Anatomical Priors for Weakly Supervised Osteophyte Detection and Localization in Hip X-rays

Evaluating BoneFinder-Derived Guidance Under Image-Level Supervision

Iulian-Mihai Onea[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: Iulian-Mihai Onea
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Julia Olkhovskaia

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Osteophytes, bony projections associated with osteoarthritis, are traditionally identified via time-consuming and subjective manual X-ray assessment. While deep learning approaches have shown promising results in medical image analysis, relatively few methods are designed to detect the presence and localization of these osteophytes, particularly in settings where only image-level labels are available and precise pixel-level annotations are missing.

This work investigates whether anatomical priors derived from landmark points can improve weakly supervised osteophyte detection and localization in hip X-rays when only image-level labels are available. We propose modified ResNet18 architectures that integrate anatomical guidance that highlights likely osteophyte regions.

We evaluate the proposed models across varying training data sizes. The results show that models with anatomical guidance generally outperformed the baseline models, with the most consistent improvements observed in classification metrics, while localization results are less conclusive. Additionally, experiments performed without guidance during testing led to a reduced classification performance. Overall, the results suggest that anatomical priors can provide useful complementary information for weakly supervised osteophyte detection, although they do not fully compensate for limited training data. Moreover, the benefit of the guidance information varies across architectures and training set sizes.

1 Introduction

Osteoarthritis (OA) is a common degenerative joint disorder in which cartilage gradually deteriorates and the underlying bone undergoes structural changes. It represents the most common form of arthritis, being a major global health problem and affecting 237 million people worldwide in 2015 (3.3% of the world's population) [1]. This number is expected to reach 765 million by 2060 [2]. The disease often leads to chronic pain, stiffness, and reduced mobility, making it the fourth leading cause of disability, particularly among older adults.

One of the key indicators assessed when looking for OA is the presence of osteophytes—bony projections that develop along joint margins, such as in the knees, spine or hip, detected through X-ray imaging. However, interpreting these scans is time-consuming and can vary between clinicians, especially because osteophytes can differ significantly in size [3]. While more advanced cases are relatively easy to identify, early-stage osteophytes are often small and subtle, making them difficult to detect reliably.

One way to address this challenge is to develop models that can learn from weak labels, where training images are annotated only with coarse information, such as whether an osteophyte is present or absent, rather than precise locations. Obtaining such image-level labels is less time-consuming and expensive than collecting detailed expert annotations, making it possible to leverage larger datasets. Previous work related to learning from weak labels applied to radiologic image analysis has already demonstrated that deep learning models could be trained to use data with incomplete or inexact annotations to obtain relevant results (e.g. [4]).

One form of weak supervision can be represented by anatomical guidance, highlighting regions of interest rather than the exact location where osteophytes occur. Since osteophytes tend to appear in anatomical regions around the joints, providing a model with information about relevant anatomical structures may help it focus on clinically meaningful regions. There have been some studies that incorporated location-specific information in the detection of the presence of osteophytes and showed that this could be a good approach when working with 2D grayscale X-rays. For example, Daneshmand et al. [5] defined regions of interest for the bones that were fed into the ResNet18 for predictions. This paper investigates highlighting potential regions where osteophytes occur (rather than the whole bone) and including them as additional information in the actual ResNet18 architecture pipeline.

Therefore, the main question of this research paper is to find the following: **Can anatomical priors improve the spatial plausibility and robustness of weakly supervised osteophyte localization under limited image-level supervision?**

We investigate this in experiments on hip X-ray data from two public osteoarthritis datasets. After the initial preprocessing of the data, the different models are trained to detect osteophytes on various amounts of data (100%, 90%, 75%, 50%, 25%) and both prediction and localization are compared between the modified model and its baseline variant.

2 Methodology

2.1 Architectural choices

Several deep-learning methods have been proposed to detect or grade hip osteoarthritis on X-ray images (e.g. [6, 7]). However, incorporating region awareness (for where osteophytes can occur from such X-

rays) alongside weak data into measuring the robustness of a model to see if this helps predictions remains unexplored.

This paper investigates whether adding anatomical guidance provided by BoneFinder [8] actually improves the robustness of a model when predicting if there is an osteophyte in a hip image by looking at clinically relevant regions where osteophytes can occur.

BoneFinder is an automatic landmark detection tool for radiographic images that estimates anatomically meaningful landmark points around the bones in the X-rays. These landmark points provide a representation of the underlying anatomy and can be used as prior information indicating clinically relevant regions. Therefore, we use these landmark points to derive a guidance matrix that would be used as anatomical guidance in the network. A more comprehensive representation of how the guidance is obtained is done in Section 3.2.

This is done by using a modified Convolutional Neural Network (CNN). The tested hypothesis is that adding a BoneFinder-based matrix can compensate for reducing training data, achieving a comparable overlap and accuracy/MCC (mentioned in section 3.6) with the original 3-channel ResNet18.

The architecture change that we employ is modifying the first convolutional layer of the network to accept a different number of channels (such as condensing the standard three image channels into one, or adding an extra channel for spatial guidance). The rest of the architecture remains the same for the other layers (including the initialization of the pretrained weights). This change is possible as the results of the computations after the first layer are combined before the information goes into the next layer. By modifying the number of channels, we allow the introduction of different data into the network by one of the channels. This idea has already been explored in literature when additional information was needed. For example Dwivedi et al. [9] use an additional channel to add Near-infrared information to the RGB ResNet model.

2.2 Modifying pretrained network weights

The models used in this work are initialized from pretrained ResNet18 [10] weights. However, the original ResNet18 architecture expects a three-channel RGB input, whereas the X-ray images used in this study are grayscale. Furthermore, some experiments incorporate an additional guidance channel, resulting in a four-channel input. Modifying the number of input channels creates a mismatch between the pretrained first-layer weights and the new input representation. Therefore, an initialization strategy is required that allows the network to benefit from the pretrained weights while accommodating the modified input structure.

For a single spatial location of a 3-channel RGB model, the output of the first layer is described by the formula:

$$y = w_R x_R + w_G x_G + w_B x_B \tag{1}$$

Here w_X represents the weight for channel X and x_X is the input given in that channel.

