Otimização de Modelos de Visão Computacional via Quantização para Detecção de Pólipos em Tempo Real

Davi de Jesus Teixeira[1], Carlos Eduardo Gonçalves de Oliveira[2], Gustavo Novack Viana Lima[1], Rian de Souza Santos[1], Ricardo Augusto Pereira Franco[3]

[1]Centro de Excelência em Inteligência Artificial (CEIA)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás, Brasil

[2]Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás, Brasil

[3]Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Goiás, Brasil

dvjteixeira@gmail.com, carlosedgonc@gmail.com,

novack@discente.ufg.br, rian.souza@discente.ufg.br,

ricardofranco@ufg.br

Abstract. *This work evaluates YOLO model variants for real-time colorectal polyp detection, with and without quantization. The models were trained and evaluated on a dataset of colonoscopy images on a GPU, adopting 60 FPS as the real-time threshold. Quantization increased inference speed between 20.1% and 121.4% across all model variants, with a minor impact on metrics. YOLOv8m FP16 and YOLOv11m FP16 presented the best balance between accuracy and inference speed: mAP@0.5 of 0.904 and 0.903, Recall of 0.916 and 0.921, at 96.9 and 83.0 FPS, respectively. The results demonstrate that FP16 quantization is a safe and effective optimization strategy for real-time polyp detection.*

Resumo. *Este trabalho avalia variantes do modelos YOLO para a detecção de pólipos colorretais em tempo real, com e sem quantização. Os modelos foram treinados e avaliados em um conjunto de imagens de colonoscopia em GPU, adotando 60 FPS como limiar de tempo real. A quantização aumentou a velocidade de inferência entre 20,1% e 121,4% em todas as variantes do modelo, com pequeno impacto nas métricas. Os YOLOv8m FP16 e YOLOv11m FP16 apresentaram o melhor equilíbrio entre acurácia e velocidade de inferência: mAP@0.5 de 0,904 e 0,903, Recall de 0,916 e 0,921, a 96,9 e 83,0 FPS, respectivamente. Os resultados demonstram que a quantização FP16 é uma estratégia de otimização segura e eficaz para a detecção de pólipos em tempo real.*

1. Introdução

O câncer colorretal (CCR) é uma doença heterogênea que se desenvolve a partir de mutações genéticas em lesões benignas [Instituto Nacional de Câncer (INCA) 2023].

Globalmente, estima-se mais de 1,9 milhão de casos e 904 mil mortes por CCR ao ano, tornando-o o terceiro câncer mais incidente e o segundo em mortalidade [Bray et al. 2024]. No Brasil, a projeção aponta 45.630 casos anuais no triênio 2023–2025, colocando a doença em terceiro lugar entre os cânceres mais incidentes no país. [Instituto Nacional de Câncer (INCA) 2023].

Apesar de invasivo, o exame de colonoscopia, ao permitir a identificação e remoção de adenomas (pólipos pré-cancerosos), contribui para reduzir significativamente as taxas de incidência do CCR [Bray et al. 2024]. Uma métrica amplamente utilizada para avaliar a qualidade do exame é a taxa de detecção de adenomas (ADR), para a qual estudos têm demonstrado que valores mais elevados estão associados a menores riscos de CCR e de mortalidade pela doença [Corley et al. 2014]. O *American College of Gastroenterology* (ACG) e a *American Society for Gastrointestinal Endoscopy* (ASGE) tem recomendado um ADR maior ou igual a 35% como limiar mínimo de qualidade [ASGE/ACG 2024].

Contudo, existem dificuldades para a identificação de pólipos ao se realizar o exame de colonoscopia. A variabilidade morfológica entre os pacientes, as diferenças de iluminação ou o contraste entre diferentes exames de colonoscopia, artefatos decorrentes do movimento do equipamento, o desfoco ou mesmo a presença de resíduos podem ser consideradas algumas das dificuldades existentes [Rex et al. 2024]. A Figura 1 apresenta exemplos de pólipos obtidos durante um exame de colonoscopia.

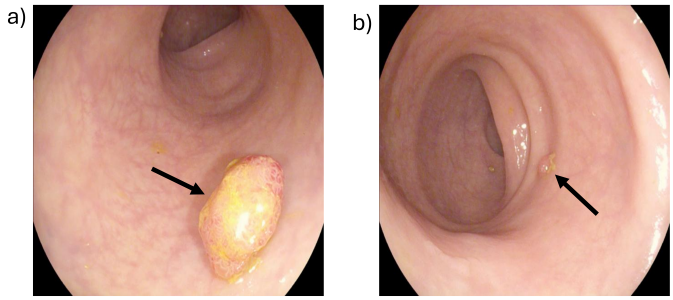

Figura 1. Exemplos de pólipos que podem ser encontrados num exame de colonoscopia [Borgli et al. 2020].

