Evaluating Baseline and Anatomically Guided Preprocessing for Weakly Supervised Hip Osteophyte Classification

Ege Yarar
Supervisor(s): Jesse Krijthe, Gijs van Tulder
[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: Ege Yarar
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Julia Olkhovskaia

Abstract

Weakly supervised osteophyte classification in hip X-ray images is challenging because only image-level labels are available, providing no explicit information about osteophyte location. However, anatomical landmarks can be used to identify regions where osteophytes are most likely to occur and guide the model towards clinically relevant structures. At the same time, broader anatomical context may also contain useful information for classification. As a result, it remains unclear whether models benefit more from broad anatomical context or from localized regions centered on anatomically relevant structures.

This project evaluates whether anatomically guided preprocessing can improve weakly supervised hip osteophyte classification compared to a baseline preprocessing approach. Hip X-rays from the Osteoarthritis Initiative (OAI) and CHECK datasets were processed using two strategies: broad femoral head centered crops and localized landmark based crops generated using BoneFinder anatomical landmarks. ResNet-18 models were trained for binary osteophyte classification and evaluated using ROC-AUC. We further hypothesized that anatomically guided preprocessing would be particularly beneficial when training data is limited, as focusing on clinically relevant regions may improve data efficiency. To investigate this, additional experiments were conducted using reduced training set sizes (50%, 25%, and 10% of the available training data).

Unexpectedly, the results show that the baseline preprocessing approach consistently achieved higher classification performance than the anatomically guided approach across all evaluated anatomical regions, despite using lower resolution crops than the landmark-guided approach. For example, the baseline model achieved an ROC-AUC of 0.889 for superior femoral osteophyte classification, whereas the corresponding landmark-based model achieved an ROC-AUC of 0.783. Reducing the training set size generally reduced performance for both approaches.

These findings suggest that localized landmark based crops do not necessarily improve weakly supervised osteophyte classification and that broader anatomical context may provide important information to predict accurately. Future work could investigate alternative localization strategies and more precise osteophyte annotations.

The source code used in this study is publicly available at:
`https://github.com/egeyarar/osteophyte-classification`

1 Introduction

Osteoarthritis [1] is a common joint disease that affects millions of people worldwide and is mostly diagnosed using X-ray imaging. One of the key indicators of osteoarthritis is the presence of osteophytes, which are small bone growths that form around joints such as the hip. Detecting these structures can be difficult, especially in the early stages of the disease, because they are often small and not clearly visible. This makes it important to develop methods that can support more reliable analysis of X-ray images.

In recent years, deep learning models, particularly convolutional neural networks (CNNs), have been widely used in medical image analysis and have achieved strong results. However, obtaining detailed spatial annotations, such as pixel-level segmentations or precise location annotations, is often expensive and time consuming [2]. In the datasets used in this project, osteophyte labels are available for predefined anatomical regions of the hip joint, including femoral superior, femoral inferior, acetabular superior, and acetabular inferior locations. Nevertheless, the annotations do not provide precise spatial information about the exact extent or boundaries of the osteophytes. As a result, the problem remains weakly supervised and the model must learn relevant image features without explicit localization annotations.

Although deep learning methods have achieved strong results in medical image analysis [3], it remains unclear whether incorporating anatomical knowledge can improve weakly supervised osteophyte classification. Osteophytes develop at specific anatomical locations around the hip joint, suggesting that focusing on clinically relevant regions may help the model learn more relevant features while reducing distracting image content. This challenge is particularly relevant for pelvic X-rays, where osteophytes occupy relatively small regions while the surrounding image contains a large amount of anatomical information that may not be directly related to osteophyte formation.

This motivates the comparison between broad anatomical crops and localized anatomically-guided crops. Broad crops preserve more information about the overall joint structure but also include potentially irrelevant image content. Localized crops focus the model on regions where osteophytes are expected to occur, but may remove useful contextual information. Evaluating this trade-off helps determine whether anatomically guided preprocessing improves classification performance and data efficiency.

In previous works, anatomical guidance has been incorporated into medical image analysis in several ways. Some approaches use anatomical segmentations or anatomically defined regions of interest to isolate structures before classification [4], while others incorporate attention mechanisms that guide the network towards clinically relevant regions during training [5]. Landmark detection methods have also been widely used to localize anatomical structures and support downstream image analysis tasks [6].

While these studies demonstrate that anatomical information can improve medical image analysis, they incorporate anatomical guidance in fundamentally different ways. Segmentation based approaches require additional annotations that are often unavailable, whereas attention based approaches introduce architectural modifications that may themselves influence performance. Consequently, it remains unclear whether anatomical localization alone, introduced through preprocessing rather than model design, is sufficient to improve weakly supervised osteophyte classification. This uncertainty motivates this study. For this reason, the current study focuses exclusively on anatomical guidance through preprocessing rather than architectural modifications. This allows the effect of anatomical localization itself to be investigated without introducing additional model complexity or supervision requirements.

The main hypothesis of this work is that restricting the input image to anatomically relevant regions can reduce unnecessary image context and encourage the model to focus on structures that are more directly associated with osteophyte formation. If this assumption is correct, anatomically-guided preprocessing may improve classification performance compared to broader hip centered crops. Furthermore, focusing on anatomically relevant regions may improve data efficiency by enabling the model to learn useful representations from fewer training examples.

To investigate this hypothesis, this project evaluates anatomically guided preprocessing using BoneFinder pointfiles for weakly supervised osteophyte classification in hip X-ray images. Two preprocessing strategies are compared: a baseline method using broad femoral head centered crops and a landmark-guided method using localized crop regions around osteophyte related anatomical structures. The project uses hip X-ray images from the OAI [7] and CHECK [8] datasets and evaluates whether anatomically guided preprocessing affects classification performance and data efficiency.

