RadImageNet: A Large-scale Radiologic Dataset for Enhancing Deep Learning Transfer Learning Research

Yang Yang (yang.yang@mssm.edu)  
Icahn School of Medicine at Mount Sinai  
https://orcid.org/0000-0002-2841-4243

Xueyan Mei  
Icahn School of Medicine at Mount Sinai

Philip Robson  
Icahn School of Medicine at Mount Sinai

Brett Marinelli  
Icahn School of Medicine at Mount Sinai

Mingqian Huang  
Icahn School of Medicine at Mount Sinai

Amish Doshi  
Icahn School of Medicine at Mount Sinai

Adam Jacobi  
Icahn School of Medicine at Mount Sinai

Katherine Link  
Icahn School of Medicine at Mount Sinai

Thomas Yang  
Icahn School of Medicine at Mount Sinai

Chendi Cao  
Icahn School of Medicine at Mount Sinai

Ying Wang  
University of Oklahoma

Hayit Greenspan  
Icahn School of Medicine at Mount Sinai

Timothy Deyer  
East River Medical Imaging

Zahi Fayad  
Icahn School of Medicine at Mount Sinai  
https://orcid.org/0000-0002-3439-7347

Biological Sciences - Article
Abstract

Most current medical imaging Artificial Intelligence (AI) relies upon transfer learning using convolutional neural networks (CNNs) created using ImageNet, a large database of natural world images, including cats, dogs, and vehicles. Size, diversity, and similarity of the source data determine the success of the transfer learning on the target data. ImageNet is large and diverse, but there is a significant dissimilarity between its natural world images and medical images, leading Cheplygina to pose the question, “Why do we still use images of cats to help Artificial Intelligence interpret CAT scans?”. We present an equally large and diversified database, RadImageNet, consisting of 5 million annotated medical images consisting of CT, MRI, and ultrasound of musculoskeletal, neurologic, oncologic, gastrointestinal, endocrine, and pulmonary pathologies over 450,000 patients. The database is unprecedented in scale and breadth in the medical imaging field, constituting a more appropriate basis for medical imaging transfer learning applications. We found that RadImageNet transfer learning outperformed ImageNet in multiple independent applications, including improvements for bone age prediction from hand and wrist x-rays by 1.75 months (p<0.0001), pneumonia detection in ICU chest x-rays by 0.85% (p<0.0001), ACL tear detection on MRI by 10.72% (p<0.0001), SARS-CoV-2 detection on chest CT by 0.25% (p<0.0001) and hemorrhage detection on head CT by 0.13% (p<0.0001). The results indicate that our pre-trained models that are open-sourced on public domains will be a better starting point for transfer learning in radiologic imaging AI applications, including applications involving medical imaging modalities or anatomies not included in the RadImageNet database.

Introduction

ImageNet\textsuperscript{2,3} is a dataset comprising millions of images of the natural world. ImageNet, as an open-sourced dataset, has been a central resource to derive sophisticated models in computer vision. Transfer learning\textsuperscript{1} is a common deep learning approach whereby a model designed for one problem can be reused to initiate a different but related task in machine learning. Due to the lack of annotated images and limited resources of computing power to train new models from scratch, transfer learning has become a popular method in deep learning for researchers to transfer the knowledge gained from pre-trained models to a related problem and thus speed up the training process with fewer input data and improve the performance and generalizability of a deep learning model\textsuperscript{11}. Transfer learning with models trained using ImageNet has been extensively explored in medical imaging AI applications. The architectures of VGG\textsuperscript{12}, ResNet\textsuperscript{13–15}, Inception networks\textsuperscript{16–19}, MobileNet\textsuperscript{20}, and DenseNet\textsuperscript{21} pre-trained with ImageNet have been widely adopted and used in medical imaging applications for COVID-19 diagnosis on chest CT\textsuperscript{22}, classification of fibrotic lung disease\textsuperscript{23}, and classification of skin cancer\textsuperscript{24}. Despite the high performance of many medical imaging models pre-trained with ImageNet, successful transfer learning requires reasonably large sample size, diversity of images, and similarity between the training image database and the target application images. While ImageNet meets the size and diversity criteria, there remains a significant dissimilarity between the training images in ImageNet and the medical images in
the new task that represents an important limitation. The development of transfer learning strategies to bridge that gap is an active medical imaging machine learning research area.

RadImageNet provides millions of annotated advanced medical images from various modalities demonstrating numerous different pathologies and can be used to develop pre-trained models for predictive tasks in image-based medicine. We propose that a pre-trained model based solely on medical radiographic features from a vast medical imaging database will provide more appropriate feature representation for image-based predictive problems in medicine than pre-trained models derived from the natural images in ImageNet.

In this study, we describe a large-scale, diverse medical imaging dataset, RadImageNet, to generate pre-trained convolutional neural networks trained solely from medical imaging to be used as the basis of transfer learning for medical imaging applications. We compare the pre-trained weights derived from RadImageNet and ImageNet on multiple medical imaging use cases, including target image modalities and anatomies not included in the training RadImageNet database. We show that the pre-trained networks generated from RadImageNet exceed the performance of pre-trained models developed from ImageNet. Furthermore, we show how a medical image recognition problem with a small medical image dataset could benefit from pre-trained weights derived from RadImageNet. This provides evidence that the pre-trained weights from RadImageNet can be transferable across multiple modalities, anatomies, and pathologies. Figure 1 illustrates an overview of this study.

