Context-aware dual-task deep network for concurrent bone segmentation and clinical assessment to enhance shoulder arthroplasty preoperative planning

Luca Marsilio, Andrea Moglia, Alfonso Manzotti, Pietro Cerveri

All*Abstract*—**Goal: Effective preoperative planning for shoulder joint replacement requires accurate glenohumeral joint (GH) digital surfaces and reliable clinical staging. Methods: xCEL-UNet was designed as a dual-task deep network for humerus and scapula bone reconstruction in CT scans, and assessment of three GH joint clinical conditions, namely osteophyte size (OS), joint space reduction (JS), and humeroscapular alignment (HSA). Results: Trained on a dataset of 571 patients, the model optimized segmentation and classification through transfer learning. It achieved median root mean squared errors of 0.31 and 0.24 mm, and Hausdorff distances of 2.35 and 3.28 mm for the humerus and scapula, respectively. Classification accuracy was 91 for OS, 93 for JS, and 85% for HSA. GradCAM-based activation maps validated the network's interpretability. Conclusions: this framework delivers accurate 3D bone surface reconstructions and dependable clinical assessments of the GH joint, offering robust support for therapeutic decision-making in shoulder arthroplasty.**

Index Terms—**Assisted preoperative planning; Deep learning; Explainable AI; Shoulder arthroplasty; Shoulder bone segmentation.**

***Impact Statement*- Integrating a multi-task CNN with explainable AI techniques can provide trustworthy segmentation and classification outcomes, potentially enhancing the reliability and adoption of AI-driven solutions in surgical planning.**

I. Introduction

OSTEOARTHRITIS (OA) is a degenerative condition affecting bones and cartilage, often resulting in changes to the bony surfaces, including osteophyte development, bone density loss, and joint spaces narrowing [1], [2]. In the shoulder, the glenohumeral (GH) joint comprises the humeral head and the scapula glenoid surface (i.e. the humeral socket). Primary OA leads to cartilage deterioration, causing a reduction in the GH joint space [3]. OA progression may lead to direct contact between the humeral head and its socket, culminating in impingement, inflammation, pain, and limited joint mobility. As OA advances, osteophytes may develop in the antero-inferior portion of the humeral head and extend downward [2]. The constant bone rubbing flattens the glenoid and further advances osteophyte formation along its boundaries, disrupting GH joint functionality over time [4]. Pathological humeroscapular alignments, such as subluxation or eccentricity, exacerbate joint instability and OA progression [5], [6]. Identifying these conditions is crucial for shoulder joint treatment, as it enables effective preoperative planning and drives the selection of the most suitable surgical implant, between anatomical and reverse [7], [8]. Furthermore, personalized surgical instruments (PSIs) proved effective in decreasing surgical time and enhancing postoperative bone alignment [9], [10]. PSIs for shoulder arthroplasty are patient-specific cutting jigs based on digital 3D models of the humerus and scapula obtained from medical image data [11], [12]. These aids facilitate proper implant sizing and cutting plane definition, reducing the risk of implant loosening [13]. However, irregular bone profiles pose challenges for accurate boundary delineation, requiring advanced image processing techniques [14]. Deep learning tools, particularly convolutional neural networks (CNNs), held promise to automate image processing and analysis in orthopedics [15]–[17]. Encoder-decoder architectures like the UNet and nnUNet have been effective in identifying osseous regions and soft tissues in 2D and 3D scans [15], [16], [18]–[21]. CNNs have also been studied for OA staging and treatment prediction [22]–[24]. In this context, core research gaps may be synthesized. First, irregular bone profiles and pathological changes complicate accurate boundary delineation and segmentation using traditional imaging techniques. Advanced image processing tools are needed to handle these complexities effectively. Second, while personalized surgical instruments (PSIs) have shown promise in improving surgical outcomes, their reliance on manual or semi-automated workflows for image analysis can be time-consuming and prone to variability. Third, although CNNs have demonstrated success in image segmentation and OA staging, existing models often lack multi-task capabilities, which are essential for simultaneously analyzing multiple clinical conditions affecting the GH joint. Finally, despite the effectiveness of CNNs, their "black-box" nature hinders clinical trust. Explainable AI tools, such as GradCAM, are not fully integrated into workflows to provide interpretable diagnostic insights.

To address these gaps, this study introduces a novel multi-task deep learning framework, xCEL-UNet, designed specifically for automated analysis of shoulder CT scans (Fig. 1). The network predicts the proximal humerus and scapula segmentation, concurrently with the staging of three different clinical

[1]L. Marsilio, A. Moglia, and P. Cerveri are with the Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
[2]P. Cerveri is also with the Department of Industrial and Information Engineering, University of Pavia, Pave, Italy.
[3]A. Manzotti is with Hospital ASST FBF-Sacco, Milan, Italy.

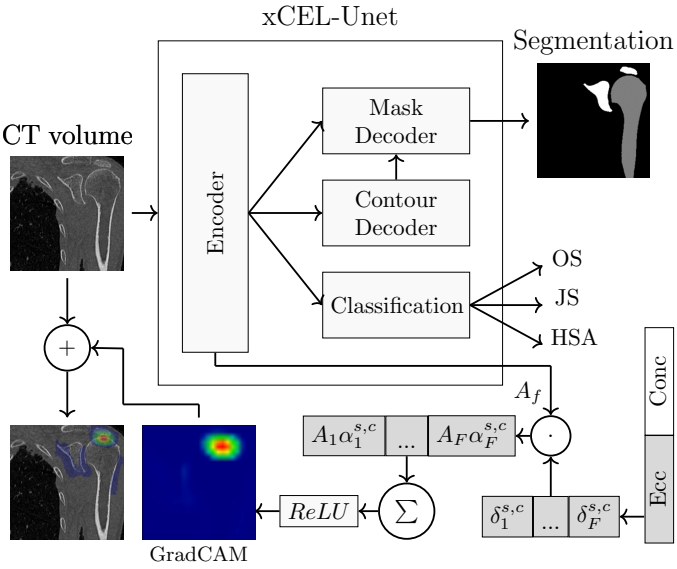

Fig. 1. xCEL-UNet model. It performs shoulder bone segmentation and clinical assessment of the glenohumeral joint, providing interpretability of the prediction by means of GradCAM.

conditions affecting the glenohumeral joint, namely the osteophyte size (OS), the GH joint space narrowing (JS), and the humeroscapular alignment (HSA). Each condition was stratified into multiple severity classes, comprehensively analyzing the shoulder bone pathologies. In addition, a gradient class activation map-based (GradCAM) [25] module was incorporated to produce visual explanation maps of the diagnostic classification (see subsections II-D and II-G for detailed technicality). By integrating segmentation and classification tasks into a single framework, xCEL-UNet offers a comprehensive analysis of GH joint conditions. Automated segmentation and severity staging enable more accurate and efficient preparation for shoulder arthroplasty, facilitating personalized treatment and reducing surgical risks. Incorporating GradCAM visualization enhances model transparency, helping clinicians understand the basis of diagnostic predictions and fostering trust in AI-assisted workflows. The xCEL-UNet leverages encoder-decoder architectures and multi-task learning to address complex pathological presentations, pushing the boundaries of existing CNN applications in orthopedics. As such, novel contributions can be summarized as:

- innovative xCEL-UNet deep learning framework[1], for bone shoulder segmentation, 3D surface reconstruction, and GH joint clinical assessment,
- GradCAM-based explainability module to enhance model interpretability and trustworthiness;
- validation across diverse demographic groups, clinical conditions, varying levels of disease severity, and bone morphological heterogeneity.