A common initialization when working with a 3-channel model and grayscale X-ray images is to triple the image into the 3 channels to mimic the RGB structure. Therefore the formula above becomes:

$$y = w_R x + w_G x + w_B x = x(w_R + w_G + w_B) \tag{2}$$

From this equation we can see that we can simplify this model to a one channel model that has its weight set to the sum of the weights that would be given for the three-channel, keeping the initial mathematical model the same (before backpropagation which will update three weights for the three-channel model and only one for the one-channel). Moreover, extending this model to a 4 channel architecture where the initial weight of the new channel is 0 also preserves at the first epoch the mathematical model:

$$y = w_R x + w_G x + w_B x + 0 * z = w_R x + w_G x + w_B x \tag{3}$$

Aside from the same initialization in the first epoch as the model without guidance to keep the same mathematical model, initializing with weight zero has several advantages. If we initialized with some other weight the model might try to guess complex textures which are nonexistent in our proxy guidance information or the weights for the new channel might reach to dominate the pretrained weights of the other channels.

Following this initialization, the weights of the newly added channel are updated via standard backpropagation during training, becoming integrated into the learning process of the existing architecture.

2.3 Models

ResNet18 serves as the backbone architecture in this study. This choice has been made because it provides a strong baseline for image classification that prioritizes simplicity, reproducibility, and interpretability over maximum raw performance. It is a light-weight version in the ResNet family that is used when the dataset is not too big and has been extensively used

in medical imaging (e.g. [5]). A comparison between different architectures for detecting OA from MRI scans using transfer learning (pretrained networks) is also done by Yeoh et al. [11], where the ResNet models (particularly shallower networks like ResNet18 and ResNet34) achieved the highest values for the analyzed metrics.

This model also supports easy modification of its layers and presents convolutional feature maps near the output which are used to observe model attention.

In this paper we train different models to output a binary label (is there an osteophyte or not). Based on this, we analyze the following models based on the first channel initialization:

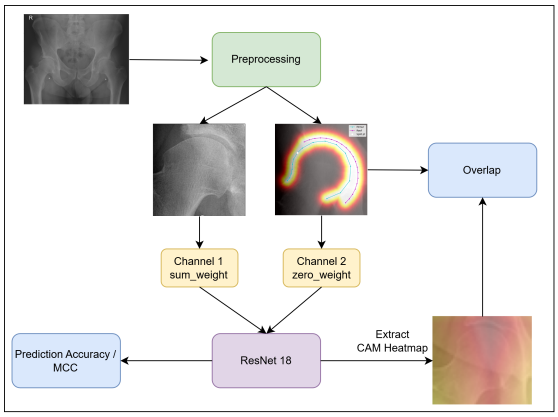
1. 1-channel (1ch) - a common initialization when working with grayscale data. For example, this has been explored by Seiger et al. [12]. When it comes to the initialization of this channel's weights, we set it to the sum of the pretrained RGB first layer weights to keep the initial step consistent with the baseline 3-channel one

2. 2-channel (1ch+guidance) - an image channel and one guidance channel to add the prior information that we need. Again the initialization of the image channel weight is the sum of the RGB weights (so the only change compared to the 1-channel model is adding location information). To start the same as the 1-channel model, we initialize the weight for the guidance channel to zero.

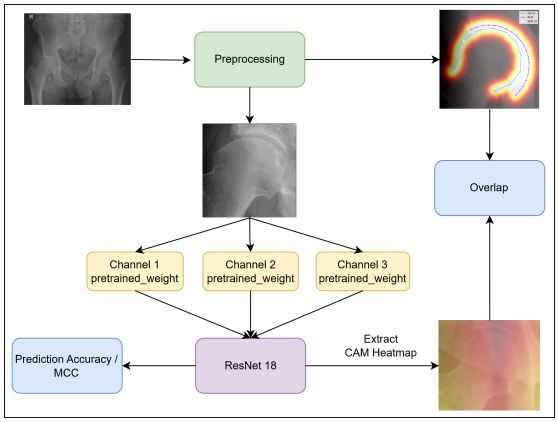
3. 3-channel (3ch) - three channels containing the X-ray information tripled in the three channels. The initialization of the weights is using the standard RGB weights of the pretrained model.

4. 4-channel (3ch+guidance) - three channels containing the X-ray information tripled in the three channels and one guidance channel to add the prior information that we need. The same as above, the weights for the image channels are the standard pretrained RGB ones and for the guidance we start with weight zero.

These modifications of the channels are also described for the 2-channel and the 3-channel original architecture in Figure 1 and for the 1-channel and 4-channel in Appendix A.



(a) The pipeline for the 1ch+guidance modified model



(b) The pipeline for the 3ch original model

Figure 1: Comparison of the 2-channel and 3-channel ResNet pipelines.

3 Experimental setup

3.1 Data

The datasets used in this study are provided by the Osteoarthritis Initiative (OAI; https://nda.nih.gov/oai) and the CHECK [13] datasets, using only the hip X-rays. Both consist of up to 5 scans per patient (5 visits) and each scan may contain different metadata. The split of the data was 70% training, 15% validation and 15% testing, considering the whole dataset.

We also used a .csv file with all the metadata extracted beforehand per patient (empty if some of the data is missing). The values analyzed in this paper alongside the unique identifier are the visit, the side (left or right indicating the side of the hip), and the grades for the following four regions where hip osteophytes can occur: superior femur, inferior femur, superior acetabulum, inferior acetabulum. Alongside

these files, we also had access to the pointfiles associated with these datasets (files containing landmark coordinates), generated by the BoneFinder tool.

The analyzed models are adjusted to make binary predictions (is there an osteophyte in an image or not). This binary prediction is derived from the metadata in the images. As the X-rays offer information about the grade of an osteophyte in different regions (OARSI grades 0-3), our classification metric was defined as "1" if any of the four target regions have an osteophyte with grade at least 1 ($\max(Z1, Z2, Z3, Z4) > 0$) and "0" if the zones do not present an osteophyte ($\max(Z1, Z2, Z3, Z4) = 0$). The images that do not have complete labels and the only present labels have a score of zero are therefore omitted as we cannot guarantee if an osteophyte is there or not and would be costly to label a patient as healthy when they have an osteophyte present.

