

Quantifying Data Efficiency in YOLO-Based Femur Segmentation for Axial T1-Weighted MRI

Patrick Gerard Sujeeth

Vellore Institute of Technology

I R Praveen Joe

praveen.joe@vit.ac.in

Vellore Institute of Technology

Research Article

Keywords: YOLO, femur segmentation, MRI, instance segmentation, transfer learning, data efficiency, data augmentation, deep learning, musculoskeletal imaging, Dice Similarity Coefficient

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10418013/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantifying Data Efficiency in YOLO-Based Femur Segmentation for Axial T1-Weighted MRI

Patrick Gerard Sujeeth¹ and I R Praveen Joe²

¹ *School of Mechanical Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India*

² *School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India*

Corresponding author: *I R Praveen Joe, School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India. Email: praveen.joe@vit.ac.in*

Abstract

Background: Automated femur segmentation in T1-weighted MRI is clinically important yet annotation-intensive. Most clinical settings lack the large labelled datasets that conventional deep learning models require. Systematic investigation of how performance scales with training data across architecturally distinct YOLO-family models is currently absent from the literature. Extensive research to surpass the current state of art in terms of segmentation accuracy can improve clinical outcomes by automating femur segmentation in MRI.

Methods: Three YOLO-family instance segmentation architectures — YOLOv8m-seg, YOLOv11m-seg, and YOLOv26m-seg — were trained across five dataset sizes (N = 40, 100, 200, 400, 800) using a two-phase transfer learning protocol. Forty T1-weighted axial MRI slices of the femur, acquired on a Siemens 3T scanner, were synthetically expanded via an anatomically validated augmentation pipeline. All models were evaluated on a fixed 100-image synthetic test set using the Dice Similarity Coefficient (DSC), Intersection over Union (IoU), 95th-percentile Hausdorff Distance (HD95), Average Symmetric Surface Distance (ASSD), mAP@0.5, and F1

score. Statistical significance was assessed using Mann–Whitney U tests, Wilcoxon signed-rank tests, bootstrap 95% confidence intervals, and Cohen's d effect sizes.

Results: At $N = 800$, all three architectures converged to a strong and practically almost indistinguishable performance: YOLOv8m-seg (DSC = 0.960 ± 0.114), YOLOv11m-seg (DSC = 0.963 ± 0.091), and YOLOv26m-seg (DSC = 0.963 ± 0.092); all pairwise Mann–Whitney $p > 0.20$ with Cohen's $d < 0.03$. In the data-scarce regime, YOLOv11m-seg was the only architecture exhibiting monotonically increasing DSC across all conditions, achieving acceptable performance (DSC > 0.87) at $N = 100$. YOLOv8m-seg collapsed anomalously at $N = 100$ (DSC = 0.193), and YOLOv26m-seg failed at $N = 40$ (DSC = 0.163). Inference times were 59-66 ms/image (YOLOv8m-seg), 65-67 ms/image (YOLOv11m-seg), and 70-72 ms/image (YOLOv26m-seg).

Conclusions: YOLOv11m-seg offers the best balance of data efficiency, segmentation accuracy, and inference speed in the data-scarce regime. YOLO-family models can achieve high segmentation accuracy with as few as 100 annotated slices when transfer learning and anatomically relevant augmentation are applied. Architecture selection is the dominant variable below $N = 100$ and becomes almost negligible above $N = 400$.

Keywords: YOLO; femur segmentation; MRI; instance segmentation; transfer learning; data efficiency; data augmentation; deep learning; musculoskeletal imaging; Dice Similarity Coefficient

Introduction

Automated bone delineation from medical images has established clinical utility in pre-operative planning, implant sizing, and biomechanical modelling [1]. Femur segmentation from MRI is a representative problem: the anatomy is clinically indispensable, annotation is resource-intensive, and most clinical research units possess far fewer labelled cases than conventional deep learning models require [2].

Femur segmentation from T1-weighted MRI is non-trivial. Signal intensity overlaps between cortical bone and adjacent soft-tissue structures, and the femur's curved geometry across successive axial slices compounds algorithm complexity [3]. Statistical shape models [1] and atlas-based approaches [3] addressed some of this

complexity but require extensive template libraries. Deep learning delivered substantial accuracy improvements, particularly with U-Net [4] and its variants, but typically demands large annotated datasets.

Three architectural advances make this an opportune moment to revisit the data-efficiency question. First, YOLOv11 and YOLOv26 substantially evolved from the original YOLO detection paradigm [5], incorporating improved multi-scale feature fusion and enhanced segmentation heads [6,7,8]. Second, the clinical push toward point-of-care AI demands real-time inference alongside accuracy [2]. Third, YOLO-family models have recently demonstrated effectiveness on structured medical imaging targets including brain tumours [9], knee menisci [10], and patellar height [11].

What is absent from the published literature is a systematic, controlled investigation of how femur segmentation performance scales with labelled training data across architecturally distinct YOLO variants. Published bone segmentation results for MRI either examine large-dataset performance without exploring the low-N regime, or investigate data efficiency without architecture comparison.

This study addresses this gap with the following research questions:

- i. How does increasing labelled training data from $N = 40$ to $N = 800$ images affect femur segmentation accuracy across YOLOv8m-seg, YOLOv11m-seg, and YOLOv26m-seg?
- ii. Which architecture demonstrates superior data efficiency at the smallest N ?
- iii. What is the inference speed cost of architectural advancement?

Methods

Dataset

The base dataset comprised 40 T1-weighted axial MRI slices of the hip acquired from a single healthy adult female volunteer on a Siemens 3T MRI scanner via the IMAIOS e-Anatomy platform [12]. Each slice represents a distinct axial cross-section of the femur at 1265×506 pixel resolution. Detailed imaging parameters were not readily available from the platform and are acknowledged as a limitation. Polygon-based

annotations of the femoral cortex were performed using LabelMe [13], with each slice containing at least one closed polygon delineating the complete femur contour. All augmented images are derived from one anatomy, one scanner, and one imaging protocol; this is explicitly a methods study quantifying data efficiency.

