# Improving Generalizability in X-Ray Segmentation of the femur

## Evaluating the Impact of Traditional Data Augmentation Techniques on the generalizability across Datasets

**Roland Bockholt**

**Supervisor(s): Jesse Krijthe, Gijs van Tulder, Myrthe van den Berg**

**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, 2024

Name of the student: Roland Bockholt
Final project course: CSE3000 Research Project
Thesis committee: Jesse Krijthe, Gijs van Tulder, Myrthe van den Berg , Xucong Zhang

## Abstract

An accurate segmentation model for hip components could improve the diagnosis of Osteoarthritis, a prevalent age-related condition affecting joints. A significant challenge in developing effective and robust segmentation models are the domain differences across various datasets. In this study, we investigate the impact of different data augmentation and preprocessing techniques on the generalizability of femur segmentation models across datasets. Using two labeled datasets, we evaluate the performance of a U-Net segmentation model, focusing on the effectiveness of augmentations like image flipping, random rotations, blur, contrast, and brightness adjustments. Our findings reveal that certain augmentations, particularly random rotations of up to 15 degrees, vertical image flipping and light blurring, significantly improve the model's generalization to another data set, reducing boundary errors and enhancing segmentation accuracy. These results underscore the potential of targeted data augmentations in developing robust, generalizable models for hip joint component segmentation.


# 1 Introduction

Osteoarthritis (OA) is a prevalent age-related medical condition, characterized by the progressive deterioration of joint structures, particularly in weight-bearing joints such as the knees or hips. The diagnosis of OA is made on the basis of multiple factors, with the Joint Space Width (JSW) remaining the primary criterion for diagnosing OA.The joint space width (JSW) is the distance between the bones in a joint. Clinicians currently measure the JSW manually , which is time-consuming and prone to errors. A more precise and robust measurement of the JSW in radiographs is therefore an essential step in order to improve the accuracy of the diagnosis and tracking the progression of the condition. One important aspect of this are the precise segmentations of the hip joint components that create an accurate bone surface outline.

A common problem when it comes to generalisation of a segmentation model to other data sets are domain differences [1]. Domain differences are variations that can exist between datasets originating from different sources. These differences can significantly affect the performance of machine learning models. Understanding and addressing domain differences is important for developing robust and generalizable models.
In the context of x-ray images, domain differences can arise from different sources. Differences in the equipment used (e.g., different X-ray machines or settings), lighting conditions, and image resolution can introduce variations. For example, two hospitals might use different X-ray machines that produce images with different levels of contrast and noise. Variability in patient populations, such as age, gender, body type, and medical history, can affect the appearance of medical images. Domain differences can lead to a domain shift [2], where a model trained on one dataset performs poorly on another due to the variations between the domains. This can result in reduced accuracy and lower generalizability.
Data augmentation techniques have been shown to avoid over fitting and help generalization by enlarging the data set and introduce sample greater sample diversity [3] [4]. We can use data augmentation methods to help bridge domain differences by manipulating the training data in ways that simulate the variations present in the target domain.
In this paper we will study the effects of different data augmentation techniques on the generalisability of a segmentation model of femurs in X-ray images to different a dataset.

# 2 Methods and Materials

## 2.1 General Approach

To evaluate the effectiveness of augmentation methods in generalizing to a different dataset, we need two labeled datasets and a segmentation model. First, we train the model on one dataset and then test its segmentation accuracy on both the training dataset and the other dataset. If we observe a performance difference, we apply data augmentations to the training dataset, retrain the model on the augmented data, and reevaluate its performance on both datasets.

## 2.2 Data augmentations

Data augmentation offers numerous benefits for generalizing to different datasets. By exposing the model to augmented versions of the same data, it reduces the likelihood of memorizing specific details of the training images. Instead, it learns more generalized features that are useful across various datasets. Data augmentation can introduce variations in the training data through random transformations, mimicking the diversity found in real-world data [5]. This helps the model learn to handle a broader range of scenarios. For example, augmentations can simulate different conditions such as lighting changes, different angles, and noise, which the model might encounter in new datasets. This enhances the model's robustness and performance, even when the test data differs from the training data.
When considering how training datasets can differ, it's important to note that different hospitals use varying equipment and calibration protocols, affecting image focus, brightness, contrast, and noise levels. The position and orientation of the patient within the X-ray image can also vary. Additionally, there is variability in patient populations, such as a higher prevalence of hip implants or differences in gender distribution.
To address these differences, we propose a set of data augmentations that can simulate these variations. For example, augmentations that randomly change blur, brightness, contrast, and noise can help address equipment-related differences. However, factors like differences in patient population cannot be addressed by these augmentations.
Augmentations also force the model to focus on learning invariant and discriminative features. For instance, rotating or

flipping an image helps the model recognize objects and patterns regardless of their orientation or position, enhancing its ability to generalize. Given the limited scope of this project, we will focus on the following data augmentations:

