Instance Segmentation in Medical Imaging: A Comparative Study of CNN and Transformer-Based Models in a Teledermatology Study-Case

Rodrigo P. S. Ribeiro[1,2], **Aldo von Wangenheim**[1,2]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC) – Florianópolis – SC – Brazil

[2]LabTelemed – Hospital Universitário
Universidade Federal de Santa Catarina (UFSC) – Florianópolois, SC – Brazil

ribeiro.rodrigo@posgrad.ufsc.br, aldo.vw@ufsc.br

Abstract. *The rapid evolution of instance segmentation models necessitates empirical comparisons to guide their adoption in critical domains like medical imaging. This study evaluates four state-of-the-art architectures—Mask R-CNN, Mask2Former, YOLOv11, and YOLOv12 on a teledermatological dataset annotated for compliance-driven segmentation of rulers and patient information tags. Results demonstrate that transformer-based and hybrid models (Mask2Former, YOLOv11) significantly outperform traditional CNNs in precision-driven metrics (AP75), highlighting their suitability for medical applications. This work provides actionable insights for model selection in healthcare, emphasizing the balance between accuracy and computational efficiency.*

1. Introduction

The integration of artificial intelligence (AI) into medical imaging has revolutionized diagnostic and telemedicine workflows, particularly in resource-constrained domains like teledermatology. A critical challenge in teledermatological image analysis is ensuring compliance with standardized protocols, such as the inclusion of calibration rulers or patient identification tags. Initial efforts in this specific domain employed Mask R-CNN, a convolutional neural network (CNN) based architecture, for segmenting rulers and tags in teledermatological images, achieving an average precision (AP) of 95.23 (AP50) and 75.52 (AP75) on the test set [Ribeiro and von Wangenheim 2024]. While these results were promising, the rapid evolution of instance segmentation models, particularly the emergence of transformer-based architectures and hybrid CNN frameworks, warrants a re-evaluation of state-of-the-art methodologies.

These recent advancements in deep learning have introduced paradigm-shifting architectures. Transformers, initially popularized in natural language processing [Vaswani et al. 2017], have demonstrated exceptional performance in vision tasks by capturing long-range dependencies through self-attention mechanisms [Dosovitskiy et al. 2020]. Concurrently, iterations of YOLO (You Only Look Once) have optimized the trade-off between speed and accuracy, with YOLOv11 and YOLOv12 incorporating advanced attention mechanisms and scaled architectures to improve accuracy [Khanam and Hussain 2024, Jocher et al. 2024]. These innovations address critical limitations of earlier models, such as Mask R-CNN's reliance on region proposals, which can struggle with small objects or small overlapping instances [He et al. 2017]. For instance, Mask2Former, a transformer-based model, has demonstrated state-of-the-art results in instance segmentation across diverse benchmarks [Cheng et al. 2022]. These developments are particularly relevant to medical imaging, where precision at high Intersection over Union (IoU) thresholds (e.g., AP75) is vital for diagnostic reliability; even marginal gains in segmentation accuracy could reduce manual oversight and improve diagnosis and compliance workflows [Litjens et al. 2017, Isensee et al. 2021, Esteva et al. 2021]. While transformer-based models excel in capturing global context, their computational overhead raises concerns about scalability [Dosovitskiy et al. 2020]. Conversely, CNN-based architectures like YOLO prioritize speed but risk undersegmenting complex structures. By evaluating these trade-offs, this study provides actionable insights for selecting models tailored to medical imaging's unique demands.

2. Objectives

This study revisits the original experiment on teledermatological compliance task [Ribeiro and von Wangenheim 2024], benchmarking Mask R-CNN against Mask2Former, YOLOv11, and YOLOv12 using the same dataset and metrics, aiming to:

- Benchmark performance disparities between CNN, Hybrid and transformer-based models.
- Provide actionable insights for model selection in medical imaging pipelines.

By doing so, we answer three questions:

- (1) Do modern architectures significantly outperform Mask R-CNN in high-precision segmentation?
- (2) Which model best balances AP50 (broad detection) and AP75 (strict localization) for medical imaging and compliance automation?
- (3) What paradigm offer the best trade-off between precison and resources?