The RadImageNet Database

The RadImageNet dataset includes 5 million annotated CT, MRI, and ultrasound images of musculoskeletal, neurologic, oncologic, gastrointestinal, endocrine, and pulmonary pathology. For direct comparison with ImageNet (the initial size for the ImageNet challenge was 1.4 million images), we collected the most frequent modalities and anatomies on the same scale. The RadImageNet dataset was collected between January 2005 and January 2020 from 131,872 patients at an outpatient radiology facility in New York City. Each study was annotated by a board-certified fellowship-trained radiologist. As part of the interpretation of each study, the reading radiologist chose images representative of the pathology shown in each exam. The pathology demonstrated on each of these “key images” was annotated, and a region of interest was created to identify the imaging findings. These annotations were extracted from the key images and provided the basis for the RadImageNet classes. The portions of the RadImageNet database used for comparison to ImageNet consist of three radiologic modalities, eleven anatomies, and 165 pathologic labels (Fig. 2a and Extended Data Table 1). Inception-Res-Net-v2 \(^{16-19}\), ResNet50 \(^{13-15}\), DenseNet121 \(^{21}\), and InceptionV3 \(^{16}\) convolution neural network architectures were generated from the data in RadImageNet. We stratified the RadImageNet dataset by patient ID, allowing no overlaps in either training, validation, and test sets. The dataset was split into 75% training set, 10% validation set, and 15% test set. The performance of the models was reported on the test set. Furthermore, we randomly sampled 2,016 images from the test set and compared the model performance
to three senior sub-specialized fellowship-trained radiologists who were uninvolved with the labeling of the images in RadImageNet (Extended Data Fig. 1).

The RadImageNet Inception-Res-Net-v2, ResNet50, DenseNet121 and InceptionV3 models achieved top-1 accuracies of 78.86%, 75.04%, 74.47% and 76.13% respectively and top-5 accuracies of 95.52%, 94.30%, 93.93%, and 95.23% respectively (Fig. 2b).

**Comparison Of Radimagenet And Imagenet Pre-trained Models**

The four aforementioned RadImageNet models were used as the pre-trained models for five medical imaging applications to compare their performance to the ImageNet pre-trained models. We applied the pre-trained models to transfer learning problems using publically available datasets, including bone age prediction on hand and wrist x-rays⁶ pneumonia detection in ICU patients on chest radiographs⁷; ACL tear detection on MRI⁸; SARS-CoV-2 detection on chest CT⁹; and hemorrhage detection on head CT¹⁰. These applications were selected to evaluate the capabilities of RadImageNet models on multiple applications that included modalities, anatomies, and labels that were and were not contained in the RadImageNet database.

**Bone Age Prediction**

For bone age prediction on hand and wrist x-rays, the RadImageNet-trained models showed significant reduction in mean absolute errors compared to the ImageNet-trained models. For RadImageNet-trained Inception-Res-Net-v2, ResNet50, DenseNet121 and InceptionV3 networks, (Fig. 3a) mean absolute errors were 10.42 months (mean absolute deviation (MAD) = 6.89; P = 0.0071), 11.12 months (MAD = 6.99; P < 0.0001), 10.97 months (MAD = 6.91; P < 0.0001), and 10.25 months (MAD = 6.54; P < 0.001 ) respectively, whereas for ImageNet-trained Inception-Res-Net-v2, ResNet50, DenseNet121 and InceptionV3 networks mean absolute errors were 11.01 months (MAD = 7.03), 15.22 months (MAD = 10.18), 12.26 months (MAD = 7.87), and 11.29 months (MAD = 7.20) respectively. The RadImageNet models also demonstrated narrower standard deviation on modified Bland-Altman plots²⁵²⁶ (Extended Data Fig. 2)

**Pneumonia Detection**

For pneumonia detection on chest radiographs, the RadImageNet-trained models showed significant improvement in the area under curve of receiver operating characteristics (AUROC) compared to the ImageNet-trained models. For RadImageNet-trained Inception-Res-Net-v2, ResNet50, DenseNet121 networks the AUROC was 87.42% (95% confidence interval (CI) 86.22%, 88.63%; P = 0.0032), 86.81% (95% CI 85.55%, 88.07%; P = 0.00042), 86.93% (95% CI 85.68%, 88.19%; P = 0.013) respectively, whereas for ImageNet-trained Inception-Res-Net-v2, ResNet50 and DenseNet121 networks the AUROC was 86.34% (95% CI 85.04%, 87.54%), 85.48% (95% CI 84.16%, 86.80%) and 86.02% (95% CI 84.72%, 87.32%) respectively (Fig. 3b and Extended Data Fig. 3a). The InceptionV3 network of RadImageNet achieved an AUROC of 86.51% (95% CI 85.24%, 87.78%; P = 0.36) that showed a non-significant difference from the InceptionV3 network of ImageNet that showed an AUROC of 86.15% (95% CI 84.84%, 87.45%).
ACL Tear Detection

For ACL tear detection on MRI, due to the limited size of the dataset, we conducted 5-fold cross validation for this application. The performance of the models was reported on the mean and standard deviation on the five folds of test sets (Fig. 3c and Extended Data Fig. 3b). The RadImageNet Inception-Res-Net-v2, ResNet50, DenseNet121 and InceptionV3 demonstrated AUROC of 92.27% ± 3.45% (P < 0.0001), 87.11% ± 3.10% (P < 0.0001), 90.41% ± 5.90% (P = 0.12) and 92.33% ± 3.13% (P < 0.0001) respectively while the ImageNet Inception-Res-Net-v2, ResNet50, DenseNet121 and InceptionV3 showed AUROC of 83.80% ± 6.49%, 81.72% ± 3.61%, 90.27% ± 2.08% and 60.20% ± 9.78% respectively.

