Automatic Segmentation of Hip Osteophytes in DXA Scans using U-Nets

Document Version
Submitted manuscript

Automatic Segmentation of Hip Osteophytes in DXA Scans using U-Nets

R Ebsim[1], BG Faber[2,3], F Saunders[4], M Frysz[2,3], J Gregory[4],
NC Harvey[5,6], JH Tobias[2,3], C Lindner[1], and T Cootes[1]

[1] Division of Informatics, Imaging and Data Sciences, University of Manchester, UK
[2] Musculoskeletal Research Unit, University of Bristol, UK
[3] Medical Research Council Integrative Epidemiology Unit, University of Bristol, UK
[4] Centre for Arthritis and Musculoskeletal Health, University of Aberdeen, UK
[5] Medical Research Council Lifecourse Epidemiology Centre, University of Southampton, UK
[6] NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospitals Southampton NHS Foundation Trust, Southampton, UK

Abstract. Osteophytes are distinctive radiographic features of osteoarthritis (OA) in the form of small bone spurs protruding from joints that contribute significantly to symptoms. Identifying the genetic determinants of osteophytes would improve the understanding of their biological pathways and contributions to OA. To date, this has not been possible due to the costs and challenges associated with manually outlining osteophytes in sufficiently large datasets. Automatic systems that can segment osteophytes would pave the way for this research and also have potential clinical applications. We propose, to the best of our knowledge, the first work on automating pixel-wise segmentation of osteophytes in hip dual-energy x-ray absorptiometry scans (DXAs). Based on U-Nets, we developed an automatic system to detect and segment osteophytes at the superior and the inferior femoral head, and the lateral acetabulum. The system achieved sensitivity, specificity, and average Dice scores ($\pm$std) of (0.98, 0.92, 0.71$\pm$0.19) for the superior femoral head [793 DXAs], (0.96, 0.85, 0.66$\pm$0.24) for the inferior femoral head [409 DXAs], and (0.94, 0.73, 0.64$\pm$0.24) for the lateral acetabulum [760 DXAs]. This work enables large-scale genetic analyses of the role of osteophytes in OA, and opens doors to using low-radiation DXAs for screening for radiographic hip OA.

Keywords: Computational anatomy · U-nets · Osteophytes segmentation · Osteophytes detection · Automated osteoarthritis risk assessment.

1   Introduction

Osteoarthritis (OA) is a degenerative disease in which bones and surrounding soft tissue of the affected joint deteriorate, leading to pain and loss of function. Hip OA incidence is rising [6, 7] due to increasing aging and obesity in populations [4, 5]. Total hip replacement (THR) is the preferred treatment for end-stage disease which is associated with substantial care costs. For example, in England and Wales there are 150 THRs per 100,000 of population per year [7]. Semi-quantitative grading of hip OA can be done using Kellgren-Lawrence (KL) [8] or Croft [9] scoring which examine the presence and severity of radiographic characteristic features, including joint-space narrowing (JSN), osteophytes, subchondral scleroisis, and cysts. These criteria are inherently subjective [10], making clinical application difficult [11–13]. Although severity grading is traditionally performed by manual inspection of standard radiographs, Yoshida et al. [17] found that this could be performed as accurately using DXA scans. A recent semi-automated scoring system [14] on DXA-derived radiographic hip OA (rHOA) showed strong relationships with symptoms and was predictive of THR. The classifier considered automated joint space measurements and semi-automated osteophyte area calculations. Other semi-automated approaches [15, 16], incorporating osteophytes within a statistical shape model (based on manually checked and corrected point positions) also showed an association with future THR and with change in hip shape over 6-12 months. Semi-automated scoring is time-consuming and costly. Fully automated methods would overcome this and remove, as far as possible, the element of subjectivity.

This work is based on the U-Net architecture [26] which has been widely-used with great success in the domain of image segmentation. A U-Net is a deep convolutional encoder-decoder architecture with various layers at different levels forming a U shape. In addition to the upsampling operators used to increase the resolution on their outputs, U-Nets differ from conventional fully convolutional neural network for having skip connections between the encoding and decoding paths, which improves the use of context and localisation during learning [26].