Diante disso, o uso de modelos de Inteligência Artificial (IA) para a detecção de pólipos demonstrou um aumento significativo no ADR e no número médio de pólipos detectados quando comparado ao grupo sem uso de IA em um estudo controlado randomizado [Wang et al. 2019]. Além disso, uma revisão sistemática reportou que há um aumento de aproximadamente 20% no ADR na utilização de IA em exames de colonoscopia [Makar et al. 2025].

Recentemente, avanços na área de IA, sobretudo na subárea de visão computacional, impulsionados pelas redes neurais profundas, contribuíram para a detecção de pólipos. A abordagem de detecção de pólipos *frame* a *frame* é particularmente promissora, pois pode fornecer sinais visuais (como caixas delimitadoras) com latência compatível ao fluxo de operação dos exames de colonoscopia [Jha et al. 2021]. Entre os modelos de IA que podem ser citados para essa abordagem, há a família de modelos *YOLO* (*You Only Look Once*), que se destacam por conciliar alta acurácia e alta velocidade de inferência, características importantes para aplicações em tempo real [Redmon et al. 2016].

Nesse sentido, a quantização de modelos surge como uma técnica importante, permitindo a redução da precisão dos valores numéricos de pesos e ativações das redes neurais e, consequentemente, a largura de banda de memória e a latência, ao custo de impactos mínimos na acurácia [Gholami et al. 2021, Jacob et al. 2018]. Em particular, a precisão FP16 é vantajosa, pois, além de permitir uma melhoria considerável na velocidade de inferência, apresenta impacto mínimo na acurácia do modelo comparado a outras precisões numéricas (ex.: FP32 ou INT8) [Gholami et al. 2021, Jacob et al. 2018]. No contexto prático, para o propósito de quantização, pode-se utilizar o *framework TensorRT*, que, além de permitir a quantização para precisões numéricas reduzidas, é capaz de realizar a fusão otimizada de camadas e a otimização de *kernels*, considerando o hardware-alvo, fatores que em conjunto contribuem para viabilizar a implantação clínica em tempo real [NVIDIA 2024].

Diante do contexto exposto, o objetivo deste trabalho é analisar a viabilidade da detecção de pólipos colorretais em tempo real mediante quantização de modelos da família *YOLO* (YOLOv8, YOLOv9 e YOLOv11) para FP16. Para este fim, são avaliadas as variantes *nano* (n), *medium* (m) e *extra large* (xl) de cada modelo, buscando identificar as configurações que conciliem confiabilidade diagnóstica e desempenho computacional superior a 60 FPS. A adoção desse limiar fundamenta-se no orçamento de latência estabelecido pela literatura clínica: sistemas CADe de colonoscopia devem operar abaixo de 100 ms para ser considerados em tempo real [Kader et al. 2026]. Esse orçamento deve acomodar não apenas a inferência, mas todo o pipeline: pré-processamento, transferência de dados entre CPU e GPU e renderização das caixas delimitadoras. Um requisito de inferência a 30 FPS (33 ms) deixaria margem estreita para esses custos adicionais, com risco de violação do limiar clínico. Adota-se, portanto, 60 FPS (16,7 ms por *frame*) como requisito de inferência, garantindo que o ciclo completo de processamento opere com folga segura abaixo dos 100 ms estabelecidos.

Nesse sentido, este trabalho está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados à detecção de pólipos colorretais com modelos de visão computacional; na Seção 3 é descrita a metodologia experimental proposta; os resultados da comparação entre as variantes dos modelos YOLO com e sem quantização são apresentados e discutidos na Seção 4; por fim, a Seção 5 apresenta as conclusões obtidas.

2. Trabalhos Relacionados

A detecção automática de pólipos colorretais tem sido impulsionada pela contínua evolução das redes neurais profundas. Estudos como o de [Jha et al. 2021], destacaram o *YOLOv4* por seu equilíbrio entre precisão e velocidade, fornecendo base para otimizações subsequentes, como a exploração de *backbones* customizados por [Pacal and Karaboga 2021] e a inclusão de mecanismos de atenção no *YOLOv5* por [Wan et al. 2021]. Mais recentemente, a arquitetura *YOLOv8* foi avaliada por [Lalinia and Sahafi 2024], que identificaram a variante *medium* (*YOLOv8m*) como altamente promissora, alcançando elevadas métricas de detecção de pólipos (precisão de 95.6%, recall de 91.7% e F1-score de 92.4%).

Para além de alta performance diagnóstica, a adequação desses modelos para a prática clínica requer inferência em tempo real utilizando *hardware* com recursos limitados. Assim, [Carrinho and Falcao 2023] demonstraram que o uso do *framework NVIDIA*

TensorRT para reduzir a precisão numérica do *YOLOv4* (para níveis como FP16 e INT8) é capaz de aumentar drasticamente a velocidade de processamento (FPS) sem incorrer em um decréscimo significativo na acurácia (mAP@0.5). Além de viabilizar esse salto de desempenho computacional, o estudo sugere que a técnica de quantização pode atuar como um regularizador benéfico para a generalização do modelo.