SARS-CoV-2 Detection

For SARS-CoV-2 detection on chest CT, the RadImageNet-trained models showed significant improvement in the AUROC compared to the ImageNet-trained models. For RadImageNet-trained Inception-Res-Net-v2 and ResNet50 networks the AUROC was 99.88% (95% CI 99.84%, 99.92%; P < 0.0001) and 99.89% (95% CI 99.85%, 99.92%; P < 0.0001) respectively whereas for ImageNet-trained Inception-Res-Net-v2 and ResNet50 the AUROC was 99.28% (95% CI 99.17%, 99.40%) and 99.59% (95% CI 99.47%, 99.71%) respectively (Fig. 3d and Extended Data Fig. 3c). The RadImageNet DenseNet121 and InceptionV3 networks achieved an AUROC of 99.67% (95% CI 99.56%, 99.78%; P = 0.29) and 99.74% (95% CI 99.66%, 99.83%; P = 0.74) respectively that showed a non-significant difference from the ImageNet DenseNet121 and InceptionV3 networks that achieved an AUROC of 99.62% (95% CI 99.55%, 99.68%) and 99.73 (95% 99.64, 99.82%) respectively.

Hemorrhage Detection

For hemorrhage detection on head CT, the RadImageNet-trained models illustrated a significant improvement in the AUROC compared to the ImageNet-trained models. For RadImageNet-trained Inception-Res-Net-v2, ResNet50 and InceptionV3 networks the AUROC was 96.29% (95% CI 96.14%, 96.43%; P = 0.0090), 96.31% (95% CI 96.17%, 96.45%; P < 0.0001), and 96.40% (95% CI 96.26%, 96.54%; P = 0.00064) respectively whereas for ImageNet-trained Inception-Res-Net-v2, ResNet50 and InceptionV3 networks the AUROC was 96.16% (95% CI 96.02%, 96.31%), 96.06% (95% CI 95.91%, 96.22%) and 96.23% (95% CI 96.09%, 96.38%) respectively (Fig. 3e and Extended Data Fig. 4). The RadImageNet DenseNet121 achieved an AUROC of 96.22% (95% CI 96.07%, 96.37%; P = 0.14) that showed a non-significant difference from the ImageNet DenseNet121 which achieved an AUROC of 96.14% (95% CI 95.99%, 96.30%).

The gradient class activation map (Grad-cam) was used to demonstrate the features learned by the algorithms. We present the Grad-cam (Fig. 4) for the paired algorithms on each of the five applications to visualize the distinguishing features captured by the models. Grad-cam images of successful predictions of the RadImageNet and ImageNet models were used to compare the learned features.

Potential Clinical Applications
The pre-trained models trained from RadImageNet can improve predictive performance and generalizability for medical imaging applications. We simulated multiple scenarios of different medical imaging applications with multiple modalities, anatomies, and pathologies where RadImageNet models inherited some or none of the knowledge about that application. These applications show that the pre-trained RadImageNet models demonstrated significant improvement compared to the ImageNet pre-trained models. These outcomes suggest that the RadImageNet pre-trained models can improve medical imaging applications where transfer learning is needed. Moreover, gradient class activation maps suggest the interpretation of the RadImageNet models more closely conforms to the regions of interest as defined by radiologists.

Discussion

The key determinants of successfully developing models using transfer learning are to have source data that demonstrates a certain level of similarity to the target data, diversity of image type, and a large training sample size. Studies\textsuperscript{22–24} have shown that ImageNet pre-trained models that have a large number of classes and large-scale, despite the low similarity to medical data, demonstrate a high recognition rate in medical imaging analysis. If a large sample size or diverse source data is missing, source data with higher similarity to the target data can also lead to success\textsuperscript{28–32}. Models developed from the RadImageNet dataset combine the positive attributes of both methods. RadImageNet consists of 5 million annotated images. In this study, a subset of 1.4 million images was used to match the size of the ImageNet database. Moreover, RadImageNet data are more similar to the target medical imaging data and include 165 classes of target image from multiple modalities and anatomies.

The bone age hand and wrist x-rays dataset is relatively small, consisting of 12,611 images, and demonstrates the lowest similarity to the RadImageNet dataset since the no hand and wrist x-ray studies were contained in the dataset. Despite the reduced similarity of the modality of the target data compared to the modalities in the source data, the four RadImageNet models all resulted in a smaller mean absolute error (P < 0.05) than the ImageNet models. This indicates that while the modalities differed, the underlying features in RadImageNet were more useful than those in ImageNet, suggesting more broad applicability to medical imaging of models derived from RadImageNet than ImageNet.

Three out of four RadImageNet models showed significant improvement in AUROC (P < 0.05) than the ImageNet models for pneumonia detection on chest radiographs in ICU patients. This dataset contains 26,684 images, which suggests that a larger size of the target data may compensate for the lack of similarity in source data as pneumonia chest radiographs were not present in the RadImageNet database. This is in contrast to the ACL tear dataset, which is extremely small with only 1,021 images. Three out of four RadImageNet models demonstrated a higher mean AUROC (P < 0.0001) and smaller standard deviation on 5-fold cross-validation. RadImageNet contained both the modality (MRI) and a similar class (ACL injury) to the target data indicating that source data similarity can contribute to extraordinary performance with even a small dataset.
Two out of four RadImageNet models outperformed ImageNet models (P < 0.05) on SARS-CoV-2 detection on chest CT. RadImageNet contained the same modality (chest CT) and a similar label (pulmonary consolidation) which likely helped the RadImageNet models outperform the ImageNet models. However, the similarity of the images was likely compensated by the large size of the target SARS-CoV-2 dataset (58,766 images), accounting for the similar performance of the remaining models, which showed a non-significant difference between RadImageNet and ImageNet. The performance on the intracranial hemorrhage data on head CT was likely due to a similar phenomenon. Three out of four RadImageNet models illustrated significant improvement (P < 0.05). These results are likely due to a combination of a similar label (acute intra-axial hemorrhage) and anatomy (brain) being included in the RadImageNet dataset, whereas the ImageNet models were able to compensate due to the large target data set (573,614 images). This again suggests that the underlying image features in RadImageNet are transferable to other medical pathologies other than those included in the database.

