Osteoporosis International (2022) 33:355–365
https://doi.org/10.1007/s00198-021-06130-y

ORIGINAL ARTICLE

Detecting hip osteoarthritis on clinical CT: a deep learning application based on 2-D summation images derived from CT

R. K. Gebre[1] · J. Hirvasniemi[2] · R. A. van der Heijden[2] · I. Lantto[3,4] · S. Saarakkala[1,4,5] · J. Leppilahti[3,4] · T. Jämsä[1,4,5]

Received: 23 December 2020 / Accepted: 20 August 2021 / Published online: 2 September 2021

Abstract

Summary We developed and compared deep learning models to detect hip osteoarthritis on clinical CT. The CT-based summation images, CT-AP, that resemble X-ray radiographs can detect radiographic hip osteoarthritis and in the absence of large training data, a reliable deep learning model can be optimized by combining CT-AP and X-ray images.

Introduction In this study, we aimed to investigate the applicability of deep learning (DL) to assess radiographic hip osteoarthritis (rHOA) on computed tomography (CT).

Methods The study data consisted of 94 abdominopelvic clinical CTs and 5659 hip X-ray images collected from Cohort Hip and Cohort Knee (CHECK). The CT slices were sequentially summed to create radiograph-like 2-D images named CT-AP. X-ray and CT-AP images were classified as rHOA if they had osteoarthritic changes corresponding to Kellgren-Lawrence grade 2 or higher. The study data was split into 55% training, 30% validation, and 15% test sets. A pretrained ResNet18 was optimized for a classification task of rHOA vs. no-rHOA. Five models were trained using (1) X-rays, (2) downsampled X-rays, (3) combination of CT-AP and X-ray images, (4) combination of CT-AP and downsampled X-ray images, and (5) CT-AP images.

Results Amongst the five models, Model-3 and Model-5 performed best in detecting rHOA from the CT-AP images. Model-3 detected rHOA on the test set of CT-AP images with a balanced accuracy of 82.2% and was able to discriminate rHOA from no-rHOA with an area under the receiver operating characteristic curve (ROC AUC) of 0.93 [0.75–0.99]. Model-5 detected rHOA on the test set at a balanced accuracy of 82.2% and classified rHOA from no-rHOA with an ROC AUC of 0.89 [0.67–0.97].

Conclusion CT-based summation images that resemble radiographs can be used to detect rHOA. In addition, in the absence of large training data, a reliable DL model can be optimized by combining CT-AP and X-ray images.

Keywords Classification · Computed tomography · Deep learning · Hip osteoarthritis · Radiology

Introduction

Hip osteoarthritis (OA) is a degenerative disorder characterized by the progressive loss of articular cartilage, subchondral sclerosis, subchondral cysts, osteophytosis, and altered shape of the hip joint bones [1–3]. Some of the risk factors of hip OA include age, gender, developmental disorders, heredity, bone mineral density (BMD), body mass index (BMI), smoking, and heavy physical activity [1, 4–8]. Assessment of hip OA involves radiographic investigation and clinical diagnosis of the joint [1–3, 9]. The most common assessment method of radiographic hip OA (rHOA) is the Kellgren and Lawrence (KL) severity grading [1, 2]. Clinical computed tomography (CT) has lower resolution than plain radiographs but provides detailed three-dimensional (3-D)

✉ R. K. Gebre
robel.gebre@oulu.fi

[1] Research Unit of Medical Imaging, Physics and Technology, University of Oulu, Oulu, Finland

[2] Department of Radiology & Nuclear Medicine, Erasmus MC University Medical Center, Rotterdam, the Netherlands

[3] Division of Orthopaedic and Trauma Surgery, Oulu University Hospital, Oulu, Finland

[4] Medical Research Center, University of Oulu and Oulu University Hospital, Oulu, Finland

[5] Diagnostic Radiology, Oulu University Hospital, Oulu, Finland

information on the joint bones [10, 11]. Consequently, due to the differences between the 3-D CT and two-dimensional (2-D) X-ray images, the KL grading is not directly applicable to CT. Although there have been some previous attempts to introduce CT-based hip OA severity grading [10, 11], there is still no widely accepted gold standard. Furthermore, grading hip OA severity on CTs can prove to be difficult and time consuming [11] especially for pelvic and hip studies that use clinical CTs, e.g., to investigate the relationship between proximal femur or acetabular fractures and hip OA. Hence, developing an automatic method of detecting the presence or absence of rHOA on CTs could be useful for such studies and enable efficient analysis of large datasets.

In the recent years, deep learning (DL) convolutional neural networks (CNNs) have been applied to automatically detect OA from plain radiographs of the hip [12–14]. The hip OA DL studies either adapted a rHOA classification based on the KL grades as training classes [12] or used binary classification to train their models [13, 14]. In addition, these studies collect large quantities of X-ray images to train, validate, and test their CNNs either from OA-based cohorts [12] or from hospitals [13, 14]. On the other hand, unlike the X-ray-based OA DL studies with access to OA cohort data, there are no large and freely available CT-based OA cohorts that can be used as sources of training data for DL OA studies. Therefore, the aim of this study was to develop and validate DL models using X-ray images and CT 2-D summation images, both separately and combined as training data, to assess rHOA on CT.

Materials and methods

Training data

Computed tomography images

The CT data consisted of abdominopelvic clinical images ($n = 94$, age range 50–95 years) that were scanned using standard protocols and obtained from the picture archiving and communication system of Oulu University Hospital, Oulu, Finland [15, 16]. The dataset consisted of 26 females (mean age ± standard deviation (SD): 69 years ± 14 years) and 68 males (67 years ± 9 years). A research permit (220/2017) was obtained from the Northern Ostrobothnia Hospital District, and written informed consent was not required due to the register-based study design. The pixel sizes and slice thicknesses of the CTs used in this study were 0.74 mm ± 0.09 mm and 0.80 mm ± 0.32 mm, respectively.

Anteroposterior (AP) radiograph-like 2-D images, referred to as CT-AP, were created from the CT slices (see "Creating AP radiograph-like images from CT data" section). These images were manually graded (RvdH, a senior resident in Radiology with musculoskeletal sub-specialization) into two classes, i.e., rHOA vs no-rHOA, based on the Kellgren and Lawrence (KL) grading of the radiographic OA features present in the image. The CT-AP images were classified into two classes based on the rHOA features that corresponded to those found in KL2 or higher (Fig. 1). A binary classification was chosen so that a clear delineation between the radiographic features could be learned by the DL model.

