Segmentation of hip joint anatomy structures from radiographic images

Lenka Blažková
Department of Biomedical Engineering
FEEC, Brno University of Technology
Brno, Czech Republic
xblazk01@vutbr.cz

***Abstract*—This paper deals with the problem of a hip joint segmentation in radiographic images with the use of a deep learning approach. The paper is focused on training nnU-Net models and creating an original dataset that contains 150 radiographs, 100 training and 50 test images. There are six trained models, five from cross-validation training and one trained on all training data. All models are evaluated on the test dataset using the Dice score for individual labels and the combined mean Dice score for the image. The best-performing model was the model trained on all training images. The most challenging labels for segmentation were those representing the Köhler teardrop and the space between the femoral head, teardrop and acetabulum due to their size and variability observed across the dataset.**

***Index Terms*—hip joint, femur, pelvis, radiography, radiographic image, segmentation, deep learning, convolutional neural network, U-Net, nnU-Net.**

I. Introduction

The hip joint is a critical structure in the human musculoskeletal system, essential for movement and maintaining whole-body posture, and radiography is one of the most common imaging methods used for upper femur and pelvis imaging. It is typically employed as an initial step in a joint examination due to its availability, short duration, and relatively low radiation exposure. Further procedures, surgery planning or aftercare also often rely only on it. For instance, primary hip arthroplasty is frequently performed based solely on visual assessments of X-ray images (2D radiographs) and the surgical field. However, hip preoperative planning is crucial for the success of the surgery and the patient's recovery. Planning must be precise to restore function and enhance the patient's quality of life. Therefore, accurate determination of the correct anatomical landmarks is essential to selecting the correct component type, size, or shape. [1], [2]

M. Kim et al. in [3] dealt with preoperative planning for successful total hip arthroplasty and proposed a deep-learning and rule-based algorithm for optimal hip prosthesis determination. A 2D U-Net supplemented with a classification convolutional neural network was used as a segmentation model. Another research group, led by W. Xu [4], held research regarding hip joint pathologies in infants' radiographs. The research specialized in the localization of developmental dysplasia lesions and proposed an algorithm utilizing joint segmentation via Feature Pyramid Network with ResNet50. L. Chen et al. [5] dealt with segmenting dysplasia lesions as well, specifically using 2D ultrasound images and a cascaded fully convolutional neural network.

In addition, other research groups propose hip joint segmentation approaches but from 3D data. A fully automated algorithm for hip joint segmentation was proposed by J. J. Kim et al. in [6]. It employed the complementary use of a patient-specific optimal thresholding together with a watershed algorithm. C. Chu et al. in [7] also published a fully automatic CT segmentation approach but with the integration of fast random forest regression-based landmark detection, multi-atlas segmentation, and articulated statistical shape model fitting. P. Xu et al. [8] used 3D CT scans and the MultiPlanar U-Net with transfer learning. They embraced model training on publicly available and poorly annotated data with only a few accurate training scans for fine-tuning.

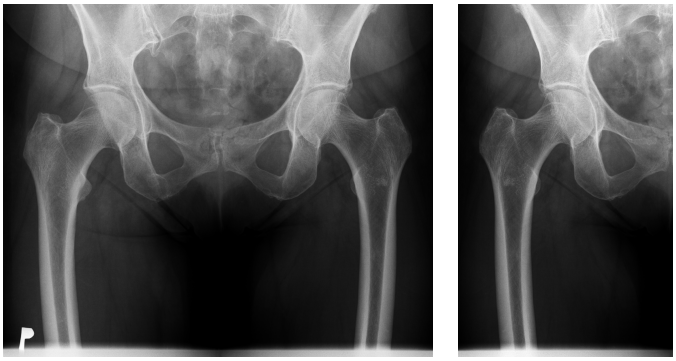
II. Image Dataset

There were two datasets containing 2D radiographs found, the first one by M. Kim et al. mentioned in [3] and the second one by R. Zhao et al. from [9]. However, neither of these two datasets was publicly available and the data could not be retrieved even upon the request. Therefore, the creation of an original dataset was initiated.

83 plain anteroposterior (AP) radiographs were obtained from the DICOM database. All images were anonymised and included the entire pelvis and both upper femurs, e.g. Fig. 1a. The created original dataset itself consists of 150 images of one joint, see Fig. 1b. The remaining joints were not eligible because of already implanted hip replacement prostheses, as well as fixation components like plates and screws used for fracture and other treatments.

The dataset contains data from computed radiography acquired between the years 2010 and 2011. The image sizes range from 1223×1943 pixels to 2140×3520 pixels, and the image spacings are 0.1×0.1 mm, 0.15×0.15 mm or 0.175×0.175 mm per pixel.