Image preprocessing

All images were pre-processed through a four-step pipeline. (Figure 1)

- i. Percentile intensity normalisation: The image was clipped to the [1st, 99th] percentile range and normalised, with $\varepsilon = 1 \times 10^{-8}$.

$$I_n = \frac{\text{clip}(I - p_1, 0, p_{99} - p_1)}{p_1 - p_1 + \varepsilon} \quad (1)$$

- ii. CLAHE enhancement: Contrast-Limited Adaptive Histogram Equalization [14] (clipLimit = 2.0, tileGridSize = 8×8) to enhance local cortical edge contrast without amplifying noise in homogenous medullary fat regions.
- iii. Geometric normalization: Bilinear resize to 1265 × 506 px with proportional polygon vertex scaling.

$$s_x = \frac{w_{native}}{w_{orig}} \quad (2)$$

$$s_y = \frac{h_{native}}{h_{orig}} \quad (3)$$

- iv. Channel replication: Pre-processed greyscale image was replicated a cross three channels to satisfy COCO-pretrained YOLO backbone input requirements.

[Insert Figure 1]

Data augmentation pipeline

The augmentation pipeline was designed to generate visual diversity while preserving anatomical plausibility. Geometric transforms were applied simultaneously to images and polygon annotations using the Albumentations library [15] with KeypointParams, ensuring pixel-level annotation consistency without additional annotation cost. Each

polygon was flattened into an ordered list of (x, y) keypoints; the composed transform was applied; keypoints were regrouped into polygons (polygons with fewer than three surviving vertices were discarded).

The MRI_Augment pipeline comprised multiple components:

- i. Discrete multi-angle rotation (-45° to $+30^\circ$, $p = 0.80$): 10 pre-defined rotation spans were generated and BORDER_REFLECT padding was used to eliminate black corners.
- ii. Horizontal flip ($p = 0.50$): Flipping the image along the vertical axis to simulate orientation variability.
- iii. Vertical flip ($p = 0.20$): Flipping the image along the horizontal axis to simulate orientation variability between left and right femur.
- iv. Random brightness/contrast (± 0.25 , $p = 0.70$): Simulates inter-scanner intensity variation.
- v. Gaussian noise (variance $\in [0, 400]$, $p = 0.60$): Simulates scanner thermal noise.
- vi. Gaussian blur (kernel $\in [3, 7]$, $p = 0.30$): Simulates partial volume effects.

Experimental conditions and dataset generation

1000 valid augmented images were generated from the 40 pre-processed sources (global seed = 42) and partitioned into a training bank ($n = 800$) and a global test pool ($n = 200$), from which randomized 100 images (fixed seed = 42) were held out as the unknown synthetic test set, excluded from all training, validation, and hyperparameter tuning across every model and experimental setups.

Five experimental conditions were created (Table 1). For Setups 2–5, augmented images were sampled without replacement from the training bank. The 40 original MRI images were included in all setups, with an 80:20 train/validation split applied within each pool.

[Insert Table 1]

Table 1. Experimental dataset-size conditions. The 100-image unknown synthetic test set is identical across all conditions and all models.

Setup	N	Originals	Augmented	Train / Val
1	40	40	0	32 / 8
2	100	40	60	80 / 20
3	200	40	160	160 / 40
4	400	40	360	320 / 80
5	800	40	760	640 / 160

YOLO model architectures

Three instance segmentation variants from the YOLO family were evaluated: YOLOv8m-seg [6], YOLOv11m-seg [7], and YOLOv26m-seg [8]. All used the medium ('m') variant to ensure approximately matched parameter counts. All were initialised from COCO-pretrained checkpoints [16].

Two-phase transfer learning protocol

A two-phase transfer learning protocol was used to prevent catastrophic forgetting of pretrained feature representations.

Phase 1 (Backbone Freezing, 20 epochs): The first 10 backbone layers were frozen with initial learning rate ($lr_0 = 10e-3$) and learning factor ($lrf = 0.10$), without early stopping. This preserves COCO-learned low-level features when as few as 32 training images are available.

Phase 2 (Full Fine-Tuning, 80 epochs): All layers unfrozen with $lr_0 = 10e-4$, $lrf = 0.01$ (cosine annealing), early stopping (patience = 10), and weight decay = $5 \times 10e-4$. All hyperparameters are summarised in Table 2.

[Insert Table 2]

Table 2. Fixed training hyperparameters across all models and dataset-size conditions.

Parameter	Value	Rationale
Image size	1280 × 512	MRI aspect ratio (divisible by 32)
Batch size	4	Tesla T4 VRAM constraint
Phase 1 epochs	20	Head adaptation
Phase 2 epochs	80	Full convergence

lr_0 (Phase 1)	10^{-3}	Head training
lr_0 (Phase 2)	10^{-4}	Conservative fine-tuning
Weight decay	5×10^{-4}	L2 regularisation
Rectangular training	True	Preserves MRI aspect ratio
Global seed	42	Reproducibility

Evaluation metrics

Four segmentation quality metrics were computed per test image by comparing predicted binary masks against ground-truth masks [17]. These metrics were chosen specifically to evaluate the performance of the model as they are overlap and boundary distance-based metrics:

- i. Dice Similarity Coefficient (DSC):

$$DSC = \frac{2 | P \cap G |}{| P | + | G |} \quad (4)$$

Measures the spatial overlap between two areas or volumes. This is one of the most basic and standard metrics used in medical image segmentation evaluations.

- ii. Intersection over Union (IoU):

$$IoU = \frac{| P \cap G |}{| P \cup G |} \quad (5)$$

Measures the size of the intersection divided by the size of the union of the two masks. This is also considered to be a standard metric for object detection.

- iii. 95th-percentile Hausdorff Distance (HD95):

$$HD95(A,B) = P95(H(A,B)) \quad (6)$$

Measures the maximum distance between the predicted boundary and the true boundary, but discards the top 5% of the largest distances (outliers).

- iv. Average Symmetric Surface Distance (ASSD):

$$ASSD = \frac{1}{2} \left(\frac{\sum_{p \in P_s} d(p, G_s)}{|P_s|} + \frac{\sum_{g \in G_s} d(g, P_s)}{|G_s|} \right) \quad (7)$$

Measure the overall contour precision and boundary smoothness by calculating the average of all shortest distances from the edge points of the prediction to the true edge and vice versa.