1. Image Flipping
2. Random Rotation
3. Random Blur
4. Random Contrast Adjustments
5. Random Brightness Adjustments

We will apply these augmentations with varying levels of intensity, such as different degrees of rotation and blur, to determine which levels of augmentation work best.

### 2.3 Measuring Differences between Datasets

Given the potential differences we highlighted earlier and the data augmentations designed to address these discrepancies, it is important to quantify these differences in order to asses if they exist and how strong they are. Specifically, we need a method to compare the brightness, contrast, and blurriness between the datasets.
A straightforward approach is to compare the intensity distributions between the two datasets. Differences can be observed by visually inspecting the histogram, as well as by comparing the mean and standard deviation.

As measure for Blur, we can calculate the mean variance of the Laplacian for each data set. The variance of the Laplacian is a widely-used method for quantifying image blurriness [6]. The Laplacian operator highlights regions of rapid intensity change, such as edges. By applying the Laplacian operator to an image, we measure the variability of pixel intensity changes.
High variance can indicate a large number of rapid intensity changes, corresponding to a sharp image with well-defined edges where as low variance can indicate fewer rapid intensity changes, corresponding to a blurry image where edges are less distinct and details are lost. However, there are limitations to using the Laplacian variance metric. The Laplacian operator, is sensitive to noise. This sensitivity can lead to an inflation of variance, falsely indicating sharper image quality.
Additionally, images with extensive uniform regions, such as plain backgrounds, tend to exhibit low variance regardless of their actual blurriness. Moreover, when comparing datasets, variations in image content, lighting conditions, and other factors can influence the computed variance. These variations may confound assessments of image blurriness, complicating the interpretation and comparison of results.

### 2.4 Segmentation Model

The U-Net architecture is a widely used convolutional neural network (CNN) designed for image segmentation tasks [7] . It was originally developed for biomedical image segmentation but has since been applied to various segmentation problems across different domains [8][9][10]. The architecture consists of a symmetric encoder-decoder structure with skip connections between corresponding layers in the encoder and decoder paths [7]. The symmetric encoder-decoder architecture of U-Net allows for the effective extraction of both low-level and high-level features. The encoder path captures the context of the input image, while the decoder path allows for precise localization, important for accurate segmentation of joint components.
Additionally, U-Net employs skip connections between corresponding layers in the encoder and decoder paths, which help preserve spatial information and details[11]. This is particularly important for X-ray images, where small structures and subtle variations must be accurately captured. U-Net is also highly adaptable to different segmentation tasks. Its architecture can be easily modified and extended to incorporate additional layers or different types of input data. Numerous studies and applications have successfully used U-Net for similar tasks, showing its effectiveness. This includes segmentation of bone structures and other anatomical features in radiographs [12][13][14], which resembles our application.

While there are other suitable models such as SegNet[15], DeepLab[16], or adaptations of the UNet architecture like ResU-Net [17] or TransUNet [18], we chose the basic model. This decision was mainly driven by the abundance of available resources, tutorials, and code samples.

### 2.5 Evaluation Metrics

To evaluate the accuracy of model we employ two different metrics: Jaccard index and Hausdorff distance. The Jaccard index J between two sets $A$ and $B$ is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The Hausdorff distance $d_H$ between two sets $A$ and $B$ is defined as:

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \inf_{b \in B} d(a, b), \sup_{b \in B} \inf_{a \in A} d(a, b) \right\}$$

where:

- $\sup$ denotes the supremum (least upper bound),
- $\inf$ denotes the infimum (greatest lower bound),
- $d(a, b)$ represents the distance between points $a$ and $b$ in the metric space.

