Osteoarthritis and Cartilage Open 8 (2026) 100821

Contents lists available at ScienceDirect

Osteoarthritis and Cartilage Open

journal homepage: www.elsevier.com/journals/osteoarthritis-and-cartilage-open/2665-9131

Deep learning-based classification of interphalangeal finger joints in erosive hand osteoarthritis

Zakaria Oubbi [c,*,1], Gust Verbruggen [d,**,1], Sam Leroux [a,b], Pieter Simoens [a,b]

[a] IDLab, Department of Information Technology at Ghent University, Ghent, Belgium
[b] IMEC, Interuniversity Microelectronics Centre, Ghent, Belgium
[c] ML6, Ghent, Belgium
[d] Osteoarthritis Imaging LLC, Nazareth, Belgium

1. Introduction

After osteoarthritis (OA) affects the knee and hip joints, it most commonly develops in the hand.

A significant proportion of the patients with hand OA suffer from the erosive type of OA. The prevalence of erosive hand OA ranged from 3% in men to 10% in women of Caucasian origin [1]. This distinct form of OA affects primarily the interphalangeal (IP) finger joints and is rapidly progressing and destructive [2–6]. Therefore, many patients with erosive hand OA experience significant discomfort and functional disability [7–10]. Accurate staging of the structural damage sustained by the affected joint tissues is essential for developing an effective therapeutic approach for these patients.

The tissue changes in erosive hand OA include resorption of articular cartilage usually preceding osteolysis and resorption of the subchondral bone endplate. At the same time osteolytic changes occur in the epiphyseal subchondral bone area [11,12]. Articular tissue destruction is then followed by reparative features such as remodeling of the subchondral bone plate and the formation of bony nodules at the margins of the affected IP joints.

On radiographs, in addition to classic OA features such as joint space narrowing, subchondral bone plate sclerosis, subchondral cysts, and osteophyte formation, IP joints affected by erosive OA exhibit more aggressive changes. These include a complete loss of joint space, which typically precedes or coincides with the collapse of the subchondral bone plate. This destructive phase is followed by a reparative stage, characterized by the reappearance of a new subchondral plate covered with fibrocartilaginous tissue within the previously eroded joints. These erosive and remodeling changes can occur concurrently in both proximal and distal IP joints, often in an unpredictable order within the same patient. As a result, the disease may remain active for several decades, progressing toward a terminal remodeling phase in all affected joints [11]. To standardize the assessment of erosive hand OA progression, the five-stage Verbruggen-Veys (V&V) radiographic scoring system was developed [11]. This ordinal system classifies IP joints into five categories: normal (N), static with typical OA features (S), joint space loss (J), erosive (E), and remodeling (R), representing a characteristic progression: $N \prec S \prec J \prec E \prec R$. The defining features of erosive hand OA include alterations in the subchondral plate, specifically resorption (E) and subsequent repair (R).

Progression through these anatomical stages has been documented earlier [13–15] Over one year, around 4.0–7.0% of N, S or J joints in placebo patient cohorts showed erosive (E) progression. Erosive evolution was higher in interphalangeal joints showing signs of clinical inflammation, i.e. swelling [13]. Another 5.0%–8.0% E joints from placebo treated patients in these reports showed remodeling (R) within 48 weeks of follow-up.

Conventional radiography remains the primary imaging modality for diagnosing and monitoring OA. Increasingly, computer-aided diagnosis systems assist in radiographic interpretation by providing an objective second opinion, reducing the reliance on multiple expert readers. For erosive hand OA, an automated approach to V&V scoring could improve diagnostic efficiency and consistency. Given the discrete nature of the labeling scale, this problem is well-suited for machine learning-based classification. Deep learning is especially suitable due to the use of image data as input. In particular, convolutional neural networks have shown strong performance in medical image analysis, including applications to musculoskeletal disorders. To address data limitations, transfer learning with pre-trained models can be employed [16–18]. Furthermore, the ordinal nature of the V&V scoring system presents an opportunity to explore ordinal classification techniques, which can

* Corresponding author.
** Corresponding author.
E-mail addresses: zakaria.oubbi@hotmail.com (Z. Oubbi), gvbrugge@gmail.com (G. Verbruggen), sam.leroux@ugent.be (S. Leroux), pieter.simoens@ugent.be (P. Simoens).
[1] ZO and GV share first authorship.

https://doi.org/10.1016/j.ocarto.2026.100821
Received 28 November 2025; Accepted 11 May 2026

potentially capture the inherent relationship between disease stages more effectively than traditional multi-class classification methods.

To our knowledge, no prior work has leveraged deep learning to automate the scoring of erosive hand OA. This study aims to fill that gap by developing and evaluating a deep learning-based automated system for scoring erosive hand OA using full-hand radiographs. We employed a two-step approach, first detecting individual IP joints using YOLOv8 [19], and then classifying them according to the V&V scoring system using a fine-tuned ResNet-50 architecture [20]. Critically, we investigated the impact of data augmentation, class imbalance mitigation, and transfer learning strategies on model performance. We explored the potential of improving model performance by leveraging the ordinal nature of the V&V scores through an ordinal classification framework, specifically Conditional Ordinal Regression for Neural Networks (CORN) [21]. By developing a reliable automated scoring system, we aim to improve the efficiency and consistency of IP OA assessment, ultimately facilitating more effective clinical management and research.

2. Materials and methods

DATASET: SOURCE AND CONSTRUCTION: The datasets utilized consisted of full hand radiographs that originated from three controlled pharmacological interventional studies [13–15]. The three studies included independent patient cohorts of 60, 85, and 100 participants, with no patient participating in more than one study. Patients were selected based on IP joints showing J and/or E phases.

Posteroanterior full hand radiographs were taken at baseline and after 24 and 48 weeks during these studies. For each patient and timepoint, eight DIP and eight PIP joints were obtained, providing a total of 11104. Detailed information is given in a supplementary table.

Radiographs were provided in JPEG/TIFF format to enable the reading on PCs of the different readers.

The protocol allowed the creation of a machine-learning-ready dataset containing 11096 IP finger joint images and their respective V&V scores. Five IP joints with surgical intervention were excluded.