Construindo sobre essa base, este estudo agrega à literatura científica ao investigar a viabilidade das variantes *nano*, *medium* e *extra large* de modelos YOLO sob o rigoroso limiar de 60 FPS, focando na otimização via quantização FP16 utilizando o *TensorRT*. Dessa forma, ao incluir arquiteturas mais recentes e bem estabelecidas, como o *YOLOv9* e o *YOLOv11*, os resultados deste trabalho fornecem diretrizes atualizadas para a implementação de IA em exames de colonoscopia em tempo real, considerando uma GPU de médio porte para inferência e garantindo alta precisão diagnóstica com latência minimamente perceptível ao especialista.

3. Metodologia

Este estudo foi conduzido por meio de uma metodologia experimental e sistemática para avaliar e comparar o desempenho computacional e a performance das variantes *nano* (n), *medium* (m) e *extralarge* (xl) dos modelos YOLOv8, YOLOv9 e YOLOv11, com e sem a aplicação de quantização para FP16.

A tarefa definida foi a detecção de pólipos colorretais em tempo real, conforme sumarizado na Figura 2. O processo compreende a preparação e partição do conjunto de dados consolidado, o treinamento e validação das diferentes arquiteturas YOLO, a aplicação de quantização para FP16 e, por fim, a execução da inferência em hardware específico para a extração das métricas de desempenho. Cada uma dessas etapas será detalhada nas subseções a seguir.

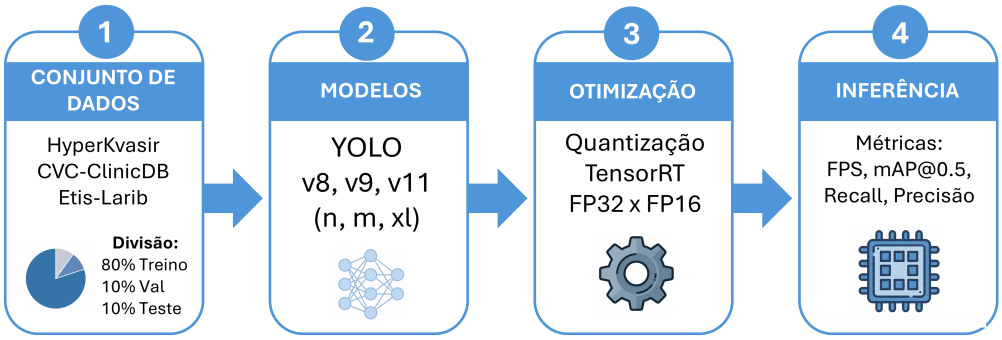

Figura 2. Visão geral da metodologia experimental.

3.1. Conjunto de Dados

Para o treinamento e a avaliação dos modelos, foi construído um conjunto de dados a partir da união de três fontes públicas, amplamente reconhecidas e utilizadas para o *benchmarking* de algoritmos de detecção de pólipos. As fontes selecionadas foram:

- HyperKvasir [Borgli et al. 2020]: Contém 1.000 imagens de pólipos colorretais acompanhadas por máscaras de segmentação e caixas delimitadoras (*bounding bo-*

xes). Diferentemente dos demais conjuntos utilizados, as imagens do HyperKvasir constituem *frames* curados de exames de colonoscopia, selecionados de modo independente em vez de extraídos de sequências contínuas. As imagens foram coletadas no Hospital Bærum, na Noruega, e as anotações foram validadas por gastroenterologistas experientes.

- CVC-ClinicDB [Bernal et al. 2015]: Composto por 612 *frames* extraídos de 29 sequências de vídeo de colonoscopia. Cada imagem possui uma máscara de segmentação de pólipo correspondente.
- ETIS-Larib Polyp DB [Huang et al. 2024]: Consiste em 196 *frames* extraídos de vídeos de colonoscopia. Originado de uma colaboração entre o Hospital Lariboisière (França) e o laboratório ETIS, cada imagem é acompanhada por uma máscara binária de *ground truth* que delimita a região do pólipo.

Ao combinar estas três fontes, obteve-se um *dataset* final com um total de 1.808 imagens únicas. Para a tarefa de detecção de objetos, foram utilizadas as caixas delimitadoras fornecidas pelo *dataset HyperKvasir*. Para os *datasets CVC-ClinicDB* e *ETIS-Larib*, que disponibilizam apenas máscaras de segmentação, as caixas delimitadoras foram geradas a partir das coordenadas extremas de cada máscara.

O particionamento dos dados foi realizado ao nível de imagem, de forma aleatória, na proporção de 80% para treinamento (1.446 imagens), 10% para validação (181 imagens) e 10% para teste (181 imagens).