Contributions: To our knowledge, this is the first work on automating pixel-based hip osteophyte segmentation. Our system automatically detects osteophytes with high accuracy and segments osteophytes at the superior and the inferior femoral head, and the lateral acetabulum. It has the following potential:

- It is an essential step towards identifying osteophytes genetic variants which have not been explored before. Manual osteophyte detection/segmentation for genetic studies is not feasible. That is because genetic studies need very large-scale datasets given the low prevalence of osteophytes.
- Due to the low radiation dose and strong relationships recently-associated [14] between DXA-derived rHOA features (JSN and osteophyte sizes) with OA symptoms and prediction of THR, DXAs could potentially be used to screen low risk clinical populations and assess their risk of subsequent THR. This could help stratify patients who should be targeted with weight loss and physiotherapy interventions. Previous research automated JSN calculations in hip DXAs, but osteophyte segmentation has not been automated before.

2   Related work

There are few published methods in the area of automated rHOA classification. Some research focused on detecting rHOA (i.e. binary classification of healthy vs. diseased) as shown in radiographs [18–20] and in CTs [20] by fine-tuning pre-trained deep learning models. Other work [21] also used transfer learning to evaluate multiple individual features of rHOA, including femoral osteophytes (FOs), and acetabular osteophytes (AOs). Each of these two features were graded as absent, mild, moderate, or severe. Lateral/medial and superior/inferior sides were combined by picking the most severe grade for each feature. The reported performance on two test sets was as follows: (1) sensitivity [74.6-83% for FOs, 69.9-76% for AOs], (2) specificity [91-91.1% for FOs, 76-76.4% for AOs], and (3) F1-score(%) [82.6-87.4% for FOs, 65.1-72.1% for AOs]. Few studies considered segmenting osteophytes in particular. The KOACAD algorithm [22] considered only one of the regions in which knee osteophytes develop. The determined osteophyte area was on the medial tibial margin and found as the medial prominence over the smoothly extended outline of the tibia. The area under ROC curve (AUC) for detecting OA by the osteophyte area was 0.65. Thomson et al. [23, 24] detected knee OA by first segmenting knee bones including osteophytes implicitly [23] and explicitly [24] using a Random Forest Regression Voting Constrained Local Model [25] and then trained random forest classifiers on derived shape and texture information to detect OA. Their system achieved an AUC of 0.85 and 0.93 for detecting knee osteophytes and OA, respectively.

3   Method

An example of a left hip DXA with no osteophytes is given in Fig. 1(a,b) with its outline annotated with 85 points.

The areas where osteophytes develop are around point 15 (inferior femoral head osteophytes), point 30 (superior femoral head osteophytes), and point 78 (lateral acetabulum osteophytes). A diseased hip having osteophytes on all three sites is shown in Fig. 1(c,d).

U-Nets were trained to predict the segmentation mask for a given patch from a radiograph. Patches were cropped around the three sites. One U-Net was trained from scratch for each site. Our implementation has only small changes from the original U-Net architecture [26]. We used an input image size of 128x128 pixels, outputting segmentation maps of the same size as the input image with an osteophytes class probability at each pixel. To eliminate the effect of different hip sizes/scales in cropping the areas of interest, points {12,39} in Figure 1(b) were used to define a reference length. The width/height of each pixel in the reference frame is a proportion (dp) of the reference length. By visually inspecting the resulting patches of different dp values and ensuring a good coverage of the area of interest, we set dp to 0.005 for the inferior femoral head, 0.006 for the superior femoral head, and 0.01 for the acetabulum. All cropped patches were normalised. Examples of input patches and their corresponding ground truth masks are shown in Figure 2.

