RAM-W600: A Multi-Task Wrist Dataset and Benchmark for Rheumatoid Arthritis

Songxiao Yang[1*] **Haolin Wang**[2*] **Yao Fu**[2] **Ye Tian**[3] **Tamotsu Kamishima**[2]
Masayuki Ikebe[2] **Yafei Ou**[1†] **Masatoshi Okutomi**[1]

[1] Institute of Science Tokyo, Tokyo, Japan
[2] Hokkaido University, Sapporo, Japan
[3] The University of Tokyo, Tokyo, Japan

Abstract

Rheumatoid arthritis (RA) is a common autoimmune disease that has been the focus of research in computer-aided diagnosis (CAD) and disease monitoring. In clinical settings, conventional radiography (CR) is widely used for the screening and evaluation of RA due to its low cost and accessibility. The wrist is a critical region for the diagnosis of RA. However, CAD research in this area remains limited, primarily due to the challenges in acquiring high-quality instance-level annotations. (i) The wrist comprises numerous small bones with narrow joint spaces, complex structures, and frequent overlaps, requiring detailed anatomical knowledge for accurate annotation. (ii) Disease progression in RA often leads to osteophyte, bone erosion (BE), and even bony ankylosis, which alter bone morphology and increase annotation difficulty, necessitating expertise in rheumatology. This work presents a multi-task dataset for wrist bone in CR, including two tasks: (i) wrist bone instance segmentation and (ii) Sharp/van der Heijde (SvdH) BE scoring, which is the first public resource for wrist bone instance segmentation. This dataset comprises 1048 wrist conventional radiographs of 388 patients from six medical centers, with pixel-level instance segmentation annotations for 618 images and SvdH BE scores for 800 images. This dataset can potentially support a wide range of research tasks related to RA, including joint space narrowing (JSN) progression quantification, BE detection, bone deformity evaluation, and osteophyte detection. It may also be applied to other wrist-related tasks, such as carpal bone fracture localization. We hope this dataset will significantly lower the barrier to research on wrist RA and accelerate progress in CAD research within the RA-related domain.
Benchmark & Code: github.com/YSongxiao/RAM-W600
Data & Dataset Card: huggingface.co/datasets/TokyoTechMagicYang/RAM-W600

1 Introduction

The wrist is a highly complex joint that facilitates a wide range of motion and bears substantial mechanical loads during daily activities. Due to its anatomical complexity and functional demands, the wrist is particularly susceptible to various pathological conditions [17]. Among these, rheumatoid arthritis (RA) is a common and debilitating autoimmune disease that frequently affects the wrist joint early in its progression [63]. It is marked by joint swelling and tenderness, which progressively leads to joint destruction and significant disability. Radiographic analysis plays a pivotal role in the diagnosis and management of RA, with joint space narrowing (JSN) progression and bone erosion (BE) serving as key markers for evaluating and tracking disease progression [2, 56]. However, traditional radiographic assessment heavily relies on the radiologist's expertise and subjective interpretation

*Equal Contribution.
†Corresponding Author (ou.y.ac@m.titech.ac.jp)

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

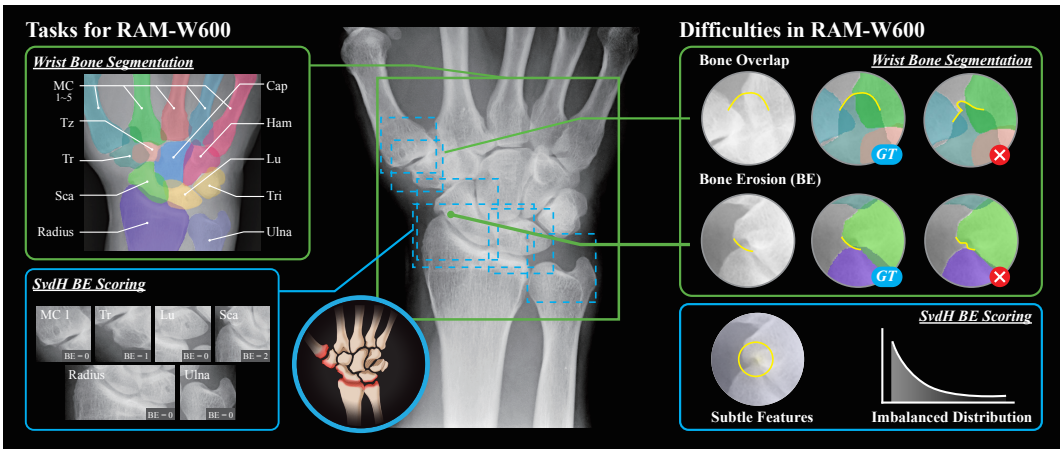


Figure 1: Overview of the RAM-W600 dataset, designed for wrist bone segmentation and SvdH BE scoring tasks. (MC 1 to 5: Metacarpal 1st to 5th; Tz: Trapezoid; Tr: Trapezium; Sca: Scaphoid; Radius: DistalRadius; Cap: Capitate; Ham: Hamate; Lu: Lunate; Tri: Pisiform & Triquetrum; Ulna: DistalUlna)

to detect subtle pathological features, which is time-consuming and often associated with limited accuracy and sensitivity. As a result, the development of computer-aided diagnostic (CAD) systems has attracted growing interest from both academic and industrial communities [65, 35, 74, 75].

Accurate segmentation of wrist bones is critically important in medical image analysis, as it serves as a foundational step for numerous downstream tasks essential to the diagnosis and management of RA. These tasks include, but are not limited to, the evaluation of bone deformities, detection of osteophytes, and assessment of JSN. For example, in bone deformity analysis, precise segmentation is required to extract geometric features such as bone angles, alignment, and morphological irregularities across longitudinal scans [29]. In osteophyte detection, segmented bone contours help identify abnormal bony outgrowths that are often hard to distinguish in raw radiographs due to anatomical overlap [54]. Similarly, accurate quantification of JSN depends on the precise boundaries between adjacent bones [32, 53, 76]. Segmentation errors can lead to incorrect inter-bone distance measurements, which are essential for monitoring disease progression.

However, the annotation process of a large-scale dataset is highly challenging and labor-intensive due to the anatomical complexity of the wrist and different pathological changes in the wrist bones. As shown in Fig. 1, (i) Obscured edges due to overlapping structures. The wrist, a structurally complex joint system, features tightly interlocked carpal bones [5]. This configuration frequently leads to overlapping phenomena in conventional radiography (CR), which significantly complicates the identification of each bone's outer edges. (ii) Morphological alterations resulting from pathological conditions. Due to the progression of RA and other pathological changes, BE, JSN, and osteophyte formation can affect certain bones or joints to varying degrees, often leading to substantial alterations in bone morphology [68, 18, 30]. Moreover, these factors may interact in diverse and combinatorial ways, further complicating the consistency and accuracy of annotations.

Sharp/van der Heijde (SvdH) BE scoring [71] is a widely recognized task in the automated diagnosis of RA. Nevertheless, it remains highly challenging due to difficulties in both annotation and model training. On the annotation side [59, 72], (i) accurate annotation demands specialized rheumatological expertise, as assessing the severity of BE is inherently complex. (ii) The process is subjective and prone to substantial inter-observer variability, resulting in inconsistent and uncertain ground truth labels. This subjectivity and ambiguity undermine the quality of supervision available for model training. From a training perspective, (i) the task is further complicated by a severe class imbalance, as cases of high-grade erosion are underrepresented in most clinical datasets. (ii) The pathological features of BE are often subtle, highly localized, small in scale, and demonstrate minimal variation across severity levels, thereby posing substantial challenges for automated detection and classification. Collectively, these factors render SvdH BE scoring a challenging task in developing robust and generalizable deep learning models for RA assessment.

Table 1: Comparison between RAM-W600 and the publicly available annotated datasets. Ann/Img: Annotations per image.

Modality	Dataset	Year	Images (Ann/Img)	Resolution (mm/pixel)	Age (Mean±SD)	Tasks		Purpose
						Mask	Score	
CR	Halabi et al. [24]	2019	14236 (15)	-	0.35	✓		BAA
	Sun et al. [66]	2022	674 (31)	-	-		✓	RA
	Ours (RAM-W600)	2025	618 (15) + 800 (6)	0.15*	49.86±20.26	✓	✓	RA
CT	Moore et al. [49]	2007	30 (15)	-	26.25±3.33	✓		-

BAA: Bone Age Assessment; *: Internal cohorts only.

In this paper, we introduce **R**heumatoid **A**rthritis **M**odeling-**W**rist 600 (RAM-W600), a multi-task dataset for wrist bone in conventional radiography. It comprises 1048 wrist conventional radiographs of 388 patients from six medical centers. Among them, 618 high-resolution wrist radiographs are provided with expert-verified instance-level annotations for wrist bone segmentation, along with 4800 SvdH BE scores. This dataset is expected to support a wide range of downstream tasks, such as anatomical structure localization, erosion progression analysis, and automated disease staging, thereby contributing to the broader advancement of computer-aided diagnosis in RA. Our primary contributions are threefold:

- **First Multi-Task dataset for RA**: RAM-W600 is the first public large-scale dataset dedicated to both segmentation and SvdH BE scoring tasks, providing a valuable benchmark for developing and validating deep learning algorithms in conventional radiographs. Its multi-institutional composition ensures diversity in acquisition conditions, enhancing the generalizability of trained models.
- **High-quality annotations**: We provide high-quality pixel-level annotations of the wrist bones, including careful handling of overlapping region boundaries, and SvdH BE score in the region of interest (ROI).
- **Comprehensive benchmarks**: We present a benchmark for wrist bone instance segmentation and SvdH BE scoring, enabling standardized evaluation and comparison of algorithms for automated RA assessment.

2 Related Works

2.1 Hand Radiographic Datasets

Although hand radiographic data are relatively easy to acquire, the complex anatomical structure of the hand and the inherent limitations of current imaging techniques present significant challenges for accurate annotation. As shown in Table 1, these challenges are further intensified in disease-specific applications, such as the diagnosis and monitoring of RA, where high-quality, expert-annotated datasets remain scarce. Earlier efforts produced computed tomography (CT) datasets with segmentation masks [49], but these were limited by small sample sizes. More recently, Halabi et al. [24] released a large-scale CR dataset annotated with segmentation masks; however, its utility is confined to pediatric bone age assessment and limited to selected phalangeal regions. In contrast, RA-specific datasets such as that of Sun et al. [66] provide severity scores but lack pixel-wise annotations, thereby constraining their applicability to tasks requiring precise image segmentation.

2.2 Wrist Bone Segmentation

Radiological analysis of the wrist bones is central to the study of hand-related disorders. In particular, image segmentation plays an important role and holds significant value for both clinical practice and research, as summarized in Table 2. Notable progress has been made in wrist bone segmentation using various imaging modalities, including CT and magnetic resonance imaging (MRI). Early studies employed mathematical modeling techniques to achieve relatively mature segmentation outcomes on CT and MRI scans [4, 20]. With recent advances in deep learning, both 2D and 3D segmentation of wrist bones in CT and MRI has further matured [80, 69, 57, 62, 58], enabling more specialized investigations into disease-induced bone pathologies. In contrast, research on wrist bone segmentation

Table 2: Summary of recent works on wrist segmentation. Ann/Img: Annotations per image.

Modality	Works	Year	Backbone	Dataset	Images (Ann/Img)	Age (Mean±SD)	Objects			Purpose
							F	C	UR	
CR	Yang et al. [78]	2021	ResNet	Private	720 (2)	36±13			✓	BMD
	Kang et al. [34]	2022	Mask R-CNN	Private	702 (10)	-		✓	✓	-
	Lee et al. [38]	2023	SAM	Private	192 (7)	-	✓		✓	BMD
	Du et al. [16]	2024	GRU-Unet	[24] & Private	2000 (13)	-	✓		✓	BAA
CT	Anas et al. [4]	2016	-	[49] & Private	60 (15)	-	✓	✓	✓	-
	Sebro et al. [62]	2022	-	Private	196 (17)	64.9±8.7	✓	✓	✓	BMD
4DCT	Teule et al. [69]	2024	nnU-Net	Private	19 (9)	-		✓	✓	-
MRI	Foster et al. [20]	2018	-	Private	160 (8)	47.1±9.25		✓		OA
	Radke et al. [57]	2021	CNN	Private	56 (8)	30.7±13.6		✓	✓	LWI
	Yiu et al. [80]	2024	nnU-Net	Private	80 (15)	54±12	✓	✓	✓	RA(BME)
	Raith et al. [58]	2025	3D U-Net	Private	15 (8)	27.8±3.11		✓		-

F: Finger Bones; **C**: Carpal Bones; **UR**: Radius and Ulna Bones;
BMD: Bone Mineral Density; **BAA**: Bone Age Assessment;
RA(BME): Rheumatoid Arthritis with Bone Marrow Edema; **LWI**: Ligamentous Wrist Injuries.

Table 3: Summary of recent works on RA-related scoring. Ann/Img: Annotations per image.

Modality	Works	Year	Backbone	Dataset	Images (Ann/Img)	Patients	Age (Mean±SD)	Tasks	
								SvdH BE	Others
CR	Hirano et al. [27]	2019	CNN	Private	216 (15)	108	64.9±4.87	✓	SvdH JSN
	Ureten et al. [70]	2020	CNN	Private	180 (2)	180	-		RA & HC
	Maziarz et al. [46]	2021	Unet	[66]	674 (31)	562	-		Damage
	Hioki et al. [26]	2021	Yolo V3	Private	50 (4)	-	-		Destruction
	Miyama et al. [48]	2022	DNN	Private	226 (31)	40	61.5±11.6	✓	SvdH JSN
	Wang et al. [73]	2022	Yolo	Private	915 (30)	400	>20		mTSS
	Sun et al. [66]	2022	DNN	[66]	674 (31)	562	-	✓	SvdH JSN
	Bo et al. [6]	2024	ResNet	Private	3818 (10)	-	-	✓	SvdH JSN
	Lien et al. [41]	2025	Yolo V7	Private	823 (30)	-	>20		mTSS
HR-pQCT	Folle et al. [19]	2022	GradCAM	Private	932 (3)	617	45±15		HC & RA & PsA
MRI	Schlereth et al. [61]	2024	CNN	Private	211 (66)	112	54.1±12.4	✓	osteitis & synovitis

mTSS: modified total Sharp Score; **PsA**: psoriatic arthritis; **HC**: healthy controls.

from radiographs is still limited. Although several deep learning-based methods have been proposed [78, 34, 16, 38], few studies focus on complex pathological conditions such as RA. Due Due to the limitations of CR imaging, its two-dimensional nature causes anatomical overlap, tissue superposition, and low contrast, which make it difficult to identify bone boundaries and anatomical structures. In addition, although CR is more accessible and cost-effective than CT or MRI, accurate annotation is still difficult, especially in cases with active osteoarticular lesions. As a result, there are few high-quality, publicly available annotated datasets. This lack of data makes it hard to train and evaluate reliable segmentation models.

Consequently, achieving high-precision wrist bone segmentation in radiographs of patients with complex pathological conditions remains a critical challenge. Addressing this issue holds substantial potential for advancing efficient and user-friendly clinical decision support systems.

2.3 Detection and Assessment of BE

The SvdH scoring system has been widely used to evaluate various joint abnormalities in RA. As summarized in Table 3, an increasing number of automated methods have been developed in recent years to facilitate RA radiograph scoring. These approaches are typically based on the SvdH system and aim to assess key indicators such as JSN, BE, and the modified total Sharp score (mTSS). Most models are trained and validated on private datasets. Earlier studies primarily employed convolutional neural networks (CNNs) for feature extraction and classification [28, 48, 66, 6]. Recently, object detection-based models have been introduced [26, 41], enabling the integration of lesion localization and scoring within end-to-end pipelines and enhancing both automation and usability. Some studies have explored RA classification and severity assessment using scoring systems other than SvdH [70, 46, 26]. Furthermore, research on automated RA assessment has expanded to encompass various imaging modalities, including MRI [61] and high-resolution peripheral quantitative computed tomography (HR-pQCT) [19], along with the investigation of alternative scoring methods and evaluation standards, thereby further advancing the field of RA imaging analysis.

In summary, the wrist joint is one of the most anatomically complex and diagnostically significant regions in RA radiographs, offering substantial clinical and research value. Notably, the integration of precise wrist bone segmentation and lesion scoring within a multi-task learning framework has emerged as a key direction in advancing automated RA analysis. However, publicly available hand CR datasets remain significantly limited, particularly those focused on the wrist. Most datasets lack high-precision segmentation masks specifically annotated for the wrist region, and their corresponding BE scores are often incomplete or missing. This limits their suitability for RA-specific research, which requires high-quality, multi-dimensional annotated data. Therefore, the development of a wrist-focused CR dataset with detailed anatomical annotations and validated clinical scores is essential for the progress of intelligent RA imaging assessment.

3 Overview of Dataset

Ethical Considerations RAM-W600 dataset is in compliance with the guidelines of the Declaration of Helsinki and obtained approval from the Ethics Committee of Hokkaido University (approval number: 24-104) and Institute of Science Tokyo (approval number: A24672). All radiographs included in this dataset were collected with informed consent for research use and public release.

