



HAL
open science

A Deep Learning-based Algorithm for Automatic Detection of Perilunate Dislocations in Frontal Wrist Radiographs

Negin Majzoubi, Rémi Allègre, Cédric Wemmert, Philippe Liverneaux

► **To cite this version:**

Negin Majzoubi, Rémi Allègre, Cédric Wemmert, Philippe Liverneaux. A Deep Learning-based Algorithm for Automatic Detection of Perilunate Dislocations in Frontal Wrist Radiographs. *Hand Surgery and Rehabilitation*, 2024, 43 (4), pp.101742. 10.1016/j.hansur.2024.101742 . hal-04602786

HAL Id: hal-04602786

<https://hal.science/hal-04602786v1>

Submitted on 5 Jun 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Deep Learning-based Algorithm for Automatic Detection of Perilunate Dislocations in Frontal Wrist Radiographs

Negin Majzoubi^a, Rémi Allègre^a,
Cédric Wemmert^a, Philippe Liverneaux^{a,b}

^a*ICube, Université de Strasbourg, CNRS, France*

^b*Department of Hand Surgery, Strasbourg University Hospitals, France*

Abstract

Perilunate dislocations are a rare but serious pathology, often undetected in the emergency setting. In this study, a Deep Learning algorithm is proposed to automatically detect perilunate dislocations in frontal radiographs. A total of 374 annotated frontal wrist radiographs, comprising 345 normal and 29 pathological ones from adolescents and adults aged 16 and above with skeletal maturity, were utilized to train, validate, and test two YOLOv8 deep neural models. The first model is responsible for detecting the carpal region, and the second for segmenting a region between Gilula's 2nd and 3rd arcs. The output of the segmentation model, trained multiple times with varying random augmentations, is then given a probability to be normal or pathological through ensemble averaging. On the considered dataset, the proposed algorithm achieves an overall F1-score of 0.880. The F1-score reaches 0.928 on the normal subgroup with a precision of 1.0, and 0.833 on the pathological subgroup with a recall (or sensitivity) of 1.0, demonstrating that the diagnosis of perilunate dislocations can be improved through automatic analysis of frontal radiographs.

Keywords: Medical imaging, Computer-aided diagnosis, Deep Learning, Perilunate dislocations

PACS: 87.57.R-, 87.57.N-

2000 MSC: 92C55, 68U10

1. Introduction

Perilunate Dislocations (PLDs) are lesions that alter the anatomical relationships of the two rows of the carpus as a result of trauma [1]. They may be isolated or associated with fractures [2]. Diagnosis is made on standard

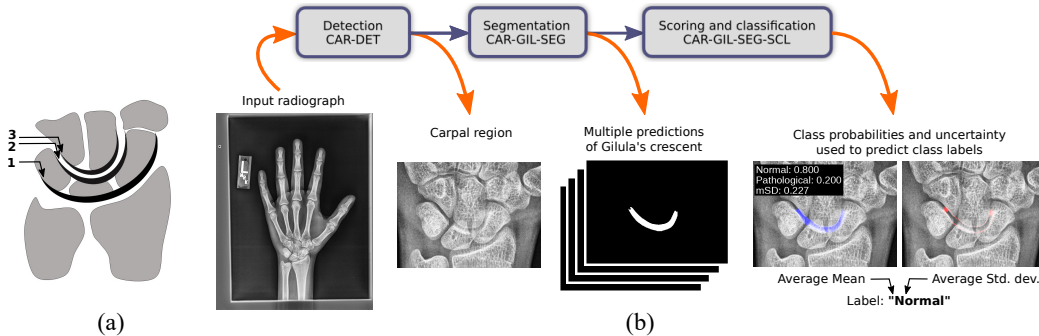


Figure 1: (a): Illustration of Gilula’s arcs, used to assess the alignment of the carpal bones. In this work, we focus on the crescent bounded by the 2nd and 3rd arcs. (b): Processing pipeline of our method. From an input frontal wrist radiograph, our algorithm outputs a distribution over the two classes "Normal" and "Pathological", together with an uncertainty value.

radiographs of the wrist from front and side [3]. PLDs are rare (less than 10% of wrist injuries) but serious if not treated promptly, because of the risk of disabling functional sequelae [4]. The diagnosis goes undetected in emergency settings in 25% of all cases, especially in isolated forms, as reading frontal radiographs requires a certain amount of experience [5, 6]. Lateral radiographs are often misinterpreted [7, 8, 9].

Deep Learning (DL) has transformed image analysis over the last decade, especially in medical imaging [10], and finds multiple applications in diagnostic radiology [11, 12, 13, 14, 15]. Pridgen et al. [16] have introduced a Deep Learning-based algorithm for automatic detection of PLDs on lateral wrist radiographs, that proceeds to the detection of the lunate and classifies the radiographs as normal or pathological. To our knowledge, this is the first attempt of a DL algorithm to deal with this task. A challenge is the small size of the available datasets, which is frequent in clinical settings and impacts the design of machine learning algorithms [17].