X-ray images

We used X-ray images collected from the Cohort Hip Cohort Hip (CHECK) database as part of the training set to develop the DL model. CHECK is a multicenter prospective cohort study formed by the Dutch Arthritis Foundation, to investigate the clinical, biochemical, and radiographical signs and symptoms of hip and knee OA [17]. A total of 1002 subjects (age range 45–65 years) participated for a 10-year follow-up from October 2002 to September 2005. Some of the inclusion criteria for the hip OA were pain, morning stiffness lasting less than 60 min, and a first consult with a general practitioner at or within 6 months from the onset of the symptoms. Some of the exclusion criteria were previous malignancies, pathologies, and comorbidities preventing follow-up [17]. Medical ethics guidelines were followed where all participants provided written informed consent [17]. The spatial resolution of the X-ray images was of pixel size 0.16 mm ± 0.02 mm.

The hip OA classification in the CHECK study was based on KL severity grading [2]. The KL grading scheme categorized radiographic OA features are as KL0 (none), KL1 (doubtful), KL2 (minimal), KL3 (moderate), and KL4 (severe). In this study, we used a binary classification system where KL0s were classified in the no-rHOA class, while KL2 and KL3 were categorized into one class of rHOA (Fig. 1).

A total of 10,020 hip joints from five patient visit time points (baseline, year 2, year 5, year 8, and year 10) were collected. Subsequently, 4360 joints were excluded due to unavailable KL grading, hip replacement, and insufficient image quality (artifacts or underexposed). In addition, KL1s were excluded because of the doubtful and ambiguous radiographic OA features. Lastly, the final dataset of X-ray images separated into two classes consisted of 3671 hip joints in the no-rHOA class and 1988 hip joints in the rHOA class.

Creating AP radiograph-like images from CT data

To detect rHOA features on CT data, we used a method for summing the slices to form an AP-style image, which we referred to as CT-AP. First, the CT slices were thresholded between − 150 Hounsfield units (HU) and + 600 HU. This

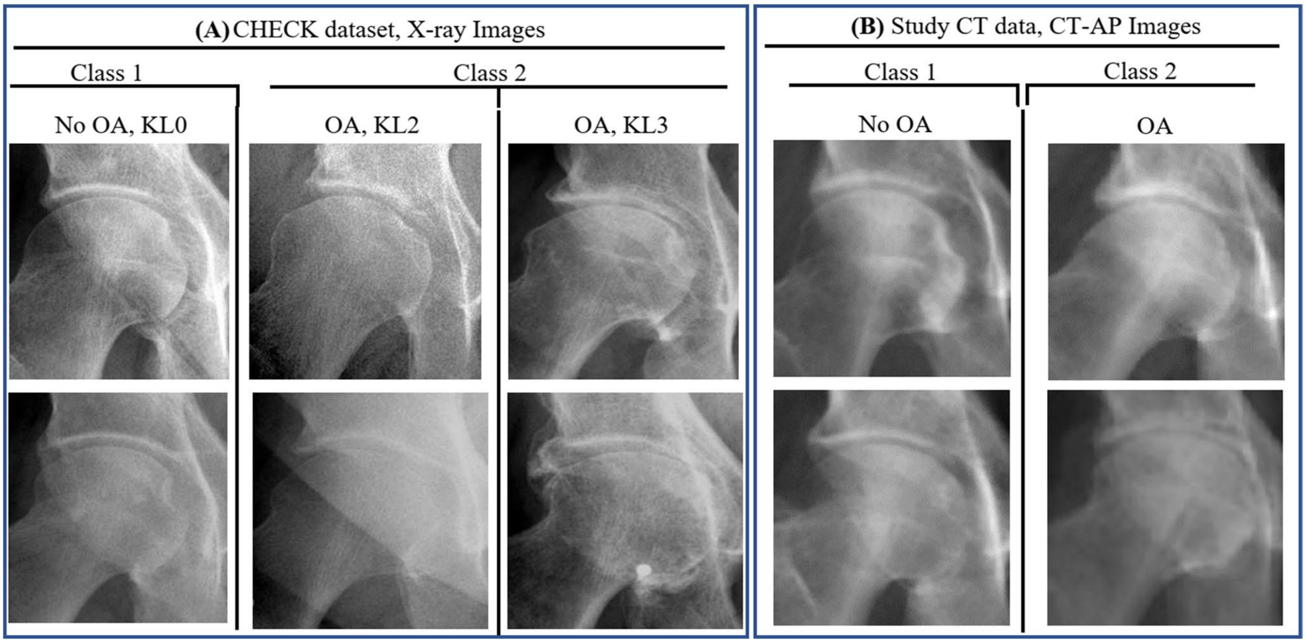
(A) CHECK dataset, X-ray Images
Class 1
Class 2
No OA, KL0
OA, KL2
OA, KL3
(B) Study CT data, CT-AP Images
Class 1
Class 2
No OA
OA

Fig. 1 Examples of the images used in the two classes for training the deep learning models: class 1 = no-rHOA and class 2 = rHOA. **A** The CHECK dataset X-ray images that were KL graded as part of the cohort study. **B** The CT-AP images that were manually OA graded in a binary classification as part of this study

custom HU threshold was chosen instead of the full HU range (− 1024 HU to + 1650 HU) because of the better quality of the resulting CT-AP images. In addition, with the full HU range, the CT-AP images showed a single narrow peak histogram, while a clear contrast of peaks and valleys could be seen in the histogram of the images created using the custom HU range (Supplementary Fig. 1).

Patient positioning at the time of scanning was not standardized since the scans were taken in clinical settings. Hence, before creating the CT-AP images, the slices were realigned to an AP (coronal) plane [15, 16]. Throughout this text "slices" refer to the coronal images reconstructed from the raw axial slices that were initially generated by the CT scanner. The realignment operation was conducted using a reference Cartesian coordinate system in Mimics and 3-Matic (Materialise, Leuven BE, Belgium) [15, 16]. Realignment was accomplished by first creating the AP plane on the surface of the 3-D pelvic models using the anatomical landmarks, Anterior Superior Iliac Spine (ASIS), and Pubic Tubercles (PT) and then parallelly reorienting this AP plane to the reference XY plane [15, 16]. Then, the reconstructed slices were re-sliced to be aligned to the AP plane [16]. The complete procedure for realigning the slices can be found here [16].

After the thresholding and realignment operations, the coronal slices were summed to form the final CT-AP image. These slices used for summation were selected manually where blank slices found at the top and bottom of the CT stack, slices containing partially scanned anatomy, and bad quality slices affected by motion artifacts were excluded. Summation was done sequentially, i.e., the next slice was added to the sum of slices before it and so forth (Fig. 2). This technique is similar to the sliding thin-slab maximum intensity projection technique originally introduced by Napel et al. for visualizing blood vessels and airways [18]. Other studies have also adapted this technique for reducing information overload in thin-section scans [19, 20].

