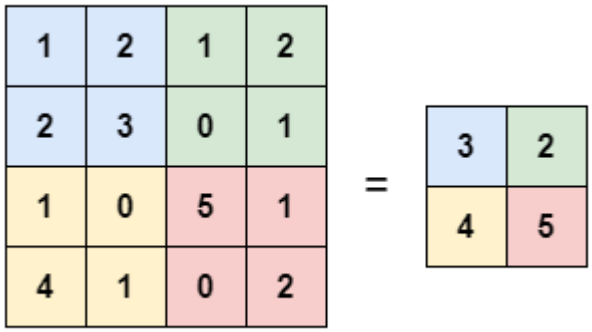
Fonte: Adaptado de (ALZUBAIDI *et al.*, 2021).

Em seguida, são adicionadas funções de ativação para garantir propriedades não lineares, necessárias para que o modelo consiga aprender representações complexas, como a modelagem de diversos tipos de dados de alta dimensão. Dessa forma, as funções de ativação garantem a capacidade não-linear ao processamento realizado pelas redes neurais, sendo a Tangente Hiperbólica (Tanh) e a Unidade Linear Retificada (ReLU) as funções mais comumente utilizadas (SHARMA *et al.*, 2020; VAZ; BALAJI, 2021).

Após as camadas de convolução, são adicionadas camadas de *pooling*, responsáveis por reduzir as dimensões das camadas de entrada. Essa redução introduz uma invariância espacial à medida que diminui o número de parâmetros a serem aprendidos pela rede, e, consequentemente, aumenta a agilidade no treinamento. Logo, o método de *pooling* mais utilizado, denominado *max-pooling*, seleciona o valor mais alto em uma vizinhança para aplicar a um tensor de saída. A Figura 6 ilustra a operação de *pooling* em uma matriz de tamanho 4 x 4, por meio de um filtro de *pooling* de tamanho 2 x 2, neste procedimento são selecionados os maiores valores da vizinhança (ALZUBAIDI *et al.*, 2021; VAZ; BALAJI, 2021).

Por fim, para realizar a tarefa de classificação, são acrescentadas FC. Estas camadas estabelecem conexões entre cada neurônio e os neurônios das camadas anteriores. Nesse contexto, as FC são responsáveis por combinar as saídas geradas pelas camadas convolucionais ou de *pooling* em um vetor unidimensional, que consiste em probabilidades de cada característica pertencer a um rótulo. A camada totalmente conectada final possui um número de nós de saída

Figura 6 – Representação da operação *Max-Pooling*.



Fonte: Adaptado de (UNIVERSITY, 2022).

igual ao número de classes do problema de classificação, seguida por uma função não linear. A escolha da função de ativação final depende do problema em questão, sendo a função de ativação sigmóide utilizada em tarefas de classificação binária e a função softmax aplicada em tarefas de classificação multiclasse (YAMASHITA *et al.*, 2018; VAZ; BALAJI, 2021).

2.3.1 Treinamento da Rede

O processo de treinamento de uma rede neural consiste em encontrar os pesos ideais para os nós em uma camada para que o aprendizado ocorra. No treinamento são definidos os *kernels* das camadas convolucionais e os pesos nas FC. No treinamento de redes neurais, a inicialização dos parâmetros, *kernels* e pesos é realizada com valores aleatórios. A entrada da rede consiste em vetores de recursos do conjunto de dados de treinamento. Para avaliar o desempenho do modelo é utilizada uma função de perda, responsável pelos cálculos dos erros em cada saída. O algoritmo *backpropagation* é definido para otimização de algoritmo, por meio da alteração dos parâmetros em cada nó usando gradiente descendente. Logo, os *kernels* e os pesos são atualizados de acordo com o valor da perda definido pelo algoritmo de otimização *backpropagation*. Após repetidas iterações, são calculados os parâmetros ideais que fornecem perda mínima no algoritmo.

2.3.2 Transfer Learning

A técnica de *Transfer Learning* é uma metodologia que se concentra na transferência de conhecimento entre domínios. A técnica é aplicada para melhorar o desempenho dos alunos-alvo em um determinado domínio, transferindo informações e conhecimento contido em um

domínios diferentes, mas relacionados (ZHUANG *et al.*, 2021).

O *Transfer Learning* é uma técnica particularmente valiosa, visto que reduz a dependência de um grande volume de dados do domínio-alvo para construir alunos-alvo eficazes. O conceito central da aprendizagem por transferência é inspirado nas habilidades cognitivas humanas, onde o conhecimento adquirido em uma tarefa pode ajudar na execução de uma tarefa relacionada de forma mais eficiente, por exemplo, uma pessoa habilidosa em tocar violino pode achar mais fácil aprender piano porque ambos são instrumentos musicais e compartilham características em comum (WEISS *et al.*, 2016; ZHUANG *et al.*, 2021).

Dessa forma, em certos cenários, obter dados que correspondam ao espaço de características e à distribuição esperada pode ser difícil e caro. Isso ocorre porque pode haver um suprimento limitado de dados de treinamento no domínio de destino, ou esses dados podem ser caros para coletar e rotular, ou até mesmo inacessíveis. Portanto, surge a necessidade de criar um modelo de alto desempenho para o domínio de destino por meio de um domínio de origem relacionado, sendo essa a motivação principal para o uso do aprendizado por transferência (WEISS *et al.*, 2016).

2.3.3 Arquiteturas CNN

Atualmente, existem diversas arquiteturas de redes neurais convolucionais, projetadas para atender diferentes necessidades, como o número de camadas, a quantidade de classes e o volume de imagens. Essas variações permitem um aprendizado mais robusto e eficaz, adaptando-se às particularidades de cada tarefa.

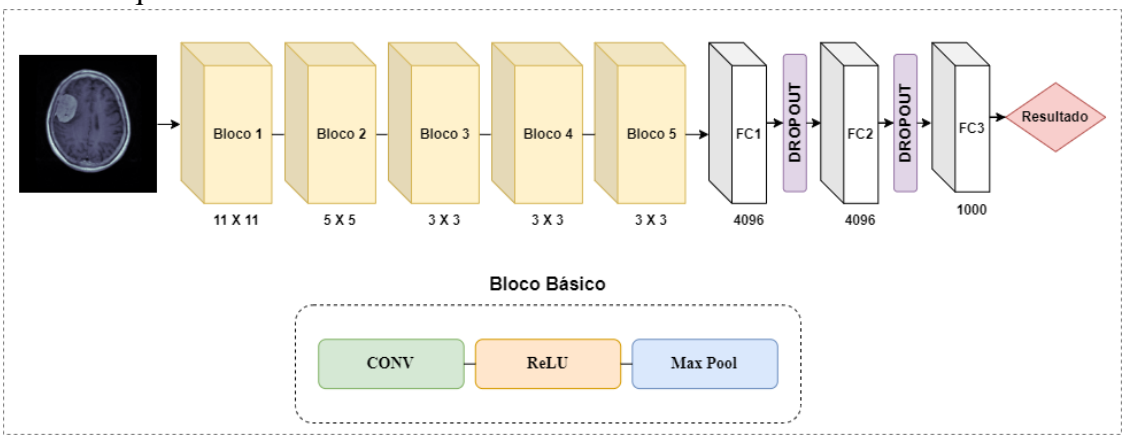
2.3.3.1 AlexNet

A AlexNet é uma clássica arquitetura de rede neural convolucional, desenvolvida por Krizhevsky *et al.* (2012). A rede consiste em camadas de convolução, camadas de *pooling* máximo e camadas densamente conectadas. Tais componentes atuam em conjunto como blocos de construção básicos, fundamentais para o aprendizado de características complexas durante o treinamento. Além disso, a rede AlexNet possui como proposta a sua implementação em duas Unidades de Processamento Gráfico (GPUs), o que permite uma aceleração significativa no treinamento do modelo, tornando viável o processamento de grandes quantidades de dados e o ajuste de milhões de parâmetros (KRIZHEVSKY *et al.*, 2012).

A Figura 7 apresenta o esquema da arquitetura da rede AlexNet. Nesse contexto, a

rede é composta por 60 milhões de parâmetros e 650.000 neurônios. A rede apresenta cinco camadas convolucionais, algumas das quais seguidas por camadas de *max-pooling* e três FC com um softmax final para realizar a classificação de 1.000 classes distintas. Para acelerar o treinamento são empregados neurônios não saturados e uma implementação de GPU da operação de convolução. Ademais, o método de regularização *dropout* é utilizado nas FC para mitigar o problema de *overfitting*. O *dropout* consiste em definir como zero a saída de cada neurônio oculto com uma probabilidade de 0,5. Os neurônios que são "dropados" não participam da retropropagação, forçando a rede a aprender características mais robustas que são eficazes em combinação com diversos subconjuntos aleatórios de outros neurônios (KRIZHEVSKY *et al.*, 2012).

Figura 7 – Arquitetura AlexNet.



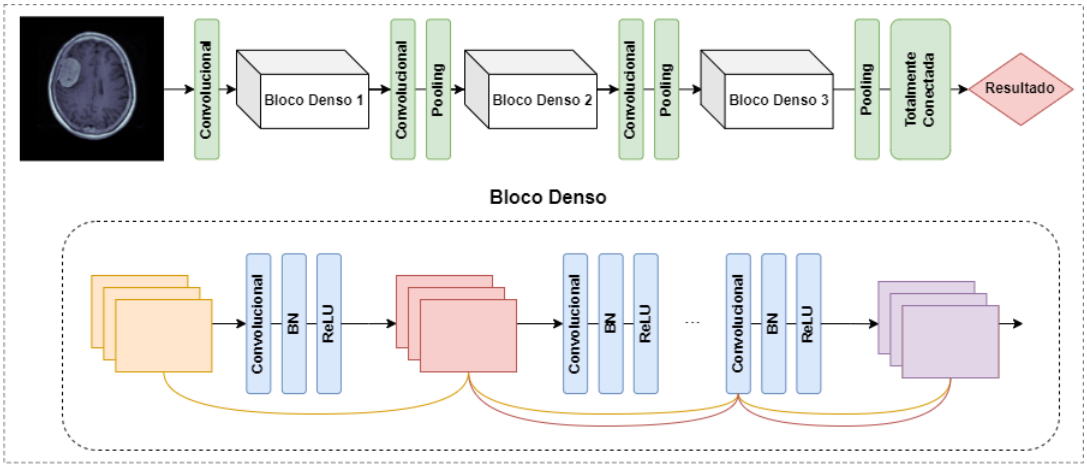
Fonte: Adaptado de (KRIZHEVSKY *et al.*, 2012).

2.3.3.2 DenseNet

A Rede Convolucional Densa (DenseNet) é uma arquitetura de *CNN* projetada para melhorar o fluxo de informações entre as camadas, por meio da introdução de conexões diretas de qualquer camada com todas as camadas subsequentes. Isto posto, a arquitetura melhora significativamente o problema de desaparecimento de gradientes, incentiva a reutilização de características e diminui substancialmente o número geral de parâmetros (HUANG *et al.*, 2018). A Figura 8 apresenta esquematicamente a arquitetura da rede DenseNet.

Dessa forma, uma rede convolucional tradicional que compreende L camadas, apresenta L conexões entre as camadas. Todavia, a rede DenseNet possui conexões diretas com todas as camadas subsequentes, o que resulta em $\frac{L(L+1)}{2}$ conexões. Consequentemente, uma camada l

Figura 8 – Arquitetura DenseNet.



Fonte: Adaptado de (HUANG *et al.*, 2018).

recebe os mapas de características de todas as camadas anteriores e realiza a concatenação antes da aplicação da transformação não linear $H_l(\cdot)$, conforme apresentada na equação:

$$x_l = H_l([x_0, x_1, \ldots, x_{\ell-1}]), \tag{2.1}$$

onde $[x_0, x_1, \ldots, x_{\ell-1}]$ indica à concatenação dos mapas de características produzidos nas camadas de 0 até $l-1$.

Logo, a função $H_l(\cdot)$ é uma função composta de três operações consecutivas, definidas como normalização em lote (BN), seguida por uma ReLU e uma convolução 3×3 (Conv). A união dessas três funções compõe o chamado Bloco Denso. Além disso, a rede é organizada em blocos densos e em camadas de transição. As camas de transição consistem em camadas de convolução 1×1 e de *pooling* 2×2, responsáveis pela redução da amostragem de camadas, que alteram o tamanho dos mapas de características.

Nesse contexto, existem variantes da rede DenseNet tradicional, como, Densenet121, DenseNet169 e DenseNet20. A principal diferença entre essas variantes é o número de camadas que compõem cada arquitetura, o que impacta diretamente na capacidade de aprendizado da rede. Logo, a DenseNet121 é uma rede mais leve, eficaz para aplicações com os recursos computacionais limitados. Já as redes DenseNet169 e DenseNet201, por possuírem mais camadas, são capazes de capturar informações mais complexas dos dados, entretanto, exigem mais memória e maior poder de processamento, sendo mais apropriadas para cenários com dados abundantes e necessidade de alta performance.

2.3.3.3 EfficientNet

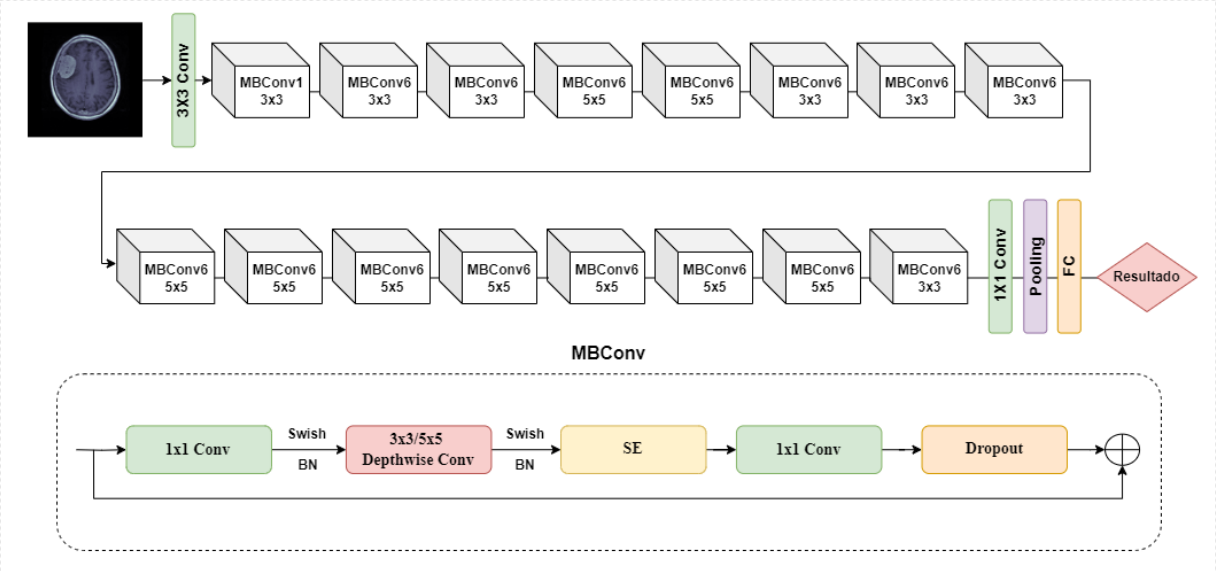
A EfficientNet, desenvolvida por Tan (2019), engloba uma família de arquiteturas de *CNN*s, projetadas para serem altamente eficientes em termos de consumo de recursos computacionais. Ademais, a arquitetura EfficientNet escala uniformemente as dimensões de profundidade, largura e resolução da rede com um conjunto de coeficientes de escala fixos. A profundidade refere-se ao número de camadas, a largura ao número de canais em cada camada e a resolução ao tamanho das imagens de entrada. O método de escala composta é utilizado com o intuito de que se a imagem de entrada for maior, a rede precisará de mais camadas para aumentar o campo receptivo e mais canais para capturar padrões mais refinados na imagem (TAN, 2019).

A Figura 9 apresenta a estrutura da arquitetura da rede EfficientNet. A EfficientNet emprega blocos MBConv (*Mobile Inverted Bottleneck Convolution*), inspirados na MobileNet. A rede utiliza convoluções separáveis em profundidade e convoluções 1x1 para reduzir o número de operações necessárias, enquanto preservam a capacidade de aprendizado da rede. Além disso, a arquitetura incorpora a função de ativação *Swish*, que melhora o desempenho em tarefas de aprendizado profundo em comparação com a ReLU tradicional, ao oferecer uma função diferenciável em todos os pontos. A rede também adiciona a normalização em lotes (BN), que consiste em uma técnica usada para normalizar as ativações em uma camada, ajudando a estabilizar e acelerar o treinamento de redes neurais. O módulo *Squeeze and Excitation* (SE) permite que a rede aprenda a ponderar a importância de diferentes canais de entrada de acordo com o contexto da tarefa. Por fim, o *dropout* é utilizado para mitigar o *overfitting* durante o treinamento. A combinação desses elementos torna a EfficientNet uma arquitetura poderosa e eficiente, adequada para uma ampla variedade de aplicações de visão computacional (TAN, 2019).

2.3.3.4 GoogLeNet

A GoogLeNet, proposta por Szegedy *et al.* (2015), é uma arquitetura baseada no modelo Inception, que introduziu uma abordagem inovadora ao permitir que a rede escolha entre diferentes tamanhos de filtros convolucionais em cada bloco, maximizando a eficiência computacional. A rede foi projetada para ser altamente prática e eficiente, permitindo que a inferência seja executada em dispositivos individuais, incluindo até mesmo aqueles com recursos computacionais limitados e com baixa memória. A estrutura da GoogLeNet é composta por

Figura 9 – Arquitetura EfficientNet.



Fonte: Adaptado de (TAN, 2019).

22 camadas de profundidade com uma entrada de tamanho 224×224. No total, a rede utiliza aproximadamente 100 blocos de construção independentes. Além disso, é utilizado o conceito de pooling médio antes das FC (SZEGEDY *et al.*, 2015).

Desse modo, devido à profundidade relativamente grande da rede, a capacidade de propagar gradientes de volta por todas as camadas de maneira eficaz tornou-se uma preocupação considerável. Para lidar com a propagação eficiente dos gradientes através da rede profunda, foram introduzidos classificadores auxiliares conectados a camadas intermediárias. Durante o treinamento, a perda desses classificadores auxiliares é adicionada à perda total da rede, com um peso de desconto, o que melhora a propagação dos gradientes. No entanto, esses classificadores são descartados durante a inferência. Especificamente, a rede auxiliar inclui (SZEGEDY *et al.*, 2015):

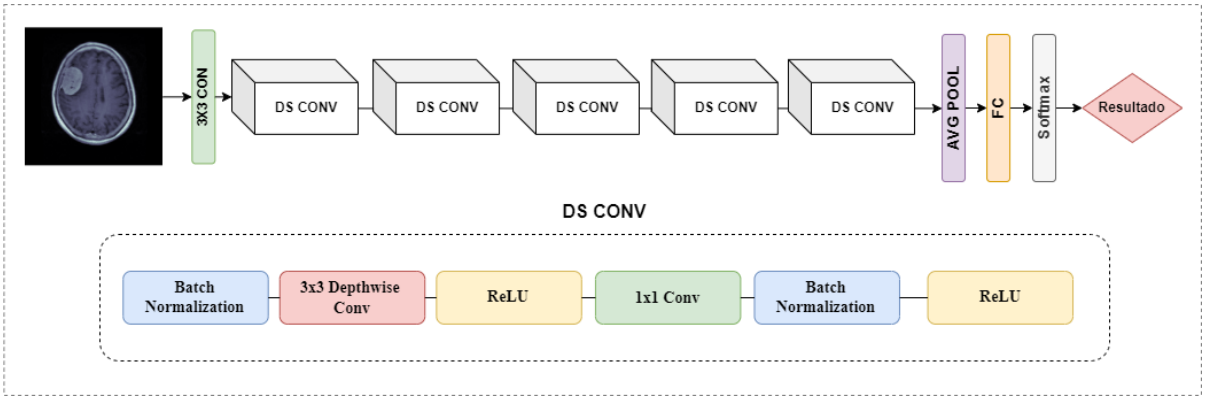
- Uma camada de pooling médio com filtro 5×5 e passo 3;
- Uma convolução 1×1 com 128 filtros para redução de dimensão e ativação linear retificada;
- Uma camada totalmente conectada com 1024 unidades e ativação linear retificada;
- Uma camada de abandono com 70% de taxa de saídas descartadas;
- Uma camada linear com perda softmax para classificação.

2.3.3.5 *MobileNet*

A classe de modelos MobileNets, introduzida por Howard *et al.* (2017), foi proposta como uma arquitetura de redes neurais profundas simplificada e leve, com foco em reduzir o custo computacional e o tamanho do modelo, por meio do uso de convoluções separáveis em profundidade. Este método consiste em aplicar um filtro em cada canal de entrada e é composto de duas etapas: convoluções em profundidade e convoluções pontuais. Na primeira etapa, um filtro é aplicado a cada canal de entrada individualmente, o que reduz significativamente a complexidade computacional. Enquanto, na convolução pontual, uma operação simples 1×1, é usada para combinar linearmente as saídas das convoluções em profundidade, o que resulta em uma representação compacta e eficiente dos dados (HOWARD *et al.*, 2017).

A estrutura do MobileNet é construída inteiramente com base em convoluções separáveis em profundidade. A Figura 10 ilustra a arquitetura da MobileNet, composta por 28 camadas. A arquitetura da rede inclui convoluções regulares, camadas de normalização em lotes (*batch normalization*), função de ativação ReLU, convoluções em profundidade e convoluções pontuais 1 × 1. Cada camada de convolução é seguida por uma normalização em lotes e por uma função de ativação ReLU, com exceção da camada final totalmente conectada que não possui a não linearidade e alimenta a camada *softmax* para classificação. Além disso, uma camada de *pooling* médio é aplicada para reduzir a resolução espacial antes da camada totalmente conectada, o que contribui para a eficiência do modelo (HOWARD *et al.*, 2017).

Figura 10 – Arquitetura MobileNet.



Fonte: Adaptado de (HOWARD *et al.*, 2017).

Desse modo, as versões da MobileNet se diferenciam pelas suas abordagens para otimização e eficiência. A MobileNetV2, introduzida por Sandler *et al.* (2018), faz uso do bloco de convolução invertida com conexões residuais e expansão linear, o que permite uma

maior eficiência computacional e preservação de informações. Já a MobileNetV3, proposta por Howard *et al.* (2019), utiliza uma combinação de técnicas avançadas de automação de design de rede neural (NAS) e otimizações manuais para alcançar uma maior eficiência em termos de velocidade e precisão.

2.3.3.6 ResNet

As Redes Residuais (ResNet), introduzidas por He *et al.* (2016), são uma estrutura de aprendizado residual elaboradas para facilitar o treinamento de redes que são substancialmente mais profundas, mitigando problemas causados pelo crescimento da profundidade da rede, como saturação da acurácia e aumento do erro de treinamento. Para solucionar o problema da degradação, a arquitetura ResNet introduz uma estrutura de blocos residuais. Em uma rede neural tradicional, as camadas são empilhadas e se encaixam diretamente em um mapeamento subjacente desejado. Já na ResNet, as camadas se encaixam em um mapeamento residual. A saída de um bloco residual $H(x)$ pode ser representada por:

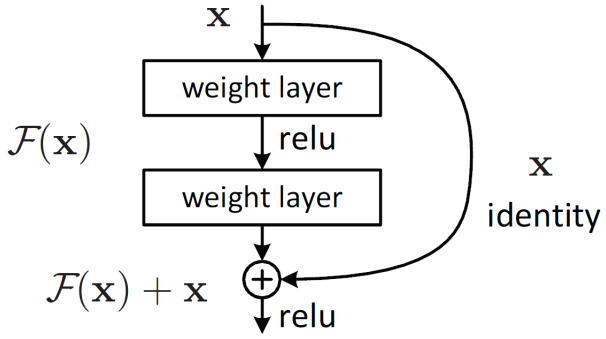
$$H(x) = F(x) + x, \tag{2.2}$$

onde $F(x)$ representa o mapeamento residual aprendido pela rede e o termo de identidade x permite que o gradiente flua mais facilmente.

A Figura 11 apresenta a representação de um bloco residual, composto por uma sequência de camadas de camadas de convolução, seguidas por operações de normalização e ativação, como a ReLU. A principal característica dos blocos residuais é a conexão de atalho (*skip connection*), que permite que a entrada original do bloco seja somada à sua saída antes de ser passada para a próxima camada, evitando a degradação do desempenho em redes neurais profundas. Além disso, a ResNet utiliza uma camada chamada *Global Average Pooling (GAP)* para reduzir a dimensionalidade dos dados antes da camada final de classificação (HE *et al.*, 2016).

Por fim, as variações da arquitetura ResNet, como a ResNet18, ResNet50 e ResNet101, diferem pelo número de camadas e blocos residuais, possuindo 18, 50 e 101 camadas, respectivamente, o que confere maior capacidade de representação e aprendizado nas redes mais profundas.

Figura 11 – Arquitetura ResNet.



Fonte: Adaptado de (HE *et al.*, 2016).

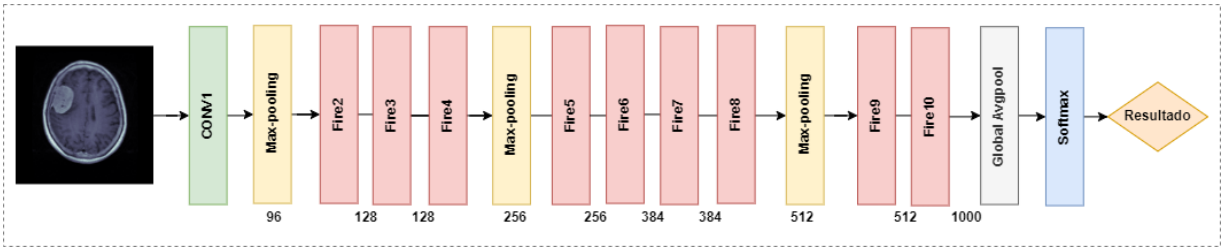
2.3.3.7 SqueezeNet

A SqueezeNet, introduzida por Iandola *et al.* (2016), foi proposta como uma pequena arquitetura de rede neural convolucional. A rede é capaz de garantir alta precisão em tarefas de classificação, ao mesmo tempo que utiliza menos parâmetros, tornando-se mais viável para implementações em dispositivos com recursos de memória limitados.

Para garantir a precisão e eficiência com a baixa quantidade de parâmetros aplicados, a SqueezeNet aborda três estratégias principais. Primeiramente, os tradicionais filtros de convolução 3 x 3 são substituídos por filtros 1 x 1, que são menos exigentes em termos de parâmetros. Em seguida, são reduzidos o número de canais de entrada nas camadas convolucionais por meio de camadas de compressão, que alimentam camadas de expansão com uma combinação de filtros 1 x 1 e 3 x 3. Essas duas estratégias permitem a redução criteriosa do número de parâmetros, enquanto preserva a precisão da rede. Ademais, para maximizar a precisão da SqueezeNet, é aplicada uma abordagem para atrasar a aplicação da subamostragem na rede, permitindo que as camadas convolucionais operem em mapas de ativação maiores (IANDOLA *et al.*, 2016).

A Figura 12 apresenta a arquitetura da SqueezeNet. A rede inicia com uma camada de convolução autônoma (conv1), seguida por oito módulos *Fire*. O módulo *Fire* é composto por uma uma camada de convolução *squeeze*, que utiliza exclusivamente filtros 1 x 1, e uma camada de expansão, que combina filtros de convolução 1 x 1 e 3 x 3. Além disso, a SqueezeNet aplica a operação de *max-pooling* com um passo de 2, otimizando assim a precisão em um orçamento limitado de parâmetros disponíveis (IANDOLA *et al.*, 2016).

Figura 12 – Arquitetura SqueezeNet.

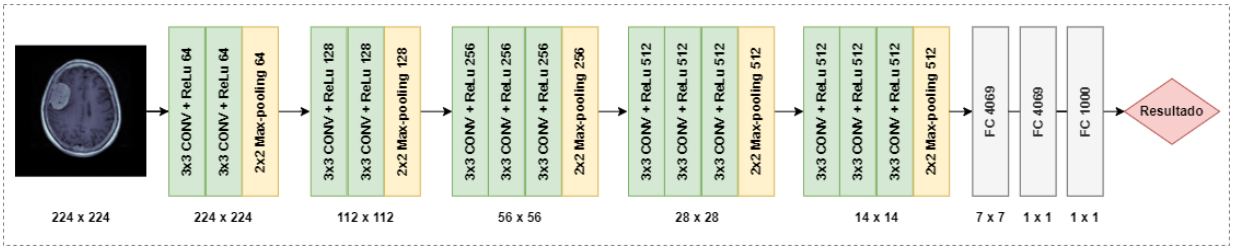


Fonte: Adaptado de (IANDOLA *et al.*, 2016).

2.3.3.8 VGG

Introduzidas por Simonyan e Zisserman (2014), a Visual Geometry Group (VGG) é uma clássica arquitetura de redes neurais convolucionais. A rede se baseia em uma análise do impacto da profundidade da rede convolucional para o reconhecimento de imagens de larga escala. O principal foco da arquitetura VGG é a profundidade crescente de camadas de peso, enquanto utiliza camadas convolucionais simples e uniformes, com pequenos filtros de convolução (3x3). As variantes da arquitetura VGG diferenciam-se no número de camadas totais, sendo a VGG-16 e a VGG-19 compostas por 16 e 19 camadas convolucionais, respectivamente. A imagem 13 apresenta a arquitetura da rede VGG.

Figura 13 – Arquitetura VGGNet.



Fonte: Elaborado pela autora.

Durante o treinamento, a rede VGGNet recebe como entrada imagens RGB de tamanho fixo de 224 x 224 pixels. A imagem passa por uma sequência de camadas convolucionais, com filtros 3 x 3 e passo de convolução fixado em 1 pixel, o que permite capturar com precisão a orientação espacial, como esquerda, direita, cima, baixo e centro. Os filtros de convolução são usados para transformar a entrada linearmente. Além disso, todas as camadas ocultas são equipadas com a função de ativação ReLU, que adiciona o componente de não-linearidade ao modelo, para o aprendizado de padrões complexos nos dados (SIMONYAN; ZISSERMAN, 2014).

Ademais, a rede inclui camadas de *max-pooling* que acompanham algumas camadas convolucionais. O *max-pooling* é realizado por um filtro de 2 x 2 pixels com passo 2, o que reduz a dimensionalidade espacial da entrada. Desse modo, a arquitetura VGGNet possui uma pilha de camadas convolucionais, na qual a profundidade da rede varia conforme a versão da rede. Por fim, os blocos convolucionais são seguidos por três FC. As duas primeiras camadas são compostas por 4096 canais cada, enquanto a última camada, responsável por realizar a classificação, possui 1000 canais, correspondentes às classes de saída (SIMONYAN; ZISSERMAN, 2014).

2.4 Segmentação Semântica

A segmentação semântica consiste em atribuir um rótulo de categoria a cada pixel de uma imagem, sendo uma tarefa fundamental, porém desafiadora, no campo da visão computacional. Em suma, o objetivo da segmentação semântica é dividir uma imagem em subconjuntos mutuamente exclusivos, onde cada subconjunto representa uma região significativa da imagem original. Nesse contexto, diversos métodos têm alcançado resultados promissores utilizando redes neurais profundas. Geralmente, ao fornecer um número suficiente de imagens e seus respectivos mapas de rótulos como dados de treinamento, uma rede neural profunda é treinada para aprender a correspondência entre um rótulo semântico e suas variadas aparências visuais (HAO *et al.*, 2020). Logo, no contexto de segmentação semântica, as arquiteturas do tipo *encoder-decoder* desempenham um papel fundamental, permitindo que a rede aprenda representações complexas da imagem.

Dessa forma, a arquitetura *encoder-decoder* é amplamente utilizada para o aprendizado de sequência para sequência. A arquitetura é composta por dois módulos principais, denominados, como, codificador (*encoder*) e decodificador (*decoder*). O codificador é responsável por processar a sequência de entrada, aprendendo e extraindo características relevantes ao codificar os aspectos essenciais da imagem. Durante esse processo, o codificador transforma a imagem em um conjunto de vetores de contexto que capturam as informações mais importantes em relação aos rótulos verdadeiros correspondentes. Esses vetores de contexto são então utilizados pelo decodificador, que reconstrói a saída desejada a partir das características codificadas, permitindo a geração de uma sequência de saída precisa e informada (KUSAKUNNIRAN *et al.*, 2023).

Ademais, o módulo decodificador é responsável por transformar a representação codificada de volta à sua forma original, realizando a reconstrução da imagem segmentada. Esse

processo envolve mapear os recursos extraídos pelo codificador, aplicando operações como convolução separável, normalização de lote, ReLU e *upsampling*. Além disso, o decodificador utiliza os índices de *max-pooling* memorizados dos mapas de recursos do codificador correspondente para aprimorar a precisão da reconstrução, integrando informações da saída anterior e do bloco codificador correspondente (BADRINARAYANAN *et al.*, 2017; KUSAKUNNIRAN *et al.*, 2023). Logo, existem diversos modelos decodificadores, por exemplo, UNet, UNet++ e FPN, utilizados neste trabalho.

