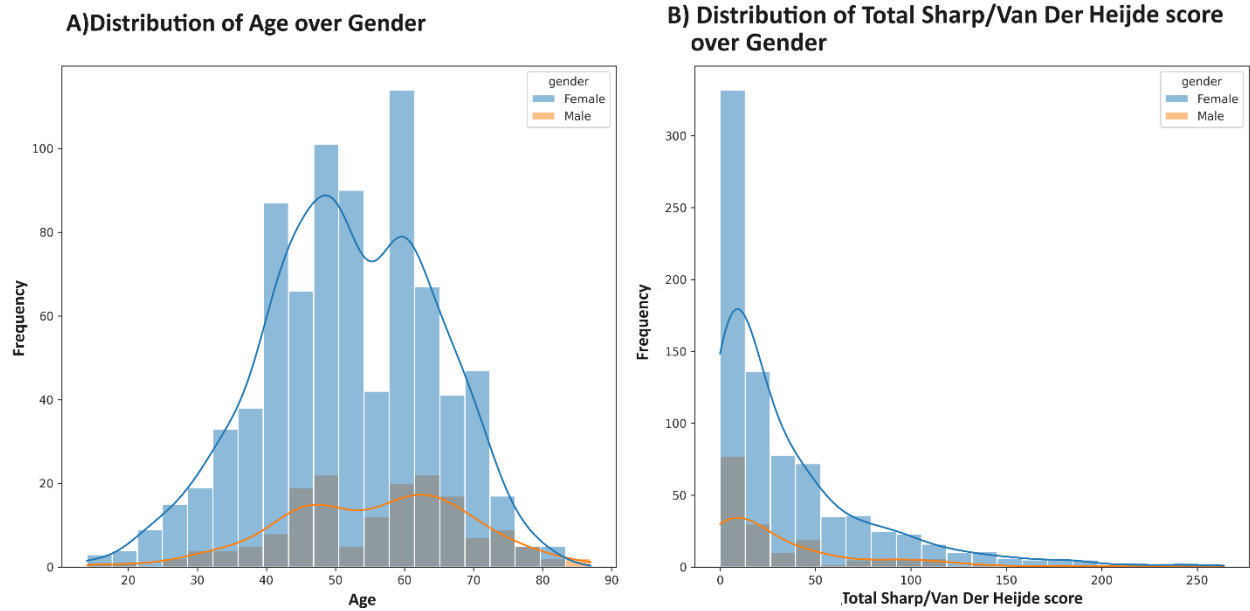


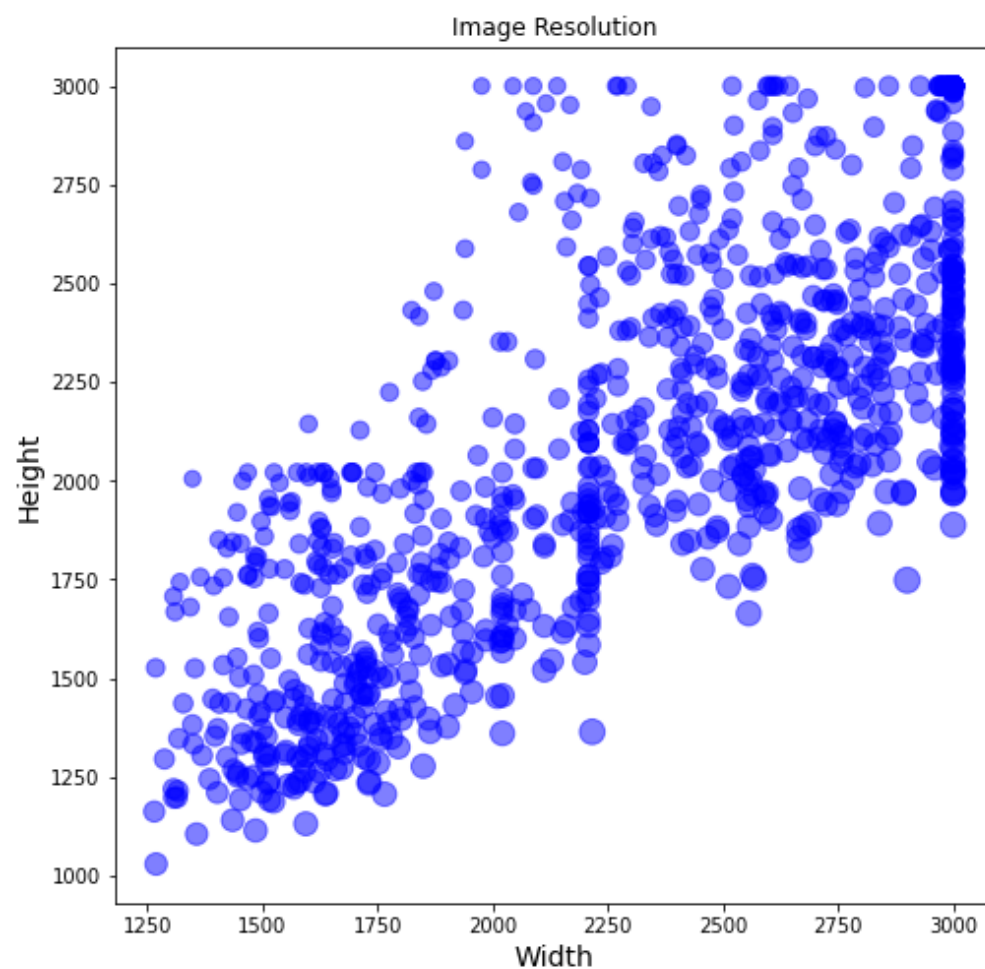
Multistage Deep Learning Methods for Automating Radiographic Sharp Score Prediction in Rheumatoid Arthritis

Hajar Moradmand, Ph.D., Department of