(a)   (b)   (c)   (d)
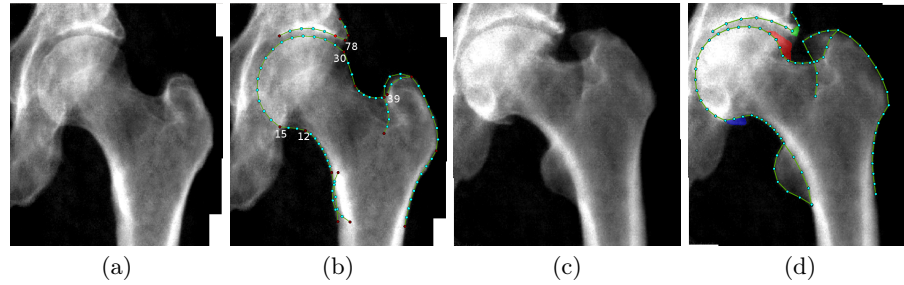
Fig. 1. (a,b): Example DXA of healthy hip showing the joint outline annotated with 85 points. Red points are key anatomical landmarks. Points {12,39} define a reference length. Each pixel in the normalised cropped patches has a width/height equal to a proportion (dp) of the reference length. (c,d): Hip DXA with features of radiographic hip OA (blue: inferior femoral head osteophyte; red: superior femoral head osteophyte; green: acetabular osteophyte).

Fig. 2. Top: Osteophyte sites of the diseased hip in Fig. 1(d). Middle: Input patches to U-Nets. Bottom: their corresponding segmentation masks (ground truth).

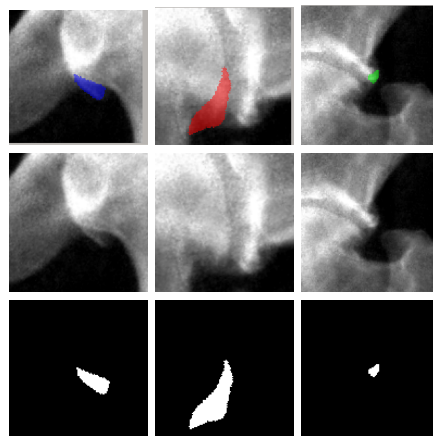## 4   Experiments

4.1   Data

The UK Biobank (UKB) is a prospective cohort study with phenotypic and genetic data collected on approximately 500,000 individuals from the UK, aged between 40 and 69 years at the time of recruitment (between 2006 – 2010)[1]. High resolution iDXA (iDXA GE-Lunar, Madison, WI, USA) scans are being collected as part of the Imaging Enhancement study [2]. As of April 2021, DXA images were available for 42,441 participants, of which 41,160 had left hip images. We excluded 820 left hip scans/participants due to either poor image quality, image error or withdrawal of consent, resulting in a dataset of 40,340 participants [mean age 63.7 years (range 44–82 years)]. Osteophytes were present in 4,013 (10 %) participants/images. Manual osteophyte markups/segmentations were agreed by two experienced annotators (BGF & FS). The markup repeatability was tested after > two months from the first review on a set of 500 DXAs selected to include 20% with osteophytes. The inter-rater kappa values were between 0.80-0.91 for the presence of osteophytes, and the concordance correlation coefficients were between 0.87-0.92 for osteophyte size, depending on osteophyte site.

Data partitions are summarised in Table 1. Per osteophyte site, we randomly selected negative examples to match the number of images with osteophytes to balance the dataset. We split the data randomly into training, validation and test sets. Manual point placements were available for all data. Automatic point placements were available for a subset of all images; obtained fully automatically by applying a Random Forest-based landmark detection system as described in [14] (available from www.bone-finder.com). The training and validation sets used the manual point placements and were balanced (i.e. positive examples (with osteophytes) = negative examples). The focus was on developing a fully automatic system and we thus only included images with automatic point placements in the test sets, leading to slightly imbalanced test sets.

Table 1. Data partitions used to train, validate, and test a U-Net for each osteophyte site. Values are expressed as number of left hip scans/participants (OPs = osteophytes). Some images contained multiple osteophytes and were used for more than one site.