Cropping and contrast adjustment of hip joints

Hip joints of the CHECK and CT-AP images were localized using a semi-automatic method involving faster-RCNN object detector and AlexNet network architecture. The faster RCNN object detector is a variant of the RCNN object detector with a network architecture especially suited for object detection purposes [21]. Briefly, the main faster-RCNN network region proposal network (RPN) gathers information present in the ground truth training images, such as the size and aspect ratio, in order to predict bounding boxes around the object [21]. Here, the RPN was created in MATLAB using the functions regionProposalLayer, rpnClsLayers, and rpnRegLayers and then finally trained using trainFasterRCNNObjectDetector.

In our study, the ground truth used to train the detector was created by manually drawing a rectangular region of interest (ROI) around the hip joints ($n = 3000$) of randomly

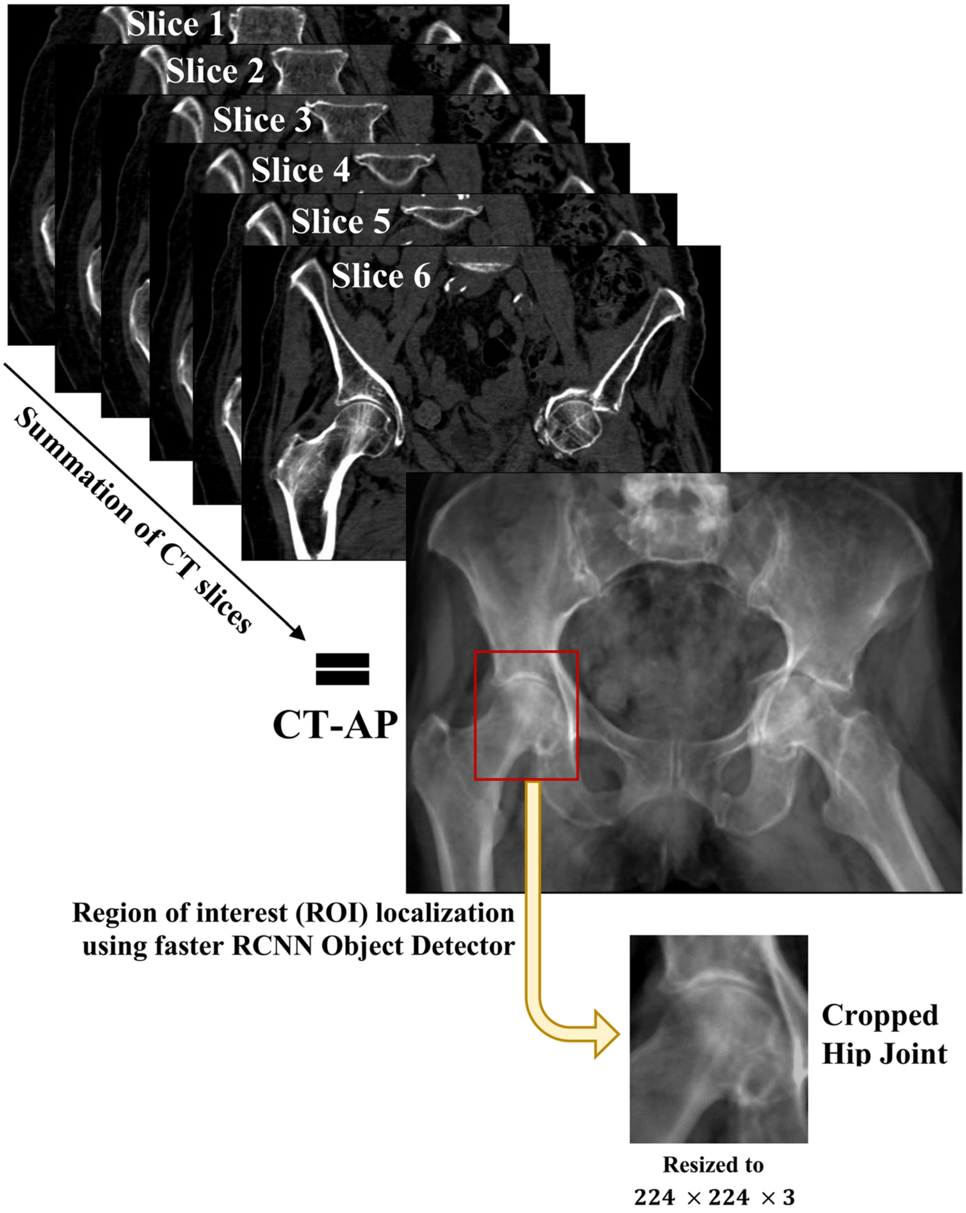

Fig. 2 Summation of CT slices to form the 2-D image referred to as CT-AP. The process of cropping the hip joints by localizing the region of interest using faster RCNN object detector is also shown

selected CHECK and CT-AP images. The *imageLabeler* function in MATLAB was used for this purpose. To standardize and simplify the labeling process, only the left sides were labeled where the AP images were cut at the vertical midline and the right sides reflected to left. In order to annotate the hip joint, a rectangular ROI was drawn to include the femoral head in the center of the ROI. The left border of the ROI was drawn to include the femoral neck up to but not reaching the greater trochanter, while the top and bottom borders were drawn to include the superolateral and inferolateral joint spaces of the acetabulum, respectively. The right border of the ROI was drawn to include the iliopectineal line. These annotations were performed by RKG (supervised by JH), who has 4 years of experience working on radiographic images of the hip and pelvis.

The labeled images in the ground truth were split into 90% ($n=2700$) and 10% ($n=300$) to be used for training and testing the detector, respectively. The object detector performed on the test set at an average precision of 0.99, F1-score of 0.67, and intersection of union (IoU) of 0.85 ± 0.07. It should be noted that the goal of using this detector was not to produce ROIs that were identical to one another, but rather to simplify the joint localization process and produce images with some differences so that the main DL models can learn more features as a result of this variability. In addition, due to processing capacity limitations, complex network architectures other than AlexNet or the inclusion of more training images were not possible. In practice, the ROI that was automatically detected by the network was visually inspected to ensure only the hip joint was localized before cropping.

To standardize the size of the ROIs, after cropping, the rectangular ROIs were made into squares by clipping the difference between their height and width. This operation did not result in loss of information. The resulting ROIs of the CHECK X-ray images were of size 603 pixels ± 194 pixels. Likewise, the ROIs of the CT-AP images resulted in sizes of 95 pixels ± 14 pixels which were then resized to 100 pixels × 100 pixels. Lastly, contrast adjustment was done on the CHECK X-rays by saturating the bottom and the top 1% of the pixel values. However, due to the HU thresholding conducted before slice summation, and to avoid saturating the images, contrast adjustment was not done on the CT-AP images.