These five clinical applications show that RadImageNet pre-trained models, despite varying levels of similarity and diversity to the target medical imaging data, demonstrate superiority to ImageNet pre-trained models and hold promise to aid development and clinical translation of medical imaging artificial intelligence. Immediate adoption of these models can be achieved by open sourcing them on public domains (https://github.com/BMEII-AI/RadImageNet).

Our proposed RadImageNet models do have limitations. First, one major limitation of this study is that only a single sequence per patient was provided for assessment. Many pathologies require additional sequences and/or adjacent images for accurate diagnosis. Second, the images presented may contain multiple pathologies but only one label. The annotating radiologists may only label the major findings of the key diagnosis while not exhaustively annotating all other pathologies demonstrated on the image. Third, we provided the radiologists with full-resolution images while the RadImageNet models utilized lower resolution images in algorithm development due to processing limitations. These lower resolution images may obscure small areas of pathology. Finally, the number of classes in the limited RadImageNet dataset used for comparison to ImageNet was less than the number in ImageNet.

In future studies, higher spatial-resolution images could result in higher performance for recognition of smaller foci of pathology. The number of classes of pathology in RadImageNet can be further expanded to match the number of classes in ImageNet. The success of ImageNet is in part due to the number of classes available to discriminate between objects. For example, “dog” is expanded to encompass “Husky”, “Golden Retriever”, etc. Moreover, performance could be improved by introducing the regions of interest, as defined by radiologists, to highlight pathological appearance in the images, as well as by providing additional sequences and/or adjacent images. In addition, pre-trained models for CT only or MRI only derived from the RadImageNet dataset can be developed for CT or MRI applications compared to the comprehensive RadImageNet models. Finally, more studies of fine-tuning the pre-trained models compared to the standard pre-trained models will be further analyzed.
In conclusion, RadImageNet (5 million annotated CT, MRI, and ultrasound images) and the associated pre-trained models illustrate the important role of a database with a higher degree of similarity between the source images and the target application as a starting point for transfer learning approaches in medical imaging analysis. We believe the proposed RadImageNet pre-trained models, based on a large-scale, diverse, and high-quality annotated dataset of medical images with a high degree of similarity to the target applications, could improve the recognition rate and visualizations of other medical imaging CNN-based transfer learning applications.

Declarations

Competing interests

Timothy Deyer

Dr. Timothy Deyer is the co-founder of RadImageNet LLC.

Author contributions

X.M., T.D., Y.Y., and Z.A.F contributed to the design and conception of the study. X.M. and T.D. contributed to data collection and acquisition. X.M., T.Y., C.C., K.E.L., and T.D. contributed to the analysis and interpretation of the data. M.H., A.J., and A.D. contributed to reading reader studies. X.M. and T.D. contributed to drafting the manuscript. X.M. and Y.W. performed the statistical analysis. All authors contributed to revising the manuscript.

Acknowledgments

We would like to acknowledge multiple contributors for this project: Drs. Richard Katz, Morton Scheneider, Steven Albert, Alison Haimes, Stephen Greenberg, Douglas Decorato, Gavin Duke, Paul Choi, Sean Herman, Robert Ludwig, Gwen Harris, Adam Wilner, Mark Pinals, Nicole Lee, Clyde Hershon, Michelle Klein, and Barbara Braffman at East River Medical Imaging who annotated the studies; David Vazquez and Justin Ponquinette from Department of Information Technology at East River Medical Imaging who helped collect and transfer the data. We thank Chenyu Liu from Icahn School of Medicine at Mount Sinai who helped analyze tabulated data and make figures.

Online Methods

Study participants.

The study was approved by the institutional review board of East River Medical Imaging (data provider) and the Icahn School of Medicine at Mount Sinai in New York (data receiver). The institutional review boards waived the requirement for written informed consent for this retrospective study, which evaluated de-identified data and involved no potential risk to patients. To avert any potential breach of
confidentiality, no link between the patients, the data provider, and the data receiver was made available.
A third party issued a certification of de-identified data transfer from the data provider to the data receiver.

We collected a total of 203,341 CT (52,691), MRI (142,422), and ultrasound (8,228) studies from 131,872
patients between 1 January 2005 and 31 January 2020 where patients had diagnostic scans at East
River Medical Imaging in New York.

**Multimodal and multi-anatomy components**

CT exams of the chest, abdomen, and pelvis; MRI exams of the shoulder, knee, ankle, foot, spine, knee, hip
and brain; and ultrasound studies of the thyroid, abdomen, and pelvis were collected in the curation of the
RadImageNet database. For each study, the radiologist annotated key images with a corresponding
diagnosis. To create a reasonable number of classes, the annotations were grouped by pathology and
imaging appearance resulting in a total of 11 anatomies from 3 modalities and 165 diagnostic labels. For
example, ACL tears and ACL sprains on MRI were combined to a single class - MRI, Knee, ACL injury (see
Extended Data Table 3).