3.1 Image and Annotation

The dataset consisted of 1048 hand posteroanterior projection (PA) radiographs from 207 patients with RA and 181 patients without RA. The images were obtained from six different institutions: Hokkaido Medical Center for Rheumatic Diseases (HMCRD) (Sapporo, Japan), Sapporo City General Hospital (SCGH) (Sapporo, Japan), Hokkaido University (HU) (Sapporo, Japan), Digital Hand Atlas (DHA) from the University of Southern California (CA, US) [10], Bone Tumor X-ray Radiograh Dataset (BTXRD) from Monash University (Melbourne, Australia) [79], and FracAtlas (FA) from Islamic University of Technology (Gazipur, Bangladesh) [1]. Each institution has its own CR systems, and the dataset is managed using the digital imaging and communications in medicine (DICOM) standard, with the detailed information of imaging parameters referred to Table 6.

We employed specialized imaging processing methodologies to systematically construct wrist joint data. Initially, image cropping techniques were applied to focus on the wrist region, effectively eliminating interference from extraneous anatomical structures. Annotation was performed by a dedicated team consisting of a radiological technologist and two clinically experienced experts, including a board-certified radiologist with 25 years of experience and an orthopedic doctor with 5 years of clinical practice. This multidisciplinary expertise ensured that the annotations were both medically accurate and clinically relevant. For the segmentation task, initial contours were delineated by the radiological technologist and subsequently verified by the radiologist. For the classification task, three annotators independently assigned labels, and any discrepancies were resolved through discussion and consensus. Based on this protocol, the annotation comprised three principal components:

- **Anatomical Structure Annotation**: Precise contour delineation was performed for 14 wrist bones, including the first-fifth metacarpals (MC1-5), trapezium (Tr), trapezoid (Tz), scaphoid (Sca), lunate (Lu), capitate (Cap), hamate (Ham), pisiform & triquetrum (Tri), distal radius (Radius), and distal ulna (Ulna). A multi-label annotation strategy was implemented to independently mark each osseous structure.
- **Bone Location Annotation**: The SvdH BE scoring system focuses on five key joint regions: Metacarpal 1st, Trapezoid, Scaphoid, Lunate, Distal Radius, and Distal Ulna. We performed ROI annotations on these areas.
- **SvdH BE Scoring Annotation**: BE assessment was conducted using the SvdH scoring system, specifically targeting five critical articular groups: Metacarpal 1st, Trapezoid, Scaphoid, Lunate, Distal Radius, and Distal Ulna. This systematic evaluation focused on quantifying erosive changes at these predetermined anatomical sites.

With the division of these images, a comprehensive annotation pipeline was adopted, including professional annotators and strict inspection procedures. Further details of the data division and annotation can be found in Sec. B.

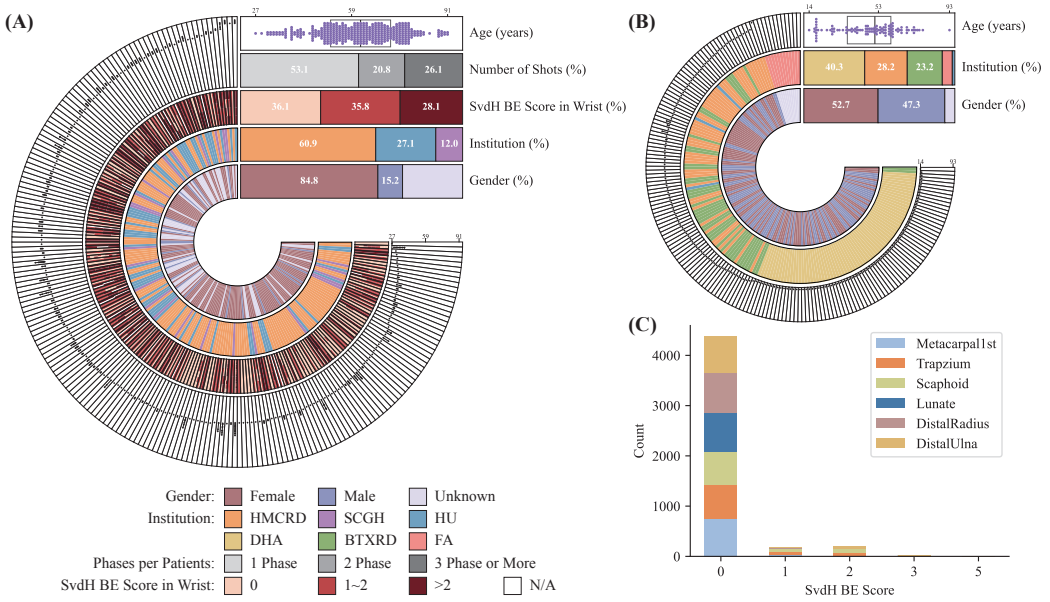


Figure 2: Distribution and Statistics for the age, gender, institution, number of shots, and BE scores in the RAM-W600 dataset. (A) Circular overview of the RA cohort. Each bar around the circular plot represents a unique patient. The concentric layers from inner to outer encode: (i) Gender distribution. (ii) Institution distribution. (iii) SvdH BE scores in both wrists for each study. Patients with multiple shots are represented multiple times in this layer. (iv) The patient's age at each acquisition. (B) Circular overview of the Non-RA cohort. Similar to (A), each bar around the circular plot represents a unique patient. (C) Distribution of SvdH BE scores by joint surface.

3.2 Statistics of RAM-W600

We present statistical analyses of the RAM-W600 dataset to characterize both the RA cohort and the Non-RA cohort. Key attributes, including patient demographics (age, gender), institutional sources, follow-up frequency (phase), and BE scores, were systematically examined. In addition, joint-specific BE score distributions were compared across anatomical locations. Detailed statistics are summarized in Fig. 2. The RA cohort (A) collected from HMCRD, SCGH, and HU primarily consists of female patients. This pattern accords with epidemiologic evidence, since RA occurs most frequently in women between 30 and 50 years of age [55, 37]. Most wrist joints in this cohort are annotated with an SvdH BE score of 0, indicating minimal erosive changes, while non-zero scores remain relatively uncommon. In addition, most patients underwent only a single imaging phase, and the age distribution spans a broad range. The Non-RA cohort (B) includes healthy controls from HMCRD and HU, as well as additional cohorts from DHA, BTXRD, and FA. This group exhibits a more balanced gender ratio and also shows a broad age distribution. Joint-level SvdH BE annotations in (C) reveal a highly imbalanced distribution across joint surfaces, with the vast majority of joint faces assigned a score of 0. Higher scores, such as 3 or 5, are nearly absent. Such an imbalanced distribution has been commonly reported in clinical cohorts [9, 33]. With advances in medical care, early detection and the effective use of disease-modifying treatments have markedly reduced the number of patients progressing to late-stage RA, making high BE scores increasingly rare in modern cohorts. Moreover, CR imaging is primarily performed to monitor early and moderate stages of RA, while advanced stages are less frequently imaged in current clinical practice.

4 Experiments and Benchmarks

4.1 Wrist Bone Segmentation

To evaluate wrist bone instance segmentation performance, we tested a series of widely used supervised architectures and their variants on the RAM-W600 dataset, as well as recent foundation

Table 4: Instance segmentation results obtained on the Test set. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	DSC ↑ (%)			NSD ↑ (%)			VOE ↓ (%)	MSD ↓ (pix)	Params (M)	Time (ms)
	BE	nonBE	All	BE	nonBE	All				
Supervised Models										
Unet [60]	96.70±0.05	96.83±0.09	96.79±0.08	83.59±0.41	83.27±0.58	83.36±0.52	6.13±0.14	1.83±0.10	7.94	13.57
DeepLabV3 [12]	96.55±0.03	96.86±0.02	96.78±0.02***	82.17±0.28	82.89±0.23	82.69±0.19*	6.20±0.03	1.37±0.01	26.00	9.19
FPN [42]	96.59±0.07	96.85±0.07	96.78±0.07***	81.45±0.68	81.83±0.63	81.73±0.64	6.19±0.13	1.38±0.02	23.15	8.43
PSPNet [82]	95.30±0.05	95.55±0.07	95.48±0.06*	71.58±0.46	71.02±0.47	71.17±0.45	8.52±0.10	2.05±0.04	21.49	4.46
DeepLabV3+ [13]	96.78±0.01	97.01±0.03	96.95±0.02***	83.56±0.17	83.73±0.27	83.68±0.20	5.87±0.04	1.31±0.02	22.43	5.57
SegResNet [50]	96.48±0.21	96.64±0.20	96.60±0.20*	81.78±1.25	81.79±1.14	81.79±1.17	6.50±0.37	1.79±0.21	1.60	4.93
Unet++ [83]	97.21±0.02	97.37±0.04	97.33±0.03*	86.85±0.26	87.04±0.26	86.99±0.23	5.15±0.06	1.36±0.07	2.41	14.83
SegFormer [77]	96.82±0.06	97.09±0.02	97.01±0.03***	84.24±0.46	84.65±0.20	84.53±0.25	5.74±0.06	1.28±0.00	21.87	5.04
TransUNet [11]	97.50±0.04	97.67±0.06	97.62±0.05***	89.20±0.24	89.59±0.36	89.48±0.33	4.60±0.10	1.05±0.03	105.91	22.05
UKAN [39]	96.74±0.06	96.98±0.05	96.91±0.05***	83.15±0.22	83.41±0.16	83.33±0.16	5.93±0.10	1.34±0.04	6.36	10.30
UMambaBot [45]	97.40±0.04	97.58±0.02	97.53±0.03**	88.77±0.23	88.94±0.18	88.89±0.20	4.76±0.05	1.13±0.01	4.42	15.12
UMambaEnc [45]	97.44±0.05	97.61±0.03	97.56±0.03**	88.92±0.31	89.17±0.29	89.10±0.28	4.71±0.06	1.11±0.02	4.58	16.44
SwinUMamba [43]	**97.65±0.02**	**97.80±0.02**	**97.75±0.02****	**90.56±0.12**	**90.77±0.15**	**90.71±0.14**	**4.35±0.03**	**1.06±0.05**	59.89	38.52
Foundation Models										
SAM (box) [36]	88.91±5.59	88.67±4.80	88.74±5.01	65.91±6.06	63.82±7.32	64.40±7.03	18.45±5.23	4.25±1.46	641.09	193.47
SAM (pt) [36]	80.18±7.10	80.46±11.13	80.38±10.14	55.56±9.93	55.84±12.06	55.76±11.47	28.42±10.82	18.21±16.08	641.09	32.72
MedSAM (box) [44]	85.07±2.05	85.06±2.69	85.07±2.52	39.91±6.46	38.38±7.28	38.81±7.07	25.15±3.59	5.97±1.19	93.74	99.48

Time: Inference time per image on RTX 4090 GPU.
Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).
Mann-Whitney U Test between BE & nonBE, *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

models. The supervised architectures included Unet [60], DeepLabV3 [12], FPN [42], PSPNet [82], DeepLabV3+ [13], SegResNet [50], Unet++ [83], SegFormer [77], TransUNet [11], UKAN [39], UMambaBot [45], UMambaEnc [45], SwinUMamba [43] , while the foundation models comprised SAM [36] and MedSAM [44]. In line with standard practice, segmentation performance was quantified using Dice Similarity Coefficient (DSC) [15]; Normalized Surface Dice (NSD) [52]; Volumetric Overlap Error (VOE) [67]; Mean Surface Distance (MSD) [67]; and Relative Absolute Volume Difference (RAVD) [67]. The threshold for the NSD was set to 2 pixels.

Implementation details The dataset was split according to the configuration shown in Table 7 (a) in Sec. B.4. BE and Non-BE cases were stratified using the SvdH BE score, where radiographs with a total BE score greater than 0 were considered BE cases. Cases were stratified based on the SvdH BE score, where radiographs with a total score greater than zero were classified as BE, while those with a score of zero were classified as non-BE. For supervised models, all experiments were repeated five times on a single NVIDIA RTX 4090 GPU using five fixed random seeds (1024, 2025, 3407, 4096, and 5214) to ensure reproducibility, whereas foundation models were evaluated by a single inference run without repetition. All radiographs were resized to 512×512 pixels and used as input to the model. Model training employed the AdamW optimizer with a weight decay of 1e-2. The initial learning rate was set to 1e-4 and decayed according to a cosine annealing schedule (CosineAnnealingLR). Training was carried out for 100 epochs using a batch size of 8 and standard data augmentation techniques.

Benchmark results The results shown in Table 4 demonstrate that mainstream supervised models achieve outstanding performance in terms of DSC, with the highest value reaching 97.75% (SwinUMamba), indicating robust overlap accuracy in global segmentation regions. However, NSD values remain comparatively low (peak: 90.71%), with significant variations across models, highlighting persistent challenges in bone boundary delineation. This limitation is closely tied to the inherent complexities of wrist bone segmentation: inter-bone occlusions leading to blurred boundaries, and BE regions characterized by abnormal texture and edge variations, which further exacerbate segmentation difficulty. Meanwhile, group analysis reveals statistically significant differences ($p < 0.05$–0.001) in DSC between BE and nonBE samples for most models, confirming the detrimental impact of BE on segmentation performance. In contrast, the NSD metric exhibited no statistically significant differences between groups. This discrepancy may stem from the heightened sensitivity of NSD to boundary errors and the larger variance in boundary-related discrepancies within the dataset, underscoring the intrinsic difficulty in handling bone edges. In addition, foundation models such as SAM and MedSAM achieved lower DSC ($\leq$ 88.7% for SAM and 85.1% for MedSAM) and

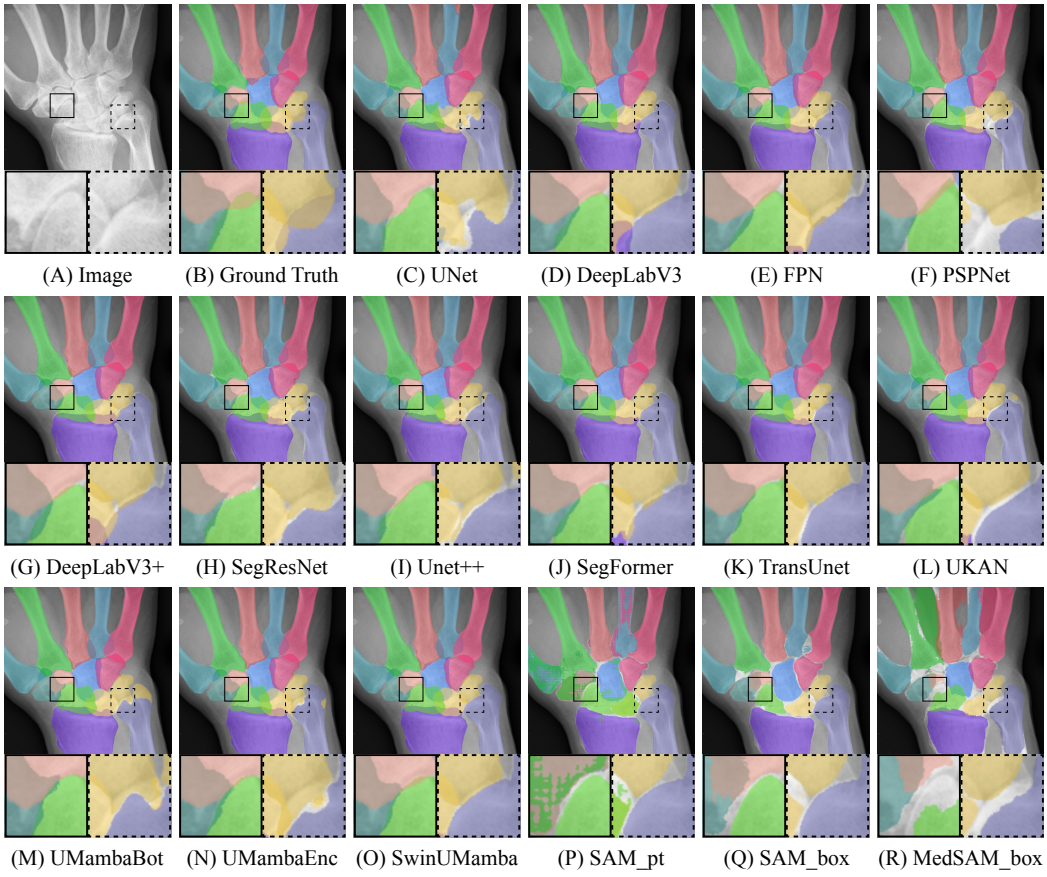


Figure 3: Wrist bone segmentation visualization results. The solid box indicates segmentation challenges caused by BE, while the dashed box represents difficulties arising from bone overlap.

NSD ($\leq$ 75.0%) compared with supervised models, further demonstrating the limited adaptability of general-purpose segmentation priors to the specialized task of wrist bone delineation. In conclusion, the primary bottleneck in this task lies in improving model robustness for complex bone boundaries.

