UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

ANTONIO ERMESON PEREIRA ALVES

AVALIAÇÃO DE DESEMPENHO DE REDES NEURAIS PROFUNDAS PARA SEGMENTAÇÃO PULMONAR EM TOMOGRAFIAS COMPUTADORIZADAS DO TÓRAX

FORTALEZA

2026

ANTONIO ERMESON PEREIRA ALVES

AVALIAÇÃO DE DESEMPENHO DE REDES NEURAIS PROFUNDAS PARA SEGMENTAÇÃO PULMONAR EM TOMOGRAFIAS COMPUTADORIZADAS DO TÓRAX

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Paulo César Cortez

FORTALEZA

2026

ANTONIO ERMESON PEREIRA ALVES

AVALIAÇÃO DE DESEMPENHO DE REDES NEURAIS PROFUNDAS PARA SEGMENTAÇÃO PULMONAR EM TOMOGRAFIAS COMPUTADORIZADAS DO TÓRAX

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Computação.

Aprovado em: 29 de Janeiro de 2026.

BANCA EXAMINADORA

Prof. Dr. Paulo César Cortez (Orientador)
Universidade Federal do Ceará (UFC)

Dra. Débora Ferreira de Assis (Coorientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Bruno Riccelli dos Santos Silva
Universidade Federal do Ceará (UFC)

Me. Pedro Crosara Motta
Universidade Federal do Rio de Janeiro (UFRJ)

Ma. Andressa Gomes Moreira
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

A Deus pela minha vida, pela vida dos meus e por Seu amor, bem como pelas oportunidades diárias de conversão verdadeira que Ele me concede. O agradeço também pelas abundantes graças e milagres concedidos.

À Nossa Senhora, Maria Santíssima, por Sua constante intercessão e Sua presença tão fiel e amorosa na minha vida. Suas virtudes são exemplos belíssimos e seus cuidados maternos aliviam qualquer tempestade.

A minha família, em especial à minha mãe, Carmilda Pereira, e minha madrinha, Antonia de Fátima. A educação que me deram fez-me ser o que sou hoje e estar onde estou.

A minha namorada, Jeisilyane Silveira, por estar presente em inúmeros momentos de minha graduação, fossem alegres ou tristes, e por sempre me apoiar a ser uma pessoa melhor. Conhecê-la durante a graduação foi sem dúvidas uma das melhores coisas que poderia me ter ocorrido.

Aos amigos que conheci no LESC, tanto membros da graduação quanto da pós-graduação (alguns já doutores). Devo uma enorme parte dos meus conhecimentos atuais a vivência com essas pessoas tão boas que Deus colocou em minha vida.

Ao meu orientador, o Prof.Dr. Paulo César Cortez, por suas importantes contribuições em minha trajetória, desde o segundo semestre (onde a primeira oportunidade de iniciação científica representou um fator decisivo para minha persistência no curso) até os dias de hoje.

À Divisão de Benefícios e Moradia (DIBEM) da PRAE, pela manutenção do programa de Residências Universitárias da UFC. Este importante programa de assistência estudantil representa um segundo fator decisivo para minha persistência no curso.

“Vinde a mim, todos os que estais cansados e oprimidos, e eu vos aliviarei. Tomai sobre vós o meu jugo, e aprendei de mim, que sou manso e humilde de coração; e encontrareis descanso para as vossas almas. Porque o meu jugo é suave e o meu fardo é leve.”

(Mt 11, 28-30)

RESUMO

As doenças respiratórias figuram entre as principais causas de morte em todo o mundo, o que torna essenciais métodos de diagnóstico por imagem precisos, como a Tomografia Computadorizada (TC). Neste contexto, a segmentação semântica do pulmão desempenha um papel indispensável no fluxo de sistemas de Diagnóstico Assistido por Computador (CADe). O presente trabalho tem como objetivo avaliar, de forma sistemática, o desempenho de diferentes arquiteturas de Aprendizado Profundo, abrangendo tanto Redes Neurais Convolucionais (CNNs) quanto *Vision Transformers*, na tarefa de segmentação pulmonar. Para os experimentos, utilizou-se o conjunto de dados LOCCA, composto por volumes do HCU e do TASK06. As imagens foram submetidas a um pipeline de pré-processamento que incluiu janelamento, CLAHE e filtro da mediana. Foram comparadas as arquiteturas UNet++, DeepLabv3+, DPT e SegFormer, utilizando a técnica de validação cruzada K-Fold (k=5) com três repetições para garantir a robustez dos dados. As métricas de avaliação incluíram IoU, F1-Score, Sensibilidade, Especificidade, além de tempo de processamento e consumo de VRAM. Os resultados indicaram que, embora os modelos apresentem equivalência estatística nas métricas de qualidade de segmentação, houve divergência significativa no custo computacional. A arquitetura DeepLabv3+ destacou-se por apresentar o melhor equilíbrio entre desempenho e eficiência de recursos, enquanto o modelo DPT apresentou o maior custo de tempo e memória.

Palavras-chave: Segmentação Semântica. Tomografia Computadorizada de Tórax. Doenças Pulmonares. Aprendizado Profundo. Avaliação de Desempenho.

[1]

[1] Código fonte disponível em https://github.com/ermeson-alves/meu_tcc

ABSTRACT

Respiratory diseases rank among the leading causes of death worldwide, making accurate imaging-based diagnostic methods, such as Computed Tomography (CT), essential. In this context, lung semantic segmentation plays a crucial role in the workflow of Computer-Aided Diagnosis (CADe) systems. The objective of this study is to systematically evaluate the performance of different Deep Learning architectures, encompassing both Convolutional Neural Networks (CNNs) and Vision Transformers, in the task of lung segmentation. The experiments were conducted using the LOCCA dataset, composed of volumes from HCU and TASK06. The images were processed through a preprocessing pipeline that included windowing, CLAHE, and median filtering. The architectures UNet++, DeepLabv3+, DPT, and SegFormer were compared using the K-Fold cross-validation technique (k=5) with three repetitions to ensure data robustness. Evaluation metrics included IoU, F1-score, Sensitivity, Specificity, as well as processing time and VRAM consumption. The results indicated that, although the models exhibited statistical equivalence in segmentation quality metrics, there were significant differences in computational cost. The DeepLabv3+ architecture stood out by providing the best balance between performance and resource efficiency, while the DPT model showed the highest time and memory cost.

Keywords: Semantic Segmentation. Chest Computed Tomography. Pulmonary Diseases. Deep Learning. Performance Evaluation.

[2]

[2] Code available at https://github.com/ermeson-alves/meu_tcc

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CADe	*Computer-Aided Detection*
CLAHE	Equalização de histograma adaptativa com limitação de contraste
CNNs	*Convolutional Neural Networks*/Redes Neurais Convolucionais
DL	*Deep Learning*/Aprendizado Profundo
FIRS	Fórum Internacional de Sociedades Respiratórias
MAs	Mecanismos de Atenção
PET-TC	Tomografia por Emissão de Pósitrons
TC	Tomografia Computadorizada

SUMÁRIO

1 INTRODUÇÃO

De acordo com relatório de 2024 da Organização Mundial de Saúde (OMS) (World Health Organization, 2024), as 10 principais causas de morte em todo o mundo se dividem em dois grandes temas: cardiovascular e respiratório. No mesmo relatório consta um ranking das 10 principais causas de morte (considerando o invervalo de 2000 a 2021) em todo o mundo, no qual destacam-se: COVID-19 (2º lugar), Doença Pulmonar Obstrutiva Crônica (4º lugar), infecções do trato respiratório inferior (5º lugar) e cânceres de traqueia, brônquios e pulmão (6º lugar).

As doenças pulmonares também são conhecidas por doenças das vias aéreas e prejudicam estruturas anatômicas essenciais para a respiração humana. Algumas dessas doenças mais relevantes são: pneumonia, tuberculose, COVID-19, Câncer de Pulmão, Doença Pulmonar Obstrutiva Crônica (DPOC), fibrose pulmonar etc. De acordo com o Fórum Internacional de Sociedades Respiratórias (*FIRS*), cerca de 200 milhões de pessoas sofrem de DPOC e 3,2 milhões morrem por ano. De forma análoga, anualmente, a pneumonia é responsável pela morte de 2,4 milhões de pessoas e o câncer de pulmão por 1,8 milhões (sendo o mais letal dos cânceres) (Forum of International Respiratory Societies, 2021).

1.1 Tipos de Imagens Médicas Utilizadas

As doenças pulmonares podem ser diagnosticadas por meio de exames de imagem como Raio-X, Ressonância Magnética (RM), Tomografia por Emissão de Pósitrons (*PET-TC*) e a Tomografia Computadorizada (*TC*). Cada uma das formas de gerar imagem apresenta suas particularidades e seu grau de detalhamento.

Esses exames desempenham um papel importante no diagnóstico e tratamento de doenças pulmonares; a detecção precoce é uma das estratégias para reduzir a mortalidade desses tipos de doenças (HU *et al.*, 2020). Além disso, Hu *et al.* (2020) explicam que o primeiro passo para diagnóstico de doenças pumonares a partir de *TC* é **delimitação de regiões pulmonares**.

O Raio-X consiste na emissão de feixes em direção ao corpo do paciente. A intensidade da cor preta determina se esses raios foram ou não bloqueados. A imagem gerada por esse procedimento possui duas dimensões e as estruturas corporais ficam sobrepostas. Apesar de ser uma modalidade preferida entre os profissionais de saúde, com relação ao seu baixo custo e carga ínfima de radiação, há maiores dificuldades para a identificação de pequenas lesões características.

A técnica com maior grau de detalhamento consiste na *TC*. Ela consiste na realização imagens em de fatias milimétricas que posteriormente são processadas e se tornam uma imagem em 3 dimensões. Além desse procedimento minucioso, a substância contraste (a base de iodo) pode ser consumida, o que destaca as estruturas como vasos sanguíneos, dessa maneira, permitindo observar pequenas lesões e diferenciar tumores, inflamações e cicatrizes. Assim, facilitando o diagnóstico de embolia pulmonar, tumores, infecções profundas e sangramentos.

O uso de *TC* por modelos de IA, é justificado pela formação detalhada das imagens, a minuciosidade em sua interpretação e a variedade de possíveis mazelas identificadas pela técnica.

1.2 Diagnóstico auxiliado por Computador

Sistemas de detecção auxiliada por computador, ou "*Computer-Aided Detection* (*CADe*) *systems*", têm auxiliado especialistas a identificar doenças pulmonares e determinar a progressão dessas doenças com maior precisão (KUMAR *et al.*, 2024). Eles estabelecem um pipeline de dados que reduz a carga de trabalho de especialistas e minimizam os erros humanos através da geração automática de relatórios.

Os sistemas *CADe* colaboram para o alcance de metas do *FIRS*, bem como da OMS, pois com uma maior quantidade de sistemas como esse o alcance das metas "acesso universal a serviços saúde de qualidade" e "diagnóstico precoce de doenças respiratórias" podem ser alcançadas. Os experimentos e investigações do presente trabalho contribuem para a construção de sistemas *CADe*.

1.3 Objetivos

O avanço do Aprendizado Profundo na área médica resultou em uma ampla diversidade de arquiteturas baseadas em CNNs e, mais recentemente, em Transformers, cada uma apresentando diferentes níveis de desempenho e custo computacional (GUPTA *et al.*, 2022; SHAH *et al.*, 2025). No contexto da segmentação pulmonar, a inexistência de uma arquitetura universalmente superior torna necessária a realização de avaliações comparativas e sistemáticas, considerando simultaneamente métricas de qualidade da segmentação e eficiência computacional (JAVED *et al.*, 2024; KUMAR *et al.*, 2024).

Nessa lógica, como objetivo geral, esse trabalho pretende realizar uma avaliação

de desempenho de redes neurais profundas baseadas em CNNs e/ou Transformers na tarefa de segmentação semântica pulmonar multiclasse, a fim de entender que método traz os melhores resultados ao passo que seu custo computacional é justificável.

Os objetivos específicos desse trabalho são:

- Implementar e treinar arquiteturas de Aprendizado Profundo baseadas em CNNs e Vision Transformers para a segmentação pulmonar multiclasse.
- Avaliar o desempenho das arquiteturas UNet++, DeepLabv3+, DPT e SegFormer utilizando métricas consolidadas da literatura, como IoU, F1-score, Sensibilidade e Especificidade.
- Comparar o custo computacional dos modelos avaliados, considerando tempo de treinamento, tempo de inferência e consumo de memória gráfica (VRAM).
- Identificar a arquitetura que apresenta o melhor equilíbrio entre qualidade de segmentação e eficiência computacional no contexto avaliado.