Furthermore, all images in the dataset have a corresponding ground truth segmentation mask created by a non-medical health professional, according to Act No. 96/2004 Coll. [10]. Each mask includes five labels, depicted in Fig. 2. The region of the first label (L1) contains the femoral head, the medullary cavity, and the greater trochanter; the second label region (L2) includes the inferior neck, the diaphyseal cortex, and the



(a) (b)

Fig. 1: Example of plain AP radiography images from one patient: (a) full image showing both hip joints and (b) image of only one hip joint (image from the original dataset)

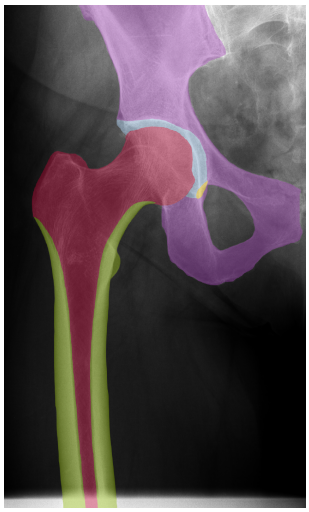


Fig. 2: Example of the segmentation mask with five labels: (L1, *red*) femoral head, medullary cavity and greater trochanter; (L2, *green*) inferior neck and diaphyseal cortex and lesser trochanter; (L3, *yellow*) Köhler teardrop; (L4, *blue*) region between femoral head, teardrop and acetabulum; and (L5, *purple*) pelvic bone

lesser trochanter; the third region (L3) is the Köhler teardrop; the fourth (L4) covers the area between the femoral head, the teardrop, and the acetabulum; and the fifth region (L5) represents the pelvic bone.

The entire dataset is divided into training and test subsets, with the training set consisting of 100 images and the test set comprising of 50 images. To ensure data integrity, images from a single patient image are not included in both subsets simultaneously.

III. Segmentation Model

The used segmentation model architecture is the nnU-Net by F. Isensee et al. [11], available at https://github.com/mic-dkfz/nnunet. It is a deep learning-based segmentation method, more specifically a convolutional neural network, comprising of an encoder-decoder U-Net architecture that automatically adapts its hyperparameters to suit diverse datasets. In this particular instance, the 2D nnU-Net model is selected, since the dataset consists of 2D images.

The data are preprocessed to the voxel spacing of $1.0\times0.15\times0.15$ mm and the shape is standardized to a size of $1\times2320\times1414$ pixels. In addition, the data processing includes Z-score normalization applied uniformly.

The specific model architecture for the used dataset is based on the plain convolution U-Net with nine stages and employs from 32 to 512 features per stage. Each stage incorporates two convolutional layers with a kernel size of 3×3. The network architecture also leverages stride adjustments in conjunction with instance normalization and the Leaky ReLU activation function. The model training is carried out with a batch size of 2, a patch size of 1536×1024 pixels and optimized using the Dice score.

IV. Results and Discussion

There are 6 trained models, five obtained via five-fold cross-validation and one trained on the whole training dataset. All the models were trained for 1000 epochs. The model performance was evaluated on the test dataset using the Dice score metric. The results of the different models were evaluated along all labels as a mean Dice score and also in the individual label categories. The mean Dice score results overview for each model can be seen in Fig. 3 and in Table I.

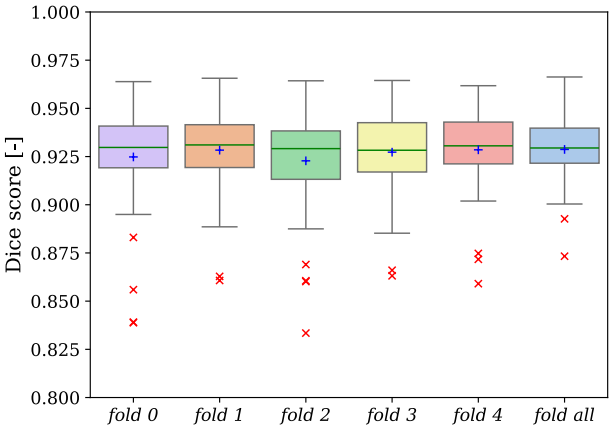
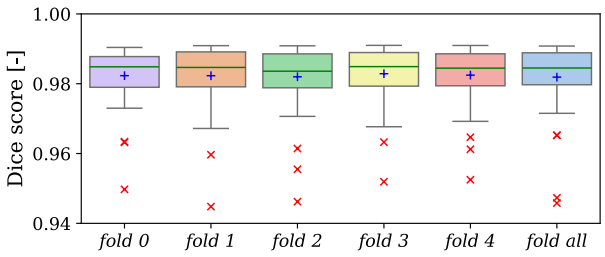