A. *Related works*

Traditional segmentation methods, such as thresholding, region growing, and statistical shape models have been investigated to achieve accurate bone CT segmentation. However, the development of joint-related pathological conditions, including osteoarthritis and osteoporosis, further complicates the boundary delineation by blurring intensity contrasts and deteriorating bone profiles [19], [26], [27]. For this reason, UNet-like deep CNNs were largely proposed in medical image segmentation to overcome these limitations [15], [28], [29]. Their encoder-decoder design, with skip connections, captures image features at different resolutions enhancing the anatomical identification. These architectures were extensively studied for the segmentation of knee, ankle, and shoulder bones in CT and MRI scans, and vertebral bodies in CTs [22], [30], [31]. Reduced segmentation accuracy and model generalization ability were assessed when training networks with small datasets [32], [33]. Pre-training UNet models or deploying adversarial regularization have shown promise in overcoming these limitations [34]. However, the same drawbacks might persist even after increasing the dataset numerosity [35]. Likewise, CNNs were proposed to evaluate cartilage osteonecrosis in knee X-ray images using the Kellgren-Lawrence scoring, achieving results comparable to expert operators [17], [22]. In [36], the authors tested various pre-trained models, such as ResNet and DenseNet, for discriminating fracture/non-fracture conditions in X-ray images. A multitask deep learning model was investigated for grading hip osteoarthritis in 4368 radiographs [37]. In [23] the rotator cuff muscle degeneration was analyzed on a 95-patient CT dataset using CNNs, demonstrating accuracy comparable to expert raters. However, comprehensive evaluation of shoulder conditions necessitates addressing osteophyte and humeral head shift assessments, which were neglected in the above-mentioned studies. Nonetheless, none of the previous studies faced morphology and diagnostic evaluation in a holistic approach. Context-aware segmentation and diagnostic classification methods have gained significant attention in the biomedical field as a single model can learn multiple related tasks by sharing representations. The contextual information from one task can inform another, enhancing overall performance. In [38], the authors proposed a multi-class segmentation of the aorta based on a UNet model, improved with context-aware self-attention. Likewise, the segmentation of hepatic vessel was improved by incorporating into a UNet model devoted modules to exploit the spatial vessel development in adjacent CT slices [39]. In [40], the authors increased the quality of hand bone segmentation in ultrasound images by weighting attention mechanism able to better learn differences among the different bones. Nonetheless, most of the deep learning-based literature contributions lacked strategies to increase clinical prediction transparency. Saliency maps were proposed to produce visual explanation maps assigning each pixel to a value representing its relevance to the prediction of a certain class [41]. Specifically, Grad-CAMs were proposed as a general method for evaluating the coherence of the predicted classification in a large class of CNN-based models [25]. In orthopedics, they were deployed to explain the predictions of bone mineral density in CT [42], to improve the reliability of radiographic fracture classification [43], and to enhance bone tumor classification in the proximal femur [44].

[1]source code available at https://github.com/LucaMarsilio/xCEL_UNet.git

II. Materials and Methods

A. Dataset description

607 axial CT scans were provided by MEDACTA International SA (Castel San Pietro, TI, Switzerland), in pseudo-anonymized form. The original images were acquired in preparation for PSI-based total (TSA) or reverse (RSA) arthroplasty interventions, all performed between 2021 and 2022. In detail, a five-to-one ratio in favor of RSA intervention was observed. The dataset was multi-centric, spanning North America (30%), Europe and the Middle East (48%), Asia Pacific (20%), and Latin America (2%). CT scans, characterized by 512×512 pixels, 330 slices on average, variable pixel size from 0.30 to 0.98 mm, and variable axial slicing from 0.30 to 2.5 mm, were acquired with different equipment, including General Electric, Varian, and Philips. Patients, aged 74±11 years, were diagnosed with different shoulder-related clinical conditions, including primary GH osteoarthritis, osteonecrosis, inflammatory arthritis, osteophyte development, and post-traumatic degenerative disease. Alongside CT scans, the scapula and proximal humerus reference surfaces were included. Two clinical operators, with more than 10 years of radiological expertise, produced and revised the bony surfaces with Mimics (v.16.0, Materialise, Leuven, Belgium). Two different humerus surfaces were available for each patient. The first one represented the original proximal humerus morphology. The second was a manually elaborated version where osteophytes and deformities were cleared to reproduce the physiological humeral head anatomy. While the original surface was crucial to designing the custom-cutting implant and planning the contact areas between the bone and corresponding jigs, the modified version was necessary to plan the optimal prosthesis positioning and size.

B. Ethical and Regulatory Compliance

The personal patient information in the dataset was completely unavailable, with each case identified by an alphanumeric code. Clinical data and images were encoded by MEDACTA with a two-stage anonymization. This method is used to safeguard clinical data, including images, in compliance with the GDPR (General Data Protection Regulation). A separate department in MEDACTA (trusted party) retains the ability to re-identify individuals using a key or decoding mechanism. Thus, all the training and testing stages described in this paper were compliant with current regulations in medical data management.

C. Training set preparation

Among the 607 cases, 36 were excluded because bones presented internal metal components such as screws, implants, and plates, whose analysis was beyond the aim of this study. Therefore, the present investigation was conducted with the remaining 571 cases, comprising 274 females and 297 males, and 300 right and 271 left shoulders. CT preprocessing was carried out to standardize the dataset samples. In detail, voxel normalization was performed to scale their intensity between 0 and 1. Since each scan was derived from different scanning machinery, voxel values were converted to the Hounsfield Unit (HU) range, ensuring a range between -1024 and 2500, representing air to dense cortical bone [45]. Values were then shifted to positive units and normalized between 0 and 1 for consistency. The 571 cases were randomly split into 485 (85%) as training and validation set, and 86 (15%) as test set. Two different groups were generated from the original dataset, one for CT segmentation and the other for staging three GH-related conditions (see Supplementary Materials, Fig. 1). The first one (DSeg) included the preprocessed shoulder CT scans and their corresponding segmentation labels. To reduce the computational overhead, cropping was applied to the CT volumes in the axial, coronal, and sagittal views to eliminate all slices where the proximal humerus and scapula labels were not available. In addition, a patch-based method was deployed to augment the training set size while keeping the original voxel resolution. Specifically, cropped CTs were patched into sub-volumes of size 160×160×160, with a variable overlapping degree (25% on average), depending on their initial spatial size. The second dataset (DCls) consisted of CT volumes focused on the humeral head and glenoid surface, and their corresponding segmentation and classification labels. The automatic extraction of these GH-centered bounding boxes from the original shoulder CT scans rested upon a prior method originally developed and tested for proximal femur head [11]. The number of CT scans in the DCls set was finally doubled by data augmentation flipping the originated sub-volumes in the sagittal plane. The humeral head osteophytes severity degree (OS) was staged into three classes, according to the Samilson-Prieto grading system [46], [47], highlighting increasing osteophyte size. Grade 0 revealed small-size (s_o<3 mm), grade 1 medium-size (3<s_o<7 mm), and grade 2 large-size (s_o>7 mm) osteophytes. Automated OS labeling was achieved by computing the maximum distance (in millimeters) between the osteophyte-cleared and morphologic humerus reference surfaces for each dataset case. The GH joint space (JS) was manually identified following the Kellgren-Lawrence grading system [1]. Three classes were tagged for each dataset case according to the residual JS including physiological joint space (grade 0), slightly narrowed joint space (grade 1), and non-detectable joint space (grade 2). Finally, HSA was assessed by looking at the humeral head shift from the glenoid surface in cranio-caudal direction by identifying either concentric-physiological or eccentric-pathological humeral head alignment [48]. The entire labeling procedure was supervised and revised by an orthopedic surgeon with more than 25 years of clinical practice (A.M.) (Table I).