Training the deep learning model

The DL model, ResNet18, was trained for assessing the presence or absence of rHOA. Different types of CNNs such as VGG [13], ResNet [22], and DenseNet [12] have previously been used in OA-related studies. In this study, ResNet18 was chosen after considering the graphics processing unit (GPU) capacity and time vs predictive accuracy when compared to other models. ResNet18 has a network architecture based on the ResNet CNNs and takes input data of size (224 × 224 × 3); details on the network architecture can be found in the original publication by He K et al. [23].

Five models (Model-1 to Model-5) were trained to investigate the performance of the different training data to detect rHOA. Model-1 was trained on the X-ray images from the CHECK cohort while Model-2 was trained on the same X-ray images that were downsampled to sizes of 100 pixels × 100 pixels to match the ROI sizes of the CT-AP images. Model-3 was trained with a combination of the CT-AP images and the X-ray images used in Model-1, while Model-4 was trained using the CT-AP images and the downsampled X-ray images from Model-2. Model-5 was trained using the CT-AP images only.

Network training was done on a 64-bit Windows 10 Enterprise (Microsoft Corporation) 32 GB RAM CPU (AMD-Ryzen 12-Core) computer with a 16 GB integrated single GPU (NVIDIA GeForce GTX 1070 Ti, NVIDIA).

Data partitioning

To train the networks, depending on the model, an input dataset of the cropped hip joints from the CHECK X-rays ($n = 5659$) and from the CT-AP images ($n = 94$) were used in combination or separately. This overall dataset containing both the CHECK and CT-AP images was split into 55% training ($n = 3158$), 30% validation ($n = 1722$), and 15% test ($n = 873$) (Table 1). Furthermore, the CT-AP images within the overall dataset were split into partitions of 50% training ($n = 46$), 25% validation ($n = 24$), and 25% test ($n = 24$) sets (Table 1). The validation sets were not used for training the models, and their purpose was to evaluate models' performances during the training process. In addition, the test sets were used to evaluate the performance of the models on unseen data.

Data augmentation

Data augmentation is a method of transforming training data so that a CNN can learn newer features without the need to collect more images [24, 25]. It has been shown to increase performance and predictive accuracy of task [24, 25]. Here, minimal data augmentation was performed on the training and validation partitions. These augmentations included rotation by randomly chosen angles between − 5 and + 5°, random horizontal and vertical translations by a distance between − 5 and + 5 pixels, and uniform isotropic scaling by a randomly chosen factor between 0.75 and 1.2. For the test set, augmentation was not done.

Transfer learning and training options

Transfer learning (TL) training technique was employed for training the DL model, ResNet18. TL is the technique where

Table 1 Data partitions used to train, validate, and test ResNet18 to predict radiographic hip osteoarthritis (rHOA). Model-1 and Model-2 were trained with unprocessed and downsampled CHECK X-ray images, respectively. Model-3 and Model-4 were trained on a combination of the CT-AP and X-ray images where the X-ray images were similar to the ones used in Model-1 and Model-2, respectively. Model-5 was trained solely on the CT-AP images. The overall images are the total of Cohort Hip and Cohort Knee (CHECK) X-ray and CT-AP images

Data partitions	CHECK		CT-AP		Combined		Overall (%)
	Model-1 and Model-2		Model-5		Model-3 and Model-4		
	rHOA	no-rHOA	rHOA	no-rHOA	rHOA	no-rHOA	
Training	1093	2019	29	17	1122	2036	3158 (55%)
Validation	597	1101	15	9	612	1110	1722 (30%)
Test	298	551	15	9	313	560	873 (15%)
Total	1988	3671	59	35	2047	3706	5753 (100%)

CNNs pre-trained on natural images such as ImageNet are fine-tuned to fit a new task such as OA classification [12, 13, 22, 26–28]. In this study, the ResNet18 was pretrained on ImageNet. In practice, TL involves replacing the final three layers, i.e., the fully connected, soft max, and classification layers to fit the new task which in our case is a binary class classification of the presence or absence of rHOA. After end-layer replacement and adjusting for number of classes, the network was fine-tuned by varying the training options. Here, we used the adaptive moment (ADAM) gradient-based optimizer [29], a mini-batch size of 32 for Model-1 to Model-5. An initial learning rate of 10^{-4} was used for all models. The maximum epochs were 40 for Model-3 and Model-4, 85 for Model-1 and Model-2, and 100 for Model-5. The learning drop factor was 0.1 for Model-5 and 0.05 for Model-1 to Model-4. The learning drop period was 10 epochs for all models. Hence, given these training options, the networks were trained for a maximum of 8245 iterations for Model-1 and Model-2, 4040 iterations for Model-3 and Model-4, and 1000 iterations for Model-5. The validation accuracy was taken was at every 53rd iteration for Model-1 to Model-4 and the 8th iteration for Model-5. Every epoch was shuffled during each iteration to increase the randomness in the data while the same seed number was used for all models for repeatability. Lastly, visualizations of the features with a positive contribution for the predicted class were shown using occlusion sensitivity [30].

A custom MATLAB (version R2019b, The MathWorks, Inc., Natick, MA, USA) script was written to perform the operations. The codes written for training the deep learning models can be found in our publicly available GitHub repository (https://github.com/MIPT-Oulu/Radiographic-Hip-OA-DL-Trainer).

Statistical analyses

To evaluate the models' classification performance, the accuracy, precision, recall, and F1-scores were determined from the predicted and true classes. Furthermore, a receiver operator characteristic curve (ROC) was created to determine the area under the curve (AUC). In addition, due to the unbalanced classes in the dataset, balanced accuracy and precision-recall curves AUC were also determined. These statistical analyses were done in MATLAB.

Results

Performance of the deep learning models to detect radiographic hip OA on X-ray images

When analyzing the classification performances of Model-1 to Model-4 to discriminate rHOA from no-rHOA on the test sets, they performed at ROC AUC values of 0.98 [95% confidence interval [CI]: 0.97–0.98], 0.97 [0.96–0.98], 0.98 [0.97–0.99] and 0.94 [0.97–0.98], respectively (Table 2). For Model-5 that was trained on only the CT-AP images, the classification performance to discriminate rHOA on the X-ray test sets was at a ROC AUC of 0.69 [0.66–0.72] for the unprocessed images and 0.58 [0.54–0.61] for the downsampled images (Table 2). In addition, to detect the presence or absence of rHOA on the test sets, Model-1 to Model-4 performed at accuracies of 92.3%, 90.2%, 91.9%, and 91.3%, respectively (Table 2). Similarly, Model-1 to Model-4 performed at precisions and F1-scores of 0.91 and 0.92, 0.88 and 0.89, 0.91 and 0.91, and 0.91 and 0.91, respectively (Table 2). Furthermore, Model-5 was able to detect the presence or absence of rHOA on the test sets of the unprocessed and the downsampled X-ray images at accuracies, precisions, and F1-scores of 49.1%, 0.59 and 0.61, and 60.5%, 0.55, and 0.55, respectively (Table 3).