1.4 Organização do Trabalho

Este trabalho está organizado da seguinte forma: o Capítulo 1 apresenta a introdução, contextualizando o problema, os objetivos e a motivação do estudo. O Capítulo 2 discute a fundamentação teórica, abordando conceitos de redes neurais convolucionais, arquiteturas baseadas em Transformers, segmentação semântica e técnicas de pré-processamento. No Capítulo 3 são apresentados os trabalhos relacionados, destacando abordagens recentes e relevantes da literatura. O Capítulo 4 descreve a metodologia adotada para a avaliação de desempenho, incluindo o conjunto de dados, o pipeline experimental, as métricas e os testes estatísticos. O Capítulo 5 apresenta e discute os resultados obtidos. Por fim, o Capítulo 6 reúne as conclusões do trabalho e aponta direções para pesquisas futuras.

2 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo serão apresentados os conceitos fundamentais para o entendimento deste trabalho. Inicialmente, os conceitos relacionados a *Convolutional Neural Networks*/Redes Neurais Convolucionais (*CNNs*) são abordados, levando-se em consideração seu histórico e estrutura. Conceitos de aumento de dados e aprendizagem por transferência também são explorados. Posteriormente, o foco é dado para as arquitetura baseadas em *Transformers*, as quais surgem do contexto de processamento de linguagem natural e vêm sendo progressivamente adaptadas para tarefas de visão computacional. Em seguida, conceitos de Segmentação Semântica, bem como a apresentação breve de algumas redes para este tipo de tarefa, são apresentados. Por fim, são apresentadas as técnicas de pré-processamentos utilizadas nesse trabalho e a importância de cada uma.

2.1 Redes Neurais Artificiais Convolucionais

CNNs são modelos de Inteligência Artificial agrupados no ramo conhecido como *Deep Learning*/Aprendizado Profundo (*DL*). O termo *DL* surge do fato de que modelos presentes nessa categoria possuem um grande número de parâmetros e também um grande número de camadas ocultas (as que estão entre a camada de entrada e a de saída). Ao longo das camadas, tais modelos conseguem reconhecer padrões cada vez mais complexos e possuem um número enorme de conexões entre os neurônios artificiais. Somado a isso, o número de dados de treinamento deve ser bem maior para um reconhecimento de padrões adequado, dái o termo profundo (GUPTA *et al.*, 2022). É comum utilizar um diagrama de Venn para mostrar a relação entre IA, Aprendizado de Máquina e Aprendizado profundo. Um diagrama como esse é exibido na Figura 1.

Uma característica fundamental do aprendizado profundo é a extração de características dos dados de entrada de forma automatizada. Anteriormente, com o paradigma de aprendizado de máquina clássico, inúmeros métodos deveriam ser testados para extrair informações quantitativas relevantes sobre esses dados, mas com o surgimento do aprendizado profundo, grande parte dessa etapa foi facilitada (GUPTA *et al.*, 2022). Os sistemas de DL permitem transformar imagens de entrada em saídas úteis em diferentes tarefas como: classificação, detecção e segmentação semântica (MARGERIE-MELLON; CHASSAGNON, 2023).

Figura 1 – Diagrama de Venn relacionando IA, ML e DL.

Fonte: Elaborada pelo autor.

De acordo com Gonzalez e Woods (2018):

> Uma imagem pode ser definida como uma função bidimensional, $f(x,y)$, em que x e y são coordenadas espaciais (plano), e a amplitude de f em qualquer par de coordenadas (x,y) é chamada de intensidade ou nível de cinza da imagem nesse ponto.

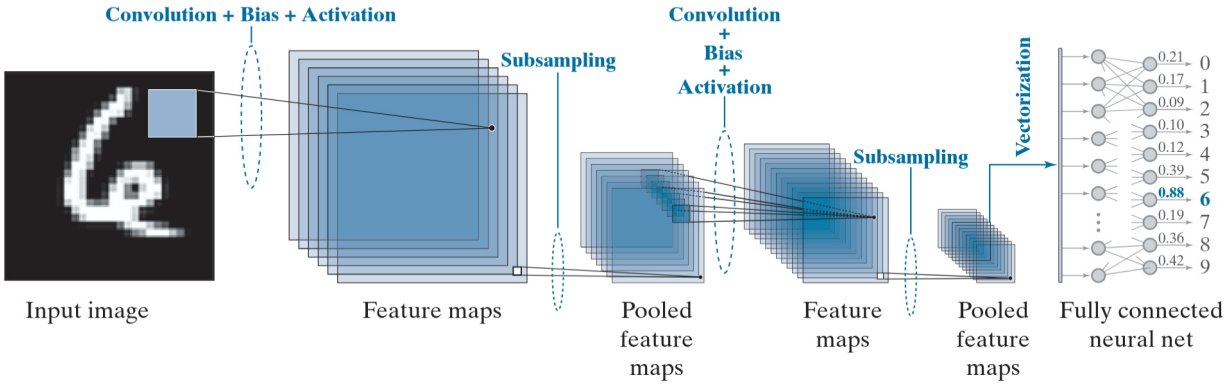
Vale ressaltar que tal definição se estende facilmente para imagens com mais de duas dimensões, ou seja: sempre que há um conjunto de imagens 2D relacionadas. Devido a bidimensionalidade das imagens, muitos dos métodos tradicionais de aprendizado de maquina são incompatíveis, dado que esperam uma entrada unidimensional. Por essa razão, as CNNs se tornaram fundamentais para diferentes atividades envolvendo imagens, seja classificação, segmentação semântica, detecção etc (KOUTOULAKIS *et al.*, ; ARCHANA; JEEVARAJ, 2024; ANTONELLI *et al.*, 2022).

Gonzalez e Woods (2018) definem a convolução de um filtro $w(x,y)$ m x n com uma imagem $f(x,y)$ é dada pela equação 2.1, em que $a = (m-1)/2$ e $b = (n-1)/2$. Essa operação é fundamental no ramo do processamento digital de imagens por permitir a realização de filtragem no plano de pixels da imagem (ou domínio espacial). Além disso, essa é a base por trás das camadas convolucionais de CNNs.

$$w(x,y) \star f(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t) f(x-s, y-t) \tag{2.1}$$

Conforme fora supracitado, CNNs em geral possuem um grande número de camadas, mas também há diferentes tipos de camadas como: camada de convolução, camada de *pooling*, camada totalmente conectada, etc. Entender a diferença entre tais camadas é fundamental. A Figura 2 exibe uma arquitetura base de CNN.

Figura 2 – Arquitetura básica de uma Rede Neural Convolucional (CNN), aplicada a um problema de classificação.



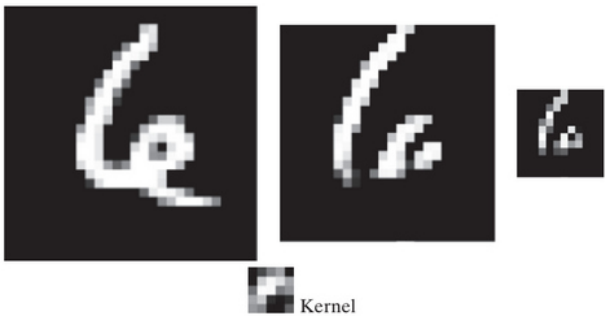
Fonte: (GONZALEZ; WOODS, 2018).

A camada de convolução é composta por um volume de *kernels* (ou filtros de convolução) que são deslocados pelos dados da camada anterior e funcionam principalmente como extratores de características ou ativações em posições espaciais das imagens (O'SHEA; NASH, 2015). Destaca-se que o "aprendizado" das *CNNs* ocorre justamente no ajuste de valores presentes nos kernels. Esse é o motivo pelo qual O'Shea e Nash (2015) usam o termo "**kernels aprendíveis**".

Percebe-se que ao longo das camadas as imagens perdem resolução espacial e ganham profundidade (há um volume cada vez mais denso de mapas de características). Nas primeiras camadas, os filtros de convolução conseguem capturar características espaciais mais simples (como bordas, linhas e texturas simples) já nas camadas finais é possível capturar padrões mais complexos devido a aplicação sequencial de filtros convolucionais. A Figura 3 ilustra bem a captura de linhas diagonais com um kernel de convolução, seguida de uma redução espacial via agrupamento (pooling).

Um conceito muito importante apresentado por Gonzalez e Woods (2018) é o de "**campos receptivos**", que se trata de uma vizinhança de pixels da imagem de entrada de mesma resolução que um kernel de convolução aplicado e cujas coordenadas espaciais também correspondem as coordenadas espaciais dos pixels deste kernel, em uma determinada iteração da convolução.

Figura 3 – Captura de linhas diagonais e redução espacial de uma imagem de entrada.



Fonte: (GONZALEZ; WOODS, 2018).

A camada de pooling serve para reduzir a dimensionalidade dos dados dados e, consequentemente, a carga de processamento ao mesmo passo que preserva características espaciais relevantes (O'SHEA; NASH, 2015). Como exemplo, um kernel de pooling 2x2 é deslocado pelos dados de entrada e apenas o pixel de maior valor é considerado (max-pooling).

Na camada totalmente conectada há neurônios que estão diretamente conectados com os neurônios das camadas adjacentes. Inclusive, isso é análogo a forma como os neurônios estão distribuídos em redes neurais tradicionais (O'SHEA; NASH, 2015). Após essa camada, há a saída da CNN tradicional que pode ser vista como um conjunto de probabilidades para cada uma das classes em questão, partindo do problema tradicional de classificação.

2.1.1 Aumento de Dados

O aumento de dados (do inglês: *Data Augmentation)* é um conjunto de métodos para gerar novos dados sintéticos a partir dos dados existentes e é útil para problemas como os da engenharia biomédica. Quando os dados de origem são imagens, é comum o uso de transformações geométricas e de valores de intensidade dos pixels (a fim de gerar diferentes perpectivas ou simular condições diversas de iluminação).

Devido a necessidade de uma grande quantidade de dados para treinamento de modelos de aprendizado profundo, muitas vezes a precisão para problemas da área médica é afetada. É comum que haja certos dados sensíveis nesse contexto e também um número reduzido de especialistas dispostos a anotar os dados. Somado a isso, o processo de anotação sem um software adequado é muito trabalhoso.

2.1.2 Transfer Learning

O *transfer learning* é uma estratégia amplamente utilizada em tarefas de visão computacional e consiste em reutilizar modelos previamente treinados em grandes bases de dados como ponto de partida para um novo problema. Dessa forma, ao empregar um modelo pré-treinado, é possível aproveitar esse conhecimento previamente adquirido, reduzindo o tempo de treinamento e melhorando a capacidade de generalização do modelo, especialmente em cenários com conjuntos de dados limitados.

No contexto da segmentação de imagens, o transfer learning é geralmente aplicado ao encoder da arquitetura, cuja função é extrair representações semânticas progressivamente mais abstratas a partir da imagem de entrada. A utilização de um encoder pré-treinado contribui para uma extração de características mais robusta e estável, além de reduzir o risco de overfitting. O decodificador, por sua vez, é treinado especificamente para a tarefa de segmentação, sendo responsável por reconstruir os mapas de segmentação na resolução original da imagem.

Neste trabalho, a MobileNetV2 (SANDLER *et al.*, 2019) foi fixada como encoder da rede de segmentação devido à sua elevada eficiência computacional. Essa arquitetura foi projetada com foco na redução do número de parâmetros e do custo de processamento, empregando convoluções separáveis em profundidade e blocos residuais invertidos. Como resultado, a MobileNetV2 oferece um bom equilíbrio entre desempenho e eficiência, tornando-se adequada para aplicações que demandam menor tempo de treinamento e inferência, sem comprometer significativamente a qualidade das segmentações produzidas.

2.2 Arquiteturas baseadas em Transformers

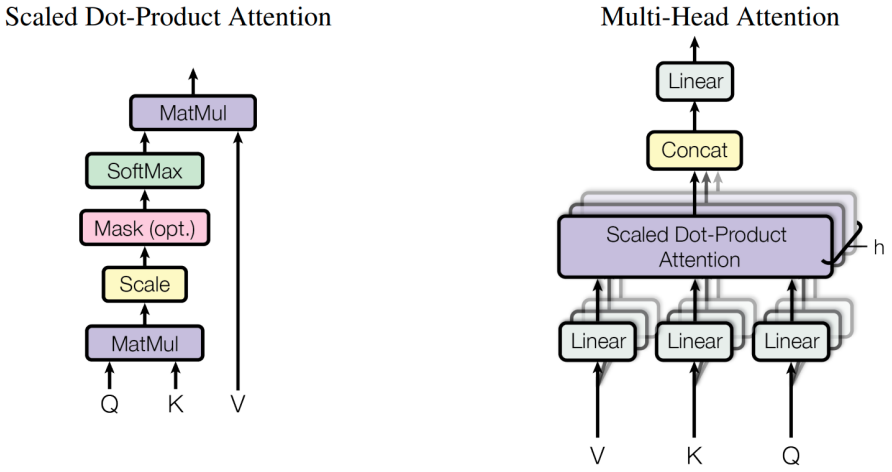
Apesar das vantagens oferecidas pelas *CNNs*, percebe-se uma rápida ascensão de modelos de *DL* baseados em *Transformers* para segmentação semântica de imagens nos últimos anos. Essa arquitetura, introduzida por Vaswani *et al.* (2017) surgiu no contexto de problemas de transdução, mais especificamente para tradução de textos, quando redes neurais recorrentes e redes neurais tradicionais eram o estado da arte.