In this paper, we propose a DL-based algorithm for the automatic detection of isolated PLDs in *frontal* radiographs. Inspired by Gilula’s carpal arcs [18], we have developed a three-stage algorithm named **ADELUC** (Algorithm for DEtection of LUxation of the Carpus). This algorithm 1) identifies the carpal region, 2) attempts to segment a region enclosed by Gilula’s 2nd and 3rd arcs, termed *Gilula’s crescent*, and 3) gives a probability to the segmented region to be "Normal" or "Pathological" together with an uncertainty value, as well as a label predicted from these values (Fig.1). Our approach incorporates two deep neural models, trained, validated, and tested using 374 annotated frontal wrist radiographs, which include 345 normal and 29 pathological images. Despite the data imbalance, the proposed algorithm

achieves an overall F1-score of 0.880 on this dataset. The F1-score reaches 0.928 on the normal subgroup with a precision of 1.0, and 0.833 on the pathological subgroup with a recall (or sensitivity) of 1.0, meaning *absence of false negatives* on this subgroup, supporting the conclusion that our method can enhance the diagnostic accuracy of PLDs from frontal radiographs. The subsequent sections of this paper elaborate on our methodology and discuss the results obtained.

2. Material and methods

2.1. Gilula’s arcs and crescent

The conventional method for diagnosing PLDs from frontal radiographs involves the utilization of Gilula’s arcs [18], as depicted in Fig. 1(a). Intact and smoothly continuous arcs suggest no luxation, while any irregularities or interruptions in these arcs indicate potential PLD. In line with this diagnostic method, our goal was to develop a DL-based algorithm capable of automating both the drawing and interpretation of Gilula’s arcs. However, it is pertinent to note that most deep neural models are tailored predominantly for vision tasks, focusing primarily on *bounded regions* within images and analyzing pixel intensities. To address this, we have introduced the *Gilula’s crescent* concept, which is defined as the area enclosed by the 2nd and 3rd arcs, demarcating the two rows of carpal bones. In normal cases, Gilula’s crescent appears continuous and smooth, whereas in the presence of PLDs, it may show irregularities or interruptions.

The regular appearance of Gilula’s crescent in the most common class of cases, i.e., normal cases, is leveraged to counter the challenge of data imbalance. Our hypothesis is twofold: firstly, a model trained to segment Gilula’s crescent would reliably produce a complete and continuous crescent in normal cases, and in contrast, reveal an absent, partial, or disconnected crescent in the less frequent pathological cases. Secondly, we have anticipated that this distinct visual differentiation would facilitate the effective training of a classification model.

2.2. Dataset and annotations

The dataset utilized in this study encompasses 374 frontal wrist radiographs. It includes 54 radiographs from adult patients at Strasbourg University Hospital, comprising 25 normal and 29 pathological cases. Additionally, the dataset features 320 normal radiographs of adolescents aged 16 to 19, sourced from the public 2017 RSNA Pediatric Bone Age Challenge [19]. The skeletal maturity of the participants has been checked for each image. Consequently, our dataset consists of 345 normal and 29 pathological radiographs.

These high-resolution images exhibit diverse dimensions, with pixel counts ranging from approximately 1M to 5M. A notable aspect of this dataset is its relatively small size and significant imbalance between the two classes.

For annotation purposes, each radiograph in the dataset has undergone manual labeling, which involves three key processes:

1. Classification of the radiograph as either "Normal" or "Pathological".
2. Delineation of the carpal region using an axis-aligned bounding box.
3. In cases classified as "Normal", Gilula's crescent have been segmented manually by drawing a polygon. Conversely, for "Pathological" cases, this segmentation step has been omitted.

These annotations has been facilitated by various image labeling tools, with our choice being the open-source software Label Studio [20].

In dealing with pathological radiographs, the decision to leave the segmentation of Gilula's crescent empty is deliberate. This approach diverges from the traditional method of learning distinct "Normal" and "Pathological" classes. Instead, our method concentrates on learning to detect and accurately segment normal Gilula's crescents, which are more prevalent in our dataset. For pathological images in the test set, we anticipate either an absence of detection or the emergence of noisy segmentation results. In contrast, for normal images, we expect to observe complete and accurately segmented Gilula's crescents.

2.3. Deep neural architecture

In this study, we have implemented YOLO (You Only Look Once) [21, 22, 23], a widely recognized deep neural network-based algorithm initially conceived for object detection in images. YOLO operates by solving a regression problem across a grid of patches at multiple scales. Over time, its capabilities have expanded to include classification and semantic segmentation tasks [24, 25]. The latest iteration used in our study is YOLOv8.

2.4. Algorithm stages

Our algorithm processes each input radiograph in three distinct stages: 1) Detection of the carpal region, 2) Segmentation of Gilula's crescent using multiple segmentation models, and 3) Scoring and classification of the segmented result w.r.t. "Normal" and "Pathological" classes. The detection and segmentation stages utilize YOLOv8 models initially trained on the COCO dataset, a standard in image detection and segmentation (Sections 2.4.1 and 2.4.2). These models are subsequently fine-tuned with images from our dataset, which undergo preprocessing and augmentation offline (Section 2.4.3). To assess the influence of segmented Gilula's crescents on

diagnostic accuracy, we have trained classification models using images from the carpal region with only the class labels as annotations (Section 2.4.4). Detailed information about the data splits, training parameters, and online augmentations can be found in Table 1 and Section Appendix A.