Radiation Oncology, University of Maryland
School of Medicine, Baltimore, MD, USA

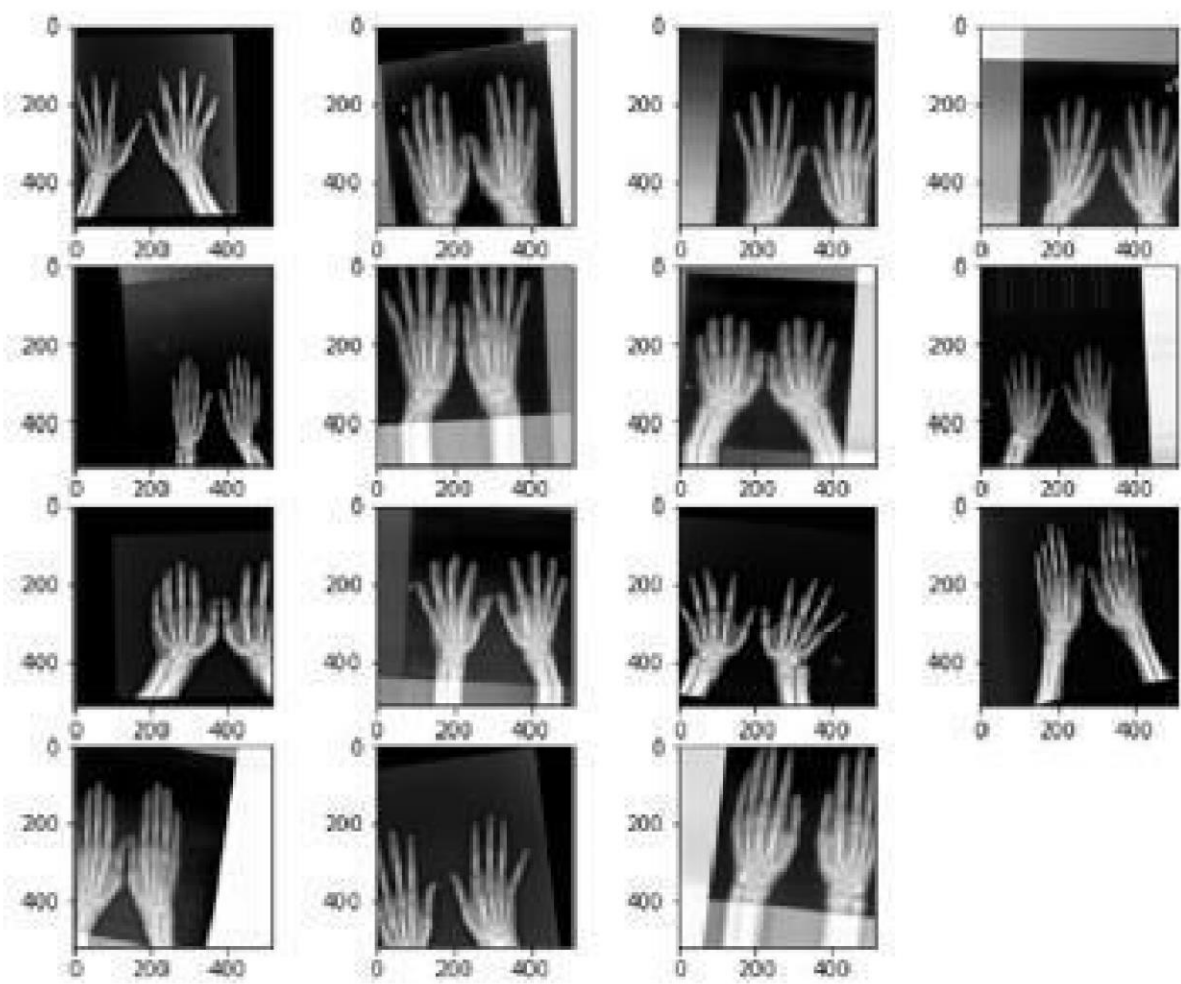
Lei Ren, Ph.D., Department of Radiation Oncology, University of Maryland School of
Medicine, Baltimore, MD, USA