3.2. Modelos de Detecção

Os modelos investigados foram as variantes *n*, *m* e *xl* dos modelos *YOLOv8*, *YOLOv9* e *YOLOv11* [Jocher et al. 2023]. As variantes *nano*, *medium* e *extra large* foram selecionadas por representarem, respectivamente, os extremos inferior e superior do espectro de complexidade arquitetural e um ponto intermediário equilibrado, permitindo caracterizar o comportamento dos modelos ao longo de toda a curva de *trade-off* entre eficiência computacional e acurácia diagnóstica. As variantes *small* e *large*, por ocuparem posições intermediárias já cobertas por esse recorte, foram omitidas para evitar redundância experimental sem perda de representatividade analítica. Para cada modelo, o processo de treinamento foi inicializado utilizando pesos pré-treinados no *dataset COCO* (*Common Objects in Context*). Subsequentemente, os modelos foram submetidos a um processo de *fine-tuning* com o conjunto de dados de treinamento, especializando-os na tarefa de detecção de pólipos.

3.3. Ambiente Experimental e Treinamento

Todos os experimentos foram conduzidos em um mesmo ambiente computacional para assegurar a consistência das medições.

O treinamento dos modelos foi realizado em uma estação de trabalho equipada com uma GPU NVIDIA RTX 4090 (24 GB de memória RAM), um processador *Intel Core* i9-13900K e 64 GB de memória RAM. As medições de inferência e desempenho foram conduzidas em um hardware distinto do utilizado no treinamento — um notebook equipado com uma GPU NVIDIA GeForce RTX 3070 (8 GB de VRAM), um processador *Intel Core* i7-11800H e 16 GB de memória RAM. Durante a fase de treinamento, as imagens de entrada foram redimensionadas para 640×640 *pixels*, ocorrendo por 100

épocas no total, com um *batch size* de 16. Foi utilizado o otimizador Adam com um *learning rate* de 1×10^{-3} e um *cosine scheduler* para o decaimento do *learning rate*.

Posteriormente à etapa de treinamento, foi realizada a quantização a FP16 dos modelos utilizando-se o *framework NVIDIA TensorRT*. O modelo base (sem quantização) foi mantido à parte para comparações posteriores.

3.4. Métricas de Avaliação

A avaliação dos modelos fundamentou-se em métricas de desempenho computacional e eficácia diagnóstica, permitindo uma análise quantitativa do equilíbrio entre a velocidade de processamento e a precisão na detecção. O desempenho computacional foi mensurado pela taxa de Quadros por Segundo (FPS), definida como a razão entre o total de imagens processadas e o tempo total de execução. Conforme discutido anteriormente, este estudo adota o limiar de 60 FPS como referência para garantir a fluidez necessária em procedimentos intervencionistas no exame de colonoscopia.

Para a avaliação da performance diagnóstica, as predições foram classificadas em Verdadeiros Positivos (TP), Falsos Positivos (FP) e Falsos Negativos (FN), baseando-se no critério de *Intersection over Union* (IoU). O IoU quantifica a sobreposição entre a caixa delimitadora predita (B_{inf}) e a de referência (B_{ref}), conforme expresso na Equação 1:

$$\text{IoU} = \frac{\text{Área}(B_{inf} \cap B_{ref})}{\text{Área}(B_{inf} \cup B_{ref})} \tag{1}$$

Neste trabalho, adotou-se o limiar de IoU ≥ 0.5 para definir uma detecção como TP. A partir dessa classificação, derivam-se a Precisão, que indica a confiabilidade das predições do modelo, e o *Recall*, que reflete a capacidade do sistema em não omitir lesões existentes (métrica de suma importância clínica para a redução da taxa de pólipos não detectados). Tais métricas são calculadas conforme a Equação 2 e a Equação 3:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Por fim, a performance global dos modelos foi sintetizada através do *mean Average Precision* (mAP) com limiar de 0.5 (mAP@0.5). O mAP representa a área sob a curva Precisão-*Recall*, fornecendo uma medida robusta da acurácia média do detector.

4. Resultados e Discussão

Os resultados obtidos, sumarizados na Tabela 1 e ilustrados nas Figuras 3 e 4, revelam um impacto expressivo da quantização FP16 sobre o desempenho computacional dos modelos avaliados, com ganhos de velocidade que variam de +20.1% (YOLOv8n) a +121.4% (YOLOv9xl)[1]

[1]Cabe ressaltar que o YOLOv9 não adota a nomenclatura *nano*, *medium* e *extra large*; para este trabalho, foram selecionadas as variantes de complexidade equivalente às denominações n, m e xl dos demais modelos, sendo elas *YOLOv9t*, *YOLOv9m* e *YOLOv9e*, respectivamente. Por simplicidade, ao longo deste artigo, essas variantes serão referidas como *nano*, *medium* e *extra large*.

Tabela 1. Performance das variantes dos modelos YOLOv8, YOLOv9 e YOLOv11 na GPU NVIDIA RTX 3070 com e sem quantização a FP16.