2.4.1. Preprocessing and offline augmentations

To minimize biases, all radiographs have undergone preprocessing. Right-hand images have been mirrored to resemble left-hand radiographs. Contrast enhancement was applied to each image using Contrast Limited Adaptive Histogram Equalization (CLAHE) [26]. Furthermore, to address the class imbalance in our dataset, offline augmentation have been performed on the pathological radiographs within the training set. For each pathological radiograph, 16 variants have been created by applying moderate random elastic deformations (with σ approximately 5% of the shortest image dimension and α set to this dimension) and random rotations within a $\pm 5^\circ$ range.

2.4.2. Carpal region detection

The carpal region detection model, designated as **CAR-DET**, is an adaptation of the pre-trained YOLOv8x detection model, which boasts 68.2 million parameters. For training CAR-DET, we have selected 500 images from our dataset. This subset comprises 245 normal radiographs and 15 pathological radiographs. Each pathological radiograph has been duplicated 17 times, incorporating the previously mentioned variations, resulting in a total of 255 pathological images.

To bolster the model’s robustness, we employ a range of online augmentations during training. These include scaling, rotation, and modifications to image intensity and contrast. The parameters for these augmentations have been randomly selected from uniform distributions and applied to all training images in each epoch.

The loss function of the YOLOv8x model, utilized in CAR-DET, is a composite of several loss components. This includes a bounding box loss and a class loss. The model outputs an axis-aligned bounding box that encircles the detected carpal region, along with a corresponding confidence value. We have established a confidence threshold at 0.5; thus, any detection with a confidence level below this threshold is disregarded.

2.4.3. Segmentation of the Gilula’s crescent

In the second stage, multiple YOLOv8x-seg models (71.8M parameters), named **CAR-GIL-SEG**, have been trained using 260 images of the carpal region detected by CAR-DET. This includes 245 normal radiographs annotated with Gilula’s crescents, and 15 pathological radiographs with empty

annotations, augmented with their offline-produced variations, for a total of 255 pathological radiographs.

Relying on multiple deep neural models, instead of a single one, has proven to result in high-confidence predictions by providing reliable uncertainty estimation, e.g. through prediction variance [27, 28, 29]. In our case, we have trained 40 segmentation models, starting from the pre-trained YOLOv8x-seg model. Each training consists in fine-tuning the weights of this model using the complete training dataset. The order in which the training images are enrolled in the batches, as well as the applied random online augmentations, are made different for each training by selecting different seed values. The same online augmentations as for CAR-DET have been applied.

The loss function of the YOLOv8x-seg model is a weighted sum of multiple losses, including a bounding box loss, a class loss, and a segmentation loss. The segmentation result is a binary mask with a confidence value based on the detection of the segmented object. We set the confidence threshold value to 0.5, which means that segmentation results with low confidence are eliminated.

2.4.4. Segmentation scoring and classification

For a given test radiograph, we assign a score, namely a probability value, to each of the two classes "Normal" and "Pathological", which is achieved by averaging the 40 predicted segmentations on a per-pixel basis. We also compute an uncertainty value as the standard deviation of these predictions. As a final step, we assign a label based on the class probabilities and the uncertainty value. We will refer to this method built on top of the results of CAR-GIL-SEG as **CAR-GIL-SCL**.

Given an input image, let M_i denote the i -th binary segmentation mask, for $i \in [0, K]$, where $K = 40$ in our experiments. Let U denote the union of these binary masks, i.e. $U = \cup_{i=1}^K M_i$. For each pixel location \mathbf{p} in U , we compute $\mu(\mathbf{p}) = \text{E}[M_i(\mathbf{p})]$, with $M_i(\mathbf{p}) \in \{0, 1\}$. The probability value assigned to the "Normal" class is $p_{\text{SEG}}(\text{"Normal"}) = \text{E}[\mu(\mathbf{p})]$. For the "Pathological" class, we set $p_{\text{SEG}}(\text{"Pathological"}) = 1 - p_{\text{SEG}}(\text{"Normal"})$.

The uncertainty value is computed as the mean of per-pixel standard deviations. We first compute $\sigma(\mathbf{p}) = \sqrt{\text{E}[(M_i(\mathbf{p}) - \text{E}[M_i(\mathbf{p})])^2]}$ for each pixel location \mathbf{p} in U . The value of $\sigma(\mathbf{p})$ ranges between 0.0 and 0.5. Then, the uncertainty value is given by $\text{mSD} = 2 \times \text{E}[\sigma(\mathbf{p})]$, where mSD stands for "mean Standard Deviation".

We finally predict a class label upon the following condition:
If $\max(\text{mSD}, p_{\text{SEG}}(\text{"Pathological"})) > 0.5$ **then** set label "Pathological";
Otherwise, set label "Normal".

This condition means that the outcome label will be "Pathological" in case

of an mSD value greater than 0.5 or in case of a probability greater than 0.5 for the "Pathological" class. The label will be "Normal" in other cases.

2.4.5. Classification models for comparison

For the sake of comparison, we have trained classification models based on pre-trained YOLOv8x-cls models (57.4M parameters), termed **CAR-CL**. Training has been performed on the images of the carpal region detected by CAR-DET, keeping the same dataset splits as for the previous models. To be consistent with the segmentation approach, we have trained 40 classification models using different seed values to get predictions of higher confidence than with a single model, together with an uncertainty value, on which we rely to assign labels. This method will be denoted by **CAR-CL-SCL**. For a given image, we compute probabilities for "Normal" and "Pathological" that correspond to the average numbers of predicted labels "Normal" and "Pathological", respectively, denoted as $p_{\text{CL}}(\text{"Normal"})$ and $p_{\text{CL}}(\text{"Pathological"})$. The uncertainty value is the standard deviation of the predicted labels, denoted as SD. Then we use the following condition to assign labels:

If $\max(\text{SD}, p_{\text{SEG}}(\text{"Pathological"})) > 0.5$ **then** set label "Pathological";
Otherwise, set label "Normal".