We also contribute to two general gaps:

- comparative analyses between CNN, hybrid and transformer-based models in medical imaging with a real-world study case;
- the need to validate whether newer architectures can surpass traditional CNNs in clinically relevant metrics (AP75).

By evaluating these models on a standardized dataset, this work provides empirical insights into the evolving landscape of instance segmentation, with implications for clinical deployment across resource-constrained settings, emphasizing robustness under stringent IoU thresholds.

3. Material and Methods

3.1. Dataset

To compare performance between the models, we employed a Lesion Info Tag and Ruler Dataset used in the teledermatology of Santa Catarina State Integrated Telemedicine and Telehealth System (STT/SC) for automatic protocol assesment [Ribeiro and von Wangenheim 2024]. This dataset contains 14,238 annotated images distributed in 2 classes used to identify rulers and patient information tags in teledermatological images. The data was split into a training set (70%), validation set (15%), and test set (15%). The distribution of labeled data is provided in Table 1.

	Train	**Validation**	**Test**	**Total**
Labeled as Tag	10,107	2,180	2,178	14,465
Labeled as Ruler	10,040	2,150	2,139	14,329
Qty. images	9,966	2,136	2,136	14,238

Table 1. Labels and image distribution across subsets.

3.2. Models and Setup

All models were trained using the PyTorch framework on a single NVIDIA H100 80GB GPU. With the exception of the YoloV12 model (which at the time of this experiment had no pre-training data for segmentation), all other models used pre-training on the ImageNet [Deng et al. 2009] or COCO dataset [Lin et al. 2014]. In order to reduce training time, we employed a method of super-convergence known as One Cycle, in which the learning rate gradually grows to an upper limit and then starts to decay tending to zero instead of being fixed all the time and the network momentum varies with the learning rate [Smith and Topin 2019]; other specific hyperparameters of the training of each model are described in Table 2. Average Precision (AP) metrics were used to assess segmentation accuracy at IoU thresholds of 0.50 and 0.75, reflecting moderate and strict segmentation criteria [Padilla et al. 2020]. For the evaluation of the results we used the official COCOAPI[1] for python to generate the metrics. For the inference time and VRAM allocation, in order to better reflect small-scale implementation, we used a mid-tier consumer-grade GPU: NVIDIA RTX 3060 12GB.

	MaskRCNN	Mask2Former	Yolo11	Yolo12
Epoch	50	50	50	50
Batch	48	16	160	160
Input Size	800-1333	800-1333	640	640
Backbone	ResNet50	Swin-Baseline	Yolo11-n	Yolo12-n
Optimizer	SGD	AdamW	AdamW	AdamW
Learning Rate	Adaptative	Adaptative	Adaptative	Adaptative
Starting LR	0.00005	0.00005	0.00005	0.00005
Upper bound LR	0.0001	0.0001	0.0001	0.0001
LR Policy	OneCycle	OneCycle	OneCycle	OneCycle
Momentum	0.85∼0.90	0.85∼0.90	0.85∼0.90	0.85∼0.90
Pre-trained	ImageNet+COCO	COCO	COCO	None
Arch	CNN	Transformer	Hybrid	Hybrid
Task	Instance Seg.	Instance Seg.	Instance Seg.	Instance Seg.
Year	03/2017	06/2022	11/2024	02/2025

Table 2. Training specification and models information

4. Results

The experiment compared four models: Mask R-CNN (CNN), Mask2Former (transformer), YOLO11, and YOLO12—using identical datasets and evaluation protocols. The

[1] https://github.com/cocodataset/cocoapi

results for the AP50 metric are shown in Table 3, and the results for the AP75 metric can be seen in Table 4.