Given the relatively large area we need to segment, substantial boundary errors can have a relatively small impact on the Jaccard index. Therefore, we include the Hausdorff distance, which emphasizes boundary accuracy and can, in some cases, provide a more insightful measure of performance compared to the Jaccard index.

## 3 Experiment

### 3.1 Data

We utilize two datasets in our study: the OAI [19] and the CHECK [20] datasets. Both datasets contain X-ray images of the hips of osteoarthritis patients. The OAI dataset is significantly larger, with a total of 12294 images, compared to the

CHECK dataset, which includes 3707 images. Images were taken over multiple years, the total number of patients that participated in the CHECK study is 1002, where as in the OAI study, 4796 people participated. For both of the datasets we have obtained the segmentation from Bonefinder [21] which serves as our ground truth. Bonefinder is a fully automatic software tool designed to outline and segment skeletal structures from 2D radiographs.

### 3.2 Preprocessing

For preprocessing, we start by loading the DICOM image. We use the BoneFinder segmentation to crop the image to the general region of the femur. Although we utilize the ground truth for this step, it is worth noting that a model could be trained to perform this cropping, the detailed outlines of the segmentation are harder to obtain than finding the general area of interest. Next, we use the segmentation obtained from BoneFinder to create a binary segmentation mask of both cropped femurs. Finally, we resize both the image and the mask to 256x256 pixels. In 1 we can see an example of the processed image and the corresponding binary mask.

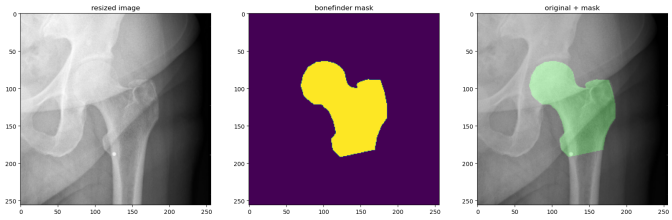


Figure 1: Left:cropped and resized image, Middle: cropped and resized binary mask, Right: mask overlaid on the image

### 3.3 Model Training

The specific UNET-model and code to train the model was taken from a project on Github to segment lungs[22]. The preprocessed images and masks are divided into three sets: a training set comprising 70 percent of the data, a validation set containing 10 percent, and a test set containing 20 percent. We use the negative log likelihood (NLL) as the loss function, a common choice for deep learning models. Each model is trained for 30 epochs using the Adam optimizer, which is known for its efficiency and effectiveness in training deep learning models, with a learning rate of 0.0005. The model is saved whenever the validation loss is smaller than the previous minimum validation loss. From our observations, the validation loss typically starts to rise continuously after about 10 to 20 epochs, so we run the training for 30 epochs to ensure convergence.

### 3.4 Data augmentations

The following data augmentation techniques will be applied to the training data of the CHECK data set in order to asses its influence on the generalisation. For each augmentation, a new model with the specific augmentation will be trained on CHECK and evaluated on CHECK and OAI data set. The augmentation levels for blur, contrast, and brightness adjustment were manually selected to realistically simulate strong, medium, and small changes.

1. Image Flipping: The image and the corresponding are vertically flipped.
2. Random Rotation: Random rotations are applied to both the image and the mask within different ranges
   - small: rotation of up to 5 degree
   - medium: rotations of up to 15 degree
   - high: rotations of up to 25 degree
3. Random Blur: A gaussian blur kernel of size 5 is applied with random sigma values of different ranges.
   - small: sigma between 0.1 and 1
   - medium: sigma between 0.5 and 2
   - high: sigma between 1 and 3
4. Random Contrast Adjustment: the brightness of an image is scaled by a random number of different ranges
   - small: between 0.9 and 1.1
   - medium: between 0.7 and 1.3
   - high: between 0.5 and 1.5
5. Random Contrast Adjustment: the contrast of an image is scaled by a random number of different ranges
   - small: between 0.9 and 1.1
   - medium: between 0.7 and 1.3
   - high: between 0.5 and 1.5

In Figure 2 we can see selected levels of augmentation applied to a resized and cropped image from the CHECK data set.

