Optimizing labeling

The limits of weakly supervised osteophytes severity grading and localization in Hip X-Rays

Alvin Ye[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: Alvin Ye
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Julia Olkhovskaya

Abstract

Osteophytes are bony protrusions that are key radiographic indicators of hip osteoarthritis (OA), but grading their severity in specific hip locations is a time consuming process that requires an expert. In many cases it is expensive to scale datasets with location annotated severity labelling by experts, where as weak labels, containing only the global presence of osteophytes is much easier to attain. This paper investigates whether such weak global label can improve localized severity grading through a multitask deep learning framework.

We study a ResNet-18 based convolutional network that shares and updates its weights across two output heads, a global binary classification head and four regional ordinal heads for femur superior, femur inferior, acetabulum superior and acetabulum inferior. The model is trained under four supervision strategies: a strong-only configuration using only quadrant-level labels, a masked baseline that incorporates weakly labelled negatives via label propagation and ignores weak positives in the local loss, and two Multi-Instance Learning variants that use a Noisy-OR loss to propagate weak positive labels to the quadrants. We systematically vary the ratio of weak to strong labels and evaluate performance using quadratic weighted Cohen's kappa as the primary metric.

Experiments show that the masked baseline with weak labels improves regional kappa score compared to the strong-only configuration, while MIL variants fail to outperform the baseline and can degrade performance at higher weak-to-strong ratios. We further observe that selecting checkpoints by minimal joint validation loss underestimates achievable kappa score, due to faster convergence of the global task, whereas selecting by maximal kappa score yields substantially better localized grading. Overall the findings highlight the trade off between localization and classification performance in weakly supervised multitask learning pipelines for regional osteophytes grading in hip X-Rays.

1 Introduction

Osteoarthritis (OA) is a frequently occurring disease that develops osteophytes, which are bone protrusions on the joints. In 2020, about 595 million people globally had a form of osteoarthritis, with predictions till 2050 expecting an increase [1]. While there are multiple skeletal locations where osteophytes can manifest, this study focusses specifically on the hip area. In medical images, evaluating hip osteoarthritis requires more than a simple binary diagnosis of disease presence. Clinicians rely on localized severity grading across distinct structural regions of the hip joint around the femoral head. More specifically, the superior and inferior quadrant of the acetabular and femoral head. Grading the multi-class severity of osteophytes within these specific quadrants is key for diagnosis and management, however annotating the location and severity is a time-consuming and expensive process that requires an expert.

Consequently, medical image analysis and specifically deep learning models often face a trade-off between the abundance of coarse, global annotations and the scarcity of expensive, region-specific expertise. While clinical datasets often contain expert annotated severity grading, scaling these datasets becomes expensive due to intensive labour needed to grade new images. This paper investigates a specific weakly supervised application: whether a pool of easily acquired, 'weak' global labels (indicating the presence of osteophytes) can be leveraged to improve localized, multi-class severity grading when paired with a limited set of 'strong' region-specific data. For the rest of this paper we will denote these two data classes respectively as 'weak' and 'strong'.

Weak labels only contain whether the image have osteophytes or not, thus information about the grade of severity is missing. When we incorporate this in a pipeline, it is hard for the model to learn severity grading thus providing noise in gradients. In the paper we will evaluate the weak-strong framework on hip X-ray images for osteophytes grading, exploring the balance between maximizing label volume and managing the noisy information acquired from weak supervision.

Previously, it has been demonstrated that machine learning models like CNNs can accurately determine osteoarthritis with high accuracy [2]. Furthermore, semi-supervised frameworks have proven beneficial in mitigating annotation scarcity in general radiology [3]. Although weakly supervised models could offer a potential solution to data scarcity [4], there remains a knowledge gap in applying these methods to localized hip severity grading.

To address this, the main question we want to answer is whether utilising the weak labels by adding them to a strong dataset can improve classification performance and internal model representation for a multi task pipeline with a joint loss function. Specifically, we are interested in whether the integration of weak data alters regional severity grading performance compared to a strong data only baseline. Furthermore, we are interested in identifying whether a saturation threshold exists beyond which adding more weak global labels ceases to benefit severity grading accuracy due to increased noise. Finally we are interested in whether this weak to strong data ratio shifts the network focus, evaluating if and to what extent it increases or decreases the area of focus when analysed through spatial attention methods like Grad-CAM.

The main contributions of this paper are the following:

1. An empirical validation of a multi task loss function that balances strong and weak data, specifically identifying the threshold at which the addition of weak data ceases to improve severity classification.

2. A visual analysis using Grad-CAM spatial attention maps to demonstrate how different weak supervision strategies affect the area of focus. Showing a trade off in grading accuracy and localisation.

The remainder of this paper is structured as follows: Section 2 provides the background and a more detailed explanation of the problem, as well as our methodology, outlining the specifics of the loss function we are considering. Section 3 describes the experimental setup and dataset-preprocessing. Section 4 displays results and visualizes the spatial attention maps. Section 5 reflects on the responsible research and reproducibility aspects of this study. Finally, Sections 6 and 7 provide a discussion of the findings, conclude the research, and outline potential areas for future work.

2 Methodology and Problem Description

This section formally defines the computational constraints of mixing weak and strong labels, defines our architectural choices and concludes with the hypotheses of the study.

2.1 Problem Description

2.1.1 The problem and knowledge gap

As established in section 1, the problem is not that there is not enough data, the problem lies in asymmetry in the quality of data, where strong data contains much more information than weak data. To be precise, there are two types of data that we can define:

1. **Strongly Labeled Subset (D_s)**: These images contain multi class severity grades in four areas of the hip (acetabular inferior, acetabular superior, femoral inferior, femoral superior). The severity grades can be represented as a vector $\mathbf{y} = [y_1, y_2, y_3, y_4]$ where each location of the hip $y_q \in \{0, 1, 2, 3\}$ corresponds to the OARSI severity grade [5].
2. **Weakly Labeled Subset (D_w)**: These images contain only a single global binary image label $Y^w \in \{0, 1\}$ that indicates the presence of an osteophyte within the image.

The dataset composition is then quantified by the weak-to-strong ratio ρ:

$$\rho = \frac{|D_w|}{|D_s|} \tag{1}$$

Under this formulation, $\rho = 0$ represents the strong-only baseline.