D. Segmentation module of the xCEL-UNet

The segmentation and classification modules of the xCEL-UNet (Fig. 1) share a common encoder path featuring a sequence of convolutional blocks (convolutional, ReLU activation, and max-pooling layers), with the same characteristics of the original CEL-UNet, described in [20], which is our reference segmentation architecture for this work. It includes three processing blocks, doubling at each one the number of feature maps, initially set to eight. The convolutional filter size and stride were 3×3×3 and 1×1×1, respectively, while max-

TABLE I. GH osteoarthritic-related condition labeling. The first column describes the pathological condition, namely osteophyte size (OS), GH joint space (JS), and humeroscapular alignment (HSA). The second and third columns show the labeling criteria for each multi-class task and the provided index. The last column reports the frequency of each class in the dataset for the corresponding task.

Condition	Criteria	Index	Frequency [%]
OS	<3 mm	0	31.1
	3-7 mm	1	36.1
	>7 mm	2	32.8
JS	Physiological	0	38.2
	Narrowed	1	27.2
	Non-detectable	2	34.6
HSA	Concentric	0	56.1
	Eccentric	1	43.9

pooling uses 2×2×2 filters. Unlike the UNet, the xCEL-UNet was characterized by two parallel decoder branches, namely the mask decoder (MD) for region segmentation, and the contour decoder (CD) for edge detection. Both branches performed upsampling via transpose convolution, with the number of feature maps halving at each decoding block. Skip connections linked the encoder to both decoder branches. The final layer of the MD branch was a 1×1×1 convolution with three output channels (background, proximal humeral bone, scapula) and Softmax activation. Unidirectional skip connections from each block in the contour decoder to the corresponding block in the mask decoder were enabled. The training of the segmentation module was based on the optimization of both $\mathcal{L}_r$ and $\mathcal{L}_c$ loss functions as:

$$\begin{aligned} \mathcal{L}_r &= 1-(\alpha\cdot\mathcal{D}+(1-\alpha)\cdot\mathcal{C}) \\ \mathcal{L}_c &= 1-(\beta\cdot\mathcal{C}+(1-\beta)\cdot\hat{\mathcal{C}}) \end{aligned} \tag{1}$$

where $\mathcal{C}$ and $\hat{\mathcal{C}}$ are the distance cross-entropy and reverse distance cross-entropy terms, $\mathcal{D}$ is the Dice score, and α and β parameters weight the contribution of the Dice and cross-entropy terms.

E. Classification module of the xCEL-UNet

The xCEL-UNet classification module consisted of the encoder, shared with the segmentation one, a 3D global average pooling, and three classification branches, tailoring the network for each pathological condition. They embedded two consecutive dense layers with 64 and 16 units and a ReLU activation function. The output of two of the three-stage classification tasks, namely OS and JS, were dense layers with three output neurons with Softmax activation function, while the binary neuron for the HSA prediction featured sigmoid activation functions. During the training of the classification task, categorical and binary cross-entropy loss functions were chosen for the categorical and binary predictions, respectively. To balance the uneven training label frequency for each task (Table I), each loss function was weighted with a parameter K_c to balance the representation of every class, following Eq.2:

$$\mathcal{K}_c = \frac{\frac{1}{N_c}}{\sum_{i=1}^{C}\left(\frac{1}{N_i}\right)} \tag{2}$$

where c, N_c, and C were the current class for the specific classification task, the number of total cases for each class, and the number of classes, respectively.

F. xCEL-UNet training: transfer learning and fine tuning

The xCEL-UNet training was performed in two sequential steps. In the first stage, the segmentation module (encoder branch, mask decoder and contour decoder, cfr. Fig. 1) was trained using the DSeg dataset. The training process utilized the ADAM optimizer (Adaptive Moment Estimation) with a learning rate of 10^{-4}. The parameter α in the $\mathcal{L}_r$ loss (eq. 1) was initially set to 1 and reduced by a factor of 0.005 per iteration until reaching a value of 0.5 at the 100^{th} iteration, after which it remained constant. This scheduling strategy was designed to enable the Dice loss to dominate weight learning in the early stages, gradually incorporating the influence of cross-entropy based on the distance-weighted map, thereby reducing the prominence of the Dice loss component over time. To balance the two contributions in the $\mathcal{L}_c$ loss (eq. 1), the parameter β was defined as the ratio of shape boundary voxels to the total number of voxels within the batch. Early stopping was applied to prevent overfitting, terminating the training after 40 consecutive epochs with no improvement in validation loss. In the second stage, the training of the clinical staging module (classification branch, cfr. Fig. 1) harnessed four distinct strategies (Table II), employing the DCls dataset. The first utilized a transfer-learning (TL) approach, where the encoder, bottleneck, and decoder branches were frozen (i.e. their weights were not retrained). This method aimed to determine if the high and low-level features learned during segmentation training could be leveraged for identifying GH osteoarthritic-related conditions in the classification task. The other three strategies involved fine-tuning specific sections of the segmentation network: the network bottleneck (FT-B), the encoder (FT-E), and the entire network (FT-N). This time, the number of re-trained weights in the segmentation module varied depending on the training setup. The analysis sought to evaluate whether modifying the segmentation module during the second training phase could improve feature extraction for

TABLE II. xCEL-UNet training setup summary. The transfer-learning-based method (TL) entails just the optimization of the classification branch layers, while the three fine-tuning-based setups re-train different portions of the segmentation network, including the whole encoder (FT-E), the bottleneck (FT-B), and the whole network (FT-N).

Setup	**Encoder**	**Bottleneck**	**Decoder**	**Classification**
TL	frozen	frozen	frozen	trainable
FT-B	frozen	trainable	frozen	trainable
FT-E	trainable	trainable	frozen	trainable
FT-N	trainable	trainable	trainable	trainable

GH condition classification while maintaining the accuracy of the segmentation outputs. Computations were performed on a 32-core CPU and Nvidia A100-PCIe GPU with 40 GB RAM.

G. GradCAM and visual explainability

To investigate the network's ability to learn the clinical context, this work employed the GradCAM algorithm proposed in [25] and extended it to the multi-class classification. This implementation produced a 3D activation heatmap overlaid onto the corresponding CT volume to visualize which shoulder regions were important for the specific GH clinical classification (see Fig. 1). In detail, the entire procedure involved: 1) the forward pass, processing the CT scan with the xCEL-UNet; 2) the selection of the output score y corresponding to the class set s (e.g. HSA in Fig. 1) and the specific class c (e.g. eccentric); 3) the backward pass to compute the gradients $\delta_f^{s,c}$ of the score y for the target class s,c with respect to the feature maps A_f of the bottleneck layer (f ranged between 1 and 64); 4) the global average of these gradients across the bottleneck spatial dimension ($10 \times 10 \times 10$) to obtain the weights $\alpha_f^{s,c}$ for each feature map; 5) the weighted sum of the feature maps, 6) the ReLU activation to focus on the positive activations, 7) upsampling the heatmap to the CT volume size and overlaying. Mathematically, the gradient was computed for each voxel (i,j,k) as:

$$\delta_f^{s,c} = \frac{\partial y^{s,c}}{\partial A_{f,i,j,k}} \tag{3}$$

and then globally averaged across the spatial dimensions as:

$$\alpha_f^{s,c} = \frac{1}{Z} \sum_{i=1}^{10} \sum_{j=1}^{10} \sum_{k=1}^{10} \delta_f^{s,c} \tag{4}$$

where $Z = 1000$ was the feature map voxel number. Finally, the class activation map was obtained as:

$$L_{GradCAM}^{s,c} = \mathrm{ReLU}\left(\sum_{f=1}^{64} \alpha_f^{s,c} A_{f,i,j,k} \right) \tag{5}$$

