Convolutional Neural Network-Based Automatic Measurement of Joint Space Width to Predict Radiographic Severity and Progression of Knee Osteoarthritis

James Chung Wai Cheung
Hong Kong Polytechnic University
Yiu Chow TAM
Hong Kong Polytechnic University
Lok Chun CHAN
Hong Kong Polytechnic University
Ping Keung CHAN
Queen Mary Hospital
Chunyi WEN (chunyi.wen@polyu.edu.hk)
Hong Kong Polytechnic University

Research Article

Keywords: neural network, knee osteoarthritis, radiography

Posted Date: February 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-221004/v1

Abstract

Objectives

To develop a deep convolutional neural network (CNN) for the segmentation of femur and tibia on plain x-ray radiographs, hence enabling an automated measurement of joint space width (JSW) to predict the severity and progression of knee osteoarthritis (KOA).

Methods

A CNN with ResU-Net architecture was developed for knee X-ray imaging segmentation. The efficiency was evaluated by the Intersection over Union (IoU) score by comparing the outputs with the annotated contour of the distal femur and proximal tibia. By leveraging imaging segmentation, the minimal and multiple JSWs in the tibiofemoral joint were estimated and then validated by radiologists' measurements in the Osteoarthritis Initiative (OAI) dataset using Pearson correlation and Bland–Altman plot. The estimated JSWs were deployed to predict the radiographic severity and progression of KOA defined by Kellgren-Lawrence (KL) grades using the XGBoost model. The classification performance was assessed using F1 and area under receiver operating curve (AUC).

Results

The network has attained a segmentation efficiency of 98.9% IoU. Meanwhile, the agreement between the CNN-based estimation and radiologist's measurement of minimal JSW reached 0.7801 ($p < 0.0001$). Moreover, the 32-point multiple JSW obtained the highest AUC score of 0.656 to classify KL-grade of KOA. Whereas the 64-point multiple JSWs achieved the best performance in predicting KOA progression defined by KL grade change within 48 months, with AUC of 0.621. The multiple JSWs outperform the commonly used minimum JSW with 0.587 AUC in KL-grade classification and 0.554 AUC in disease progression prediction.

Conclusion

Fine-grained characterization of joint space width of KOA yields comparable performance to the radiologist in assessing disease severity and progression. We provide a fully automated and efficient radiographic assessment tool for KOA.

Introduction

Knee Osteoarthritis (KOA) is a prevalent musculoskeletal disease that is a leading cause of chronic pain and disability in older adults. Clinical diagnosis of KOA relies on plain radiography; Kellgren-Lawrence (KL) grading system is widely deployed in current practice to subjectively describe the severity and

progression of radiographic OA[1]. Joint space width (JSW) is a primary indicator for the integrity of articular cartilage and the severity of KOA[2]. Osteoarthritis Research Society International (OARSI) atlas[3] has been recently established for feature-specific measurement of JSW; however, similar to KL-Grade, the subjectivity of the individuals becomes detrimental to the repeatability and reproducibility of measurement[4]. There has been a growing interest in the development of automated computer-aided methods for consistent quantification of joint space information on plain radiographs for diagnostics and prognostics of KOA.

One of the most commonly used quantities for characterization of the radiographic severity of KOA is minimum joint space width (mJSW). The key to the automatic estimation lies in the accurate segmentation of femur and tibia plateau[1]. The earlier computer-aided approaches were built on traditional methods such as edge detection filters and other statistical algorithms[1,5,6]. Such naive approaches either failed to address the 3D joint structure projection onto 2D images, resulting in the identification of irrelevant bone edges[7], hence inaccurate joint space width estimation or required prior parameterization to roughly localize the bone regions on every images, leading to lack of automation[8].

Recently, deep learning has emerged with superior performance in extracting sophisticated features from a wide variety of data types[9]. By leveraging such an approach, a number of recent OA studies have yielded great success in the analysis of KOA progression prediction[10], total knee replacement (TKR) prediction based on MRI[11], human tissue segmentation[12]. However, to our best knowledge, only a little research has been done in an attempt to identify a smooth, continuous contour of the knee joint for accurate and fine-grained characterization of the tibiofemoral joint space. [13,14] and [7] both leverage low-cost labels to identify only the coarse landmarks, instead of a detailed contour of the knee joint. [15,16] employed the convolutional neural networks (CNN) to create a bounding box to localize the joint space for subsequent detailed grading. The above approaches leverage deep learning or other advanced machine learning methods to generate rough landmarks or region-of-interest (ROI) for various subsequent applications. However, these coarse-grained localizations do not favor detailed quantification joint space features. As a result, a new approach, which could output fine-grained bone contour while being capable of distinguishing relevant edge structures under the 3-D projection in the 2-D radiographic image, is of great need.

To this end, in this paper, we first develop a deep neural network based on the ResU-Net [15] architecture which performs automatic segmentation of the tibia and femur. Subsequently, the performance of our ResU-Net approach is compared with the other deep learning-based image segmentation techniques, including CUMedVision [17,18], DeepLabV3 [19,20], and U-Net [21]. Second, with the identification of the tibial and femoral bone contour, pixel-wise quantitative measurements are made to calculate the knee JSW. In particular, apart from the mJSW defined in the medial compartment, the smooth and continuous contours obtained allow for the calculation of multiple JSWs at fixed locations in the tibiofemoral joint. In that sense, it is inferred that not only richer 1-dimensional information regarding the bone margin could be retrieved; together, they could characterize the whole joint shape which may effectively enhance the

detection of radiographic OA, as inspired by Bayramoglu *et al.*'s recent work [22]. To validate the JSW calculation by our proposed algorithm, we compared our results with the measurements by radiologists from the Osteoarthritis Initiative (OAI) database. Finally, in pursuit of demonstrating the added values of the multi-point JSWs generated by our approach, we compared its prediction prowess towards radiographic severity and progression of KOA defined by Kellgren-Lawrence (KL) grades with the mJSW measured by our method and clinical practitioners, respectively.

