HHS Public Access

Author manuscript

J Magn Reson Imaging. Author manuscript; available in PMC 2020 December 01.

Published in final edited form as:

J Magn Reson Imaging. 2020 December ; 52(6): 1607–1619. doi:10.1002/jmri.27001.

Deep Learning for Lesion Detection, Progression, and Prediction of Musculoskeletal Disease

Richard Kijowski, MD[1,*], **Fang Liu, PhD**[1], **Francesco Caliva, PhD**[1], **Valentina Pedoia, PhD**[2]

[1]Department of Radiology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, USA;

[2]Department of Radiology, University of California at San Francisco School of Medicine, San Francisco, California, USA

Abstract

Deep learning is one of the most exciting new areas in medical imaging. This review article provides a summary of the current clinical applications of deep learning for lesion detection, progression, and prediction of musculoskeletal disease on radiographs, computed tomography (CT), magnetic resonance imaging (MRI), and nuclear medicine. Deep-learning methods have shown success for estimating pediatric bone age, detecting fractures, and assessing the severity of osteoarthritis on radiographs. In particular, the high diagnostic performance of deep-learning approaches for estimating pediatric bone age and detecting fractures suggests that the new technology may soon become available for use in clinical practice. Recent studies have also documented the feasibility of using deep-learning methods for identifying a wide variety of pathologic abnormalities on CT and MRI including internal derangement, metastatic disease, infection, fractures, and joint degeneration. However, the detection of musculoskeletal disease on CT and especially MRI is challenging, as it often requires analyzing complex abnormalities on multiple slices of image datasets with different tissue contrasts. Thus, additional technical development is needed to create deep-learning methods for reliable and repeatable interpretation of musculoskeletal CT and MRI examinations. Furthermore, the diagnostic performance of all deep-learning methods for detecting and characterizing musculoskeletal disease must be evaluated in prospective studies using large image datasets acquired at different institutions with different imaging parameters and different imaging hardware before they can be implemented in clinical practice.

MACHINE LEARNING, and more specifically deep learning, is one of the most widely used forms of artificial intelligence, which has revolutionized computer science, especially computer vision, and has quickly expanded into a much broader range of science and engineering disciplines, including medical imaging.[1] Recent surveys on deep learning in medical imaging have shown a wide variety of applications in different imaging subspecialties including neuro, lung, abdomen, cancer, breast, and cardiac imaging.[2,3] Applications in musculoskeletal imaging are emerging but remain relatively limited

[*]Address reprint requests to: R.K., Department of Radiology, University of Wisconsin School of Medicine and Public Health, Clinical Science Center-E3/311, 600 Highland Ave., Madison, WI 53792-3252. rkijowski@uwhealth.org.

compared to other fields. Deep learning has been used for various musculoskeletal imaging applications including tissue segmentation,[4–7] image reconstruction,[8–10] and disease detection. The use of deep learning for disease detection would be especially important for clinical radiologists, as it could potentially maximize diagnostic performance while reducing subjectivity and errors due to distraction and fatigue. This review article will provide a general overview of deep learning in medical imaging. The article will also provide a summary of the previously published peer-reviewed articles on the current clinical applications of deep learning for lesion detection, progression, and prediction of musculoskeletal disease on radiographs, computed tomography (CT), magnetic resonance imaging (MRI), and nuclear medicine.

Overview of Deep Learning in Medical Imaging

Artificial intelligence can refer to any machine that performs tasks that typically require human knowledge in a manner that resembles how humans solve problems. Machine learning is a branch of artificial intelligence that aims to automatically extract patterns or features out of data and to use such data trends in the form of experience. In this way, machine learning can make informed decisions with a measurable performance that improves over time. Conventional machine-learning techniques such as regression, clustering, shallow classifiers such as support vector machine, and random forest have been used in medical imaging for decades. While the classification abilities of these approaches can be remarkable, all methods rely on the effort invested in preprocessing and transforming data and designing feature extractors for use in pattern learning. However, such feature engineering typically requires domain knowledge and does not result in the extraction of the true essence of the data being analyzed. Furthermore, the results lack scalability, and the acquired knowledge is difficult to transfer to new domains.[11–13]

A requisite for the recent advancement of artificial intelligence was the design of systems smart enough to independently understand the surrounding world and to learn representations of the data that could support the extraction of meaningful information without a-priori features definition. This meaningful information is analogous to concepts, which are generally structured in hierarchies: complicated concepts can be interpreted as the fusion of many simple concepts at different levels of abstraction. This is the core idea of deep learning, which is based on the use of a specific algorithm architecture called a neural network. Neural networks have proven to be useful as they *obey the universal approximation theorem* and allow mapping input features to outputs by ways of matrix multiplication operations. In other words, the theorem states that when provided with enough neurons and nonlinear activation functions, neural networks can approximate any function that maps from any finite dimensional discrete space to another.[1]

The design of the neural network was initially inspired by the interconnected neurons in the brain and how they function to perform complex cognitive tasks. A neural network consists of input, hidden, and output layers. The input layer receives the data that is to be processed, while the output layer provides the results of the desired task. Multiple hidden layers are included in the neural network between the input and output layers and are interconnected with each other by weights similar to the real interconnected neurons in the brain. Thus,

deep learning usually refers to a deep structure with a large number of hidden layers. Unique features of this architecture are its flexibility and scalability, as it can range from a simple shallow structure with just a few layers to an ultrasophisticated deep structure with hundreds of layers that can perform complex functions well beyond the capabilities of conventional machine-learning techniques.[1]

The convolutional neural network (CNN) is the main neural network architecture used in computer vision and subsequently in medical imaging. A CNN consists of a series of convolutional layers interspersed with a series of pooling layers. Each convolution layer consists of one or more filters or kernals, which analyze features of regional voxels in the input image to generate feature maps. Units of the feature maps are connected to one another and to the units of feature maps in other convolutional layers by filter weights. An activation function, typically a rectified linear unit,[14] replaces all negative values in the feature map to zero, which allows only pertinent features to pass from one layer to the next. The pooling layer is then used to downsample the feature maps and merge sematically similar features by either taking the maximum value (max pooling) or average value (average pooling) of adjacent units in the feature map. This pooling process reduces the number of parameters and computations in the neural network and makes the network less sensitive to distortions and translations in the input image.

The CNN connects multiple convolution layers together in a deep network that can detect a hierarchy of features that are increasingly more abstract and complex. For example, the filters in the first few layers may be used to identify intensity variations, edges, and patterns in the input image at a rudimentary level. Next, new sets of filters at intermediate-level layers can then process the output from low-level features to construct composite information such as local shapes and patterns. This process repeats and continues until a high-level layer, where multiple global features are combined to represent a unique set of patterns in the image. The multiple convolution layers of the CNN are then connected to a fully-connected layer. This output layer of the CNN uses a softmax activation function to combine the features extracted from the different convolution layers in order to classify the input image into different categories.[1]

The CNN typically undergoes supervised training, in which a large number of paired images and reference classification labels are divided into separate training, validation, and testing groups. However, unsupervised or weakly-supervised training can also be applied for training the CNN for applications in which collecting the paired data and reference labels is challenging, expensive, or not possible. In supervised training, a back-propagation procedure is performed on the training group that is used to determine the most optimal filter weights. The CNN can be trained using random numbers as the initial weights, or it can be initialized through a process called transfer learning, in which the initial weights are obtained from a dataset other than the task-specific one. Since the networks could have millions of potential weights for some applications, transfer learning is able to reduce the demands for abundant training data and improve training efficiency.[15,16] The CNN uses the input images in the training group to determine the output classifications utilizing the initial model weights, which are then compared to the output classifications of the reference labels to quantify the error. The CNN then adjusts its internal weights in order to reduce the estimated error. Since

the training process is performed iteratively through all images in the training group, the CNN estimation on the validation group determines how many total iterations or epochs provide an optimal training performance. This estimation on data other than the training group ensures that the training can be generalized to other image datasets without the model being overly adapted to the training data (ie, overfitting). Finally, once the training is complete, the filter weights of the CNN are fixed, and the testing group is used to assess the diagnostic performance of the deep-learning model in the real-world scenario.

Clinical Applications of Deep Learning in Musculoskeletal Imaging

Detection of Internal Derangement of the Knee Joint on MRI

Deep-learning methods have been used to detect cartilage lesions within the knee joint on MRI. Liu et al[17] used an encoder-decoder VGG-16 CNN to segment cartilage on sagittal fat-suppressed T_2-weighted 2D fast spin-echo (FSE) images and extract cartilage image patches in ~100 regions of interest placed on the articular surface of the femur and tibia in 175 subjects. The extracted patches were then analyzed by a VGG-16 classification CNN to determine the presence or absence of cartilage lesions using the interpretation of an experienced radiologist as the reference standard. Two individual evaluations were performed using separate hold-out testing groups consisting of 1310 cartilage patches. The machine had an area under the curve (AUC) on receiver operating characteristic (ROC) analysis of 0.92 for detecting cartilage lesions with a sensitivity and specificity of 84% and 85%, respectively, for evaluation 1 and an AUC of 0.91 with a sensitivity and specificity of 81% and 88%, respectively, for evaluation 2. Furthermore, there was good intraobserver agreement between the two evaluations, with a kappa value of 0.76 (Fig. 1).[17] Pedoia et al[18] used a U-Net CNN to segment patellar cartilage on sagittal fat-suppressed proton density-weighted 3D FSE images in 1478 subjects. The segmented cartilage was then analyzed by a custom-made classification CNN to determine the presence or absence of cartilage lesions using the interpretation of an experienced radiologist as the reference standard. Using a hold-out testing group consisting of 222 patellar cartilage surfaces, the machine had an AUC of 0.88 for detecting cartilage lesions with sensitivity and specificity of 80%.[18]

Deep-learning methods have been used to detect meniscal tears on MRI. Roblot et al[19] used a fast-region CNN to first segment meniscus on sagittal fat-suppressed T_2-weighted 2D FSE images in 1123 subjects and then determine the presence or absence of meniscal tears using the interpretation of experienced radiologists as the reference standard. Using a hold-out testing group consisting of meniscus segments in 700 subjects, the machine had an AUC of 0.94 for detecting meniscal tears.[19] Couteaux et al[20] used a similar deep-learning approach to detect meniscal tears on fat-suppressed T_2-weighted 2D FSE images as part of the French Radiology Society Challenge. The Challenge included a training group consisting of images from 1128 subjects and a hold-out testing group consisting of images from 700 subjects with the interpretation of experienced radiologists as the reference standard. A fast-region CNN was used to segment meniscus and then determine the presence or absence of meniscal tears. The neural network was coupled to a shallow ConvNet CNN to classify the orientation of the tear as vertical or horizontal. The machine had an AUC of 0.91 for detecting and classifying meniscal tears.[20] Pedoia et al[18] used a U-Net CNN to segment meniscus on

sagittal fat-suppressed proton density-weighted 3D FSE images in 1478 subjects. The segmented meniscus was then analyzed by a custom-made classification CNN to determine the presence or absence of meniscal tears using the interpretation of an experienced radiologist as the reference standard. Using hold-out testing groups consisting of 887 meniscus segments, the machine had an AUC of 0.89 for detecting meniscal tears with sensitivity and specificity of 90% and 82%, respectively (Fig. 2).[18]

