Annotation-Efficient Osteophyte Severity Estimation in Hip X-rays

Combining Binary Presence Labels with Limited OARSI Grade Supervision

David-Andrei Gogoana[1]
Supervisor(s): Jesse Krijthe[1]**, Gijs van Tulder**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: David-Andrei Gogoana
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Julia Olkhovskaia

Abstract

Detailed OARSI grading of osteophytes, an important radiographic indicator of hip osteoarthritis, is expensive because it requires expert annotation, whereas coarser binary presence labels are far easier to obtain. This study investigates how effectively these binary labels can be combined with a limited number of graded labels to estimate ordinal osteophyte severity in hip X-ray crops, and whether the choice of which samples to grade matters. We formulate the task as cumulative ordinal regression over four anatomical locations per hip, in which binary labels supervise the presence threshold and graded labels supervise the higher severity thresholds, while thresholds with no available grade are left unsupervised. A binary-only baseline detected osteophyte presence well and produced confidence scores that rose with true grade, but could not resolve the higher grades. A few graded labels enabled ordinal expected-severity estimates and reduced macro-averaged mean absolute error, with the largest gains at the smallest budgets and diminishing returns beyond. Comparing score-stratified sampling against random selection of the graded subset, the score-based strategy was competitive but not consistently better, indicating that most of the benefit comes from adding graded supervision rather than from how the samples are chosen. All results are reported on a held-out test set, averaged over three seeds. Combining many binary labels with relatively few graded labels is a promising way to reduce expert annotation burden while still producing useful ordinal severity estimates.

1 Introduction

Osteoarthritis is a common degenerative joint disease that can cause pain, stiffness, and reduced mobility [1]. In hip osteoarthritis, structural disease progression is commonly assessed from radiographic X-ray images. Radiographic scoring systems evaluate features such as joint-space narrowing, sclerosis, deformity, and osteophytes. Osteophytes are bony outgrowths near the margins of the joint and are important radiographic indicators of osteoarthritis presence and severity [2]. In the OARSI atlas, osteophyte severity is described using ordered categories ranging from absence to severe osteophyte formation [2].

Manual grading of osteophyte severity is clinically meaningful, but it is also expensive. It requires expert interpretation of radiographic structures, and the distinction between neighbouring severity levels can be subtle. This becomes a practical limitation in large osteoarthritis studies, where many images must be assessed consistently. Automated image analysis could reduce this annotation burden, but fully supervised deep learning methods usually require large numbers of detailed labels [3]. In medical imaging, such labels are often costly because they require expert time and domain-specific knowledge.

A possible way to reduce this burden is to use labels with different levels of detail. In osteophyte assessment, a binary label only indicates whether an osteophyte is present, while a graded label also describes its severity. Binary labels are less informative, but they are cheaper to obtain and can still supervise the distinction between absence and presence. Graded labels are more informative, but collecting them for every image may be unnecessary if a smaller graded subset is sufficient. This suggests a mixed-supervision setting, where many coarse labels are combined with fewer detailed labels. Related weakly supervised learning approaches have worked well for detection or presence tasks in radiology and medical imaging when full expert annotations are unavailable or expensive [4, 5].

A second relevant aspect is that osteophyte severity is ordinal. The categories have a natural order: confusing a mild osteophyte with a moderate osteophyte is less severe than confusing absence with severe disease. Standard multi-class classification does not explicitly

use this ordering. Ordinal regression methods instead model ordered outcomes and have been used when prediction targets represent severity, stage, or rank rather than independent classes [6, 7]. This makes ordinal modelling a natural fit for OARSI severity estimation.

Although weak supervision and ordinal regression have each been studied, they are typically treated separately. Weakly supervised work in radiology often uses coarse image-level labels or incomplete labels, but usually targets detection or presence rather than ordered severity [5, 4]. Ordinal-regression methods, by contrast, model ordered outcomes but generally assume that full graded labels are available for all training samples [6, 7, 8]. This motivates a weakly supervised ordinal-learning problem: in a realistic annotation-limited setting, many samples may have only a coarse binary osteophyte-presence label, while only a smaller subset receives detailed OARSI grading.

The challenge is that binary labels supervise whether an osteophyte is present, but they do not directly distinguish mild, moderate, and severe osteophytes. This study therefore uses a fully graded dataset to simulate annotation-limited conditions. During training, OARSI grades are exposed only for a selected subset of samples, while the remaining samples are treated as binary-labelled. This allows us to study how much graded supervision is needed for ordinal osteophyte severity estimation.

A second question arises if detailed grading can still be assigned: instead of accepting the graded subset as fixed, can we choose which binary-labelled samples should receive full OARSI grading? This connects the setting to active learning, where annotation is directed towards samples expected to be informative for the model [9]. In this study, this idea is tested by comparing random selection with score-stratified selection based on the binary model output.

This study investigates the following research question:

> *How effectively can binary osteophyte-presence labels be combined with a limited number of OARSI graded labels for ordinal osteophyte severity estimation in hip X-rays?*

We address this question through four subquestions. First, we examine whether a model trained only on binary presence labels produces prediction scores that are associated with true OARSI severity, even though it is not explicitly trained to distinguish severity grades. Second, we evaluate how severity estimation changes as the number of graded annotations increases. Third, we compare whether selecting samples for grading using a binary-model score improves annotation efficiency compared with random selection, which matters if detailed grading can be targeted. Fourth, we analyse whether performance differs across severity thresholds and anatomical osteophyte locations.