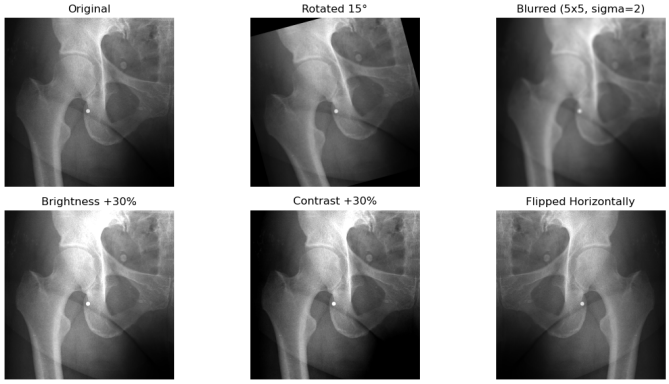


Figure 2: Example of augmentations applied to a cropped and resized image of CHECK

### 3.5 Evaluation

As previously mentioned, our evaluation metrics include the Jaccard index for measuring overlap and the Hausdorff distance for assessing border accuracy. We calculate the average and standard deviation of both metrics on two data sets: the 20 percent of the training data used to train the model and the entirety of the other data set.

# 4 Results

## 4.1 Differences between data sets

In Figure 3 we can see differences within the brightness distribution of the two data sets. A difference in average

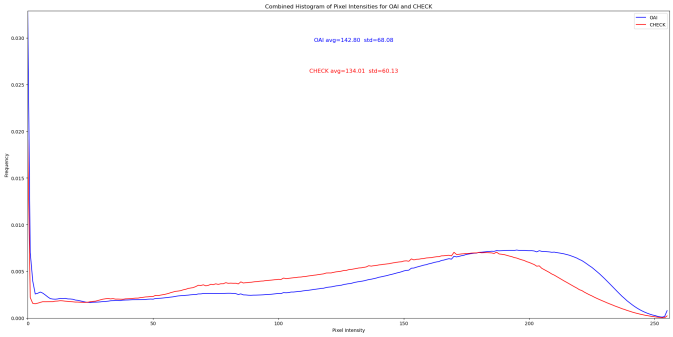


Figure 3: Histograms of normalized and resized OAI and CHECK data set

brightness as well as the standard deviation can also be seen.

In Figure 4 we can see the significant differences between the mean variance of the Laplacian between the CHECK and OAI data set. In the results for the baseline

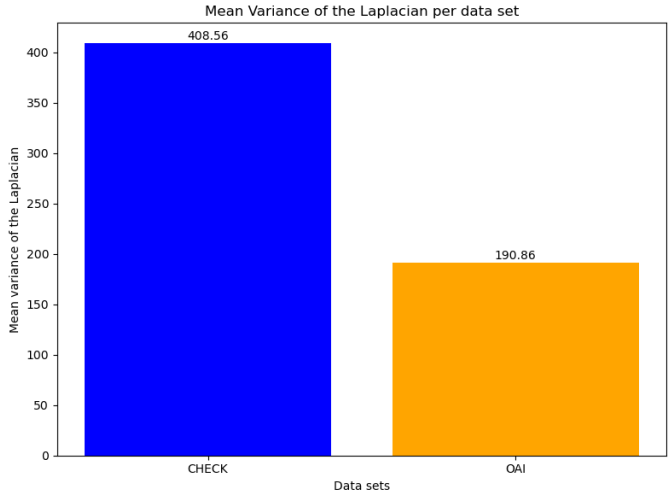


Figure 4: Difference in Mean of variance of Laplacian across data sets

model, which was trained without any data augmentation, we observe that the model trained on the CHECK data exhibits a notable difference in generalization performance within its own dataset compared to the other dataset. Although the differences in the Jaccard indecies are relatively small but noticeable, the Hausdorff distance reveals a significant disparity between the two datasets.

The model trained on the OAI dataset achieves better results across all evaluation metrics and datasets. Additionally, the difference in performance between evaluations on the OAI dataset and the CHECK dataset is significantly smaller.

| Trained on | Tested on | Jaccard | Hausdorff |
|---|---|---|---|
| CHECK | CHECK | 0.967042 | 6.83842 |
| CHECK | OAI | 0.955002 | 10.9427 |
| OAI | OAI | 0.97302 | 4.91234 |
| OAI | CHECK | 0.97319 | 5.13456 |

Table 1: Results of Jaccard and Hausdorff distances for different training and testing data sets.

This suggests that the model generalizes more effectively when trained on the OAI data.

