BMC Medical Imaging

RESEARCH

Open Access

Comparing YOLO and U-net deep learning algorithms in chronic wound image segmentation

Indrani Marchal[1*], Zhara Ali[1†], Haroun Ammi[1†], Steven Smet[2], Lies Van de Voorde[2], Carolina Varon[3], Michel-Antony Ngan Yamb[1], Jhonny Alexander Yunda Sangoluisa[4], Francois Quitin[1] and Antoine Nonclercq[1]

Abstract

Chronic wound analysis is crucial for effective wound care, necessitating precise and efficient segmentation techniques. This study explores the use of YOLOv8 and YOLO11, renowned deep learning algorithms for object detection, in medical image segmentation. A comprehensive evaluation was conducted comparing different versions of YOLOv8 and YOLO11 with the benchmark U-Net model using the FUSeg and Wound Data databases. Both databases were used simultaneously for training and testing to achieve a cross-dataset validation, ensuring the robustness of our models, and highlighting the importance of having data from different sources to represent the problem in various clinical contexts, thereby allowing AI models to be more adaptable and generalizable. Results show that both YOLO models significantly outperform U-Net in terms of segmentation accuracy, generalization, and inference speed. YOLOv8n achieved the highest performance with an IoU of 71.7%, a precision of 83.8%, a recall of 77.9%, and a DSC of 79.3%, and, while YOLO11s, the best performing of the YOLO11 models (IoU of 70.1%, Precision of 81.9%, Recall of 79.9%, and DSC of 78.5%), distinguished itself for its stability between thresholds and robustness across datasets. This research confirms that modern YOLO architectures offer fast, accurate, and robust solutions for automated wound segmentation, laying the groundwork for further development in AI-driven wound analysis and diagnosis.

Keywords Wound segmentation, Deep learning, YOLO, U-net, Chronic wound, Cross-dataset, AI tool

Introduction

The wound treatment market is a growing sector, reaching a value of $21.5 billion in 2023 [1]. This expansion is driven by the prevalence of chronic wounds, which affect 1 to 2% of individuals at least once in their lifetime [2]. Chronic wounds manifest in various forms, including ulcers (venous, arterial, diabetic foot, and pressure), infectious wounds, ischemic wounds, and surgical wounds. Poor blood flow, caused by conditions such as diabetes, atherosclerosis, pressure, and venous insufficiency, is a key factor hindering wound healing [3]. For instance, diabetic foot ulcers (DFUs) alone affect approximately 15 to 25% of diabetic patients and are a leading

†Zhara Ali and Haroun Ammi contributed equally to this work.

*Correspondence:
Indrani Marchal
indrani.marchal@ulb.be
[1]Department of Bio- Electro- and Mechanical Systems (BEAMS), Ecole Polytechnique de Bruxelles, Université Libre de Bruxelles, Avenue Franklin Roosevelt 50, Brussels 1050, Belgium
[2]Universitair Ziekenhuis Gent (UZGent), C. Heymanslaan 10, Gent 9000, Belgium
[3]Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven (KULeuven), Kasteelpark Arenberg 10, Leuven 3001, Belgium
[4]Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTM), Ecole Polytechnique de Louvain (EPL), Université Catholique de Louvain (UCLouvain), Rue Archimède 1, Louvain 1348, Belgium

cause of non-traumatic lower limb amputation [4, 5]. Venous leg ulcers (VLUs) and pressure ulcers (PUs) are other significant types of chronic wounds. Altogether, these conditions significantly impair patients' quality of life and impose a heavy financial burden on healthcare systems due to their long healing times, high recurrence rates, and associated complications [6, 7].

Traditional wound assessment, such as visual inspection and palpation, is inherently subjective and prones to variability. Although clinicians examine features as wound size, depth, edges, periwound skin, wound bed, and signs of infection, these evaluations remain imprecise and inconsistent. More objective methods, such as tissue cultures or blood tests, can offer more reliable insights but usually require longer processing times and are therefore not frequently clinically performed.

In contrast, automated image-based systems could provide standardized and objective wound assessments, offering clinicians accurate metrics to monitor healing and adjust treatment plans. As in many other clinical domains where segmentation is used to isolate anatomical structures or pathological regions on CT or MRI, this process [8, 9] enables the precise extraction of wound features, including area, shape, and color-key indicators for monitoring the progression of healing. These systems also have the potential to reduce clinical workloads, improve consistency in treatment decisions, and enable better care in remote or underserved regions [10]. For example, darkened skin may indicate infection, exposed muscle or bone may suggest an advanced stage of the wound, and granulation tissue (bright red) may indicate healing [11]. These features can then be integrated into decision algorithms (e.g., neural network) to support diagnosis and follow-up. Recent data from dermatology [12], ophthalmology [13], and oncology [9] show that convolutional neural networks (CNNs) can match or even surpass the performance of experts in detection and classification tasks, reinforcing the feasibility of AI-assisted decision support in clinical practice.

Several segmentation methods have been applied in this context. Traditional segmentation algorithms such as the Sobel operator [14], the Canny filter [15], the Otsu's threshold [16], the region-growing method [17], the K-Means algorithm [18], and watershed segmentation [19] have been widely used. However, as in other areas of medical imaging, these classical approaches often lack the spatial detail required for reliable clinical interpretation [20], and their performance often degrades under variable lighting, scale, and imaging conditions, limiting their generalization capabilities [21]. Recent reviews of deep learning–based medical image segmentation have reinforced these limitations, highlighting issues of dataset bias, cross-domain variability, and limited clinical deploy ability [13, 20]. These studies emphasise that achieving robustness across heterogeneous imaging conditions remains a central challenge for contemporary segmentation architectures.

With the emergence of deep learning (DL), more robust segmentation methods have emerged. Architectures such as SegNet [22], U-Net [23], and Facebook's DeepLab [24] - all based on encoder-decoder structures - have laid the groundwork for modern medical image segmentation. Among these, U-Net, specifically designed for biomedical applications, remains a gold standard, and is still widely used as a baseline in recent SOTA studies across tasks such as colorectal polyp delineation [25] or brain tumor [26] segmentation, often extended with multi-resolution blocks or attention modules to improve boundary extraction. In wound analysis, DL models have been widely applied to classification, detection, and segmentation tasks using CNNs [27], Visual Geometry Group (VGG) [28], and their iterations such as Fast-RCNN [29], Faster-R-CNN [30], ensemble methods [31], and Fully Convolutional Networks (FCNs) [32], while hybrid architectures combining CNN and transformer components have been explored in parallel fields for improved feature representation without compromising inference efficiency.

Numerous studies have confirmed the effectiveness of these models in various wound imaging contexts. Alabdulhafith et al. [33] combined ResNet34 with U-Net for precise wound segmentation. Doğru et al. [34] tested three structurally distinct variants of U-Net based on CNN to segment in vitro wound microscopy images, reaching a mean Dice Similarity Coefficient (DSC) between 95.8% and 96.8%, demonstrating the high performance of U-Net for specific biomedical tasks. Moyà-Alcover et al. [35] developed a multitask U-Net for the simultaneous segmentation of wounds and staples in abdominal surgery images, reaching an intersection-over-union score (IoU) of 71.1%. Lightweight U-Net variants have also been developed for mobile applications, as shown by Borst et al. [36], while hybrid architectures combining CNN and transformer components have been explored in parallel fields for improved feature representation without compromising inference efficiency [37].

Other advanced DL models have also been successfully applied to wound-related tasks. Liu et al. [38] used Mask R-CNN for pressure injury object detection, segmentation, and staging, achieving an average precision of 60.3% for stage recognition, exceeding the performance of medical personnel. Jacobson et al. [39] developed an AI system that combines ultrasound and RGB images to segment and classify burn injuries with an accuracy of over 84% globally. Finally, Jishnu et al. [40] used a conditional generative adversarial network (cGAN) called AFSegGAN for wound segmentation and morphological parameter estimation. This model was validated on

the MICCAI 2021 foot ulcer dataset, achieving an IoU of 99.07% and a DSC of 93.11%, which showcases its effectiveness in wound segmentation for remote healthcare applications.

In parallel, real-time computer vision has progressed with the YOLO (You Only Look Once) model, first introduced by Redmon et al. in 2015 [41]. YOLO processes an entire image in a single step to predict bounding boxes and class probabilities, making it faster than other regression models. Through successive versions (YOLOv1 to v5 under Redmon and YOLOv6 to YOLO11 under Ultralytics), YOLO improved in speed, precision, and adaptability to diverse visual tasks.

Initial applications of YOLO in wound care focused on detection and classification. For instance, Amin et al. [42] combined YOLOv2 with ShuffleNet to localize and classify DFUs, and they validated their model on their proprietary dataset, achieving a notable mean Average Precision (mAP) score of 94%. Anisuzzaman et al. [43] developed a mobile app based on YOLOv3 for wound localization, achieving a mAP of 93.9% on the AZH Wounds database. Other apps incorporating the YOLO algorithm have been developed [44, 45], including that of Han et al. [46], which integrated YOLOv3 with ResNet-101, MobileNetv2, and Darknet-53 to classify ulcer wounds according to the Wagner scale, resulting in an mAP of 91.95%. Adnan et al. [47] extended YOLOv3 with additional layers to classify wounds, achieving a remarkable classification accuracy of 99%.

However, due to the lack of native segmentation support in early YOLO versions, researchers often combined detection with external segmentation modules such as U-Net and ShuffleNet [48]. For instance, Carrión et al. [49] attempted to evaluate and track the wound size of wounded mice using the YOLOv3 model, with the segmentation part performed by a U-Net-based algorithm. Similarly, Wang et al. [50] combined YOLOv3 with the encoder-decoder MobileNetsV2 model to develop a lightweight, smartphone-compatible system, achieving a DSC of 90.47%.

The introduction of YOLOv4 and YOLOv5 enabled further performance improvements and integration into more sophisticated frameworks. For example, Kucharski et al. [51] proposed a hybrid system using YOLOv4 and the DETR Vision Transformer for DFU detection, coupled with a U-Net for segmentation, achieving a DSC of 72%. Still in the scope of the DFUC, Yap et al. [52] conducted a large benchmark of ulcer detection methods [12], that includes three variants of the Faster R-CNN method, an ensemble method, YOLOv3, YOLOv5, EfficientDet, and a Cascade Attention Network, showing that YOLOv3 and YOLOv5 performed competitively with a higher mAP for YOLOv2 and a higher F1-score and precision for YOLOV5. The CO2Dnet framework proposed by Monroy et al. [53] incorporates YOLOv4 to detect and crop images around the wound region in leprosy patients, feeding them into the U-Net segmentation model, and achieves an F1-score of 80.3%. They performed an ablation study that highlighted the benefit of integrating detection, segmentation, and data augmentation. Wound detection using YOLO extends beyond ulcer wounds, incorporating wounds such as skin burns and cutaneous leishmaniasis. Ferdinand et al. [54] successfully detected and classified skin burns based on the degree of burn (1st, 2nd, or 3rd) and achieved the best results with YOLOv5l with an mAP score at 50% equal to 83.1% compared to YOLOv5m, YOLOv5x, and YOLOv7. Leishmaniasis wound detection and classification were achieved by Noureldeen et al. [55] using the YOLOv5s model, resulting in an average classification accuracy of 70%. In a recent study by Aldughayfiq et al. [56], the detection and classification of PUs using the YOLOv5 algorithm demonstrated promising results, with an overall mAP of 76.9%.

Despite these achievements, it was only with the release of YOLOv8 in 2023 by Ultralytics [57] that native segmentation functionality was integrated into the YOLO architecture. YOLOv8 introduces anchor-free detection heads, eliminating the need for manually defined anchor boxes (as in YOLOv5) [31]. It provides for reducing the number of boxes and accelerating post-processing loads [58]. Additionally, it replaces the C3 block with a C2f module from YOLOv5. In the prediction step, YOLOv8 separates detection and classification to optimize the accuracy and efficiency of the output model [58–60]. These innovations allow YOLOv8 to handle detection, classification, and segmentation tasks within a single framework, making it highly suitable for medical image segmentation.

Building on YOLOv8, Ultralytics recently released YOLO11 as a lightweight yet powerful update. Although no peer-reviewed publication has yet described its architecture in detail, it reportedly achieves higher inference speed and segmentation accuracy, with a 22% reduction in parameters compared to YOLOv8 [61]. Both YOLOv8 and YOLO11 have demonstrated promising results in real-time segmentation tasks [62, 63], making them particularly promising for time-sensitive medical applications where both accuracy and responsiveness are critical. However, their application to wound segmentation remains unexplored, justifying the current study.