The dataset was divided into training end test sets at the patient level, ensuring that all images from a given patient were assigned exclusively to one subset to prevent data leakage. Each trial was split using the same approach, preserving proportional representation across training and test sets. The final training and test sets contained 9792 and 1304 joint images, respectively, corresponding to 27 patients in the test set. A stratified approach maintained class distribution in each subset to address class imbalance. Percentage distribution for the anatomical phases N, S, J, E and R was 28%, 35%, 7%, 8% and 22%, respectively.

SCORING METHOD: The V&V categorical scoring system assesses the progression of erosive hand OA through five distinct anatomic phases: N (Normal), S (Stationary (non-erosive)), J (Joint Space Loss), E (Erosive), and R (Remodeling). The phases reflect the evolution of the affected IP finger joint, from normal to eroded and remodeled states, and follow an ordinal relationship, representing the sequential progression of the disease: $N \prec S \prec J \prec E \prec R$ [11]. Each phase is characterized by specific features observed in imaging. The defining features of erosive hand OA include alterations in the subchondral plate, specifically resorption (E) and subsequent repair (R). (Fig. 1).

LABELING: V&V scores were given for all patients' 16 IP joints at baseline and at each study time point. Consistency in scoring across datasets was maintained as scoring was done independently by 2 experienced readers in the Ghent rheumatology department. Intra- and inter-reader reproducibility were both excellent. Labels assigned to the IP joints from the 3 studies were adopted in the current dataset for subsequent supervised deep learning analysis.

IMPLEMENTATION: This work utilizes the open-source machine learning libraries PyTorch [22] and PyTorch Lightning [23] for implementing the neural networks. Additional software libraries include YOLOv8 [19] for IP joint detection and PyTorch Image Models (timm) [24] for loading the pretrained ResNet-50.

IP FINGER JOINT DETECTION: Erosive hand OA affects individual IP finger joints, which necessitates an approach that can effectively utilize full-hand radiographs as input. The strategy used was inspired by similar work for rheumatoid arthritis and involves a two-step process where joint detection and scoring are treated as separate tasks [25,26]. This separating detection from scoring allows models to be more precisely optimized for each task, potentially improving performance. Joint detection was performed using YOLOv8 [19], pretrained on the COCO dataset [27] and finetuned on a manually annotated subset of 100 full-hand X-ray images. While the model distinguishes DIP and PIP joints separately with high precision and recall (Fig. 2A), it does not inherently assign a joint index, i.e., the corresponding finger.

A custom algorithm was therefore implemented to determine hand orientation (left or right) based on bounding box positions, enabling automatic indexing. The detected bounding boxes were used to automatically crop regions of interest (PIP and DIP joints) from full-hand radiographs, forming the V&V dataset. The bounding box's largest dimension was used for cropping on square dimensions, with an additional fixed 15-pixel margin (applied uniformly to all joints) to retain relevant joint features and support data augmentation. Cropping was performed on the original-resolution radiographs; therefore, the margin was constant in pixel units and does not correspond to a fixed physical size (mm). DIP and PIP joints were modeled jointly, as the V&V scoring system applies uniformly across IP joints. These cropped images, along with their labels, served as inputs for the joint scoring models (Fig. 2B).

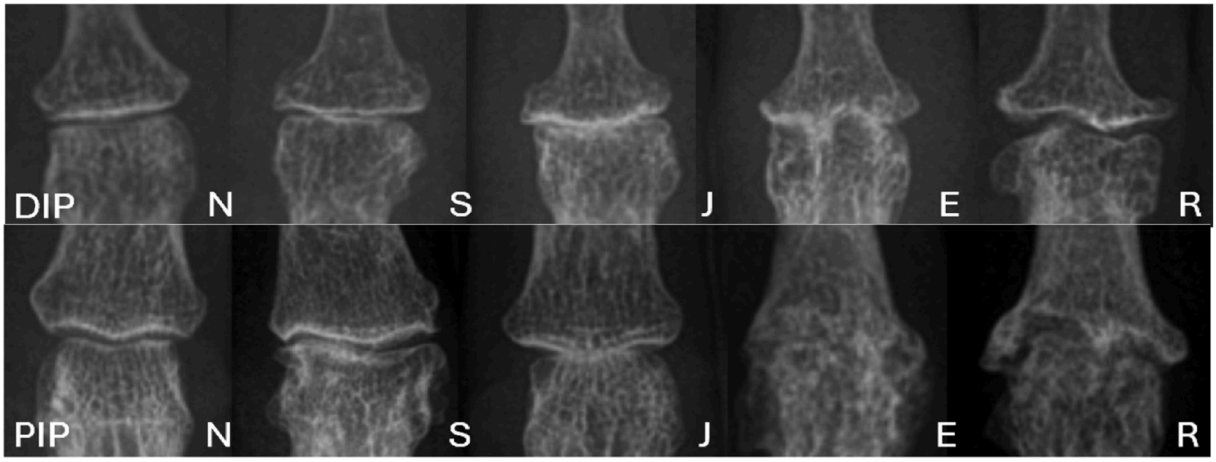

Fig. 1. Sequential anatomical phases in a distal (D) and proximal (P) IP finger joint.