The main research question addressed in this work is:

> How does anatomically guided preprocessing affect hip osteophyte classification performance compared to a baseline preprocessing approach?

To answer this question, three aspects are investigated. First, the research observes which anatomical crop regions provide the most useful information for osteophyte classification. Second, the effect of reducing the amount of available training data is evaluated to determine how classification performance changes under limited data conditions. Finally, the research investigates whether anatomically guided preprocessing provides advantages when only a small number of training examples are available.

To investigate these questions, CNN [9] models are trained for binary osteophyte classification using multiple anatomical target regions. Model performance is evaluated using ROC-AUC [10] scores. In addition, experiments are performed using 100%, 50%, 25%, and 10% of the available training data to evaluate the effect of dataset size on model performance.

The rest of the paper is structured as follows. Section 2 describes the methodology. Section 3 describes the experimental setup. Section 4 reports the results. Section 5 discusses responsible research considerations. Section 6 provides a discussion of the findings. Finally, Section 7 concludes the paper and outlines directions for future work.

2 Methodology

The methodology of this project consists of dataset preparation, preprocessing, anatomical landmark integration, CNN training, and evaluation. The proposed approach is designed to be independent of a specific deep learning framework and can be applied to other medical image classification tasks where anatomical landmarks are available.

2.1 Dataset Preparation

The experiments use hip X-ray images from the OAI dataset and the CHECK cohort dataset. Together, these datasets contain approximately 22,000 hip X-ray images with osteophyte related labels. The labels indicate the presence or absence of osteophytes in different anatomical regions of the hip joint.

The project starts from the original DICOM X-rays and performs all preprocessing steps within the developed pipeline, including image normalization, anatomical cropping, resizing, and HDF5 dataset generation.

All available images are included during dataset generation regardless of which osteophyte labels are present. For each image, only the available annotations are stored in the HDF5 dataset. Images for which preprocessing fails due to invalid landmark configurations, missing point files, or other data quality issues are excluded. Approximately 2% of the images are removed during this process.

2.2 Defining Anatomical Crop Locations

BoneFinder provides anatomical landmark annotations describing the geometry of the hip joint [6]. Each X-ray contains 160 landmark points in total (80 per hip), corresponding to anatomical structures of the femur, acetabulum, and surrounding pelvic anatomy. The renumbered BoneFinder landmark convention provided by the preprocessing framework is used throughout this project.

Two preprocessing strategies are used in this project: a baseline approach based on broad femoral head centered crops and a landmark-guided approach based on localized anatomical crop regions. BoneFinder landmarks are used differently in each strategy to determine crop locations.

For the baseline approach, BoneFinder landmarks are used to fit a femoral head circle. The center of this circle is used as the crop center, resulting in a broad crop around the hip joint.

For the landmark-guided approach, crop centers are defined using specific groups of BoneFinder landmarks corresponding to anatomical regions where osteophytes commonly occur [11, 12]. Four anatomical regions are considered: femoral superior, femoral inferior, acetabular superior, and acetabular inferior. The crop center is calculated as the mean position of the selected landmark group.

Using anatomically defined landmarks ensures that crop regions are positioned consistently across subjects despite differences in anatomy and image acquisition.

2.3 Preprocessing Pipelines

A custom preprocessing pipeline is implemented to convert the raw DICOM images into HDF5 datasets suitable for CNN training. The preprocessing process includes image loading, pixel spacing standardization, anatomical cropping, landmark validation, image resizing, and HDF5 dataset generation.

Two different preprocessing pipelines are implemented and compared throughout the experiments.

- **Baseline preprocessing pipeline:** The baseline approach generates broad anatomical crops centered around the femoral head. The crop location is determined using the fitted femoral head circle obtained from BoneFinder landmarks. The cropped images are resized to 224×224 pixels before being stored in HDF5 format. This preprocessing strategy preserves a larger amount of surrounding anatomical context and serves as the reference approach for comparison.
- **Landmark-guided preprocessing pipeline:** The proposed approach uses BoneFinder anatomical landmarks to generate localized crop regions centered around osteophyte-related anatomical structures. Four anatomical regions are considered: femoral superior, femoral inferior, acetabular superior, and acetabular inferior. For each region, crop centers are determined using predefined landmark groups and images are cropped around these locations before being resized to 224×224 pixels.

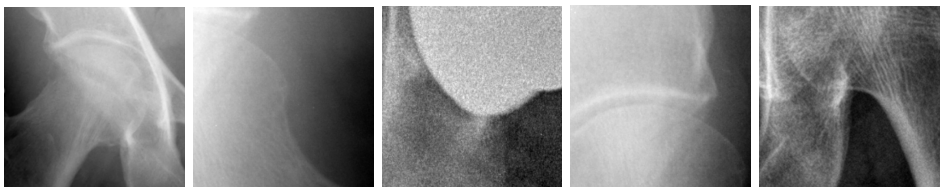


Figure 1: From left to right: baseline, femoral superior, femoral inferior, acetabular superior and acetabular inferior crop regions.

Figure 1 shows examples of the crop regions generated by both preprocessing pipelines. Compared to the baseline crop, the landmark-guided crops focus on substantially smaller anatomical regions while preserving structures relevant to osteophyte detection.

Both preprocessing pipelines start from the same raw DICOM images and apply identical normalization and dataset generation procedures. The pipelines differ both in the anatomical

region used for cropping and in the effective image resolution. The baseline approach uses larger femoral head centered crops generated at 0.4 mm/pixel, whereas the landmark-guided approach uses localized crops generated at 0.2 mm/pixel. As a result, the landmark-guided crops provide a more zoomed in view of the anatomical structures, while the baseline crops preserve a broader view.