Other recent YOLO versions, including YOLOv9 and YOLOv10, introduced further architectural refinements but remain limited to detection tasks. The architecture consists of a backbone feature extractor, followed by detection heads responsible for predicting bounding boxes and class probabilities [64]. It builds on this by incorporating two innovations: the Programmable

Gradient Information (PGI) framework, which enhances gradient reliability during training and improves prediction precision; and the Generalized Efficient Layer Aggregation Network (GELAN), which enhances gradient reliability and feature aggregation [58]. YOLOv10 released in 2024, removes the traditional non-maximum suppression post-processing step, which significantly accelerates inference [57–59]. It employs a dual-assignment training strategy to enrich positive samples during learning [59, 65]. However, neither version supports native segmentation and was thus excluded from our comparative evaluation.

While numerous wound segmentation algorithms have been proposed, none have truly distinguished themselves in the literature or been adopted in clinical practice [66]. A major limitation is the lack of generalizability across datasets, often due to variability in image acquisition conditions, such as lighting, resolution, or context [67]. As a result, many models often perform well on their training data but exhibit significant performance drops on external datasets. This domain overfitting problem has been highlighted by Monroy et al. [53], who tested their framework on two distinct datasets and evaluated four training–validation combinations, revealing substantial variability in performance.

To overcome the limitations of subjective assessment and improve generalizability, this study proposes an artificial intelligence (AI)-based approach for wound segmentation using DL algorithms to automate and standardize the segmentation of chronic wound images. Among existing approaches, U-Net remains the benchmark architecture for biomedical image segmentation due to its encoder-decoder structure and high performance across a wide range of clinical tasks. In contrast, YOLOv8 and YOLO11 enable real-time detection and segmentation in a single, unified framework, offering a fundamentally different paradigm that is particularly relevant for fast, point-of-service, or mobile applications. Automated segmentation is clinically valuable as it enables objective estimation of wound area, quantitative monitoring of lesion progression, and facilitates remote monitoring in telemedicine workflows.

We benchmark U-Net, a widely accepted reference model in medical image segmentation, against YOLOv8 and YOLO11, the only iterations in the YOLO series that natively support segmentation tasks, making them strong challengers in the context of real-time medical applications. To our knowledge, neither YOLOv8 nor YOLO11 has been previously applied to the segmentation of chronic wounds. Beyond the introduction of these recent architectures, this study is the first to propose a systematic cross-dataset benchmarking of YOLO-based segmentation models in the context of chronic wound analysis, enabling a controlled assessment of their intrinsic generalization capabilities across distinct clinical image sources.

All models are trained and tested on two large public datasets (FUSeg and Wound Data), using a cross-dataset validation strategy to assess both segmentation accuracy and robustness. Ultimately, this study aims to identify models that combine performance, speed, and generalization to support the development of clinically applicable AI tools for wound care. To achieve this, we adopt a controlled comparison framework in which complete segmentation architectures are evaluated under identical training conditions, ensuring a consistent and clinically relevant assessment of model behaviour.

Materials and methods

Databases

This study uses two of the largest publicly available wound datasets: the FUSeg database [43] and the Wound Data database from Kaggle [68].

The FUSeg database includes 1109 foot ulcer images from 889 patients in .jpg format and was collected over two years at the AZH Wound and Vascular Center in Milwaukee, Wisconsin. The images were captured with a Canon S×620 HS digital camera and an iPad Pro. Each image is annotated by a wound specialist and focuses solely on the wound region and its surrounding skin (Fig. 1). Since this dataset was created for a segmentation challenge, the test set does not contain any labels, leaving 831 images available for training and validation.

The Wound Data dataset contains 1210 foot ulcer images, likewise, acquired at the AZH Wound and Vascular Center using a Canon S×620 HS camera and an iPad Pro under uncontrolled illumination conditions, with various backgrounds. All images were zero-padded and had a unique size of 512×512 pixels. These data are already divided into train, validation, and test files. The train file contains 810 images, while the validation and test files each contain 200 images; however, only the train and validation files have their corresponding masks. It was then decided to remove the test files and retain only the labeled images, resulting in a total of 1010 images for this dataset (Fig. 2).

The two datasets were merged and divided into training, validation, and test sets in a 64-16-20 ratio. To mitigate data-split bias, a 5-fold cross-validation approach was employed. For the FUSeg database, the train, validation, and test sets contain 492, 169, and 170 images, respectively, in each fold. As for the Wound Data database, it contains 486, 162, and 162 images, respectively. To ensure reproducibility, the main acquisition and annotation characteristics are summarised in Table 1. Because neither dataset provides patient-level information, the split was performed at the image level. Each image appears in only one of the training, validation, or

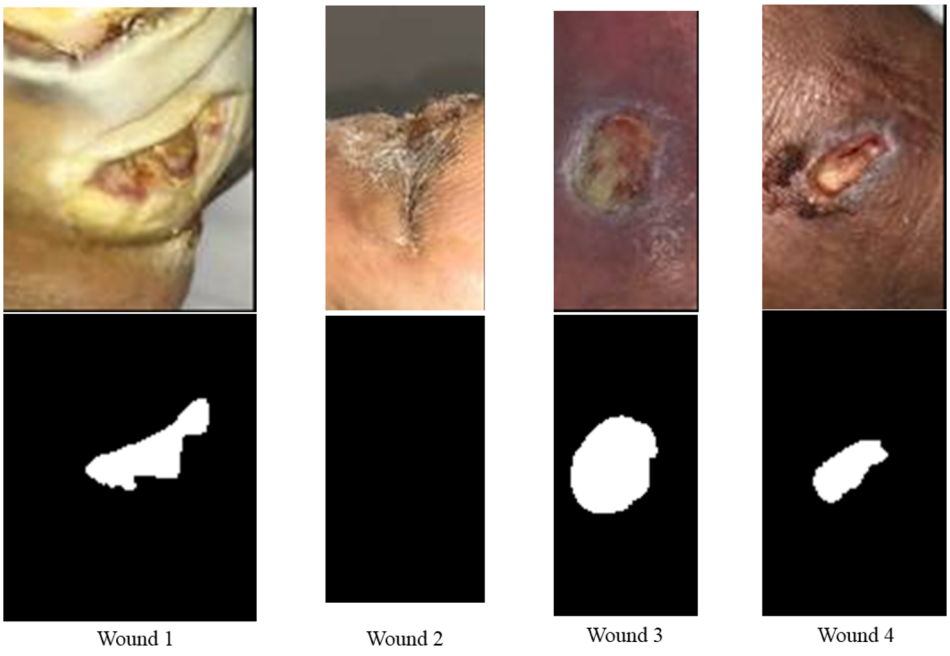

Fig. 1 Annotated images from the FUSeg database

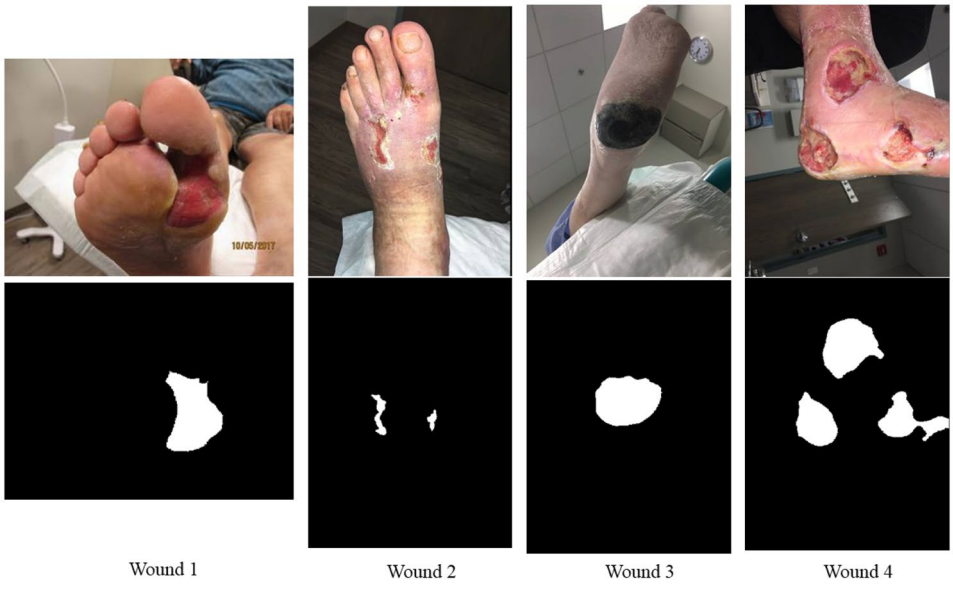

Fig. 2 Annotated images from the Wound Data database

test sets. However, images from the same wound may still occur across different folds.

Metrics

Model performance was assessed using standard metrics derived from the confusion matrix. These metrics include IoU, precision, recall, and DSC.

IoU (Eq. 1) is the ratio of the overlapping area to the union area of predicted and ground truth regions. For YOLO and U-Net models, IoU is computed from pixel-wise overlap between the predicted segmentation mask and the reference wound mask.

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} \tag{1}$$

IoU provides a global accuracy measure of the segmentation performance, integrating both false positives and false negatives. It serves as a key metric for comparing overall segmentation quality between models.

Table 1 Detailed acquisition, resolution, and annotation characteristics of the FUSeg and Wound Data datasets

Attribute	FUSeg Dataset	Wound Data Dataset
Number of images	1109 images; 831 labelled (train/val)	1210 images; 1010 labelled (train/val)
Imaging modality	Camera (Canon S×620 HS), iPad Pro	Camera (Canon S×620 HS), iPad Pro
Image type	RGB photographs (.jpg)	RGB photographs (.png)
Resolution range	Heterogeneous (device-dependent)	Heterogeneous (device-dependent)
Acquisition conditions	Uncontrolled illumination; variable backgrounds	Uncontrolled illumination; variable backgrounds
Annotation protocol	Manual segmentation masks reviewed by wound-care specialists	Two-stage annotation: outlining by trained assistants, reviewed by two MDs and two medical assistants; all non-epithelialised tissue labelled as wound
Mask availability	Binary segmentation masks	Binary segmentation masks
Known limitations	Limited contextual diversity	Limited contextual diversity

Precision (Eq. 2) measures the ratio between the correctly classified samples and all the samples classified as positive.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Where TP (True Positive) refers to a pixel belonging to the positive class being classified correctly, and FP (False Positive) refers to a pixel belonging to the negative class but being classified as positive. In the U-Net and YOLO models, TP and FP values are defined at the pixel level based on the predicted binary mask. In this study, precision is used to evaluate the ability of the models to avoid over-segmentation, i.e., to ensure that the predicted wound regions are not excessively extended beyond the true wound boundaries.

Recall (Eq. 3) is the ratio between the correctly classified samples and all the positive samples in the set.

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

Where FN (False Negative) refers to a pixel belonging to the positive class but being classified as negative. Recall is critical in this context to assess the models' sensitivity in capturing the entirety of the wound area, which is essential in clinical applications to avoid missing pathological regions.

DSC (Eq. 4) measures the similarity between predicted and ground truth segmentation masks. For all models, DSC is calculated using pixel-wise overlap between the predicted binary mask and the annotated wound mask.

$$DSC = \frac{2 * Area\ of\ overlap}{Pixels\ in\ prediction + Pixels\ in\ ground\ truth} \tag{4}$$

The DSC balances precision and recall into a single metric, making it particularly useful for comparing models' abilities to both identify and localize the wound area accurately.

Metrics were computed across thresholds ranging from 0.1 to 1.0 in increments of 0.1.

Statistical test

Model performance was evaluated using five-fold cross-validation for each model and training–testing configuration. For every fold, IoU, DSC, precision, and recall were computed at the threshold that maximised the mean IoU across folds. These fold-level performances were then summarised using the mean, standard deviation (SD), and the corresponding 95% confidence interval (CI), computed as:

$$mean \pm 1.96 * SD/\sqrt{5} \tag{5}$$

Per-fold statistics for all models and configurations are provided in Suplementary file S1.

The non-parametrical McNemar's test [69] was applied to a 2×2 contingency table [70] based on TP, TN (True Negative), FP, FN. This analysis evaluates whether two models systematically disagree on specific pixels and complements the continuous metrics reported in the main text.

The table categorizes:

- pixels correctly classified by both algorithms,
- pixels misclassified by YOLO only,
- pixels misclassified by U-Net only, and
- pixels misclassified by both algorithms.

Since McNemar evaluates binary disagreement, it does not directly quantify segmentation quality. Therefore, to statistically compare model performance under identical cross-validation splits, we performed paired analyses of IoU and DSC values at the fold level. Given the small sample size ($n = 5$ folds), the Wilcoxon test was used as the primary non-parametric test. Paired t-tests were also calculated as a complementary analysis. A p-value < 0.05

was considered statistically significant. Comparisons by fold and *p*-values are presented in Supplementary file S2. Paired statistical tests (Wilcoxon, paired t-test) for YOLOv8 vs U-Net and YOLO11 vs U-Net.

Models

YOLOv8 and YOLO11 models were selected to achieve wound segmentation. The results obtained with these models are compared to a benchmark U-Net model. Both models come in multiple versions (-l, -m, -n, -s, -x), which differ in the number of trainable parameters, as summarized in Table 2. Training was performed from scratch without external pretraining, and no architectural modifications were applied. The native Ultralytics segmentation heads were used throughout; anchor configuration was not required, as both YOLOv8 and YOLO11 operate in an anchor-free formulation.