H. Result comparison, metrics, and statistical analysis

The four xCEL-UNet training setups were evaluated for segmentation, 3D reconstruction, and classification outcomes. The first two tasks were further compared to the results achieved by the CEL-UNet architecture to assess deviations from a traditional segmentation network training approach. The original CEL-UNet, backbone of the xCEL-UNet, was also compared to state-of-the-art nnUNet architecture to evaluate its raw performances. Dice score was computed to measure the intersection over union network performances against segmentation labels, while precision and recall are responsive for both over- and under-segmentation errors, respectively. The 3D volumes of each prediction were built exploiting a custom marching cube-based automated algorithm [11]. Reconstruction errors were evaluated by computing the root mean squared error (RMSE) and the Hausdorff distance, considering the average and maximum distance between the reference and predicted surfaces. In detail, a one-way analysis was carried out by comparing the distances between each vertex of the target mesh and its closest from the predicted mesh. The statistical analysis of segmentation and 3D reconstruction results was performed using the non-parametric Friedman test, followed by Wilcoxon Signed-Rank tests with Bonferroni correction for post-hoc analysis. A p-value below 0.05 was considered indicative of statistical difference between competitive models. Accuracy, precision, recall, and F1-score computation were carried out to identify the best training setup for the classification tasks. In addition, the confusion matrix of the most promising approach was showed provide a broader evaluation of its classification performances.

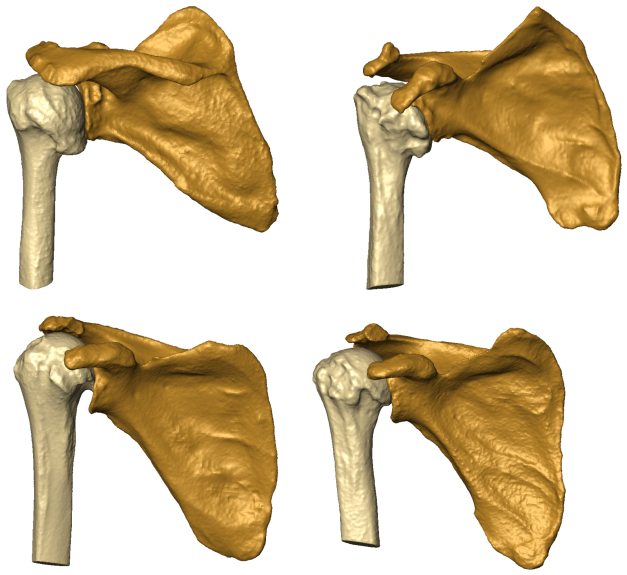

Fig. 2. Four reconstructed test set cases (0005, 0102, 0328, and 0628) showing different morphological structures and pathological conditions.

III. Results

A. Segmentation and 3D reconstruction

The segmentation and reconstruction of the humerus and scapula by the xCEL-UNet was accurate across a wide range of morphologies (Fig. 2). As an example, the upper left image depicts large deformations of the humeral head due to osteophyte development. Likewise, the upper right one represents distributed osteophytes on the humeral head, with null intra-articular space. Both bottom images displayed fewer osteophytes, but eccentric humeral heads. The comparison between the xCEL-UNet against the nnU-Net [15] segmentation outcomes showed competitive dice results in the range of 99% (p=0.0002) and 98% (p=0.0005) for the humerus and scapula, respectively (see Supplementary Material, Table 2). Considering the xCEL-UNet variants, the results computed with the TL strategy were by definition identical to the one achieved with the original CEL-UNet, as both the encoder and the decoder branches were not re-trained (Fig. 3). The outcomes of the three xCEL-UNet fine-tuning variants were similar to the TL approach for the humerus dice score (p>0.5), while for the scapula they showed larger interquartile ranges,

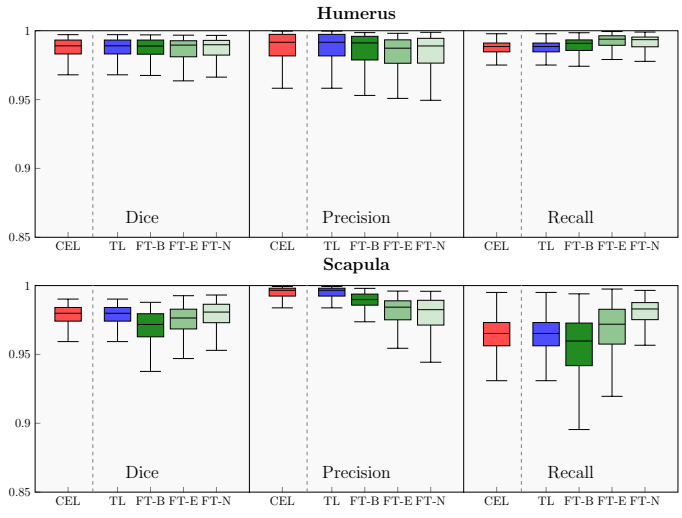

Fig. 3. Dice, precision, and recall for humerus (*above*) and scapula (*below*). Red boxplots display the CEL-UNet scores for the three metrics, the blue ones the xCEL-UNet trained with a transfer learning (TL) approach, while three tones of green depict the different fine-tuning strategies, re-training only the bottleneck alongside the classification layers (dark green, FT-B), the encoder (green, FT-E), and the whole network (light green, FT-N).

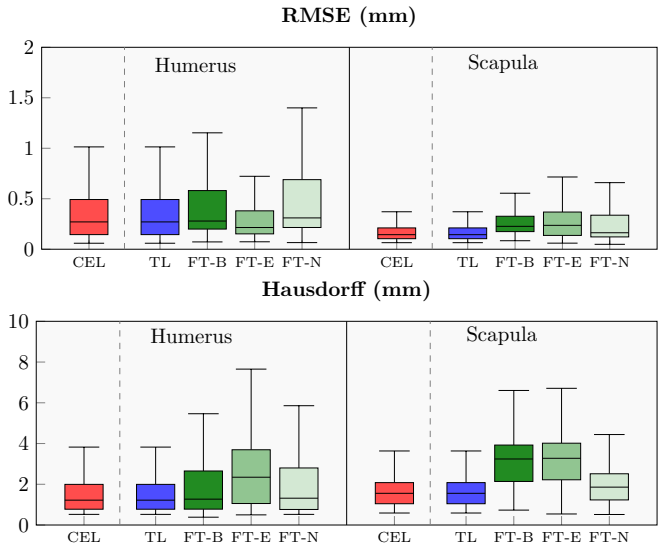

Fig. 4. Root mean squared error (*above*) and Hausdorff distance (*below*) for the humerus and scapula 3D reconstruction. Red boxplots display the CEL-UNet results, the blue ones the xCEL-UNet with the transfer learning (TL) strategy, while three tones of green depict the different fine-tuning strategies, re-training only the bottleneck with the classification layers (dark green, FT-B), the encoder (green, FT-E), and the whole network (light green, FT-N).

with the FT-B model showing significantly lower metrics (p=0.0007). The precision results highlighted again similar performances in the humerus segmentation across the three variants. However, FT-E and FT-N models were significantly worse in the scapula than those achieved by the CEL-UNet. Regarding recall, FT-E and FT-N scores were significantly greater than the CEL-UNet ones in the humerus (p=0.0002 and p=0.0001, respectively), while for the scapula, FT-N showed the best results overcoming again the CEL-UNet. The 3D reconstruction errors saw a RMSE for the xCEL-UNet variants confirming high segmentation quality for the humerus, similar to that of the original CEL-UNet, featuring median values less than 0.3 mm with an IQR ranging from (0.15-0.48). The FT-E and FT-N variants showcased the best and worst results, respectively, with 0.21 mm (0.15-0.38) and 0.31 mm(0.22-0.69) (Fig. 4). The RMSE for the scapula was significantly lower than the ones of the humerus (p=0.0001). For the CEL-UNet, the median Hausdorff distance error was in the range of 1.5 mm, for both the humerus and scapula. Among the three variants, FT-E showcased the worst results for the scapula, featuring a median RMSE and Hausdorff distance of 0.24 mm and 3.28 mm.

