Trabalho de Conclusão de Curso

Análise Comparativa entre Detection Transformer e YOLOv8 para Detecção Precoce de Nódulos Pulmonares

Victor Mafra de Holanda Ferraz

orientado por
Prof. Dr. Marcelo Costa Oliveira

Universidade Federal de Alagoas
Instituto de Computação
Maceió, Alagoas
25 de março de 2024

UNIVERSIDADE FEDERAL DE ALAGOAS
Instituto de Computação

ANÁLISE COMPARATIVA ENTRE DETECTION TRANSFORMER E YOLOV8 PARA DETECÇÃO PRECOCE DE NÓDULOS PULMONARES

Trabalho de Conclusão de Curso apresentado ao Instituto de Computação da Universidade Federal de Alagoas como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação.

Victor Mafra de Holanda Ferraz

Orientador: Prof. Dr. Marcelo Costa Oliveira

Banca Examinadora:

Thales Miranda de Almeida Vieira	Prof. Dr., IC-UFAL
Lucas Lins de Lima	Prof. Me., UNIP - Ribeirão Preto

Maceió, Alagoas
25 de março de 2024

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Helena Cristina Pimentel do Vale – CRB4 –661

F381a Ferraz, Victor Mafra de Holanda.
Análise comparativa entre detection transformer e yolov8 para detecção precoce de nódulos pulmonares / Victor Mafra de Holanda Ferraz. - 2024.
39 f : il.

Orientador: Marcelo Costa Oliveira.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2024.

Bibliografia: f. 33-39.

1. Detecção de objetos. 2.. Transformers. 3. Redes neurais. 4. Nódulo pulmonares - Diagnóstico clínico. I. Título.

CDU: 681.5:616-07

Agradecimentos

À minha família, em especial à minha mãe, Marta, meu pai, Monteiro, e minha irmã, Isadora, que sempre estiveram ao meu lado e me deram apoio, não importando o quão difícil foi a caminhada até aqui;

À minha namorada, Nathália, que entrou na minha vida em um dos momentos mais complicados que já vivi e me deu forças e apoio para terminar esta estapa;

A todos os meus amigos, em especial ao Erick e a Priscila, que estão comigo nesta vida a quase 20 anos e nunca sairam do meu lado, não importando o que acontecesse;

Ao meu colega de curso Yuri Dimitri, que desde o quarto ou quinto período é minha dupla em todas as disciplínas, sendo parte fundamental do meu êxito em muitas delas;

A todos os membros do Grupo do DotA da minha cidade natal, Delmiro Gouveia, que estão diariamente comigo, sendo uma companhia muito importante que se tornaram grandes amigos, muito além de um jogo online;

A todos os professores que me ajudaram ao longo do percurso da faculdade, que me deram a base do conhecimento acadêmico e profissional e que acreditaram no meu potencial;

Aos professores Thales Vieira e Tiago Vieira, que foram meus orientadores nos ciclos de PIBIC, me dando oportunidades e apresentando uma área de atuação que hoje sou apaixonado;

Ao professor Marcelo, por todo o apoio que me deu durante a pesquisa e no desenvolvimento deste TCC;

Aos professores Thales Vieira e Lucas Lins, pelo interesse e disponibilidade em participar da banca examinadora;

A todos os funcionários do IC.

Maceió, 25 de Março de 2024.

Resumo

O câncer de pulmão (CP) é o segundo tipo mais prevalente de câncer em todo o mundo e o mais mortal, sendo responsável por uma em cada cinco mortes relacionadas ao câncer globalmente. As chances de sobrevivência para pacientes diagnosticados com esse tipo de câncer aumentam consideravelmente quando o diagnóstico é feito precocemente, com a taxa de sobrevivência de 5 anos chegando a até 70%. Radiologistas realizam o diagnóstico de CP por meio de imagens de Tomografia Computadorizada (TC), mas esse diagnóstico é uma tarefa complexa e sujeita a erros. Por meio de ferramentas auxiliadas por computador, esse processo de diagnóstico pode ser automatizado com o intuito de auxiliar o profissional, reduzindo o tempo e o esforço para os especialistas, além de melhorar a confiança no diagnóstico. O objetivo deste trabalho foi avaliar e comparar a eficácia das arquiteturas de Redes Neurais Convolucionais (CNN) e *Transformer* na detecção de pequenos nódulos pulmonares ($\leq$15mm), onde a questão orientadora da pesquisa deste trabalho foi "Qual é o impacto do tamanho dos nódulos pulmonares na precisão de detecção das arquiteturas CNN e *Transformer*?". O conjunto de dados utilizado foi baseado no banco de dados público LUNA16, filtrando o conjunto de testes para incluir apenas cortes de TC com nódulos com até 15mm. Os modelos escolhidos para nossas comparações foram o YOLOv8, uma CNN considerada estado-da-arte em detecção de objetos, e o DEtection TRansformer (DETR), que combina a arquitetura de transformer com uma camada CNN, onde obtivemos resultados como mAP50 = 0,70, sensibilidade = 0,91 e λ = 0,85 para o DETR e mAP50 = 0,90, sensibilidade = 0,83 e Λ = 0,77 para o YOLOv8. Também avaliamos o impacto do tamanho do nódulo no desempenho de ambos os modelos, onde o desempenho do YOLOv8 foi impactado pela diminuição do tamanho dos nódulos, enquanto o DETR continuou a mostrar resultados satisfatórios independentemente de quão pequenos os nódulos fossem.

***Palavras-chave*: Detecção de Objetos, Transformers, Redes Neurais, Nódulo Pulmonares, Diagnóstico Médico**

Abstract

Lung cancer (LC) is the second most prevalent type of cancer worldwide and the deadliest, accounting for one in every five cancer-related deaths globally. The chances of survival for patients detected with this type of cancer increase considerably when the diagnosis is made early, with the 5-year survival rate reaching up to 70%. Radiologists perform LC diagnosis through Computed Tomography (CT) images, but such diagnosis is a complex and error-prone task. Through computer-aided tools, this diagnostic process can be automated in order to assist the professional, reducing time and effort for specialists, as well as improving confidence in the diagnosis. The objective of this work was to evaluate and compare the effectiveness of Convolutional Neural Network (CNN) and Transformer architectures in detecting small lung nodules ($\leq$15mm), where the guiding research question of this work was "What is the impact of the size of lung nodules on the detection accuracy of CNN and Transformer architectures?". The dataset used was based on the public database LUNA16, filtering the test set to include only sections with nodules with up to 15mm. The models chosen for our comparisons were YOLOv8, a CNN considered state-of-the-art in object detection, and DEtection TRansformer (DETR), which combines the transformer architecture with a CNN layer, where we obtained results such as mAP50 = 0.70, Sensitivity = 0.91 and Λ = 0.85 for the DETR and mAP50 = 0.90, Sensitivity = 0.83 and Λ = 0.77 for the YOLOv8. We also assessed the impact of nodule size on the performance of both models, where the performance of YOLOv8 was impacted by the decrease in nodules size, while DETR continued to show satisfactory results regardless of how small the nodules were.

***Keywords*: Object Detection, Transformers, Neural Networks, Pulmonary Nodules, Medical Diagnostic**

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

$p_{interp(r)}$	Ponto p interpolado no recall r
Λ	Métrica de avaliação da curva FROC aumentada
d	Diâmetro do nódulo

Lista de Abreviações

CP Câncer de Pulmão

TC Tomografia Computadorizada

CAD do inglês *Computer-Aided Diagnosis*

DL do inglês *Deep Learning*

CNN do inglês *Convolutional Neural Network*

YOLO do inglês *You Only Look Once*

PLN Processamento de Linguagem Natural

ViT do inglês *Vision Transformer*

ANN do inglês *Artificial Neural Network*

BB do inglês *Bounding Box*

DETR do inglês *DEtection TRansformer*

FFN do inglês *Feed-forward Network*

IoU do inglês *Intersection over Union*

TP do inglês *True Positive*

FP do inglês *False Positive*

FN do inglês *False Negative*

FROC do inglês *Free-Response Receiver Operating Characteristic*

AUC do inglês *Area Under the Curve*

FPR do inglês *False Positive Rate*

GT do inglês *Ground Truth*

Sumário

Capítulo 1

Introdução

As estatísticas globais do câncer indicam que o câncer de pulmão (CP) é o segundo tipo de câncer mais prevalente em todo o mundo, com 2,2 milhões de novos casos e o mais letal, com 1,8 milhão de mortes relatadas no ano de 2020, representando aproximadamente um em cada 10 (11,4%) diagnósticos de câncer e um em cada 5 (18%) mortes por câncer em todo o mundo. Nos homens, o CP é a principal causa de morbidade e mortalidade por câncer, enquanto nas mulheres, ocupa o terceiro lugar em incidência, atrás do câncer de mama e cólon, e o segundo em mortalidade, atrás apenas do câncer de mama (OMS (2020)). Entre as principais causas do CP estão o tabagismo, a poluição do ar, a exposição a produtos químicos e influências genéticas, e aproximadamente metade dos pacientes diagnosticados com CP morrem dentro do primeiro ano após a descoberta da doença, frequentemente devido a diagnósticos em estágios avançados que afetam outras partes do corpo (Lima et al. (2019)). O diagnóstico precoce do câncer de pulmão é essencial para aumentar as chances de sobrevivência do paciente, com uma taxa de sobrevida de 5 anos alcançando 70% quando diagnosticado no estágio I (Blandin Knight et al. (2017)), onde, segundo os resultados obtidos por Wisnivesky et al. (2005), caso um nódulo menor que 15 milímetros seja cancerígeno, ele tem uma grande chance de ainda estar no estágio I da doença. No Brasil, no entanto, estima-se que 70% dos casos de CP sejam diagnosticados em estados avançados (estágio II) ou metastáticos (estágio IV), com apenas 9% diagnosticados no estágio I, um número menor em comparação com alguns países desenvolvidos (Araujo et al. (2018)).

A Tomografia Computadorizada (TC) é a principal ferramenta usada por radiologistas para detectar nódulos pulmonares, pois fornece imagens 3D de alta resolução com contraste, refletindo diferenças na intensidade, textura e forma de tumores. No entanto, o diagnóstico do câncer de pulmão usando imagens de TC apresenta vários desafios e limitações para profissionais da área. Os nódulos podem ser muito pequenos, localizados em estruturas anatômicas complexas da região pulmonar (por exemplo, vasos e pleura) e podem ter contraste semelhante ao tecido pulmonar. Os profissionais de radiologia frequentemente enfrentam outras dificuldades, como fadiga e pressão de tempo no

diagnóstico, condições externas adversas como ruído de imagem e iluminação fraca no ambiente de trabalho e um aumento contínuo no volume de imagens, o que pode levar a falhas diagnósticas (Lima et al. (2019); Patz et al. (2014); Degnan et al. (2019)).

A falha em detectar um nódulo potencialmente cancerígeno pode ter sérias consequências tanto para o paciente quanto para o radiologista e é uma das causas mais comuns de reclamações por negligência contra especialistas (Emani et al. (2019)). Estudos também mostram que os nódulos são diagnosticados corretamente por radiologistas apenas 68% das vezes, aumentando para 82% quando examinados por dois radiologistas (Nasrullah et al. (2019)). Para melhorar o diagnóstico de imagens médicas, os sistemas de Diagnóstico Auxiliado por Computador (CAD) são ferramentas importantes que podem fornecer suporte para a tomada de decisões dos radiologistas, funcionando como uma segunda opinião (Halder et al. (2020)). As ferramentas CAD podem automatizar o processo de diagnóstico, reduzindo o tempo e o esforço necessários para a análise, além de melhorar a confiabilidade e a repetibilidade da tarefa (Ferreira et al. (2018); Choi and Choi (2013)). Os sistemas CAD envolvem tanto a localização de lesões em imagens médicas (CADe) quanto a classificação de lesões como malignas ou benignas (CADx), o que traz vários desafios para o campo (Firmino et al. (2016)).

Atualmente, as técnicas de Aprendizado Profundo, do ingles *Deep Learning* (DL), são o estado da arte em aplicações de CAD para detecção de nódulos pulmonares (Halder et al. (2020)), e vários estudos (McBee et al. (2018); Adams et al. (2021); Halder et al. (2020)) demonstraram o potencial do DL na detecção de nódulos pulmonares, bem como a eficácia do uso de sistemas CAD em centros de saúde. Entre as arquiteturas de DL, as Redes Neurais Convolucionais (CNNs) surgiram como a principal arquitetura no campo de informática médica devido aos resultados excepcionais obtidos em visão computacional (Ravì et al. (2016)). Vários estudos (Ardila et al. (2019); Cui et al. (2020); Li and Fan (2020)) usando modelos CNN para detecção de nódulos mostraram excelentes resultados. No entanto, apesar dos avanços nos modelos CNN, o número de falsos positivos na detecção de nódulos pulmonares permanece um desafio para a implementação de sistemas em clínicas médicas (Liang et al. (2021); Shaukat et al. (2019)).

Entre as CNNs, o modelo *You Only Look Once* (YOLO) (Redmon et al. (2016)) é o estado da arte em detecção de objetos, sendo um algorítmo de detecção de ponta a ponta em tempo real que tem ganhado significativa atenção na comunidade de visão computacional. A aplicação do YOLO no domínio médico tem despertado interesse devido à sua capacidade de detectar e localizar estruturas anatômicas (Mortada et al. (2023); Zeng et al. (2023)), lesões (Baccouche et al. (2021); Santos et al. (2022)), tumores (Montalbo (2020); Rong et al. (2023); Safdar et al. (2020)) e outros objetos médicos clinicamente relevantes (Zhou et al. (2023)). A adoção do YOLO em aplicações médicas tem o potencial de melhorar a precisão e eficiência do diagnóstico médico, o que pode ter um impacto significativo nos resultados dos pacientes (Qureshi et al. (2023)).