Detection-level metrics like mAP@0.5, precision, recall, F1 were obtained from YOLO's built-in validation pipeline. Per-image wall-clock inference time (ms) was measured via `model.predict()` on a Tesla T4 GPU, averaged over the 100-image test set.

Statistical analysis

Inter-architecture performance differences were assessed using the Mann–Whitney U test (unpaired comparisons) and the Wilcoxon signed-rank test (paired comparisons on the same test images), both with $\alpha = 0.05$. Cohen's d quantified effect size magnitude (Small: $|d| < 0.5$; Medium: $0.5 \leq |d| < 0.8$; Large: $|d| \geq 0.8$). Spearman ρ quantified the correlation between N and mean DSC per architecture. Bootstrap 95% confidence intervals (1000 iterations) were computed for mean DSC at each (N, model) condition. No multiple-comparison correction was applied given the exploratory nature of the study and all comparisons are evaluated on the same 100 image test set.

Implementation details

All experiments were implemented in Python 3.12 with PyTorch 2.10.0+cu128, CUDA 12.8, Ultralytics 8.4.45, and Albumentations 1.4.24. Training and evaluation were conducted on Google Colab using a Tesla T4 GPU (15.6 GB VRAM). Global seed = 42 was applied to Python random, NumPy, and PyTorch RNGs.

Results

Quantitative segmentation performance

Figure 2 & Table 3 presents the complete quantitative results for all three models across all five experimental setups, evaluated on the 100-image unknown synthetic test set.

[Insert Table 3]

Table 3. Segmentation performance on the 100-image unknown synthetic test set. DSC: Dice Similarity Coefficient; IoU: Intersection over Union; HD95: 95th-percentile Hausdorff Distance (pixels, lower is better); ASSD: Average Symmetric Surface Distance (pixels, lower is better); mAP@0.5: mask mean average precision at IoU threshold 0.50; F1: F1 Score; ms/img: mean per-image inference time. Bold = best value per column. *Poor DSC.

Model	N	DSC ↑	IoU ↑	HD95 ↓	ASSD ↓	mAP@0.5 ↑	F1 ↑	ms/img
YOLOv8-seg	40	0.757	0.669	135.7	30.76	0.904	0.869	65.7
	100	0.193	0.177	76.74	28.52	0.461	0.470	59.0*
	200	0.943	0.916	6.034	1.877	0.984	0.974	63.0
	400	0.948	0.925	4.755	1.540	0.995	0.990	62.9
	800	0.960	0.938	3.971	1.335	0.995	0.994	66.3
YOLOv11-seg	40	0.640	0.524	204.6	50.10	0.712	0.738	66.4
	100	0.873	0.824	50.36	12.03	0.934	0.927	65.3
	200	0.886	0.842	37.77	9.308	0.947	0.933	65.1
	400	0.950	0.925	6.928	1.888	0.983	0.974	66.3
	800	0.963	0.939	3.973	1.301	0.975	0.978	67.4
YOLOv26-seg	40	0.163	0.156	10.70	2.874	0.415	0.386	72.0
	100	0.877	0.847	10.83	3.050	0.928	0.934	70.5
	200	0.936	0.910	6.058	1.902	0.983	0.970	72.4
	400	0.911	0.884	8.067	2.405	0.954	0.954	71.9
	800	0.963	0.939	3.786	1.289	0.993	0.974	72.4

[Insert Figure 2]

Performance as a function of dataset size

YOLOv11-seg is the only architecture exhibiting a monotonically increasing DSC and IOU across all five conditions (Table 4): 0.640 (N = 40), 0.873 (N = 100), 0.886 (N = 200), 0.950 (N = 400), 0.963 (N = 800), confirmed by Spearman $\rho = 1.00$ ($p < 0.0001$). The steepest gain occurs between N = 40 and N = 100 (+36.5%). HD95 mirrors this by falling from 204.6 px at N = 40 to 50.4 px at N = 100, indicating a non-linear improvement characteristic of the model crossing a representational adequacy threshold. Beyond N = 200, the improvement characteristic reduces, signalling saturation. Figure 3 shows the high inter-quartile range at small N and the convergence of distributions at N = 800. Circles represent outliers.

[Insert Figure 3]

YOLOv8-seg exhibited a striking 74.6% collapse from $N = 40$ (DSC = 0.757) to $N = 100$ (DSC = 0.193), confirmed by $mAP@0.5$ (0.461 vs. 0.904). The model recovered fully at $N = 200$ (DSC = 0.943) (Figure 4), then improved smoothly to 0.960 at $N = 800$ (Spearman $\rho = 0.90$, $p = 0.037$). The Spearman $\rho = 0.90$ ($p = 0.037$) captures the overall improvement characteristic of the model, but masks the anomalous behaviour at $N = 100$.

[Insert Figure 4]

YOLOv26-seg collapsed at $N = 40$ (DSC = 0.163, $mAP = 0.415$), with rapid recovery to 0.877 at $N = 100$ (+438.8%). A non-monotonicity appeared at $N = 400$ (DSC = 0.911), dipping below the $N = 200$ value (0.936) before recovering at $N = 800$ (0.963) (Spearman $\rho = 0.90$, $p = 0.037$).

[Insert Table 4]

Table 4. Gain across models and experimental setups. Abs. gain = $DSC_N - DSC_{N-1}$. % Gain relative to the previous DSC. $N = 40$ is the baseline.