2.4 CNN Training

The classification experiments are performed using a ResNet-18 convolutional neural network [13].

For the baseline preprocessing strategy, a single network is trained using the large femoral head centered crop. This network predicts osteophyte presence for all four anatomical target regions simultaneously.

For the landmark-guided preprocessing strategy, four separate networks are trained. Each network receives a localized crop corresponding to a specific anatomical region (femoral superior, femoral inferior, acetabular superior, or acetabular inferior) and predicts osteophyte presence for that region only.

This setup enables a direct comparison between a single model that has access to broader anatomical context and models that focus on more precise localized anatomical regions.

3 Experimental Setup

This section describes the experimental setup used throughout the study. Details regarding the software and hardware environment, dataset splits, preprocessing procedures, model configuration, training process, and evaluation methodology are provided to ensure reproducibility and facilitate a fair comparison between preprocessing strategies.

3.1 Software and Hardware Environment

All experiments were implemented in Python using PyTorch, PyTorch Lightning, and MONAI. Training and evaluation were performed on the DelftBlue high-performance computing cluster.

3.2 Dataset Splits

Dataset splitting was performed at the subject level to prevent data leakage between training, validation, and test sets. Subject identifiers were first sorted and subsequently shuffled using a fixed random seed of 123 to ensure reproducibility.

The dataset was divided into approximately 70% training, 15% validation, and 15% test subjects. Additional experiments were performed using reduced training subsets containing 50%, 25%, and 10% of the original training subjects while keeping the validation and test sets unchanged.

3.3 Preprocessing Configuration

The baseline preprocessing dataset was generated using a target pixel spacing of 0.4 mm/pixel and fixed-size crops of 224×224 pixels.

For the landmark-guided preprocessing strategies, images were resampled to a target pixel spacing of 0.2 mm/pixel and cropped to 224×224 pixels. Separate datasets were generated for the femoral superior, femoral inferior, acetabular superior, and acetabular inferior anatomical regions.

All generated datasets were stored in HDF5 format and used for subsequent training and evaluation experiments.

3.4 Model Configuration

All experiments used a ResNet-18 architecture pretrained on ImageNet. The final classification layer was adapted for binary osteophyte classification.

Input X-rays were resized to 224×224 pixels and replicated to three channels before being provided to the network.

ResNet-18 was selected because it provides a good balance between model complexity and computational efficiency. In addition, ResNet architectures are widely used in medical image analysis and have demonstrated strong performance on X-ray classification tasks, making them a suitable baseline for comparing preprocessing strategies.

3.5 Training Procedure

All experiments used a pretrained ResNet-18 architecture initialized with ImageNet weights. Models were trained using the Adam optimizer with a learning rate of 0.001 and a mini-batch size of 16. Training was performed for a maximum of 20 epochs.

Binary cross-entropy loss with logits (BCEWithLogitsLoss)[14] was used as the training objective. For a single prediction, the loss is defined as

$$L_i = -\left(y_i \log(\sigma(x_i)) + (1 - y_i) \log(1 - \sigma(x_i))\right)$$

where x_i represents the model output logit, y_i the corresponding binary ground-truth label, and σ the sigmoid activation function.

During training, the loss was averaged over all valid labels in a mini-batch. Missing annotations were excluded from the loss computation.

Model checkpoints were saved throughout training. For each experiment, the checkpoint achieving the lowest validation loss was selected for final evaluation on the test set.

3.6 Evaluation Metrics

The performance of the model is evaluated using ROC-AUC. In addition, accuracy, precision, and recall are reported to provide a more complete classification evaluation performance.

Because not all images contain annotations for every osteophyte target region, only samples with available annotations for the corresponding target are included when computing evaluation metrics. As a result, the number of evaluated samples differs slightly between anatomical regions.

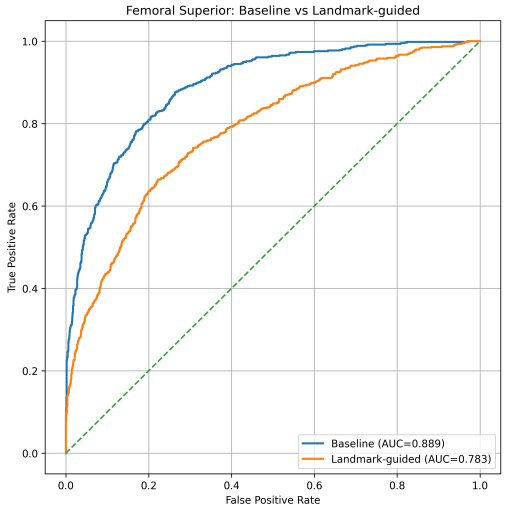
The baseline and landmark-guided preprocessing pipelines are compared using identical dataset splits, network architectures, optimization settings, and evaluation procedures. This ensures that observed performance differences can be attributed primarily to the preprocessing strategy.

Because osteophyte positive samples are less common than negative samples, class imbalance may affect threshold dependent metrics such as accuracy, precision, and recall. ROC-AUC was therefore used as the primary evaluation metric, as it evaluates discrimination performance across all classification thresholds.

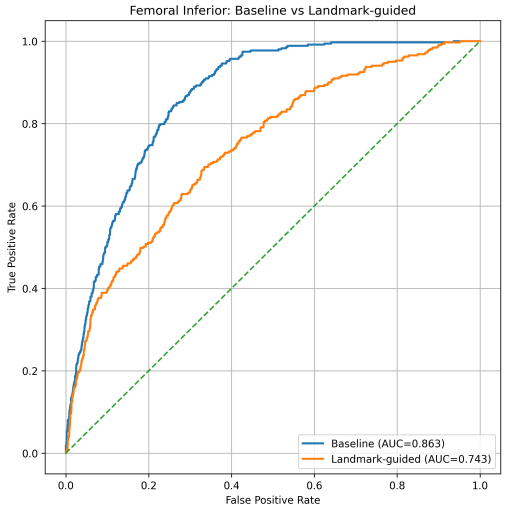
4 Results

The results are presented in three parts. First, the baseline and landmark-guided preprocessing strategies are compared directly. Second, the performance of the different anatomical regions is compared. Finally, the effect of reducing the amount of available training data is analyzed.