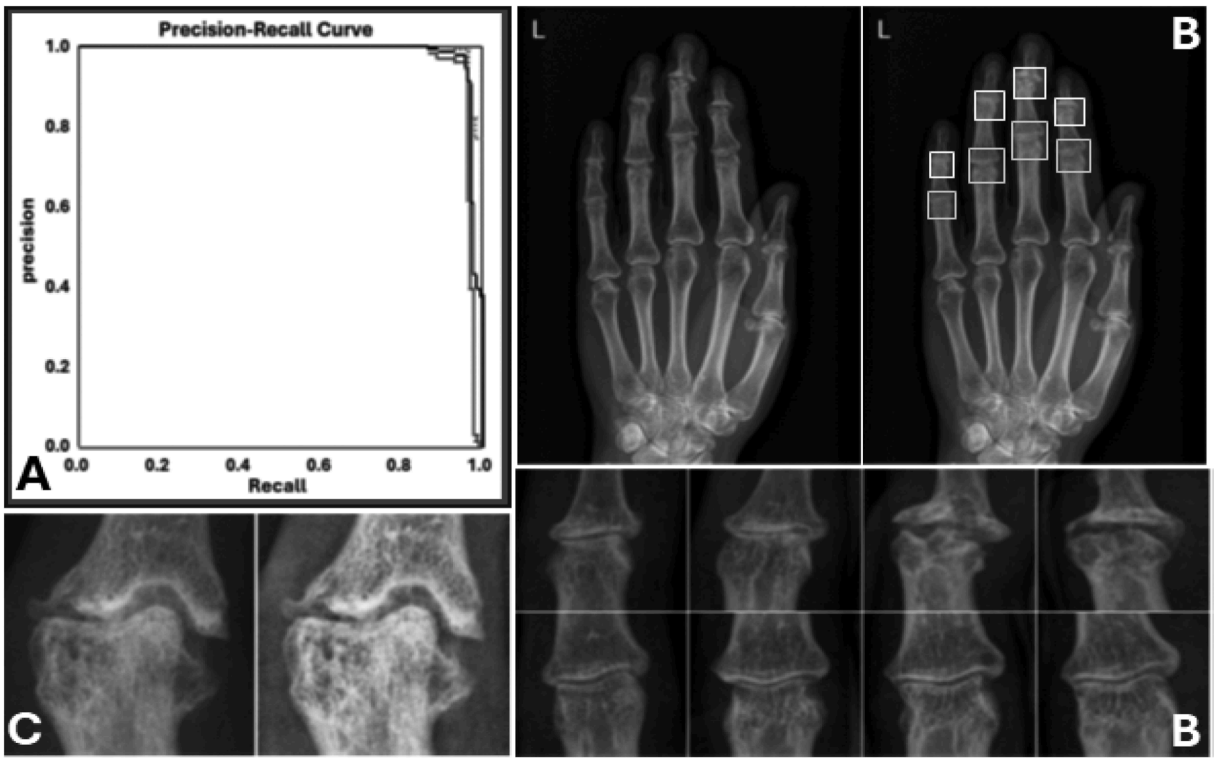

Fig. 2. A- Precision-recall curve of the YOLOv8 joint detector on an unseen test set of 32 full hand radiographs. B- The joint extraction process. The left upper image shows the original hand radiograph while the right upper image shows the same radiograph with the joint predictions from the joint detection model. These joints are cropped on the bounding boxes resulting in the 8 joint images at the bottom. C - CLAHE's effect on a joint image.

IP FINGER JOINT SCORING - MODEL TRAINING: In all experiments, a ResNet-50 model pretrained on the ImageNet dataset [28] was used. The output layer was adapted to the five-score classification of the V&V scoring system. Starting with the conventional (i.e. vanilla) classification, we explored different strategies for data augmentation such as random rotations, horizontal flips, translations, and scaling [29,30]. Some studies suggest using Contrast Limited Adaptive Histogram Equalization (CLAHE) [29] to enhance contrast in X-ray images [31–33]. CLAHE adjusts contrast locally while limiting noise amplification, improving image clarity and potentially aiding classification. To assess its impact, the augmentation pipeline was tested with and without CLAHE. The clip limit, which controls contrast intensity, was empirically set to 2. Fig. 2C illustrates CLAHE's effect on a joint image. We then explored different finetuning strategies where increasingly larger parts of the model were finetuned. Finally, we moved from conventional classification to ordinal classification. To implement ordinal classification, we tested one framework: CORN [21]. This framework is model-agnostic and compatible with any neural network, allowing us to adapt the pretrained ResNet-50's architecture for ordinal classification. The same dataset splits, preprocessing steps, and data augmentation methods as those employed for conventional classification were applied.

STATISTICAL ANALYSIS: For evaluating the results obtained from automating the V&V scoring system, unless stated otherwise, all metrics are reported in macro-averaged form. In this approach, metrics are calculated independently for each class and then averaged with equal weight across classes, ensuring that minority classes contribute equally despite class imbalance.

During the pretraining phase, a 5-fold stratified cross-validation scheme was used to obtain averaged results. Given the extensive training time required for finetuning large model segments, a systematic approach was adopted for the finetuning phase. First, the ResNet-50 was pretrained using the best configuration across three independent runs, each with a different random seed. Each pretrained model was then finetuned using its corresponding seed. Final performance was averaged across the three runs to obtain a more robust estimate. This approach was applied consistently to both conventional and ordinal classification experiments.

Ground-Truth Verification: To assess potential ambiguity in ground-truth labels, a subset of misclassified joints from the test set was re-evaluated. Fifty misclassified joints (10 per anatomical phase) were randomly selected. Sample identifiers were pseudorandomized to conceal study origin, patient identity, and trial information. An expert reader independently re-scored the selected joints without access to the original labels or model predictions. Agreement between the expert, ground-truth labels, and model predictions was quantified using percentage agreement and Cohen's kappa coefficient.

3. Results

3.1. MODEL TRAINING

Conventional classification: As a baseline, we only finetuned the output layer of the model. We explored different hyperparameter combinations using the ADAM optimizer [34] ($\beta1 = 0.9$, $\beta2 = 0.999$, no AMSGrad) with a batch size of 32. A 5-fold stratified cross-validation was applied, averaging evaluation metrics Area-Under-Curve (AUC) score and cross-entropy loss across all five runs. Each run used a different random seed, but the same random seeds were chosen across the different hyperparameter configurations for fair comparison. Hyperparameter combinations tested are shown in Table 1. Overfitting occurred before decay had an impact, and Learning rate (LR) = 1e-4 provided the most stable results, making it the chosen configuration.

Overfitting and Class Imbalance: Phases J and E, typically the most critical stages in the progression of erosive hand OA, were underrepresented in the dataset. We tested three strategies to mitigate this imbalance (Table 2). All experiments used the optimal hyperparameter configuration from the previous section, with a baseline comparison of the configuration without any imbalance or overfitting strategies. The

Table 1
Hyperparameter search for the ADAM optimizer.