2.4.0.1 U-Net

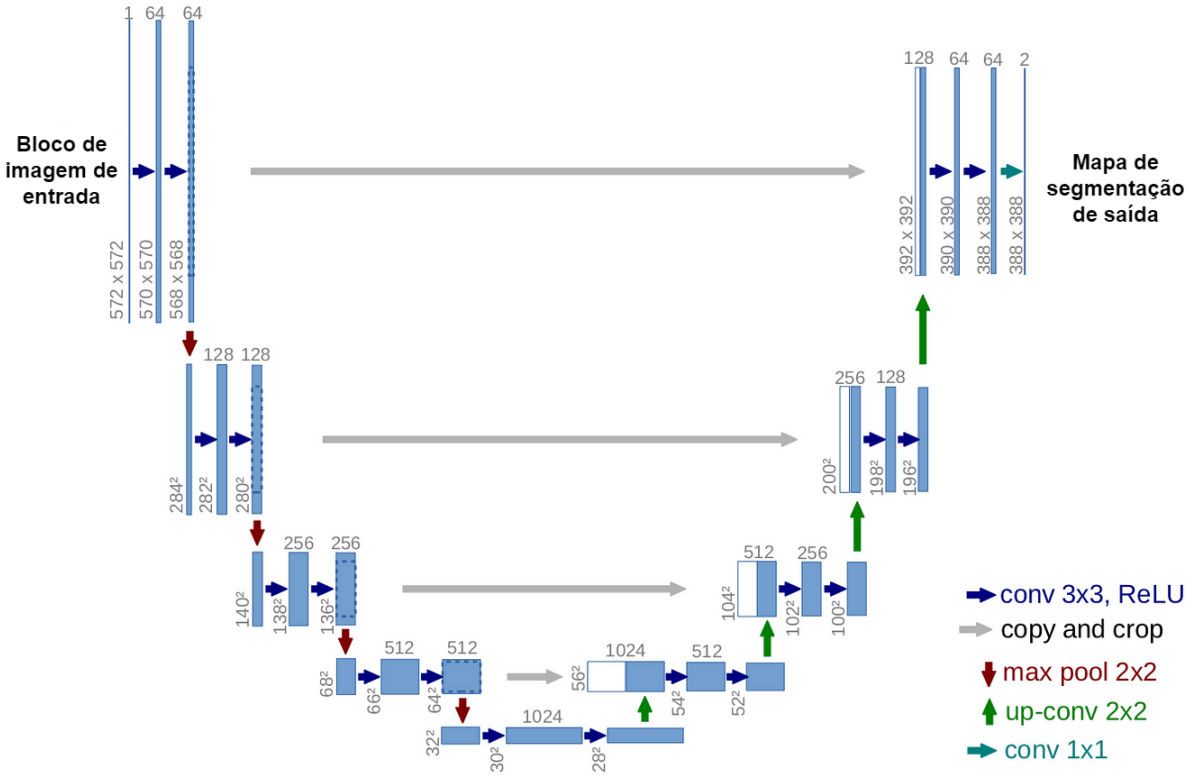
A U-Net é uma arquitetura de aprendizado profundo desenvolvida por Ronneberger *et al.* (2015), foi projetada especificamente para realizar a segmentação de imagens biomédicas. Além disso, a rede destaca-se pela sua flexibilidade e pela capacidade de produzir previsões precisas e detalhadas, mesmo com a quantidade limitada de dados de treinamento. A estrutura da arquitetura U-Net consiste em um caminho de contração (*encoder*) e em um caminho de expansão (*decoder*) (RONNEBERGER *et al.*, 2015).

O caminho de contração da U-Net apresenta camadas codificadoras capazes de extrair informações da imagem e reduzir a resolução espacial da entrada. Nesse processo, são aplicadas camadas convolucionais seguidas de funções de ativação ReLU e operações de *pooling*. As camadas convolucionais capturam as características locais importantes, enquanto o *pooling* reduz a dimensão espacial das características, ampliando o campo de visão da rede. Ademais, a estrutura de expansão realiza a decodificação dos dados, após a etapa de codificação. O principal objetivo é recuperar a resolução original da imagem e refinar a segmentação. As camadas decodificadoras, também conhecidas como camadas de *upsampling*, aumentam a resolução dos mapas de características enquanto aplicam operações convolucionais para melhorar a precisão dos detalhes (RONNEBERGER *et al.*, 2015; WANG *et al.*, 2023).

A Figura 14 ilustra a arquitetura da rede U-Net, composta por um caminho de contração (lado esquerdo) e por um caminho de expansão (lado direito). A etapa de contração consiste em uma rede convolucional tradicional, onde são aplicadas duas camadas de convolução 3 x 3, seguidas por uma uma ReLU e uma operação de *pooling* máximo 2 x 2 para *downsampling*. Já no caminho de expansão, são aplicadas camadas *upsampling* para aumentar a resolução dos mapas de características, seguidas por uma convolução 2 x 2 e uma concatenação com o mapa de características correspondente. Em seguida, são aplicadas duas convoluções 3 x 3, cada

uma acompanhada por uma função de ativação ReLU. Na camada final, uma convolução 1 x 1 é usada para mapear cada vetor de recursos de 64 componentes para o número desejado de classes (RONNEBERGER *et al.*, 2015).

Figura 14 – Arquitetura U-Net.



Fonte: Adaptado de (RONNEBERGER *et al.*, 2015).

2.4.0.2 UNet++

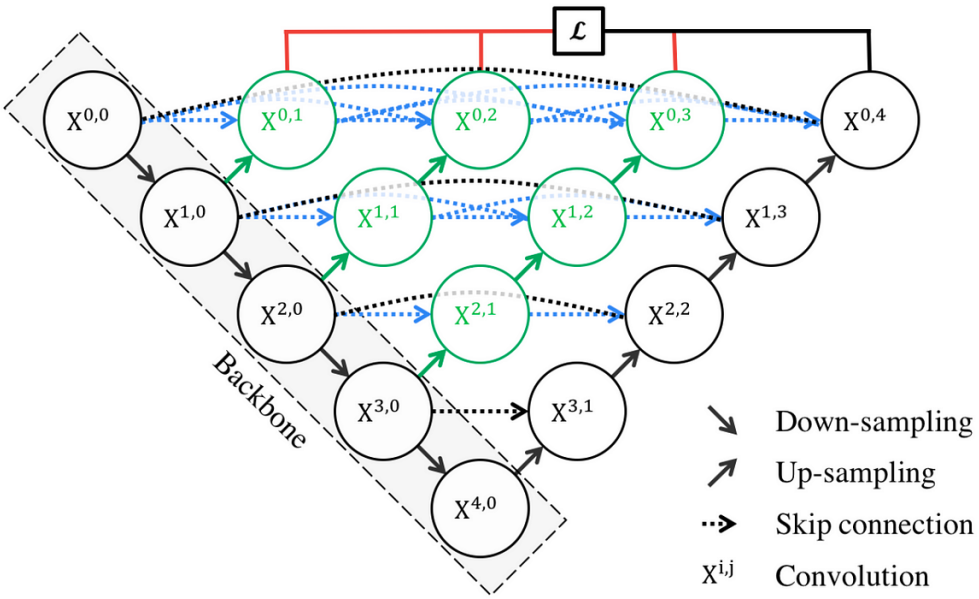
A arquitetura UNet++, desenvolvida por Zhou *et al.* (2018), é uma evolução da arquitetura U-Net, projetada como uma poderosa arquitetura para segmentação de imagens médicas. Desse modo, enquanto a U-Net tradicional já se destaca por sua capacidade de segmentar imagens com alta precisão, a UNet++ introduz melhorias significativas para abordar alguns dos desafios presentes na arquitetura tradicional. A UNet++ é composta de uma rede codificadora-decodificadora profundamente supervisionada, onde as sub-redes codificadora e decodificadora são conectadas por meio de uma série de caminhos de salto densos e aninhados, que conectam essas duas sub-redes de maneira mais eficaz (ZHOU *et al.*, 2018).

Na UNet tradicional os mapas de características do codificador são recebidos dire-

tamente no decodificador por meio de conexões de salto simples, o que pode resultar em uma lacuna semântica entre as informações do codificador e do decodificador, o que pode dificultar a reconstrução de detalhes refinados na segmentação. Por outro lado, a UNet++ resolve esse problema redesenhando essas conexões de salto. Nesse contexto, os mapas de características são alimentados por um bloco de convolução denso antes de serem recebidos no decodificador. Isso facilita a aproximação do nível semântico dos mapas de características do codificador ao do decodificador, o que favorece o aprendizado e resulta em segmentações mais precisas (ZHOU *et al.*, 2018).

A Figura 15 apresenta a arquitetura da UNet++, composta por caminhos de salto aninhados e uma supervisão profunda, que podem ser identificadas em azul/verde e vermelho, respectivamente. Enquanto, os componentes U-Net originais são mantidos em preto. Desse modo, os caminhos de salto aninhados são conexões que passam por blocos de convolução densos antes de conectar o codificador ao decodificador. Esses blocos convolucionais adicionais ajudam a reduzir a lacuna semântica, o que permite o recebimento de informações mais refinadas para a reconstrução da imagem. Além disso, a supervisão profunda consiste em uma técnica em que a rede é treinada considerando várias saídas intermediárias, não apenas a final. Essa abordagem fornece uma forma de regularização ao modelo, o que auxilia na obtenção de resultados mais precisos e robustos durante o processo de segmentação (ZHOU *et al.*, 2018).

Figura 15 – Arquitetura UNet++.



Fonte: (ZHOU *et al.*, 2018).

2.4.0.3 FPN

O decodificador *Feature Pyramid Network* (FPN), desenvolvido por Lin *et al.* (2017), emprega uma arquitetura assimétrica com um caminho de baixo para cima (*bottom-up*) e de cima para baixo (*top-down*) para tarefas de detecção e segmentação de objetos. Na fase inicial do caminho *bottom-up*, a imagem de entrada passa por múltiplas camadas convolucionais, extraindo recursos de baixo nível, como bordas e texturas, que são codificados em mapas de recursos em várias escalas. Esses mapas são combinados através de conexões laterais com camadas convolucionais 1x1, preservando a consistência semântica entre diferentes escalas, o que é crucial para a precisão na detecção e segmentação. O caminho *top-down* melhora a resolução espacial dos mapas de recursos gerados no processo *bottom-up*, realizando upsampling dos mapas para aumentar sua resolução espacial, o que aprimora a localização de objetos. As conexões laterais mesclam os recursos dos dois caminhos, e convoluções 3x3 são aplicadas para evitar efeitos de aliasing. A arquitetura FPN, com sua capacidade de gerar mapas de recursos em várias resoluções e com rica semântica, é essencial para tarefas como segmentação semântica e detecção de objetos, assegurando tanto precisão espacial quanto semântica (LIN *et al.*, 2017; SHAREN *et al.*, 2024).

2.5 Pré-processamento

O pré-processamento de imagens consiste em transformar os dados de imagem brutos em dados de imagem limpos, com objetivo de aprimorar a qualidade da imagem e reduzir artefatos indesejados (BOW, 2002). As imagens brutas coletadas de centros de digitalização podem conter ruídos indesejados. Dessa forma, o pré-processamento é uma importante etapa para análise de imagens de ressonância magnética, que visa, a remoção de artefatos e o aprimoramento das imagens. O pré-processamento envolve o redimensionamento de imagem, a conversão das imagens para escala de cinza, a remoção de ruído e a melhora da qualidade e nitidez, para produzir uma imagem na qual as minúcias podem ser detectadas corretamente (PERUMAL; VELMURUGAN, 2018).

2.5.1 Redimensionamento

O redimensionamento de imagens é uma técnica comumente utilizada para garantir a compatibilidade de exibição em diferentes dispositivos e melhorar a eficiência dos algoritmos,

especialmente em função da resolução espacial relativamente pequena. Além disso, o redimensionamento é crucial para assegurar que as imagens de entrada estejam compatíveis com as redes neurais durante o treinamento. Existem diversas abordagens para redimensionamento, como corte, dimensionamento e deformação, entre outros métodos. Nesse contexto, o dimensionamento, em particular, é definido por um mapeamento homogêneo entre os pixels da imagem original e os pixels da imagem alvo, sendo a interpolação dos pixels da imagem original a técnica mais comum para realizar esse ajuste (LIN *et al.*, 2014; TALEBI; MILANFAR, 2021).

2.5.2 Escala de Cinza

A escala de cinza em imagens digitais consiste que o valor de cada pixel representa apenas a informação de intensidade da luz, ou seja, cada pixel recebe um nível de tons de cinza que apresenta a luminância do pixel. Em suma, a imagem contém apenas as cores preto, branco e cinza, nas quais o cinza tem vários níveis. Em imagens em escala de cinza , o valor de cada pixel está relacionado ao número de bits de dados usados para representar o pixel. O valor da imagem cinza é geralmente representado por 8 bits, ou seja, a combinação de oito números binários representa o valor de pixel de um pixel (TAN; JIANG, 2013; LIU, 2020). Logo, a transformação de imagens em escala de cinza serve para simplificar a análise visual e o processamento de imagens.

2.5.3 Binarização

O método de binarização ou limiarização é uma das questões fundamentais no processamento digital de imagens. A técnica de binarização de imagens é comumente utilizada para segmentar regiões com homogeneidade diferente em imagens em tons de cinza. Em suma, a binarização realiza o agrupamento com base nos níveis de intensidade de pixels em um histograma de imagem. O processo consiste em transformar uma imagem colorida ou em escala de cinza em uma imagem binária, ou seja, com duas cores, preto e branco (DU; HE, 2023).

2.5.4 Filtragem bilateral

A filtragem bilateral, introduzida por Tomasi e Manduchi (1998), é uma técnica de processamento de imagem usada para suavizar imagens, reduzir o ruído indesejado e preservar a nitidez das bordas. A filtragem bilateral leva em consideração tanto a proximidade espacial

quanto a similaridade de intensidade dos pixels. Nesse contexto, ao deslocar-se pela imagem, o filtro substitui cada pixel pela média ponderada dos pixels vizinhos, onde os pesos são determinados pela distância espacial e pela diferença de intensidade entre os pixels. Dessa forma, o filtro bilateral padrão é o produto de uma máscara gaussiana espacial, que considera a proximidade entre os pixels, e uma máscara gaussiana de alcance, que leva em conta as diferenças de intensidade (JEME; JEROME, 2023).

2.5.5 Equalização de Histograma Adaptável Limitado por Contraste

A Equalização de Histograma Adaptável Limitado por Contraste (CLAHE) é uma técnica de processamento de imagem desenvolvida, desenvolvida por Zuiderveld (1994), para aumentar o contraste de imagens de forma adaptativa. O método consiste em dividir a imagem em várias regiões não sobrepostas, onde cada uma passa por uma equalização de histograma individual, limitando o contraste para evitar amplificação de ruídos em áreas homogêneas. Após o cálculo dos histogramas para cada região, estes são redistribuídos com base em um limite de corte, e as funções de distribuição cumulativa (CDF) são usadas para mapear os pixels em tons de cinza. Dessa forma, a técnica CLAHE aplica uma combinação linear dos mapeamentos das quatro regiões mais próximas para melhorar o contraste sem introduzir artefatos indesejados, principalmente, nas regiões das bordas e nos cantos da imagem (REZA, 2004; MUSA *et al.*, 2018).

2.6 Aumento de Dados

O Aumento de Dados (AD) ou *data augmentation* é uma técnica comumente utilizada para expandir o tamanho e a variabilidade de um conjunto de dados de treinamento, melhorar a capacidade de generalização dos modelos e mitigar problemas de *overfitting* (CHLAP *et al.*, 2021). Arquiteturas de redes neurais exigem um grande volume de dados para realizar um treinamento efetivo e alcançar resultados precisos. No entanto, os dados, especialmente em análise de imagens médicas, são frequentemente escassos ou custosos de coletar. Portanto, o principal objetivo das técnicas de AD é gerar novas amostras de dados e mitigar a escassez de dados rotulados (TAYLOR; NITSCHKE, 2017; CHLAP *et al.*, 2021).

O aumento de dados consiste em aplicar transformações no conjunto de imagens, seja transformações geométricas ou fotométricas. As transformações geométricas são responsáveis

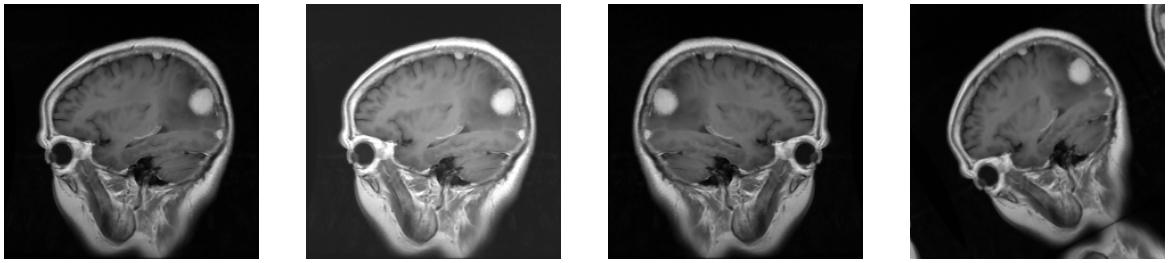
por alterar a geometria da imagem por meio do mapeamento dos pixels individuais para novas posições. São exemplos de transformações geométricas a rotação, giro, zoom e deslocamento. Por outro lado, as transformações fotométricas consistem na aplicação de efeitos como ruído, brilho e saturação (TAYLOR; NITSCHKE, 2017).

Logo, é possível gerar conjunto de dados expandido por meio das transformações aplicadas nas imagens de entrada:

- **Brilho e Contraste**: Consiste em realizar modificações aleatórias no brilho e contraste da imagem;
- **Inversão**: Aplica uma inversão horizontal ou vertical na imagem;
- **Deslocamentos, escalas e rotações**: Mudanças na posição, escala e rotação das imagens.

A Figura 16 apresenta exemplos da aplicação de transformações de imagens para a realização da técnica de aumento de dados. Na Figura 16(a) é apresentada a imagem original de entrada. Em seguida, a Figura 16(b) exibe alterações no brilho e contraste da imagem. A Figura 16(c) apresenta um exemplo de inversão horizontal na imagem de entrada. Por fim, na Figura 16(d) são aplicadas técnicas de deslocamento, alteração de escala e rotação na imagem.

Figura 16 – Aplicação de transformações em imagens.



(a) Original. (b) Brilho e Contraste. (c) Inversão Horizontal. (d) Deslocamento, escala e rotação.

Fonte: Elaborado pela autora.

2.7 *Grid Search*

O desempenho de modelos de aprendizado de máquina é determinado em grande parte pela escolha apropriada dos hiperparâmetros associados aos algoritmos. Os hiperparâmetros são variáveis externas ao modelo e representam valores no processo de treinamento, como a taxa de aprendizado, otimizador, número de camadas ocultas e tamanho de cada camada (MARINOV; KARAPETYAN, 2019).

O *Grid Search* é uma abordagem sistemática que consiste em buscar os melhores conjuntos de hiperparâmetros no espaço de pesquisa e em criar todas as combinações possíveis. Embora a técnica de ajuste de hiperparâmetros forneça garantias, como, a maximização da eficácia dos modelos, também apresenta desvantagens significativas. Por exemplo, para uma otimização com um grande número de hiperparâmetros, são necessárias inúmeras combinações, o que resulta em em um volume extenso de cálculos e tempo gasto (MARINOV; KARAPETYAN, 2019; ALIBRAHIM; LUDWIG, 2021).

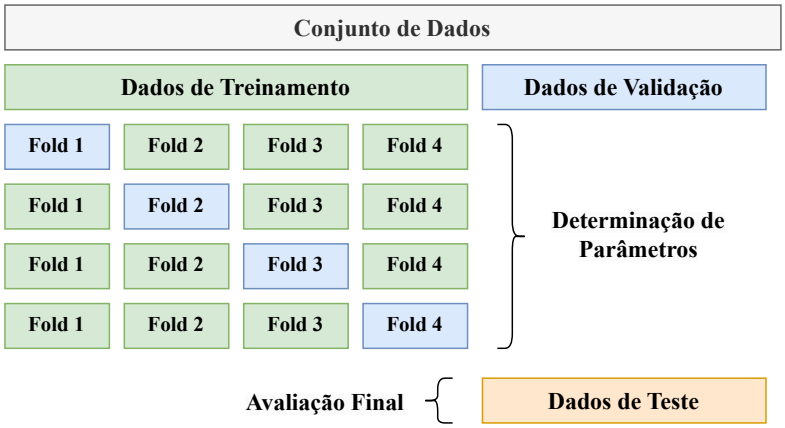
2.8 Validação Cruzada

A validação cruzada é um método de reamostragem de dados que consiste em particionar o conjunto de dados em subconjuntos de treinamento e validação e estimar a precisão preditiva dos dados de validação pelo modelo obtido a partir do conjunto de dados de treinamento. O treinamento de dados é usado para criar o modelo, enquanto os dados de teste são usados para encontrar a precisão dos modelos (WANG; CHAOVALITWONGSE, 2011; ROSHINTA *et al.*, 2023).

Dessa forma, a validação cruzada é uma técnica comumente utilizada para avaliar a capacidade de generalização de modelos e evitar o *overfitting*. Problemas de *overfitting* ocorrem quando um modelo se ajusta muito bem ao conjunto de dados de treinamento, mas não se ajusta aos dados de validação. Em suma, a validação cruzada consiste em reduzir a variabilidade da avaliação do modelo, por meio da realização de testes em diferentes partições do conjunto de dados original. Logo, o resultado da avaliação de um modelo preditivo é a média dos testes realizados (WANG; CHAOVALITWONGSE, 2011; BERRAR, 2018).

A validação cruzada *K-fold* é uma abordagem comumente utilizada para realizar múltiplas etapas de validação cruzada. O método consiste em particionar aleatoriamente o conjunto de dados em K subconjuntos, denominadas *folds*. A validação cruzada é repetida K vezes, na qual, cada vez um dos subconjuntos é reservado como dados de validação, e os subconjuntos $K-1$ restantes são os conjuntos de dados de treinamento. O resultado da validação é a média dos K resultados. A abordagem garante que cada observação é utilizada para validação exatamente uma vez (WANG; CHAOVALITWONGSE, 2011; BERRAR, 2018; ROSHINTA *et al.*, 2023). A Figura 17 ilustra o funcionamento da validação cruzada *K-fold* para $K = 4$, em que os dados são divididos em 4 *folds*.

Figura 17 – Validação Cruzada *K-fold* para $K = 4$.



Fonte: Adaptado de (SCIKIT-LEARN, 2007 - 2022).

2.9 Métricas de Avaliação

Para uma avaliação precisa do desempenho dos algoritmos de *DL*, é essencial empregar métricas de avaliação apropriadas. Dessa forma, um conjunto de métricas são utilizadas para determinar se o modelo está de fato atendendo aos requisitos estabelecidos.

2.9.1 Métricas para Classificação

Tarefas de classificação são fundamentais na categorização dos dados em diferentes classes. A classificação binária indica que uma observação pode pertencer a um de duas classes possíveis. Enquanto, as tarefas de classificação multiclasse abrangem a categorização em mais de duas classes (KOLO, 2011).

2.9.1.1 Matriz de Confusão

A matriz de confusão é uma ferramenta poderosa para avaliação de desempenho em tarefas de classificação. Trata-se de uma tabela cruzada, onde suas linhas registram o número de ocorrências para a classe real e as colunas apresentam os rótulos previstos por um classificador (TING, 2010; HEYDARIAN *et al.*, 2022).

O principal objetivo da matriz de confusão é visualizar e analisar a distribuição de rótulos previstos correta e incorretamente. Além disso, é utilizada para calcular os valores de Verdadeiro Positivo (VP) Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN), essenciais para o cálculo das principais métricas de avaliação (GRANDINI *et al.*, 2020; HEYDARIAN *et al.*, 2022):

- **Verdadeiro Positivo (VP)**: Número de exemplos corretamente identificados como pertencentes a uma determinada categoria pelo modelo;
- **Verdadeiro Negativo (VN)**: Número de exemplos identificados corretamente como não pertencentes a uma determinada categoria em análise;
- **Falso Positivo (FP)**: Número de exemplos erroneamente identificados como pertencentes a uma determinada classe, ou seja, são os elementos classificados incorretamente na coluna;
- **Falso Negativo (FN)**: Número de exemplos da erroneamente identificados como não pertencentes à categoria principal. Identificam os elementos classificados incorretamente na linha da classe.

Desse modo, a matriz de confusão apresentada na Tabela 2, está organizada para prever a ocorrência ou ausência de um evento, com a classe "B" sendo o foco de referência.

Tabela 2 – Matriz de Confusão multiclasse. A classe "B" é o foco de referência.

	Classe Predita A	**Classe Predita B**	**Classe Predita C**	**Classe Predita D**
Classe Verdadeira A	VN	FP	VN	VN
Classe Verdadeira B	FN	VP	FN	FN
Classe Verdadeira C	VN	FP	VN	VN
Classe Verdadeira D	VN	FP	VN	VN

Fonte: Adaptado de (GRANDINI *et al.*, 2020).

Logo, as seguintes equações representam as métricas de avaliação, derivadas da matriz de confusão e essenciais para a determinação do desempenho de modelos de classificação (NASER; ALAVI, 2020; TING, 2010; HEYDARIAN *et al.*, 2022):

- **Acurácia (Acc)**: Determina o desempenho geral do modelo, por meio da análise da proporção do número de previsões corretas para o número total de elementos.

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

- **Precisão (Prec)**: Refere-se à capacidade do modelo em identificar de forma precisa os casos positivos, através da análise da proporção de observações positivas classificadas corretamente em relação ao total de casos previstos como positivos.

$$Prec = \frac{VP}{VP + FP} \tag{2.4}$$

- **Sensibilidade ou *Recall***: Mede a capacidade do modelo em identificar precisamente os casos positivos, mesmo que também classifique erroneamente alguns casos negativos como positivos. Corresponde à taxa de acerto na classe positiva, também chamada de taxa de verdadeiros positivos.

$$Recall = \frac{VP}{VP + FN} \tag{2.5}$$

- ***F1-Score*** **(F_1)**: Define a média harmônica das métricas de precisão e *recall*, considerando o mesmo grau de importância para as duas medidas.

$$F_1 = \frac{2 \times Prec \times Recall}{Prec + Recall} \tag{2.6}$$

- **Especificidade (Esp)**: Indica a taxa de acerto na classe negativa, ou seja, mede as proporções de observações negativas que são verdadeiros positivos. É uma métrica essencial para avaliar o desempenho de um modelo em problemas em que a identificação correta dos casos negativos é crucial.

$$Esp = \frac{VN}{VN + FP} \tag{2.7}$$

2.9.2 Métricas para Segmentação

A pesquisa em IA teve um rápido crescimento com modelos de *DL*, especialmente no campo de segmentação de imagens médicas. No entanto, a avaliação de desempenho de modelos de segmentação de imagens carece de uma análise confiável e demonstra viés estatístico por implementação ou uso incorreto de métricas. Além disso, a avaliação da segmentação pode ser uma tarefa complexa, visto que é necessária para medir a precisão da classificação, bem como a exatidão da localização. Portanto, é crucial a escolha de métricas apropriadas, dentre as disponíveis na literatura, a fim de aumentar a confiabilidade da pesquisa no campo da segmentação de imagens médicas (MÜLLER *et al.*, 2022).

Nesse contexto, a maioria das métricas de segmentação (com exceção da distância de Hausdorff) são baseadas no cálculo de uma matriz de confusão. Assim, no contexto de segmentação de imagens, pode-se definir a classe positiva como sendo pixels correspondentes ao objeto de interesse, enquanto à classe negativa são atribuídos os pixels de fundo:

- **Verdadeiro Positivo (VP)**: Número de exemplos da classe positiva segmentados corretamente;
- **Verdadeiro Negativo (VN)**: Número de exemplos da classe negativa segmentados corretamente;
- **Falso Positivo (FP)**: Número de exemplos da classe negativa erroneamente classificados como exemplos da classe positiva;
- **Falso Negativo (FN)**: Número de exemplos da classe positiva classificados incorretamente como exemplos da classe negativa.

Dessa forma, uma extensa variedade de métricas de avaliação no contexto de segmentação de imagens pode ser encontrada na literatura, dentre elas (TAHA; HANBURY, 2015; MÜLLER *et al.*, 2022):

- **Acurácia (Acc)**: Indica o número de previsões corretas, em comparação com o número total de previsões. No entanto, o uso apenas da métrica de acurácia (Eq. 2.3) é desencorajado, visto que, pode resultar em uma pontuação alta ilegítima, principalmente, em casos de desequilíbrio de classes.
- ***F1-Score*** **(F_1)**: Também conhecida como Coeficiente de Similaridade de Dados (DSC), define a média harmônica entre as métricas de precisão e *recall*. Trata-se da principal métrica para validação e interpretação de desempenho de modelos de segmentação de imagens médicas.

$$F_1 = DSC = \frac{2 \times VP}{2 \times VP + FP + FN} \tag{2.8}$$

- **IoU**: Conhecida também como Índice de Jaccard, é calculada como a interseção entre área da região predita e da região verdadeira, dividida pela união dessas áreas. Em suma, a IoU entre dois conjuntos indica a sobreposição entre eles, dividida pela sua união (JACCARD, 1912).

$$IoU = \frac{\text{Área da Interseção}}{\text{Área da União}} = \frac{TP}{TP + FP + FN} \tag{2.9}$$

- **DH**: É uma métrica baseada na distância espacial que avalia a similaridade entre dois conjuntos de pontos, como uma verdade fundamental e uma segmentação prevista. Essa métrica permite a pontuação da similaridade de localização, concentrando-se na delimitação dos limites. Em resumo, trata-se de uma medida

que quantifica a semelhança entre dois conjuntos de pontos, sendo comumente usada para avaliar a precisão da segmentação em tarefas de detecção de contornos ou bordas.

$$d(A,B) = \max\left(\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(b,a)\right) \tag{2.10}$$

onde A e B são conjuntos de pontos e $d(a,b)$ representa a métrica de distância entre os pontos a e b.

2.10 Testes Estatísticos

Os testes estatísticos são ferramentas utilizadas para compreender a significância dos resultados, identificar as possíveis diferenças entre os resultados dos modelos e determinar se os resultados são estatisticamente válidos. Tais testes podem ser divididos em duas categorias principais: Paramétricos e Não Paramétricos. Os testes paramétricos são aplicados quando os dados seguem uma distribuição normal e possuem variância e desvio padrão iguais. Logo, se os dados não satisfazem os critérios de um teste paramétrico, então, são atribuídos testes não paramétricos (NEIDEEN; BRASEL, 2007).

2.10.1 Shapiro-Wilk

O teste de *Shapiro-Wilk* é um teste de distribuição usado para verificar a suposição de normalidade de um conjunto de dados (SHAPIRO; WILK, 1965). Na análise estatística, um teste com um nível de significância de 5% ($\alpha = 0,05$), é considerado para uma distribuição normal. Para verificar se uma hipótese é válida e estatisticamente significativa, é usado o valor p. Caso a hipótese nula esteja correta, então o valor p é a probabilidade de obter resultados tão extremos quanto os resultados observados do teste estatístico. Um valor p maior que o nível de significância ($\alpha = 0,05$) indica que a hipótese nula é aceita, caso contrário é rejeitada (BARUAH *et al.*, 2020).

Dessa forma, a Hipótese Nula (H_0) e a Hipótese Alternativa (H_A) para o teste de *Shapiro-Wilk* são considerados como:

- H_0: A distribuição segue a normalidade;
- H_A: A distribuição não segue a normalidade.

2.10.2 Levene

O teste de Levene é aplicado para avaliar a suposição de igualdade de variâncias entre dois ou mais grupos de dados (BROWN; FORSYTHE, 1974). Dessa forma, um valor p significativo maior que o nível de significância de 5% ($\alpha = 0,05$) indica que a homogeneidade entre as populações não é significativamente diferente. Enquanto, um valor de p menor do que $0,05$ determina que a homogeneidade entre as populações é significativamente diferente (OTHMAN *et al.*, 2022).

As hipótese nula (H_0) e a hipótese alternativa (H_A) para o teste de Levene são definidas como:

- H_0: A distribuição segue a homogeneidade entre as populações;
- H_A: A distribuição não segue a homogeneidade entre as populações.

2.10.3 ANOVA

A Análise de Variância (ANOVA) (GREENHOUSE; GEISSER, 1959) é um teste paramétrico aplicado para determinar se existe uma diferença significativa entre as médias das métricas dos grupos avaliados. A ANOVA é utilizada quando todas as populações atendem aos requisitos de normalidade e homogeneidade (GREENHOUSE; GEISSER, 1959).