Results

Reliability of the annotations Before training our deep learning model for knee bone segmentation on plain radiographic images, we first assessed the reliability of the annotations in the dataset. The mJSW measurements obtained from the annotated data is further compared to the radiologists' measurement extracted from OAI dataset to produce a baseline of interobserver error. The mean interobserver error is 0.483 mm, with a standard deviation of 0.661 mm, and an R2 value of 0.9565. The intra-class correlation coefficient (ICC) was used to test the agreement of inter-observer measurement 23. The ICC between OAI measurement and contour annotator is 0.812, showing that the mJSW measurements have high consistency with the measurements by radiologists. Bone segmentation performance comparison The segmentation accuracy of the four segmentation methods (i.e. CUMed-vision, U-Net, DeepLab V3, and ResU-Net) were compared in Table 1. The segmentation masks produced by the four networks and the ground truth are shown in Figure 2. Both ResU-Net and DeeplabV3 achieved the highest mean IoU score of 0.989, outperforming the other two candidates. Validation loss of ResU-Net is lower than DeeplabV3 (0.006<0.011), showing that form the former model outperforms DeeplabV3 in terms of validation loss. Finally, it was noticed that the overfitting score of DeeplabV3 is higher than that of ResU-Net which indicates its greater tendency of undesirable over-fitting. As a result, the ResU-Net was conceived as the best model in terms of both performance and robustness. Automated measurement of joint space width As the ResU-Net has demonstrated its superiority over the other CNN architectures in this automatic segmentation task on plain radiographs, it was selected as the algorithm to outline the bone contour for subsequent joint space measurements using the CV2 package from python. We then employed the algorithm to segment 4,216 knees (2108 X-ray images) then automatically calculated their mJSW in the medial compartment, with estimated numerical values ranging from 0mm to 7.16 mm, with a mean of 3.53 mm, and a standard deviation of 1.35 mm. In order to access the validity of our automated estimation, we additionally harvested the JSW measurements by clinical doctors or radiologists of those 4,216 knees from the OAI dataset. The measurements' values range from 0 mm to 7.744 mm, with a mean of 3.68 mm and a standard deviation of 1.36 mm. To examine the performance of our proposed deep learning-based automated JSW measurement algorithm, we first performed a linear regression analysis between the mJSW in the medial compartment measured by radiologists which were obtained from the OAI database and that estimated by our proposed deep learning-based automated method with the automatic measurement method (Figure 4a). A significant correspondence was observed among them with an R2 value of 0.6086 and a Pearson correlation of 0.7801 ($p< 0.0001$). Moreover, the Bland-Altman plot 24,25 between the two measurements was also plotted (Figure 4b), which indicated a low

mean difference (d = 0.61 mm), while most of the data were within the 95% confidence interval (±1.76 mm) around the mean difference. This indicated a good agreement between the results obtained by the automatic quantitative JSW estimation and measurement by radiologists. Prediction of KOA severity and progression The accurate JSW measurements enable further study of morphological factors in the severity and progression of OA. KL-grade is a semi-quantitative clinical criterion widely used for the diagnosis of OA, which reflects the severity of OA. The mJSW observes the narrowest points between the tibia and femur plateau in the medial compartment, and act as a monitoring factor for the joint space narrowing (JSN) condition. Nonetheless, this measurement only quantifies the JSW at one single site, which may overlook the whole joint morphological information. Encouraged by our deep learning approach, where continuous contours of the knee joint could be accurately identified, it is possible to measure the JSW at multiple points simultaneously. In the experiment, 16 points were chosen from both the lateral and medial compartments at a fixed interval. Based on the bone contour identified by our ResU-Net, the algorithm automatically calculated the JSWs at all 16 sites at the same time. Additionally, to demonstrate the added value of using 16-point JSWs over the use of single-point mJSW, they were compared side-by-side in the prediction of KL-grade. Table 2 shows that using the 16-point JSWs in place of the mJSW, improves both macro F1 (from 0.311 to 0.402) and AUC scores (from 0.587 to 0.624) significantly in the classification of KL-grades. The measurements by radiologists obtained from the OAI database were also benchmarked with the automatically measured JSWs. Despite having higher prediction scores than the computer-aided estimation in both single-point and 16-points cases, the results still indicated a consistent trend in the classification of KL-grades. Alongside, the 16-point and mJSWs at baseline were deployed to predict the OA progression defined by the increase in KL-grade from unaffected to affected condition within the future 48-month period. Significant prediction improvements in both metrics (Table 3) were observed when replacing single-point mJSW with 16-point JSWs, where the macro F1 and AUROC scores increased from 0.484 to 0.544 and 0.554 to 0.583 respectively, while a similar trend was also observed from the radiologists' measurements. Finally, by leveraging the continuous contours of the tibia and femur output by our ResU-Net model, we further divided the joint space into equally spaced regions with several different densities and hence the 8, 32, and 64-point JSWs were calculated and subsequently employed for prediction of KL-grade and OA prediction. Figure 5a and b both revealed general increasing trends of the AUC score as the number of JSWs increase. Specifically, in the classification of KL-grades, the prediction performance levels off at 32 points of JSW. This might indicate that 64 points of JSW do not provide more additional information than the 32-point JSWs. On the other hand, the prediction performance increases strictly as a greater number of JSWs are involved. It is noteworthy that in both classifications, the optimal CNN-estimated JSWs yield a similar classification score as the radiologist-measured 16-point JSWs.