**Normal studies**

To better investigate the characteristics of abnormal key images for model development, we queried
8,528 normal studies based on radiology reports for the aforementioned modal and anatomical images.
Each normal study was further confirmed by a board-certified radiologist (T.D.). All associated diagnostic
sequences and images were included.

**RadImageNet model development**

To develop the pre-trained models from RadImageNet, we trained 4 different convolutional neural
networks without importing weights from existing models, namely, Inception-Res-Net-v216-19,
ResNet5013-15, DenseNet12121, and InceptionV316. The dataset was split into 75% training set, 10%
validation set, and 15% test set stratified by patient ID. The associated images from the same patient
were in the same set.

Rather than importing the weights from existing models, we randomly initialized the weights to train the
individual models. A global average pooling layer, a dropout layer at a rate of 0.4, and the output layer
activated by the softmax function were added after the convolutional neural networks. The models
returned a list of probabilities that the image corresponded to one of the 165 labels.

**RadImageNet model performance**

To investigate the performance of the AI models trained on RadImageNet, the top-1 accuracy and top-5
accuracy were calculated. The highest probability of all predicted categories was used to calculate the
top-1 accuracy, whereas the top 5 highest probabilities were used to evaluate the top-5 accuracy.

**Image preprocessing (window leveling)**
If window and leveling parameters were set by the annotating radiologist, these were used when transforming the data from DICOM to PNG. Otherwise, a recommended window was used for CT based on the study and reconstruction kernel (for example, a chest CT reconstructed with a lung kernel image had a standard lung window applied) or an auto window generated for MRI images. Detailed CT window and leveling data can be found in Extended Data Table 4.

**Reader Studies**

To further evaluate the RadImageNet models’ performance, we compared model performance to three senior subspecialized radiologists, a neuroradiologist (A.D, clinical experience 12 years), a cardiothoracic radiologist (A.J., clinical experience 10 years), a musculoskeletal radiologist (M.H., clinical experience 10 years). Studies were randomly sampled from the test set to create reader studies in around 2,000 images. If the number of images in a category was greater than 13 images, then a random selection of 14 images from that category was selected for reader studies. Otherwise, all images within that category were presented for reader studies. Radiologist performances were compared to the aforementioned four neural networks trained on RadImageNet according to the expert’s specialty. Individual anatomy accuracy was reported for the comparison of radiologists and RadImageNet models.

**Transfer learning applications**

To compare the performance of convolutional neural networks created from RadImageNet and ImageNet datasets, we used Inception-Res-Net-v2, ResNet50, DenseNet121, and InceptionV3 models on five applications. All of the paired models were structured with the same parameters and layers for direct comparison with respect to pre-trained weight from RadImageNet and ImageNet.

**Bone age prediction on hand and wrist x-rays:** this dataset was obtained from RSNA Pediatric Bone Age Challenge. The dataset was split into 75% training set, 10% validation set, and 15% test set. The mean absolute error was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the linear function were introduced after the last layer of the pre-trained models. A total of 50 epochs were trained, whereas the models with the lowest mean absolute error in such epochs were saved for further evaluation and comparison. A modified Bland-Altman plot for the absolute difference between the ground truth and prediction was used to assess the consistency of mean absolute error and evaluate model performance.

**Pneumonia detection on chest radiographs:** the dataset was acquired from the RSNA Pneumonia Detection Challenge to identify pneumonia on chest radiographs (CXR) from ICU patients. Instead of building deep learning models to localize pneumonia on CXRs, we created a classification model for the detection of pneumonia. The cases provided with bounding box information were considered pneumonia cases, while the subjects with no bounding box information were considered non-pneumonia cases. The dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of
40 epochs were trained, whereas the models with the lowest validation loss in such epochs were saved for further evaluation and comparison.

**Knee ACL tear detection on MRI:** this dataset was requested from the Stanford MRNet dataset for ACL and Meniscus Tear Detection. The original dataset contained all images from knee MRIs with and without ACL tears. To more easily compare results between the RadImageNet and ImageNet derived models, we manually selected the sagittal images containing either a normal or torn ACL (images were selected by T.D., a musculoskeletal radiologist with 12 years of clinical experience). The labels of “tear” or “no-tear” were maintained from the original dataset. A total of 4 studies were excluded due to poor image appearance due to excessive susceptibility and/or motion. Model performance on the modified dataset was conducted using 5-fold cross-validation due to the small size of the dataset. In each fold, the dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer were introduced after the last layer of the pre-trained models. A total of 50 epochs were trained. The models with the lowest validation loss in each epoch were saved for further evaluation and comparison.

**SARS-CoV-2 detection on chest CT:** the dataset was obtained from the China National Center for Bioinformation. Zhang et al. provided key images demonstrating both proven COVID-19 and community-acquired pneumonia where normal chest CT scans were not included. These labels were used to develop the models. We stratified the dataset by patient ID. The dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of 40 epochs were trained. The models with the lowest validation loss in such epochs were saved for further evaluation and comparison.

**Hemorrhage detection on head CT:** this dataset was obtained from the RSNA intracranial hemorrhage detection. Slice level labels (intracranial hemorrhage or no intracranial hemorrhage) were provided for each study. Due to the imbalance of the dataset, the non-hemorrhage images from positive studies were excluded for model development. The dataset was stratified by the patient ID and was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of 30 epochs were trained. The models with the lowest validation loss in such epochs were saved for further evaluation and comparison.