Osteophyte Site	Training Set (with OPs)	Validation Set (with OPs)	Test Set (with OPs)
Inferomedial femoral head	1400 (700)	230 (115)	409 (200)
Superolateral femoral head	1700 (850)	286 (143)	793 (351)
Lateral acetabulum	3760 (1880)	400 (200)	760 (335)

4.2   Implementation details

U-Nets were trained for 15 epoches on NVIDIA Tesla V100 SXM2 16GB, with Keras/Tensorflow implementation (Python 3.7 for GPUs). The best performing

models on the validation sets during training were saved. The sparse categorical cross entropy loss function was optimized with Adam [28] (default parameter values used), and weights were initialised with He normal initializer [27]. We used dropout rate, with probability 0.3, to reduce overfitting, and zero padding to control the shrinkage of dimension due to convolution. The training and validation datasets were augmented with seven additional samples per image including random rotation (within ± 0.1 radians) and scaling (in the range of ±10%). No augmentation was performed for the test sets.

4.3   Results

The performance of the trained U-Nets were evaluated for (1) osteophyte detection (i.e. osteophyte presence vs absence) in terms of sensitivity and specificity; and (2) osteophyte segmentation by average Dice scores (±std). The results are summarised in Table 2.

Table 2. Performance of osteophyte segmentation by U-Nets (1-4), each trained/tested for one osteophyte site. All U-Nets (1-4) had the same architecture, and were trained/-validated on their corresponding datasets with the same augmentation procedure. U-Net(4) only differs in introducing random pixel displacements when augmenting the training set.

	Model	Sensitivity	Specificity	F1-score	Dice score (±std)
1	Inferior femoral head	0.96	0.85	0.91	0.66±0.24
2	Superior femoral head	0.98	0.92	0.94	0.71±0.19
3	Lateral acetabulum	0.67	0.84	0.71	0.46±0.38
4	Lateral acetabulum	0.94	0.73	0.82	0.64±0.24

Our initial results for the lateral acetabulum osteophytes (U-Net(3) in Table 2) were poorer than the performance we achieved at the other sites. Therefore, we investigated the reason by studying two possibilities: *Firstly, could it be underfitting due to insufficient training examples?* To test this, we conducted experiments with an increased number of samples in the training set (by decreasing the size of the test set). However, the low performance pattern was persistent even with a significantly larger training set. *Secondly, could it be related to the automatic point placements?* The U-Net was trained on patches centered on the manual placement of point 78, but tested on new images in which all points were placed automatically. To accommodate for the differences between manual and automatic point placements, we conducted experiments including additional data augmentation. We augmented the training data with 19 additional samples per image including the above described rotation and scaling augmentations as well as random displacements of the centres (point 78) of the training patches. A random displacement was set to be within 0.05 of the patch width/height (about ± 6 pixels) in x- and y-axes. The latter model (U-Net(4) in Table 2) yielded an

increase in performance for the lateral acetabulum osteophytes (sensitivity=0.94, specificity=0.73, Dice=0.64±0.24).

Fig. 3 shows the binary confusion matrices for detecting osteophytes at the three different sites.
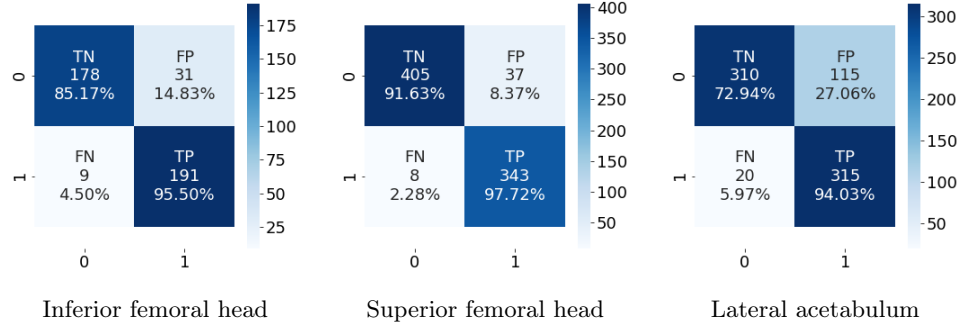

Fig. 3. Performance of three U-Nets on detecting osteophytes at three specific sites. For the lateral acetabulum, random pixel displacements were introduced when augmenting the training set.