Visualization Some representative results are shown in Fig. 3. Compared to the ground truth, mainstream supervised networks exhibit performance degradation in segmenting bone edges with multi-layer occlusions, a challenge that becomes particularly pronounced under complex occlusion scenarios. Current models also demonstrate notable inconsistency, lacking reliable solutions to address this issue effectively. Furthermore, in the context of BE (RA), most existing architectures fail to adequately capture the inward collapse of bone edges caused by erosive changes. However, networks incorporating Mamba-based architectures show partial improvements in handling such morphological distortions, as evidenced by comparative analyses. Visualization results further corroborate the persistent challenges in this segmentation task, primarily attributed to bone overlaps and erosion-induced structural anomalies. These factors collectively lead to fragmented or inaccurate edge predictions, emphasizing the need for dedicated architectural innovations. In contrast, foundation models such as SAM and MedSAM exhibit less precise boundary localization and frequent edge discontinuities, underscoring their limited adaptability to the fine-grained requirements of wrist bone segmentation.

Unlike natural images or other medical imaging modalities such as MRI and CT, CR captures the cumulative attenuation of X-rays along their path, resulting in grayscale representations of internal structures. This often leads to overlapping anatomical features and blurred boundaries in two-dimensional images. Moreover, pathological BE caused by RA can induce notable morphological changes in bone structure, further complicating segmentation. Traditional image processing and segmentation techniques often struggle to accurately delineate overlapping bone boundaries or detect

Table 5: BE & nonBE classification results obtained on the Test set. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	BACC↑ (%)	F1-Score↑ (%)	DOR↑	ACC↑ (%)	SEN↑ (%)	SPC↑ (%)	PRE↑ (%)	Params	Time (ms)
MobileViT [47]	**52.64±0.61**	11.85±0.48	1.82±0.19	81.42±0.87	21.06±0.93	84.23±1.13	9.31±0.76	4.94M	4.53
ResNet [25]	51.75±1.02	10.89±1.06	1.16±0.41	78.27±1.31	23.10±2.54	80.40±1.60	7.79±0.74	0.70M	1.99
MobileNet [31]	47.84±2.52	10.79±1.98	0.89±0.38	74.08±6.31	17.02±4.60	78.66±7.31	9.07±2.99	0.69M	1.72
LeViT [23]	49.29±0.69	6.73±1.90	1.51±1.45	84.17±2.46	8.49±3.57	90.09±3.73	8.99±5.68	7.01M	2.65
EfficientFormer [40]	50.63±1.86	**12.40±2.43**	1.06±0.31	72.04±3.45	**27.90±8.73**	73.37±5.23	8.82±0.63	3.25M	3.63
MedMamba [81]	50.83±1.00	6.91±3.51	**5.89±9.66**	86.56±4.48	8.94±7.45	92.73±6.55	**11.56±7.98**	14.45M	6.06
ConvKAN [7]	49.26±0.84	3.49±3.13	0.44±0.37	**87.42±4.55**	3.82±4.89	**94.70±6.32**	6.56±7.09	3.49M	29.96

Time: Inference time per image on RTX 4090 GPU.

morphological abnormalities resulting from pathological alterations. To address these challenges, future research may benefit from exploring multi-scale feature fusion strategies and advanced edge refinement techniques. Given the relatively fixed spatial arrangement of bones, incorporating global contextual information could be particularly advantageous for improving segmentation accuracy.

4.2 Classification of BE

The advanced binary classification methods of BE were evaluated on the RAM-W600 dataset. The selected classification models included MobileViT [47], ResNet [25], MobileNet [31], LeViT [23], EfficientFormer [40], MedMamba [81], and ConvKAN [7]. In line with standard practice, classification performance was quantified using balanced accuracy (BACC) [8], F1-score [14], diagnostic odds ratio (DOR) [22], accuracy (ACC) [21], sensitivity (SEN) [3], specificity (SPC) [3], and precision (PRE) [64].

Implementation details The dataset was split according to the configuration shown in Table 7 (b) in Sec. B.4. BE classification was performed on a joint-surface basis, focusing on the six joint surfaces of clinical interest. A joint surface was labeled as BE if its corresponding SvdH BE score was greater than 0. All experiments were repeated five times on a single NVIDIA RTX4090 GPU using five fixed random seeds (1024, 2025, 3407, 4096, 5214) to ensure reproducibility. All ROIs were resized to 224×224 pixels and used as input to the model. Model training utilized the AdamW optimizer with a weight decay of 1e-2. The initial learning rate was set to 1e-6 and decayed using a cosine annealing schedule (CosineAnnealingLR). Training was performed for 100 epochs with a batch size of 16 and standard data augmentation techniques.

Benchmark results The results in Table 5 reveal that mainstream models achieve only modest performance in terms of BACC and F1-score, with the best results reaching 52.64% (MobileViT) and 12.40% (EfficientFormer), respectively, indicating limited robustness in distinguishing BE from nonBE cases. In contrast, the DOR exhibits considerable variability across models, peaking at 5.89 (MedMamba). Notably, some models (e.g., ConvKAN) achieve relatively high specificity (94.70%) while suffering from extremely low sensitivity (3.82%), reflecting a strong bias toward negative predictions. This inconsistency across metrics underscores the difficulty of the task, likely stemming from extreme class imbalance and the subtle radiographic presentation of BE. The confusion matrices in Fig. 4 further illustrate this imbalance, showing that all models consistently perform better on the majority class (nonBE) than on the minority class (BE), highlighting the inherent challenge of detecting subtle BE features.

Future research should further focus on enhancing the model's ability to detect subtle BE features under highly imbalanced data conditions. In clinical practice, early or mild BE lesions typically exhibit low visibility, presenting as small and inconspicuous regions that are easily confounded by overlapping bones, imaging artifacts, or noise. Although advanced BE lesions are more prominent in size, they often co-occur with other RA manifestations such as joint space narrowing and osteophyte formation, introducing additional sources of interference. These challenges collectively complicate the end-to-end scoring process for BE across different stages of the disease. To improve model performance on such difficult samples, future efforts may explore targeted augmentation strategies for minority classes or develop architectures capable of extracting weak pathological signals. Such

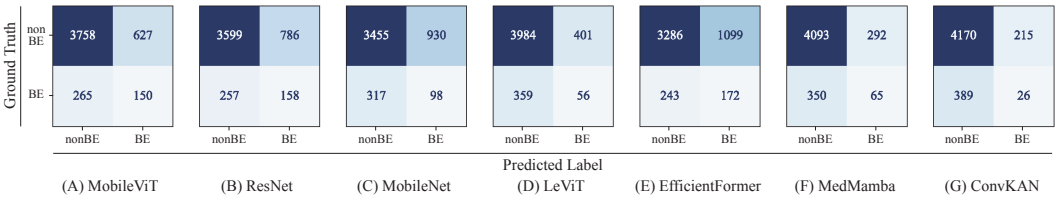


Figure 4: BE & nonBE confusion matrix results for classification of BE.

advancements would enhance both the sensitivity and robustness of RA imaging assessment tools, thereby promoting their clinical applicability and translational value.

5 Conclusions and Limitations

We have introduced RAM-W600, the first publicly available multi-task CR dataset for RA assessment, which encompasses two key tasks: wrist bone segmentation and SvdH BE localization and scoring. RAM-W600 has provided high-quality pixel-level annotations for the anatomically complex wrist region, which often presents with severe bone overlapping and erosive changes. In addition to detailed annotations, the dataset includes benchmark results for both segmentation and BE scoring tasks. Experimental findings have demonstrated the considerable challenges posed by these tasks, including the accurate delineation of bones in the presence of occlusion and erosion in the segmentation task, and the robust scoring of affected joints in the grading task. By establishing RAM-W600 and its associated benchmarks, we have offered a valuable resource for advancing research in medical image analysis. This dataset has opened new avenues for the development and validation of robust CAD systems and holds promise for improving diagnostic accuracy and clinical decision-making in the management of RA.

Despite its contributions to advancing CAD for RA, the RAM-W600 dataset has several limitations. First, the RA cases are primarily derived from a single geographic region and a relatively homogeneous ethnic population, which may limit the generalizability of models trained on the dataset to more diverse clinical settings. This lack of demographic variability could reduce the robustness of model performance across different populations. Second, the distribution of SvdH BE scores is imbalanced, with certain score levels notably underrepresented. This imbalance poses challenges for both training and evaluation, particularly in learning fine-grained disease severity and ensuring consistent performance across all stages of RA progression.

For wrist bone segmentation, future research should focus on developing dedicated network architectures that incorporate multi-scale contextual information and boundary-sensitive mechanisms. Such designs are essential to address the challenges posed by anatomical complexity and projection-induced overlap in wrist radiographs, particularly for achieving accurate delineation in regions affected by bone overlap and BE. Regarding the SvdH BE scoring task, early-stage lesions often present weak radiographic signals and are obscured by overlapping structures, while advanced-stage cases commonly exhibit coexisting RA-related features, resulting in complex local characteristics. In addition, the highly imbalanced distribution of BE samples continues to hinder lesion recognition in current approaches. To overcome these limitations, it is crucial to design model components capable of extracting subtle pathological features, thereby improving sensitivity and robustness in detecting early-stage BE. Advancements in these directions are expected to significantly enhance the automation of RA wrist image analysis and reinforce its clinical utility in diagnosis and longitudinal disease monitoring.

Acknowledgments

This work was supported by JST BOOST and JST SPRING, Japan Grant Number JPMJBS2426, JPMJSP2106 and JPMJSP2180.

References

[1] Iftekharul Abedeen, Md Ashiqur Rahman, Fatema Zohra Prottyasha, Tasnim Ahmed, Tareque Mohmud Chowdhury, and Swakkhar Shatabda. Fracatlas: A dataset for fracture classification, localization and segmentation of musculoskeletal radiographs. *Scientific data*, 10(1):521, 2023.

[2] Daniel Aletaha and Josef S Smolen. Diagnosis and management of rheumatoid arthritis: a review. *Jama*, 320(13):1360–1372, 2018.

[3] Douglas G Altman and J Martin Bland. Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, 308(6943):1552, 1994.

[4] Emran Mohammad Abu Anas, Abtin Rasoulian, Alexander Seitel, Kathryn Darras, David Wilson, Paul St John, David Pichora, Parvin Mousavi, Robert Rohling, and Purang Abolmaesumi. Automatic segmentation of wrist bones in ct using a statistical wrist shape + pose model. *IEEE transactions on medical imaging*, 35(8):1789–1801, 2016.

[5] Anil K Bhat, Bhaskaranand Kumar, and Ashwath Acharya. Radiographic imaging of the wrist. *Indian journal of plastic surgery: official publication of the Association of Plastic Surgeons of India*, 44(2):186, 2011.

[6] Zhiyan Bo, Laura C Coates, and Bartłomiej W Papież. Deep learning models to automate the scoring of hand radiographs for rheumatoid arthritis. In *Annual Conference on Medical Image Understanding and Analysis*, pages 398–413. Springer, 2024.

[7] Alexander Dylan Bodner, Antonio Santiago Tepsich, Jack Natan Spolski, and Santiago Pourteau. Convolutional kolmogorov-arnold networks. *arXiv preprint arXiv:2406.13155*, 2024.

[8] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010.

[9] Karin Bruynesteyn, Désirée van der Heijde, Maarten Boers, Ariane Saudan, Paul Peloso, Harold Paulus, Harry Houben, Bridget Griffiths, John Edmonds, Barry Bresnihan, et al. Determination of the minimal clinically important difference in rheumatoid arthritis joint damage of the sharp/van der heijde and larsen/scott scoring methods by clinical experts and comparison with the smallest detectable difference. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, 46(4):913–920, 2002.

[10] Fei Cao, HK Huang, Ewa Pietka, and Vicente Gilsanz. Digital hand atlas and web-based bone age assessment: system design and implementation. *Computerized medical imaging and graphics*, 24(5):297–307, 2000.

[11] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[12] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[14] Nancy Chinchor and Beth M Sundheim. Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.

[15] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[16] Hongbo Du, Hai Wang, Chunlai Yang, Luyando Kabalata, Henian Li, and Changfu Qiang. Hand bone extraction and segmentation based on a convolutional neural network. *Biomedical Signal Processing and Control*, 89:105788, 2024.

[17] Jörg Eschweiler, Jianzhang Li, Valentin Quack, Björn Rath, Alice Baroncini, Frank Hildebrand, and Filippo Migliorini. Anatomy, biomechanics, and loads of the wrist joint. *Life*, 12(2):188, 2022.

[18] Fatemeh Ezzati and Parham Pezeshk. Radiographic findings of inflammatory arthritis and mimics in the hands. *Diagnostics*, 12(9):2134, 2022.

[19] Lukas Folle, David Simon, Koray Tascilar, Gerhard Krönke, Anna-Maria Liphardt, Andreas Maier, Georg Schett, and Arnd Kleyer. Deep learning-based classification of inflammatory arthritis by identification of joint shape patterns—how neural networks can tell us where to "deep dive" clinically. *Frontiers in Medicine*, 9:850552, 2022.

[20] Brent Foster, Anand A Joshi, Marissa Borgese, Yasser Abdelhafez, Robert D Boutin, and Abhijit J Chaudhari. Wrist: A wrist image segmentation toolkit for carpal bone delineation from mri. *Computerized Medical Imaging and Graphics*, 63:31–40, 2018.

[21] Jerome Friedman. The elements of statistical learning: Data mining, inference, and prediction. *(No Title)*, 2009.

[22] Afina S Glas, Jeroen G Lijmer, Martin H Prins, Gouke J Bonsel, and Patrick MM Bossuyt. The diagnostic odds ratio: a single indicator of test performance. *Journal of clinical epidemiology*, 56(11):1129–1135, 2003.

[23] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12259–12269, 2021.

[24] Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[26] Yuri Hioki, Koji Makino, Kensuke Koyama, Hirotaka Haro, and Hidetsugu Terada. Evaluation method of rheumatoid arthritis by the x-ray photograph using deep learning. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 444–447. IEEE, 2021.

[27] T Hirano, M Nishide, N Nonaka, J Seita, K Ebina, K Sakurada, et al. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. rheumatol adv pract. 2019. available from:; 3.

[28] Toru Hirano, Masayuki Nishide, Naoki Nonaka, Jun Seita, Kosuke Ebina, Kazuhiro Sakurada, and Atsushi Kumanogoh. Development and validation of a deep-learning model for scoring of radiographic finger joint destruction in rheumatoid arthritis. *Rheumatology advances in practice*, 3(2):rkz047, 2019.

[29] Stefanie Hirsiger, Andreas Schweizer, Junichi Miyake, Ladislav Nagy, and Philipp Fürnstahl. Corrective osteotomies of phalangeal and metacarpal malunions using patient-specific guides: Ct-based evaluation of the reduction accuracy. *Hand*, 13(6):627–636, 2018.

[30] Jan Lucas Hoving, Rachelle Buchbinder, Stephen Hall, Gary Lawler, Peter Coombs, Stephen McNealy, Paul Bird, and David Connell. A comparison of magnetic resonance imaging, sonography, and radiography of the hand in patients with early rheumatoid arthritis. *The Journal of rheumatology*, 31(4):663–675, 2004.

[31] Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[32] Yinghe Huo, Koen L Vincken, Desiree Van Der Heijde, Maria JH De Hair, Floris P Lafeber, and Max A Viergever. Automatic quantification of radiographic wrist joint space width of patients with rheumatoid arthritis. *IEEE Transactions on Biomedical Engineering*, 64(11):2695–2703, 2017.

[33] LMA Jansen, IE Van der Horst-Bruinsma, D Van Schaardenburg, PD Bezemer, and BAC Dijkmans. Predictors of radiographic joint damage in patients with early rheumatoid arthritis. *Annals of the rheumatic diseases*, 60(10):924–927, 2001.

[34] Bo-kyeong Kang, Yelin Han, Jaehoon Oh, Jongwoo Lim, Jongbin Ryu, Myeong Seong Yoon, Juncheol Lee, and Soorack Ryu. Automatic segmentation for favourable delineation of ten wrist bones on wrist radiographs using convolutional neural network. *Journal of Personalized Medicine*, 12(5):776, 2022.

[35] Kathryn M Kingsmore, Christopher E Puglisi, Amrie C Grammer, and Peter E Lipsky. An introduction to machine learning and analysis of its use in rheumatic diseases. *Nature Reviews Rheumatology*, 17(12):710–730, 2021.

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[37] Tore K Kvien, Till Uhlig, Sigrid Ødegård, and Marte S Heiberg. Epidemiological aspects of rheumatoid arthritis: the sex ratio. *Annals of the New York academy of Sciences*, 1069(1):212–222, 2006.

[38] Hyungeun Lee, Ung Hwang, Seungwon Yu, Chang-Hun Lee, and Kijung Yoon. Osteoporosis prediction from hand and wrist x-rays using image segmentation and self-supervised learning. *arXiv preprint arXiv:2311.06834*, 2023.

[39] Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4652–4660, 2025.

[40] Yanyu Li, Geng Yuan, Yang Wen, Ju Hu, Georgios Evangelidis, Sergey Tulyakov, Yanzhi Wang, and Jian Ren. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, 35:12934–12949, 2022.

[41] Chung-Yueh Lien, Hao-Jan Wang, Cheng-Kai Lu, Tzu-Hsuan Hsu, Woei-Chyn Chu, and Chien-Chih Lai. Deep learning with an attention mechanism for enhancing automated modified total sharp/van der heijde scoring of hand x-ray images in rheumatoid arthritis. *Journal of Medical and Biological Engineering*, pages 1–9, 2025.