Learning rate (LR)	Decay	Loss	AUC
1e−3	0	1.13 ± 0.04	0.77 ± 0.02
1e−3	1e−3	1.12 ± 0.03	0.77 ± 0.02
1e−4	0	**1.11 ± 0.03**	**0.78 ± 0.02**
1e−4	1e−3	1.11 ± 0.03	0.77 ± 0.02
1e−5	0	1.13 ± 0.03	0.77 ± 0.02
1e−5	1e−3	1.14 ± 0.03	0.77 ± 0.02

Metrics averaged over 5 runs, Mean ± SD given.

Table 2
Class imbalance and overfitting strategy.

Strategy	Loss	AUC
Baseline (no strategy)	1.11 ± 0.03	0.78 ± 0.02
ROS with augmentation	1.16 ± 0.03	0.79 ± 0.01
ROS with augmentation + CLAHE	**1.16 ± 0.03**	**0.80 ± 0.01**
Class weighting	1.21 ± 0.04	0.77 ± 0.02
ROS with augmentation + class weighting	1.57 ± 0.04	0.77 ± 0.01

Metrics averaged over 5 runs, Mean ± SD given. AUC: Area Under Curve; ROS: Random Oversampling; CLAHE: Contrast Limited Adaptive Histogram Equalization.

best results came from random oversampling combined with data augmentation that included CLAHE as a step in the pipeline. The class-penalty strategy using class weights in cross-entropy loss performed poorly. It is important to note that the minimum loss was higher with imbalance and overfitting strategies, as the model was still in the pre-training phase and had limited capacity. Without these strategies, the model focused on overrepresented classes, leading to higher “apparent" performance.

Transfer learning: To improve the performance of the model, we experimented with finetuning larger parts of the model. We found that finetuning the entire ResNet-50 yielded the best performance, suggesting the dataset was large enough to support deep tuning. However, increasing the number of trainable parameters also heightened the risk of overfitting. As finetuning depth increased, overfitting occurred earlier, as reflected in the epochs of minimal validation loss (Table 3). For instance, when finetuning only the first residual block (i.e. the block before the output layer), minimal loss occurred at epoch 342, whereas for the full model (at the same learning rate of 1e-5), it occurred at epoch 24. This indicates that deeper finetuning was necessary due to the substantial difference between the source (natural images) and target (hand radiographs) domains.

We also experimented with adjusting the learning rate. Lowering it for configurations such as finetuning two residual blocks led to worse results. However, increasing it to 1e-4, despite causing overfitting after just three epochs, significantly improved performance. This suggests that lower learning rates may have trapped the model in local minima, preventing effective adaptation. Additionally, the substantial domain shift likely required more aggressive weight updates. The best-performing model was selected based on the best evaluation metrics, and its confusion matrices on the train and test sets are presented in Fig. 3a and b.

Table 3
Finetuning up to different block-depths in the ResNet-50.

Block depth	Learning rate (LR)	Epoch min. Loss	Loss	AUC
B1	1e-5	342	1.060 ± 0.010	0.850 ± 0.002
B1-2	1e-5	79	1.040 ± 0.018	0.864 ± 0.008
	1e-6	470	1.053 ± 0.014	0.850 ± 0.003
B1-3	1e-5	76	0.975 ± 0.016	0.875 ± 0.002
B1-4	1e-5	76	0.967 ± 0.012	0.883 ± 0.002
B1-15	1e-5	24	0.892 ± 0.013	0.903 ± 0.001
B1-16	1e-5	24	0.889 ± 0.008	0.901 ± 0.001
	1e-4	3	**0.846 ± 0.013**	**0.911 ± 0.001**

Metrics averaged over 3 runs, Mean ± SD given. LR: Learning Rate; AUC/Area Under Curve.

Ordinal classification: Finally, we used the CORN framework to implement ordinal classification which is a more natural fit for the ordinal scoring scale. We explored the same hyperparameter space used for conventional classification. The ADAM optimizer [34] ($\beta1 = 0.9$, $\beta2 = 0.999$, no AMSGrad) was applied with a batch size of 32, a 5-fold stratified cross-validation scheme, and consistent random seeds.

Since full fine-tuning previously yielded the best performance, the CORN model was fully fine-tuned, with the results presented in Table 4 for comparison. Confusion matrices from the best-performing runs are shown in Fig. 3a and b (vanilla classifier) and Fig. 3c and d (CORN classifier).

4. Results on a balanced test set with verified labels

Class distribution has a direct impact on the performance of a model. To investigate this effect we constructed a balanced test set of 608 samples drawn from the full dataset, including only joints with non-debatable scores. This set contained roughly 120 samples per class, with PIP and DIP joints equally represented across all scores. A corresponding training set of 6186 samples was created from non-test patients to prevent data leakage, though it remained unbalanced. An instance of the best-performing model configuration, CORN, was trained on this set for three runs, with the averaged results: MAE 0.370 ± 0.0110 and RMSE 0.731 ± 0.0041.

Surprisingly, the model performed slightly better on this new test set than the reference model did on the original test set, despite training on 40% less data. This indicates that employing non-debatable criteria for score assignment, together with a balanced class distribution, has a significant effect on model performance. For reproducibility, the data set and the labels are

given in a supplemental file: DOI: https://doi.org/10.5281/zenodo.17297344.

5. ANALYSIS of MODEL mispredictions

We examined the errors made by CORN, our preferred model. CORN misclassified 439 instances in the test set, with errors distributed across different datasets (Fig. 4A). The most misclassified score was S, followed by R and N, which aligns with class distribution: S being the largest, followed by N and R. Despite using random oversampling, early overfitting limited the model's exposure to S samples, potentially contributing to its misclassification. Additionally, N–S confusion discussed further below likely played a role in S being the most wrongly predicted class. Finally, cautious review of these misclassifications showed that human error caused few of these wrong scores. Moving the original labels to spreadsheet documents for automation caused some errors to appear in the ground-truth data. Reviewing individual misclassified samples revealed that other misreadings were, however, justifiable.