Vaswani *et al.* (2017) descrevem que tais arquiteturas partem do paradigma codificador-decodificador utilizado anteriormente, onde o codificador atua mapeando uma sequência de símbolos $(x_1, ..., x_n)$ para uma "sequência de representações contínuas"$(z_1, ..., z_m)$. O decodificador, por sua vez, gera a sequência de símbolos de sáida $(y_1, ..., y_m)$.

Uma grande mudança trazida pelos *Transformers* foi uma arquitetura completamente baseado nos Mecanismos de Atenção (*MAs*), método inspirado na característica da atenção humana de focar nos detalhes mais importantes de algo. Os *MAs* cumprem a função crucial de priorizar determinas partes de uma entrada, isto é: a importância relativa de cada parte.

Além disso, *MAs* permitem a paralelização no processamento das entradas, algo extremamente significativo, dado que anteriormente tal processamento ocorria de forma sequencial. (VASWANI *et al.*, 2017), definem atenção como o mapeamento de uma consulta e um conjunto de pares chave-valor para uma saída, em que ambos os componentes são vetores, dessa forma eles explicam que a saída é como uma soma ponderada dos valores, dependente da consulta e das respectivas chaves. A Figura 4 exibe dois tipos de função de atenção.

Figura 4 – Atenção de produto escalar escalonado (esquerda); atenção multi-cabeças com camadas paralelas de atenção (direita).

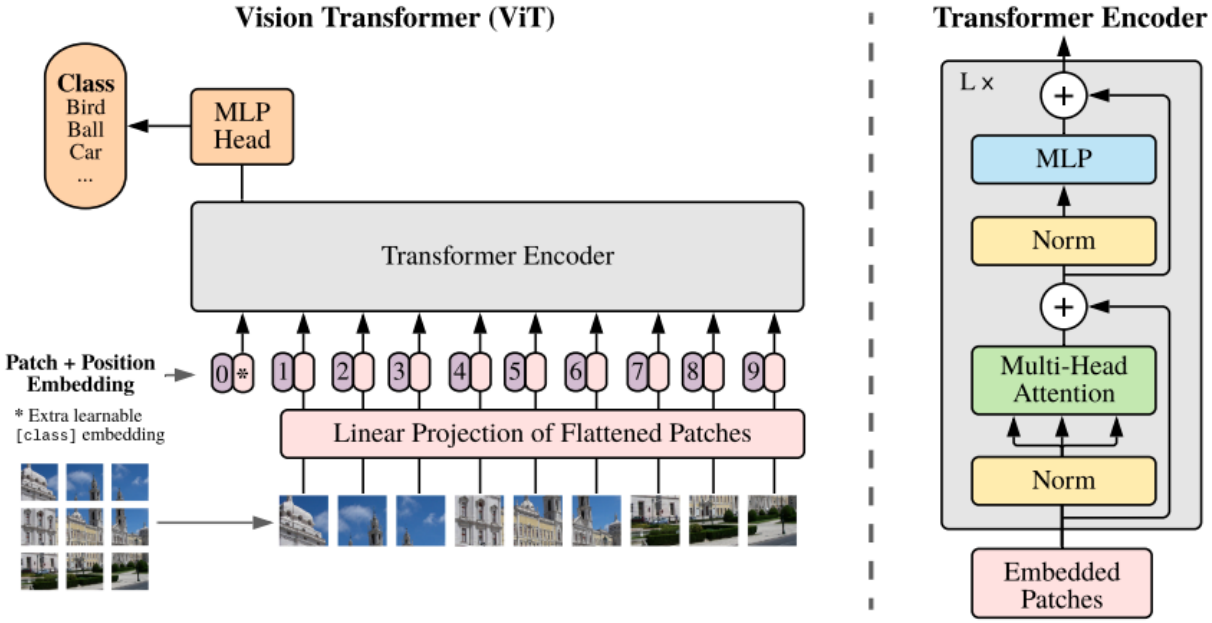


Fonte: (VASWANI *et al.*, 2017).

Em sistemas de visão computacional, *MAs* são úteis por permitirem a captura de relações entre **campos receptivos** mais distantes um do outro, quando comparados com a operação de convolução. De acordo com Shah *et al.* (2025), no que tange a segmentação semântica de imagens médicas, o fato de que os órgãos não são restritos a um campo receptivo pequeno implica que os *Transformers* são cruciais.

Apesar de as arquiteturas baseadas em *Transformers* terem sido amplamente utilizados em tarefas de processamento natural de linguagem, Ranftl *et al.* (2021) propuseram os *Vision Transformers* que, com base na Figura 5, funcionam dividindo as imagens em *pathes* de tamanho fixo para projeção linear e incorporação no bloco codificador de um *Transformer*. Isso é essencial para aplicar tais arquiteturas em tarefas de Segmentação Semântica de imagens, por exemplo.

Figura 5 – Visão geral da arquitetura *Vision Transformers*.



Fonte: (RANFTL *et al.*, 2021).

2.3 Segmentação Semântica

A segmentação semântica é uma das tarefas computacionais presentes no ramo de Visão Computacional. Acima foi apresentada uma arquitetura de CNN básica para a tarefa computacional de classificação (Figura 2), onde, dada uma imagem de entrada, obtêm-se como saída uma distribuição de probabilidades para determinados rótulos.

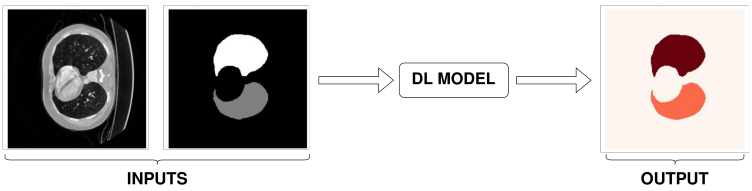
Acontece que para muitos sistemas não basta apenas atribuir rótulos às imagens, é necessário destacar regiões de interesse de forma automatizada. Para isso, a entrada passa a ter uma máscara com a verdade sobre a região de interesse e, combinada com a imagem de entrada, deve resultar em uma máscara equivalente predita pelo modelo de DL.

Essa tarefa é muito importante no contexto da engenharia biomédica, visto que frequentemente os especialistas buscam por lesões em forma de artefatos contidos nas imagens clínicas. De acordo com Zhang *et al.* (2025), o papel que a segmentação semântica desempenha nesse contexto é indispensável.

Como forma de exemplificação, a imagem abaixo mostra como a segmentação semântica pode ser aplicada a este trabalho:

Muitos dos modelos utilizados nos últimos anos para esse propósito são inspirados no paradigma codificador-decodificador (o mesmo mencionado anteriormente no contexto de *Transformers*). Tradicionalmente, baseados em *CNNs* o codificador é responsável por uma redução espacial das imagens e conversão da entrada em volumes densos de mapas de características.

Figura 6 – Processo de segmentação semântica de imagens aplicado à segmentação pulmonar.

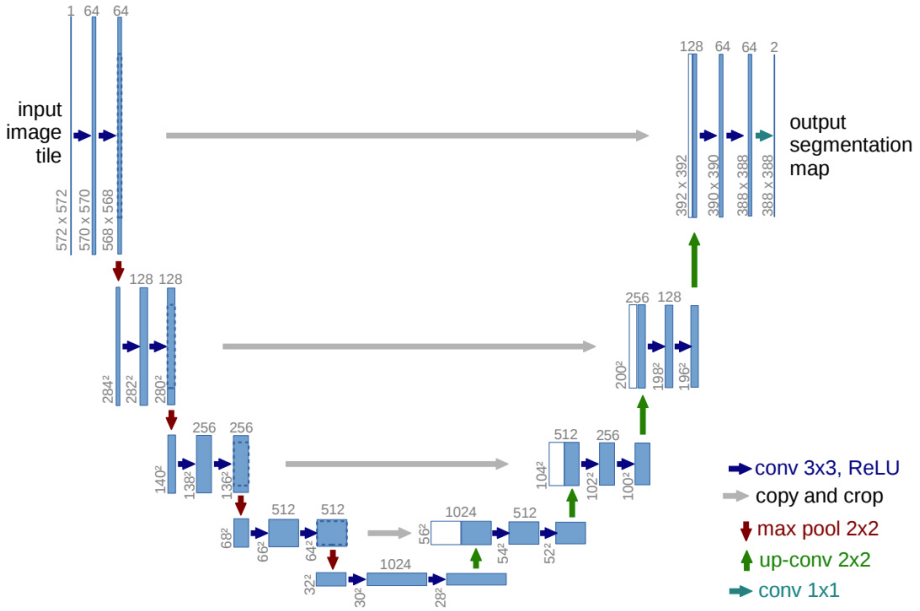


Fonte: Elaborada pelo autor.

A saída do codificador, portanto, é rica em semântica. Já o decodificador, atua na reconstrução da imagem para posterior obtenção da região de interesse predita (para isso é fundamental a operação de convolução reversa).

O trabalho de Ronneberger *et al.* (2015) representa um marco temporal considerável, por ser um dos primeiros trabalhos a utilizar *CNNs* sobre o paradigma supracitado. Tais autores apresentaram a arquitetura U-Net (Figura 7) para segmentação de imagens biomédicas. A simetria dessa arquitetura explica o porquê de seu nome e permite a visualização clara de um bloco codificador seguido do bloco decodificador. Sua simplicidade permite uma melhor compreensão de modelos mais modernos e, inclusive, os modelos utilizados neste trabalho (sendo a UNet++ uma evolução desta arquitetura base).

Figura 7 – Arquitetura U-Net, um modelo consolidado no processo de segmentação semântica.



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

2.3.1 UNet++

A UNet++ (ZHOU *et al.*, 2018) surge como uma evolução direta da UNet tradicional, partindo da observação de que a maioria das arquiteturas do tipo codificador–decodificador compartilham um elemento central: as conexões de atalho (skip connections). A proposta da UNet++ é justamente repensar essas conexões, introduzindo atalhos densos e aninhados, com camadas de convolução adicionais ao longo desses caminhos. A hipótese central do modelo é que o grande desafio da UNet está na diferença semântica entre os mapas de características do codificador (mais abstratos) e do decodificador (mais espaciais). Ao inserir convoluções intermediárias nos atalhos, a UNet++ busca aproximar semanticamente esses mapas, tornando o processo de otimização mais simples e estável durante o treinamento.

Na prática, isso resulta em uma arquitetura mais profunda e conectada, onde cada nível do decodificador recebe informações progressivamente refinadas vindas do codificador, em vez de um único salto direto como na UNet tradicional. Além disso, a UNet++ introduz o conceito de supervisão profunda, permitindo gerar saídas intermediárias em diferentes níveis da rede, o que contribui para maior estabilidade no treinamento e melhor convergência. Esse modelo foi proposto especialmente no contexto de segmentação de imagens médicas, um cenário crítico onde pequenas melhorias na qualidade da segmentação podem ter grande impacto nos resultados clínicos.

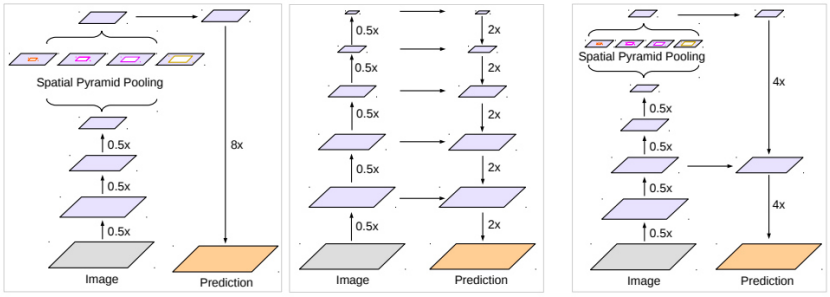
2.3.2 DeepLab-v3+

Essa arquitetura apresentada por Chen *et al.* (2018) surgiu a partir da DeepLab (CHEN *et al.*, 2017), uma rede neural profunda utilizada também para segmentação semântica, onde os autores defendem o uso de “agrupamento espacial piramidal” e Convoluções Atrous. Esse tipo de convolução é utilizada para aumento do campo receptivo sem necessariamente um aumento no kernel. Isso é uma alternativa as “camadas deconvolucionais”, dado que estas requerem mais rempo e memória, conforme Chen *et al.* (2017).

Já o “agrupamento espacial piramidal” serve principalmente para lidar com variações na escala dos dados, o que pode beneficiar tanto o treinamento de CNNs quanto a inferência. Ao final dessa operação, a saída são ricas informações semânticas densas. No entanto, Chen *et al.* (2018) afirmam que mesmo com esse beneficio, as informações detalhadas sobre os objetos de interesse são perdidas.