Discussion

In this study, we have proposed a novel deep learning-based approach for automated bone segmentation in the knee joint on radiographic images. Different from the previous works such as BoneFinder [26] and KNEEL [7], which only identify discontinuous landmarks on the bone margin, our proposed deep learning

model outputs continuous bone contours, allowing characterization of tibiofemoral joint-space shape in higher resolution [27,28]. Four different prominent neural network architectures, including CUMedVision [18], DeepLab V3 [19], U-Net [21], and ResU-Net-18 [29], designed specifically for image segmentation were explored and compared for our application. Lastly, the ResU-Net-18 architecture was selected for its high performance (average IoU of 98.9%). We further demonstrated the robust estimation of the JSWs using our trained network, while such estimations do not only agree well with the measurements by radiologists, but also readily applicable for prediction of KOA severity and progression risk in the future 48-month based on KL-grading system [10,30].

Instead of merely estimating the minimal JSW in the medial compartment of the tibiofemoral joint, which is known as a common clinical practice in KOA diagnosis, with the continuous contour output by our knee segmentation network, it paves the way for measuring JSWs at multiple fixed locations simultaneously. The experimental results indicated that multi-point JSWs is a significantly better predictor over the single-point mJSW in the classification of KOA severity as well as prediction of disease progression defined by the KL-grading system. Moreover, our results also pointed out that increasing the density of the JSW estimations further enhances the classification performances in both KL-grade and KL-defined radiographic OA progression. It could be explained by the fact that incorporation of multiple JSW measurements at different locations along the bone contour would provide more information in the characterization of the tibiofemoral joint's global morphology, which was previously shown to associate with the OA severity[22,31,32]. On top of that, we have further corroborated that joint morphology could also be a valuable predictor of KOA progression.

Previous attempts on applying the traditional computer-vision segmentation approach, which relies on handcrafted features, such as edge detection filters[1] and active contour method[6] for segmentation, the former one detects every edges on the radiograph using the first-order gradient; however, could not distinguish the anterior and posterior edge of the tibial articular surface, where the bright bands of subchondral cortical bone of the tibial plateau and femoral condyle instead of the outermost edge visualized on the radiographs are essential for the measurement of JSW[33] (*Figure 3*). Meanwhile, the latter method's performance relies heavily on the prior curve parameterization by users to roughly locate the regions of interest, which is usually image-specific, thus leading to the lack of automation during the segmentation process[8,34]. On the other hand, deep neural networks have a large number and automatic feature filter generation, hence allowing the model to learn more complex image details and anatomical structures, instead of simple edges and boundaries[27] automatically. Furthermore, this class of models was recently shown to outperform another decision tree-based segmentation technique, BoneFinder[7,35]. Specifically, our deep learning-based bone segmentation approach is superior to the existing approaches in a way that it produces continuous contour of the tibial plateau and femoral condyle rather than discrete landmarks[7,35,36], and is capable of accurately identifying the relevant tibial contour for JSW measurements. This allows preservation of pixel-level boundary information in the tibiofemoral joint, hence beneficial to the extraction of fine-grained morphological details such as multiple JSWs.

The ResU-Net-18 architecture was selected as the backbone of our deep knee segmentation network owing to its high performance and resistance to overfitting compared to the other three candidates. This network enables the low-level details to be passed across the hidden layers to the final output layer, while its residual blocks extract higher-level features hence reducing the overfitting problem as well as ensuring a better fusion of different levels of image features. Additionally, the model adopts atrous convolution, which allows a larger receptive field to be detected[20], thus being beneficial to large image segmentation in our case. On the other hand, the original ResU-Net-50 network was further carefully modified by reducing its number of hidden-layers from 50 to 18 to cater to our mono-color, low-variation bone segmentation task, such modification would effectively reduce the risk of over-fitting in the model[29].

Conclusion

In this work, we present a novel deep learning-based approach that automatically detects the bone contours with high accuracy in the knee joint. By leveraging the continuous contours, the JSWs were measured in an automated manner which are comparable to radiologist-level measurements. We further demonstrated the capability of our algorithm to provide nice characterization of the global joint-space shape by estimating the JSWs at multiple fixed locations, which is time-consuming, if not impractical in the regular clinical settings. And we found that such quantities are more effective than the commonly used mJSW in classifying the OA severity and the prediction of disease progression. As a result, our method provides a computer-aided tool to the clinical practitioners that could facilitate the KOA diagnosis and prognosis with the fully automated, accurate, and efficient computation of the joint-space parameters.