[42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[43] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024.

[44] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

[45] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[46] Krzysztof Maziarz, Anna Krason, and Zbigniew Wojna. Deep learning for rheumatoid arthritis: Joint detection and damage scoring in x-rays. *arXiv preprint arXiv:2104.13915*, 2021.

[47] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[48] Kazuki Miyama, Ryoma Bise, Satoshi Ikemura, Kazuhiro Kai, Masaya Kanahori, Shinkichi Arisumi, Taisuke Uchida, Yasuharu Nakashima, and Seiichi Uchida. Deep learning-based automatic-bone-destruction-evaluation system using contextual information from other joints. *Arthritis Research & Therapy*, 24(1):227, 2022.

[49] Douglas C Moore, Joseph J Crisco, Theodore G Trafton, and Evan L Leventhal. A digital database of wrist bone anatomy and carpal kinematics. *Journal of biomechanics*, 40(11):2537–2542, 2007.

[50] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI brainlesion workshop*, pages 311–320. Springer, 2018.

[51] Arjun Nanduri, Alison Kim, Carolyn Nolan, Jesse Dubey, and Andrew Barbera. Triquetrum fracture with pisiform dislocation. *Orthopedic Reviews*, 14(2):32339, 2022.

[52] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. *Journal of medical Internet research*, 23(7):e26151, 2021.

[53] Yafei Ou, Prasoon Ambalathankandy, Ryunosuke Furuya, Seiya Kawada, Tianyu Zeng, Yujie An, Tamotsu Kamishima, Kenichi Tamura, and Masayuki Ikebe. A sub-pixel accurate quantification of joint space narrowing progression in rheumatoid arthritis. *IEEE Journal of Biomedical and Health Informatics*, 27(1):53–64, 2023.

[54] Benjamin Schultz Overgaard, Anders Bossel Holst Christensen, Lene Terslev, Thiusius Rajeeth Savarimuthu, and Søren Andreas Just. Artificial intelligence model for segmentation and severity scoring of osteophytes in hand osteoarthritis on ultrasound images. *Frontiers in Medicine*, 11:1297088, 2024.

[55] Alexander Pfeil, Joachim Böttcher, Bettina E Seidl, Jens-Peter Heyne, Alexander Petrovitch, Torsten Eidner, Hans-Joachim Mentzel, Gunter Wolf, Gert Hein, and Werner A Kaiser. Computer-aided joint space analysis of the metacarpal-phalangeal and proximal-interphalangeal finger joint: normative age-related and gender-specific data. *Skeletal radiology*, 36(9):853–864, 2007.

[56] Raj Ponnusamy, Ming Zhang, Zhiheng Chang, Yue Wang, Carmine Guida, Samantha Kuang, Xinyue Sun, Jordan Blackadar, Jeffrey B Driban, Timothy McAlindon, et al. Automatic measuring of finger joint space width on hand radiograph using deep learning and conventional computer vision methods. *Biomedical signal processing and control*, 84:104713, 2023.

[57] Karl Ludger Radke, Lena Marie Wollschläger, Sven Nebelung, Daniel Benjamin Abrar, Christoph Schleich, Matthias Boschheidgen, Miriam Frenken, Justus Schock, Dirk Klee, Jens Frahm, et al. Deep learning-based post-processing of real-time mri to assess and quantify dynamic wrist movement in health and disease. *Diagnostics*, 11(6):1077, 2021.

[58] Stefan Raith, Matthias Deitermann, Tobias Pankert, Jianzhang Li, Ali Modabber, Frank Hölzle, Frank Hildebrand, and Jörg Eschweiler. Multi-label segmentation of carpal bones in mri using expansion transfer learning. *Physics in Medicine and Biology*, 2025.

[59] Eric EJ Raven, Michel PJ van den Bekerom, Annechien Beumer, and C Niek van Dijk. Radiocarpal and midcarpal instability in rheumatoid patients: a systematic review. *The Open Orthopaedics Journal*, 9:246, 2015.

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[61] Maja Schlereth, Melek Yalcin Mutlu, Jonas Utz, Sara Bayat, Tobias Heimann, Jingna Qiu, Chris Ehring, Chang Liu, Michael Uder, Arnd Kleyer, et al. Deep learning-based classification of erosion, synovitis and osteitis in hand mri of patients with inflammatory arthritis. *RMD open*, 10(2):e004273, 2024.

[62] Ronnie Sebro and Cynthia De la Garza-Ramos. Machine learning for opportunistic screening for osteoporosis from ct scans of the wrist and forearm. *Diagnostics*, 12(3):691, 2022.

[63] Kassem Sharif, Alaa Sharif, Fareed Jumah, Rod Oskouian, and R Shane Tubbs. Rheumatoid arthritis in review: Clinical, anatomical, cellular and molecular points of view. *Clinical Anatomy*, 31(2):216–223, 2018.

[64] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[65] Berend C Stoel, Marius Staring, Monique Reijnierse, and Annette HM van der Helm-van Mil. Deep learning in rheumatological image interpretation. *Nature Reviews Rheumatology*, 20(3):182–195, 2024.

[66] Dongmei Sun, Thanh M Nguyen, Robert J Allaway, Jelai Wang, Verena Chung, Thomas V Yu, Michael Mason, Isaac Dimitrovsky, Lars Ericson, Hongyang Li, et al. A crowdsourcing approach to develop machine learning models to quantify radiographic joint damage in rheumatoid arthritis. *JAMA network open*, 5(8):e2227423–e2227423, 2022.

[67] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15:1–28, 2015.

[68] Bachir Taouli, Souhil Zaim, Charles G Peterfy, John A Lynch, Alexander Stork, Ali Guermazi, Bo Fan, Kenneth H Fye, and Harry K Genant. Rheumatoid arthritis of the hand and wrist: comparison of three imaging techniques. *American Journal of Roentgenology*, 182(4):937–943, 2004.

[69] EHS Teule, N Lessmann, EPA van der Heijden, and S Hummelink. Automatic segmentation and labelling of wrist bones in four-dimensional computed tomography datasets via deep learning. *Journal of Hand Surgery (European Volume)*, 49(4):507–509, 2024.

[70] Kemal Üreten, Hasan Erbay, and Hadi Hakan Maraş. Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network. *Clinical rheumatology*, 39:969–974, 2020.

[71] DMFM Van der Heijde. How to read radiographs according to the sharp/van der heijde method. *The Journal of rheumatology*, 27(1):261–263, 2000.

[72] DMFM Van der Heijde, T Dankert, F Nieman, R Rau, and M Boers. Reliability and sensitivity to change of a simplification of the sharp/van der heijde radiological assessment in rheumatoid arthritis. *Rheumatology*, 38(10):941–947, 1999.

[73] Hao-Jan Wang, Chi-Ping Su, Chien-Chih Lai, Wun-Rong Chen, Chi Chen, Liang-Ying Ho, Woei-Chyn Chu, and Chung-Yueh Lien. Deep learning-based computer-aided diagnosis of rheumatoid arthritis with hand x-ray images conforming to modified total sharp/van der heijde score. *Biomedicines*, 10(6):1355, 2022.

[74] Haolin Wang, Yafei Ou, Prasoon Ambalathankandy, Gen Ota, Pengyu Dai, Masayuki Ikebe, Kenji Suzuki, and Tamotsu Kamishima. Bls-gan: A deep layer separation framework for eliminating bone overlap in conventional radiographs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7674–7681, 2025.

[75] Haolin Wang, Yafei Ou, Prasoon Ambalathankandy, Gen Ota, Pengyu Dai, Masayuki Ikebe, Kenji Suzuki, and Tamotsu Kamishima. Layer separation: Adjustable joint space width images synthesis in conventional radiography. *arXiv preprint arXiv:2502.01972*, 2025.

[76] Haolin Wang, Yafei Ou, Wanxuan Fang, Prasoon Ambalathankandy, Naoto Goto, Gen Ota, Taichi Okino, Jun Fukae, Kenneth Sutherland, Masayuki Ikebe, et al. A deep registration method for accurate quantification of joint space narrowing progression in rheumatoid arthritis. *Computerized Medical Imaging and Graphics*, 108:102273, 2023.

[77] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[78] Fan Yang, Xin Weng, Yuehong Miao, Yuhui Wu, Hong Xie, and Pinggui Lei. Deep learning approach for automatic segmentation of ulna and radius in dual-energy x-ray imaging. *Insights into Imaging*, 12:1–9, 2021.

[79] Shunhan Yao, Yuanxiang Huang, Xiaoyu Wang, Yiwen Zhang, Ian Costa Paixao, Zhikang Wang, Charla Lu Chai, Hongtao Wang, Dinggui Lu, Geoffrey I Webb, et al. A radiograph dataset for the classification, localization, and segmentation of primary bone tumors. *Scientific Data*, 12(1):88, 2025.

[80] Chungwun Yiu, James Francis Griffith, Fan Xiao, Lin Shi, Bingjing Zhou, Su Wu, and Lai-Shan Tam. Automated quantification of wrist bone marrow oedema, pre-and post-treatment, in early rheumatoid arthritis. *Rheumatology Advances in Practice*, 8(3):rkae073, 2024.

[81] Yubiao Yue and Zhenzhang Li. Medmamba: Vision mamba for medical image classification. *arXiv preprint arXiv:2403.03849*, 2024.

[82] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[83] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, proceedings 4*, pages 3–11. Springer, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [NA]
   (b) Did you include complete proofs of all theoretical results? [NA]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section A.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1 and Section 4.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1 and Section 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Section B.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Section A.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [NA]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [NA]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [NA]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA]

A RAM-W600 Data Access and Format

The data can be accessed on HuggingFace at https://huggingface.co/datasets/TokyoTechMagicYang/RAM-W600. The dataset has a permanent DOI: https://doi.org//10.57967/hf/5328. The benchmark and code can be accessed on Github at https://github.com/YSongxiao/RAM-W600.

The dataset is organised in two main folders (Segmentation/ and BE_SvdH_Prediction/) corresponding to two tasks. The dataset structure is shown as follows:

```
RAM-W600/
|-- JointLocationDetection/
|   |-- images/               # Contains all input images in BMP format
|       |-- 0145_0004_L.bmp
|       |-- 0145_0004_R.bmp
|       |-- ...
|   |-- Joints.coco.json      # Ground-truth annotations for joints'
    ↪ locations
|-- BoneSegmentation/
|   |-- images/                   # Contains all input images in BMP
    ↪ format
|       |-- 0001_0001_L.bmp
|       |-- 0001_0001_R.bmp
|       |-- ...
|   |-- masks/                    # Contains corresponding masks in
    ↪ NumPy (.npy) format
|       |-- train/
|           |-- 0006_0001_L.npy
|           |-- ...
|       |-- val/
|           |-- 0001_0001_R.npy
|           |-- ...
|       |-- test/
|           |-- 0002_0001_L.npy
|           |-- ...
|-- SvdHBEScoreClassification/
|   |-- train/
|       |-- 0003_0001_L/
|           |-- DistalRadius.bmp
|           |-- DistalUlna.bmp
|           |-- ...
|       |-- ...
|   |-- val/
|       |-- 0001_0001_R/
|           |-- DistalRadius.bmp
|           |-- DistalUlna.bmp
|           |-- ...
|       |-- ...
|   |-- test/
|       |-- 0005_0001_L/
|           |-- DistalRadius.bmp
|           |-- DistalUlna.bmp
|           |-- ...
|       |-- ...
|   |-- JointBE_SvdH_GT.json      # Ground-truth annotations for joint
    ↪ BE scores
|-- Metadata.xlsx       # Metadata for the dataset
```

- BoneSegmentation/images/: Contains all original images in BMP format. Each file is named as [PatientID]_[StudyID]_[L/R].bmp, where L and R indicate the left or right hand, respectively.

- `BoneSegmentation/masks/`: Contains the corresponding segmentation masks stored as NumPy arrays (.npy). The masks are organized into `train/`, `val/`, and `test/` subsets, with filenames matching the corresponding images.
- `JointLocationDetection/images/`: Contains all original images in BMP format. Each file is named as `[PatientID]_[StudyID]_[L/R].bmp`, where L and R indicate the left or right hand, respectively.
- `JointLocationDetection/Joints.coco.json`: A JSON file containing ground-truth annotations for the joint scores, indexed by case identifiers. The format of entries in JSON file is shown as follows:


```
{
  "images": [
    {
      "id": 0,
      "file_name": "0334_0001_R.bmp",
      "width": 600,
      "height": 600
    },
    ...
  ],
  "annotations": [
    {
      "id": 3281,
      "image_id": 546,
      "category_id": 1,
      "bbox": [170.0, 305.88, 235, 235],
      "area": 23680.95,
      "segmentation": [],
      "iscrowd": 0
    },
    ...
  ],
  "categories": [
    {
      "id": 1,
      "name": "DistalRadius",
      "supercategory": "joint"
    },
    {
      "id": 2,
      "name": "DistalUlna",
      "supercategory": "joint"
    },
    ...
  ]
}
```


  Each entry in the `images` list represents a wrist radiograph, while the `annotations` list contains bounding box annotations for individual joints, identified by their `category_id`. The `categories` section maps category IDs to specific joint names such as `Lunate`, `Scaphoid`, and `Trapezium`.
- `SvdHBEScoreClassification/train/val/test/`: Each subset contains folders named as `[PatientID]_[StudyID]_[L/R]`, representing individual cases. Inside each folder are six ROI images in BMP format, each corresponding to different joint surfaces.
- `SvdHBEScoreClassification/JointBE_SvdH_GT.json`: A JSON file containing ground-truth annotations for the joint scores, indexed by case identifiers. The format of entries in JSON file is shown as follows:


```
{
    "identifier": "0035_0001_L",
    "patient_id": "0035",
    "study_id": "0001",
```

```
    "hand": "L",
    "joints": {
      "Metacarpal1st": 0,
      "Trapezium": 0,
      "Scaphoid": 0,
      "Lunate": 0,
      "DistalRadius": 0,
      "DistalUlna": 0
    }
}
```


- `Metadata.xlsx`: An Excel file containing patient-, study-, and image-level metadata. It provides identifiers, demographic attributes, institutional sources, imaging parameters, and clinical reference scores. The key columns are described as follows:
  - `Mapped Image Stem`: A normalized identifier of each radiographic study in the format `XXXX_XXXX`. This stem represents the study itself rather than a direct image file. The corresponding radiographs are determined by appending the hand side (`_L` or `_R`) to the stem, which specifies the left or right hand image.
  - `PatientID`: An anonymized patient identifier, allowing multiple studies from the same individual to be grouped.
  - `StudyID`: An anonymized study identifier, denoting examinations at different time points.
  - `IsRA`: Binary flag for rheumatoid arthritis status (1 = RA patient, 0 = non-RA control).
  - `PatientSex`: Patient sex, recorded as `M` (male), `F` (female) or `O` (unknown).
  - `PatientAge`: Age at the time of the study, expressed in years (e.g., 59.5).
  - `InstitutionName`: Source institution where the radiograph was acquired (e.g., HMCRD, SCGH, HU).
  - `StudyDate (Days)`: Relative day of the study, with baseline examination set to 0.
  - `ImagerPixelSpacing`: In-plane resolution of the image in millimeters, recorded as `[row spacing, column spacing]`.
  - `[Rows, Columns]`: Image resolution in pixels.
  - `L / R`: Indicators for whether valid SvdH scores are available for the left or right hand (1 = available, 0 = unavailable).
  - `SvdH_L / SvdH_R`: Total Sharp/van der Heijde erosion scores for the left and right hands.
  - `Joint-specific scores`: Integer scores for six anatomical regions (Metacarpal1st, Trapezium, Scaphoid, Lunate, DistalRadius, DistalUlna), recorded separately for left (_L) and right (_R) hands. Higher scores indicate more severe erosion.

B Detailed Information of RAM-W600

B.1 License and Attribution

The conventional radiographs and associated annotations (segmentation masks and SvdH BE scores) in the dataset are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

For proper attribution when using this dataset in any publications or research outputs, please cite with the DOI.

Suggested Citation: Yang, S., Wang, H., Fu, Y., Tian, Y., Kamishima, T., Ikebe, M., Ou, Y., & Okutomi, M.(2025). RAM-W600: A Multi-Task Wrist Dataset and Benchmark for Rheumatoid Arthritis. `https://doi.org/10.57967/hf/5328`

B.2 Data Rights Compliance and Issue Reporting

We are committed to complying data protection rights in accordance with relevant regulations, including but not limited to the General Data Protection Regulation (GDPR). All personally identifiable

Table 6: Radiographic imaging configuration parameters

	HU	HMCRD	SCGH	DHA	BTXRD	FA
Model	-	Radnext 32	KXO-50G	IPI LAB(Secondary)	-	-
Manufacturer	FUJIFILM	HITACHI	TOSHIBA	Array(Secondary)	-	FUJIFILM & Philips
Aluminum filter (mm)	NO	0.5	NO	-	-	-
Tube voltage (kV)	-	50	45	-	-	-
Tube current (mA)	-	100	250	-	-	-
Exposure time (mSec)	-	25	14	-	-	-
Source to image (cm)	-	100	100	-	-	-
Resolution (mm/pixel)	0.15	0.15	0.15	-	-	-
Image size (pixel)	2010×1670	2010×1490	2010×1490	1744×2126	-	-
Bit depth (bit)	16	10	10	16	-	-

HU: Faculty of Health Sciences, Hokkaido University.
HMCRD: Hokkaido Medical Center for Rheumatic Diseases, Japan.
SCGH: Sapporo City General Hospital, Japan.
DHA: Digital Hand Atlas, University of Southern California, US.
BTXRD: Bone Tumor X-ray Radiograph Dataset, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Australia.
FA: FracAtlas, Islamic University of Technology, Bangladesh.