	MaskRCNN	Mask2Former	Yolo11	Yolo12
Validation	96.01	96.90	97.00	96.60
Test	96.33	96.57	97.38	96.70

Table 3. Comparative of AP scores across models with treshold $\geq$ 0.50

	MaskRCNN	Mask2Former	Yolo11	Yolo12
Validation	78.48	84.18	85.31	82.6
Test	80.35	84.55	86.08	83.03

Table 4. Comparative of AP scores across models with treshold $\geq$ 0.75

	MaskRCNN	Mask2Former	Yolo11	Yolo12
Allocated VRAM	1.8Gb	11.7Gb	1.3Gb	1.3Gb
Full Inference Time	498ms	1589ms	566ms	699ms

Table 5. GPU VRAM allocation and inference time in milliseconds of all models in a 3024x4032 pixels color image with NVidia RTX 3060 12Gb

4.1. Performance Analysis at IoU Limits

Analyzing lenient metrics and more restrictive metrics is crucial for a good understanding of how the model behaves in a given task and to identify gaps for improvement. Our experiment shows that all models surpassed the original Mask R-CNN baseline, particularly in AP75 (Table 4). YOLO11 emerged as the top performer, achieving test AP50/AP75 of 97.38/86.08, followed by Mask2Former (96.57/84.55). Notably, even the updated training of Mask R-CNN showed gains (AP75: +4.83), likely due to implementation refinements, but was still outperformed.

Lenient Metric (AP50)

Our study implies that all models excel at coarse segmentation (**AP50**), where the tested models achieved high scores ($>$ 95% on both sets), with marginal differences (Table 3). YOLO11 led (test AP50: 97.38), followed by Mask2Former (96.57) and YOLO12 (96.70). Mask R-CNN improved slightly over its original implementation (test AP50: 96.33 vs. 95.23). This indicated that in a situation where maximum precision of segmentation is not required, the models perform very closely without a significant advantage. The stable scores across validation and test sets of all tested models, reiforce the similar capability of these models on a less restricted metric.

Strict Metric (AP75)

The analysis of the **AP75** (Table 4) demonstrated significant disparities. YOLO11 outperformed all models (test AP75: 86.08), followed by Mask2Former (84.55) and YOLO12 (83.03). Mask R-CNN lagged (80.35), despite a 4.8 point improvement over its prior iteration. Highlighting a significant advantage in instance segmentation accuracy for newer and hybrid models. The AP75 metric, which demands tighter segmentation boundaries, is clinically significant, and the gap between AP50 and AP75 emphasizes a critical consideration: while high AP50 satisfies basic compliance checks, AP75 aligns better with clinical rigor. For instance, misaligned ruler annotations could skew lesion measurements, leading to diagnostic inaccuracies. A ruler segmented at IoU 0.75 guarantees its scale is fully visible, avoiding cropping errors that could invalidate lesion size calculations. YOLO11's 86.08 AP75 represents a 14% reduction in segmentation errors compared to the original Mask R-CNN, potentially enhancing medical image analysis reliability and suggesting it misses fewer edge cases, reducing the need for manual correction.

4.2. Model Architecture Analysis

Understanding how components of a model affect the outcome is essential for making decisions about which model to use when considering efficiency and preicison limits.

Our analysis indicates that Yolo11's leading performance likely stems from its hybrid architecture, combining efficient depth-wise convolutions with attention modules, anchor-free detection and advanced feature pyramids [Khanam and Hussain 2024, Jocher et al. 2024]. Its AP50 (97.38) also suggests exceptional recall, critical for minimizing false negatives in medical imaging. Such designs enhance feature discrimination in instance segmentation, evident in its 86.08 AP75. The new entry in Yolo series (YOLO12) improved with Area Attention, a new self-attention approach that processes large receptive fields more efficiently, along with an improved feature aggregation module based on ELAN (Residual Efficient Layer Aggregation Network) [Tian et al. 2025], these characteristics helped Yolo12 reach a state-of-the-art score even without pre-trained data, this lack of pre-trained data can also explain the slightly lower scores compared to Yolo11, though it remains superior to Mask R-CNN, suggesting room for improvements while also emphasizing the value of transfer learning in medical imaging, where datasets are often smaller and more heterogeneous [Esteva et al. 2021]. Mask R-CNN's lower AP75 (80.35) reflects its reliance on region proposal networks (RPNs), which introduce localization errors during proposal generation [He et al. 2017].