Fig. 4B illustrates four misclassified joints, where, in some cases, the model's prediction appeared more reasonable than the ground-truth label. For instance, the joint (A) in Fig. 4B, labeled as N, seems to align more closely with the model's prediction of R. Since these joints retained the same ground-truth labels across 3 trial periods, the model accumulated 13 errors from these four joints alone. Further analysis showed that 9 out of 27 test-set patients had 20 or more misclassifications each, collectively accounting for over half of all errors. This suggested that many ground-truth labels may be ambiguous or require re-evaluation.

As described in the Methods section, fifty misclassified joints were re-scored by an expert reader to evaluate potential ground-truth ambiguity. Percentage of agreement between the expert and the ground-truth labels was 46% (Cohen's Kappa coefficient: 0.32). Percentage of agreement

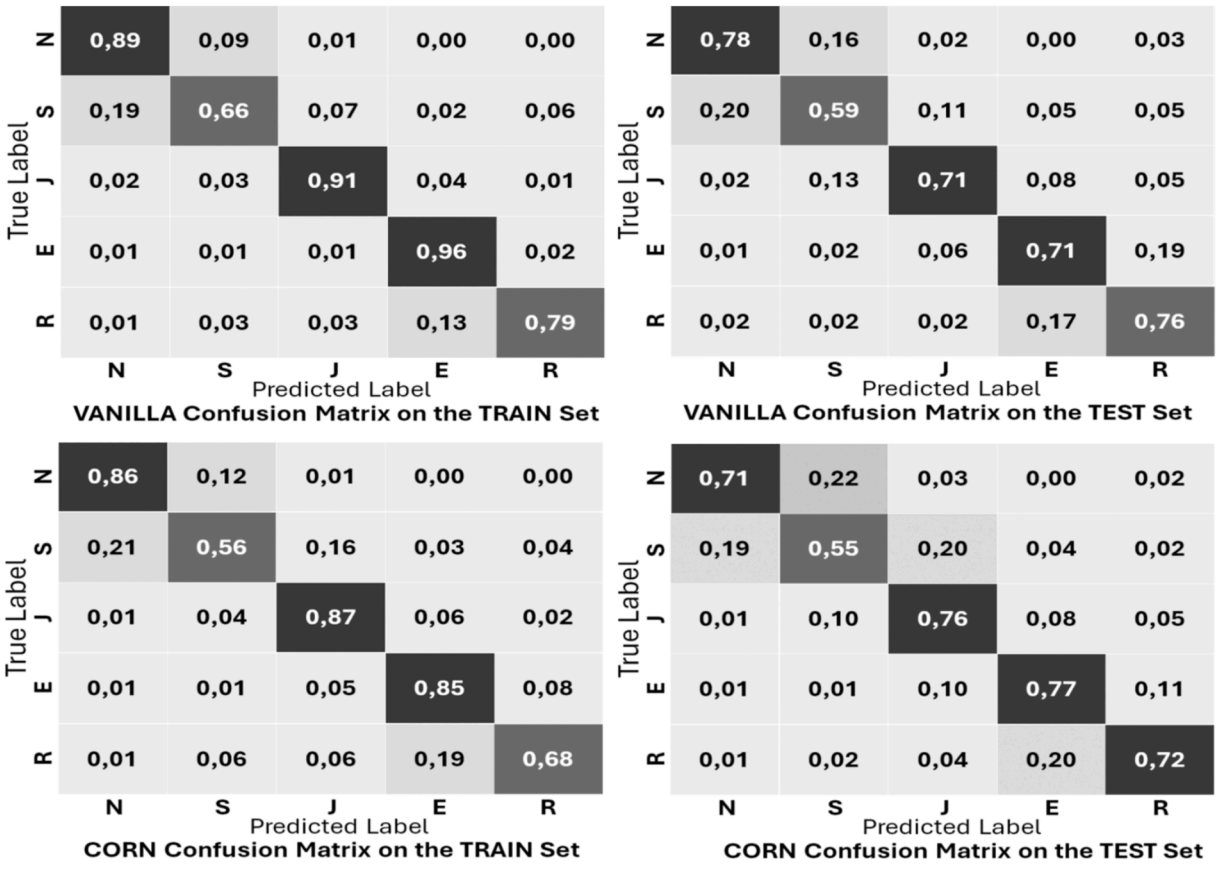

Fig. 3. Confusion Matrices of the best performing models for classification.

Table 4
Comparison of the optimal results for each model type.

Model	LR	MAE	RMSE	ACC
Baseline	1e−4	0.583 ± 0.0389	0.964 ± 0.0419	0.583 ± 0.0435
Vanilla	1e−4	0.398 ± 0.0110	0.811 ± 0.0249	**0.704 ± 0.0093**
CORN	1e−4	**0.393 ± 0.0078**	**0.791 ± 0.0190**	0.700 ± 0.0064

Metrics averaged over 3 runs, Mean ± SD given. A baseline model, trained without pretrained weights, random oversampling (ROS), or data augmentation, is included for reference.
LR: Learning Rate; MAE: Mean Absolute Error; RMSE: Root Mean Squared Error; ACC: Accuracy.

between the expert and the model predictions was 48% (Cohen's Kappa coefficient: 0.33). The percentage of agreement between the ground-truth labels and the model was 0%, as the 50 joints were all misclassified. As shown, re-scoring showed similar agreement with the model and with the ground-truth labels. This suggests that the model's actual performance may surpass its reported metrics, as some errors may reflect that many joints have labels that are subject to interpretation.

6. Discussion

Erosive hand OA is a fast-advancing type of OA that affects the interphalangeal finger joints.

Radiographs reveal progression through distinct anatomical phases, each phase marking its progression and providing a scoring system presenting unique opportunities for diagnosis and assessment [11].

Machine learning has hitherto been applied in radiology for image classification. Previous research has used deep learning to localize joints on hand radiographs [35,36] and to differentiate types of joint diseases, e.g. osteoarthritis and rheumatoid arthritis [37,38].