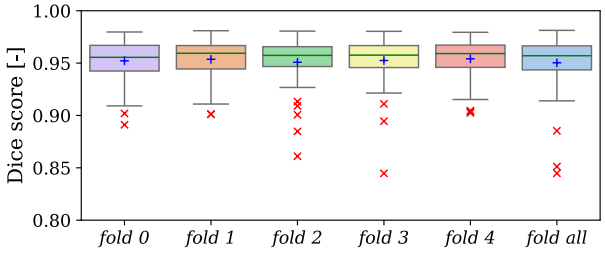
Fig. 3: Box and whisker plots of mean Dice score evaluation of different models on test dataset: mean (blue +), median (green line), 25^{th}–75^{th} percentile (coloured box), extreme values (black whiskers) and individual outliers (red ×)

TABLE I: Evaluation of the different models

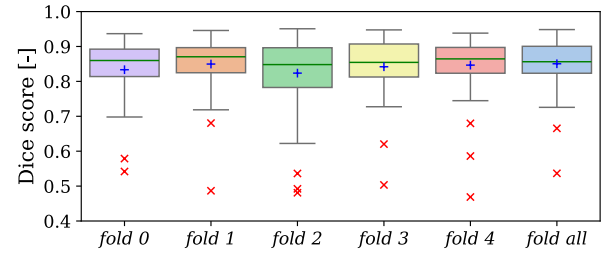
Dice score	*fold 0*	*fold 1*	*fold 2*	*fold 3*	*fold 4*	*fold all*
mean val.	0.9248	0.9283	0.9228	0.9273	0.9285	**0.9287**
std val.	0.0263	0.0212	0.0262	0.0212	0.0207	**0.0176**
median	0.9297	**0.9310**	0.9291	0.9283	0.9306	0.9295
max val.	0.9639	0.9656	0.9643	0.9645	0.9618	**0.9663**
min val.	0.8388	0.8607	0.8334	0.8631	0.8590	**0.8733**



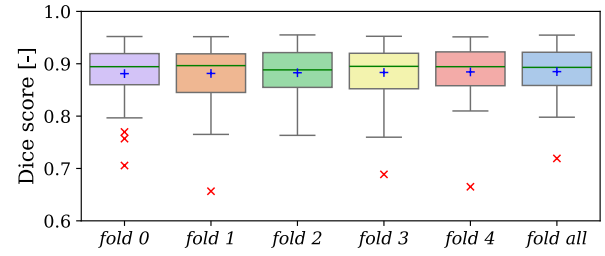
(a) Evaluation of label L1



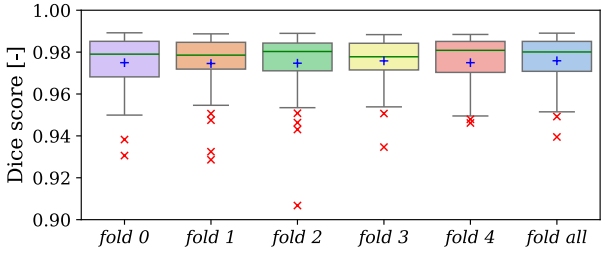
(b) Evaluation of label L2



(c) Evaluation of label L3



(d) Evaluation of label L4



(e) Evaluation of label L5

Fig. 4: Box and whisker plots of individual labels Dice score evaluation of different models on test dataset: mean (blue +), median (green line), 25th–75th percentile (coloured box), extreme values (black whiskers) and individual outliers (red ×)

Individual models were also evaluated according to the different labels, and the graphical representation can be seen in Fig. 4. The best Dice score from all models has L1, above 0.94 for all data and models. L1 evaluation results are followed by L5 and L2 results. The results are connected to the sizes of the regions, which are proportionally larger than the rest of the labels. It can be assumed that they are visually simpler to be evaluated from a radiographic image.

On the contrary, L3 has the smallest Dice scores, which corresponds with the overall small label region size. L4 also has a lower Dice score, which can be a consequence of its proximity to L3. If an error is made at L3, it is likely to be made at L4. The main reasons are high variability in the region size, shape and location. Moreover, they highly depend on the positioning of the patient during the image acquisition (the rotation of the femur and the tilt of the pelvic bone) and are the most affected by joint disorders. Therefore, the ground truth annotations may vary in terms of label placement.