Para solucionar isso, nessa arquitetura os autores adicionaram um módulo decodificador simples no modelo arquitetural anterior, o que permite recuperar informações sobre os limites da região de interesse. Em síntese, a Deeplabv3 é aproveitada como um poderoso módulo codificador e combinada com o decodificador comentado anteriormente. O processo é exibido na Figura 8

Figura 8 – Etapas para construção da Deeplabv3+. Esquerda: fluxo da operação de "agrupamento espacial piramidal" empregada na Deeplabv3. Centro: estrutura com codificador-decodificador posteriormente combinada com o "agrupamento espacial piramidal". Direita: Arquitetura final.



Fonte: Elaborada pelo autor.

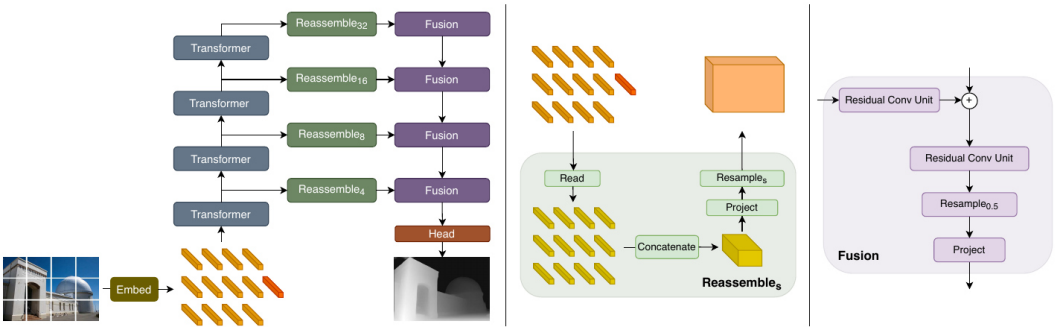
2.3.3 DPT

Ranftl *et al.* (2021) introduziram uma nova arquitetura baseada em *Vision Transformers* para predição densa (Figura 9), ou seja: tarefas de visão computacional que exigem a geração de previsões de granulação fina (a nível de pixel) e de resolução total, como a segmentação semântica. Os autores explicam que os tokens mantém correspondência de 1 para 1 com os *patches* de entrada, o que auxilia na persistência da resolução espacial da incorporação inicial. Esse ponto é crítico para a arquitetura proposta, porque os autores defendem que a perda de resolução espacial ao longo das camadas de uma CNN fazem com que a DPT tenha previsões mais detalhadas e globalmente coerentes.

O decodificador nessa arquitetura atua realizando fusões progressivas das predições densas finais. Com a operação de remontagem proposta no trabalho (Equação 2.2, onde t representa tokens de camadas arbitrárias), torna-se possível lidar adequadamente com os tokens de entrada do modelo e também gerar dados intermediários equivalentes aos mapas de características. Após isso, são realizadas as fusões supracitadas.

$$Reassemble_S^{\hat{D}}(t) = (Resample_s \circ Concatenate \circ Read)(t) \tag{2.2}$$

Figura 9 – Resumo gráfico da arquitetura DPT.



Fonte: (RANFTL *et al.*, 2021).

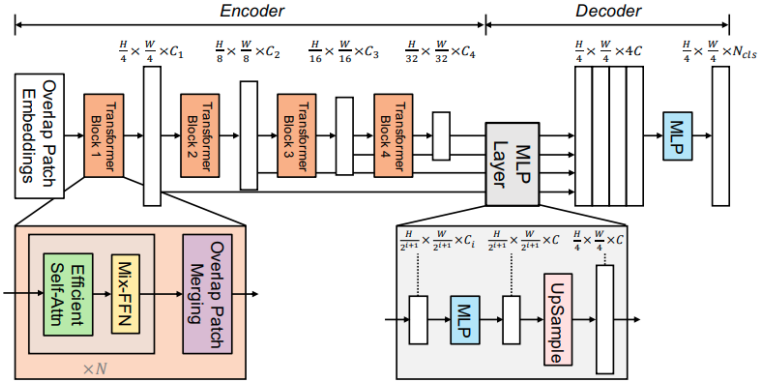
2.3.4 Segformer

Essa arquitetura, proposta por Xie *et al.* (2021), também parte dos *Vision Transformers* (RANFTL *et al.*, 2021) e foca principalmente em simplicidade, eficiência e performance. A grande mudança desse trabalho foi introduzir um decodificador completamente baseado em MLP's (redes neurais básicas e leves) para ganho de performance. Os autores partem da premissa que, como o codificador baseado em *Transformer* possui um grande campo receptivo, é possível dispensar decodificadores pesados.

Um aspecto interessante desse trabalho é que a passagem dos dados pelo codificador segue uma estrutura hierárquica semelhante a evolução dos mapas de características ao longo das camadas de uma CNN. Para isso, os autores projetaram uma série de "*Mix Transformers*" (arquiterura baseada em *Vision Transformers* com mesma arquitetura, mas tamanhos diferentes.

A Figura 10 ilustra bem o codificador e decodificador da arquitetura *SegFormer*:

Figura 10 – Arquitetura SegFormer.



Fonte: (XIE *et al.*, 2021).

2.4 Pré-processamento

Gonzalez e Woods (2018) apresentam passos fundamentais para o processamento de imagens. Dentre esses passos, há o realce de imagens, que é o processo de adequação de imagens originais de entrada para uma aplicação específica. Além disso, também há o processo de restauração de imagens, cujo objetivo reside na melhora visual das imagens.

Em síntese, é fundamental que sistemas de visão computacional tenham um bloco de pré-processamento para tornar os dados de entrada mais adequados para o objetivo final. Diversas técnicas podem ser empregadas nesse sentido. Abaixo, algumas dessas técnicas, aplicadas neste trabalho, são discutidas.

2.4.1 Operação de Janelamento (Windowing)

Ao contrário das imagens tradicionais de 8 bits, onde os valores de intensidade variam de 0 a 255, as TCs possuem um range maior de intensidades onde é adotada a escala Hounsfield, no intuito de destacar diferentes substâncias a partir das intensidades.

A operação de janelamento consiste no ajuste de contraste de imagens onde é definido um valor de intensidade máximo e mínimo, com a apoio da escala Hounsfield. Seja f(x,y) uma imagem qualquer, l o menor valor de intensidade desejado para a "janela" e u o maior valor de intensidade desejado, essa operação pode ser descrita pela equação 2.4. Nesse trabalho, isso será útil para destacar a região pulmonar e os limiares serão definidos conforme literatura.

$$f(x,y) = \begin{cases} u & \text{, se } f(x,y) < u \\ l & \text{, se } f(x,y) > l \\ f(x,y) & \text{, c.c.} \end{cases} \tag{2.3}$$

2.4.2 CLAHE

A Equalização de histograma adaptativa com limitação de contraste (*CLAHE*) é uma técnica avançada de Equalização de Histogramas de imagem para aprimoramento do contraste de imagens digitais. De acordo com Gonzalez e Woods (2018), a distribuição de intensidades de uma imagem (ou histograma) traz informações valiosas sobre o nível de contraste da mesma. Percebe-se que imagens com contraste reduzido têm o histograma menos uniforme.

Conforme Gonzalez e Woods (2018), a equalização de histograma geral é definida a partir da Equação (1), onde r_k é um valor de intensidade em [0, L-1] e o termo aplicado no somatório representa a probabilidade de ocorrência desse valor na imagem atual.

$$T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j) \tag{2.4}$$

Diferentemente da equalização de histograma global, a *CLAHE* realiza a equalização de forma adaptativa e local, dividindo a imagem em pequenas regiões não sobrepostas, denominadas tiles. Para cada uma dessas regiões, um histograma é calculado e equalizado individualmente, permitindo o realce de detalhes locais e a melhoria do contraste em imagens que apresentam variações significativas de iluminação.

Na implementação da OpenCV (BRADSKI, 2000), a CLAHE é controlada pelos parâmetros $clipLimit$ e $tileGridSize$. O $clipLimit$ restringe a amplificação do contraste em cada tile, reduzindo a intensificação de ruído, enquanto o $tileGridSize$ define a divisão da imagem em regiões locais para o cálculo dos histogramas.

2.4.3 Filtro da Mediana

O filtro da mediana é uma técnica de suavização no domínio espacial amplamente utilizada para a redução de ruídos impulsivos, como o ruído do tipo “sal e pimenta” (GONZALEZ; WOODS, 2018). Nesse método, o valor de cada pixel é substituído pela mediana dos valores de intensidade presentes em uma vizinhança ao seu redor, definida por um kernel de dimensões fixas, como 3×3 ou 5×5.

Por utilizar a mediana em vez da média, esse filtro apresenta maior robustez a valores extremos, preservando melhor as bordas e estruturas relevantes da imagem. A aplicação do filtro ocorre por meio do deslocamento do kernel sobre a imagem de origem, sendo que cada pixel da imagem resultante, localizado no centro do kernel, é obtido a partir da operação de mediana calculada sobre a vizinhança correspondente (processo análogo ao da operação de convolução).

3 TRABALHOS RELACIONADOS

Para uma melhor tomada de decisões acerca dos experimentos, foi realizada uma seleção de trabalhos relevantes na literatura. Para a seleção desses artigos, priorizou-se periódicos bem renomados, com fator de impacto significativo e/ou artigos atuais e relevantes, seja pelo número de citações ou pela robustez do trabalho e/ou suas contribuições para o tema.

Nesse sentido, a literatura evidencia a ampla adoção de *TC* do tórax em tarefas de segmentação e classificação de doenças pulmonares, bem como a utilização de diferentes arquiteturas baseadas em CNNs e, mais recentemente, em *Transformers*. Os trabalhos analisados exploram tanto a segmentação pulmonar quanto a segmentação de lesões como etapas fundamentais em sistemas de diagnóstico auxiliado por computador, além de empregarem diferentes estratégias de pré-processamento, conjuntos de dados, métricas e protocolos experimentais. Assim, a análise comparativa desses estudos permite contextualizar as escolhas metodológicas deste trabalho e justificar as arquiteturas avaliadas.

Gite *et al.* (2023) realizaram uma avaliação de desempenho das arquiteturas FCN, SegNet, U-Net e UNet++ aplicadas à imagens de raio-X do Tórax. A comparação foi estalecida a partir do problema de segmentação pulmonar, como nesse trabalho. Os autores defendem que os resultados da tarefa de classificação de algumas doenças pulmonares são significativamente aprimorados com uma etapa anterior de extração de regiões de interesse a partir da segmentação com redes neurais. Como forma de avaliar as arquiteturas propostas, foram utilizadas as métricas IoU médio, especificidade, e sensibilidade além da precisão.

Diversos trabalhos têm utilizado imagens de *TC* do tórax seja para tarefas computacionais de classificação ou segmentação semântica (DLAMINI *et al.*, 2023; SAID *et al.*, 2023; DUTANDE *et al.*, 2022; SHAH *et al.*, 2023; JALALI *et al.*, 2021). Said *et al.* (2023) utilizam esse tipo de imagem para construção de um sistema com duas etapas: segmentação pulmonar alicerçada na arquitetura UNETR (rede neural baseada em *Transformers* cuja simetria entre codificador e decodificador remete à tradicional U-Net) e classificação para diagnóstico automatizado de câncer de pulmão.

Cheng *et al.* (2024) compararam arquiteturas baseadas em CNNs com arquiteturas baseadas em *Transformers*, com relação a detecção de variados tipos de doenças pulmonares, a partir de imagens de raio-X. O *dataset* utilizado apresentava caixas delimitadoras das lesões características, criadas por um radiologista experimente. Para avaliar a qualidade da deteção a métria IoU foi empregada. Acurácia, precisão, sensibilidade e F1-score foram utilizadas na

avaliação do bloco de classificação. Após coleta de resultados os autores concluíram que os modelos baseados em CNN superaram modelos baseados em *Transformers*, mas afirmam que a proporção do conjunto de dados tem impacto direto sobre isso.

Jalali *et al.* (2021) propuseram a ResBCDU-Net com base em CNNs para realizar segmentação pulmonar, com o objetivo principal de beneficiar futuros trabalhos de detecção de doenças pulmonares a partir de TC's. Nesse trabalho a tese de que um processo de segmentação anterior à detecção é importante foi apresentada mais uma vez, sendo que os autores consideram essa etapa como "inseparável" no processo de diagnóstico automatizado. Para construção do modelo proposto, os autores se alicerçam em arquiteturas como a U-Net, ResNet34 e FCN. Uma série de pré-processamentos foram aplicados até obterem um coeficiente DiCE de 97,15% e F1-score de 98,52%, como limiariação através da operação de janelamento, erosão, fechamento e preenchimento de máscaras preditas.