After diagnosing joint disease, grading its severity is crucial as it guides the clinician's treatment options. Deep learning methods graded the severity of joint damage in patients with rheumatoid arthritis [25,26,36]. However, deep learning has not yet been employed to automate the scoring of erosive hand OA. This study addresses that gap by applying deep learning to automate the ordinal V&V scoring system.

Our dataset included radiographs from 245 subjects taken at three intervals, covering over 10,000 IP finger joints. X-ray images obtained from roughly equal numbers of patients enabled the development of deep learning methods elsewhere. Hirano et al. used 216 radiographs from 108 patients with rheumatoid arthritis; 11160 images of PIP and MCP joints were used to assess joint space narrowing and erosions [25]. Peng et al. X-rayed 344 hands and developed scores for arthritis injuries [38]. Most of these studies applied augmentation techniques to extract more data from limited datasets.

Our approach compared conventional and ordinal classification using a pretrained ResNet-50. Both methods performed well, with deep tuning and class imbalance strategies proving effective. As shown, CORN and the conventional “vanilla” classifier performed similarly. Despite their comparable overall performance, differences emerged in the confusion matrices: the vanilla classifier better distinguished N from S joints, whereas CORN was more effective at separating E from R joints. However, distinguishing between certain successive anatomical phases remained challenging. Both models frequently misclassified N and S joints, with S joints often incorrectly labeled as N, and less commonly as J. Misclassifications between E and R joints were also frequent. These observations are explored further below.

CORN was chosen as the preferred model for three reasons: (1) it performed better on underrepresented scores J and E, which are crucial for assessment; (2) it exhibited less overfitting than the vanilla classifier, as observed in the confusion matrices; and (3) it achieved slightly better

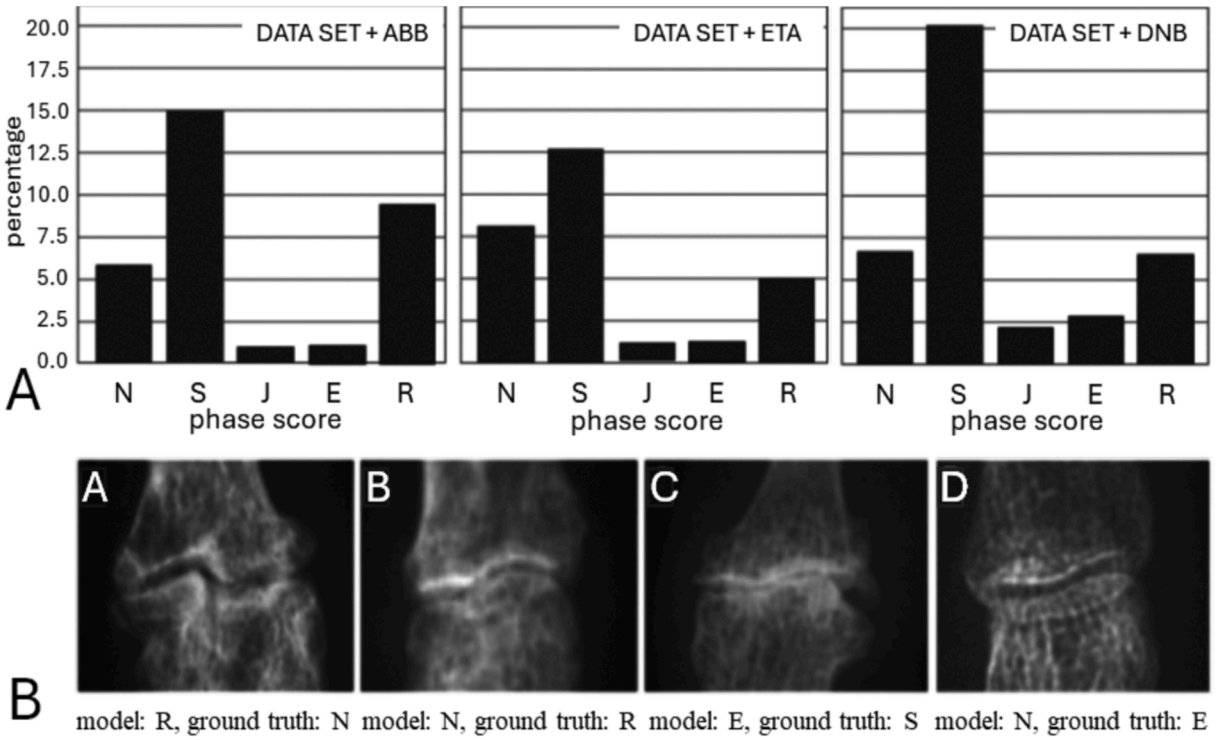

Fig. 4. A Distribution of the mispredictions per dataset normalized over all mispredictions. Of the 439 mispredictions, 171 (39%) came from the DNB dataset, 144 (33%) from the ADA dataset, and 124 (28%) from the ETA dataset. Most mispredictions happened to the S-joints. B Examples of model mispredictions: IP joint A: erroneous label in the ground-truth; IP joint C: misprediction of the model.

MAE and RMSE scores, which quantify prediction accuracy for ordinal labels.

The CORN classifier performed best on classes J and E (erosive OA joints), achieving accuracies of 76% and 77%, respectively. This finding is important as the presence of $\geq$1 IP joint in the J or E phase of the anatomical phase scoring system is a key inclusion criterion for patients in trials exploring the effects of pharmacological agents on structure modification in erosive hand OA.

The model's performance (in between 70 and 80% confidence) is in line with the inherent reader's hesitancy while deciding on specific anatomical phases of affected IP finger joints. The subtle changes in the anatomy of non-affected N IP joints and those showing the first signs of OA may be dubious to discern. The frequent occurrence of N–S confusion is due to the minimal visual changes in the joints during this phase transition. Regarding S-J confusion, it was observed that severely narrowed and barely visible joint spaces often lead to misclassification. Expert readers assess joint space narrowing ideally by comparing with the contralateral IP joint. Such comparison is not done by the ordinal classification model. Misclassification of IP joints by the ordinal classifier arises from the difficulty in determining when a joint transitions from one to another anatomical phase. Likewise, these arguable anatomical changes lead to the inter- and intrareader variabilities by experienced readers.

