misclassification between non-neighboring grades, only occurred in 0.3% for FOS, 2% for AOS, and 3% for JSN. A review of these cases by a board-certified musculoskeletal radiologist, showed that most of these either had inaccurate gradings or reduced image quality as demonstrated in the examples in Figure 2A and 2B. Gradient-weighted class activation maps as shown in Figure 2C demonstrate that the model focused on the region of the osteoarthritis feature for its assessment.

Conclusions: In this study we demonstrated that a deep learning based approach allowed for automated severity grading of radiographic features of hip OA. Clear grading errors were rare ranging from 0.3 to 3% in FOS, AOS, and JSN. The performance of the model detecting and grading those features was comparable to previously reported radiologist grading reliabilities. This model might aid radiologists in clinical practice in reading hip radiographs and improve detection and grading accuracy of radiographic hip OA features.

560
DEEP LEARNING PREDICTS KNEE OSTEOARTHRITIS PROGRESSION FROM PLAIN RADIOGRAPHS

A. Tiulpin[1], S. Klein[2], S. Bierma-Zeinstra[3,4], J. Thevenot[1], J. van Meurs[5], E. Oei[6], S. Saarakkala[1,7]. [1] *Res. Unit of Med. Imaging, Physics and Technology, Univ. of Oulu, Oulu, Finland, Oulu, Finland;* [2] *BioMed. Imaging Group Rotterdam, Depts. of Med. Informatics & Radiology, Erasmus MC, Univ. Med. Ctr. Rotterdam, Rotterdam, Netherlands;* [3] *Dept. of Gen. Practice, Erasmus MC, Univ. Med. Ctr. Rotterdam, Rotterdam, Netherlands;* [4] *Dept. of Orthopedics, Erasmus MC, Univ. Med. Ctr. Rotterdam, Rotterdam, Netherlands;* [5] *Dept. of Internal Med., Erasmus MC, Univ. Med. Ctr. Rotterdam, Rotterdam, Netherlands;* [6] *Dept. of Radiology & Nuclear Med., Univ. Med. Ctr. Rotterdam, Rotterdam, Netherlands;* [7] *Dept. of Radiology, Oulu Univ. Hosp., Oulu, Finland*

Purpose: Knee Osteoarthritis (OA) is currently diagnosed using clinical examination and plain radiography. Imaging-based diagnostic methods can be enhanced using novel quantitative methods, however, one clinically relevant challenge to address in OA is the prediction of its progression as well as the detection of early changes. In this study, we present a novel, fully-automatic Deep Learning (DL)-based method for OA progression prediction from plain radiographs. The method is compared to a reference approach based on logistic regression and externally validated on an independent test set.

Methods: We utilized two publicly available datasets - Osteoarthritis Initiative (OAI) and Multicenter Osteoarthritis Study (MOST). Bilateral radiographs were acquired with 10 degrees beam angles. Data from the baseline examination in OAI were used for training our method (2,738 subjects / 5,139 knees) and the baseline images from MOST were used for independent testing (1,352 subjects / 2,491 knees). Among these, the test set had 960 knees which did and 1531 knees which did not progress during the 84 months. We defined OA progression as an increase of Kellgren-Lawrence (KL) grade during the follow-up visits. Increase from KL4 to total knee replacement was also considered as a progression. Instead of using a binary outcome (non-progressor / progressor), we split the labels into different sub-classes: 0 = no progression within 84 months, 1 = progression earlier than 60 months, and 2 = progression between 60 and 84 months (including both follow-ups). Such label organization allowed to harmonize MOST and OAI data, thus, enabling the possibility of independent testing of the method. While predicting three classes, at the test stage, we summed the probabilities for classes 1 and 2 to obtain a probability of OA progression for a given knee. In order to assess the added value of our method, we first implemented a reference model: a logistic regression (LR) with regularization, which used Age, Sex, BMI and the KL grade for the given knee as input variables. With this reference model, we directly predicted dichotomous outcome (combined classes 1 and 2). We also trained LR models separately on Age, Sex, BMI and KL grade factors, to assess their individual associations with OA progression. Before training our DL-based approach, we first cropped the knee joints with the help of BoneFinder software to 140x140mm regions of interest, performed global contrast normalization, and rotated the knees images to horizontally align the tibial plateau using the landmarks from BoneFinder. We used a multi-task transfer learning and constructed a Deep Convolutional Neural Network with an ImageNet pre-trained model SE-ResNeXt 50 32x4d, which acted as an image feature extractor. Subsequently, two parallel tasks were trained using the extracted convolutional features - one to predict the KL grade for the current (baseline) image, and the other to predict the three aforementioned progression classes. Besides only predicting progression, we investigated the interpretability of our approach by using GradCAM - a state-of-the-art technique, allowing to examine decisions of neural networks. To assess the performance of the method, we used Area Under the Receiver Operating Characteristic Curve (AUC) and calculated the 95% confidence intervals using bootstrapping. Both for the reference LR and our proposed DL method, we used 5-fold subject-wise cross-validation (CV) to optimize the hyperparameters. To statistically compare the ROC curves for LR and DL, we utilized a DeLong's test.