O *Transformer* (Vaswani et al. (2017)) é uma arquitetura codificador-decodificador que revolucionou o campo do Processamento de Linguagem Natural (PLN) ao apresentar uma estrutura mais simples sem a necessidade de convoluções. Seu principal componente é o mecanismo de atenção, que utiliza dependências globais entre entradas e saídas. O *transformer* tornou-se o estado da arte em tarefas de tradução e desde então tem sido utilizado em diversas áreas (Jumper et al. (2021); Radford et al. (2021); Zandie and Mahoor (2023)). Nesse contexto, o *Vision Transformer* (ViT) (Dosovitskiy et al. (2020)) surgiu como uma adaptação do *Transformer* para tarefas de visão computacional, funcionando em combinação com arquiteturas CNN tradicionais ou substituindo-as completamente. Vários modelos inspirados no ViT original surgiram posteriormente e alcançaram excelentes resultados em tarefas de detecção de objetos (Carion et al. (2020)), segmentação (Wang et al. (2021)), geração de imagens sintéticas (Jiang et al. (2021)), entre outras tarefas.

Os modelos ViT têm atraído interesse significativo da comunidade médica para serem adaptados para aplicações de imagens médicas devido aos excelentes resultados obtidos em tarefas de visão computacional e tornaram-se um tema recorrente em conferências e periódicos da área (Shamshad et al. (2023)). Na área de detecção de nódulos usando ViT, Zhu et al. (2022) propuseram uma arquitetura de ponta a ponta que utiliza uma rede residual em forma de U em combinação com o mecanismo de atenção e alcançaram 95% de sensibilidade na detecção de nódulos pulmonares com um número significativamente menor de parâmetros em comparação com modelos CNN, reduzindo também o número de falsos positivos. Niu and Wang (2022) propuseram um modelo de região baseado em ViT 3D para identificar nódulos pulmonares em um conjunto de regiões candidatas. O modelo proposto obteve resultados superiores (3% de melhoria) na detecção de nódulos em comparação com modelos CNN 3D de última geração.

1.1 Objetivo geral

Apesar da ampla gama de estudos sobre o uso de CAD na detecção de nódulos pulmonares descritos anteriormente, a identificação precoce desses nódulos, especialmente aqueles com até 15 milímetros, ainda é um problema que requer investigação adicional. Nesse contexto, o principal objetivo deste trabalho foi avaliar e comparar a eficácia das arquiteturas de Rede Neural Convolucional (CNN) e *Transformer* na detecção de pequenos nódulos pulmonares ($\leq$15mm), sempre tentando responder o questionamento "Qual é o impacto do tamanho dos nódulos pulmonares na precisão de detecção das arquiteturas CNN e *Transformer*?".

1.1.1 Objetivos Específicos

1. Geração da base de imagens em 2D para uso no modelo de detecção de objetos;
2. Filtragem do conjunto de validação, onde apenas nódulos com diâmetro menor a 15mm foram mantidos;
3. Adaptação do Transformer para o uso com a base de dados de nódulos pulmonares;
4. Treinamento dos modelos YOLO e Transformer e geração dos pesos para inferência;
5. Plotagem das métricas utilizadas na validação dos modelos;
6. Avaliação e comparação dos resultados.

1.2 Organização do trabalho

Este trabalho foi organizado em capítulos que detalham os passos seguidos durante a pesquisa e utilização dos modelos, detalhando conceitos computacionais e definições matemáticas que servem como base para este modelo.

O Capítulo 2 aprofunda-se em conceitos de redes neurais, explicando o funcionamento e arquitetura dos modelos YOLOv8 e DETR, além de apresentar as métricas de avaliação de desempenho utilizadas, junto a alguns conceitos fundamentais.

O Capítulo 3 detalha a obtenção da base de dados utilizada neste trabalho, a organização do conjunto de dados de treinamento e de teste, da a motivação por trás da escolha dos modelos e por fim detalha como foi feito o treinamento dos modelos.

O Capítulo 4 traz a análise e a discussão dos principais resultados obtidos por ambos os modelos na detecção de nódulo pulmonares precoces, apresentando tabelas, gráficos e figuras que demonstram bem os resultados obtidos.

O Capítulo 5 traz os pensamentos finais em relação ao desenvolvimento e aos resultados, discutindo o panorama dos modelos em relação à detecção de nódulo pulmonares precoces, e as perspectivas futuras da aplicação.

Capítulo 2

Fundamentação Teórica

2.1 Aprendizado de máquina

O aprendizado de máquina é um campo da inteligência artificial que tem revolucionado diversas áreas, possibilitando que sistemas computacionais aprendam padrões a partir de dados e realizem tarefas sem serem explicitamente programados para tal (Mitchell (1997)).

Esse campo abrange diferentes paradigmas, destacando-se o aprendizado supervisionado, no qual os algoritmos aprendem a partir de exemplos rotulados (Bishop and Nasrabadi (2006)), o aprendizado não supervisionado, que busca identificar padrões em dados não rotulados (Hastie et al. (2009)), e o aprendizado por reforço, no qual os agentes interagem com o ambiente e aprendem com as consequências de suas ações (Sutton and Barto (2018)). Tais abordagens têm sido amplamente aplicadas em áreas como diagnósticos médicos, reconhecimento de voz e visão computacional, demonstrando a importância e a versatilidade do aprendizado de máquina na atualidade (Murphy (2012)).

Durante as últimas décadas, a área de estudo do aprendizado de máquina trouxe uma variedade enorme de avanços em algoritmos de aprendizado e técnicas de pré-processamento eficientes. Um desses avanços foi a evolução das redes neurais artificiais, do inglês *Artificial Neural Network* (ANN) em direção às arquiteturas de redes neurais profundas com capacidades de aprendizado melhoradas, resumidas no termo aprendizado profundo, do inglês *Deep Learning* (DL) (Goodfellow et al. (2016)). A figura 2.1 demonstra a hierarquia entre as áreas citadas nesta seção.

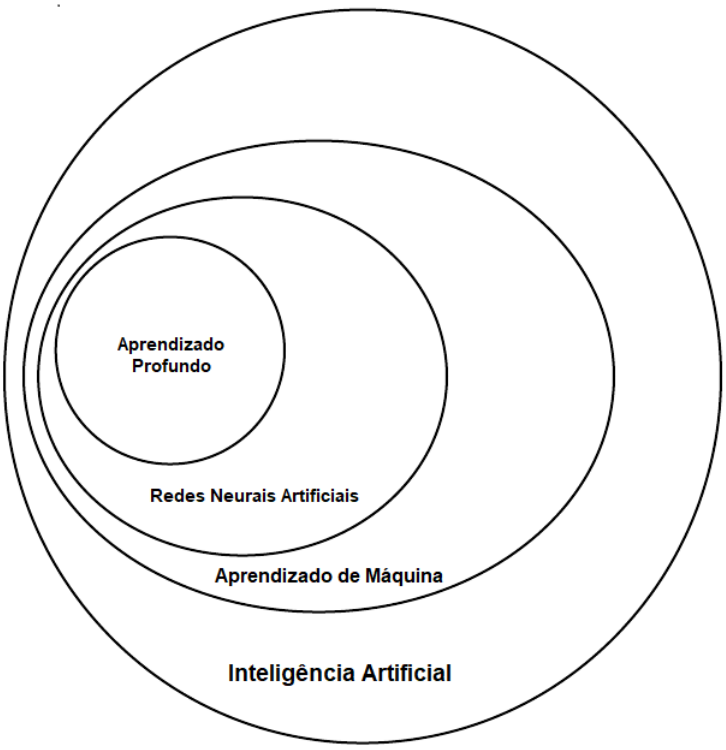


Figura 2.1: Representação em diagrama de Venn da hierarquia na área de aprendizado de máquina. Fonte: Modificado de Goodfellow et al. (2016).

2.2 Redes Neurais Convolucionais

Redes Neurais Convolucionais (CNNs) operam através de uma série de camadas especializadas que são capazes de aprender e extrair padrões visuais complexos em dados de natureza espacial, como imagens. As camadas convolucionais, o componente fundamental das CNNs, aplicam operações de convolução sobre os dados de entrada, o que permite detectar padrões locais, como bordas, texturas e formas. Essas camadas são seguidas por camadas de pooling, que reduzem a dimensionalidade dos dados, preservando as características mais importantes. Finalmente, camadas totalmente conectadas são empregadas para realizar a classificação ou regressão com base nas características extraídas anteriormente. Um dos marcos pioneiros na aplicação de CNNs foi a LeNet-5, proposta por LeCun et al. (1998), que introduziu pela primeira vez a arquitetura de CNNs e demonstrou sua eficácia na classificação de dígitos escritos à mão. Com suas capacidades de aprendizado de características complexas e robustas, as CNNs continuam a ser uma ferramenta essencial em uma variedade de aplicações, desde reconhecimento de objetos em imagens médicas até detecção de objetos em vídeos de vigilância. Um exemplo da rede criada por LeCun et al. (1998), a LeNet, pode ser visto abaixo:

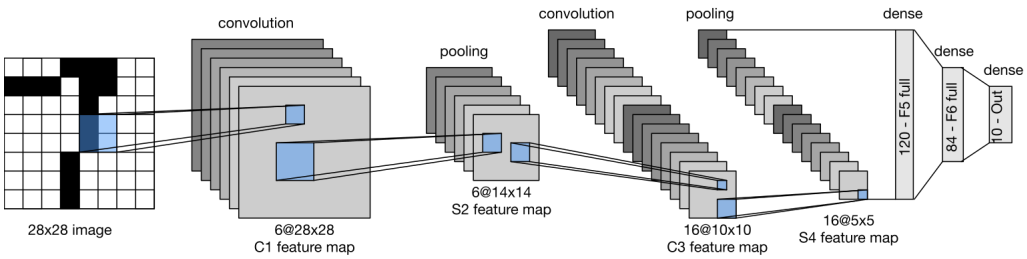


Figura 2.2: Exemplo da arquitetura LeNet. Fonte: Zhang et al. (2021).

2.3 *You Only Look Once* (YOLO)

A rede neural YOLO é uma arquitetura de detecção de objetos que se destaca por sua eficiência e precisão proposta por Redmon et al. (2016). Segundo o autor, o YOLO é projetado para realizar detecções em tempo real, permitindo que sistemas automatizados identifiquem objetos em imagens de forma rápida e eficiente. O método YOLO aborda o problema de detecção de objetos como uma única tarefa de regressão, onde uma rede neural é treinada para prever caixas delimitadoras e probabilidades de classe simultaneamente em uma única passagem pela imagem. Isso é destacado por Redmon and Farhadi (2018) ao afirmarem que o YOLO é capaz de gerar detecções precisas e rápidas, tornando-o adequado para uma variedade de aplicações em tempo real. A arquitetura YOLO tem evoluído ao longo do tempo, onde hoje se encontra na sua oitava versão (YOLOv8), obtendo uma melhora tanto em performance quanto em precisão em relação as suas versões anteriores (Jocher et al. (2023)), como pode ser visto na figura 2.3.

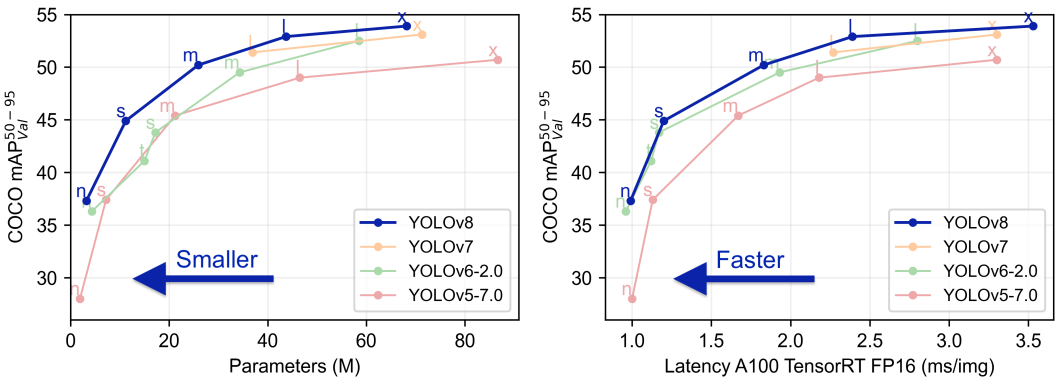


Figura 2.3: Gráficos comparando as versões do YOLO em tamanho, velocidade e precisão. Fonte: Jocher et al. (2023).

2.3.1 Arquitetura

A arquitetura da rede é inspirada no modelo GoogLeNet para classificação de imagens (Szegedy et al. (2015)). A rede é composta por 24 camadas convolucionais seguidas por 2 camadas totalmente conectadas, onde as camadas convolucionais iniciais da rede extraem características da imagem enquanto as camadas totalmente conectadas preveem as probabilidades de saída e as coordenadas. Foram utilizadas camadas de redução 1 × 1 seguidas por camadas convolucionais 3 × 3 (Redmon et al. (2016)). A arquitetura da rede pode ser observada na figura 2.4.