2.1.2 Architecture and two headed pipeline

To evaluate the interaction between varying label types, we design a multi-task convolutional learning framework utilizing a shared backbone. The network takes an image and produces a spatial feature map, which it routes into two concurrent prediction heads: a global binary classification branch tracking joint-wide osteoarthritis markers, and four independent localized ordinal branches designed to grade osteophyte severity across specific anatomical hip quadrants (femur superior, femur inferior, acetabulum superior, and acetabulum inferior), as schematically illustrated in Figure 1. Both heads operate on the same spatial feature map each producing gradients specific to its task, which jointly updates the shared backbone. By routing training images dynamically through this shared architecture based on whether they are weak or strong labels, the pipeline optimizes a joint multi-task loss function.

2.2 Methodology and Loss Functions

To isolate the impact of label ratios, we implement a modular pipeline centered around the multi-headed backbone optimization mentioned in 2.1.2 with a custom loss function. To train our network simultaneously on both global and regional features, we optimize a joint multi-task objective function, $\mathcal{L}_{\text{total}}$. This total loss is simply the sum of our global disease detection loss and the four individual localized regional grading losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \sum_{q=1}^{4} \mathcal{L}_{\text{ordinal}}^{(q)} \tag{2}$$

Where $q \in \{1, 2, 3, 4\}$ represents the index of the four distinct spatial quadrants of the hip joint. Specifically, each value of q is directly related to one individual regional grade output head: q=1: Femur Superior q=2: Femur Inferior q=3: Acetabulum Superior q=4: Acetabulum Inferior

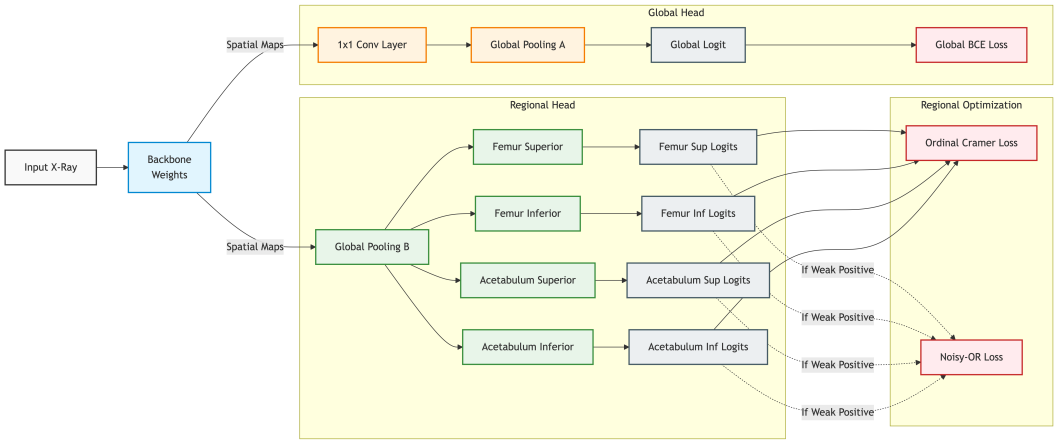

Figure 1: Network architecture, with shared weights and two heads, the global head optimizing healthy vs unhealthy using Binary Cross Entropy and a regional head optimizing for localized severity .

2.2.1 Global Detection Loss

The global binary classification head determines the overall presence of osteoarthritis across the entire hip joint. It is optimized using standard Binary Cross-Entropy (BCE) over a batch of N_g samples:

$$\mathcal{L}_{\text{global}} = -\frac{1}{N_g}\sum_{j=1}^{N_g}\left[Y_{g,j}\log(\sigma(\hat{Y}_{g,j})) + (1-Y_{g,j})\log(1-\sigma(\hat{Y}_{g,j}))\right] \tag{3}$$

where $Y_{g,j} \in \{0,1\}$ represents the ground-truth global status (0 for healthy, 1 for diseased), $\hat{Y}_{g,j}$ denotes the predicted raw logit, and $\sigma(\hat{Y}_{g,j})$ is the sigmoid mapping function.

2.2.2 Localized Regional Ordinal Loss via Cramer Distance

While the global head pipeline can output binary labels, evaluating individual hip quadrants requires transitioning from binary detection into a more complex severity grading pipeline. Standard classification loss functions treat a misrepresentation equally. This means misclassifying a normal joint (Grade 0) as severe osteophytes (Grade 3) would carry the exact same penalty as misclassifying it as doubtful (Grade 1).

The following method solves the equal penalty issue by comparing the **Cumulative Distribution Functions (CDFs)** of our predicted probabilities (P_c) and the true expert labels (Y_c), which mathematically corresponds to the Cramer distance. [6] This method here is preferred above standard cross entropy given that the distance is quadratically squared, meaning a predicting a Grade 0 for a Grade 3 has a higher penalty than 1 for a Grade 3.

First, we convert standard probabilities into cumulative probabilities up to a specific severity grade c:

$$P_c = \sum_{k=0}^{c} p_k \quad \text{and} \quad Y_c = \sum_{k=0}^{\mathbf{c}} y_{\text{one\_hot},k} \tag{4}$$

Here, p_k is the model's predicted probability for grade k, and $y_{\text{one\_hot},k}$ is the probability an expert gave to a certain grade. Note that the sum of these probabilities is one.

Given multiple samples will be combined into one batch, for a batch of N_o valid samples, the ordinal loss ($\mathcal{L}_{\text{ordinal}}$) is computed as the average sum of squared differences between these two sums per step:

$$\mathcal{L}_{\text{ordinal}} = \frac{1}{N_o} \sum_{j=1}^{N_o} \sum_{c=0}^{K-2} (P_{j,c} - Y_{j,c})^2 \tag{5}$$

Here, $K = 4$ represents our four clinical severity stages (Grades 0, 1, 2, and 3). The inner loop stops at $K - 2$ because the cumulative probabilities at the final grade ($K - 1$) are guaranteed to always equal 1, making their difference zero.

By calculating the difference in the cumulative space, the loss automatically captures the distance between classes. Squaring these differences ensures that wider misclassifications (e.g., predicting Grade 0 instead of Grade 3) result in a significantly higher gradient penalty than near-misses.