3.2 Preprocessing

Since the datasets used contain large hip X-ray images taken from the OAI and CHECK datasets, and the standard pretrained ResNet18 requires input images of 224x224 pixels, some preprocessing is needed.

First, the X-ray image format is verified. The image must contain some information about the pixel spacing (size) and the width of a pixel must be equal to its length. Then images are transformed into Monochrome2 format (white = maximum intensity, black = minimum intensity) or ignored if the original one was not either Monochrome1 or Monochrome2.

In order to prepare the images, the BoneFinder tool is used. This is present to give precise bone location landmarks as no location information is present in the analyzed data. The second step in the preprocessing pipeline is normalizing the pixels of the images as different scanners might output different pixel spacing, so all the images have been rescaled to 0.4mm/pixel.

Given the location of the femoral head by BoneFinder, the third step of the preprocessing is cropping the image to be the size of 224x224 pixels centered on the femoral head (found by using the algorithm of Algebraic circle fit by Taubin [14]). This produces two images (for right and left hip) and in order to have a similar image structure we flip the left hip image.

3.3 Guidance generation

The anatomical guidance is obtained using the BoneFinder tool. Points are sampled around the femoral head and on the circle defined by the acetabular roof. They are used to draw a region of interest (this represents the most likely region where osteophytes can occur). This is done for every image so the anatomical prior varies slightly, but the resulting matrix is be the same size as the cropped image. Analyzing different images led to choosing 16 points from the BoneFinder proximal femur ones (range 13-28) and sampling equally spaced points on the fitted circle for the sourcil (range -0.65π radians to 0.27π radians). To generate the spatial guidance information from these sampled points, a binary matrix is first constructed. This is done by connecting 3 regions: a line represented by the first 7 points of the femur, a polygon connecting the sourcil points reversed and the points 19 to 27 of the femur and a line representing the last 2 points of the femur. From this, the values of the pixels are calculated as a Euclidean Distance to the closest pixel with value "1" in the matrix. Then the matrix values are remapped into the [0,1] range via a Gaussian function where the standard deviation was initialized to a baseline of $\sigma = 8$, where "0" distance corresponds to a value of "1", and the value decays smoothly close to 0 as distance grows. A complete visualization of the preprocessing pipeline can be found in Appendix B.

3.4 Experimental design

The steps in the pipeline are the following: images are preprocessed and are split into training, validation and testing. Then different modified models that include anatomical guidance from BoneFinder are trained to predict the binary label. From the last convolutional layer, the feature maps for each image are extracted to be used to generate a Grad-CAM heatmap for the image that is used to compute the overlap metric. The results of this first phase of experiments determine whether the model learns from the addition of the guidance information in terms of the MCC and overlap. After training the models, the checkpoint with the smallest validation loss is chosen to compute the metrics for. We then varied the amount of training data (validation and testing remained the same to give conclusive results under the same conditions). The percentages of training data also analyzed are 90%, 75%, 50% and 25%. By observing how our metrics evolve across these splits, it is determine whether BoneFinder guidance can compensate for reduced training data. The pipeline can also be seen in Figure 1 and Appendix A.

3.5 Parameters

The parameters that have not been discussed so far are the ones used for training. Therefore, all of the

training has been done using 30 epochs with validation after each one and a stopping condition if the validation loss has not improved for 7 epochs (experimentally it was found that the model reaches the best validation loss after less than 10 epochs). The chosen learning rate for the models is 0.001 and the batch size is set to 32. The Adaptive Moment Estimation (Adam) optimizer has been used to update the model's parameters.

The loss function is Binary Cross Entropy with Logits. This is chosen to be minimized because the overlap metric is only computed during evaluation. Furthermore, MCC depends on thresholded class predictions. As a result, two models may obtain the same MCC despite producing substantially different confidence scores. In contrast, Binary Cross Entropy operates directly on the predicted logits and provides a continuous optimization objective that can be efficiently minimized using gradient-based methods. The formula for this loss is:

$$\mathcal{L}(z, y) = -\frac{1}{N} \sum_{i=1}^{N} \Big[y_i \cdot \log \sigma(z_i) + (1 - y_i) \cdot \log(1 - \sigma(z_i)) \Big] \tag{4}$$

where $\sigma(z_i) = \frac{1}{1+e^{-z_i}}$, z_i is the raw logit, $y_i \in \{0, 1\}$ is the ground truth label, and N is the batch size.

3.6 Metrics

The analyzed metrics are based on different measurements based on the predictions of the model, the rates of true positives (tp), true negatives (tn), false positives (fp), false negatives (fn) and mean overlaps. Therefore those were:

- **accuracy** representing how many images were correctly classified as having or not having osteophytes
- **overlap function** - used to see how much of the model's attention (Grad-CAM heatmap) is within the region of interest defined by the BoneFinder landmarks. This is defined as $\frac{\sum_{i,j} H(i,j) \cdot M(i,j)}{\sum_{i,j} H(i,j)}$, where H is the normalized Grad-CAM, M is the binary guidance matrix (value in the matrix above 0.5 become 1 and the rest are 0) and i and j are coordinates. This metric was chosen regardless of the prediction outcome as even when the model predicts that there is no osteophyte in the image it should look at the specified region to reach a conclusion
- **Matthews correlation coefficient** (MCC) defined as $\frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp+fp)(tp+fn)(tn+fp)(tn+fn)}}$ - used as it is a metric that penalizes systematic bias (ex. always predicting negatives) especially useful for imbalanced classes, being considered one of the most informative metrics

These metrics were chosen due to the nature of the datasets (imbalanced datasets) and because most of them are standard metrics used in medical applications [15].

3.7 Resources and tools

To help with the model training and storage of the data, the DelftBlue High-Performance Computing (HPC) cluster [16] has been used, mainly the "compute" partition for CPU jobs (48 CPUs and 185 GB of RAM per node) and the "gpu-a100-small" partition for small GPU jobs (no more than 1 GPU with a maximum of 10 GB video RAM, and no more than 2 CPU cores). So all non-training related scripts were run on the CPU partition (preprocessing, splitting the data into train, test and validation) and the rest on the GPU partition (training, metrics calculation).