Their architecture comprises [71–73]:

- Backbone: Extracts meaningful features from the input image.
  For both YOLO models, the modified CSP (Cross Stage Partial) includes a C2f block for YOLOv8, which focuses on efficiency and preserving the feature map. It also includes a C3k2 block for YOLO11, which improves feature representation while using fewer parameters than the C2f block.
- Neck: Acts as a bridge between the backbone and the head, performing feature fusion and integrating contextual information.
  Both models employ a Spatial Pyramid Fast Fusion (SPFF) module, designed to extract features at different scales from different regions of an image. YOLO11 further introduces a C2PSA attention module to improve the model's focus on critical areas of an image.
- Head: Generates the final outputs, including bounding boxes, class probabilities, and, when operating in segmentation mode, instance mask coefficient used to construct the shotput segmentation masks.
  For both YOLO models, it is composed of multiple convolutional layers fully connected to process the features extracted by the backbone In segmentation mode, YOLOv8 and YOLO11 generate instance masks for each detected region, along with class labels and confidence scores. While YOLO still relies on a grid-based detection mechanism to locate objects, the segmentation task extends this prediction by identifying the specific pixels belonging to each lesion rather than just their bounding boxes [74].

Table 2 Number of parameters in the different YOLOv8 and YOLO11 versions

YOLO version	Number of parameters
YOLOv8n	3.2
YOLOv8s	11.2
YOLOv8m	25.9
YOLOv8l	43.7
YOLOv8x	68.2
YOLO11n	2.6
YOLO11s	9.4
YOLO11m	20.1
YOLO11l	24.3
YOLO11x	56.9

The U-Net model is a convolutional network specifically designed for segmenting biomedical images, developed by Ronneberger et al. [23] in 2015. The U-Net follows an encoder-decoder architecture. The encoder extracts high-level features by progressively reducing spatial dimensions, while the decoder restores the original dimensions, integrating high-resolution features.

In this study, the U-Net configuration used consists of four encoding stages and four decoding stages composed of paired convolutional blocks connected by skip connections, with a bottleneck layer at the deepest level. Each block uses 3×3 convolutions (Fig. 3) with ReLU activation and 'same' padding. The sub-sampling is performed via 2×2 max-pooling operation, and oversampling is performed by nearest neighbour interpolation followed by feature-map concatenation. The number of filters doubles at each downsampling step (16–32–64–128) and reaches 256 filters in the bottleneck. The decoder symmetrically halves the number of filters at each level. The final segmentation layer applies a 1×1 convolution with sigmoid activation to produce a binary wound mask. No normalization layers or dropout were integrated. The model was trained using the binary cross-entropy loss, the Adam optimizer, and a learning rate of 0.001. A complete layer-by-layer description, including the number of filters per block, the depth of the architecture and the total parameter count, is provided in Table 3 to ensure full reproducibility.

Both models were trained from scratch on an Intel Xeon Gold 5218 R CPU @ 2.10 GHz (2 processors) with 256 GB RAM and the NVIDIA RTX A4000 GPU was used to accelerate the processing time. Python 3.8.0 and the official Ultralytics implementation were used for YOLOv8 and YOLO11 [57]. The U-Net model was trained on a Jupyter Notebook.

All models were trained using the Adam optimizer (initial learning rate of 0.001). To ensure consistency, a fixed input resolution of 640×640 pixels was used for all models; Images were resized and normalized prior to training; no data augmentation was applied. YOLOv8 and

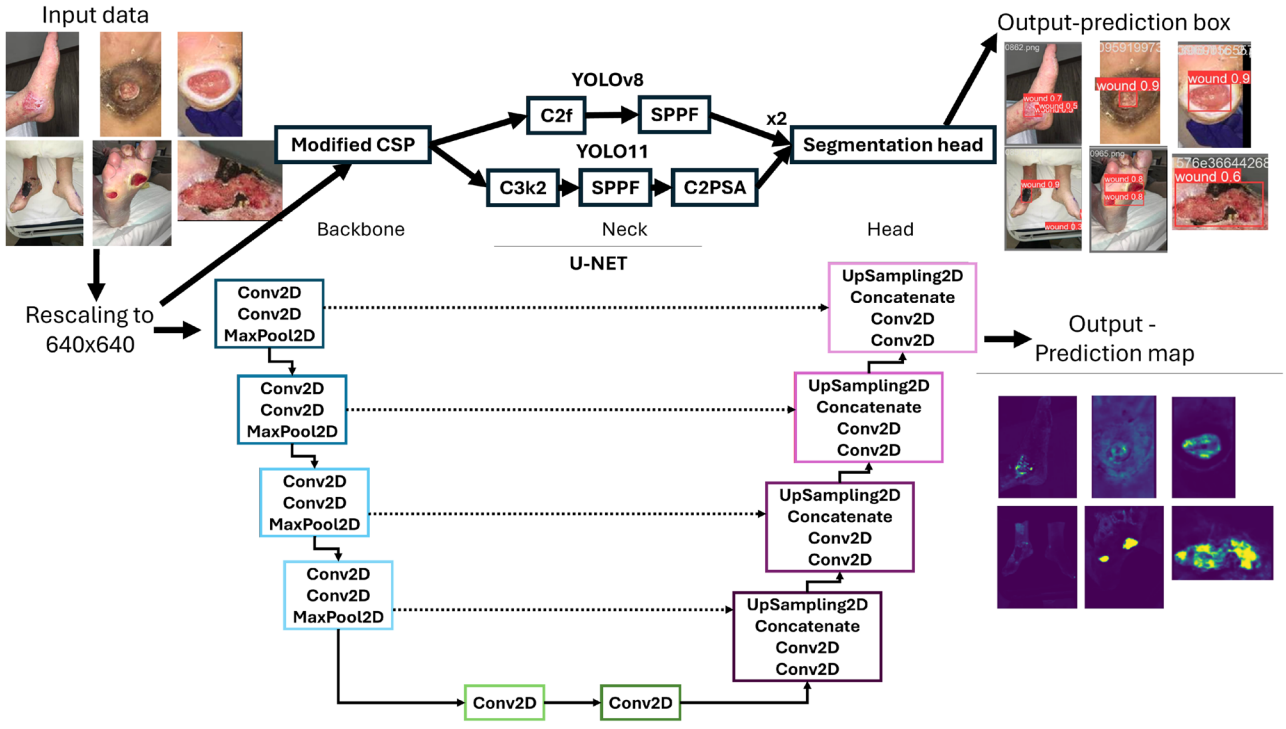

Fig. 3 Segmentation workflow (with a simplified representation of the YOLOv8 and YOLO11 models), where modified CSP is a convolutional neural network and backbone for object detection, C2f is the faster implementation of C2 (CSP bottleneck with two convolutions), C3k2 is an enhancement of the CSP block, SPPF is spatial Pyramid pooling Fast module, C2PSA is a CSP with spatial attention block, Conv2D are the convolutional 2D layers, and MaxPool2D are the max pooling 2D layers

Table 3 U-Net configuration [23, 75]

Layer	Description	Operation	Number of parameters	Filters
Input	Input normalization and spatial scaling (640 × 640)	Image rescaling	0	0
Encoder Block 1	Extraction of low-level spatial features	Conv2D Conv2D MaxPooling2D	448; 2,320; 0	16
Encoder Block 2	Extraction of mid-level features; spatial reduction	Conv2D Conv2D MaxPooling2D	4,640; 9248; 0	32
Encoder Block 3	Extraction of increasingly abstract patterns	Conv2D Conv2D MaxPooling2D	18,496; 36,928; 0	64
Encoder Block 4	High-level semantic encoding	Conv2D Conv2D MaxPooling2D	73,856; 147,584; 0	128
Bottleneck	Deep semantic representation (latent features)	Conv2D Conv2D	295,168; 590,080	256
Decoder Block 1	Recovery of spatial resolution + fusion with encoder features	UpSampling2D Concatenate Conv2D Conv2D	0; 0; 442,496; 147,584	128
Decoder Block 2	Spatial refinement and feature consolidation	UpSampling2D Concatenate Conv2D Conv2D	0; 0; 110,656; 36,928	64
Decoder Block 3	Progressive restoration of fine-grained details	UpSampling2D Concatenate Conv2D Conv2D	0; 0; 2,768,016; 9,248	32
Decoder Block 4	Final resolution enhancement before output	UpSampling2D Concatenate Conv2D Conv2D	0; 0; 6,928; 2,320	16
Output	Pixel-wise probability estimation	Conv2D	17	1

YOLO11 were trained with disabled pretrained weights, ensuring complete comparability with U-Net. Detailed training configurations, including random seeds, worker settings, and additional hyperparameters, are provided in Table 4.

Workflow

The segmentation workflow can be summarized as follows (Fig. 3):

1. Preprocessing: the images were resized to 640 × 640 pixels for all models and normalized to the [0,1] range, following the preprocessing conventions of both Ultralytics and the U-Net implementation. No data augmentation was applied.
2. Model Outputs:

   - YOLOv8 or YOLO11 (in segmentation mode) first produced bounding boxes with class labels and

Table 4 Comparison of model configurations for YOLOv8, YOLO11 and U-Net

Parameter	YOLOv8	YOLO11	U-Net
Model variants	n, s, m, l, x	n, s, m, l, x	Single architecture
Main objective	Real-time object detection + segmentation	Real-time object detection + segmentation	Pixel-wise semantic segmentation
Input resolution	640 × 640 px	640 × 640 px	640 × 640 px
Pretraining	None (trained from scratch)	None (trained from scratch)	None (trained from scratch)
Architectural modifications	Native Ultralytics segmentation head	Native Ultralytics segmentation head	Canonical U-Net encoder–decoder
Anchor configuration	Anchor-free	Anchor-free	Not applicable
Backbone / Encoder	Modified CSP with C2f block	Modified CSP with C3k2 block	Four-level encoder (Conv2D–Conv2D–MaxPool2D)
Neck / Bottleneck	SPFF	SPFF + C2PSA	Conv2D–Conv2D
Decoder / Head	Ultralytics mask head	Ultralytics mask head	Four-level decoder (UpSampling2D–Concat–Conv2D–Conv2D)
Optimizer	Adam	Adam	Adam
Initial learning rate	0.001	0.001	0.001
Batch size	16	16	16
Max epochs	100	100	100
Early stopping	Patience = 20	Patience = 20	Patience = 20
Seed	18	18	42
Workers	16	16	Not applicable (custom Keras generator)

Table 5 Cross-dataset combinations

Test	Training set	Testing set
1	Fuseg	Fuseg
2	Fuseg	Wound Data
3	Wound Data	Wound Data
4	Wound Data	Fuseg
5	Fuseg & Wound Data	Fuseg
6	Fuseg & Wound Data	Wound Data

confidence scores containing potential wounds. For each detection, the models also generate a corresponding segmentation mask. These masks were binarized using thresholds from 0.1 to 1.0, in increments of 0.1.

- U-Net generated a probability map with pixel values ranging from 0 to 1, where each pixel has a higher or lower likelihood of belonging to the wound. The same range of thresholds (0.1 to 1.0) was used to binarize the predictions and evaluate performance.

3. Metrics Computation: the predictions were segmented using a threshold, and the performance metrics were averaged over 5-fold cross-validation. The metric values considered were those given by the threshold with which the highest IoU score was achieved.

Six tests were performed using cross-dataset training and testing for each model (YOLOv8, YOLO11, and U-Net). The combinations are summarized in Table 5 :

Results

The results of the tests described in Table 5 are summarized in Tables 6, 7, 8, 9, 10 and 11 and can be visualized on the metric graphs (Figs. 4, 5 and 6). The tables are color-coded to indicate the closeness of the results from the different models, ranging from dark (closest) to light green (close) and light (far) to dark orange (furthest).

