Annotation-Efficient Osteophyte Severity Estimation in Hip X-rays

David-Andrei Gogoana

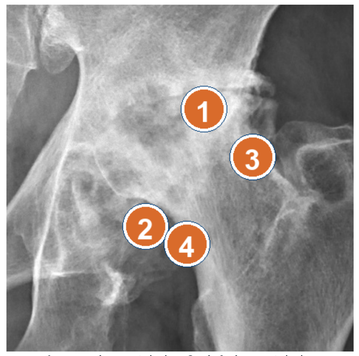
Delft University of Technology, EEMCS Faculty
Supervisors: Gijs van Tulder and Jesse Krijthe

1. Background & Motivation

Hip osteoarthritis is a degenerative joint disease that can cause pain, stiffness and reduced mobility. In radiographic studies, the presence of the disease is commonly assessed via hip X-ray imaging. One important imaging sign is an **osteophyte**: a bony outgrowth near the joint margin.

Osteophytes are useful because they indicate both the presence and severity of osteoarthritis. In this project, each hip crop is assessed at four anatomical osteophyte sites. The OARSI atlas grades each site using ordered severity categories [1]:

0 = absent 1 = small 2 = medium 3 = large



1 superior acetabular **2** inferior acetabular
3 superior femoral **4** inferior femoral

Approximate region markers on illustrative hip OA radiograph (not study data).

Source: Mikael Hagstrom, Wikimedia Commons, CC0.

Annotation bottleneck. Detailed OARSI grading requires expert radiographic assessment. The reader must inspect subtle shape differences around the joint margin and distinguish neighbouring grades. This is clinically meaningful, but it is slow and expensive when repeated across large image cohorts and multiple anatomical locations.

A simpler binary label, by contrast, only records whether an osteophyte is present. It is less informative than a full grade, but it still captures the first severity threshold: absence versus presence.

Motivation. The aim is to reduce annotation time and cost without discarding severity estimation. Instead of requiring every training sample to receive a detailed grade, this project asks whether many binary presence labels can be combined with a much smaller set of expert OARSI grades.

2. Research Question

How effectively can binary osteophyte-presence labels be combined with limited OARSI grades for ordinal severity estimation in hip X-rays?

Guiding subquestions:

1. Do scores from a binary-presence model carry information about true OARSI severity?
2. How does severity-estimation performance change as the number of graded annotations increases?
3. Does score-stratified selection for grading improve annotation efficiency compared with random selection?
4. Where does performance differ across severity thresholds and anatomical osteophyte locations?

3. Methodology

The model treats OARSI grades as ordered severity levels rather than four unrelated classes. For each hip location, it asks three cumulative yes/no questions:

$$q_1 = P(y \geq 1) \text{ presence}, \quad q_2 = P(y \geq 2) \text{ moderate-or-worse}, \quad q_3 = P(y \geq 3) \text{ severe}.$$

These thresholds are ordered, so the probabilities stay consistent: $q_1 \geq q_2 \geq q_3$. The continuous severity estimate used for ordinal error is:

$$\hat{y} = q_1 + q_2 + q_3$$

Hip X-ray Crop → Ordinal Thresholds → Masked Loss → Expected Severity Score

Label source	What the loss is allowed to learn
Binary negative	no osteophyte, so all thresholds are negative: $(0, 0, 0)$
Binary positive	osteophyte present, so only $y \geq 1$ is known: $(1, ?, ?)$
Full OARSI grade	exact ordinal target is known, e.g. grade 2 gives $(1, 1, 0)$

The **masked ordinal loss** lets each label train only what it actually tells us. Binary labels supervise presence, while full OARSI grades supervise all severity thresholds.

Binary Labels Anchor Presence + Graded Labels Calibrate Severity

- Binary-positive samples answer only the first question: is an osteophyte present?
- Graded samples also answer whether the osteophyte is moderate-or-worse or severe.
- Unknown severity thresholds are ignored, not guessed or imputed.

4. Experiments

Dataset and split
Experiments use 224×224 hip X-ray crops from the CHECK and OAI studies. Each crop has four OARSI osteophyte locations. Splits are subject-level, so images from the same person do not appear in both training and test data.