O teste estatístico incorpora médias e variâncias para determinar a estatística do teste (F). A estatística de teste é então usada para determinar se os grupos de dados são iguais ou diferentes. Em suma, o teste avalia a diferença entre os métodos utilizados por mais de dois grupos e utiliza o cálculo para determinar se as variáveis reais se desviam da média geral da variável dependente (NEIDEEN; BRASEL, 2007; SINGH *et al.*, 2021).

Existem dois tipos de hipóteses fornecidas pela ANOVA:

- H_0: Não há diferença entre os grupos e há igualdade entre suas médias;
- H_A: Há diferença entre seus grupos e também entre suas médias.

2.10.4 Tukey

O teste de *Tukey* é um um teste *pós-hoc* realizado após a ANOVA se a estatística de teste F for significativa (TUKEY, 1949). O teste de *Tukey* é um teste de análise de variância que determina se há diferença significativa entre as médias dos grupos, onde são realizadas duas comparações de cada vez. O processo mais popular para comparação de médias entre pares de

grupos em uma população é o *Tukey* HSD (MEFTAH *et al.*, 2018; SRAVANI *et al.*, 2022). O teste de *Tukey* é determinado como:

$$HSD = q\sqrt{\frac{MS}{n}}, \tag{2.11}$$

onde HSD é a diferença honestamente significativa, MS é o valor quadrático médio calculado pela ANOVA e n é o número de amostras em grupos individuais.

O teste de *Tukey* é também um teste de hipótese, onde as hipóteses nula e alternativa são determinadas como:

- H_0: As médias analisadas são semelhantes;
- H_A: As médias são significativamente diferentes.

2.10.5 Friedman

O teste de *Friedman* é uma alternativa robusta e não paramétrica ao teste ANOVA, utilizado quando pelo menos uma das populações não atende às suposições de normalidade e homogeneidade de variâncias. O teste não paramétrico é aplicado em situações onde as mesmas unidades experimentais são medidas repetidamente em diferentes condições ou ao longo do tempo, conhecidas como medidas repetidas. O teste de *Friedman* avalia se há diferenças significativas entre essas medições repetidas, permitindo uma análise estatística confiável mesmo na ausência de distribuição normal dos dados (FRIEDMAN, 1937).

O objetivo é testar a hipótese de que todos os grupos dependentes j possuem distribuições idênticas (WILCOX, 2003):

$$H_0 : F_1(x) = ... = F_j(x). \tag{2.12}$$

Logo, as hipóteses nula H_0 e alternativa H_A para o teste de Friedman são dadas como:

- H_0: Não há diferença significativa entre os grupos dependentes;
- H_A: Há uma diferença significativa entre os grupos dependentes.

2.10.6 Nemenyi

O teste de *Nemenyi* é um teste *pós-hoc* realizado após o teste não paramétrico de Friedman. Trata-se de um método de comparação múltipla utilizado para identificar quais

grupos apresentam diferenças significativas entre si. O teste de *Nemenyi* consiste em realizar comparações entre pares de grupos, com o objetivo de identificar onde ocorrem as diferenças significativas entre esses grupos (NEMENYI, 1963).

Portanto, as duas hipóteses que contemplam o teste de *Nemenyi* são:

- H_0: Não há diferença significativa entre os grupos comparados;
- H_A: Há uma diferença significativa entre pelo menos dois dos grupos comparados.

2.11 Validação Externa

A validação de modelos de *ML* consiste em uma das etapas cruciais seja em atividades de classificação ou segmentação. A validação interna refere-se ao processo de estimar o desempenho dos modelos, por meio do particionamento do conjunto de dados original em partições menores para validação e teste (ZHANG *et al.*, 2020). Todavia, o protocolo de validação interna pode não ser suficientemente válido para ambientes críticos, por exemplo, análise de imagens médicas. Portanto, é essencial que os modelos sejam robustos e confiáveis em contextos diferentes daqueles em que os dados de treinamento foram obtidos (BLEEKER *et al.*, 2003; HERNANDEZ-BOUSSARD *et al.*, 2020; CABITZA *et al.*, 2021).

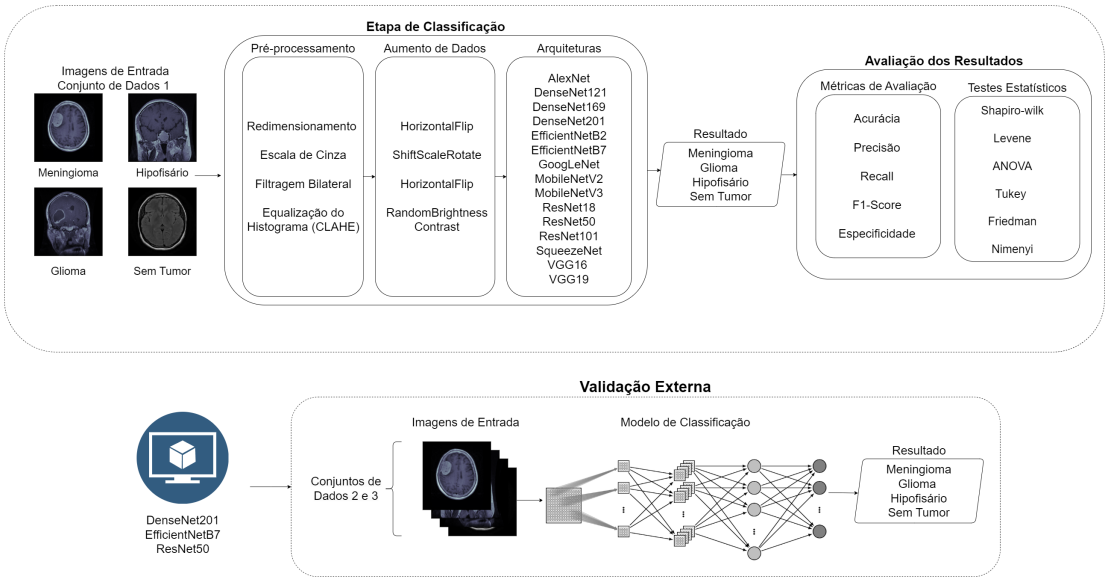
Logo, para garantir a generalização e a robustez de um modelo, a validação externa é incentivada. Esse tipo de validação consiste em utilizar dados externos, ou seja, novos conjuntos de dados provenientes de fontes diferentes daquelas utilizadas na criação do modelo, a fim de garantir que o desempenho e confiabilidade sejam reproduzidos em diferentes contextos (CABITZA *et al.*, 2021).

3 METODOLOGIA

Nesta seção, está concentrada a metodologia utilizada para a elaboração deste projeto. Serão explorados os detalhes sobre o fluxo dos experimentos realizados. Para isso, serão abordados os conjuntos de dados empregados, as técnicas de pré-processamento e aumento de dados, além da abordagem de *grid search* para otimização de hiperparâmetros. Também serão abordadas as etapas de classificação e segmentação, a utilização de validação cruzada, as métricas de avaliação e os testes estatísticos utilizados. Por fim, serão descritos os *softwares* e bibliotecas utilizados ao longo do desenvolvimento.

Este projeto consiste em duas etapas principais: a classificação de tumores cerebrais em Meningioma, Glioma, Hipofisário e em casos sem tumor, e a segmentação desses tumores em imagens de RM. A Figura 18 apresenta as etapas realizadas no fluxo de trabalho proposto. A priori, na etapa de classificação, foram aplicadas técnicas de pré-processamento nas imagens de entrada e realizado o aumento de dados do conjunto de imagens de treinamento. Em seguida, fez-se o treinamento de arquiteturas de redes neurais pré-treinadas. Ao final dessa etapa, foi desenvolvida a análise dos resultados, por meio de métricas de avaliação e testes estatísticos. Por fim, para avaliar o desempenho e robustez dos modelos, foi aplicada uma validação externa com novos conjuntos de dados.

Figura 18 – Fluxograma de trabalho proposto para etapa de classificação.

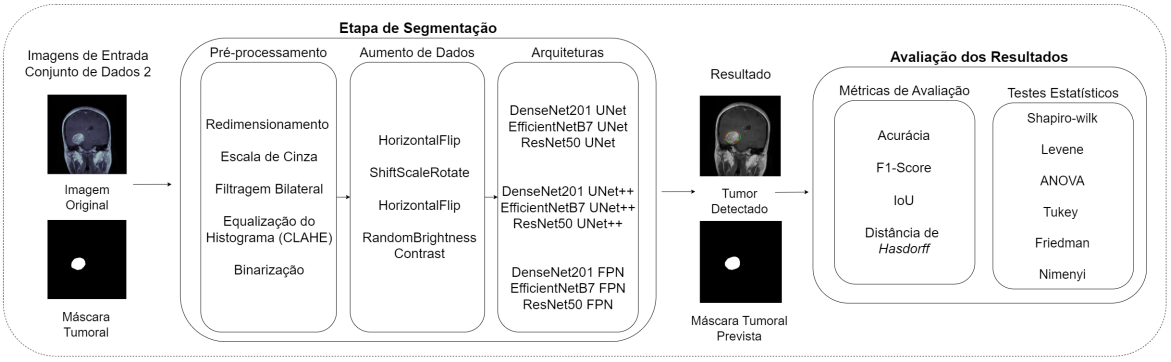


Fonte: Elaborado pela autora.

Em seguida, na etapa de segmentação de tumores cerebrais, apresentada na Figura

19, é empregada uma base de dados composta por imagens de RM de tumores cerebrais e as máscaras tumorais correspondentes a cada imagem. O conjunto de dados é submetido a técnicas de aumento de dados, para ampliar a variabilidade do banco de dados. Ademais, as imagens são fornecidas como entrada para as redes codificadora e decodificadora. As redes codificadoras foram determinadas com base nos resultados da classificação das imagens. Em suma, nessa etapa, é realizada a segmentação da região tumoral, com base nas máscaras tumorais previstas pelas redes testadas.

Figura 19 – Fluxograma de trabalho proposto para etapa de segmentação.



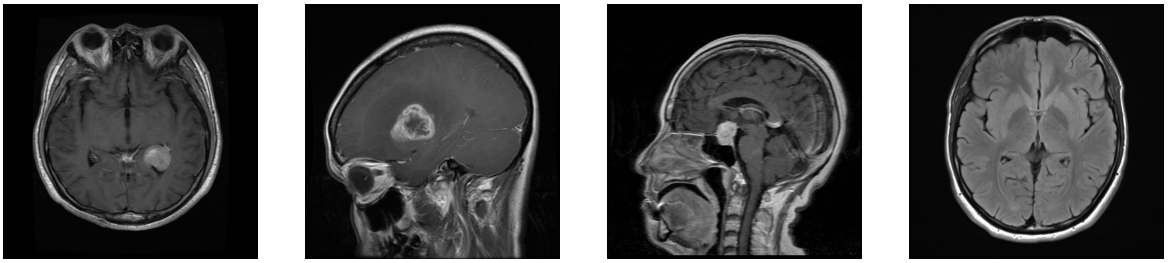
Fonte: Elaborado pela autora.

3.1 Base de dados

Na etapa de classificação, foi utilizado o conjunto de dados *Brain Tumor Classification (MRI)* disponível publicamente na literatura (BHUVAJI *et al.*, 2020). Este conjunto de dados foi empregado para realizar o treinamento das arquiteturas para a classificação de imagens. A base de dados é composta por 3264 imagens de RM com contraste aprimorado no formato JPEG. As imagens são categorizadas de acordo com os tipos de tumor cerebral: Meningioma (937 imagens), Glioma (926 imagens), Hipofisário (901 imagens) e casos Sem Tumor (500 imagens). A Figura 20 apresenta as imagens relacionadas às diferentes classes que compõem a base de dados.

Na etapa de segmentação, foi utilizado um conjunto de dados público da literatura *Figshare* (CHENG, 2017). O banco de dados possui 3064 imagens de RM com contraste, obtidas de 233 pacientes e classificadas em três classes: Meningioma (708 imagens), Glioma (1426 imagens) e Hipofisário (930 imagens). A base de dados é composta pelas coordenadas de pontos discretos da borda do tumor. Dessarte, para cada imagem do banco existe uma imagem binária

Figura 20 – Amostras da base de dados *Brain Tumor Classification (MRI)*.

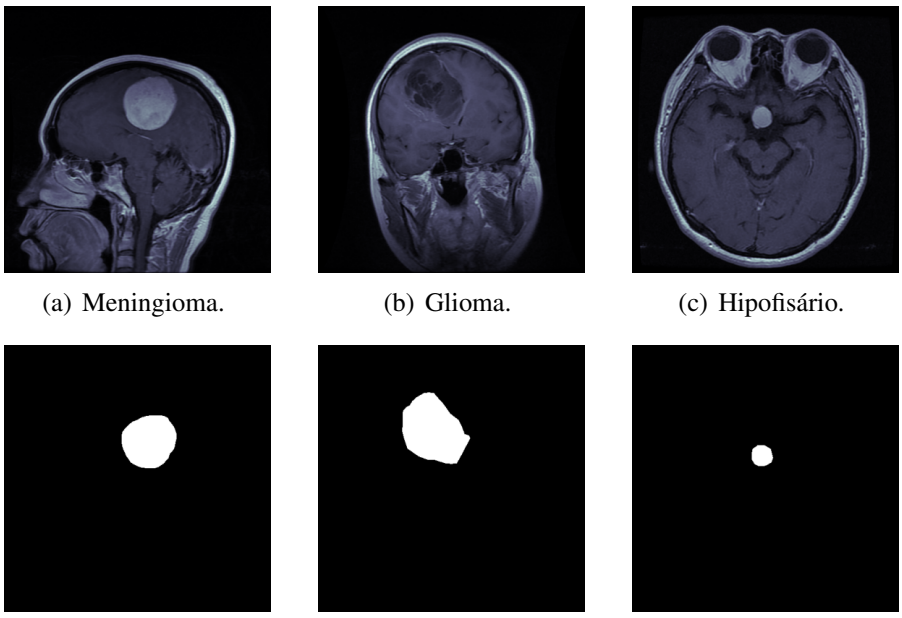


(a) Meningioma. (b) Glioma. (c) Hipofisário. (d) Sem Tumor.

Fonte: (BHUVAJI *et al.*, 2020).

da máscara tumoral correspondente. A Figura 21 apresenta as imagens relacionadas aos três tipos de tumores e as suas respectivas máscaras. A primeira linha exibe três exemplos representativos de imagens do banco de dados, referente aos tumores meningioma, glioma e hipofisário. Posteriormente, na segunda linha são apresentadas as máscaras tumorais correspondentes a cada uma dessas imagens.

Figura 21 – Amostras da base de dados *Figshare*. A Linha 1 apresenta as imagens originais e a Linha 2 as máscaras tumorais correspondentes.



(a) Meningioma. (b) Glioma. (c) Hipofisário.

(d) Máscara Meningioma. (e) Máscara Glioma. (f) Máscara Hipofisário.

Fonte: (CHENG, 2017).

Para a validação externa dos modelos de classificação, foi empregada a combinação de dois conjuntos de dados públicos. Para isso, foi utilizado novamente o conjunto de dados Figshare (CHENG, 2017) combinado com a base de dados Br35H (CHAKRABARTY, 2017). Dessa forma, o conjunto de dados Br35H inclui imagens de RM divididas em duas classes, com

tumor e sem tumor. Para esse trabalho, foram selecionadas todas as imagens da classe sem tumor, totalizando 1.500 imagens, para compor a validação externa. A Tabela 3 apresenta um compilado das informações dos bancos de dados empregados neste trabalho.

Tabela 3 – Informações sobre os conjuntos de dados.

Base de Dados	Classificação	Validação	Segmentação	Classes	Total
Brain Tumor Classification (BHUVAJI *et al.*, 2020)	✓	✗	✗	Meningioma, Glioma Hipofisário e Sem Tumor.	3.264
Figshare (CHENG, 2017)	✗	✓	✓	Meningioma, Glioma e Hipofisário.	3.064
Br35H (CHAKRABARTY, 2017)	✗	✓	✗	Sem Tumor.	1.500

Fonte: Elaborado pela autora.

3.2 Pré-processamento dos Dados

A etapa de pré-processamento de imagens consiste em transformar os dados brutos de imagem em dados de imagem limpos, com objetivo de aprimorar a qualidade da imagem e reduzir artefatos indesejados. Neste projeto, as imagens provenientes dos bancos de dados foram convertidas para escala de cinza, redimensionadas para o tamanho 224 x 224 pixels e salvas no formato PNG.

Em seguida, foi aplicado a técnica de filtragem bilateral, a fim de suavizar a imagem, reduzir o ruído indesejado e preservar a nitidez das bordas. Ademais, para melhorar o contraste das imagens, foi realizada a CLAHE. Por fim, para as máscaras tumorais provenientes do banco de dados *Figshare*, aplicou-se a técnica de limiarização simples, a fim de binarizar tais imagens.

3.3 Aumento de Dados

Para melhorar a capacidade de generalização dos modelos, expandir a variabilidade do banco de dados e mitigar problemas de *overfitting*, foram utilizadas técnicas de transformações em imagens. Vale ressaltar que as transformações foram aplicadas exclusivamente nas imagens de treinamento, logo, não foram aplicadas para os conjuntos de validação e teste.

Portanto, foram utilizadas transformações afins, por exemplo, mudanças no brilho e contraste (*RandomBrightnessContrast*), inversão horizontal (*HorizontalFlip*), mudança de

escala, translação e rotação da imagem (*ShiftScaleRodar*). A Tabela 4 apresenta os parâmetros utilizados para cada transformação.

Tabela 4 – Técnicas e parâmetros de aumento de dados.

Técnicas	Parâmetros
RandomBrightnessContrast	Limite de brilho = $0,3$, Limite de contraste = $0,3$, Probabilidade de aplicar a Transformada: $p = 0,5$
HorizontalFlip	Probabilidade da imagem ser invertida: $p = 0,5$
ShiftScaleRodar	Deslocamento = $0,05$, Escala = $0,1$, Rotação = 15, Probabilidade de aplicar a Transformada: $p = 0,5$

Fonte: Elaborado pela autora.

3.4 *Grid Search*

Para otimizar o desempenho dos modelos de classificação, utilizou-se a técnica *Grid Search*. Essa abordagem consiste em definir os melhores conjuntos de hiperparâmetros para cada arquitetura utilizada, com objetivo de maximizar a eficácia dos modelos. Nesse contexto, definiu-se os valores para os hiperparâmetros: taxa de aprendizagem, otimizador, tamanho do lote (*batch size*) e decaimento de peso (*weight decay*).

Em suma, a taxa de aprendizagem garante que o modelo atinja a convergência de maneira eficaz. O otimizador é responsável por ajustar os pesos do modelo durante o treinamento. O *batch size* define a quantidade de exemplos de treinamento usados em uma iteração. Já o *weight decay* adiciona uma penalidade à função de custo do modelo, a fim de evitar o *overfitting*. A Tabela 5 apresenta o conjunto de valores possíveis para cada hiperparâmetro na construção dos modelos.

Tabela 5 – Hiperparâmetros otimizados com *Grid Search*.

Hiperparâmetros	Etapa	Valores
Taxa de Aprendizagem	Classificação	$[0.0001, 0.0005, 0.001]$
Otimizador	Classificação	[Adam, RMSprop, SGD]
Batch Size	Classificação	$[16, 32, 64]$
Weight Decay	Classificação	$[0.00001, 0.00005, 0.0001]$

Fonte: Elaborado pela autora.

3.5 Classificação

Para o reconhecimento dos tipos de tumores Meningioma, Glioma, Hipofisário e em casos sem tumor, realizou-se o processo de classificação por meio de arquiteturas *CNN* pré-treinadas. No processo de classificação, um modelo é responsável por definir a probabilidade de uma saída pertencer à mesma classe de uma determinada entrada. Entretanto, o treinamento de uma arquitetura CNN do zero é um processo extremamente demorado. Para mitigar essa questão, é comum a aplicação da técnica de *Transfer Learning*.

Nesse projeto, foram utilizados modelos *CNN* pré-treinados em conjuntos de dados, aplicando o conhecimento já adquirido. Logo, para realizar a classificação dos tipos de tumores, foram aplicadas as arquiteturas AlexNet, DenseNet121, DenseNet169, DenseNet201, EfficientNetB2, EfficientNetB7, GoogLeNet, MobileNetV2, MobileNetV3, ResNet18, ResNet50, ResNet101, SqueezeNet, VGG16 e VGG19, com pesos pré-treinados do ImageNet.

Nesse contexto, foram definidos diversos hiperparâmetros para otimizar o desempenho dos modelos. Aplicou-se a técnica *grid search* para ajustar os valores dos hiperparâmetros, incluindo taxa de aprendizagem (LR), otimizador, *batch size* e *weight decay* (WD). Além disso, cada modelo foi treinado por 50 épocas, o que representa o número de iterações durante o treinamento. Foi definida a função de perda Entropia Cruzada, amplamente empregada em problemas com duas ou mais classes. Por fim, foi estabelecida uma paciência (*patience*) igual a 5, o que indica o número de épocas consecutivas em que, na ausência de melhora no desempenho do modelo, o treinamento pode ser interrompido. A Tabela 6 apresenta o conjunto de valores definidos para cada hiperparâmetro na construção dos modelos.

3.6 Segmentação

Para realizar a tarefa de detecção e segmentação do tumor cerebral nas imagens de ressonância magnética, foi utilizado o conceito de codificadores e decodificadores (do inglês, *Encoder-decoder*). O codificador, em geral, é uma rede de classificação pré-treinada, responsável por codificar a imagem de entrada em representações de recursos em vários níveis diferentes. Em contrapartida, o decodificador, projeta semanticamente os recursos discriminativos aprendidos pelo codificador no espaço de pixels para obter uma classificação densa.

Neste trabalho, utilizou-se uma combinação de arquiteturas codificadoras e decodificadoras para analisar as estruturas para segmentação da área tumoral. Os codificadores utilizados

Tabela 6 – Hiperparâmetros definidos para os modelos de classificação.

Arquiteturas	LR	Otimizador	*Batch Size*	WD	Épocas	Função de Perda	*Patience*
AlexNet	0,001	Adam	32	0,00001	50	Entropia Cruzada	5
DenseNet121	0,0001	Adam	32	0,0001	50	Entropia Cruzada	5
DenseNet169	0,0001	Adam	32	0,0001	50	Entropia Cruzada	5
DenseNet201	0,0001	Adam	32	0,0001	50	Entropia Cruzada	5
EfficientNetB2	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
EfficientNetB7	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
GoogLeNet	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
MobileNetV2	0,001	Adam	32	0,0001	50	Entropia Cruzada	5
MobileNetV3	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
ResNet18	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
ResNet50	0,0001	RMSprop	32	0,00001	50	Entropia Cruzada	5
ResNet101	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
SqueezeNet	0,0001	Adam	32	0,0001	50	Entropia Cruzada	5
VGG16	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5
VGG19	0,0005	Adam	32	0,0001	50	Entropia Cruzada	5

Fonte: Elaborado pela autora.

no experimento foram DenseNet201, EfficientNetB7 e ResNet50, enquanto, os decodificadores foram as redes UNet, UNet++ e FPN. Durante a etapa de segmentação, os modelos *encoder-decoder* recebem como entrada uma imagem de RM e uma máscara correspondente à área tumoral, utilizada para o treinamento. Como resultado, as redes retornam uma máscara tumoral prevista, que indica a localização do tumor na imagem de RM.

Para o treinamento dos modelos, aplicou-se 50 épocas, com uma paciência (*patience*) de 5 épocas. Além disso, para os modelos codificadores, utilizou-se os mesmos hiperparâmetros definidos na etapa de classificação (Tabela 6). Já para os modelos decodificadores, aplicou-se a função de perda *Lovasz*, projetada para otimizar a métrica de IoU, frequentemente usada em segmentação semântica. A Tabela 7 apresenta um resumo dos hiperparâmetros utilizados para o treinamento dos modelos de segmentação.

3.7 Validação Cruzada

Neste procedimento foi utilizada a validação cruzada *k-fold*. Nessa abordagem, os dados foram divididos em 80% a 20% para treinamento e teste, respectivamente, com 10% dos dados de treinamento alocados para validação. Para a tarefa de classificação o conjunto de dados foi particionado aleatoriamente em $k = 10$ *folds*. Já para a etapa de segmentação utilizou-se um $k = 5$ *folds* para particionar o banco de dados. Por fim, para definir o desempenho de cada modelo é realizada média do desempenho em cada um dos *folds*.

Tabela 7 – Hiperparâmetros definidos para os modelos de segmentação.

Hiperparâmetros	Etapa	Valores
Taxa de Aprendizagem	Segmentação	$[0,0001; 0,0005]$
Otimizador	Segmentação	[Adam; RMSprop]
Batch Size	Segmentação	[32]
Weight Decay	Segmentação	$[0,00001; 0,0001]$
Função de Perda	Segmentação	[*Lovasz*]
Patience	Segmentação	[5]

Fonte: Elaborado pela autora.

3.8 Análise dos Resultados

No desenvolvimento de algoritmos de *DL* é imprescindível a utilização de métricas de avaliação adequadas para avaliar o desempenho dos modelos. A qualidade do treinamento é definida pelo resultado das métricas, logo, é crucial a escolha de métricas que avaliem corretamente se o modelo está atendendo aos objetivos propostos.

3.8.1 Métricas de Avaliação

Em tarefas de classificação, o valor VP refere-se aos casos corretamente identificados como pertencentes a uma determinada categoria pelo modelo, ou seja, são os tumores de uma determinada classe que o modelo identificou corretamente. Os FP são os casos erroneamente identificados como pertencentes a uma determinada classe. Os VN indicam os casos identificados corretamente como não pertencentes a uma categoria em análise. Finalmente, os FN representam os casos erroneamente identificados como não pertencendo à categoria principal.

No contexto de classificação de imagens de tumores cerebrais, foram utilizadas as métricas: acurácia, precisão, *recall*, *F1-score* e especificidade. Portanto, a métrica acurácia é usada para determinar o desempenho geral do modelo, enquanto a precisão refere-se à capacidade do modelo em identificar de forma precisa os casos positivos. O *recall*, também chamada de sensibilidade, indica que o modelo é capaz de identificar a maioria dos casos positivos, mesmo que também classifique erroneamente alguns casos negativos. Além disso, o *F1-score* é a média harmônica das métricas de precisão e *recall*, considerando a importância das duas as métricas. Por fim, a especificidade é uma métrica essencial para avaliar o desempenho de um modelo em

problemas em que a identificação correta dos casos negativos é crucial.

Ademais, no cenário de segmentação de imagens o valor VP refere-se aos pixels do tumor segmentados corretamente. Já o valor VN indica os pixels de fundo segmentados corretamente. O FP refere-se aos pixels de fundo erroneamente classificados como pixels de tumor. E por fim, o FN indica os pixels de tumor classificados incorretamente como pixels de fundo. Dessa forma, para avaliar os modelos de segmentação foram utilizadas as métricas: acurácia, F1-Score, intersecção sobre união (IoU) e Distância de *Hasdorff*.

Logo, a acurácia indica o número de previsões corretas, em comparação com o número total de previsões. O *F1-score* é a principal métrica utilizada para validar o desempenho de modelos de segmentação. A IoU é uma métrica amplamente utilizada em tarefas de segmentação, por denotar a interseção da área da região predita com a área da região real, dividida pela união dessas áreas. Por fim, a Distância de *Hasdorff* é uma medida que quantifica a semelhança entre dois conjuntos de pontos, sendo comumente usada para avaliar a precisão da segmentação em tarefas de detecção de contornos ou bordas.

3.8.2 Testes Estatísticos

Para compreender a significância dos resultados e identificar as possíveis diferenças entre os resultados dos modelos de classificação, foi conduzido uma análise estatística abrangente, por meio dos seguintes testes:

- ***Shapiro-wilk***: Testar a suposição de normalidade das médias das métricas de avaliação para cada um dos modelos;
- ***Levene***: Avaliar a suposição de variâncias iguais;
- **ANOVA**: Utilizado se todas as populações atenderem ao requisito de normalidade e homogeneidade. O teste ANOVA é aplicado para determinar se existe uma diferença significativa entre as médias das métricas das arquiteturas
- ***Tukey***: Considera quais pares específicos de métricas apresentam diferenças significativas entre si;
- ***Friedman***: Aplicado quando pelo menos uma das populações não atende ao critério de normalidade e homogeneidade. O teste não paramétrico de *Friedman* determina se existe diferença significativa entre as médias;
- ***Nemenyi***: O *pós-hoc* de *Nemenyi* é responsável por identificar quais grupos são significativamente diferentes entre si.

A Figura 22 apresenta os procedimentos adotados para conduzir a análise estatística. Dessa forma, o fluxograma exibe os testes estatísticos utilizados para avaliar a significância dos resultados. São analisadas separadamente as métricas de avaliação de cada modelo *CNN*, incluindo acurácia, precisão, *recall*, *F1-Score* e especificidade, para os modelos de classificação. Assim como, são consideradas as métricas de acurácia, *F1-Score*, IoU e Distância de *Hasdorff*, para o cenário de segmentação de imagens. Cada pontuação é o resultado retornado pelos K folds analisados, ou seja, se o treinamento for realizado com um $K = 10$, cada modelo terá 10 valores para a métrica em questão. O fluxo de trabalho previsto no fluxograma consiste em:

1. O primeiro passo é testar a suposição de normalidade para as médias das métricas de avaliação para cada um dos modelos por meio do teste de *Shapiro-wilk*;
2. Em seguida, avaliar a suposição de variâncias iguais para cada um dos modelos pelo teste de *Levene*;
3. Se todas as populações atenderem ao requisito de normalidade e homogeneidade é utilizado o teste ANOVA para determinar se existe uma diferença significativa entre as médias das métricas dos modelos avaliados e aplica-se o *pós-hoc* de *Tukey* para determinar quais pares específicos de métricas apresentam diferenças significativas entre si;
4. Caso contrário, se pelo menos uma das populações não atende ao critério de normalidade e homogeneidade deve-se aplicar o teste não paramétrico de *Friedman*, a fim de determinar se existe diferença significativa entre as médias. Por fim, caso seja utilizado o teste de Friedman, aplica-se o *pós-hoc* de *Nemenyi*, para identificar quais grupos são significativamente diferentes entre si.

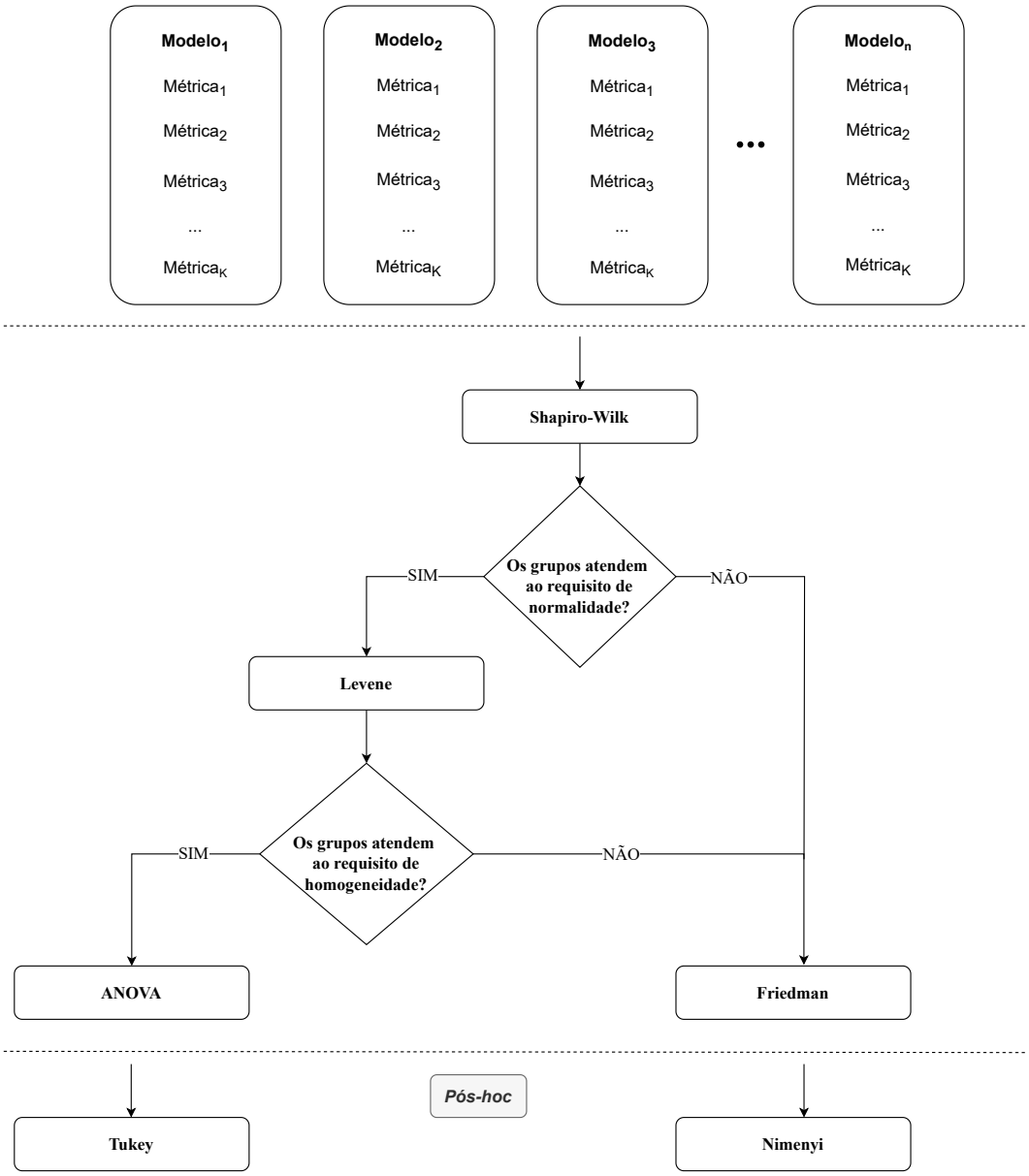
3.9 Validação Externa

Para garantir a generalização e a robustez dos modelos de classificação, foi realizada uma validação externa. Para isso, foram combinados dois conjuntos de dados de imagens de ressonância magnética, não utilizados durante o treinamento, validação e teste dos modelos. Desse modo, o objetivo é avaliar o desempenho dos modelos para situações distintas.