Performances of the deep learning models to detect radiographic hip OA on the CT-AP images

Model-1 and Model-2 that were trained on only X-ray images detected rHOA on the test sets of the manually graded CT-AP images at accuracies of 50.0%, F1-scores of 0.65, and precisions of 0.60, respectively (Table 3 and Fig. 3). In addition, Model-1 and Model-2 classified rHOA from no-rHOA at ROC AUCs of 0.73 [0.49–0.89] and 0.63 [0.37–0.84], respectively (Table 3 and Fig. 3B).

Model-3 and Model-4 which were trained on the combination of CT-AP and X-ray images detected rHOA on the CT-AP test sets at accuracies of 83.3% and 75.0%, precisions of 0.82 and 0.80, and F1-scores of 0.82 and 0.80, respectively (Table 3 and Fig. 3). In addition, Model-3 and Model-4 were able to classify rHOA from no-rHOA at ROC AUCs of 0.93 [0.72–0.99] and 0.87 [0.64–0.97], respectively (Table 3 and Fig. 3B).

When assessing the performance of Model-5 that was trained on only the CT-AP images to detect rHOA on its test set, the accuracy was 83.3% and the precision and F1-score were 0.82 and 0.82 (Table 3 and Fig. 3). In addition, Model-5 was able to classify rHOA from no-rHOA at ROC AUC of 0.89 [0.67–0.97] (Table 3 and Fig. 3B).

Visualization of the learned features for Model 3 is shown in Fig. 4. Outputs for the training process for the five models are shown in Supplementary Figs. 2–6.

Discussion

In the present study, we optimized DL models for assessing radiographic hip osteoarthritis (rHOA) from computed tomography images. A two-step method was developed in

Table 2 Performances of the five models optimized for detecting radiographic hip osteoarthritis on the X-ray images within the validation and test datasets. The CT-AP images were created by sequentially summing the CT slices. Model-1 was trained with unprocessed X-ray images. Model-2 was trained similarly with downsampled X-rays to resemble CT-AP images. Model-3 and Model-4 were trained on a combination of the CT-AP and X-ray images where the X-ray images were similar to the ones used in Model-1 and Model-2, respectively. Model-5 was trained solely on the CT-AP images

Trained models	Accuracy	Balanced accuracy	Precision	Recall	F1-score	PR AUC [95% CI]	ROC AUC [95% CI]
X-ray images of the validation dataset							
Model-1	93.3	92.1	0.92	0.93	0.93	0.96 [0.95–0.97]	0.98 [0.98–0.99]
Model-2	90.6	89.2	0.89	0.89	0.89	0.94 [0.92–0.95]	0.97 [0.96–0.97]
Model-3	92.7	92.3	0.92	0.92	0.92	0.95 [0.93–0.95]	0.98 [0.97–0.98]
Model-4	90.4	89.2	0.89	0.89	0.89	0.94 [0.93–0.95]	0.94 [0.97–0.98]
X-ray images of the test dataset							
Model-1	92.3	88.5	0.91	0.92	0.92	0.96 [0.95–0.98]	0.98 [0.97–0.98]
Model-2	90.2	88.1	0.88	0.90	0.89	0.95 [0.93–0.96]	0.97 [0.96–0.98]
Model-3	91.9	91.1	0.91	0.91	0.91	0.95 [0.93–0.96]	0.98 [0.97–0.98]
Model-4	91.3	90.8	0.91	0.90	0.91	0.93 [0.91–0.95]	0.97 [0.96–0.98]
Model-5†	49.1	58.9	0.59	0.64	0.61	0.53 [0.50–0.56]	0.69 [0.66–0.72]
Model-5‡	60.5	55.2	0.55	0.56	0.55	0.57 [0.52–0.62]	0.58 [0.54–0.61]

†Performance of Model-5 on the X-ray images used in Model-1

‡Performance of Model-5 on the downsampled X-ray images used in Model-2

PR AUC, area under the precision recall curve; *ROC AUC*, area under the receiver operating characteristics curve; *CI*, confidence interval

Table 3 Performances of the five models optimized for detecting of radiographic hip osteoarthritis on the CT-AP images within the validation and test datasets. The CT-AP images were created by sequentially summing the CT slices. Model-1 was trained with unprocessed X-ray images. Model-2 was trained similarly with downsampled X-rays to resemble CT-AP images. Model-3 and Model-4 were trained on a combination of the CT-AP and X-ray images where the X-ray images were similar to the ones used in Model-1 and Model-2, respectively. Model-5 was trained solely on the CT-AP images

Trained models	Accuracy	Balanced accuracy	Precision	Recall	F1-score	PR AUC [95% CI]	ROC AUC [95% CI]
CT-AP images of the validation dataset							
Model-3	79.2	79.0	0.79	0.78	0.78	0.90 [0.79–0.94]	0.93 [0.73–0.98]
Model-4	87.5	83.3	0.83	0.92	0.87	0.88 [0.76–0.94]	0.91 [0.69–0.99]
Model-5	83.3	83.3	0.83	0.82	0.83	0.76 [0.52–0.91]	0.93 [0.71–0.99]
CT-AP images of the test dataset							
Model-1	50.0	60.0	0.60	0.71	0.65	0.79 [0.63–0.93]	0.73 [0.49–0.89]
Model-2	50.0	60.0	0.60	0.71	0.65	0.72 [0.55–0.89]	0.63 [0.37–0.84]
Model-3	83.3	82.2	0.82	0.82	0.82	0.89 [0.79–0.94]	0.93 [0.72–0.99]
Model-4	75.0	80.0	0.80	0.80	0.80	0.87 [0.77–0.94]	0.87 [0.64–0.97]
Model-5	83.3	82.2	0.82	0.82	0.82	0.80 [0.57–0.92]	0.89 [0.67–0.97]

PR AUC, area under the precision recall curve; *ROC AUC*, area under the receiver operating characteristics curve; *CI*, confidence interval

which we first created an AP-style image from the CT study data, by sequentially summing the slices and constituting them into the training data. Hence, five different models were developed by varying the training data used. Their performances were evaluated on a separate CT-AP test set. The DL model trained solely on the CT-AP images (Model-5) was able to detect and classify rHOA on the CT-AP images with a balanced accuracy of 82.2% and ROC AUC of 0.89. In addition, the model trained on a combination of the CT-AP and unprocessed X-ray images (Model-3) had a balanced accuracy of 82.2% and ROC AUC of 0.93.