B. Clinical classification

The pure transfer learning (TL) of the CEL-UNet segmentation weights, alongside the classification module training, proved ineffective in staging the three pathologies, achieving results just above the random classification thresholds (Table III). Conversely, the fine-tuning variants provided better results for the three class sets. The best results for the osteophyte size (OS) and the joint space (JS) detection were achieved by retraining the whole encoder (FT-E), with accuracy and precision values of 0.91 and 0.93, and recall and F1-

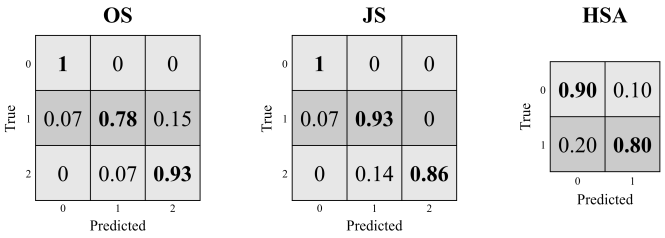

Fig. 5. Confusion matrix of the GH osteoarthritic-related condition classification computed with the FT-E xCEL-UNet.

score of 0.90 and 0.93, respectively, consistently outscoring the other strategies. Conversely, the greatest performance for the HSA assessment was computed with the bottleneck fine-tuning (FT-B), reaching accuracy, precision, recall, and F1-score of 0.91, 0.93, 0.89, and 0.91, respectively. Overall, FT-N was less reliable than the other two variants, especially for the OS prediction, with a 74% accuracy. Based on these findings, the FT-E approach was selected as the optimal training strategy for this multi-task, multi-class classification problem. Accordingly, the FT-E confusion matrix (Fig. 5) was computed to provide broader insights into the classification performance. It demonstrates how the non-pathological condition staging for OS and JS (i.e., small/no osteophyte and physiological joint space, respectively) was always estimated at 100%, while the medium-size identification was the most challenging assessment, with a true positive rate of 78%. For the HSA index, a slight bias toward the concentric condition prediction was observed, with a true positive rate of 90%.

This article has been accepted for publication in IEEE Open Journal of Engineering in Medicine and Biology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/OJEMB.2025.3527877

TABLE III. Accuracy, Precision, Recall, and F1-score for the three GH osteoarthritic-related condition classification, osteophyte size (OS), joint space (JS), and humeroscapular alignment (HSA) with the four different training setups of the xCEL-UNet, including transfer learning (TL), and the fine-tuning of the bottleneck (FT-B), encoder (FT-E), and the whole network (FT-N).

Training	**Accuracy**			**Precision**			**Recall**			**F1 Score**		
	OS	JS	HSA	OS	JS	HSA	OS	JS	HSA	OS	JS	HSA
TL	0.33	0.47	0.58	0.35	0.31	0.67	0.32	0.45	0.40	0.33	0.37	0.50
FT-B	0.83	0.80	**0.91**	0.82	0.80	**0.93**	0.82	0.79	**0.89**	0.82	0.79	**0.91**
FT-E	**0.91**	**0.93**	0.85	**0.91**	**0.93**	0.90	**0.90**	**0.93**	0.80	**0.90**	**0.93**	0.85
FT-N	0.74	0.86	0.90	0.74	0.87	0.93	0.74	0.85	0.87	0.74	0.85	0.90

C. GradCAM-based xCELUnet interpretability

The GradCAM activation maps were generated for the FT-E xCEL-UNet model which achieved the best classification scores against the other three training strategies. For HSA, a strong coherence between the eccentric pathological condition (i.e., humeral head shifting upward in the coronal plane) and the generated heatmap was found, with maximum activation in the intersection regions between the humeral head and the scapula acromion (see Supplementary Materials, Fig.2). The radial colormap distribution showed its highest values in the narrowed intra-articular space (A, B, and C), with lower intensity activation in the surrounding areas. Conversely, the concentric HSA prediction resulted in a less homogenous activation pattern (see Supplementary Materials, Fig. 3). Likewise, the network decision-making process leading to a correct OS and JS staging demonstrated strong coherence with the clinical features of the corresponding pathological condition of interest (Fig. 6). In detail, the highest activations registered for the large-size osteophyte classification were predominantly in the osteophyte regions (A, C, and E) of the humeral head, while for the narrowed JS identification they were mainly confined to the contact areas between the glenoid and the humeral head (B, D, and F), with less or no activation in the osteophyte regions. The comparison of case 0362 JS and OS visual explanations provided valuable insights into the selectivity for the two different specific clinical conditions. Interestingly, the visual analysis of the OS classification confirmed the network ability to discriminate the three gradings with specific activation patterns. For the grade 0 staging (see Supplementary Materials, Fig. 4 - case 0221), the absence of relevant osteophytes was explained featuring a smooth activation distributed across the overall distal humerus. For grade 1 (see Supplementary Materials, Fig. 4 - case 0158), the presence of a medium-size osteophyte corresponded to an activation map very localized on the specific region. For grade 2, the activation map focused coherently on the large region affected by the osteophytes spanning both the top and lateral surface of the humeral head (see Supplementary Materials, Fig. 4 - case 0362).

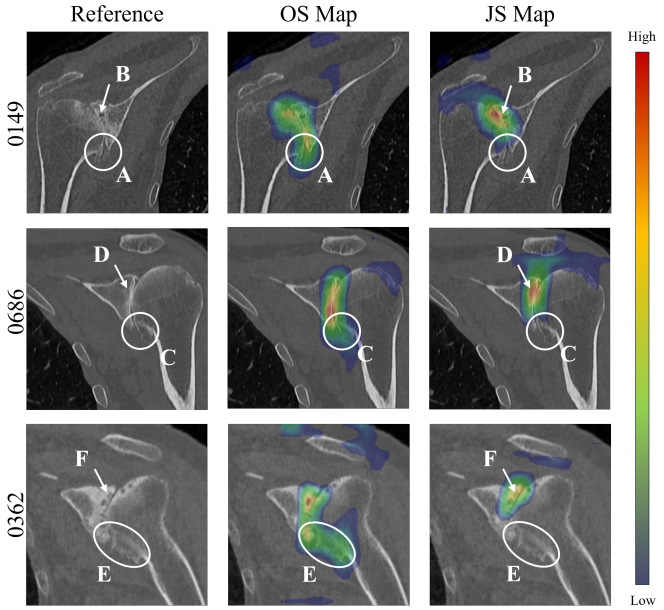

Fig. 6. OS and JS GradCAM-generated activation maps (second and third column, respectively) overlaid on top of their corresponding input CT (coronal view). A, C, and E areas highlight large humeral osteophytes, while B, D, and F arrows the non-detectable joint space between the humerus and the glenoid.

IV. Discussion

The humerus and scapula segmentation can be more challenging than other bones. With a pathological GH joint space reduction and concurrent osteophyte development, the humeral head is wrapped around the glenoid, further complicating the identification and delineation of the two bone boundaries. The proposed xCEL-UNet was retrospectively validated for humerus and scapula segmentation using a private dataset of 571 patients from different ethnic groups, characterized by a large variability of GH pathological conditions, enhancing the robustness and generalization capability of the model. The patients were diagnosed with different shoulder-related clinical conditions including primary GH joint osteoarthritis and osteonecrosis, inflammatory arthritis, and post-traumatic degenerative disease. Some patients featured subluxation conditions with eccentricity of the humeral head. Others reported morphological variations due to osteophyte formation. All these ensured that the model performance was aligned with real-world medical scenarios. The labeling process for all three clinical conditions was overseen and refined by an orthopedic surgeon with over 25 years of experience. Each condition

TABLE IV. Comparison of with state-of-the art papers dealing with shoulder bone segmentation.