**Gradient weighted class activation mapping**

To understand the model interpretability, we used gradient weighted class activation mapping (Grad-CAM) to visualize where the models make predictions in an image. The Grad-CAM highlights the important regions in an image by using the gradients of the target layer that flows into the final convolutional layer to generate a localization map. For both RadImageNet and ImageNet models, the output layer was the target layer, whereas the conv_7b_ac, conv5_block3_out, relu, and mixed10 were
selected as the final convolutional layer to generate the Grad-CAM for Inception-ResNetv2, ResNet50, DenseNet121, and InceptionV3 networks, respectively.

**Statistical analysis**

The normality of the distribution of predicted bone age on hand and wrist x-rays was confirmed by the Shapiro test. The paired t-test was used to calculate the two-sided P-value for mean absolute error in bone age prediction. The Delong method was used to evaluate the 95% confidence interval of AUROC and to calculate the two-sided P-value for the comparison of RadImageNet and ImageNet models. Statistical significance was defined as a P-value < 0.05. The statistics of AUC comparisons were computed in the pROC package in R. The Shapiro test and paired t-test were performed in the statsmodels package in Python.

**References**

1. Pan, S.J., & Yang, Q. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, 1345–1359 (2010).
2. Deng, J., et al. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255 (2009).
3. Russakovsky, O., et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115, 211–252 (2015).
4. Cheplygina, V. Cats or CAT scans: transfer learning from natural or medical image source datasets? ArXiv, abs/1810.05444 (2018).
5. Cheplygina, V., Bruijne, M.D., & Pluim, J. Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis, 54, 280–296 (2019).
6. Halabi, S., et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology, 290, 498–503 (2019).
7. RSNA Pneumonia Detection Challenge | Kaggle. Kaggle.com. Retrieved 17 July 2019, from https://www.kaggle.com/c/rsna-pneumonia-detection-challenge.
8. Bien, N., et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Medicine, 15 (2018).
9. Zhang, K., et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell, 181, 1423–1433.e11 (2020).
10. Flanders, A., et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. Radiology: Artificial Intelligence, 2(3) (2020).
11. Agrawal, P., Girshick, R., & Malik, J. Analyzing the performance of multilayer neural networks for object recognition. In European conference on computer vision, 329–344. (2014)
12. Simonyan, K., & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556 (2015).

13. He, K., Zhang, X., Ren, S., & Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015 IEEE International Conference on Computer Vision (ICCV), 1026–1034 (2015).

14. He, K., Zhang, X., Ren, S., & Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778 (2016).

15. He, K., Zhang, X., Ren, S., & Sun, J. Identity mappings in deep residual networks. In European conference on computer vision, 630–645 (2016).

16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2818–2826 (2016).

17. Krizhevsky, A., Sutskever, I., & Hinton, G.E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60, 84–90 (2012).

18. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. AAAI (2017).

19. Ioffe, S., & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning, 448–456 (2015).

20. Howard, A.G., et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. ArXiv, abs/1704.04861 (2017).

21. Huang, G., Liu, Z., & Weinberger, K.Q. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2261–2269 (2017).

22. Mei, X., et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nature Medicine, 26, 1224–1228 (2020).

23. Walsh, S., Calandriello, L., Silva, M., & Sverzellati, N. Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study. The Lancet. Respiratory medicine, 6, 837–845 (2018).

24. Esteva, A., et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, 115–118 (2017).

25. Altman, D., & Bland, J. Measurement in Medicine: The Analysis of Method Comparison Studies. The Statistician, 32, 307–317 (1983).

26. Bland, J., & Altman, D. STATISTICAL METHODS FOR ASSESSING AGREEMENT BETWEEN TWO METHODS OF CLINICAL MEASUREMENT. The Lancet, 327, 307–310 (1986).

27. Selvaraju, R.R., et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 IEEE International Conference on Computer Vision (ICCV), 618–626 (2017).

28. Shi, B., et al. Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning. In Medical Imaging 2018: Computer-Aided Diagnosis.
International Society for Optics and Photonics, **10575**, 105752R, (2018).

29. Wong, K., Syeda-Mahmood, T., & Moradi, M. Building medical image classifiers with very limited data using segmentation networks. Medical Image Analysis, **49**, 105–116 (2018).

30. Lei, H., et al. A deeply supervised residual network for HEp-2 cell classification via cross-modal transfer learning. Pattern Recognit., **79**, 290–302 (2018).

31. Tajbakhsh, N. et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Transactions on Medical Imaging, **35**, 1299–1312 (2016).

32. Zhang, R., et al. Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features From Nonmedical Domain. IEEE Journal of Biomedical and Health Informatics, **21**, 41–47 (2017).

33. **Online methods**

34. **Study participants.**

35. The study was approved by the institutional review board of East River Medical Imaging (data provider) and the Icahn School of Medicine at Mount Sinai in New York (data receiver). The institutional review boards waived the requirement for written informed consent for this retrospective study, which evaluated de-identified data and involved no potential risk to patients. To avert any potential breach of confidentiality, no link between the patients, the data provider, and the data receiver was made available. A third party issued a certification of de-identified data transfer from the data provider to the data receiver.

36. We collected a total of 203,341 CT (52,691), MRI (142,422), and ultrasound (8,228) studies from 131,872 patients between 1 January 2005 and 31 January 2020 where patients had diagnostic scans at East River Medical Imaging in New York.