Studies have used deep-learning methods to detect anterior cruciate ligament (ACL) tears on MRI. Liu et al[21] used coupled CNNs consisting of LeNet-5 and YOLA to isolate the ACL on sagittal proton density-weighted and fat-suppressed T_2-weighted 2D FSE images in 175 subjects with ACL tears and 175 subjects without ACL tears. The isolated ACLs were then analyzed by a DenseNet classification CNN to determine the presence or absence of ACL tears using arthroscopy as the reference standard. Using a hold-out testing group consisting of 100 isolated ACLs, the machine had an AUC of 0.98 for detecting ACL tears with sensitivity and specificity of 96% and 98%, respectively (Figs. 3 and 4).[21] Chang et al[22] used a U-Net CNN to isolate the ACL on coronal proton density-weighted 2D FSE images in 130 subjects with ACL tears and 130 subjects without ACL tears. The isolated ACLs were then analyzed by a ResNet classification CNN to determine the presence or absence of ACL tears using the interpretation of an experienced radiologist as the reference standard. Using a hold-out testing group consisting of 60 isolated ACLs, the machine had an AUC of 0.97 for detecting ACL tears with sensitivity and specificity of 100% and 93%, respectively.[22]

The previous studies used coupled deep-learning pipelines to detect internal derangement of the knee joint on MRI, with the first CNN isolating the structures of interest and the second classification CNN determining the presence or absence of pathology within the isolated joint structures (Fig. 5). In all studies, the machine had similar diagnostic performance as human readers for evaluating the same hold-out testing groups (Fig. 6).[17,18,21,22] In an alternative approach, Bien et al[23] used a single custom-made "MRNet" classification CNN to analyze nonsegmented axial fat-suppressed proton density-weighted 2D FSE, coronal T_1-weighted 2D FSE, and sagittal fat-suppressed T_2-weighted 2D FSE images in 1370 subjects with 319 ACL tears and 508 meniscal tears to determine the presence or absence of ACL and meniscal tears using the interpretation of experienced radiologists as the reference standard (Fig. 7). Using a hold-out testing group consisting of 120 MRI examinations, the machine had a sensitivity and specificity of 76% and 97%, respectively, for detecting ACL tears and a sensitivity and specificity of 71% and 74%, respectively, for detecting meniscal tears. The diagnostic performance of the machine was significantly lower ($P < 0.05$) than the diagnostic performance of human readers evaluating the same hold-out testing group.[23] The results suggest that isolation of individual joint structures may be a burdensome and time-consuming first step in the deep-learning pipeline, but may be necessary to maximize diagnostic performance for detecting internal derangement of the knee joint on MRI.

Detection of Osseous Metastatic Disease on CT, MRI, and Nuclear Medicine

Deep-learning methods have been used to localize vertebral body metastases on CT and MRI of the spine using metastatic lesions outlined by experienced radiologists as the reference standard. Roth et al[24] used a custom-designed computer-assisted detection (CAD)

method combining attenuation thresholding, region growing, and watershed algorithms to isolate individual vertebrae segments on axial CT images in 49 subjects with 539 sclerotic lesions and five control subjects without metastatic disease. The isolated vertebrae segments were then analyzed by a custom-designed "DropConnect" classification CNN to determine the presence or absence of metastatic disease. Using 5-fold crossvalidation, the machine had an AUC of 0.83 for localizing large metastatic lesions more than 3 cm in diameter, with 9.5 false positives per subject at a sensitivity of 90%.[24] Chmelik et al[25] used a custom-designed CAD method to segment the entire spine on sagittal CT images of 31 subjects with 1046 lytic lesions and 1135 sclerotic lesions. Orthogonal plane voxel-based masks were created from the isolated spine, which were then analyzed by a custom-designed, voxel-based classification CNN to determine the presence or absence of metastatic disease within each tissue voxel. Using a hold-out testing group consisting of 24,000 segmented tissue voxels, the machine had an AUC of 0.80 for detecting lytic meta-static lesions and an AUC of 0.72 for detecting sclerotic metastatic lesions with the voxels. For object-wise evaluation, the machine had 45.6 false positives per subject at 92% sensitivity for localizing small metastatic lesions less than 1.5 mm in diameter and 5.9 false positives per subject at 99% sensitivity for localizing large metastatic lesions more than 3 cm in diameter.[25] Wang et al[26] used a custom-designed classification CNN to analyze nonsegmented sagittal fat-suppressed T_2-weighted 2D FSE images of the spine in 26 subjects with metastatic disease. The images were preprocessed through multiresolution transformation prior to analysis to create image datasets with different resolutions tailored to detect metastatic disease in different regions of the spine. Using 4-fold crossvalidation, the machine had 0.4 false positives per subject at 90% sensitivity for localizing metastatic lesions (Fig. 8). However, the number and sizes of the metastatic lesions, which would influence diagnostic performance, was not specified in the study.[26]

Deep-learning methods have also been used to localize whole-body osseous metastatic disease on combined positive emission tomography and computer tomography (PET-CT). Xu et al[27] used two enhanced V-Net CNNs that were cascaded to build a W-shaped framework to learn the volumetric feature representation of the skeleton and to differentiate between normal bone and bone metastases using both the PET and CT images. The training data consisted of 2000 image patches with and 2000 image patches without meta-static lesions outlined by an experienced radiologist as the reference standard from PET-CT images of 12 subjects with multiple myeloma. Using 3-fold crossvalidation, the machine had 73% sensitivity and 99% specificity for detecting meta-static lesions throughout the axial and appendicular skeleton. Furthermore, the diagnostic performance of the deep-learning approach was superior to multiple traditional machine learning methods including random forest classifier, k-nearest neighbor, and support vector machine.[27]

Detection of Spine Degenerative Disk Disease and Infection on MRI

Deep-learning methods have been used to detect degenerative disc disease and infection of the spine on MRI. Jamaludin et al[28,29] used a custom-designed CAD method to isolate vertebral body and disc segments on sagittal T_2-weighted 2D FSE images of the lumbar spine in 2009 subjects. The isolated vertebral body and disc segments were then analyzed by a VGG-M classification CNN to determine the presence or absence of various findings of

degenerative disc disease, including Pfirrmann grade, disc narrowing grade, central canal stenosis, vertebral body endplate defects and marrow changes, and spondylolisthesis using the interpretation of experienced radiologists as the reference standard (Fig. 9). Using a hold-out testing group consisting of 1200 isolated vertebral body and disc segments, the interrater agreement between the machine and the human reader ranged between 70% for assigning a Pfirrmann grade and 93% for determining the presence or absence of vertebral body marrow changes. Furthermore, the interreader agreement between the machine and the human reader was similar to the intrareader agreement between the same human reader evaluating the same vertebral body and disc segments at different timepoints (Table 1).[28,29] Kim et al[30] used a custom-designed classification CNN to analyze nonsegmented axial T_2-weighted 2D FSE images in 80 subjects with tuberculous spondylitis and 80 subjects with pyogenic spondylitis using biopsy as the reference standard. For network training, experienced radiologists placed regions of interest around the pathologic abnormalities on the images. Using 4-fold crossvalidation, the machine had an AUC of 0.80 for distinguishing between tuberculous and pyogenic spondylitis with a sensitivity and specificity of 85% and 68%, respectively. The machine had similar diagnostic performance as human readers evaluating the same hold-out testing groups.[30]

Estimation of Pediatric Bone Age on Radiographs

Multiple deep-learning methods have been described for estimating pediatric bone age using anterior–posterior hand radio-graphs, with some algorithms already approved by the United States Food and Drug Administration (US FDA) for use in clinical practice. Larson et al[31] published one of the earliest and largest studies, which used 14,036 hand radiographs to train a custom-designed classification CNN to estimate pediatric bone age with the interpretation of experienced radiologists as the reference standard. Using a hold-out testing group consisting of 200 hand radiographs, the root mean square and mean absolute difference between the bone age estimates provided by the machine and the bone age estimates provided by the human readers were 0.63 years and 0.50 years, respectively.[31] A recent study performed by Halabi et al[32] described the results of the Radiological Society of North America Machine Learning Challenge for estimating pediatric bone age using a training group consisting of 14,036 hand radiographs and a hold-out testing group consisting of 200 hand radio-graphs, with the interpretation of experienced radiologists as the reference standard. The Challenge had 109 participants with a 4.2-month mean absolute difference between the bone age estimates provided by the top-ranked machine and the bone age estimates provided by the human readers. The top-ranked machine used an Inceprion V3 classification CNN to analyze pixel information on the hand radiographs, which was then concatenated with gender information in a joint training model.[32]

Detection of Fractures on Radiographs and CT

Multiple deep-learning methods have been used to detect fractures on radiographs. Most studies have used open-source CNNs and large training datasets for detecting fractures in multiple body parts including the hip,[33–36] shoulder,[36,37] wrist,[36,38–40] and ankle[36,41] using the interpretation of experienced radiologists as the reference standard. Diagnostic performance varied but was generally high for all studies, with AUCs ranging between 0.90 and 0.99, sensitivities ranging between 73% and 99%, specificities ranging between 73%

and 97%, and accuracies ranging between 75% and 96% (Table 2). One study deserves particular mention. Lindsey et al[36] used a modified U-Net classification CNN to detect fractures in 11 body parts utilizing a training group consisting of 135,845 radiographs. Using a hold-out testing group consisting of 300 randomly chosen radiographs, the machine had an AUC of 0.99 for detecting fractures with a sensitivity and specificity of 94% and 95%, respectively (Fig. 10). The same hold-out testing group was used to evaluate the diagnostic performance of emergency medicine physicians for detecting fractures with and without use of the machine. The clinicians experienced an average 47% reduction in the misinterpretation rate when using the machine to aid in interpretation of the radiographs.[36]