The contribution of this work is an empirical evaluation of annotation-efficient osteophyte severity estimation under mixed supervision. We compare models trained with different fixed proportions of binary and graded supervision, evaluate the effect of increasing the graded annotation budget, and compare score-based sample selection with random sample selection. The budget is varied across separate experiments to estimate annotation efficiency; it is not increased during a single training run.

The rest of this paper is organised as follows. Section 2 describes the mixed-supervision ordinal method. Section 3 presents the dataset, experimental setup, and evaluation metrics. Section 4 reports the results. Section 5 discusses responsible research considerations. Section 6 interprets the findings, discusses limitations, and outlines future work.

2 Methodology

2.1 Location-Specific Osteophyte Severity as an Ordinal Task

The goal of this study is to estimate osteophyte severity from hip X-ray images while reducing the amount of full OARSI grade supervision required during training. For each image x_i, osteophyte severity is predicted at four anatomical locations: superior acetabular, inferior acetabular, superior femoral, and inferior femoral. Let l index the anatomical location. The desired target is the OARSI grade

$$y_{i,l} \in \{0, 1, 2, 3\},$$

where grade 0 indicates absence of an osteophyte and grades 1, 2, and 3 indicate increasing osteophyte severity.

A binary osteophyte-presence label is defined as

$$b_{i,l} = \begin{cases} 0, & y_{i,l} = 0, \\ 1, & y_{i,l} > 0. \end{cases}$$

The binary label indicates whether an osteophyte is present, but it does not distinguish between mild, moderate, and severe osteophytes. In this study, the binary labels are derived deterministically from the available OARSI grades, so they should be interpreted as noiseless coarse versions of the same annotation source rather than as independently collected binary annotations. To study annotation-limited learning, we formulate the task as a mixed-supervision simulation: each sample provides a binary presence label, while only a subset G provides the detailed severity grade. Grades outside G are considered unobserved by the method. The task is therefore to learn an ordinal severity model from abundant coarse supervision and scarce detailed supervision, under simulated partial grading.

2.2 Two-Stage Mixed-Supervision Pipeline

The method consists of two training stages and an annotation-selection step. First, a binary osteophyte-presence model is trained using all available binary labels. This model learns the first ordinal threshold, corresponding to $y \geq 1$, and is used in two ways: as the binary-only baseline and as an initialization for ordinal training.

Second, after binary pretraining on all training samples, a subset of training samples is selected to provide full OARSI grade supervision. We compare two selection strategies for this subset: random sampling and score-stratified sampling based on the pretrained binary model output. The samples not selected for full grading remain binary-labelled only.

Third, the mixed-supervision ordinal model is trained using both label types. Samples in the graded subset supervise all ordinal thresholds, while binary-only samples supervise only the thresholds justified by their binary labels. This is implemented with a masked ordinal loss, so that unavailable severity thresholds do not contribute to training.

2.3 Ordinal Severity Formulation

OARSI grades are ordered categories rather than independent classes: grade 2 is closer to grade 3 than to grade 0. Standard four-class classification ignores this ordering, so severity estimation is instead formulated as cumulative ordinal regression. Rather than predicting a

single class, the model predicts, for each anatomical location, the probability that the grade exceeds each severity level:

$$q_{i,l,1} = P(y_{i,l} \geq 1 \mid x_i), \quad q_{i,l,2} = P(y_{i,l} \geq 2 \mid x_i), \quad q_{i,l,3} = P(y_{i,l} \geq 3 \mid x_i),$$

corresponding to osteophyte presence, moderate-or-worse severity, and severe osteophytes.

For each location the network produces a single scalar severity score $f_{i,l}$, together with three location-specific cutpoints $c_{l,1} < c_{l,2} < c_{l,3}$ that are ordered by construction. The threshold logits are obtained by comparing the score against each cutpoint,

$$a_{i,l,k} = f_{i,l} - c_{l,k}, \qquad q_{i,l,k} = \sigma(a_{i,l,k}),$$

where σ is the sigmoid function. Using one score per location with ordered cutpoints, rather than three independent outputs, guarantees that the threshold probabilities are non-increasing, $q_{i,l,1} \geq q_{i,l,2} \geq q_{i,l,3}$, so the cumulative interpretation stays consistent.

This cumulative parameterisation is also what makes mixed supervision natural. A binary presence label corresponds exactly to the first threshold ($y \geq 1$), so binary and graded labels supervise the same per-location score at different cutpoints: a binary label constrains only the presence threshold, whereas a graded label constrains all three. A single regression output could not absorb a binary present/absent label as supervision of one specific severity boundary.

The expected severity score is the sum of the threshold probabilities,

$$\hat{y}_{i,l} = q_{i,l,1} + q_{i,l,2} + q_{i,l,3},$$

which is the expected OARSI grade under this model, using the identity $\mathbb{E}[y] = \sum_k P(y \geq k)$ for a non-negative ordinal variable. This gives a continuous estimate between 0 and 3. Because the thresholds are ordered, a high $P(y \geq 3)$ also forces $P(y \geq 1)$ and $P(y \geq 2)$ to be high, so a near-maximal $q_{i,l,3}$ drives $\hat{y}_{i,l}$ towards 3 rather than towards a low value. The expected severity score is the primary model output used for ordinal error evaluation.

