# Anatomy-Aware Masked Autoencoders for Hip Osteoarthritis Classification in X-ray Images

**Jasper Christiaan van Beusekom**[1]

**Supervisors: 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 23, 2025

Name of the student: Jasper Christiaan van Beusekom
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Michael Weinmann

## Abstract

Self Supervised Learning (SSL) has been shown to effectively utilise unlabelled data for pre-training models used in down-stream medical tasks. This property of SSL enables it to use much larger datasets when compared to supervised models, which require manually labelled data. Medical classification tasks often require the identification of patterns inside a small Region Of Interest (ROI) known to be relevant for radiographic diagnosis. This contrasts standard image classification tasks, which generally rely on broader patterns. To guide a model in learning such anatomically relevant features, we investigated the hip osteoarthritis classification performance of a ROI-guided Masked Autoencoder (MAE) with a Convolutional Neural Network (CNN)-based architecture. Unlike conventional MAEs, which learn latent features by reconstructing randomly masked images, our alternative uses generated anatomical landmarks to exclusively mask the ROI or background. Contradicting similar research on Vision Transformer (ViT)-based MAEs, random masking outperformed our ROI-guided alternatives, revealing a fundamental difference in what drives performance for the two architectures, and guiding future research on more sophisticated ROI-guided masking strategies. The code is available on GitHub: https://github.com/Jasperdetweede/AnatAMAE/


## 1 Introduction

Osteoarthritis (OA) is a degenerative joint disease that is progressively affecting more individuals [14]. Over time, the disease causes a gradual breakdown of the tissues in the affected joint, leading to stiffness and pain. The radiographic diagnosis of osteoarthritis is typically performed by trained medical personnel through the examination of X-ray images, a process that is both costly and time-consuming. To mitigate these issues, automation using machine learning has been proposed [11].

However, many machine learning models require a sizeable number of labelled examples to function properly. In medical settings, acquiring this labelled data is especially time-consuming and subjective, contradicting the reason for using machine learning in the first place. In contrast, unlabelled data is far more abundant, as it does not require human annotation. Self-Supervised Learning (SSL) has emerged as a promising approach that effectively utilises this unlabelled data for medical tasks [7, 1], as it uses only input data to extract latent features.

SSL operates by creating a supervised task where the target is generated from input data instead of provided. This paper focusses on self-predictive SSL methods, which learn by reconstructing original versions of masked, transformed or contrasted data.

A distinct characteristic making classification tasks in medical imaging particularly challenging is the presence of a relatively small Region Of Interest (ROI) within otherwise uninformative areas. Where regular imaging tasks generally require the model to find broader patterns, medical classification often requires capturing localized structures, like hairline fractures or early-stage tumours. For hip osteoarthritis classification, these include reduced cartilage thickness around the femoral head and the presence of bone abnormalities called osteophytes [8]. Figure 1 shows a healthy hip, and one affected by severe osteoarthritis.

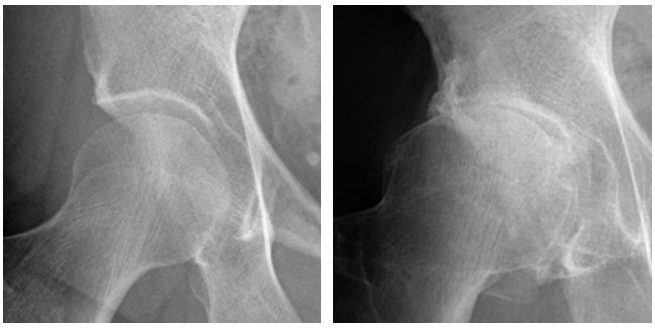


Figure 1: Comparison between a healthy hip (left) and one affected by severe osteoarthritis (right). The affected hip shows strong signs of joint space narrowing.

Given the smaller ROIs, it would be valuable to guide a model's attention toward these anatomically relevant regions, known to contain important diagnostic information. While several approaches exist with the potential to leverage such domain knowledge, this paper will focus on Masked Autoencoders (MAEs), which learn by attempting to reconstruct masked images and remain largely unexplored in this context.