Test 1 & 3 - trained on one database, tested on the same database

The results, illustrated in Fig. 4 and summarized in Table 6 for the FUSeg database, show a significant performance gap between the YOLO models and the U-Net model. The U-Net model achieves an IoU of 43.2% (95% CI: 39–47%), considerably lower than all versions of YOLO. Although its precision reaches 70.4% (95% CI: 66–75%), it masks a lower sensitivity (recall = 63.2%), and the DSC remains limited to 62.2% (95% CI: 54–71%). This suggests an insufficient ability to accurately capture wound regions, possibly due to difficulties in modeling irregular edges or a lack of spatial contextualization in the U-Net architecture. In comparison, the YOLOv8 models demonstrate consistent performance, with an IoU ranging from 68.7% to 70.6%, representing an absolute improvement of more than 25% over the U-Net model. The differences between the different versions are minimal (less than 2.5% for IoU and DSC, and less than 4% for recall), indicating a high degree of stability between model sizes. This similarity of results means that a version can be selected based on the available resources, without any noticeable impact on the quality of the segmentation. For YOLO11, version n emerged as the best version, with an

Table 6 Average results, trained and tested on FUSeg. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.2	43.2	70.4	63.2	62.2
YOLOv8n	0.4	69.4	80.4	76.8	76.7
YOLOv8s	0.2	70.6	80.6	79	78.1
YOLOv8m	0.3	69.3	81.3	75.4	76.6
YOLOv8l	0.2	70.6	80.6	79	78.1
YOLOv8x	0.4	68.7	80	75.8	75.9
YOLO11n	0.3	**70.7**	**81.9**	**79.4**	**78.6**
YOLO11s	0.2	57.9	73.1	66.3	66.3
YOLO11m	0.3	67.3	80.4	76	75.7
YOLO11l	0.3	66.7	79.8	74.9	74.9
YOLO11x	0.4	51.3	67.9	60.1	60.2

Table 7 Average results, trained and tested on wound data. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.5	48.9	58.3	67.3	56
YOLOv8n	0.3	69.2	**82.3**	75.6	77.3
YOLOv8s	0.3	69	81.6	76.1	77.1
YOLOv8m	0.4	69.5	79.6	78.8	77.6
YOLOv8l	0.4	67.9	77.8	76.8	75.9
YOLOv8x	0.3	68.1	80.5	74.9	76.2
YOLO11n	0.3	**72.1**	82.2	**81.9**	**80.4**
YOLO11s	0.3	50.7	62.3	59.8	57.8
YOLO11m	0.3	71.4	81.9	81.7	79.9
YOLO11l	0.3	71.2	81.8	81.6	79.8
YOLO11x	0.4	55.2	70.7	64.9	64.3

Table 8 Average results, trained on FUSeg – tested on wound data. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.4	15.4	24.3	50	21.6
YOLOv8n	0.6	38.7	76.4	42.8	46.5
YOLOv8s	0.9	**40.5**	79.5	**44.7**	48.7
YOLOv8m	0.6	40.2	77.1	44.3	**48.9**
YOLOv8l	0.6	35.2	78	38.7	42.4
YOLOv8x	0.6	31.4	77.8	34.7	38.3
YOLO11n	0.9	33.5	**79.7**	36.5	41.7
YOLO11s	0.9	31.7	73.5	35.6	39.8
YOLO11m	0.9	31.5	77.5	34.7	39
YOLO11l	0.9	32.6	77.9	36.2	40.8
YOLO11x	0.9	25.1	66	28.8	32.2

IoU of 70.7% (95% CI: 70–72%), a precision of 81.9% (95% CI: 80–85%), a recall of 79.4% (95% CI: 78–80%) and a DSC of 78.6% (95% CI: 77–80%), reflecting an excellent compromise between sensitivity and specificity.

Similar trends were observed in the Wound Data database (Table 7). YOLO11n was the best performer, with an IoU of 72.1% (95% CI: 70–75%), a precision of 82.2% (95% CI: 80–84%), a recall of 81.9% (95% CI: 79–85%) and a DSC of 80.4% (95% CI: 78–83%). Among the YOLOv8

Table 9 Average results, trained on wound data – tested on FUSeg. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.1	36.1	66.1	49.8	45.7
YOLOv8n	0.3	53.2	57.7	60.5	58.2
YOLOv8s	0.1	52	54.9	61.9	57.4
YOLOv8m	0.3	58.5	**65.3**	65.9	63.9
YOLOv8l	0.2	56.1	61.9	63.2	61.6
YOLOv8x	0.3	54.2	58.5	62	59.4
YOLO11n	0.1	**59.6**	63.5	**71.5**	**66.1**
YOLO11s	0.1	39.7	46.8	49.1	45.6
YOLO11m	0.1	55.7	59.7	67.6	62.2
YOLO11l	0.1	57	61.4	68.8	63.6
YOLO11x	0.1	35.7	42.3	48.5	42

Table 10 Average results, trained on both databases - tested on FUSeg. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.1	38.1	57.9	61.1	48.4
YOLOv8n	0.4	73	81.5	80.6	79.7
YOLOv8s	0.1	**79.4**	88.5	**86.1**	**86.2**
YOLOv8m	0.4	73.7	81.5	81	80.3
YOLOv8l	0.4	73.1	81.1	80.6	79.7
YOLOv8x	0.4	73	80.7	80.9	79.7
YOLO11n	0.2	73.1	82.6	81.7	80.5
YOLO11s	0.2	73.6	82.8	81.9	80.8
YOLO11m	0.1	72.3	**83**	81.2	80.1
YOLO11l	0.2	72.3	82.6	81.1	79.9
YOLO11x	0.1	61.5	71.7	71.3	69.1

Table 11 Average results, trained on both databases - tested on wound data. Only the threshold with the best results is presented. The highest metric scores are in bold

Model	Best threshold	IoU (%)	Precision (%)	Recall (%)	DSC (%)
U-Net	0.8	57.3	72.6	78	65
YOLOv8n	0.3	**71.7**	**83.8**	77.9	**79.3**
YOLOv8s	0.2	70.9	80.2	**79**	78
YOLOv8m	0.3	71.1	82.9	77.6	78.8
YOLOv8l	0.3	70.8	83.1	76.9	78.6
YOLOv8x	0.4	70.9	80.3	79	78.6
YOLO11n	0.4	63.2	80.2	71.9	72.1
YOLO11s	0.4	70.1	81.9	79.9	78.5
YOLO11m	0.4	64.8	80.5	73.5	73.6
YOLO11l	0.6	63.1	81.2	71.5	72
YOLO11x	0.4	55.5	75.1	64.1	64.4

models, YOLOv8m achieves the best results with an IoU of 69.5% (95% CI: 68–71%), while all the models in this family exceed an IoU of 67.9%, maintaining a good balance between precision and recall. However, the U-Net model achieves an IoU of 48.9% (95% CI: 46–52%) at best, confirming its limitations in contexts richer in visual content. The performance of the different versions of YOLO11 was more mixed, with an IoU ranging from 50.7% (YOLO11s) to 72.1% (YOLO11n). The worst-performing versions (YOLO11s and x) had accuracies and

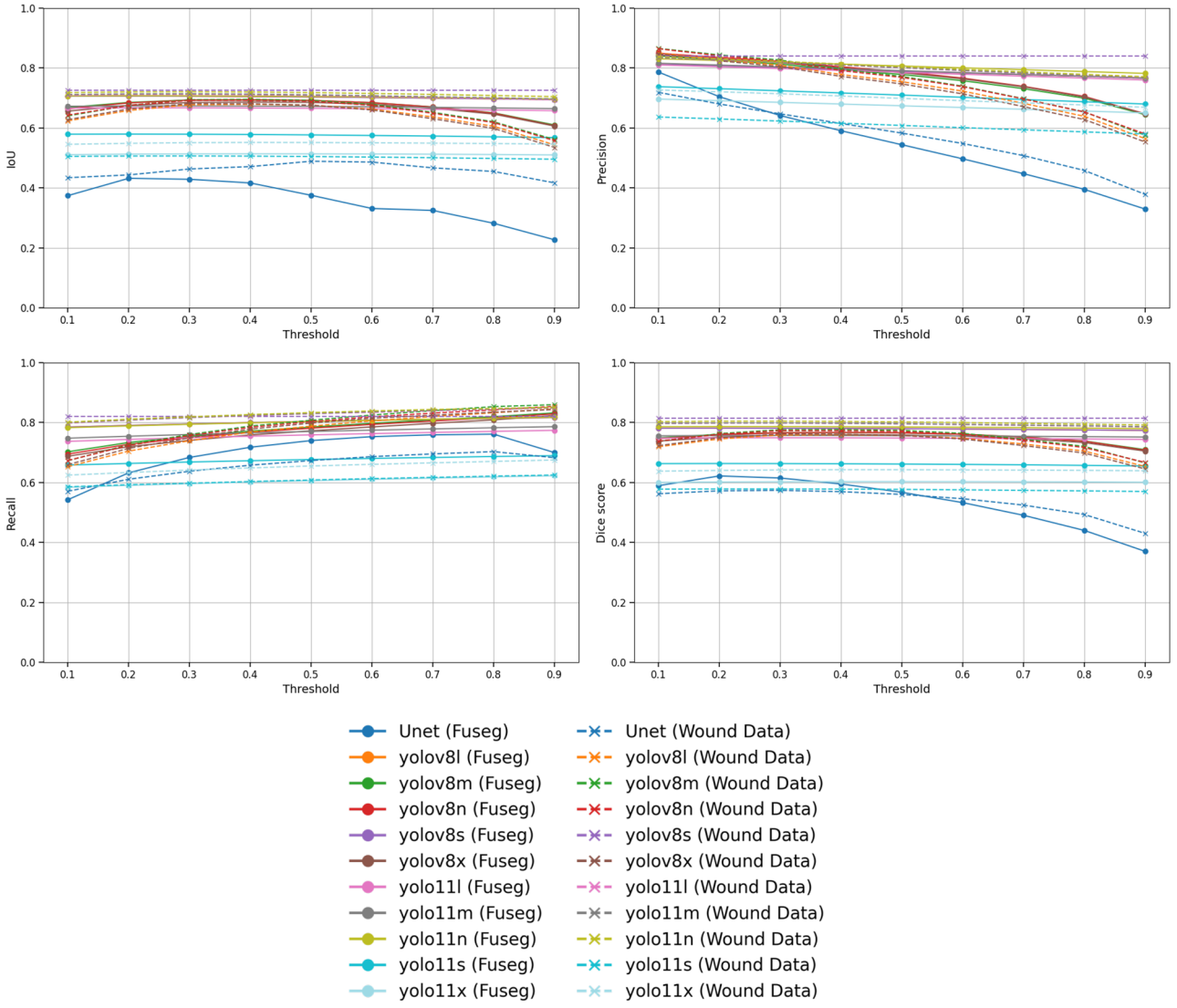

Fig. 4 Results when trained and tested on the same database (the color refers to the database used). Metric graphs averaged over 5 folds (top right: IoU, top left: precision, bottom right: recall, bottom left: dice coefficient) for Test 1 and 3

recalls of less than 70.7%, suggesting overlearning or a structural mismatch with the complexity of the Wound Data images.

The curves in Fig. 4 confirm these observations for Tests 1 and 3. The U-Net model plots are more irregular, more sensitive to threshold variations, and perform significantly worse on all metrics. The YOLO11 models show the best inter-threshold stability, with an average variation in IoU of just 1.13%, compared with 11% for YOLOv8 and 16.7% for U-Net. This remarkable stability underlines the robustness and effectiveness of YOLO11 for segmentation tasks.

Finally, fold-level paired tests confirmed that these differences are statistically meaningful. In both Test 1 and Test 3, the best YOLOv8 and YOLO11 variants achieved significantly higher IoU and DSC than U-Net (paired t-tests, $p < 0.01$ for most comparisons). Complete per-fold metrics and p-values are reported in Supplementary file S1 and S2.

Tests 2 & 4 - trained on one database, tested on another

As shown in Fig. 5 and Table 8, this test reveals the limitations of cross-domain generalization when models are trained on focused images, such as those from FUSeg. U-Net performed poorly, with an IoU of 15.4% (95% CI: 6–25%) and a DSC of 21.6% (95% CI: 9–34%). The YOLOv8 models are more resilient, with an IoU ranging from 31.54% to 40.25% and a DSC ranging from 38.3% to 48.9%. YOLOv8s performed best, with an IoU of 40.5% (95% CI: 35–45%), a precision of 79.5% (95% CI: 81–90%), a recall of 44.7% (95% CI: 37–47%) and a DSC of 48.7% (95% CI: 43–54%).The YOLO11 models exhibit