Study	Cases/Images	Region/Condition	Modality	Segm Accuracy	Class Accuracy
[49]	31	Humeral head/acetabulum	MRI	0.88	N/A
[50]	116	Scapula	CT	0.97	N/A
[35]	485	Scapula	CT	0.87	N/A
[21]	16	Upper body bones	CT	0.85	N/A
[51]	500	Humerus head	MRI	0.91	N/A
xCELUnet	571	Humerus/Scapula	CT	0.98	0.90

was categorized into multiple severity levels, allowing for a thorough and detailed assessment of shoulder bone pathologies. The qualitative results showed the potential clinical impact of the xCEL-UNet model for shoulder bone segmentation (see Supplementary Materials, Table 2). The network was capable of handling a wide range of joint morphological deformations, such as the absence of intra-articular cartilage, abnormal humeral head positioning, and bone osteophytes (see Fig. 2). The fast reconstruction of the shoulder bone surface sensibly reduces the time required for diagnosis, provides a meaningful three-dimensional comprehension of the overall pathological condition, and supports the decision for the identification of the optimal treatment. Once clinically validated, the proposed deep learning method may feature a substantial clinical impact by improving diagnostic accuracy and surgical options in the treatment of glenohumeral joint arthritis. As far as classification is concerned, pure transfer learning yielded poor results, indicating that features learned during segmentation were insufficient for accurate clinical staging without retraining. Fine-tuning the network encoder alongside training the classification branch was the best trade-off for concurrent segmentation and classification tasks, achieving the highest classification outcomes (Table III) while maintaining high segmentation (cfr. Fig. 3) and 3D reconstruction (cfr. Fig. 4) quality. GradCAM enhanced the interpretability of the xCEL-UNet model, showing coherence between predicted classifications and clinical context through activation maps, highlighting the model decision-making process (cfr. Fig. 6, and Supplementary Materials Fig.2, 3, and 4). Our findings are supported by the literature (Table IV). A CNN, trained on a dataset of 95 shoulder CT scans, achieved a 97% Dice similarity coefficient for humerus segmentation [52]. Furthermore, a UNet-based glenoid segmentation from 237 CT scans featuring anterior shoulder dislocation was investigated against a 248 control group dataset, reaching a segmentation accuracy of 96% [35], similar to our results computed over the entire scapula. Deep neural networks, such as the ResNet18, were proposed as diagnostic support systems for distal humerus fracture with sensitivity results in the range of 61%, but with a high specificity (95%) making them particularly useful for identifying these lesions [53]. A multi-class bone segmentation pipeline was presented in [21], achieving an overall dice score of 85% across 126 bone classes. However, the study's generalizability was limited, as the training and testing were conducted on only 16 upper-body postmortem CT scans. Additionally, the dataset did not include orthopedic patients, meaning the humerus and scapula were not consistently affected by osteoarthritic pathologies. From a clinical translation perspective, the segmentation tool's accuracy makes it viable for PSI manufacture and planning, with the entire inference pipeline running below 15 seconds on a dedicated GPU-based cluster computer [54]. The inference (segmentation, surface reconstruction, and clinical prediction) was tested on a consumer laptop (Intel i7 processor, 24GB RAM) showing a computational time of approximately 130 seconds on average, still compatible with orthopedic planning practice. In the light of the results, we remark that this work provided enhanced precision in anatomical segmentation and reconstruction, improved diagnostic accuracy and decision-making, and supported interpretability for clinical adoption of deep learning models. The accurate segmentation of the humerus and scapula and the ability to handle morphological variations are critical for preoperative planning. The xCEL-UNet results demonstrates its reliability in capturing fine anatomical details. This may have implications in the pre-operative implant design as surgeons can use these detailed reconstructions to select and customize prosthetic components, particularly in shoulder arthroplasty, where precise alignment and fit are crucial. In addition, the CEL-UNet's robustness in cases with osteophytes, joint narrowing, or eccentric alignment underscores its potential to improve planning in complex cases, reducing intraoperative uncertainty. The classification results, especially with the FT-E fine-tuning approach, demonstrated high accuracy in staging all the three critical OA-related features. This may have implications in the early identification of pathological changes like large osteophytes or narrowed joint spaces that can guide earlier intervention, potentially delaying disease progression. In addition, such findings may help to select tailored treatment plans according to severe against mild-to-moderate OA, suggesting joint replacement surgery and more conservative therapies, respectively [4], [55]. Lastly, the model's ability to reliably identify concentric vs. eccentric HSA conditions can help stratify surgical risks. Eccentric conditions often indicate rotator cuff dysfunctions or glenoid wear, which may require more complex surgical techniques or grafting. The GradCAM maps demonstrated the model's ability to focus on clinically relevant regions, such as inter-articular spaces for JS and osteophyte sites for OS. This may have implications in increasing the clinical confidence of the operator in the transparent use of AI tools [56]. These maps might also serve as educational tools, helping less experienced orthopedics understand key diagnostic markers of the GH osteoarthritis. However, these findings must be interpreted within the study's limitations. In detail, just a binary HSA condition was considered, whereas the humeral head

eccentricity (subluxation) is typically described in posterior, anterior, inferior, and superior directions. This choice was driven by the study's focus on patients with degenerative shoulder arthritis. Moreover, despite an extensive dataset of 571 patients, it lacked clinical conditions such as bone loss, and prior surgeries. Additionally, a bias is registered towards an older population (the median age was 74 years) reflecting instances more susceptible to bone deterioration and decreased turnover. Although the xCEL-UNet demonstrated strong performance in general cases, its application to rare pathologies or extreme deformities may require further fine-tuning and data augmentation to ensure robustness. Effective implementation in clinical practice will depend on close interdisciplinary collaboration between surgeons, radiologists, and engineers to validate AI predictions under real-world conditions. Additionally, compliance with healthcare regulations and ethical standards is essential to guarantee the safe deployment of these models. Looking ahead, the xCEL-UNet holds promise for intraoperative applications, such as integration into surgical navigation systems for real-time guidance, and for longitudinal monitoring, enabling clinicians to track disease progression and adjust treatment plans over time.

V. Conclusions

We demonstrated that a dual-task deep network can effectively perform CT segmentation and clinical assessment simultaneously. This pilot study is the first to apply a UNet-like architecture to the scapula and humerus segmentation while retraining it to classify three clinical conditions affecting the glenohumeral joint. GradCAM analysis confirmed that the network consistently learned the context of GH joint clinical conditions. These findings suggest a significant advancement in AI-based decision tools, improving clinical interpretation of the GH joint and aiding in the selection of appropriate prosthetic and surgical strategies, making it more viable for clinical implementation.

VI. Acknowledgments

The authors would like to thank Medacta International SA for providing images and clinical data. This work was partially supported by P.E. PE00000013-FUTURE ARTIFICIAL INTELLIGENCE RESEARCH (FAIR), Italian Ministry of Research and University.

VII. Conflict of interest

The authors declare no conflict of interest.

References

[1] J. H. Kellgren and J. S. Lawrence. Radiological assessment of osteo-arthrosis. *Annals of the rheumatic diseases*, 16:494–502, Dec 1957.

[2] Kiyohisa Ogawa, Atsushi Yoshida, and Hiroyasu Ikegami. Osteoarthritis in shoulders with traumatic anterior instability: Preoperative survey using radiography and computed tomography. *Journal of Shoulder and Elbow Surgery*, 15(1):23–29, 2006.