There is limited literature on the application of 2-D summation images to grade hip OA, either manually or automatically. The one prominent example is the study by Turmezei et al. that introduced a CT-based hip OA scoring system by making use of multiplanar reconstruction [11]. The CT-AP image used in our study is similar to the digitally reconstructed radiograph (DRR) since both rely on the sliding

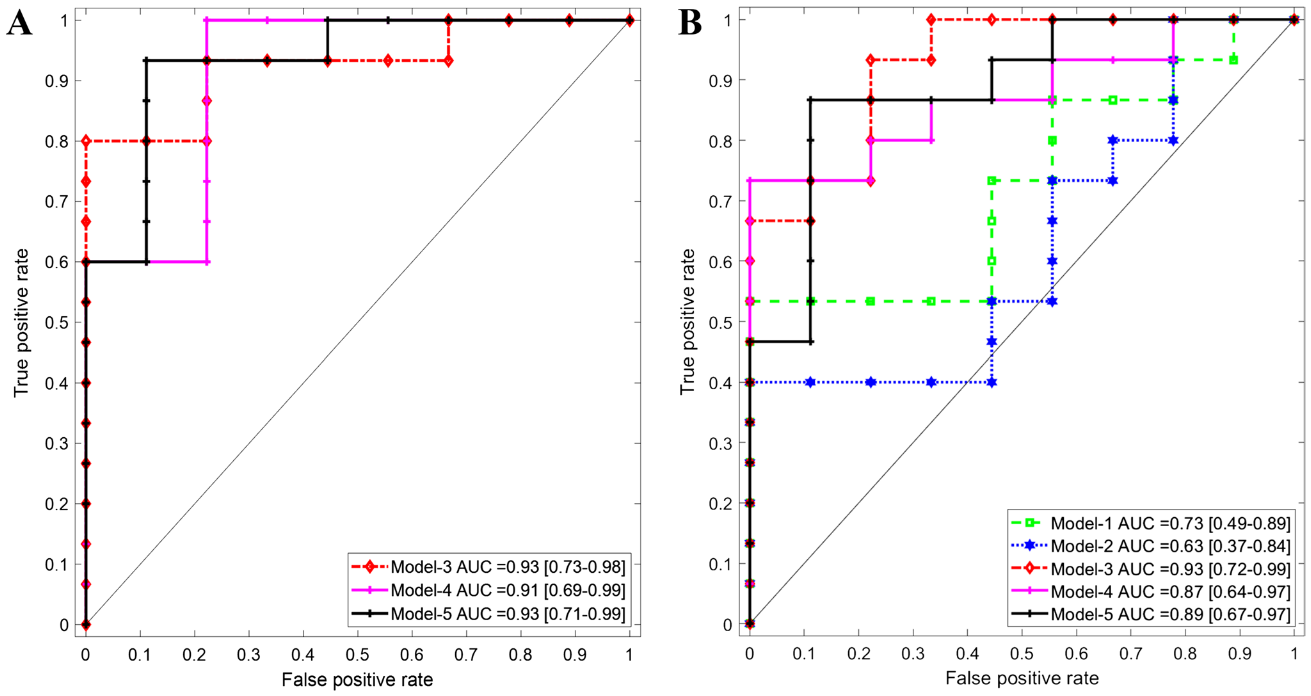

Fig. 3 Classification performances of the models on the CT-AP images shown as area under the curve (AUC) values of the receiver operator characteristic (ROC) curves. **A** ROC AUCs of Model-3, Model-4, and Model-5 for the validation dataset. **B** ROC AUCs for all the five models of the test dataset. Model-1 was trained with unprocessed X-ray images and Model-2 with only the downsampled X-rays images. Model-3 and Model-4 were trained on a combination of the CT-AP and X-ray images where the X-ray images were similar to the ones used in Model-1 and Model-2, respectively. Model-5 was trained solely on the CT-AP images

thin-slab maximum intensity projection technique [18]. Turmezei et al. based their hip OA grading on three separate assessments of joint space narrowing, osteophytes, and subchondral cysts [10, 11]. Their scoring system was diverse, and they reported high reliability, especially for their composite three-class scoring system [11]. Another more recent example is the OsteoArthritis Computed Tomography (OACT) score that was developed by Gielis et al. for assessing structural OA in large joints and the spine [10]. They modified the scoring system that was developed by Turmezei et al. to a simplified four-grade system in order to conserve time and increase reliability [10].

In our study, Model-1, Model-2, and Model-5 were developed to investigate the effectiveness of the pretrained ResNet18 CNN architecture to learn and to detect rHOA features on the CT-AP images, and vice versa. From the results of Model-1 and Model-2, it is possible that the models trained with only X-ray images were able to detect some rHOA features similarly found in the CT-AP images and that there could be some similarities between the CT-AP and the X-ray images. In addition, whether the X-ray images in the training set were unprocessed or downsampled did not have a noticeable effect on the two models' performances, which was also the case for Model-3 and Model-4. This was primarily due to the necessary resizing operation, which was performed prior to training, to meet the network input size requirement (224 pixels × 224 pixels), i.e., images were downsampled for Model-1 and Model-3, and upsampled for Model-2 and Model-4, resulting in similar performances for the models. In addition, even though there were fewer CT-AP images in Model-3 and Model-4 compared to the large number of X-ray images, the models' performance to detect rHOA on CT-APs improved noticeably. The increased performance of the combined image models could be explained by the specific network training options applied, data variability, and data quality of the CHECK KL-grading and CT-APs' KL-based classifications. Furthermore, Model-5 was able to detect some rHOA features on the unprocessed and downsampled X-ray images with similar performance. Unlike Model-1 and Model-2, Model-5's performances on the X-ray images were low, possibly due to the small number of training data used.