2.2.3 Weak Supervision Variants

To optimize our shared network simultaneously across distinct labels, samples are dynamically routed through different supervisory rules which define our control groups:

1. **The Lower Bound (Strong-Only Configuration)**: This configuration represents the baseline performance floor, utilizing exclusively the strongly labeled subset (D_s) where $\rho = 0$ in (1). The local ordinal heads are supervised solely by expert-annotated, quadrant-specific severity grades utilizing the Cramer distance defined in (5). This ensures the model learns foundational fine-grained feature extractors without the potential regularization or noise introduced by coarse-grained global labels. However, this lower bound suffers from data scarcity as the model is limited to only 1000 images, increasing the risk of overfitting and bounding its generalizability.

2. **The Masked Baseline (Naive Weak Supervision)**: The Masked Baseline incorporates the weakly labeled subset (D_w) but handles weak positives and negatives asymmetrically. For weak negatives ($Y_g = 0$), absolute label propagation is applied, actively penalizing the local heads to enforce a healthy state ($y = [0, 0, 0, 0]$) via (5). However, for weak positives ($Y_g = 1$), the local ordinal loss is entirely masked out, ignoring them in the local loss computation. Consequently, weak positives only contribute gradients to the shared backbone via the global Binary Cross-Entropy (BCE) head. This configuration acts as a critical ablation study to determine if global supervision alone provides sufficient regularization for the shared feature extractor to improve fine-grained regional grading.

3. **The Dual-Head MIL (Noisy-OR Positive Propagation)**: In contrast to the Masked Baseline, the Dual-Head MIL actively attempts to propagate the coarse weak positive labels ($Y_g = 1$) to the specific spatial quadrants utilizing a Multi-Instance Learning Noisy-OR loss configuration. We hypothesize the global disease presence as a Noisy-OR probabilistic function of the localized quadrant health predictions. Let $p_0^{(q)}$ be the model's predicted probability that quadrant q is free of disease (Grade 0). The joint probability that the scan is globally diseased is formulated as:

$$P(\text{diseased}) = 1 - \prod_{q=1}^{4} p_0^{(q)} \tag{6}$$

We directly minimize the negative log-likelihood of this global prediction for weakly positive samples:

$$\mathcal{L}_{\text{weak\_positive}} = -\log\left(1 - \prod_{q=1}^{4} p_0^{(q)} + \epsilon\right) \tag{7}$$

where $\epsilon = 1 \times 10^{-7}$ ensures numeric stability. This enables the model to update localized weights without explicit region labels. Under this architecture, we train multiple ratios on this combined loss function and evaluate them.

4. **The Pure MIL Configuration**: This strategy acts as an architectural ablation test. It completely deactivates the standard global binary classification loss $\mathcal{L}_{\text{global}}$, forcing all weak global labels to propagate gradients exclusively through the Noisy-OR loss function formalized in (7).

5. **The Oracle (Upper Bound)**: To establish the theoretical ceiling of the architecture under optimal conditions, the Oracle configuration treats the entire dataset as if it were 100% strongly annotated acting as a situation with no annotation scarcity. This allows us to quantify the exact performance gap between our weakly supervised approximations and the theoretical ceiling.

2.2.4 Hypothesis and Research Objectives

Thus with all our methods defined, the study considers three hypotheses:

- **Hypothesis 1**: Introducing weak data D_w in a joint multitask pipeline improves the models ability to accurately predict location and severity compared to a strong baseline.

- **Hypothesis 2**: There exists an optimal saturation threshold $\rho_{\max}$, beyond which injecting more weak global labels skews the model towards coarse global features, thus degrading the performance of the model in predicting fine grained severity grades.

- **Hypothesis 3**: As the weak-to-strong ratio ρ is increased, the spatial attention boundaries will contract, indicating a shift from the general scanning of features to a more refined localization profile.

3 Experimental Setup

This section outlines the empirical environment, data workflow and resources used to compute and predict severity labels for hip X-Ray images. For the experimental setup, we used the aggregated dataset of CHECK and OAI datasets which contain pelvic anatomy of patients with their respective visits. As for the shared backbone used we choose ResNet-18 [7] given the model is faster to train, and reaches desired prediction capability faster than other models like ResNet-50 which are more complex. The ResNet-18 uses a 512 channel 7x7 spatial feature map.

3.1 BoneFinder Preprocessing

To eliminate irrelevant pelvic anatomy and standardize the input space for the shared network backbone, images undergo an automated region of interest (ROI) extraction pipeline leveraging the external BoneFinder framework [8]. The algorithm automatically maps continuous landmark points along the primary structural boundaries of the proximal femur, centering specifically on the femoral head. Images are subsequently cropped around these generated coordinates to isolate the four clinical quadrants of interest: the superior and inferior margins of both the acetabulum and the femoral head. All cropped inputs are resized to a uniform resolution 224x224 to maintain consistent feature map spatial dimensions across the final convolutional layers.

3.2 Subject-Level Data Split

To ensure robust generalization and prevent model overfitting, dataset partitioning is executed strictly at the subject level using a unique `subject_id` identifier rather than a naive per-scan split. Medical imaging datasets frequently include bilateral scans and longitudinal follow-up visits from individual patients. Splitting images arbitrarily would distribute highly correlated anatomical baselines and physiological priors across both the training and validation sets, causing the network to memorize patient-specific characteristics rather than generalized markers of marginal osteophyte formation.

3.3 Data Definitions

Both CHECK and OAI contain strong images, thus we have created our own weak labels by processing all but 1000 images with the following rule: If any one of the labels > 0 then 1 else 0. The total data pool thus looks as the following:

- **Strongly Labeled Subset (D_s):** A base cohort of 1,000 images containing expert, quadrant-level multi-class severity grades.
- **Weakly Labeled Subset (D_w):** A pool of $\approx$14,000 images carrying only a single joint-wide binary presence label.

During evaluation, the baseline strongly supervised dataset remains locked, while the weak-to-strong ratio ρ scales up to the full dataset limits in ratio's 1:0.5, 1:1, 1:2.5, 1:5, 1:all to map performance trajectories under varying degrees of label asymmetry.

3.4 Distributed Computational Environment

Given the high throughput of matrix operations required to systematically train and evaluate dozens of distinct ratio combinations, all training routines are parallelized and deployed within a distributed environment. The pipeline utilizes high-performance A100 GPU (80GB) compute nodes on the DelftBlue High-Performance Computing (HPC) cluster [9].