Fig. 5 shows two examples of the segmentation masks from the trained models, where some segmentation errors are present. Fig. 5a shows incorrect segmentation of L1 (particularly the femoral head), L4 and L3, due to the femoral head projection medial to the ilioischial line (acetabular protrusion). Another suboptimal segmentation of L1 and L4 can be observed in Fig. 5b, where the hip joint exhibits signs of osteoarthritis.

Upon further research, it was determined that the best model is the *fold all* model. Its results are shown in Table II. It has the best mean Dice score for labels L3, L4 and L5 (0.85047, 0.88497 and 0.97587). The mean Dice score for L2 is 0.9503, which is just 0.004 lower than *fold 4* model, the best for L2. The mean Dice score for L1 is only 0.001 lower than the best score from *fold 3* model. The mean Dice score of L1 is 0.9819.

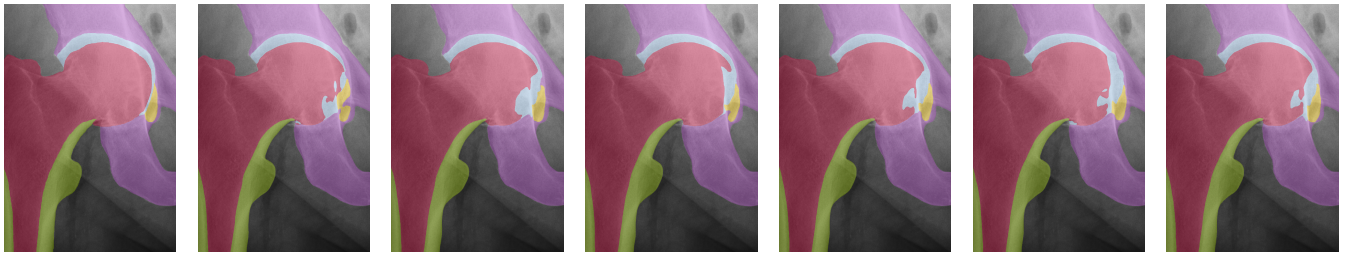
TABLE II: Evaluation of the *fold all* model across labels

Dice score	*L1*	*L2*	*L3*	*L4*	*L5*
mean val.	0.9819	0.9503	0.8505	0.8850	0.9759
std val.	0.0096	0.0271	0.0741	0.0456	0.0118
median	0.9845	0.9570	0.8563	0.8931	0.9801
max val.	0.9908	0.9813	0.9484	0.9547	0.9891
min val.	0.9458	0.8448	0.5366	0.7192	0.9395

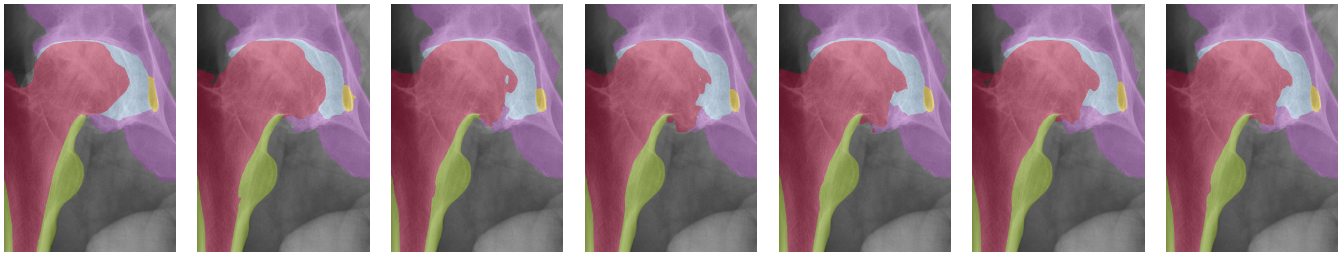
V. Conclusion

This paper aimed to train a segmentation model for key anatomical structures of the hip joints which has a high potential for hip joint analysis during medical examinations or surgery planning. The model demonstrates considerable potential for use in medical examinations, the diagnosis of hip joint disorders and the improvement of pre-operative planning.

The trained model was the nnU-Net framework by F. Isensee et al. [11] with automatically determined hyper-parameters. The training and test data were from an original radiograph dataset explicitly created for this research. A total of six models were trained, and their evaluation using the Dice score metric confirmed the effectiveness of the models.



