VIDEOLARINGOSCOPIA ASSISTIDA POR INTELIGÊNCIA ARTIFICIAL: AVALIAÇÃO DA YOLOV11 PARA SEGMENTAÇÃO DA GLOTE

MARCELO CLASEN RIBEIRO[1], TIAGO THOMPSEN PRIMO[2], ANA CRISTINA BEITIA KRAEMER MORAES[3], RAFAEL GUERRA LUND[4], MARILTON SANCHOTENE DE AGUIAR[5]

[1] *Universidade Federal de Pelotas – mcribeiro@inf.ufpel.edu.br*
[2] *Universidade Federal de Pelotas – tiago.primo@inf.ufpel.edu.br*
[3] *Universidade Católica de Pelotas – anacristinabkmoraes@gmail.com*
[4] *Universidade Federal de Pelotas – rafael.lund@gmail.com*
[5] *Universidade Federal de Pelotas – marilton@inf.ufpel.edu.br*

1. INTRODUÇÃO

A videolaringoscopia consolidou-se como uma ferramenta indispensável na medicina contemporânea, especialmente em salas de cirurgia e unidades de terapia intensiva (UTI), desempenhando papel crucial na intubação endotraqueal (WU et al., 2025). A utilização desse exame possibilita uma visualização superior das vias aéreas em comparação aos laringoscópios convencionais, oferecendo maior segurança e precisão ao procedimento, além de apoiar profissionais em situações críticas em que o acesso à traqueia apresenta obstáculos.

A falha na intubação logo na primeira tentativa pode elevar o risco de eventos adversos, como aspiração, hipotensão, arritmia ou trauma de tecidos moles (KIM et al., 2013). Nesse cenário, a identificação automática de estruturas anatômicas da laringe, em especial da glote, pode auxiliar o profissional na execução adequada da intubação, ao fornecer informações visuais em tempo real sobre a região de interesse e reduzir a probabilidade de insucessos. Técnicas baseadas em *deep learning* têm demonstrado elevada eficácia no ambiente clínico, atingindo desempenho de estado da arte em tarefas de detecção, classificação e segmentação de imagens médicas (WANG et al., 2021).

Entre as arquiteturas de visão computacional voltadas à detecção e segmentação, a família YOLO (*You Only Look Once*) (REDMON et al., 2015) destaca-se por aliar desempenho competitivo à capacidade de processamento em tempo real. Suas versões mais recentes incorporam a segmentação de instâncias, permitindo não apenas a detecção, mas também o delineamento preciso das áreas de interesse em imagens médicas. No contexto da videolaringoscopia, essa abordagem possibilita a apresentação das estruturas relevantes ao médico durante o exame, sem comprometer a fluidez do procedimento.

O objetivo deste estudo é aplicar e avaliar a rede YOLOv11 (JOCHER; QIU, 2024) na segmentação da glote em imagens de videolaringoscopia. A investigação propõe-se a analisar o desempenho do modelo por meio de métricas amplamente utilizadas na área de segmentação médica, como IoU e Dice, além de mensurar o tempo de inferência, verificando sua viabilidade em aplicações em tempo real.

2. METODOLOGIA

Para o desenvolvimento deste estudo, empregou-se o conjunto de dados BAGLS (*Benchmark for Automatic Glottis Segmentation*) (GÓMEZ et al., 2020), amplamente utilizado em pesquisas sobre análise automática da laringe. O *dataset* é composto por exames de videolaringoscopia provenientes de sete instituições dos Estados Unidos e da União Europeia, totalizando 640 vídeos. Esses vídeos foram divididos em conjuntos de treino e teste, contendo 570 e 70 vídeos, respectivamente. Posteriormente, foram selecionados entre 50 e 100 quadros de cada vídeo,

resultando em 55.750 quadros para treino e 3.500 para teste, os quais foram anotados por três especialistas, originando máscaras binárias da glote.

Os experimentos foram realizados por meio da abordagem de *hold-out validation*, na qual 20% dos quadros do conjunto de treino foram aleatoriamente reservados para validação. Ressalta-se que os quadros do conjunto de teste foram extraídos de vídeos distintos daqueles utilizados no treino ou na validação, assegurando a inexistência de sobreposição entre os conjuntos e prevenindo o vazamento de dados na avaliação final do modelo. A Tabela 1 apresenta a distribuição dos dados nos subconjuntos de treino, validação e teste.

Tabela 1: Distribuição dos quadros nos conjuntos de treino, validação e teste.