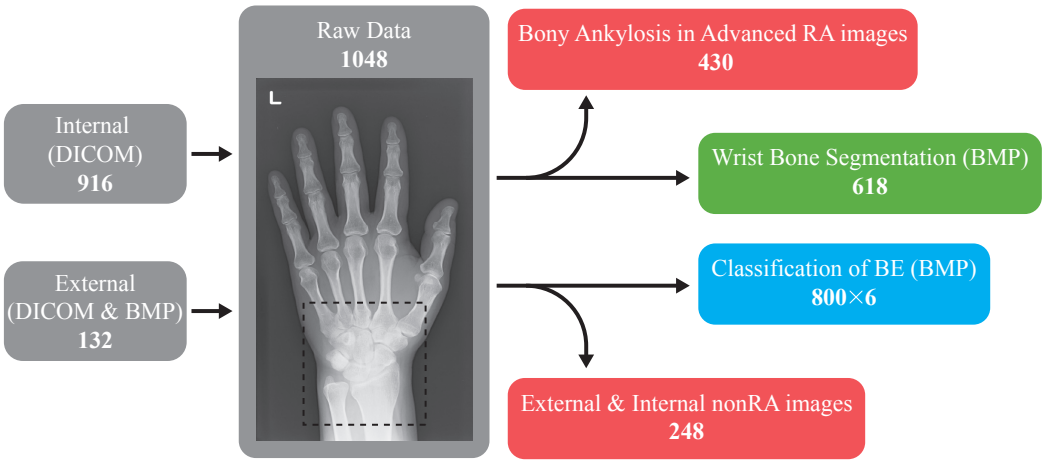


Figure 5: A total of 1048 DICOM-format wrist radiographs were collected, including 916 internal cases from our institutions and 132 external cases from three different sources. Within the internal cohort, 116 images were identified as non-RA, while the remaining were RA cases. All 132 external images were non-RA. After filtering, 430 advanced RA cases with bony ankylosis were excluded. The final dataset was used for two primary tasks: wrist bone instance segmentation (618 BMP images) and BE classification (800 images × 6 joint areas). The external non-RA images were used exclusively for comparison purposes.

information (PII) has been removed through anonymization techniques. If any individual represented in the dataset wishes to have their data removed, we provide a clear and accessible process for issue reporting and resolution via our GitHub repository. Concerned parties are encouraged to contact the authors directly through the contact form linked on the GitHub page. Upon receiving a request, we will engage with the individual to verify their identity and promptly remove the relevant data entries from the dataset.

B.3 Data Acquisition

Radiographs were collected from six institutions with varying imaging configurations, including differences in equipment models, acquisition settings, and image resolutions, as shown in Table 6.

Table 7: Joint score distribution across train, valid, and test sets of RAM-W600.

(a) ***Wrist Bone Segmentation*** Dataset

Score	MC 1	Tr	Sca	Lu	Radius	Ulna
Train Set						
0	414	384	373	421	423	400
1	5	8	14	2	0	9
2	6	30	34	2	2	15
3	0	3	4	0	0	1
5	0	0	0	0	0	0
Valid Set						
0	66	61	56	69	66	59
1	1	3	6	0	3	0
2	2	5	7	0	0	10
3	0	0	0	0	0	0
5	0	0	0	0	0	0
Test Set						
0	117	110	110	121	120	116
1	4	4	0	1	1	2
2	3	8	11	2	3	6
3	0	2	3	0	0	0
5	0	0	0	0	0	0

(b) ***SvdH BE Scoring*** Dataset

Score	MC 1	Tr	Sca	Lu	Radius	Ulna
Train Set						
0	493	357	338	531	541	470
1	55	117	123	11	8	37
2	11	83	79	8	4	38
3	0	2	16	5	6	0
5	0	0	3	4	0	4
Valid Set						
0	67	42	49	74	68	58
1	10	20	11	2	6	11
2	4	18	16	1	3	8
3	0	1	3	0	4	3
5	0	0	2	4	0	1
Test Set						
0	140	90	94	157	151	123
1	18	39	36	2	7	21
2	2	28	26	0	1	13
3	0	3	4	0	1	3
5	0	0	0	1	0	0

Table 8: Institution score distribution across train, valid, and test sets of RAM-W600.

(a) ***Wrist Bone Segmentation*** Dataset

Score	HMCRD	SCGH	HU	DHA	BTXRD	FA
Train Set						
NonRA	540	0	24	318	210	36
0	1038	208	41	0	0	0
1	26	9	3	0	0	0
$\geq$2	76	17	4	0	0	0
Valid Set						
NonRA	24	0	12	48	24	12
0	212	21	24	0	0	0
1	13	0	0	0	0	0
$\geq$2	21	3	0	0	0	0
Test Set						
NonRA	96	0	0	72	48	24
0	354	100	0	0	0	0
1	8	4	0	0	0	0
$\geq$2	28	10	0	0	0	0

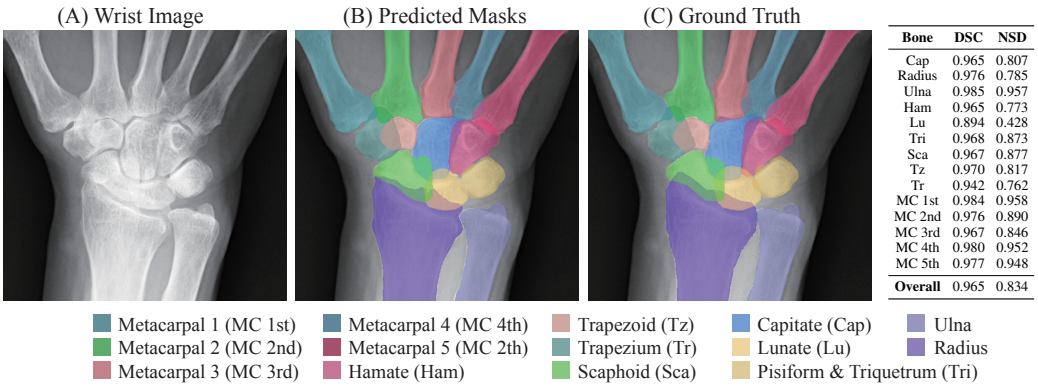
(b) ***SvdH BE Scoring*** Dataset

Score	HMCRD	SCGH	HU
Train Set			
0	1628	353	744
1	93	32	226
$\geq$2	103	41	128
Valid Set			
0	256	18	84
1	15	5	40
$\geq$2	23	1	44
Test Set			
0	408	92	255
1	21	8	94
$\geq$2	39	8	35

B.4 Data Pre-Processing

In the pre-processing pipeline of RAM-W600 (Fig. 5), we first localize the ROI around the wrist across all 1048 DICOM-format hand radiographs. For the wrist bone segmentation task, we exclude 430 images exhibiting bony ankylosis associated with advanced RA, resulting in a curated subset of 618 BMP-format images for segmentation. For the BE classification task, we exclude only 248 non-RA images from the internal and external cohorts, retaining 800 RA cases from our internal dataset. These cases are subsequently converted to BMP format, and six joint-level crops are extracted per image, yielding a total of 4800 samples for BE classification.

The wrist bone segmentation dataset and the SvdH BE scoring dataset are split independently. To prevent data leakage and reduce potential bias, we randomly partition the cases into training, validation, and test sets based on unique patient IDs using an approximate ratio of 70%/10%/20%. Table 7 summarizes the distribution of the six wrist joints (1st Metacarpal, Trapezium, Scaphoid, Lunate, Radius, and Ulna) across scores 0–5 within the training, validation, and test subsets of both datasets. The majority of joints are assigned an SvdH BE score of 0, while those with a score of 5 are



Bone	DSC	NSD
Cap	0.965	0.807
Radius	0.976	0.785
Ulna	0.985	0.957
Ham	0.965	0.773
Lu	0.894	0.428
Tri	0.968	0.873
Sca	0.967	0.877
Tz	0.970	0.817
Tr	0.942	0.762
MC 1st	0.984	0.958
MC 2nd	0.976	0.890
MC 3rd	0.967	0.846
MC 4th	0.980	0.952
MC 5th	0.977	0.948
Overall	0.965	0.834

Figure 6: Wrist bone segmentation.(A) Original wrist radiograph. (B) Predicted instance segmentation masks. (C) Ground truth annotations. The right panel reports the segmentation performance per bone.

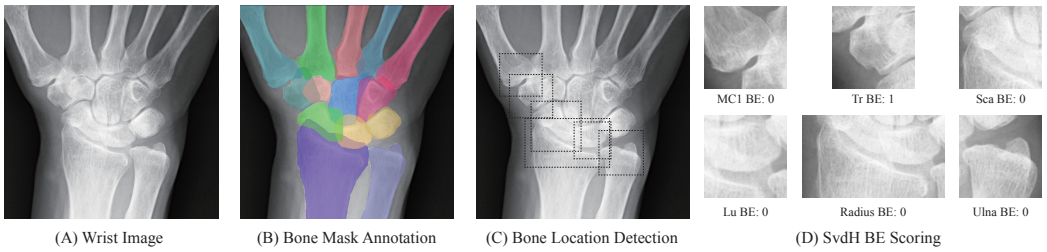


Figure 7: Image input and annotation.(A) Raw wrist radiograph. (B) Instance bone segmentation mask annotation (C) bone location annotations for target regions. (D) SvdH BE scores assigned to each joint region.

extremely rare. In both tasks, the number of joints decreases as the score increases, resulting in a clearly imbalanced distribution.

In addition, Table 8 further details the institution-wise distribution of cases across the two datasets. Table 8a presents the distribution for the wrist bone segmentation task, where both RA and nonRA cases are included, while Table 8b shows the corresponding distribution of joint scores in the SvdH BE scoring dataset. This breakdown highlights the contribution of each collaborating institution and illustrates how score imbalance manifests across different sources and subsets.

B.5 Dataset Maintenance

As the authors and maintainers of this dataset, we affirm that while the dataset is self-contained and does not depend on any external links or content, we may provide future updates, such as adding new cases or incorporating additional tasks. These potential updates aim to enhance the dataset's value while maintaining its long-term usability.

B.6 Wrist Bone Segmentation

Wrist bone segmentation from radiographs is a critical prerequisite for downstream tasks such as joint localization, morphological analysis, and BE scoring in RA assessment. As illustrated in Fig. 6, this task involves delineating multiple overlapping and irregularly shaped carpal and metacarpal bones, which often exhibit low contrast and anatomical ambiguity in radiographs. Accurate segmentation enables reliable quantification of structural features and supports automated interpretation in clinical workflows.

In this task, we annotate 14 distinct wrist bones, including both carpal, metacarpal components and Distal Radius & Distal Ulna. Notably, the Pisiform and Triquetrum bones are difficult to distinguish in clinical practice due to their overlapping appearance and low visibility on standard radiographs. Consequently, it is challenging to evaluate them as independent diagnostic regions [51].

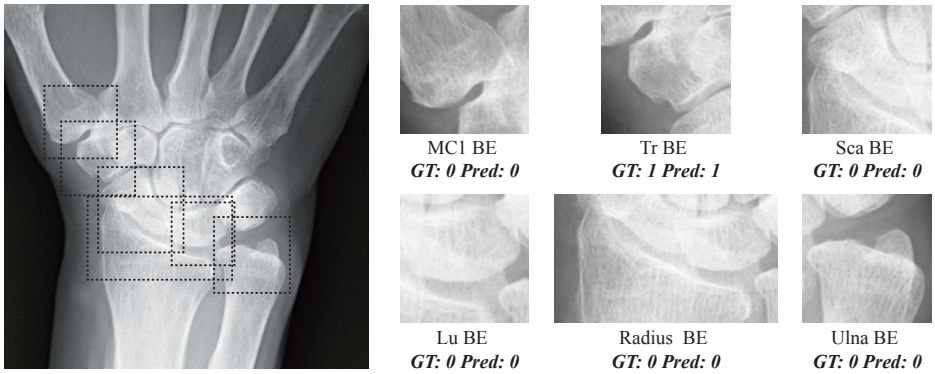


Figure 8: BE and nonBE Classification. The left panel shows six annotated joint regions used for BE classification. The right panels display each joint with ground truth (GT) and predicted (Pred) SvdH BE scores.

Therefore, we merge these two structures into a single category during annotation to reflect their practical indistinguishability. The input to the segmentation model is the wrist ROI cropped from the radiograph, and the output and ground truth are a pixel-wise mask for each annotated bone, as illustrated in Fig. 7.

B.7 Classification of BE

BE classification is a key component of the SvdH scoring system, widely adopted in clinical practice for evaluating joint damage in RA. As illustrated in Fig. 8, this task involves identifying subtle pathological changes in individual carpal bones from radiographs, such as cortical breaks and irregular bone surfaces. The classification task is particularly challenging due to the subtlety of erosion features and the high degree of anatomical overlap in wrist joints. Accurate BE detection is essential for automated RA scoring systems and downstream severity assessment, yet remains difficult for both traditional and deep learning models, especially under class imbalance and in early-stage lesions.

In this task, we annotate the SvdH BE scores for six joint surfaces within the wrist. To formulate the problem as a binary classification task, all joint surfaces with non-zero scores were treated as positive cases (i.e., exhibiting BE), while those with a score of zero were treated as negative cases (i.e., without BE). The input to the model is the ROI corresponding to an individual joint surface, and the output is a probability distribution over the two classes, representing the model's confidence in the presence or absence of BE, as shown in Fig. 7.

C Detailed Analysis of Experimental Results

C.1 Wrist Bone Segmentation

C.1.1 Overall Segmentation Results

Table 9 and Table 10 report the segmentation results in terms of DSC and NSD, which together provide complementary perspectives on overlap accuracy and boundary precision. A consistent observation across both metrics is that supervised models substantially outperform foundation models, with Mamba-based architectures leading the performance on nearly all joints. This gap highlights that current foundation models are not sufficient for wrist bone segmentation, making supervised baselines essential for establishing reliable performance benchmarks.

In terms of DSC (Table 9), supervised models achieve very high overlap accuracy, often above 97% across most bones. SwinUMamba yields the strongest results, surpassing all other methods in almost every region. For example, it achieves 98.84% on the Radius, 98.74% on the MC1, and 98.64% on the Metacarpal 5th. UMambaEnc and UMambaBot also deliver competitive results, particularly on large and structurally less ambiguous bones such as the Distal Ulna (98.81% and 98.75%). Transformer-

Table 9: DSC performance on all joints. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Cap	Radius	Ulna	Ham	Lu	Tri	Sca
Supervised Models							
UNet	96.24±0.11	98.17±0.05	98.20±0.06	96.25±0.09	95.29±0.19	96.37±0.08	96.80±0.13
DeepLabV3	96.49±0.08	98.32±0.04	97.98±0.02	96.26±0.04	95.47±0.06	96.20±0.06	96.68±0.06
FPN	96.32±0.08	98.21±0.03	98.22±0.06	96.08±0.09	95.43±0.12	96.26±0.13	96.45±0.12
PSPNet	94.91±0.13	97.39±0.15	96.87±0.20	94.78±0.05	93.51±0.20	94.93±0.09	94.69±0.26
DeepLabV3+	96.50±0.08	98.39±0.01	98.44±0.03	96.21±0.06	95.52±0.03	96.55±0.04	96.73±0.07
SegResNet	95.68±0.31	97.97±0.15	98.16±0.03	95.84±0.22	95.10±0.25	96.00±0.20	96.60±0.24
UNet++	97.06±0.08	98.46±0.06	98.55±0.13	96.75±0.04	95.91±0.13	97.11±0.09	97.28±0.05
SegFormer	96.63±0.06	98.34±0.04	98.37±0.04	96.35±0.08	95.78±0.09	96.56±0.08	96.84±0.08
TransUNet	97.41±0.10	98.71±0.07	98.88±0.02	97.18±0.03	96.51±0.06	97.30±0.06	97.73±0.02
UKAN	96.32±0.16	97.97±0.73	98.44±0.06	96.26±0.16	95.64±0.09	96.36±0.10	96.72±0.08
UMambaBot	97.32±0.04	98.55±0.05	98.75±0.02	97.04±0.04	96.26±0.07	97.23±0.09	97.59±0.03
UMambaEnc	97.33±0.06	98.56±0.12	98.81±0.03	97.00±0.05	96.38±0.10	97.40±0.08	97.59±0.10
SwinUMamba	**97.58±0.02**	**98.84±0.02**	**98.91±0.06**	**97.29±0.05**	**96.66±0.09**	**97.50±0.06**	**97.85±0.03**
Foundation Models							
SAM (box)	90.67±3.68	92.49±2.35	93.21±11.63	87.29±3.57	83.38±5.17	92.72±3.71	87.64±5.09
SAM (pt)	75.64±27.72	92.04±6.54	96.56±5.28	73.56±22.31	81.30±13.21	91.08±8.03	79.09±18.78
MedSAM (box)	82.54±6.60	90.63±2.88	94.91±4.64	82.43±5.70	80.06±5.91	88.18±4.11	82.05±6.47