## 4.2 Generalisation within the CHECK data set

In Table 2 we can see the evaluation of the effects of data augmentation on the generalisation within the own data set.

| **Augmentation** | **Jaccard** | **Std Jaccard** | **Hausdorff** | **Std Hausdorff** |
|---|---|---|---|---|
| none | 0.9670 | 0.0376 | 6.8420 | 13.4496 |
| **Augmentation** | **Δ Jaccard** | **Δ Std Jaccard** | **Δ Hausdorff** | **Δ Std Hausdorff** |
| flipped | +0.0025 | +0.0007 | -2.2085 | -9.0319 |
| rotation high | +0.0008 | +0.0009 | -1.9327 | -8.3420 |
| rotation medium | +0.0021 | -0.0005 | -2.1332 | -9.2301 |
| rotation low | +0.0009 | -0.0007 | -1.9811 | -8.5597 |
| brightness high | -0.0002 | -0.0005 | -1.6609 | -6.9942 |
| brightness medium | +0.0006 | +0.0036 | -1.9259 | -7.8763 |
| brightness low | -0.0004 | -0.0002 | -1.2404 | -4.3286 |
| contrast high | +0.0002 | +0.0024 | -1.8227 | -8.4019 |
| contrast medium | -0.0003 | +0.0009 | -1.5686 | -6.8507 |
| contrast low | +0.0001 | +0.0020 | -1.8699 | -9.0481 |
| blur high | +0.0005 | +0.0023 | -1.8193 | -8.1567 |
| blur medium | +0.0005 | +0.0010 | -1.9202 | -9.3682 |
| blur low | +0.0005 | +0.0029 | -1.9358 | -8.8604 |

Table 2: Performance metrics for different augmentations in the CHECK dataset tested on CHECK

Flipping the images resulted in noticeable improvements. It increased the average Jaccard index, indicating better overlap between the predicted and true masks, the average Hausdorff distance decreased, suggesting more precise and consistent boundary predictions. This method also reduced the variability in performance, as seen from the lower standard deviations.

Rotation at various intensities also showed positive effects. High, medium, and low rotation augmentations improved the average Jaccard index, but did not significantly reduce the standard deviation, with medium rotation yielding the highest improvements among the three. The average Hausdorff distance and its standard deviation was reduced across all rotation settings, with the medium rotation providing the best improvements.
Brightness adjustments produced mixed results. Medium brightness adjustments slightly improved the Jaccard index and Hausdorff distance. In contrast, high and low brightness adjustments primarily enhanced the Hausdorff distance but resulted in overall poorer performance in the other metrics.

High and medium contrast adjustments improved the Hausdorff distance, but medium contrast adjustments resulted in a slightly higher improvement of the Jaccard index

. Low contrast adjustments did not enhance performance except for a lower Hausdorff distance.

Blur augmentations generally improved model performance. High blur settings provided moderate improvements in the Jaccard index, along with a reduction in the Hausdorff distance. Medium and low blur settings also enhanced performance, with medium blur showing the best overall improvement in segmentation accuracy and boundary precision.

Among the tested augmentation techniques, flipping and medium rotation showed the most substantial improvements in both segmentation accuracy and boundary precision. Medium brightness and blur adjustments also slightly improved performance, suggesting these methods help the model generalize better within the same dataset. Conversely, contrast adjustments were the least effective.
We observe that augmentations significantly reduced the Hausdorff distance, indicating substantial improvements in boundary precision. In contrast, the impacts on the Jaccard index were more subtle, showing only modest improvements. The standard deviations for the Jaccard index generally did not improve; instead, they slightly increased, reflecting a minor rise in performance variability.

### 4.3 Generalisation to OAI data set

In Table 3 we can see the evaluation of the effects of data augmentation on the generalisation to another data set. Flipping images led to notable improvements. The average