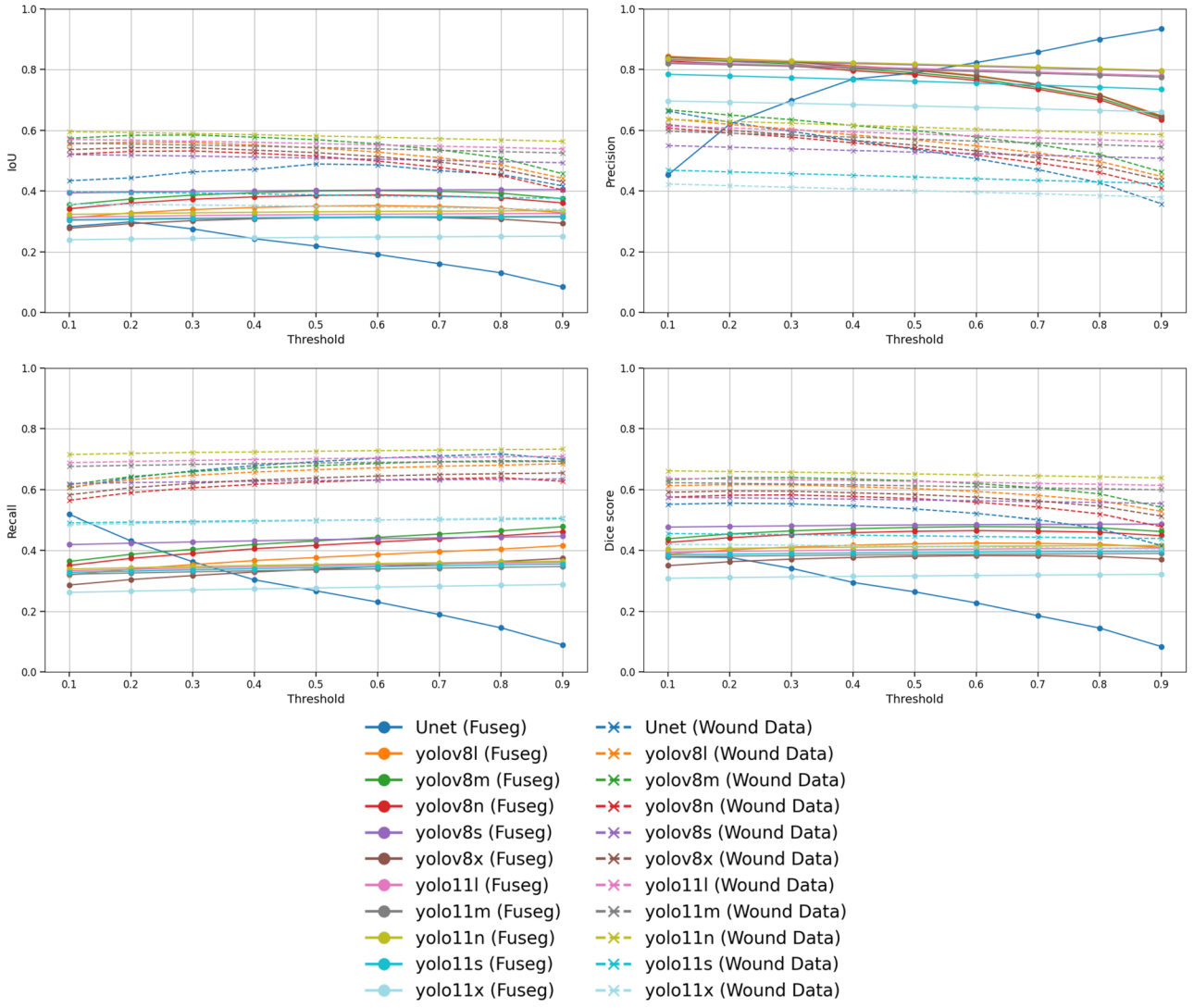

Fig. 5 Results when trained on one database (the color refers to the database used), tested on another database. Metric graphs averaged over 5 folds (top right: IoU, top left: precision, bottom right: recall, bottom left: dice coefficient) for Test 2 and 4

higher heterogeneity, with overall scores lower than those of YOLOv8. Their IoU performance ranges from 25.1% (95% CI: 21–29%) to 33.5%, with YOLO11x being the worst model, achieving a precision of 66% (95% CI: 57–75%), a recall of 28.8% (95% CI: 24–34%), and a DSC of 32.2% (95% CI: 27–37%). In general, all the models suffered a significant drop in performance compared to test 1 (Sect. 3), reflecting a difficulty in effectively transferring the knowledge acquired on FUSeg to more complex and contextual images, such as those from Wound Data.

In Test 4 (Table 9), performance improved significantly compared with Test 2. U-Net improved with an IoU of 48.9% and a DSC of 53.6%, confirming that learning on images rich in visual context, such as those from Wound Data, facilitates generalization to simpler images, such as those from FUSeg. The YOLOv8 models confirmed their robustness, with an IoU ranging from 52% to 58.5% and a DSC ranging from 57.4% to 63.9%, with the highest scores achieved by YOLOv8m. YOLO11 models show more variable performance, with an IoU ranging from 35.7% to 59.6%. YOLO11n in particular stands out with an IoU of 59.6%, a DSC of 66.1%, and a good precision-recall balance, outperforming all the other models. The m and l versions remain competitive, while YOLO11s and x, as seen in Test 3 (Sect. 3), exhibit marked instability (with an IoU of 39.7% and 35.7% respectively).

A comparison of the first four tests reveals the consistent superiority of YOLO models over U-Net, regardless of the scenario. YOLOv8m stands out for its consistency, while YOLO11n excels in certain specific configurations. U-Net exhibits significant difficulty in generalizing from one domain to another, with a loss of IoU of almost 15% when transitioning from an intra-dataset test to an inter-dataset test. The YOLOv8 models maintain much better

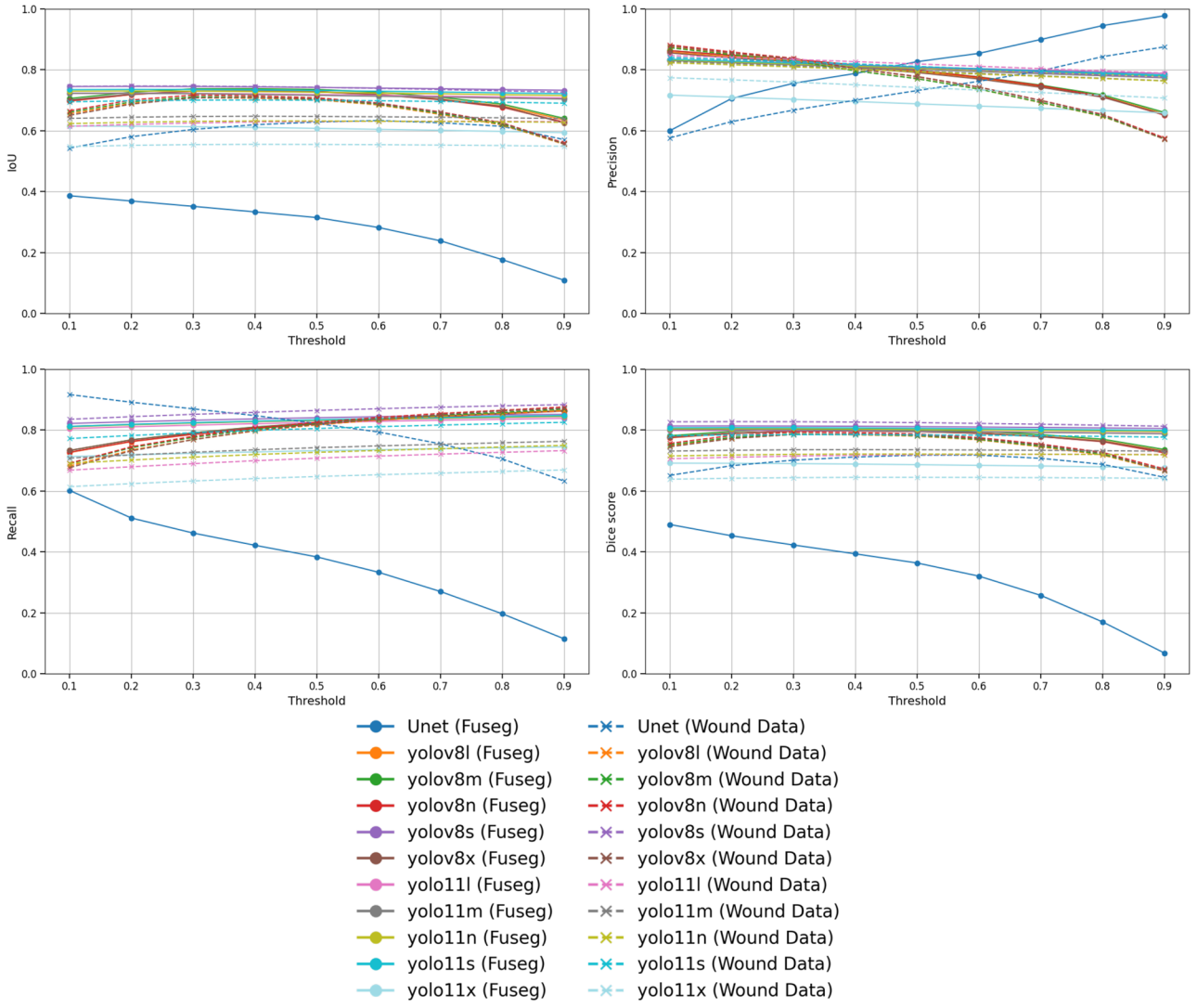

Fig. 6 Results when trained on both databases and tested on one database (the color refers to the database used). Metric graphs averaged over 5 folds (top right: IoU, top left: precision, bottom right: recall, bottom left: dice coefficient) for Test 5 and 6

performance than U-Net, despite a moderate 10–15% drop in IoU when changing domains. YOLO11 exhibits greater variability across versions, but stands out for its remarkable inter-segment stability, with only 1.9% variation in IoU, compared with 8.2% for YOLOv8 and 19.4% for U-Net, as shown in Fig. 5. This stability makes YOLO11 particularly suitable for clinical applications, where fine-tuning of the decision threshold is not always possible. So while YOLOv8 guarantees high accuracy, YOLO11 offers increased robustness to domain changes.

Fold-level statistical tests reflect the increased variability induced by domain shift. Paired t-tests favoured YOLOv8 and YOLO11 over U-Net for IoU and DSC in several configurations ($p < 0.05$). This pattern indicates that, although absolute performance differences remain substantial, variability across datasets widens confidence intervals and should be considered when interpreting generalization performance.

Test 5 & 6 - trained on both databases, tested on one database

In Test 5, the models were trained on the merged FUSeg and Wound Data databases and then evaluated on FUSeg only, simulating a realistic multi-source learning scenario aimed at improving robustness, as shown in Fig. 6 and Table 10. U-Net shows only a marginal improvement over Test 1 (Sect. 3), with an IoU of 38.1% (95% CI: 36–40%) and a DSC of 48.4% (95% CI: 46–51%), confirming that its architectural limitations persist, even with enriched training data. In contrast, all YOLOv8 models show significant performance improvements, with YOLOv8s leading the way (IoU = 78% (95% CI: 76–80%), DSC = 85% (95% CI: 83–87%)). The YOLO11 models cover a larger

range (IoU from 61.5% to 73.6%), but YOLO11s stands out with the best overall result: a precision of 82.8% (95% CI: 81–84%), a recall of 81.9% (95% CI: 80–83%), and a DSC of 80.8% (95% CI: 80–82%). The gap with U-Net becomes particularly marked, exceeding 30% in IoU and DSC, making the latter unsuitable for clinical use. These results highlight the importance of data availability for ML tasks.

Test 6 evaluates the opposite scenario, which presents a larger anatomical context and increased visual variability. The results are presented in Fig. 6 and Table 11. U-Net performed best here, with an IoU of 57.3% (95% CI: 45–69%) and a DSC of 65% (95% CI: 55–79%), indicating that learning on diversified data enables it to adapt more effectively to the visual complexity of this game. However, its results are still significantly lower than those of the YOLO models, with a difference of up to 20% in IoU. YOLOv8n performed best in this test (IoU = 71.7% (95% CI: 71–73%), DSC = 79.3% (95% CI: 78–80%)), closely followed by the other variants (IoU between 70.8% and 71.7%, and DSC between 78.6% and 79.3%). YOLO11s also stands out, with an IoU of 70.1% (95% CI: 67–73%), a precision of 81.9% (95% CI: 79–85%), a recall of 79.9% (95% CI: 76–84%), and a DSC of 78.5% (95% CI: 76–81%). However, the performance of the YOLO11 versions remains more dispersed, with IoU ranging from 55.5% to 70.1%. While YOLOv8 achieves a higher IoU overall, YOLO11 retains better inter-threshold stability, an important factor for clinical applications that require constant reliability without fine adjustments to thresholds. YOLO11 varies little between thresholds (1.35%), compared with 12.6% for YOLOv8 and 20.21% for U-Net.

Fold-level statistical analyses confirmed the magnitude of these improvements. For both datasets, paired t-tests showed highly significant gains ($p < 0.001$) in IoU and DSC for the best YOLOv8 and YOLO11 models compared with U-Net, with Wilcoxon tests yielding consistent conclusions despite the small sample size ($n = 5$).

Finally, Fig. 7 provides a qualitative summary of the segmentation outputs obtained in Tests 1–6. Each row corresponds to one test, and each column to U-Net, YOLOv8n and YOLO11s. Images for Tests 1, 4 and 5 originate from FUSeg, whereas those for Tests 2, 3 and 6 come from Wound Data database. This overview visually illustrates the effect of training-domain differences: U-Net often produces fragmented or unstable masks under domain shift, while YOLO-based models generate more coherent boundaries, particularly when trained on both datasets.

Interestingly, models trained on the Wound Data database consistently achieved higher IoU and DSC scores when tested on the FUSeg database compared to models trained on FUSeg and tested on the Wound Data database. This asymmetry suggests that training on visually more complex and contextually rich images enhances the model's ability to generalize to simpler, more localized images.

Training models on both databases (Test 5 and 6) (Sect. 3) generally yielded better results than training on a single database. This highlights the importance of data diversity and variability for machine learning algorithms.