The performances of all the models developed in this study are well comparable to prior X-ray-based hip OA DL studies [12–14]. For instance, Xue et al. investigated the diagnostic value of DL in hip OA by training a VGG-16 CNN on 420 hip X-ray images categorized into a binary class (*normal* vs *abnormal*) [14]. They reported a high accuracy of 92.8% and ROC AUC of 0.94 and reported the classification performance of their model as being comparable

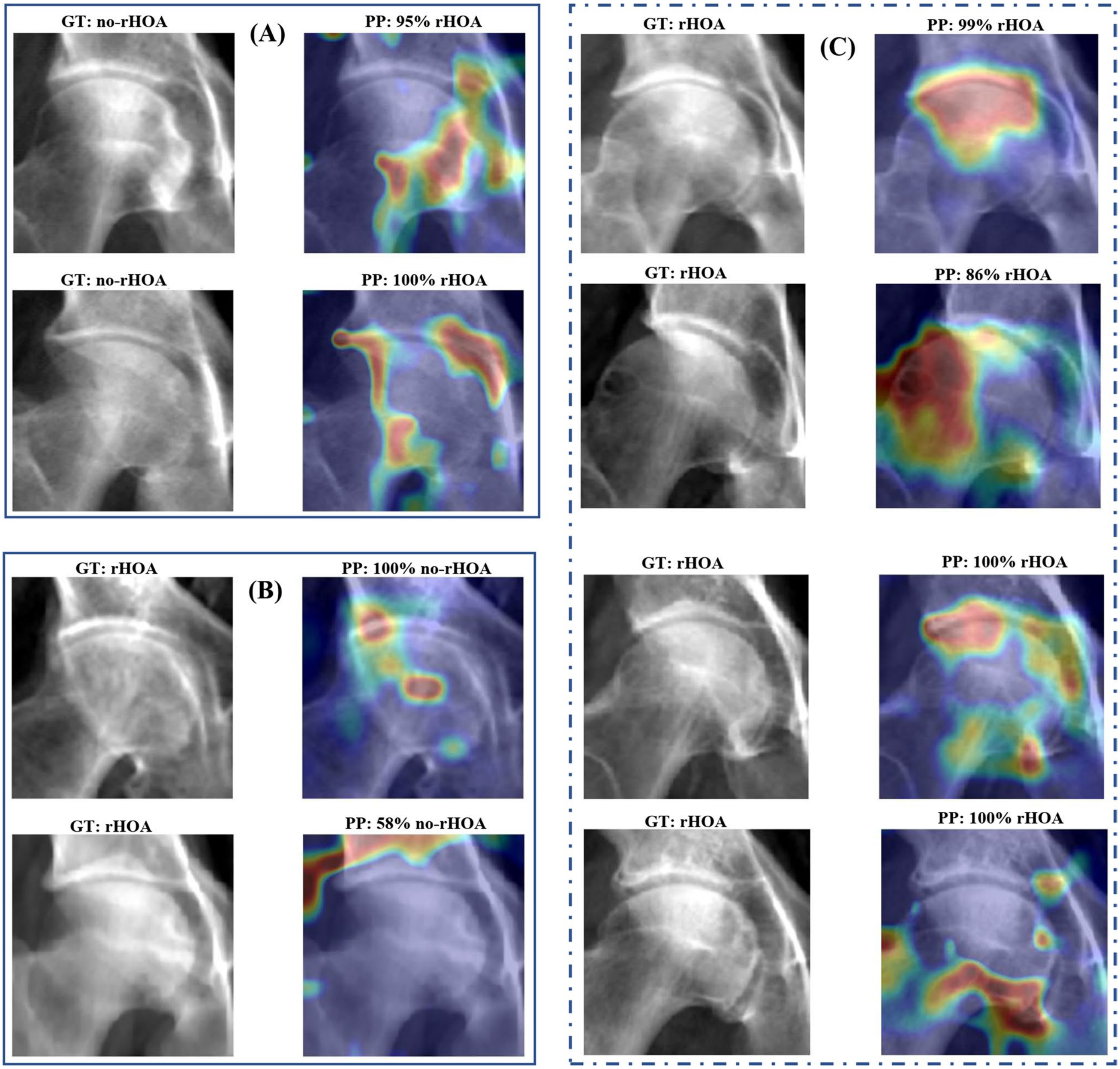

Fig. 4 Examples of radiographic hip osteoarthritis (rHOA) and no-rHOA features learned by ResNet18 that was trained using a combination of CT-AP and X-ray images. The images shown are CT-AP images. The prediction probabilities (PP) are on the top right, and the ground truth (GT) from the manual grading is on the top left sides of the images. The red bright areas indicate learned features contributing more to the highest predicted probability class. **A** Misclassification of no-rHOA as rHOA. **B** Misclassification of rHOA as no-rHOA. **C** The different learned features which were accurately classified into the rHOA class

to that of an experienced attending physician [14]. Recently, von Schacky et al. using the Osteoarthritis initiative (OAI) data ($n = 15{,}364$ hip joints) developed a DenseNet model to evaluate different OA features and reported an F1-score of 63% for predicting KL grades [12]. Another recent study by Üreten K. et al. using 868 hips collected from hospital developed a VGG model for a binary OA classification and reported a 90% accuracy [13].

In our study, the visualizations of the contributing image features suggest that joint space narrowing, changes in the shape of the joints, and osteophytes were learned by Model-3 as different rHOA features. Accurate detection and misclassification of these features were also indicated at the correct anatomical locations on the CT-AP images (Fig. 4). Further studies to understand the different criteria that were learned and used by the models in comparison to the known OA scoring systems could be beneficial.

Our study has some limitations. First, a small number of CT-AP images were used to train Model-5. Although the high-performance results were found for Model-5, further studies with more data are needed to confirm the findings. Second, we used a binary classification for OA. Since the objective of this study was to show the applicability of DL to detect hip OA on CT and since there was a small number of CT-AP data, further division other than the current binary class would have affected the data points in the individual classes as well as the models' performances. DL models for multiclass OA classification could be developed in the future with more training data. Third, other known risk factors of hip OA were not considered. Fourth, since the objective of our study was to develop and validate a methodological pipeline for future CT-based DL OA studies and considering that we are using a binary classification, the OA grading was done by one rater which could have biased the CT-AP grading. Lastly, some of the CT-AP images were blurrier than typical plain radiographs due to the summation operation and pixel size of CT. However, the performance and visualization of the contributing image features indicate the impact of the blurring on the results was low.

In conclusion, we were able to develop DL models to assess rHOA on CT data by creating a 2-D summation image of the slices. The motivation of this study was the gap in the availability of a large volume of CT data for DL-based hip OA studies. In our study, we showed a network such as ResNet18 which had been pretrained on a different set of images such as ImageNet can be optimized for detection of rHOA using transfer learning. Future hip or pelvic CT-based studies that aim to investigate hip OA can further adapt the method presented in this study. For instance, by first training a network on X-ray images to adjust the pretrained weights and then retraining this network on CT data, it is possible to achieve higher performance and reliability. Furthermore, such automatic DL models can be advantageous in saving time and resources. Although the initial training can be time consuming, once a reliable model is validated, the detection of rHOA features is extremely fast and can be streamlined to analyze entire datasets in a very short time.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00198-021-06130-y.