Along with Yolo's dominance in AP75, Mask2Form highlights the value of self-attention mechanisms in capturing fine-grained spatial relationships. By unifying mask classification via pixel-level transformer decoders [Cheng et al. 2022], Mask2Former mitigates fragmented masks common in complex backgrounds, critical for medical image segmentation where overlapping edges occur. This characteristic also helps in segmenting small, irregularly shaped objects, a task in which normally is a challenge for CNNs such as MaskRCNN, which prioritize local features through convolutional filters. This aligns with studies showing transformers superiority in pixel-level tasks [Carion et al. 2020].

Yolo11 and Yolo12 demonstrated unmatched efficiency, using 1.3GB VRAM and achieving inference times of 566ms and 699ms, respectively. In contrast, Mask2Former required 11.7GB VRAM and 1589ms—prohibitively high for clinics using mid-tier hardware (Table 5). This efficiency stems from YOLO's CNN backbone, which employs strided convolutions and spatial pyramid pooling to reduce computational redundancy [Khanam and Hussain 2024, Tian et al. 2025]. Mask2Former's transformer-based attention mechanisms, while powerful, incur quadratic memory costs with high-resolution inputs [Cheng et al. 2022]. Mask R-CNN, despite being outperformed by YOLO11, showed balanced metrics and moderate resource usage (VRAM 1.8 GB) and speed (498 ms), retaining relevance in environments prioritizing stability over peak performance.

4.3. Results Discussion

The superior performance of pretrained models (YOLO11, Mask2Former) highlights the importance of leveraging large-scale datasets like COCO and ImageNet. Pretraining imbues models with generalized feature extraction capabilities, mitigating overfitting in specialized medical tasks and reducing training time [Esteva et al. 2021]. YOLO12 respectable performance despite training from scratch suggests that newer architectures are increasingly data efficient, though they can still benefit from foundational knowledge transfer.

The narrowing gap between AP50 and AP75 (e.g., YOLO11: $\triangle$11.3 vs. Original Experiment Mask R-CNN: $\triangle$19.71) indicates that newer models excel not just in detection but in geometric accuracy. This aligns with medical imaging requirements, where both localization and boundary precision are of great importance.

Despite improvements, all models exhibit a 10–12 point drop between AP50 and AP75, suggesting residual challenges in mask refinement for maximum precision. Potential factors can include annotation noise (e.g., ambiguous annotation) or rare subclasses (e.g., the same object with various shapes and colors).

The marked improvement in AP75 scores (e.g., +10.56 for YOLO11 over the original Mask R-CNN) is clinically significant. In teledermatology, precise ruler segmentation ensures accurate lesion measurement. Beyond dermatology, these findings advocate for transformer-based or hybrid CNN models in tasks requiring pixel-level precision, such as tumor segmentation in radiology. Also, Yolo11 balanced speed (566 ms) and precision, making it ideal for high-throughput clinics.

4.4. Limitations

Dermatology and teledermatology limitations:

- The lack of a similar external dataset prevented cross-validation.
- Demographic biases in the dataset were not addressed. Such biases could lead to inequitable performance across patient populations, a critical ethical issue in medical AI that remains unexplored.

Others limitations to be considered when extrapolating to different medical imaging fields:

- Instance segmentation on different medical field may impose different challenges and approaches.
- Some medical field may use complementary metrics such as Dice score, sensitivity, or specificity for a more holistic view of performance, particularly in error-sensitive medical contexts.

While the study provides valuable insights into the performance of modern instance segmentation models in teledermatology, these limitations emphasize the need for more comprehensive evaluations. Further experimentations should prioritize diverse datasets, external validation, quantization techniques to optimize transformer models for clinical hardware, and ethical audits to bridge the gap between technical performance and real-world clinical utility. Addressing these gaps will enhance the trustworthiness and applicability of AI-driven tools in medical imaging.

5. Conclusion

This study advances the field of medical imaging by empirically validating the suitability of modern architectures for teledermatology and beyond. YOLO11 emerges as the optimal choice, delivering state-of-the-art precision (AP75: 86.08) while operating efficiently on consumer-grade GPUs. Its success underscores that CNN-hybrid based models, when optimized for both accuracy and speed, remain indispensable in healthcare — a domain where resource constraints and diagnostic accuracy are crucical.