Examples of segmentation outputs are given in Fig. 4 for various Dice scores.

5 Discussion and Conclusion

We have developed an automatic system to detect and segment osteophytes at three sites in hip DXAs. Our results show that the system achieves high performance in detecting osteophytes at the superior and the inferior femoral head, and the lateral acetabulum, outperforming related work on automatically grading hip osteophytes in standard radiographs [21]. To the best of our knowledge, this is the first system to automate pixel-wise segmentation of hip osteophytes.

We achieved the best overall results for the superior femoral head osteophytes. This is likely to be because of the nature of the surrounding area within the image as there are fewer overlapping projections compared to the other two osteophyte sites. This could have also affected the accuracy of the automated point placement favorably.

It has been shown that the three osteophyte sites have potentially independent relationships with hip pain, consistent with roles in partially-independent biomechanical pathways [29]. For this reason our work considered each femoral head osteophyte site separately as opposed to the single combined femoral head osteophyte feature in [21].

For the acetabular osteophytes, we showed that enriching the training set with augmented examples of random displacements from the centre of the area
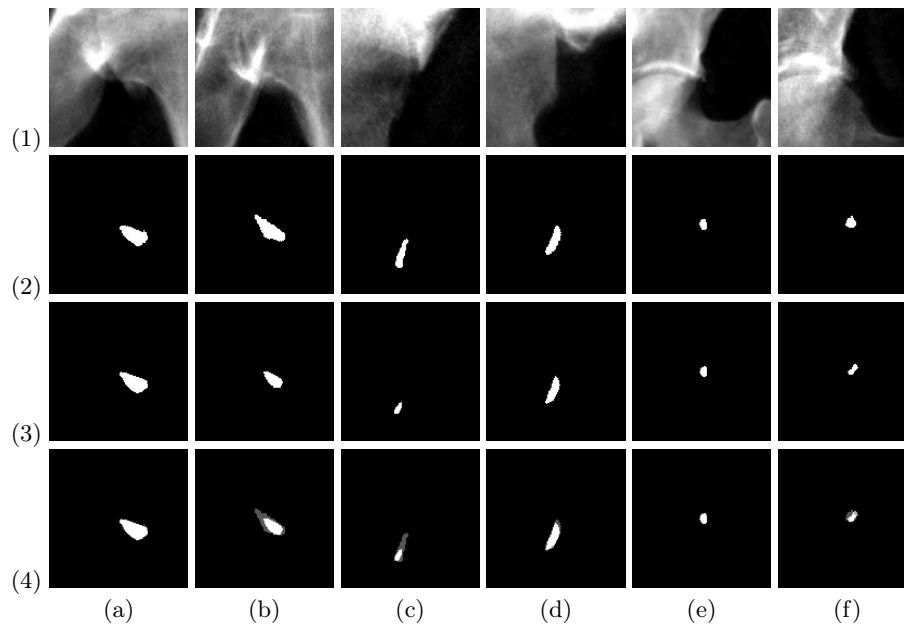
Fig. 4. Examples of osteophyte segmentation results. To visualise the overlap between true and predicted masks, the summations of true masks with two times of predicted masks were plotted. Row (1): Input test patches to U-Nets (based on automatic point placements). Row (2): Manual masks (ground truth). Row (3): Automatically predicted masks. Row (4): The overlap of true and predicted mask are the brightest pixels. (a,b) are examples of inferior femoral head osteophytes with Dice scores (0.95, 0.64). (c,d) are examples of superior femoral head osteophytes with Dice scores (0.42, 0.83). (e,f) are examples of lateral acetabulum osteophytes with Dice scores (0.93, 0.59).