Funding Open access funding provided by University of Oulu including Oulu University Hospital. This study was financially supported by CINOP Global (NICHE/ETH/246) funded by EP-Nuffic.

The CHECK-cohort study is funded by the Dutch Arthritis Foundation. Involved are Erasmus Medical Center Rotterdam; Kennemer Gasthuis Haarlem; Leiden University Medical Center; Maastricht University Medical Center; Martini Hospital Groningen/Allied Health Care Center for Rheumatology and Rehabilitation Groningen; Medical Spectrum Twente Enschede /Ziekenhuisgroep Twente Almelo; Reade Center for Rehabilitation and Rheumatology; St. Maartens-kliniek Nijmegen; University Medical Center Utrecht and Wilhelmina Hospital Assen.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Lespasio MJ, Sultan AA, Piuzzi NS et al (2018) Hip osteoarthritis: a primer. Perm J 22:89–94. https://doi.org/10.7812/TPP/17-084
2. Kellgren JH, Lawrence JS (1957) Radiological assessment of osteo-arthrosis. Ann Rheum Dis 16:494–502. https://doi.org/10.1136/ard.16.4.494
3. Croft P, Cooper C, Wickham C, Coggon D (1990) Defining osteoarthritis of the hip for epidemiologic studies. Am J Epidemiol 132:514–522. https://doi.org/10.1093/oxfordjournals.aje.a115687
4. Chen FP, Fu TS, Lin YC, Fan CM (2018) Risk factors and quality of life for the occurrence of hip fracture in postmenopausal women. Biomed J 41:202–208. https://doi.org/10.1016/j.bj.2018.04.001
5. Cumming RG, Klineberg RJ (1993) Epidemiological study of the relation between arthritis of the hip and hip fractures. Ann Rheum Dis 52:707–710. https://doi.org/10.1136/ard.52.10.707
6. Arden NK, Griffiths GO, Hart DJ et al (1996) The association between osteoarthritis and osteoporotic fracture: the Chingford study. Br J Rheumatol 35:1299–1304. https://doi.org/10.1093/rheumatology/35.12.1299
7. Hunter DJ, Mcdougall JJ, Keefe FJ et al (2009) The symptoms of OA and the genesis of pain. Rheum Dis Clin North Am 34:1–19. https://doi.org/10.1016/j.rdc.2008.05.004
8. Banks E, Reeves GK, Beral V et al (2009) Hip fracture incidence in relation to age, menopausal status, and age at menopause: prospective analysis. PLoS Med 6.https://doi.org/10.1371/journal.pmed.1000181
9. Terjesen T, Gunderson RB (2012) Radiographic evaluation of osteoarthritis of the hip: an inter-observer study of 61 hips treated for late-detected developmental hip dislocation. Acta Orthop 83:185–189. https://doi.org/10.3109/17453674.2012.665331
10. Gielis WP, Weinans H, Nap FJ et al (2021) Scoring osteoarthritis reliably in large joints and the spine using whole-body ct: osteoarthritis computed tomography-score (oact-score). J Pers Med 11:1–13. https://doi.org/10.3390/jpm11010005
11. Turmezei TD, Fotiadou A, Lomas DJ et al (2014) A new CT grading system for hip osteoarthritis. Osteoarthr Cartil 22:1360–1366. https://doi.org/10.1016/j.joca.2014.03.008

12. von Schacky CE, Sohn JH, Liu F et al (2020) Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. Radiology 295:139–145. https://doi.org/10.1148/radiol.2020190925
13. Üreten K, Arslan T, Gültekin KE et al (2020) Detection of hip osteoarthritis by using plain pelvic radiographs with deep learning methods. Skeletal Radiol 49:1369–1374. https://doi.org/10.1007/s00256-020-03433-9
14. Xue Y, Zhang R, Deng Y et al (2017) A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS ONE 12:1–9. https://doi.org/10.1371/journal.pone.0178992
15. Gebre RK, Hirvasniemi J, Lantto I et al (2019) Structural risk factors for low-energy acetabular fractures. Bone 127:334–342
16. Gebre RK, Hirvasniemi J, Lantto I et al (2020) Discrimination of low-energy acetabular fractures from controls using computed tomography-based bone characteristics. Ann Biomed Eng. https://doi.org/10.1007/s10439-020-02563-4
17. Wesseling J, Boers M, Viergever MA et al (2016) Cohort profile: Cohort Hip and Cohort Knee (CHECK) study. Int J Epidemiol 45:36–44. https://doi.org/10.1093/ije/dyu177
18. Napel S, Rubin GD, Jeffrey RB (1993) Sts-mip: A new reconstruction technique for ct of the chest. J Comput Assist Tomogr 17:832–838. https://doi.org/10.1097/00004728-199309000-00036
19. Lee KH, Hong H, Hahn S et al (2008) Summation or axial slab average intensity projection of abdominal thin-section CT datasets: can they substitute for the primary reconstruction from raw projection data? J Digit Imaging 21:422–432. https://doi.org/10.1007/s10278-007-9067-y
20. von Falck C, Galanski M, Shin HO (2010) Informatics in radiology: sliding-thin-slab averaging for improved depiction of low-contrast lesions with radiation dose savings at thin-section CT. Radiographics 30:317–326. https://doi.org/10.1148/rg.302096007
21. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031
22. Tiulpin A, Thevenot J, Rahtu E et al (2018) Automatic knee osteoarthritis diagnosis from plain radiographs: a deep learning-based approach. Sci Rep 8:1–10. https://doi.org/10.1038/s41598-018-20132-7
23. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-Decem:770–778. https://doi.org/10.1109/CVPR.2016.90
24. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6.https://doi.org/10.1186/s40537-019-0197-0
25. Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. arXiv
26. Xu J, Zhou C, Lang B, Liu Q (2017) Deep learning for histopathological image analysis: towards computerized diagnosis on cancers. In: Advances in Computer Vision and Pattern Recognition. pp 73–95
27. Shin HC, Lu L, Kim L et al (2016) Interleaved text/image deep mining on a large-scale radiology database for automated image interpretation. J Mach Learn Res 17:1–31
28. Gupta A, Ayhan MS, Maida AS (2013) Natural image bases to represent neuroimaging data. 30th Int Conf Mach Learn ICML 2013 28:2024–2031
29. Kingma DP, Ba JL (2015) Adam: A method for stochastic optimization. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc, pp 1–15
30. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Computer vision – ECCV 2014. Springer International Publishing, Cham, pp 818–833

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.