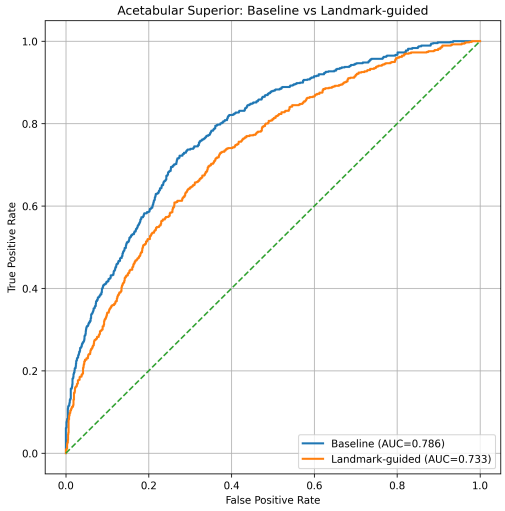
4.1 Comparison Between Baseline and Landmark-Guided Preprocessing



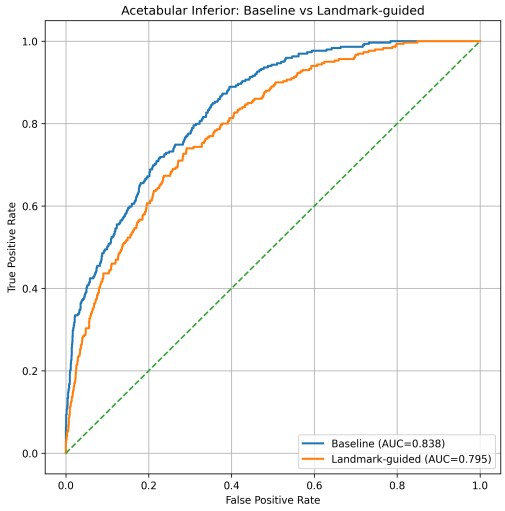
(a) Femoral superior



(b) Femoral inferior



(c) Acetabular superior



(d) Acetabular inferior

Figure 2: Comparison of ROC curves obtained using the baseline and landmark-guided preprocessing strategies for each anatomical region. Results are shown for models trained on the full training dataset.

Across all anatomical regions, the baseline preprocessing strategy achieved higher ROC-AUC scores than the corresponding landmark-guided approach. The highest overall performance was obtained using the baseline femoral superior crop, which achieved a ROC-AUC score of 0.889. In comparison, the best landmark-guided result was obtained using the acetabular

inferior region, reaching a ROC-AUC score of 0.795.

Table 1: Best ROC-AUC scores obtained for each preprocessing strategy using the full training dataset.

Method	Region	ROC-AUC
Baseline	Femoral Superior	0.889
Baseline	Femoral Inferior	0.863
Baseline	Acetabular Inferior	0.838
Baseline	Acetabular Superior	0.786
Landmark-guided	Femoral Superior	0.783
Landmark-guided	Femoral Inferior	0.743
Landmark-guided	Acetabular Inferior	0.795
Landmark-guided	Acetabular Superior	0.733

These results indicate that anatomically-guided preprocessing did not improve classification performance. The baseline approach achieved higher ROC-AUC scores in all four anatomical regions. The largest performance differences were observed for the femoral superior and femoral inferior regions, where the baseline approach outperformed the landmark-guided approach by more than 0.10 ROC-AUC points. The smallest difference was observed for the acetabular inferior region, where the landmark-guided approach achieved its strongest performance. Nevertheless, the baseline approach remained superior in all comparisons, suggesting that preserving broader anatomical context is beneficial for osteophyte classification.