This condition means that the outcome label will be "Pathological" in case of an SD value greater than 0.5 or in case of a probability greater than 0.5 for the "Pathological" class. The label will be "Normal" in other cases.

3. Results and discussion

3.1. Implementation details

We developed the ADELUC algorithm using Python. The training and testing of the models involved in this algorithm have been conducted on an NVidia Quadro RTX 6000 GPU, which is equipped with 24 GB of RAM. The total training time for each model is approximately 20 minutes. Notably, the peak memory usage during the training of the segmentation model reaches about 15 GB.

Once the models are trained, the process of detecting the carpal region, predicting segmentation, and scoring and classifying a new radiograph is executed in a matter of seconds. It is important to note that the majority of this execution time is attributed to loading the models rather than the processing itself.

3.2. Evaluation

The performance of our algorithm has been evaluated based on a unique set of test images, including 30 normal radiographs and 10 pathological ones, and are summarized in Table 2.

3.2.1. Detection

Despite the limited data, the performance of CAR-DET is observed to be high on the test dataset, with a F1-score reaching 1.0 in terms of instance detection on the normal images, with a confidence threshold of 0.5 (Fig. 2). The $mAP@0.5:0.95$, that considers the IoU (Intersection over Union) of the bounding boxes of the detected regions for multiple confidence thresholds between 0.5 and 0.95, reaches 0.691 on the overall dataset, while it is 0.734 on the normal subgroup, and 0.543 on the pathological subgroup. This metric shows that, although a detection is obtained for each radiograph, there may be some slight variations in the coordinates of the detected region. However, we have observed that these variations have no negative impact on the detections, that tend to have more proximity to the carpal region than the provided annotations.

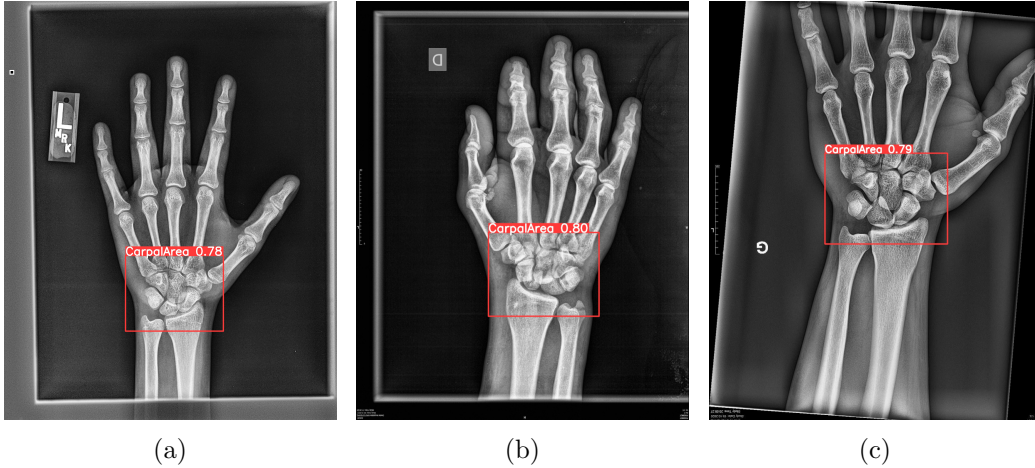


Figure 2: (a) demonstrates the detection of the carpal region on a normal radiograph, while (b) and (c) show this detection on pathological radiographs. In all cases, the carpal region is identified with a confidence level greater than 0.5.

3.2.2. Segmentation

In our test dataset, the normal images predominantly exhibit continuous and smoothly segmented Gilula’s crescents, with minimal uncertainty observed in almost all cases. As illustrated in Fig. 3, the clear detection of normality is evident in images (a)-(c). The minor standard deviations observed are primarily due to variations at the ends of the segmented Gilula’s crescent. However, for images (d)-(f), the mean Standard Deviation (mSD) value exceeds 0.5, indicating a significant deviation from the training dataset norms and suggesting these cases may require additional scrutiny by radiologists.

Conversely, for the pathological images, the CAR-GIL-SEG model demonstrates varied segmentation outcomes, as depicted in Fig. 4. These outcomes include 1) absence of segmentation, 2) a connected but abnormal shape, and 3) a disconnected shape. When segmentation occurs, it typically results in a high score for the pathological class, accompanied by substantial uncertainty. The images (a)-(e) exhibit a strong indication of pathology, with both the class probabilities and mSD values corroborating the pathological assessment. In the case of image (f), while the average segmentation result leans towards a normal classification, the elevated mSD value signals a possible pathological condition.

Regarding the label assignment by our CAR-GIL-SEG-SCL method, it is worth mentioning that it reaches a precision of 1.0 on the normal radiographs, and a recall (or sensitivity) of 1.0 on the pathological radiographs. This means that the classification method involves *no false positives* on the normal subgroup, and *no false negative* on the pathological subgroup. In other words, our method does not miss any pathological cases on the test dataset, which is of primary importance in a clinical setting, however at the cost of a few normal radiographs that are falsely detected as potentially pathological.