To optimize algorithmic sustainability and minimize the environmental footprint of long-running cluster jobs, data loading pipelines were optimized to prevent starvation and disk I/O bottlenecks. Storing the preprocessed, cropped tensor representations directly within local node caches significantly reduced training runtimes, preventing the GPUs from consuming unnecessary idle energy while waiting on disk reads.

3.5 Training Protocol and Hyperparameters

The entire pipeline is implemented using the PyTorch Lightning framework, enforcing a strict deterministic configuration to support open science.

- **Deterministic Controls:** Pseudo-random number generators across Python, NumPy, and PyTorch are pinned to a static global seed of 123.
- **Data Augmentation:** To improve regularized feature extraction, geometric data augmentations are restricted to random rotations up to $\pm 10^{\circ}$.
- **Hyperparameters:** The learning rate was kept at 0.0001 for all experiments. All experiments ran for 50 epochs.
- **Batch Dynamics:** Batch sizes are kept entirely uniform at 32 across all experimental scales of ρ to prevent gradient variance from altering optimization dynamics between baseline and maximum data expansion models.
- **Model Selection:** Normally, the model is selected by the model with the maximum performance and minimal validation loss. In our case, we maximize the models mean Quadratic Weighted Kappa score [10] across the 4 hip locations.

3.6 Evaluation Metric

Standard classification accuracy is a flawed metric for imbalanced, ordinal medical datasets. Thus, the localized regional grading performance is primarily evaluated utilizing the Quadratic Weighted Kappa [10], which adjusts for by-chance agreements and penalises larger misclassification quadratically. QWK score ranges from 0 to 1, where 0 means completely random guesses, and 1 means perfect agreement. Since we want the highest severity prediction accuracy, we try to maximise the mean Kappa score.

4 Results

The models selected on the maximum Kappa weights from the validation set were evaluated on a fully unseen test split to avoid validation set biases. Table 1 shows the following: Oracle achieved a maximum mean Kappa of 0.8586, the Masked Baseline achieved 0.5836 in its peak performance, and the Dual-Head MIL improved from 0.4809 to 0.5058 using the full data set. This shows that weak labels can indeed improve performance and that

naive masking provides better regularization over MIL architectures, specifically for severity grading. The peak values were based on the validation set and held for the Masked Baseline and Pure MIL. However, the Dual-Head MIL reached its peak in the test set at 2500 added weak images compared to 1000 in the validation set, degrading in performance similarly to the validation set afterwards (see Appendix B). This anomaly is likely due to statistical variance with this specific weight distribution being a better fit to the test set samples.

Model	ρ **(Weak Samples)**	**Test Mean Kappa**
Oracle	100% Strong	0.8586
Masked Baseline	0	0.4809
Masked Baseline	ALL (∼15k) (Peak)	**0.5836**
Dual-Head MIL	0	0.4809
Dual-Head MIL	1000 (Val-Selected Peak)	0.5123
Dual-Head MIL	2500 (Test-Absolute Peak)	0.5577
Dual-Head MIL	ALL (∼15k)	0.5058
Pure MIL	0	0.4769
Pure MIL	500 (Peak)	0.5311
Pure MIL	ALL (∼15k)	0.4502

Table 1: Evaluation of localized ordinal severity grading (Quadratic Weighted Cohen's Kappa) under varying scales of weak supervision. Model weights were selected by maximizing validation Kappa to bypass the multi-task global loss bottleneck. The Masked Baseline scales consistently without degrading, outperforming the MIL variants which peak early and subsequently collapse due to noisy gradients.

4.1 Scaling Weak Supervision: Baseline versus MIL

To see how the models perform under different ratios of scarcity, we scaled the weak-to-strong ratio (ρ) from 0 (1,000 strong samples only) to the full dataset limits.

The Masked Baseline scaled exceptionally well, improving its mean Kappa on the test set from 0.4809 ($\rho = 0$) to a peak of 0.5836 when utilizing all available weak samples. In contrast, the Dual-Head MIL peaked prematurely on the validation set at 1000 weak samples and even the best test accuracy (at 2500 weak samples) failed to beat the Masked Baseline before degrading in performance to 0.5058. Finally, the Pure MIL configuration which entirely ignored the global BCE head experienced significant performance degradation, peaking at 0.5311 and dropping to 0.4502 on the full dataset as can be seen in Table 1. This performance loss is likely due to the information weak labels contain and the way Kappa score is calculated. Weak labels do not contain a severity grade limiting the information we can extract, thus introducing noisy gradients which could negatively impact severity prediction.

4.2 Spatial Attention Analysis (Grad-CAM)

To investigate the focus of the trained models, Gradient-weighted Class Activation Mapping (Grad-CAM) [11] was executed on the final convolutional layer of the ResNet-18 backbone. Figure 2 illustrates the visual attention fields generated for the specific *femur superior* classification head on a representative right hip scan. This image has only one osteophyte in the

image, of grade 3. The GradCAM images reveal what the model looks at when it is forced to show the grade 3 features it extracted.

The qualitative visualization reveals a split between severity grading and localization across different supervision strategies.

- **Baseline (0 Weak):** Generates a flat activation field. This model is trained on only 1000 strong images and categorizes the sample as healthy, thus not indicating any activation map.
- **Oracle (Upper Bound):** Shows a highly widespread pattern covering both the superior femoral boundary and the nearby acetabular area. This broad activation suggests that under full label supervision the model extracts highly contextual features that span the entire joint space showing reliance on secondary features rather than focusing on the quadrant it is predicting for.
- **Masked Baseline (ALL Weak):** While achieving the highest quantitative ordinal grading metrics (Test Mean Kappa of 0.5836), the spatial attention map reveals a similar anatomical shortcut to that of the Oracle model. The heatmap isolates the centralised acetabular and superior joint space rather than the localised superior femoral neck. This indicates that the global Binary Cross-Entropy regularizer guides the shared backbone to extract general markers of osteoarthritis across the joint, but fails to enforce region specificity on the individual classification head.
- **Dual-Head MIL (ALL Weak):** Exhibits the most anatomically precise localization profile. The heatmap tightly encompasses the true clinical site of pathology, the superior margin of the femoral head and neck junction. This provides visual confirmation that the Noisy-OR loss formulation successfully penalises spatial misalignments by routing positive disease gradients directly to the correct anatomical region of interest.
- **Pure MIL (ALL Weak):** Produces an overly constricted attention map centered completely in the interior femoral head, entirely missing the correct osteophyte boundaries. Without the anchoring gradient of the global BCE head, the Noisy-OR formulation misses the general features of osteophytes, focusing on a tight subset of pixels and explaining its poor generalizability.