4.2 Comparison of Anatomical Regions

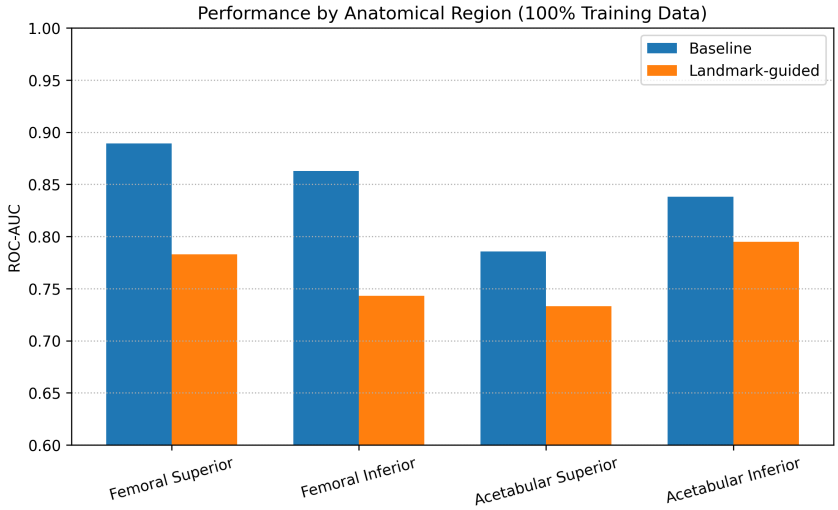


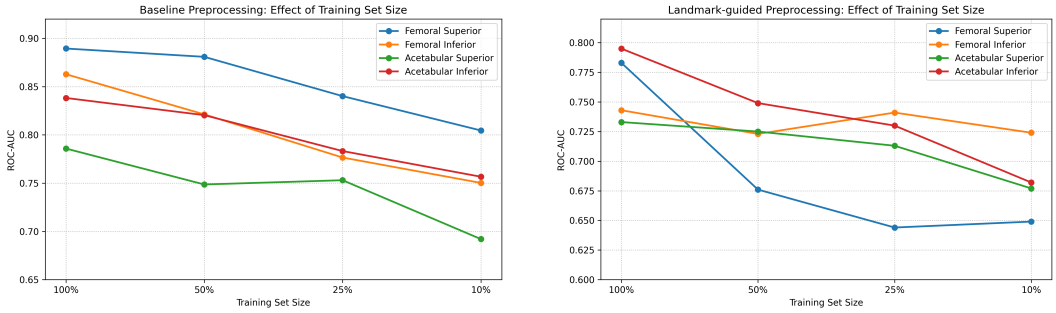
Figure 3: ROC-AUC scores obtained using the baseline and landmark-guided preprocessing strategies for each anatomical region when trained on the full dataset.

Femoral superior achieved the highest performance under the baseline preprocessing strategy, whereas acetabular inferior achieved the highest performance among the landmark-guided approaches. Acetabular superior consistently produced the lowest ROC-AUC scores.

The smallest performance gap between the two preprocessing strategies was observed for the acetabular inferior region, where the landmark-guided approach achieved its strongest result. In contrast, the largest differences were observed for the femoral superior and femoral inferior regions.

4.3 Data Efficiency Analysis

To investigate data efficiency, experiments were repeated using 50%, 25%, and 10% of the original training subjects.



(a) Baseline preprocessing

(b) Landmark-guided preprocessing

Figure 4: Effect of training set size on ROC-AUC for the baseline and landmark-guided preprocessing strategies.

For the baseline approach, all anatomical regions showed a gradual decrease in ROC-AUC as training set size decreased. Femoral superior consistently achieved the highest performance across all training set sizes.

For the landmark-guided approach, the effect of reducing training data differed between anatomical regions. Femoral superior showed the largest performance reduction, whereas femoral inferior remained relatively stable across all training set sizes.

The results indicate that larger training datasets generally improve performance for both preprocessing strategies. However, the magnitude of the performance decrease varies across anatomical regions. In particular, femoral inferior appears more robust to reductions in training data than the other landmark-guided regions.

5 Responsible Research

This project uses hip X-ray datasets from the OAI and CHECK cohorts. Since all personal information has been removed from the datasets, no patient identifiable information is used during preprocessing, training, or evaluation.

The generated datasets and trained models are used only for this research project. The data and results are stored securely and used solely for academic purposes.

There are several limitations to this work. First, the available labels only indicate whether an osteophyte is present and do not provide its exact location. Therefore, the

models are trained in a weakly supervised setting and may learn image features that are not directly related to osteophytes. Although the proposed landmark-guided preprocessing aims to reduce this problem, it cannot completely eliminate it. Second, the OAI and CHECK datasets may contain demographic or clinical biases. As a result, models trained on these datasets may not generalize equally well to other populations or imaging settings.

Furthermore, performance was evaluated primarily using aggregate metrics. Consequently, potential differences in performance across demographic subgroups could not be assessed within the scope of this study.

To support reproducibility, all preprocessing steps, model configurations, dataset splits, and evaluation procedures were implemented in Python using PyTorch, PyTorch Lightning, and MONAI. Training, validation, and test splits were created at the subject level to avoid data leakage and ensure a fair evaluation of model performance. In addition, the complete source code and experiment configurations have been made publicly available. Nevertheless, exact reproduction of the results may still be affected by differences in hardware, software versions, and stochastic training behaviour.

Finally, the models developed in this project are intended for research purposes only and should not be used for clinical decision making without further validation.

AI based tools were used to assist with code debugging and to improve the clarity and structure of the written report. All suggestions were reviewed and verified by the author.

6 Discussion

The results show that anatomically-guided preprocessing did not outperform the baseline preprocessing strategy. Although the landmark-guided crops focused on clinically relevant anatomical regions, the baseline femoral head centered crops consistently achieved higher ROC-AUC scores.

One possible explanation is that the baseline crops preserve a larger amount of surrounding anatomical context. While osteophytes are localized structures, their presence may be associated with broader structural changes throughout the hip joint. Osteophytes occurring in one anatomical region may also be accompanied by changes in other regions of the joint. As a result, the model may benefit from observing a larger anatomical field of view rather than focusing exclusively on a single localized region. By restricting the input to small crops, the landmark-guided preprocessing strategy may remove contextual information that contributes to classification performance. However, the current experiments do not directly reveal which image regions are used by the models when making predictions. Additional analyses using visualization techniques could help determine whether the baseline models rely on broader anatomical context or whether the landmark-guided models focus more strongly on the localized crop regions.

At the same time, localized crops may also have some advantages. By focusing on a specific anatomical region, the model is encouraged to learn features that are directly related to that region instead of relying on information from other parts of the image. This may make the model focus more on the structures that are relevant for osteophyte detection. In addition, the relatively stable performance of the femoral inferior region when the amount of training data was reduced suggests that anatomically-guided preprocessing may still be useful in situations where only limited training data is available, even though it did not outperform the baseline approach in this study.