Specifically, this research paper aims to explore the adaptation of SSL models through the lens of masked autoencoders, by proposing an anatomy-aware masked autoencoder for hip osteoarthritis detection. This will be investigated by introducing two contrasting masking strategies: one that masks only within the ROI, and another that masks only outside the ROI. These strategies will be compared against two baseline methods across varying mask sizes. Each configuration will be evaluated on classification performance.

## 2 Related Work

The masked autoencoder was first introduced in 2021 [6], making it a relatively young self-supervised learning approach. Consequently, the conducted research in this area is limited. In general, it has been shown that masked autoencoders benefit from context-aware modifications [5, 15, 3, 9], especially when tailored to a specific medical task [16, 13].

**Non-Medical Context-Guided Masking** - Comparable studies in non-medical contexts show that replacing random masking with a context-guided approach significantly increases accuracy. Various papers have suggested different strategies for context-based masking, like graph-cut segmentation [5] and attention maps created by a trained generator [3].

**Medical Text-Guided Masking** - In the medical field, Xie et al. [15] developed MedIM, which identifies key words and

sentences in the radiology reports accompanying medical images, and uses them to determine which areas to mask. Concretely, the model was trained on key words and their visual counterpart from the image, and tasked to map them close together in a shared embedding space. Subsequently, the areas named in the report are masked more often. This strategy was shown to consistently increase performance for various tasks.

**Medical Anatomy-Aware Masking** - A conceptually similar method to the one in this paper was proposed by Zheng et al. [16]. They used the different intensities from a CT-scan as a segmentation, which formed the base for their masks. This experiment showed that for a segmentation task, anatomy-aware masking significantly improves the accuracy of an MAE. Another concept similar to the one in this paper was proposed by Szijártó et al. [13]. Here, ROI-awareness was shown to increase the accuracy of a masked autoencoder for ultrasound video. Specifically, they applied a binary segmentation to ultrasound video to mask everything except the ROI. Subsequently, the ROI was masked using random masking. This nearly doubled the accuracy when compared to their baseline.

In sum, recent work has shown that swapping out purely random masks for ones guided by anatomy [16], medical reports [15] or ROI detections [13] consistently helps MAEs learn features that matter for medical tasks.

This research will apply a similar anatomy-guided approach to the task of hip osteoarthritis classification, with the goal of learning how performance is affected by different masking strategies. One noticeable difference between this paper and the ones discussed is that, where others predominantly use a ViT architecture, this paper employs a CNN-based autoencoder.

# 3 Methodology

## 3.1 Model Architecture

The masked autoencoder consists of an encoder and a decoder. The encoder processes a masked input image to extract a latent feature representation, which the decoder then uses to reconstruct the original image. Each epoch, the reconstruction loss is calculated from the difference between the original and reconstructed images, and used to update the model parameters. Following the pre-training, the encoder is extracted and a classifier is attached to its output. Subsequently, supervised learning with unmasked images is utilised to train the classifier and fine-tune the encoder. Figure 2 shows a visual representation of the model architecture.

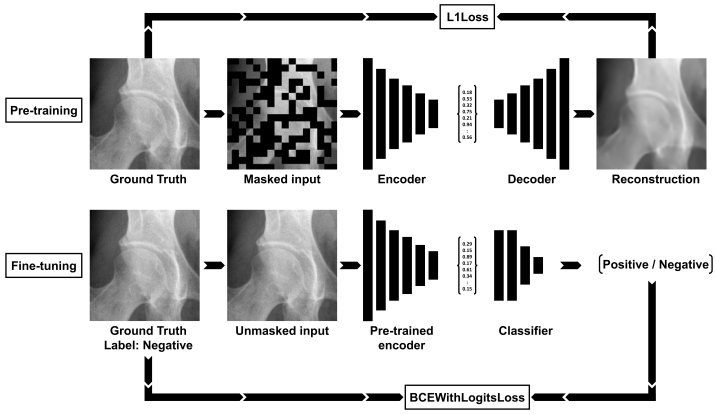



Figure 2: Overview of the model architecture. The first cycle illustrates the pre-training pipeline; the second shows the fine-tuning pipeline. L1Loss calculates the linear, pixel-wise difference between the ground truth image and reconstruction. BCEWithLogitsLoss calculates the loss between raw model output and a binary ground truth.