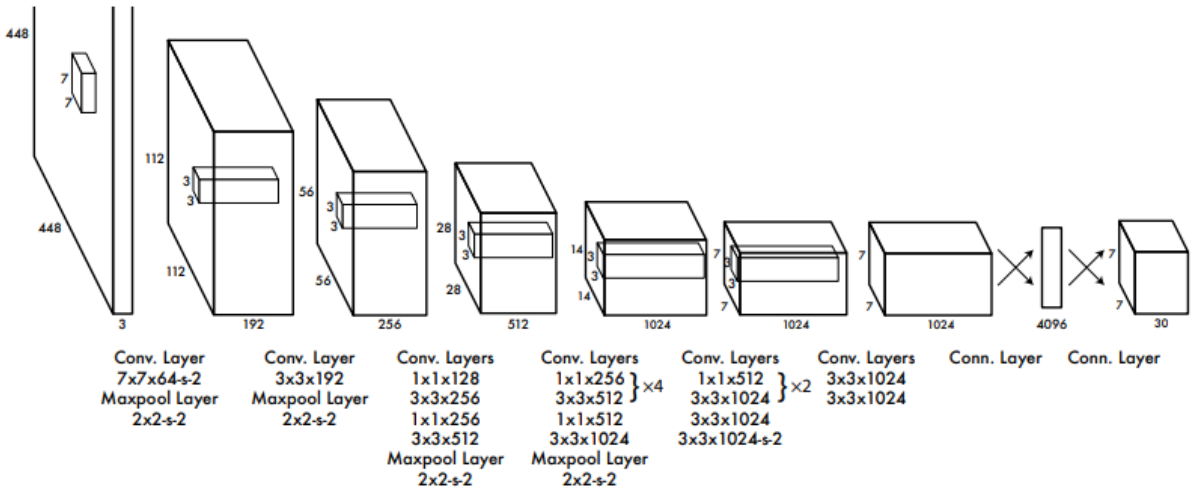


Figura 2.4: Arquitetura da primeira versão do YOLO. Constituida por 24 camadas convolucionais seguidas por 2 camadas totalmente conectadas. Camadas convolucionais alternadas de 1 × 1 reduzem o espaço de características das camadas anteriores. Fonte: Redmon et al. (2016).

2.3.2 YOLOv8

O YOLOv8 utiliza uma arquitetura de base semelhante a suas versões anteriores, com algumas alterações na camada CSP, agora chamado de módulo C2f. O módulo C2f (*bottleneck* parcial entre estágios com duas convoluções) combina características de alto nível com informações contextuais para melhorar a precisão da detecção. Nesta versão, é adotado um modelo sem âncoras com uma cabeça desacoplada para processar tarefas de objetividade, classificação e regressão de forma independente. Esse design permite que cada ramo se concentre em sua tarefa específica e melhore a precisão geral do modelo. Na camada de saída, eles utilizam a função sigmoide como função de ativação para a pontuação de objetividade, representando a probabilidade de que a caixa delimitadora, do inglês *Bounding Box* (BB) contenha um objeto. A função softmax é utilizada para as probabilidades de classe, representando as probabilidades dos objetos pertencerem a cada

classe possível (Terven et al. (2023)). O YOLOv8 utiliza as funções de perda CIoU (Zheng et al. (2020)) e DFL (Li et al. (2020)) para a perda de BB e entropia cruzada binária para a perda de classificação. Essas funções de perda têm melhorado o desempenho na detecção de objetos, especialmente ao lidar com objetos menores (Terven et al. (2023)). A figura 2.5 demonstra de maneira detalhada a arquitetura do YOLOv8. A função de perda CIoU é definida como:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{2.1}$$

onde IoU representa a intersecção sobre a união entre as caixas de detecção (BB) da predição do modelo e do *ground truth* (GT), definido em 2.10. $\mathbf{b}$ e $\mathbf{b}^{gt}$ representam o ponto central das BBs de detecção e de GT, enquanto ρ representa a distancia euclidiana entre esses pontos e c representa o comprimento diagonal da menor caixa que cobre as duas BB. α é um parâmetro positivo de compensação e v mede a consistência da proporção de aspecto, definidos como:

$$v = \frac{4}{\pi^2}(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h})^2 \tag{2.2}$$

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{2.3}$$

onde w^{gt} e h^{gt} representam a largura e altura da BB do GT e h e t representam a largura e altura da BB da detecção. Já a função DLF é definida como:

$$\mathbf{DFL}(S_i, S_{i+1}) = -((y_{i+1} - y)log(S_i) + (y - y_i)log(S_{i+1})) \tag{2.4}$$

onde S_i, S_{i+1} e y são definidos como:

$$S_i = \frac{y_{i+1} - y}{y_{i+1} - y_i} \tag{2.5}$$

$$S_{i+1} = \frac{y - y_i}{y_{i+1} - y_i} \tag{2.6}$$

$$y = \int_{-\infty}^{+\infty} \delta(x - y)xdx \tag{2.7}$$

onde δ representa a função Delta de Dirac (Balakrishnan (2003)). O YOLOv8 utiliza a função de ativação SiLU ao final dos blocos convolucionais, onde essa função é definida como:

$$SiLU(x) = x * \sigma(x) \tag{2.8}$$

onde σ representa a função sigmoid logistica, definida como:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.9}$$

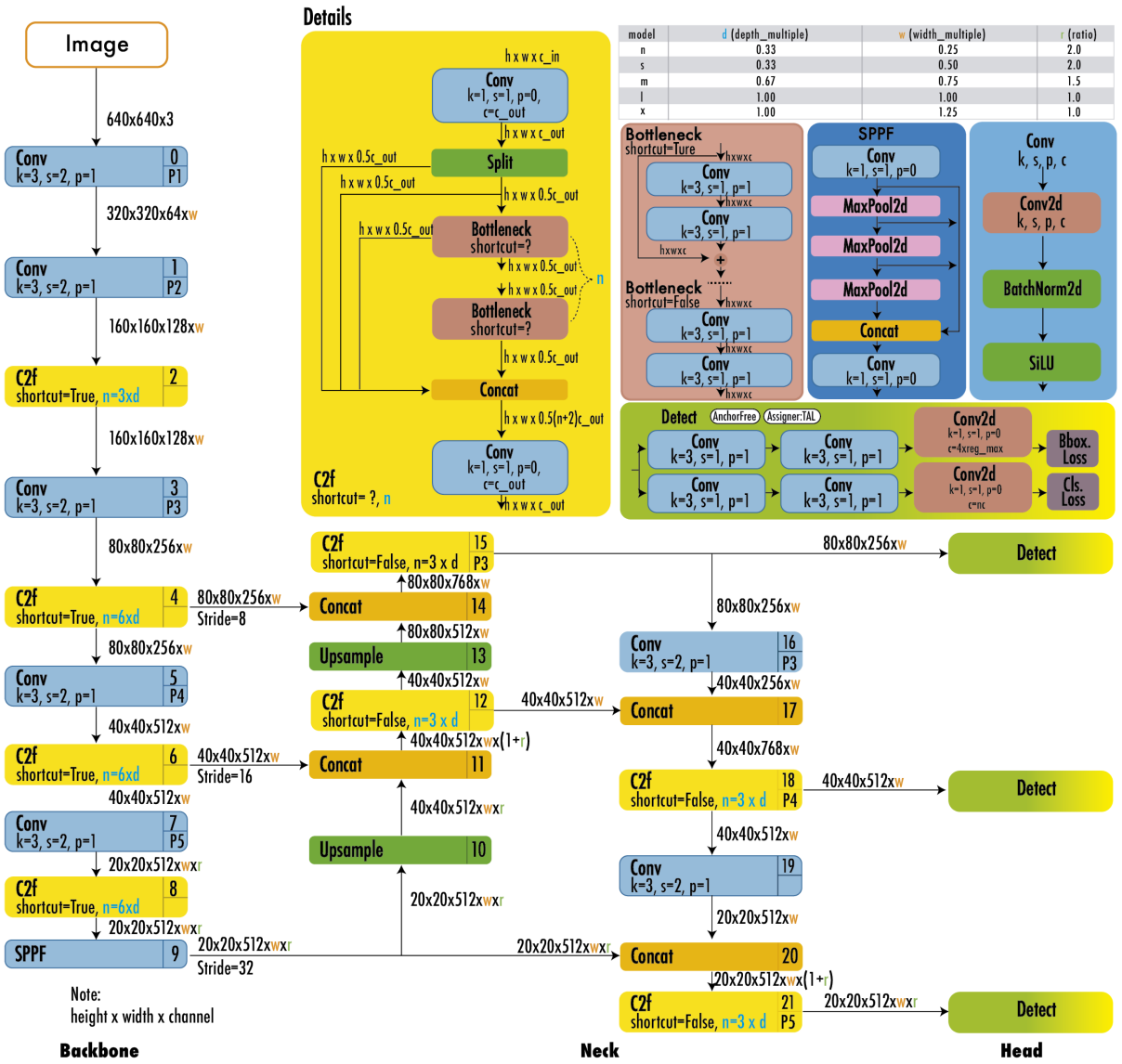


Figura 2.5: Arquitetura do YOLOv8. Utiliza um *backbone* CSPDarknet53 modificado. O módulo C2f substitui a camada CSP usada no YOLOv5. Uma camada de pirâmide espacial de pooling rápido (SPPF) acelera a computação ao agrupar características em um mapa de tamanho fixo. Cada convolução possui normalização em lote e ativação SiLU. A cabeça é desacoplada para processar tarefas de objetividade, classificação e regressão de forma independente. Fonte: Terven et al. (2023).

2.4 DEtection TRansformers (DETR)

Seguindo o exemplo do ViT, pesquisadores do Facebook Research foram mais além e propuseram em 2020 o modelo DETR, o primeiro modelo de detecção de objetos utilizando transformers Carion et al. (2020). A metodologia proposta considera a detecção de objetos como um problema de predição direta de conjuntos. O DETR é capaz de predizer todos os objetos paralelamente, e utiliza uma função de perda que realiza a correspondência bipartida (*bipartite matching*) entre os objetos preditos e os objetos de ground-truth, baseado no método Húngaro Kuhn (1955). O modelo DETR infere um número fixo N de predições por imagem a cada vez que passa pelo bloco de decoder, onde N é definido como sendo um número significantemente maior do que o número típico de objetos na imagem. A função de perda produz a correspondência bipartida ótima entre as predições e os objetos de *ground-truth*, o que significa que apenas uma caixa correspondente ao objeto é considerada, com perda mínima.

2.4.1 Arquitetura

O DETR possui uma arquitetura simples, que contém três elementos principais: Um backbone CNN, um bloco transformer encoder-decoder, e uma rede feed-forward (FFN), que pode ser visualizado na figura 2.6. O backbone CNN é usado para extrair representações reduzidas das características das imagens. Em sequência encoding posicionais são adicionados para produzir um token que é passado como entrada para o transformer. O modelo conta ainda com blocos encoder-decoder, que utilizam embeddings posicionais (object queries) para observar partes específicas da imagem e predizer os objetos (com boxes e classe) ou "no object". Por fim, uma rede feed-forward (FFN) é responsável pela predição final do modelo.

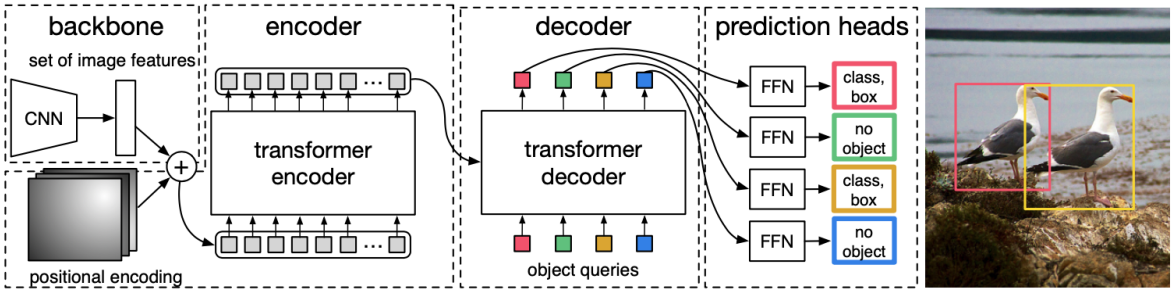


Figura 2.6: Arquitetura geral do DETR. Fonte: Carion et al. (2020).

2.5 Métricas de Avaliação de Desempenho

Em detecção de objetos, diversas métricas são utilizadas para avaliar o desempenho dos modelos. Em geral, são adotados padrões utilizados em competições, como COCO e PASCAL VOC. Nesta são descritos alguns conceitos fundamentais e métricas utilizadas na detecção de objetos, sumarizado por Padilla et al. (2020).

2.5.1 Conceitos Fundamentais

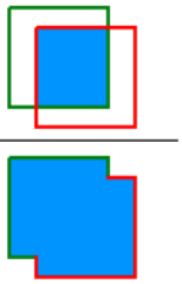
- **Nível de confiança**: Cada previsão feita pelo modelo é associada a uma pontuação que indica a confiança com a qual a previsão foi feita. Na detecção de objetos, mais especificamente, indica a confiança de que a caixa de marcação, do inglês *bounding box* (BB) contém um objeto.