Model	N	DSC	Abs. Gain	% Gain
YOLOv8-seg	40	0.757	0	Baseline
	100	0.193	-0.565	-74.6
	200	0.943	0.75	390
	400	0.948	0.005	0.5
	800	0.960	0.013	1.3
YOLOv11-seg	40	0.640	0	Baseline
	100	0.873	0.233	36.5
	200	0.886	0.013	1.5
	400	0.950	0.064	7.2
	800	0.963	0.014	1.4
YOLOv26-seg	40	0.163	0	Baseline
	100	0.877	0.714	438.8
	200	0.936	0.06	6.8
	400	0.911	-0.0258	-2.8
	800	0.963	0.052	5.7

Statistical analysis Results

At N = 40 and N = 100, all pairwise and unpaired performance differences were highly significant (all Mann–Whitney and Wilcoxon $p < 0.001$) with large effect sizes ($|d| > 1.5$ at N = 40; Cohen's $d = 2.19$ for v11 vs. v8 at N = 100; Table 5). At N = 200, YOLOv26-seg and YOLOv8-seg were statistically indistinguishable. At N = 800, no unpaired comparison reached significance. The Wilcoxon paired test detected residual per-image differences at N = 800 for v11 vs v8 and v26 vs v8, but with negligible effect size (Cohen's $d < 0.03$). This is consistent with the nature of the test reflecting the greater statistical power of the paired test to detect small effects. Though statistically detectable, it is practically inconsequential and not evidence against architectural convergence.

[Insert Figure 5]

Bootstrap 95% Cis (1000 iterations) for mean DSC (Figure 5) narrows substantially from N = 40 to N = 800 (YOLOv8 [0.934, 0.979]; YOLOv11 [0.942, 0.979]; YOLOv26 [0.944, 0.978]) reflecting improved accuracy. The three models are nearly indistinguishable at maximum dataset size, confirming statistical convergence. (Figure 6)

[Insert Table 5]

Table 5. Pairwise statistical significance of DSC differences between architectures. MW: Mann–Whitney U test (unpaired); W: Wilcoxon signed-rank test (paired); * $p < 0.05$; ns: not significant; S: Small effect size; L: Large effect size

N	Comparison	MW p-value	Wilcoxon p-value	Cohen's d
40	yolov11 vs yolov26	3.18E-16*	1.56E-13*	1.499888 (L)
	yolov11 vs yolov8	0.00013*	2.86E-06*	-0.43864 (S)
	yolov26 vs yolov8	4.72E-18*	1.26E-14*	-1.86977 (L)
100	yolov11 vs yolov26	0.001125*	9.93E-06*	-0.01313 (S)
	yolov11 vs yolov8	7.43E-25*	3.70E-17*	2.193651 (L)
	yolov26 vs yolov8	2.83E-25*	6.33E-16*	2.087442 (L)
200	yolov11 vs yolov26	1.05E-11*	1.11E-16*	-0.24451 (S)
	yolov11 vs yolov8	7.76E-12*	2.36E-17*	-0.2914 (S)
	yolov26 vs yolov8	0.79657 (ns)	0.659 (ns)	-0.04116 (S)
800	yolov11 vs yolov26	2.97E-08*	4.49E-14*	0.201165 (S)
	yolov11 vs yolov8	0.757253 (ns)	0.562 (ns)	0.011282 (S)
	yolov26 vs yolov8	1.62E-09*	3.91E-16*	-0.18941 (S)

800	yolov11 vs yolov26	0.811702 (ns)	0.804 (ns)	0.0045 (S)
	yolov11 vs yolov8	0.305362 (ns)	0.003038*	0.026419 (S)
	yolov26 vs yolov8	0.202582 (ns)	0.001674*	0.022384 (S)

Inference speed Result

YOLOv8-seg maintained a consistent speed advantage across all N conditions (~59-66 ms/image) relative to YOLOv11-seg (~65-67 ms/image) and YOLOv26-seg (~70-72 ms/image), representing a 4% and 12% advantage for the older architectures respectively. (Table 3)

[Insert Figure 6]

Discussion

Principal findings

The central finding of this study is that architecture choice in the data-scarce regime has quantifiable, practically significant consequences. The cost of choosing the wrong architecture is not merely a performance degradation but, for YOLOv8-seg at $N = 100$, an effective failure. YOLOv11m-seg is the only architecture that improves monotonically across all dataset sizes and is therefore the recommended choice for datasets of 100 - 800 images.

At $N \leq 40$, YOLOv8-seg is the most reliable choice (DSC = 0.757), where YOLOv11-seg yields 0.640 and YOLOv26-seg essentially fails (0.163). This counterintuitive result (the oldest architecture performs best with fewest images) reflects a well-established principle: more expressive models require more data to avoid overfitting, not always offset by richer pretrained initialisations [18]. At $N = 100$, YOLOv11-seg and YOLOv26-seg attain $\text{DSC} \geq 0.85$, while YOLOv8-seg collapses. Our data warrant a precise recommendation: for $N = 100$, do not use YOLOv8-seg. At $N \geq 200$, all three architectures are competitive (Figure 7); the choice depends on whether inference speed (YOLOv8-seg vs YOLOv11-seg ~4% faster) or learning predictability (YOLOv11-seg) is prioritised.

[Insert Figure 7]

YOLOv8-seg anomaly at N = 100

Three non-mutually-exclusive hypotheses are offered for this collapse, explicitly labelled as such since no ablation experiments have been conducted. Hypothesis 1 (Distributional boundary sensitivity): the 60 augmented images in Setup 2 introduce a distributional shift that may affect YOLOv8's feature pyramid design more than YOLOv11's improved neck architecture. Hypothesis 2 (Training instability): training curves show evidence of early catastrophic forgetting; a learning rate of 10^{-4} may be too aggressive at this data quantity. Hypothesis 3 (Early stopping under high variance): with only 20 validation images, the patience = 10 stopping criterion may terminate training before recovery from an early instability spike.

YOLOv26-seg at N = 40

The near-failure of YOLOv26-seg at N = 40 (DSC = 0.163) reflects effective underfitting of COCO-pretrained features to T1 MRI when insufficient fine-tuning gradient is available. The dissociation between mAP = 0.415 (detection-level performance) and DSC = 0.163 (boundary delineation) suggests that YOLOv26 built on YOLOv11 requires substantially more task-specific supervision to calibrate polygon regression than its detection module [19].

Ceiling performance and distributional considerations

All three architectures converging to DSC ≈ 0.96 at N = 800 is impressive, but requires careful interpretation. The test set was drawn from the same augmentation pipeline and source images as training data. What the models have learned is the augmented synthetic distribution derived from one individual's anatomy. Generalisation to different subjects, scanners, or pathological femora is entirely untested. The shrinking DSC standard deviation from N = 40 (SD ≈ 0.27 – 0.36) to N = 800 (SD ≈ 0.09 – 0.11) is a genuine clinical finding: high-variance segmentation is unsafe regardless of mean performance.