Multiple deep-learning methods have been used to detect spine fractures on CT using the interpretation of experienced radiologists as the reference standard. Raghavendra et al[42] used a custom-designed classification CNN to analyze nonsegmented sagittal CT scans of the entire spine in 100 subjects with fracture and 60 subjects without fracture of the thoracolumbar vertebral bodies. Using a hold-out testing group consisting of 210 CT images with fracture and 126 CT images without fracture, the machine had a sensitivity and specificity of 100% and 98%, respectively, for detecting vertebral body fractures.[42] Tomita et al[43] used coupled neural networks to analyze nonsegmented sagittal CT scans of the entire spine in 713 subjects with fracture and 719 subjects without fracture of the thoracolumbar vertebral bodies. A ResNet classification CNN was first used for feature extraction followed by a recurrent neural network module to aggregate the extracted features and make the final diagnosis. Using a hold-out testing group consisting of 129 CT scans, the machine had an AUC of 0.91 for detecting vertebral body fractures with a sensitivity and specificity of 85% and 96%, respectively.[43] However, both studies only classified the presence or absence of a fracture for the entire spine, which raises questions regarding the ability of the deep-learning methods to localize the exact site of injury. Roth et al[44] used a custom-designed CAD method combining multiatlas label fusion and edge mapping algorithms to isolate individual vertebrae segments on axial CT scans in 18 subjects with 55 displaced posterior element fractures and five control subjects without fracture. The isolated vertebrae segments were then analyzed by a custom-designed classification CNN to determine the presence or absence of fracture. Using a hold-out testing group consisting of isolated vertebrae segments from six patients with fractures, the machine had an AUC of 0.86 for detecting posterior element fracture with a sensitivity of 71% or 81% at 5 or 10 false positives per patient, respectively.[44]

A deep-learning method has also been described for detecting and characterizing calcaneal fractures on CT. Pranata et al[45] used a combined CNN and CAD approach to analyze 683 nonsegmented CT images with fracture and 1248 nonsegmented CT images without fracture from an unspecified number of subjects with calcaneus fractures. A ResNet classification CNN was first used to classify the axial, coronal, and sagittal images in the CT scans into fracture and nonfactors categories using the interpretation of an experienced radiologist as the reference standard. A CAD method consisting of speedup robust features, canny edge detection, and contour tracing algorithms were then used to detect the exact location of the fractures on the CT images. Using a hold-out testing group consisting of 136 CT images with fracture and 250 CT images without fracture, the machine had 98% accuracy for localizing the calcaneal fractures.[45]

Detection and Characterization of Osteoarthritis on Radiographs and MRI

The Kellgren–Lawrence (KL) system is widely used to assess the severity of osteoarthritis (OA) on knee and hip radio-graphs in clinical practice and research studies.[46] Two recent studies have used deep-learning methods to automatically assign a KL grade on anterior–posterior knee radiographs using large image datasets from the Osteoarthritis Initiative (OAI) and Multi-Center Osteoarthritis Study (MOST) and the interpretation of experienced radiologists as the reference standard. Tiulpin et al[47] made use of knee joint symmetry by adopting a Siamese classification CNN for assessing the severity of knee OA. The features learned from the medial and lateral sides of the knee were shared between two paths in the network and were eventually concatenated using a final fully connected layer to assign a KL grade. The model was trained using images from MOST and evaluated on a hold-out testing group consisting of 5960 knees radiographs from the OAI. The machine had an average multiclass accuracy of 67% and quadratic kappa coefficient of 0.83 for assigning a KL grade.[47] Conversely, Norman et al[48] used a modified DenseNet CNN to assign a KL grade on knee radiographs. The DenseNet architecture was modified to allow inclusion of demographic and clinical factors into the prediction model including age, gender, body mass index (BMI), and pain and disability scores. The variables were fed into the model as a four-dimensional vector, which was transformed into a 32-dimensional vector by a fully connected layer and then concatenated onto the flattened image output by DenseNet. Using a hold-out testing group consisting of 621 knee radio-graphs from the OAI, the machine had a quadratic kappa coefficient of 0.83 for assigning a KL grade. The sensitivity for detecting no OA (KL grades 0 and 1), mild OA (KL grade 2), moderate OA (KL grade 3), and severe OA (KL grade 4) was 84%, 70%, 69%, and 86%, respectively with corresponding specificity of 86%, 84%, 97%, and 99%, respectively (Fig. 11).[48] The kappa coefficient of 0.83 was identical to the kappa coefficient reported by Tiulpin et al[47] and similar to the kappa coefficients reported for inter-observer and intraobserver agreement for assigning a KL grade to knee radiographs by human readers.[49] In a smaller study, Xue et al[50] used a VGG-16 classification CNN to assign a KL grade to assess the severity of hip OA on anterior-posterior pelvic radiographs using the interpretation of an experienced radiologist as the reference standard. Using a hold-out testing group consisting of 83 pelvic radiographs, the machine had an AUC of 0.94 for determining the presence (KL grades 0 and 1) or absence (KL grades 2, 3, and 4) of radiographic OA with a sensitivity and specificity of 95% and 91%, respectively.[50]

A recent study by Pedoia et al[51] used a deep-learning method combined with voxel-based relaxometry for the analysis of T_2 relaxation time maps to determine the presence or absence of radiographic knee OA using the interpretation of experienced radiologists as the reference standard. A shallow random forest classifier model trained on handcrafted features consisting of average cartilage T_2 values on different articular surfaces of the knee joint was compared to a DenseNet classification CNN model trained on the raw T_2 data. Using a hold-out testing dataset consisting of 658 MRI examinations, the random forest classifier model trained with features extracted with simple linear pattern decomposition had an AUC of 0.77 for detecting radiographic OA, while the DenseNet model had a significantly higher ($P < 0.05$) AUC of 0.83. The study highlighted the ability of a deep-learning approach to exploit

uncovered information in T_2 relaxation time maps drastically underused by machine-learning analysis due to oversimplified feature extraction.[51]

Miscellaneous Applications on Radiographs

Chee et al[52] used a modified ResNet classification CNN to analyze anterior–posterior hip radiographs in 1892 subjects with femoral head osteonecrosis and 855 subjects without femoral head osteonecrosis using the interpretations of MRI examinations by experienced radiologists as the reference standard. Using a hold-out testing group consisting of 250 hip radiographs, the machine had a sensitivity and specificity of 85% and 91%, respectively, for detecting femoral head osteonecrosis, which was similar to the diagnostic performance of human readers.[52] England et al[53] used a DenseNet classification CNN to analyze lateral elbow radiographs in 882 children with a history of trauma to detect an elbow joint effusion using the interpretation of experienced radiologists as the reference standard. Using a hold-out testing group consisting of 96 elbow radiographs with effusion and 33 elbow radiographs without effusion, the machine had an AUC of 0.94 for detecting an elbow joint effusion with a sensitivity and specificity of 91%.[53]

Conclusion and Future Directions

There are many emerging applications of deep learning in musculoskeletal imaging. The most promising applications are in the use of deep-learning approaches for the interpretation of radiographs. Multiple deep-learning algorithms have already been approved by the US FDA for estimating pediatric bone age on hand radiographs. Deep-learning methods have also shown tremendous success for detecting fractures, assessing the severity of knee and hip OA, identifying femoral head osteonecrosis, and detecting elbow joint effusion on radiographs. However, no deep-learning methods have been described for detecting neoplastic, inflammatory, infectious, or metabolic processes in bone or for identifying soft-tissue abnormalities. Furthermore, all recently described deep-learning methods have been designed to perform a single task. To be useful in clinical practice, multiple deep-learning algorithms will need to be combined in a single pipeline to evaluate every possible abnormality in bone and soft tissue on radiographs, similar to the interpretation of the imaging studies by clinical radiologists.

Deep-learning methods have also shown promising results for detecting fractures on CT and osseous metastatic disease on CT and nuclear medicine. Deep-learning approaches for detecting pathology using these imaging modalities is more challenging than using radiographs, as it typically requires evaluating a large number of image slices. For this reason, the preliminary results have been less encouraging, with a higher number of false positive and false negative interpretations. Furthermore, there has been little work performed to determine the ability of deep-learning approaches to characterize abnormalities detected on CT or nuclear medicine. For example, a deep-learning method may be able to identify an osseous lesion within the spine on CT, but can it perform the next step in the evaluation process and determine whether the lesion represents metastatic disease rather than a benign hemangioma or bone island? This classification process requires a great deal of experience that may be beyond the capabilities of even the deepest neural network.

The use of deep-learning approaches for detecting musculoskeletal disease on MRI is especially challenging, as it often requires analyzing complex abnormalities on multiple slices of image datasets acquired in different planes with different tissue contrasts. Furthermore, the variability in image quality on MRI is much greater than on radiographs, CT, and nuclear medicine due to the use of a wide variety of scanners, pulse sequences, and imaging parameters in clinical practice. Nevertheless, there has been promising preliminary results on the use of deep-learning methods for detecting internal derangement and OA of the knee and degenerative disk disease, infection, and metastatic disease of the spine on MRI. However, current deep-learning approaches have limitations that need to be addressed. For example, most of the deep-learning methods used to detect internal derangement of the knee require isolating the structures of interest before determining the presence or absence of pathology, which is a burdensome and time-consuming first step that may be difficult to perform due to the variable image quality and tissue contrast on MRI. Furthermore, no deep-learning methods have been described for detecting internal derangement in more difficult to image joints such as the wrist, hip, ankle, or elbow. In addition, deep-learning approaches have shown promising preliminary results for determining the presence or absence of central canal stenosis of the spine on MRI. However, can even the best neural network trained using large image datasets be able to determine with high accuracy and repeatability the exact degree of central canal, lateral recess, and neural foraminal stenosis at each level of the spine? These are questions that need to be addressed in future studies.