In conclusion, while the Masked Baseline provides better ordinal severity classification due to global joint regularisation, the Dual-Head MIL shows greater anatomical localisation in our instance. This demonstrates a trade-off in weakly supervised medical imaging: Multi-Instance Learning results in better localisation at the cost of having its severity grading weakened due to gradient noise.

4.3 Model Selection

During the empirical evaluation, there was an optimisation conflict between the joint multi-task validation loss ($\mathcal{L}_{total}$) and the targeted localised performance. When evaluating model checkpoints strictly by the minimum joint validation loss, the reported Kappa scores performed worse by showing a lower mean Kappa on the validation set. For instance, the Oracle reference achieved a minimum validation loss at step 478 yielding a mean Kappa of 0.5110; however, by step 21075, the mean Kappa maximized at 0.6107.

This discrepancy arises from the difference in complexity of the distinct network heads. The global binary classification task (BCE loss) optimises rapidly and subsequently begins

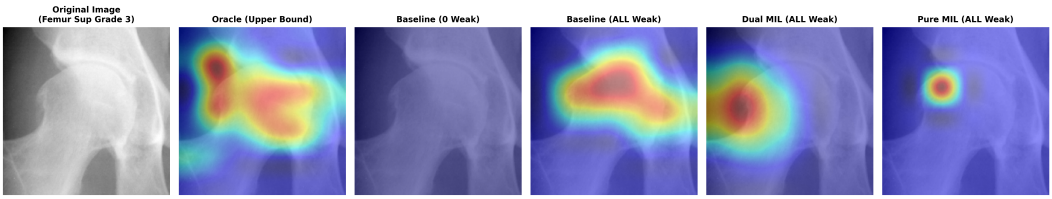

Figure 2: Grad-CAM heatmaps for the femur superior head on a representative diseased right hip scan (Subject: OAI-9883115, visit V00, actual femur superior osteophyte grade: 3, others: 0). The Oracle represents the upper-bound target attention. Comparing Dual MIL and Pure MIL against the Masked Baseline configurations demonstrates how incorporating weak supervision through Noisy-OR MIL formulations affects model localisation and spatial focus in regions of interest.

to overfit, artificially inflating the total validation loss early. At the same time, the more complex local ordinal heads require significantly more epochs to extract fine-grained severity features. Thus, relying on the joint loss for early stopping negatively affects the learning trajectory of the localized ordinal heads. To resolve this multi-task bottleneck, all subsequent model selections were performed by directly maximizing the `val/mean_kappa`, bypassing the negative effect of using validation loss as the model weights decider.

5 Responsible Research and Reproducibility

Using deep learning architectures within radiology demands strict following of ethical research practices, resource sustainability, and algorithmic transparency. This section reflects on the ethical implications of data handling, the carbon footprint of high-performance computing (HPC) workflows, and the technical safeguards implemented to ensure complete reproducibility.

5.1 Ethical Considerations and Clinical Safety

The utilization of medical imaging datasets inherently involves vulnerable patient data, necessitating strict privacy and clinical safety boundaries. While the dataset used in this study was fully anonymized prior to acquisition, preventing the extraction of protected health information (PHI), the algorithmic behavior introduces ethical considerations:

- **The Risk of Diagnostic Shortcuts:** As visually uncovered via Grad-CAM analysis in Figure 2, the Masked Baseline model achieves high grading accuracy by exploiting contextual joint-space markers rather than isolating the true location of the osteophyte. In a clinical deployment scenario, relying on such shortcuts introduces ethical risks, as the model may output a correct severity grade based on irrelevant or secondary radiographic artifacts, potentially ignoring co-occurring diseases.
- **Gradient Instability and False Negatives:** The optimization collapse observed in the Pure MIL and Dual-Head MIL configurations at high weak-to-strong ratios (ρ) shows a safety concern. Forcing a model to route gradients through an unstable Noisy-OR gate can increase false-negative rates in complex, multi area disease states.

In medical AI, a false negative weighs significantly more than a false positive, given that false negatives can equal late diagnosis and stop early prevention.

5.2 Algorithmic Sustainability and Environmental Footprint

Executing high-throughput deep learning pipelines across thousands of medical images introduces non-trivial environmental costs due to the carbon footprint of GPU compute infrastructure. Responsible engineering requires minimizing unnecessary compute cycles through pipeline profiling and optimization on shared academic infrastructure, such as the DelftBlue HPC cluster.
Initially our batch runs suffered from severe CPU starvation and disk I/O bottlenecks, causing the GPU's to stall and consume idle energy while waiting on the images. By modifying our dataset and storing it in cache, the runtime significantly dropped, reducing energy usage.

5.3 Reproducibility Protocol and Open Science

To ensure that the empirical trajectories, baseline configurations, and results reported in Table 1 can be independently verified, we enforce strict deterministic controls across our PyTorch Lightning codebase:

1. **Deterministic Seeding:** Pseudo-random number generators across Python, NumPy, and PyTorch were pinned to a static global seed (123). This guarantees that geometric data augmentations (e.g., random rotations up to $\pm 10^{\circ}$) and random weight initializations remain identical across replicated runs.

2. **Data Partitioning Auditing:** To prevent data leakage and arbitrary distribution shifts, data splitting was managed deterministically via pre-computed CSV files tracking strict `subject_id` isolation. The exact ratio matrices (ρ), configuration shell scripts (`run_experiment_baseline.sh`, etc.), and model checkpoint files have been preserved to allow complete auditing of the empirical pipeline.

5.4 Use of AI Assistance

The author acknowledges the use of large language models (LLMs) to assist with brainstorming, rewriting and verifying spelling. Furthermore AI tools were used to assist with coding, and validate implementation decisions. These tools were not used for automation of scientific research, the author has reviewed code changes and content. The use of AI assistance solely supported the work and research. Responsibilities for the final work remain with the author.