The code for this project was implemented in Python. Execution on the DelftBlue supercomputer was facilitated using Bash scripts. The implementation is based on the PyTorch deep learning framework with PyTorch Lightning used for structured training and experiment management. Additional libraries include NumPy for numerical computations and pytorch-gradcam for visualizing model attention maps.

4 Results

4.1 Main results

To keep the comparison fair, we looked at the comparisons between 1ch and 1ch+guidance followed by comparisons between 3ch and 3ch+guidance. This is done as the pairs described have the same base architecture, the only difference being the addition of the guidance information.

First, for classification (Figure 2), comparing 1ch and 1ch+guidance, the latter performed slightly better or the same as 1ch in terms of accuracy (the exception is when we had 25% data). MCC mostly mirrored this behavior of the models, the difference being that at 100% where 1ch achieved a slightly better value. An interesting observation is that both models achieved slightly higher MCC at 75% of the training data than at 90%. Since the MCC trend

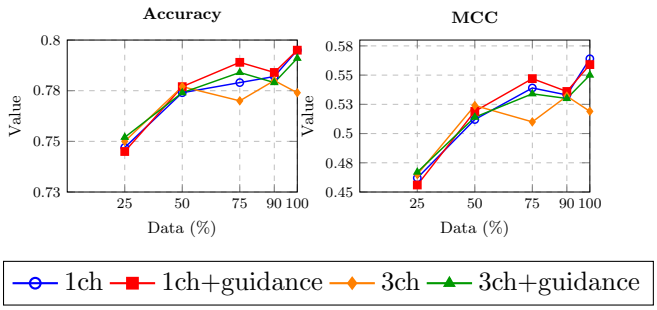


Figure 2: Classification performance across different training set sizes for the four models analyzed

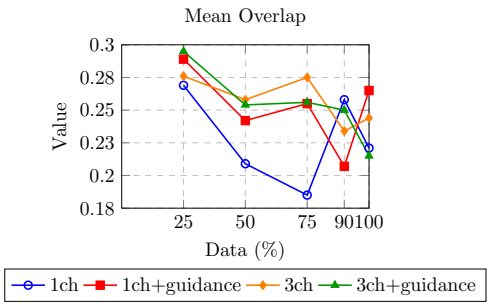


Figure 3: Localization performance measured using mean overlap for the four models analyzed

is not strictly monotonic with respect to dataset size, this suggests that factors such as the particular train-validation-test split or optimization dynamics may have influenced the results in addition to the amount of available training data. However, the improvements obtained by adding the guidance channel are relatively small, particularly at lower data percentages. This suggests that the additional guidance information alone may be insufficient to fully compensate for the reduction in available training data.

Second, comparing 3ch and 3ch+guidance, some improvements can be observed both in terms of accuracy and MCC when adding the guidance (the 3ch+guidance model performs better at 25% , 75% and 100%). Interestingly, the 3ch+guidance model trained on 75% of the data achieved higher performance than the 3ch model trained on both 90% and 100% of the data. This indicates that the BoneFinder guidance provides complementary information that can partially offset the negative impact of reduced training data.

The guidance had a higher impact for 3ch+guidance compared to 3ch than 1ch+guidance compared to 1ch as the curves are closer together than the ones with multiple channels, suggesting that the effectiveness of the guidance information may depend on the underlying image representation.

Comparing all architectures, the 1ch+guidance model generally achieved the strongest performance across most data percentages. This indicates that the addition of BoneFinder guidance can provide a meaningful benefit for the classification task and may represent an effective way of incorporating anatomical prior knowledge into the model.

For localization, overlap generally decreased at higher data percentages (Figure 3), although the trend is not strictly monotonic. With only 25% of the data, the network may rely on broader image regions, resulting in higher overlap with the predefined region of interest, despite lower classification performance. As more training data become available, the learned representations appear to become more focused on discriminative features, which may reduce agreement with the predefined regions. This suggests that localization overlap and classification performance do not necessarily improve together.

On one hand, when we compare the respective architecture pairs we observe that 1ch+guidance achieved greater overlap than 1ch for most of the percentages, suggesting that the addition of guidance information improves the spatial alignment between the model attention and the regions of interest for most data percentages (except 90%). On the other hand, the effect of the guidance became less consistent in the 3ch–3ch+guidance setting (3ch slightly outperformed 3ch+guidance in terms of overlap for several data percentages, although the difference was not consistent), suggesting that the interaction between three-image inputs and guidance information may be more complex than in the one-image case.

Finally, in some cases, lower MCC values coincide with higher overlap scores (and vice versa), suggesting a potential trade-off between classification performance and localization performance. However, this relationship is not consistent across all models and data percentages, and should therefore be interpreted with caution.

Full tables with the results including standard deviations are available in Appendix C.

4.2 Initializations

As an additional experiment we have tested how the initialization for the 3ch model changes if we choose the weights for the 3 channels not to be the standard pretrained weights, but a mean of those (3ch_mean). The initial mathematical model becomes:

$$y = 3*\left(\frac{w_R + w_G + w_B}{3}\right)*x = x(w_R+w_G+w_B) \quad (5)$$

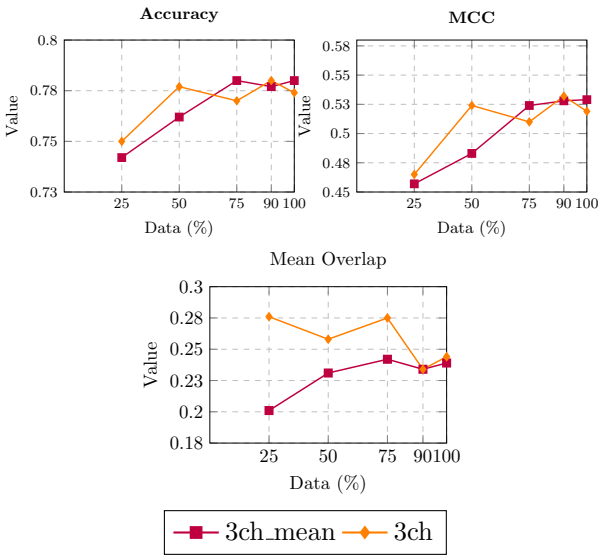


Figure 4: Classification performance (Accuracy and MCC, top) and localization performance (Mean Overlap, bottom) across varying training set sizes for the 3ch and 3ch_mean models.