For secondary hard-grade analyses, the cumulative probabilities can be converted into per-grade probabilities:

$$P(y = 0) = 1 - q_1, \quad P(y = 1) = q_1 - q_2, \quad P(y = 2) = q_2 - q_3, \quad P(y = 3) = q_3,$$

where the image and location indices are omitted for readability. These are treated as secondary because the main task is ordinal severity estimation, for which the expected severity and threshold probabilities are more informative than an exact class assignment.

2.4 Shared Backbone and Location-Specific Heads

The model uses a shared convolutional backbone followed by separate prediction heads for the four anatomical osteophyte locations. This follows a hard-parameter-sharing design: the backbone learns image features shared across locations, while each location-specific head produces the threshold predictions for one anatomical site [10]. The backbone uses residual-style convolutional blocks inspired by residual learning [11].

The same architecture is used for binary pretraining and mixed ordinal training. In the binary stage, each location head predicts only the first threshold, $P(y \geq 1)$. In the ordinal stage, each head predicts all three cumulative thresholds, $P(y \geq 1)$, $P(y \geq 2)$, and $P(y \geq 3)$. The architecture is kept fixed across all annotation budgets and sampling strategies so that differences in performance can be attributed to the supervision setting rather than to model design.

2.5 Binary Pretraining

The first training stage uses only binary osteophyte-presence labels. For each image and anatomical location, the model predicts whether the OARSI grade is greater than or equal to 1. The binary target is therefore

$$z_{i,l,1} = b_{i,l}.$$

The model is trained with binary cross-entropy on the first threshold. For a logit a and target z, the binary cross-entropy with the sigmoid applied to the logit is

$$\ell(a, z) = -\big[\, z \log \sigma(a) + (1 - z) \log\big(1 - \sigma(a)\big) \,\big],$$

and the binary pretraining loss is

$$\mathcal{L}_{binary} = \frac{1}{N} \sum_{i,l} \ell(a_{i,l,1}, b_{i,l}).$$

The higher thresholds $P(y \geq 2)$ and $P(y \geq 3)$ are not used during binary pretraining. Therefore, the binary model is not treated as a full ordinal severity model. Its output is interpreted only as an osteophyte-presence score. Whether this score is associated with true OARSI severity is evaluated empirically in the results.

The binary model is also used to initialize the mixed-supervision ordinal models. This allows ordinal training to start from image features that already detect osteophyte presence.

2.6 Mixed-Supervision Ordinal Loss

The mixed-supervision stage uses all binary labels together with the full OARSI grades that are available. Let G denote the set of training samples for which a full OARSI grade is available. For samples in G, all three ordinal thresholds are supervised. For the remaining samples, only the binary presence label is available.

For a graded sample, the ordinal target vector is

$$(z_1, z_2, z_3) = \begin{cases} (0,0,0), & y = 0, \\ (1,0,0), & y = 1, \\ (1,1,0), & y = 2, \\ (1,1,1), & y = 3. \end{cases}$$

For a binary-only sample, the target depends on the binary label. If $b = 0$, the sample is known to have no osteophyte, so all thresholds are known negative:

$$(z_1, z_2, z_3) = (0, 0, 0).$$

If $b = 1$, the sample is known only to have grade at least 1. The first threshold is supervised as positive, while the higher thresholds are unknown:

$$(z_1, z_2, z_3) = (1, ?, ?).$$

The unknown entries are masked out of the loss.

A binary mask $m_{i,l,k}$ indicates whether threshold k is supervised for image i and location l. Each visible threshold contributes a cross-entropy term in which positive targets carry a class-imbalance weight $p_{l,k}$,

$$\ell_p(a_{i,l,k}, z_{i,l,k}) = -\Big[\, p_{l,k}\, z_{i,l,k} \log q_{i,l,k} + (1 - z_{i,l,k}) \log\big(1 - q_{i,l,k}\big) \,\Big],$$

with $q_{i,l,k} = \sigma(a_{i,l,k})$. These terms are combined using the mask, a source weight $s_{i,l,k}$, and a threshold weight τ_k:

$$\mathcal{L}_{mixed} = \frac{\sum_{i,l,k} m_{i,l,k}\, s_{i,l,k}\, \tau_k\, \ell_p(a_{i,l,k}, z_{i,l,k})}{\sum_{i,l,k} m_{i,l,k}\, s_{i,l,k}\, \tau_k}.$$

The threshold weights $\tau = (1, 1.25, 1.5)$ place more emphasis on the higher-severity thresholds. These weights were chosen heuristically and kept fixed across all experiments; their independent effect was not ablated. The source weight $s_{i,l,k}$ reflects label reliability: it is 1 for a threshold supervised by a full OARSI grade, 0.75 for a threshold inferred from a binary label, and $0.75 \times 0.35 \approx 0.26$ for the $y \geq 2$ and $y \geq 3$ negatives inferred from binary-negative samples, which are the least informative. The class-imbalance weight is capped,

$$p_{l,k} = \min\big(30,\ \max(1,\ N^-_{l,k}/N^+_{l,k})\big),$$

where $N^+_{l,k}$ and $N^-_{l,k}$ are the numbers of visible positive and negative targets at location l and threshold k; this prevents the rare positives at the higher-severity thresholds from being overwhelmed by the abundant negatives.