Subconjunto	**Treino**	**Validação**	**Teste**
Quadros	44.600	11.150	3.500

Durante o pré-processamento, foi necessário adequar os dados ao formato exigido pela rede. As imagens originais, que apresentavam resoluções distintas, foram ajustadas por meio da técnica de *letterboxing*, preservando a proporção original e adicionando bordas para atingir o tamanho desejado, evitando distorções que poderiam comprometer o desempenho do modelo. As máscaras binárias correspondentes à glote foram convertidas para o padrão requerido pela YOLO, no qual os rótulos são representados por valores inteiros a partir de zero e as estruturas são descritas por polígonos que delimitam seus contornos, garantindo compatibilidade com o processo de treinamento.

A rede de segmentação adotada para os experimentos foi a YOLOv11n-seg, versão recente da família YOLO que incorpora a segmentação de instâncias. Essa arquitetura foi escolhida por aliar desempenho competitivo à elevada eficiência computacional, permitindo inferência em tempo real mesmo em hardwares com recursos limitados. As versões mais atuais da YOLO são disponibilizadas pela Ultralytics, cujos modelos de segmentação são pré-treinados no conjunto COCO-Seg, uma variação do COCO que reúne 330 mil imagens distribuídas em 80 categorias, com anotações detalhadas para segmentação.

A configuração dos hiperparâmetros empregados durante o treinamento seguiu, em grande parte, os valores padrão fornecidos pelos autores, incluindo o otimizador *SGD*, selecionado automaticamente de acordo com as configurações da arquitetura, com taxa de aprendizado de 0,01 e *momentum* de 0,9. Os parâmetros definidos manualmente neste estudo incluem o número de épocas, estabelecido em 50, o uso de *early stopping* configurado com paciência de 10 épocas, a resolução de 640$\times$640 e o *batch size* de 32, determinado de modo a explorar de forma eficiente a capacidade computacional disponível.

Para a avaliação dos resultados, adotaram-se as métricas *Intersection over Union* (IoU) e *Dice Similarity Coefficient* (DSC), amplamente empregadas na área de segmentação de imagens médicas. Ambas foram calculadas a partir das inferências realizadas sobre o conjunto de teste, utilizando um código em *Python* desenvolvido para comparar as máscaras preditas pela rede com as máscaras de referência disponibilizadas no BAGLS.

3. RESULTADOS E DISCUSSÃO

Os experimentos foram realizados em um computador equipado com processador Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz, 32 GB de RAM e uma GPU

NVIDIA TITAN Xp com 12 GB de VRAM. A Tabela 2 apresenta os valores de *IoU* e *Dice* obtidos no conjunto de teste pelo modelo proposto, em comparação com os resultados da U-Net reportados pelos autores do *dataset*. Ressalta-se que os valores de *Dice* e o modelo treinado não foram disponibilizados pelos autores, o que dificulta uma comparação mais precisa entre as arquiteturas.

Tabela 2: Comparação de métricas de segmentação IoU e Dice entre YOLOv11n-seg e U-Net no conjunto de teste.

Modelo	IoU	Dice
U-Net (GÓMEZ et al., 2020)	**0,799**	–
YOLOv11n-seg	0,670	0,766

Ao analisar os resultados, verifica-se que a rede YOLO11vn-seg apresentou desempenho inferior ao modelo U-Net. Contudo, destaca-se que a principal vantagem das redes YOLO reside na capacidade de processamento em tempo real, o que implica um *trade-off* em relação ao desempenho quando comparadas a arquiteturas tradicionais de segmentação de imagens médicas, como a U-Net. Para possibilitar a avaliação desse aspecto, a Tabela 3 apresenta os tempos de inferência dos dois modelos, medidos em milissegundos (ms), tanto em CPU quanto em GPU.

Tabela 3: Comparação do tempo de inferência (ms) entre YOLOv11n-seg e U-Net em GPU e CPU.

Modelo	GPU (ms)	CPU (ms)
U-Net (GÓMEZ et al., 2020)	**24ms**	2.120ms
YOLOv11n-seg	25ms	**474ms**

Ao examinar os tempos de inferência, constata-se que, na GPU, a YOLOv11n-seg apresentou 25 ms, ligeiramente acima dos 24 ms reportados para a U-Net. Cabe salientar que os autores utilizaram uma RTX Titan com *Tensor Cores*, executando a U-Net por meio do *TensorFlow/Keras*, o que provavelmente contribuiu para acelerar a inferência da rede. Em contraste, a YOLOv11n-seg foi avaliada em uma TITAN Xp, sem suporte a *Tensor Cores*, o que indica que a diferença observada decorre, sobretudo, das características do hardware. Na CPU, a YOLOv11n-seg apresentou tempo de inferência de 474 ms, substancialmente inferior aos 2.120 ms da U-Net, evidenciando a eficiência da arquitetura YOLO em cenários com hardware mais limitado.