Model	Tr	Tz	MC1	MC2	MC3	MC4	MC5
Supervised Models							
UNet	95.69±0.19	94.14±0.08	98.14±0.09	97.72±0.27	96.87±0.09	97.30±0.07	97.90±0.07
DeepLabV3	95.53±0.08	93.98±0.07	97.98±0.02	97.77±0.08	97.00±0.03	97.30±0.02	97.89±0.03
FPN	95.43±0.11	93.86±0.07	97.90±0.10	97.96±0.05	97.14±0.09	97.53±0.06	98.06±0.05
PSPNet	94.12±0.09	92.90±0.08	96.67±0.08	96.94±0.02	95.90±0.12	96.17±0.05	96.96±0.14
DeepLabV3+	95.67±0.06	94.04±0.07	98.21±0.02	98.01±0.04	97.18±0.02	97.62±0.03	98.18±0.02
SegResNet	95.36±0.30	94.07±0.13	97.93±0.25	97.22±1.10	96.86±0.13	97.48±0.17	98.05±0.09
UNet++	96.16±0.08	94.62±0.07	98.39±0.09	98.32±0.07	97.52±0.04	98.03±0.03	98.44±0.05
SegFormer	95.75±0.06	94.24±0.10	98.16±0.03	98.07±0.04	97.20±0.04	97.72±0.02	98.17±0.03
TransUNet	96.54±0.04	95.05±0.09	98.64±0.06	98.45±0.15	97.65±0.08	98.16±0.16	98.49±0.21
UKAN	95.70±0.05	94.20±0.07	98.16±0.07	97.96±0.08	97.16±0.06	97.69±0.05	98.19±0.05
UMambaBot	96.42±0.07	94.85±0.05	98.56±0.04	98.42±0.03	97.69±0.04	98.21±0.02	98.55±0.02
UMambaEnc	96.50±0.06	94.90±0.11	98.60±0.08	98.44±0.04	97.69±0.05	98.13±0.08	98.54±0.06
SwinUMamba	**96.67±0.02**	**95.14±0.04**	**98.74±0.04**	**98.57±0.01**	**97.84±0.02**	**98.31±0.01**	**98.64±0.01**
Foundation Models							
SAM (box)	83.89±7.05	85.81±6.99	97.13±1.06	91.08±18.93	88.96±14.95	78.39±31.39	89.68±20.65
SAM (pt)	66.05±22.68	59.37±19.22	96.90±1.74	83.90±20.71	86.35±17.22	76.50±22.18	67.02±25.35
MedSAM (box)	82.43±5.49	74.59±9.48	92.76±6.80	87.06±6.66	85.95±4.00	80.18±8.83	87.14±6.02

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

based models including TransUNet and SegFormer maintain stable performance but are slightly behind the Mamba-based architectures, while CNN-based methods such as UNet and DeepLabV3 show moderate accuracy with greater fluctuations across regions. In sharp contrast, foundation models lag considerably, with SAM (pt) and MedSAM recording DSC values 10 to 20 points lower than supervised models; for instance, SAM (pt) achieves only 75.64% on the Capitate and 76.50% on the Metacarpal 4th.

Turning to NSD (Table 10), which emphasizes surface-level boundary precision, the same trend persists but the performance gap becomes even more pronounced. SwinUMamba again dominates with top results across nearly all joints, showing notable advantages on small and challenging structures such as the Trapezium and Triquetrum where precise boundary delineation is critical. UMambaEnc and UMambaBot remain highly competitive, confirming their robustness across both large bones and finer anatomical details. By comparison, CNN-based models exhibit larger drops in NSD despite acceptable DSC values, indicating difficulties in capturing fine boundary details for complex joint shapes. Foundation models perform the worst under this metric, with SAM- and MedSAM-based approaches consistently trailing far behind, underscoring their limited ability to recover accurate anatomical boundaries in wrist bone segmentation. These findings indicate that overlap-based DSC alone may mask boundary errors, and NSD is necessary to reveal clinically critical differences in fine anatomical structures.

Table 10: NSD performance on all joints. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Cap	Radius	Ulna	Ham	Lu	Tri	Sca
Supervised Models							
UNet	75.30±1.04	82.16±0.43	94.05±0.44	78.34±0.66	74.95±0.64	82.69±0.37	83.13±1.17
DeepLabV3	74.77±0.94	82.18±0.78	93.29±0.15	76.32±0.66	74.40±0.56	80.29±0.52	80.74±0.53
FPN	73.08±0.88	81.14±0.47	93.94±0.39	74.62±1.01	73.40±0.94	80.32±1.44	78.43±1.33
PSPNet	59.43±1.20	70.27±1.72	83.14±2.08	62.68±0.50	61.55±0.94	68.70±1.09	63.22±1.85
DeepLabV3+	75.29±0.94	83.44±0.32	96.17±0.28	76.28±0.58	74.05±0.35	83.56±0.38	81.78±0.89
SegResNet	70.71±2.29	80.15±1.62	93.92±0.35	75.30±1.38	73.49±1.73	79.72±1.86	81.10±1.95
UNet++	82.04±0.82	85.22±1.26	96.05±0.41	81.97±0.35	78.28±0.86	88.78±0.36	87.00±0.60
SegFormer	76.97±0.83	82.99±0.71	95.77±0.27	78.06±1.01	76.49±0.81	84.17±0.55	82.60±0.96
TransUNet	84.82±0.94	89.16±0.70	98.16±0.15	85.98±0.43	82.21±0.69	90.12±0.33	90.96±0.33
UKAN	74.13±1.17	81.26±2.53	95.25±0.41	77.15±1.03	75.89±0.87	82.19±0.82	81.38±0.79
UMambaBot	84.19±0.41	87.19±0.94	97.59±0.12	84.93±0.24	81.00±0.48	90.22±0.73	89.99±0.14
UMambaEnc	84.17±0.68	87.62±1.47	97.80±0.20	84.55±0.68	81.52±0.70	91.02±0.45	90.06±0.55
SwinUMamba	**86.71±0.20**	**90.86±0.33**	**98.24±0.30**	**87.31±0.47**	**83.75±0.58**	**92.29±0.37**	**91.99±0.17**
Foundation Models							
SAM (box)	57.24±14.60	61.51±8.00	82.09±14.87	46.82±11.76	56.04±11.46	65.82±16.59	63.67±10.05
SAM (pt)	46.48±28.63	62.62±12.69	91.68±12.18	32.01±20.09	54.22±16.56	66.38±21.09	53.52±22.49
MedSAM (box)	20.81±13.22	45.31±13.06	75.01±19.14	18.82±10.95	38.20±17.00	37.38±16.40	26.78±12.72

Model	Tr	Tz	MC1	MC2	MC3	MC4	MC5
Supervised Models							
UNet	74.47±1.52	70.46±0.69	92.91±0.59	90.64±1.62	84.13±0.64	90.07±0.50	93.75±0.15
DeepLabV3	72.75±0.80	68.31±0.55	92.25±0.11	91.87±0.38	85.46±0.14	91.30±0.19	93.75±0.29
FPN	71.15±0.97	66.85±1.08	90.36±0.78	91.89±0.38	84.73±0.44	90.79±0.44	93.48±0.59
PSPNet	59.64±1.22	59.15±0.47	81.49±0.62	84.69±0.26	75.78±1.35	81.37±0.44	85.32±0.62
DeepLabV3+	73.86±0.61	68.78±0.70	93.53±0.27	92.48±0.23	85.45±0.19	91.69±0.14	95.20±0.18
SegResNet	71.99±2.26	70.53±1.33	91.49±1.42	88.49±3.81	83.46±0.67	90.48±1.14	94.20±0.58
UNet++	77.99±0.81	74.07±0.49	94.44±0.70	94.19±0.48	87.59±0.21	93.95±0.30	96.22±0.27
SegFormer	74.57±0.54	70.75±0.86	93.40±0.36	93.31±0.15	86.16±0.36	92.89±0.13	95.31±0.21
TransUNet	81.55±0.30	77.19±0.81	96.20±0.85	95.62±0.35	88.73±0.39	95.77±0.10	96.27±3.00
UKAN	73.97±0.65	69.99±0.92	92.62±0.35	91.38±0.59	85.07±0.41	91.83±0.36	94.59±0.44
UMambaBot	81.20±0.67	75.52±0.66	95.95±0.22	94.99±0.32	88.83±0.21	95.63±0.10	97.22±0.08
UMambaEnc	82.07±0.50	76.15±1.04	96.34±0.33	95.13±0.38	88.73±0.38	95.07±0.62	97.14±0.25
SwinUMamba	**83.21±0.13**	**77.95±0.61**	**97.34±0.28**	**96.15±0.15**	**89.92±0.16**	**96.32±0.06**	**97.91±0.06**
Foundation Models							
SAM (box)	47.55±12.22	43.56±16.30	88.81±5.26	78.79±15.78	65.68±11.48	65.60±20.89	78.44±20.84
SAM (pt)	34.34±20.19	23.77±17.87	88.52±6.65	66.97±23.86	60.10±17.24	54.50±23.73	45.50±27.15
MedSAM (box)	32.61±10.31	15.90±12.79	61.62±12.88	48.02±17.31	40.55±15.58	38.63±14.70	43.69±13.67

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

Table 11: DSC performance on representative wrist bones. (Mann-Whitney U test between BE & nonBE, *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$).

Model	Radius			Ulna			Lunate		
	BE	nonBE	P	BE	nonBE	P	BE	nonBE	P
Supervised Models									
UNet	97.96±0.06	98.25±0.05	**	98.31±0.03	98.16±0.08		94.76±0.08	95.49±0.23	***
DeepLabV3	98.13±0.07	98.40±0.04	***	97.92±0.02	98.00±0.02		94.61±0.15	95.81±0.05	***
FPN	98.04±0.06	98.28±0.02	***	98.24±0.06	98.21±0.06		94.77±0.14	95.69±0.13	***
PSPNet	97.28±0.13	97.44±0.17	*	96.93±0.23	96.85±0.20		93.30±0.17	93.59±0.25	**
DeepLabV3+	98.21±0.03	98.46±0.02	***	98.45±0.03	98.44±0.03		94.85±0.09	95.79±0.05	***
SegResNet	97.79±0.12	98.05±0.16	**	98.24±0.12	98.14±0.08		94.73±0.22	95.25±0.26	***
UNet++	98.39±0.05	98.48±0.07		98.68±0.10	98.49±0.15		95.18±0.11	96.20±0.14	***
SegFormer	98.17±0.05	98.40±0.04	***	98.35±0.04	98.37±0.05		95.07±0.12	96.06±0.08	***
TransUNet	98.63±0.06	98.75±0.07	***	98.89±0.02	98.87±0.02		95.82±0.09	96.78±0.05	***
UKAN	97.79±0.80	98.04±0.69	**	98.44±0.12	98.44±0.03		94.75±0.10	95.98±0.10	***
UMambaBot	98.54±0.07	98.55±0.06		98.52±0.05	98.84±0.02		95.69±0.04	96.48±0.09	***
UMambaEnc	98.54±0.09	98.57±0.13		98.66±0.08	98.86±0.03		95.81±0.12	96.60±0.10	***
SwinUMamba	98.80±0.03	98.86±0.02	*	98.97±0.02	98.88±0.08		95.95±0.10	96.94±0.09	***
Foundation Models									
SAM (box)	92.51±2.43	92.48±2.34		95.68±2.22	92.24±13.54		83.28±4.56	83.42±5.41	
SAM (pt)	92.25±5.13	91.95±7.04		97.55±2.37	96.18±6.02		83.16±5.49	80.58±15.15	
MedSAM (box)	90.32±3.48	90.75±2.62		95.09±3.38	94.84±5.06		78.68±6.17	80.59±5.75	

Model	Scaphoid			Trapezium			MC1		
	BE	nonBE	P	BE	nonBE	P	BE	nonBE	P
Supervised Models									
UNet	96.86±0.15	96.77±0.12		95.30±0.20	95.84±0.21	*	98.15±0.04	98.13±0.12	
DeepLabV3	96.56±0.10	96.73±0.06		94.88±0.13	95.78±0.08	***	97.91±0.03	98.01±0.02	*
FPN	96.22±0.14	96.54±0.12	*	94.85±0.12	95.66±0.10	***	97.92±0.10	97.89±0.11	
PSPNet	94.35±0.44	94.82±0.25	**	93.83±0.12	94.23±0.13		96.52±0.13	96.74±0.08	
DeepLabV3+	96.69±0.05	96.74±0.08		95.05±0.05	95.91±0.07	***	98.19±0.01	98.22±0.03	
SegResNet	96.54±0.31	96.62±0.21		94.91±0.32	95.54±0.30	**	97.98±0.20	97.91±0.27	*
UNet++	97.37±0.06	97.24±0.06	*	95.59±0.11	96.38±0.07	***	98.46±0.11	98.36±0.08	
SegFormer	96.66±0.15	96.91±0.06		95.19±0.07	95.96±0.07	**	98.18±0.03	98.16±0.04	
TransUNet	97.73±0.03	97.73±0.03		96.03±0.08	96.74±0.05	***	98.68±0.03	98.62±0.07	*
UKAN	96.59±0.12	96.77±0.07		95.15±0.08	95.91±0.05	**	98.23±0.05	98.13±0.09	*
UMambaBot	97.65±0.05	97.57±0.03	*	95.81±0.07	96.65±0.09	***	98.63±0.03	98.54±0.05	*
UMambaEnc	97.62±0.13	97.59±0.11		95.87±0.12	96.75±0.06	***	98.68±0.03	98.57±0.11	*
SwinUMamba	97.92±0.04	97.82±0.03		96.17±0.08	96.87±0.03	***	98.76±0.04	98.74±0.05	
Foundation Models									
SAM (box)	87.82±5.20	87.58±5.07		82.26±8.43	84.52±6.37		97.17±1.08	97.12±1.06	
SAM (pt)	76.62±22.58	80.05±17.13		65.13±22.88	66.40±22.71		96.56±2.10	97.03±1.58	
MedSAM (box)	81.55±6.35	82.25±6.55		81.69±6.59	82.71±5.01		93.37±5.55	92.52±7.24	

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

C.1.2 Impact of BE

Table 11 and Table 12 report segmentation outcomes for representative wrist bones with and without BE. Both DSC and NSD indicate that joints affected by BE are harder to segment. The gap is generally small for DSC, typically about 1%, but it is more evident for NSD, commonly 2% to 6%. Across both metrics, supervised methods outperform foundation models. Since BE alters bone morphology through erosion and deformation, analyzing BE versus nonBE groups allows us to assess whether such pathological changes affect segmentation accuracy.

For DSC (Table 11), supervised approaches achieve very high accuracy above 95% in both the BE and nonBE groups. SwinUMamba attains the best scores on nearly all bones, remaining between 98% and 99% in nonBE cases and only slightly lower in BE cases, for example 98.0% on the Distal Radius and 97.9% on the Ulna. UMambaEnc and UMambaBot also maintain stable performance above 97%. By contrast, CNN baselines such as UNet and DeepLabV3 yield slightly lower values, typically 96% to 97%, and foundation models perform worse overall with DSC around 90% to 95%, regardless of BE status.

Table 12: NSD performance on representative wrist bones. (Mann-Whitney U test between BE & nonBE, *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$).