Comparison with existing literature

Direct numerical comparison across studies is complicated by differences in imaging protocol, evaluation sets, and anatomical targets. Ambellan et al. [20] demonstrated DSC approaching 0.98 for knee bone segmentation using combined statistical shape

models and CNN features on the Osteoarthritis Initiative dataset. Lerch et al. [21] achieved DSC of 0.95–0.97 for the proximal femur using 3D CNN on 49 subjects. Liu et al. [22] compared segmentation performance of U-Net (0.932 ± 0.066), Attention U-Net (0.954 ± 0.065), U-KAN (0.949 ± 0.090), SAM 2 (0.950 ± 0.035). Montin et al. [23] reported DSC = 0.89 for femur segmentation under data augmentation. Our 2D YOLO results (DSC ≈ 0.96 at $N = 800$) are competitive with these volumetric approaches, with the important caveat that our test set shares its source distribution with training data.

Limitations

This study has four principal limitations. First, the single-subject, single-scanner dataset means that DSC = 0.96 at $N = 800$ reflects performance on synthetically augmented copies of 40 source slices from one healthy female volunteer on one Siemens 3T scanner, not evidence of cross-population generalisability. Second, the test set was drawn from the same augmentation pipeline and source anatomy as training data; a true external validation set is required for any generalisability claim. Third, no external baseline comparison (U-Net, nnU-Net [24]) was performed on the same dataset. Fourth, the $N = 100$ YOLOv8-seg failure mechanism remains mechanistically unresolved, with only hypotheses rather than ablation evidence available.

Conclusions

With the right architecture, YOLO-family models require substantially less labelled data than conventional approaches to produce highly accurate femur segmentation from T1-weighted axial MRI. At $N = 800$, all three architectures achieve strong, statistically comparable performance (DSC ≈ 0.96 , HD95 < 4 px, ASSD < 1.4 px). In the data-scarce regime, architecture choice is the dominant variable.

YOLOv11m-seg is the recommended choice for datasets of 100-800 images: it is the only architecture exhibiting monotonically increasing DSC, attains DSC ≥ 0.85 at $N = 100$, and achieves large effect-size advantages at the critical low- N transition point.

YOLOv8-seg should not be used near $N = 100$ due to its anomalous collapse, but remains useful at $N \leq 40$. YOLOv26-seg is viable from $N = 100$ onward but exhibits unpredictable intermediate non-monotonicity.

Future Direction

These findings rest on a single-subject, single scanner, single-experimental run and synthetic dataset. Multi-subject, multi-site validation, pathological extension, and 3D volumetric integration are the necessary next steps to attain cross-population generalisability and establish clinical applicability.

List of Abbreviations

- **AI:** Artificial Intelligence
- **ASSD:** Average Symmetric Surface Distance
- **CLAHE:** Contrast-Limited Adaptive Histogram Equalization
- **CNN:** Convolutional Neural Network
- **COCO:** Common Objects in Context (dataset)
- **DSC:** Dice Similarity Coefficient
- **GPU:** Graphics Processing Unit
- **HD95:** 95th-percentile Hausdorff Distance
- **IoU:** Intersection over Union
- **mAP:** Mean Average Precision
- **MRI:** Magnetic Resonance Imaging
- **SD:** Standard Deviation
- **YOLO:** You Only Look Once

Declarations

Ethics approval: This study involved archival imaging data from a single healthy adult volunteer. No patient data, protected health information, or clinical interventions were

involved. The research was conducted in accordance with institutional guidelines. No additional institutional review was required.

Consent to participate: Not applicable

Availability of data and materials: Source code will be available upon reasonable request. The 40 original MRI slices were sourced from publicly accessible archival imaging data — IMAIOS e-Anatomy [12], hip MRI. The analytical scripts are available from the author on reasonable request.

Competing interests: The authors declare that they have no competing interests.

Funding: No external funding was received for this study.

Authors' contributions: PGS: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. IRPJ: Supervision, Validation, Writing – Review & Editing. All authors read and approved the final manuscript.

Acknowledgements: The authors acknowledge the IMAIOS e-Anatomy platform for providing access to the archival MRI data used in this study. The authors thank the Google Colab team for providing the computational resources.

References

1. Heimann T, Meinzer HP. Statistical shape models for 3D medical image segmentation: a review. *Medical image analysis*. 2009 Aug 1;13(4):543-63.
2. Litjens G, Kooi T, Bejnordi B, Setio A, Ciompi F, Ghahfarooian M, et al. A Survey on Deep Learning in Medical Image Analysis. *arXiv (Cornell University)*. 2017.
3. Seim H, Kainmueller D, Zachow S, Lamecker H, Heller M, Hege H. Automatic segmentation of the pelvic bones from CT data based on a statistical shape model. In: *Eurographics. Proceedings Of Eurovis X a Short Paper*; 2008. p. 93–100.
4. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv (Cornell University)*. 2015;:234–241.
5. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection. *Institute Of Electrical Electronics Engineers*; 2016. p. 779–788.
6. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8. 2023. Available from: <https://docs.ultralytics.com/models/yolov8/>. Accessed 05 April 2026.
7. Jocher G, Ultralytics Team. Ultralytics YOLO11. 2024. Available from: <https://docs.ultralytics.com/models/yolo11/>. Accessed 05 April 2026.
8. Jocher G, Qiu J. Ultralytics YOLO26. 2026. Available from: <https://docs.ultralytics.com/models/yolo26/>. Accessed 05 April 2026.