6 Discussion

The results of this study show how weakly supervised, multi-task models behave when trained on hip X-rays. Most importantly, our experiments show that Multi-Instance Learning (MIL) in this configuration is not actually the best approach for grading osteophyte severity. Instead, the data shows a trade-off, where the best accuracy does not mean the best localisation.

6.1 Multi-Instance Learning Configurations

Our results show that adding Multi-Instance Learning (MIL) actually made the model perform worse in predicting severity. The Dual-Head MIL model peaked early and then got steadily worse as we added more data, ultimately losing to our simple Masked Baseline.

This happens due to two major reasons. First, there is a mismatch between the mathematical rules we gave to the model and real life. The Noisy-OR rule assumes that the model only needs to find the single worst spot to decide if a hip is diseased. In reality, osteoarthritis usually occurs on multiple parts of the hips at the same time. Forcing the network to route its training signals exclusively through whichever single quadrant looks worst at that moment causes the local heads to have conflicting gradient updates. The gradients shift erratically back and forth between different regions from one training step to the next, which destabilizes the stable patterns learned from the clean, expert-labelled data.

Secondly, the weak positive labels lack specific severity grading. When the model tries to minimise the loss, the Cramer loss penalises larger classification gaps more significantly. Given that the majority of the dataset is of Grade 0 and Grade 1, the model chooses the path of least resistance and predicts Grade 1 to satisfy the positive constraint.

The Pure MIL model, which removed the global head broke down even more. Without the global "healthy versus sick" anchor to keep its training stable, the model struggled to learn from the data. As we can see in the Grad-CAM visuals, the model's focus completely deteriorated. Instead of looking at the joint margins where bone spurs form, the model focused on a very small area of the image that does not correspond to the likely location of the osteophyte. Because its focus became so narrow and misplaced, it failed to capture the target location.

6.2 Severity Grading versus Spatial Localization

Another discovery of this study shows a classification versus localization distinction, where optimising for one performance metric leads to degradation of another, specifically the Masked Baseline has the better classification performance (Mean Kappa of 0.5836) where as the Dual-Head MIL has the better visualization, focussing on a more compact spot on the right area.

While this trade-off shows a divergence in performance on different metrics, we must address one important factor. While the Masked Baseline performs better than the other models on classification, its spatial attention maps show it exploiting an anatomical shortcut. It mimics the oracle model, focussing its attention around the joint space to determine the severity rather than the protrusion it is tasked with detecting, showing its attention remains diffuse rather than focussed. Conversely, the Noisy-OR constraint forces the shared backbone to focus too much on a single point, ignoring other prominent osteophytes, possibly resulting in an overall decrease in mean Kappa score.

6.3 The use of healthy samples

Scaling the amount of weak data in the Masked Baseline improves the baseline of 0.4809 test mean Kappa at $\rho = 0$ to a peak of 0.5836 at full scale. Because the Masked Baseline ignores weak positive labels (unhealthy), this improvement cannot be caused by the model becoming better at recognising osteophytes. Instead, this improvement is purely driven by the healthy samples. Given the flood of new data and healthy joints, the model likely becomes stricter in determining the boundary between healthy and unhealthy. In return, the model is able

to more accurately predict determine grade 1 from grade 0, as can be seen in Appendix A, which shows better predictions grade 0/1 improving the kappa score.

6.4 Limitations: The Impact of Too Much Healthy Data

A major limitation of this study comes from the imbalance in our clinical data. Real-world medical datasets like CHECK and OAI contain many healthy patients or people with very early-stage anomalies, likely due to advanced osteoarthritis being rarer. This means that our model did not get the chance to train on very severe cases of osteophytes, which are easier to spot than smaller cases.

Another limitation of the study is that our most successful model ignores the weak positive labels completely. This means that there are thousands of images that the model does not train on, while these images could potentially improve the model.

Finally, the reported results were derived from a single seed (123). Consequently, the smaller, individual performance differences should be interpreted with caution as they may fall in between the variance of runs. The larger trends however, such as the divergence between Masked Baseline and degrading MIL variants are more robust to this uncertainty.

7 Conclusions

This study systematically evaluated the potential and boundaries of weakly supervised multi-task deep learning architectures for regional osteophyte severity grading and anatomical localization in hip X-Rays.

By analyzing our models across multiple training ratios, our experiments suggests answers to our hypotheses and research questions. First, our results demonstrate that introducing weak data into a joint multi-task pipeline can significantly improve regional grading performance, but this outcome is heavily dependent on the chosen architecture. The Masked Baseline achieved a robust test score of 0.5836 Kappa compared to the strong-only baseline score of 0.4809, proving that simple global labels can close the gap toward the perfect-label Oracle ceiling of 0.8586. This performance trajectory shows that a network does not necessarily need thousands of examples of severe disease to grade it well. Instead, it can improve significantly based on healthy anatomy alone through the use of weak negative labels. Conversely, trying to actively train on weak positive labels using a standard MIL approach introduces noisy data that misses severity information, degrading performance.

Second, we observed from our findings that our models are indeed affected by noisy data, but only when using explicit positive label propagation. For the MIL configurations, validation performance peaks early at 1,000 weak samples (0.5294 Kappa) and steadily degrades as more data is added. This occurs likely because at early weak-to-strong ratios, the weak samples provide some form of regularization, however when the ratio grows, the missing severity grade information on the weak positive labels push the model towards predicting Grade 1 to take the path of least resistance (given the dataset contains more Grade 1 images). For the Masked Baseline, however, the model continued to scale effectively to maximum data limits without hitting a performance ceiling.

Finally, our spatial analysis did not provide direct support to the third hypothesis. Rather than the weak-to-strong ratio causing a consistent contraction of the attention field, spatial focus appeared to be reliant on the supervision strategy. The Dual-Head MIL concentrated on a compact area near the femoral site, while the Masked Baseline maintained a diffuse, joint-wide focus like the Oracle model. This more precise localization however,

did not provide a better grading accuracy. The Masked Baseline achieved a higher kappa score despite its diffuse focus. This reveals a trade-off between getting the right grade and looking at the right spot.

8 Future Work

The trade-off between severity grading and localization opens several crucial avenues for future investigation.