The mask ensures that the model is supervised only by the labels that are actually available. For a binary-positive sample without a graded label, the true grade (1, 2, or 3) is unknown, so the thresholds $y \geq 2$ and $y \geq 3$ are excluded from its loss and only the presence threshold is supervised. For binary-positive samples without grades, positive evidence for the $y \geq 2$ and $y \geq 3$ thresholds comes only from the graded subset G, while binary-negative samples can still provide weak negative supervision for these thresholds.

2.7 Annotation-Selection Strategies

The formulation above only requires a graded subset G; it does not assume how that subset was obtained. In a standard weakly supervised setting, G could simply be the set of samples that already has detailed labels. However, one may also ask what happens if additional detailed grading can be requested for specific X-rays. In that case, the question is which binary-labelled samples should be selected for grading. We compare two ways of constructing G under the same grading budget.

Random sampling. Random sampling constructs G by selecting samples uniformly from the training set. This represents the simplest way to allocate a limited grading budget: it does not require a pretrained binary model or an additional scoring step before annotation. It therefore provides both a practical baseline for real-world use and a reference point for evaluating whether a more targeted strategy is worth the extra complexity.

Score-stratified sampling. Score-stratified sampling uses the binary presence model to assign each training sample a score. The score is based on the predicted probability of osteophyte presence, aggregated across anatomical locations. Samples are then selected across the score distribution, so that the graded subset includes both likely osteophyte-positive cases and lower-score cases that preserve coverage of the dataset. This strategy uses only binary-model predictions and does not use hidden severity grades.

Both strategies produce a graded subset G. Samples in G provide detailed severity supervision, while samples outside G continue to provide only binary supervision.

3 Experimental Setup

3.1 Dataset

The data were prepared as follows. The dataset, provided for this research project, combines hip X-ray images from the CHECK and OAI osteoarthritis studies. Each sample is a 224×224 grayscale hip crop corresponding to one subject, visit, and hip side. Osteophytes are scored at four anatomical locations: superior acetabular, inferior acetabular, superior femoral, and inferior femoral. These four locations are treated as related prediction sites for the same severity-estimation task. Unless stated otherwise, reported metrics are averaged across the four anatomical locations.

The data were split at the subject level, so that all images from a given subject appear in only one of the training, validation, and test splits, preventing subject leakage across splits. The resulting split contained 15,306 training samples, 3,316 validation samples, and 3,293 test samples. Complete OARSI grade annotations were available for 12,660 training, 2,736 validation, and 2,741 test samples. The grade distribution was highly imbalanced: grade 0 accounted for approximately 84.01% of graded locations, grades 1 and 2 for about 11.61% and 3.97%, and grade 3 for only 0.41%.

3.2 Model Architecture

All experiments used the same convolutional architecture described in Section 2, kept fixed across the binary baseline and all mixed-supervision ordinal models so that performance differences reflect the supervision setting rather than model design. The model has 4.33 million trainable parameters, of which 4.16 million (96.1%) belong to the shared backbone. The location-specific components are comparatively small: a spatial attention module (33.7k parameters), the four severity heads (134.1k parameters in total), and the ordered cutpoints (12 parameters). The architecture therefore relies mainly on shared feature learning, with lightweight location-specific modules for anatomical adaptation and ordinal prediction.

3.3 Annotation Budgets and Sampling Strategies

The mixed-supervision experiments used fixed graded annotation budgets of

$$64,\ 128,\ 256,\ 512,\ 1024,\ 2048,\ 4096,\ 8192,$$

together with a full-supervision setting using all 12,660 graded training samples; the zero-budget setting is the binary-only baseline. Each budget defines a separate experiment: the graded subset is selected once and held fixed during training rather than being grown during a run. The resulting curve over budgets is therefore an evaluation design, not a training curriculum in which graded samples are gradually added. The budgets are nested, so each larger budget contains all samples from the smaller ones.

The two selection strategies are those defined in Section 2. For score-stratified selection, binary-positive training samples were ranked by the binary-model severity proxy and divided into five quantile bins. Samples were drawn in a round-robin manner across these bins, giving approximately even coverage of the score distribution, while reserving about 15% of each budget for binary-negative calibration cases.

3.4 Training and Model Selection

The binary model was trained first using all available binary osteophyte-presence labels. This binary model served two roles. First, it provided the zero-budget baseline, allowing us to evaluate what can be learned from binary presence labels alone. Second, it provided the scoring function used by the score-stratified sampling strategy. The mixed-supervision loss could in principle be trained from scratch, so this study does not isolate the independent effect of binary pretraining. Instead, the binary stage is part of the experimental design needed to compare binary-only, random mixed supervision, and score-stratified mixed supervision under the same framework.

For each annotation budget and sampling strategy, mixed-supervision ordinal models were trained using three random seeds, each initialized from the binary pretrained model and trained with the masked ordinal loss of Section 2. Optimisation used AdamW [12, 13], mixed precision, gradient clipping, and a cosine learning-rate schedule. The same preprocessing, architecture, loss formulation, and evaluation pipeline were kept fixed across budgets and sampling strategies.

The validation set was used for early stopping and model selection. After these choices were fixed, the held-out test set was used for final evaluation. Unless stated otherwise, all results in Section 4 are reported on the test set as mean ± standard deviation over three seeds.