9. Almufareh MF, Imran M, Khan A, Humayun M, Asim M. Automated Brain Tumor Segmentation and Classification in MRI Using YOLO-Based Deep Learning. *IEEE Access*. 2024;12:16189–16207.
10. Şimşek MA, Sertbaş A, Sasani H, Dinçel YM. Automatic Meniscus Segmentation Using YOLO-Based Deep Learning Models with Ensemble Methods in Knee MRI Images. *Applied Sciences*. 2025;15:2752.
11. Kwolek K, Grzelecki D, Kwolek K, Marczak D, Kowalczewski J, Tyrakowski M. Automated patellar height assessment on high-resolution radiographs with a novel deep learning-based approach. *World J Orthop*. 2023;14:387–398.
12. Micheau A, Hoa D. e-Anatomy Atlas. IMAIOS. 2026. Available from: <https://www.imaios.com>. doi:10.37019/e-anatomy.
13. Wada K. Labelme: Image polygonal annotation with Python [Internet]. GitHub; 2018 [cited 2026]. Available from: <https://github.com/wkentaro/labelme>. doi:10.5281/zenodo.5711226.
14. Pizer SM, Amburn EP, Austin JD, Cromartie R, Geselowitz A, Greer T, ter Haar Romeny B, Zimmerman JB, Zuiderveld K. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*. 1987 Sep 1;39(3):355-68.
15. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA. Albumentations: Fast and Flexible Image Augmentations. *Information*. 2020;11:125.
16. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In *European conference on computer vision 2014 Sep 6* (pp. 740-755). Cham: Springer International Publishing.
17. Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging*. 2015;15:29.
18. Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding Transfer Learning for Medical Imaging. *arXiv (Cornell University)*. 2019.
19. Jocher G, Qiu J, Liu M, Lyu S, Akyon F, Kalfaoglu M. Ultralytics YOLO26: Unified Real-Time End-to-End Vision Models. 2026.
20. Ambellan F, Tack A, Ehlke M, Zachow S. Automated segmentation of knee bone and cartilage combining statistical shape knowledge and convolutional neural networks: Data from the Osteoarthritis Initiative. *Medical Image Analysis*. 2018;52:109–118.
21. Lerch TD, Zeng G, Boschung A, Meier MK, Steppacher SD, Tannast M, et al. Magnetic Resonance Imaging–Based 3-Dimensional Models of the Pelvis and Hip Using Machine Learning for Automatic Bone Segmentation in a Dynamic Hip Impingement Simulation. *Orthopaedic Journal of Sports Medicine*. 2025;13:23259671251334138.
22. Liu M, Chen Y, Tian A, Wu X, Shen M, Gong T, Lee J. Performance analysis of deep learning models for femur segmentation in mri scan. In *2025 IEEE Conference on Artificial Intelligence (CAI) 2025 May 5* (pp. 447-450). IEEE.
23. Montin E, Deniz CM, Kijowski R, Youm T, Lattanzi R. The impact of data augmentation and transfer learning on the performance of deep learning models for the segmentation of the hip on 3D magnetic resonance images. *Informatics in Medicine Unlocked*. 2024;45:101444.
24. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2020;18:203–211.

Figures

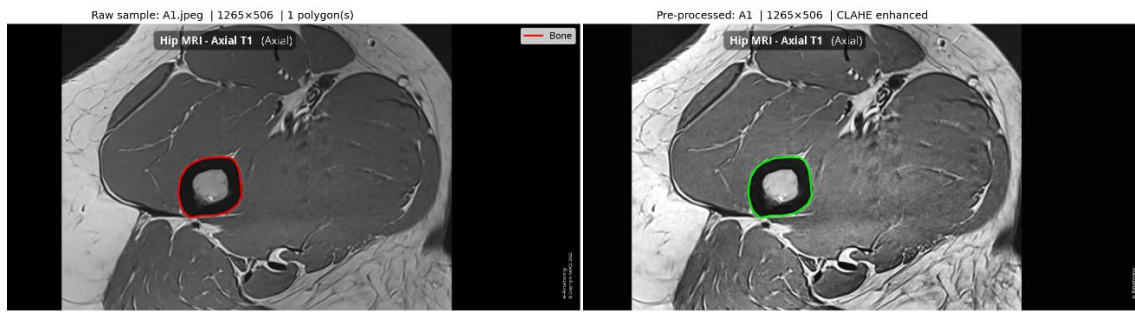


Figure 1. Representative femur T1 axial MRI slice before (Left) and after (Right) the pre-processing pipeline. CLAHE enhancement visibly improves the cortical edge contrast relative to the homogeneous medullary fat signal, facilitating subsequent polygon annotation and model learning.