Variante	Modelo	Quantização	FPS	Precisão	Recall	mAP@0.5
Nano (n)	YOLOv8	Base (FP32)	87.6	**0.930**	**0.906**	0.891
		FP16	**105.2**	**0.930**	**0.906**	0.890
	YOLOv9	Base (FP32)	46.3	0.859	0.895	0.888
		FP16	82.1	0.859	0.890	0.884
	YOLOv11	Base (FP32)	72.5	0.882	0.900	0.888
		FP16	87.8	0.883	**0.906**	**0.894**
Medium (m)	YOLOv8	Base (FP32)	63.7	0.865	0.906	0.892
		FP16	**96.9**	0.875	0.916	**0.904**
	YOLOv9	Base (FP32)	40.6	0.852	0.900	0.880
		FP16	66.4	0.851	0.895	0.876
	YOLOv11	Base (FP32)	61.5	**0.895**	0.895	0.886
		FP16	83.0	0.850	**0.921**	0.903
Extra Large (xl)	YOLOv8	Base (FP32)	38.7	0.828	0.885	0.866
		FP16	**78.7**	0.837	0.890	0.869
	YOLOv9	Base (FP32)	21.5	0.825	0.864	0.839
		FP16	47.6	0.825	0.864	0.839
	YOLOv11	Base (FP32)	36.2	0.837	**0.916**	**0.898**
		FP16	70.6	**0.841**	**0.916**	0.896

Na variante *nano*, os modelos YOLOv8 e YOLOv11 já superam o limiar de 60 FPS mesmo na configuração base FP32, atingindo 87.6 e 72.5 FPS, respectivamente. Com a aplicação da quantização FP16, esses valores sobem para 105.2 FPS (+20.1%) e 87.8 FPS (+21.1%), consolidando-os como opções viáveis para aplicações em tempo real com mínima degradação de acurácia — as variações de mAP@0.5 são inferiores a 0.001 em ambos os casos. O YOLOv9n, por sua vez, com apenas 46.3 FPS na configuração base, não atinge o limiar de viabilidade em FP32, porém com a quantização FP16 obtém 82.1 FPS (+77.3%), tornando-se competitivo, ainda que ao custo de uma leve redução no *Recall* (0.895 para 0.890) e no mAP@0.5 (0.888 para 0.884). Em termos de acurácia diagnóstica na variante Nano, o YOLOv8n destaca-se com a maior Precisão (0.930) e mAP@0.5 (0.890 em FP16), enquanto o YOLOv11n apresenta o melhor *Recall* na configuração FP16 (0.906), o que é particularmente relevante do ponto de vista clínico, dado que falsos negativos implicam diretamente na taxa de pólipos perdidos.

Na variante *medium*, o cenário apresenta nuances importantes. O YOLOv8m e o YOLOv11m atingem marginalmente o limiar de 60 FPS em FP32 (63.7 e 61.5 FPS, respectivamente), mas com a quantização FP16 alcançam desempenhos significativamente superiores, de 96.9 FPS (+52.1%) e 83.0 FPS (+35.0%), respectivamente. Notavelmente, ambos apresentam melhora simultânea nas métricas diagnósticas com a quantização: o YOLOv8m eleva seu mAP@0.5 de 0.892 para 0.904, e o YOLOv11m de 0.886 para 0.903, sugerindo que o processo de otimização via TensorRT contribuiu positivamente para a generalização desses modelos. O YOLOv9m, em contraste, opera em 40.6 FPS na base, abaixo do limiar, e mesmo com FP16 atinge apenas 66.4 FPS (+63.5%), superando o limiar por margem estreita, enquanto apresenta leve queda em Precisão (0.852 para 0.851) e mAP@0.5 (0.880 para 0.876). Dentre todos os modelos e variantes avaliados, o YOLOv8m FP16 (0.904) e o YOLOv11m FP16 (0.903) registraram os maiores valores absolutos de mAP@0.5, posicionando-se como as configurações de melhor equilíbrio

entre eficácia diagnóstica e desempenho computacional.

Na variante *extra large*, nenhum dos três modelos supera o limiar de 60 FPS na configuração base FP32 — o YOLOv8xl opera a 38.7 FPS, o YOLOv11xl a 36.2 FPS e o YOLOv9xl a apenas 21.5 FPS, o menor valor registrado em todo o experimento. Com a quantização FP16, o YOLOv8xl e o YOLOv11xl tornam-se viáveis, atingindo 78.7 FPS (+103.4%) e 70.6 FPS (+95.0%), respectivamente. O YOLOv9xl, apesar do maior ganho relativo de toda a avaliação (+121.4%), alcança apenas 47.6 FPS em FP16, permanecendo abaixo do limiar estabelecido e, portanto, inviável para uso clínico em tempo real na GPU avaliada. Em termos de acurácia, o YOLOv11xl FP16 apresenta o maior mAP@0.5 entre as variantes *extra large* (0.896), com um *Recall* de 0.916, indicando que modelos maiores tendem a ser mais sensíveis à detecção de lesões, porém com custo computacional significativamente maior.

De maneira geral, observa-se que o YOLOv9 apresenta consistentemente o menor desempenho computacional em todas as variantes, independentemente da aplicação da quantização. O YOLOv8 demonstra o melhor desempenho computacional absoluto na configuração base FP32, especialmente nas variantes *nano* e *medium*. Já o YOLOv11 se destaca por apresentar os maiores ganhos relativos de mAP@0.5 com a quantização FP16 nas variantes *nano* e *medium*, além de combinar alto *Recall* com velocidades superiores ao limiar de 60 FPS, o que o posiciona como uma arquitetura de grande potencial clínico.