[3] Michael Khazzam, Albert O. Gee, and Michael Pearl. Management of glenohumeral joint osteoarthritis. *The Journal of the American Academy of Orthopaedic Surgeons*, 28:781–789, Oct 2020.

[4] Lawrence Lo, Scott Koenig, Natalie L. Leong, Brian B. Shiu, S. Ashfaq Hasan, Mohit N. Gilotra, and Kenneth C. Wang. Glenoid bony morphology of osteoarthritis prior to shoulder arthroplasty: what the surgeon wants to know and why. *Skeletal radiology*, 50:881–894, May 2021.

[5] Adam Sassoon, Bradley Schoch, Peter Rhee, Cathy D. Schleck, William S. Harmsen, John W. Sperling, and Robert H. Cofield. The role of eccentric and offset humeral head variations in total shoulder arthroplasty. *Journal of Shoulder and Elbow Surgery*, 22:886–93, Jul 2013.

[6] Vani J. Sabesan, Mark Callanan, Ari Youderian, and Joseph P. Iannotti. 3d ct assessment of the relationship between humeral head alignment and glenoid retroversion in glenohumeral osteoarthritis. *The Journal of Bone and Joint Surgery. American Volume*, 96:e64, April 2014.

[7] Florian M. Buck, Bernhard Jost, and Juerg Hodler. Shoulder arthroplasty. *European radiology*, 18(12):2937–2948, December 2008.

[8] Jonelle Petscavage-Thomas. Preoperative planning and postoperative imaging in shoulder arthroplasty. *Seminars in musculoskeletal radiology*, 18:448–462, September 2014.

[9] Jong-Keun Seon, Hyeong-Won Park, Seung-Hyun Yoo, and Eun-Kyoo Song. Assessing the accuracy of patient-specific guides for total knee arthroplasty. *Knee Surgery, Sports Traumatology, Arthroscopy*, 24(11):3678–3683, November 2016.

[10] John W Noble Jr, Chris A Moore, and Ning Liu. The value of patient-matched instrumentation in total knee arthroplasty. *The Journal of arthroplasty*, 27(1):153–155, 2012.

[11] Pietro Cerveri, Costanza Sacco, Gianluca Olgiati, Alfonso Manzotti, and Guido Baroni. 2d/3d reconstruction of the distal femur using statistical shape models addressing personalized surgical instruments in knee arthroplasty: A feasibility analysis. *The International Journal of Medical Robotics + Computer Assisted Surgery : MRCAS*, 13(4), Dec 2017.

[12] Adolph V Lombardi Jr and Keith R Berend. Patient-specific approach in total knee arthroplasty. *Orthopedics (Online)*, 31(9):927, 2008.

[13] Marco Mandolini, Agnese Brunzini, Giulia Facco, Alida Mazzoli, Archimede Forcellese, and Antonio Gigante. Comparison of three 3d segmentation software tools for hip surgical planning. *Sensors*, 22(14):5242, 2022.

[14] Ajay Goud, Dmitri Segal, Pejman Hedayati, John J. Pan, and Barbara N. Weissman. Radiographic evaluation of the shoulder. *European Journal of Radiology*, 68(1):2–15, October 2008.

[15] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Communication*, 18(2):203–211, December 2021.

[16] Guangbin Wang and Yaxin Han. Convolutional neural network for automatically segmenting magnetic resonance images of the shoulder joint. *Computer Methods and Programs in Biomedicine*, 200:105862, Mar 2021.

[17] Pauline Shan Qing Yeoh, Khin Wee Lai, Siew Li Goh, Khairunnisa Hasikin, Yan Chai Hum, Yee Kai Tee, and Samiappan Dhanalakshmi. Emergence of deep learning in knee osteoarthritis diagnosis. *Computational intelligence and neuroscience*, 2021:4931437, 2021.

[18] Berk Norman, Valentina Pedoia, and Sharmila Majumdar. Use of 2d u-net convolutional neural networks for automated cartilage and meniscus segmentation of knee mr imaging data to determine relaxometry and morphometry. *Radiology*, 288(1):177–185, 2018.

[19] Davide Marzorati, Mattia Sarti, Luca Mainardi, Alfonso Manzotti, and Pietro Cerveri. Deep 3d convolutional networks to segment bones affected by severe osteoarthritis in ct scans for psi-based knee surgical planning. *IEEE Access*, 8:196394–196407, 2020.

[20] Matteo Rossi, Luca Marsilio, Luca Mainardi, Alfonso Manzotti, and Pietro Cerveri. Cel-unet: Distance weighted maps and multi-scale pyramidal edge extraction for accurate osteoarthritic bone segmentation in ct scans. *Frontiers in Signal Processing*, 2:857313, 2022.

[21] Eva Schnider, Antal Huck, Mireille Toranelli, Georg Rauter, Magdalena Müller-Gerbl, and Philippe C. Cattin. Improved distinct bone segmentation from upper-body ct using binary-prediction-enhanced multi-class inference. *International Journal of Computer Assisted Radiology and Surgery*, 17:2113–2120, Nov 2022.

[22] Kevin A. Thomas, Lukasz Kidziński, Eni Halilaj, Scott L. Fleming, Guhan R. Venkataraman, Edwin H. G. Oei, Garry E. Gold, and Scott L. Delp. Automated classification of radiographic knee osteoarthritis severity using deep neural networks. *Radiology: Artificial Intelligence*, 2(2):e190065, March 2020.

[23] Elham Taghizadeh, Oskar Truffer, Fabio Becce, Sylvain Eminian, Stacey Gidoin, Alexandre Terrier, Alain Farron, and Philippe Büchler. Deep

learning for the rapid automatic quantification and characterization of rotator cuff muscle degeneration from shoulder ct datasets. *European radiology*, 31:181–190, Jan 2021.

[24] Anish G. Potty, Ajish S. R. Potty, Nicola Maffulli, Lucas A. Blumenschein, Deepak Ganta, R. Justin Mistovich, Mario Fuentes, Patrick J. Denard, Paul M. Sethi, Anup A. Shah, and Ashim Gupta. Approaching artificial intelligence in orthopaedics: Predictive analytics and machine learning to prognosticate arthroscopic rotator cuff surgical outcomes. *Journal of Clinical Medicine*, 12(6):2369, March 2023.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[26] Sozan Mohammed Ahmed and Ramadhan J. Mstafa. A comprehensive survey on bone segmentation techniques in knee osteoarthritis research: From conventional methods to deep learning. *Diagnostics*, 12(3), 2022.

[27] André Klein, Jan Warszawski, Jens Hillengaß, and Klaus H Maier-Hein. Automatic bone segmentation in whole-body ct images. *International Journal of Computer Assisted Radiology and Surgery*, 14(1):21–29, 2019.

[28] Hyojune Kim, Keewon Shin, Hoyeon Kim, Eui sup Lee, Seok Won Chung, Kyoung Hwan Koh, and Namkug Kim. Can deep learning reduce the time and effort required for manual segmentation in 3d reconstruction of mri in rotator cuff tears? *PLoS ONE*, 17(10), 1 2022.

[29] Cheng Chen, Siyu Qi, Kangneng Zhou, Tong Lu, Huansheng Ning, and Ruoxiu Xiao. Pairwise attention-enhanced adversarial model for automatic bone segmentation in ct images. *Physics in medicine and biology*, 68, Jan 2023.

[30] Abhinav Suri, Brandon C. Jones, Grace Ng, Nancy Anabaraonye, Patrick Beyrer, Albi Domi, Grace Choi, Sisi Tang, Ashley Terry, Thomas Leichner, Iman Fathali, Nikita Bastin, Helene Chesnais, and Chamith S. Rajapakse. A deep learning system for automated, multi-modality 2d segmentation of vertebral bodies and intervertebral discs. *Bone*, 149:115972, Aug 2021.