3.5 Evaluation Metrics

The primary metric is macro-averaged mean absolute error (macro-MAE), following the macro-averaged error measures proposed for ordinal regression under class imbalance by Baccianella et al. [14]. For each anatomical location, the absolute error is first averaged separately within each true OARSI grade and then averaged across grades:

$$\mathrm{MAE}_l^{\mathrm{macro}} = \frac{1}{4}\sum_{g=0}^{3}\frac{1}{N_{l,g}}\sum_{i:y_{i,l}=g}|\hat{y}_{i,l}-y_{i,l}|.$$

This differs from ordinary mean absolute error over the full sample. Ordinary MAE would weight grades in proportion to how often they occur, so the abundant grade-0 cases would dominate the metric. The nested sums avoid this by first computing the error within each grade and then giving the four grades equal weight. This is important for the present dataset because the higher OARSI grades, especially grade 3, are rare.

The final macro-MAE is obtained by averaging over the four anatomical locations:

$$\mathrm{MAE}^{\mathrm{macro}} = \frac{1}{4}\sum_{l}\mathrm{MAE}_l^{\mathrm{macro}}.$$

Since OARSI grades range from 0 to 3, we also report

$$\mathrm{Quality} = 1 - \frac{\mathrm{MAE}^{\mathrm{macro}}}{3},$$

where higher values indicate better severity estimation.

To analyse ordinal severity boundaries, we report AUROC and average precision (AP) for $y \geq 1$, $y \geq 2$, and $y \geq 3$ [15, 16]. For each threshold, AP evaluates how well the model ranks true positive cases above negative cases using the predicted cumulative probability $P(y \geq k)$;

it does not require assigning a hard final grade. These thresholds correspond to osteophyte presence, moderate-or-worse osteophytes, and severe osteophytes. AP is especially relevant for $y \geq 3$ because severe osteophytes are rare and AP remains sensitive under strong class imbalance. We do not rely on hard-grade class assignment, since collapsing the continuous expected severity into a single class discards the ordinal information the model is trained to capture.

4 Results

4.1 Mixed Supervision Improves Severity Estimation

Adding a small graded subset enabled the model to produce ordinal severity estimates rather than only a presence score, with most of the macro-MAE improvement appearing at the smallest annotation budgets. The binary-only baseline provides the starting point for this comparison. Trained only on presence labels, corresponding to the first threshold $P(y \geq 1)$, it reached a macro-averaged MAE of 1.005 ± 0.020 and a presence AUROC of 0.824 ± 0.009 on the test set. Its mean predicted presence probability increased with true grade, from 0.283 ± 0.016 for grade 0 to 0.914 ± 0.027 for grade 3, indicating that the binary score contains severity-associated information. However, because this model is never trained to separate grades 1, 2, and 3, the score remains a proxy for severity rather than a true ordinal estimate.

Macro-MAE should be interpreted carefully because the binary-only baseline predicts only a presence score and is therefore structurally limited to values in $[0, 1]$. This makes it a useful reference for what binary supervision alone can express, but not a fully fair ordinal baseline for grades 0–3. Within the ordinal models, 64 graded samples reached a macro-MAE of 0.688 ± 0.024, compared with 0.602 ± 0.037 under full supervision (Table 1; Figure 1). The remaining gap from 64 samples to full supervision was therefore 0.086 macro-MAE, while larger budgets approached the full-supervision result with diminishing returns. Small non-monotonic differences between adjacent budgets, such as 4096 versus 8192 samples, fall within seed variability.

Because macro-MAE weights all grades equally, these improvements are especially relevant for the rarer higher grades. The larger partial budgets already contained much of the available high-grade supervision: the 4096- and 8192-sample score-stratified subsets included roughly 1,190 and 1,370 grade-≥ 2 labels, and about 130 and 150 grade-≥ 3 labels, respectively. This helps explain why the largest partial budgets approached the full-supervision result.

Table 1: Score-stratified test performance across graded annotation budgets (mean ± SD over three seeds). The binary model is the zero-budget baseline. Lower macro-MAE is better; higher quality and AP are better. For the binary baseline, AP for $y \geq 2$ and $y \geq 3$ uses the presence score as a severity proxy.

Budget	Fraction	Macro-MAE	Quality	AP $y \geq 2$	AP $y \geq 3$
Binary	0.000	1.005 ± 0.020	0.665 ± 0.007	0.241 ± 0.011	0.075 ± 0.026
64	0.005	0.688 ± 0.024	0.771 ± 0.008	0.231 ± 0.035	0.072 ± 0.033
128	0.010	0.665 ± 0.027	0.778 ± 0.009	0.229 ± 0.012	0.092 ± 0.045
256	0.020	0.704 ± 0.088	0.765 ± 0.029	0.224 ± 0.012	0.089 ± 0.042
512	0.040	0.654 ± 0.018	0.782 ± 0.006	0.237 ± 0.013	0.122 ± 0.067
1024	0.081	0.673 ± 0.060	0.776 ± 0.020	0.232 ± 0.013	0.089 ± 0.040
2048	0.162	0.639 ± 0.031	0.787 ± 0.010	0.245 ± 0.029	0.095 ± 0.036
4096	0.324	0.613 ± 0.012	0.796 ± 0.004	0.260 ± 0.016	0.150 ± 0.056
8192	0.647	0.616 ± 0.033	0.795 ± 0.011	0.269 ± 0.019	0.164 ± 0.071
Full	1.000	0.602 ± 0.037	0.799 ± 0.012	0.266 ± 0.016	0.136 ± 0.060