To fix the limitations of the Noisy-OR formulation, future architectures must move away from assigning a global label to just a single most significant area. Implementing a *Soft-OR* probabilistic formulation, or training flexible *Attention-Based* pooling modules [12] would allow the network to share learning signals across multiple quadrants at the same time. This adjustment would allow the model to better capture the real world, where osteoarthritis commonly comes in multiple areas at once.

Furthermore, future work should not have to guess the location completely in the dark. By leveraging software like BoneFinder to limit the area where the model can look at, we force the models attention map around the bone edges. This would separate the model's focus from the global branch, preventing the network from relying on joint-space shortcuts while keeping its final grading scores high.

Finally, a promising direction involves the use of medical vision-language foundation models to automate the extraction of weak data from unstructured text reports [13]. Rather than compressing rich clinical descriptions down to a binary "healthy vs. diseased" indicator, natural language processing can extract nuanced, location specific notes. Feeding these text descriptions directly through our local heads, would provide much richer training signals bypass the current data scarcity bottleneck entirely.

References

[1] Jaimie D Steinmetz, Garland T Culbreth, Lydia M Haile, Quinn Rafferty, Justin Lo, Kai Glenn Fukutaki, Jessica A Cruz, Amanda E Smith, Stein Emil Vollset, Peter M Brooks, et al. Global, regional, and national burden of osteoarthritis, 1990–2020 and projections to 2050: a systematic analysis for the global burden of disease study 2021. *The Lancet Rheumatology*, 5(9):e508–e522, 2023.

[2] Yanping Xue, Rongguo Zhang, Yufeng Deng, Kuan Chen, and Tao Jiang. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. *PloS one*, 12(6):e0178992, 2017.

[3] Huy Hoang Nguyen, Simo Saarakkala, Matthew B. Blaschko, and Aleksei Tiulpin. Semixup: In- and out-of-manifold regularization for deep semi-supervised knee osteoarthritis severity grading from plain radiographs. *IEEE Transactions on Medical Imaging*, 39(12):4346–4356, 2020.

[4] Leo Misera, Gustav Müller-Franzes, Daniel Truhn, and Jakob Nikolas Kather. Weakly supervised deep learning in radiology. *Radiology*, 312(1):e232085, 2024.

[5] Roy D Altman and GE Gold. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and cartilage*, 15:A1–A56, 2007.

[6] Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE, June 2016.

[8] Claudia Lindner, S Thiagarajah, JM Wilkinson, the arcOGEN Consortium, and TF Cootes. Robust and accurate shape model matching using random forest regression-voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1474, 2013.

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

[10] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220, 1968.

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[12] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *CoRR*, abs/1802.04712, 2018.

[13] Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025.

A Figures

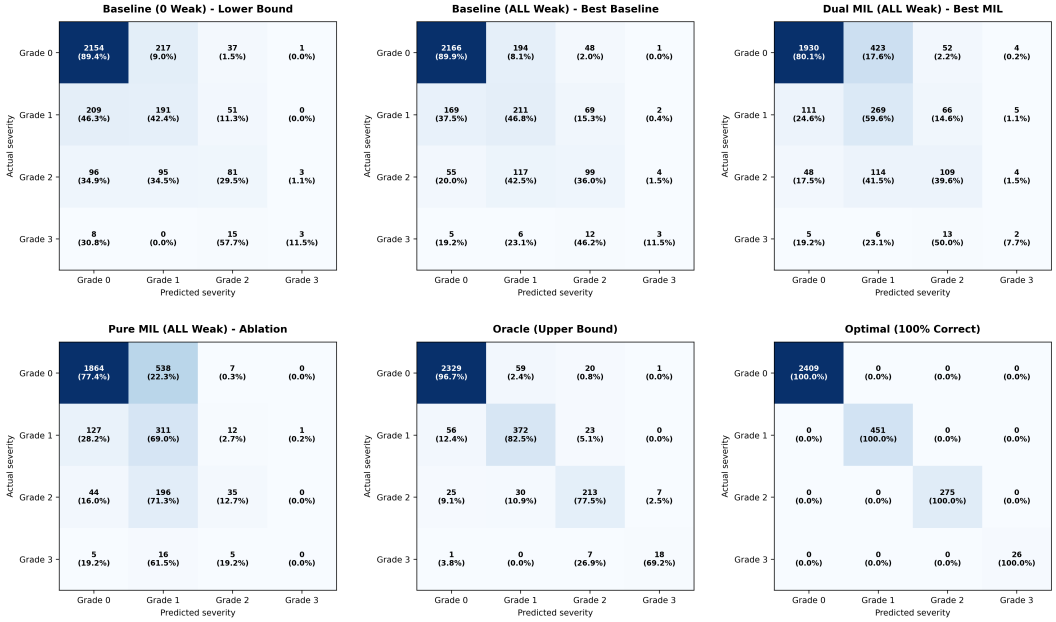

Figure 3: Confusion matrix containing the different architectures evaluated on the test set

B Complete Results

Table 2: Validation Performance Selected by Minimizing Joint Multi-Task Loss ($\mathcal{L}_{total}$)

Configuration / Run	Best Step	Val Loss	Global Acc	Femur Sup	Femur Inf	Acet. Sup	Acet. Inf	Mean κ
Oracle Reference								
100% Strong	478	1.0059	0.8359	0.7128	0.5100	0.4715	0.3498	**0.5110**
Baseline (No MIL)								
bounds_1000_0	63	1.2362	0.7704	0.2951	0.1148	0.2071	-0.0025	**0.1536**
bounds_1000_500	187	1.1621	0.7903	0.5960	0.3237	0.2774	0.1452	**0.3356**
bounds_1000_1000	125	1.1827	0.7751	0.2691	0.0904	0.3309	0.0000	**0.1726**
bounds_1000_2500	329	1.1388	0.7970	0.4364	0.1149	0.2758	0.2200	**0.2618**
bounds_1000_5000	934	1.1304	0.7846	0.6601	0.4903	0.4120	0.3860	**0.4871**
bounds_1000_ALL	1915	1.0943	0.8235	0.6822	0.3689	0.3506	0.0094	**0.3528**
Dual-Head MIL								
bounds_1000_0	63	1.2362	0.7704	0.2951	0.1148	0.2071	-0.0025	**0.1536**
bounds_1000_500	187	1.2119	0.8083	0.5850	0.3608	0.3760	0.2473	**0.3923**
bounds_1000_1000	125	1.3195	0.7751	0.4393	0.0314	0.2210	0.0338	**0.1814**
bounds_1000_2500	439	1.2189	0.7761	0.5642	0.1423	0.3769	0.1859	**0.3173**
bounds_1000_5000	560	1.2429	0.8207	0.4803	0.3808	0.2548	0.1644	**0.3201**
bounds_1000_ALL	1915	1.1789	0.8359	0.7002	0.3587	0.4304	0.0380	**0.3818**
Pure MIL (No Global)								
bounds_1000_0	159	0.6329	0.5351	0.5693	0.4392	0.2990	0.1714	**0.3697**
bounds_1000_500	187	0.6420	0.5721	0.5768	0.1994	0.3863	0.2560	**0.3546**
bounds_1000_1000	125	0.6805	0.4421	0.2346	0.0123	0.3237	0.0257	**0.1491**
bounds_1000_2500	1099	0.6813	0.4649	0.5166	0.4118	0.3700	0.3708	**0.4173**
bounds_1000_5000	934	0.7294	0.5047	0.5941	0.4585	0.4220	0.3819	**0.4641**
bounds_1000_ALL	2873	0.7193	0.5835	0.5434	0.2993	0.3668	0.0868	**0.3241**