Two options were considered for implementation of the MAE: a Convolutional Neural Network (CNN) and a Vision Transformer (ViT). ViTs function by dividing an image into patches, which are serialized and shown to the encoder accompanied by positional embeddings. In contrast, CNNs receive the complete input data without division into patches. ViTs have frequently been shown to outperform CNNs in image classification tasks [12]. Nevertheless, CNNs often achieve comparable performance in such cases [4]. Additionally, using a conceptually simpler model reduces the risk of confounding factors influencing the results. Given that this study primarily aims to compare masking strategies, the masked autoencoder was implemented using a CNN architecture.

## 3.2 Masking

Before being passed to the autoencoder, images are masked following a predefined strategy. This paper focuses on the design and effect of that masking approach. To this end, we distinguish between random patch masking, the prevailing standard, and ROI-guided masking.

**Random Masking** - Random masking is the strategy utilised by He et al. in the original paper on masked autoencoders [6]. Random masking divides the image in patches of a fixed size and masks a fixed percentage. A masking percentage of 75% is generally found to be optimal [6], although some studies report improved performance with slightly lower ratios [9].

**ROI-guided Masking** - ROI-guided masking is a hypernym encompassing any strategy that considers the region of interest when determining which areas of the image to mask. The version used in this paper defines the ROI to be the anatomical area known to be most relevant in radiographic diagnosis of hip osteoarthritis: the outline of the femoral head [8]. Specifically, two ROI-guided strategies were considered: one masks part of the ROI while leaving the background intact; the other masks part of the background while preserving the ROI. Figure 6 depicts the ROI for the input data.

# 4 Experiment

## 4.1 Training Data

The Cohort Hip and Cohort Knee (CHECK) dataset [2] was used for both pre-training and fine-tuning. This population-based dataset covers 1002 participants from the Netherlands during a maximum of five visits [T0, T2, T5, T8, T10] over ten years. Participants were selected based on the presence of symptomatic osteoarthritis of the hip and/or knee. Although

X-ray images of various joints are present in the dataset, this study was limited to images of the hip. Additionally, only measurements with a Kellgren-Lawrence label for each hip were used, resulting in 3.359 images and 6.718 hips from 940 participants. Internally, the data is stored in DICOM file format, together with metadata required for preprocessing, like pixel density and photometric interpretation.

**BoneFinder** - For all used images, a set of landmarks was generated to outline anatomically significant parts of the image. To this end, BoneFinder [10], a machine learning tool, was used. An example of such landmarks can be seen on the output images in figure 3.

**Kellgren-Lawrence Grading** - The Kellgren-Lawrence (KL) system is the most commonly used radiographic classification system for osteoarthritis of the hip joint. It scores hips on an integer range of [0,4], where a score of 0 means no features of osteoarthritis are present. A score of 4 is defined by the presence of large bone abnormalities called osteophytes, joint space narrowing and severe sclerosis. The CHECK dataset provided KL-scores for all individual hips.

## 4.2 Preprocessing

All images were preprocessed to standardise the model input. This included ensuring all images were stored using MONOCHROME1 photometric interpretation, and with a pixel density of 0.1 mm/pixel. The centre of the femoral head was then calculated by averaging the coordinates of all landmarks outlining it. Each image was cropped to 1024×1024 pixels centred on this point and subsequently downsampled to 256×256 to balance image clarity and computational feasibility. For every modification to the image, the landmarks were translated accordingly. Figure 3 depicts a conceptual outline of the preprocessing pipeline.

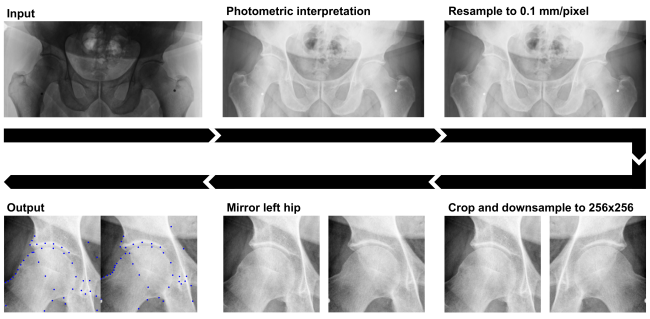



Figure 3: Overview of the preprocessing pipeline. The input is an X-ray image of the hip from the CHECK dataset. Sequentially, photometric interpretation and resampling are applied, followed by cropping centred on the femoral head, downsampling and mirroring of the left hip. The blue dots on the output represent the generated anatomical landmarks.