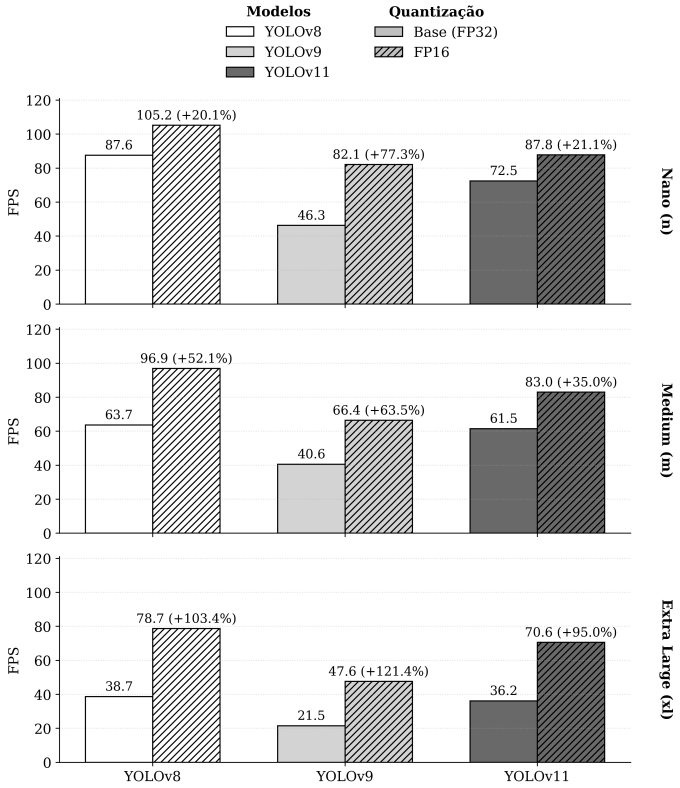

Figura 3. Comparação da velocidade de inferência, medida em FPS, entre os modelos YOLOv8, YOLOv9 e YOLOv11.

A análise de *trade-off* entre mAP@0.5 e FPS, ilustrada na Figura 4, evidencia que

as configurações YOLOv8m FP16 e YOLOv11m FP16 ocupam a região de maior interesse, combinando acurácia diagnóstica superior a 0.900 com velocidades de inferência que garantem ampla margem acima do limiar de 60 FPS adotado neste trabalho.

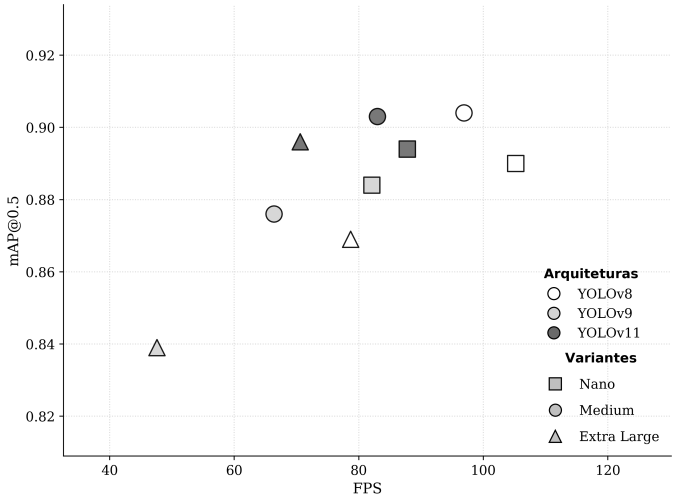

Figura 4. Análise do *trade-off* entre a eficácia diagnóstica (mAP@0.5) e o desempenho computacional (FPS) para as variantes quantizadas em FP16.

A quantização a FP16 via TensorRT demonstrou ser uma estratégia altamente eficaz para a aceleração de inferência em todos os modelos e variantes avaliados, com ganhos que chegam a dobrar ou até mais que dobrar a taxa de FPS em relação à configuração base FP32, como observado no YOLOv8xl (+103.4%), YOLOv11xl (+95.0%) e YOLOv9xl (+121.4%). Esses resultados são consistentes com a literatura, que aponta a redução da precisão numérica de FP32 para FP16 como uma das abordagens mais eficientes para otimização de inferência em GPUs modernas, dada a capacidade dessas arquiteturas de processar operações em meia precisão com *throughput* significativamente superior [Gholami et al. 2021, Jacob et al. 2018].