Split	Images	Fully graded
Train	15,306	12,660
Validation	3,316	2,736
Test	3,293	2,741

Annotation-budget simulation
Experiments simulate limited expert grading from a grade-complete dataset. Every training crop retains its binary osteophyte-presence label, but only samples in the fixed budget retain the full OARSI grade.

Binary Only	Learns presence from all binary labels
Mixed Ordinal	Uses all binary labels plus the visible graded subset
Full Supervision	Uses all 12,660 graded training samples

Evaluation focus
Performance is measured on the held-out, fully graded test set. The primary metric is macro-MAE on expected severity $\hat{y}$, with AP/AUROC for $y \geq 1$, $y \geq 2$ and $y \geq 3$. Results are reported as mean $\pm$ SD over three seeds.

5. Results

64 graded samples recover 78% of the macro-MAE gain

Binary only	→	64 grades	→	Full grading
1.005		0.688		0.602
macro-MAE		macro-MAE		macro-MAE

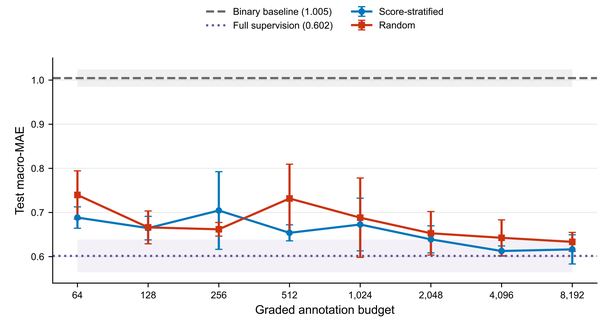


Figure 1. Macro-MAE falls quickly with a small graded subset, then approaches full supervision more gradually.

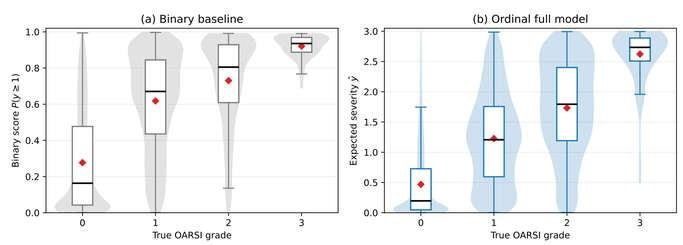


Figure 2. Binary scores rise with true OARSI grade, but neighbouring positive grades still overlap.

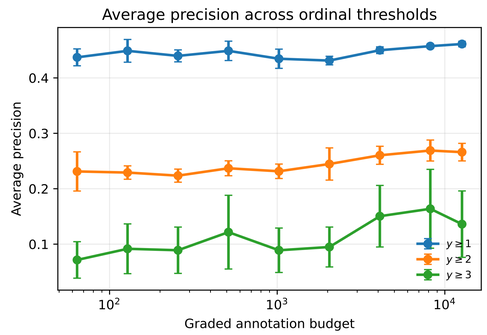


Figure 3. AP is strongest for $y \geq 1$. Severe-threshold performance remains unstable because grade 3 is rare.

6. Discussion & Conclusion

The results suggest that binary osteophyte-presence labels are not merely coarser substitutes for OARSI grades. They provide useful supervision for the first ordinal threshold, $y \geq 1$. This helps explain why the mixed model improves sharply at small graded budgets: binary labels establish presence, while full grades calibrate the moderate and severe boundaries. However, the violin and box plots show that neighbouring positive grades overlap, so binary confidence alone cannot reliably separate mild, moderate and severe osteophytes. The AP analysis also shows that performance is strongest for presence detection and weakest for the rare severe threshold, where grade-3 scarcity limits learning.

Conclusion. A small set of expert OARSI grades, combined with many binary labels, can recover most of the average-error improvement while keeping the remaining clinical risk around rare severe cases explicit.

Reference. [1] Altman and Gold, *Osteoarthritis and Cartilage*, 2007.

d.gogoana@student.tudelft.nl