In conclusion, there have been many recent advances in the use of deep-learning methods in musculoskeletal imaging. The excellent preliminary results of deep-learning approaches for the interpretation of radiographs suggests that these methods may soon be available for use in clinical practice. However, much additional work is needed to develop more accurate and efficient deep-learning techniques to detect and characterize musculoskeletal disease on CT, MRI, and nuclear medicine. This may be aided by future technological advances, including the development of deeper neural networks and more sophisticated graphic processing units that would allow the use of extremely large image datasets for machine training. Regardless of the imaging modality used, the diagnostic performance and repeatability of deep-learning methods for detecting and characterizing musculoskeletal disease must be evaluated in prospective studies using large testing datasets acquired at different institutions with different imaging parameters and different imaging hardware before they can be implemented in clinical practice. Furthermore, improvements in diagnostic performance and efficiency of clinical radiologists when using the deep-learning technology needs to be assessed. The great deal of additional work needed for the translation of current deep-learning methods into clinical practice clearly indicates that the job security of clinical radiologists is not in jeopardy, at least for the time being.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521: 436–444. [PubMed: 26017442]
2. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60–88. [PubMed: 28778026]
3. Shen D, Wu G, Suk H-I. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221–248. [PubMed: 28301734]

4. Liu F, Zhou Z, Jang H, Samsonov A, Zhao G, Kijowski R. Deep convolutional neural network and 3D deformable approach for tissue segmentation in musculoskeletal magnetic resonance imaging. Magn Reson Med 2018;79:2379–2391. [PubMed: 28733975]
5. Zhou Z, Zhao G, Kijowski R, Liu F. Deep convolutional neural network for segmentation of knee joint anatomy. Magn Reson Med 2018;80: 2759–2770. [PubMed: 29774599]
6. Norman B, Pedoia V, Majumdar S. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology 2018;172322.
7. Liu F SUSAN: Segment unannotated image structure using adversarial network. Magn Reson Med 2019;81:3330–3345. [PubMed: 30536427]
8. Liu F, Samsonov A, Chen L, Kijowski R, Feng L. SANTIS: Sampling-Augmented Neural neTwork with Incoherent Structure for MR image reconstruction. Magn Reson Med 2019;82:1890–1904. [PubMed: 31166049]
9. Hammernik K, Klatzer T, Kobler E, et al. Learning a variational network for reconstruction of accelerated MRI data. Magn Reson Med 2017;79: 3055–3071. [PubMed: 29115689]
10. Liu F, Feng L, Kijowski R. MANTIS: Model-Augmented Neural neTwork with Incoherent k-space Sampling for efficient MR parameter mapping. Magn Reson Med 2019;82:174–188. [PubMed: 30860285]
11. Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC. Machine learning in medical imaging. IEEE Signal Process Mag 2010;27:25–38. [PubMed: 25382956]
12. Giger ML. Machine learning in medical imaging. J Am Coll Radiol 2018;15:512–520. [PubMed: 29398494]
13. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. RadioGraphics 2017;37:505–515. [PubMed: 28212054]
14. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: Proc 27th Int Conf Mach Learn 2010;807–814.
15. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans Med Imaging 2016;35:1299–1312. [PubMed: 26978662]
16. Shin H-C, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. ArXiv e-prints 2016;1602:1285–1298.
17. Liu F, Zhou Z, Samsonov A, et al. Deep learning approach for evaluating knee MR images: Achieving high diagnostic performance for cartilage lesion detection. Radiology 2018;172986.
18. Pedoia V, Norman B, Mehany SN, Bucknor MD, Link TM, Majumdar S. 3D convolutional neural networks for detection and severity staging of meniscus and PFJ cartilage morphological degenerative changes in osteoarthritis and anterior cruciate ligament subjects. J Magn Reson Imaging 2019;49:400–410. [PubMed: 30306701]
19. Roblot V, Giret Y, Bou Antoun M, et al. Artificial intelligence to diagnose meniscus tears on MRI. Diagn Interv Imaging 2019;100:243–249. [PubMed: 30928472]
20. Couteaux V, Si-Mohamed S, Nempont O, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging 2019;100:235–242. [PubMed: 30910620]
21. Liu F, Guan B, Zhou Z, et al. Fully-automated diagnosis of anterior cruciate ligament tears on knee MR images using deep learning. Radiol Artif Intell 2019 [Epub ahead of print].
22. Chang PD, Wong TT, Rasiej MJ. Deep learning for detection of complete anterior cruciate ligament tear. J Digit Imaging 2019 [Epub ahead of print].
23. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018;15:e1002699. [PubMed: 30481176]
24. Roth HR, Yao J, Lu L, Stieger J, Burns JE, Summers RM. Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications. Cham, Switzerland: Springer; 2015:3–12.

25. Chmelik J, Jakubicek R, Walek P, et al. Deep convolutional neural network-based segmentation and classification of difficult to define metastatic spinal lesions in 3D CT data. Med Image Anal 2018;49:76–88. [PubMed: 30114549]
26. Wang J, Fang Z, Lang N, Yuan H, Su M-Y, Baldi P. A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks. Comput Biol Med 2017;84:137–146. [PubMed: 28364643]
27. Xu L, Tetteh G, Lipkova J, et al. Automated whole-body bone lesion detection for multiple myeloma on 68Ga-pentixafor PET/CT imaging using deep learning methods. Contrast Media Mol Imaging 2018;2018: 2391925. [PubMed: 29531504]
28. Jamaludin A, Kadir T, Zisserman A. SpineNet: Automated classification and evidence visualization in spinal MRIs. Med Image Anal 2017;41:63–73. [PubMed: 28756059]
29. Jamaludin A, Lootus M, Kadir T, et al. ISSLS prize in bioengineering science 2017: Automation of reading of radiological features from magnetic resonance images (MRIs) of the lumbar spine without human intervention is comparable with an expert radiologist. Eur Spine J 2017;26:1374–1383. [PubMed: 28168339]
30. Kim K, Kim S, Lee YH, Lee SH, Lee HS, Kim S. Performance of the deep convolutional neural network based magnetic resonance image scoring algorithm for differentiating between tuberculous and pyogenic spondylitis. Sci Rep 2018;8:13124. [PubMed: 30177857]
31. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018;287:313–322. [PubMed: 29095675]
32. Halabi SS, Prevedello LM, Kalpathy-Cramer J, et al. The RSNA pediatric bone age machine learning challenge. Radiology 2019;290:498–503. [PubMed: 30480490]
33. Adams M, Chen W, Holcdorf D, McCusker MW, Howe PD, Gaillard F. Computer vs human: Deep learning versus perceptual training for the detection of neck of femur fractures. J Med Imaging Radiat Oncol 2019;63:27–32. [PubMed: 30407743]
34. Cheng C-T, Ho T-Y, Lee T-Y, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol 2019;29:5469–5477. [PubMed: 30937588]
35. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. Skeletal Radiol 2019; 48:239–244. [PubMed: 29955910]
36. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A 2018;115: 11591–11596. [PubMed: 30348771]
37. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018;89:468–473. [PubMed: 29577791]
38. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: Transfer learning from deep convolutional neural networks. Clin Radiol 2018;73:439–445. [PubMed: 29269036]
39. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop 2017;88:581–586. [PubMed: 28681679]
40. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 2019;1:e180001.
41. Kitamura G, Chung CY, Moore BE. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging 2019;32:672–677. [PubMed: 31001713]
42. Raghavendra U, Bhat NS, Gudigar A, Acharya UR. Automated system for the detection of thoracolumbar fractures using a CNN architecture. Futur Gener Comput Syst 2018;85:184–189.
43. Tomita N, Cheung YY, Hassanpour S. Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans. Comput Biol Med 2018;98:8–15. [PubMed: 29758455]
44. Roth HR, Wang Y, Yao J, Lu L, Burns JE, Summers RM. Deep convolutional networks for automated detection of posterior-element fractures on spine CT. In: Med Imaging 2016 Comput Diagnosis. Vol. 9785 Tourassi GD, Armato SG, eds. 2016;97850P http://

proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2217146. Accessed December 10, 2018.

45. Pranata YD, Wang K-C, Wang J-C, et al. Deep learning and SURF for automated classification and detection of calcaneus fractures in CT images. Comput Methods Programs Biomed 2019;171:27–37. [PubMed: 30902248]
46. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16:494–502. [PubMed: 13498604]
47. Tiulpin A, Thevenot J, Rahtu E, Lehenkari P, Saarakkala S. Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Sci Rep 2018;8:1727. [PubMed: 29379060]
48. Norman B, Pedoia V, Noworolski A, Link TM, Majumdar S. Applying densely connected convolutional neural networks for staging osteoarthritis severity from plain radiographs. J Digit Imaging 2018 [Epub ahead of print].
49. Klara K, Collins JE, Gurary E, et al. Reliability and accuracy of cross-sectional radiographic assessment of severe knee osteoarthritis: Role of training and experience. J Rheumatol 2016;43:1421–1426. [PubMed: 27084912]
50. Xue Y, Zhang R, Deng Y, Chen K, Jiang T. A preliminary examination of the diagnostic value of deep learning in hip osteoarthritis. PLoS One 2017;12:e0178992. [PubMed: 28575070]
51. Pedoia V, Lee J, Norman B, Link TM, Majumdar S. Diagnosing osteoarthritis from T2 maps using deep learning: An analysis of the entire Osteoarthritis Initiative baseline cohort. Osteoarthr Cartil 2019;27: 1002–1010.
52. Chee CG, Kim Y, Kang Y, et al. Performance of a deep learning algorithm in detecting osteonecrosis of the femoral head on digital radiography: A comparison with assessments by radiologists. Am J Roentgenol 2019;213:155–162.
53. England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM. Detection of traumatic pediatric elbow joint effusion using a deep convolutional neural network. Am J Roentgenol 2018;211:1361–1368. [PubMed: 30300006]