Khomduean *et al.* (2023) utilizaram as arquiteturas 3D-UNet, DenseNet e ResNet para segmentação automática de lesões pulmonares relacionadas à COVID-19, utilizando o Escore Total de Gravidade (*Total Severity Score* - TSS) (CHUNG *et al.*, 2020) e o coeficiente Dice como métricas de avaliação. Os autores afirmam que os métodos do aprendizado profundo auxiliam especialistas e aumentam a precisão da segmentação de lesões pulmonares. Após a conversão de dos dados proprietários para JPEG, foi aplicado um *resize* para ajuste de resolução para 256x256 e posteriormente CLAHE para aprimoramento de contraste. Destaca-se que os autores treinaram 2 modelos principais, o primeiro foi responsável pela segmentação de lóbulos pulmonares (segmentação multiclasse) e o segundo foi responsável por uma segmentação binária de lesões. Isso ocorreu porque, conforme os mesmos, imagens sem regiões extrapulmonares são preferíveis para o treinamento do modelo de lesões.

Hu *et al.* (2020) apresentaram uma abordagem para segmentação pulmonar baseada na rede Mask R-CNN e obtiveram uma acurácia de $97,68 \pm 3,42\%$ com tempo médio de execução de 11,2s. Os autores destacam a importância de TC's no fluxo de trabalho de especialistas para diagnóstico de doenças pulmonares, alegando que esse formato de imagem aumenta a precisão de sistemas de visão computacional. Além disso, é discutido que especialistas devem primeiramente delimitar regiões pulmonares para posterior diagnóstico clínico. A rede Mask R-CNN foi combinada com métodos tradicionais de aprendizado de máquina no intuito de segmentar de forma automática a região pulmonar, onde a combinação que mais se destacou foi Mask R-CNN com SVM.

Park *et al.* (2023) utilizaram Tomografia por Emissão de Pósitrons (PET TC) para treinamento de uma arquitetura U-Net de dois estágios, com intuito de aprimorar o desempenho da segmentação de câncer de pulmão. Os resultados indicaram que o método proposto superou a arquitetura UNet 3D convencional. Os autores destacam que a rede UNet é uma das mais utilizadas no contexto de segmentação semântica de imagens médica. O coeficiente Dice foi utilizado como métrica e para concluir que ele era significativamente diferente, com relação ao método proposto e a UNet tradicional, foi empregado o teste *t de Student*.

Com o objetivo de sintetizar e comparar de forma estruturada os principais trabalhos relacionados, a Tabela 1 apresenta uma visão consolidada dos estudos analisados, destacando o tipo de imagem utilizado, os objetivos, as arquiteturas empregadas, os conjuntos de dados, as métricas de avaliação, os protocolos experimentais e as técnicas de pré-processamento adotadas. Essa organização permite identificar convergências e divergências metodológicas na literatura, bem como posicionar o presente trabalho em relação aos estudos existentes.

Percebe-se que frequentemente tais trabalhos utilizam a divisão de dados simples conhecida como *Hold-out*, devido a isso nesse trabalho uma abordagem de validação cruzada baseada em KFold é explorada. Além disso, os testes estatísticos podem fortalecer conclusões e foi observado pouca frequência nos trabalhos em questão. Objetivou-se utilizar tais ferramentas estatísticas para uma melhor discussão acerca dos resultados quantitativos. Por fim, no presente trabalho são exploradas tanto arquiteturas baseadas em *CNNs* quanto arquiteturas baseadas em *Transformers*.

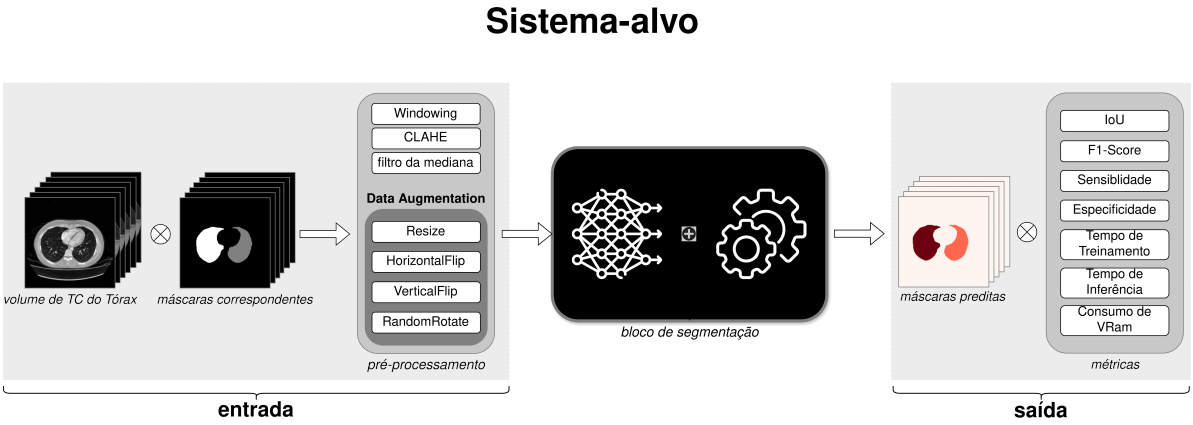
Tabela 1 – Conjunto de trabalhos relacionados.

Referência	Tipo de Imagem	Objetivo	Modelos Utilizados	Dataset	Métricas	Testes estatísticos?	Validação externa?	Divisão dos dados	Pré-processamento
(GITE *et al.*, 2023)	raio-X	Segmentação pulmonar	U-Net++.; FCN; SegNet; UNet.	Montgomery County X-ray set; Shenzhen Hospital X-ray set	Accuracy; IoU; Precision; Specificity; Dice; Sensitivity; Recall.	✗	✗	Hold-out	Conversão para PNGs; Resizing.
(SAID *et al.*, 2023)	TC	Segmentação; Classificação.	UNETR	TASK06 (Decathlon)	DICE; Sensitivity; Specificity; Accuracy.	✗	✗	Hold-out	-
(JALALI *et al.*, 2021)	TC	Segemtação pulmonar	Propõe Res BCDU-Net; UNet como base; (ResNet-34 como codificador)	LIDC-IDRI	Accuracy; Precision; Recall; F1-score; Dice.	✗	✗	Hold-out	Limiarização com HU; Erosão; Fechamento; Detecção de bordas.
(CHENG *et al.*, 2024)	raio-X	Detecção de lesões com ou sem classificação binária.	YOLOX; Dynamic R-CNN; RetinaNet; TFA; FSCE.	Dataset privado de: EDA-Hospital	Accuracy; IoU; Precision; Recall; F1-score.	✗	✗	Hold-out duplo; Divisão para o método de detecção com poucos exemplos.	Transformação logarítmica da intensidade dos pixels nas imagens DICOM; Normalização com média e σ do imagenet.
(KHOMDUEAN *et al.*, 2023)	TC	Segmentação pulmonar; Segmentação de lesões.	3D-UNet; Encoders: DenseNet e ResNet	Dataset privado de: Hospital Chulabhorn	DICE; PI; TSS; Hausdorff distance.	✓	✗	Hold-out	Resizing; CLAHE.
(DLAMINI *et al.*, 2023)	TC	Detecção de lesões; Segmentação de lesões.	YOLOv4; Kmeans.	Dataset do TCIA	IoU; Precision; Sensitivity; DICE; F1-score.	✗	✗	Hold-out	Filtro da média para redução de ruído; Filtro Gaussiano.
(SHAH *et al.*, 2023)	TC	Classificação	CNNs 2D prórias	LUNA16	Accuracy; Precision Recall.	✗	✗	Hold-out	Resizing; Conversão para JPEG;.
(DUTANDE *et al.*, 2022)	TC	Segmentação de lesões	Propoem DRS-CNN 1 e 2; FCN; SegNet; Unet.	TASK06 (Decathlon); StructSeg 2019.	Dice; Hausdorff distance; Sensibility; Precision.	✗	✓	Hold-out	Windowing (-1024, 2400); Geração de patches.
(HU *et al.*, 2020)	TC	Segmentação pulmonar	Mask R-CNN;	Dataset privado do hospital universitário Walter Cantídio	Position adjustment; Dice; Accuracy; Sensitivity; Specificity.	✗	✗	Hold-out	nenhum
(PARK *et al.*, 2023)	PET/TC	Segmentação de lesões	Arquitetura UNet de 2 estágios	Dados privados	Dice	✗	✗	Hold-out	-

4 METODOLOGIA PARA A AVALIAÇÃO DE DESEMPENHO

Como forma de realizar uma avaliação de desempenho das CNN's de forma sistemática, é crucial haver uma boa definição do sistema (com entradas, blocos de processamento e saídas) além de entendermos os principais parâmetros (características essenciais da carga de trabalho ou do sistema) e fatores do mesmo (parâmetros que serão variados para uma investigação do impacto na saída final) (JAIN, 1991). Essa seção discorre sobre isso, além de apresentar decisões fundamentais de design de experimentos e métricas de avaliação. Um fluxograma do sistema avaliado é apresentado na Figura 11.

Figura 11 – Sistema proposto.



Fonte: Elaborada pelo autor.

Como entrada para o sistema, tem-se os volumes de *TC* de cada paciente, bem como suas respectivas máscaras rotuladas por profissionais especialistas. Em seguida, na camada de pré-processamento, algoritmos como janelamento, CLAHE, filtro da mediana, redimensionamento, HorizontalFlip, VerticalFlip e RandomRotate são aplicados, visando trazer mais variabilidade para o conjunto de treinamento. Em seguida, tem-se o bloco segmentador, o qual consiste de diversas redes objetivando a segmentação semântica do pulmão, bem como de sua etapa de treinamento. Após esta etapa, os modelos treinados entram em modo de predição, de forma a automaticamente segmentar os pulmões, dada um exame de entrada. Por fim, métricas de avaliação como IoU, F1-Score, Sensibilidade, Especificidade, Tempo de treinamento/Inferência e consumo de VRAM são utilizadas para a avaliação do desempenho de cada modelo.

4.1 Conjunto de Dados de Entrada

A entrada para o sistema em questão serão as *TC* do tórax do conjunto de dados LOCCA (RIBEIRO *et al.*, 2025) convertidas em PNG. Com o intuito de extrair o pulmão para maximizar os resultados de trabalhos futuros na segmentação de lesões (dado que isso implica remover informações irrelevantes, como os tecidos de osso e rim, por exemplo (WONG *et al.*, 2022)), esse trabalho propõe inicialmente uma segmentação semântica do pulmão e as CNN's serão avaliadas com relação a isso.

O conjunto de dados LOCCA (RIBEIRO *et al.*, 2025) consiste em 30 volumes de *TC* de pacientes com COVID-19 do Hospital de Clínicas da Unicamp (HCU) e 30 volumes de *TC* do TASK06, um dos conjuntos de dados públicos disponibilizados pelo desafio *Medical Segmentation Decathlon* (ANTONELLI *et al.*, 2022). Os 60 volumes resultantes contém anotações associadas para cada lóbulo pulmonar feitas manualmente por especialistas. Cada lóbulo pulmonar recebeu um rótulo numérico: 0 para o fundo, 1 para LUL, 2 para LLL, 3 para RUL, 4 para RML e 5 para RLL (RIBEIRO *et al.*, 2025). Os rótulos 1 e 2 formam o pulmão esquerdo e os rótulos 3, 4 e 5 formam o pulmão direito.

Em relação aos volumes do HCU, os dados que inicialmente estavam no formato DICOM foram anonimizados e convertidos para NIFTI, com o intuito de facilitar o compartilhamento dos mesmos. Além disso, os autores afirmam terem truncado os valores de intensidade das imagens para a faixa de Unidades Housfield [-1024, 600]. Isso facilita a visualização da região pulmonar e de fissuras, como explicado no tópico 2.4.1.

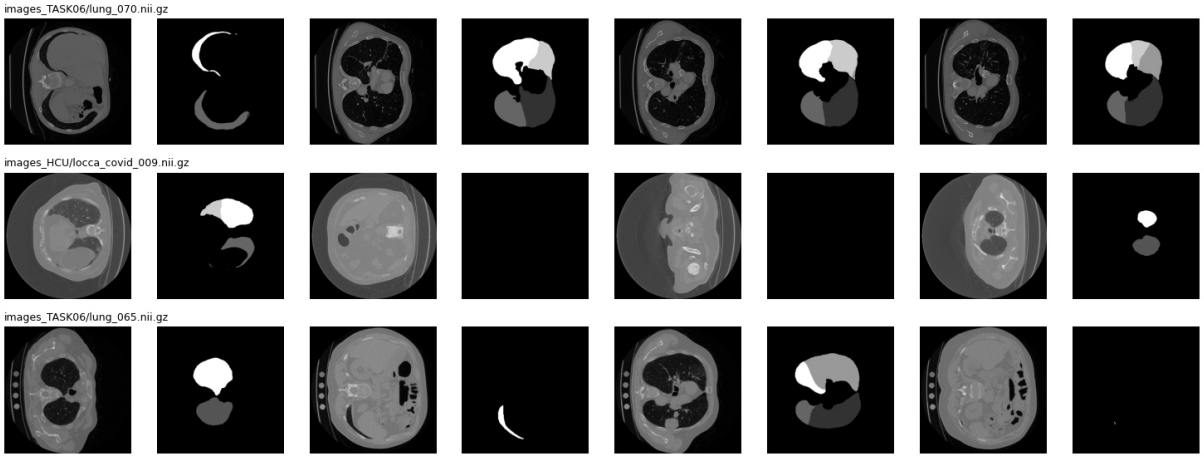
A Figura 12 exibe uma amostra de alguns volumes do conjunto de dados LOCCA, em forma de matriz de *slices*. Cada linha na matriz representa um volume escolhido aleatoriamente. As colunas ímpares possuem *slices* aleatórias do volume em questão, onde as máscaras correspondentes são exibidas nas colunas pares, logo em seguida.