Methods

Dataset and preprocessing

All radiographic images being used were retrieved from Osteoarthritis Initiative (OAI) database (https://data-archive.nimh.nih.gov/oai). In this study, we just focus on the bilateral X-ray images from the baseline cohort which consist of a total of 4216 images. The patient's age ranged from 47-79, with a median of age 61. In the preprocessing pipeline, the 16-bit DICOM images were first normalized using global contrast normalization and a histogram truncation between the 5th and 99th percentiles. These images were being downscaled to 1024*1024 pixels for both training and inferencing. Out of the 4216 images, 100 bilateral radiographs (200 knees) were chosen randomly. The masks were being annotated by two observers using Computer Vision Annotation Tool (https://github.com/openvinotoolkit/cvat) and were cross-checked to refine the annotations. Among all the annotated data, 90% were being used for training, while 10% were used for validation. It has been reported that bilateral knee OA patients demonstrated larger interlimb kinematic asymmetry that may lead to different severity of OA among their limbs[37]. As a result, the wearing rate of both legs might be different and could be biased towards one of the legs in the population, thus potentially leading to the model overfitting. Given this, horizontal flipping

of the X-ray images as a means of data augmentation was employed to improve the model generalization and reduce the bias.

Bone segmentation using deep neural network

In our automated JSW estimation approach, we first employed a deep learning model to perform bone segmentation on plain radiographic images. To this end, four deep convolutional neural network models including U-Net[21], CUMedVision[16], ResU-Net[29], and DeepLabV3[19,20] were selected for producing segmentation of the X-ray images.

U-Net is a class of neural networks designed for image segmentation that extends the fully convolutional net (FCN)[17] by adding skip connections from encoder layers to decoder layers to facilitate backpropagation through different convolutional layers and hence reducing gradient vanishing problem. This type of network has been widely applied to medical image segmentation, such as knee menisci segmentation from MRI[38], and knee cartilage tracking[39].

CUMedvision is a variant of FCN, which uses multi-level feature fusion to integrate both high-level and low-level features, making it excels in identifying objects with huge size differences on the image[16].

On the other hand, ResU-Net is another variant of U-Net, with the addition of residual blocks and skip connections [40]. The residual blocks in ResU-Net further assist in propagating low-level details to higher network layers, thereby facilitating more fine-grained segmentation of objects (*Figure 1*). Instead of the structure defined in the original work, a low complexity version of ResU-Net using 18 residual layers in place of 50 were applied as the network backbone, which accommodates a lower memory usage for training and better performance in radiographic images.

DeeplabV3 further extends ResU-Net by using dilated convolution, context module, spatial pyramid pooling, etc[20]. For the hyperparameters and network structures in Deeplab V3 and U-Net, we employed the default settings from PyTorch 1.7.0. While CUMedvision is in its settings following the original paper.

The four selected models are all in an encoder-decoder architecture [17], in which for each pixel, the neural networks classify whether it belongs to one of the four categories: femur, fibula, tibia, or background with a probability between 0 to 1 with a sigmoid function in the output layer. We compared their performance and subsequently selected the best performing model with the highest mean Intersect over Union (IoU) score.

Model training

In the training procedure of the four models, the deep network is implemented using PyTorch version 1.7.0. Adam optimizer with a learning rate of 0.001 was used, which provides a tradeoff between training time and accuracy. Weight decay with 1e-5 was used. Also, the early stopping strategy was also applied, which terminate the training when there is no loss improvement for 10 epochs to prevent overfitting.

Backpropagation optimizes parameters by minimizing the loss function using first-order gradient. All four network uses Binary Cross-Entropy (BCE) as the loss function, which aims to maximize the log-likelihood for correct predictions of the classes of each pixel.

See formula 2 in the supplementary files.

To tackle the issue of limited data, data augmentation was applied to improve the model generalization ability. Histogram normalization was used to maintain consistency across different image sets which were taken by different observers and equipment. Alongside, saturation and contrast jitter, translation, and random flipping were also applied in the augmentation process. Whereas the rotation and the horizontal shift of the images were +/- 5 degree and +/- 10%, respectively.

Quantitative measurement

Following the output of masks indicating the femur and tibia from the deep neural network, a program for automated calculation of JSWs was derived. Firstly, contours are being extracted from the femoral and tibial masks generated with Canny filters using OpenCV 3 package in Python. The horizontal distance of the extracted tibial plateau contour was normalized to a scale of 1. We denote this scale as a variable x. (*Figure 3*) The multi-point JSWs measurement was calculated in the range x=0.15~0.30 (lateral compartment) and x=0.7~0.9. at 0.05, 0.025, 0.0125 and 0.00625 intervals for 8-point, 16-point, 32-point and 64-point JSWs respectively. While for the mJSW, pixel distance between all pairs of pixels in the two contour segments of the condyles and tibial plateau were computed in the range x=0.7~0.9 (medial compartment) and finally the minimum distance was identified as the mJSW. The measurements were further normalized to a millimeter-scale using the flexion beams. To validate the estimation accuracy, we compared the mJSW calculated by our approach against the radiologists' measurements from the OAI database. Their correspondence was quantified using Pearson correlation and the difference was visualized by a Bland-Altman plot[25].