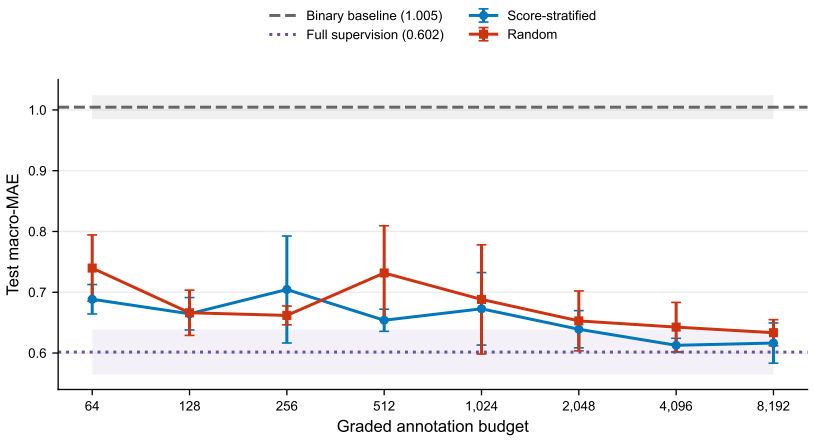

Figure 1: Test macro-MAE across graded annotation budgets for random and score-stratified selection. Error bars show standard deviation over three seeds. Horizontal reference lines show the binary baseline and the full-supervision model.

The same pattern is visible in how the predicted scores track the true grade (Figure 2). The binary presence score rises with grade but saturates and overlaps across the positive grades. The ordinal full model instead produces an expected severity that increases monotonically with grade, with mean expected severities of 0.469, 1.226, 1.732, and 2.623 for grades 0–3. Grade 0 separates clearly from the positive grades, while neighbouring positive grades, especially 1 and 2, remain closer together. Figure 2 therefore supports the main interpretation from macro-MAE: binary supervision learns a useful presence signal, but graded supervision is needed to estimate ordinal severity.

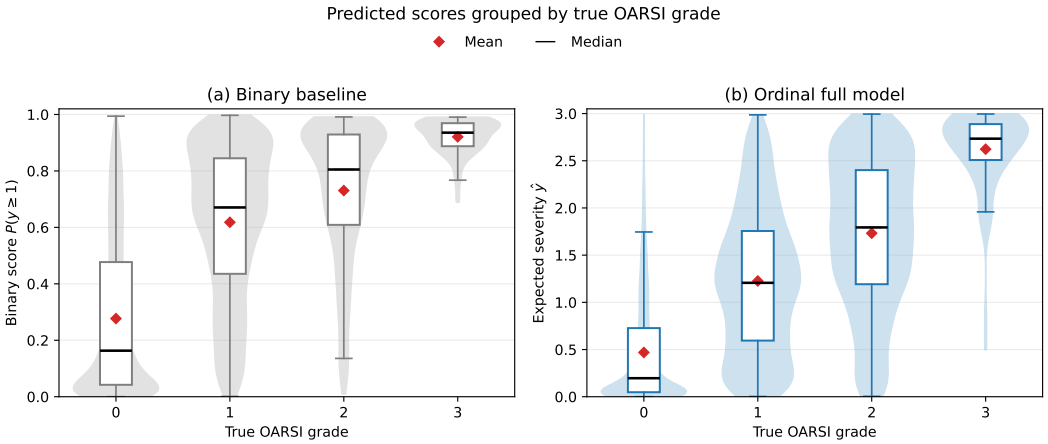

Figure 2: Predicted scores grouped by true OARSI grade on the test set over three seeds. Violins show the score distribution, boxes the interquartile range, red diamonds the mean, and black lines the median.

4.2 Severity Thresholds Differ in Difficulty

The three severity thresholds showed different levels of difficulty (Figure 3). Performance for $y \geq 1$ was high and changed little across budgets, as expected because this threshold is directly supported by binary labels. The higher thresholds depended more strongly on graded supervision and therefore better reflect the added value of mixed supervision.

The higher-threshold AP results were much less pronounced than the macro-MAE improvements. For score-stratified models, AP for $y \geq 2$ changed only from 0.231 ± 0.035 with 64 graded samples to 0.269 ± 0.019 with 8192 graded samples, compared with 0.266 ± 0.016 for full supervision. Several partial-budget models were below the binary presence-score proxy on this metric. AP for $y \geq 3$ was lower and more variable, reaching 0.164 ± 0.071 at 8192 samples compared with 0.136 ± 0.060 for full supervision. These results indicate that graded supervision mainly improved the scale of the expected-severity estimates, while gains in ranking moderate-or-worse and severe cases were modest. AUROC was high for the higher thresholds in the full model (0.863 ± 0.005 for $y \geq 2$ and 0.965 ± 0.018 for $y \geq 3$), but AP gives a more conservative view under strong class imbalance.