The findings challenge the assumption that transformer-based models universally outperform CNNs in medical tasks. While Mask2Former achieves respectable AP75 (84.55), its computational demands render it impractical for widespread deployment. This dichotomy highlights the need for context-aware model selection: transformers may excel in data-rich, compute-unconstrained environments, but CNNs like YOLO11 offer a pragmatic balance for real-world medical applications.

Our main contributions include but aren't limited to:

- **Empirical Validation of Architectural Advances**: YOLO11 and Mask2Former achieved state-of-the-art AP75 scores (86.08 and 84.55, respectively), highlighting their suitability for medical applications where boundary accuracy is crucial.
- **Framework for Model Selection**: The results advocate for transformer-based or hybrid models in medical imaging pipelines, balancing accuracy and efficiency, affirming the potential of attention mechanisms in medical tasks.
- **Clinical Deployment**: For real-time applications (e.g., teledermatology compliance checks), YOLO11's low latency and VRAM usage are advantageous. For offline tasks requiring pixel-perfect accuracy, Mask2Former remains viable if infrastructure permits.
- **Clinical Relevance**: Improved segmentation reliability directly enhances automated compliance systems, reducing manual oversight and standardizing teledermatological assessments.
- **Broader Implications for Medical AI**: The findings extend beyond dermatology, informing model selection in radiology, pathology, and endoscopic imaging, where efficiency and precision are equally crucial.

While the study focused on a teledermatology dataset, its implications can be extended to broader medical imaging (e.g., radiology, histopathology) where similar tradeoffs between precision and efficiency exist. The findings underscore the importance of adopting cutting-edge architectures in healthcare AI. As the field evolves, continuous benchmarking against clinical metrics (e.g., AP75) will ensure that technological advancements translate to tangible improvements in patient care.

As healthcare systems increasingly adopt AI-driven compliance pipelines, model selection must prioritize architectures that balance precision and generalizability. This study support the use of transformer-based models in scenarios demanding high IoU thresholds, while acknowledging YOLO's suitability for real-time applications. Future work should explore model lightweighting for edge devices and domain-specific pretraining to further bridge the gap between computational research and clinical deployment.

6. Ethics and Dataset Availability

The dataset used in this study was approved by the ethics committee at UFSC (decision nr. 5.183.897) and patients involved signed a term of consent. The study was performed as part of the regular research and development activities of the Santa Catarina State Integrated Telemedicine and Telehealth System (STT/SC) and thus received implicit support from the Government of the State of Santa Catarina. This work was also supported by Fundação de Amparo a Pesquisa e Inovação do Estado de Santa Catarina (FAPESC). Authors state that there are no conflicts of interest. Since the dataset employed in this study contains images where the identity of the patients could be determined, the dataset can unfortunately not be made available publicly. The trained weights of these models are publicly available at https://codigos.ufsc.br/ribeiro.rodrigo/dermacritic.public .

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. https://arxiv.org/abs/2005.12872.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1280–1289.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., and Socher, R. (2021). Deep learning-enabled medical computer vision. *npj Digital Medicine*, 4(1).

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Isensee, F., Jaeger, P., Kohl, S., Petersen, J., and Maier-Hein, K. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9.

Jocher, G., Qiu, J., and Chaurasia, A. (2024). Ultralytics YOLO 11. https://github.com/ultralytics/ultralytics.

Khanam, R. and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. https://arxiv.org/abs/2410.17725.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft COCO: Common Objects in Context.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

Padilla, R., Netto, S. L., and da Silva, E. A. B. (2020). A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242.

Ribeiro, R. d. P. e. S. and von Wangenheim, A. (2024). Automated image quality and protocol adherence assessment of examinations in teledermatology: First results. *Telemedicine and e-Health*, 30(4):994–1005. PMID: 37930716.

Smith, L. N. and Topin, N. (2019). Super-convergence: very fast training of neural networks using large learning rates. In Pham, T., editor, *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, page 36. SPIE.

Tian, Y., Ye, Q., and Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.