KOA severity and progression prediction

After the development of an automated JSWs measuring system, we randomly sampled 1760 bilateral X-ray images together with their corresponding KL-grades assessed by the radiologists from the OAI database (those used for training and validation of the segmentation models were excluded) and employed the algorithm to output the mJSW and multi-point JSWs accordingly. We defined the KOA severity using the 5-grade KL-grading system. A XGBoost model which is a tree-based method capable of capturing nonlinearity within the data structure [41], was trained using the estimated JSWs as input to classify the severity of KOA. The optimal hyperparameters of the model were obtained using grid-search with 5-fold cross-validation. From which the maximum depth, alpha, and lambda parameters were found to be 30, 1, and 1 respectively. In the next experiment, the disease progression was defined as an increase in KL-grade from unaffected (grade 0 and 1) to the confirmed case (grade 2 to 4) within 48 months. Moreover, samples that showed no progression and dropped out of the study before the 48-month follow-up were viewed as data with missing labels and were subsequently ruled out. After the selection, there

remain 945 pairs of knees. The grid-search procedure with 5-fold cross-validation was repeated for this experiment and the most optimal hyperparameters of the XGBoost model were identified to be maximum depth=25, alpha=0.5, and lambda=1. Both experiments were conducted with an 8:2 train-test split. We evaluated the model performance with the test set using the macro F1 and average area under receiver operating curve (AUC) scores for severity classification. Whereas the disease progression prediction, F1, and AUC scores were used. Lastly, in pursuit of comparing the performance of our CNN-based JSW estimation and those measured by radiologists in the prediction of disease severity and progression, we repeated the above experiments using the mJSW and 16-point JSWs from the OAI dataset.

Declarations

Acknowledgments:

This work was supported by Research Grants Council of Hong Kong Early Career Scheme (PolyU 251008/18M), PROCORE-France/Hong Kong Joint Research Scheme (F-PolyU504/18) and also Health and Medical Research Fund Scheme (01150087#, 15161391#, 16172691#).

Author Contributions:

YT, LC, JC, CW and PC conceived this review. YT and LC conducted literature search, systemic review and data analyses. All authors contributed to the writing of the manuscript and approved the final version.

Conflict of interests:

The authors have no relevant competing interests to disclose.

Data Availability Statement:

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

1 Oka, H. *et al.* Fully automatic quantification of knee osteoarthritis severity on plain radiographs. *Osteoarthritis and Cartilage* **16**, 1300-1306 (2008).

2 Gale, D. *et al.* Meniscal subluxation: association with osteoarthritis and joint space narrowing. *Osteoarthritis and Cartilage* **7**, 526-532 (1999).

3 Altman, R. D. & Gold, G. Atlas of individual radiographic features in osteoarthritis, revised. *Osteoarthritis and cartilage* **15**, A1-A56 (2007).

4 Tiulpin, A. & Saarakkala, S. Automatic grading of individual knee osteoarthritis features in plain radiographs using deep convolutional neural networks. *arXiv preprint arXiv:1907.08020* (2019).

5 Shamir, L. *et al.* Early detection of radiographic knee osteoarthritis using computer-aided analysis. *Osteoarthritis and Cartilage***17**, 1307-1312 (2009).

6 Gornale, S. S., Patravali, P. U. & Manza, R. R. Detection of osteoarthritis using knee x-ray image analyses: a machine vision based approach. *International Journal of Computer Applications***145** (2016).

7 Tiulpin, A., Melekhov, I. & Saarakkala, S. in *Proceedings of the IEEE International Conference on Computer Vision Workshops.* 0-0.

8 Petroudi, S., Loizou, C., Pantziaris, M. & Pattichis, C. Segmentation of the common carotid intima-media complex in ultrasound images using active contours. *IEEE transactions on biomedical engineering***59**, 3060-3069 (2012).

9 Goodfellow, I. *Deep learning*. (The MIT Press, 2016).

10 Tiulpin, A. *et al.* Multimodal Machine Learning-based Knee Osteoarthritis Progression Prediction from Plain Radiographs and Clinical Data. *Scientific Reports***9**, 20038, doi:10.1038/s41598-019-56527-3 (2019).

11 Wang, T., Leung, K., Cho, K., Chang, G. & Deniz, C. M. in *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track.*

12 Tiulpin, A., Finnilä, M., Lehenkari, P., Nieminen, H. J. & Saarakkala, S. in *International Conference on Advanced Concepts for Intelligent Vision Systems.* 131-138 (Springer).

13 Lindner, C. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 181-189 (Springer).

14 Thomson, J., O'Neill, T., Felson, D. & Cootes, T. in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 127-134 (Springer).

15 Antony, J., McGuinness, K., Moran, K. & O'Connor, N. E. in *International conference on machine learning and data mining in pattern recognition.* 376-390 (Springer).

16 Norman, B., Pedoia, V., Noworolski, A., Link, T. M. & Majumdar, S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. *Journal of digital imaging***32**, 471-477 (2019).

17 Badrinarayanan, V., Kendall, A. & Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence***39**, 2481-2495 (2017).

18 Chen, H., Qi, X. J., Cheng, J. Z. & Heng, P. A. in *Thirtieth AAAI conference on artificial intelligence.*

19 Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence***40**, 834-848 (2017).

20 Chen, L.-C., Papandreou, G., Schroff, F. & Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

21 Ronneberger, O., Fischer, P. & Brox, T. in *International Conference on Medical image computing and computer-assisted intervention.* 234-241 (Springer).

22 Bayramoglu, N., Nieminen, M. T. & Saarakkala, S. A Lightweight CNN and Joint Shape-Joint Space (JS2) Descriptor for Radiological Osteoarthritis Detection. *arXiv preprint arXiv:2005.11715* (2020).

23 Bartko, J. J. The intraclass correlation coefficient as a measure of reliability. *Psychological reports***19**, 3-11 (1966).

24 Bland, J. M. & Altman, D. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet***327**, 307-310 (1986).

25 Giavarina, D. Understanding bland altman analysis. *Biochemia medica: Biochemia medica***25**, 141-151 (2015).

26 Lindner, C. *et al.* in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 353-360 (Springer).

27 Hesamian, M. H., Jia, W., He, X. & Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging***32**, 582-596 (2019).

28 Mahapatra, D., Ge, Z., Sedai, S. & Chakravorty, R. 73-80 (Springer International Publishing).

29 Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S. & Pal, C. in *Deep Learning and Data Labeling for Medical Applications* 179-187 (Springer, 2016).