Apesar de apresentar um IoU médio de 0,670, inferior ao desempenho de 0,799 da U-Net, o tempo médio de inferência obtido na GPU possibilita uma visualização em tempo real de aproximadamente 40 quadros por segundo, velocidade adequada para aplicações clínicas que demandam acompanhamento contínuo das estruturas anatômicas durante o procedimento. Esses resultados evidenciam o *trade-off* clássico entre desempenho e velocidade, demonstrando que a YOLOv11n-seg oferece inferência rápida e eficiente, apropriada para cenários clínicos em tempo real, ainda que isso implique uma discreta perda na precisão da segmentação.

4. CONCLUSÕES

Neste estudo, foi avaliado o desempenho da rede YOLOv11 na tarefa de segmentação da glote em vídeos provenientes de diferentes hospitais. O modelo apresentou métricas satisfatórias, com valores de *IoU* e *Dice* que, embora inferiores

aos reportados em pesquisas anteriores com a U-Net, demonstraram viabilidade prática para aplicações em tempo real, uma vez que o YOLOv11 alcançou taxas de processamento superiores a 40 FPS. Essa característica o distingue de arquiteturas tradicionais, que, apesar de oferecerem maior acurácia, não possibilitam execução eficiente durante exames clínicos. Ressalta-se, entretanto, que a comparação direta entre os modelos é limitada, pois os autores do estudo de referência não disponibilizaram o modelo treinado, impossibilitando a avaliação sob condições idênticas. Assim, os resultados reforçam a importância de considerar não apenas a precisão, mas também a aplicabilidade prática no desenvolvimento de soluções de segmentação em ambientes clínicos. Nesse sentido, este estudo evidencia o potencial de arquiteturas leves para aplicações em tempo real, apontando caminhos promissores para investigações futuras que busquem conciliar eficiência computacional e maior acurácia.

5. AGRADECIMENTOS

O presente trabalho contou com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – Brasil, Código de Financiamento 001. Agradece-se igualmente ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte a esta pesquisa e pelo auxílio financeiro concedido e bolsas de pesquisa (Processos CNPq: 406417/2024-5 e 382388/2025-9).

6. REFERÊNCIAS BIBLIOGRÁFICAS

GÓMEZ, P.; KIST, A. M.; SCHLEGEL, P.; BERRY, D. A.; CHHETRI, D. K.; DÜRR, S.; ECHTERNACH, M.; JOHNSON, A. M.; KNIESBURGES, S.; KUNDUK, M. et al. Bagls, a multihospital benchmark for automatic glottis segmentation. **Scientific data**, Nature Publishing Group UK London, v. 7, n. 1, p. 186, 2020.

JOCHER, G.; QIU, J. **Ultralytics YOLO11**. 2024. Disponível em: <https://github.com/ultralytics/ultralytics>.

KIM, C.; KANG, H. G.; LIM, T. H.; CHOI, B. Y.; SHIN, Y.-j.; CHOI, H. J. What factors affect the success rate of the first attempt at endotracheal intubation in emergency departments? **Emergency Medicine Journal**, British Association for Accident and Emergency Medicine, v. 30, n. 11, p. 888–892, 2013. ISSN 1472-0205.

REDMON, J.; DIVVALA, S. K.; GIRSHICK, R. B.; FARHADI, A. You only look once: Unified, real-time object detection. **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**, p. 779–788, 2015.

WANG, J.; ZHU, H.; WANG, S.-H.; ZHANG, Y.-D. A review of deep learning on medical image analysis. **Mobile Networks and Applications**, Springer, Berlin, Germany, v. 26, n. 1, p. 351–380, 2021.

WU, J.; GUO, W.; CHEN, Z.; HU, H.; LI, H.; ZHANG, Y.; HUANG, J.; LIU, L.; XU, Z.; XU, T.; ZHOU, M.; ZHU, C.; CUI, H.; XU, W.; ZOU, Z. A segmentation network based on cnns for identifying laryngeal structures in video laryngoscope images. **Computerized Medical Imaging and Graphics**, v. 124, p. 102573, 2025. ISSN 0895-6111.