- **IoU (Interseção sobre união)**: Representa o nível de sobreposição entre a BB predita e o *ground truth*, que seria valor verdadeiro esperado. Quanto melhor for o encaixe entre as caixa, mais próximo o valor do IoU estará de 1, e quanto mais desajustado, mais próximo de 0. O cálculo é feito dividindo a área de interseção pela área de união entre uma BB predita (BB_p) e uma BB de *ground truth* (BB_{gt}).

$$IoU = \frac{area(BB_p \cup BB_{gt})}{area(BB_p \cap BB_{gt})} \tag{2.10}$$

  Uma representação visual do IoU é mostrado na figura a seguir:

$$IOU = \frac{\text{area of overlap}}{\text{area of union}} =$$



Uma classificação pode ser classificada como verdadeiro positivo (TP), falso positivo (FP), falso negativo (FN) ou verdadeiro negativo (TN), onde esta última categoria não é utilizada no contexto de detecção de objetos. Essas categorias são definidas como:

- **Verdadeiro positivo (TP)**: Uma detecção correta. Detecção com IoU $\geq$ *threshold*.
- **Falso positivo (FP)**: Uma deteção errada. Detecção com IoU $<$ *threshold*.
- **Falso negativo (FN)**: *Ground truth* não detectado.

2.5.2 Métricas de Detecção de Objetos

Algumas das métricas mais populares utilizadas na detecção de objetos são:

- **Precisão**: É a capacidade que um modelo tem de evitar rotular amostras negativas como positivas. O cálculo da precisão é feito dividindo o número de verdadeiros positivos (TP) pela soma entre verdadeiros positivos (TP) e falsos positivos (FP).

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

- **Recall (Sensitivity)**: Se trata da capacidade que o modelo tem de detectar todas as amostras positivas. O cálculo é feito dividindo o número de verdadeiros positivos (TP) pela soma entre verdadeiros positivos (TP) e falsos negativos (FN).

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{2.12}$$

- **F1-score**: Também conhecido como F-score ou f-measure, o F1-score pode ser definido como sendo a média ponderada entre a precisão e o recall, com valores variando em 0 e 1.

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{2.13}$$

- ***Mean Average Precision* (mAP)**: O AP (*Average Precision*) é um valor numérico que ajuda a comparar diferentes modelos de detecção de objetos. A definição geral de precisão média (*average precision*) é como sendo a área abaixo da curva precisão-recall. Para lidar com o efeito zig-zag da curva, primeiro é realizado a interpolação da precisão utilizando diferentes valores de recall. O método mais popular é utilizar 11 pontos para a interpolação: [0,0.1,0.2,...,1].

$$AP = \frac{1}{11} \sum_{r \epsilon \{0, 0.1, ..., 1\}} p_{interp(r)} \tag{2.14}$$

  O método de avaliação COCO usa um método de interpolação de 101 pontos para cálculo de AP junto com a média de mais de dez limites de IoU. AP@[.5:.95] corresponde a uma média de AP para IoU de 0.5 a 0.95 com intervalos de 0.05. O COCO utiliza ainda as métricas AP@0.50 e AP@0.75, que são os mAPs para valores de IoU de 0.5 e 0.75, denotados como mAP50 e mAP75, respectivamente.

- **Curva FROC**: A curva FROC (*Free-Response Receiver Operating Characteristic curve*) é definida como a relação entre a sensibilidade do modelo (sensitivity) e sua taxa de falso positivo (FPR), onde pode-se calcular a área sob a curva (AUC), que

é uma excelente maneira de descrever a efetividade de um sistema de detecção de objetos, onde um maior valor de AUC indicaria uma melhor performance do modelo.

Além de todas essas métricas, também foi utilizada uma métrica proposta por Bandos et al. (2009), onde segundo o autor, embora a área sob a curva FROC resuma o desempenho do sistema FROC para todos os limiares de decisão, em algumas situações, a área sob a curva FROC pode ser considerada um índice subótimo ou, pior, um resumo potencialmente enganoso do desempenho geral do sistema. Para lidar com esse problema, este mesmo autor propôs um método para melhorar essa métrica adicionando uma ampliação da curva FROC até um valor especificado de FPR e uma nova curva representando um modelo ingênuo que apenas tentaria adivinhar os valores, chamada de curva *Guessed*. Assim, a nova métrica seria a área sob a curva FROC ampliada e acima da curva *Guessed*, sendo esse valor denotado por Λ. Este método pode ser visualizado na figura 2.7, onde FAUC representa a AUC da curva FROC original, G seria a AUC da curva *Guessed* e a linha pontilhada seria a ampliação da curva FROC.

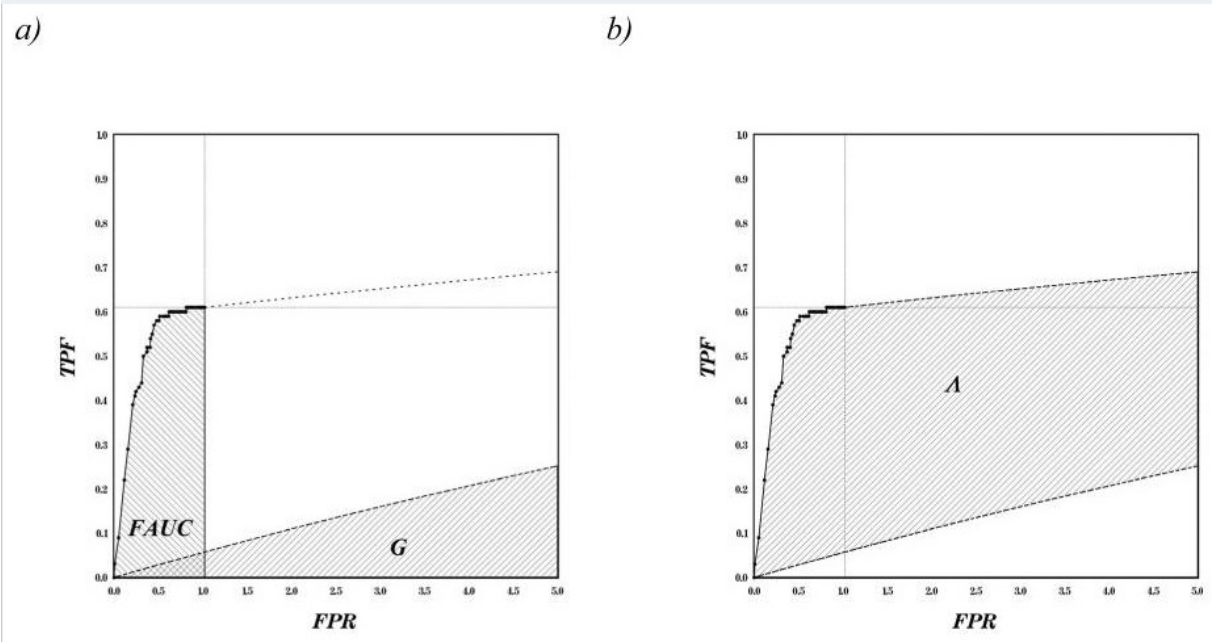


Figura 2.7: Demonstração visual da métrica Λ. Fonte: Bandos et al. (2009).

Capítulo 3

Metodologia

O esquemático geral da metodologia aplicada deste trabalho é apresentado na Figura 3.2. O conjunto de dados usado para o treinamento do modelo foi gerado a partir do LUNA16, que contém TCs de tórax em 3D. O conjunto de dados foi pré-processado (Figura 3.1-A), e cortes específicas de cada TC foram selecionadas para compor o conjunto de dados de treinamento; além de delimitar as BB de cada nódulo, a seção 3.1 descreve a base de dados e os passos para obtenção do conjunto de dados para o treinamento do modelo (Figura 3.1-B). O conjunto de dados foi então dividido em conjuntos de treinamento e teste, com uma proporção de 80/20, respectivamente. Após isso, uma análise dos tamanhos dos nódulos foi realizada nos cortes presentes no conjunto de teste, mantendo apenas as cortes onde os diâmetros dos nódulos eram inferiores a 15 mm (Figura 3.1-C), conforme descrito na seção 3.2. Posteriormente, dois modelos foram treinados: DETR e YOLOv8, passando por um ciclo onde ambos os modelos foram treinados, os resultados avaliados e melhorias nos hiperparâmetros foram feitas para aprimorar o desempenho dos modelos (Figura 3.1-D), a seção 3.3 elucida a motivação por trás da escolha dos modelos usados neste trabalho e a seção 3.4 ilustra a configuração e execução do treinamento para ambos os modelos. Finalmente, uma análise e comparação dos resultados (Figura 3.1-E) obtidos usando as métricas descritas na seção 2.5.2 foram conduzidas. Neste trabalho, a linguagem de programação Python (versão 3.7) foi utilizada, juntamente com os modelos DETR (Carion et al. (2020)) e YOLOv8 (Ultralytics (2023)), além das bibliotecas Pytorch (Paszke et al. (2019)), Skimage (Van der Walt et al. (2014)), Scipy (Virtanen et al. (2020)), Monai (Cardoso et al. (2022)) e Ultralytics (Jocher et al. (2023)).

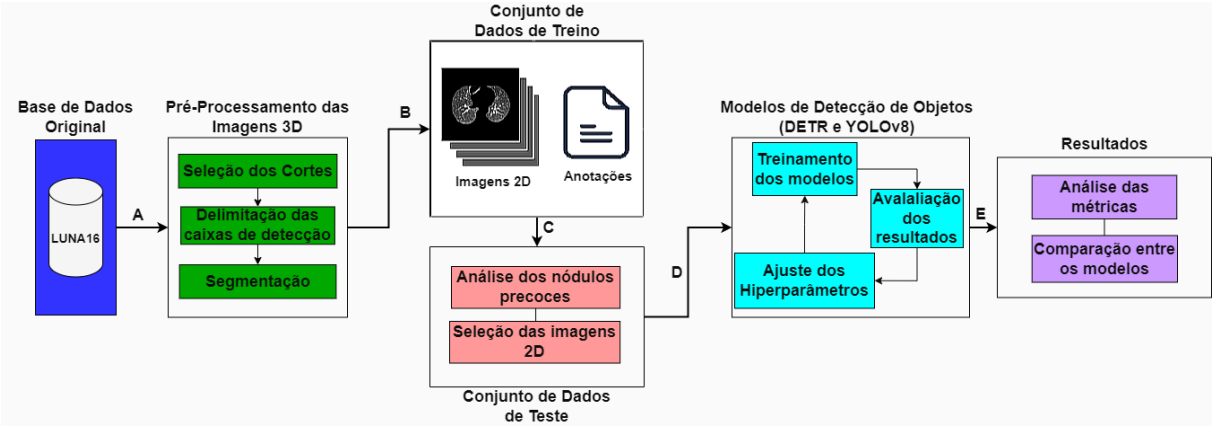


Figura 3.1: Esquemática metodológico aplicado neste trabalho. Fonte: autor, 2024.

3.1 Base de Dados

A base de dados selecionada para este estudo foi o LUNA16 (LUng Nodule Analysis 2016) (Setio et al. (2017)), um conjunto de dados publicamente disponível de exames de tomografia computadorizada (TC) da região torácica de pacientes identificados com nódulos pulmonares, contendo 888 exames de TC e 1186 anotações de nódulos.

Neste trabalho, utilizamos imagens 2D devido ao custo computacional associado ao processamento de imagens 3D e ao fato de que um dos modelos analisados (YOLOv8) não suporta imagens 3D. Com isso, o conjunto de dados original utilizado neste trabalho foi o mesmo criado por Nilson Carvalho em seu trabalho de conclusão de curso (CARVALHO (2022)), onde as imagens originalmente estavam no formato DICOM e as anotações do nódulos em um arquivo CSV, contendo as coordenadas referentes ao centro do nódulo (X, Y e Z), raio do nódulo (em milímetros) e ID único do exame. A partir disto, o autor citado anteriormente selecionou apenas os cortes dos exames de TC onde os nódulos estavam presentes, incluindo multiplos cortes de um mesmo nódulo, não apenas o corte central, e fez um pré-processamento nestas imagens, buscando eliminar ruídos, artefatos e outras informações irrelevantes dos exames para tentar focar apenas nos órgãos que são objeto do estudo. Com isso, a base de dados foi disponibilizada para a utilização neste trabalho pelo próprio autor, através de uma pasta no Google Drive, contendo as imagens 2D compactadas em arquivos .ZIP e as anotações dos nódulos nos formatos COCO e YOLO.

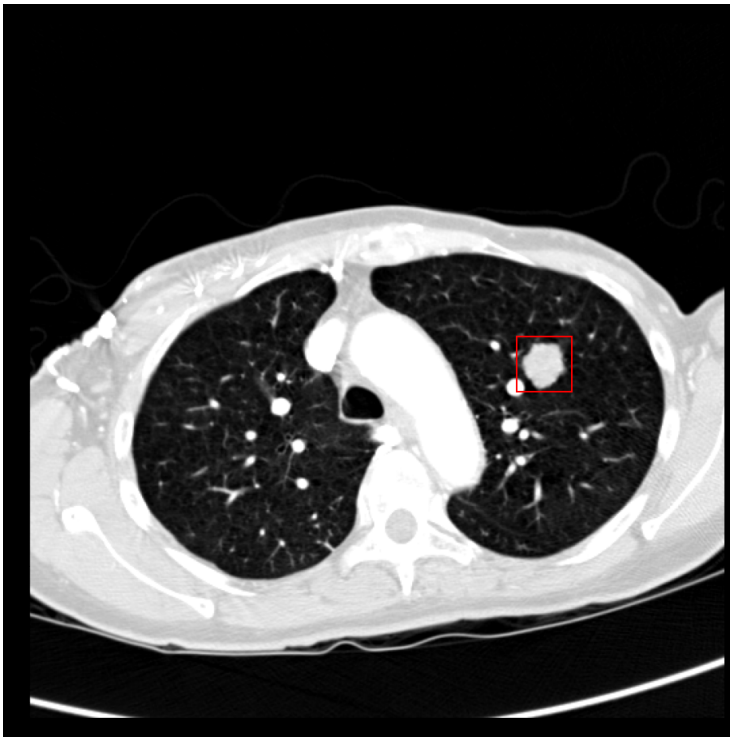


Figura 3.2: Região torácica com nódulo pulmonar em evidência. Fonte: CARVALHO (2022).