3.2.3. Comparison with classification

The classification model CAR-CL and scoring CAR-CL-SCL exhibit lower performance compared to the CL-CAR-GIL-SCL method. On the normal subgroup, the precision only attains 0.882, whereas the recall (or sensitivity) on the pathological subgroup is 0.600. This means that the classification results in many false negatives on the pathological subgroup. This result supports the conclusion that our method based on the segmentation of Gilula’s crescents is a more robust approach than the direct classification of images of the carpal region for the detection of the PLDs.

3.3. Discussion

The results presented in this study show that it is possible to develop an Artificial Intelligence tool to improve the diagnosis of PDLs from frontal wrist radiographs by leveraging the larger availability of normal radiographs w.r.t. pathological radiographs. However, this study remains limited by the relatively small dataset used, and the reduced diversity of the image origins (Strasbourg University Hospital and from RSNA Pediatric Bone Age Challenge [19]). Confronting our trained models to external tests datasets would be a necessary step towards considering clinical application of the proposed method, as it is known that the performance of DL models is subject to variations when applied to novel datasets [30]. To overcome the lack of data availability, another line of work would be to produce artificial images,

e.g. from parametric 3D models, or from generative models. The challenge then would be to produce realistic images, which remains a largely unexplored research area.

Also, another future work could be to conduct a study mixing our method for the analysis of frontal radiographs with the method proposed by Pridgen et al. [16] for the analysis of lateral radiographs.

4. Conclusion

In this paper, we have introduced a DL algorithm designed to automatically detect PLDs in frontal wrist radiographs. Our methodology employs two YOLOv8 deep neural models that have been trained, validated, and tested using a dataset comprising 374 annotated radiographs. This dataset includes 345 normal images and 29 pathological images. The algorithm we propose achieves a F1-score of 0.880 on this dataset. The F1-score reaches 0.928 on the normal subgroup with a precision of 1.0, and 0.833 on the pathological subgroup with a recall (or sensitivity) of 1.0, meaning absence of false negatives on this subgroup, supporting the conclusion that our method has the potential to improve the diagnostic accuracy of PDLs.

	Normal	Pathological	Total
Set \ size	345	29	374
Training w/o aug.	245	15	260
Training w/ aug.	245	255	500
Validation	70	4	74
Test	30	10	40

Table 1: Data splits used for training and evaluating the performance of the models involved in our processing pipeline.

5. Acknowledgments

This work is part of the HealthTech Interdisciplinary Thematic Institute of the ITI 2021-2028 program of the University of Strasbourg, CNRS and Inserm. This work was supported by funding from the French Investments for the Future Program managed by the ANR under references ANR-10-IDEX-0002 (IdEx Unistra), SFRI (STRAT’US project, ANR-20-SFRI-0012) and ANR-10-IAHU-02 (IHU Strasbourg).