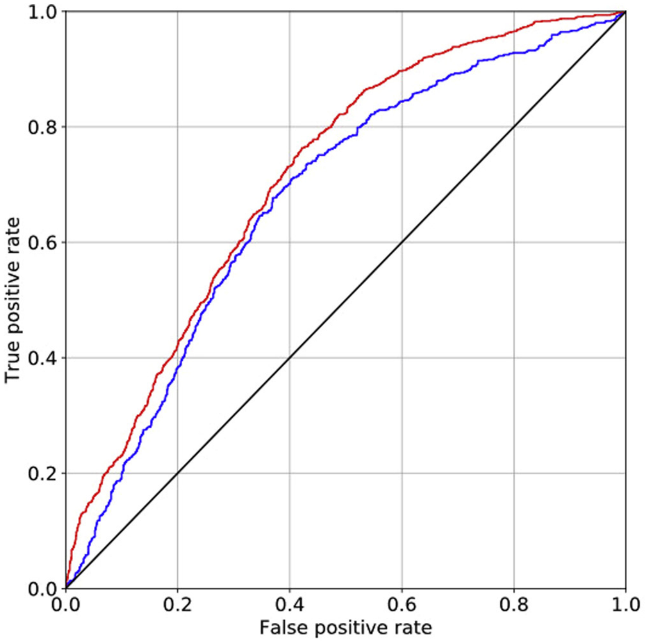

Figure 1. Comparison of the developed Deep Learning-based and the reference methods using Receiver Operating Characteristic (ROC) curves. Our model is based solely on knee X-rays and the reference method is a logistic regression trained on Age, Sex, BMI and the KL grade. Blue lines indicate the reference method and the red lines indicate our approach. The Area Under the ROC curve (AUC) for our method was 0.71 [0.69-0.73] and the reference method yielded an AUC of 0.68 [0.66-0.70] (DeLong's p-value < 1e-6).

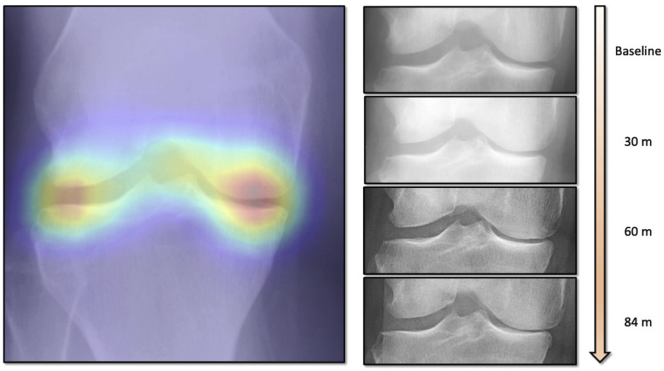

Figure 2. Example of OA progression detection from our model for a knee graded as doubtful osteoarthritis (KL grade 1). On the left, the locations on the baseline image that positively correlated with the output (as determined using GradCAM) are shown. On the right, the knee joints imaged during the follow-up examinations are shown. From this figure, it can be seen that the model recognised the early signs of progression already at baseline.

Results: The reference method yielded an AUC of 0.68 [0.66-0.70]. Age, Sex, BMI and KL grade individually yielded AUCs of 0.56 [0.54-0.58], 0.53 [0.51-0.55], 0.59 [0.57-0.61] and 0.65 [0.63-0.67], respectively. In contrast, our method yielded an AUC of 0.71 [0.69-0.73]. The ROC curves for the reference and proposed method are presented in Figure 1. The difference between them was found statistically significant (p-value < 1e-6). When inspecting the decisions made by our method using GradCAM, we found that in multiple cases the model is able to depict visually unobservable signs of progression in KL0 and KL1 cases. An example of such finding is presented in Figure 2.