Supplementary Figure S1. Distribution of Age and Sharp Score over Gender. a) Histogram of age by gender with KED curves shows a peak of around 40-50 and 60-70 years among female and male patients. b) Histogram of Overall Sharp Score over gender shows the unbalance of scores with a higher frequency of lower scores for both genders. Kernel density estimation (KDE) curves provide a smooth representation of the distributions, with histograms indicating the frequency distribution of age and Overall Sharp Score values.



Supplementary Figure S2. The scatter plot shows the different width and height of the original hand X-ray images in our database.



Supplementary Figure S3. Sample of data augmentation of re-orientation model (stage 1).

Hand Segmentation Using U-Net:

Semantic segmentation is a technique used to divide an image into different regions, assigning a label to each pixel to identify the type of object it represents. This process provides a high-level understanding of the image by classifying pixels into broad categories, such as "tree" or "cat," based on visual features and the context within the scene.

Semantic vs. Instance Segmentation:

While semantic segmentation labels each pixel with a general category, instance segmentation takes it a step further. It not only identifies the object class for each pixel but also distinguishes between individual instances of that class. For example, in an image containing multiple cats, instance segmentation would separately identify each cat, providing more detailed object information.

The U-Net Architecture for Image Segmentation:

The U-Net architecture, introduced by Ronneberger et al. (1), is a neural network specifically designed for image segmentation tasks, particularly in the biomedical field. It uses an encoder-decoder structure with skip connections that bridge the corresponding layers in both the encoder and decoder pathways.

Encoder: The encoder, or contracting path, consists of multiple convolutional layers that reduce the spatial dimensions of the input image while simultaneously increasing the number of feature channels. This process allows the network to capture more complex and abstract features at various scales.

Decoder: The decoder, or expansive path, consists of transposed convolutional layers (upsampling layers) that restore the image's spatial dimensions. This part of the network takes the compressed information from the encoder and progressively refines it to achieve pixel-level segmentation.