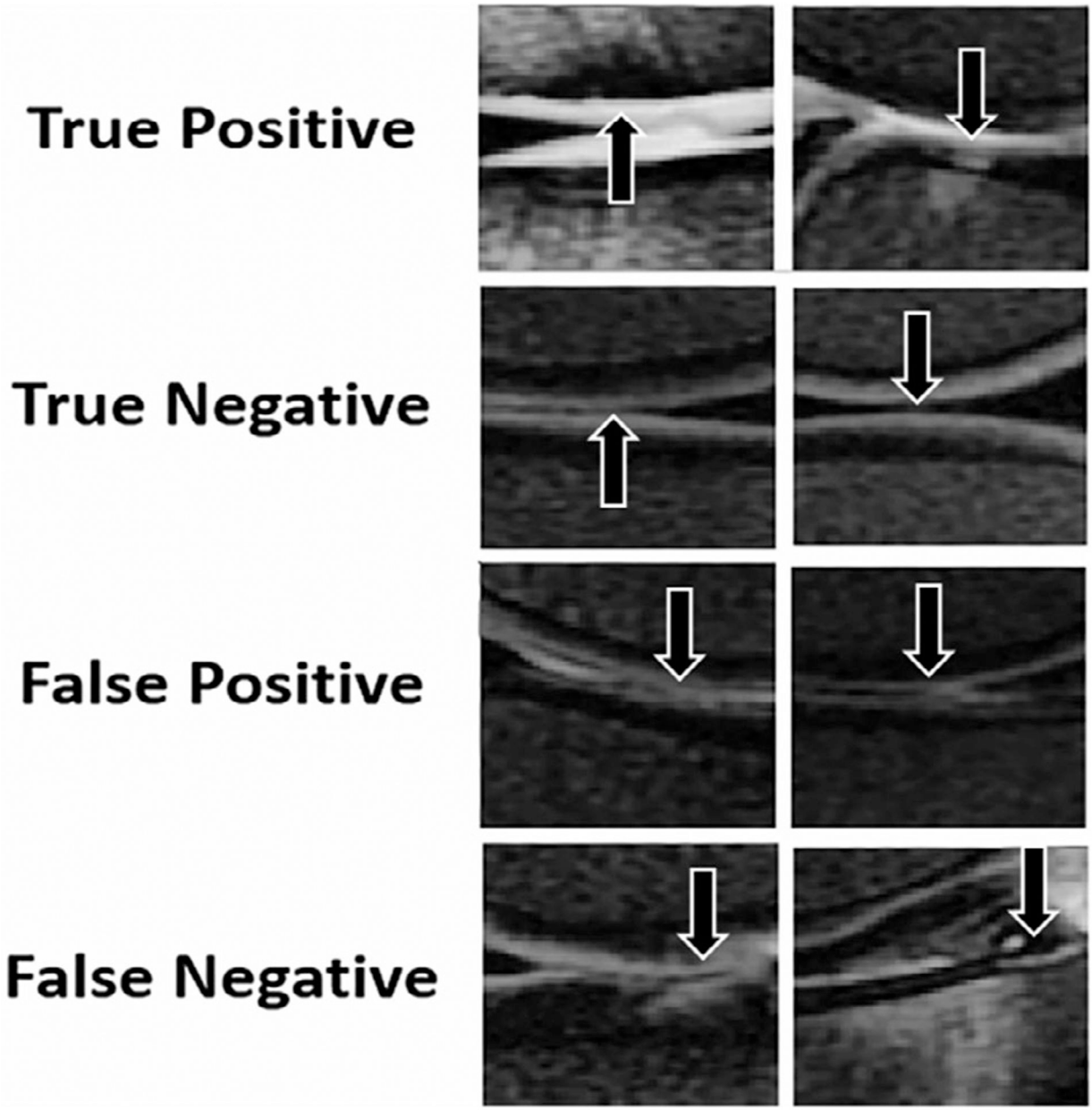

FIGURE 1:
Examples of true-positive, true-negative, false-positive, and false-negative interpretations of a deep-learning method for detecting cartilage lesions within the knee joint on MRI (arrows) using sagittal fat-suppressed T_2-weighted 2D FSE images and the interpretation of an experienced radiologist as the reference standard. Figure obtained from a study performed by Liu et al.[17]

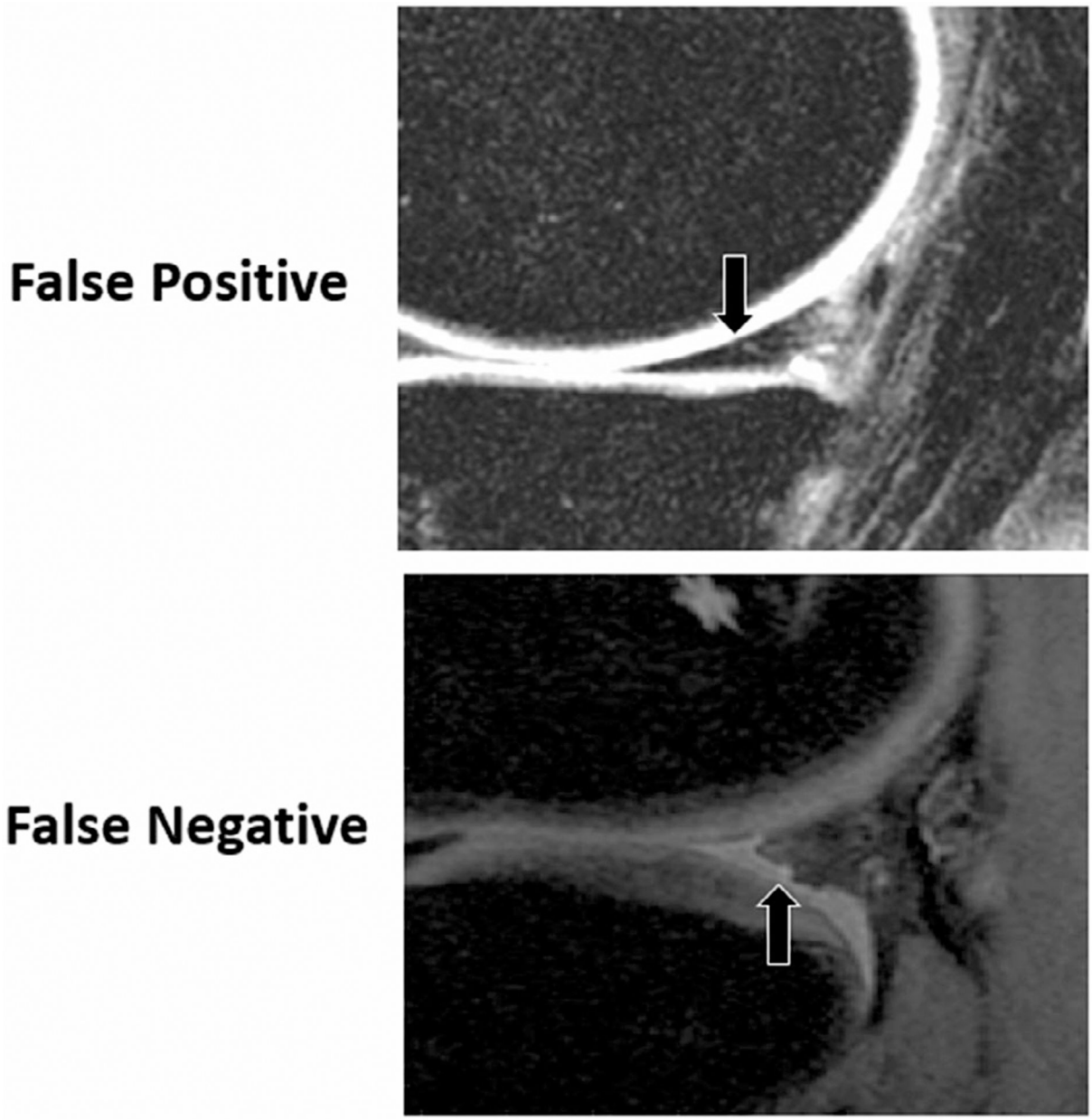

FIGURE 2:
Examples of false-positive and false-negative interpretations of a deep-learning method for detecting meniscal tears within the knee joint on MRI (arrows) using sagittal fat-suppressed proton density-weighted 3D FSE images and the interpretation of an experienced radiologist as the reference standard. Figure obtained from a study performed by Pedoia et al.[18]

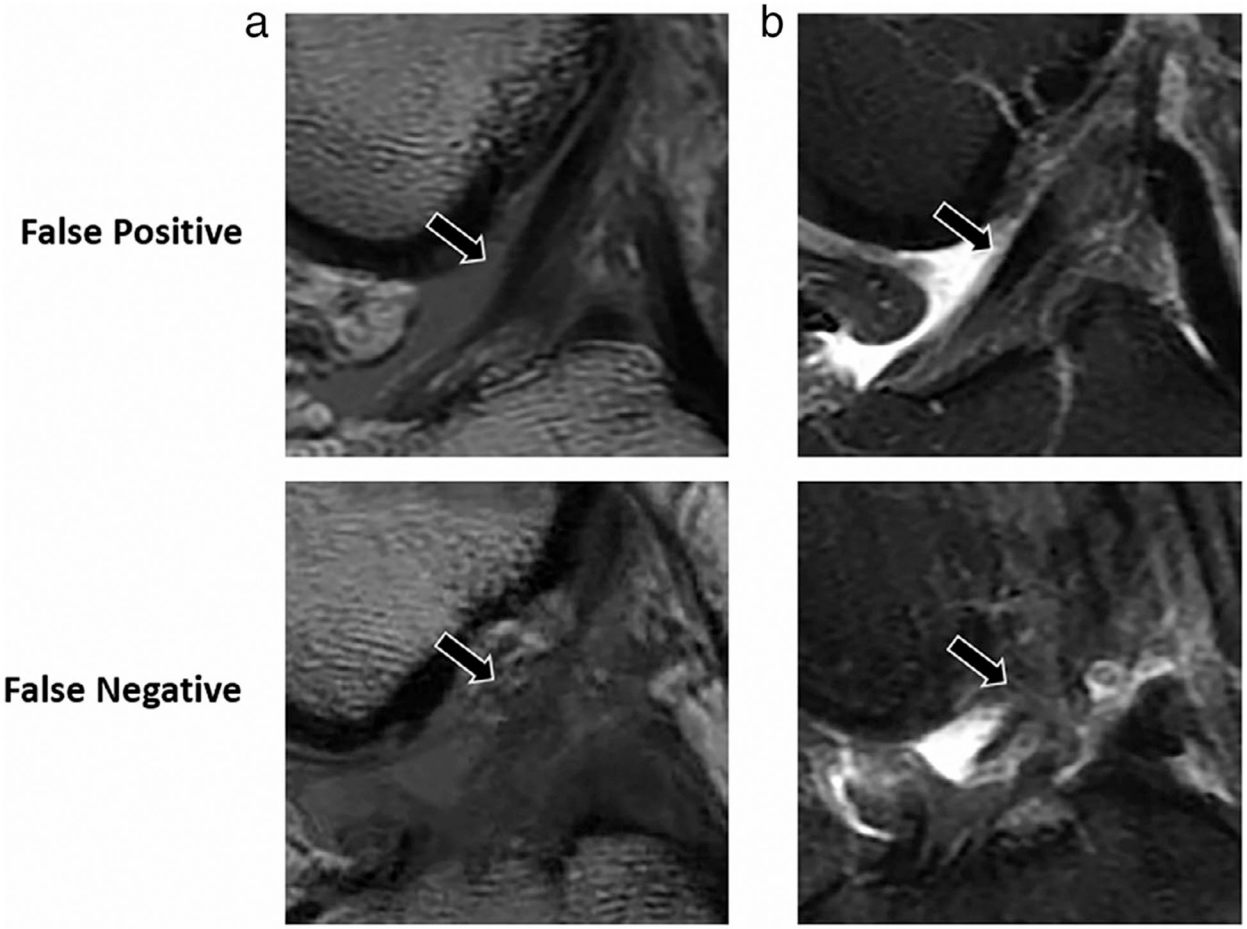

FIGURE 3:
Examples of false-positive and false-negative interpretations of a deep-learning method for detecting meniscal tears within the knee joint on MRI (arrows) using sagittal **(a)** proton density-weighted and **(b)** fat-suppressed T_2-weighted 2D FSE images and arthroscopy as the reference standard. Figure obtained from a study performed by Liu et al.[21]