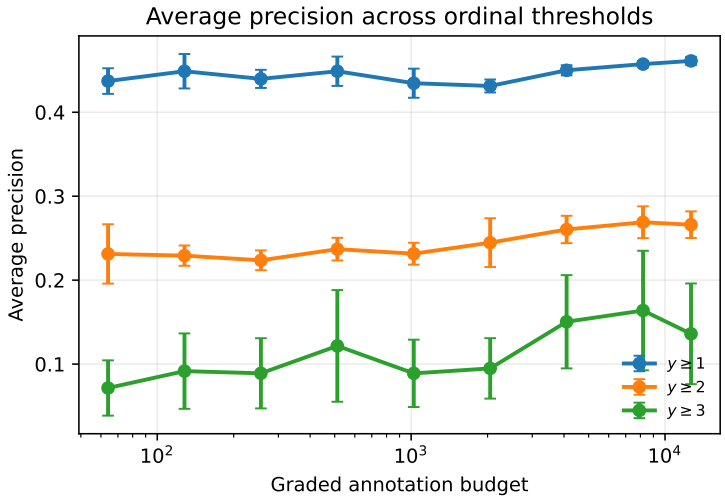

Figure 3: Average precision across ordinal thresholds and graded annotation budgets. Error bars show standard deviation over three seeds.

4.3 Sampling Strategy Has a Smaller Effect than Graded Supervision

The choice of which samples to grade had a smaller effect than the amount of graded supervision. In Figure 1, score-stratified sampling usually achieved lower mean macro-MAE than random sampling, but the advantage was not uniform. Random sampling was better at the 256-sample budget, and the two strategies were nearly identical at 128 samples.

The main conclusion is therefore that adding graded labels matters more than the specific selection strategy used here. Score-stratified sampling can give a modest advantage, especially at some smaller budgets, but it also requires an additional binary-model scoring step. Random sampling is simpler and remained competitive across the tested budgets.

4.4 Performance Differs by Anatomical Location

Performance varied across anatomical sites (Table 2). The superior femoral location was easiest, with the lowest macro-MAE and the strongest high-threshold AP, while the inferior acetabular location was hardest. The averaged results should therefore be read as aggregate performance across four related but unequally difficult prediction problems.

This location-level variation is important for interpreting the overall results. The mixed-supervision model improves severity estimation under limited graded supervision, but its reliability is not uniform across anatomical sites. Results for severe osteophytes should also remain cautious because grade 3 is rare, especially when performance is broken down by location.

Table 2: Location-specific test performance for the full model (mean ± SD over three seeds). Lower macro-MAE is better; higher AP is better.

Location	Macro-MAE	AP $y \geq 2$	AP $y \geq 3$
Inferior acetabular	0.743 ± 0.112	0.120 ± 0.024	0.030 ± 0.034
Inferior femoral	0.553 ± 0.051	0.242 ± 0.014	0.057 ± 0.022
Superior acetabular	0.593 ± 0.026	0.252 ± 0.037	0.108 ± 0.063
Superior femoral	0.518 ± 0.006	0.451 ± 0.024	0.349 ± 0.131

5 Responsible Research

5.1 Ethical Considerations

This study uses medical imaging data, which requires careful handling because radiographs and associated metadata may contain sensitive patient information. The dataset, provided for this research project, combines data from the CHECK and OAI osteoarthritis studies. It was kept in a secure research storage environment on the DelftBlue cluster throughout the project and was not downloaded to personal devices or redistributed. The study does not attempt to identify individuals, and all experiments use image crops and structured metadata needed only for training and evaluation.

The model is developed for research-oriented severity estimation, not clinical deployment, and the results should not be read as evidence of clinical readiness. Automated severity estimation can support large-scale research by reducing annotation burden, but clinical use would require further validation, prospective evaluation, calibration, and expert review. Because incorrect predictions could be harmful if used directly for diagnosis or treatment, the model should be considered a research or decision-support tool rather than an autonomous diagnostic system.

5.2 Dataset Bias and Generalisation

Although the combined CHECK and OAI data provide many samples, they may not represent all patient populations, imaging protocols, or clinical settings. Differences in scanner type, acquisition, patient positioning, and demographic composition can all affect performance, so the model may not generalise equally to other hospitals or populations.

Class imbalance is a further limitation: grade 0 dominates and grade 3 is rare, which affects both training and evaluation, since a model can score well overall while performing poorly on rare grades. We mitigate this in reporting by using macro-MAE and threshold-level metrics (AP and AUROC) rather than ordinary accuracy, but grade-3 results should still be interpreted cautiously. Detection difficulty also varies by anatomical location, so averaged metrics can hide location-specific differences; we therefore store and report location-specific results. For this reason, the reported results should be interpreted as evidence about annotation efficiency within the CHECK/OAI experimental setting, not as evidence that the model is ready for clinical use or will generalise unchanged to other imaging sources.

5.3 Limitations of Weak Supervision

A binary-positive label indicates only that an osteophyte is present, not whether it is small, medium, or large, so binary-positive samples cannot supervise the $y \geq 2$ and $y \geq 3$ thresholds. As a result, learning to distinguish positive higher-severity cases depends strongly on

the comparatively small graded subset, and the quality of severity estimation is sensitive to which samples that subset contains.

In this simulation, binary labels are noiseless because they are derived from the available OARSI grades, but real independently collected binary labels could be noisy or inconsistent with detailed grades. To reduce this risk we audited the dataset before training with four checks: a split-size check, verifying that the train, validation, and test counts matched the dataset index; a subject-leakage check, verifying that no subject appeared in more than one split; a missing-label check, verifying that samples without a grade were correctly flagged so they could be masked rather than silently treated as grade 0; and a binary–graded consistency check, verifying that every binary-negative location had grade 0 and every binary-positive location had grade at least 1. Finally, score-stratified selection may overrepresent cases the binary model scores confidently and underrepresent unusual ones; the budgets include lower-score samples for calibration, but future work should compare against uncertainty-based and oracle sampling.