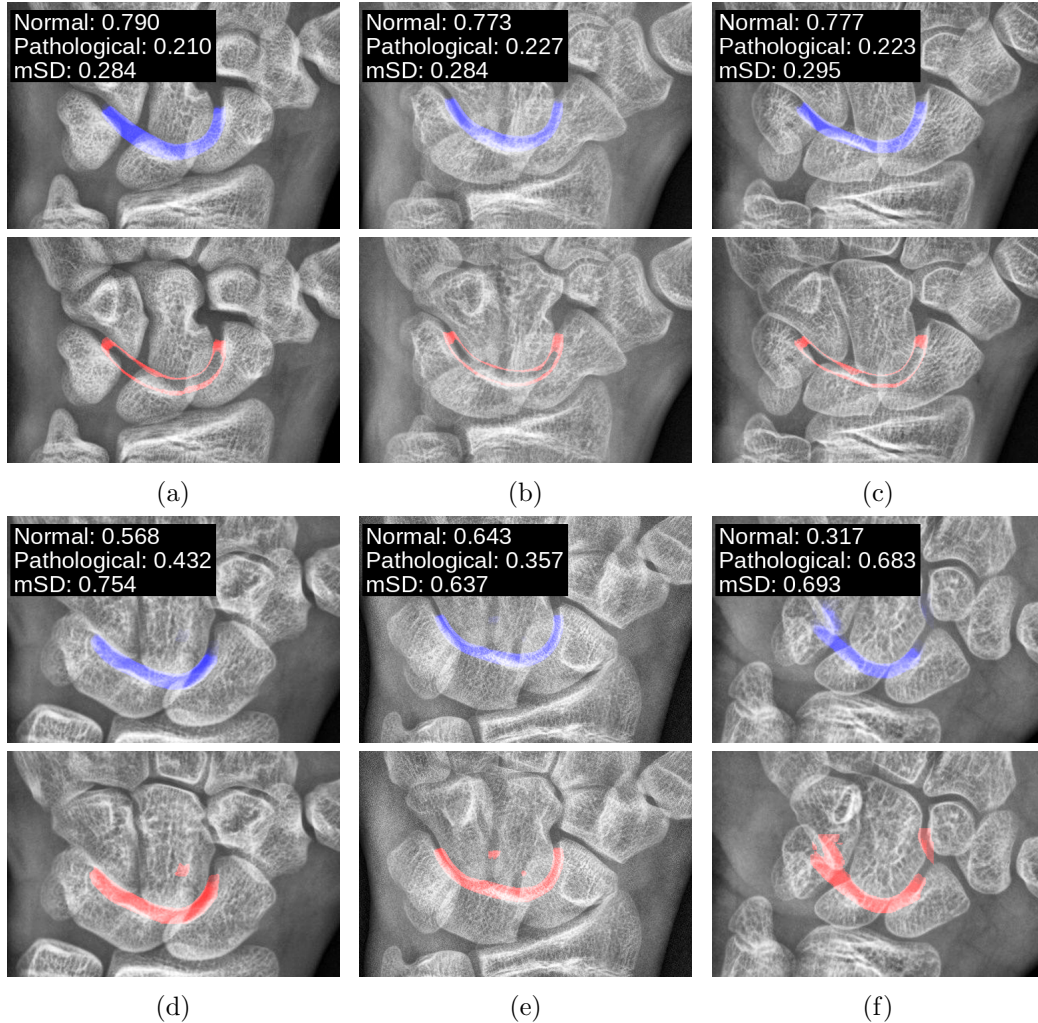


Figure 3: Display of ADELUC algorithm results on normal radiographs. Each pair of images comprises: (top row) the segmentation outcome and scoring by CAR-GIL-SEG, and (bottom row) the standard deviation of the predicted segmentations. In the top row images, the intensity of the overlaid segmentation masks corresponds to the mean values $\mu(\mathbf{p})$, whereas in the bottom row, it reflects the standard deviation values $\sigma(\mathbf{p})$. Images (a)-(c) demonstrate a clear indication of normality. Conversely, images (d)-(f) display elevated mean Standard Deviation (mSD) values, signifying high uncertainty and indicating the need for further examination by a radiologist.

Appendix A. Data splits and training parameters

The data splits for our study are detailed in Table 1, as seen on the left side. It is important to note that the same images have been utilized across the train, validation, and test sets for all models and throughout all experiments.