30 Neumann, G. *et al.* Location specific radiographic joint space width for osteoarthritis progression. *Osteoarthritis and cartilage***17**, 761-765 (2009).

31 Minciullo, L. & Cootes, T. in *2016 23rd International Conference on Pattern Recognition (ICPR).* 3787-3791 (IEEE).

32 Haverkamp, D. J., Schiphof, D., Bierma-Zeinstra, S. M., Weinans, H. & Waarsing, J. H. Variation in joint shape of osteoarthritic knees. *Arthritis & Rheumatism***63**, 3401-3407 (2011).

33 Dupuis, D. *et al.* Precision and accuracy of joint space width measurements of the medial compartment of the knee using standardized MTP semi-flexed radiographs. *Osteoarthritis and cartilage***11**, 716-724 (2003).

34 Chan, T. F. & Vese, L. A. Active contours without edges. *IEEE Transactions on image processing***10**, 266-277 (2001).

35 Lindner, C. *et al.* Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE transactions on medical imaging***32**, 1462-1472 (2013).

36 Gielis, W. *et al.* An automated workflow based on hip shape improves personalized risk prediction for hip osteoarthritis in the CHECK study. *Osteoarthritis and cartilage***28**, 62-70 (2020).

37 Mills, K., Hettinga, B. A., Pohl, M. B. & Ferber, R. Between-limb kinematic asymmetry during gait in unilateral and bilateral mild to moderate knee osteoarthritis. *Archives of Physical Medicine and Rehabilitation***94**, 2241-2247 (2013).

38 Zahra, E., Ali, B. & Siddique, W. Medical Image Segmentation Using a U-Net type of Architecture. *arXiv preprint arXiv:2005.05218* (2020).

39 Dunnhofer, M. *et al.* Siam-U-Net: encoder-decoder siamese network for knee cartilage tracking in ultrasound images. *Medical image analysis***60**, 101631 (2020).

40 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770-778.

41 Chen, T. & Guestrin, C. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* 785-794.

Tables

Table 1. Segmentation performance of different deep learning models

Models	Mean IoU	Validation Loss	Training Loss	Over-fitting (
CUMedVision	0.973	0.047	0.008	0.830
U-Net	0.594	0.410	0.409	0.002
Deeplab V3	**0.989**	0.011	0.005	0.545
ResU-Net-18 (ours)	**0.989**	0.006	0.004	0.333

Table 2. KL-grade classification performance using mJSW and 16-point JSWs from radiologists' measurement or CNN-based estimation using XGBoost model. The error represents the 95% confidence interval.

		Macro Average F1	Average AUC
CNN-based estimation	mJSW (single-point)	0.311 (±0.020)	0.587 (±0.017)
	16-point JSWs	0.402 (±0.030)	0.624 (±0.017)
Radiologist's measurement	mJSW (single-point)	0.337 (±0.027)	0.609 (±0.022)
	16-point JSWs	0.454 (±0.024)	0.655 (±0.014)

Table 3. KL-progression prediction performance using mJSW and 16-point JSWs from radiologists' measurement or CNN-based estimation using XGBoost model. The error represents the 95% confidence interval.

		Average F1	Average AUC
CNN-based estimation	mJSW (single-point)	0.484 (±0.041)	0.554 (±0.039)
	16-point JSWs	0.544 (±0.032)	0.583 (±0.040)
Radiologist's measurement	mJSW (single-point)	0.480 (±0.041)	0.551 (±0.024)
	16-point JSWs	0.562(±0.044)	0.613 (±0.018)

Figures

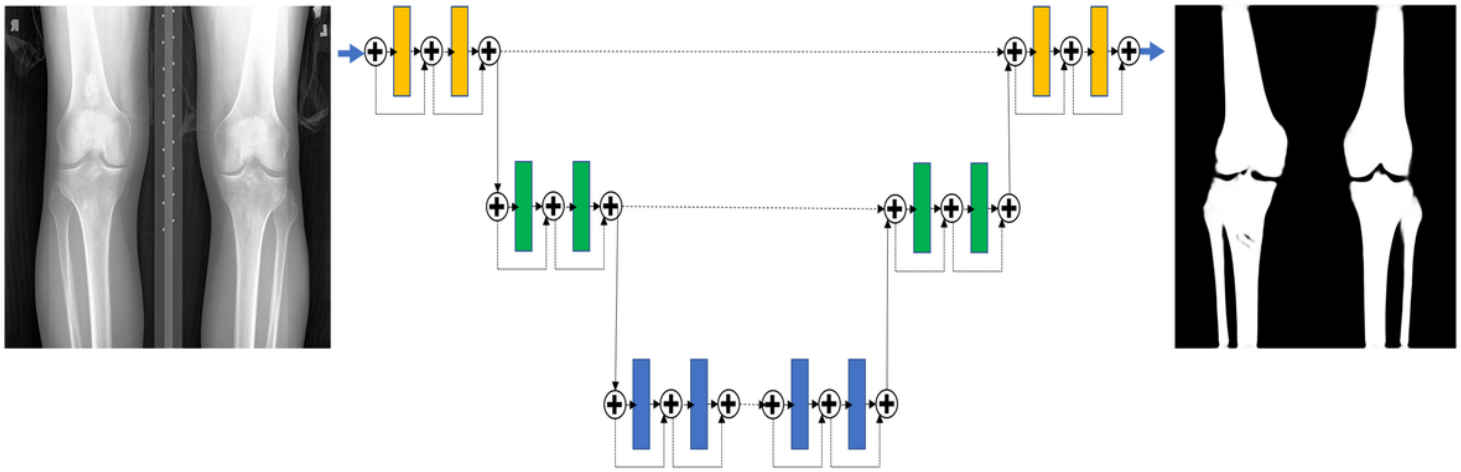

Figure 1

The ResU-Net-18 architecture of our knee joint segmentation convolutional neural network