Conclusions: In this study, we presented a novel method to predict knee OA progression solely from X-ray images. We found that our model outperforms a traditional logistic regression model. Potentially, the performance of the method can further be improved by incorporating injury history and symptomatic data. We expect the developed approach to be helpful in better patient selection for OA drug trials. The source codes of the developed method will be publicly released to allow full reproducibility of the experiments.

561

MRI-BASED MULTI-TASK DEEP LEARNING FOR CARTILAGE LESION SEVERITY STAGING IN KNEE OSTEOARTHRITIS

B. Astuto Arouche Nunes [1], I. Flament [1], R. Shah [1,2], M. Bucknor [1], T. Link [1], V. Pedoia [1,2], S. Majumdar [1,2]. [1] *Univ. of California San Francisco, Dept. of Radiology and BioMed. Imaging, San Francisco, CA, USA;* [2] *Ctr. for Digital Hlth.Innovation, UCSF, San Francisco, CA, USA*

Purpose: Semi quantitative scoring systems, such as the Whole-Organ Magnetic Resonance Imaging Score (WORMS) or MRI Osteoarthritis Knee Score (MOAKS), have been developed in an attempt to standardize Knee MRI reading. Despite grading systems being widely used in research setting the clinical application is hampered by the time and level of expertise needed to reliably perform the reading making the automation of this task appealing for a smoother and faster clinical translation. The goal of this study is to fill this void by capitalizing on recent developments in Deep Learning (DL) applied to medical imaging. Specifically, we aim to: i) create models to identify cartilage lesions (CLs) and assess its severity, ii) identify the presence of bone marrow edema lesions (BMELs), ii) combine the two models in a multi-task automated and scalable fashion to improve assessment accuracy.

Methods: 1,435 knee MRI from subjects with and without OA were collected from three previous studies (age = 42.79 ± 14.75 years, BMI = 24.28 ± 3.22Kg/m2, 48 males, 52 females). All studies used a high-resolution 3D fast spin-echo (FSE) CUBE sequence TR/TE = 1500/26.69ms, field-of-view = 14cm, matrix = 512-by-512, slice-thickness = 0.5mm, bandwidth = 50kHz).A 3D V-Net neural network (NN) architecture was used to learn segmentations of the 6 cartilage compartments using 480 manually segmented volumes as training/test data. In order to optimize the segmentation task, we utilized two V-net architectures. The first performed segmentations for 5 classes (Figure 1A), namely femur, tibia and patella cartilage, one class for meniscus and one for background (BG). The second V-net (Figure 1B), solves the problem of assigning 11 labels to the compartments segmented by the first V-net. The 11 classes are: patella, trochlea, medial and lateral tibia, medial and lateral femur cartilage, 4 menisci and BG). After applying the segmentation to the entire dataset, bounding boxes around the 6 cartilage compartments were extracted, resulting in 8,610 cartilage volumes of interest (cVOIs) (Figure 1C). cVOIs were randomly divided with a 65/20/15% split into training, validation, and holdout sets, keeping the distributions of lesion severity per compartment. 3 classes labeled were generated as follows: (1) No Lesion NL (WORMS 0 and 1), Partial Thickness Lesion - PT (WORMS 2, 3 and 4) and (3) Full Thickness Lesion - FT (WORMS 2.5, 5 and 6). Randomly generated 3-axis rotational (±25degrees) and zooming (±a factor of 20%) image augmentations were performed (Figure 1C). MOAKS grading can also be used for our study, however WORMS grades were available for all cVOI.

The distribution of lesions varies for each compartment, being Patella the one where we find the most balanced dataset throughout the lesion severity classes. Nonetheless, during preprocessing augmentation and up sampling were used, together with class weights applied to the loss functions during training to mitigate the unbalancing issue.

The lesion classification problem was divided in 3 steps: I) 3D CNN automatic CL severity 3-class classification (Figure 1E), II) automatic BMEL 2-class classification (same 3D CNN architecture used for CL classification, but with a 2-class output - Figure 1F) and III) The final optimal combination of the outputs of both DL networks were combined with demographics data and fed as input to a Gradient Boost classifier (Figure 1G), where a final lesion severity staging solution was output and applied to a holdout set.