3.10 Ambiente de Desenvolvimento

O ambiente de desenvolvimento para os experimentos foi configurado em um sistema operacional Ubuntu Linux 24.04 LTS, com hardware que incluiu uma placa de vídeo NVIDIA

Figura 22 – Fluxograma dos testes estatísticos aplicados. As etapas são repetidas para cada uma das métricas de avaliação utilizadas. São utilizados nos testes os valores retornados em cada *fold* para a métrica analisada.



Fonte: Elaborado pela autora.

GeForce RTX 3090 com 24 GB de RAM dedicada, um processador Intel Core i9-14900K e 128 GB de memória.

3.11 Linguagens e Bibliotecas

O desenvolvimento das aplicações foi realizado em *Python*[1], uma linguagem *open-source* amplamente utilizada, especialmente por sua vasta gama de bibliotecas voltadas ao desenvolvimento de modelos de *DL*. Para o pré-processamento das imagens, foi utilizada a biblioteca *OpenCV*[2]. O treinamento dos modelos foi conduzido com o uso das bibliotecas *Scikit-learn*[3], *PyTorch*[4], *PyTorch Lightning*[5].

Além disso, bibliotecas como *NumPy*[6], *Matplotlib*[7] e *Seaborn*[8], foram empregadas para operações numéricas e criação de gráficos, facilitando a visualização e análise dos resultados. Para os testes estatísticos, foram utilizadas as bibliotecas *SciPy*[9], Statsmodels[10] e Scikit-posthocs[11].

[1] Disponível em: <https://www.python.org/>. Acesso em: 30 agosto 2024.
[2] Disponível em: <https://opencv.org/>Acesso em: 30 agosto 2024.
[3] Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 30 agosto 2024.
[4] Disponível em: <https://pytorch.org/>. Acesso em: 30 agosto 2024.
[5] Disponível em: <https://lightning.ai/docs/pytorch/stable/>. Acesso em: 30 agosto 2024.
[6] Disponível em: <https://numpy.org/>. Acesso em: 30 agosto 2024.
[7] Disponível em: <https://matplotlib.org>. Acesso em: 30 agosto 2024.
[8] Disponível em: <https://seaborn.pydata.org/>. Acesso em: 30 agosto 2024.
[9] Disponível em: <https://scipy.org/>. Acesso em: 30 agosto 2024.
[10] Disponível em: <https://www.statsmodels.org/stable/index.html.Acesso em: 30 agosto 2024.
[11] Disponível em: <https://scikit-posthocs.readthedocs.io/en/latest/>. Acesso em: 30 agosto 2024.

4 RESULTADOS

Esta seção apresenta uma análise e discussão detalhada dos resultados referentes à metodologia empregada neste projeto para a classificação e segmentação de tumores cerebrais em imagens de ressonância magnética.

4.1 Classificação

Para realizar a treinamento do modelo e a classificação dos dados, a base de dados foi dividida em 80% para treinamento e 20% para teste, com 10% dos dados de treinamento alocados para validação e utilizada a validação cruzada *k-fold*. O desempenho dos modelos para a tarefa de classificação multiclasse foi analisado quantitativamente pelas métricas de acurácia de teste, precisão, *recall*, *F1-Score*, especificidade e tempo de treinamento.

4.1.1 Análise de Métricas

No contexto de classificação, um modelo que apresenta taxas detalhadas para as métricas apresentadas é de extrema importância para a eficácia e confiabilidade do diagnóstico e para garantir um sistema abrangente e preciso. Dessa forma, uma alta acurácia indica a capacidade do modelo em identificar corretamente tanto os tumores quanto os casos sem tumor. A precisão mede a proporção de verdadeiros positivos entre todas as previsões positivas, assegurando que a classificação do tumor seja confiável. O *recall* analisa a capacidade do modelo em identificar todos os casos de tumores, o que diminui as chances de um tumor não ser verificado. O *F1-Score*, é a média harmônica entre as métricas precisão e *recall*, o que sugere a habilidade do modelo em ter um desempenho adequado mesmo se houver um desequilíbrio entre as classes analisadas. Já a especificidade mede a capacidade do modelo em identificar corretamente os casos sem tumor, evitando diagnósticos incorretos em pacientes saudáveis. Por fim, o tempo indica o tempo em segundos para realizar o treinamento de cada um dos modelos.

A Tabela 8 exibe os resultados para as métricas de avaliação dos cinco modelos avaliados. Os resultados obtidos sugerem que os modelos avaliados têm potencial promissor para distinguir os tumores cerebrais em Meningioma, Glioma, Hipofisário e casos sem tumor. Dentre os modelos analisados, o EfficientNetB7 obteve as maiores taxas gerais, em comparação com as demais redes, com uma acurácia de teste de 97,68%, precisão de 97,63%, *recall* de 97,69%, *F1-Score* de 97,64% e especificidade de 99,21%. Ademais, as arquiteturas DenseNet201

e ResNet50 apresentaram desempenho semelhante em relação às métricas analisadas, na qual, a arquitetura DenseNet201 apresentou uma acurácia de 97,25%, precisão de 97,36%, *recall* de 97,08%, *F1-Score* de 97,20% e especificidade de 99,07%. Enquanto, a rede ResNet50 obteve uma acurácia e 97,11%, precisão de 97,30%, *recall* de 96,97%, *F1-Score* de 97,12% e especificidade de 99,02%.

Tabela 8 – Resultados da classificação de tumores cerebrais

Arquitetura	ACC (%)	Prec (%)	*Recall (%)*	F1 (%)	ESP (%)	Tempo (s)
AlexNet	84,53 ± 0,81	84,66 ± 0,01	84,71 ± 0,01	84,47 ± 0,87	94,81 ± 0,24	359,27 ± 21,56
DenseNet121	96,74 ± 1.01	96,64 ± 0,01	96,64 ± 0,01	96,61 ± 1.05	98,9 ± 0,34	614,81 ± 3,13
DenseNet169	96,2 ± 1,01	96,11 ± 0,01	95,79 ± 0,01	95,88 ± 1,28	98,74 ± 0,32	675,94 ± 85,69
DenseNet201	97,25 ± 0,47	97,36 ± 0,01	97,08 ± 0,01	97,2 ± 0,54	99,07 ± 0,16	1108,59 ± 12,12
EfficientNetB2	96,46 ± 1,38	96,33 ± 0,02	96,61 ± 0,01	96,42 ± 1, 51	98,8 ± 0,47	731,79 ± 140,28
EfficientNetB7	**97,68 ± 0,55**	**97,63 ± 0,01**	**97,69 ± 0,01**	**97,64 ± 0,54**	**99,21 ± 0,19**	3664,82 ± 233,09
GoogLeNet	96,76 ± 0,97	96,88 ± 0,01	96,65 ± 0,01	96,73 ± 0,87	98,91 ± 0,33	345,69 ± 57,43
MobileNetV2	83,24 ± 0,6	82.72 ± 0,0	84,14 ± 0,01	83,07 ± 0,49	94,34 ± 0,25	**244,41 ± 12,99**
MobileNetV3	95,56 ± 2.82	95,9 ± 0,02	95,51 ± 0,03	95,58 ± 2.79	98,52 ± 0,88	**237,64 ± 83,97**
ResNet18	96,93 ± 0,51	96,94 ± 0,0	96,74 ± 0,0	96,82 ± 0,42	98,97 ± 0,18	361,47 ± 10,83
ResNet50	97,11 ± 0,52	97,3 ± 0,0	96,97 ± 0,01	97,12 ± 0,5	99,02 ± 0,17	581,81 ± 70,17
ResNet101	96,36 ± 0,39	96,54 ± 0,01	96,06 ± 0,01	96,25 ± 0,46	98,78 ± 0,13	1081,47 ± 27,71
SqueezeNet	95,03 ± 0,8	95,5 ± 0,01	94,4 ± 0,01	94,82 ± 0,99	98,35 ± 0,24	282,76 ± 1,27
VGG16	96,41 ± 0,73	96,59 ± 0,01	96,29 ± 0,01	96,4 ± 0,75	98,79 ± 0,24	1481,73 ± 393,73
VGG19	96,53 ± 0,75	96,67 ± 0,01	96,36 ± 0,01	96,46 ± 0,84	98,83 ± 0,24	1838,98 ± 275,81

Fonte: Elaborado pela autora.

Todavia, as redes AlexNet e MobileNetV2 alcançaram as menores pontuações em termos métricas, em relação às arquiteturas analisadas. A AlexNet apresenta uma acurácia e 84,53%, precisão de 84,66%, *recall* de 84,71%, *F1-Score* de 84,47% e especificidade de 94,81%. Em comparação, a MobileNetV2 possui uma acurácia de 83,24%, precisão de 82,72%, *recall* de 84,14%, *F1-Score* de 83,07% e especificidade de 94,34%. Esses resultados indicam que, embora ambas as redes tenham desempenho comparável, a AlexNet geralmente apresenta ligeira vantagem em todas as métricas. No entanto, ambas as redes estão abaixo do desempenho observado em outras arquiteturas mais avançadas, como DenseNet201 e EfficientNetB7, indicando a necessidade de avaliar o desempenho das arquiteturas para obter melhores resultados na tarefa de classificação.

Em termos de tempo de treinamento, a EfficientNetB7, devido à sua complexidade, ao elevado número de parâmetros e à profundidade das camadas, apresentou o maior tempo de treinamento, com 3664,82 segundos. No entanto, a arquitetura alcançou as melhores métricas de desempenho. Em contrapartida, a SqueezeNet, projetada como uma arquitetura compacta com poucos parâmetros, quando comparada às demais redes, ideal para dispositivos com recursos de

memória limitados, foi a rede que exigiu o menor tempo de treinamento, com apenas 282,76 segundos. Além disso, a rede ResNet50 mostrou-se uma alternativa viável, visto que apresentou boas taxas de desempenho e um tempo de treinamento de 581,81 segundos, inferior ao tempo necessário para realizar o treinamento da rede EfficientNetB7.

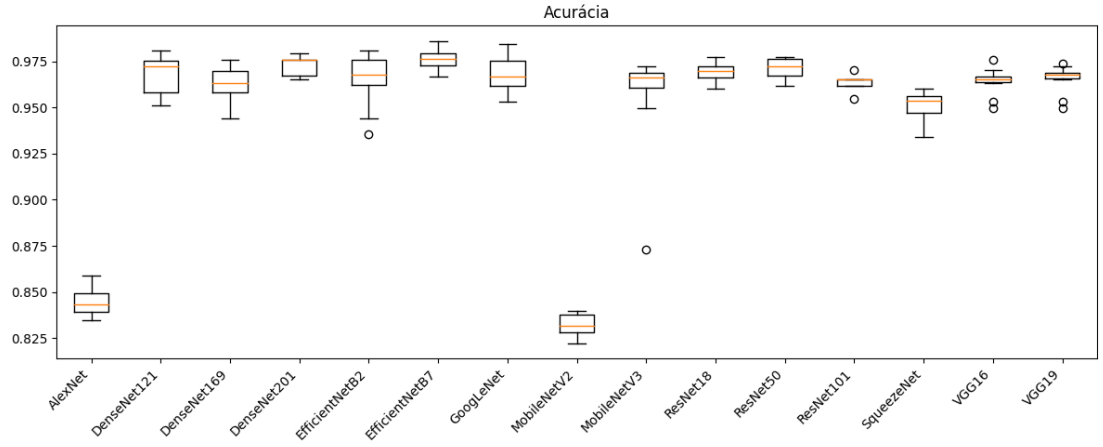
Além disso, a Figura 23 ilustra os *boxplots* com os resultados para as métricas acurácia, *F1-Score* e especificidade, referentes à classificação de tipos de tumores cerebrais. Dessa forma, ao considerar as métricas as redes como DenseNet201, EfficientNetB7 e ResNet50 se destacam com os maiores valores em todas essas métricas, em relação às demais arquiteturas. Isso sugere a eficácia das redes para a tarefa de classificação de tumores sugerindo que são eficazes para a tarefa de classificação de tumores. Em contraste, as redes AlexNet e MobileNetV2 exibem desempenho inferior, com valores menores para as métricas. Ademais, apresentam inconsistência, com variabilidade significativa em suas métricas, o que indica que essas redes podem não ser a melhor escolha para a tarefa de classificação.

A Figura 24 apresenta o boxplot para a métrica de tempo de treinamento. Em termos de tempo de treinamento, as redes AlexNet, MobileNetV2 e SqueezeNet e destacam-se por seu curto tempo de treinamento, enquanto a rede EfficientNetB7 e as variantes VGG16 e VGG19 apresentam tempos significativamente mais longos. Isso indica que redes mais complexas e profundas tendem a exigir mais tempo para treinamento. Além disso, redes como DenseNet201 e ResNet50 apresentam um equilíbrio entre tempo de treinamento e desempenho, oferecendo altas métricas de classificação com um tempo de treinamento moderado. Tais informações são essenciais para a escolha da arquitetura de rede neural mais adequada e que atende às necessidades específicas na tarefa de classificação de tumores.

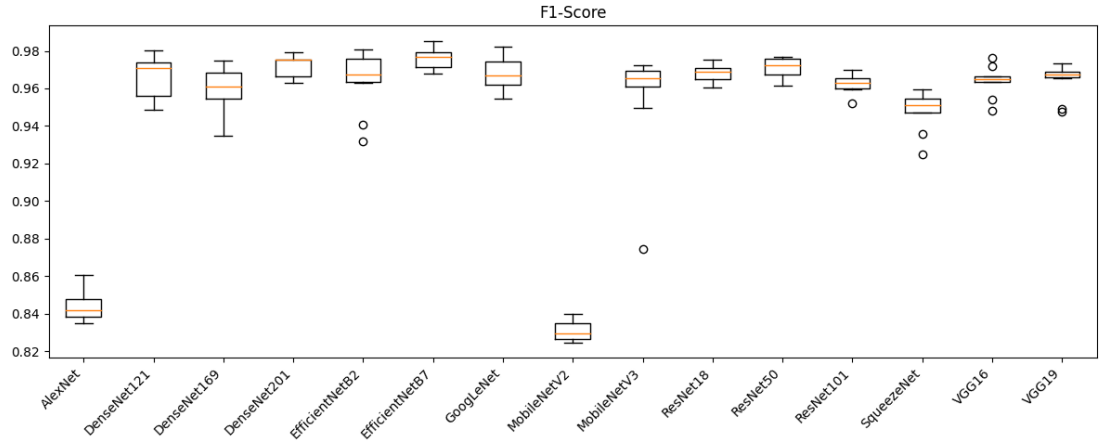
4.1.2 Testes Estatísticos

Para avaliar a significância dos resultados e identificar possíveis diferenças entre as médias das métricas utilizadas, foi conduzida uma análise abrangente por meio de testes estatísticos. Dessa forma, para validar as premissas de normalidade dos dados, foi realizado o teste de *Shapiro-wilk*. Em seguida, foi conduzido o teste de *Levene* para avaliar a homogeneidade entre populações. Entretanto, as hipóteses nulas de normalidade em todas as populações e de homocedasticidade dos dados foram rejeitadas. Logo, como os dados não atenderam aos requisitos de normalidade e homogeneidade, foram aplicados os testes não paramétricos de *Friedman* e o *pós-hoc* de *Nemenyi*. Em vista disso, o teste paramétrico da ANOVA e o *pós-hoc*

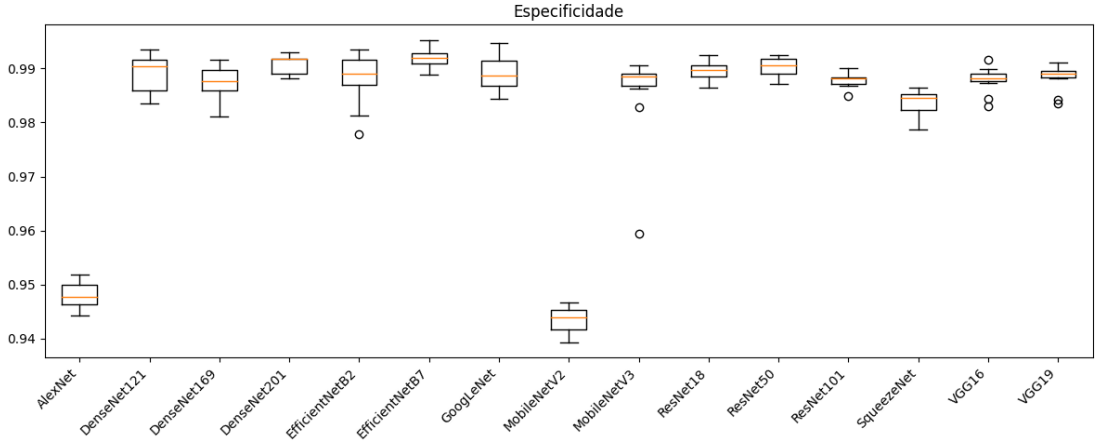
Figura 23 – Métricas da classificação de tumores cerebrais.



(a) Acurácia.



(b) F1-Score.

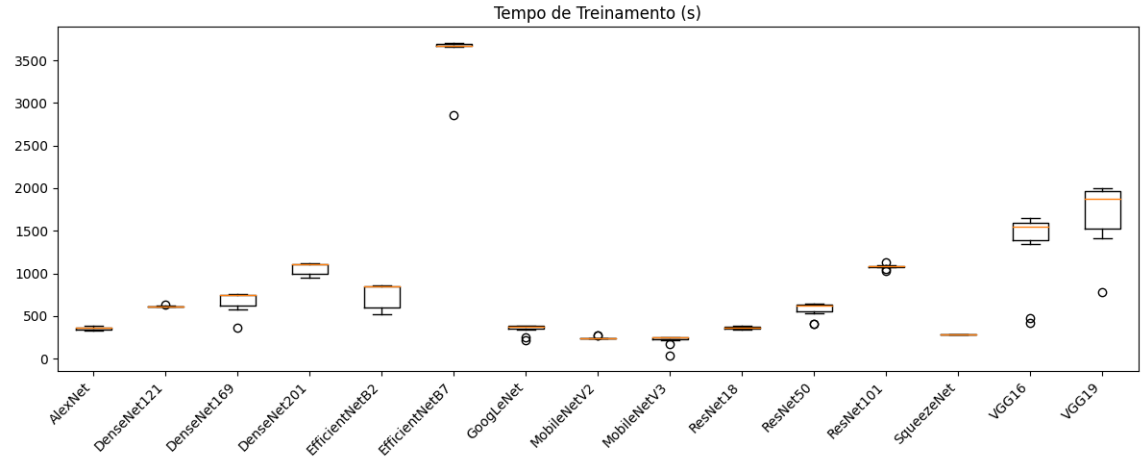


(c) Especificidade.

Fonte: Elaborado pela autora.

de *Tukey*, utilizados quando as suposições de normalidade dos dados e a homogeneidade das variâncias são atendidas, não foram contemplados.

Figura 24 – Análise de tempo de treinamento.



Fonte: Elaborado pela autora.

O teste de *Friedman* foi utilizado para determinar se existem diferenças significativas entre os valores médios das métricas de avaliação para todos os modelos de classificação analisados, além do tempo para realizar o treinamento de cada um dos modelos. A Tabela 9 exibe os resultados do teste não paramétrico de *Friedman*, com um nível de significância de 5%. É possível observar que em todos os casos obteve-se um valor p inferior a 0,05. O valor p é uma medida estatística que representa a probabilidade de observar um resultado igual ou mais extremo do que o observado, assumindo que a hipótese nula seja verdadeira. Neste contexto, um valor p inferior a 0,05 indica que deve-se rejeitar a hipótese nula, ou seja, existem diferenças significativas entre os métodos avaliados. Portanto, conclui-se que há diferenças significativas no desempenho dos modelos de classificação avaliados. Isso indica que pelo menos um dos modelos possui desempenho estatisticamente distinto em relação aos outros, em termos de métricas de avaliação.

Tabela 9 – Resultados do teste de *Friedman* para os modelos de classificação.

	Acc	**Prec**	***Recall***	**F1**	**Esp**	**Tempo**
Estatística de Teste	40,43	38,68	41,84	39,88	40,51	48.11
Valor *p*	$1{,}22e^{-07}$	$2{,}74e^{-07}$	$6{,}32e^{-07}$	$1{,}57e^{-07}$	$1{,}17e^{-07}$	$3{,}36e^{-09}$

Fonte: Elaborado pela autora.

Logo, devido à significância dos resultados obtidos no teste de Friedman, foi utilizado, em seguida, o *pós-hoc* de *Nemenyi* para inferir quais diferenças são estatisticamente relevantes. O teste foi utilizado para determinar a diferença estatística entre os pares de métodos, e, assim, avaliar o desempenho dos modelos aplicados.

A Tabela 10 apresenta a análise do teste de *Nemenyi* para a métrica acurácia, com um nível de significância de 5%. Estão destacados os resultados para qual o valor *p* é inferior a 0,05. Isso sugere que a hipótese nula deve ser rejeitada, ou seja, existem diferenças estatísticas entre os pares de métodos analisados. Pode-se observar que, para a métrica de acurácia, a arquitetura AlexNet apresenta diferenças significativas em relação às redes DenseNet201 e EfficientNetB7. Entre essas redes, a AlexNet exibe a menor pontuação de acurácia. Além disso, a rede MobileNetV2 também apresenta diferenças significativas em relação às redes DenseNet201, EfficientNetB7 e ResNet50, registrando o menor valor de acurácia entre todas as arquiteturas.

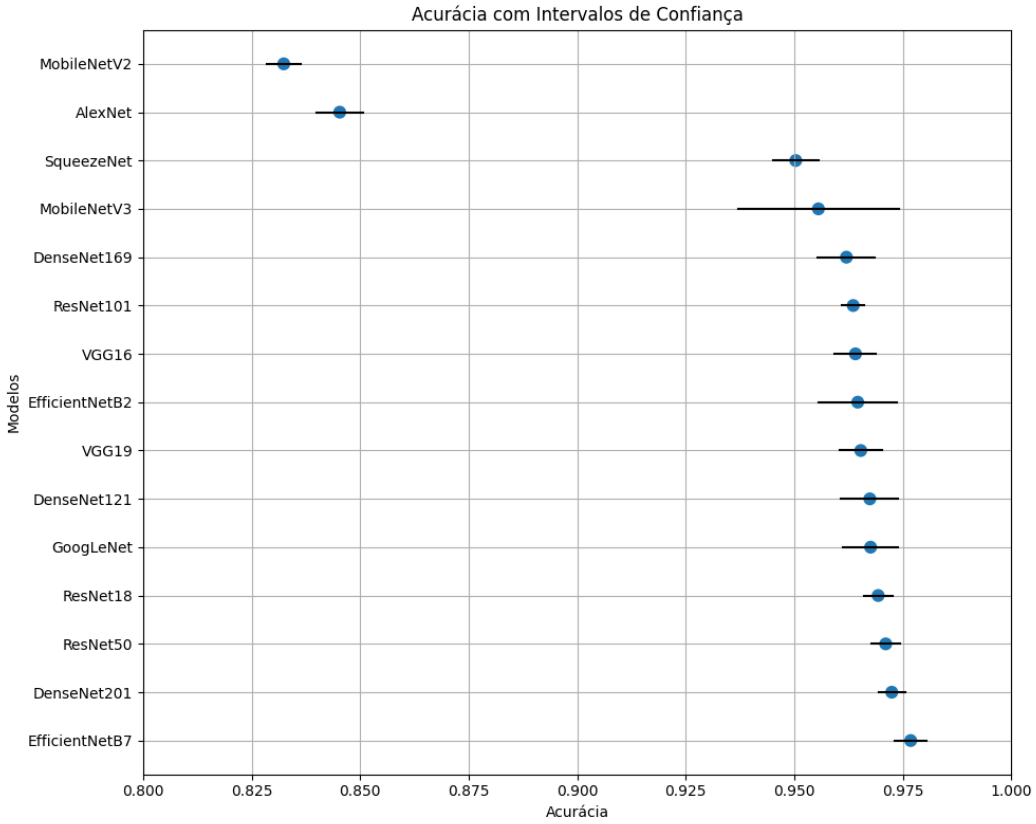
Tabela 10 – Resultados do teste de *Nemenyi* para a métrica Acurácia para os diferentes modelos de classificação.

	Acurácia														
	AlexNet	**DenseNet121**	**DenseNet169**	**DenseNet201**	**EfficientNetB2**	**EfficientNetB7**	**GoogLeNet**	**MobileNetV2**	**MobileNetV3**	**ResNet18**	**ResNet50**	**ResNet101**	**SqueezeNet**	**VGG16**	**VGG19**
AlexNet	1,0	0,31	0,87	**0,03**	0,44	**0,001**	0,41	1,0	0,85	0,21	0,07	0,92	0,99	0,79	0,58
DenseNet121	0,31	1,0	0,99	0,99	1,0	0,99	1,0	0,15	0,99	1,0	1,0	0,99	0,87	1,0	1,0
DenseNet169	0,87	0,99	1,0	0,99	1,0	0,82	1,0	0,7	1,0	0,99	0,99	1,0	0,99	1,0	1,0
DenseNet201	**0,032**	0,99	0,99	1,0	0,99	1,0	0,99	**0,009**	0,99	1,0	1,0	0,98	0,35	0,99	0,99
EfficientNetB2	0,44	1,0	1,0	0,99	1,0	0,99	1,0	0,24	1,0	1,0	0,99	0,99	0,93	1,0	1,0
EfficientNetB7	**0,001**	0,99	0,82	1,0	0,99	1,0	0,99	**0,0002**	0,84	0,99	0,99	0,73	0,06	0,89	0,97
GoogLeNet	0,4	1,0	1,0	0,99	1,0	0,99	1,0	0,21	1,0	1,0	1,0	0,99	0,91	1,0	1,0
MobileNetV2	1,0	0,15	0,7	**0,009**	0,24	**0,0002**	0,21	1,0	0,67	0,08	**0,02**	0,79	0,99	0,59	0,36
MobileNetV3	0,85	0,99	1,0	0,99	1,0	0,4	1,0	0,67	1,0	0,99	0,99	1,0	0,99	1,0	1.0
ResNet18	0,21	1,0	0,99	1,0	1,0	0,99	1,0	0,08	0,99	1,0	1,0	0,99	0,77	0,99	1,0
ResNet50	0,07	1,0	0,99	1,0	0,99	0,99	1,0	**0,02**	0,99	1,0	1.0	0,99	0,52	0,99	0,99
ResNet101	0,92	0,99	1,0	0,97	0,99	0,73	0,99	0,79	1,0	0,99	0,99	1,0	0,99	1,0	1,0
SqueezeNet	0,99	0,86	0,99	0,35	0,93	0,06	0,91	0,99	0,99	0,77	0,52	0,99	1,0	0,99	0,97
VGG16	0,79	1,0	1,0	0,99	1,0	0,89	1,0	0,59	1,0	0,99	0,99	1,0	0,99	1,0	1,0
VGG19	0,58	1,0	1,0	0,99	1,0	0,97	1,0	0,36	1,0	1,0	0,99	1,0	0,97	1,0	1,0

Fonte: Elaborado pela autora.

A Figura 25 apresenta uma comparação das acurácias médias e seus respectivos intervalos de confiança para os diferentes modelos treinados. Os pontos azuis indicam as acurácias médias, enquanto as linhas pretas horizontais representam os intervalos de confiança, que fornecem uma estimativa da incerteza em torno dessas acurácias. Dessa forma, o intervalo de confiança é um indicador da variabilidade da acurácia estimada para cada modelo.

Figura 25 – Comparação entre valores de acurácia médias e os seus respectivos intervalos de confiança para os diferentes modelos de classificação.



Fonte: Elaborado pela autora.

Em suma, intervalos de confiança mais curtos, como os observados para as redes ResNet101, DenseNet201, EfficientNetB7 e ResNet50, indicam estimativas mais precisas. Em contrapartida, intervalos de confiança mais largos, como é o caso da rede MobileNetV3, sugerem maior incerteza. Além disso, a ausência de sobreposição entre os intervalos de confiança reafirma as diferenças significativas entre os modelos. Logo, os intervalos de confiança das redes MobileNetV2 e AlexNet não se sobrepõem com os de DenseNet201, EfficientNetB7 e ResNet50, o que reforça a conclusão do teste de Nemenyi sobre a diferença significativa entre esses modelos.

Além disso, foi realizado o teste de *Nemenyi* para analisar o desempenho dos modelos

em relação à métrica *F1-Score*. Dessa forma, utilizou-se a métrica *F1-Score*, em detrimento das métricas de precisão e *recall*, devido a capacidade da métrica em fornecer uma visão equilibrada entre essas duas métricas, uma vez que o *F1-Score* é a média harmônica entre a precisão e o *recall*. Isto posto, essa análise é importante, pois, precisão e o *recall* podem fornecer informações parciais ou desequilibradas sobre o desempenho dos modelos. Portanto, ao considerar o *F1-Score*, obtém-se uma medida mais representativa da eficácia geral dos modelos em identificar corretamente as classes de interesse.

A Tabela 11 exibe o resultado para o teste de *Nemenyi* calculado para a métrica *F1-Score*. Os valores destacados, em que o valor *p* é inferior a 0,05 a um nível de significância de 5%, indica que a hipótese nula deve ser rejeitada, e consequentemente, os valores de *F1-Score* para os pares de modelos analisados são significativamente diferentes. Novamente, a arquitetura AlexNet apresenta diferenças significativas para a métrica de *F1-Score* quando comparada às redes DenseNet201 e EfficientNetB7. Ademais, a rede MobileNetV2 registra valores significativamente diferentes quando as médias da métrica *F1-Score* são analisadas em comparação aos resultados médios para as redes DenseNet201, EfficientNetB7 e ResNet50.

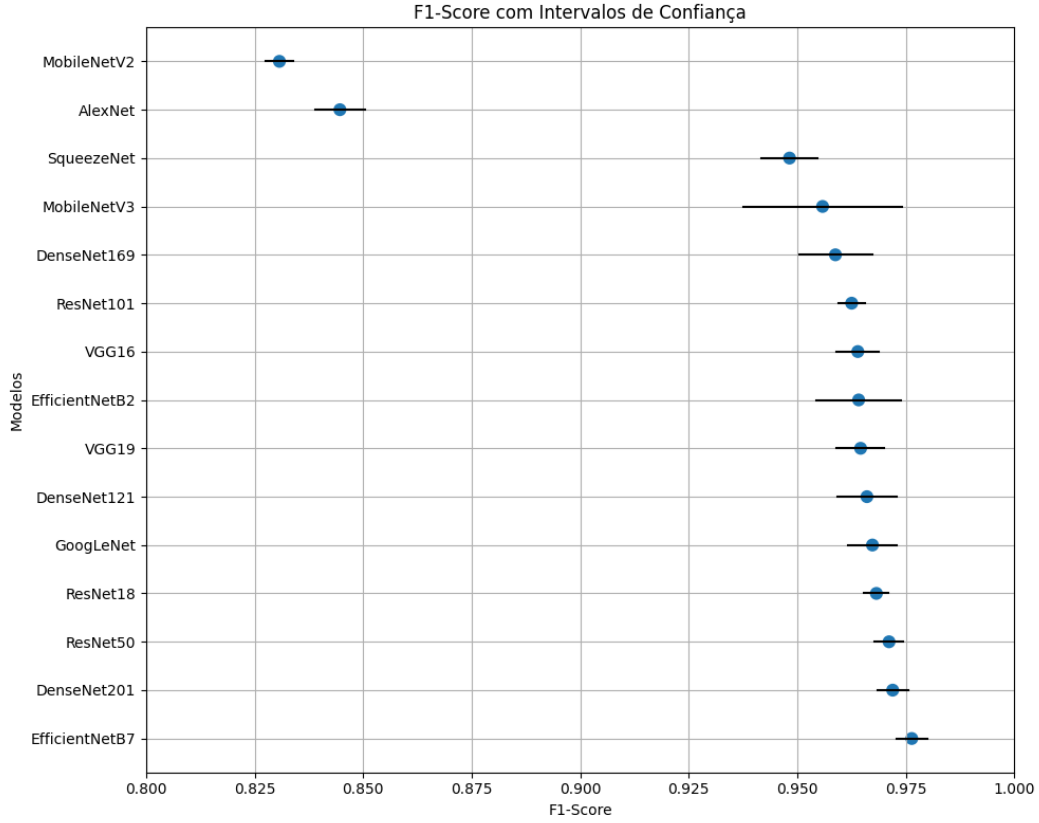
Tabela 11 – Resultados do teste de *Nemenyi* para a métrica *F1-Score* para os diferentes modelos de classificação.

