10.1148/radiol.2020192091

Appendix E1

Image Preprocessing

Images were extracted from DICOM files and were inverted if the image type on DICOM was labeled as MONOCHROME1. Since image resolution is not uniform among radiographs, we applied bicubic spline interpolation to obtain a fixed resolution of 0.2×0.2 mm^2. We applied global contrast normalization (1) and histogram truncation between the fifth and 99th percentiles to reduce the impact of noise (10). From interpolated images, individual knee images were extracted from bilateral posteroanterior fixed-flexion knee radiographs. We extracted 1024 by 1024 pixels from radiographs covering either the left or the right knee joint using the center of bounding boxes provided by Tiulpin et al (10) and accessed from a github repository (https://github.com/MIPT-Oulu/DeepKnee/tree/master/Dataset). These bounding boxes were generated using a model based on support vector machines and necessary manual corrections on mislabeled bounding boxes. The resulting image arrays were replicated to 3 channels to mimic RGB input requirement on ResNet models and they were saved in HDF5 format to be used in model training and evaluation,

Deep Learning Model

Residual network (ResNet) architectures have been successfully employed in a range of image recognition tasks providing accurate classification. We used a publicly available ResNet with 34 layers (ResNet34) model (23) that is pretrained on ImageNet2012 dataset using 1.28-million images that reports a top-1 error of 26.7% and a top-5 error of 8.58%. Transfer learning on the pretrained ResNet34 model was performed by fine-tuning the weights of this network with knee radiographs. The output represented the probability of TKR within nine years given an input radiograph. The probability of TKR within nine years was a risk factor in the statistical analysis. We implemented our deep learning models based on ResNet34 architecture with changes on the average pooling layer and fully connected layer. Kernel size of the average pooling operator was changed from 7 to 28 to enable the use of original resolution image of size 1024 × 1024. The fully connected layers size was changed from 1000 to 2 for all models to identify the binary classification of patients with TKR versus controls. An additional fully connected layer of size 5 was added on the DL-TL-MT model to identify the KL-grade of the radiograph. These final classification layers were randomly initialized (2) and they used a softmax activation function to compute two separate probabilities (multitask learning): i) $p(y \mid \boldsymbol{x})$, the probability of TKR (y) and ii) $p(KL \mid \boldsymbol{x})$, the probability of KL-grade (*KL*) given an input knee radiograph $(\boldsymbol{x})$.

Model training was performed by updating the weights of these networks with knee radiographs using a cross-entropy loss functions on TKR and KL-grade prediction tasks. The adaptive moment estimation (Adam) optimizer (3) with default running average and eps parameters were used. Grid search for learning rate was used between 1e-2 and 1e-5 with a 10-fold decrease; the optimal learning rate used in the study was 1e-4. Batch size of 8 was used for 200 epochs. The model selection was performed based on the best accuracy on predicting the TKR on a validation set within 200 epochs. Random cropping to extract 896 × 896 pixel images and random horizontal flipping were used for data augmentation during training. For validation, we used center cropping.

Model Selection and Evaluation

To identify an optimal model, we implemented a seven-fold nested cross-validation (CV) scheme to optimize hyperparameters of the learning algorithm. The purpose of nested CV is to identify learning parameters that generalize well across the population samples we learn from in each fold. In nested CV, stratified random sampling used to partition the 728 patients and controls in the matched subcohort into seven disjoint groups, with each having 52 TKR patients and 52 controls, and the patients in each fold were consistent among all our trained models. Each of the seven groups served as a test set to assess the performance of a prediction model (outer loop). DL models were identified using 624 patients and controls (trained and validated). The prediction model was derived using a set of hyperparameters applied to the training set of 520 patients and controls from the other five groups combined and validated on set of 104 patients and controls (inner loop). In this way, six separate prediction models were derived for each test set, with each model applied to predict the TKR outcome of patients and controls in a test set with data independent of that used to derive models. DL models were tested on the 104 patients and controls not used for either training or validation of the models. The test set does not contribute data to the derivation of the "best fit" model, and as such the test data are indeed independent of the data used to fit the model. In each test set, TKR probability and KL-grade predictions were averaged from six models (developed within inner loop) and they are used in statistical analysis. The results reported in this work were from an independent group of patients and controls who were not used for training/validating models. The use of nested CV eliminated the bias that can be introduced by conventional CV due to hyperparameter tuning implemented during training.

"TestSets" folder under code repository (https://github.com/denizlab/oai-xray-tkr-klg) provides .csv files that provides subject IDs for each patent and control who are included seven separate groups for nested CV. Filenames match with the "Test Set Numbers" defined in Table 5 of the paper. In addition to the subject IDs from OAI study, separate columns identify TKR status of patients (0: controls, 1:patients underwent TKR within 9 years from baseline), knee side (0: Left Knee, 1: Right knee), KL grade from the patient's knee and strata (from case-control matching).

Visualization of Regions Affecting Prediction of DL models

We employed visualization tools to identify the regions in which the trained DL models use to make a decision with high impact. Gradient-weighted class activation mapping (Grad-CAM (27)) method was used for interpreting convolutional network behavior. Given the desired output class, Grad-CAM provides a heat map of where the network is more activated and "focused on" without the need for retraining the existing network. Grad-CAM generates a heat map by calculating the weighted average of activations from all kernels after the last layer of convolution. The weight is determined by back-propagating the selected class through fully connected layers until they reach the layer where the activations are aggregated. The weight also goes through a ReLU function to filter out negative values for a clear representation of positively activated areas. The weighted average is then overlaid on top of the original images for visualization. We calculated heatmaps from each model trained in the inner CV and heatmaps are spatially averaged to obtain a single heatmap used in Figure 4.

1. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge, Mass: MIT Press, 2016.

2. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Aistats. 2010;9:249–256. http://machinelearning.wustl.edu/mlpapers/paper_files/AISTATS2010_GlorotB10.pdf. Accessed DATE.

3. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. CoRR. 2014; abs/1412.6:1–15. http://arxiv.org/abs/1412.6980. Published 2014. Accessed DATE.