Model	Radius			Ulna			Lunate		
	BE	nonBE	P	BE	nonBE	P	BE	nonBE	P
Supervised Models									
Unet	80.57±0.62	82.78±0.51	*	94.12±0.21	94.02±0.56		72.24±0.69	76.00±0.86	***
DeepLabV3	79.68±1.15	83.15±0.70	***	93.01±0.43	93.40±0.21		71.71±1.13	75.45±0.57	***
FPN	78.61±0.91	82.12±0.43	***	93.73±0.24	94.02±0.47		72.70±1.30	73.68±1.12	
PSPNet	69.94±1.87	70.40±1.80		83.78±2.46	82.90±1.97		61.34±1.35	61.63±1.10	
DeepLabV3+	81.31±0.45	84.27±0.37	***	95.96±0.33	96.25±0.27		72.56±0.67	74.63±0.59	*
SegResNet	78.70±1.60	80.71±1.67	*	94.07±1.06	93.86±0.20		71.80±1.58	74.15±1.80	*
Unet++	84.84±1.10	85.36±1.34		96.49±0.37	95.88±0.47		74.94±0.86	79.57±0.94	***
SegFormer	81.01±0.73	83.76±0.71	***	95.72±0.35	95.78±0.29		74.84±1.33	77.14±0.64	*
TransUNet	87.96±0.72	89.63±0.72	**	97.79±0.11	98.31±0.17		79.20±0.97	83.39±0.73	***
UKAN	79.61±2.39	81.90±2.60	**	94.99±0.55	95.35±0.35		73.07±1.11	76.98±0.82	***
UMambaBot	86.83±1.42	87.33±0.88		96.66±0.29	97.95±0.12	**	78.85±0.23	81.84±0.64	***
UMambaEnc	87.19±1.27	87.79±1.58		97.16±0.20	98.05±0.24	*	79.42±0.86	82.34±0.72	**
SwinUMamba	90.63±0.50	90.95±0.35		98.45±0.21	98.15±0.36		81.05±0.58	84.81±0.60	***
Foundation Models									
SAM (box)	63.34±7.41	60.80±8.14		85.47±7.28	80.78±16.78		55.49±10.55	56.26±11.84	
SAM (pt)	63.67±11.79	62.22±13.06		93.45±8.22	91.00±13.38		54.76±13.71	54.01±17.61	
MedSAM (box)	44.22±16.33	45.73±11.62		74.88±18.11	75.06±19.63		32.76±15.99	40.32±16.99	*

Model	Scaphoid			Trapezium			MC1		
	BE	nonBE	P	BE	nonBE	P	BE	nonBE	P
Supervised Models									
Unet	84.69±1.41	82.52±1.14	**	73.35±1.63	74.90±1.58		93.61±0.34	92.64±0.71	**
DeepLabV3	80.82±1.11	80.71±0.61		70.01±1.07	73.81±0.80	**	92.29±0.33	92.24±0.16	
FPN	77.30±1.43	78.88±1.33		69.27±1.39	71.88±0.81	*	91.02±0.70	90.10±0.81	*
PSPNet	61.72±2.67	63.81±1.73		60.33±1.06	59.37±1.37		81.94±1.14	81.31±0.44	
DeepLabV3+	81.97±0.51	81.71±1.06		71.63±0.46	74.73±0.75	*	93.89±0.31	93.39±0.31	
SegResNet	80.83±2.88	81.20±1.61		70.73±2.73	72.48±2.10		92.39±0.91	91.14±1.64	**
Unet++	87.84±0.61	86.68±0.67		75.68±0.99	78.89±0.85	*	95.06±0.77	94.19±0.72	
SegFormer	81.93±1.63	82.86±0.74		72.84±0.88	75.24±0.57		93.91±0.32	93.21±0.43	
TransUNet	90.86±0.27	91.01±0.45		79.43±0.43	82.38±0.37	*	96.53±0.52	96.08±0.99	
UKAN	81.20±0.80	81.45±0.90		72.05±0.67	74.71±0.70		93.46±0.44	92.29±0.42	*
UMambaBot	90.60±0.38	89.75±0.09		78.82±0.33	82.13±0.90	**	96.31±0.23	95.81±0.32	
UMambaEnc	90.56±0.97	89.86±0.70		79.46±0.70	83.09±0.55	**	96.83±0.13	96.14±0.44	
SwinUMamba	92.47±0.36	91.80±0.13		81.13±0.53	84.02±0.17	*	97.48±0.18	97.28±0.32	
Foundation Models									
SAM (box)	63.44±11.32	63.76±9.58		47.29±14.25	47.64±11.42		89.71±5.22	88.46±5.27	
SAM (pt)	51.73±25.27	54.21±21.42		33.50±21.40	34.66±19.81		88.06±7.41	88.70±6.37	
MedSAM (box)	26.40±14.98	26.92±11.81		35.81±10.66	31.37±9.96	*	63.48±13.71	60.90±12.55	

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

For NSD (Table 12), the difference between BE and nonBE cases is more pronounced, especially in the Lunate and Trapezium. SwinUMamba again delivers the highest accuracy, reaching about 90% to 97% in nonBE cases and only a few percentage points lower in BE cases. UMambaEnc and UMambaBot remain competitive, whereas CNN baselines generally fall to the low 80s in BE joints. Foundation models show the weakest boundary accuracy, often dropping below 70% in BE cases. For example, MedSAM records only 32.8% on the Lunate with BE. These findings indicate that BE degrades boundary precision more than volumetric overlap, and that supervised Mamba family models, particularly SwinUMamba, are the most resilient to these challenges. This suggests that accurate segmentation in BE-affected regions remains a critical challenge for clinical applicability, as these areas are most relevant for disease monitoring and treatment decisions.

Table 13: Instance segmentation results on overlapping regions. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	DSC ↑ (%)	NSD ↑ (%)	VOE ↓ (%)	MSD ↓ (pix)	MSD Fail Rate (%)
Supervised Models					
UNet	65.80±0.83	66.60±1.30	46.10±0.88	2.77±0.15	3.28±0.31
DeepLabV3	68.60±0.36	67.01±0.45	43.49±0.36	2.27±0.04	3.07±0.10
FPN	67.11±0.39	64.68±0.82	45.32±0.45	2.41±0.03	2.83±0.20
PSPNet	61.12±0.54	54.42±0.60	52.07±0.55	3.10±0.07	4.20±0.15
DeepLabV3+	68.02±0.29	66.39±0.38	44.21±0.35	2.31±0.02	2.69±0.05
SegResNet	57.30±5.65	58.63±7.12	53.38±4.36	3.22±0.74	8.51±10.72
UNet++	70.16±0.68	71.31±0.57	41.46±0.62	2.16±0.04	2.81±0.21
SegFormer	68.61±0.15	67.94±0.35	43.37±0.20	2.30±0.04	2.63±0.13
TransUNet	73.27±1.01	75.66±0.51	37.70±0.90	1.91±0.04	2.61±0.23
UKAN	62.68±1.86	63.85±1.85	48.73±1.72	2.67±0.44	3.72±0.51
UMambaBot	72.70±0.18	74.55±0.40	38.55±0.17	2.08±0.18	2.60±0.11
UMambaEnc	72.45±0.47	74.67±0.51	38.79±0.52	1.97±0.03	2.77±0.15
SwinUMamba	**74.45±0.25**	**77.15±0.20**	**36.25±0.27**	**1.83±0.02**	2.69±0.20
Foundation Models					
SAM (box)	3.78±2.83	2.51±1.88	97.09±2.34	9.49±7.79	89.96
SAM (pt)	3.41±1.90	2.58±1.72	97.91±1.29	49.63±23.80	61.13
MedSAM (box)	5.32±4.16	3.34±2.52	96.31±3.12	12.10±5.43	75.79

Foundation models: one inference (mean ± std across cases, MSD Fail Rate excluded).
Supervised models: five runs (mean ± std across runs).

C.1.3 Segmentation of Overlapping Regions

Although overall DSC and NSD values are high, visual inspection reveals that overlapping bones remain problematic, motivating a focused evaluation on these regions. Table 13, Table 14, and Table 15 demonstrate that overlapping wrist bones are particularly difficult to segment. SwinUMamba achieves the best performance across all metrics. In Table 13, it reaches 74.5% DSC and 77.2% NSD, while also obtaining the lowest VOE (36.3%) and the lowest MSD (1.83 pixels). The failure rate remains below 3%, indicating strong robustness. In contrast, foundation models almost completely fail in this scenario, with DSC values below 6% and NSD values below 4%. These results show that such models cannot effectively separate closely packed structures without task-specific training.

For DSC in Table 14, SwinUMamba ranks first in nearly all pairwise regions. For example, it achieves 88.1% on the Radius–Lunate interface and 89.7% on the Capitate–Scaphoid interface. These results are about 1% to 3% higher than those of UMambaEnc and UMambaBot. CNN-based models such as UNet and DeepLabV3 are usually more than 10% lower. TransUNet is the only Transformer baseline that approaches the Mamba-based models, with 86.3% on Distal Radius–Lunate and 85.0% on Trapezium–Meracarpal 2nd, but its performance is less consistent. SegResNet produces the lowest results, especially on Hamate-related overlaps, where its accuracy falls below 50%.

For NSD in Table 15, a similar trend is observed, with boundary effects more evident. SwinUMamba again achieves the best results, reaching 87.1% on the Hamate–MC5 interface and 86.6% on the Capitate–Scaphoid interface, while remaining above 79% even on difficult regions such as Trapezium–Trapezoid. UMambaEnc and UMambaBot follow closely, usually within 2% to 3% of SwinUMamba. TransUNet shows mixed performance; for instance, it improves to 69.9% on Capitate–Metacarpal 3rd but lags behind on other pairs due to underestimation of overlap. CNN baselines drop further, often to 60% to 70%, and foundation models show the weakest performance, with NSD consistently below 5% on almost all overlapping regions.

These results suggest that overlapping regions remain the most challenging aspect of wrist bone segmentation. Future work should focus on developing specialized strategies to improve performance in these areas, such as overlap-aware loss functions, boundary refinement modules, or targeted data augmentation. Enhancing segmentation accuracy in overlapping regions will be critical for achieving reliable and clinically applicable models. Improving overlap segmentation is especially important for clinical reliability, since diagnostic assessment often depends on accurate separation of adjacent bones in crowded anatomical areas.

Table 14: Overlap DSC performance on overlapping regions. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Cap-Sca	Cap-Tz	Cap-MC3	Radius-Lu	Radius-Sca	Ham-MC4	Ham-MC5
Supervised Models							
Unet	84.47±0.70	49.31±1.78	39.78±7.58	81.30±0.63	76.06±0.94	43.49±7.90	85.74±0.28
DeepLabV3	84.85±0.26	50.78±2.23	49.65±0.70	82.73±0.73	76.18±0.88	51.81±1.82	85.86±0.20
FPN	83.57±0.40	48.86±1.14	45.76±2.51	81.15±0.55	75.50±0.67	50.23±1.67	85.03±0.43
PSPNet	75.76±0.74	49.85±1.10	35.00±1.90	73.82±0.44	69.53±1.68	39.91±1.21	80.24±0.54
DeepLabV3+	84.04±0.47	50.46±0.59	49.34±1.83	82.75±0.12	76.73±0.15	51.92±1.86	85.48±0.18
SegResNet	83.38±1.53	31.14±14.43	19.11±12.34	78.97±1.51	74.26±1.49	22.44±14.80	85.21±1.15
Unet++	86.99±0.27	55.20±1.52	50.89±1.80	84.07±0.49	78.59±0.45	52.37±2.20	87.51±0.41
SegFormer	85.12±0.39	49.64±0.86	49.08±1.20	83.03±0.54	77.59±0.83	50.73±0.98	86.34±0.23
TransUNet	88.96±0.19	59.28±1.15	51.39±10.98	88.04±0.26	83.01±0.44	57.01±1.23	88.92±0.19
UKAN	84.74±0.54	38.25±13.95	28.58±10.77	66.13±36.37	74.94±2.83	42.81±10.79	86.04±0.35
UMambaBot	88.58±0.14	59.13±0.97	56.24±0.91	85.74±0.64	81.01±0.60	56.31±1.06	88.48±0.16
UMambaEnc	88.71±0.54	58.56±0.82	54.49±2.11	86.25±0.34	80.90±0.90	53.31±3.81	88.26±0.19
SwinUMamba	**89.71±0.13**	**60.01±0.52**	**57.51±0.64**	**88.09±0.13**	**83.40±0.27**	**59.59±0.78**	**89.48±0.19**
Foundation Models							
SAM (box)	1.00±5.86	0.03±0.34	0.00±0.00	0.54±5.94	0.40±4.38	0.00±0.00	0.14±0.65
SAM (pt)	2.09±7.30	0.69±4.10	0.12±0.54	0.84±6.05	0.66±5.06	0.31±1.13	1.21±3.53
MedSAM (box)	8.88±21.35	0.43±4.64	0.45±3.44	2.14±10.70	4.94±14.02	0.11±0.83	5.48±13.45

Model	Lu-Sca	Sca-Tr	Tr-Tz	Tz-MC1	Tr-MC2	Tz-MC2	MC2-MC3
Supervised Models							
Unet	74.14±0.39	66.03±0.96	88.07±0.31	68.08±0.88	73.79±10.99	26.74±6.07	64.24±1.09
DeepLabV3	73.72±0.50	66.20±0.72	88.08±0.20	69.26±0.28	80.35±0.42	33.18±0.83	67.74±0.49
FPN	72.27±0.65	66.26±0.96	87.49±0.21	65.91±1.20	78.46±0.65	33.33±1.96	65.71±0.65
PSPNet	63.95±1.10	61.35±0.39	85.16±0.30	58.42±1.39	72.12±0.61	31.25±1.58	59.39±1.66
DeepLabV3+	72.65±0.43	66.50±0.43	88.01±0.14	66.68±0.92	79.07±0.59	32.02±1.14	66.62±0.77
SegResNet	72.70±1.67	65.25±1.25	88.00±0.22	53.52±18.05	58.48±33.39	12.36±8.09	57.38±4.50
Unet++	75.60±0.72	69.77±1.13	88.99±0.09	70.08±1.25	81.46±0.70	32.46±2.67	68.26±1.16
SegFormer	73.72±0.79	67.87±0.73	88.39±0.17	70.48±0.64	80.39±0.32	31.28±0.74	66.85±0.34
TransUNet	77.80±0.26	72.16±0.50	89.95±0.20	74.58±0.58	83.90±0.16	**39.27±3.11**	71.50±0.59
UKAN	73.87±0.83	65.82±1.42	88.39±0.17	63.17±3.31	79.20±1.10	25.24±3.39	60.31±7.51
UMambaBot	77.67±0.33	71.11±0.65	89.44±0.15	74.14±0.85	83.14±0.32	36.43±0.77	70.32±0.55
UMambaEnc	77.75±0.35	**72.17±0.51**	89.79±0.19	73.97±0.44	82.90±0.43	36.79±1.79	70.38±0.77
SwinUMamba	**79.70±0.60**	71.94±0.41	**90.21±0.13**	**77.24±0.95**	**84.53±0.06**	38.71±1.21	**72.18±0.68**
Foundation Models							
SAM (box)	0.52±4.92	0.00±0.00	46.72±32.90	0.00±0.00	1.02±6.01	0.29±2.25	0.00±0.00
SAM (pt)	1.14±3.49	1.24±4.44	31.73±17.32	0.05±0.31	3.85±5.32	1.93±6.69	0.10±0.53
MedSAM (box)	1.77±6.18	0.33±1.76	38.47±28.17	2.50±8.19	3.97±11.70	0.80±3.33	0.42±2.58

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

Table 15: Overlap NSD performance on overlapping regions. The best results in each column are highlighted in **bold**, and the second-best values are underlined.

Model	Cap-Sca	Cap-Tz	Cap-MC3	Radius-Lu	Radius-Sca	Ham-MC4	Ham-MC5
Supervised Models							
UNet	74.70±1.62	68.99±1.00	64.07±4.10	72.51±0.89	73.19±0.84	53.46±4.02	78.07±1.33
DeepLabV3	72.88±0.76	67.01±1.18	67.08±0.65	71.83±1.73	70.15±1.77	57.32±0.46	77.14±0.32
FPN	68.41±1.18	67.91±0.74	65.33±1.09	69.49±1.27	68.44±1.54	56.62±1.90	74.32±1.35
PSPNet	51.44±1.21	62.87±0.98	55.30±0.83	56.47±0.98	57.62±2.00	44.42±0.59	61.76±1.50
DeepLabV3+	71.16±1.40	67.90±1.31	67.72±0.75	71.37±0.62	70.40±0.47	58.40±0.58	75.64±0.59
SegResNet	71.48±3.92	58.18±12.06	41.02±24.53	69.84±1.65	70.77±1.59	36.11±21.44	77.50±1.42
UNet++	80.33±0.76	71.79±0.98	69.30±0.51	77.23±1.18	75.37±0.84	60.32±1.06	82.17±0.89
SegFormer	73.21±1.41	69.25±0.75	67.99±0.62	71.79±0.81	71.10±1.89	59.48±0.92	79.04±0.71
TransUNet	84.69±0.57	74.89±1.31	69.91±4.83	84.13±0.16	82.37±0.43	66.49±1.64	85.67±0.60
UKAN	73.75±0.77	60.99±13.29	56.68±7.33	59.72±26.29	69.73±2.49	52.76±4.84	78.49±1.11
UMambaBot	83.93±0.41	74.89±0.31	72.12±0.42	80.31±1.66	79.30±1.69	64.90±1.73	84.62±0.42
UMambaEnc	84.32±1.41	74.88±1.06	71.52±0.59	81.15±1.32	79.16±1.40	62.80±3.55	84.25±0.35
SwinUMamba	**86.58±0.46**	**75.22±0.37**	**73.42±0.43**	**85.37±0.39**	**83.53±0.52**	**67.44±0.51**	**87.13±0.49**
Foundation Models							
SAM (box)	0.98±5.41	0.16±1.73	0.00±0.00	0.26±2.84	0.47±4.71	0.00±0.00	1.31±5.52
SAM (pt)	1.32±4.61	1.34±4.08	0.50±1.97	0.80±4.61	1.22±5.00	0.76±2.40	2.97±5.18
MedSAM (box)	3.64±8.67	0.17±1.89	0.73±5.19	2.14±8.44	4.62±10.37	0.15±1.52	6.18±12.19

Model	Lu-Sca	Sca-Tr	Tr-Tz	Tz-MC1	Tr-MC2	Tz-MC2	MC2-MC3
Supervised Models							
UNet	68.15±0.60	68.29±0.93	63.58±1.52	72.96±1.32	57.88±10.60	50.17±5.65	66.45±1.39
DeepLabV3	66.52±0.91	67.51±1.73	62.41±1.17	72.29±0.84	63.09±0.89	52.59±1.35	67.01±0.45
FPN	62.54±1.43	67.31±0.62	59.06±1.30	67.25±1.30	58.59±1.86	52.31±1.63	64.68±0.82
PSPNet	48.54±1.87	57.64±1.25	50.13±1.02	58.50±1.24	48.71±0.77	48.97±0.85	54.42±0.60
DeepLabV3+	64.20±0.79	68.70±1.26	62.06±0.44	69.61±1.23	60.82±1.13	52.35±0.62	66.39±0.38
SegResNet	65.16±2.72	67.23±1.39	63.21±1.64	59.98±14.16	45.19±26.18	31.04±18.28	58.63±7.12
UNet++	71.29±1.61	72.70±1.66	67.06±0.48	75.67±1.43	67.84±1.59	55.13±1.63	71.30±0.57
SegFormer	66.06±1.79	70.09±1.39	64.00±0.80	74.29±0.97	63.59±1.00	52.11±0.89	67.94±0.34
TransUNet	76.90±0.89	75.67±0.68	71.88±0.80	81.25±0.57	73.38±0.35	**58.16±2.20**	73.89±0.73
UKAN	68.05±1.24	68.51±0.89	63.42±0.81	66.16±3.47	62.21±2.08	48.77±3.72	63.86±1.86
UMambaBot	74.86±0.71	75.95±0.72	70.21±0.56	80.82±1.34	72.76±0.73	55.81±0.85	74.55±0.39
UMambaEnc	75.08±1.18	77.10±0.39	71.72±0.86	81.01±0.68	72.31±1.40	56.19±0.98	74.67±0.51
SwinUMamba	**78.88±0.86**	**77.27±0.61**	**73.40±0.64**	**84.71±0.98**	**75.61±0.41**	56.37±0.83	**75.24±0.64**
Foundation Models							
SAM (box)	0.33±2.64	0.00±0.00	27.66±19.73	0.00±0.00	1.79±7.52	0.86±4.99	0.00±0.00
SAM (pt)	1.67±4.55	1.85±5.79	12.41±12.48	0.45±1.77	5.24±8.63	4.72±9.65	0.36±1.65
MedSAM (box)	0.73±3.84	0.60±2.87	15.03±14.29	5.15±10.61	3.97±11.70	0.80±3.33	0.70±3.29

Foundation models: one inference (mean ± std across cases).
Supervised models: five runs (mean ± std across runs).