The differences between the 3ch and 3ch_mean models are more pronounced in the localization task than in classification performance (Figure 4). In particular, the 3ch model consistently achieved higher mean overlap across most data percentages, suggesting that pretrained RGB weights may provide a more effective initialization for learning spatially meaningful features aligned with the region of interest.

In terms of accuracy and MCC, the differences between both models were less consistent, with neither variant showing a clear advantage across all training set sizes. Notably, the 3ch_mean model remained competitive and even slightly exceeded the standard 3ch model in some higher-data settings.

Overall, these results indicate that while pretrained RGB initialization may provide benefits for learning spatial localization patterns, its impact on classification performance is less pronounced. However there is no clear evidence that the 3ch_mean model outperforms the 3ch model.

The full measurements with standard deviations for the 3ch_mean model can be found in Appendix D

4.3 Missing guidance

Our evaluation setup relies on the continuous availability of BoneFinder guidance, so anatomical prior information is incorporated into all test set images. We also looked at what happens when we remove this guidance. Such a comparison for the 1ch+guidance model and 3ch+guidance model compared to their no image variants (2ch_nobf and 4ch_nobf) can be seen in Figure 5.

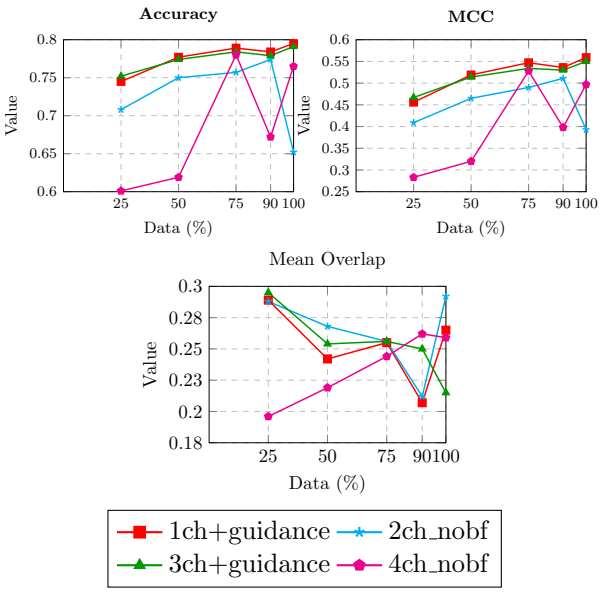


Figure 5: Classification performance (Accuracy and MCC, top) and localization performance (Mean Overlap, bottom) across varying training set sizes for the 1ch+guidance, 2ch_nobf, 3ch+guidance and 4ch_nobf models.

We see that both the models having the image removed at test time acquired a lower value for accuracy and MCC at every percentage of data. This suggests that the models learn to rely on the guidance channel during training and are negatively affected when this information is no longer available at test time. In contrast, the overlap metric was considerably less affected by the removal of the guidance information. For the 2ch_nobf model, overlap remained comparable to, and in some cases exceed that of the original 1ch+guidance model. A similar trend was observed for the 4ch_nobf model at higher data percentages. These findings indicate that the absence of guidance primarily impacts classification performance, while its effect on localization behavior is less pronounced.

5 Responsible research

5.1 Ethical Considerations

The CHECK and OAI datasets contain medical imaging data derived from human subjects and therefore require careful handling to protect patient privacy and confidentiality. Although the datasets were made

available for research purposes, they contain sensitive health-related information. To minimize privacy risks, all data used in this study were anonymized prior to access. Furthermore, the data were stored and processed exclusively on the DelftBlue computing cluster, which provides controlled access to authorized researchers. The datasets were not distributed to unauthorized third parties, and no attempts were made to identify individual participants from either the original radiographs or the processed data. These measures were taken to ensure that the research complied with the ethical requirements associated with handling medical imaging data.

A second ethical consideration concerns the potential use of the developed models in clinical practice. While the proposed approach investigates whether anatomical priors can improve the localization of clinically relevant regions and support osteophyte detection, the obtained results do not demonstrate sufficient reliability for autonomous clinical decision-making. Osteophyte detection remains a challenging task, particularly for subtle early-stage cases, and prediction errors may occur. Consequently, the model should be viewed as a research tool that may provide supporting information to clinicians rather than as a replacement for expert medical judgment. Further validation, robustness testing, and clinical evaluation would be required before such a system could be considered for deployment in a real-world healthcare setting.

5.2 Reproducibility of the method

All the experiments were performed using the same seeds everywhere (123, 42 and 2026) to make results reproducible. Therefore splitting and sub-sampling the training set could be done the same every time to use the same data. The hyperparameters for training were also fixed and only the model architecture and amount of training data has been changed. One limitation could come from the fact that the research has been done using only three fixed seeds due to the huge amount of time taken for training the different models. Also, only a limited amount of hyperparameter variation has been analyzed due to these computational constraints.

To make this research reproducible, all steps have been documented (including preprocessing and the parameters used) and can be followed to obtain the same results.

To aid in reproducibility and validation of the research, the model was configured to save predictions, checkpoints, per batch and per epoch loss and accuracy metrics, and metrics for all of the analyzed models and percentages of data.

Some minor nondeterminism could be introduced by the GPU operations, making results vary slightly and different versions of the used frameworks and libraries might lead to changes in performance.

5.3 Dataset biases

The provided datasets pose an unbalanced split of patients having and not having osteophytes, guiding the model towards predicting the negative class more often than the positive one. That is one reason why MCC was analyzed alongside accuracy.

The experiments do not explicitly account for demographic or clinical factors such as age, sex, ethnicity, body mass index, disease history, or other parameters that might be relevant in the appearance of osteophytes. Therefore the results should be taken with care as the X-rays can come from people with similar backgrounds and not be representative for the entire population of the globe. Since the data comes from specific cohorts, the performance might also not generalize to scanners that differ substantially from those used in this research.

The data used was labeled by experts so it is subject to human error and interpretability. Different experts might have slightly different grading criteria which can therefore introduce some bias. Moreover, less severe cases might get mislabeled more often introducing uncertainty into the ground truth labels.