| **Augmentation** | **Jaccard** | **Std Jaccard** | **Hausdorff** | **Std Hausdorff** |
|---|---|---|---|---|
| none | 0.9550 | 0.0456 | 11.0472 | 21.6011 |
| **Augmentation** | **Δ Jaccard** | **Δ Std Jaccard** | **Δ Hausdorff** | **Δ Std Hausdorff** |
| flipped | +0.0059 | -0.0060 | -4.9325 | -11.9413 |
| rotation high | +0.0046 | -0.0067 | -5.0488 | -12.5593 |
| rotation medium | +0.0059 | -0.0075 | -4.9767 | -12.0267 |
| rotation low | +0.0033 | -0.0048 | -4.3427 | -10.5684 |
| brightness high | +0.0026 | -0.0025 | -5.0405 | -13.9138 |
| brightness medium | +0.0034 | -0.0037 | -5.1056 | -13.5936 |
| brightness low | +0.0028 | -0.0061 | -4.1101 | -9.3770 |
| contrast high | +0.0033 | -0.0037 | -4.8256 | -12.5132 |
| contrast medium | +0.0005 | +0.0029 | -4.2099 | -11.1701 |
| contrast low | +0.0034 | -0.0054 | -4.8653 | -12.5176 |
| blur high | +0.0031 | -0.0051 | -4.8749 | -13.0071 |
| blur medium | +0.0038 | -0.0046 | -5.1504 | -13.7956 |
| blur low | +0.0035 | -0.0039 | -5.1933 | -13.6012 |

Table 3: Performance metrics for different augmentations in the CHECK dataset tested on OAI.

Jaccard index increased. The average Hausdorff distance decreased significantly, showing more precise and consistent boundary predictions. This method also resulted in a lower standard deviation, indicating a lower performance variability.

The rotation augmentations had varying degrees of effectiveness. High, medium, and low rotations all led to improvements in the Jaccard index, with medium rotation yielding the highest improvements among the three. The Hausdorff distance was consistently reduced across all rotation settings, with medium rotation showing the best scores across all metrics

Brightness adjustments produced mixed outcomes. High brightness adjustments slightly improved the Jaccard index and also led to a notable reduction in the Hausdorff distance. Medium brightness adjustments provided the best scores across all the metrics. Low brightness adjustments resulted in similar Jaccard index improvements as the high adjustment with lower standard deviations, but a significantly worse Hausdorff distance.

Contrast adjustments also showed varied results. High and low contrast adjustments led to improvements in the Jaccard index, but low contrast adjustments were more effective in reducing the Hausdorff distance. Medium contrast adjustments did not significantly enhance performance and even increased the standard deviation of the Jaccard index, except for the Hausdorff distance.
Blur augmentations generally improved model performance. High blur settings provided moderate improvements in the Jaccard index, along with a reduction in the Hausdorff distance. Medium and low blur settings also enhanced performance, with medium blur showing the best overall improvement in both segmentation accuracy and boundary precision.

Among the augmentation techniques tested, flipping, medium and high rotation provided the most substantial improvements in both segmentation accuracy and boundary precision. Medium brightness and medium and low blur adjustments also proved effective, with the low blur augmentation scoring the lowest average Hausdorff distance. Low contrast adjustments were the least effective.
All data augmentation techniques improved the results across all metrics, and most also lowered standard deviations, indicating more robust and consistent segmentation.

## 5 Discussion

As shown in Table 1, the model trained on the OAI dataset generalizes much better to another dataset. A key factor is that the OAI dataset is roughly three times larger, providing a more diverse set of data. All models, regardless of the training or evaluation set, perform quite well on the Jaccard index. This can be partly explained by the fact that the femur is relatively large, so even significant boundary changes do not greatly affect these overlap metrics. However, boundary accuracy varies more significantly, as reflected by the Hausdorff score.
We observe that all data augmentations, tested on both the original and the new dataset, significantly improved the Hausdorff distance. Figure 5 illustrates a case with relatively high Jaccard index but a much worse Hausdorff distance, highlighting the disparity in boundary precision. Hausdorff distance is highly susceptible to outliers, and the reduced Hausdorff distance likely results from the increased sample size of images. However, some augmentation methods are more effective than others.
When it comes to generalization to the OAI dataset, medium