3.2 Organização do Conjunto de Dados

Para a criação do nosso conjunto de dados, o banco de dados foi dividido em conjuntos de treinamento e teste, com 80% para treinamento e 20% para teste. Mitigamos vazamentos de dados garantindo que imagens do mesmo nódulo não fossem incluídas nos conjuntos de treinamento e teste. No total, o conjunto de treinamento contém 3.255 imagens, enquanto o conjunto de teste contém 832, onde o diâmetro dos nódulos d pertence ao intervalo $d \in [3, 32]$, em milímetros. As imagens foram salvas em sua resolução original (512 x 512) com 8 bits.

Posteriormente, foi realizada uma análise dos nódulos no conjunto de teste, selecionando apenas os cortes das TCs onde os nódulos detectados tinham um diâmetro inferior a 15mm, caracterizando-os como nódulos pequenos. Esse processo envolveu separar todas os cortes pertencentes ao mesmo exame de TC no conjunto de teste e verificar se o maior nódulo neste conjunto de seções tinha um diâmetro menor que 15 mm. Assim, todo o conjunto de cortes foi adicionado à versão final do conjunto de dados, enquanto aqueles que falharam nesta verificação foram removidos. Consequentemente, das 832 imagens originalmente no conjunto de teste, apenas 587 permaneceram, que foram utilizadas para validação dos modelos. Além disso, foi feita uma subdivisão no conjunto de dados onde foram selecionados três intervalos (I1, I2 e I3) de diâmetros dos nódulos (d), onde a

quantidade de cortes em cada intervalo por ser vista na tabela 3.1.

Diâmetro (d)	**Cortes (Conjunto de Teste)**
I1: 15mm $\geq d >$ 10mm	172
I2: 10mm $\geq d >$ 5mm	385
I3: 5mm $\geq d$	30

Tabela 3.1: Tabela mostrando a disposição das imagens dentro dos intervalos determinados. Fonte: autor, 2024.

3.3 Escolha dos Modelos

Neste trabalho, nosso objetivo foi comparar os resultados de dois modelos de detecção significativamente distintos. O primeiro é um tipo de rede neural que tem ganhado popularidade nos últimos anos, conhecido como *Transformer*, neste caso, o *DEtection TRansformer* (DETR) Carion et al. (2020). O outro modelo escolhido foi o *You Only Look Once* (YOLO) Jocher et al. (2023), uma rede neural mantida pela Ultralytics que é bem conhecida e estabelecida no campo de detecção de objetos, atualmente em sua oitava versão, que foi utilizada neste estudo.

O DETR emprega a arquitetura *Transformer* em conjunto com uma CNN e simplifica o pipeline de detecção, não requerendo camadas personalizadas, tornando-o facilmente reproduzível em qualquer framework que contenha classes padrão de CNN e Transformer Carion et al. (2020). A escolha deste modelo se deu pelo fato do DETR ter sido o primeiro modelo a integrar um bloco *Transformer* no pipeline de detecção de objetos e por possuir extensa discussão em fóruns, o que facilita o entendimento do modelo e sua adaptação a diferentes cenários.

A versão mais recente do YOLO no momento da realização deste trabalho (YOLOv8), foi escolhida porque, como mencionado anteriormente na seção 2.3, é um modelo aclamado na área de detecção de objetos e considerado estado-da-arte. Este modelo tem demonstrado excelentes resultados em diversos domínios e mostrou consideráveis melhorias de desempenho em comparação com suas versões anteriores, como já foi apresentado anteriormente na figura 2.3.

3.4 Treinamento dos Modelos

O treinamento para ambos os modelos foi conduzido em um servidor remoto executando o sistema operacional Linux (Ubuntu 20.04 LTS) e equipado com uma placa gráfica NVIDIA Tesla T4 com 16GB de memória. O procedimento inicial para o treinamento envolveu a criação de um *fork* do repositório DETR no GitHub para fazer as modificações necessárias e adaptar o modelo para detecção de nódulos pulmonares. O número de

camadas foi mantido o mesmo que no modelo original. A primeira modificação envolveu a alteração do número de consultas de objeto, que é a quantidade de objetos previstos por imagem, onde o número foi reduzido para 10, uma vez que o número de nódulos na mesma imagem geralmente é muito menor que 100 (valor padrão). Outros hiperparâmetros do modelo foram readaptados conforme o treinamento progredia.

Uma das transformações realizadas pelo modelo antes do treinamento para melhor generalização envolve o redimensionamento aleatório das imagens, seguindo uma lista de tamanhos. O tamanho máximo para ampliação foi definido como 800 x 800. O tamanho do lote foi configurado para 8. O modelo foi então treinado por 400 épocas. O valor de queda da taxa de aprendizado também foi aumentado de 200 para 1000 devido ao número relativamente pequeno de imagens em nosso banco de dados. A tabela 3.2 expõe os hiperparâmetros utilizados no treinamento do modelo.

Descrição do parâmetro	**Valor**
Número de camadas encoder	6
Número de camadas decoder	6
Tamanho dos embeddings	256
Número de slots para detecção	10
Dropout aplicado no transformer	10^{-1}
Número de attention heads usadas para atenção	8

Tabela 3.2: Tabela mostrando os hyperparâmetros utilizados no treinamento do DETR. Fonte: autor, 2024.

Continuando com o treinamento e utilizando os hiperparâmetros identificados, conseguimos treinar o modelo por 400 épocas, com cada época levando em média 10 minutos para ser concluída, totalizando aproximadamente 66 horas de treinamento.

O treinamento do YOLOv8 foi bem mais simples, já que a Ultralytics fornece um pacote Python que simplifica muito o processo de treinamento e uso do modelo. O treinamento foi configurado para rodar por 400 épocas, semelhante ao DETR. No entanto, um mecanismo de parada precoce foi implementado, onde se nenhuma melhoria fosse observada no treinamento nas últimas 50 épocas, o treinamento seria interrompido. Consequentemente, o treinamento do modelo foi interrompido após 126 épocas, indicando que o melhor modelo foi observado já na época 76. O treinamento para este modelo ocorreu relativamente rápido em comparação com o DETR, levando não mais do que 2 horas.

Devido à facilidade e rapidez do treinamento, decidimos treinar um segundo modelo YOLOv8, desta vez aplicando aumento de dados ao conjunto de treinamento, onde este modelo será chamado de YOLOv8AUG a partir de então. Rotações de 90, 180 e 270 graus foram aplicadas a imagens aleatórias no conjunto original, aumentando assim o conjunto de treinamento de 3.255 para 9.469 imagens. O treinamento deste modelo também foi

conduzido com as mesmas configurações do anterior, sendo interrompido após 77 épocas e levando aproximadamente 80 minutos para ser concluído.

Capítulo 4

Resultados e Discussões

Neste capítulo serão apresentados os resultados obtidos através da metodologia apresentada no capítulo anterior para a detecção de nódulos pulmonares precoces em imagens de TC.

4.1 Resultados do Treinamento

A Figura 4.2 apresenta os resultados do modelo DETR para erro de classe e mAP50 obtidos após o treinamento por 400 épocas. Nosso modelo DETR é representado pela curva laranja. Os valores de erro de classe exibidos na imagem não são exatos devido ao uso de menos previsões por imagem (10) do que o modelo original (100). No entanto, a curva é útil para observar que o erro de classe está diminuindo ao longo do treinamento. A segunda métrica mostrada é o mAP50, que se refere à precisão média usando um IoU de 0,5, onde podemos observar que a curva exibe uma leve tendência ascendente até as últimas épocas registradas, indicando que com algumas épocas adicionais de treinamento, as caixas delimitadoras poderiam se tornar ainda mais precisas. Podemos observar os resultados do treinamento doo modelos YOLOv8 e YOLOv8AUG na Figura 4.1, comportando-se de maneira semelhante aos gráficos do DETR, no entanto, alcançando valores de mAP50 mais altos dentro de uma faixa de épocas relativamente menor, onde o pico é alcançado 50 épocas antes do fim do treinamento, graças ao mecânismo de parada precoce implementado. Para comparar os resultados de forma mais direta, a Tabela 4.1 contém as métricas descritas em 2.5.2 para todos os três modelos treinados. Assim, é fácil observar que, em relação à precisão do modelo (mAP50), o YOLOv8 se destaca consideravelmente em comparação com o DETR, onde o modelo YOLOv8AUG alcançou mAP50 = 0,896. Outra comparação interessante pode ser feita observando a equação 2.13, que relaciona sensibilidade e precisão, servindo como um indicador geral mais confiável da eficácia do modelo, onde os resultados do YOLOv8 se destacam mais uma vez.

Model	mAP50	f1-Score
DETR	0.703	0.772
YOLOv8	0.844	0.805
YOLOv8AUG	**0.896**	**0.848**

Tabela 4.1: Tabela exibindo os valores das métricas mAP50 e F1-Score descritas na Seção 2.5.2 para todos os modelos, com o valor ótimo para cada métrica destacado em negrito. Fonte: autor, 2024.

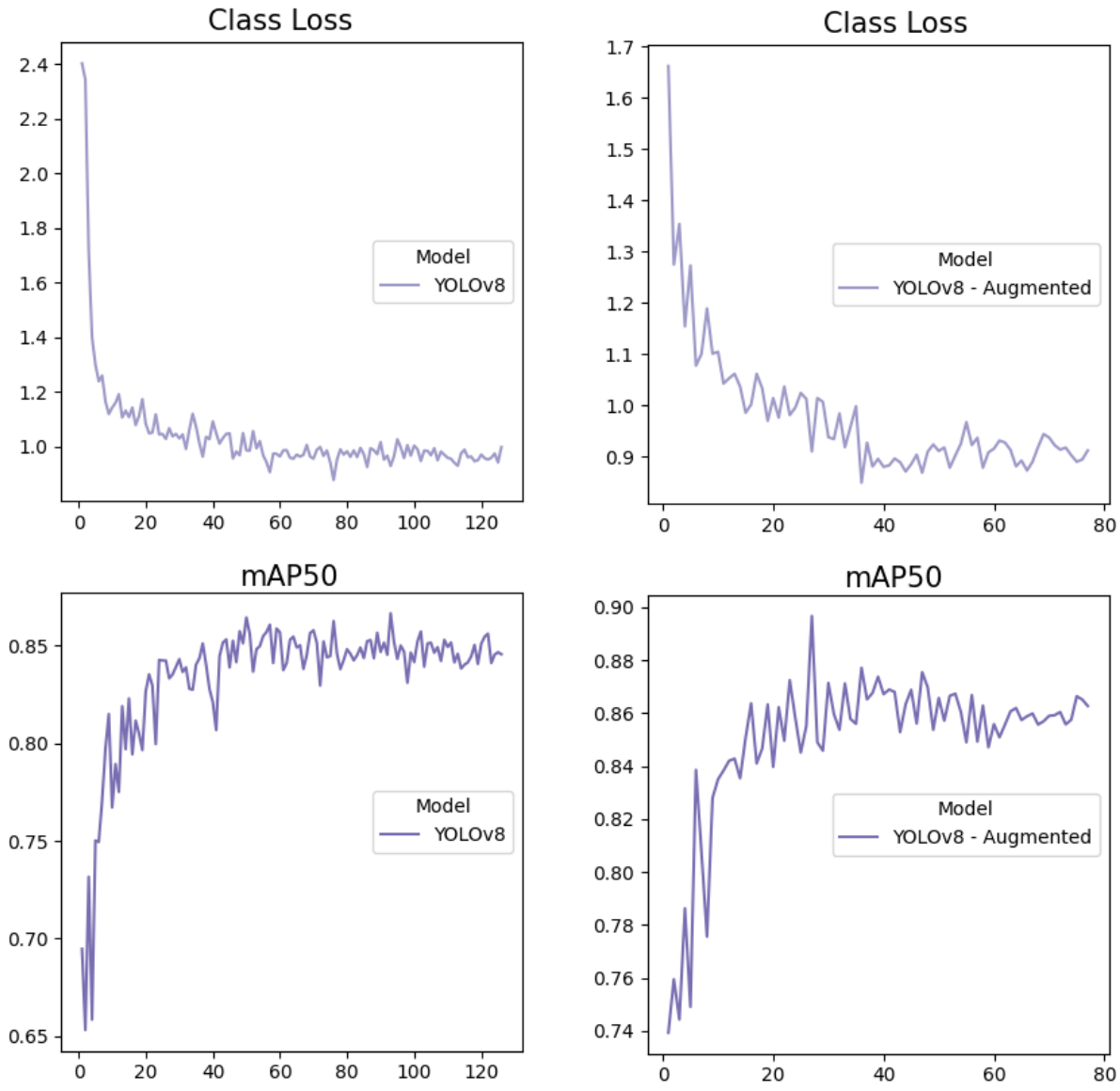


Figura 4.1: Plotagem dos logs de treinamento dos modelos YOLOv8 e YOLOv8AUG. Fonte: autor, 2024.