Finally, regarding computing time, the YOLO11 model achieved an average inference speed of 16.6 ms, while YOLOv8 reached 21.5 ms. In contrast, the U-Net model took an average of 120.3 ms, making it seven times slower than YOLO11. This further highlights the advantages of using the YOLO segmentation module (See Table 12).

Discussion

Overall, the results of the six tests confirm the consistent superiority of the YOLO models over the U-Net reference model for segmenting chronic wounds. Despite its proven track record in biomedical segmentation, it suggests that its encoder–decoder structure may lack the architectural depth and contextual modeling required to segment irregular wound contours in more heterogeneous datasets. YOLOv8m and YOLOv8n emerge as the most consistent performers, achieving top scores in four out of six tests, with YOLOv8s reaching the highest IoU (71.7%) and DSC (79.3%) in Test 6 (Sect. 3). YOLO11n and YOLO11s also prove highly competitive, with YOLO11s standing out in the multi-source training configurations (Tests 5 and 6) for its excellent balance between precision, recall, and DSC. Although YOLOv8 achieves better raw scores overall, the YOLO11 models exhibit better inter-threshold stability, which is crucial for reliability and repeatability, thereby reducing dependence on the threshold parameter.

Beyond numerical performance, the fold-level statistical analyses support the robustness of the YOLO architectures. Paired t-tests systematically showed that the best-performing YOLOv8 and YOLO11 variants achieved significantly higher IoU and DSC values than U-Net in most configurations (typically $p < 0.01$ when models were trained and tested within the same domain, and $p < 0.05$ in cross-domain settings). These results confirm that the observed performance gaps are not attributable to sampling variability across folds. Given the limited number of folds ($n = 5$), Wilcoxon signed-rank tests were also computed as a non-parametric complement. As expected in small-sample settings, Wilcoxon tests were more conservative and did not always reach statistical significance, despite effect sizes clearly favouring YOLO. This discrepancy reflects the lower statistical power of rank-based methods when the number of paired observations is small, rather than an absence of true performance differences. Importantly, both tests aligned in directionality across all comparisons, consistently

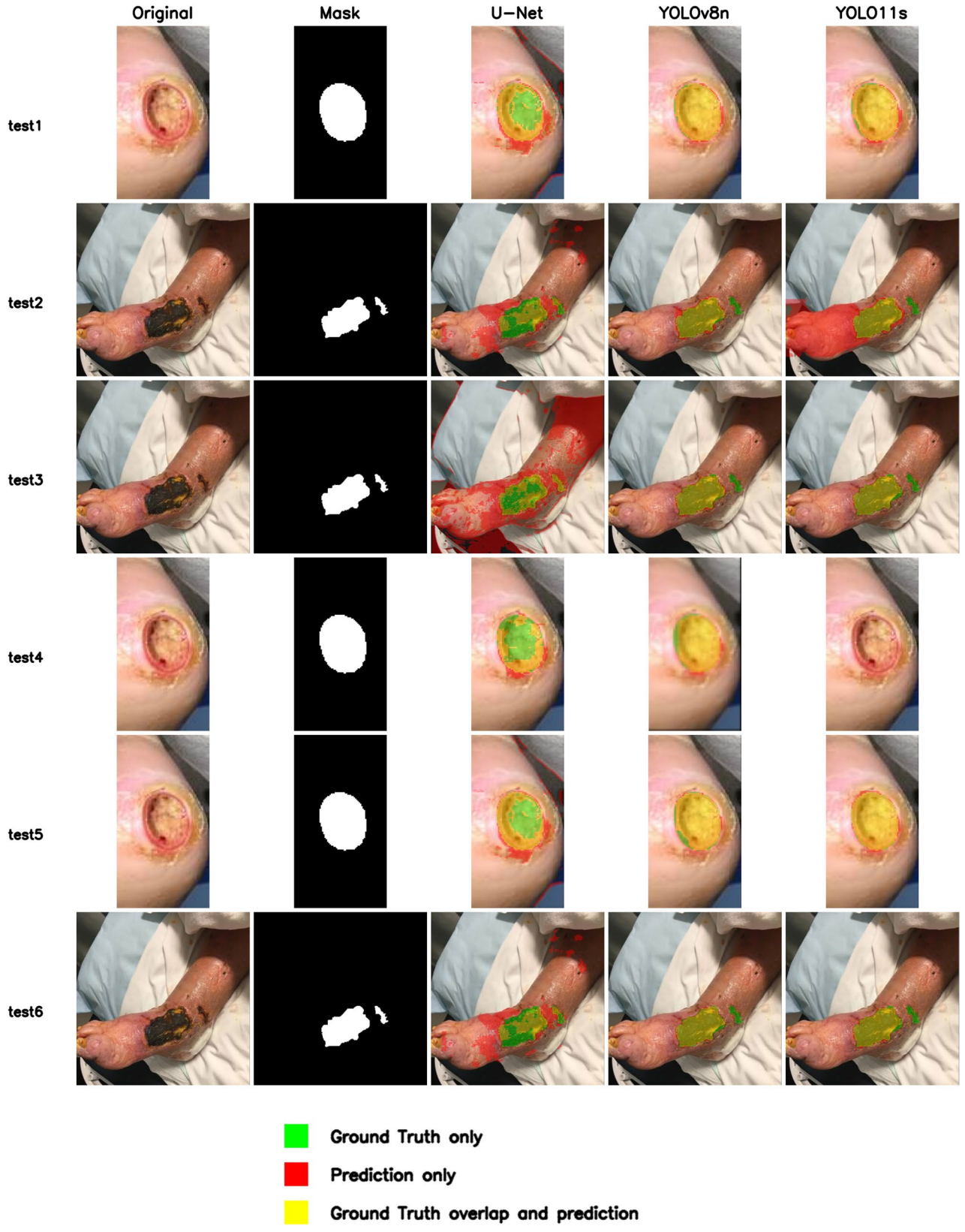

Fig. 7 Qualitative comparison across Tests 1-6. Each row corresponds to one training-testing configuration and each column reports the predicted segmentation masks produced by U-Net, YOLOv8n, and YOLO11s (overlaid against the ground truth)

Table 12 Global comparison of the best U-Net, YOLOv8 and YOLO11 configurations in the most realistic setting (multi-source training, evaluation on wound data – Test 6)

Model	Parameters (M)	IoU (%)	DSC (%)	Precision (%)	Recall (%)	Inference time (ms)
U-Net	1.9	57.3	65.0	72.6	78.0	120.3
YOLOv8s	11.2	74.6	82.8	83.3	84.4	21.5
YOLO11s	9.4	70.1	78.5	81.9	79.9	16.6

indicating better segmentation performance for YOLO models. Taken together, the confidence intervals, fold-level means, and paired statistical tests all support the conclusion that YOLO architectures outperform U-Net in terms of robustness across folds and training conditions, even when accounting for the inherent variability introduced by cross-dataset generalisation.

The FUSeg database differs from Wound Data in its content: FUSeg images focus solely on the wound and its region of interest, with no additional context, whereas Wound Data pictures encompass the entire foot. This structural difference may indicate an overfitting to the specific characteristics of FUSeg, resulting in a lack of generalization when applied to unseen data from the Wound Data database. It highlights the importance of datasets encompassing a broad range of wound types and characteristics. Furthermore, the current models were trained on publicly available datasets captured under standardized conditions (e.g., close-up images with controlled lighting, angles, and zoom). However, images frequently exhibit strong colour variations, irregular wound boundaries, shadows and non-uniform illumination.These factors introduce substantial visual variability that can hinder segmentation robustness and exacerbate domain-shift effects. The qualitative results presented in Fig. 7 illustrate some of these challenges, particularly for U-Net, which is more sensitive to such variations. To address these limitations, strategies such as multi-dataset training, domain adaptation techniques, and ongoing model validation in real-world conditions should be considered [76]. Similar findings were reported by Monroy et al., who also identified domain overfitting issues in cross-dataset evaluations [53]. Although they did not investigate dataset fusion during training, our results clearly show that combining diverse datasets significantly enhances segmentation performance and generalisability.

Merging the FUSeg and Wound Data datasets demonstrated a significant improvement in model performance. Indeed, when comparing the YOLOv8n results when training and testing on FUSeg alone with the ones when training on the fusion of FUSeg and Wound Data and testing on FUSeg, the IoU increased from 69.4% to 71.7% when trained on the merged datasets and tested on FUSeg, with precision and recall improving from 80.4% and 76.8% to 83.8% and 77.9%, respectively and a DSC from 76.7% to 79.3%. Similarly, YOLO11n improved from an IoU of 70.7% to 73.1%, and achieved higher precision (from 81.9% to 82.6%), recall (from 79.4% to 81.7%), and DSC (from 78.6% to 80.5%). This enhancement underscores the importance of dataset diversity in enhancing generalization and model adaptability. Cross-dataset validation (training on one database and testing on another) further underscored the challenges of generalization across datasets with different characteristics. A similar study by Monroy et al. [53] also observed the limitations of generalizing models between two databases, but did not explore the benefits of dataset merging. Our results confirm that combining diverse datasets is a key strategy for improving segmentation performance.

Compared to earlier versions of the YOLO architecture, YOLOv8 and YOLO11 achieved superior performance in both accuracy and generalisability. Notably, the latest YOLO model used in the literature is the YOLOv7 version, as employed by Ferdinand et al. for skin burn detection and classification [54], where the model achieved a precision of 46.3% and a recall of 67.2%. Older YOLO models, such as YOLOv2 combined with ShuffleNet [42], achieved a high mAP score of 94% for DFU detection on a proprietary dataset. YOLOv3 performs similarly, with an mAP of 93.9% on the AZH Wounds dataset [43]. The hybrid YOLOv4 model used for DFU detection on the DFUC dataset achieved a DSC of 55.6% [51], while YOLOv5 achieves a precision of 78.1% and a recall of 68.5% for PU detection on a proprietary dataset [56]. Our YOLOv8 and YOLO11 models achieved superior metrics with an IoU of 74.5% and 73.6%, a precision of 82.6% and 82.8%, a recall of 82.7% and 81.9%, and a DSC of 81.4% and 80.8%, respectively, for DFU segmentation on the combined AZH Wounds and FUSeg datasets. This comparison demonstrates that new YOLO versions outperform their predecessors in chronic wound segmentation. However, it is essential to acknowledge that differences exist in the metrics used for performance evaluation and the datasets employed for training and testing. Such differences can affect the direct comparability of results, highlighting the need for standardized benchmarks in the field While recent 2023–2025 studies have explored advanced U-Net and CNN–Transformer variants in domains such as polyp [25] and brain tumour segmentation, they do not address native YOLO-based segmentation of chronic wounds nor cross-dataset evaluation on heterogeneous wound images. Our benchmark of U-Net, YOLOv8 and YOLO11 therefore provides a complementary perspective focused on robustness and deploy

Table 13 Comparison of segmentation and detection performance of different YOLO versions reported in the literature and the best YOLOv8/YOLO11 results from this study

Study	Model	Wound type	Dataset	Metrics
Amin et al. [42]	YOLOv2 + ShuffleNet	Diabetic foot ulcer	Proprietary	mAP = 94.0%
Anisuzzaman et al. [43]	YOLOv3	Diabetic foot ulcer	AZH Wounds	mAP = 93.9%
Kucharski et al. [51]	Hybrid YOLOv4	Diabetic foot ulcer	DFUC dataset	DSC = 55.6%
Aldughayfiq et al. [56]	YOLOv5	Pressure ulcer	Proprietary	p = 78.1%, R = 68.5%, mAP = 76.9%
Ferdinand et al. [54]	YOLOv7	Skin burns	Kaggle + Roboflow	p = 46.3%, R = 67.2%
Wang et al. [50]	MobileNetV2 + U-Net	Diabetic foot ulcer	FUSeg	p = 89.0%, R = 91.3%, DSC = 90.5%
This study	**YOLOv8s**	Diabetic foot ulcer	AZH Wounds + FUSeg	IoU = 74.5%, p = 82.6%, R = 82.7%, DSC = 81.4%
This study	**YOLO11s**	Diabetic foot ulcer	AZH Wounds + FUSeg	IoU = 73.6%, p = 82.8%, R = 81.9%, DSC = 80.8%

ability. To provide a clearer quantitative overview, Table 13 summarises the best-performing models reported in the literature for chronic-wound analysis together with the best YOLOv8 and YOLO11 results from our study.