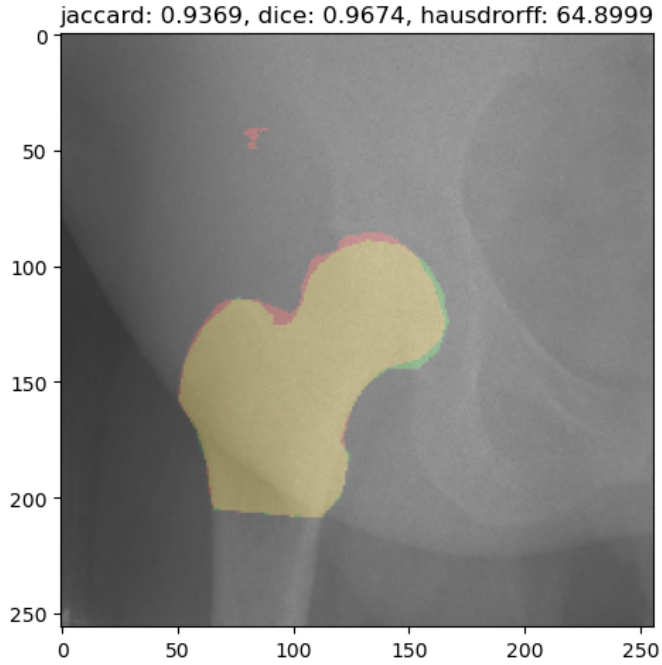


Figure 5: Example segmentation of the base model trained on CHECK on an OAI image.Yellow: overlap between ground truth and predicted mask, Red: predicted mask, Green: ground truth

and low blur augmentations and medium and high brightness adjustments are particularly effective. These augmentations may effectively eliminate outliers. The most effective augmentations target the differences between the two datasets, likely by removing outliers in the new dataset. When testing on the original dataset, flipping is the most effective in reducing the Hausdorff distance. This could be because flipping introduces realistic variations for the CHECK dataset, while the OAI dataset is more diverse, and augmentations targeting differences better simulate outlier behavior, thus reducing the Hausdorff distance most effectively.

Overall, data augmentations improve segmentation performance on the OAI dataset more than on the CHECK dataset. Figures 3 and 4 show differences between the datasets, explaining why augmentations targeting these differences may not perform as well on the original dataset, as they introduce images with characteristics that are different from that dataset, increasing overall variance.

For augmentations with different levels, we see that slight blurring retains most important details, such as edges and textures, while introducing some variability, helping the model recognize features under slightly different conditions without losing critical information. Medium blur introduces a moderate amount of blur, possibly still preserving essential details while challenging the model to generalize better. Too much variability due to high blur can make the training data less representative of real-world scenarios, leading to poorer generalization.

Similarly, for brightness augmentation, medium brightness adjustments perform best within both the OAI and CHECK datasets. Low adjustments may not provide enough variability, while high adjustments may be too unrealistic. Medium brightness adjustments simulate real-world conditions better.

For contrast adjustments, high and low adjustments perform best, possibly for similar reasons.

These arguments do not apply to rotation augmentation. Low rotations perform poorly, while medium and high rotations perform best. In the OAI dataset, rotations of 15 or 25 degrees appear unrealistic. High rotations introduce substantial changes, possibly preventing the model from memorizing specific details and orientations of the training images, forcing it to learn more general features applicable across different datasets.

## 6 Conclusions and Future Work

We can conclude that traditional data augmentations effectively enhance the generalizability of segmentation models to different datasets, improving segmentation accuracy and reducing variability. Among the tested augmentations, random rotations of 15 degrees or more and horizontal flipping were the most effective, followed by medium blur, medium brightness, and low contrast adjustments. The effectiveness of brightness, contrast, and blur adjustments depends on the strength of the modifications and the target dataset. Within the training dataset, all augmentations reduced the Hausdorff distance but had a minimal impact on the Jaccard index, leading to increased variability in segmentation accuracy.

Future work could explore other models and datasets, evaluate combinations of augmented methods, and test additional augmentation techniques like adding random noise. Applying different augmentations to subsets of the data rather than uniformly across the entire dataset would also be an interesting approach to investigate.

## 7 Responsible Research

This section outlines our approach to maintaining data integrity and reproducibility in our study.

### 7.1 Data Integrity

The datasets used in this study are sourced from the Osteoarthritis Initiative (OAI) and the Cohort Hip and Cohort Knee (CHECK) study, both of which are reputable and well-regarded sources. These datasets can be accessed by researchers and have been used in medical research already, ensuring their reliability and quality.
To preserve data integrity during preprocessing, we follow a standardized process. This includes consistent methods for loading DICOM images, applying BoneFinder segmentation, and creating binary segmentation masks.
We ensure that all data handling processes, including storage and transfer, are secure and comply with data protection regulations.