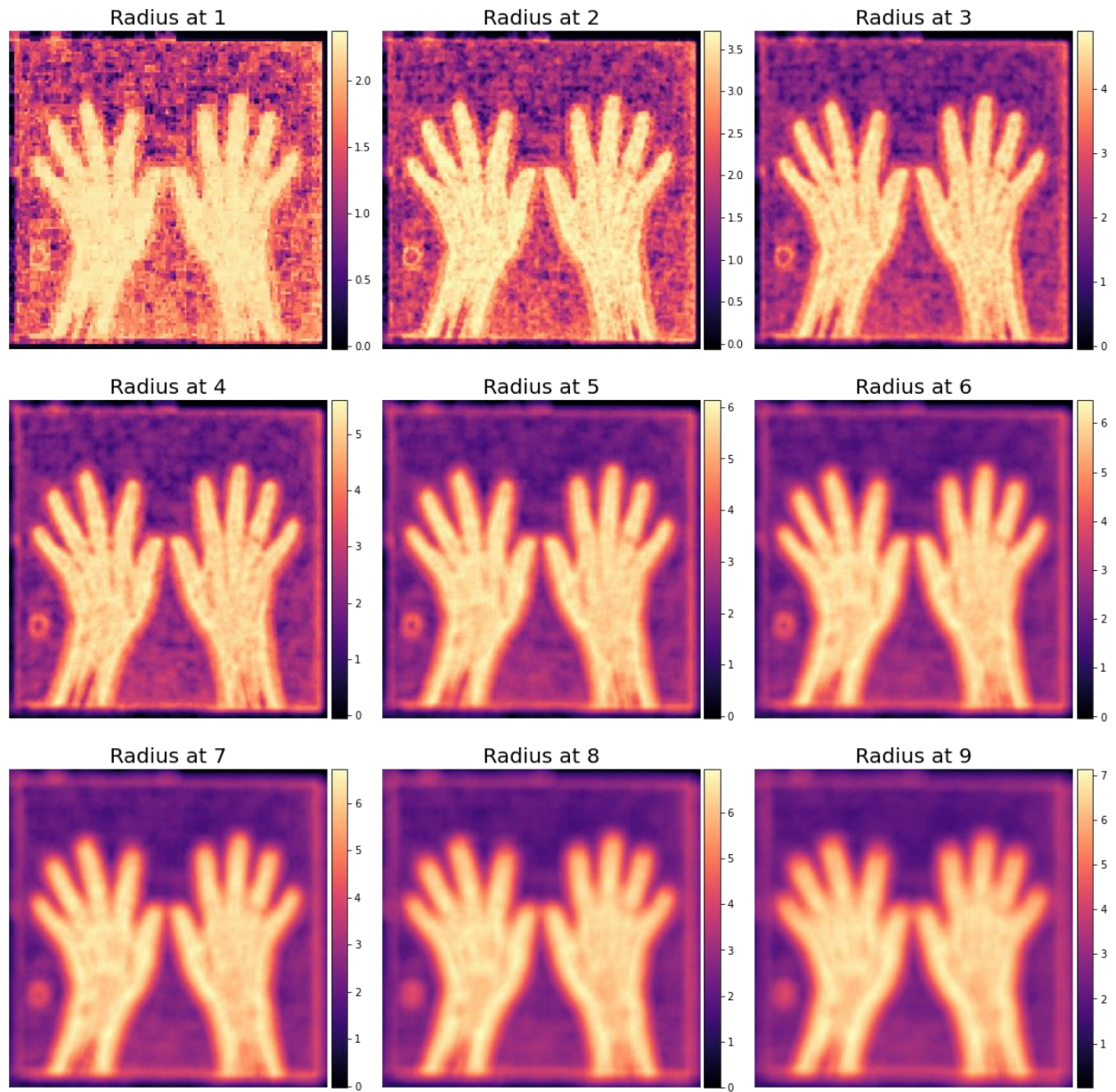
Skip Connections: A unique aspect of U-Net is its skip connections, which link the encoder layers to their corresponding decoder layers. These connections combine high-resolution, detailed information from the early stages of the encoder with the abstracted information in the decoder. This integration helps improve segmentation accuracy and prevent information loss during the upsampling process.

U-Net has become one of the most popular architectures for biomedical image segmentation since its introduction (2,3).