Significant methodological differences were previously highlighted by comparing the U-Net and YOLO models. For example, Wang et al. [50] achieved higher precision, recall, and DSC (89.04%, 91.29%, and 90.15%, respectively) using a U-Net model trained on the FUSeg dataset. In contrast, our study achieved the best U-Net scores of only 70.4% precision, 63.2% recall, and 62.2% DSC. This notable difference can be attributed to the advanced techniques employed by Wang et al., including data augmentation, transfer learning with pre-trained weights from the Pascal VOC dataset [77], and post-processing steps such as hole filling and small region removal, which enhanced segmentation accuracy by suppressing noise and refining boundaries. By contrast, our study aimed to compare the YOLOv8, YOLO11, and U-Net models under fair conditions, with both models trained from scratch on the same limited dataset, using identical preprocessing and without data augmentation or post-processing, to isolate the intrinsic contribution of each architecture. This controlled setup enables a fair assessment of baseline performance, demonstrating that YOLO achieves robust and competitive performance. While these overall results highlight the superior performance of YOLO-based models, it is nevertheless important to consider the limitations specific to each architecture. YOLOv8 and YOLO11 sometimes miss very small or low-contrast lesions, particularly when the wound occupies only a small region of the image or blends with the surrounding skin. This behaviour stems from their detection-oriented architecture, in which subtle structures may not generate sufficient activation to produce a reliable segmentation mask. Conversely, U-Net tends to over-segment the wound area by including adjacent erythematous, calloused, or inflammatory skin regions. This reflects its pixel-wise prediction strategy and spatial smoothing, which favour sensitivity but can reduce boundary specificity. These complementary error modes help explain the observed differences: YOLO achieves higher precision, whereas U-Net maintains increased sensitivity at the cost of occasional overestimation in complex skin contexts. YOLOv8 and YOLO11 consistently outperformed U-Net for chronic wound segmentation. This advantage can be attributed to YOLOv8 and YOLO11's architectural complexity, which integrates efficient feature extraction mechanisms for accurate detection and segmentation. Furthermore, YOLOv8 and YOLO11 achieved, respectively, an average inference time of 21.5 ms and 16.6 ms, significantly faster than U-Net's 120.3 ms, making them strong candidates for integration into real-time clinical tools.

Our comparative analysis used identical training settings across models, including a fixed image size (640 × 640) and batch size (16), as well as the default YOLO hyperparameters. This unified configuration was chosen to avoid bias in favor of larger models, which tend to converge more slowly, and to allow for a direct comparison of performance under consistent conditions. While we acknowledge that the number of epochs could be optimized for each architecture, the primary objective was not to reach absolute peak performance, but to compare their relative performance and robustness of each model in a standardized setting. In line with the controlled scope of this benchmark, models were evaluated without transfer learning, domain adaptation, or training on reduced data. This design ensures that the reported cross-dataset results reflect the intrinsic generalization ability of each architecture rather than the effects of fine-tuning or sampling strategies specific to each dataset. Likewise, the scope of this benchmark was intentionally restricted to YOLO-based one-stage architectures and the U-Net reference model, excluding other competitive frameworks such as Mask R-CNN, DeepLabV3+, or U-Net++ to maintain a focused and methodologically coherent comparison between real-time and encoder–decoder paradigms.

YOLO11 is proposed as an improved version of YOLOv8. Although our results indicate that YOLOv8 slightly outperforms YOLO11, when comparing the best performances of the two models on the same test. Additionally, YOLO11 is more stable than YOLOv8 models

across different thresholds, reducing variability and dependency on the threshold parameter. This is crucial for minimizing errors and ensuring consistent performance in clinical applications.

Automating wound diagnosis through AI-driven segmentation tools has significant clinical implications. Tools such as YOLOv8 and YOLO11 can be integrated into mobile applications or software used by healthcare professionals. These tools would capture images, automatically segment the injured area, and extract relevant morphological features (surface, shape, colour). This type of automated analysis could considerably enhance existing tools, such as the ImitoWound application, which has already been used to measure the surface area of wounds to monitor their evolution [78]. By integrating segmentation algorithms, it would be possible to further automate diagnosis and improve the accuracy and reproducibility of assessments. Such a system would facilitate longitudinal monitoring, telemedicine, and continuity of care while reducing the need for face-to-face consultations. Ultimately, these systems could also combine the visual information extracted with additional clinical data (type of wound, pathological context, and evolution) to support a more comprehensive and personalized diagnosis. This approach would help to optimise care, speed up therapeutic decisions, and reduce the burden on healthcare systems.

A significant issue generally encountered with all DL models in healthcare is their interpretability. Clinicians need to understand how these algorithms make decisions to trust and use them effectively. Techniques such as Layer-wise Relevance Propagation (LRP) [79] or Grad-CAM [80] visualize the areas of the image that have most influenced the model's prediction, generating heatmaps that indicate the regions of interest activated during processing. Other tools, such as SHapley Additive exPlanations (SHAP) [81] and Local Interpretable Model-Agnostic Explanations (LIME) [82], complementary global or local insights into model behaviour. Integrating such interpretability frameworks in future developments may help assess how segmentation models focus on clinically relevant wound structures and support their adoption in real-world clinical settings.

In the case of YOLO algorithms, these techniques could be adapted to better visualise the regions activated during detection and segmentation, paving the way for a better understanding of their decisions. Incorporating these approaches into future developments would increase the transparency of the models, facilitating their integration into real clinical environments.

Conclusion

This study comprehensively evaluates the YOLOv8 and YOLO11 algorithms for chronic wound segmentation, establishing their superiority over the benchmark U-Net model. Both YOLO architectures delivered high and consistent performance across six tests, with YOLOv8n and YOLOv8m achieving the best results in terms of accuracy, while YOLO11s and YOLO11n stood out for their inter-threshold stability, a key factor for clinical reliability. The best-performing models achieved IoU scores of up to 71.7%, precision and recall rates exceeding 78%, and a DSC above 79.3%, confirming their clinical relevance. Training on a merged dataset combining FUSeg and Wound Data led to the best outcomes, highlighting the importance of data diversity in improving model generalization. In contrast, cross-dataset experiments revealed persistent challenges in domain transfer, particularly when models trained on cropped images (FUSeg) were tested on more complex, contextual ones (Wound Data). These results underline the need to train segmentation models on visually varied and representative clinical datasets to ensure robustness and adaptability in real-world scenarios.

This research not only confirms the effectiveness of YOLOv8 and YOLO11 in segmentation tasks but also lays the groundwork for more comprehensive AI-driven wound diagnosis systems that integrate advanced tools for greater clinical impact.

Future work should aim to validate these findings in broader, real-world contexts and across diverse clinical settings through large-scale, multi-institutional studies. Expanding datasets to include images acquired under heterogeneous conditions, as well as integrating additional multimodal wound-related metadata (e.g., wound aetiology, tissue composition, patient comorbidities, and temporal evolution), would enable a more comprehensive and clinically meaningful assessment of model performance. Approaches that support large-scale training while respecting patient confidentiality, such as privacy-preserving and federated learning frameworks, could also play an important role in enabling collaborative model development across institutions. Beyond chronic foot ulcers, these architectures could be adapted to other dermatological and tissue assessment tasks, including burn severity grading, pressure ulcer staging, or ulcer morphology classification, thereby extending their clinical utility across specialties. In parallel, systematic ablation studies of YOLOv8 and YOLO11 architectural modules (e.g., C2f, C3k2, C2PSA, SPFF) should be investigated to quantify their individual contribution to segmentation performance and to clarify the design trade-offs that govern accuracy, stability, and computational efficiency. Such analyses would help determine which components are essential for robust wound segmentation and which

may be simplified or optimized for clinical deployment. It should also explore domain adaptation, transfer learning, and reduced-data training regimes to assess how these architectures behave under the data scarcity and domain shifts typical of clinical imaging. Together, these developments would clarify which architectural components are essential, how performance degrades or is preserved under constrained conditions, and how to optimise these models for real-world deployment. Beyond methodology, broader clinical validation will be necessary. Expanding datasets to include images acquired under heterogeneous real-world conditions, and incorporating multimodal wound metadata (e.g., wound aetiology, tissue composition, patient comorbidities, temporal evolution), would enable more comprehensive and clinically grounded performance assessment. Approaches such as privacy-preserving or federated learning could further support large-scale, multi-institutional collaboration while maintaining patient confidentiality. Future studies should further extend the benchmark to additional state-of-the-art frameworks, such as Mask R-CNN, DeepLabV3+, and U-Net++, to contextualize YOLO's performance within the broader medical imaging landscape. In summary, this study represents a significant step forward in AI-driven wound care, demonstrating how advanced segmentation algorithms, such as YOLOv8 and YOLO11, can transform the management of chronic wounds.

List of abbreviations

DFU	Diabetic foot ulcer
VLU	Venous leg ulcer
PU	Pressure ulcer
DL	Deep learning
CNN	Convolutional Neural Network
VGG	Visual Geometry Group
FCN	Fully Convolutional Network
DSC	Dice Similarity Coefficient
IoU	Intersection-over-union score
cGAN	Conditional generative adversarial network
YOLO	You OnlyLook Once
mAP	Mean Average Precision
PGI	Programmable Gradient Information
GELAN	Generalized Efficient Layer Aggregation Network
AI	Artificial intelligence
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
SPFF	Spatial Pyramid Fast Fusion
CSP	Cross Stage Partial
LRP	Layer-wise Relevance Propagation
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-026-02266-7.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements
Non applicable.

Author contributions
I.M. – conceptualization of the study, algorithm implementation, data processing, figure and table generation, and manuscript writing. Z.A. – conceptualization of the study, algorithm implementation and data processing, figure generation, and manuscript writing. H.A. – conceptualization of the study, algorithm implementation and data processing. S.S. and L.V.V. - medical expertise and analysis. C.V. algorithm expertise and analysis. M.A.N.Y., and J.A.Y.S. – manuscript review and feedback. F.Q. and A.N. – general supervision, interpretative analysis and manuscript review. All authors read and approved the final manuscript before submission.

Funding
This study was founded by AI-enabling Smart Wound carE Electronics Patches 2 (AI-SWEEP-2) project, with the support of the Belgian government, represented by Mr. Mathieu Michel, Secretary of State for Digitalization, responsible for Administrative Simplification, Privacy Protection, and Building Management, Deputy to the Prime Minister.

Data availability
The databases used in this project are publicly available and do not require specific access. The results of cross validation are provided within the supplementary information files (Supplementary file S1).

Declarations

Ethics approval and consent to participate
Non applicable.

Consent for publication
Non applicable.

Competing interests
The authors declare no competing interests.

Received: 7 August 2025 / Accepted: 3 March 2026
Published online: 19 March 2026

References

1. Wound Care Market by Product, Devices, Wounds, End User, and Region – Global Forecast to 2028. 2023.
2. Martins-Green M. Cutaneous chronic wounds: a worldwide silent epidemic. Open Access Gov. 2023;38:51–53. https://doi.org/10.56367/OAG-038-10647.
3. Nunez K. What you need to know about the causes of and treatments for skin ulcers. healthline. 2019.
4. Snyder RJ, Kirsner RS, Warriner RA, Lavery LA, Hanft JR, Sheehan P. Consensus recommendations on advancing the standard of care for treating neuropathic foot ulcers in patients with diabetes. Ostomy Wound Manage. 2010;56:1–24.
5. Yazdanpanah L, Shahbazian H, Nazari I, Arti HR, Ahmadi F, Mohammadianinejad SE, et al. Incidence and risk factors of diabetic foot ulcer: a population-based diabetic foot cohort (adfc study)-two-year follow-up study. Int J Endocrinol 2018. 2018;7631659. https://doi.org/10.1155/2018/7631659.
6. Patel NP, Labropoulos N, Pappas PJ. Current management of venous ulceration. Plast Reconstructive Surg. 2006;117:254–60. https://doi.org/10.1097/01.prs.0000222531.84704.94.
7. Landi F, Onder G, Russo A, Bernabei R. Pressure ulcer and mortality in frail elderly people living in community. Arch Gerontol Geriat. 2007;44(Suppl 1):217–23. https://doi.org/10.1016/j.archger.2007.01.030.
8. Banerjee T, Singh DP, Kour P, Swain D, Mahajan S, Kadry S, et al. A novel unified inception-u-net hybrid gravitational optimization model (uigo) incorporating automated medical image segmentation and feature selection