Furthermore, even with a detailed definition of each anatomical phase, interphalangeal finger joints advancing through these successive phases may show considerable overlapping in phase characteristics. An IP finger joint in the S phase showing a 90 % reduced joint space width strictly does not meet the definition of an J phase joint when some erosion of the subchondral plate already may appear. Technically, such an IP joint can be classified as an E joint. Likewise, a partly restored E phase IP joint can only be classified as an R phase joint when the subchondral plates entirely reappeared.

Lastly, it should be emphasized that radiographs in the clinical studies were collected at set intervals. The same joint then appears multiple times in the dataset. This has its implications: unlike the model, clinicians can use prior scores to make informed assessments. If a joint was previously classified as J, it can only transition to J, E, or R in subsequent periods. However, the model treats each radiograph as an independent sample, lacking longitudinal context. This limitation may affect performance, as the model cannot leverage temporal progression information that clinicians naturally use. Future research could explore how to incorporate longitudinal data into model training to enhance predictive accuracy.

It should be emphasized, however, that guidelines for clinical trials on hand osteoarthritis generally require radiographs to be blinded to sequence. For most studies, radiographs of each patient are presented to the readers in random order, with the date of acquisition masked. In studies that include inactive controls, some argue that blinding to sequence is not necessary as long as the readers are blinded to treatment allocation [39,40].

In conclusion, automating the V&V scoring system addresses two significant challenges: the time-consuming nature of manual scoring and the complexity that restricts its use to highly trained experts. The selection of patients for inclusion in therapeutic studies requires the participation of one or more experienced and well-trained readers. Our results showed that trained models performed comparably to expert clinicians on test sets, while completing assessments in a fraction of the time, and even reducing the risk of human error. Using a pretrained ResNet-50 and comparing conventional and ordinal classification approaches, we demonstrated that deep learning methods can support experts in the assessment process. Finally, such an AI assisted classification system will reduce the costs associated with therapeutical trials and enable large-scale studies.

Author contributions

Author contributions GV, ZO, PS and SL conceived the study. GV selected and labeled the data. ZO prepared deep learning strategy and

analyzed the data. GV and ZO drafted the manuscript with critical input from all authors. PS and SL provided strategic direction and senior guidance. All authors critically reviewed and approved the final manuscript.

Ethical approval

Not applicable.

Funding

None.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgments

This work was part of Zakaria Oubbi's master dissertation submitted to obtain the academic degree of master of science in computer science engineering. Assignment of Verbruggen-Veys scores to the IP joints in previous the studies [13–15] was done by G. Verbruggen and R. Wittoek.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ocarto.2026.100821.

References

[1] I.K. Haugen, M. Englund, P. Aliabadi, J. Niu, M. Clancy, T.K. Kvien, et al., Prevalence, incidence and progression of hand osteoarthritis in the general population: the Framingham osteoarthritis study, Ann. Rheum. Dis. 70 (2011) 1581–1586.

[2] R.M. Stecher, H. Hauser, Heberden's nodes. VII. The roentgenological and clinical appearance of degenerartive joint disease of the fingers, Am. J. Roentgenol. 59 (1948) 326–337.

[3] L. Punzi, R. Ramonda, P. Sfriso, Erosive osteoarthritis, Best Pract. Res. Clin. Rheumatol. 18 (2004) 739–758.

[4] W.Y. Kwok, M. Kloppenburg, F.R. Rosendaal, J.B. van Meurs, A. Hofman, S.M. A. Bierma-Zeinstra, Erosive hand osteoarthritis: its prevalence and clinical impact in the general population and symptomatic hand osteoarthritis, Ann. Rheum. Dis. 70 (2011) 1238–1242.

[5] T. Vanhaverbeke, L. Pardaens, R. Wittoek, Natural disease progression in finger osteoarthritis: results from a 10 year follow-up cohort, Scand. J. Rheumatol. 49 (2020) 498–504.

[6] D.C. Crain, Interphalangeal osteoarthritis. Characterized by painful, inflammatory episodes resulting in deformity of the proximal and distal articulations 175 (12) (1961) 1049–1053, https://doi.org/10.1001/jama.1961.03040120011003.

[7] I. Kjeken, H. Dagfinrud, B. Slatkowsky-Christensen, P. Mowinckel, T. Uhlig, T. K. Kvien, et al., Activity limitations and participation restrictions in women with hand osteoarthritis: patients' descriptions and associations between dimensions of functioning, Ann. Rheum. Dis. 64 (2005) 1633–1638.

[8] J. Bijsterbosch, I. Watt, I. Meulenbelt, F.R. Rosendaal, T.W.J. Huizinga, M. Kloppenburg, Clinical burden of erosive hand osteoarthritis and its relationship to nodes, Ann. Rheum. Dis. 69 (2010) 1784–1788.

[9] I.K. Haugen, B. Slatkowsky-Christensen, P. Boyesen, D. van der Heijde, T.K. Kvien, Cross-sectional and longitudinal associations between radiographic features and measures of pain and physical function in hand osteoarthritis, Osteoarthr. Cartil. 21 (2013) 1191–1198.

[10] R. Wittoek, B. Vander Cruyssen, G. Verbruggen, Predictors of functional impairment and pain in erosive osteoarthritis of the interphalangeal joints: comparison with controlled inflammatory arthritis, Arthritis Rheum. 64 (2021) 1430–1436.

[11] G. Verbruggen, E.M. Veys, Numerical scoring systems for the anatomic evolution of osteoarthritis of the finger joints, Arthritis Rheum. 39 (1996) 308–320.

[12] G. Verbruggen, R. Wittoek, B. Vander Cruyssen, D. Elewaut, Morbid anatomy of 'erosive osteoarthritis' of the interphalangeal finger joints: an optimised scoring system to monitor disease progression in affected joints, Ann. Rheum. Dis. 69 (2010) 862–867.

[13] G. Verbruggen, R. Wittoek, B. Vander Cruyssen, D. Elewaut, TNF-blockade for the treatment of erosive osteoarthritis of the interphalangeal finger joints: a double blind, randomized trial on structure modification, Ann. Rheum. Dis. 71 (2012) 891–898.