Vale ressaltar que para simplificação e menor custo computacional atrelado a leitura dos arquivos NIFTI's, todas as slices foram convertidas para o formato de dados PNG após o pré-processamento.

4.2 Técnicas de Pré-Processamento

Após a análise exploratória de dados foi possível observar uma diferença expressiva quanto aos valores de intensidade de volumes dos subconjuntos TASK06 e HCU, como ilustrado

Figura 12 – Amostra do conjunto de dados LOCCA.



Fonte: Elaborada pelo autor.

nas linhas 1 e 2 da Figura 12. Por essa razão, e com o intuito de destacar o pulmão, a operação de janelamento (2.4.1) foi aplicada. Foram definidos como limiares para essa operação -1000 e 650, levando em conta os trabalhos presentes na literatura.

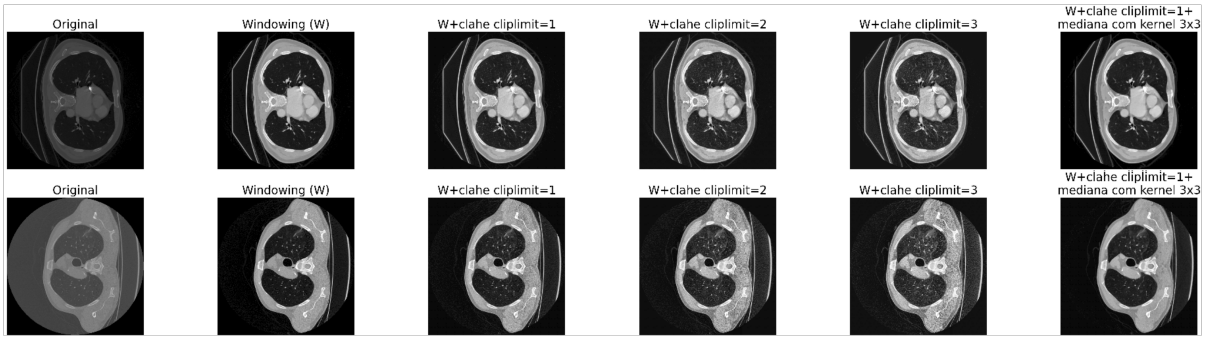
Além disso, objetivou-se aprimorar o contraste para realçar a região pulmonar, com relação ao restante da imagem, pois a região pulmonar tende a níveis mais baixos de intensidade enquanto os componentes anatômicos em sua volta (como costela, veia braquiocefálica, esôfago, etc) tendem a níveis mais altos de intensidade, como ilustrado na Figura 12.

Assim, aplicou-se CLAHE (2.4.2) com os parâmetros $clipLimit = 1$ e $titleGridSize$ de 8x8. Os valores desses parâmetros foram definidos experimentalmente, selecionando volumes aleatórios, aplicando a operação de janelamento e calculando o resultado para $clipLimit = 1$, $clipLimit = 2$ e $clipLimit = 3$.

Tais experimentos iniciais exibiram resultados em que a região extrapulmonar indicou a presença de ruído do tipo “sal e pimenta”, portanto foi decidido aplicar o filtro da mediana com *kernel* 3x3 após CLAHE com $clipLimit = 1$, visto que acima de 1 o método intensificava além do admissível os valores de intensidades das estruturas internas, o que poderia ir contra o objetivo inicial ao aplicar o CLAHE. A Figura 13 exibe algumas imagens resultantes do processo descrito.

Após essa etapa foi possível definir o bloco de pré-processamento como a sequência: operação de janelamento, CLAHE ($clipLimit = 1$ e $titleGridSize$ 8x8) e filtro da mediana com *kernel* 3x3. Um conjunto maior de imagens resultantes com alta resolução pode ser acessado em: https://drive.google.com/file/d/1TUw6iHjnJ2E3QRENFzZSqjaSHTvN8LTm/view?usp=sharing.

Figura 13 – Imagens originais e pré-processadas a partir de cada um dos algoritmos: Windowing, CLAHE e mediana, bem como a combinação de ambos.



Fonte: Elaborada pelo autor.

4.3 Aumento de Dados

Nesse trabalho prezou-se apenas por transformações geométricas, devido os dados de entrada serem pré-processados na presente análise. A Tabela 2 exibe as transformações escolhidas.

Tabela 2 – Transformações utilizadas para aumento de dados.

Transformação	**Probabilidade**
Resize $\rightarrow (256, 256)$	100%
HorizontalFlip	50%
VerticalFlip	50%
RandomRotate (90°)	50%
ToTensor	100%

Fonte: o autor.

Esse conjunto de transformações é necessário para permitir uma melhor generalização das predições dos modelos para as diferentes *TC* do conjunto de teste. A rotação de 90° combinada com inversões faz com que determinado pulmão mude de localização e assim há uma maior generalização no conjunto de treino, visto que em outros conjuntos de dados não necessariamente o pulmão direito, por exemplo, se encontra na mesma localização que aquela do conjunto LOCCA.

Destaca-se que para o conjunto de teste, em cada iteração, apenas *Resize* e *ToTensor* foram aplicadas. A ideia é tornar o conjunto de testes sempre inalterável para realizar comparações justas.

4.4 Parâmetros e Fatores

Com base na abordagem de avaliação de desempenho proposta por Jain (1991), ao longo do Desing dos Experimentos foram investigados quais seriam os principais parâmetros do sistema (ou características do sistema e de sua entrada). Observou-se que para um sistema de visão computacional há uma grande quantidade de parâmetros que podem afetar as métricas de interesse, mas o custo computacional atrelado a inclusão de vários parâmetros torna isso a abordagem inviável.

Estima-se que os experimentos realizados tiveram duração total de 30h, o que pode ser facilmente confirmado com a interface da biblioteca TensorBoard (como será explicado na seção 4.5). A adição de um simples fator com 2 níveis poderia elevar o tempo total dos experimentos para 60h.

A partir dessa observação inicial observou-se que uma boa delimitação de escopo seria crucial para obter resultados satisfatórios de avaliação no invervalo de tempo disponível para esse trabalho. Devido a este motivo, parâmetros como função de perda ou otimizador não foram definidos como **fatores** (parâmetros que possuem seus valores associados alterados ao longo dos experimentos). A Tabela 3 exibe uma lista de **parâmetros** elencados.

Tabela 3 – Principais parâmetros do sistema.

Parâmetro	**Nível Fixado**
SO	Ubuntu 24.04.3 LTS
CPU	Intel i7-13700 (13ª geração)
GPU	NVIDIA GeForce RTX 3060 Lite
RAM	32 GB
vRAM	12 GB
Python	3.13.7
Encoder	MobileNetV2
Nº de épocas	5
Batch size	8
Resolução	256×256

Fonte: o autor.

4.5 Treinamento e Teste

Optou-se por utilizar como base para o trabalho em questão a biblioteca Python *Segmentation Models Pytorch* (IAKUBOVSKII, 2019), essa padronização possibilitou que a forma de implementação dos modelos tivesse influência mínima nos resultados, além de acelerar os experimentos.

Todos os modelos foram avaliados considerando a função de perda *Dice* (adaptada para o problema multiclasse) e otimizador Adam, com taxa de aprendizado de 0,0002. Com relação a divisão de dados para treinamento, optou-se por aplicar o KFold com $k = 5$, isso permite obter conclusões mais sólidas, umas vez que uma mesma métrica é computada para 5 diferentes configurações de repartição dos dados. Usar hold-out poderia levar a conclusões dependentes da forma como os dados foram divididos.

Ressalta-se que a divisão em *folds* foi realizada a nível de volumes e não de *slices*, como forma de impedir que um mesmo volume contasse com *slices* no conjunto de treino e no de teste ao longo dos experimentos. Isso impede vazamento de informações do conjunto de teste para o conjunto de treinamento.

Outrossim, com o intuito de reduzir os efeitos de variação aleatória das métricas em questão optou-se por 3 execuções do KFold. As repetições implementadas favorecem o cálculo de estatísticas descritivas (como a média e desvio padrão), além de permitir conclusões mais sólidas sobre os desempenhos dos modelos de aprendizado profundo escolhidos (visto que a forma como os pesos evoluem está sujeita a aleatorização).

Como forma de organizar resultados experimentais para análise posterior, durante o treino e teste as métricas foram registradas com a biblioteca Python TensorBoard. Os principais resultados são acessíveis em https://tccmetrics.lesc-dev.fun e também nos arquivos de log da TensorBoard. Recomenda-se o uso da expressão regular "**.***" (quantidade variável de qualquer caractere) para buscas. Por exemplo: "unet.*repet1.*" lista todos os resultados da repetição 1 para a rede UNet++. Houveram 15 medições para cada modelo (k=5 e 3 repetições), o que resultou em 60 arquivos de *checkpoint*.

O consumo de VRAM também foi visto como um fato crucial para análise. Durante o treinamento foram realizados logs da saída do utilitário "nvidia-smi" a cada 30s. Após cruzamento com os horários armazenados em eventos da TensorBoard, observou-se que o consumo de um modelo tendia a um valor com pouca variação ao longo do tempo.

4.6 Métricas de avaliação

Nesta avaliação de desempenho, foram utilizadas métricas baseadas na consolidada matriz de confusão, considerando os termos *True Positives* (TP), *False Positives* (FP), *False Negatives* (FN) e *True Negatives* (TN). Essas métricas permitem quantificar tanto a qualidade da sobreposição espacial entre as regiões segmentadas quanto a capacidade do modelo em identificar

corretamente as classes de interesse. Neste trabalho, adotou-se a *Intersection over Union* (IoU), amplamente empregada em tarefas de segmentação de imagens, bem como as métricas auxiliares Sensibilidade (Recall), Especificidade e F1-score, que fornecem uma análise complementar do desempenho do modelo sob diferentes perspectivas.

No contexto da segmentação pulmonar multiclasse, essas métricas são aplicadas de forma independente para cada classe, considerando-se, em cada caso, os termos da matriz de confusão obtidos a partir da comparação pixel a pixel entre a predição do modelo e a segmentação de referência.

Após isso, é possível agregar os resultados de cada classe para análise com alguma estratégia de redução. Optou-se por uma estratégia de redução ponderada, onde o fundo teve peso 2 e ambos os pulmões tiveram peso igual a 4 (tais valores foram definidos empiricamente e pensados de tal foram que a soma dos pesos fosse 10). A justificativa para isso é que o fundo representa a maior parte de uma *slice* de *TC*, então é importante que predições certas de pulmão tenham mais interferência positiva no cálculo de métricas do que predições certas de fundo.

4.6.1 Intersecção sobre União (IoU)

IoU (Equação 4.1) mede o grau de sobreposição entre a região prevista pelo modelo e a região de referência (ground truth). Essa métrica penaliza simultaneamente falsos positivos e falsos negativos, sendo amplamente utilizada na avaliação de tarefas de segmentação por fornecer uma medida direta da qualidade espacial da predição.

$$IoU = \frac{TP}{TP + FP + FN} \tag{4.1}$$

4.6.2 Sensibilidade

A Sensibilidade (Equação 4.2) avalia a capacidade do modelo em identificar corretamente os pixels pertencentes à classe de interesse. Valores elevados indicam que a maior parte dos verdadeiros positivos foi corretamente detectada, sendo especialmente relevante em aplicações onde a omissão de regiões importantes é crítica.

$$Sensibilidade = \frac{VP}{FP + FN} \tag{4.2}$$

4.6.3 Especificidade

A Especificidade (Equação 4.3) quantifica a capacidade do modelo de identificar corretamente os pixels que não pertencem à classe de interesse. Essa métrica reflete o quão bem o modelo evita falsos alarmes, ou seja, classificar incorretamente regiões negativas como positivas.

$$Especificidade = \frac{VN}{VN + FP} \tag{4.3}$$

4.6.4 F1-Score

Por fim, F1-score (Equação 4.4) é a média harmônica entre precisão e sensibilidade, fornecendo uma medida balanceada do desempenho do modelo. Ele é particularmente útil em cenários com desbalanceamento entre classes, pois considera simultaneamente erros de omissão e de comissão.

$$F1 - score = \frac{2TP}{2TP + FP + FN} \tag{4.4}$$

4.7 Teste de Friedman

Ao longo dos experimentos houveram indícios de proximidade entre os resultados de cada modelo para as diferentes métricas principais. Como forma de averiguar a existência de uma diferença significativa entre os grupos em questão (as quatro arquiteturas avaliadas), foi aplicado o *Teste de Friedman* (FRIEDMAN, 1937). O objetivo por trás disso foi fortalecer futuras conclusões acerca dos resultados, já analisados com estatísticas descritivas e BoxPlots.