	F1-Score														
	AlexNet	**DenseNet121**	**DenseNet169**	**DenseNet201**	**EfficientNetB2**	**EfficientNetB7**	**GoogLeNet**	**MobileNetV2**	**MobileNetV3**	**ResNet18**	**ResNet50**	**ResNet101**	**SqueezeNet**	**VGG16**	**VGG19**
AlexNet	1,0	0,37	0,92	**0,04**	0,37	**0,001**	0,39	1,0	0,8	0,31	0,053	0,94	0,99	0,77	0,58
DenseNet121	0,37	1,0	0,99	0,99	1,0	0,99	1,0	0,17	1,0	1,0	0,99	0,99	0,89	1,0	1,0
DenseNet169	0,93	0,99	1,0	0,98	0,99	0,72	0,99	0,78	1,0	0,99	0,98	1,0	0,99	1,0	1,0
DenseNet201	**0,04**	0,99	0,98	1,0	0,99	0,99	0,99	**0,01**	0,99	1,0	1,0	0,97	0,39	0,99	0,99
EfficientNetB2	0,37	1,0	0,99	0,99	1,0	0,99	1,0	0,17	1,0	1,0	0,99	0,99	0,89	1,0	1,0
EfficientNetB7	**0,001**	0,99	0,72	0,99	0,99	1,0	0,99	**0,0002**	0,88	0,99	0,99	0,67	0,053	0,91	0,97
GoogLeNet	0,39	1,0	0,99	0,99	1,0	0,99	1,0	0,18	1,0	1,0	0,99	0,99	0,90	1,0	1,0
MobileNetV2	1,0	0,17	0,78	**0,01**	0,17	**0,0002**	0,18	1,0	0,57	0,13	**0,01**	0,82	0,99	0,53	0,33
MobileNetV3	0,80	1,0	1,0	0,99	1,0	0,88	1,0	0,57	1,0	1,0	0,99	1,0	0,99	1,0	1,0
ResNet18	0,31	1,0	0,99	1,0	1,0	0,99	1,0	0,13	1,0	1,0	1,0	0,99	0,85	1,0	1,0
ResNet50	0,053	0,99	0,98	1,0	0,99	0,99	0,99	**0,01**	0,99	1,0	1,0	0,98	0,43	0,99	0,99
ResNet101	0,94	0,99	1,0	0,97	0,99	0,67	0,99	0,82	1,0	0,99	0,98	1,0	0,99	1,0	1,0
SqueezeNet	0,99	0,89	0,99	0,39	0,89	0,053	0,90	0,99	0,99	0,85	0,43	0,99	1,0	0,99	0,96
VGG16	0,77	1,0	1,0	0,99	1,0	0,90	1,0	0,53	1,0	1,0	0,99	1,0	0,99	1,0	1,0
VGG19	0,58	1,0	1,0	0,99	1,0	0,97	1,0	0,33	1,0	1,0	0,99	1,0	0,96	1,0	1,0

Fonte: Elaborado pela autora.

A Figura 26 exibe os resultados para a métrica *F1-Score* e os seus respectivos intervalos de confiança. Dessa forma, é possível observar que as arquiteturas MobileNetV2 e AlexNet apresentam as menores taxas, enquanto, as redes ResNet50, DenseNet201 e EfficientNetB7 obtem os maiores valores para a métrica *F1-Score*. Além disso, a não sobreposição entre os intervalos de confiança dessas métricas ressalta que existem diferenças significativas entre os modelos, conforme analisado pelo teste estatístico de *Nemenyi*.

Figura 26 – Comparação entre valores de *F1-Score* médias e os seus respectivos intervalos de confiança para os diferentes modelos de classificação.



Fonte: Elaborado pela autora.

Além disso, a Tabela 12 apresenta os resultados do teste pós-hoc de Nemenyi para a métrica de especificidade. A métrica avalia a taxa de acertos para a classe negativa, sendo interessante na avaliação de casos sem tumor. De acordo com o teste de Nemenyi, foram identificadas diferenças significativas ao comparar as médias de especificidade da rede AlexNet com as redes DenseNet201, EfficientNetB7 e ResNet50. É possível observar que as redes DenseNet201, EfficientNetB7 e ResNet50 apresentaram os maiores valores de especificidade, indicando uma maior capacidade de avaliar com precisão os casos negativos. Em contrapartida, a rede AlexNet não apresenta um desempenho tão sofisticado quanto as demais analisadas.

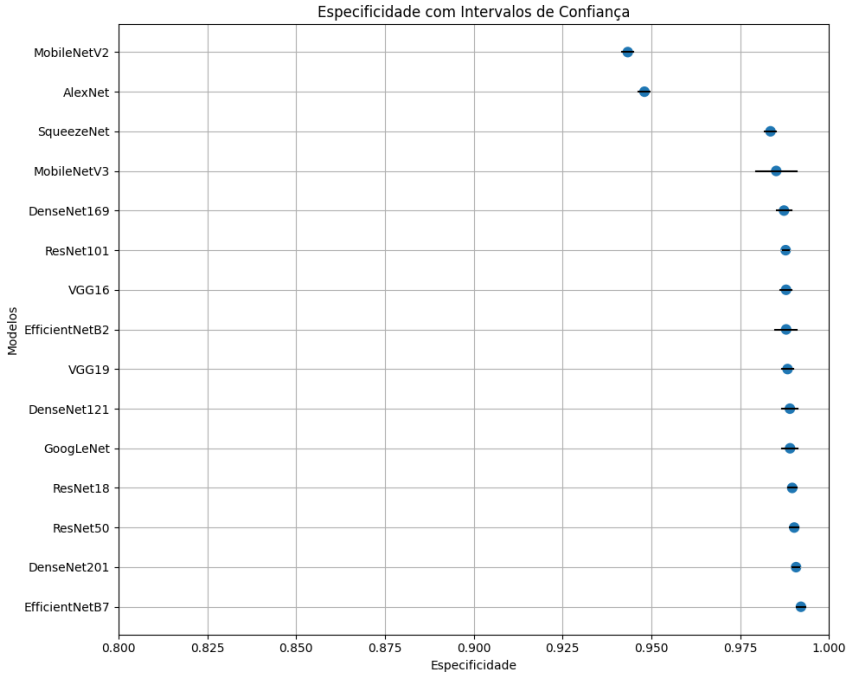
Tabela 12 – Resultados do teste de *Nemenyi* para a métrica Especificidade para os diferentes modelos de classificação.

Especificidade															
	AlexNet	**DenseNet121**	**DenseNet169**	**DenseNet201**	**EfficientNetB2**	**EfficientNetB7**	**GoogLeNet**	**MobileNetV2**	**MobileNetV3**	**ResNet18**	**ResNet50**	**ResNet101**	**SqueezeNet**	**VGG16**	**VGG19**
AlexNet	1,0	0,18	0,79	**0,007**	0,34	**0,0003**	0,28	0,8	0,8	0,08	**0,03**	0,85	0,99	0,70	0,44
DenseNet121	0,18	1,0	0,99	0,99	1,0	0,99	1,0	0,99	0,99	1,0	0,99	0,99	0,76	0,99	0,99
DenseNet169	0,79	0,99	1,0	0,97	0,99	0,76	0,99	1,0	1,0	0,99	0,99	1,0	0,99	1,0	0,99
DenseNet201	**0,007**	0,99	0,97	1,0	0,99	0,99	0,99	0,97	0,97	0,99	1,0	0,95	0,16	0,98	0,99
EfficientNetB2	0,34	1,0	0,99	0,99	1,0	0,98	1,0	0,99	0,99	0,99	0,99	0,99	0,89	0,99	1,0
EfficientNetB7	**0,0003**	0,99	0,76	0,99	0,98	1,0	0,98	0,74	0,74	0,99	0,99	0,68	**0,02**	0,83	0,95
GoogLeNet	0,28	1,0	0,99	0,99	1,0	0,98	1,0	0,99	0,99	0,99	0,99	0,99	0,85	0,99	1,0
MobileNetV2	0,80	0,99	1,0	0,97	0,99	0,74	0,99	1,0	1,0	0,99	0,99	1,0	0,99	1,0	0,99
MobileNetV3	0,80	0,99	1,0	0,97	0,99	0,74	0,99	1,0	1,0	0,99	0,99	1,0	0,99	1,0	0,99
ResNet18	0,08	1,0	0,99	0,99	0,99	0,99	0,99	0,99	0,99	1,0	1,0	0,99	0,58	0,99	0,99
ResNet50	**0,03**	0,99	0,99	1,0	0,99	0,99	0,99	0,99	0,99	1,0	1,0	0,99	0,36	0,99	0,99
ResNet101	0,85	0,99	1,0	0,95	0,99	0,68	0,99	1,0	1,0	0,99	0,99	1,0	0,99	1,0	0,99
SqueezeNet	0,99	0,76	0,99	0,16	0,89	**0,02**	0,85	0,99	0,99	0,58	0,36	0,99	1,0	0,99	0,94
VGG16	0,70	0,99	1,0	0,98	0,99	0,83	0,99	1,0	1,0	0,99	0,99	1,0	0,99	1,0	1,0
VGG19	0,44	0,99	0,99	0,99	1,0	0,95	1,0	0,99	0,99	0,99	0,99	0,99	0,94	1,0	1,0

Fonte: Elaborado pela autora.

A Figura 27 apresenta os valores da métrica de especificidade para cada arquitetura de classificação, juntamente com seus respectivos intervalos de confiança. Entre as redes analisadas, a MobileNetV2 obteve a menor taxa de especificidade, seguida pela AlexNet. Embora a MobileNetV2 tenha apresentado as menores taxas e os intervalos de confiança não se sobreponham quando comparada com outras redes, o teste de Nemenyi não identificou diferenças significativas entre a MobileNetV2 e as demais arquiteturas analisadas.

Figura 27 – Comparação entre valores de especificidade e os seus respectivos intervalos de confiança para os diferentes modelos de classificação.



Fonte: Elaborado pela autora.

A Tabela 13 apresenta os resultados obtidos pelo teste pós-hoc de Nemenyi, analisando o tempo de treinamento dos modelos de classificação. As redes MobileNetV3 e MobileNetV2, que demonstraram os menores tempos de treinamento, obtiveram diferenças significativas em comparação com a maioria das redes avaliadas. Vale ressaltar que, apesar do menor tempo de treinamento, a MobileNetV3 obteve pontuações quantitativas superiores às da MobileNetV2. Em contrapartida, a rede EfficientNetB7, devido à sua maior complexidade e ao número elevado de camadas, registrou o maior tempo de treinamento, embora tenha alcançado as melhores taxas nas métricas quantitativas. Outras alternativas viáveis incluem as redes DenseNet201 e ResNet50, que mostraram boas métricas quantitativas e não apresentaram diferenças significativas em relação ao tempo de treinamento quando comparadas à maioria das arquiteturas de classificação.

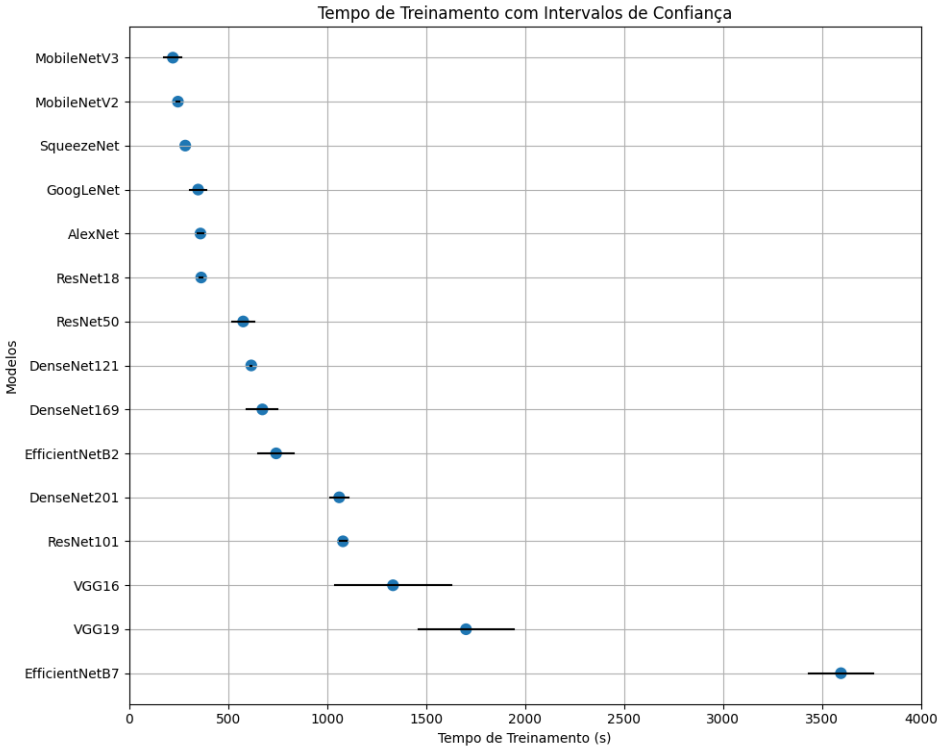
Tabela 13 – Resultados do teste de *Nemenyi* para a métrica de Tempo de Treinamento para os diferentes modelos de classificação.

Tempo de Treinamento															
	AlexNet	**DenseNet121**	**DenseNet169**	**DenseNet201**	**EfficientNetB2**	**EfficientNetB7**	**GoogLeNet**	**MobileNetV2**	**MobileNetV3**	**ResNet18**	**ResNet50**	**ResNet101**	**SqueezeNet**	**VGG16**	**VGG19**
AlexNet	1,0	0,99	0,99	0,53	0,98	**0,02**	1,0	0,99	0,99	1,0	0,99	0,60	0,99	0,48	0,14
DenseNet121	0,99	1,0	1,0	0,99	0,99	0,61	0,99	0,62	0,62	0,99	1,0	0,99	0,93	0,99	0,92
DenseNet169	0,99	1,0	1,0	0,99	1,0	0,75	0,99	0,47	0,47	0,99	1,0	0,99	0,86	0,99	0,97
DenseNet201	0,53	0,99	0,99	1,0	0,99	0,99	0,48	**0,01**	**0,01**	0,53	0,99	1,0	0,12	1,0	0,99
EfficientNetB2	0,98	0,99	1,0	0,99	1,0	0,84	0,98	0,35	0,35	0,98	0,99	0,99	0,78	0,99	0,98
EfficientNetB7	**0,02**	0,61	0,75	0,99	0,84	1,0	**0,01**	**0,00001**	**0,00001**	**0,02**	0,60	0,99	**0,0007**	0,99	0,99
GoogLeNet	1,0	0,99	0,99	0,48	0,98	**0,01**	1,0	0,99	0,99	1,0	0,99	0,55	0,99	0,43	0,12
MobileNetV2	0,99	0,62	0,47	**0,01**	0,35	**0,00001**	0,99	1,0	1,0	0,99	0,64	**0,02**	0,99	**0,011**	**0,0006**
MobileNetV3	0,99	0,62	0,47	**0,01**	0,35	**0,00001**	0,99	1,0	1,0	0,99	0,64	**0,02**	0,99	**0,01**	**0,0006**
ResNet18	1,0	0,99	0,99	0,53	0,98	**0,02**	1,0	0,99	0,99	1,0	0,99	0,60	0,99	0,49	0,14
ResNet50	0,99	1,0	1,0	0,99	0,99	0,60	0,99	0,64	0,64	0,99	1,0	0,99	0,94	0,99	0,92
ResNet101	0,60	0,99	0,99	1,0	0,99	0,99	0,55	**0,02**	**0,02**	0,60	0,99	1,0	0,16	1,0	0,99
SqueezeNet	0,99	0,93	0,86	0,12	0,78	**0,0007**	0,99	0,99	0,99	0,99	0,94	0,16	1,0	0,104	**0,01**
VGG16	0,48	0,99	0,99	1,0	0,99	0,99	0,43	**0,01**	**0,01**	0,49	0,99	1,0	0,104	1,0	0,99
VGG19	0,14	0,92	0,97	0,99	0,98	0,99	0,12	**0,0006**	**0,0006**	0,14	0,92	0,99	**0,01**	0,99	1,0

Fonte: Elaborado pela autora.

Por fim, a Figura 28 apresenta os diferentes tempos de treinamento para cada modelo de classificação e os seus respectivos intervalos de confiança. As redes MobileNetV3 e MobileNetV2 obtiveram os menores tempos de treinamento, resultado da sua estrutura mais leve e do número reduzido de camadas. Entretanto, como já mencionado, embora a MobileNetV2 exija menos tempo de treinamento, apresentou as menores pontuações para as métricas de avaliação, assim como a AlexNet. Em contrapartida, a rede EfficientNetB7, que obteve as melhores taxas para as métricas quantitativas, exigiu o maior tempo de treinamento, o que reflete um custo computacional mais elevado, devido à complexidade e estrutura da rede.

Figura 28 – Comparação entre valores de tempo de treinamento e os seus respectivos intervalos de confiança para os diferentes modelos de classificação.



Fonte: Elaborado pela autora.

4.1.3 Matriz de Confusão

Para realizar uma análise detalhada das taxas de acerto e erro, evidenciando a eficácia das redes em classificar corretamente as amostras em suas respectivas categorias, foram criadas as matrizes de confusão para os modelos. A matriz de confusão é uma ferramenta essencial para compreender o comportamento do modelo em cada classe específica, destacando pontos fortes e áreas que necessitam de melhorias. A Figura 29 exibe as matrizes de confusão para os modelos

EfficientNetB7 e DenseNet201, que obtiveram as melhores pontuações em termos de métricas, como também, para as redes AlexNet e MobileNetV2, que tiveram os menores resultados.

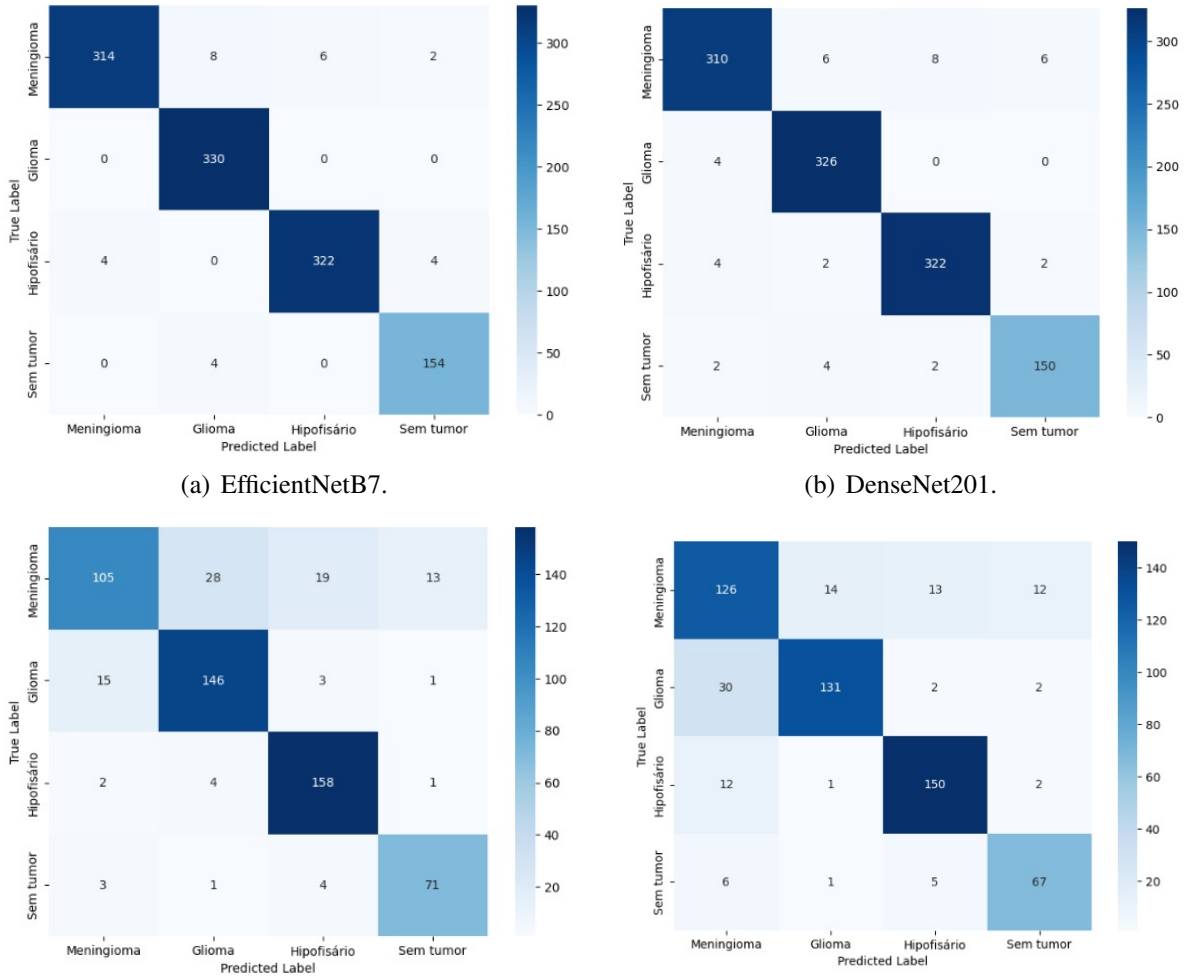
Destaca-se a matriz de confusão que resume os resultados de classificação para a rede EfficientNetB7 (Figura 29(a)). O classificador obteve um desempenho significativo em classificar as quatro classes, com elevado número de verdadeiros positivos e baixo valor de falsos positivos. Em suma, o modelo EfficientNetB7 classificou corretamente 314 instâncias para a classe Meningioma, 330 instâncias para a classe Glioma, 322 instâncias para a classe Hipofisário e 154 para a classe sem tumor. Ademais, a rede DenseNet201 (Figura 29(b)) apresentou desempenho comparável, classificando corretamento 310 exemplos para a classe Meningioma, 326 instâncias para a classe Glioma, 322 instâncias para a classe Hipofisário e 150 para a classe sem tumor. Em contrapartida, as redes AlexNet (Figura 29(c)) e MobileNetV2 (Figura 29(d)) apresentaram uma elevada taxa de falsos positivos, especialmente, para as classes Meningioma e Glioma.

4.1.4 Monitoramento de Desempenho

Ademais, a Figura 30 ilustra os gráficos com o comportamento das acurácias e das perdas obtidas durante o treinamento das redes neurais EfficientNetB7 e DenseNet201 ao longo de 50 épocas. A arquitetura EfficientNetB7 (Figura 30(a)) apresenta uma acurácia de treinamento que aumenta rapidamente e se estabiliza próxima de 1,0, enquanto a acurácia de validação se estabiliza em torno de 0,9. Além disso, a perda de treinamento diminui rapidamente e estabiliza em valores muito baixos. Já a perda de validação diminui de forma mais lenta e se estabiliza próximo de 0,2. A rede DenseNet201 (Figura 30(b))) apresenta comportamento similar , com ambas acurácias de treinamento e validação se aproximando de 1,0, e perdas de treinamento e validação estabilizando em valores baixos, inferiores a 0,2.

Em seguida, a Figura 31 apresenta as curvas de acurácia e perda dos treinamentos das redes AlexNet e MobileNetV2 ao longo de 50 épocas. As redes AlexNet e MobileNetV2, em comparação às arquiteturas EfficientNetb7 e DenseNet201, apresentam maiores variações e diferenças entre as curvas de treinamento e validação. A acurácia de treinamento da AlexNet se estabiliza em torno de 0,8, enquanto a acurácia de validação possui valores menores. As perdas de treinamento e validação também diminuem, mas apresentam maiores flutuações. A MobileNetV2 segue um padrão similar, com acurácias estabilizando em torno de 0,8 e perdas se estabilizando em valores mais altos para validação. Em suma, as redes EfficientNetb7 e

Figura 29 – Matriz de Confusão dos modelos EfficientNetB7, DenseNet201, AlexNet e MobileNetV2 para a classificação de tumores cerebrais.



(a) EfficientNetB7. (b) DenseNet201.

(c) AlexNet. (d) MobileNetV2.

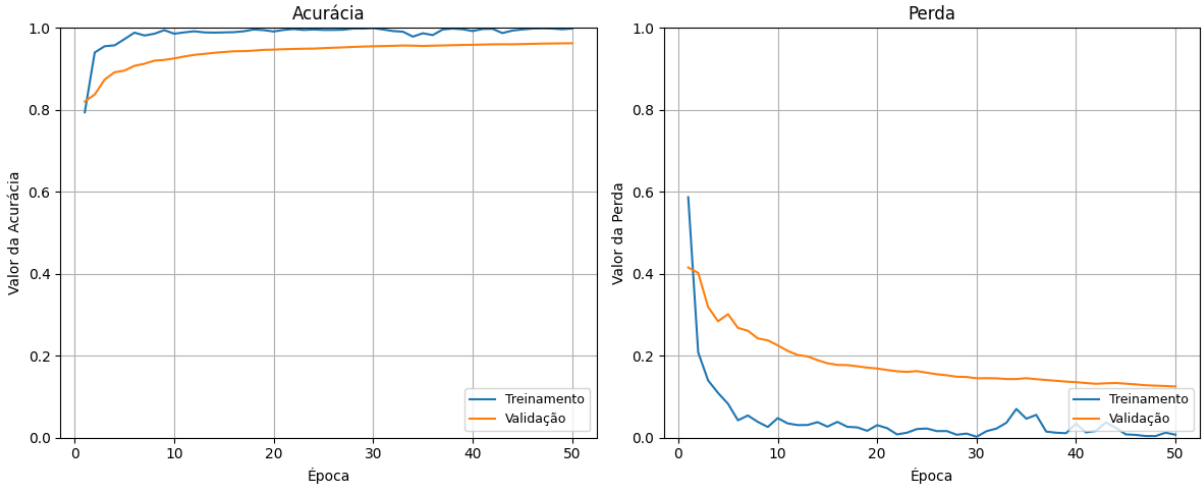
Fonte: Elaborado pela autora.

DenseNet201 apresentam menor diferença entre perdas de treinamento e validação, indicando menor *overfitting*, enquanto AlexNet e MobileNetV2 apresentam maior variabilidade, sugerindo maior potencial de *overfitting*.

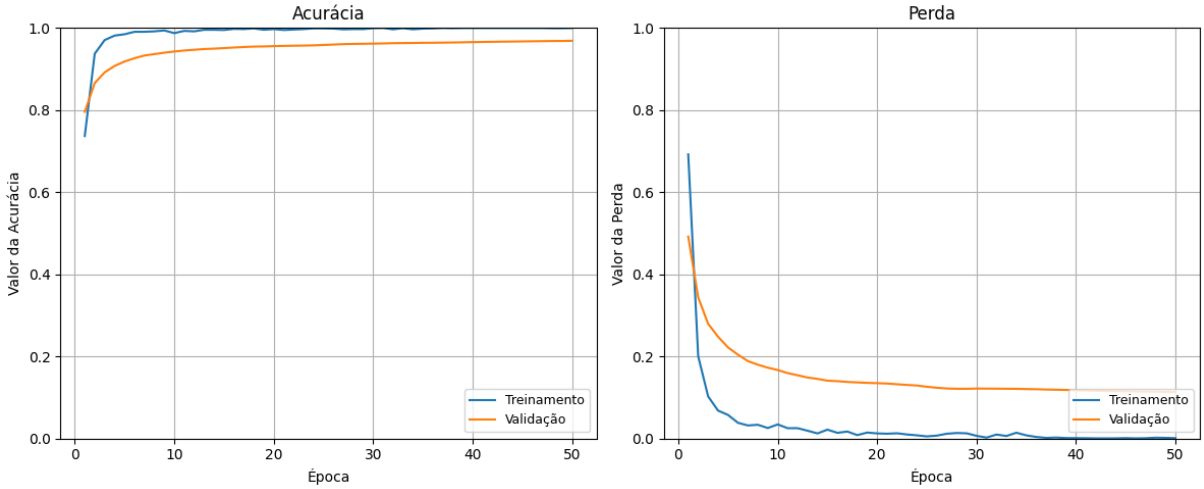
4.1.5 Validação Externa

Por fim, foi realizado o procedimento de validação externa para avaliar a capacidade de generalização das arquiteturas DenseNet201, EfficientNetB7 e ResNet50. A Tabela 14 apresenta os resultados dos modelos avaliados. As arquiteturas apresentaram resultados semelhantes e competitivos na validação externa para os conjuntos de dados Figshare e Br35H. A rede EfficientNetB7, apresentou as melhores pontuações para as métricas analisadas, com 99,01% de acurácia, precisão de 98,79%, *recall* de 98,91%, *F1-Score* 98,85% e 99,66% de

Figura 30 – Acurácias e perdas obtidas pelos modelos durante o treinamento para as redes EfficientNetB7 e DenseNet201.



(a) EfficientNetB7.



(b) DenseNet201.

Fonte: Elaborado pela autora.

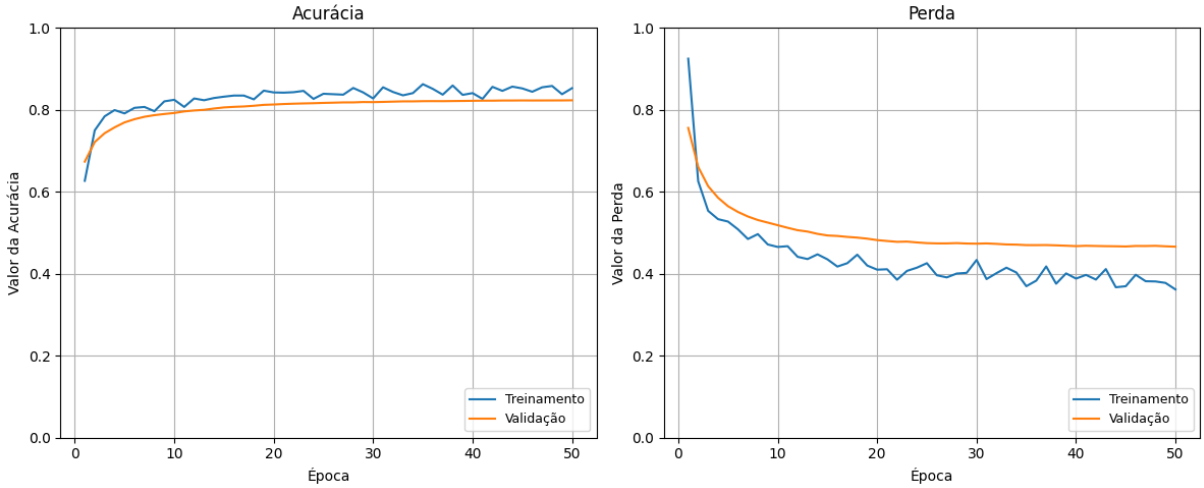
especificidade. Vale ressaltar que a validação externa desempenha um papel fundamental na comparação de arquiteturas CNN, pois simula situações do mundo real, permitindo analisar qual modelo melhor generaliza para novas amostras. Os resultados da validação externa foram possivelmente melhores devido às características mais simples do banco de dados.

Tabela 14 – Resultado da validação externa da classificação de tumores cerebrais.