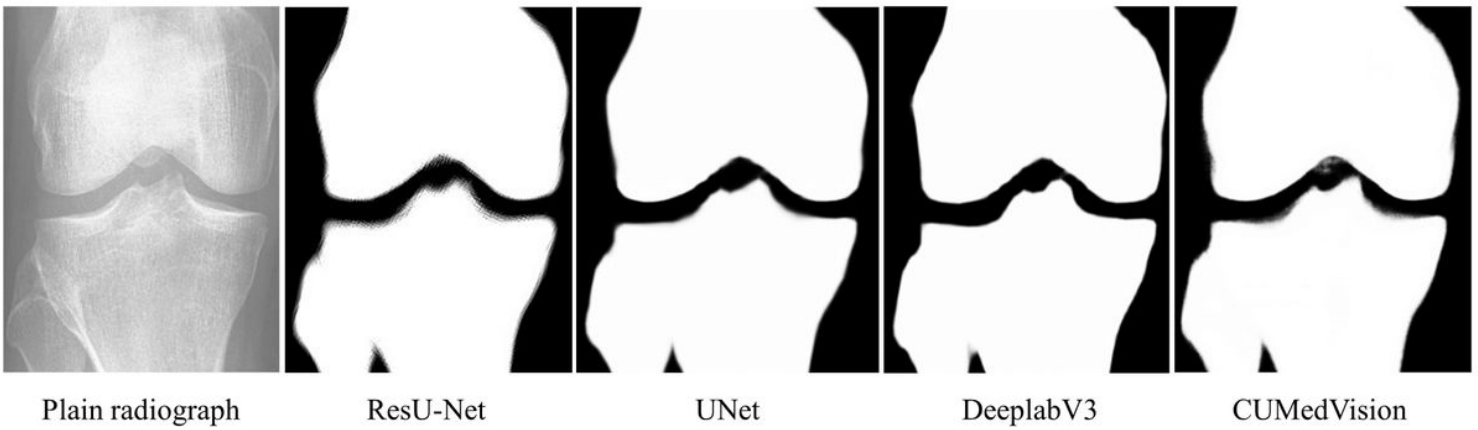

Figure 2

Comparison of masks produced by different semantic segmentation network architectures.

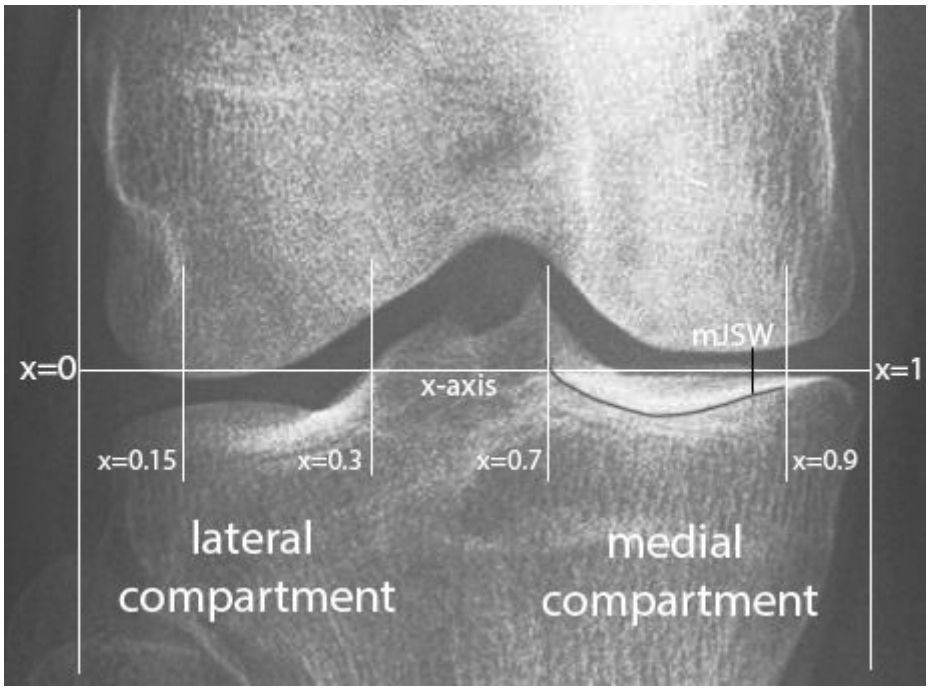

Figure 3

The definition of minimum joint space width (mJSW).