O *Teste de Friedman* é um tipo de teste de hipótese não paramétrico comumente utilizado para analisar variações entre grupos de dados que não atendem às suposições de normalidade, homogeneidade de variâncias e que são dependentes.

As hipóteses nula (H_0) e alternativa (H_A) para o *teste de Friedman* são:

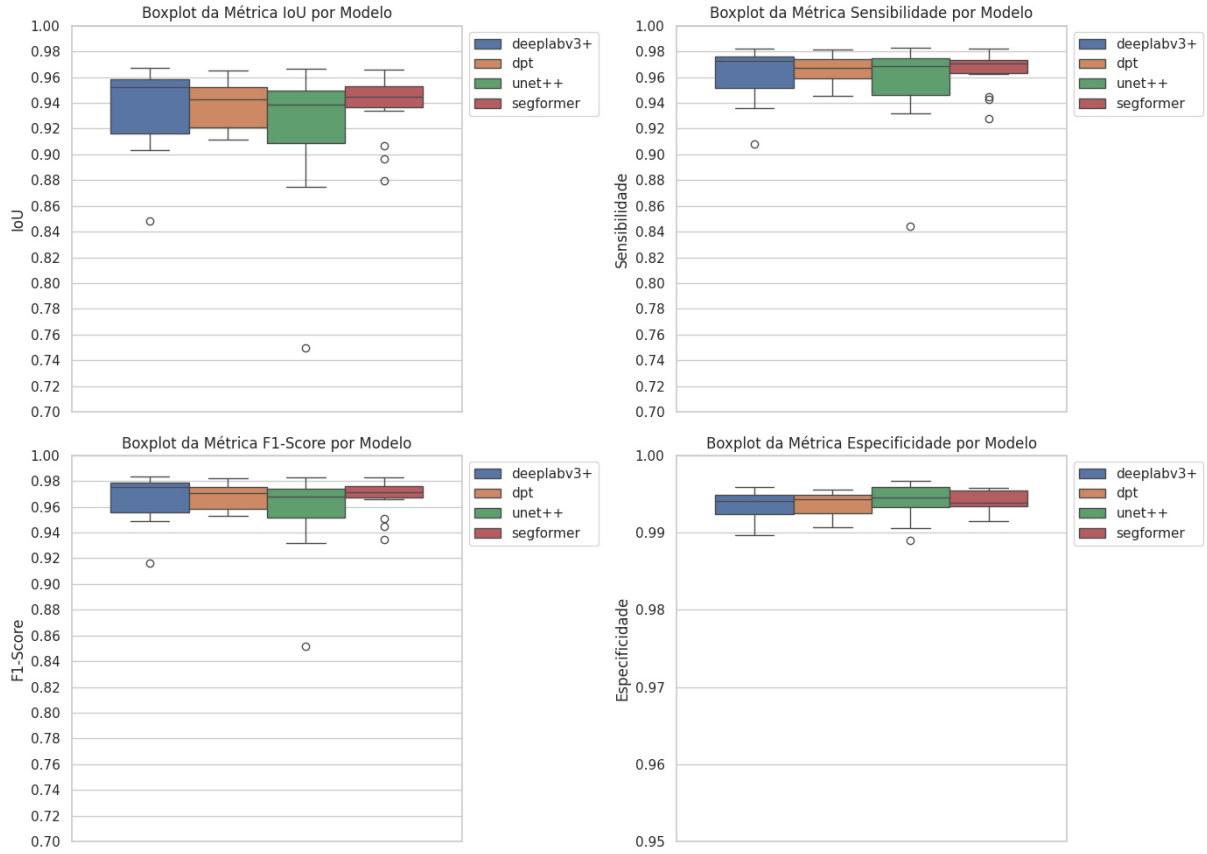
- H_0: Não há diferença estatisticamente significativa entre os grupos dependentes comparados.
- H_A: Há uma diferença significativa entre os grupos dependentes comparados.

Dessa forma, adotando um nível de significância de 5%, um valor p maior que 5% implica aceitação de H_0.

5 RESULTADOS E DISCUSSÕES

A partir da análise dos boxplots presentes na Figura 14, observa-se de forma descritiva que os modelos avaliados apresentam desempenhos globais bastante semelhantes, com medianas próximas e intervalos interquartis sobrepostos na maioria das métricas. Esse comportamento indica que, do ponto de vista médio dessas métricas, há uma tendência de equivalência entre os modelos.

Figura 14 – BoxPlots das principais métricas: IoU, Sensibilidade, F1-score e Especificidade, respectivamente.



Fonte: Elaborada pelo autor.

Nota-se ainda que a especificidade na Figura 14 apresenta valores elevados e pouca variabilidade entre os modelos, assim como esperado, visto que a identificação correta dos verdadeiros negativos tende a ser facilitada pelo predomínio de regiões de fundo nas imagens de *TC*.

Entretanto, a análise da variabilidade dos resultados revela diferenças relevantes entre os modelos. A presença de outliers e a dispersão observada em métricas como IoU e F1-score indicam que, embora o desempenho médio seja alto, alguns modelos apresentam maior

instabilidade em determinados volumes, o que pode ser crítico em aplicações clínicas.

Para verificar se as diferenças observadas entre os modelos são estatisticamente relevantes, foi aplicado o teste não paramétrico de Friedman às principais métricas de avaliação, conforme apresentado na seção 4.7. A Tabela 4 apresenta a estatística do teste e os respectivos p-valores para cada métrica, permitindo analisar a existência de diferenças significativas no desempenho dos modelos.

Tabela 4 – Resultados do teste de Friedman para cada métrica principal.

	IoU	**Sensibilidade**	**Especificidade**	**F1**	**DiceLoss**
Estatística	2.3600	0.76	2.1200	2.3600	0.8400
P-valor	0.5011	0.8590	0.5479	0.5011	0.8399

Fonte: o autor.

Todos os p-valores superaram o nível de significância 5%, ou seja, não há indícios de rejeitar a hipótese nula. Isso sugere que, apesar das variações observadas nos valores médios e na dispersão das métricas, o desempenho global dos modelos pode ser considerado estatisticamente equivalente sob as condições analisadas.

Ao analisar também os tempos de treinamento e inferência observa-se menos equivalência entre os modelos, principalmente pelo fato de a rede DPT ter tido maior custo de tempo, tanto no treino quanto no teste, como apresentado na Figura 15. Ainda assim, observa-se que a SegFormer (uma arquitetura baseada em *Transformers*, como a DPT) se destacou no tempo de treinamento, com desempenho relativo competitivo mesmo utilizando um codificador transformer. Isso pode ser explicado pela escolha arquitetural de decodificador baseado em MLPs (2.3.4).

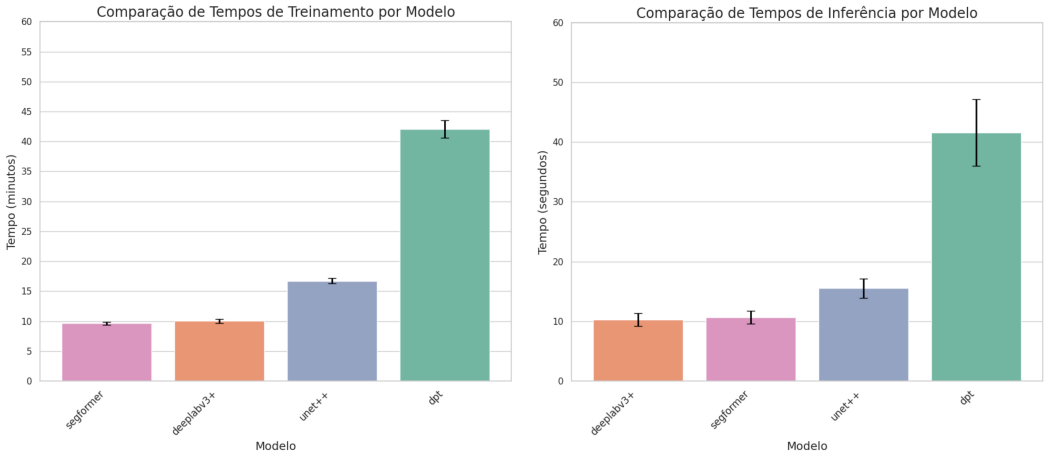
Percebe-se que com auxílio das métricas secundárias vRAM, tempo de treino e tempo de teste (ligadas mais diretamente ao ambiente de computação), foi possível que a DeepLabv3+ tivesse maior destaque.

Tabela 5 – Amostras de consumo de vRAM.

Modelo	**Amostra 1 (GB)**	**Amostra 2 (GB)**	**Amostra 3 (GB)**	**Média (GB)**
unet++	2.30	2.29	2.27	2.29
deeplabv3+	1.68	1.66	1.66	1.67
segformer	1.82	1.82	1.82	1.82
dpt	3.88	4.03	4.09	4.00

Fonte: o autor.

Figura 15 – Resultados relacionados ao tempo. Tempos de treinamento e de inferência, respectivamente.

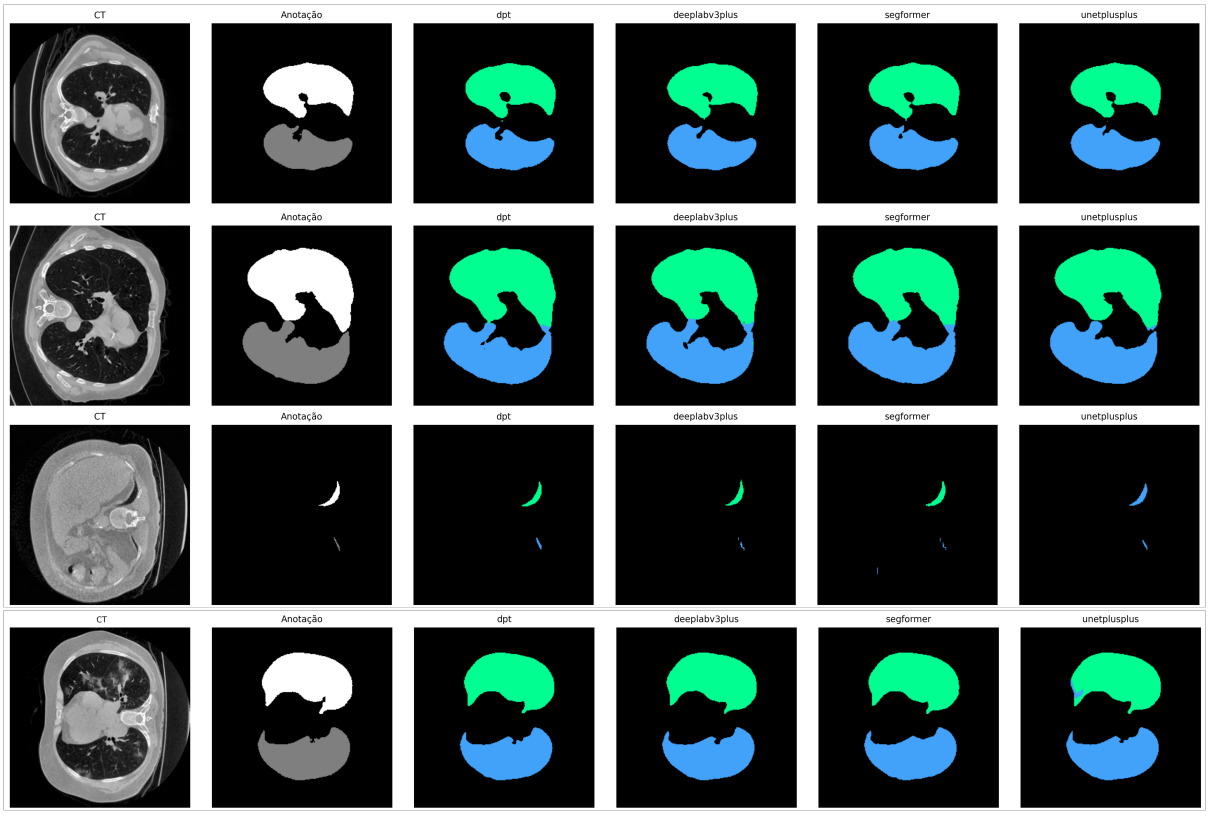


Fonte: Elaborada pelo autor.

Os resultados sugerem um melhor equilíbrio entre eficiência computacional e desempenho para a arquitetura DeepLabv3+, tornando essa arquitetura particularmente atrativa para aplicações práticas que demandam bom desempenho aliado a menor custo de recursos.