5.4 Use of Generative AI

This work has been aided in some of the parts by Generative AI models. The artificial intelligence was used to clarify some topics that were not well understood from the found papers, for connecting topics, for checking grammar mistakes or rephrasing ambiguous phrases, for helping with LaTeX formatting and as an aid for the coding process. All of the AI contribution has been reviewed and verified and it has not been used to make scientific decisions.

6 Discussion and Conclusion

This paper investigated whether anatomical priors obtained from BoneFinder can improve weakly supervised osteophyte classification and localization under image-level supervision. The results show that incorporating guidance information generally improves classification performance, with guided models achieving higher accuracy and MCC than their counterparts with the same underlying architecture. However, these improvements are relatively modest

and do not fully compensate for reductions in training data, indicating that anatomical priors should be viewed as complementary information rather than a replacement for additional annotated samples.

The impact of guidance on localization was less consistent. While some guided models achieved higher overlap with the predefined regions of interest (e.g. 1ch+guidance), the improvements were more variable than those observed for classification performance. Furthermore, experiments comparing different initialization strategies suggest that preserving pretrained RGB weights is more beneficial for localization than collapsing the pretrained filters into a single averaged representation.

An additional analysis showed that removing the BoneFinder guidance during testing leads to a noticeable decrease in classification performance while having a considerably smaller effect on localization overlap. This suggests that the guidance helps the model identify finer details, rather than changing which part of the body it looks at.

Overall, the results indicate that anatomical priors can provide a meaningful benefit for weakly supervised osteophyte analysis, particularly for classification, but cannot consistently substitute for missing data. The effectiveness of these priors depends on both the underlying architecture and the availability of guidance information during inference, highlighting important considerations for practical deployment.

One important limitation of the experiments could come from the fact that BoneFinder does not represent the ground truth, and on some images this could give wrong results, therefore damaging the learning of the model. BoneFinder also does not give osteophyte annotations. That's why it is used to define the plausible regions where osteophytes could occur. These regions again do not represent ground truth, so there could be a better delimitation done by an expert in the medical field by defining different boundaries or how the guidance matrix degrades further from these regions of interest. Guidance matrices depend on how BoneFinder defines the points, and the sampling for the points of this region might differ slightly between patients leading to defining less accurate regions. Feeding these regions to the model could also be a source of bias as osteophytes might be correlated with features outside of the zone of interest that are omitted in the current structure of the guidance matrices.

The overlap metrics are calculated using proxy regions as well (no ground truth defined by clinicians), therefore this value should be interpreted as an indicator of anatomical plausibility rather than precise localization accuracy.

There could be an anatomical localization bias as some regions might be more prone to osteophytes, therefore the model could learn to identify those regions and not the actual osteophytes. This could also happen due to the fact that the data also poses an imbalance in the grade of the osteophytes (more having an OARSI grade 1 than 3). One more reason for this could be the image itself that captures some regions better than the others.

To further improve this research, different experiments with varying hyperparameters should be done (e.g. different learning rate, different weight for the matrix guidance, different values for sampling the points for the guidance matrix, different rates of decay for the guidance matrix) to reach a more comprehensive result. The presented experiments should be repeated for multiple seeds (more than 3) so a better average could be made. Since this approach analyzed only the OAI and the CHECK datasets, including data from other sources might be useful to demonstrate that the network does not learn to generalize only on the analyzed group of people.

Another change could include exploring more options of modifying the ResNet architecture. Studying the influence of multiple channels with different weights could lead to more surprising results and possibly to a better understanding if this incorporation of the guidance actually helps with weakly supervised learning. Also a dual architecture with two CNN models, one processing the image and one processing the guidance information and with their outputs combined might provide a more efficient solution.

More work should be done to address the existence of BoneFinder guidance. Since this tool might not always be available, measures should be taken to be able to use this method. One thing could be training the model with a lower percentage of BoneFinder information (so not all the images have guidance), so that when the new images without BoneFinder come, the network would be ready to classify those with a great probability of success.

References

[1] Theo Vos et al. "Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015". In: *The Lancet* 388.10053 (Oct. 8, 2016), pp. 1545–1602. ISSN: 0140-6736, 1474-547X. DOI: 10.1016/S0140-6736(16)31678-6. URL: https://www.thelancet.

com/journals/lancet/article/PIIS0140-6736(16)31678-6/fulltext.

[2] Xuefeng Luo et al. "A cross-sectional study unveiling the global impact and future projections through 2060 of osteoarthritis". In: *European Journal of Public Health* 36.2 (Mar. 14, 2026), ckaf087. ISSN: 1464-360X. DOI: 10.1093/eurpub/ckaf087.

[3] R. D. Altman and G. E. Gold. "Atlas of individual radiographic features in osteoarthritis, revised". In: *Osteoarthritis and Cartilage* 15 (Jan. 1, 2007), A1–A56. ISSN: 1063-4584, 1522-9653. DOI: 10.1016/j.joca.2006.11.009. URL: https://www.oarsijournal.com/article/S1063-4584(06)00328-1/fulltext.

[4] Leo Misera et al. "Weakly Supervised Deep Learning in Radiology". In: *Radiology* 312.1 (July 2024), e232085. ISSN: 0033-8419. DOI: 10.1148/radiol.232085. URL: https://pubs.rsna.org/doi/10.1148/radiol.232085.

[5] Mitra Daneshmand et al. "Deep learning based detection of osteophytes in radiographs and magnetic resonance imagings of the knee using 2D and 3D morphology". In: *Journal of Orthopaedic Research* 42.7 (2024), pp. 1473–1481. ISSN: 1554-527X. DOI: 10.1002/jor.25800.

[6] Shenghao Xu et al. "Development and validation of a deep learning model for automatic severity grading of hip osteoarthritis: a multicenter study". In: *Annals of Medicine* 57.1 (), p. 2584361. ISSN: 0785-3890. DOI: 10.1080/07853890.2025.2584361. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12604132/.

[7] Yanping Xue et al. "A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis". In: *PLOS ONE* 12.6 (June 2, 2017), e0178992. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0178992. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178992.

[8] C. Lindner et al. "Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting". In: *IEEE Transactions on Medical Imaging* 32.8 (Aug. 2013), pp. 1462–1472. ISSN: 1558-254X. DOI: 10.1109/TMI.2013.2258030. URL: https://ieeexplore.ieee.org/document/6497663.