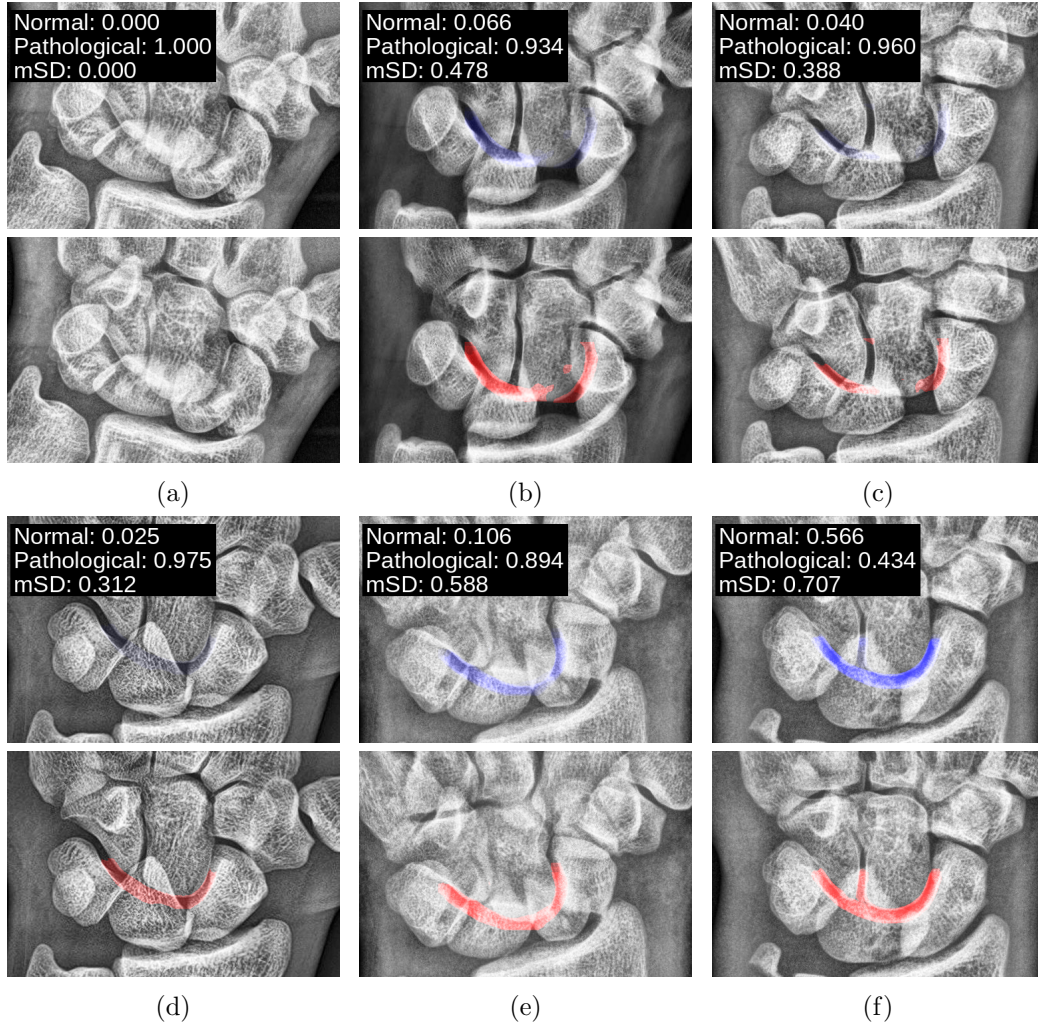


Figure 4: Representative results of the ADELUC algorithm applied to pathological radiographs. In each image pair, the top row illustrates the segmentation results and scoring achieved by CAR-GIL-SEG, while the bottom row displays the standard deviation of these segmentations. The pixel-wise intensity in the top row segmentation masks correlates with the mean values $\mu(\mathbf{p})$, and in the bottom row, it corresponds to the standard deviation values $\sigma(\mathbf{p})$. The images (a)-(e) reveal a pronounced detection of abnormality. Image (f), however, presents an interesting case: the average segmentation result leans towards a normal classification, but the heightened uncertainty, as indicated by the mSD value, suggests its categorization within the pathological class.

In line with standard practice, we have trained YOLOv8 models that had been pre-trained on conventional image processing datasets. Specifically, the YOLOv8x model (for detection) and YOLOv8x-seg model (for segmentation) have been pre-trained on the COCO dataset. Meanwhile, the YOLOv8x-cls

Normal			
Model \ Metrics	Precision	Recall	F1-score
CAR-GIL-SEG ($\times 40$)	0.864 \pm 0.149	0.867 \pm 0.150	0.864 \pm 0.144
CAR-GIL-SCL	1.000	0.867	0.928
CAR-CL ($\times 40$)	0.879 \pm 0.049	1.000 \pm 0.000	0.934 \pm 0.027
CAR-CL-SCL	0.882	1.000	0.937

Pathological			
Model \ Metrics	Precision	Recall	F1-score
CAR-GIL-SEG ($\times 40$)	1.000 \pm 0.000	0.995 \pm 0.008	0.997 \pm 0.004
CAR-GIL-SCL	0.714	1.000	0.833
CAR-CL ($\times 40$)	1.000 \pm 0.000	0.576 \pm 0.191	0.712 \pm 0.158
CAR-CL-SCL	1.000	0.600	0.750

Global			
Model \ Metrics	Precision	Recall	F1-score
CAR-GIL-SEG ($\times 40$)	0.898 \pm 0.142	0.899 \pm 0.141	0.897 \pm 0.137
CAR-GIL-SCL	0.857	0.933	0.880
CAR-CL ($\times 40$)	0.935 \pm 0.024	0.788 \pm 0.095	0.823 \pm 0.093
CAR-CL-SCL	0.941	0.800	0.843

Table 2: Performance metrics of the segmentation and classification models used in our experiments, computed on the same test set consisting of 30 normal and 10 pathological radiographs. The top and middle arrays provide the metrics for the "Normal" and "Pathological" subgroups, whereas the bottom array shows the metrics for the overall test set.

model (for classification) has been pre-trained using the ImageNet dataset.

We have standardized the batch size at 16 for all training sessions, with a maximum epoch limit set at 300. However, in practical terms, training has been halted if the loss value stabilized over the last 50 epochs. Our training process incorporates online augmentations, uniformly applied to all images in each epoch. The parameters for these augmentations have been randomly selected from uniform distributions. The applied augmentations include scaling by $\pm 10\%$, translation by $\pm 10\%$, and rotation by $\pm 20^\circ$. Considering the inherent intensity and contrast variations in radiographs, we also have implemented colorimetric augmentations in the HSV color space, with saturation and value gains of up to 20%. Notably, we have deactivated the mosaic augmentation typically used by YOLOv8, as it does not align with

our specific requirements, given the fixed spatial organization of radiographs. We also point out that we do not use dropout in our experiments.

References

- [1] J. H. Dobyns, R. L. Linscheid, W. P. C. III, Fractures and dislocations of the wrist and hand, then and now, *J. Hand Surg.* 8 (5) (1983) 687—690.
- [2] C. Garçon, B. Degeorge, B. Coulet, C. Lazerges, M. Chammas, Perilunate dislocation and fracture dislocation of the wrist: Outcomes and long-term prognostic factors, *Orthopaedics & Traumatology: Surgery & Research* 108 (5) (2022) 103332. doi:<https://doi.org/10.1016/j.otsr.2022.103332>.
URL <https://www.sciencedirect.com/science/article/pii/S1877056822001591>
- [3] T. Muller, J. Hidalgo Diaz, E. Pire, G. Prunières, S. Facca, P. Livernois, Treatment of acute perilunate dislocations: ORIF versus proximal row carpectomy, *Orthopaedics & Traumatology: Surgery & Research* 103 (1) (2017) 95–99. doi:<https://doi.org/10.1016/j.otsr.2016.10.014>.
- [4] A. Kinghorn, G. Finlayson, A. Faulkner, N. Riley, Perilunate injuries: Current aspects of management, *Injury* 52 (10) (2021) 2760–2767. doi:<https://doi.org/10.1016/j.injury.2021.09.012>.
URL <https://www.sciencedirect.com/science/article/pii/S0020138321007865>
- [5] W. R. Obermann, Wrist injuries: pitfalls in conventional imaging, *European Journal of Radiology* 22 (1) (1996) 11–21. doi:[10.1016/0720-048X\(95\)00702-R](https://doi.org/10.1016/0720-048X(95)00702-R).
- [6] J. B. Jupiter, F. A. Nunez, Jr, F. Nunez, Sr, D. L. Fernandez, A. Y. Shin, Current perspectives on complex wrist Fracture-Dislocations, *Instr Course Lect* 67 (2018) 155–174.
- [7] D. J. Sauder, G. S. Athwal, K. J. Faber, J. H. Roth, Perilunate injuries, *Orthopedic Clinics of North America* 38 (2) (2007) 279–288, wrist Trauma. doi:<https://doi.org/10.1016/j.ocl.2007.02.001>.
URL <https://www.sciencedirect.com/science/article/pii/S0030589807000041>