A Figura 16 traz exemplos de predições das arquiteturas:

Figura 16 – Conjunto de predições de cada arquitetura avaliada, bem como a máscara de anotação correspondente.



Fonte: Elaborada pelo autor.

Apesar de subjetivos, os resultados apresentados na Figura 16 trazem uma visão geral das predições de cada arquitetura para 4 *slices* escolhidas aleatoriamente. Ambas as arquiteturas tendem a diferenciar corretamente o pulmão direito do pulmão esquerdo além de conseguirem mapear a região pulmonar de interesse. Isso ilustra a tendência de equivalência, do ponto de vista de métricas principais, observada de forma quantitativa anteriormente.

A linha 2 da Figura 16 mostra que em alguns casos as predições tendem a se sobrepor, mas nesse cenário a UNet++ se destaca por uma separação mais adequada. No entanto, a linha 3 revela uma falha da UNet++ para um caso em que a região pulmonar é bem pequena (os dois artefatos da predição são considerados pulmão esquerdo - azul).

Por fim, um ponto que deve ser trazido a esta análise é o fato de que a utilização da arquitetura leve MobileNetV2 como codificador em cada bloco experimental permitiu que o conjunto de todos os 60 arquivos de *checkpoint* dos modelos, juntos, somassem 8,3Gb. Isso é algo importante a ser observado, uma vez que modelos leves são mais adequados para um ambiente computacional com restrições de memória e necessidade de maior performance.

A Tabela 6 encerra a discussão com o tamanho aproximado de arquivos de checkpoint de cada modelo. Aqui novamente a rede baseada em *Vision Transformers* SegFormer tem um destaque com relação a eficiência, o que vai de encontro a proposta inicial de seus criadores (como explicado na seção 2.3.4). Destaca-se também a rede baseada em CNNs DeepLabv3+ em 2º lugar neste ranking, arquitetura que havia tido certo destaque anteriormente.

Tabela 6 – Tamanho aproximado dos arquivos de checkpoint por modelo.

	UNet++	**DeepLabv3+**	**SegFormer**	**DPT**
Checkpoint (Mb)	82,4	53,0	35,5	385,0

Fonte: o autor.

6 CONCLUSÕES E TRABALHOS FUTUROS

Com base nos resultados expostos, foi possível concluir que embora as arquiteturas avaliadas tenham tido desempenho equivalente, com relação à métricas de avaliação comuns da literatura (o que foi confirmado com o teste de Friedman), ao considerar métricas relacionadas ao ambiente de computação é possível estabelecer preferência sobre uma arquitetura.

Por ter tido destaque não apenas na atividade de segmentação pulmonar, mas também em métricas atreladas ao custo computacional, a arquitetura DeepLabv3+ destaca-se para o problema em questão.

Os resultados apresentados reforçam que redes baseadas em *Transformers* estão se tornando cada vez mais performáticas para tarefas de visão computacional. Tal evidência é relevante pois durante muitos anos CNNs eram consideradas o estado da arte para a grande maioria de problemas da área de segmentação de imagens.

A análise do custo computacional evidenciou que o tempo de processamento e o consumo de memória gráfica variam significativamente entre as arquiteturas, o que pode impactar diretamente a viabilidade de uso desses modelos em ambientes clínicos reais ou em sistemas com recursos limitados.

Trabalhos futuros poderiam utilizar os resultados desse trabalho de segmentação pulmonar como ponto de partida para a desafiadora atividade de segmentação de lesões. Além disso, seria importante avaliar a influência de outros parâmetros na avaliação de desempenho proposta, como função de perda, otimizador e número de épocas.

REFERÊNCIAS

ANTONELLI, M.; REINKE, A.; BAKAS, S.; FARAHANI, K.; KOPP-SCHNEIDER, A.; LANDMAN, B. A.; LITJENS, G.; MENZE, B.; RONNEBERGER, O.; SUMMERS, R. M. *et al.* The medical segmentation decathlon. **Nature communications**, Nature Publishing Group UK London, v. 13, n. 1, p. 4128, 2022.

ARCHANA, R.; JEEVARAJ, P. S. E. Deep learning models for digital image processing: a review. **Artificial Intelligence Review**, v. 57, n. 1, p. 11, jan. 2024. ISSN 0269-2821, 1573-7462. Disponível em: <https://link.springer.com/10.1007/s10462-023-10631-z>.

BRADSKI, G. The OpenCV Library. **Dr. Dobb's Journal of Software Tools**, 2000.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. **DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs**. 2017. Disponível em: <https://arxiv.org/abs/1606.00915>.

CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. **Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation**. 2018. Disponível em: <https://arxiv.org/abs/1802.02611>.

CHENG, Y.-C.; HUNG, Y.-C.; HUANG, G.-H.; CHEN, T.-B.; LU, N.-H.; LIU, K.-Y.; LIN, K.-H. Deep learning-based object detection strategies for disease detection and localization in chest x-ray images. **Diagnostics**, v. 14, n. 23, p. 2636, 2024.

CHUNG, M.; BERNHEIM, A.; MEI, X.; ZHANG, N.; HUANG, M.; ZENG, X.; CUI, J.; XU, W.; YANG, Y.; FAYAD, Z. A. *et al.* Ct imaging features of 2019 novel coronavirus (2019-ncov). **Radiology**, Radiological Society of North America, v. 295, n. 1, p. 202–207, 2020.

DLAMINI, S.; CHEN, Y.-H.; KUO, C.-F. J. Complete fully automatic detection, segmentation and 3d reconstruction of tumor volume for non-small cell lung cancer using yolov4 and region-based active contour model. **Expert Systems with Applications**, Elsevier, v. 212, p. 118661, 2023.

DUTANDE, P.; BAID, U.; TALBAR, S. Deep residual separable convolutional neural network for lung tumor segmentation. **Computers in biology and medicine**, Elsevier, v. 141, p. 105161, 2022.

Forum of International Respiratory Societies. **The Global Impact of Respiratory Disease: Third Edition**. Lausanne, Switzerland: European Respiratory Society, 2021. Acessado: 2025-12-20. Disponível em: <https://firsnet.org/wp-content/uploads/2025/01/FIRS_Master_09202021.pdf>.

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.

GITE, S.; MISHRA, A.; KOTECHA, K. Enhanced lung image segmentation using deep learning. **Neural Computing and Applications**, Springer, v. 35, n. 31, p. 22839–22853, 2023.

GONZALEZ, R. C.; WOODS, R. E. **Digital Image Processing**. 4th global edition. ed. Harlow, England: Pearson Education Limited, 2018. ISBN 978-1-292-22304-9.

GUPTA, J.; PATHAK, S.; KUMAR, G. Deep learning (cnn) and transfer learning: a review. In: IOP PUBLISHING. **Journal of Physics: Conference Series**. [S.l.], 2022. v. 2273, n. 1, p. 012029.

HU, Q.; SOUZA, L. F. d. F.; HOLANDA, G. B.; ALVES, S. S.; SILVA, F. H. d. S.; HAN, T.; FILHO, P. P. R. An effective approach for ct lung segmentation using mask region-based convolutional neural networks. **Artificial intelligence in medicine**, Elsevier, v. 103, p. 101792, 2020.

IAKUBOVSKII, P. **Segmentation Models Pytorch**. [S.l.]: GitHub, 2019. <https://github.com/qubvel/segmentation_models.pytorch>.

JAIN, R. **The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling**. New York: John Wiley & Sons, 1991.

JALALI, Y.; FATEH, M.; REZVANI, M.; ABOLGHASEMI, V.; ANISI, M. H. Resbcdu-net: a deep learning framework for lung ct image segmentation. **Sensors**, MDPI, v. 21, n. 1, p. 268, 2021.

JAVED, R.; ABBAS, T.; KHAN, A. H.; DAUD, A.; BUKHARI, A.; ALHARBEY, R. Deep learning for lungs cancer detection: a review. **Artificial Intelligence Review**, Springer, v. 57, n. 8, p. 197, 2024.

KHOMDUEAN, P.; PHUAUDOMCHAROEN, P.; BOONCHU, T.; TAETRAGOOL, U.; CHAMCHOY, K.; WIMOLSIRI, N.; JARRUSROJWUTTIKUL, T.; CHUAJAK, A.; TECHAVIPOO, U.; TWEEATSANI, N. Segmentation of lung lobes and lesions in chest ct for the classification of covid-19 severity. **Scientific Reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 20899, 2023.

KOUTOULAKIS, E.; TRIVIZAKIS, E.; MARKODIMITRAKIS, E.; AGELAKI, S.; TSIKNAKIS, M.; MARIAS, K. A critical review of explainable deep learning in lung cancer diagnosis. v. 59, n. 1, p. 28. ISSN 1573-7462. Disponível em: <https://link.springer.com/10.1007/s10462-025-11445-x>.

KUMAR, S.; KUMAR, H.; KUMAR, G.; SINGH, S. P.; BIJALWAN, A.; DIWAKAR, M. A methodical exploration of imaging modalities from dataset to detection through machine learning paradigms in prominent lung disease diagnosis: a review. **BMC Medical Imaging**, Springer, v. 24, n. 1, p. 30, 2024.

MARGERIE-MELLON, C. D.; CHASSAGNON, G. Artificial intelligence: A critical review of applications for lung nodule and lung cancer. **Diagnostic and Interventional Imaging**, v. 104, n. 1, p. 11–17, jan. 2023. ISSN 22115684. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S221156842200225X>.

O'SHEA, K.; NASH, R. **An Introduction to Convolutional Neural Networks**. 2015. Disponível em: <https://arxiv.org/abs/1511.08458>.

PARK, J.; KANG, S. K.; HWANG, D.; CHOI, H.; HA, S.; SEO, J. M.; EO, J. S.; LEE, J. S. Automatic lung cancer segmentation in [18f] fdg pet/ct using a two-stage deep learning approach. **Nuclear medicine and molecular imaging**, Springer, v. 57, n. 2, p. 86–93, 2023.

RANFTL, R.; BOCHKOVSKIY, A.; KOLTUN, V. Vision transformers for dense prediction. **ArXiv preprint**, 2021.

RIBEIRO, J. A.; CARMO, D. S. D.; REIS, F.; MAGALHAES, R. S.; DERTKIGIL, S. S.; APPENZELLER, S.; RITTNER, L. Descriptor: Manually annotated ct dataset of lung lobes in covid-19 and cancer patients (locca). **IEEE Data Descriptions**, IEEE, 2025.

RONNEBERGER, O.; FISCHER, P.; BROX, T. **U-Net: Convolutional Networks for Biomedical Image Segmentation**. 2015. Disponível em: <https://arxiv.org/abs/1505.04597>.

SAID, Y.; ALSHEIKHY, A. A.; SHAWLY, T.; LAHZA, H. Medical images segmentation for lung cancer diagnosis based on deep learning architectures. **Diagnostics**, MDPI, v. 13, n. 3, p. 546, 2023.

SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; CHEN, L.-C. **MobileNetV2: Inverted Residuals and Linear Bottlenecks**. 2019. Disponível em: <https://arxiv.org/abs/1801.04381>.

SHAH, A. A.; MALIK, H. A. M.; MUHAMMAD, A.; ALOURANI, A.; BUTT, Z. A. Deep learning ensemble 2d cnn approach towards the detection of lung cancer. **Scientific reports**, Nature Publishing Group UK London, v. 13, n. 1, p. 2987, 2023.

SHAH, O. I.; RIZVI, D. R.; MIR, A. N. Transformer-based innovations in medical image segmentation: A mini review. **SN Computer Science**, Springer, v. 6, n. 4, p. 375, 2025.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

WONG, P. K.; YAN, T.; WANG, H.; CHAN, I. N.; WANG, J.; LI, Y.; REN, H.; WONG, C. H. Automatic detection of multiple types of pneumonia: Open dataset and a multi-scale attention network. **Biomedical Signal Processing and Control**, v. 73, p. 103415, 2022. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809421010120>.

World Health Organization. **The top 10 causes of death**. 2024. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>. Acessado: 2025-12-20.

XIE, E.; WANG, W.; YU, Z.; ANANDKUMAR, A.; ALVAREZ, J. M.; LUO, P. **SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers**. 2021. Disponível em: <https://arxiv.org/abs/2105.15203>.

ZHANG, L.; JINDAL, B.; ALAA, A.; WEINREB, R.; WILSON, D.; SEGAL, E.; ZOU, J.; XIE, P. Generative ai enables medical image segmentation in ultra low-data regimes. **Nature Communications**, Nature Publishing Group UK London, v. 16, n. 1, p. 6486, 2025.

ZHOU, Z.; SIDDIQUEE, M. M. R.; TAJBAKHSH, N.; LIANG, J. **UNet++: A Nested U-Net Architecture for Medical Image Segmentation**. 2018. Disponível em: <https://arxiv.org/abs/1807.10165>.