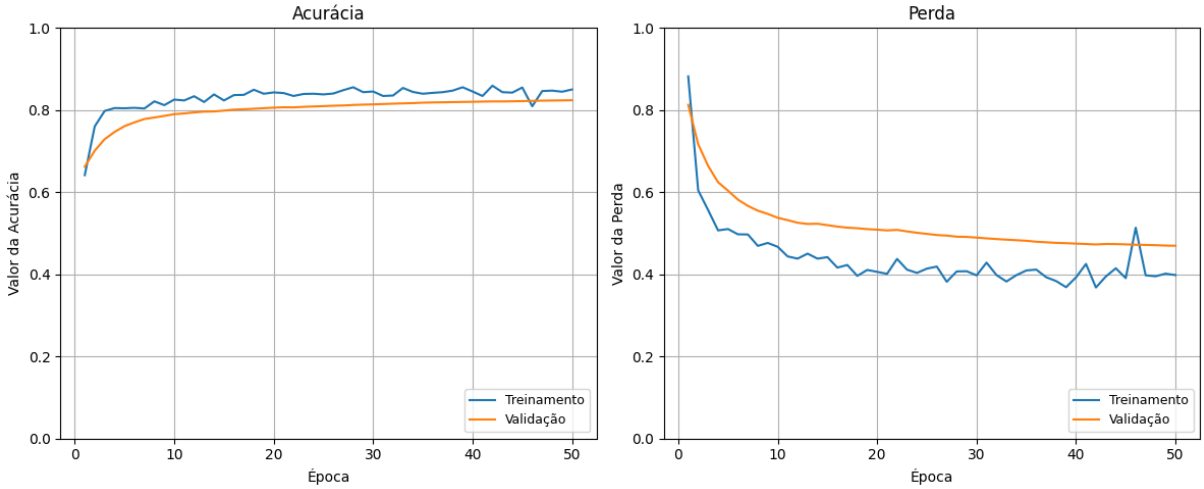
Arquitetura	**Acc (%)**	**Prec (%)**	***Recall* (%)**	**F1 (%)**	**Esp (%)**
DenseNet201	97,96	97,29	97,92	97,58	99,29
EfficientNetB7	**99,01**	**98,79**	**98,91**	**98,85**	**99,66**
ResNet50	97,85	97,23	98,11	97,63	99,23

Fonte: Elaborado pela autora.

Figura 31 – Acurácias e perdas obtidas pelos modelos durante o treinamento para as redes AlexNet e MobileNetV2.



(a) AlexNet.



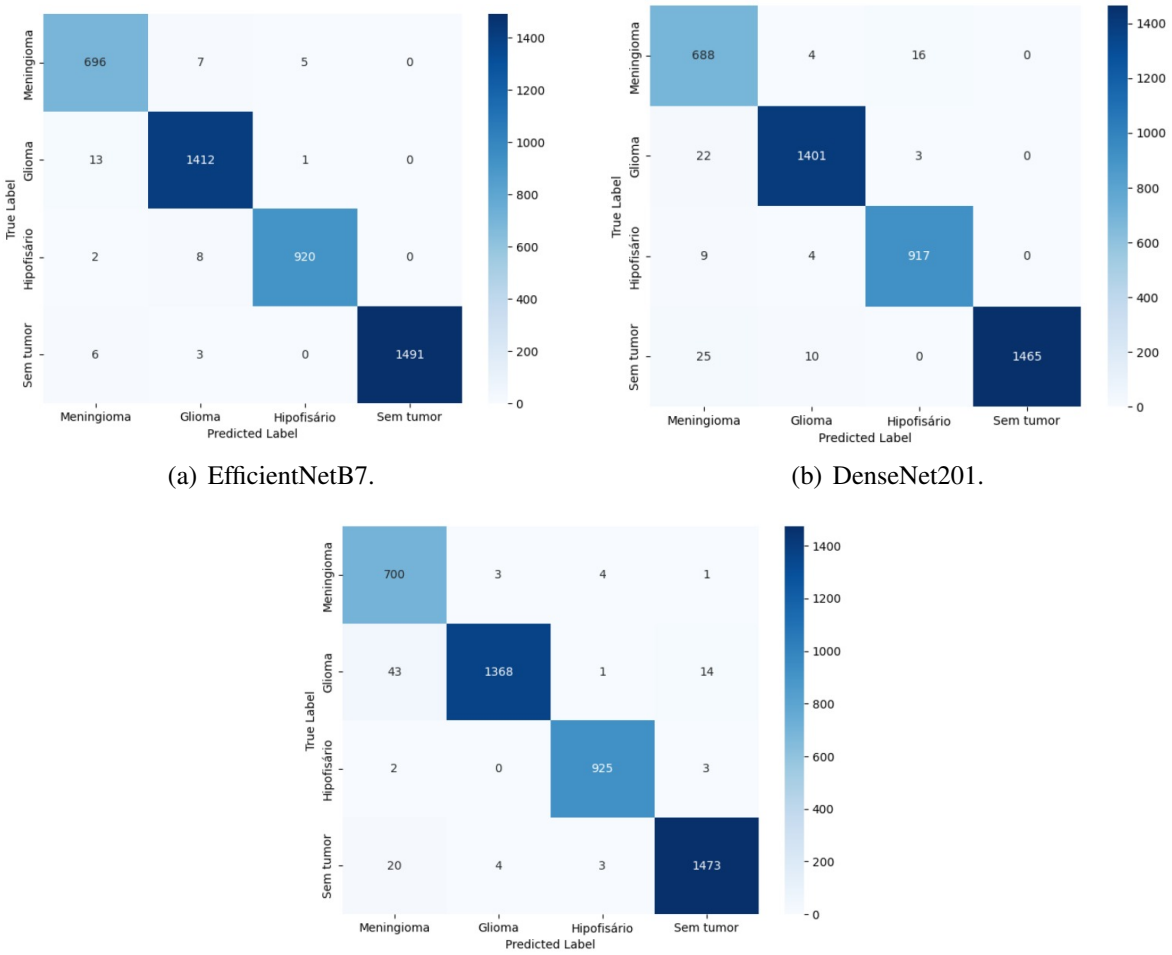
(b) MobileNetV2.

Fonte: Elaborado pela autora.

A Figura 32 exibe a matriz de confusão para o modelo EfficientNetB7 durante a validação externa. Em resumo, o modelo demonstrou um desempenho robusto ao classificar as quatro classes, com valores de verdadeiros positivos: 696 instâncias para a classe Meningioma, 1412 instâncias para a classe Glioma, 920 instâncias para a classe Hipofisário e 1491 para a classe sem tumor.

Os resultados revelam um potencial nos modelos DenseNet201, EfficientNetB7 e ResNet50 para a classificação de tumores cerebrais. No geral, não foram observadas diferenças estatisticamente significativas entre os três modelos. No entanto, destaca-se que a rede EfficientNetB7 apresentou as melhores taxas para todas as métricas avaliadas. Isso evidencia a capacidade do modelo em classificar de forma precisa os tipos de tumores e diferenciá-los de

Figura 32 – Matriz de Confusão para validação externa.



(a) EfficientNetB7.

(b) DenseNet201.

(c) ResNet50.

Fonte: Elaborado pela autora.

casos sem tumor.

4.2 Segmentação

Para a análise dos resultados da segmentação de tumores cerebrais em imagens de RM, foram empregadas medidas quantitativas como acurácia, *F1-Score*, Interseção sobre União (IoU), Distância de *Hausdorff* (DH) e tempo de treinamento. A metodologia adotada baseou-se em arquiteturas do tipo encoder-decoder. Nesse contexto, utilizou-se as redes codificadoras DenseNet201, EfficientNetB7 e ResNet50, combinadas com as redes decodificadoras UNet, UNet++ e FPN. Para realizar o treinamento das arquiteturas fez-se uso da base de dados *Figshare*, composta por imagens de ressonância magnética e das máscaras tumorais correspondentes (CHENG, 2017).

4.2.1 Análise de Métricas

A análise das métricas quantitativas é fundamental para avaliar a confiabilidade e eficácia dos modelos na segmentação de tumores cerebrais. Portanto, uma alta taxa de acurácia indica que a maioria dos pixels foi corretamente classificada, entretanto, é essencial considerar outras métricas para garantir que o modelo não esteja apenas identificando uma única classe, por exemplo, os pixels de fundo. O *F1-Score* representa a média harmônica entre o *recall* (capacidade de detectar todos os pixels do tumor) e a precisão (quantidade de pixels segmentados que realmente pertencem ao tumor). Ademais, um alto valor de IoU sugere uma sobreposição adequada entre o tumor real e o segmentado. Por fim, um baixo valor de DH indica que as bordas da segmentação prevista estão próximas das bordas do tumor real, o que reflete uma boa precisão na detecção das margens do tumor.

A Tabela 15 apresenta os resultados da segmentação da área tumoral para as arquiteturas utilizadas. No geral, as arquiteturas alcançaram resultados promissores na tarefa de segmentação de tumores cerebrais. Em especial, a combinação das redes FPN e EfficientNetB7 apresentou as melhores pontuações em termos de métricas, com 99,52% de acurácia, 85,23% para a métrica *F1-Score*, 74,29% para IoU e 4,56 para a distância de Hausdorff. Além disso, as redes UNet e ResNet50, UNet++ e DenseNet e UNet++ e EfficientNetB7 apresentam valores de desempenho semelhantes, com *F1-Scores* de 83,62%, 83,96% e 84,02%, respectivamente.

Entretanto, as métricas das redes FPN e DenseNet201 e FPN e ResNet50 obtiveram os menores valores em termos de métrica. Especificamente, a combinação das redes FPN e DenseNet201 obteve uma acurácia 98,11%, todavia, apresentou 64,23% para a métrica de *F1-Score*, 49,92% para a métrica IoU e 5,69 para a DH. Por outro lado, as arquiteturas FPN e ResNet50 alcançaram valores de 98,11% para a métrica de acurácia, 46,75% para o *F1-Score*, 30,65% para a IoU e 6,43 para a DH. Logo, embora para a métrica de acurácia as redes apresentem desempenho comparável, para as demais métricas existem diferenças mais significativas.

No contexto de tempo de treinamento, a rede UNet++ combinada com a rede DenseNet201 e com a arquitetura EfficientNetB7, apresentaram os maiores tempos de treinamento, com 7503,98 segundos e 7841,04 segundos, respectivamente. Em contrapartida, a rede FPN e ResNet50 obteve o menor tempo de treinamento, com 189,71 segundos. Entretanto, essa arquitetura apresentou as menores taxas para as métricas analisadas. Já a rede FPN e EfficientNetB7, que apresentaram as melhores pontuações, foi treinada em 6101,22 segundos. Os resultados evidenciam o equilíbrio entre o tempo de treinamento e a qualidade das métricas, destacando

Tabela 15 – Resultados da segmentação de tumores cerebrais.

Arquitetura		Acc (%)	F1 (%)	IoU (%)	DH	Tempo (s)
UNet	**DenseNet201**	97,52 ± 5,89	77,11 ± 21,08	66,11 ± 19,61	5,26 ± 1,94	3005,65 ± 653,78
	EfficientNetB7	98,04 ± 4,3	77,22 ± 19,88	65,95 ± 18,9	5,26 ± 1,81	5915,31 ± 1647,72
	ResNet50	99,48 ± 0,03	83,62 ± 1,4	71,88 ± 2,06	4,66 ± 0,08	2154,21 ± 13,8
UNet++	**DenseNet201**	99,48 ± 0,06	83,96 ± 2,01	72,41 ± 2,98	4,63 ± 0,16	7503,98 ± 25,37
	EfficientNetB7	99,48 ± 0,04	84,02 ± 1,15	72,47 ± 1.71	4,63 ± 0,13	7841,04 ± 73,79
	ResNet50	99,31 ± 0,16	77,24 ± 7,47	63,48 ± 9,12	4,97 ± 0,3	5221,59 ± 2046,18
FPN	**DenseNet201**	98,59 ± 1,12	64,23 ± 17,8	49,92 ± 19.97	5,69 ± 1,09	1286,28 ± 1131,57
	EfficientNetB7	**99,52 ± 0,05**	**85,23 ± 1,22**	**74,29 ± 1,85**	**4,56 ± 0,11**	6101,22 ± 21,55
	ResNet50	98,11 ± 0,41	46,75 ± 5,07	30,65 ± 4,36	6,43 ± 0,38	**189,71 ± 4,28**

Fonte: Elaborado pela autora.

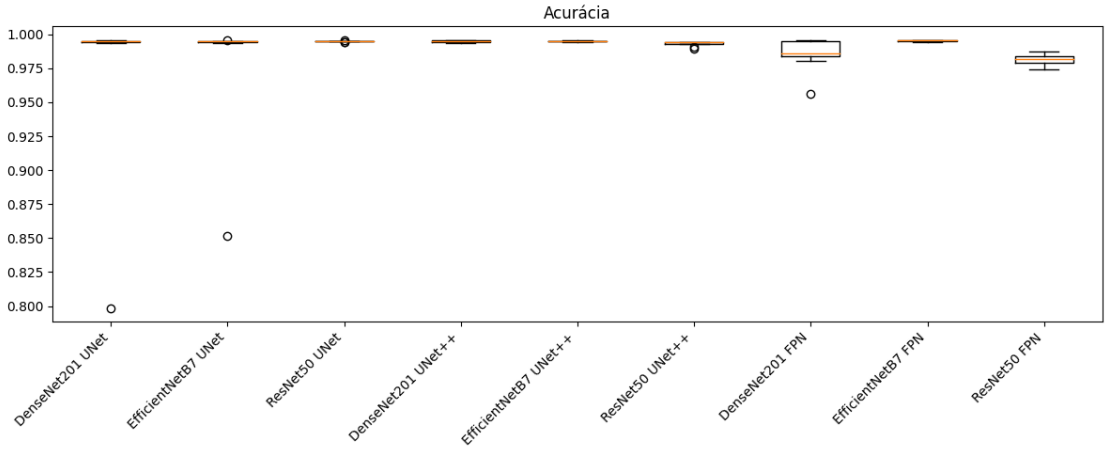
que a busca por uma maior precisão pode exigir um maior custo computacional.

Para analisar a distribuição dos resultados e comparar os diferentes modelos de segmentação treinados, foram criados os boxplots das métricas analisadas para cada modelo. A Figura 33 apresenta os *boxplots* comparativos para as métricas de acurácia e *F1-Score* para as diferentes arquiteturas utilizadas na segmentação de tumores cerebrais. Desse modo, ao considerar a métrica de acurácia (Figura 33(a)), todas as redes apresentam pontuações próximas a 100%. Vale ressaltar que a combinação das redes FPN e ResNet50 apresenta a menor acurácia dentre as redes analisadas.

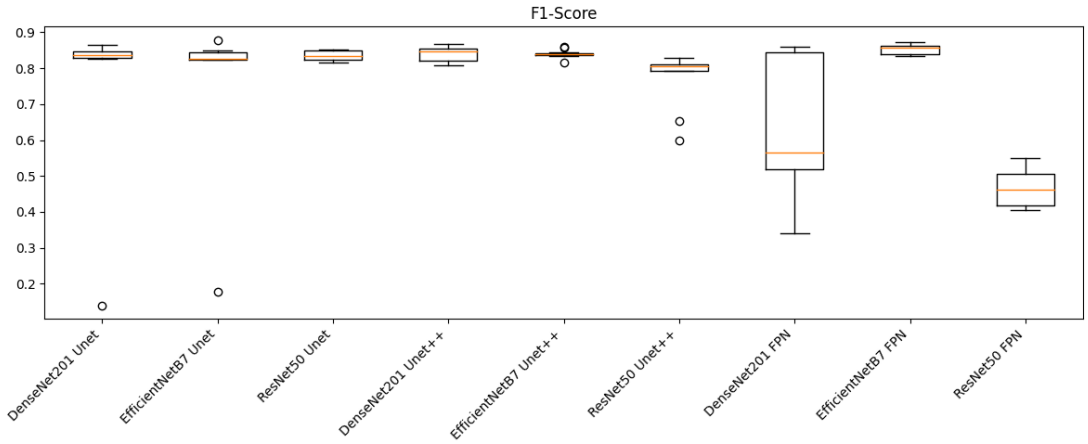
Em relação à métrica de *F1-Score* (Figura 33(b)) a combinação das redes UNet++ com EfficientNetB7 e FPN com EfficientNetB7 apresentam valores elevados e consistentes, o que demonstra o equilíbrio entre precisão e o *recall* na segmentação de tumores. Ademais, a rede FPN com ResNet50 apresentaram as taxas mais baixas para essa métrica. Por fim, é possível observar tabmém que a rede DenseNet201 com FPN apresenta uma variabilidade significativa, o que pode sugerir uma maior instabilidade no desempenho dessa rede.

Além disso, a Figura 34 exibe as comparações de desempenho entre as diferentes arquiteturas em relação às métricas de Interseção sobre União (IoU) e Distância de *Hausdorff*. A priori a métrica de Interseção sobre União (Figura 34(a)), avalia a sobreposição entre as previsões do modelo e o as máscaras verdadeiras. Novamente, é possível observar que a combinação FPN com ResNet50 apresenta os menores valores para a métrica IoU. Em contrapartida, redes como UNet ResNet50, UNet++ DenseNet201 e FPN EfficientNetB7 apresentam altas taxas para a métrica em questão com baixa variabilidade, o que pode indicar um desempenho consistente e preciso. Por outro lado, o modelo FPN DenseNet201 exibe uma maior variabilidade e um desempenho significativamente inferior em relação às demais arquiteturas, o que indica uma

Figura 33 – Métricas de Acurácia e *F1-Score* para a segmentação de tumores cerebrais.



(a) Acurácia.



(b) *F1-Score*.

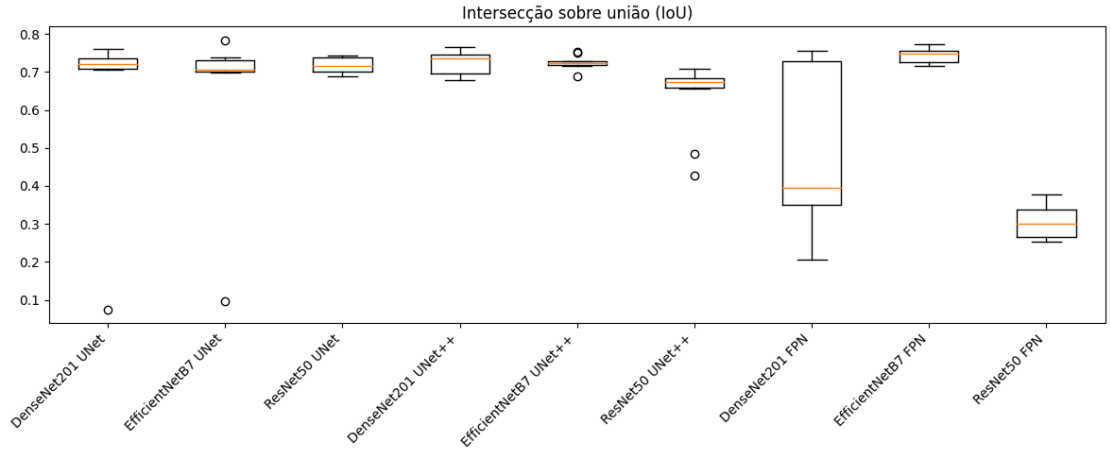
Fonte: Elaborado pela autora.

menor consistência e precisão nas previsões.

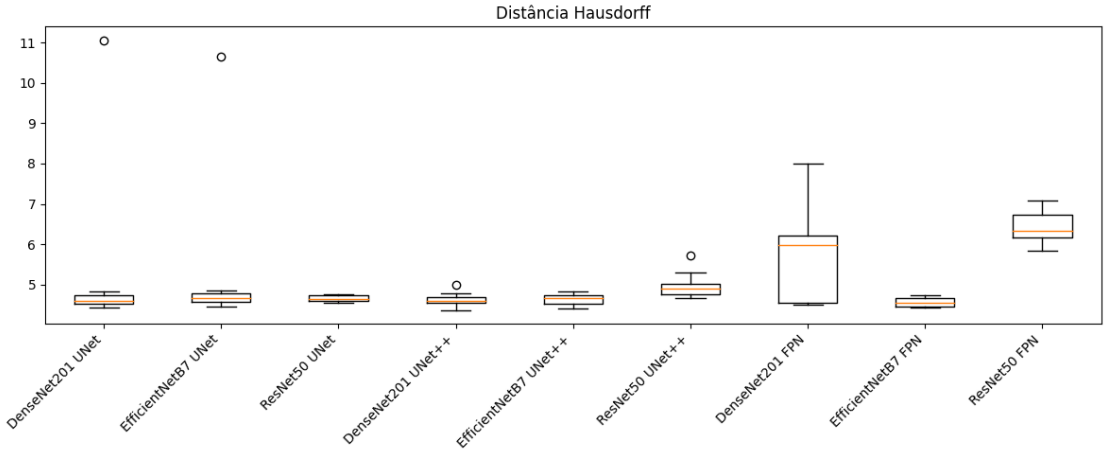
Em relação à métrica Distância de *Hausdorff* (Figura 34(b)), valores menores indicam um desempenho superior dos modelos, indicando previsões mais próximas das máscaras verdadeiras. Nesse sentido, modelos como FPN EfficientNetB7 e UNet++ EfficientNetB7 apresentam DH menores, o que indica que as previsões estão mais próximas da verdade. Em contraste, as arquiteturas FPN ResNet50 e FPN DenseNet201 apresentam maior variabilidade e valores mais altos para a Distância de *Hausdorff*, o que sugere uma menor precisão dos modelos ao realizar as previsões.

Por fim, a Figura 35 exibe o *boxplot* da distribuição do tempo de treinamento dos modelos de segmentação. As arquiteturas UNet++ com EfficientNetB7 e UNet++ com DenseNet201 se destacam como as mais demoradas, exigindo um tempo significativamente maior para concluir o treinamento. Em contapartida, a combinação FPN com ResNet50 se

Figura 34 – Métricas Intersecção sobre União (IoU) Distância de *Hausdorff* para a segmentação de tumores cerebrais.



(a) Intersecção sobre União (IoU).



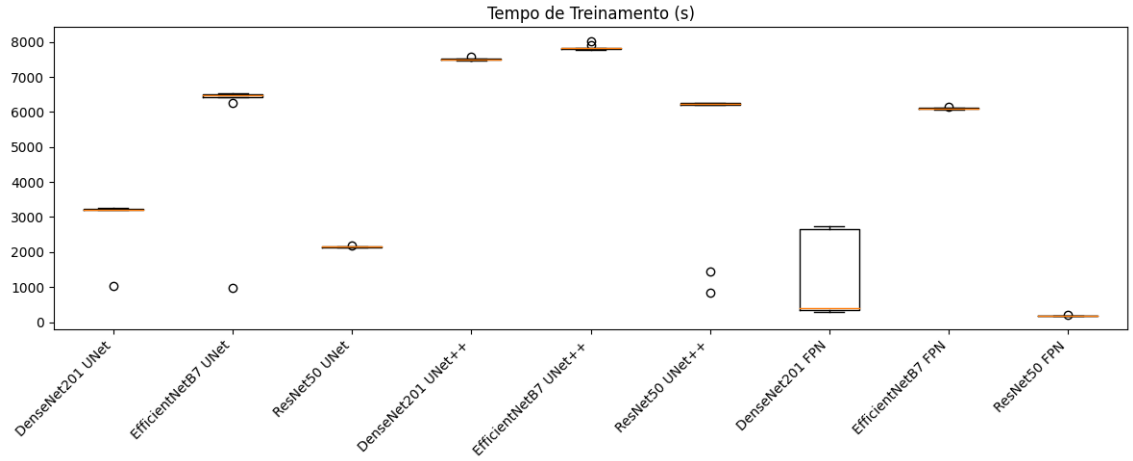
(b) Distância de *Hausdorff*.

Fonte: Elaborado pela autora.

mostrou a mais rápida, completando o treinamento em um tempo consideravelmente menor.

Logo, os *boxplots* oferecem uma visualização inicial das diferenças entre as arquiteturas, entretanto, é possível observar uma sobreposição dos *boxplots* em diversos casos. Essa sobreposição indica que as distribuições das métricas entre as diferentes arquiteturas não apresentam distinções suficientemente claras para determinar a rede que possui o melhor desempenho. Por isso, o uso de testes estatísticos é necessário para avaliar se as diferenças observadas entre as redes são estatisticamente significativas, o que permite uma conclusão mais confiável sobre a arquitetura que apresenta os melhores resultados em termos de desempenho.

Figura 35 – Análise de tempo de treinamento para os modelos de segmentação.



Fonte: Elaborado pela autora.

4.2.2 Testes Estatísticos

Os testes estatísticos foram utilizados para identificar diferenças estatisticamente significativas entre as médias das métricas dos modelos de segmentação de tumores cerebrais. A priori, para legitimar a normalidade dos dados e a homogeneidade entre as populações, foram aplicados os testes de *Shapiro-Wilk* e *Levene*, respectivamente. Novamente, as hipóteses nulas foram rejeitadas, o que aponta que os dados não atendem aos requisitos de normalidade dos dados e de homogeneidade das variâncias. Isto posto, foram estabelecidos testes não paramétricos para avaliar as diferenças estatísticas entre os resultados dos modelos.

Para verificar se existem diferenças estatisticamente diferentes entre as métricas analisadas para os modelos de segmentação aplicou-se o teste não paramétrico de *Friedman*. A Tabela 16 apresenta os resultados da estatística de teste e do valor p resultantes do teste realizado a um nível de significância de 5%. Como resultado, para todas as métricas é possível observar um valor p inferior a 0,05, o que sugere que a hipótese nula deve ser rejeitada, ou seja, existem diferenças significativas entre os métodos avaliadas. Além disso, a estatística de teste quantifica o grau de diferença entre os grupos analisados, dessa forma, quanto maior o valor da estatística de teste, maiores são as evidências de que pelo menos um dos grupos é significativamente diferente dos outros. Portanto, para todas as métricas analisadas, pelo menos uma das arquiteturas de segmentação possui desempenho estatisticamente diferente em relação aos demais modelos.

Portanto, de acordo com os resultados obtidos pelo teste de Friedman, foi utilizado o *pós-hoc* de *Nemenyi* para analisar quais os grupos possuem diferenças estatísticas entre si. A métrica de acurácia não foi analisada, pois os valores para todos os modelos foram próximos a

Tabela 16 – Resultados do teste de *Friedman* para os modelos de segmentação.

	Acc	**F1**	**IoU**	**DH**	**Tempo**
Estatística de Teste	44,37	43,12	43,12	41,46	72,21
Valor *p*	$4{,}83e^{-07}$	$8{,}33e^{-07}$	$8{,}33e^{-07}$	$1{,}70e^{-06}$	$7{,}62e^{-14}$

Fonte: Elaborado pela autora.

100%. Além disso, uma alta pontuação em termos de acurácia pode não ser um indicativo de um bom desempenho em segmentação, pois se houver uma grande quantidade de acertos em pixels de fundo, por exemplo, a taxa de acurácia resultante ainda seria elevada.

Nesse contexto, o *F1-Score* foi escolhido por equilibrar as métricas de precisão e *recall*, sendo a métrica mais relevante para a interpretação de desempenho de modelos de segmentação de imagens médicas. A Tabela 17 apresenta o resultado do pós-hoc de *Nemenyi* para o *F1-Score* para os modelos de segmentação. Os resultados representam o valor p para os pares de modelos analisados. Na tabela, foram destacados os valores inferiores a 0,05, que indicam que a hipótese nula deve ser rejeitada, concluindo-se que existem diferenças estatísticas entre os modelos.

Isto posto, para a métrica *F1-Score*, os resultados mostram que a combinação das redes FPN e ResNet50 apresenta diferenças significativas quando comparada às redes UNet++ DenseNet201, UNet++ EfficientNetB7 e FPN EfficientNetB7. Entre essas arquiteturas, a FPN e ResNet50 obteve as menores taxas para de *F1-Score*, o que indica um desempenho inferior em relação às demais arquiteturas. Vale ressaltar que ao comparar os demais pares de modelos, não foram identificadas diferenças estatisticamente significativas o que sugere que as redes analisadas possuem desempenho semelhante na tarefa de segmentação.

Tabela 17 – Resultados do teste de *Nemenyi* para a métrica *F1-Score* para os modelos de segmentação.

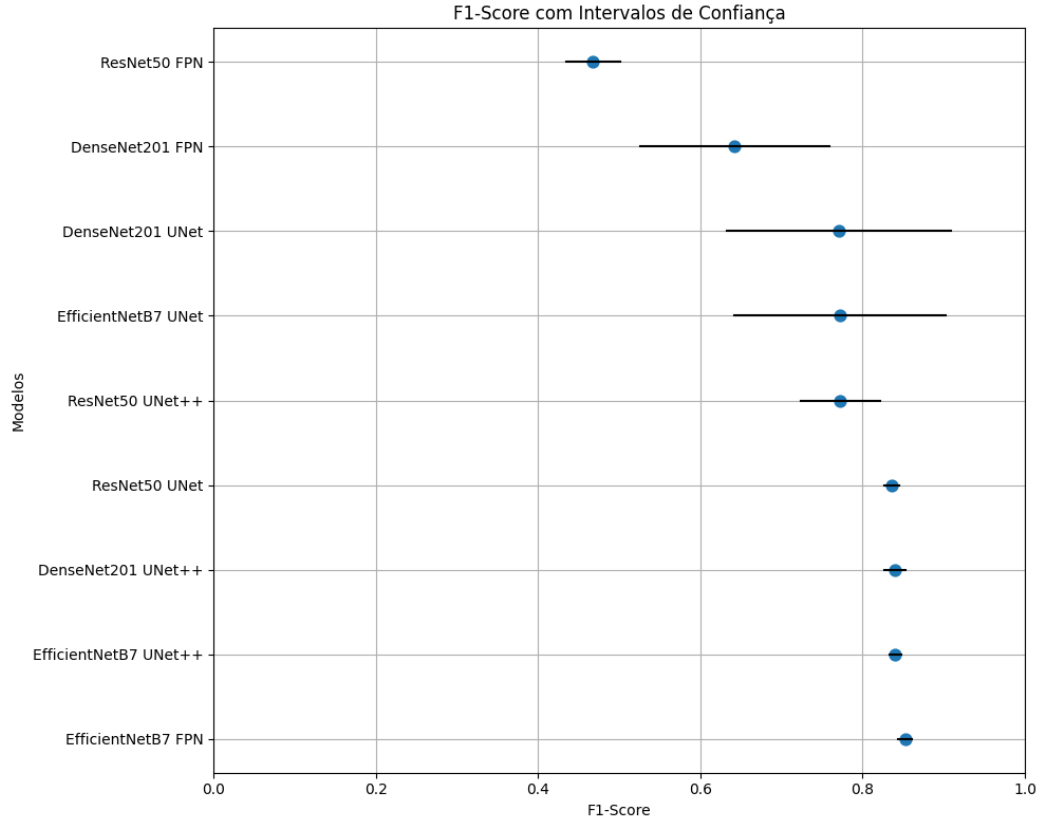
		F1-Score								
		UNet DenseNet201	**UNet EfficientNetB7**	**UNet ResNet50**	**UNet++ DenseNet201**	**UNet++ EfficientNetB7**	**UNet++ ResNet50**	**FPN DenseNet201**	**FPN EfficientNetB7**	**FPN ResNet50**
UNet	**DenseNet201**	1,0	0,99	1,0	0,99	0,99	0,72	0,97	0,95	0,08
	EfficientNetB7	0,99	1,0	0,99	0,99	0,99	0,89	0,99	0,84	0,2
	ResNet50	1,0	0,99	1,0	0,99	0,99	0,75	0,97	0,94	0,09
UNet++	**DenseNet201**	0,99	0,99	0,99	1,0	1,0	0,47	0,87	0,99	**0,02**
	EfficientNetB7	0,99	0,99	0,99	1,0	1,0	0,53	0,9	0,99	**0,03**
	ResNet50	0,72	0,89	0,75	0,47	0,53	1,0	0,99	0,052	0,97
FPN	**DenseNet201**	0,97	0,99	0,97	0,87	0,9	0,99	1,0	0,27	0,76
	EfficientNetB7	0,95	0,84	0,94	0,99	0,99	0,052	0,27	1,0	**0,0003**
	ResNet50	0,08	0,2	0,09	**0,02**	**0,03**	0,97	0,76	**0,0003**	1,0

Fonte: Elaborado pela autora.

A Figura 36 exibe os valores de *F1-Score* e os seus respectivos intervalos de confiança

para cada um dos modelos de segmentação. É possível observar intervalos de confiança mais curtos, como é o caso das redes UNet ResNet50, UNet++ DenseNet201, UNet++ EfficientNetB7 e FPN EfficientNetB7, o que sugere uma menor incerteza associada ao *F1-Score* estimado para cada modelo. Em contrapartida, para as arquiteturas FPN DenseNet201, UNet DenseNet201 e UNet EfficientNetB7, os intervalos de confiança são mais amplos, o que indica uma maior incerteza associada à tarefa de segmentação. Ademais, a ausência de sobreposição entre os intervalos de confiança da combinação das arquiteturas FPN e ResNet50 em relação às demais arquiteturas reitera que existem diferenças significativas entre as redes analisadas, conforme identificado pelo teste de *Nemenyi*.

Figura 36 – Comparação entre valores de *F1-Score* médias e os seus respectivos intervalos de confiança para os diferentes modelos de segmentação.



Fonte: Elaborado pela autora.