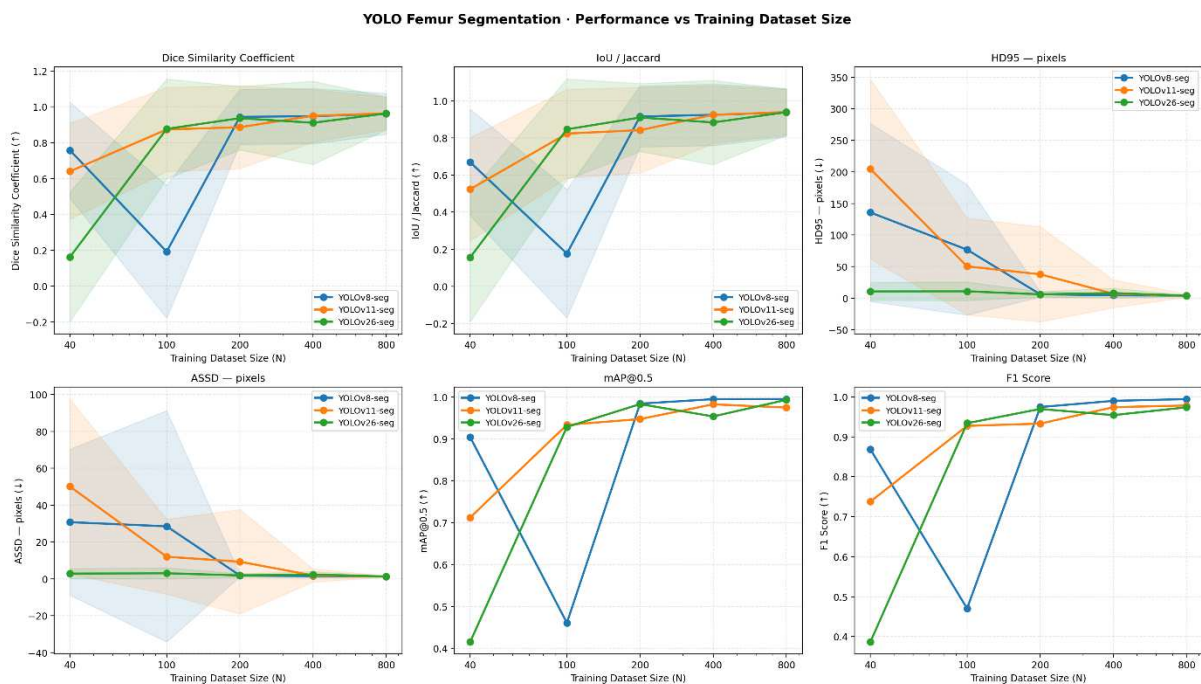


Figure 2. DSC, IOU, HD95, ASSD, mAP@0.5 and F1 score as a function of training dataset size for YOLOv8-seg (blue), YOLOv11-seg (orange), and YOLOv26-seg (green). Shaded regions: ± 1 standard deviation. The anomalous YOLOv8-seg collapse at $N = 100$ and YOLOv26-seg near-failure at $N = 40$ are clearly visible. All three converge at $N = 800$.

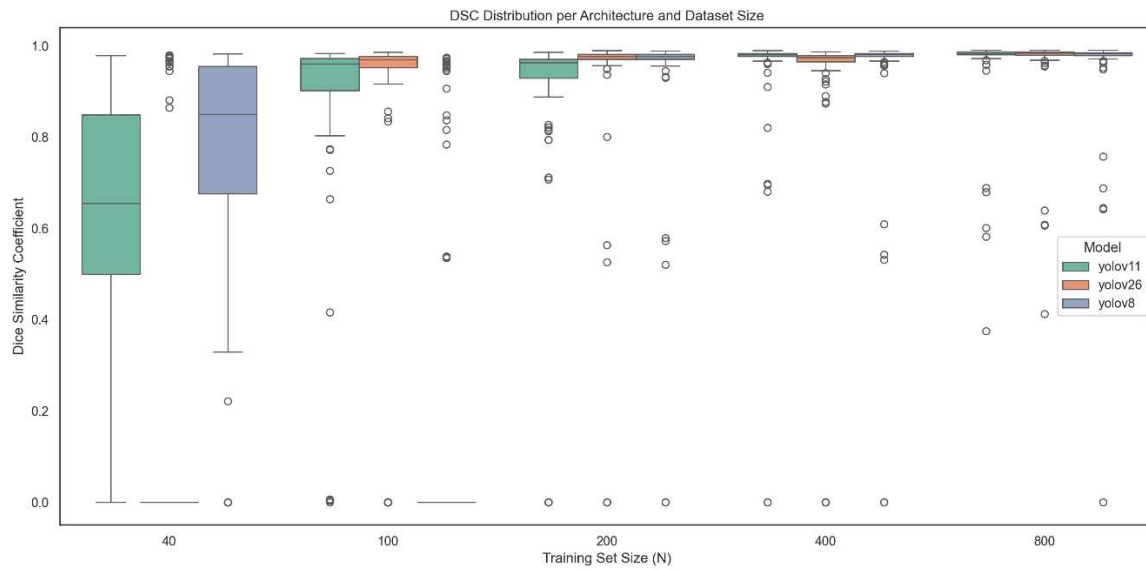


Figure 3. Boxplot distributions of DSC per architecture and dataset size.

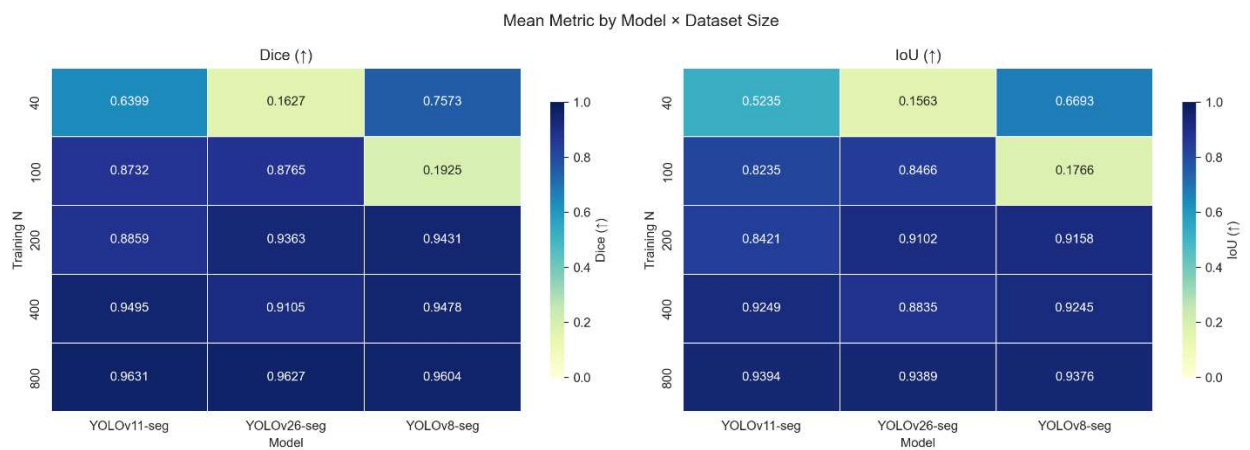


Figure 4. Heatmap of mean DSC (left) and IoU (right) across models and dataset sizes. Note the YOLOv8-seg anomaly at N = 100 and YOLOv26-seg near-failure at N = 40.