37. **Multimodal and multi-anatomy components**

38. CT exams of the chest, abdomen, and pelvis; MRI exams of the shoulder, knee, ankle, foot, spine, knee, hip and brain; and ultrasound studies of the thyroid, abdomen, and pelvis were collected in the curation of the RadImageNet database. For each study, the radiologist annotated key images with a corresponding diagnosis. To create a reasonable number of classes, the annotations were grouped by pathology and imaging appearance resulting in a total of 11 anatomies from 3 modalities and 165 diagnostic labels. For example, ACL tears and ACL sprains on MRI were combined to a single class - MRI, Knee, ACL injury (see **Extended Data Table 3**).

39. **Normal studies**

40. To better investigate the characteristics of abnormal key images for model development, we queried 8,528 normal studies based on radiology reports for the aforementioned modal and anatomical images. Each normal study was further confirmed by a board-certified radiologist (T.D.). All associated diagnostic sequences and images were included.

41. **RadImageNet model development**
42. To develop the pre-trained models from RadImageNet, we trained 4 different convolutional neural networks without importing weights from existing models, namely, Inception-Res-Net-v2\textsuperscript{16–19}, ResNet50\textsuperscript{13–15}, DenseNet121\textsuperscript{21}, and InceptionV3\textsuperscript{16}. The dataset was split into 75% training set, 10% validation set, and 15% test set stratified by patient ID. The associated images from the same patient were in the same set.

43. Rather than importing the weights from existing models, we randomly initialized the weights to train the individual models. A global average pooling layer, a dropout layer at a rate of 0.4, and the output layer activated by the softmax function were added after the convolutional neural networks. The models returned a list of probabilities that the image corresponded to one of the 165 labels.

44. **RadImageNet model performance**

45. To investigate the performance of the AI models trained on RadImageNet, the top-1 accuracy and top-5 accuracy were calculated. The highest probability of all predicted categories was used to calculate the top-1 accuracy, whereas the top 5 highest probabilities were used to evaluate the top-5 accuracy.

46. **Image preprocessing (window leveling)**

47. If window and leveling parameters were set by the annotating radiologist, these were used when transforming the data from DICOM to PNG. Otherwise, a recommended window was used for CT based on the study and reconstruction kernel (for example, a chest CT reconstructed with a lung kernel image had a standard lung window applied) or an auto window generated for MRI images\textsuperscript{33–36}. Detailed CT window and leveling data can be found in **Extended Data Table 4**.

48. **Reader Studies**

49. To further evaluate the RadImageNet models’ performance, we compared model performance to three senior subspecialized radiologists, a neuroradiologist (A.D, clinical experience 12 years), a cardiothoracic radiologist (A.J., clinical experience 10 years), a musculoskeletal radiologist (M.H., clinical experience 10 years). Studies were randomly sampled from the test set to create reader studies in around 2,000 images. If the number of images in a category was greater than 13 images, then a random selection of 14 images from that category was selected for reader studies. Otherwise, all images within that category were presented for reader studies. Radiologist performances were compared to the aforementioned four neural networks trained on RadImageNet according to the expert’s specialty. Individual anatomy accuracy was reported for the comparison of radiologists and RadImageNet models.

50. **Transfer learning applications**

51. To compare the performance of convolutional neural networks created from RadImageNet and ImageNet datasets, we used Inception-Res-Net-v2, ResNet50, DenseNet121, and InceptionV3 models on five applications. All of the paired models were structured with the same parameters and layers for direct comparison with respect to pre-trained weight from RadImageNet and ImageNet.

52. **Bone age prediction on hand and wrist x-rays**: this dataset was obtained from RSNA Pediatric Bone Age Challenge\textsuperscript{6}. The dataset was split into 75% training set, 10% validation set, and 15% test set. The
mean absolute error was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the linear function were introduced after the last layer of the pre-trained models. A total of 50 epochs were trained, whereas the models with the lowest mean absolute error in such epochs were saved for further evaluation and comparison. A modified Bland-Altman plot for the absolute difference between the ground truth and prediction was used to assess the consistency of mean absolute error and evaluate model performance.

53. **Pneumonia detection on chest radiographs**: the dataset was acquired from the RSNA Pneumonia Detection Challenge\(^7\) to identify pneumonia on chest radiographs (CXRs) from ICU patients. Instead of building deep learning models to localize pneumonia on CXRs, we created a classification model for the detection of pneumonia. The cases provided with bounding box information were considered pneumonia cases, while the subjects with no bounding box information were considered non-pneumonia cases. The dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of 40 epochs were trained, whereas the models with the lowest validation loss in such epochs were saved for further evaluation and comparison.

54. **Knee ACL tear detection on MRI**: this dataset was requested from the Stanford MRNet dataset for ACL and Meniscus Tear Detection\(^8\). The original dataset contained all images from knee MRIs with and without ACL tears. To more easily compare results between the RadImageNet and ImageNet derived models, we manually selected the sagittal images containing either a normal or torn ACL (images were selected by T.D., a musculoskeletal radiologist with 12 years of clinical experience). The labels of “tear” or “no-tear” were maintained from the original dataset. A total of 4 studies were excluded due to poor image appearance due to excessive susceptibility and/or motion. Model performance on the modified dataset was conducted using 5-fold cross-validation due to the small size of the dataset. In each fold, the dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer were introduced after the last layer of the pre-trained models. A total of 50 epochs were trained. The models with the lowest validation loss in each epoch were saved for further evaluation and comparison.