## 4.3 Evaluation

The first experiment compares four approaches. Of these four, the first and last serve as a baseline, enabling a comparison between a non-masking autoencoder, random masked autoencoder and the ROI-guided alternatives. Figure 4 visually exemplifies the four strategies when applied on a data sample. Specifically, the strategies considered during the conducted experiments were:

- No masking
- ROI masking
- Background masking
- Full masking

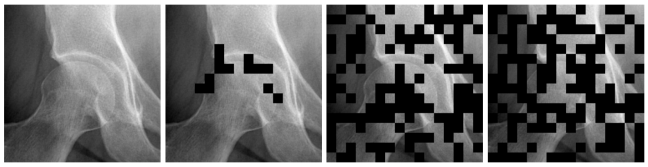


Figure 4: Overview of all masking strategies, applied on an arbitrary data sample with a patch size of 16 pixels. From left to right: 'no masking', 'ROI masking', 'background masking' and 'full masking'.

The second experiment repeats the first for three patch sizes: 8, 16 and 32 pixels. Since patch size does not affect the 'no masking' strategy, this results in ten pre-trained models over both experiments.

During fine-tuning with the pre-trained encoder, the input is not masked, allowing one consistent evaluation process across all models. Each hip was labelled positive if the Kellgren-Lawrence score was higher than 1, and labelled negative otherwise. The Receiver Operating Characteristics (ROC) metric was used to evaluate model performance. This metric gives a complete picture of performance that is independent of the classification threshold. The numerical performance of the model is then defined as the Area Under the ROC curve (AUROC), which represents the chance that the model ranks a randomly chosen positive example higher than a randomly chosen negative example.

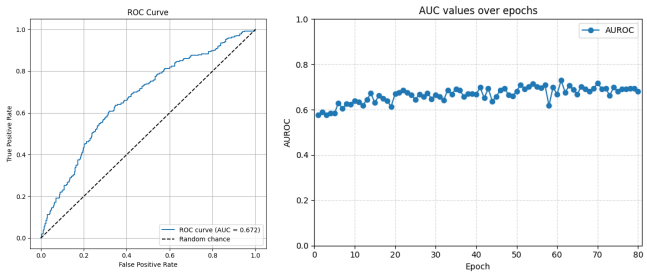



Figure 5: ROC curve (left) and AUROC over epochs (right). Note the variance in the AUROC graph due to arbitrary overfitting on the validation set. To avoid such outliers, the graph was smoothed before calculating performance.

To identify the model with the best overall performance, the AUROC was plotted for each epoch. Subsequently, this graph was smoothed with a five-epoch window, and the optimal window was selected. Finally, to avoid outliers the model with an AUROC closest to the window average was chosen. Model fine-tuning and evaluation were repeated five times per

pre-trained model, with unique seeds. Figure 5 depicts an example of an ROC curve and AUROC graph before smoothing.

**Hyperparameters** - For all experiments, all hyperparameters not under investigation were held constant and chosen based solely on reconstruction accuracy. A batch size of 16 and learning rate of 0.00025 were used to balance convergence speed with training stability. Additionally, each model was trained for 80 epochs, as preliminary tests showed all models to convergence well before this point while keeping overall runtime reasonable and avoiding overfitting. Finally, the masking ratio was set to 50% to highlight potential performance differences while avoiding an overly aggressive masking strategy that could obscure those differences.

The hyperparameters for the finetuning phase were chosen to balance performance with a fair evaluation and computational feasibility. Specifically, a learning rate of 0.0005, batch size of 16 and runtime of 80 epochs were chosen.

**Data Partitioning** - The dataset (6.718) was split into roughly 70% training data (4.754) and 30% evaluation data (1.964). To reduce confounding factors, the same split was used for each run, and each split has roughly the same number of positive labels (26%). For the fine-tuning phase, 20% of the training data was used, to simulate a real-world scenario where only part of the data is labelled.

### 4.4 Implementation Details

The autoencoder was implemented using Python and Pytorch. The encoder is build as 5 convolution layers and 4 residual blocks, using ReLU activation functions in between each layer. The decoder is the reverse of the encoder, creating a symmetrical autoencoder. The classifier head, applied during fine-tuning, is implemented as a network with one hidden layer and a single output dimension. Additionally, dropout and randomised data augmentation are applied to reduce over-fitting.