[9] Paritosh Dwivedi et al. "A Comparative Analysis of Deep Learning Architectures for Satellite Image Classification: Integrating Performance, Uncertainty, and Optimization Geometry". In: *IEEE Geoscience and Remote Sensing Letters* 23 (2026), pp. 1–4. ISSN: 1558-0571. DOI: 10.1109/LGRS.2026.3680631. URL: https://ieeexplore.ieee.org/document/11474619.

[10] Kaiming He et al. *Deep Residual Learning for Image Recognition*. Dec. 10, 2015. DOI: 10.48550/arXiv.1512.03385. arXiv: 1512.03385[cs.CV]. URL: http://arxiv.org/abs/1512.03385.

[11] Pauline Shan Qing Yeoh et al. "Transfer learning-assisted 3D deep learning models for knee osteoarthritis detection: Data from the osteoarthritis initiative". In: *Frontiers in Bioengineering and Biotechnology* 11 (Apr. 13, 2023), p. 1164655. ISSN: 2296-4185. DOI: 10.3389/fbioe.2023.1164655. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10136763/.

[12] René Seiger and Peter Fierlinger. "Predicting Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using a Vision Transformer and Hippocampal MRI Slices". In: *Bioengineering* 13.2 (Jan. 29, 2026), p. 163. ISSN: 2306-5354. DOI: 10.3390/bioengineering13020163. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12938448/.

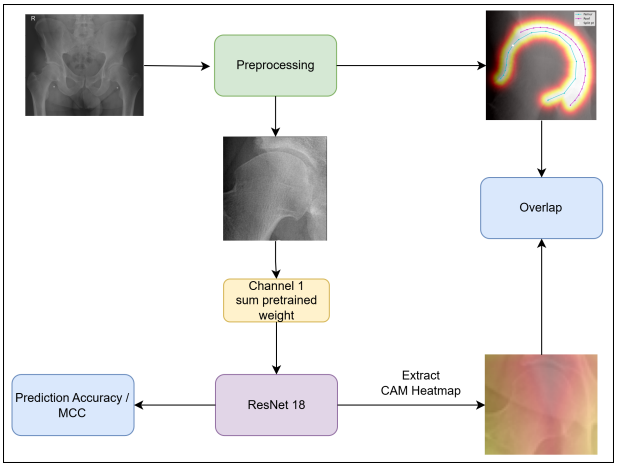
[13] J. W. J. Bijlsma and J. Wesseling. *CHECK (Cohort Hip & Cohort Knee) data of baseline and 6 to 8 years follow-up*. June 5, 2025. DOI: 10.17026/DANS-ZC8-G4CW. URL: https://lifesciences.datastations.nl/dataset.xhtml?persistentId=doi:10.17026/DANS-ZC8-G4CW.

[14] G. Taubin. "Estimation of planar curves, surfaces, and nonplanar space curves defined by implicit equations with applications to edge and range image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.11 (Nov. 1991), pp. 1115–1138. ISSN: 1939-3539. DOI: 10.1109/34.103273. URL: https://ieeexplore.ieee.org/document/103273.

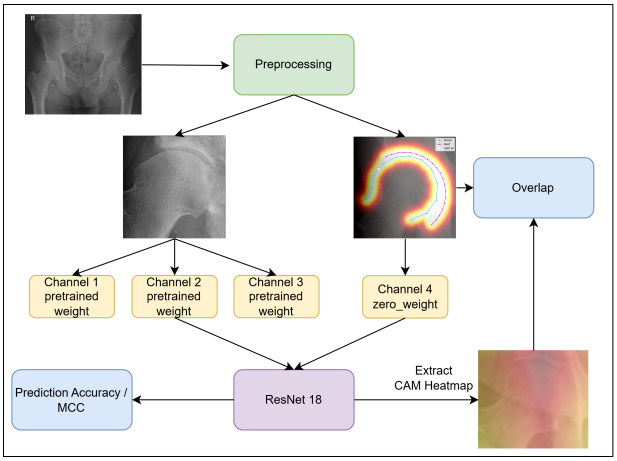
[15] Steven A. Hicks et al. "On evaluation metrics for medical applications of artificial intelligence". In: *Scientific Reports* 12 (Apr. 8, 2022), p. 5979. ISSN: 2045-2322. DOI: 10.1038/s41598-022-09954-8. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC8993826/.

[16] *DelftBluePhase2*. TU Delft. URL: https://www.tudelft.nl/dhpc/ark/delftbluephase2.

A Pipelines



(a) The pipeline for the 1-channel modified model



(b) The pipeline for the 4-channel original model

Figure 6: Pipelines for the 1-channel and 4-channel models

B Preprocessing pipeline

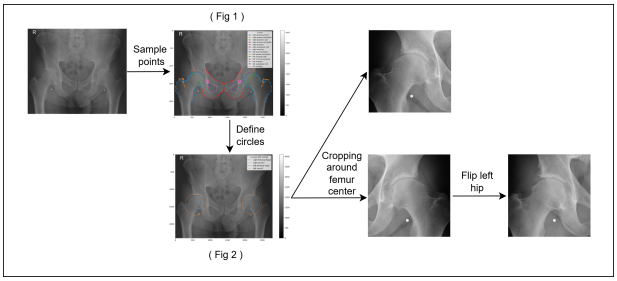


Figure 7: The preprocessing pipeline going from the full image to the region crops. (Fig 1) and (Fig 2) can be seen below.

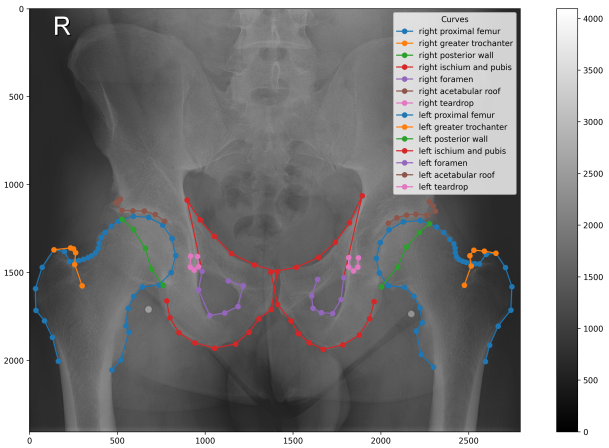


Figure 8: (Fig 1) in the diagram above representing point annotations derived from BoneFinder.