[14] M. Kloppenburg, R. Ramonda, K. Bobacz, W.Y. Kwok, D. Elewaut, T.W.I. Huizinga, et al., Etanercept in patients with inflammatory hand osteoarthritis (EHOA): a multicentre, randomised, double-blind, placebo-controlled trial, Ann. Rheum. Dis. 77 (2018) 1757–1764.

[15] R. Wittoek, G. Verbruggen, T. Vanhaverbeke, R. Colman, D. Elewaut, RANKL blockade for erosive hand osteoarthritis: a randomized placebo-controlled phase 2a trial, Nat. Med. 30 (2024) 829–836.

[16] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, et al., A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88, https://doi.org/10.1016/j.media.2017.07.005.

[17] S. Dargan, M. Kumar, M.R. Ayyagari, G. Kumar, A survey of deep learning and its applications: a new paradigm to machine learning, Arch. Comput. Methods Eng. 27 (2019) 1071–1092, https://doi.org/10.1007/s11831-019-09344-w.

[18] S.M. Anwar, M. Majid, A. Quayyum, M. Awais, M. Alnowami, M.K. Khan, Medical image analysis using convolutional neural networks: a review, J. Med. Syst. 42 (2018) 1–19. 13, https://doi.org/10.1007/s10916-018-1088-1.

[19] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics.

[20] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv: 1512.03385.

[21] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recognition Letters 2020 140 (2020) 325–331, https://doi.org/10.1016/j.patrec.2020.11.008.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, et al., Pytorch: an Imperative Style, high-performance Deep Learning Library, 2019 arXiv: 1912.01703, http://arxiv.org/abs/1912.01703.

[23] W. Falcon, J. Borovec, A. Walchli, N. Eggert, J. Schock, J. Jordan, et al., PyTorchLightning/pytorch-lightning: 0.7.6 release. https://zenodo.org/record/3828935.

[24] R. Wightman, Pytorch Image Models, 2019. https://github.com/rwightman/pytorch-image-models.

[25] 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 3 (2019) rkz047, https://doi.org/10.1093/rap/rkz047. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6921374/.

[26] J. Rohrbach, T. Reinhard, B. Sick, O. Durr, Bone erosion scoring for rheumatoid arthritis with deep convolutional neural networks, Comput. Electr. Eng. 78 (2019) 472–481. https://www.sciencedirect.com/science/article/pii/S0045790618329409.

[27] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, et al., Microsoft COCO: Common Objects in Context, 2014 arXiv:1405.0312v3, https://arXiv:1405.0312v3.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848. Miami, FL, USA.

[29] E. Goceri, Medical image data augmentation: techniques, comparisons and interpretations, Artif. Intell. Rev. 56 (11) (2023) 12 561–605, https://doi.org/10.1007/s10462-023-10453-z.

[30] N. Salem, H. Malik, A. Shams, Medical image enhancement based on histogram algorithms, Procedia Comput. Sci. 163 (2019) 300–311. https://www.sciencedirect.com/science/article/pii/S1877050919321519.

[31] S.M. Pizer, E.P. Amburn, J.D. Austin, R. Cromartie, A. Geselowitz, T. Greer, et al., Adaptive histogram equalization and its variations, Comput. Vis. Graph Image Process 39 (3) (1987) 355–368. https://www.sciencedirect.com/science/article/pii/S0734189X8780186X.

[32] S. Pizer, R. Johnston, J. Ericksen, B. Yankaskas, K. Muller, Contrast-limited adaptive histogram equalization: speed and effectiveness, in: Proceedings of the First Conference on Visualization in 95 6 References Biomedical Computing, May 1990, pp. 337–345. https://ieeexplore.ieee.org/document/109340.

[33] E.A. Tjoa, I.P. Yowan Nugraha Suparta, R. Magdalena, N. Kumalasari, The use of CLAHE for improving an accuracy of CNN architecture for detecting pneumonia, in: SHS Web of Conferences, 139, 2022 03026. https://www.shs-conferences.org/10.1051/shsconf/202213903026.

[34] D.P. Kingma, J. Ba, Adam : a method for stochastic optimization. http://arxiv.org/abs/1412.6980.

[35] S. Lee, M. Choi, H-s Choi, M.S. Park, S. Yoon, FingerNet: deep learning-based robust finger joint detection from radiographs, in: IEEE Biomedical Circuits and Systems Conference (BioCAS), 2015, pp. 1–4, https://doi.org/10.1109/BioCAS.2015.7348440.

[36] K. Maziarz, A. Krason, Z. Wojna, Deep Learning for Rheumatoid Arthritis: Joint Detection and Damage Scoring in X-rays. Deep Learning for Rheumatoid Arthritis: Joint Detection and Damage Scoring in X-rays. Workshop on AI for Public Health at ICLR, 2021, https://doi.org/10.48550/arXiv.2104.13915 arXiv/2104.13915v2.

[37] K. Üreten, H.H. Maraş, Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods, J. Digit. Imag. 35 (2) (2022) 193–199, https://doi.org/10.1007/s10278-021-00564-w.

[38] Y. Peng, X. Huang, M. Gan, K. Zhang, Y. Chen, Medical imaging Radiograph-based rheumatoid arthritis diagnosis via convolutional neural network, BMC Med. Imag. 24 (2024) 180–189, https://doi.org/10.1186/s12880-024-01362-w.

[39] E. Maheu, R.D. Altman, D.A. Bloch, M. Doherty, M. Hochberg, A. Mannoni, et al., Design and conduct of clinical trials in patients with osteoarthritis of the hand: recommendations from a task force of the osteoarthritis research society international, Osteoarthr. Cartil. 14 (4) (2006) 303–322.

[40] D.J. Hunter, N. Arden, F. Cicuttini, M.D. Crema, B. Dardzinski, J. Duryea, et al., OARSI clinical trials recommendations: hand imaging in clinical trials in osteoarthritis, Osteoarthr. Cartil. 23 (2015) 732–746, https://doi.org/10.1016/j.joca.2015.03.003.