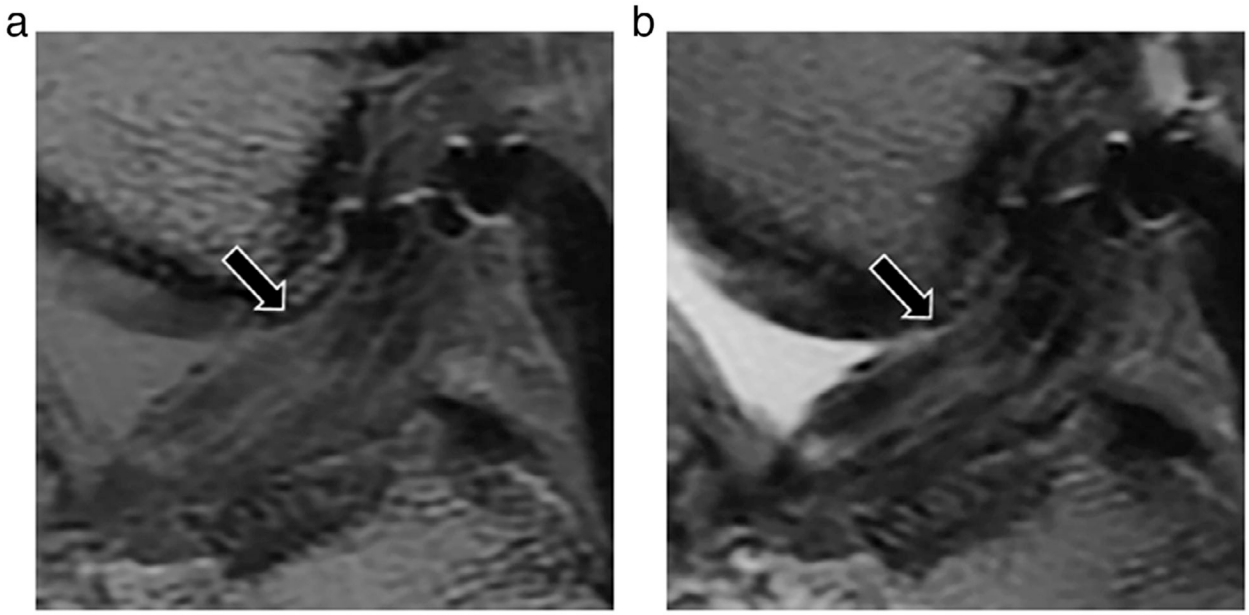

FIGURE 4:
Example of a true-positive interpretation of a deep-learning method for detecting an ACL tear within the knee joint on MRI (arrows) using sagittal **(a)** proton density-weighted and **(b)** fat-suppressed T_2-weighted 2D FSE images and arthroscopy as the reference standard. It should be noted that all clinical radiologists reviewing the same images had false-negative interpretations in this patient who had prior ACL reconstruction surgery. Figure obtained from a study performed by Liu et al.[21]

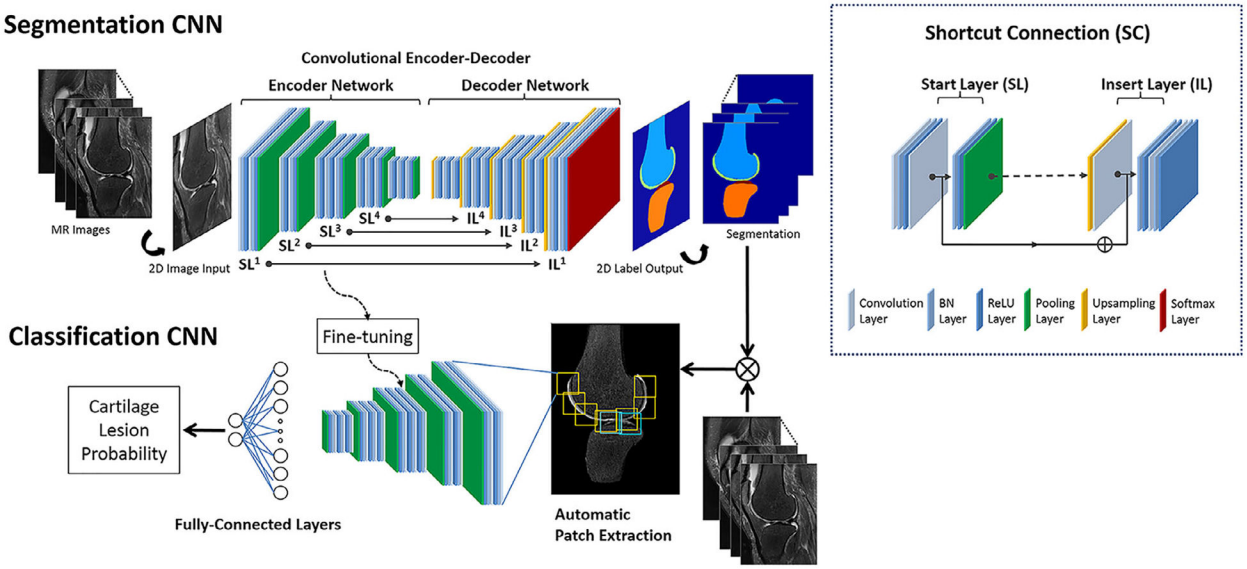

FIGURE 5:

Example of a coupled deep-learning approach for detecting cartilage lesions within the knee joint on MRI. The first CNN segmented the articular cartilage on the images, and the second classification CNN determined the presence or absence of cartilage lesions within a series of image patches extracted from the segmented cartilage. Figure obtained from a study performed by Liu et al.[17]

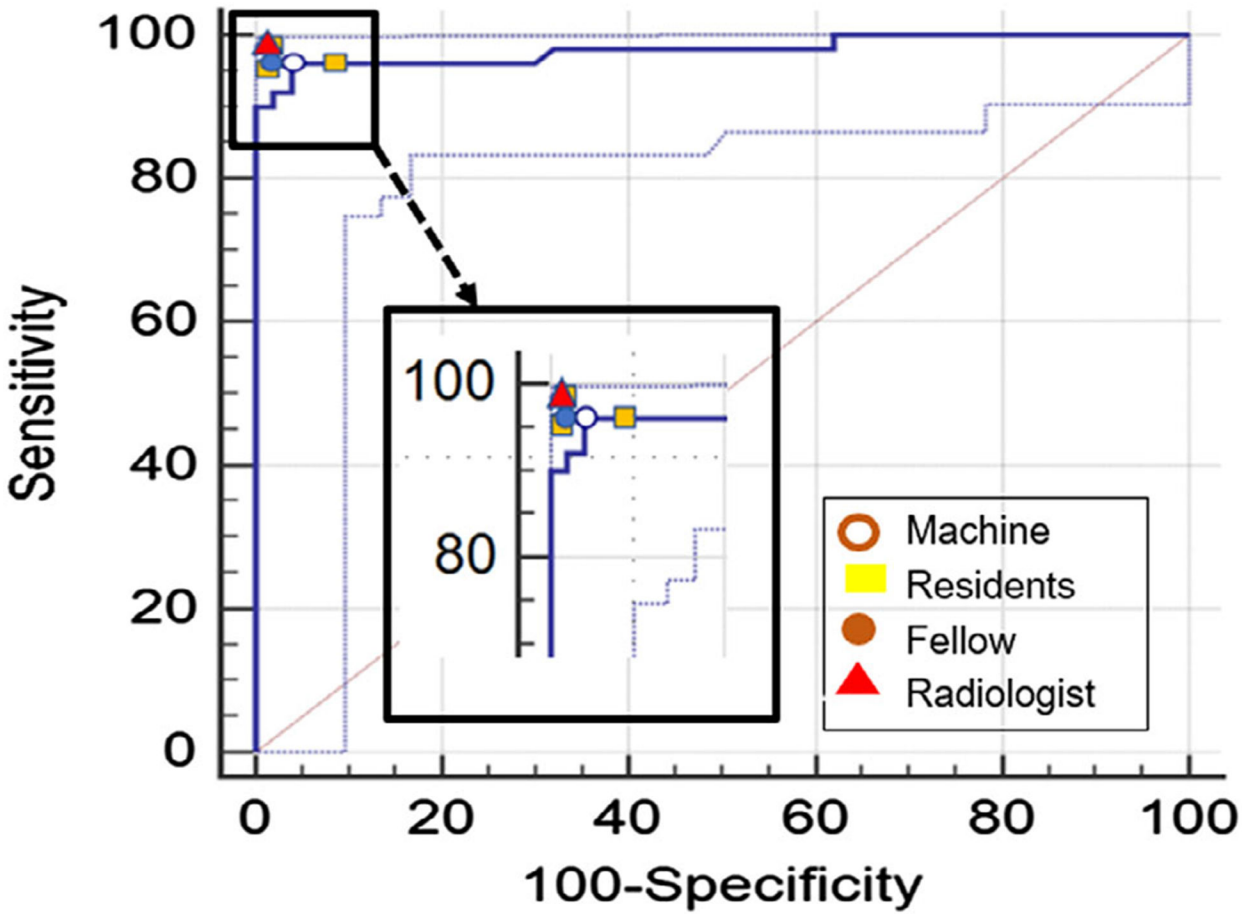

FIGURE 6:

ROC curve describing the diagnostic performance of a deep-learning method for detecting ACL tears within the knee joint on MRI using arthroscopy as the reference standard. The AUC of the machine was 0.98, indicating high overall diagnostic accuracy. The sensitivity and specificity for a musculoskeletal radiologist, a musculoskeletal radiology fellow, three radiology residents, and the machine at the optimal threshold of the Youden index are plotted. Note that the sensitivity and specificity of the human readers lie in close proximity to the ROC curve of the machine. Figure from a study performed by Liu et al.[21]