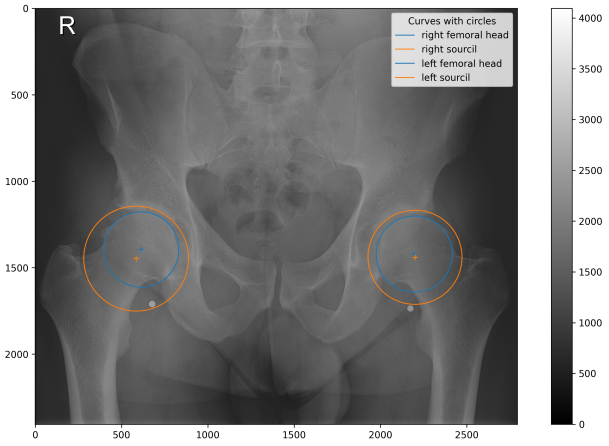


Figure 9: (Fig 2) in the diagram above representing the circles defined by the femur and acetabular roof.

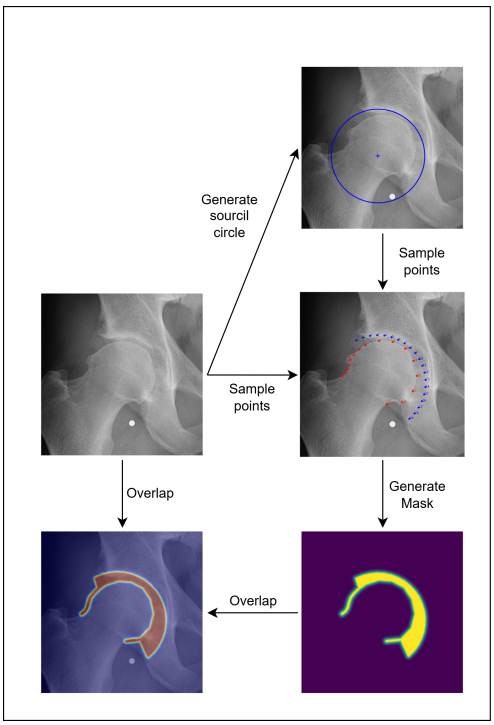


Figure 10: The definition of the guidance information

C Full results

Table 1: Performance metrics with standard deviations across different training set sizes for the **1ch** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.795 ± 0.009	0.564 ± 0.017	0.221 ± 0.022
90	0.782 ± 0.003	0.533 ± 0.007	0.258 ± 0.018
75	0.779 ± 0.022	0.539 ± 0.025	0.185 ± 0.049
50	0.774 ± 0.009	0.512 ± 0.021	0.209 ± 0.033
25	0.747 ± 0.009	0.462 ± 0.006	0.269 ± 0.028

Table 2: Performance metrics with standard deviations across different training set sizes for the **1ch+guidance** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.795 ± 0.011	0.559 ± 0.024	0.265 ± 0.022
90	0.784 ± 0.011	0.536 ± 0.024	0.207 ± 0.002
75	0.789 ± 0.009	0.547 ± 0.021	0.255 ± 0.042
50	0.777 ± 0.007	0.519 ± 0.014	0.242 ± 0.030
25	0.745 ± 0.004	0.456 ± 0.003	0.289 ± 0.091

Table 3: Performance metrics with standard deviations across different training set sizes for the **3ch** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.774 ± 0.004	0.519 ± 0.014	0.244 ± 0.045
90	0.780 ± 0.005	0.532 ± 0.017	0.234 ± 0.035
75	0.770 ± 0.017	0.510 ± 0.028	0.275 ± 0.030
50	0.777 ± 0.003	0.524 ± 0.008	0.258 ± 0.017
25	0.750 ± 0.005	0.465 ± 0.008	0.276 ± 0.071

Table 4: Performance metrics with standard deviations across different training set sizes for the **3ch+guidance** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.791 ± 0.008	0.550 ± 0.018	0.215 ± 0.033
90	0.779 ± 0.007	0.530 ± 0.016	0.250 ± 0.023
75	0.784 ± 0.005	0.534 ± 0.011	0.256 ± 0.037
50	0.774 ± 0.003	0.514 ± 0.011	0.254 ± 0.054
25	0.752 ± 0.006	0.467 ± 0.010	0.295 ± 0.051

D Full results discussion models

Table 5: Performance metrics with standard deviations across different training set sizes for the **3ch_mean** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.780 ± 0.018	0.529 ± 0.036	0.239 ± 0.032
90	0.777 ± 0.003	0.528 ± 0.008	0.234 ± 0.043
75	0.780 ± 0.007	0.524 ± 0.014	0.242 ± 0.066
50	0.762 ± 0.006	0.483 ± 0.015	0.231 ± 0.029
25	0.742 ± 0.026	0.457 ± 0.039	0.201 ± 0.052

Table 6: Performance metrics with standard deviations across different training set sizes for the **2ch_nobf** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.652 ± 0.085	0.393 ± 0.084	0.292 ± 0.074
90	0.774 ± 0.012	0.511 ± 0.027	0.212 ± 0.015
75	0.757 ± 0.020	0.490 ± 0.035	0.256 ± 0.021
50	0.750 ± 0.013	0.465 ± 0.029	0.268 ± 0.022
25	0.708 ± 0.038	0.409 ± 0.045	0.288 ± 0.076

Table 7: Performance metrics with standard deviations across different training set sizes for the **4ch_nobf** model.

Data (%)	Accuracy	MCC	Mean Overlap
100	0.765 ± 0.027	0.497 ± 0.055	0.259 ± 0.060
90	0.672 ± 0.126	0.398 ± 0.126	0.262 ± 0.038
75	0.780 ± 0.003	0.528 ± 0.008	0.244 ± 0.045
50	0.619 ± 0.145	0.320 ± 0.149	0.219 ± 0.037
25	0.601 ± 0.155	0.283 ± 0.195	0.196 ± 0.117