(point 78) improved U-Net learning in tackling point placement discrepancies, and achieved high sensitivity (0.94), compared to 0.76 in [21], as well as an improved Dice score (0.64).

In future work, we will explore whether such an augmentation as applied to the acetabular osteophytes would also improve the performance for detecting and segmenting femoral head osteophytes. In addition, we will conduct further experiments to identify whether there is room for performance improvements for the three sites by performing hyperparameter optimisation to find patch sizes, dp values, and augmentation procedures. We would also like to explore fine-tuning pre-trained U-Net models to our data and compare results.

A limitation of this work is that the system has not been replicated in an independent dataset. Further, it would be of interest to analyse the agreement of clinical experts with the automatically segmented osteophytes. This will provide insight into what Dice score is needed to yield clinical relevant results.

With respect to epidemiological studies, the system can be used to automatically obtain osteophyte segmentations for additional hips in UKB (e.g. right hips, future images) as well as for other large cohort studies such as the Rotterdam Study (RS) or the Osteoarthritis Initiative (OAI). It will also be interesting to explore how the automatically segmented osteophyte areas are related to manual osteophyte grading as well as clinical outcomes.

References

1. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., and O'Connell, J. (2018) The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203-209
2. Littlejohns, T. J., Holliday, J., Gibson, L. M., Garratt, S., Oesingmann, N., Alfaro-Almagro, F., Bell, J. D., Boultwood, C., Collins, R., and Conroy, M. C. (2020) The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nature Communications 11, 1-12
3. Chen, A., Gupte, C., Akhtar, K., Smith, P., Cobb, J.: The Global Economic Cost of Osteoarthritis: How the UK Compares. Arthritis, 2012, 698709. https://doi.org/10.1155/2012/698709
4. Prieto-Alhambra D, Judge A, Javaid MK, Cooper C, Diez-Perez A, Arden NK. Incidence and risk factors for clinically diagnosed knee, hip and hand osteoarthritis: influences of age, gender and osteoar thritis affecting other joints. Ann Rheum Dis. 2014 Sep;73(9):1659–64.

5. Yu D, Peat G, Bedson J, Jordan KP. Annual consultation incidence of osteoarthritis estimated from population-based health care data in England. Rheumatology (Oxford, England). 2015 Nov;54(11): 2051–60.

6. Kim C, Linsenmeyer KD, Vlad SC, et al. Prevalence of radiographic and symptomatic hip osteoarthritis in an urban United States community: the Framingham osteoarthritis study. Arthritis Rheumatol 2014;66(11):3013–3017.

7. National Joint Registry 17th Annual Report 2020, https://www.njrcentre.org.uk/njrcentre/Reports-Publications-and-Minutes/Annual-reports. Last accessed Feb 2022

8. Kellgren JH, Lawrence J. Radiological assessment of osteo-arthrosis. Annals of the rheumatic diseases. 1957 Dec;16(4):494.

9. Croft P, Cooper C, Wickham C, Coggon D. Defining osteoarthritis of the hip for epidemiologic studies. American journal of epidemiology. 1990 Sep 1;132(3):514-22.

10. Reijman M, Hazes JM, Pols HA, Bernsen RM, Koes BW, Bierma-Zeinstra SM. Validity and reliability of three definitions of hip osteoarthritis: cross sectional and longitudinal approach. Annals of the rheumatic diseases. 2004 Nov 1;63(11):1427-33.

11. Birrell F, Lunt M, Macfarlane G, Silman A. Association between pain in the hip region and radiographic changes of osteoarthritis: results from a population-based study. Rheumatology. 2005;44(3):337-41.