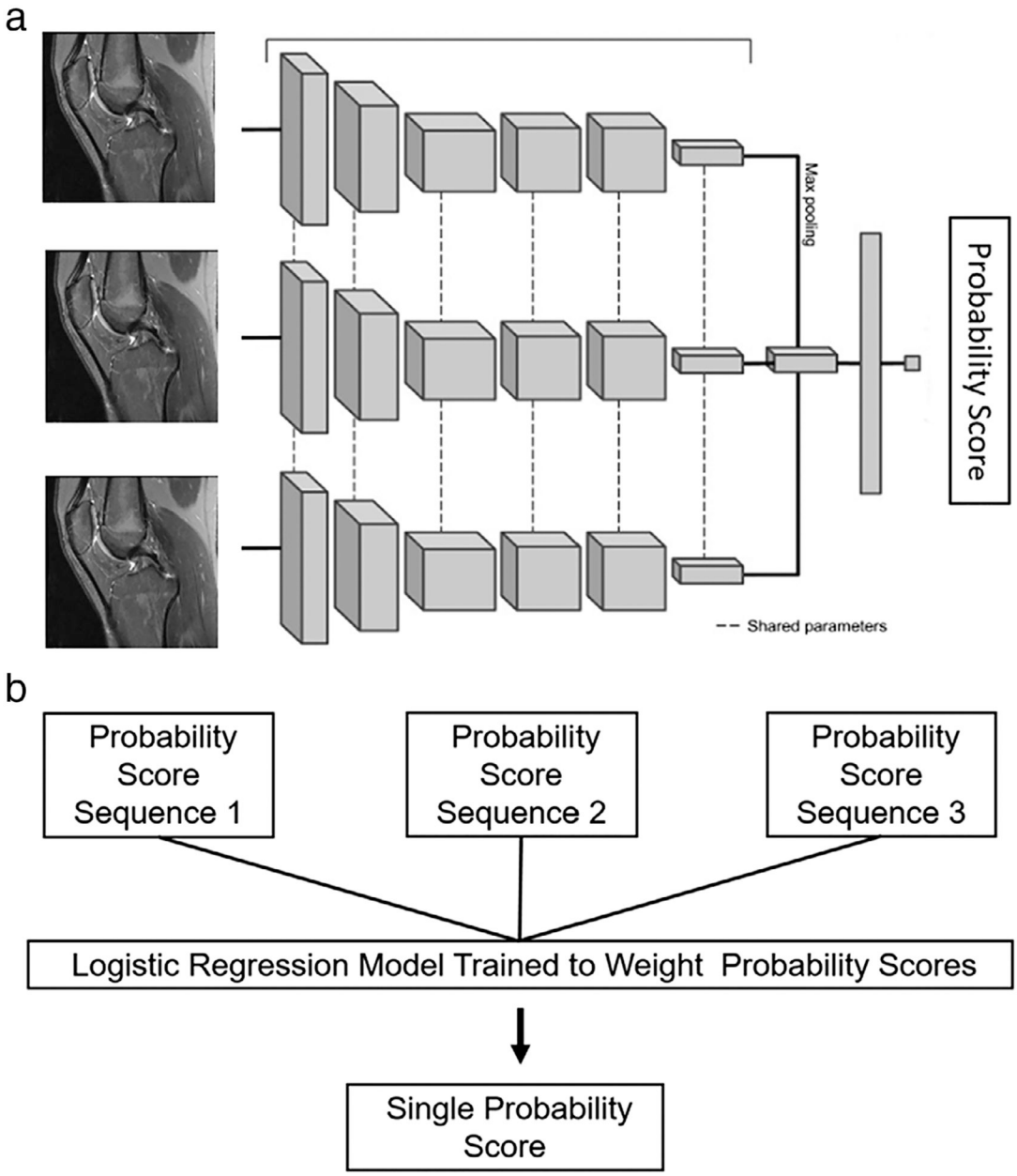

FIGURE 7:
Example of an alternative deep-learning approach for detecting ACL and meniscal tears within the knee joint on MRI. (a) The classification CNN used a series of nonsegmented images as the input and provided a probability score for the presence or absence of joint pathology for each sequence in the MRI examination. A logistic regression model was then used to combine the weighted probability scores from each individual sequence to obtain a final probability score for the presence or absence of ACL and meniscus tears. Figure obtained from a study performed by Bien et al.[23]

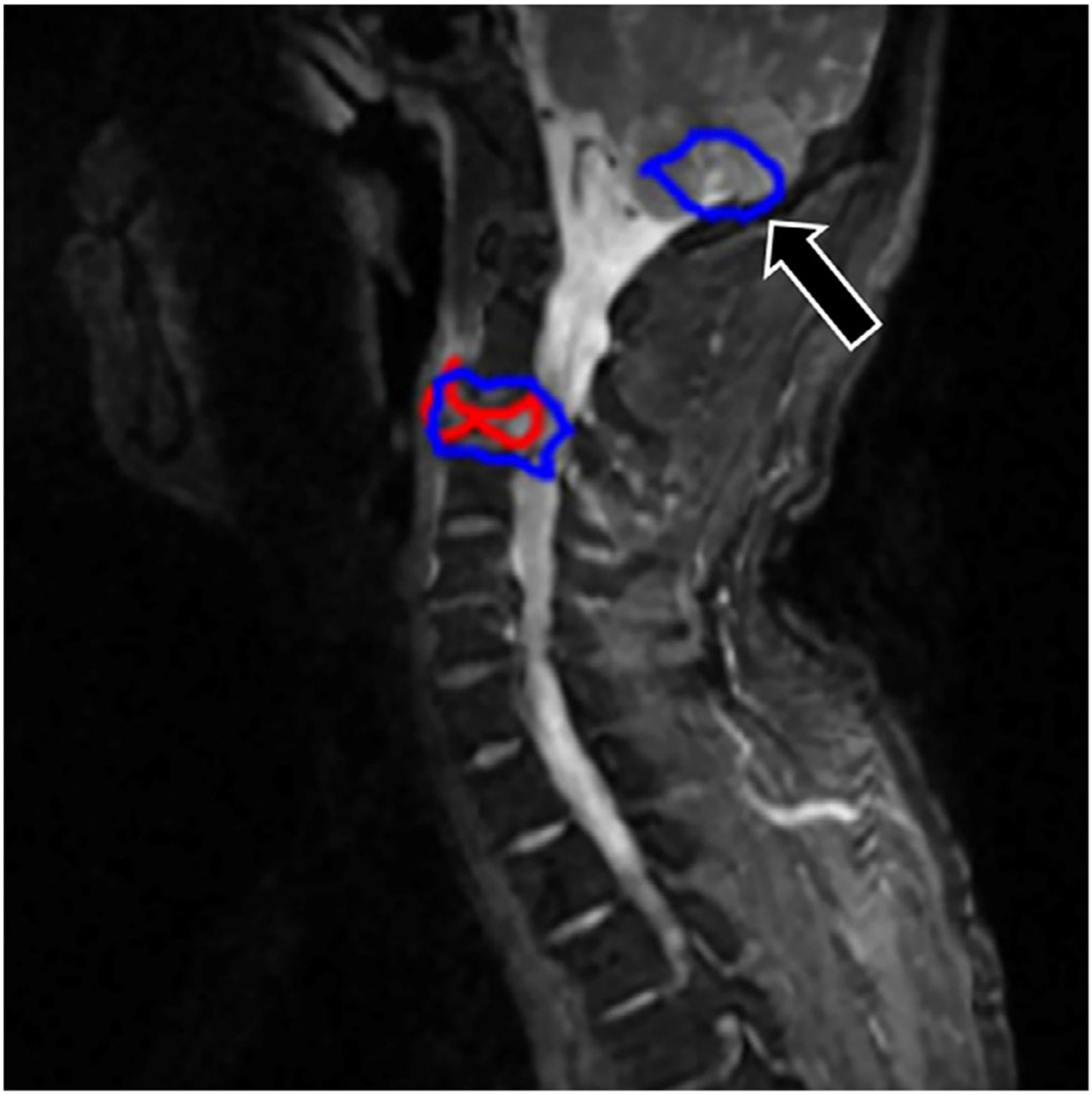

FIGURE 8:
Example of true-positive and false-positive interpretations of a deep-learning method for detecting vertebral body metastases on MRI using sagittal fat-suppressed T_2-weighted 2D FSE images. The boundaries of the metastatic lesions provided by the reference standard radiologist are marked by red contours, while the boundaries provided by the machine are marked by blue contours. Note the false-positive contour provided by the machine in the cerebellum (arrow). Figure reprinted with permission from a study performed by Wang et al. [26]

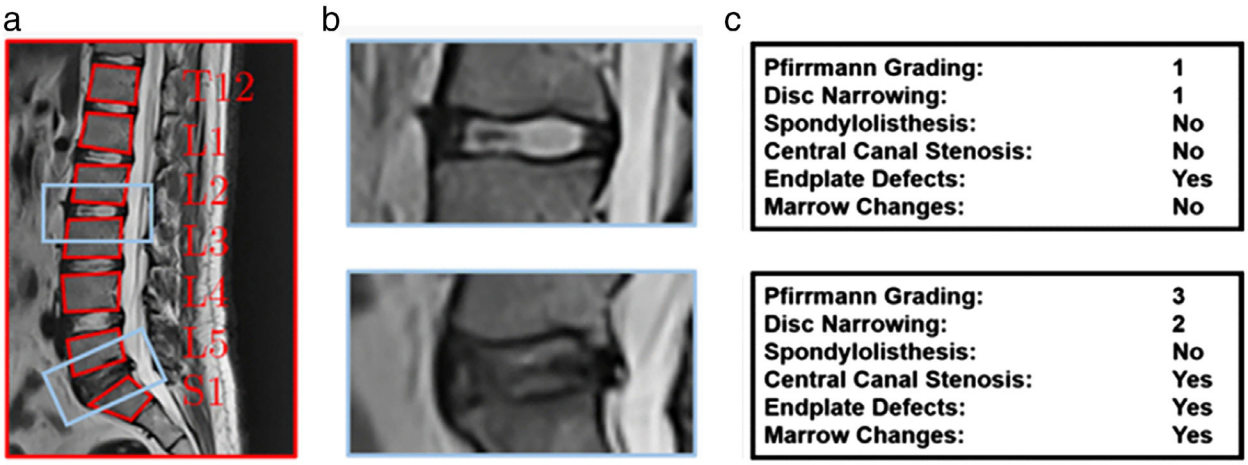

FIGURE 9:

Example of a coupled deep-learning approach for detecting degenerative disc disease within the spine on MRI. A custom-designed CAD method was used to **(a)** detect vertebrae regions (red boxes) and **(b)** extract vertebral body and disc segments (blue boxes) on sagittal T_2-weighted 2D FSE images. **(c)** The isolated vertebral body and disc segments were then analyzed by a classification CNN to determine the presence or absence of various findings of degenerative disc disease. Figure reprinted with permission from a study performed by Jamaludin et al.[29]