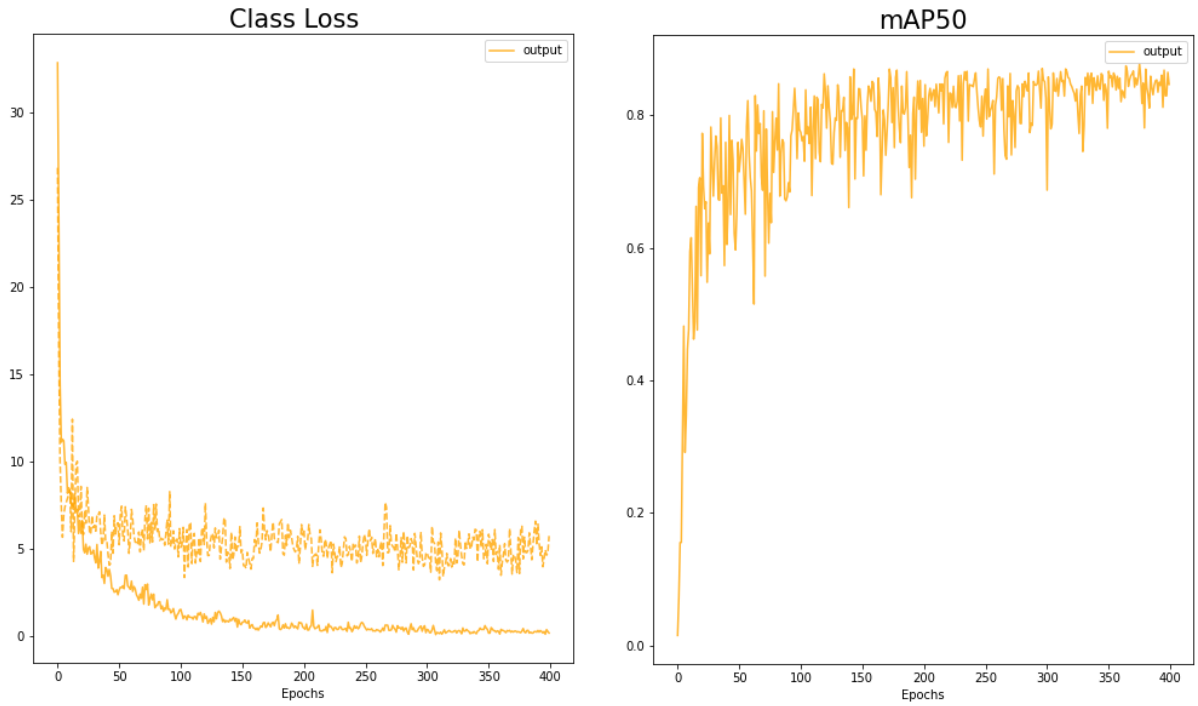


Figura 4.2: Plotagem dos logs de treinamento do modelo DETR. Fonte: autor, 2024.

4.2 Resultados da Validação

Para uma validação mais precisa dos modelos, utilizaremos a curva FROC, como descrito na Seção 2.5.2, para fornecer uma medida de comparação mais profunda do que o mAP. Utilizando o pacote Python Monai, foi possível plotar as curvas FROC para os três modelos, como pode ser visto na figura 4.3. A primeira observação são os valores de AUC dos modelos, onde o DETR claramente se destaca, e os valores de pontuação de sensibilidade (*Score*, na figura) também são superiores. No entanto, ao prestar mais atenção nas curvas, algo levemente inconsistente é notado, onde o AUC do modelo DETR é mais de duas vezes maior do que o modelo YOLOv8AUG, pois a taxa de falsos positivos (eixo x das curvas) é significativamente maior, com isso o modelo DETR apresenta uma maior métrica graças a apresentar mais erros do tipo FP, sendo isso algo bem ilógico. Consequentemente, podemos entender o que os autores de Bandos et al. (2009) queriam dizer ao afirmar que o valor de AUC da curva FROC pode ser um valor enganoso. Para resolver isso, foi criado um script para plotar a curva *Guessed* e a curva FROC aumentada seguindo as equações descritas pelos autores mencionados, truncando no valor de FPR $= 1$ para todos os modelos. Os gráficos resultantes podem ser observados na Figura 4.4, onde a métrica Λ parece refletir de forma mais concisa o que foi observado a partir das métricas anteriores, igualando o eixo x para todos os modelos, sendo assim uma comparação mais justa.

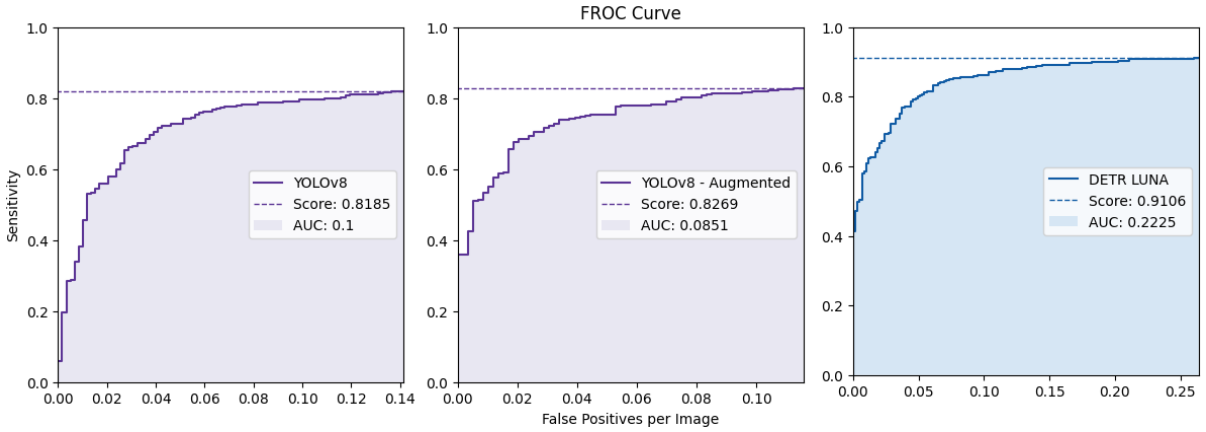


Figura 4.3: Plotagem da curva FROC para os três modelos, YOLOv8 a esquerda, YOLOv8AUG ao centro e o DETR a direita, onde a métrica *Score* representa o valor da sensibilidade de cada modelo. Fonte: autor, 2024.

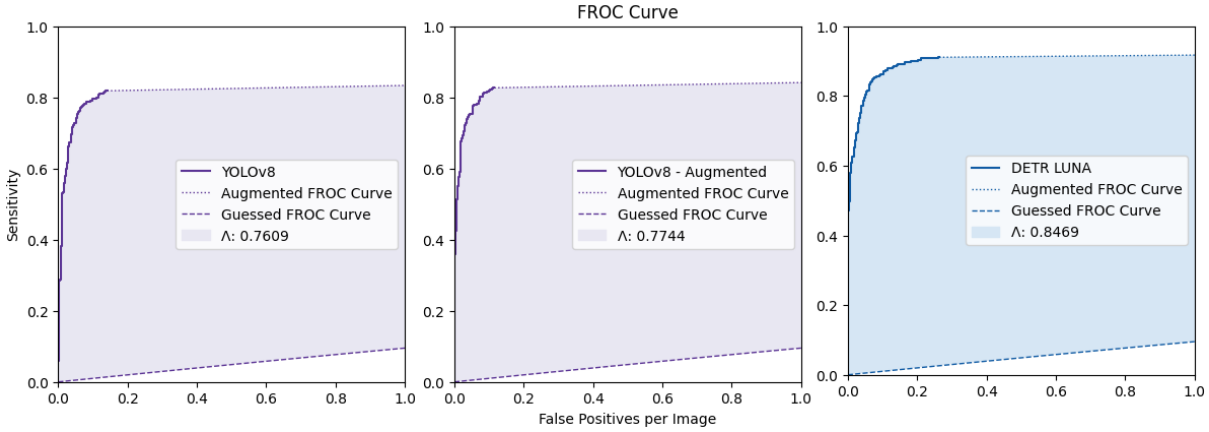


Figura 4.4: Plotagem da curva FROC melhorada segundo Bandos et al. (2009) para os três modelos, YOLOv8 a esquerda, YOLOv8AUG ao centro e o DETR a direita, onde as linhas tracejadas representam as curvas *Guessed* e as linhas pontilhadas representam o aumento das respectivas curvas FROC. Fonte: autor, 2024.

Para a comparação final entre os modelos, vamos observar os valores de TP, FP, FN, sensibilidade, AUC da FROC e Λ para todos os modelos treinados na Tabela 4.2. Usando Λ como a métrica primária para comparação de modelos, observamos que o melhor modelo foi o DETR, obtendo um $\Lambda = 0.847$, onde um modelo ideal tem um $\Lambda = 1$, enquanto o YOLOv8AUG obteve um $\Lambda = 0.774$. Uma observação importante é notar qual modelo teve menos falsos negativos, pois, em um contexto médico, esse é um tipo de erro muito significativo. Neste caso, observamos que, de um total de 595 nódulos presentes nos 587 cortes do conjunto de teste, o modelo DETR mais uma vez teve o melhor desempenho, apresentando a menor ocorrência de FN entre os três modelos, mas apresentando um alto valor de FP em comparação com o YOLOv8, que é uma relação de troca que pode ser

estudada mais detalhadamente em trabalhos subsequentes.

Model	**TP**	**FP**	**FN**	**Sensitivity**	**FROC AUC**	Λ
DETR	**542**	156	**53**	**0.910**	**0.222**	**0.847**
YOLOv8	487	84	108	0.818	0.100	0.761
YOLOv8AUG	492	**69**	103	0.827	0.085	0.774

Tabela 4.2: Tabela exibindo os valores das outras métricas descritas na Seção 2.5.2 para todos os modelos, com o valor ótimo para cada métrica destacado em negrito. Fonte: autor, 2024.

4.3 Inferências e Discussão

Para realizar inferências nas imagens, o DETR nos permite carregar o modelo genérico do repositório TorchHub e atualizá-lo com os pesos obtidos durante o treinamento. Cada inferência retorna dez consultas com dez previsões, mas retemos apenas as previsões com confiança acima de 0,8. Para o YOLOv8, o processo é simplificado usando o pacote Python da Ultralytics, onde o modelo é automaticamente salvo ao final do treinamento, e este modelo pode ser carregado diretamente para inferência, configurando a confiança acima de 0,8, similarmente ao DETR. Alguns exemplos de inferências foram salvos, onde a figura 4.5 mostra quatro cortes com as predições feitas pelo modelo DETR e a figura 4.6 mostra quatro cortes com as predições feitas pelo modelo YOLOv8AUG, onde nelas temos um nódulo com $d > 15$mm (corte superior esquerdo), um nódulo pertencente ao intervalo I1 ($15mm \geq d > 10mm$) (corte superior direito), um nódulo pertecente ao intervalo I2 ($10mm \geq d > 5mm$) (corte inferior esquerdo) e um nódulo pertecente ao intervalo I3 ($5mm \geq d$) (corte inferior direito). As caixas delimitadoras azuis representam a inferência do modelo, enquanto as caixas vermelhas mostram os valores de verdade absoluta (*ground truth*).

As detecções obtidas por todos os modelos mostraram bons resultados em termos de localização e tamanho da caixa delimitadora, aproximando-se bastante dos valores marcados pelos radiologistas, porém o YOLOv8AUG apresentou um ajuste de BB mais preciso, com maior IoU. O modelo com melhor desempenho na detecção de pequenos nódulos observados neste estudo (DETR) ocupa 1,9 GB de memória da GPU, e cada inferência leva 0,6 segundos para ser executada (1,4 segundos para realizar a inferência e plotar o resultado). Portanto, entendemos que o modelo poderia ser usado em um computador padrão com uma GPU simples.