Image resolution represents a second factor that cannot be fully separated from anatomical context in the current experimental design. The baseline dataset was generated using

a target pixel spacing of 0.4 mm/pixel, whereas the landmark-guided datasets used 0.2 mm/pixel. As a result, the baseline inputs preserve a broader anatomical field of view at lower effective resolution, while the landmark-guided inputs provide a more detailed representation of a smaller anatomical region. Since the baseline and landmark-guided pipelines were intentionally designed with different crop sizes and pixel spacings, the reported results reflect the overall effectiveness of each preprocessing strategy. Future work could investigate the individual contribution of anatomical context and image resolution by controlling these factors independently.

The experiments also revealed differences between anatomical regions. Among the landmark-guided approaches, acetabular inferior and femoral superior achieved the strongest performance on the full dataset. In contrast, acetabular superior produced the lowest ROC-AUC values in most experiments. This suggests that some anatomical regions contain more useful information for osteophyte classification than others.

A general decrease in performance was observed as the amount of available training data was reduced. However, the primary motivation for these experiments was to investigate whether anatomically-guided preprocessing could improve data efficiency under limited data conditions. The results did not provide strong evidence for this hypothesis, as the baseline preprocessing strategy remained the best performing approach across all training set sizes. Nevertheless, the landmark-guided femoral inferior region showed relatively stable ROC-AUC scores as the amount of training data decreased. This suggests that certain anatomically-guided crop regions may be more robust to limited training data, although this advantage was not sufficient to outperform the baseline approach overall.

These findings are consistent with previous work showing that broader anatomical information can be useful for osteoarthritis related image classification tasks [15]. Unlike approaches that incorporate anatomical information directly into the learning process, the current study used anatomical guidance only during preprocessing through crop selection. This distinction may explain why the current findings differ from studies reporting improvements through anatomical guidance. Methods based on attention mechanisms or segmentation guided learning can exploit anatomical information while still preserving broader contextual information, whereas the landmark-guided approach evaluated in this work restricts the available region of interest before training begins.

Several limitations should be considered. First, the experiments were performed using weakly supervised labels that indicate osteophyte presence within a region but do not provide exact osteophyte locations. Second, only a single network architecture (ResNet-18) was evaluated. Different architectures may respond differently to localized crop regions. For example, larger networks such as ResNet-50 or attention based models may be better at extracting features from small anatomical regions and could potentially benefit more from landmark-guided preprocessing. Another limitation is that the experiments were evaluated using a single train, validation, and test set split and a single random seed. Consequently, the statistical significance of the observed performance differences was not assessed. Future work could repeat the experiments using multiple random seeds, confidence intervals, and statistical significance testing to evaluate the robustness of the reported results. Despite these limitations, the consistency of the results across anatomical regions and training set sizes supports the conclusion that broader anatomical context contributes substantially to classification performance in this setting.

7 Conclusions and Future Work

The results show that anatomically-guided preprocessing did not improve overall classification performance. Across all experiments, the baseline femoral head centered preprocessing strategy achieved higher ROC-AUC scores than the landmark-guided approaches. Therefore, the results suggest that preserving broader anatomical context is more beneficial than using highly localized anatomical crops for weakly supervised osteophyte classification.

The most predictable anatomical landmark region depended on the amount of available training data. With 100% and 50% of the training data, acetabular inferior achieved the highest ROC-AUC scores of 0.795 and 0.749, respectively. When the training set was reduced to 25% and 10%, femoral inferior achieved the highest ROC-AUC scores of 0.741 and 0.724. Acetabular superior generally produced the weakest performance, while femoral superior showed a larger decrease in performance as the amount of training data was reduced. Within the current experimental setup, the acetabular inferior and femoral inferior regions appeared to provide the most discriminative information among the landmark-guided crop regions. However, this observation is specific to the datasets, preprocessing strategy, and model configuration evaluated in this study.

Classification performance generally decreased as the amount of available training data was reduced for both preprocessing strategies. However, the magnitude of the decrease differed across anatomical regions. While the baseline femoral superior region remained the strongest overall performer, the landmark-guided femoral inferior region showed relatively stable performance across all training set sizes. This suggests that the femoral inferior crop region may be more robust to reductions in training data, although it did not outperform the baseline approach in terms of overall classification performance.

The main contribution of this work is a systematic comparison between baseline femoral head centered preprocessing and anatomically guided landmark-based preprocessing for weakly supervised hip osteophyte classification. In addition, the study provides an analysis of how different anatomical crop regions behave under changing amounts of training data.

Future work could investigate alternative methods for incorporating anatomical information into deep learning models. Rather than using landmarks only for preprocessing, anatomical information could be integrated directly into the network through attention mechanisms, multi-input architectures, or segmentation-guided approaches. Such methods may allow the model to benefit from both localized anatomical information and broader contextual information simultaneously, potentially overcoming the limitations observed in the current study.

Future work could also investigate which image regions contribute most strongly to model predictions. Visualization techniques such as Grad-CAM could help determine whether baseline models benefit from broader anatomical context and whether landmark-guided models focus more specifically on osteophyte related structures. Such analyses may provide additional insight into the reasons behind the observed performance differences.

Future studies could also evaluate larger crop regions, which may preserve more anatomical context while still focusing on clinically relevant structures. In addition, different CNN architectures could be investigated, as larger networks or attention based models may be better able to extract informative features from localized anatomical regions. Finally, datasets with stronger annotations, such as pixel level osteophyte segmentations, can provide more precise supervision and help determine whether anatomically-guided approaches become more effective when exact osteophyte locations are available.

A Detailed Baseline Evaluation Results

Table 2: Baseline performance using 100% of the training data.

Region	Samples	Accuracy	Precision	Recall	AUC
Femoral Superior	3306	0.8730	0.7127	0.5293	0.8894
Femoral Inferior	3268	0.8908	0.4873	0.2181	0.8628
Acetabular Superior	2795	0.7764	0.5286	0.4635	0.7858
Acetabular Inferior	3279	0.9167	0.5747	0.3344	0.8382

Table 3: Baseline performance using 50% of the training data.