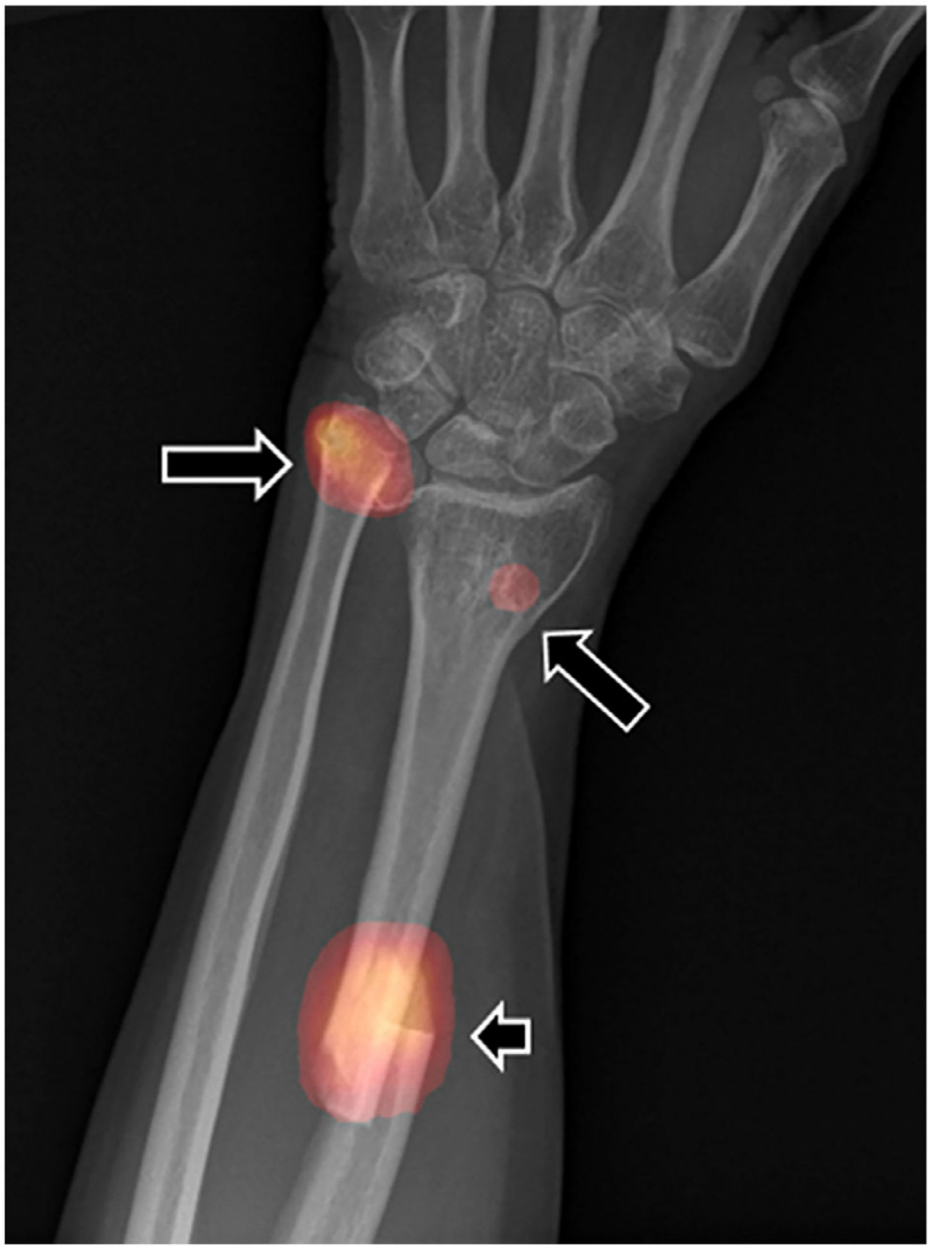

FIGURE 10:
Example of a true-positive interpretation of a deep-learning method for detecting fractures using anterior–posterior wrist radiographs and the interpretation of an experienced radiologist as the reference standard. The saliency map highlights the areas of acute (short arrow) and chronic (long arrows) fractures. While this is a true-positive interpretation, the machine could not distinguish between acute and chronic fractures and could not discern that the presence of both fracture types should raise suspicion for abuse, which was confirmed clinically. Figure provided by Hollis Potter, MD, from a study performed by Lindsey et al.[36]

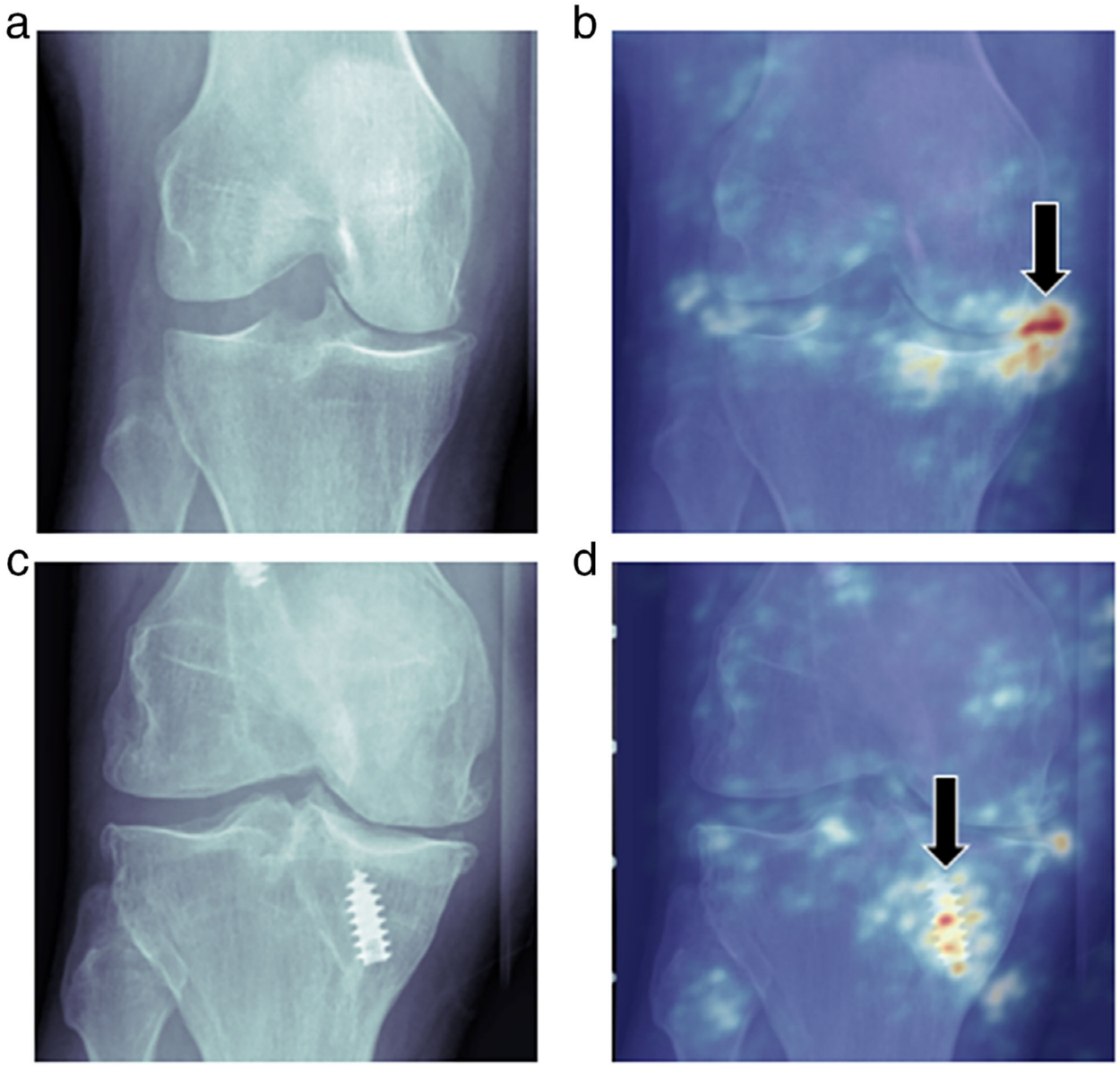

FIGURE 11: (a)
Example of a concordant interpretation of a deep-learning method for assigning a KL grade using anterior–posterior knee radiographs. Both the machine and reference standard radiologist assigned a KL grade of 2. **(b)** Corresponding saliency map shows the high probability regions on which the machine based its interpretation, which were primarily located over the medial joint space (arrow). **(c)** Example of a disconcordant interpretation of a deep-learning method for assigning a KL grade using anterior-posterior knee radiographs. The machine and reference standard radiologist assigned KL grades of 2 and 3, respectively. **(d)** Corresponding saliency map shows the high probability regions on which the machine based its interpretation, which were primarily located over a screw in the medial tibia (arrow). Figure from a study performed by Norman et al.[48]

TABLE 1.

Interreader Agreement Between a Deep Learning Method and an Experienced Radiologist and Intrareader Agreement Between the Same Radiologist for Detecting Various Findings of Degenerative Disc Disease of the Lumber Spine on MRI

MRI finding	Interreader agreement	Intrareader agreement
Pfirrman disc grade	70%	70%
Disc narrowing grade	75%	72%
Central canal stenosis	95%	80%
Upper endplate defects	87%	81%
Lower endplate defects	88%	83%
Lower endplate marrow changes	90%	93%
Lower endplate marrow changes	89%	91%
Spondylolisthesis	95%	90%

TABLE 2.

Diagnostic Performance of Deep Learning Methods for Detecting Fractures

			Machine performance			
Fracture site	**Dataset size**	**CNN used**	**AUC**	**Sensitivity**	**Specificity**	**Accuracy**
Hip 33	805	GooLeNet	0.98	NS	NS	94%
Hip 34	3,605	DenseNet	0.98	98%	84%	91%
Hip 35	3,346	VGG-16	NS	94%	97%	96%
Shoulder 37	1,891	ResNet	0.99	99%	97%	95%
Wrist 40	7,356	ResNet	0.90	98%	73%	NS
Wrist 38	1,389	Inception	0.95	90%	88%	NS
Wrist 39	256,000	VGG-16	NS	NS	NS	82%
Ankle 41	596	Xception	NS	73%	76%	75%
All Sites 36	135,409	U-Net	0.99	94%	95%	NS