[31] Luca Marsilio, Andrea Moglia, Matteo Rossi, Alfonso Manzotti, Luca Mainardi, and Pietro Cerveri. Combined edge loss unet for optimized segmentation in total knee arthroplasty preoperative planning. *Bioengineering*, 10(12):1433, December 2023.

[32] Shunjiro Noguchi, Mizuho Nishio, Masahiro Yakami, Keita Nakagomi, and Kaori Togashi. Bone segmentation on whole-body ct using convolutional neural network with novel data augmentation techniques. *Computers in Biology and Medicine*, 121:103767, June 2020.

[33] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, Xin Liu, Jie Chen, Huihui Zhou, Ismail Ben Ayed, and Hairong Zheng. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):5915, 2021.

[34] Arnaud Boutillon, Pierre-Henri Conze, Christelle Pons, Valérie Burdin, and Bhushan Borotikar. Generalizable multi-task, multi-domain deep segmentation of sparse pediatric imaging datasets via multi-scale contrastive regularization and multi-joint anatomical priors. *Medical Image Analysis*, 81:102556, Oct 2022.

[35] Qingqing Zhao, Quanlong Feng, Jianlun Zhang, Jingxu Xu, Zi xuan Wu, Chencui Huang, and Hui zong Yuan. Glenoid segmentation from ct scans based on a two-stage deep learning model for glenoid bone loss evaluation. *Journal of Shoulder and Elbow Surgery*, 32(12):e624–e635, December 2023.

[36] Fatih Uysal, Firat Hardalac, Ozan Peker, Tolga Tolunay, and Nil Tokgoz. Classification of shoulder x-ray images with deep learning ensemble models. *Applied Sciences*, 11(6):2723, March 2021.

[37] Claudio E. von Schacky, Jae Ho Sohn, Felix Liu, Eugene Ozhinsky, Pia M. Jungmann, Lorenzo Nardo, Magdalena Posadzy, Sarah C. Foreman, Michael C. Nevitt, Thomas M. Link, and Valentina Pedoia. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Radiology*, 295(1):136–145, April 2020.

[38] Yinghong Zhou, Yu Zheng, Yinfeng Tian, Youfang Bai, Nian Cai, and Ping Wang. Scan: sequence-based context-aware association network for hepatic vessel segmentation. *Medical & Biological Engineering & Computing*, 62(3):817–827, November 2023.

[39] Muhammad Imran, Jonathan R. Krebs, Veera Rajasekhar Reddy Gopu, Brian Fazzone, Vishal Balaji Sivaraman, Amarjeet Kumar, Chelsea Viscardi, Robert Evans Heithaus, Benjamin Shickel, Yuyin Zhou, Michol A. Cooper, and Wei Shao. Cis-unet: Multi-class segmentation of the aorta in computed tomography angiography via context-aware shifted window self-attention. *Computerized Medical Imaging and Graphics*, 118:102470, December 2024.

[40] Bolun Zeng, Li Chen, Yuanyi Zheng, and Xiaojun Chen. Adaptive multi-dimensional weighted network with category-aware contrastive learning for fine-grained hand bone segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):3985–3996, July 2024.

[41] Cristiano Patrício, João C Neves, and Luís F Teixeira. Explainable deep learning methods in medical image classification: A survey. *ACM Computing Surveys*, 56(4):1–41, 2023.

[42] Jeong-Woon Kang, Chunsu Park, Dong-Eon Lee, Jae-Heung Yoo, and MinWoo Kim. Prediction of bone mineral density in ct using deep learning with explainability. *Frontiers in Physiology*, 13:1061911, 2023.

[43] Zhibin Liao, Kewen Liao, Haifeng Shen, Marouska F Van Boxel, Jasper Prijs, Ruurd L Jaarsma, Job N Doornberg, Anton Van den Hengel, and Johan W Verjans. Cnn attention guidance for improved orthopedics radiographic fracture classification. *IEEE journal of biomedical and health informatics*, 26(7):3139–3150, 2022.

[44] Canyu Pan, Luoyu Lian, Jieyun Chen, and Risheng Huang. Femurtumornet: Bone tumor classification in the proximal femur using densenet model based on radiographs. *Journal of Bone Oncology*, 42:100504, 2023.

[45] Uwe Schneider, Eros Pedroni, and Antony Lomax. The calibration of ct hounsfield units for radiotherapy treatment planning. *Physics in Medicine and Biology*, 41(1):111–124, January 1996.

[46] Mohammed Elsharkawi, Balkan Cakir, Heiko Reichel, and Thomas Kappe. Reliability of radiologic glenohumeral osteoarthritis classifications. *Journal of Shoulder and Elbow Surgery*, 22(8):1063–1067, 2013.

[47] Peter Habermeyer, Petra Magosch, Christel Weiß, Nael Hawi, Sven Lichtenberg, Mark Tauber, and Bastian Ipach. Classification of humeral head pathomorphology in primary osteoarthritis: a radiographic and in vivo photographic analysis. *Journal of Shoulder and Elbow Surgery*, 26(12):2193–2199, 2017.

[48] Benjamin D. Kleim, Maximillian Hinz, Stephanie Geyer, Bastian Scheiderer, Andreas B. Imhoff, and Sebastian Siebenlist. A 3-dimensional classification for degenerative glenohumeral arthritis based on humeroscapular alignment. *Orthopaedic Journal of Sports Medicine*, 10(8):23259671221110512, Aug 2022.

[49] Michael Carl, Kaustubh Lall, Darren Pai, Eric Y. Chang, Sheronda Statum, Anja Brau, Christine B. Chung, Maggie Fung, and Won C. Bae. Shoulder bone segmentation with deeplab and u-net. *Osteology*, 4(2):98–110, June 2024.

[50] Osman Berk Satir, Pezhman Eghbali, Fabio Becce, Patrick Goetti, Arnaud Meylan, Kilian Rothenbühler, Robin Diot, Alexandre Terrier, and Philippe Büchler. Automatic quantification of scapular and glenoid morphology from ct scans using deep learning. *European Journal of Radiology*, 177:111588, August 2024.

[51] Xinhong Mu, Yi Cui, Rongpeng Bian, Long Long, Daliang Zhang, Huawen Wang, Yidong Shen, Jingjing Wu, and Guoyou Zou. In-depth learning of automatic segmentation of shoulder joint magnetic resonance images based on convolutional neural networks. *Computer methods and programs in biomedicine*, 211:106325, Nov 2021.

[52] Elham Taghizadeh, Oskar Truffer, Fabio Becce, Sylvain Eminian, Stacey Gidoin, Alexandre Terrier, Alain Farron, and Philippe Büchler. Deep learning for the rapid automatic quantification and characterization of rotator cuff muscle degeneration from shoulder ct datasets. *European Radiology*, 31(1):181–190, July 2020.

[53] Aashay Kekatpure, Aditya Kekatpure, Sanjay Deshpande, and Sandeep Srivastava. Development of a diagnostic support system for distal humerus fracture using artificial intelligence. *International Orthopaedics*, 48(5):1303–1311, March 2024.

[54] John G. Horneff and Viviana M. Serra Lopez. Preoperative planning for anatomic total shoulder arthroplasty. *The Journal of the American Academy of Orthopaedic Surgeons*, 30:e1207–e1216, October 2022.

[55] Patrick Goetti, Patrick J Denard, Philippe Collin, Mohamed Ibrahim, Adrien Mazzolari, and Alexandre Lädermann. Biomechanics of anatomic and reverse shoulder arthroplasty. *EFORT open reviews*, 6(10):918–931, 2021.

[56] Rami Ibrahim and M Omair Shafiq. Explainable convolutional neural networks: A taxonomy, review, and future directions. *ACM Computing Surveys*, 55(10):1–37, 2023.