for liver tumor detection. Sci Rep. 2025;15. https://doi.org/10.1038/s41598-025-14333-0.
9. Singh DP, Kour P, Banerjee T, Swain D. A comprehensive review of various machine learning and deep learning models for anti-cancer drug response prediction: comparative analysis with existing State of the art methods. Springer; 2025. https://doi.org/10.1007/s11831-025-10255-2.
10. Jennett PA, Hall LA, Hailey D, Ohinmaa A, Anderson C, Thomas R, et al. The socio-economic impact of telehealth: a systematic review. J Educ Chang Telemed Telecare. 2003;9:311–20. https://doi.org/10.1258/135763303771005207.
11. Nagle SM. Stevens K.A., Wilbraham, S.C Wound assessment 2024.
12. Cassidy B, Reeves ND, Pappachan JM, Gillespie D, O'Shea C, Rajbhandari S, et al. The dfuc 2020 dataset: analysis towards diabetic foot ulcer detection. touchREVIEWS Endocrinol 17. 2021;5. https://doi.org/10.17925/EE.2021.17.1.5.
13. Banerjee T, Singh DP, Kour P. Advances in deep neural, transformer learning, and kernel-based methods for diabetic retinopathy detection: a comprehensive review. Springer; 2025. https://doi.org/10.1007/s11831-025-10376-8.
14. Kanopoulos N, Vasanthavada N, Baker RL. Design of an image edge detection filter using the sobel operator. IEEE J Solid-State Circuits. 1988;23:358–67.
15. Canny J. A computational approach to edge detection. IEEE T Pattern Anal. 1986;679–98.
16. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Intell Transp Syst Syst, Man, Cybern. 1979;9:62–66. https://doi.org/10.1109/TSMC.1979.4310076.
17. Adams R, Bischof L. Seeded region growing. IEEE T Pattern Anal. 1994;16:641–47. https://doi.org/10.1109/34.295913.
18. Wu J. Advances in K-means clustering: a data mining thinking. Berlin, Heidelberg: Springer; 2012.
19. Lantuejoul C. La squelettisation et son application aux mesures topologiques des mosaiques polycristallines. 1978.
20. Singh DP, Banerjee T, Mahajan S, Chandra KR, Kumar R, Phani S, et al. A comprehensive study on deep learning models for the detection of diabetic retinopathy using pathological images. Springer; 2025. https://doi.org/10.1007/s11831-025-10315-7.
21. Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC). Springer; 2020. https://doi.org/10.1007/978-3-030-17795-9, Volume 1.
22. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. 2015.
23. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. 2015.
24. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE T Pattern Anal. 2018;40:834–48. https://doi.org/10.1109/TPAMI.2017.2699184.
25. Ahamed MF, Islam MR, Nahiduzzaman M, Chowdhury MEH, Alqahtani A, Murugappan M. Automated colorectal polyps detection from endoscopic images using multiresunet framework with attention guided segmentation. Hum-Centric Intell Syst. 2024;4:299–315. https://doi.org/10.1007/s44230-024-00067-1.
26. Ahamed MF, Hossain MM, Nahiduzzaman M, Islam MR, Islam MR, Ahsan M, et al. A review on brain tumor segmentation based on deep learning methods with federated learning techniques. Elsevier Ltd. 2023. https://doi.org/10.1016/j.compmedimag.2023.102313.
27. LeCun Y, Bengio Y: In A, MA, ed. Convolutional networks for images, speech, and time series. Cambridge, MA: MIT Press; 1995. p. 3361.
28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014.
29. Girshick R. Fast r-cnn. 2015.
30. Ren S, He K, Girshick R, Sun J. Faster r-cnn: towards real-time object detection with region proposal networks. 2015.
31. Chang Y, Kim JH, Shin HW, Ha C, Lee SY, Go T. Diagnosis of pressure ulcer stage using on-device ai. Appl Sci (Switz). 2024;14. https://doi.org/10.3390/app14167124.
32. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2014.
33. Alabdulhafith M, Mahel ASB, Samee NA, Mahmoud NF, Talaat R, Muthanna MSA, et al. Automated wound care by employing a reliable u-net architecture combined with resnet feature encoders for monitoring chronic wounds. Front. Med. 2024;11:1310137. https://doi.org/10.3389/fmed.2024.1310137.
34. Doğru D, Özdemir GD, Özdemir MA, Ercan UK, Avşar NT, Güren O. An automated in vitro wound healing microscopy image analysis approach utilizing u-net-based deep learning methodology. BMC Med Imag. 2024;24:158. https://doi.org/10.1186/s12880-024-01332-2.
35. Moyà-Alcover G, Miró-Nicolau M, Munar M, González-Hidalgo M. A multitask deep learning approach for staples and wound segmentation in abdominal post-surgical images. 2023. p. 208–19. https://doi.org/10.1007/978-3-031-39965-7_18.
36. Borst V, Dittus T, Müller K, Kounev S. Early explorations of lightweight models for wound segmentation on mobile devices. 2024.
37. Ahamed MF, Syfullah MK, Sarkar O, Islam MT, Nahiduzzaman M, Islam MR, et al. Irv2-net: a deep learning framework for enhanced polyp segmentation performance integrating inceptionresnetv2 and unet architecture with test time augmentation techniques. Sensors. 2023;23. https://doi.org/10.3390/s23187724.
38. Liu H, Hu J, Zhou J, Yu R. Application of deep learning to pressure injury staging. J Wound Care. 2024;33:368–78. https://doi.org/10.12968/jowc.2024.33.5.368.
39. Jacobson MJ, Masry ME, Arrubla DC, Tricas MR, Gnyawali SC, Zhang X, et al. Autonomous multi-modality burn wound characterization using artificial intelligence. Mil Med. 2023;188:674–81. https://doi.org/10.1093/milmed/usad301.
40. P J, K SKB, Jayaraman S. Automatic foot ulcer segmentation using conditional generative adversarial network (afseggan): a wound management system. PLoS Digit Health. 2023;2:0000344. https://doi.org/10.1371/journal.pdig.0000344.
41. Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics. 2023. https://github.com/ultralytics/ultralytics.
42. Amin J, Sharif M, Anjum MA, Khan HU, Malik MSA, Kadry S. An integrated design for classification and localization of diabetic foot ulcer based on cnn and yolov2-dfu models. IEEE Access. 2020;8:228586–97. https://doi.org/10.1109/ACCESS.2020.3045732.
43. Anisuzzaman DM, Patel Y, Rostami B, Niezgoda J, Gopalakrishnan S, Yu Z. Multi-modal wound classification using wound image and location by deep neural network. Sci Rep. 2022;12:20057. https://doi.org/10.1038/s41598-022-21813-0.
44. Silva LGR, Silva AAA, Guelfi AE, Rezende MN, Jesus Perez Alcazar J, Azevedo MT, et al. An application for wound diagnosis and treatment. Res Biomed Eng. 2022;38:629–45. https://doi.org/10.1007/s42600-022-00213-3.
45. Lau CH, Yu KH-O, Yip TF, Luk LY, Wai AKC, Sit T-Y, et al. An artificial intelligence-enabled smartphone app for real-time pressure injury assessment. Front Med Technol. 2022;4. https://doi.org/10.3389/fmedt.2022.905074.
46. Han A, Zhang Y, Li A, Li C, Zhao F, Dong Q, et al. Deep learning methods for real-time detection and analysis of wagner ulcer classification system. 2022 International Conference on Computer Applications Technology (CCAT). Guangzhou, China: IEEE; 2022, pp. 11–21). https://doi.org/10.1109/CCAT56798.2022.00010.
47. Adnan M, Asif M, Ahmad MB, Mahmood T, Masood K, Ashraf R, et al. An automatic wound detection system empowered by deep learning. J Phys Conf Ser. 2023;2547:012005. https://doi.org/10.1088/1742-6596/2547/1/012005.
48. Zhang X, Zhou X, Lin M, Sun J. Shufflenet: an extremely efficient convolutional neural network for mobile devices. 2017.
49. Carrión H, Jafari M, Bagood MD, Yang H-Y, Isseroff RR, Gomez M. Automatic wound detection and size estimation using deep learning algorithms. PLoS Comput Biol. 2022;18:1009852. https://doi.org/10.1371/journal.pcbi.1009852.
50. Wang C, Anisuzzaman DM, Williamson V, Dhar MK, Rostami B, Niezgoda J, et al. Fully automatic wound segmentation with deep convolutional neural networks. Sci Rep. 2020;10. https://doi.org/10.1038/s41598-020-78799-w.
51. Kucharski D, Kostuch A, Noworolnik F, Brodzicki A, Jaworek-Korjakowska J. DFU-Ens: end-to-end diabetic foot ulcer segmentation framework with–Vision transformer based detection. 2023. p. 101–12. https://doi.org/10.1007/978-3-031-26354-5_9.
52. Yap MH, Hachiuma R, Alavi A, Brüngel R, Cassidy B, Goyal M, et al. Deep learning in diabetic foot ulcers detection: a comprehensive evaluation. Comput Biol Med. 2021;135:104596. https://doi.org/10.1016/j.compbiomed.2021.104596.
53. Monroy B, Sanchez K, Arguello P, Estupiñán J, Bacca J, Correa CV, et al. Automated chronic wounds medical assessment and tracking framework based on deep learning. Comput Biol Med. 2023;165:107335. https://doi.org/10.1016/j.compbiomed.2023.107335.
54. Ferdinand J, Chow DV, Prasetyo SY. Automated skin burn detection and severity classification using yolo convolutional neural network pretrained model. E3S Web of Conferences. 2023. https://doi.org/10.1051/e3sconf/202342601076 426.

55. Noureldeen AM, Masoud KS, Almakhzoom OA. African Journal of Advanced Pure and Applied Sciences (AJAPAS) deep learning model for cutaneous leishmaniasis detection and classification using Yolov5. AJAPAS.
56. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Yolo-based deep learning model for pressure ulcer detection and classification. Healthcare. 2023;11:1222. https://doi.org/10.3390/healthcare11091222.
57. Jocher G, Chaurasia A, Qiu J. UltralyticsYolov. 2023;8. https://github.com/ultralytics/ultralytics.
58. Hussain M, Khanam R. In-depth review of yolov1 to yolov10 variants for enhanced photovoltaic defect detection. Solar. 2024;4:351–86. https://doi.org/10.3390/solar4030016.
59. Yolov1 to yolov10: A comprehensive review of yolo variants and their application in medical image detection. J Retailing Artif Intel Pract. 2024;7. https://doi.org/10.23977/jaip.2024.070314.
60. Alif MAR, Hussain M. Yolov1 to yolov10: a comprehensive review of yolo variants and their application in the agricultural domain. 2024.
61. Ultralytics: YOLO11 NEW. 2025. https://docs.ultralytics.com/models/yolo11/.
62. Sapkota R, Ahmed D, Karkee M. Comparing YOLOv8 and mask RCNN for object segmentation in complex orchard environments. 2024. https://arxiv.org/abs/2312.07935.
63. Sapkota R, Karkee M. Comparing yolov11 and yolov8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment. 2024.
64. Aziz F, Saputri DUE. Efficient skin lesion detection using yolov9 network. J Med Inf Technol. 2024;11–15. https://doi.org/10.37034/medinftech.v2i1.30.
65. Shah SMAH, Rizwan A, Atteia G, Alabdulhafith M. Cadfu for dermatologists: a novel chronic wounds & ulcers diagnosis system with dhunet (dual-phase hyperactive unet) and yolov8 algorithm. Healthcare (Switz). 2023;11. https://doi.org/10.3390/healthcare11212840.
66. Lucas Y, Niri R, Treuillet S, Douzi H, Castaneda B. Wound size imaging: ready for smart assessment and monitoring. Adv Wound Care. 2021;10:641–61. https://doi.org/10.1089/wound.2018.0937.
67. Goyal M, Yap MH, Reeves ND, Rajbhandari S, Spragg J. Fully convolutional networks for diabetic foot ulcer segmentation. 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2017, 618–23) https://doi.org/10.1109/SMC.2017.8122675.
68. Taher M: Wound Data. 2022. https://www.kaggle.com/datasets/mohamadtaher/wound-data/data.
69. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12:153–57. https://doi.org/10.1007/BF02295996.
70. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation. 1998;10:1895–923. https://doi.org/10.1162/089976698300017197.
71. Hidayatullah P, Syakrani N, Sholahuddin MR, Gelar T, Tubagus R. YOLOv8 to YOLO11: a comprehensive architecture In-depth comparative review a PREPRINT.
72. Pedro J. Detailed explanation of yolov8 architecture — part. 2023;1.
73. Ghosh A, Ghosh A. YOLO11: faster than you can imagine! 2024. https://learnopencv.com/yolo11/#aioseo-yolo11-architecture-and-whats-new-in-yolo11.
74. Jocher G, Qaddoumi B. Ultralytics YOLO docs. 2024. https://docs.ultralytics.com/reference/data/base/.
75. Singh DP, Banerjee T, Kour P, Malik R, Naidu GR, Durai CAD, et al. A comprehensive study of enhanced computational approaches for breast cancer classification: comparative analysis with existing State of the art methods. Springer; 2025. https://doi.org/10.1007/s11831-025-10414-5.
76. Kang Y, Zhao X, Zhang Y, Li H, Wang G, Cui L, et al. Improving domain generalization performance for medical image segmentation via random feature augmentation. Methods. 2023;218:149–57. https://doi.org/10.1016/j.ymeth.2023.08.003.
77. Everingham M, Gool LV, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. Int J Comput Vision. 2010;88:303–38. https://doi.org/10.1007/s11263-009-0275-4.
78. Ag I. Wound assessment tool - imitowound app. 2024.
79. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One. 2015;10:0130140. https://doi.org/10.1371/journal.pone.0130140.
80. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. 2016. https://doi.org/10.1007/s11263-019-01228-7.
81. Lundberg S, Lee S-I. A unified approach to interpreting model predictions. 2017.
82. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?": explaining the predictions of any classifier. 2016.

Publisher's Note