Results: The first step on CL classification was to automatically classify lesions severity only with 3D volumetric image data. Overall accuracy for that classifier was 82.79% on the holdout set. The 3D architecture applied to BMEL 2-class allowed an accuracy for 79.5%. For the shallow classifier ensemble three class WORMS model, an overall accuracy of 85.6% was achieved when combining the 3D-CNN with demographics data. The count confusion matrix can be viewed in Figure 2, along with results for the combinations of the 3 classifiers used in our pipeline. A 3rd combination is also considered when using the 3D CNN cartilage predictions, demographics data and our BMEL models together as input for the shallow classifier, where it boosted the performance to a 86.7% overall accuracy. It is worth noting that the consideration of demographics alone boosts the performance, specially decreasing mispredictions of NL as FT. When considering our BMLE model (Figure 2C) we are able to finetunes PT and FT predictions. In an attempt to interpreted better our results misclassified cases were further inspected by experts (Figure 3).

Conclusions: By combining different anatomical structures (distinct cartilage compartments) and lesion classification grading for both cartilage and BMEL, we are moving towards multitask machine learning for lesion detection. The proposed approach is weakly supervised in the sense that it learns features using only image level labels (i.e., all that is known is the presence or absence of a lesion somewhere in the 3D volume). With the proposed approach, we were able to boost performance of our final classifiers by not simply focusing on what the fine tuning of a single purpose model could offer, but rather broadly considering related tasks that could bring additional information to our classification problem.

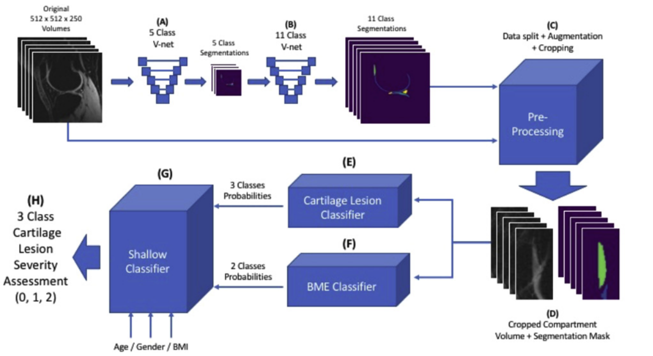

Figure 1: Fully Automated Multi-Task DL Pipeline: **(A)** 5-class cartilage compartment segmentation V-net. **(B)** The original image and its 5-class segmentations are used as input to another V-net, responsible for labeling the segmentations according to 11-classes. **(C)** Pre-processing pipeline including data splitting, bounding boxing and augmentation. **(D)** Volumes and the respective gradings are used respectively as input and labels in order to train: **(E)** 3D-CNN DL classifier to assess the presence and/or severity of a cartilage lesion, **(F)** Same 3D-CNN DL architecture used for CL classification, but with a 2-class output and trained to detect presence of BMEL. **(G)** Gradient Boost classifier, outputting **(H)** Lesion severity assessment (0:No Lesion, 1:Partial Thickness and 2:Full Thickness).

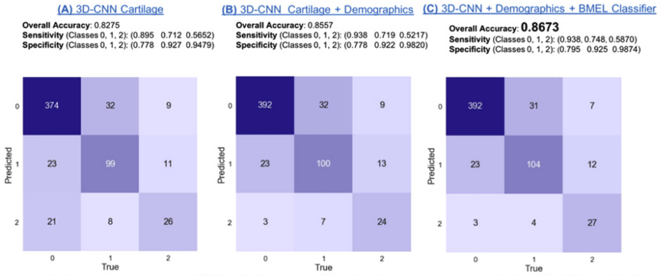

Figure 2: Accuracy assessment: **(A)** Confusion matrix(CF) showing the accuracy of the 3D-CNN CL predictions. **(B)** CF showing the accuracy of the 3D-CNN class probabilities outputs, together with patients demographics data, after passing through the Gradient Boost classifier. **(C)** CF for the 3D-CNN probabilities output, together with patients demographics data and the 2-class probability outputs from 2D-CNN, after passing through the Gradient Boost classifier.