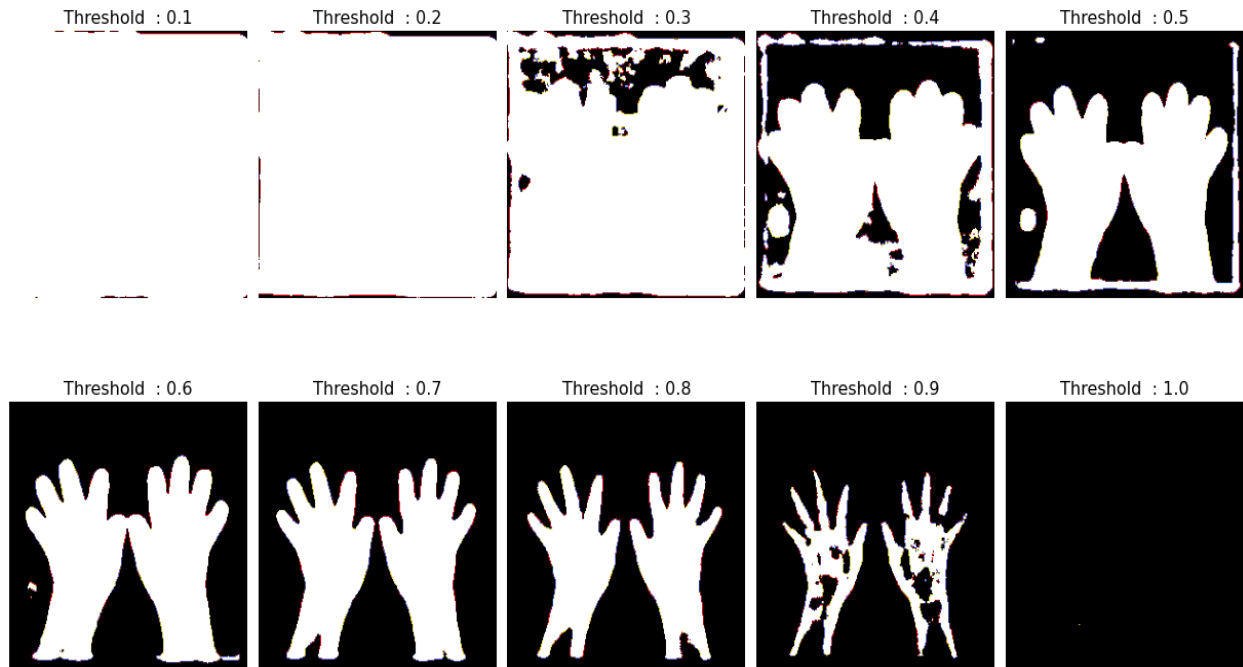
Entropy

Entropy is a statistical measure that quantifies the amount of uncertainty or randomness in a system. In the context of image processing, entropy can be used to characterize the amount of information present in an image. Specifically, image entropy is a measure of the randomness of pixel intensity values in an image. In image processing, a structuring element is a small, pre-defined shape that is used to probe and manipulate the structure of an image. The shape of the structuring element can have a significant impact on the results of various image processing operations, including entropy calculation.