5.4 Reproducibility

The dataset index stores image identifiers, splits, binary and graded labels, and missing-grade indicators. The train, validation, and test splits are fixed at the subject level, and graded budgets are saved as nested CSV files so that each larger budget contains the smaller one. The pipeline saves checkpoints, configurations, metrics, and predictions, allowing results and figures to be regenerated without retraining, and fixed random seeds are used for pretraining and budget construction.

Model training and experiment execution were performed using computational resources of the DelftBlue supercomputer provided by the Delft High Performance Computing Centre [17].

Because the project data contain a large grade-complete subset, partial annotation is simulated by exposing OARSI grades only for selected training-budget samples. All other training grades are set to missing before loss and class-weight computation. Validation grades are used for early stopping and model selection, while test grades are held out for final evaluation only. This is a correctness requirement of the experimental design rather than a property of the method itself. If hidden grades reached the loss, the partial-budget models would effectively receive more supervision than their budget allows, and the annotation-efficiency results would be optimistically biased. In a real deployment the issue does not arise, because ungraded samples genuinely have no grade to expose; here it is guarded against in code by masking grades outside the selected budget and by the consistency checks above.

5.5 Use of AI Assistance

AI tools were used during this project as support tools for brainstorming, rephrasing ideas, checking the clarity of explanations, validating implementation decisions, and assisting with coding. These tools were not used to make scientific decisions automatically or to generate experimental results. All methodological choices, code changes, result interpretations, and final manuscript content were reviewed and accepted by the author. The use of AI assistance therefore supported the research and writing process, while responsibility for the final work remains with the author.

6 Discussion and Conclusions

This study investigated whether abundant binary osteophyte-presence labels can be combined with a limited number of OARSI graded labels for ordinal severity estimation in hip X-ray crops. The results show that binary supervision is a useful but incomplete starting point: the binary-only model detects osteophyte presence and its scores increase with true grade, but it should be interpreted as a severity-associated proxy rather than as a grading model. Adding graded labels substantially improves ordinal severity estimation, especially at the smallest annotation budgets, with diminishing returns as the graded subset grows. The practical implication is that full expert grading of every training image may not be necessary to obtain useful aggregate severity estimates.

The main methodological contribution is the masked cumulative ordinal formulation. OARSI grades are ordered, so modelling them through cumulative thresholds is more natural than treating the grades as unrelated classes, although this study did not test a standard four-class classifier directly. The formulation also fits the mixed-label setting: binary labels constrain the presence threshold, while graded labels provide information about the higher severity thresholds. By masking unavailable targets, each sample contributes only the label information it actually supports.

The sampling results are more cautious. Score-stratified selection was competitive with random selection, but it was not consistently better. Most of the benefit came from adding graded labels at all, rather than from the specific choice of which samples were graded. This matters in practice because score-stratified selection requires first training a binary model and scoring the training set. If binary labels and a binary model are already available, this extra step may be reasonable; if expert grading can be assigned directly, random selection may be simpler and nearly as effective.

Several limitations remain. All experiments used one fixed convolutional architecture, so the effect of the supervision strategy was isolated at the cost of leaving absolute performance dependent on that architecture. The two-stage procedure was not compared with mixed-supervision training from scratch, so the independent effect of binary pretraining is not known. The sampling comparison was limited to random and score-stratified selection; uncertainty-based, diversity-based, or oracle grade-stratified strategies may behave differently. Finally, the fixed hip crops from CHECK and OAI may remove broader anatomical context and may not generalise to other hospitals, imaging protocols, or patient populations. Results for grade 3 should also remain cautious because severe osteophytes are rare.

Future work should evaluate more random seeds and repeated subset selections, compare score-stratified sampling with uncertainty-based and diversity-based strategies, and test whether two-stage training improves over mixed-supervision training from scratch. Further improvements may come from location-specific modelling, calibration, stronger loss weighting or oversampling for rare high-grade cases, or additional severe examples. Overall, combining many binary labels with relatively few graded labels is a promising way to reduce expert annotation burden in large-scale osteoarthritis imaging studies.

References

[1] David J. Hunter and Sita Bierma-Zeinstra. Osteoarthritis. *The Lancet*, 393(10182):1745–1759, 2019.

[2] R. D. Altman and G. E. Gold. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and Cartilage*, 15:A1–A56, 2007.

[3] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[4] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2018.

[5] L. Misera, G. Müller-Franzes, D. Truhn, and J. N. Kather. Weakly supervised deep learning in radiology. *Radiology*, 312(1):e232085, 2024.

[6] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. A neural network approach to ordinal regression. In *2008 IEEE International Joint Conference on Neural Networks*, pages 1279–1284, 2008.

[7] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–142, 1980.

[8] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4920–4928, 2016.

[9] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[10] Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[11] 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 (CVPR)*, pages 770–778, 2016.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[14] Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Evaluation measures for ordinal regression. In *2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 283–287, 2009.

[15] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[16] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

[17] Delft High Performance Computing Centre (DHPC). *DelftBlue Supercomputer (Phase 2)*, 2024. `https://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2`.