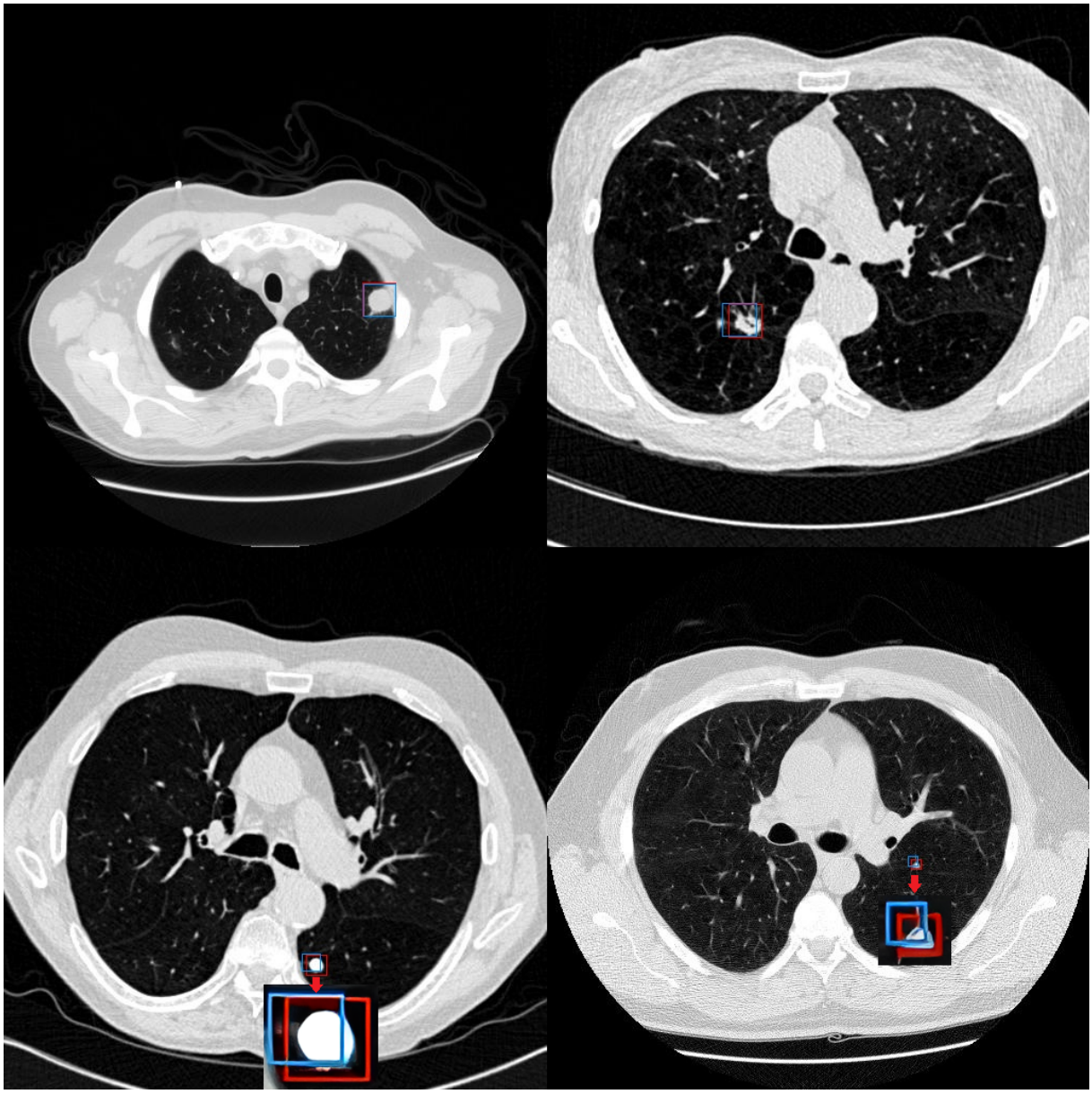


Figura 4.5: Inferências do modelo DETR (em azul) ao lado dos valores de GT (em vermelho). Seguindo a ordem da imagem da esquerda para a direita e de cima para baixo, estão apresentadas predições de um nódulo com mais de 15mm de diâmetro, e na sequência temos nódulos que pertecem aos intervalos I1, I2 e I3, respectivamente, contendo uma aproximação dos menores nódulos para melhor visualização. Fonte: autor, 2024.

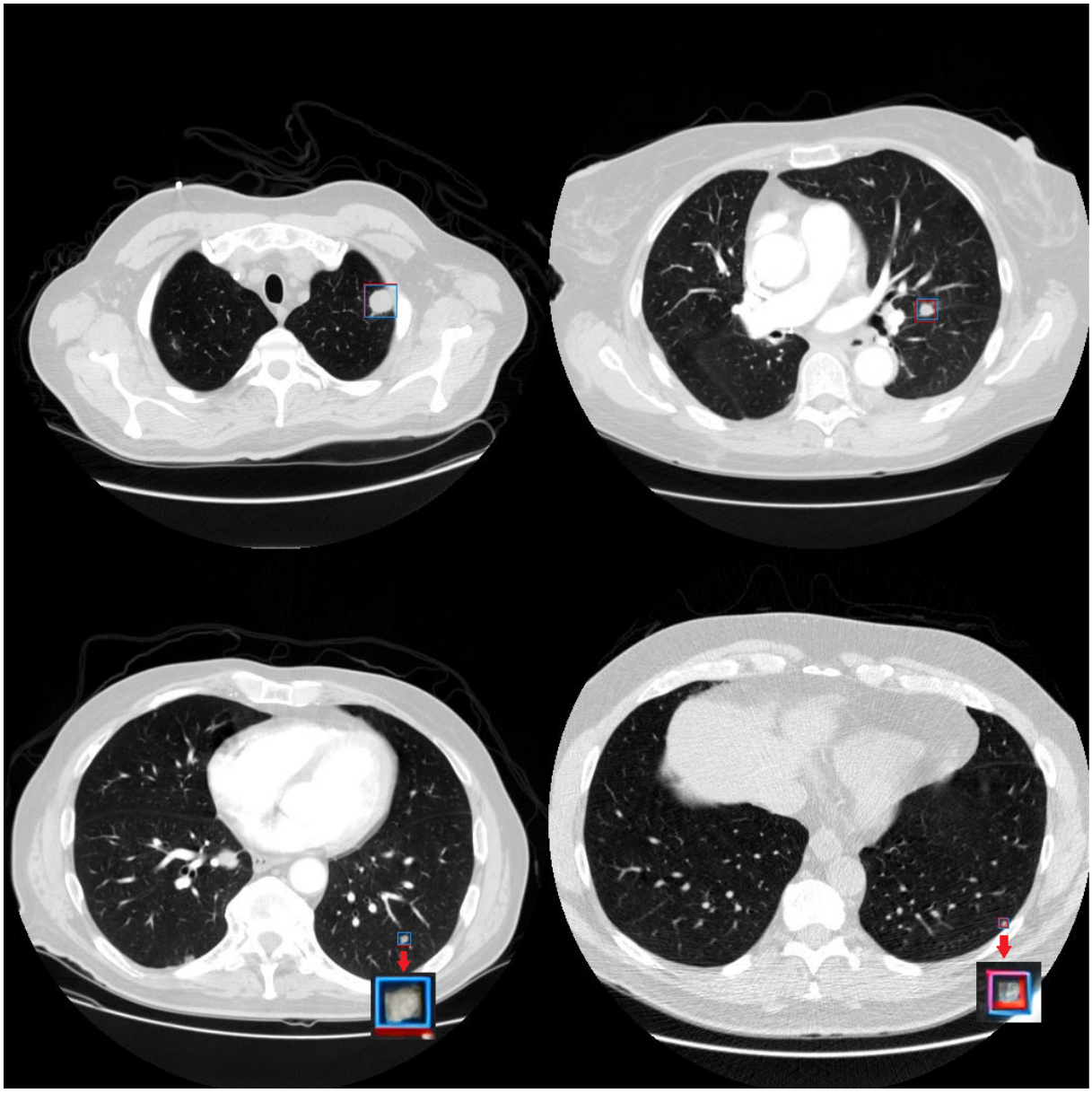


Figura 4.6: Inferências do modelo YOLOv8AUG (em azul) ao lado dos valores de GT (em vermelho). A lógica da sequência da figura 4.5 se aplica nesta também. Fonte: autor, 2024.

Ao final da Seção 3.2, foi mencionado que o conjunto de dados de teste foi dividido em três intervalos: I1, I2 e I3. É importante observar que nosso conjunto de testes continha apenas 30 imagens em I3, o que torna a detecção desses nódulos extremamente desafiadora, já que esse número reflete que no conjunto de treinamento o número de cortes com nódulo menores que 5mm também é pequeno. Além da pequena quantidade de exemplos presentes no conjunto de dados, um corte de TC com nódulos de tal tamanho apresentaria pouquíssima informação (pixels), dificultando assim a generalização dos modelos. Para responder à pergunta feita na Seção 1.1, examinaremos os valores de mAP50, sensibilidade e Λ para ambos os modelos nos intervalos I1, I2 e I3 (Tabela 4.3). À primeira vista, podemos observar que os resultados seguem a mesma lógica apresentada na Seção 4.2, onde o YOLOv8 se destaca em precisão, enquanto o DETR se destaca nos outros dois. Podemos ver claramente o impacto da redução do tamanho do nódulo no YOLOv8, onde todas as métricas diminuem à medida que os nódulos se tornam menores, enquanto o DETR consegue permanecer relativamente estável em todas as métricas, independentemente do intervalo sendo avaliado. Isso é especialmente evidente ao comparar diretamente os resultados de I1 (nódulos maiores, mais imagens) e I3 (nódulos menores, menos imagens), onde o YOLOv8 experimenta uma queda acentuada no mAP50 e uma diminuição considerável nas outras duas métricas, enquanto o DETR não apenas parece não ser afetado negativamente, mas até melhora no mAP50 e na sensibilidade, permanecendo quase inalterado no Λ. Também podemos observar a curva FROC melhorada plotada para todos os modelos nos três intervalos na figura 4.7.

Model	**mAP 50**			**Sensitivity**			Λ		
	I1	**I2**	**I3**	**I1**	**I2**	**I3**	**I1**	**I2**	**I3**
YOLOv8AUG	**0.907**	**0.902**	**0.727**	0.829	0.830	0.766	0.785	0.775	0.662
DETR	0.663	0.729	0.665	**0.858**	**0.851**	**0.900**	**0.801**	**0.790**	**0.799**

Tabela 4.3: Tabela contendo valores de mAP50, Sensitivy e Λ para ambos os modelos nos três intervalos de diâmetro, com o valor ótimo para cada métrica destacado em negrito. Fonte: autor, 2024.

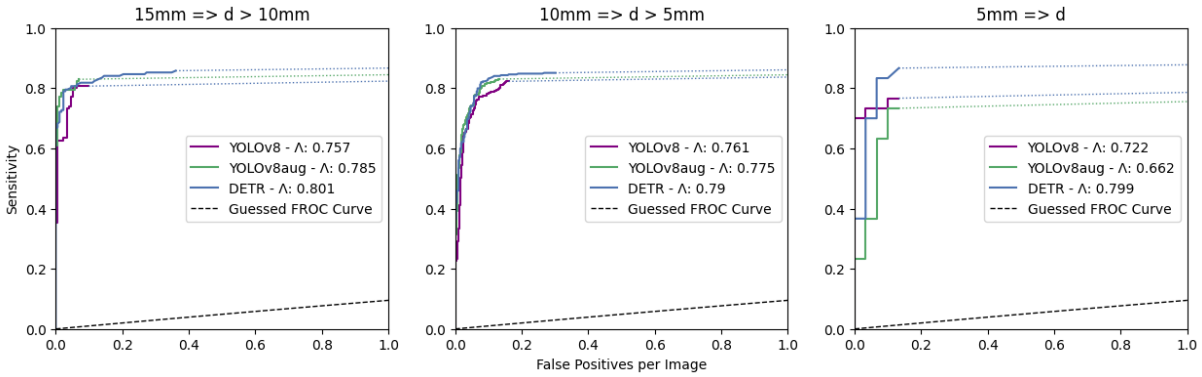


Figura 4.7: Curvas FROC melhoradas plotadas para todos os modelos nos três intervalos de diâmetro. Fonte: autor, 2024.

Para avaliar o desempenho de nossos modelos com nódulos maiores, conduzimos uma avaliação usando o conjunto de dados de teste antes de filtrar os nódulos pequenos, ou seja, com nódulos maiores e menores que 15 mm. O YOLOv8AUG alcançou um mAP50 de 0,907, semelhante à métrica em I1, e uma sensibilidade de 0,847, superior a todas as obtidas em todos os intervalos, o que reforça a dificuldade que o modelo enfrenta ao reduzir o tamanho dos nódulos. Por outro lado, o DETR alcançou um mAP50 de 0,866, significativamente superior a todos os intervalos, e uma sensibilidade de 0,899, similar a I3.

Após todas essas análises, fica claro observar que o YOLOv8 se destaca em termos de precisão, indicando que o modelo tem uma baixa taxa de falsos positivos (FP), enquanto o DETR se destaca em sensibilidade, apresentando uma baixa taxa de falsos negativos (FN). Ambos os tipos de erros são relevantes no contexto da detecção de nódulos; portanto, caberia ao profissional responsável escolher qual modelo aplicar para cada caso, dependendo de qual tipo de erro teria menos impacto.

Ao comparar os resultados obtidos por nossos modelos com modelos da literatura, nomeadamente Shah et al. (2023), que utiliza uma combinação de detecção de três CNNs 2D, compilando os resultados em uma única detecção, e Neal Joshua et al. (2021), que emprega a técnica de Ativação de Classe Ponderada por Gradiente em uma CNN 3D para detectar nódulos, observamos que nossos modelos não estão muito longe em termos de resultados. Nosso melhor modelo YOLOv8 até exibe uma precisão mais alta (88%) do que os outros (86% e 87%, respectivamente), enquanto o melhor modelo DETR permanece muito próximo em termos de sensibilidade (91%), em comparação com os outros (80% e 94%, respectivamente). Vale ressaltar que nosso conjunto de dados de teste contém apenas nódulos considerados pequenos (com até 15mm de diâmetro), enquanto os dois modelos citados não têm essa limitação, resultando em métricas mais altas. Isso enfatiza que os resultados obtidos por ambos os modelos neste estudo foram altamente satisfatórios.