Both the pre-training and fine-tuning phase use a loss function to evaluate batch-wise performance and update model parameters. During pre-training, the autoencoder minimizes the linear pixel-wise loss (L1Loss) between the original image and the reconstruction. For supervised fine-tuning, a sigmoid activation layer combined with binary cross entropy (BCEWithLogitsLoss) was applied on the classification output and the ground-truth label.

Many existing ViT-based MAEs utilise a loss function that ignores visual patches, rewarding only reconstruction accuracy for the masked patches. However, this loss function is incompatible with the zero percent masking ratio, and would, given the proposed ROI-guided masking strategies, ignore either the ROI or background entirely. Despite the potential for increased performance, this would conceptually be closer to an attention map, which was not the focus of this research.

The ROI-guided masking strategies were implemented by first defining the ROI based on a predefined subset of the generated landmarks, each representing a specific anatomical position on a bone. Figure 6 depicts an example of an extracted ROI. Subsequently, the image is divided into square patches of the specified size. For the 'ROI masking' strategy, patches overlapping one or more pixels of the ROI are selected. For the 'background masking' strategy, the exact opposite patches are extracted. In both cases, a fraction of these patches equal to the masking ratio is masked, while the remainder remains unaltered.

Additionally, the 'no masking' baseline was evaluated using the same model without masking. The 'full masking' baseline was implemented as random patch masking.

The complete codebase can be found on the GitHub repository for this paper, referenced in the abstract.

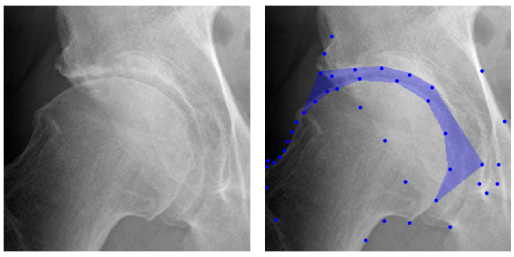


Figure 6: Visual representation of the Region Of Interest (ROI) for an arbitrary data sample. Note the blue dots, which represent generated anatomical landmarks, and the blue segmentation, representing the ROI.

## 5 Results

### 5.1 Pre-Training

Figure 7 shows an example of the image reconstruction for each strategy. Given that the pre-training loss graphs are not the focus of any experiment, this data is not shown in this paper. However, all raw data and visuals are available in the experiments folder on the aforementioned GitHub repository for this paper.

### 5.2 Down-Stream Classification

For all pre-trained encoders, the average classification performance (AUROC) over five fine-tuning runs is shown in figure 8.

First, the 'no masking' baseline is consistently outperformed by all other strategies by a minimum of 2.5%, showing the promising potential of masked autoencoders in medical classification. Secondly, for a patch size of both 8 and 16 pixels, the 'full masking' strategy outperformed the ROI-aware alternatives by 0.4-2.6%.

An interesting exception was found for a patch size of 32 pixels, where 'background masking' outperformed both 'full masking' and 'ROI masking' by roughly 0.5%. Additionally, while both ROI-aware strategies perform similar to the alternative with a 16 pixel patch size, 'full masking' drops significantly when using a 32 pixel patch size.

Although both observations could have various explanations, their combination seems to point to one in particular. In contrast to the 'ROI masking' strategy, which ensures that only a fixed proportion of the ROI is masked, the 'full masking' approach offers no such guarantee. As a result, the ROI is often either almost entirely masked or left mostly intact, both of which can hinder effective feature extraction