### 7.2 Reproducability

We have implemented several strategies to improve the reproducibility of our work. Documentation of all procedures, including data preprocessing steps, model training protocols, and evaluation methods, are maintained. This provides a clear roadmap for reproducing our experiments.

We ensure that all hyperparameters and model settings, such as learning rate, optimizer type, and number of epochs, are explicitly stated. This transparency allows others to replicate our experiments under the same conditions.

We establish and document baseline models without data augmentation to provide a point of comparison. This baseline is crucial for understanding the impact of different data augmentation techniques on model performance. We use standard evaluation metrics such as Jaccard, and Hausdorff scores, which are widely recognized and accepted in the field.

### 7.3 Use of AI

The only AI tool that we used throughout the writing of our paper was ChatGPT. We primarily used ChatGPT to ensure that a consistent writing style was used throughout the report. This meant that we wrote a section of the report, that still required refinement, and then consulted ChatGPT to assist us in achieving a more polished and cohesive text. However, we often found that the response generated by ChatGPT used overly stilted language, and usually seemed overformal. Therefore, we usually just took ideas for phrasing from ChatGPT. However, it was still a useful tool to ensure that the style of the report was coherent, that the content within paragraphs flowed well. Another aspect that we used ChatGPT for was formatting in LaTeX as well as helping to debug code.

## References


[1] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *arXiv preprint arXiv:2002.04269*, 2020.

[2] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. 2009.

[3] Teerath Kumar, Alessandra Mileo, Rob Brennan, and Malika Bendechache. Image data augmentation approaches: A comprehensive survey and future directions. 2023.

[4] Adam Tupper and Christian Gagné. Analyzing data augmentation for medical images: A case study in ultrasound images. 2024.

[5] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C. Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation, 2024.

[6] Tomasz Szandała. Convolutional neural network for blur images detection as an alternative for laplacian method. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2901–2904, 2020.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[8] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net, 2022.

[9] Fatemeh Nazem, Fahimeh Ghasemi, Afshin Fassihi, and Alireza Mehri Dehnavi. 3d u-net: A voxel-based method in binding site prediction of protein structure. *Journal of Bioinformatics and Computational Biology*, 19(02):2150006, 2021. PMID: 33866960.

[10] Mikhail E. Kandel, Yuchen R. He, Young Jae Lee, Taylor Hsuan-Yu Chen, Kathryn Michele Sullivan, Onur Aydin, M. Taher A. Saif, Hyunjoon Kong, Nahil Sobh, and Gabriel Popescu. Phase imaging with computational specificity (pics) for measuring dry mass changes in sub-cellular compartments. *Nature Communications*, 11(1), December 2020.

[11] Zongyi Li, Hongbing Lyu, and Jun Wang. Fusionu-net: U-net with enhanced skip connection for pathology image segmentation, 2023.

[12] Patrick Leydon, Martin O'Connell, Derek Greene, and Kathleen Curran. Bone segmentation in contrast enhanced whole-body computed tomography. *Biomedical Physics Engineering Express*, 8, 07 2022.

[13] Jae-Hyuk Shim, Woo Seok Kim, Kwang Gi Kim, Gi Taek Yee, Young Jae Kim, and Tae Seok Jeong. Evaluation of u-net models in automated cervical spine and cranial bone segmentation using x-ray images for traumatic atlanto-occipital dislocation diagnosis. *Scientific Reports*, 12(1):1–11, 2022.

[14] Adnan Saood and Iyad Hatem. Covid-19 lung ct image segmentation using deep learning methods: U-net versus segnet. *BMC Medical Imaging*, 21, 2021.

[15] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2016.

[16] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, 2017.

[17] Foivos I. Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:94–114, April 2020.

[18] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

[19] National Institutes of Health. https://nda.nih.gov/oai. Accessed: 2024-04-25.

[20] Janet Wesseling, Maarten Boers, Max A Viergever, Wim KHA Hilberdink, Floris PJG Lafeber, Joost

Dekker, and Johannes WJ Bijlsma. Cohort profile: Cohort hip and cohort knee (check) study. *International Journal of Epidemiology*, 45(1):36–44, 2016.

[21] 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.

[22] IlliaOvcharenko. lung-segmentation. https://github.com/IlliaOvcharenko/lung-segmentation.git, 2022. Accessed: 2024-04-25.