- [8] P. Givissis, A. Christodoulou, B. Chalidis, J. Pournaras, Neglected trans-scaphoid trans-styloid volar dislocation of the lunate. late result following open reduction and k-wire fixation, *J Bone Joint Surg Br* 88 (5) (2006) 676–680.
- [9] G. Herzberg, M. Burnier, T. Nakamura, Arthroscopic anatomy of the TFCC with relevance to function and pathology, *J. Wrist Surg.* 10 (06) (2021) 558—564.
- [10] S. Suganyadevi, V. Seethalakshmi, K. Balasamy, A review on deep learning in medical image analysis, *International Journal of Multimedia Information Retrieval* 11 (1) (2022) 19–38. doi:10.1007/s13735-021-00218-1.
URL <https://doi.org/10.1007/s13735-021-00218-1>
- [11] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng, Mura: Large dataset for abnormality detection in musculoskeletal radiographs (2018). arXiv:1712.06957.
- [12] M. He, X. Wang, Y. Zhao, A calibrated deep learning ensemble for abnormality detection in musculoskeletal radiographs, *Scientific Reports* 11 (1) (2021) 9097. doi:10.1038/s41598-021-88578-w.
URL <https://doi.org/10.1038/s41598-021-88578-w>
- [13] R. Aggarwal, V. Sounderajah, G. Martin, D. S. W. Ting, A. Karthikesalingam, D. King, H. Ashrafiyan, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, *npj Digital Medicine* 4 (1) (2021) 65. doi:10.1038/s41746-021-00438-z.
URL <https://doi.org/10.1038/s41746-021-00438-z>
- [14] N. Subramanian, O. Elharrouss, S. Al-Maadeed, M. Chowdhury, A review of deep learning-based detection methods for covid-19, *Computers in Biology and Medicine* 143 (2022) 105233. doi:<https://doi.org/10.1016/j.combiomed.2022.105233>.
URL <https://www.sciencedirect.com/science/article/pii/S0010482522000257>
- [15] F. Hardalaç, F. Uysal, O. Peker, M. Çiçeklidağ, T. Tolunay, N. Tokgöz, U. Kutbay, B. Demirciler, F. Mert, Fracture detection in wrist x-ray images using deep learning-based object detection models, *Sensors* 22 (3) (2022). doi:10.3390/s22031285.
URL <https://www.mdpi.com/1424-8220/22/3/1285>

- [16] B. Pridgen, L. von Rabenau, A. Luan, A. J. Gu, D. S. Wang, C. Langlotz, J. Chang, B. Do, Automatic detection of perilunate and lunate dislocations on wrist radiographs using deep learning, *Plastic and Reconstructive Surgery* (2023).
- [17] J. M. Johnson, T. M. Khoshgoftaar, Survey on deep learning with class imbalance, *Journal of Big Data* 6 (1) (2019) 27.
- [18] R. P. Johnson, The acutely injured wrist and its residuals, *Clin. Orthop.* (149) (1980) 33–44.
- [19] RSNA Pediatric Bone Age Challenge (2017), <https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pediatric-bone-age-challenge-2017>.
- [20] Label Studio, <https://labelstud.io>.
- [21] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection (2016). [arXiv:1506.02640](https://arxiv.org/abs/1506.02640).
- [22] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [23] YOLOv8, <https://github.com/ultralytics/ultralytics>.
- [24] D. Bolya, C. Zhou, F. Xiao, Y. J. Lee, Yolact: Real-time instance segmentation, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.
- [25] J. Terven, D. Cordova-Esparza, A comprehensive review of YOLO: From YOLOv1 and beyond (2023). [arXiv:2304.00501](https://arxiv.org/abs/2304.00501).
- [26] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: *Graphics Gems IV*, Academic Press Professional, Inc., USA, 1994, p. 474–485.
- [27] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6405–6416.

- [28] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, *Neurocomputing* 338 (2019) 34–45. doi:<https://doi.org/10.1016/j.neucom.2019.01.103>.
URL <https://www.sciencedirect.com/science/article/pii/S0925231219301961>
- [29] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, B. Cody, A. S. Mansfield, S. Rakshit, R. Bansal, M. C. Bois, A. O. Bungum, J. J. Schulte, E. E. Vokes, M. C. Garassino, A. N. Husain, A. T. Pearson, Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology, *Nature Communications* 13 (1) (2022) 6572.
- [30] A. C. Yu, B. Mohajer, J. Eng, External validation of deep learning algorithms for radiologic diagnosis: A systematic review, *Radiology: Artificial Intelligence* 4 (3) (2022) e210064. doi:[10.1148/ryai.210064](https://doi.org/10.1148/ryai.210064).