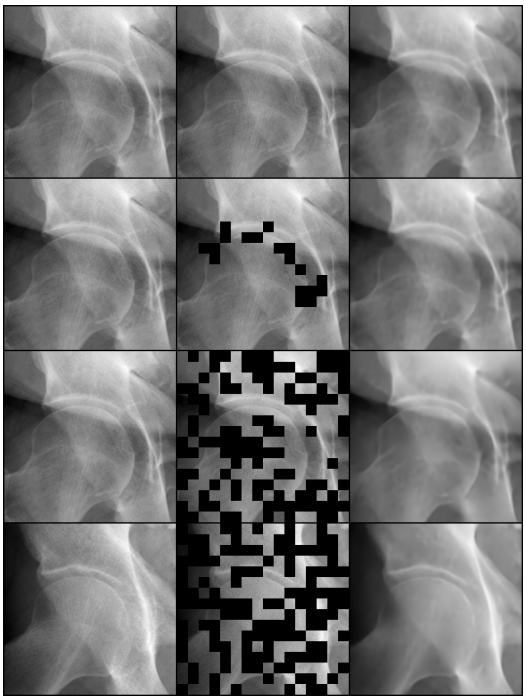
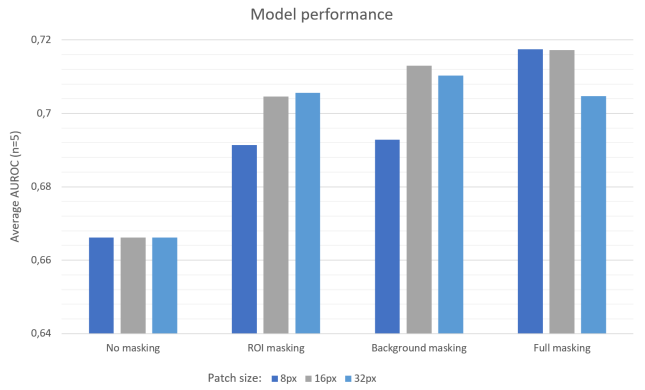


Figure 7: Pre-train reconstructions after 80 epochs for all models with a patch size of 16 pixels. From left to right: ground truth, masked input and reconstruction. From top to bottom: 'no masking', 'ROI masking', 'background masking' and 'full masking'.




| Patch size | No masking | ROI masking | Background masking | Full masking |
|---|---|---|---|---|
| 8px | 0,6662 | 0,6914 | 0,6928 | **0,7175** |
| 16px | 0,6662 | 0,7046 | 0,7130 | **0,7172** |
| 32px | 0,6662 | 0,7056 | **0,7103** | 0,7047 |

Figure 8: Average AUROC per pre-trained encoder-classifier model (n=5). Note that the range on the y-axis was limited to [0.64, 0.72] for clarity. A fully random classifier would score 0.5 on this scale, and a perfect classifier 1.0. In the table, bold text indicates the best performance per patch size.

As hypothesized, 'background masking' was observed to benefit from a larger patch size. This may be explained by the model relying more heavily on global patterns when processing background regions. Although 'ROI masking' was expected to benefit from a smaller patch size due to the presence of important fine-grained details, its performance was found largely unaffected by the variable.

Overall, a patch size of 16 pixels showed the most balanced performance. Additionally, given that the performance drop at 32-pixel suggests this patch size exceeds the optimal threshold, it could be argued that 'full masking' is the most effective masking strategy that was tested for this architecture.

# 6 Responsible Research

## 6.1 Ethical Discussion

As this research concerns both machine learning and application in the medical field, considering the ethical implications of the conducted research is of great importance.

Firstly, the real-world consequences of our model performance must be considered. For medical classification tasks, classification mistakes can often significantly affect the lives of patients. Specifically, a clear distinction has to be made between false positives, and false negatives. Where a false positive causes stress and financial loss, a false negative could deny a patient early treatment, causing suffering. In practice a trade-off between these two has to be made in the form of a classification threshold. Since our model was evaluated using a metric independent of this threshold (AUROC) we did not have to make this ethical trade-off. However, when a similar classification model were to be applied in practice, other metrics like accuracy and recall should be taken into account.

Secondly, automated systems such as ours are fundamentally paired with the ethical concern of responsibility. Morally, anybody that actively helped create and run such a model could to a degree be held responsible. However, we have little precedence for who holds legal responsibility for decisions of end-to-end automated classifiers. Until a classification system for hip osteoarthritis is proven to consistently deliver better radiographic diagnoses than trained medical personnel, we suggest these systems remain a tool, only used under human supervision.

Lastly, we acknowledge that the used dataset contains medically sensitive data. Although all participants signed informed consent statements, we are morally obliged to handle this data with care. Hence, the dataset was not uploaded to the paper's GitHub page, and the visualization of full X-ray images was limited to where it was strictly necessary. More leniency was granted for the cropped and downsampled samples from the training data, as little sensitive information can be extracted from these images.