Para avaliar a correspondência entre a área prevista pelos modelos e a área real das máscaras de segmentação, utiliza-se a métrica de IoU. Dessa forma, com o objetivo de identificar diferenças significativas na precisão de segmentação das arquiteturas analisadas, foi aplicado o teste de Nemenyi às pontuações de IoU. A Tabela 18 apresenta os resultados do *pós-hoc* de *Nemenyi* para os diferentes modelos. Os resultados indicam que existem diferenças

estatísticas ao comparar a rede FPN ResNet50 com as arquiteturas UNet++ DenseNet201 e FPN EfficientNetB7. É válido ressaltar que, entre essas redes, a FPN ResNet50 obteve as menores pontuações de IoU, enquanto a FPN EfficientNetB7 alcançou a maior taxa de IoU. Logo, em termos de precisão de segmentação, a escolha da rede FPN EfficientNetB7 é mais vantajosa em comparação à FPN ResNet50.

Tabela 18 – Resultados do teste de *Nemenyi* para a métrica Interseção sobre União (IoU) para os modelos de segmentação.

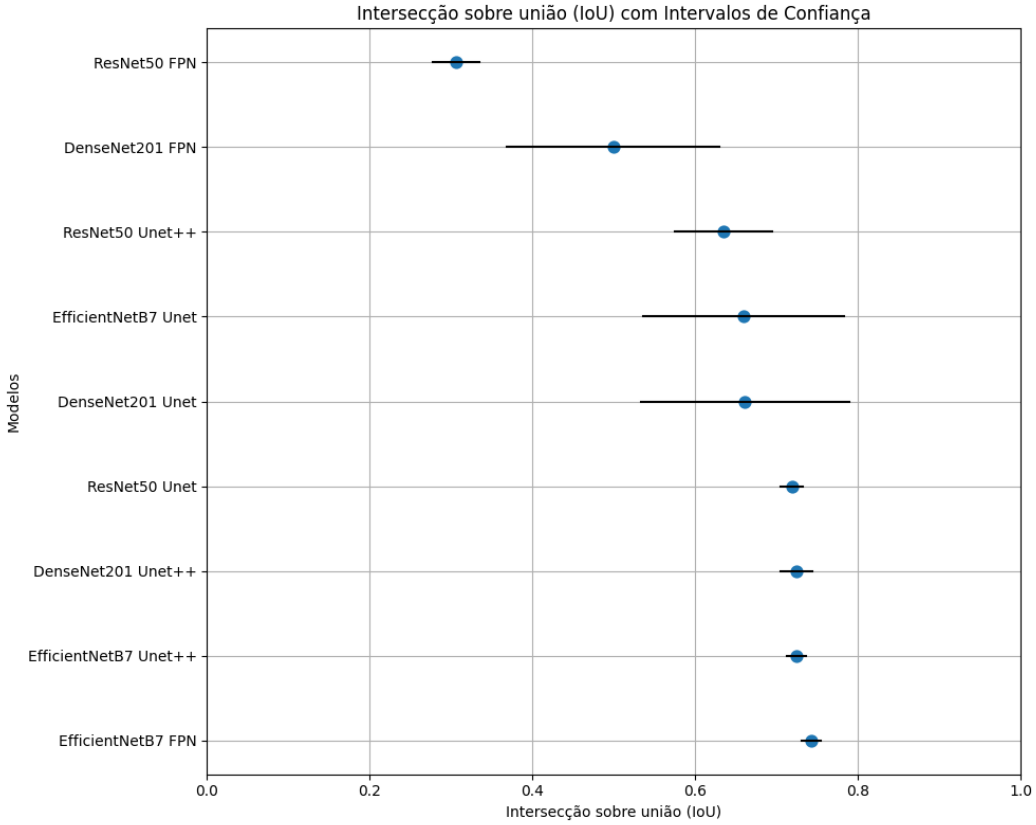
Interseção sobre União (IoU)										
		UNet DenseNet201	**UNet EfficientNetB7**	**UNet ResNet50**	**UNet++ DenseNet201**	**UNet++ EfficientNetB7**	**UNet++ ResNet50**	**FPN DenseNet201**	**FPN EfficientNetB7**	**FPN ResNet50**
UNet	**DenseNet201**	1,0	0,99	1,0	0,99	0,99	0,72	0,97	0,95	0,08
	EfficientNetB7	0,99	1,0	0,99	0,99	0,99	0,89	0,99	0,84	0,20
	ResNet50	1,0	0,99	1,0	0,99	0,99	0,75	0,97	0,94	0,09
UNet++	**DenseNet201**	1,0	0,99	0,99	1,0	1,0	0,66	0,91	0,99	**0,03**
	EfficientNetB7	1,0	0,99	0,99	1,0	1,0	0,75	0,94	0,99	0,053
	ResNet50	0,80	0,96	0,90	0,66	0,75	1,0	0,99	0,20	0,94
FPN	**DenseNet201**	0,96	0,99	0,99	0,91	0,94	0,99	1,0	0,49	0,75
	EfficientNetB7	0,99	0,92	0,97	0,99	0,99	0,20	0,49	1,0	**0,001**
	ResNet50	0,07	0,24	0,14	**0,03**	0,053	0,94	0,75	**0,001**	1,0

Fonte: Elaborado pela autora.

A Figura 37 apresenta os resultados dos modelos para a métrica IoU. Logo, corroborando com o resultado do teste de Nemenyi a rede FPN ResNet50 apresenta o menor valor de IoU, enquanto as redes UNet++ DenseNet201 e FPN EfficientNetB7 apresentam os resultados mais elevados para a métrica. Além disso, a ausência de sobreposição dos intervalos de confiança confirma as diferenças estatísticas existentes entre essas redes. Além disso, os intervalos de confiança mais curtos, observados nas redes UNet++ DenseNet201 e FPN EfficientNetB7, indicam uma menor incerteza e uma estimativa mais confiável, o que reflete uma menor variabilidade dos dados.

No contexto de segmentação de imagens, a Distância de *Hausdorff* mede o erro entre as bordas da área prevista pelo modelo e as bordas da área real. Logo, quanto menor a DH, mais precisa é a segmentação. A Tabela 19 apresenta os resultados obtidos para o pós-hoc de *Nemenyi*, realizado para identificar diferenças entre os valores resultantes para a métrica DH das arquiteturas de segmentação analisadas. Novamente, a rede FPN ResNet50 destaca-se por apresentar diferenças significativas em comparação com as redes UNet++ DenseNet201 e FPN EfficientNetB7. Vale ressaltar que a FPN ResNet50 registrou as maiores pontuações para DH, o que sugere menor precisão ao realizar a segmentação de imagens, principalmente, quando comparada às arquiteturas que obtiveram as menores pontuações para a métrica DH. As demais

Figura 37 – Comparação entre valores de Intersecção sobre União (IoU) e os seus respectivos intervalos de confiança para os diferentes modelos de segmentação.



Fonte: Elaborado pela autora.

redes não apresentaram diferenças estatísticas significativas entre si.

Tabela 19 – Resultados do teste de *Nemenyi* para a métrica Distância de *Hausdorff* para os modelos de segmentação.

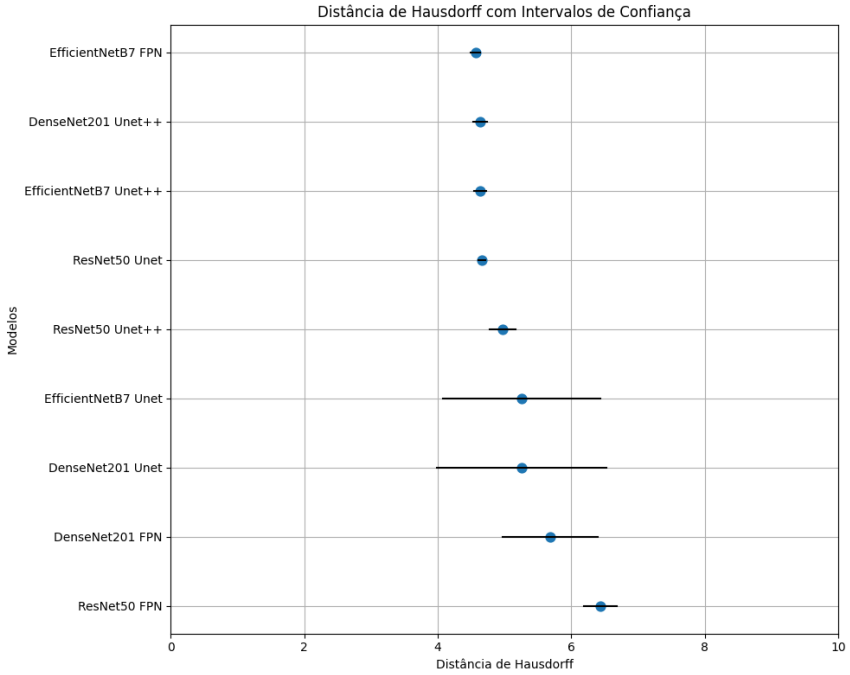
		Distância de *Hausdorff*								
		UNet DenseNet201	**UNet EfficientNetB7**	**UNet ResNet50**	**UNet++ DenseNet201**	**UNet++ EfficientNetB7**	**UNet++ ResNet50**	**FPN DenseNet201**	**FPN EfficientNetB7**	**FPN ResNet50**
UNet	**DenseNet201**	1,0	0,99	1,0	1,0	1,0	0,80	0,96	0,99	0,07
	EfficientNetB7	0,99	1,0	1,0	0,99	0,99	0,96	0,99	0,2	0,24
	ResNet50	1,0	1,0	1,0	0,99	0,99	0,90	0,99	0,97	0,14
UNet++	**DenseNet201**	1,0	0,99	0,99	1,0	1,0	0,66	0,91	0,99	**0,03**
	EfficientNetB7	1,0	0,99	0,99	1,0	1,0	0,75	0,94	0,99	0,053
	ResNet50	0,80	0,96	0,90	0,66	0,75	1,0	0,99	0,20	0,94
FPN	**DenseNet201**	0,96	0,99	0,99	0,91	0,94	0,99	1,0	0,49	0,75
	EfficientNetB7	0,99	0,92	0,97	0,99	0,99	0,20	0,49	1,0	**0,001**
	ResNet50	0,07	0,24	0,14	**0,03**	0,053	0,94	0,75	**0,001**	1,0

Fonte: Elaborado pela autora.

A Figura 38 apresenta a métrica de DH com intervalos de confiança para diferentes modelos. Nesse sentido, a FPN ResNet50 apresenta a maior taxa para DH, o que indica um menor desempenho. Em contrapartida, as redes UNet++ DenseNet201 e FPN EfficientNetB7 exibem as menores pontuações de DH. Ademais, é possível observar que não existe a sobreposição

do intervalo de confiança dessas redes com a FPN ResNet50, o que reforça o resultado obtido pelo teste de *Nemenyi*. Todavia, para as demais redes nota-se a sobreposição dos intervalos de confiança, o que indica que as variações da DH entre os modelos não são suficientemente grandes para determinar a superioridade de uma rede em relação às demais.

Figura 38 – Comparação entre valores de Distância de *Hausdorff* e os seus respectivos intervalos de confiança para os diferentes modelos de segmentação.



Fonte: Elaborado pela autora.

A Tabela 20 apresenta o resultado do teste de *Nemenyi* para o tempo de treinamento dos modelos. A arquitetura UNet++ EfficientNetB7, por apresentar o maior tempo de treinamento entre as redes analisadas, mostrou diferenças estatísticas ao ser comparada com os modelos UNet DenseNet201, UNet ResNet50, FPN DenseNet e FPN Resnet50. Vale destacar que a rede FPN EfficientNetB7, que apresentou as melhores taxas para as métricas quantitativas analisadas, não apresentou diferenças significativas ao ser comparada com as demais redes. Em contrapartida, a rede FPN Resnet50 mostrou-se estatisticamente diferente quando comparada aos modelos UNet EfficientNetB7, Unet++ DenseNet201 e UNet++ EfficientNetB7. Entretanto, apesar da rede FPN Resnet50 apresentar diferença estatística quando comparada aos demais modelos e ser a rede que obteve menor tempo de treinamento, trata-se da rede com as piores taxas para as métricas analisadas.

Por fim, a Figura 39 exibe o comportamento dos tempos de treinamento para cada

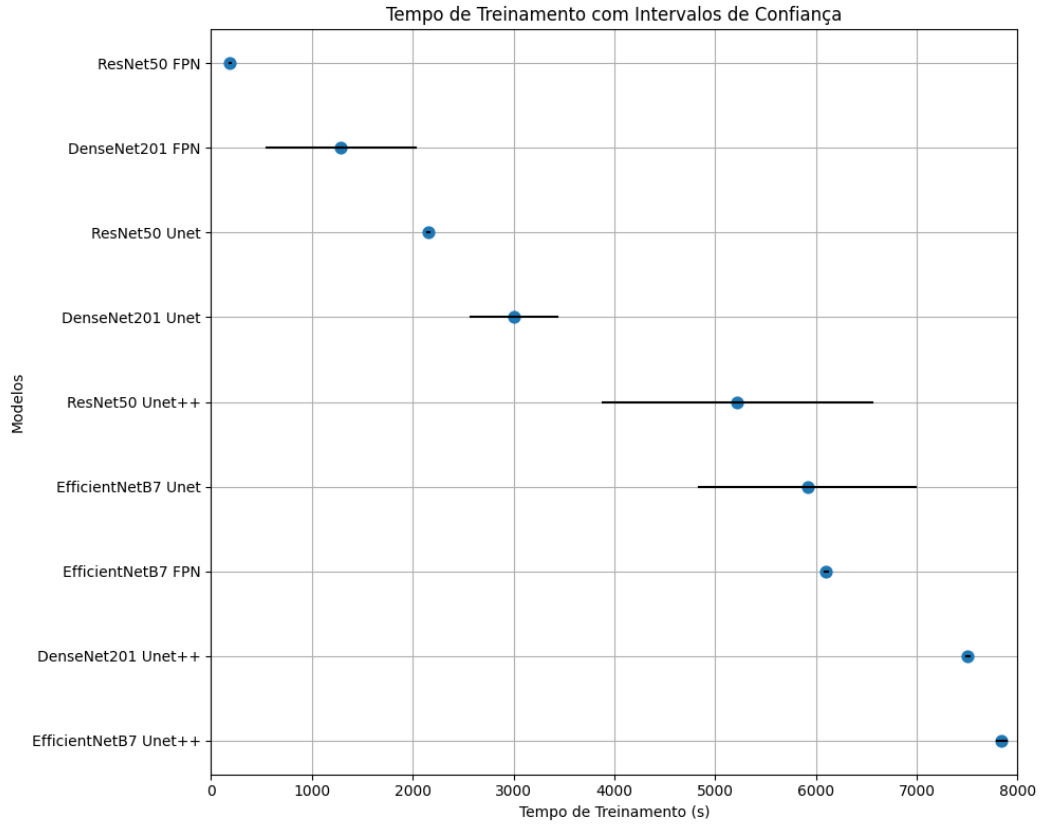
Tabela 20 – Resultados do teste de *Nemenyi* para a métrica Tempo de Treinamento.

Tempo de Treinamento										
		UNet DenseNet201	**UNet EfficientNetB7**	**UNet ResNet50**	**UNet++ DenseNet201**	**UNet++ EfficientNetB7**	**UNet++ ResNet50**	**FPN DenseNet201**	**FPN EfficientNetB7**	**FPN ResNet50**
UNet	**DenseNet201**	1,0	0,83	0,99	0,20	**0,02**	0,99	0,98	0,99	0,507
	EfficientNetB7	0,83	1,0	0,31	0,99	0,82	0,99	0,16	0,99	**0,003**
	ResNet50	0,99	0,31	1,0	**0,01**	**0,0009**	0,83	0,99	0,86	0,09
UNet++	**DenseNet201**	0,2	0,99	**0,01**	1,0	0,99	0,77	**0,005**	0,72	**0,00001**
	EfficientNetB7	**0,02**	0,82	**0,0009**	0,99	1,0	0,31	**0,0001**	0,26	**0,0000001**
	ResNet50	0,99	0,99	0,83	0,77	0,31	1,0	0,64	1,0	0,07
FPN	**DenseNet201**	0,98	0,16	0,99	**0,005**	**0,0001**	0,64	1,0	0,70	0,98
	EfficientNetB7	0,99	0,99	0,86	0,72	0,26	1,0	0,70	1,0	0,09
	ResNet50	0,507	**0,003**	0,93	**0,00001**	**0,0000001**	0,07	0,98	0,09	1,0

Fonte: Elaborado pela autora.

modelo e os respectivos intervalos de confiança. Como ressaltado anteriormente, a rede FPN Resnet50 apresentou o menor tempo de treinamento, porém, obteve as menores taxas. Além disso, não existe a sobreposição do intervalo de confiança dessa rede com as demais arquiteturas, o que comprova os resultados do teste de *Nemenyi*. Nesse contexto, a rede UNet++ EfficientNetB7 apresentou o maior tempo de treinamento, o que indica um maior custo computacional para a realização do treinamento da rede.

Figura 39 – Comparação entre valores de Tempo de Treinamento e os seus respectivos intervalos de confiança para os diferentes modelos de segmentação.



Fonte: Elaborado pela autora.

Com base nos resultados analisados, a rede FPN Resnet50 mostrou-se inferior ao analisar as métricas quantitativas para a tarefa de segmentação de imagens. Em contraste, a rede FPN EfficientNetB7 apresentou as melhores pontuações para as métricas, com diferenças estatísticas significativas apenas quando comparada com a FPN Resnet50. Portanto, pode-se concluir que a FPN Resnet50 não oferece um desempenho satisfatório. Por outro lado, as demais redes não apresentam diferenças estatísticas significativas entre si, o que sugere que as redes apresentam resultados competitivos em termos de desempenho para a tarefa de segmentação. Logo, a FPN EfficientNetB7, por obter as melhores taxas e um tempo de treinamento intermediário, se destaca como a arquitetura mais adequada para realizar a segmentação e detecção de tumores cerebrais.

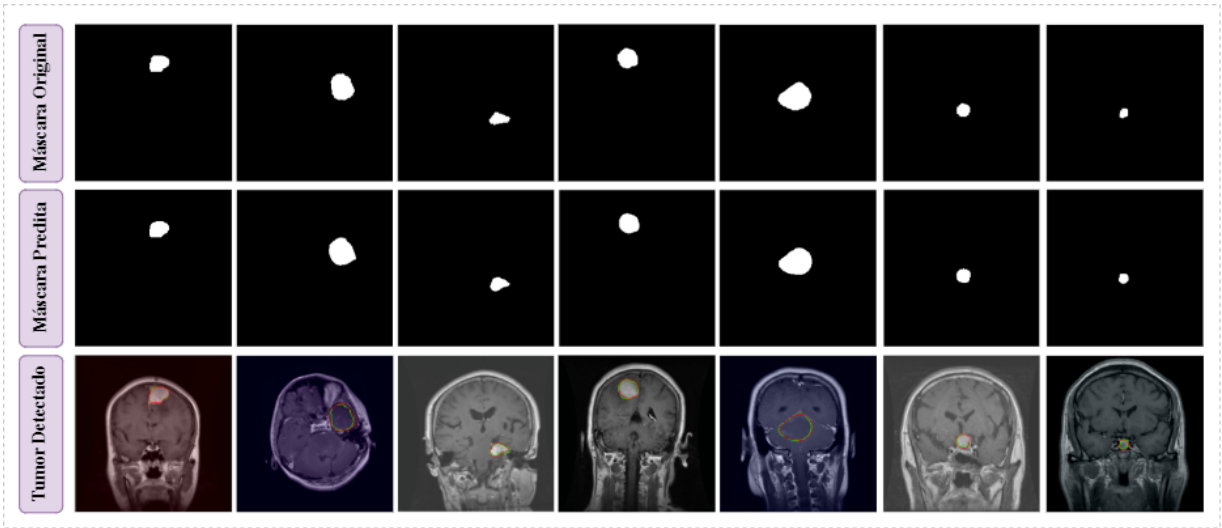
4.2.3 Resultados Visuais da Segmentação

Nesse contexto, os resultados visuais apresentam a marcação do tumor detectado pelas arquiteturas de segmentação utilizadas. A Figura 40 exibe os resultados visuais da segmentação de tumores cerebrais em imagens de RM por meio da rede FPN EfficientNetB7, visto que trata-se da rede que apresentou os melhores resultados nas análises realizadas. Desse modo, a primeira linha exibe as máscaras originais que representam a área verdadeira do tumor cerebral. A segunda linha apresenta máscaras tumorais previstas pela rede. Já a terceira linha exibe as imagens de RM com a marcação da área tumoral. Nestas imagens, o contorno contorno em verde representa a máscara tumoral original, enquanto, o contorno em vermelho indica a marcação prevista pela arquitetura.

A semelhança entre a máscara prevista com a máscara tumoral original, combinada com os resultados das métricas quantitativas, destaca o desempenho da rede FPN EfficientNetB7 na segmentação precisa dos tumores. É válido ressaltar a eficiência da rede em detectar e contornar o tumor cerebral com precisão, mesmo diante dos desafios de localização, formato e tamanho do tumor. Além disso, a rede demonstra capacidade em detectar precisamente tumores cerebrais de diferentes contrastes e até mesmo de tamanhos pequenos, conforme apresentado nos exemplos, o que é essencial para a detecção precoce dos tumores cerebrais.

Ademais, de acordo com os testes estatísticos realizados, a rede FPN EfficientNetB7 apresentou resultados comparáveis a demais arquiteturas de segmentação, o que motivou a apresentação do resultado de segmentação para outras arquiteturas. A Figura 41 exibe uma comparação entre os resultados obtidos pela rede FPN EfficientNetB7 e pela rede UNet++

Figura 40 – Resultados visuais da segmentação tumoral de imagens de RM utilizando as redes FPN EfficienteNetB7. Na primeira linha são apresentadas as máscaras originais, na segunda são exibidas as máscaras previstas pela rede e, por fim, na última linha são apresentas as imagens originais com as marcações da área tumoral. O contorno em verde representa a máscara tumoral original e o contorno vermelho representa a máscara tumoral prevista.



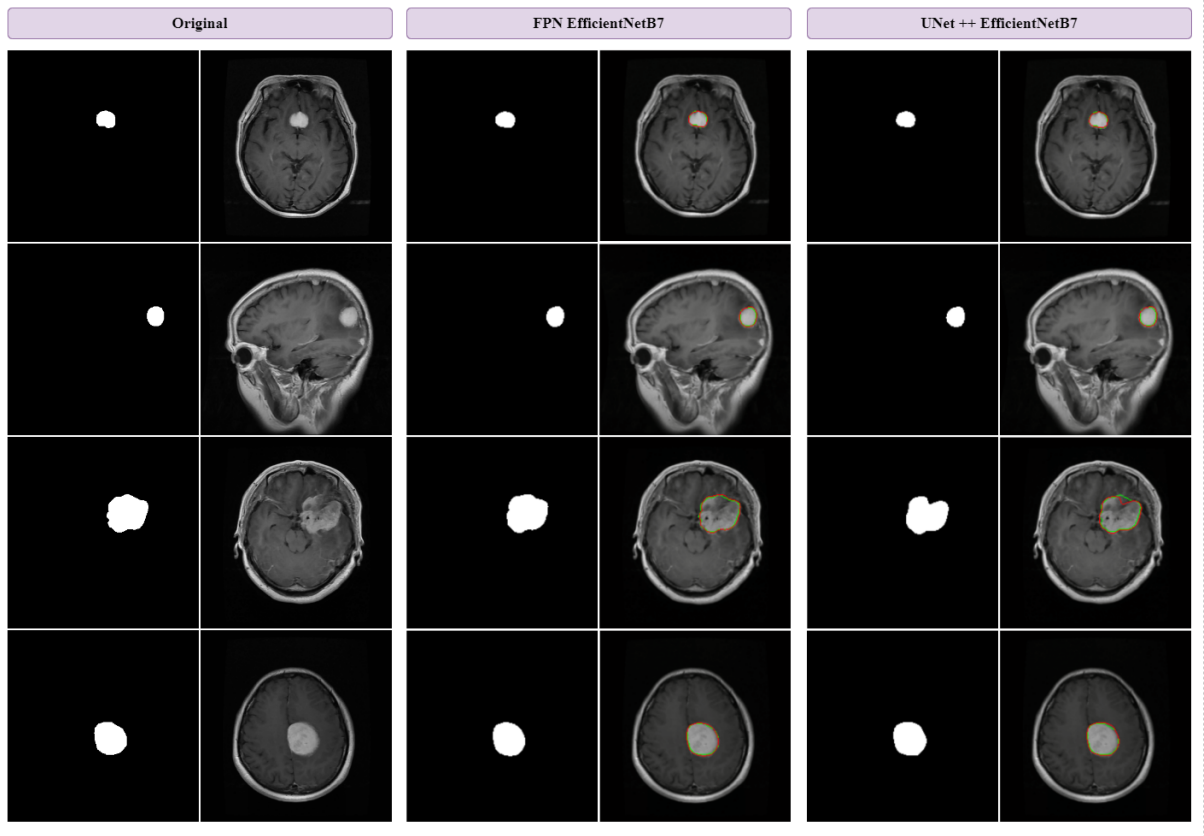
Fonte: Elaborado pela autora.

EfficientNetB7, sendo esta última a segunda rede com os melhores resultados para as métricas quantitativas.

As duas primeiras colunas da Figura 41 mostram a máscara e a imagem de RM originais. A terceira e a quarta coluna exibem a máscara prevista e a marcação da área do tumor pela rede FPN EfficientNetB7, enquanto as duas últimas colunas apresentam o resultado obtido pela rede UNet++ EfficientNetB7. Novamente, o contorno em verde indica a máscara tumoral original, e o contorno em vermelho representa a máscara prevista pela rede. Logo, ambas as arquiteturas demonstram uma capacidade e desempenho consistente de localizar tumores cerebrais. No entanto, em alguns casos, a rede UNet++ EfficientNetB7 se mostrou menos específica, como evidenciado no terceiro exemplo, onde a segmentação foi menos precisa.

Por fim, a Figura 42 apresenta os resultados da segmentação para a rede FPN ResNet50 que apresentou as piores taxas para as métricas analisadas. No exemplo, são apresentadas as máscaras e imagens de RM originais e as máscaras previstas pela arquitetura FPN ResNet50 e regiões tumorais marcadas. Desse modo, é possível observar que a rede realiza marcações imprecisas e apresenta dificuldades em identificar corretamente as áreas do tumor. As imprecisões nas segmentações evidenciam as limitações da rede, o que ressalta a importância de selecionar a arquitetura mais adequada para tarefas específicas de segmentação. Os resultados destacam

Figura 41 – Resultados visuais da segmentação tumoral de imagens de RM utilizando as redes FPN EfficienteNetB7 e UNet++ EfficienteNetB7. O primeiro grupo de imagens apresenta a máscara e a imagem de RM originais. O segundo grupo apresenta a máscara prevista pela rede FPN EfficienteNetB7 e a marcação da área do tumor e o terceiro grupo pela rede UNet++ EfficienteNetB7.



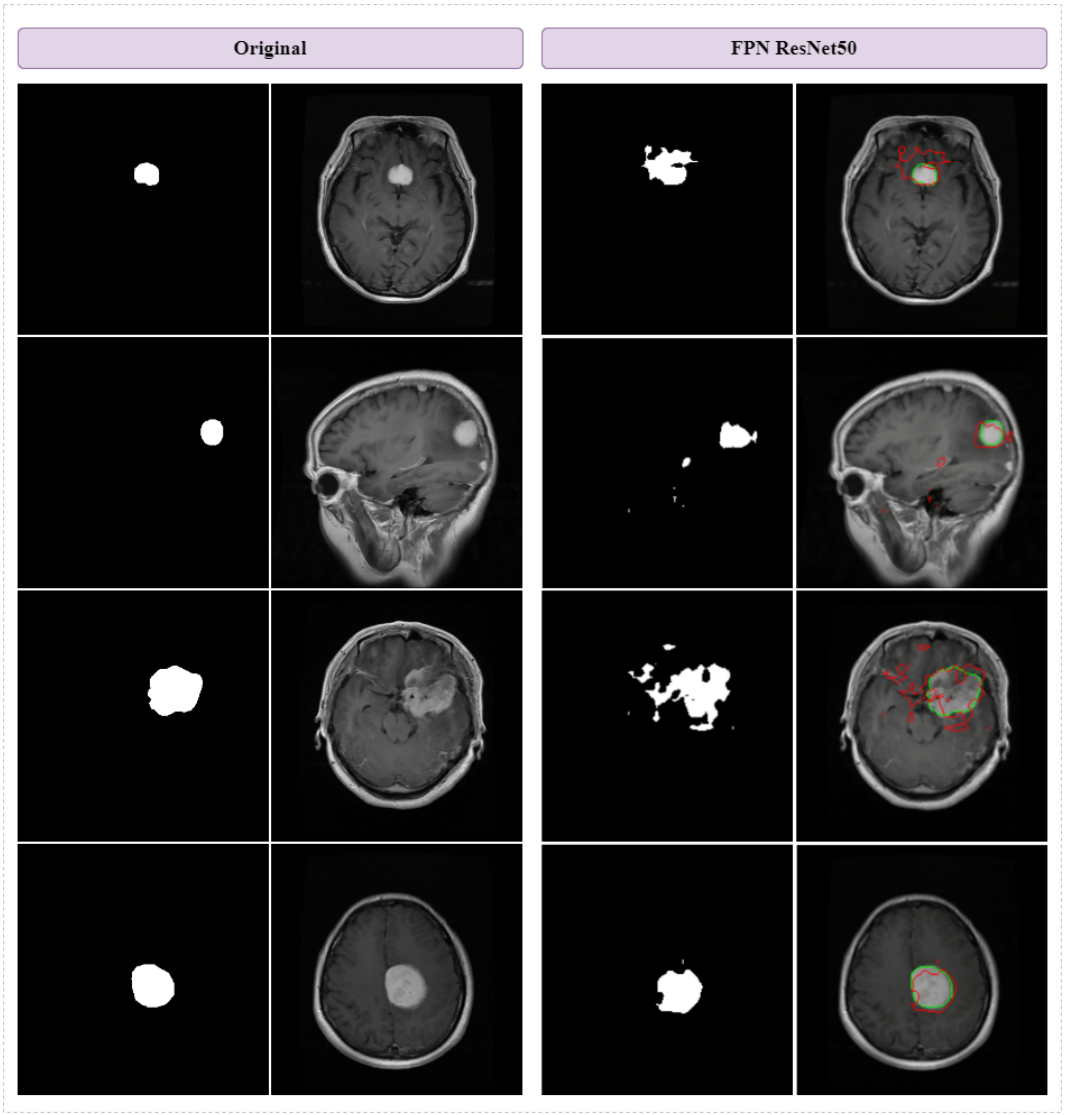
Fonte: Elaborado pela autora.

que a escolha de uma arquitetura apropriada pode fazer a diferença significativa na precisão e eficácia do modelo em identificar e delimitar os tumores cerebrais.

4.3 Sumarização dos Resultados

Neste trabalho, foi desenvolvido um fluxo de etapas para a classificação e segmentação de tumores cerebrais. O objetivo principal foi avaliar o desempenho *CNN*s, responsáveis por realizar com precisão as tarefas de classificação e segmentação. Além disso, foi conduzida uma análise estatística abrangente para detalhar os resultados e compreender o desempenho de cada modelo. Desse modo, o intuito principal seria analisar a capacidade das arquiteturas de redes neurais em classificar e segmentar tumores cerebrais, contornando os desafios existentes, como, contraste, tamanho, posição e variação da região tumoral.

Figura 42 – Resultados visuais da segmentação tumoral de imagens de RM utilizando as redes FPN e ResNet50. Na primeira linha são apresentadas as máscaras originais, na segunda são exibidas as máscaras previstas pela rede e, por fim, na última linha são apresentas as imagens originais com as marcações da área tumoral. O contorno em verde representa a máscara tumoral original e o contorno vermelho representa a máscara tumoral prevista.



Fonte: Elaborado pela autora.

A priori, a primeira etapa consistiu em classificar os tipos de tumores cerebrais em Meningioma, Glioma, Hipofisário e casos sem tumor. Para isso, foram treinados quinze diferentes modelos de aprendizado profundo, por meio da técnica de *transfer learning*. Entre eles, os modelos DenseNet201, EfficientNetB7 e ResNet50 apresentaram as melhores taxas para as métricas quantitativas analisadas. Em destaque, a rede EfficientNetB7 apresentou as pontuações mais elevadas, com uma acurácia de 97,68%, precisão de 97,63%, *recall* de 97,69%, *F1-Score* de 97,64% e Especificidade de 99,21%. No entanto, a rede exigiu o maior tempo de treinamento, totalizando 3664,82 segundos para concluir o processo.