(a1) ground truth (a2) *fold 0* (a3) *fold 1* (a4) *fold 2* (a5) *fold 3* (a6) *fold 4* (a7) *fold all*



(b1) ground truth (b2) *fold 0* (b3) *fold 1* (b4) *fold 2* (b5) *fold 3* (b6) *fold 4* (b7) *fold all*

Fig. 5: Examples of predicted segmentation masks (with mean Dice score): (a) incorrect segmentation of L1 (femoral head), L4 and L3 (0.8391, 0.8886, 0.9042, 0.8852, 0.8748 and 0.9082; acetabular protrusion); (b) incorrect segmentation of L1 (femoral head) and therefore L4 (0.8831, 0.8896, 0.8888, 0.8971, 0.9067 and 0.9021; osteoarthritis)

The *fold all* model was established as the best model. Its mean Dice score is 0.9287 with a standard deviation of 0.0176 and a median of 0.9295. In addition, this model performed best on the labels L3 and L4, showing that they were the most difficult to predict overall.

Future work on this research may focus on refining the model with a larger dataset. It may also conduct a deeper analysis of pathological cases, aim for the integration of the models into clinical workflows, and explore its application in real-time diagnostics.

Acknowledgment

I would like to express my gratitude to OR-CZ spol. s r.o. for their support and for providing the essential data that made this research possible. Additionally, I would like to extend my sincere thanks to my consultant Jan Kelča and supervisor Vratislav Harabiš for their guidance throughout this project, to Michal Nohel for his insights into the nnU-Net models, and to the authors of the nnU-Net [11]. I also acknowledge MetaCentrum and the Linux Ubuntu 22.04 system running on a machine equipped with an Nvidia Titan Xp 12GB GDDR5 graphics card for providing the computational resources necessary for this work.

References

[1] V. Adukia, K. Kulkarni, and D. K. Menon, "Clinical Examination of the Hip Joint (Basic and Surface Anatomy) with Special Tests", in *Orthopedics of the Upper and Lower Limb*, 2nd ed.Cham: Springer International Publishing, 2020, pp. 217-238. doi: 10.1007/978-3-030-43286-7_14.

[2] V. Adukia, K. Kulkarni, and D. K. Menon, "The Hip Joint", in *Orthopedics of the Upper and Lower Limb*, 2nd ed.Cham: Springer International Publishing, 2020, pp. 239-277. doi: 10.1007/978-3-030-43286-7_15.

[3] M. Kim, I.-S. Oh, and S. -J. Yoon, "Deep Learning and Computer Vision Techniques for Automated Total Hip Arthroplasty Planning on 2-D Radiographs", *IEEE Access*, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3204147.

[4] W. Xu *et al.*, "A Deep-Learning Aided Diagnostic System in Assessing Developmental Dysplasia of the Hip on Pediatric Pelvic Radiographs", *Frontiers in Pediatrics*, vol. 9, Mar. 2022, doi: 10.3389/fped.2021.785480.

[5] L. Chen *et al.*, "Femoral head segmentation based on improved fully convolutional neural network for ultrasound images", *Signal, Image and Video Processing*, vol. 14, no. 5, 2020, doi: 10.1007/s11760-020-01637-z.

[6] J. J. Kim, J. Nam, and I. G. Jang, "Fully automated segmentation of a hip joint using the patient-specific optimal thresholding and watershed algorithm", *Computer Methods and Programs in Biomedicine*, vol. 154, 2018, doi: 10.1016/j.cmpb.2017.11.007.

[7] C. Chu, C. Chen, L. Liu, and G. Zheng, "FACTS: Fully Automatic CT Segmentation of a Hip Joint", *Annals of Biomedical Engineering*, vol. 43, no. 5, 2015, doi: 10.1007/s10439-014-1176-4.

[8] P. Xu, F. Moshfeghifar, T. Gholamalizadeh, M. B. Nielsen, K. Erleben, and S. Darkner, "Auto-segmentation of Hip Joints Using MultiPlanar UNet with Transfer Learning", in *Medical Image Learning with Limited and Noisy Data*, in Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022, pp. 153-162. doi: 10.1007/978-3-031-16760-7_15.

[9] R. Zhao, H. Cai, H. Tian, and K. Zhang, "Morphological consistency of bilateral hip joints in adults based on the X-ray and CT data", *Surgical and Radiologic Anatomy*, vol. 43, no. 7, 2021, doi: 10.1007/s00276-020-02676-4.

[10] Czech Republic, *Act No. 96/2004 Coll., on Conditions for Acquisition and Recognition of Competence to Practice Non-Medical Health Professions and to Perform Activities Related to the Provision of Health Care and on Amendments to Certain Related Acts (Act on Non-Medical Health Professions)*, Collection of Laws, 2004.

[11] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-Net: a self-configuring method for deep learning-based biomedical image segmentation", *Nature Methods*, vol. 18, no. 2, 2021, doi: 10.1038/s41592-020-01008-z.