Region	Samples	Accuracy	Precision	Recall	AUC
Femoral Superior	3306	0.8482	0.5833	0.6384	0.8808
Femoral Inferior	3268	0.8874	0.4468	0.1785	0.8212
Acetabular Superior	2795	0.7857	0.5840	0.3116	0.7486
Acetabular Inferior	3279	0.9094	0.5065	0.2609	0.8204

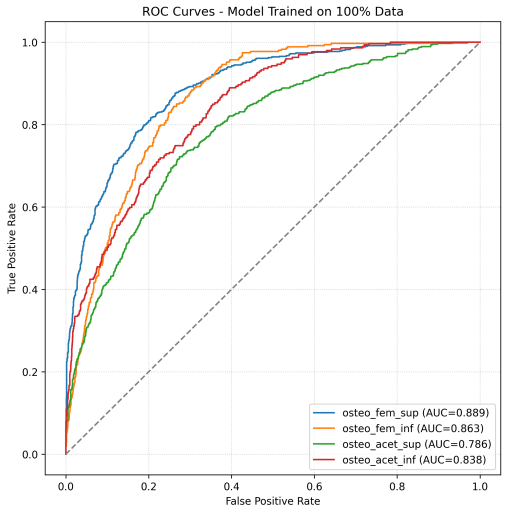
Table 4: Baseline performance using 25% of the training data.

Region	Samples	Accuracy	Precision	Recall	AUC
Femoral Superior	3306	0.8500	0.6405	0.4381	0.8401
Femoral Inferior	3268	0.8926	0.5156	0.0935	0.7765
Acetabular Superior	2795	0.7871	0.5987	0.2903	0.7530
Acetabular Inferior	3279	0.9073	0.4286	0.0502	0.7832

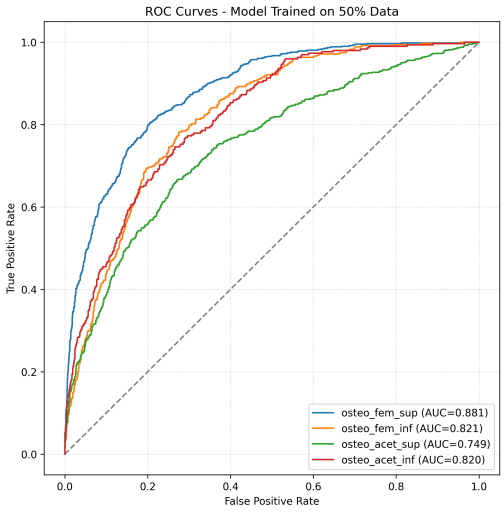
Table 5: Baseline performance using 10% of the training data.

Region	Samples	Accuracy	Precision	Recall	AUC
Femoral Superior	3306	0.8351	0.5835	0.3925	0.8045
Femoral Inferior	3268	0.8883	0.4375	0.1190	0.7502
Acetabular Superior	2795	0.7399	0.4400	0.3845	0.6920
Acetabular Inferior	3279	0.8975	0.3791	0.1940	0.7566