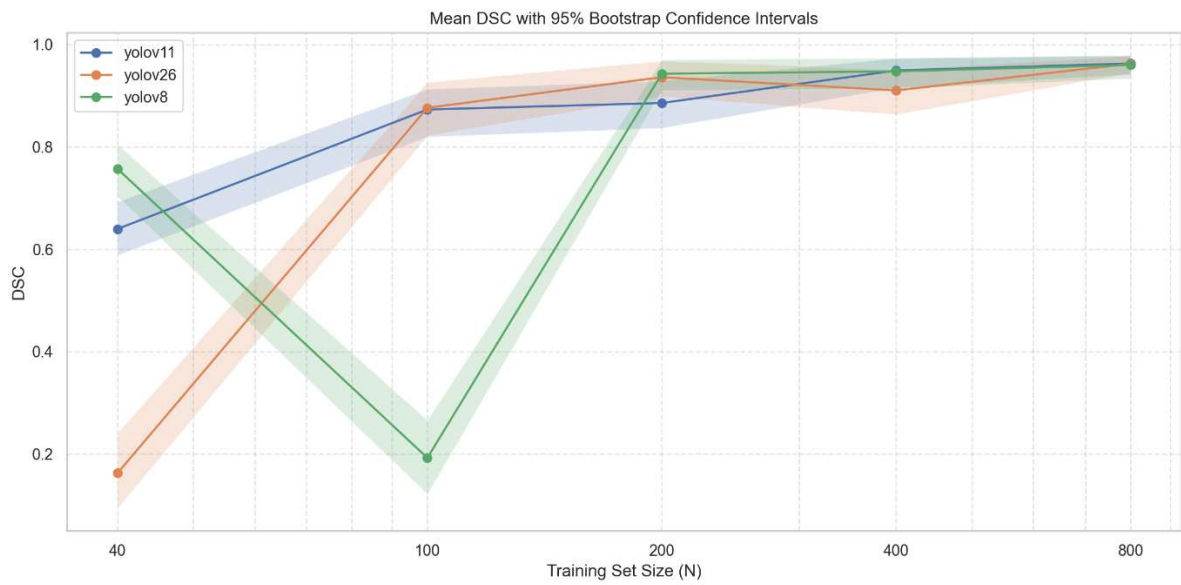


Figure 5. Mean DSC with 95% bootstrap confidence intervals (1000 iterations) per architecture across N conditions.

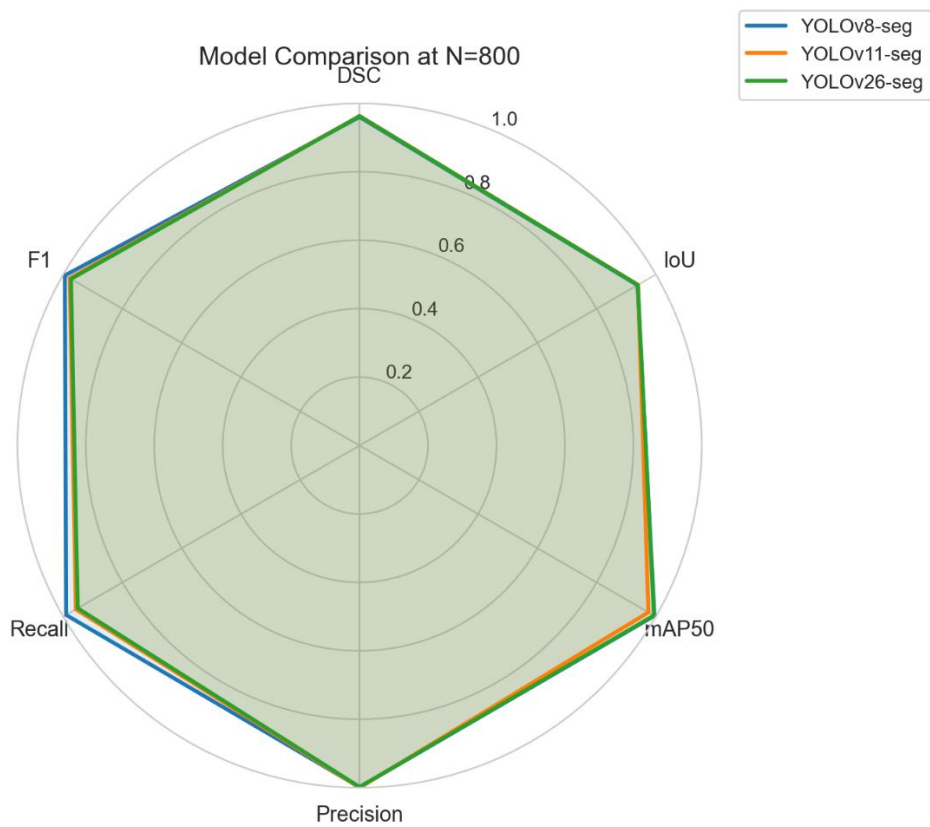


Figure 6. Radar chart comparing the three models at N = 800 across DSC, IoU, mAP@0.5, Precision, Recall, and F1.

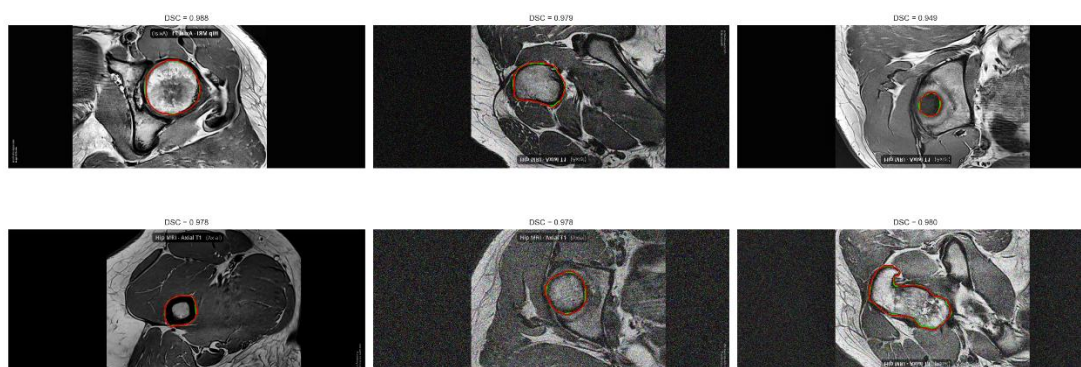


Figure 7. Representative qualitative femur contour predictions at N = 800 for YOLOv8-seg. (Left column), YOLOv11-seg (Middle column), and YOLOv26-seg (Right column) Green: ground truth; Red: prediction.