12. Iidaka T, Muraki S, Akune T, Oka H, Kodama R, Tanaka S, et al. Prevalence of radiographic hip osteoarthritis and its association with hip pain in Japanese men and women: the ROAD study. Osteoarthritis Cartilage. 2016;24(1):117-23

13. Arden NK, Lane NE, Parimi N, Javaid KM, Lui LY, Hochberg MC, et al. Defining incident radiographic hip osteoarthritis for epidemiologic studies in women. Arthritis Rheum. 2009;60(4):1052-9

14. I thought there might be a different form which I could not find. Faber, B.G., Ebsim, R., Saunders, F.R., Frysz, M., Lindner, C., Gregory, J.S., Aspden, R.M., Harvey, N.C., Davey Smith, G., Cootes, T. and Tobias, J.H., 2021. A novel semi-automated classifier of hip osteoarthritis on DXA images shows expected relationships with clinical outcomes in UK Biobank. Rheumatology.

15. Barr, R.J., Gregory, J.S., Reid, D.M., Aspden, R.M., Yoshida, K., Hosie, G., Silman, A.J., Alesci, S. and Macfarlane, G.J., 2012. Predicting OA progression to total hip replacement: can we do better than risk factors alone using active shape modelling as an imaging biomarker?. Rheumatology, 51(3), pp.562-570.

16. Barr, R.J., Gregory, J.S., Yoshida, K., Alesci, S., Aspden, R.M. and Reid, D.M., 2018. Significant morphological change in osteoarthritic hips identified over 6–12 months using statistical shape modelling. Osteoarthritis and Cartilage, 26(6), pp.783-789.

17. Yoshida, K., Barr, R.J., Galea-Soler, S., Aspden, R.M., Reid, D.M. and Gregory, J.S., 2015. Reproducibility and diagnostic accuracy of Kellgren-Lawrence grading for osteoarthritis using radiographs and dual-energy X-ray absorptiometry images. Journal of Clinical Densitometry, 18(2), pp.239-244.

18. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PloS one. 2017 Jun 2;12(6):e0178992.

19. Üreten, K., Arslan, T., Gültekin, K.E., Demir, A.N.D., Özer, H.F. and Bilgili, Y., 2020. Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiology, 49(9), pp.1369-1374.

20. Gebre, R.K., Hirvasniemi, J., van der Heijden, R.A., Lantto, I., Saarakkala, S., Leppilahti, J. and Jämsä, T., 2022. Detecting hip osteoarthritis on clinical CT: a deep

learning application based on 2-D summation images derived from CT. Osteoporosis International, 33(2), pp.355-365.

21. von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, et al. Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology. 2020;295(1):136–45.

22. Oka, H., Muraki, S., Akune, T., Nakamura, K., Kawaguchi, H. and Yoshimura, N., 2010. Normal and threshold values of radiographic parameters for knee osteoarthritis using a computer-assisted measuring system (KOACAD): the ROAD study. Journal of Orthopaedic Science, 15(6), pp.781-789.

23. Thomson J, O'Neill T, Felson D, Cootes T. Automated shape and texture analysis for detection of osteoarthritis from radiographs of the knee. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2015 Oct 5 (pp. 127-134). Springer, Cham.

24. Thomson J, O'Neill T, Felson D, Cootes T. Detecting osteophytes in radiographs of the knee to diagnose osteoarthritis. In International Workshop on Machine Learning in Medical Imaging 2016 Oct 17 (pp. 45-52). Springer, Cham.

25. Cootes, T.F., Ionita, M.C., Lindner, C. and Sauer, P., 2012, October. Robust and accurate shape model fitting using random forest regression voting. In European conference on computer vision (pp. 278-291). Springer, Berlin, Heidelberg.

26. Ronneberger, O., Fischer, P. and Brox, T., 2015, October. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

27. He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034).

28. Kingma, D.P. and Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

29. Faber, B.G., Ebsim, R., Saunders, F.R., Frysz, M., Lindner, C., Gregory, J.S., Aspden, R.M., Harvey, N.C., Smith, G.D., Cootes, T. and Tobias, J.H., 2021. Osteophyte size and location on hip DXA scans are associated with hip pain: findings from a cross sectional study in UK Biobank. Bone, 153, p.116146.