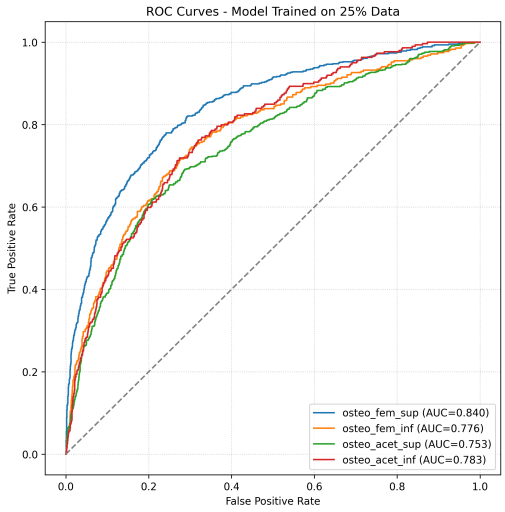
B Additional ROC Curves



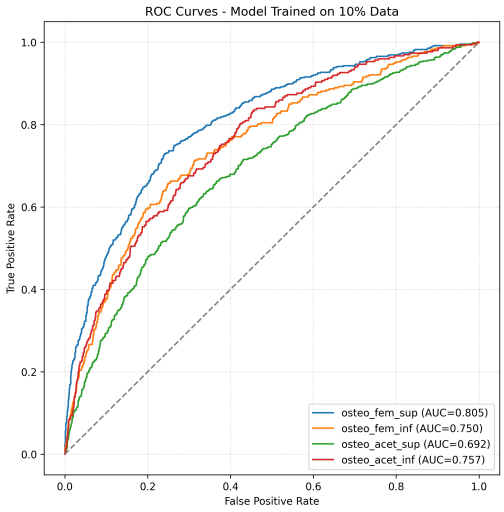
(a) 100% training data



(b) 50% training data

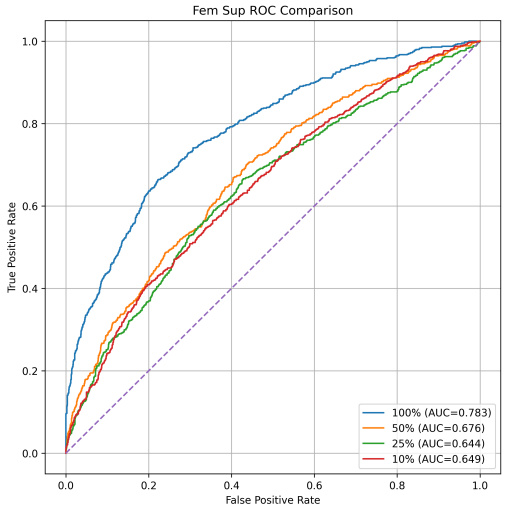


(c) 25% training data

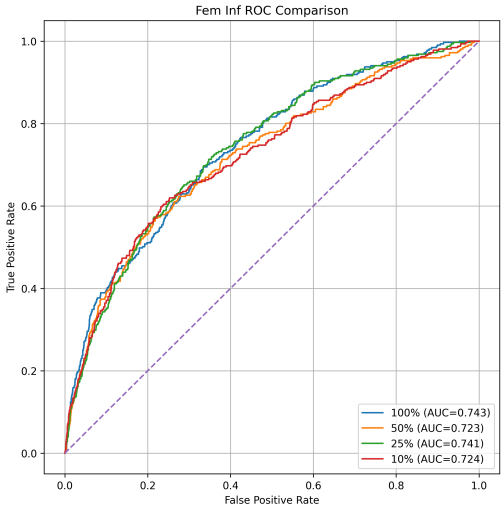


(d) 10% training data

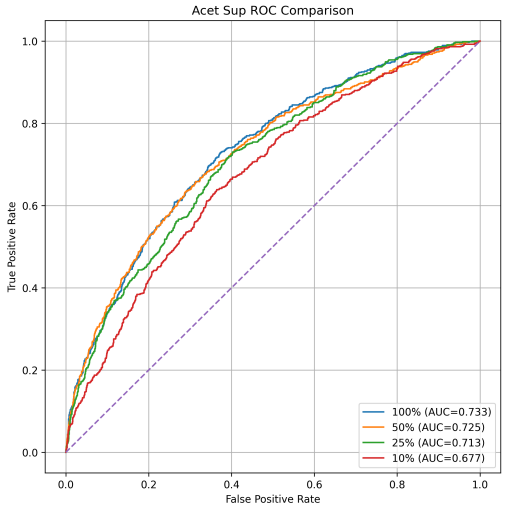
Figure 5: ROC curves obtained for the baseline preprocessing strategy at different training set sizes. The curves show the performance of all four anatomical target regions.



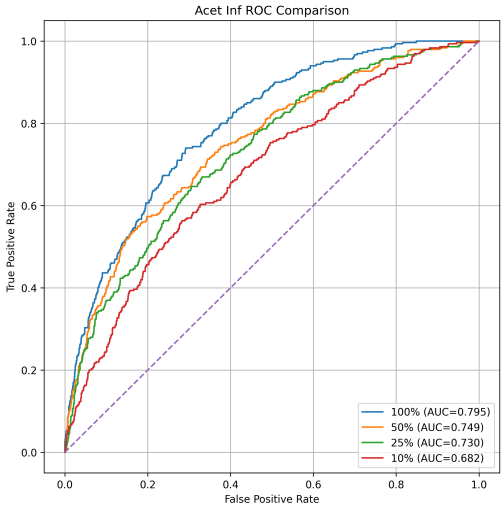
(a) Femoral superior



(b) Femoral inferior



(c) Acetabular superior



(d) Acetabular inferior

Figure 6: ROC curves obtained using landmark-guided preprocessing for different training set sizes.

C Training Configuration

All experiments were trained using the same hyperparameter configuration. The command line arguments used for training are summarized below.

Table 6: Training hyperparameters used in all experiments.

Parameter	Value
Optimizer	Adam
Learning Rate (`--lr`)	0.001
Batch Size (`--mb-size`)	16
Maximum Epochs (`--max-epochs`)	20
Random Seed (`--seed`)	123
Network Architecture	ResNet-18
Pretrained Weights	ImageNet
Input Resolution	224×224

References

[1] D. J. Hunter and S. Bierma-Zeinstra, "Osteoarthritis," *The Lancet*, 2019.

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *CVPR*, 2017.

[3] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[4] A. Tiulpin, J. Thevenot, E. Rahtu, P. Lehenkari, and S. Saarakkala, "Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach," *Scientific Reports*, 2018.

[5] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Hammerla, B. Kainz *et al.*, "Attention U-Net: Learning where to look for the pancreas," in *Medical Imaging with Deep Learning*, 2018.

[6] C. Lindner, S. Thiagarajah, J. M. Wilkinson, G. A. Wallis, and T. F. Cootes, "Fully automatic segmentation of the proximal femur using random forest regression voting," *IEEE Transactions on Medical Imaging*, 2013.

[7] Osteoarthritis Initiative, "Osteoarthritis initiative," 2024, https://nda.nih.gov/oai/.

[8] J. Wesseling *et al.*, "Cohort profile: Cohort hip and cohort knee (check) study," *International Journal of Epidemiology*, vol. 45, no. 1, pp. 36–44, 2015.

[9] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[10] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[11] B. G. Faber, R. Ebsim, F. R. Saunders *et al.*, "Osteophyte Size and Location on Hip DXA Scans Are Associated with Hip Pain: Findings from a Cross-Sectional Study in UK Biobank," *Bone*, vol. 153, p. 116146, 2021.

[12] H. Funahashi, Y. Osawa, Y. Takegami *et al.*, "Acetabular osteophyte formation in dysplastic hip osteoarthritis," *BMC Musculoskeletal Disorders*, vol. 25, 2024.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] PyTorch, "Pytorch documentation: Bcewithlogitsloss," 2025, https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html.

[15] C. E. von Schacky, J. H. Sohn, F. Liu, E. Ozhinsky, P. M. Jungmann, L. Nardo, S. C. Foreman, M. C. Nevitt, T. M. Link *et al.*, "Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs," *Radiology*, vol. 295, no. 1, pp. 136–145, 2020.