Ademais, as arquiteturas DenseNet201 e ResNet50 apresentaram desempenho comparável em relação às métricas analisadas. A rede DenseNet201 obteve uma acurácia de 97,25%, precisão de 97,36%, *recall* de 97,08%, *F1-Score* de 97,20% e especificidade de 99,07% com um tempo de treinamento de 1108,59 segundos. Por outro lado, a rede ResNet50 apresentou uma acurácia de 97,11%, precisão de 97,30%, *recall* de 96,97%, *F1-Score* de 97,12%, especificidade de 99,02% e o menor tempo de treinamento entre essas três redes, com 581,81 segundos.

Além disso, foram realizados testes estatísticos para verificar se existiam diferenças significativas entre as médias das arquiteturas analisadas. Os resultados indicaram que as redes AlexNet e MobileNetV2, que apresentaram as piores pontuações, obtiveram diferenças estatísticas em comparação com as melhores redes, o que sugere uma menor capacidade para realizar a tarefa de classificação. Diante disso, foi realizada uma validação externa com as redes DenseNet201, EfficientNetB7 e ResNet50, visto que, foram as redes que apresentaram as melhores pontuações e não obtiveram diferenças estatísticas significativas entre si.

Logo, a validação externa foi conduzida com um banco de dados não utilizado no treinamento e teste dos modelos, com o objetivo de avaliar a capacidade de generalização das arquiteturas e seu desempenho em cenários ainda não vistos. Novamente, os resultados para a validação externa confirmaram o potencial promissor das redes na classificação de tumores cerebrais, com destaque para a rede EfficientNetB7, que também apresentou as melhores pontuações para as novas amostras.

Em seguida, foi realizada a detecção da região tumoral. Nessa etapa, foi utilizado o conceito de codificadores e decodificadores. As redes codificadoras utilizadas foram a DenseNet201, EfficientNetB7 e ResNet50, devido ao desempenho considerável na etapa de classificação. Já para as redes decodificadoras utilizou-se as arquiteturas UNet, Unet++ e FPN, amplamente reconhecidas em problemas de segmentação de imagens médicas.

Nesse contexto, as redes FPN e EfficientNetB7 apresentaram os melhores resultados, com 99,52% de acurácia, 85,23% para a métrica *F1-Score*, 74,29% para IoU e 4,56 para a Distância de *Hausdorff*. Vale salientar que, por meio dos testes estatísticos, ficou comprovado que a rede ResNet50 e FPN apresentaram diferenças estatísticas consideráveis ao serem comparadas com as redes de melhor desempenho, além de obterem os piores resultados nas métricas analisadas.

Por fim, foi realizada a análise visual das arquiteturas, por meio da geração de máscara de segmentação previstas pelas redes e a marcação da área do tumor cerebral em

comparação com a marcação da área real do tumor. Em destaque, as redes FPN EfficientNetB7 e UNet++ EfficientNetB7 apresentaram resultados visuais satisfatórios, com a área do tumor detectada com precisão. Em contrapartida, a rede FPN ResNet50 exibiu resultados imprecisos e falhou capturar adequadamente os padrões nas imagens.

Portanto, este trabalho apresenta o treinamento de redes neurais e a realização de uma análise abrangente para avaliar o desempenho dos modelos na classificação e segmentação de tumores cerebrais. Por meio de uma abordagem sistemática e da realização de testes estatísticos, os resultados demonstraram que a rede EfficientNetB7 se destacou na tarefa de classificação, apresentando as melhores taxas para as métricas analisadas. Na etapa de segmentação, a combinação da rede FPN com EfficientNetB7 mostrou-se a mais eficiente, evidenciando um excelente desempenho na detecção da região tumoral.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, foi realizado um fluxo de etapas para classificação dos tipos de tumores cerebrais em Meningioma, Glioma, Hipofisário e casos sem tumor e a segmentação da região tumoral, por meio de imagens de ressonância magnética, baseado em aprendizagem profunda. Para isso, foram realizados diversos experimentos utilizando modelos robustos de *DL*. O principal objetivo consiste na análise comparativa do desempenho de *CNN*s na tarefa de classificação e segmentação de forma precisa, de forma a desenvolver um fluxo para auxiliar e viabilizar um diagnóstico mais preciso e precoce realizado pelo profissional da saúde.

Para a tarefa de classificação, o estudo apresentou resultados significativos frente ao que é explorado na literatura. Foram realizados testes com quinze redes neurais profundas pré-treinadas, por meio da técnica de *transfer learning*. Nesse contexto, as redes DenseNet201, EfficientNetB7 e ResNet50 apresentaram os melhores resultados para as métricas quantitativas de avaliação.

Além disso, foram conduzidos testes estatísticos para avaliar a significância dos resultados e identificar possíveis diferenças entre as médias dos resultados obtidos. Os resultados demonstram que as redes AlexNet e MobileNetV2, que obtiveram as piores pontuações, apresentaram diferenças significativas em relação às melhores redes, sugerindo uma menor capacidade de classificação. Em contraste, as redes DenseNet201, EfficientNetB7 e ResNet50, que obtiveram as melhores pontuações sem diferenças estatísticas significativas entre si. Vale ressaltar que as redes foram validadas externamente com um banco de dados não utilizado no treinamento. Os resultados da validação externa confirmaram o desempenho promissor das redes, destacando a EfficientNetB7 como a mais eficaz na classificação de tumores cerebrais.

Para a tarefa de segmentação, fez-se a combinação dos codificadores DenseNet201, EfficientNetB7 e ResNet50 e dos decodificadores UNet, UNet++ e FPN. A abordagem com EfficientNetB7 e FPN apresentou resultados competitivos em termos de desempenho para a detecção de tumores cerebrais. Ademais, foi conduzida uma análise abrangente por meio de testes estatísticos para avaliar a significância dos resultados dos modelos de segmentação. Por meio da análise estatística, pode-se confirmar que as arquiteturas ResNet50 e FPN apresentaram diferenças estatísticas consideráveis ao serem comparadas com as redes de melhor desempenho, o que indica que a rede não é a mais adequada para a tarefa de segmentação de tumores cerebrais, frente às demais redes analisadas.

Por fim, realizou-se a análise visual dos resultados obtidos pelas arquiteturas, por

meio da geração de máscara de segmentação da área real do tumor em comparação com a marcação da área real do tumor. Novamente, vale destacar o desempenho das redes FPN EfficientNetB7 e UNet++ EfficientNetB7, que além de apresentarem as melhores taxas para as métricas quantificativas, obtiveram resultados visuais satisfatórios, com a área do tumor detectada com precisão.

5.1 Trabalhos Futuros

Como trabalhos futuros, pretende-se integrar bancos de dados adicionais de imagens de ressonância magnética, incluindo imagens em 3D, para validar a eficiência do fluxo proposto. Além disso, serão aplicadas técnicas avançadas de pré-processamento para melhorar ainda mais o desempenho e a precisão dos modelos testados.

Ademais, deve-se testar outras arquiteturas de redes neurais profundas que sejam leves e eficientes, a fim de desenvolver um sistema de diagnóstico auxiliado por computador (CAD) para detectar, segmentar e classificar tumores cerebrais com maior precisão e performance, utilizando imagens de ressonância magnética, incluindo análises de imagens 3D para fornecer uma visão mais detalhada e abrangente do tumor.

O sistema deverá possibilitar a quantificação da área tumoral, a fim de apoiar na determinação do estágio da doença. Ao incorporar a análise de imagens 3D, espera-se que o sistema ofereça uma ferramenta ainda mais robusta para auxiliar os profissionais de saúde na realização de diagnósticos precoces e precisos, proporcionando suporte crucial na tomada de decisões clínicas.

REFERÊNCIAS

AHMAD, S.; CHOUDHURY, P. K. On the Performance of Deep Transfer Learning Networks for Brain Tumor Detection Using MR Images. **IEEE Access**, v. 10, p. 59099–59114, 2022.

ALI, S.; LI, J.; PEI, Y.; KHURRAM, R.; REHMAN, K. U.; MAHMOOD, T. A comprehensive survey on brain tumor diagnosis using deep learning and emerging hybrid techniques with multi-modal mr image. **Archives of computational methods in engineering**, Springer, v. 29, n. 7, p. 4871–4896, 2022.

ALIBRAHIM, H.; LUDWIG, S. A. Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. In: **2021 IEEE Congress on Evolutionary Computation (CEC)**. [*S. l.*: *s. n.*], 2021. p. 1551–1559.

ALZUBAIDI, L.; ZHANG, J.; HUMAIDI, A. J.; AL-DUJAILI, A.; DUAN, Y.; AL-SHAMMA, O.; SANTAMARíA, J.; FADHEL, M. A.; AL-AMIDIE, M.; FARHAN, L.; AL. et. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, 2021.

ANANTHARAJAN, S.; GUNASEKARAN, S.; SUBRAMANIAN, T.; R, V. MRI brain tumor detection using deep learning and machine learning approaches. **Measurement: Sensors**, v. 31, p. 101026, 2024. ISSN 2665-9174. Disponível em: https://www.sciencedirect.com/science/article/pii/S2665917424000023.

ASIF, S.; YI, W.; AIN, Q. U.; HOU, J.; YI, T.; SI, J. Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from mr images. **IEEE Access**, v. 10, p. 34716–34730, 2022.

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 39, n. 12, p. 2481–2495, 2017.

BARUAH, N.; SANGINENI, R.; CHAKRABORTY, M.; NAYAK, S. K. Statistical Analysis of Natural Ester based Insulating Liquid using Hypothesis Testing. In: **2020 International Symposium on Electrical Insulating Materials (ISEIM)**. [*S. l.*: *s. n.*], 2020. p. 347–350.

BERRAR, D. Cross-validation. In: SAMMUT, C.; WEBB, G. I. (Ed.). **Encyclopedia of Machine Learning and Data Mining**. 2. ed. Amsterdam: Elsevier, 2018.

BHUVAJI, S.; KADAM, A.; BHUMKAR, P.; DEDGE, S.; KANCHAN, S. **Brain Tumor Classification (MRI) [Dataset]. Kaggle.** [*S. l.*]: Kaggle, 2020. Disponível em: https://www.kaggle.com/dsv/1183165. Acesso em: 01 mar. 2024.

BINDU, J. H.; DEVI, M. U. Classification of brain tumor images using segmentation and transfer learning. In: **2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)**. [*S. l.*: *s. n.*], 2024. p. 225–232.

BLEEKER, S.; MOLL, H.; STEYERBERG, E. a.; DONDERS, A.; DERKSEN-LUBSEN, G.; GROBBEE, D.; MOONS, K. External validation is necessary in prediction research:: A clinical example. **Journal of clinical epidemiology**, Elsevier, v. 56, n. 9, p. 826–832, 2003.

BOW, S.-T. **Pattern Recognition and Image Preprocessing**. 2nd. ed. USA: Marcel Dekker, Inc., 2002. ISBN 0824706595.

BROWN, M. B.; FORSYTHE, A. B. Robust tests for the equality of variances. **Journal of the American statistical association**, Taylor & Francis, v. 69, n. 346, p. 364–367, 1974.

CABITZA, F.; CAMPAGNER, A.; SOARES, F.; GUADIANA-ROMUALDO, L. G. de; CHALLA, F.; SULEJMANI, A.; SEGHEZZI, M.; CAROBENE, A. The importance of being external. methodological insights for the external validation of machine learning models in medicine. **Computer Methods and Programs in Biomedicine**, v. 208, p. 106288, 2021. ISSN 0169-2607. Disponível em: https://www.sciencedirect.com/science/article/pii/S016926072100362X.

CHA, S. Update on Brain Tumor Imaging: From Anatomy to Physiology. **American Journal of Neuroradiology**, American Journal of Neuroradiology, v. 27, n. 3, p. 475–487, 2006. ISSN 0195-6108. Disponível em: https://www.ajnr.org/content/27/3/475.

CHA, S. Update on brain tumor imaging: from anatomy to physiology. **American Journal of Neuroradiology**, Am Soc Neuroradiology, v. 27, n. 3, p. 475–487, 2006.

CHAKRABARTY, N. **Brain MRI images for brain tumor detection**. San Francisco: Kaggle, 2017. Dataset. Disponível em: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection. Acesso em: 01 mar. 2024.

CHENG, J. Brain Tumor Dataset. **https://figshare.com/articles/dataset/brain_tumor_dataset/1512427**, 4 2017. Acesso em: 01 mar. 2024.

CHLAP, P.; MIN, H.; VANDENBERG, N.; DOWLING, J.; HOLLOWAY, L.; HAWORTH, A. A review of medical image data augmentation techniques for deep learning applications. **Journal of Medical Imaging and Radiation Oncology**, Wiley Online Library, v. 65, n. 5, p. 545–563, 2021.

CHOURMOUZI, D.; PAPADOPOULOU, E.; MARIAS, K.; DREVELEGAS, A. Imaging of brain tumors. **Surgical Oncology Clinics**, Elsevier, v. 23, n. 4, p. 629–684, 2014.

COUREUIL, M.; LÉCUYER, H.; BOURDOULOUS, S.; NASSIF, X. A journey into the brain: insight into how bacterial pathogens cross blood–brain barriers. **Nature Reviews Microbiology**, Nature Publishing Group UK London, v. 15, n. 3, p. 149–159, 2017.

DAI, C.; KANG, J.; LIU, X.; YAO, Y.; WANG, H.; WANG, R. How to classify and define pituitary tumors: recent advances and current controversies. **Frontiers in endocrinology**, Frontiers Media SA, v. 12, p. 604644, 2021.

DEANGELIS, L. M. Brain tumors. **New England journal of medicine**, Mass Medical Soc, v. 344, n. 2, p. 114–123, 2001.

DU, Z.; HE, C. Anisotropic diffusion with fuzzy-based source for binarization of degraded document images. **Applied Mathematics and Computation**, v. 441, p. 127684, 2023. ISSN 0096-3003. Disponível em: https://www.sciencedirect.com/science/article/pii/S0096300322007524.

EL-ASSIOUTI, O. S.; HAMED, G.; EL-SAADAWY, H.; EBIED, H. M.; KHATTAB, D. Regioninpaint, cutoff and regionmix: Introducing novel augmentation techniques for enhancing the generalization of brain tumor identification. **IEEE Access**, v. 11, p. 83232–83250, 2023.

EL-DAHSHAN, E.-S. A.; MOHSEN, H. M.; REVETT, K.; SALEM, A.-B. M. Computer-aided diagnosis of human brain tumor through mri: A survey and a new algorithm. **Expert Systems with Applications**, v. 41, n. 11, p. 5526–5545, 2014. ISSN 0957-4174. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417414000426.

FISHER, J. L.; SCHWARTZBAUM, J. A.; WRENSCH, M.; WIEMELS, J. L. Epidemiology of Brain Tumors. **Neurologic Clinics**, v. 25, n. 4, p. 867–890, 2007. ISSN 0733-8619. Brain Tumors in Adults. Disponível em: https://www.sciencedirect.com/science/article/pii/S0733861907000746.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. **Journal of the american statistical association**, Taylor & Francis, v. 32, n. 200, p. 675–701, 1937.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020.

GREENHOUSE, S. W.; GEISSER, S. On methods in the analysis of profile data. **Psychometrika**, Springer, v. 24, n. 2, p. 95–112, 1959.

HAO, S.; ZHOU, Y.; GUO, Y. A brief survey on semantic segmentation with deep learning. **Neurocomputing**, Elsevier, v. 406, p. 302–321, 2020.

HARSHAVARDHAN, A.; BABU, S.; VENUGOPAL, T. An improved brain tumor segmentation method from mri brain images. In: **2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT)**. [*S. l.*: *s. n.*], 2017. p. 1–7.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. **CoRR**, abs/1512.03385, 2015. Disponível em: http://arxiv.org/abs/1512.03385.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S. l.*: *s. n.*], 2016. p. 770–778.

HERNANDEZ-BOUSSARD, T.; BOZKURT, S.; IOANNIDIS, J. P.; SHAH, N. H. **MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care**. [*S. l.*]: Oxford University Press, 2020. 2011–2015 p.

HEYDARIAN, M.; DOYLE, T. E.; SAMAVI, R. MLCM: Multi-Label Confusion Matrix. **IEEE Access**, v. 10, p. 19083–19095, 2022.

HOWARD, A.; SANDLER, M.; CHU, G.; CHEN, L.-C.; CHEN, B.; TAN, M.; WANG, W.; ZHU, Y.; PANG, R.; VASUDEVAN, V. *et al.* Searching for mobilenetv3. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [*S. l.*: *s. n.*], 2019. p. 1314–1324.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. MobileNets: efficient convolutional neural networks for mobile vision applications (2017). **arXiv preprint arXiv:1704.04861**, v. 126, 2017.

HUANG, G.; LIU, Z.; MAATEN, L. van der; WEINBERGER, K. Q. **Densely Connected Convolutional Networks**. 2018. Disponível em: https://arxiv.org/abs/1608.06993.

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. **arXiv preprint arXiv:1602.07360**, 2016.

Instituto Nacional de Câncer. **Câncer do Sistema Nervoso Central**. 2022. Disponível em: https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/sistema-nervoso-central. Acesso em: 01 mar. 2024.

ISLAM, M. A.; NOSHIN, S. A.; ISLAM, M. R.; RAZY, M. F.; ANTARA, S.; REZA, M. T.; PARVEZ, M. Z. A low parametric cnn based solution to efficiently detect brain tumor cells from ultrasound scans. In: **2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)**. [*S. l.*: *s. n.*], 2023. p. 1152–1158.

JACCARD, P. The distribution of the flora in the alpine zone. 1. **New phytologist**, Wiley Online Library, v. 11, n. 2, p. 37–50, 1912.

JEME, V. J.; JEROME, S. A. Application of Bilateral Filters for Denoising Rician Noise in MRI Images. In: **2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)**. [*S. l.*: *s. n.*], 2023. p. 1291–1294.

KHAN, M. S. I.; RAHMAN, A.; DEBNATH, T.; KARIM, M. R.; NASIR, M. K.; BAND, S. S.; MOSAVI, A.; DEHZANGI, I. Accurate brain tumor detection using deep convolutional neural network. **Computational and Structural Biotechnology Journal**, Elsevier, v. 20, p. 4733–4745, 2022.

KOLO, B. **Binary and multiclass classification**. [*S. l.*]: Lulu.com, 2011.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.

KUMAR, T. S.; RASHMI, K.; RAMADOSS, S.; SANDHYA, L.; SANGEETHA, T. Brain tumor detection using svm classifier. In: **2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS)**. [*S. l.*: *s. n.*], 2017. p. 318–323.

KUSAKUNNIRAN, W.; BORWARNGINN, P.; KARNJANAPREECHAKORN, S.; THONGKANCHORN, K.; RITTHIPRAVAT, P.; TUAKTA, P.; BENJAPORNLERT, P. Encoder-decoder network with rmp for tongue segmentation. **Medical & Biological Engineering & Computing**, Springer, v. 61, n. 5, p. 1193–1207, 2023.

KUSHWAHA, V.; MAIDAMWAR, P. Btfcnn: Design of a brain tumor classification model using fused convolutional neural networks. In: **2022 10th International Conference on Emerging Trends in Engineering and Technology - Signal and Information Processing (ICETET-SIP-22)**. [*S. l.*: *s. n.*], 2022. p. 1–6.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.

LIN, T.-Y.; DOLLÁR, P.; GIRSHICK, R.; HE, K.; HARIHARAN, B.; BELONGIE, S. Feature pyramid networks for object detection. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S. l.*: *s. n.*], 2017. p. 2117–2125.

LIN, X.; MA, Y.-l.; MA, L.-z.; ZHANG, R.-l. A survey for image resizing. **Journal of Zhejiang University SCIENCE C**, Springer, v. 15, n. 9, p. 697–716, 2014.

LIU, H. Chapter 3 - rail transit collaborative robot systems. In: LIU, H. (Ed.). **Robot Systems for Rail Transit Applications**. Elsevier, 2020. p. 89–141. ISBN 978-0-12-822968-2. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128229682000036.

LIU, X.; BONNER, E. R.; JIANG, Z.; ROTH, H. R.; ANWAR, S. M.; PACKER, R. J.; BORNHORST, M.; LINGURARU, M. G. Automatic segmentation of rare pediatric brain tumors using knowledge transfer from adult data. In: **2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)**. [*S. l.*: *s. n.*], 2023. p. 1–4.

MARINOV, D.; KARAPETYAN, D. Hyperparameter Optimisation with Early Termination of Poor Performers. In: **2019 11th Computer Science and Electronic Engineering (CEEC)**. [*S. l.*: *s. n.*], 2019. p. 160–163.

MCFALINE-FIGUEROA, J. R.; LEE, E. Q. Brain Tumors. **The American Journal of Medicine**, v. 131, n. 8, p. 874–882, 2018. ISSN 0002-9343. Disponível em: https://www.sciencedirect.com/science/article/pii/S0002934318300317.

MEFTAH, A. H.; ALOTAIBI, Y. A.; SELOUANI, S.-A. Evaluation of an Arabic Speech Corpus of Emotions: A Perceptual and Statistical Analysis. **IEEE Access**, v. 6, p. 72845–72861, 2018.

METE, O.; LOPES, M. B. Overview of the 2017 who classification of pituitary tumors. **Endocrine pathology**, Springer, v. 28, p. 228–243, 2017.

MÜLLER, D.; SOTO-REY, I.; KRAMER, F. Towards a guideline for evaluation metrics in medical image segmentation. **BMC Research Notes**, Springer, v. 15, n. 1, p. 210, 2022.

MUSA, P.; RAFI, F. A.; LAMSANI, M. A Review: Contrast-Limited Adaptive Histogram Equalization (CLAHE) methods to help the application of face recognition. In: IEEE. **2018 third international conference on informatics and computing (ICIC)**. [*S. l.*], 2018. p. 1–6.

NASER, M. Z.; ALAVI, A. Insights into performance fitness and error metrics for machine learning. **CoRR**, abs/2006.00887, 2020. Disponível em: https://arxiv.org/abs/2006.00887.

NEIDEEN, T.; BRASEL, K. Understanding Statistical Tests. **Journal of Surgical Education**, v. 64, n. 2, p. 93–96, 2007. ISSN 1931-7204. Disponível em: https://doi.org/10.1016/j.jsurg.2007.02.001.

NEMENYI, P. B. **Distribution-free multiple comparisons.** [*S. l.*]: Princeton University, 1963.

NEUROCIRURGIA, I. I. de. **Gliomas**. 2024. Acessado em: 14-07-2024. Disponível em: https://inecsp.com.br/tratamentos/gliomas/.

OLVERES, J.; GONZÁLEZ, G.; TORRES, F.; MORENO-TAGLE, J. C.; CARBAJAL-DEGANTE, E.; VALENCIA-RODRÍGUEZ, A.; MÉNDEZ-SÁNCHEZ, N.; ESCALANTE-RAMÍREZ, B. What is new in computer vision and artificial intelligence in medical image analysis applications. **Quantitative imaging in medicine and surgery**, AME Publications, v. 11, n. 8, p. 3830, 2021.

O'SHEA, K.; NASH, R. An introduction to convolutional neural networks. **arXiv preprint arXiv:1511.08458**, 2015.

OTHMAN, N. A.; ZAKARIA, N. A. C.; HANAPIAH, F. A.; HASHIM, N. M.; JOHAR, K.; LOW, C. Y.; YEE, J. Quantifying the Performance of Wireless Data Acquisition System to Assess Upper Limb Spasticity. In: **2022 IEEE 5th International Symposium in Robotics and Manufacturing Automation (ROMA)**. [*S. l.*: *s. n.*], 2022. p. 1–4.

OTTOM, M. A.; RAHMAN, H. A.; DINOV, I. D. Znet: Deep learning approach for 2d mri brain tumor segmentation. **IEEE Journal of Translational Engineering in Health and Medicine**, v. 10, p. 1–8, 2022.

PADMAPRIYA, S.; DEVI, M. G. Computer-Aided Diagnostic System for Brain Tumor Classification using Explainable AI. In: **2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)**. [*S. l.*: *s. n.*], 2024. v. 2, p. 1–6.

PANNU, A. Artificial intelligence and its application in different areas. **Artificial Intelligence**, v. 4, n. 10, p. 79–84, 2015.

PEDDINTI, A. S.; MALOJI, S.; MANEPALLI, K. Evolution in diagnosis and detection of brain tumor–review. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [*S. l.*], 2021. v. 2115, n. 1, p. 012039.

PERUMAL, S.; VELMURUGAN, T. Preprocessing by contrast enhancement techniques for medical images. **International Journal of Pure and Applied Mathematics**, v. 118, n. 18, p. 3681–3688, 2018.

RAGHAVENDRA, U.; GUDIGAR, A.; PAUL, A.; GOUTHAM, T.; INAMDAR, M. A.; HEGDE, A.; DEVI, A.; OOI, C. P.; DEO, R. C.; BARUA, P. D.; MOLINARI, F.; CIACCIO, E. J.; ACHARYA, U. R. Brain tumor detection and screening using artificial intelligence techniques: Current trends and future perspectives. **Computers in Biology and Medicine**, v. 163, p. 107063, 2023. ISSN 0010-4825. Disponível em: https://www.sciencedirect.com/science/article/pii/S0010482523005280.

RAJ, J. R. F.; VIJAYALAKSHMI, K.; PRIYA, S. K.; APPATHURAI, A. Brain tumor segmentation based on kernel fuzzy c-means and penguin search optimization algorithm. **Signal, Image and Video Processing**, Springer, v. 18, n. 2, p. 1793–1802, 2024.

REHMAN, A.; NAZ, S.; RAZZAK, M. I.; AKRAM, F.; IMRAN, M. A deep learning-based framework for automatic brain tumors classification using transfer learning. **Circuits, Systems, and Signal Processing**, v. 39, 2020.

REZA, A. M. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. **Journal of VLSI signal processing systems for signal, image and video technology**, Springer, v. 38, p. 35–44, 2004.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [*S. l.*], 2015. p. 234–241.

ROSHINTA, T. A.; DINATA, I. F.; RIATMA, D. L.; SYAFI'I, M. A.; FIRDAUS, N.; A'LA, F. Y. A Comparison of Prediction Algorithms in Food Sales with Different K-Folds Cross-Validation. In: **2023 6th International Conference of Computer and Informatics Engineering (IC2IE)**. [*S. l.*: *s. n.*], 2023. p. 314–318.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S. l.*: *s. n.*], 2018. p. 4510–4520.

SARKER, I. H. Ai-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. **SN Computer Science**, Springer, v. 3, n. 2, p. 158, 2022.

SCIKIT-LEARN. **3.1. Cross-validation: evaluating estimator performance**. 2007 – 2022. https://scikit-learn.org/stable/modules/cross_validation.html.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, Oxford University Press, v. 52, n. 3-4, p. 591–611, 1965.

SHAREN, H.; JAWAHAR, M.; ANBARASI, L. J.; RAVI, V.; ALGHAMDI, N. S.; SULIMAN, W. Fdum-net: An enhanced fpn and u-net architecture for skin lesion segmentation. **Biomedical Signal Processing and Control**, Elsevier, v. 91, p. 106037, 2024.

SHARMA, S.; SHARMA, S.; ATHAIYA, A. Activation functions in neural networks. **International Journal of Engineering Applied Sciences and Technology**, v. 04, p. 310–316, 05 2020.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SINGH, R.; KUMAR, D.; SAGAR, B. B. Valuation of Significant Difference Between Various Agile Methods Using One Way ANOVA. In: **2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)**. [*S. l.*: *s. n.*], 2021. p. 1–5.

SRAVANI, C. L.; MIRIYALA, S. S.; MITRA, K. Statistical inference and analysis for efficient modeling of environmental pollution using deep neural networks. In: **2022 Eighth Indian Control Conference (ICC)**. [*S. l.*: *s. n.*], 2022. p. 385–390.

SRIVASTAVA, A.; KHARE, A.; KUSHWAHA, A. Brain tumor classification using deep learning framework. In: **2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)**. [*S. l.*: *s. n.*], 2023. p. 1–4.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [*S. l.*: *s. n.*], 2015. p. 1–9.

TAHA, A. A.; HANBURY, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. **BMC medical imaging**, Springer, v. 15, p. 1–28, 2015.

TALEBI, H.; MILANFAR, P. Learning to resize images for computer vision tasks. **CoRR**, abs/2103.09950, 2021. Disponível em: https://arxiv.org/abs/2103.09950.

TAN, L.; JIANG, J. Chapter 14 - image processing basics. In: TAN, L.; JIANG, J. (Ed.). **Digital Signal Processing (Second Edition)**. Second edition. Boston: Academic Press, 2013. p. 683–765. ISBN 978-0-12-415893-1. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780124158931000147.

TAN, M. Efficientnet: Rethinking model scaling for convolutional neural networks. **arXiv preprint arXiv:1905.11946**, 2019.

TANDEL, G. S.; BALESTRIERI, A.; JUJARAY, T.; KHANNA, N. N.; SABA, L.; SURI, J. S. Multiclass magnetic resonance imaging brain tumor classification using artificial intelligence paradigm. **Computers in Biology and Medicine**, v. 122, p. 103804, 2020. ISSN 0010-4825. Disponível em: https://www.sciencedirect.com/science/article/pii/S0010482520301724.

TAYLOR, L.; NITSCHKE, G. **Improving Deep Learning using Generic Data Augmentation**. arXiv, 2017. Disponível em: https://arxiv.org/abs/1708.06020.

TING, K. M. Confusion Matrix. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 209–209. ISBN 978-0-387-30164-8. Disponível em: https://doi.org/10.1007/978-0-387-30164-8_157.

TOMASI, C.; MANDUCHI, R. Bilateral filtering for gray and color images. In: **Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)**. [*S. l.*: *s. n.*], 1998. p. 839–846.

TUKEY, J. W. Comparing individual means in the analysis of variance. **Biometrics**, JSTOR, p. 99–114, 1949.

UNIVERSITY, S. **Convolutional Neural Networks (CNNs / ConvNets)**. 2022. Disponível em: https://cs231n.github.io/convolutional-networks/#pool. Acesso em: 01 mar. 2024.

VARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In: SN. **Proceedings of the xxix conference on graphics, patterns and images**. [*S. l.*], 2016. v. 1, n. 4.

VAZ, J. M.; BALAJI, S. Convolutional neural networks (CNNs): Concepts and applications in pharmacogenomics. **Molecular diversity**, Springer, v. 25, n. 3, p. 1569–1584, 2021.

WANG, S.; CHAOVALITWONGSE, W. A. Evaluating and comparing forecasting models. **Wiley Encyclopedia of Operations Research and Management Science, eorms0307. https://doi. org/10.1002/9780470400531. eorms0307**, 2011.

WANG, X.; HU, Z.; SHI, S.; HOU, M.; XU, L.; ZHANG, X. A deep learning method for optimizing semantic segmentation accuracy of remote sensing images based on improved unet. **Scientific reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 7600, 2023.

WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. **Journal of Big data**, Springer, v. 3, p. 1–40, 2016.

WILCOX, R. R. **Applying contemporary statistical techniques**. [*S. l.*]: Elsevier, 2003.

WULANDARI, A.; SIGIT, R.; BACHTIAR, M. M. Brain tumor segmentation to calculate percentage tumor using mri. In: **2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC)**. [*S. l.*: *s. n.*], 2018. p. 292–296.

YAMASHITA, R.; NISHIO, M.; DO, R.; TOGASHI, K. Convolutional neural networks: an overview and application in radiology. **Insights into Imaging**, v. 9, 06 2018.

YAN, C.; DING, J.; ZHANG, H.; TONG, K.; HUA, B.; SHI, S. Seresu-net for multimodal brain tumor segmentation. **IEEE Access**, v. 10, p. 117033–117044, 2022.

YOON, H. J.; JEONG, Y. J.; KANG, H.; JEONG, J. E.; KANG, D.-Y. Medical image analysis using artificial intelligence. **Progress in Medical Physics**, Korean Society of Medical Physics, v. 30, n. 2, p. 49–58, 2019.

ZHANG, J. M.; HARMAN, M.; MA, L.; LIU, Y. Machine learning testing: Survey, landscapes and horizons. **IEEE Transactions on Software Engineering**, IEEE, v. 48, n. 1, p. 1–36, 2020.

ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH, N.; LIANG, J. Unet++: A nested u-net architecture for medical image segmentation. In: SPRINGER. **Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4**. [*S. l.*], 2018. p. 3–11.

ZHU, J.; GU, C.; WEI, L.; LI, H.; JIANG, R.; SHEYKHAHMAD, F. R. Brain tumor recognition by an optimized deep network utilizing ammended grasshopper optimization. **Heliyon**, Elsevier, v. 10, n. 7, 2024.

ZHUANG, F.; QI, Z.; DUAN, K.; XI, D.; ZHU, Y.; ZHU, H.; XIONG, H.; HE, Q. A Comprehensive Survey on Transfer Learning. **Proceedings of the IEEE**, v. 109, n. 1, p. 43–76, 2021.

ZUIDERVELD, K. Contrast limited adaptive histogram equalization. In: **Graphics gems IV**. [*S. l.*: *s. n.*], 1994. p. 474–485.