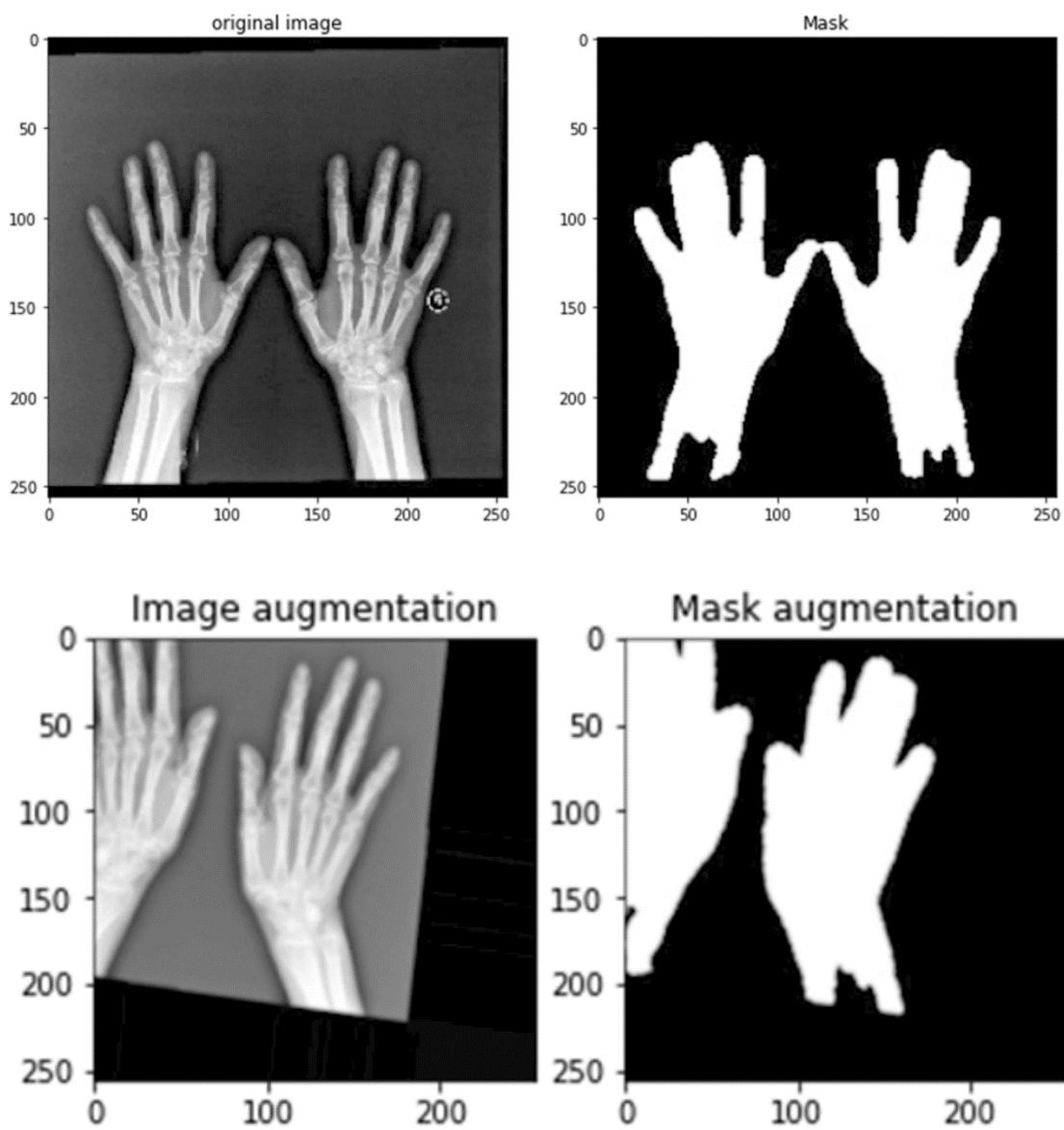
The figure illustrates the impact of threshold values on hand masking, with thresholds ranging from 0.1 to 1.0.



Supplementary Figure S4. Entropy Images. Circular neighborhoods with increasing radius (1 to 9) are used to find the optimum radius. The resulting entropy images are displayed using a color map, with the color scale shown on the right side of each subfigure. As the radius of the circular neighborhood increases, the resulting entropy images become smoother and exhibit fewer high-frequency noise artifacts. These entropy images can help detect edges and boundaries in the original grayscale image, as areas with high entropy correspond to regions of significant texture or detail.



Supplementary Figure S5. Effect of Threshold Variation on Hand Masking. The lowest threshold value (threshold 0.1) produces a mask that includes more pixels, resulting in white images. A moderate threshold value (threshold 0.5) yields a reasonably accurate hand mask, capturing the essential features of the hand. The highest threshold value (threshold 1) resulting in a black image.



Supplementary Figure S6. Sample of original images and related mask with data augmentation for hand segmentation with UNet.