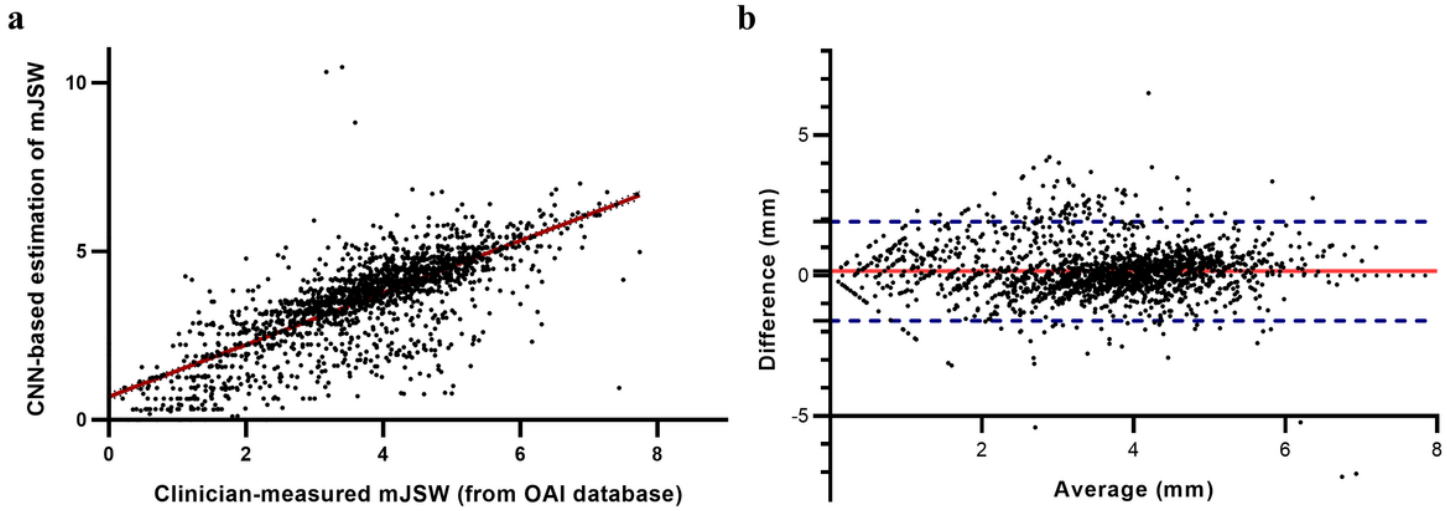

Figure 4

(a) Linear regression between mJSW measurements by radiologists or orthopedic doctors (from OAI database) and the proposed deep learning-based automated mJSW estimation approach. The regression line is colored in red with an R2 value of 0.6086, Pearson correlation being 0.7801 ($p<0.0001$). (b) Bland-Altman Plot for comparison between mJSW measurements by radiologists (from OAI database) and our CNN-based automated approach. The Blue dotted line indicates the 95% confidence interval of the

difference between the two types of measurement. The red solid line indicates the mean bias of the measurement.

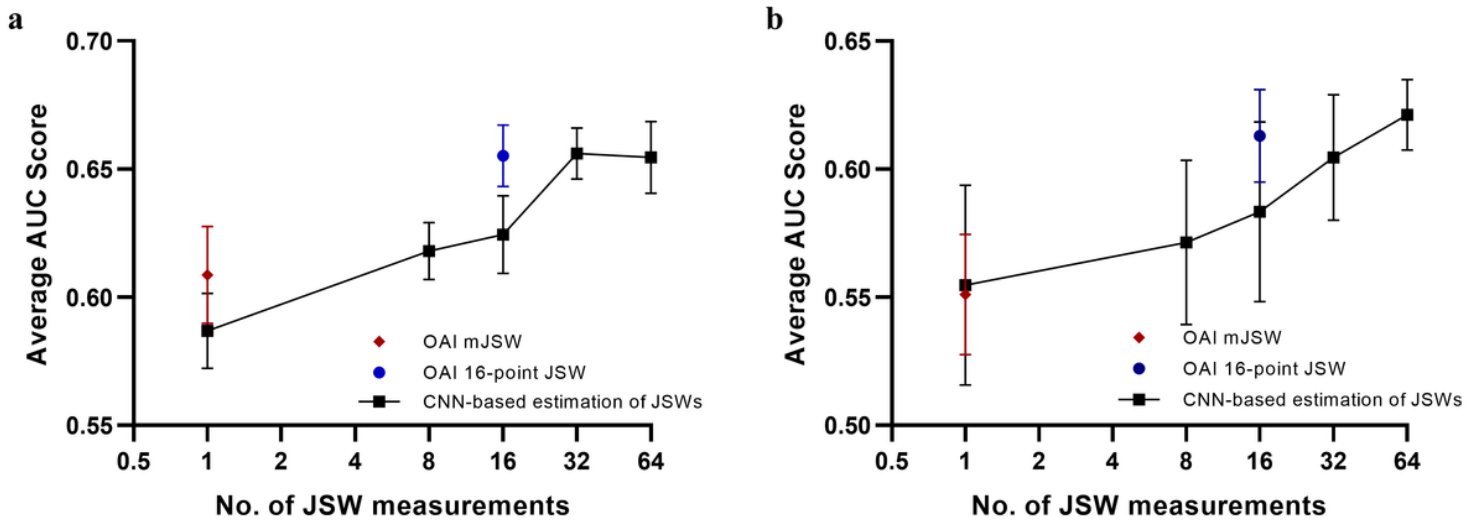

Figure 5

Performance of (a) KL-grades classification and (b) KOA progression prediction under different number of JSWs estimated by our CNN-based approach. The error bar represents the 95% confidence interval. The data points highlighted in red and blue represent the AUC scores of radiologist-measured mJSW and 16-point JSWs respectively.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- formulas.docx