Um aspecto particularmente relevante observado nos resultados é que, em diversas configurações, a quantização FP16 não apenas manteve as métricas diagnósticas, mas chegou a promover pequenas melhorias no mAP@0.5, como verificado no YOLOv8m (0.892 $\rightarrow$ 0.904), YOLOv11m (0.886 $\rightarrow$ 0.903) e YOLOv11n (0.888 $\rightarrow$ 0.894). Esse comportamento, embora contraintuitivo, pode ser atribuído ao processo de otimização de *kernels* realizado pelo TensorRT, que ao reestruturar as operações de convolução e fusão de camadas para o hardware específico pode introduzir efeitos regularizadores que beneficiam a generalização do modelo [NVIDIA 2024]. Nos casos em que houve queda nas métricas diagnósticas, as variações foram marginais e clinicamente irrelevantes — a maior redução absoluta de mAP@0.5 observada foi de 0.005 pontos no YOLOv9m (0.880 $\rightarrow$ 0.876) —, indicando que a quantização FP16 representa uma troca favorável entre velocidade e acurácia no contexto avaliado.

As três arquiteturas avaliadas — YOLOv8, YOLOv9 e YOLOv11 — apresentaram perfis de desempenho distintos ao longo das variantes analisadas. O YOLOv8 destacou-se por oferecer o maior FPS absoluto na configuração base FP32, especialmente nas variantes *nano* e *medium*, sugerindo uma arquitetura otimizada para inferência efici-

ente mesmo sem quantização. O YOLOv11, por sua vez, demonstrou a melhor relação entre ganho diagnóstico e ganho computacional com a quantização FP16, apresentando melhoras simultâneas em FPS e mAP@0.5 nas variantes *nano* e *medium*, o que pode ser reflexo de uma arquitetura mais recente e melhor adaptada às otimizações do TensorRT. O YOLOv9, em contraste, apresentou o pior desempenho computacional em todas as variantes e configurações, sem compensar essa desvantagem com ganhos expressivos em acurácia diagnóstica — seus valores de mAP@0.5 são sistematicamente inferiores aos do YOLOv8 e YOLOv11 nas variantes correspondentes, o que compromete sua indicação para o cenário clínico avaliado.

Do ponto de vista clínico, os resultados deste trabalho têm implicações diretas para a viabilidade de sistemas de detecção assistida por IA em colonoscopias em tempo real. O limiar de 60 FPS adotado foi estabelecido para garantir que o ciclo de processamento da IA não introduza latência perceptível ao examinador, assegurando que os sinais visuais de detecção — como as caixas delimitadoras — sejam apresentados em sincronia com o fluxo do exame [Jha et al. 2021]. Nesse sentido, as configurações YOLOv8m FP16 (96.9 FPS, mAP@0.5 = 0.904) e YOLOv11m FP16 (83.0 FPS, mAP@0.5 = 0.903) emergem como as mais adequadas para implantação clínica, por oferecerem ampla margem acima do limiar de 60 FPS com acurácia diagnóstica superior a 0.900. A Figura 5 apresenta exemplos visuais de inferências do modelo YOLOv8m.

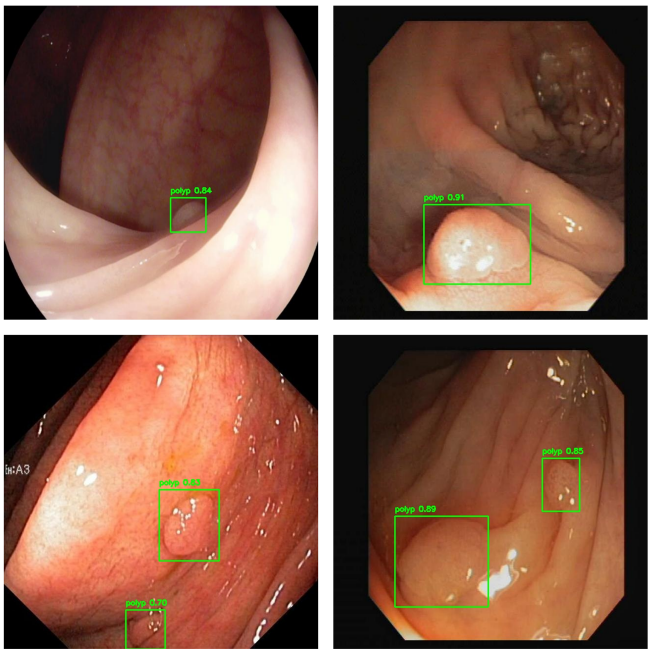

Figura 5. Exemplos visuais de inferência sobre os dados fora do conjunto de treinamento para o YOLOv8m.

A métrica *Recall* merece atenção especial nesse contexto, pois está diretamente associado à taxa de detecção de adenomas (ADR). Falsos negativos — pólipos não detectados pelo sistema — podem resultar em lesões pré-cancerosas não identificadas durante o exame, com potencial impacto na mortalidade por câncer colorretal [Corley et al. 2014]. Os modelos YOLOv11m FP16 (*Recall* = 0.921) e YOLOv11xl FP16 (*Recall* = 0.916)

apresentaram os maiores valores de *Recall* entre as variantes quantizadas sendo, dessa forma, os modelos recomendados para aplicações em tempo real.