Os modelos de detecção de objetos usando a arquitetura ViT (Transformer Visual)

são relativamente novos e apresentam alguns desafios para a comunidade de pesquisa. No campo de imagens médicas, esse desafio é particularmente ampliado pela escassez de imagens médicas, especialmente as anotadas, em comparação com outros domínios. Neste trabalho, demonstramos que um modelo de detecção de nódulos pulmonares precoces em imagens de TC usando uma arquitetura *Transformer* alcançou resultados satisfatórios em termos de mAP, sensibilidade e pontuação F1. Também superou o modelo YOLOv8 na análise baseada nas curvas FROC e exibiu caixas delimitadoras nas inferências com alinhamento satisfatório com os valores esperados.

Os modelos *Transformer*, incluindo o DETR, exigem um tempo de treinamento consideravelmente alto para alcançar os melhores resultados, tornando desafiador experimentar várias opções de hiperparâmetros - uma clara desvantagem em comparação com o YOLOv8, que oferece um processo de treinamento mais simplificado e rápido, além de um menor tempo de inferência, como pode ser visto na tabela 4.4. Os modelos *Transformer* também têm a peculiaridade de exigir uma grande quantidade de dados para alcançar resultados ótimos. Portanto, o conjunto de dados neste estudo poderia ser expandido para incluir outras bases de imagens.

Modelo	**Treinamento (por época)**	**Inferência**
DETR	10 minutos	600 milisegundos
YOLOv8	57 segundos	40 milisegundos

Tabela 4.4: Tabela demonstrando o tempo da duração de cada época de treinamento e o tempo que os modelos levam para realizar suas inferências. Fonte: autor, 2024.

Capítulo 5

Conclusão

Neste estudo, buscamos avaliar a eficácia de modelos baseados em arquiteturas CNN (YOLOv8) e *Transformer* (DETR) na detecção de pequenos nódulos pulmonares ($\leq$15mm) e comparar seus resultados, onde nosso melhor modelo DETR alcançou valores de mAP50 = 0,703, Sensibilidade = 0,91 e Λ = 0,847, resultados considerados satisfatórios, especialmente quando comparados com os resultados obtidos pelo YOLOv8, onde o DETR fica atrás apenas no mAP50. As caixas delimitadoras dos nódulos detectados também mostraram bom alinhamento com os valores de verdade absoluta. Considerando os resultados alcançados neste trabalho, podemos afirmar que um modelo com arquitetura *Transformer* é capaz de alcançar resultados similares a um modelo CNN na detecção de pequenos nódulos pulmonares, até mesmo superando o YOLOv8. Em termos do impacto do tamanho do nódulo no desempenho do modelo, observou-se que o DETR não é muito afetado pela diminuição do tamanho, alcançando bons resultados em todos os intervalos de diâmetro propostos, enquanto o desempenho do YOLOv8 diminui à medida que o tamanho do nódulo diminui. Também concluímos que ambos os modelos têm suas limitações, com o DETR exibindo mais casos de falsos positivos e o YOLOv8 mostrando mais falsos negativos. Isso é algo que deve ser levado em consideração pelos profissionais ao escolher o modelo mais adequado para cada caso.

As inferências realizadas neste trabalho foram feitas utilizando um Jupyter notebook, mas os modelos podem facilmente ser portados para serem incorporados ao pipeline de software mais robusto ou mesmo ser adicionado a uma aplicação móvel, o que abre diversas possibilidades para o seu uso.

É importante frisar que durante o desenvolvimento deste Trabalho de Conclusão de Curso, uma enorme gama de assuntos e tecnologias foram explorados, em especial nas áreas de processamento de imagens, projeto de software, aprendizagem de máquina, visão computacional e imagens médicas. O conhecimento poderá ser utilizado pela comunidade para o desenvolvimento de novas ferramentas e a continuação dos estudos na área.

5.1 Trabalhos Futuros

Seguem algumas sugestões para trabalhos futuros:

- Os modelos *Transformer* necessitam de um grande número de imagens para conseguir melhores resultados, então é importante adicionar imagens de outras bases de TCs com o objetivo de aumentar o número de amostras no treino e atingir melhores resultados.
- Testar outros hiperparâmetros no modelo DETR e treiná-lo por mais épocas para tentar otimizar os resultados.
- Desenvolver uma maneira de combinar os resultados de ambos os modelos, YOLOv8 e DETR, pois apresentam falhas complementares entre eles, podendo melhorar um resultado final.
- Fazer a comparação utilizando diferentes formas do YOLOv8 (YOLOv8n, YOLOv8s, YOLOv8m, etc.).

Referências Bibliográficas

Adams, S. J., Henderson, R. D., Yi, X., and Babyn, P. (2021). Artificial intelligence solutions for analysis of x-ray images. *Canadian Association of Radiologists Journal*, 72(1):60–72.

Araujo, L. H., Baldotto, C., Castro Jr, G. d., Katz, A., Ferreira, C. G., Mathias, C., Mascarenhas, E., Lopes, G. d. L., Carvalho, H., Tabacof, J., et al. (2018). Lung cancer in brazil. *Jornal Brasileiro de Pneumologia*, 44:55–64.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature medicine*, 25(6):954–961.

Baccouche, A., Garcia-Zapirain, B., Olea, C. C., and Elmaghraby, A. S. (2021). Breast lesions detection and classification via yolo-based fusion models. *Computers, Materials & Continua*, 69(1).

Balakrishnan, V. (2003). All about the dirac delta function (?). *Resonance*, 8(8):48–58.

Bandos, A. I., Rockette, H. E., Song, T., and Gur, D. (2009). Area under the free-response roc curve (froc) and a related summary index. *Biometrics*, 65(1):247–256.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Blandin Knight, S., Crosbie, P. A., Balata, H., Chudziak, J., Hussell, T., and Dive, C. (2017). Progress and prospects of early detection in lung cancer. *Open biology*, 7(9):170070.

Cardoso, M. J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murray, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Zalbagi Darestani, M., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B. S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P. F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M.,

Kirby, J., Cooper, L. A., Roth, H. R., Xu, D., Bericat, D., Floca, R., Zhou, S. K., Shuaib, H., Farahani, K., Maier-Hein, K. H., Aylward, S., Dogra, P., Ourselin, S., and Feng, A. (2022). MONAI: An open-source framework for deep learning in healthcare.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

CARVALHO, N. (2022). Detecção automática de nódulos pulmonares utilizando transformers.

Choi, W.-J. and Choi, T.-S. (2013). Automated pulmonary nodule detection system in computed tomography images: A hierarchical block classification approach. *Entropy*, 15(2):507–523.

Cui, S., Ming, S., Lin, Y., Chen, F., Shen, Q., Li, H., Chen, G., Gong, X., and Wang, H. (2020). Development and clinical application of deep learning model for lung nodules screening on ct images. *Scientific reports*, 10(1):13657.

Degnan, A. J., Ghobadi, E. H., Hardy, P., Krupinski, E., Scali, E. P., Stratchko, L., Ulano, A., Walker, E., Wasnik, A. P., and Auffermann, W. F. (2019). Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. *Academic radiology*, 26(6):833–845.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Emani, S., Sequist, T. D., Lacson, R., Khorasani, R., Jajoo, K., Holtz, L., and Desai, S. (2019). Ambulatory safety nets to reduce missed and delayed diagnoses of cancer. *The Joint Commission Journal on Quality and Patient Safety*, 45(8):552–557.

Ferreira, J. R., Oliveira, M. C., and de Azevedo-Marques, P. M. (2018). Characterization of pulmonary nodules based on features of margin sharpness and texture. *Journal of digital imaging*, 31:451–463.

Firmino, M., Angelo, G., Morais, H., Dantas, M. R., and Valentim, R. (2016). Computer-aided detection (cade) and diagnosis (cadx) system for lung cancer with likelihood of malignancy. *Biomedical engineering online*, 15(1):1–17.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Halder, A., Dey, D., and Sadhu, A. K. (2020). Lung nodule detection from feature engineering to deep learning in thoracic ct images: a comprehensive review. *Journal of digital imaging*, 33(3):655–677.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Jiang, Y., Chang, S., and Wang, Z. (2021). Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758.

Jocher, G., Chaurasia, A., and Qiu, J. (2023). YOLO by Ultralytics.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., and Yang, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012.

Li, Y. and Fan, Y. (2020). Deepseed: 3d squeeze-and-excitation encoder-decoder convolutional neural networks for pulmonary nodule detection. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1866–1869. IEEE.

Liang, J., Ye, G., Guo, J., Huang, Q., and Zhang, S. (2021). Reducing false-positives in lung nodules detection using balanced datasets. *Frontiers in Public Health*, 9:671070.

Lima, L. L. d. et al. (2019). Modelo computacional para classificação de nódulos pulmonares utilizando redes neurais convolucionais.

McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., Tridandapani, S., and Auffermann, W. F. (2018). Deep learning in radiology. *Academic radiology*, 25(11):1472–1480.

Mitchell, T. M. (1997). *Machine Learning*. McGraw Hill.

Montalbo, F. J. P. (2020). A computer-aided diagnosis of brain tumors using a fine-tuned yolo-based model with transfer learning. *KSII Transactions on Internet & Information Systems*, 14(12).

Mortada, M. J., Tomassini, S., Anbar, H., Morettini, M., Burattini, L., and Sbrollini, A. (2023). Segmentation of anatomical structures of the left heart from echocardiographic images using deep learning. *Diagnostics*, 13(10):1683.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nasrullah, N., Sang, J., Alam, M. S., Mateen, M., Cai, B., and Hu, H. (2019). Automated lung nodule detection and classification using deep learning combined with multiple strategies. *Sensors*, 19(17):3722.

Neal Joshua, E. S., Bhattacharyya, D., Chakkravarthy, M., and Byun, Y.-C. (2021). 3d cnn with visual insights for early detection of lung cancer using gradient-weighted class activation. *Journal of Healthcare Engineering*, 2021:1–11.

Niu, C. and Wang, G. (2022). Unsupervised contrastive learning based transformer for lung nodule detection. *Physics in Medicine & Biology*, 67(20):204001.

OMS (2020). Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000-2019. `who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death`. Accessed: 2022-05-16.

Padilla, R., Netto, S. L., and Da Silva, E. A. (2020). A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, pages 237–242. IEEE.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Patz, E. F., Pinsky, P., Gatsonis, C., Sicks, J. D., Kramer, B. S., Tammemägi, M. C., Chiles, C., Black, W. C., Aberle, D. R., Team, N. O. M. W., et al. (2014). Overdiagnosis in low-dose computed tomography screening for lung cancer. *JAMA internal medicine*, 174(2):269–274.

Qureshi, R., RAGAB, M. G., ABDULKADER, S. J., ALQUSHAIB, A., SUMIEA, E. H., Alhussian, H., et al. (2023). A comprehensive systematic review of yolo for medical object detection (2018 to 2023). *Authorea Preprints*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., and Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788.

Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rong, R., Sheng, H., Jin, K. W., Wu, F., Luo, D., Wen, Z., Tang, C., Yang, D. M., Jia, L., Amgad, M., et al. (2023). A deep learning approach for histology-based nucleus segmentation and tumor microenvironment characterization. *Modern Pathology*, 36(8):100196.

Safdar, M. F., Alkobaisi, S. S., and Zahra, F. T. (2020). A comparative analysis of data augmentation approaches for magnetic resonance imaging (mri) scan images of brain tumor. *Acta informatica medica*, 28(1):29.

Santos, C., Aguiar, M., Welfer, D., and Belloni, B. (2022). A new approach for detecting fundus lesions using image processing and deep neural network architecture based on yolo model. *Sensors*, 22(17):6441.

Setio, A. A. A., Traverso, A., De Bel, T., Berens, M. S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M. E., Geurts, B., et al. (2017). Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13.

Shah, A. A., Malik, H. A. M., Muhammad, A., Alourani, A., and Butt, Z. A. (2023). Deep learning ensemble 2d cnn approach towards the detection of lung cancer. *Scientific Reports*, 13(1):2987.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802.

Shaukat, F., Raja, G., and Frangi, A. F. (2019). Computer-aided detection of lung nodules: a review. *Journal of Medical Imaging*, 6(2):020901–020901.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Terven, J., Córdova-Esparza, D.-M., and Romero-González, J.-A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716.

Ultralytics (2023). Ultralytics yolov8 docs. `https://docs.ultralytics.com/usage/python/`. Accessed: 2023-11-22.

Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). scikit-image: image processing in python. *PeerJ*, 2:e453.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Wang, H., Zhu, Y., Adam, H., Yuille, A., and Chen, L.-C. (2021). Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474.

Wisnivesky, J. P., Yankelevitz, D., and Henschke, C. I. (2005). Stage of lung cancer in relation to its size: part 2. evidence. *Chest*, 127(4):1136–1139.

Zandie, R. and Mahoor, M. H. (2023). Topical language generation using transformers. *Natural Language Engineering*, 29(2):337–359.

Zeng, P., Liu, S., He, S., Zheng, Q., Wu, J., Liu, Y., Lyu, G., and Liu, P. (2023). Tuspm-net: A multi-task model for thyroid ultrasound standard plane recognition and detection of key anatomical structures of the thyroid. *Computers in Biology and Medicine*, page 107069.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. (2020). Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000.

Zhou, J., Zhang, B., Yuan, X., Lian, C., Ji, L., Zhang, Q., and Yue, J. (2023). Yolo-cir: The network based on yolo and convnext for infrared object detection. *Infrared Physics & Technology*, 131:104703.

Zhu, X., Wang, X., Shi, Y., Ren, S., and Wang, W. (2022). Channel-wise attention mechanism in the 3d convolutional network for lung nodule detection. *Electronics*, 11(10):1600.