## 6.2 Reproducibility and Repeatability

Repeatability refers to the likelihood of consistently achieving the same results over multiple sessions, while reproducibility concerns achieving the same results over independent studies.

**Repeatability** - If the conducted experiments were repeated under the exact same conditions, similar results will be achieved. During implementation of the model, consistent effort was put into seeding all randomised factors. Hence, using an input twice will result in the exact same results.

**Reproducibility** - The experiment section extensively describes all factors, including those that potentially affect the results. This allows an accurate recreation of the original conditions. To further improve reproducibility, the codebase used for the experiments is publicly available on the repository referenced in the abstract, along with the input commands and logs for pre-training and fine-tuning. However, the data used for this experiment is not publicly available. Although it can be downloaded on request, this limitation in availability does negatively impact the reproducibility of our findings.

# 7 Discussion and Conclusion

## 7.1 Limitations

Although the results show emerging patterns, there are limitations that need to be acknowledged. Furthermore, because of constraints in both scope and time, the effect of various factors could not be explored, leaving opportunity for follow-up research.

**Margin of Error** - Most notably, the results are limited by the number of performance samples per model. Despite each model being ran five times, the variance in these results leaves on average a significant 95% confidence interval width of 0.76% in a range of 0.43%-1.15%. This means that in some cases, the observed difference in performance falls within the margin of error and could be significantly higher or lower in practice. Nevertheless, the likelihood that the true average performances differ enough to change the overall conclusion is low, supporting the reliability of the results.

**Reconstruction and Prediction Loss** - In this experiment, each model was pre-trained for 80 epochs, based on the convergence of the reconstruction loss. This was done under the assumption that reconstruction loss and prediction accuracy would be inversely correlated. However, it is possible that a lower reconstruction loss is correlated to lower prediction accuracy. Similarly, if reconstruction loss is bottlenecked by the decoder, pre-training for longer might strengthen the embeddings and improve classification performance, while reconstruction loss has plateaued. A systematic set of experiments could be conducted in future research, to learn how reconstruction loss is related to prediction accuracy.

**Masking Ratio** - The experiments considered only two masking ratios (0% and 50%) because of limited computational resources. Despite these values being chosen with the goal of clearly showing potential differences, two points leave room for significant patterns to be overlooked due to underfitting. Repeating the conducted experiments for more masking ratios could strengthen the observed patterns, or reveal new emerging ones.

## 7.2 Discussion

The results of the conducted experiments seem to contradict the findings of similar papers. Where other papers observe improved performance with ROI-aware modifications, 'ROI masking' and 'background masking' were outperformed significantly by random masking. The two factors that most likely caused this difference are the used model architecture, and the loss function.

Firstly, while most papers use a Vision Transformer (ViT) to implement the masked autoencoder, this paper employs a convolutional neural network. Although their conceptual similarities were assumed to allow generalization of the conclusion across architectures, the conflicting results challenge this assumption.

Secondly, ViT-based MAEs are typically paired with a reconstruction loss that ignores unmasked patches, as this approach aligns well with their architecture. This strategy is less commonly used in CNN-based MAEs, due to their stronger reliance on local pixel relationships. However, such a masked-patch loss function places greater emphasis on the masked regions and could therefore significantly influence the results.

Adapting the conducted experiments to be compatible with a ViT and a masked-patch loss could provide an interesting premise for follow-up research and offer meaningful insight into the differences between CNN- and ViT-based MAEs.

Considering the broader applications of the results and proposed models, it is clear they do not achieve a competitive performance for this task, but this was expected. Real-world medical imaging applications require models far larger and more complex than the those studied in this paper. Instead, this paper focussed on gaining a more fundamental understanding of how the masking strategy affects down-stream performance. To this end, this paper showed that CNN-based MAEs do not benefit from masking only the ROI or the background. This suggests they might not benefit from ROI-guided masking at all, or not to the same degree as ViT-based models, revealing a fundamental difference in what drives performance for these two architectures. These observed results can guide the creation of more sophisticated masking strategies, and help build a more fundamental understanding of the factors that affect performance in medical classification tasks.

## 7.3 Conclusion

This paper aimed to find the effectiveness of different anatomy-aware modifications to the masked autoencoder, when applied to classification of hip osteoarthritis from X-ray images.