Table 3: Validation Performance Selected by Maximizing Localized Mean Quadratic Cohen's Kappa (val/mean_kappa)

Configuration / Run	Best Step	Val Loss	Global Acc	Femur Sup	Femur Inf	Acet. Sup	Acet. Inf	Mean κ
Oracle Reference								
100% Strong	21075	1.9796	0.8112	0.7588	0.6188	0.4802	0.5850	**0.6107**
Baseline (No MIL)								
bounds_1000_0	927	1.5956	0.7400	0.6147	0.5173	0.3996	0.3677	**0.4748**
bounds_1000_500	1644	1.7022	0.8036	0.6499	0.5613	0.3972	0.4869	**0.5238**
bounds_1000_1000	2204	1.9259	0.6983	0.6532	0.4631	0.4512	0.4813	**0.5122**
bounds_1000_2500	3519	1.7953	0.7932	0.6780	0.5861	0.4346	0.4054	**0.5260**
bounds_1000_5000	5422	1.7450	0.7789	0.6991	0.5439	0.4432	0.4060	**0.5231**
bounds_1000_ALL	5747	1.2162	0.8216	0.7318	0.6260	0.4055	0.4017	**0.5412**
Dual-Head MIL								
bounds_1000_0	927	1.5956	0.7400	0.6147	0.5173	0.3996	0.3677	**0.4748**
bounds_1000_500	2020	1.7293	0.8046	0.6523	0.5088	0.4309	0.4834	**0.5189**
bounds_1000_1000	2204	1.8179	0.7381	0.6751	0.5157	0.4536	0.4730	**0.5294**
bounds_1000_2500	5279	2.0078	0.7998	0.6730	0.4931	0.4564	0.4446	**0.5168**
bounds_1000_5000	3739	1.7665	0.7865	0.6120	0.5606	0.4292	0.4176	**0.5048**
bounds_1000_ALL	20596	1.9940	0.7941	0.6594	0.5711	0.3034	0.4423	**0.4941**
Pure MIL (No Global)								
bounds_1000_0	543	0.7162	0.4649	0.6351	0.5608	0.3607	0.4188	**0.4939**
bounds_1000_500	516	0.8193	0.5806	0.6676	0.5561	0.3466	0.4554	**0.5064**
bounds_1000_1000	1259	0.8065	0.4677	0.5371	0.4918	0.4831	0.4910	**0.5007**
bounds_1000_2500	1429	0.7925	0.5133	0.5892	0.4992	0.3895	0.3625	**0.4601**
bounds_1000_5000	934	0.7294	0.5047	0.5941	0.4585	0.4220	0.3819	**0.4641**
bounds_1000_ALL	3352	0.7703	0.5636	0.6745	0.5668	0.2072	0.4287	**0.4693**

Table 4: Performance Metrics on the Sequestered Held-Out Test Set

Configuration / Run	Global Acc	Femur Sup	Femur Inf	Acet. Sup	Acet. Inf	Mean κ
Oracle Reference						
100% Strong	0.9317	0.9107	0.8407	0.8412	0.8417	**0.8586**
Baseline (No MIL)						
bounds_1000_0	0.7636	0.5930	0.5288	0.5010	0.3007	**0.4809**
bounds_1000_500	0.8217	0.6162	0.5890	0.5007	0.4526	**0.5397**
bounds_1000_1000	0.7257	0.6476	0.4223	0.4708	0.4260	**0.4917**
bounds_1000_2500	0.8432	0.6475	0.5532	0.4987	0.4768	**0.5440**
bounds_1000_5000	0.8673	0.6795	0.5200	0.6053	0.5134	**0.5795**
bounds_1000_ALL	0.8723	0.6883	0.5933	0.5353	0.5177	**0.5836**
Dual-Head MIL						
bounds_1000_0	0.7636	0.5930	0.5288	0.5010	0.3007	**0.4809**
bounds_1000_500	0.8369	0.6596	0.5006	0.5033	0.4564	**0.5300**
bounds_1000_1000	0.7876	0.6573	0.4534	0.5003	0.4381	**0.5123**
bounds_1000_2500	0.8685	0.6864	0.5100	0.5579	0.4763	**0.5577**
bounds_1000_5000	0.8774	0.6792	0.4684	0.5967	0.3653	**0.5274**
bounds_1000_ALL	0.9140	0.6427	0.5190	0.5097	0.3520	**0.5058**
Pure MIL (No Global)						
bounds_1000_0	0.4602	0.5997	0.4949	0.4273	0.3857	**0.4769**
bounds_1000_500	0.4893	0.6498	0.5358	0.4796	0.4591	**0.5311**
bounds_1000_1000	0.5512	0.5713	0.4414	0.4941	0.4956	**0.5006**
bounds_1000_2500	0.5006	0.5981	0.4877	0.4133	0.4073	**0.4766**
bounds_1000_5000	0.4829	0.5678	0.4625	0.5114	0.3348	**0.4691**
bounds_1000_ALL	0.5537	0.6258	0.5946	0.2430	0.3376	**0.4502**