Por fim, é importante reconhecer que os experimentos foram conduzidos em hardware específico (GPU NVIDIA GeForce RTX 3070), e que os resultados de FPS podem variar em outros ambientes de execução. Contudo, dado que a RTX 3070 representa um nível de hardware acessível e representativo de estações de trabalho de médio porte, os resultados obtidos sustentam a hipótese de viabilidade clínica em tempo real para as configurações identificadas.

5. Conclusão

Este trabalho avaliou a viabilidade em tempo real de modelos de detecção de pólipos colorretais baseados na família YOLO — YOLOv8, YOLOv9 e YOLOv11 — nas variantes *nano*, *medium* e *extra large*, com e sem a aplicação de quantização a FP16 via TensorRT. Os resultados demonstraram que a quantização FP16 constitui uma estratégia eficaz e segura para acelerar a inferência em todos os cenários avaliados, com ganhos de velocidade entre +20.1% e +121.4%, sem comprometer de forma clinicamente relevante as métricas diagnósticas de Precisão, *Recall* e mAP@0.5.

Dentre as configurações avaliadas, o YOLOv8m FP16 e o YOLOv11m FP16 destacaram-se como as mais adequadas para implantação clínica, por combinarem acurácia diagnóstica superior a 0.900 de mAP@0.5 com velocidades de inferência de 96.9 e 83.0 FPS, respectivamente. O YOLOv11m FP16, em particular, apresentou o maior *Recall* entre as variantes *medium* (0.921), tornando-o especialmente relevante do ponto de vista da segurança diagnóstica. Por fim, os resultados confirmam que arquiteturas YOLO modernas quantizadas representam uma forma viável para sistemas CADe em tempo real, mesmo em *hardware* de médio porte.

6. Agradecimentos

Este trabalho foi parcialmente financiado pelo CNPq e pelo projeto Inteligência Artificial Aplicada na Detecção de Anomalias em Vídeos no Apoio à Tomada de Decisão, apoiado pelo Centro de Excelência em Inteligência Artificial (CEIA), Embrapii, ZSCAN e SEBRAE, com recursos financeiros do processo nº PEIA-2501.0109.

Referências

ASGE/ACG (2024). Asge–acg quality indicators for colonoscopy — faq. Issue Date: Oct 2024.

Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., and Vilariño, F. (2015). Cvc-clinicdb.

Borgli, H., Thambawita, V., Smedsrud, P. H., Hicks, S., and et.al. (2020). HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific Data*, 7(1):283.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., and Jemal, A. (2024). Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74(3):229–263.

Carrinho, P. and Falcao, G. (2023). Highly accurate and fast yolov4-based polyp detection. *Expert Systems with Applications*, 232:120834.

Corley, D. A., Jensen, C. D., Marks, A. R., Zhao, Y., and et.al. (2014). Adenoma detection rate and risk of colorectal cancer and death. *New England Journal of Medicine*, 370:1298–1306.

Gholami, A., Kim, S., Dong, Z., Yao, Z., Mahoney, M. W., and Keutzer, K. (2021). A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.

Huang, C.-H., Wu, H.-Y., and Lin, Y.-L. (2024). Etis-larib polyp db.

Instituto Nacional de Câncer (INCA) (2023). Câncer de cólon e reto — estimativa 2023–2025. Acesso em: 02 set. 2025.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., and et.al. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2704–2713.

Jha, D., Ali, S., Tomar, N. K., Johansen, H. D., Johansen, D., Rittscher, J., Riegler, M. A., and Halvorsen, P. (2021). Real-time polyp detection, localization and segmentation in colonoscopy using deep learning. *IEEE Access*, 9:40496–40510.

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

Kader, R., Hassan, C., Lanas, Á., et al. (2026). A novel cloud-based artificial intelligence for real-time detection of colorectal neoplasia – a randomized controlled trial (EAGLE). *npj Digital Medicine*, 9(84).

Lalinia, M. and Sahafi, A. (2024). Colorectal polyp detection in colonoscopy images using yolo-v8 network. *Signal, Image and Video Processing*, 18(3):2047–2058.

Makar, J., Abdelmalak, J., Con, D., Hafeez, B., and Garg, M. (2025). Use of artificial intelligence improves colonoscopy performance in adenoma detection: a systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 101(1):68–81.e8.

NVIDIA (2024). *TensorRT Developer Guide, Version 10.0.1*.

Pacal, I. and Karaboga, D. (2021). A robust real-time deep learning based automatic polyp detection system. *Computers in Biology and Medicine*, 134:104519.

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 (CVPR)*, pages 779–788.

Rex, D. K., Anderson, J. C., Butterly, L. F., and et al. (2024). Quality indicators for colonoscopy. *Gastrointestinal Endoscopy*, 100(3):352–381.

Wan, J., Chen, B., and Yu, Y. (2021). Polyp detection from colorectum images by using attentive yolov5. *Diagnostics*, 11(12).

Wang, P., Berzin, T. M., Glissen Brown, J. R., and et al. (2019). Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a randomized controlled study. *Gut*, 68(10):1813–1819.