We have demonstrated that for a convolutional masked autoencoder, full random patch masking outperforms masking only the ROI or background, suggesting they might not benefit from ROI-guided masking. This indicates a fundamental difference in how vision transformers and convolutional neural networks are affected by masking strategies, offering fertile ground for future research.

Secondly, a non-masked baseline was consistently outperformed by all masking alternatives, confirming the promising potential of masked autoencoders in medical classification.

Finally, although the initial hypothesis was disproven, the observed results can guide future research into more sophisticated ROI-guided masking strategies, and help build a more fundamental understanding of the factors that affect performance in masked autoencoders for medical classification tasks.

## References


[1] Mohammed Majid Abdulrazzaq, Nehad T. A. Ramaha, Alaa Ali Hameed, Mohammad Salman, Dong Keon Yon, Norma Latif Fitriyani, Muhammad Syafrudin, and Seung Won Lee. Consequential advancements of self-supervised learning (ssl) in deep learning contexts. *Mathematics*, 12(5), 2024.

[2] J.W.J. Bijlsma and J. Wesseling. Check (cohort hip & cohort knee) data of baseline and 6 to 8 years follow-up, 2015.

[3] Haijian Chen, Wendong Zhang, Yunbo Wang, and Xiaokang Yang. Improving masked autoencoders by learning where to mask, 2024.

[4] Luca Deininger, Bernhard Stimpel, Anil Yuce, Samaneh Abbasi-Sureshjani, Simon Schönenberger, Paolo Ocampo, Konstanty Korski, and Fabien Gaire. A comparative study between vision transformers and cnns in digital pathology, 2022.

[5] Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10386–10395, 2023.

[6] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[7] Shih-Cheng Huang, Anuj Pareek, Malte Jensen, Matthew P. Lungren, Serena Yeung, and Akshay S. Chaudhari. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. *npj Digital Medicine*, 6(1):74, April 2023.

[8] Jeffrey N. Katz, Kaetlyn R. Arant, and Richard F. Loeser. Diagnosis and treatment of hip and knee osteoarthritis: A review. *JAMA*, 325(6):568–578, February 2021.

[9] Yuheng Li, Tianyu Luan, Yizhou Wu, Shaoyan Pan, Yenho Chen, and Xiaofeng Yang. Anatomask: Enhancing medical image segmentation with reconstruction-guided self-masking, 2024.

[10] C. Lindner, S. Thiagarajah, J. M. Wilkinson, The arcOGEN Consortium, G. A. Wallis, and T. F. Cootes. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE Transactions on Medical Imaging*, 32(8):1462–1472, 2013.

[11] Fang Liu, Michael J. Zhou, Michael J. Samsonov, and Richard Kijowski. Predicting early symptomatic osteoarthritis in the human knee using machine learning classification of magnetic resonance images from the osteoarthritis initiative. *Journal of Orthopaedic Research*, 34(10):1646–1651, 2016.

[12] José Maurício, Inês Domingues, and Jorge Bernardino. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences*, 13(9), 2023.

[13] Ádám Szijártó, Bálint Magyar, Thomas Á. Szeier, Máté Tolvaj, Alexandra Fábián, Bálint K. Lakatos, Zsuzsanna Ladányi, Zsolt Bagyura, Béla Merkely, Attila Kovács, and Márton Tokodi. Masked autoencoders for medical ultrasound videos using roi-aware masking. In Alberto Gomez, Bishesh Khanal, Andrew King, and Ana Namburete, editors, *Simplifying Medical Ultrasound*, pages 167–176, Cham, 2025. Springer Nature Switzerland.

[14] Theo Vos, Stephen S. Lim, Cristiana Abbafati, ..., and Christopher J. L. Murray. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396(10258):1204–1222, October 2020. Publisher: Elsevier.

[15] Yutong Xie, Lin Gu, Tatsuya Harada, Jianpeng Zhang, Yong Xia, and Qi Wu. Rethinking masked image modelling for medical image representation. *Medical Image Analysis*, 98:103304, December 2024.

[16] Jie Zheng, Ru Wen, Haiqin Hu, Lina Wei, Kui Su, Wei Chen, Chen Liu, and Jun Wang. Tissue-contrastive semi-masked autoencoders for segmentation pretraining on chest CT. *CoRR*, abs/2407.08961, 2024.