YOLO (You Only Look Once) Algorithm: An Overview

The YOLO (You Only Look Once) algorithm is a widely used object detection method in computer vision, known for its efficiency and real-time processing capabilities. Unlike traditional detection methods that use complex pipelines and multiple stages, YOLO employs a single convolutional neural network (CNN) to simultaneously predict bounding boxes and class probabilities for all objects within an image in a single evaluation.

How YOLO Works:

YOLO divides the input image into a grid of cells, each responsible for detecting objects that fall within it. For each cell, the model predicts a fixed number of bounding boxes along with their respective class probabilities. Each bounding box is defined by five parameters: the x and y coordinates of the box's center, its width and height, and a confidence score. The confidence score reflects the model's certainty that the bounding box contains an object.

The size of the grid is determined by the input image dimensions and a predefined scale factor. For example, if the input image size is 416x416 pixels and the scale factor is 32, the resulting grid size would be 13x13.

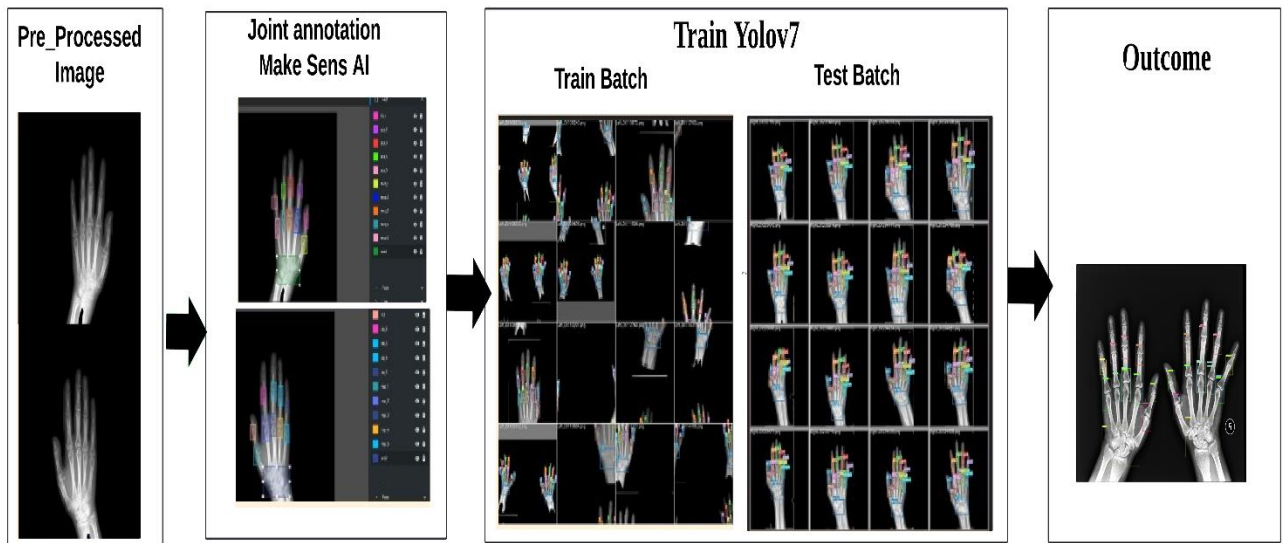
Training and Prediction:

During training, YOLO optimizes a unified loss function that balances object detection and classification accuracy. It learns to assign high confidence scores to bounding boxes that contain objects while accurately predicting the class probabilities for those objects. By treating object detection as a regression problem, YOLO directly predicts bounding boxes and class probabilities for the entire image in a single pass through the network.