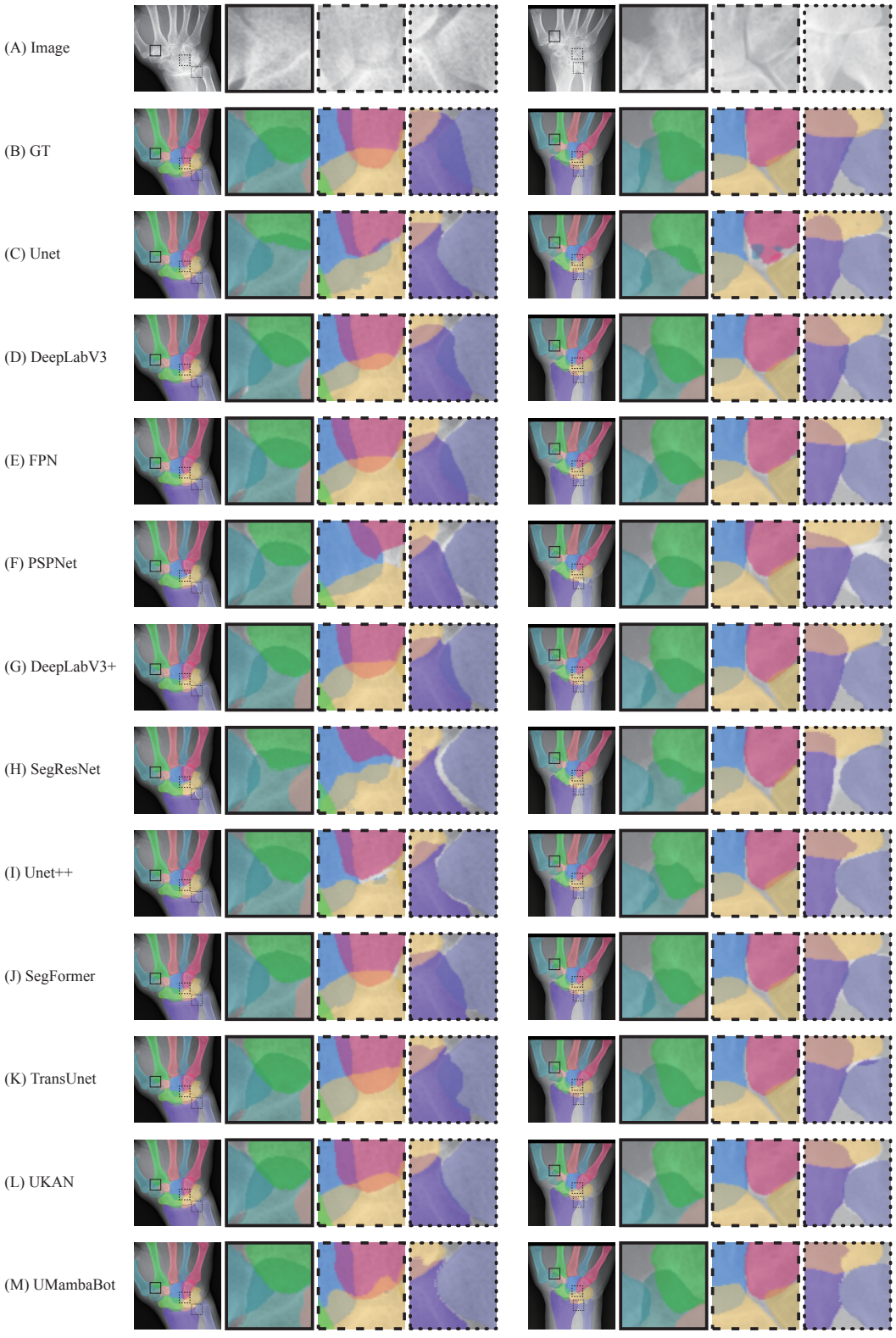


Figure 9: Additional visual results for wrist bone segmentation (Part A).

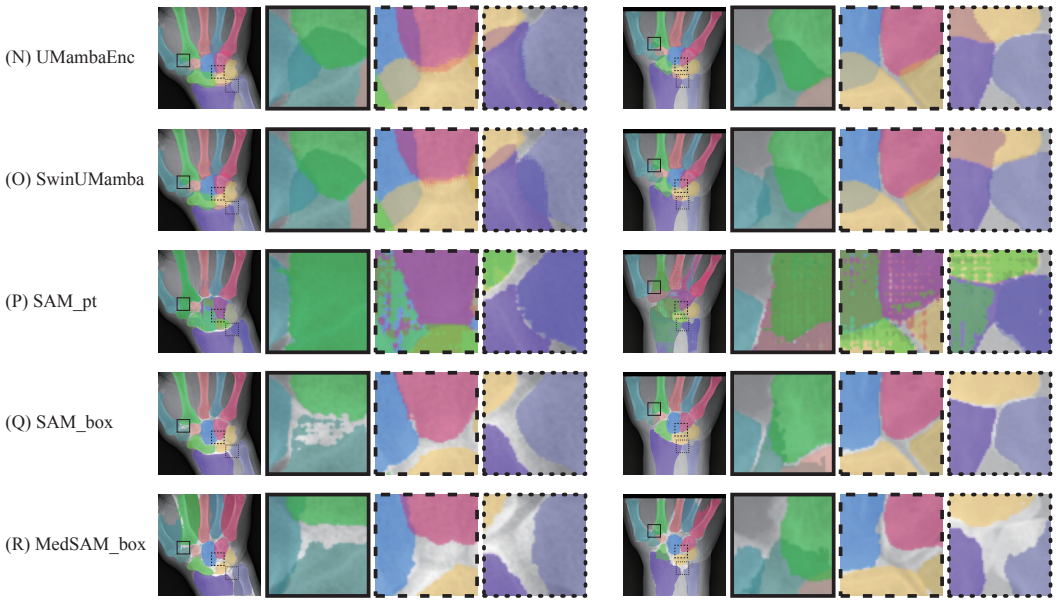


Figure 10: Additional visual results for wrist bone segmentation (Part B).

C.1.4 Qualitative Visualization

The visualization in Fig. 9 and Fig. 10 provides a detailed comparison of wrist bone segmentation results across different models. Visual inspection is crucial because numerical metrics alone may overlook local errors that are clinically significant, particularly around overlaps and pathological regions. All methods show noticeable errors compared to the ground truth, especially around boundaries where bones overlap or the image contrast is low. These problems are most pronounced in the Trapezium, Trapezoid, and Scaphoid, where irregular bone shapes and partial occlusions cause broken or incomplete predictions. In these regions, CNN-based models such as UNet, DeepLabV3, and FPN often produce blurred edges and fail to separate adjacent bones. Transformer-based models like TransUNet and SegFormer generate smoother boundaries, but they still lose fine details in overlapping areas. Mamba-based models, including UMambaEnc, UMambaBot, and SwinUMamba, show more stable performance. Their predictions follow bone contours more closely and reduce over-segmentation in crowded regions. SwinUMamba in particular achieves consistent delineation across both central and peripheral bones, with fewer gaps along thin boundaries. Nevertheless, even these models struggle in cases with severe overlap, where errors such as bone merging or missing edges remain visible. Foundation models perform poorly in visual comparison. SAM and MedSAM often produce coarse or fragmented masks that do not align with bone structures, highlighting their limitations when applied directly without fine-tuning. These models frequently miss small bones or collapse large bones into a single region, showing that task-specific supervision is essential for accurate wrist bone segmentation. In addition, BE-affected regions reveal another challenge. Across all models, bone erosion leads to irregular segmentation, with inward collapse or shape distortion often under-segmented. While some models occasionally capture these abnormalities, no architecture provides consistent results in such cases. This emphasizes the need for future work on methods that can better handle overlapping boundaries, subtle bone structures, and pathological deformations in order to achieve robust clinical applicability. Such qualitative evaluation further highlights that achieving clinically trustworthy segmentation requires not only high numerical scores but also consistent performance on challenging anatomical and pathological cases.

C.1.5 Summary and Discussion

Our quantitative and qualitative analyses lead to the following view. Current foundation models are not yet able to capture fine anatomical boundaries or resolve closely apposed bones in wrist radiographs, therefore supervised baselines remain necessary as clinically meaningful references that exploit pixel-level annotations. Although overall DSC and NSD are high, these global metrics can

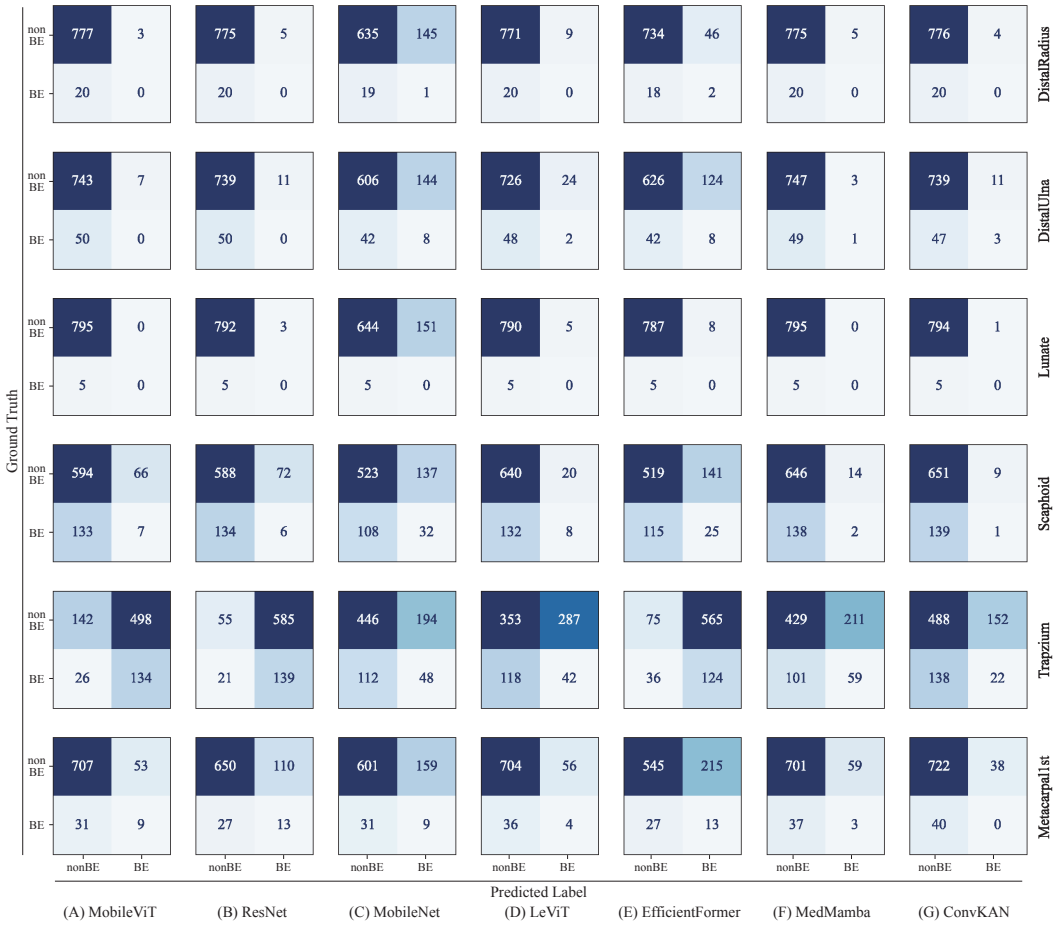


Figure 11: Confusion matrix results for classification of BE and nonBE.

conceal systematic errors that concentrate at bone interfaces and in low-contrast zones; the overlap-focused evaluation exposes these failure modes and aligns with visual inspection. Bone erosion alters local geometry and degrades boundary fidelity more than volumetric overlap, hence BE-stratified reporting is essential for clinical relevance. Among supervised methods, Mamba-based architectures strike a favorable balance between global context and local detail, whereas Transformers may sacrifice small-structure precision and CNNs struggle at complex interfaces; this mechanism-level difference explains the observed ranking across DSC and NSD. Moving toward clinical reliability will likely require a combination of architectural and procedural advances, including interface-aware or contour-consistency losses, boundary refinement and instance disambiguation modules, targeted augmentation that simulates occlusion and erosion, sampling curricula that oversample rare overlap patterns and BE cases, active learning to prioritize uncertain regions for annotation, and uncertainty estimation or test-time adaptation to mitigate distribution shift. Evaluation practice should likewise move beyond single numbers by reporting per-interface metrics, BE-stratified results, failure rates, and distance-based errors alongside DSC and NSD. Together, these directions align the benchmark with clinical priorities and outline a path toward robust and deployable wrist bone segmentation.

C.2 Classification of BE

Figure 11 shows confusion matrices for BE and nonBE classification across six representative carpal joints. Most models achieve high accuracy on nonBE cases, where predictions are strongly concentrated along the diagonal. However, substantial misclassification occurs in BE cases, reflecting the inherent difficulty of detecting pathological changes. This contrast shows that overall accuracy can be misleading, as it is dominated by the abundance of nonBE cases while failing to reflect systematic

errors in BE detection. For example, MobileNet and ResNet frequently misclassify BE samples as nonBE in the Distal Ulna and Trapezium, indicating a bias toward conservative predictions. Similar trends are observed in the Scaphoid, where BE cases are often confused with nonBE due to irregular joint boundaries. This joint-specific variability indicates that anatomical complexity directly affects classification difficulty, and a single model may not perform equally well across all regions.

More advanced architectures show partial improvements. MedMamba and ConvKAN achieve more balanced predictions, particularly in the Lunate and Distal Radius, where BE cases are identified with higher sensitivity compared to earlier CNN-based models. Nevertheless, even these models still exhibit notable false negatives, especially in challenging regions such as the Scaphoid and Trapezium. This suggests that while recent methods better capture morphological changes, robust recognition of erosive patterns remains unresolved. Clinically, missing even a small number of BE cases may delay diagnosis or underestimate disease severity, underscoring the need for higher sensitivity in BE detection. These results highlight the importance of designing models capable of learning discriminative features that generalize well to pathological variations, especially in early-stage RA where accurate BE detection is clinically critical.

Future work should address the extreme class imbalance between BE and nonBE cases through techniques such as focal loss or targeted data augmentation, and explore generative approaches for synthesizing BE-like patterns, in order to enhance the ability of models to capture subtle pathological features.

D Broader Impact

This work provides a publicly available and well-annotated multi-task wrist dataset and benchmark designed to advance research in RA diagnosis using conventional wrist radiographs. This resource enables researchers to build and evaluate advanced models for RA-related tasks with consistency and rigor. The authors do not anticipate any negative societal impacts stemming from this work. On the contrary, a positive impact may arise through the development of robust computer-aided diagnosis systems, which can facilitate early detection and monitoring of RA with reduced reliance on manual annotations. This has the potential to enhance clinical efficiency, reduce expert workload, and improve access to specialized care, particularly in under-resourced healthcare settings.