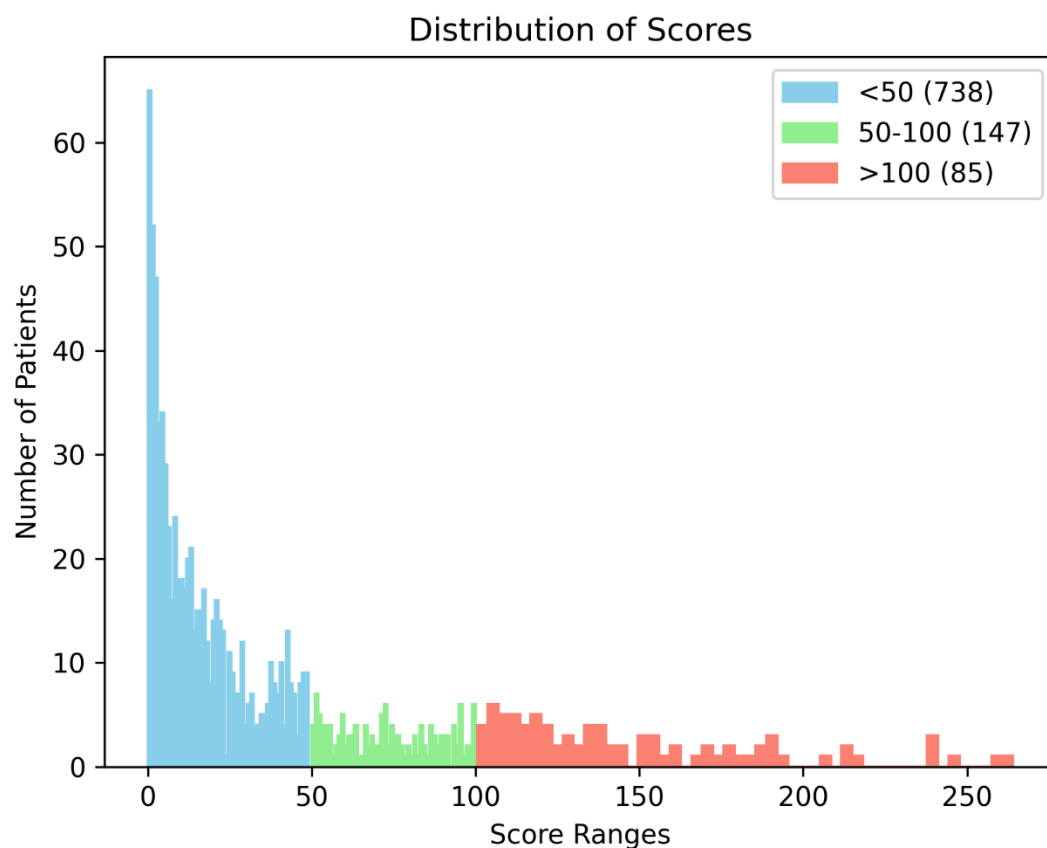
Advantages of YOLO:

YOLO offers several key benefits over traditional object detection methods:

1. **Speed:** It processes images in real time, capable of handling streaming video with a latency of less than 25 milliseconds.
2. **Simplicity:** YOLO uses a single CNN, avoiding the need for complex pipelines and separate training stages common in other methods like DPM and R-CNN.
3. **Global Context:** By examining the entire image during detection, YOLO incorporates contextual information, reducing background errors and improving detection accuracy.
4. **Generalization:** YOLO generalizes well to new domains, including natural images and artwork, outperforming other detection methods in diverse datasets.



Supplementary Figure S7. Workflow for Joint Identification using YOLOv7. The process begins with pre-processed X-ray images, followed by joint localization using the Make Sens AI tool for annotating joint positions. YOLOv7 is then trained with batches of annotated training images, which allows the model to accurately identify joints in the test images. The outcome shows successful joint identification on X-ray images, providing the foundation for further joint damage assessment.



Supplementary Figure S8. Distribution of patient scores. Histogram of patient scores categorized into three ranges: scores below 50 (738 patients, blue), scores between 50 and 100 (147 patients, green), and scores above 100 (85 patients, red). The plot indicates a higher concentration of patients with scores less than 50, gradually decreasing as scores increase.

References

1. Ronneberger O, Fischer P, Brox T, editors. U-Net: Convolutional Networks for Biomedical Image Segmentation 2015; Cham: Springer International Publishing.
2. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal.* 2017;36:61-78.
3. Yin XX, Sun L, Fu Y, Lu R, Zhang Y. U-Net-Based Medical Image Segmentation. *J Healthc Eng.* 2022;2022:4189781.