55. **SARS-CoV-2 detection on chest CT**: the dataset was obtained from the China National Center for Bioinformation\(^9\). Zhang et al. provided key images demonstrating both proven COVID-19 and community-acquired pneumonia where normal chest CT scans were not included. These labels were used to develop the models. We stratified the dataset by patient ID. The dataset was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of 40 epochs were trained. The models with the lowest validation loss in such epochs were saved for further evaluation and comparison.
56. **Hemorrhage detection on head CT**: this dataset was obtained from the RSNA intracranial hemorrhage detection\(^\text{10}\). Slice level labels (intracranial hemorrhage or no intracranial hemorrhage) were provided for each study. Due to the imbalance of the dataset, the non-hemorrhage images from positive studies were excluded for model development. The dataset was stratified by the patient ID and was split into 75% training set, 10% validation set, and 15% test set. The binary cross-entropy was selected as the loss function. A global average pooling layer, a dropout layer, and an output layer activated by the softmax function were introduced after the last layer of the pre-trained models. A total of 30 epochs were trained. The models with the lowest validation loss in such epochs were saved for further evaluation and comparison.

57. **Gradient weighted class activation mapping**

58. To understand the model interpretability, we used gradient weighted class activation mapping\(^\text{32}\) (Grad-CAM) to visualize where the models make predictions in an image. The Grad-CAM highlights the important regions in an image by using the gradients of the target layer that flows into the final convolutional layer to generate a localization map. For both RadImageNet and ImageNet models, the output layer was the target layer, whereas the conv\(_7\)b\_ac, conv5\_block3\_out, relu, and mixed10 were selected as the final convolutional layer to generate the Grad-CAM for Inception-Res-Netv2, ResNet50, DenseNet121, and InceptionV3 networks, respectively.

59. **Statistical analysis**

60. The normality of the distribution of predicted bone age on hand and wrist x-rays was confirmed by the Shapiro test\(^\text{37,38}\). The paired t-test\(^\text{39}\) was used to calculate the two-sided \(P\)-value for mean absolute error in bone age prediction. The Delong\(^\text{40}\) method was used to evaluate the 95% confidence interval of AUROC and to calculate the two-sided \(P\)-value for the comparison of RadImageNet and ImageNet models. Statistical significance was defined as a \(P\)-value < 0.05. The statistics of AUC comparisons were computed in the \texttt{pROC}\(^\text{41}\) package in R. The Shapiro test and paired t-test were performed in the \texttt{statsmodels} package in Python.

61. **References**

62. Zatz, L.M. Basic principles of computed tomography scanning. In: T.H. Newton, D.G. Potts, (Eds.), Technical Aspects of Computed Tomography. Mosby, St. Louis, 3853–3876 (1981).

63. Seeram, E. Computed Tomography, Physical Principles, Clinical Applications, and Quality Control, 4th Edition (Elsevier, 2015)

64. Turner, P., & Holdsworth, G. Commentary. CT stroke window settings: an unfortunate misleading misnomer? The British journal of radiology, 84 1008, 1061–6 (2011).

65. Hoang, J., Glastonbury, C., Chen, L., Salvatore, J.K., & Eastwood, J. CT mucosal window settings: a novel approach to evaluating early T-stage head and neck carcinoma. AJR. American journal of roentgenology, 195, 1002–1006 (2010).

66. Shapiro, S., & Wilk, M. An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52, 591–611 (1965).
67. Razali, N.M., & Wah, Y.B. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics, 2, 21–33 (2011).

68. David, H.A., & Gunnink, J.L. The Paired t Test Under Artificial Pairing. The American Statistician, 51, 9–12 (1997).

69. DeLong, E., DeLong, D., & Clarke-Pearson, D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44, 837–45 (1988).

70. Robin, X., et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics, 12, 77–77 (2010).

71. Reporting Summary

72. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Figures
Curation of a medical imaging database RadImageNet, development of pre-trained convolutional neural networks over RadImageNet, and comparison of RadImageNet pre-trained models and ImageNet pre-trained models on multiple medical imaging applications. This study was designed in five pathways. First, key images and associated diagnoses were annotated by radiologists. Second, the images and diagnoses were further grouped by modalities, anatomies, and labels according to their imaging patterns to construct the medical imaging only database RadImageNet. Third, four neural networks as pre-trained models were trained from scratch on the basis of RadImageNet. Fourth, the pre-trained models from RadImageNet and ImageNet were utilized on five medical imaging applications. Finally, comparisons of the pre-trained models and the visualizations of gradient class activation maps were evaluated.

**Figure 2**

Representative images and performance of the pre-trained networks developed from scratch of RadImageNet. a. Overview of RadImageNet modalities and anatomies. RadImageNet was constructed with CT, MRI, and ultrasound images, including CT of the chest, CT of the abdomen and pelvis, MRI of the ankle and foot, MRI of the knee, MRI of the hip, MRI of the shoulder, MRI of the brain, MRI of the spine, MRI of the abdomen and pelvis, ultrasound of the abdomen and pelvis, and ultrasound of the thyroid.
These images represent the diversity and fundamental structure of the RadImageNet database. b. The top-1 and top-5 accuracy of the Inception-ResNet-v2, ResNet50, DenseNet121 and InceptionV3 models on the test set of RadImageNet. The Inception-ResNet-v2, ResNet50, DenseNet121 and InceptionV3 networks achieved top-1 accuracies of 78.86%, 75.04%, 74.47%, and 76.13% respectively while top-5 accuracies were 95.52%, 94.30%, 93.93%, and 94.45% respectively. Extended Data Figure 1 demonstrates the comparison between these four networks and senior radiologists on a dataset randomly sampled from the test set.

Figure 3
Performance of the RadImageNet pre-trained models and ImageNet pre-trained models on five medical imaging applications. a. Comparison of the mean absolute error for bone age prediction on hand and wrist x-rays. b. Comparison of AUROC for pneumonia detection on chest radiographs. c. Comparison of AUROC for ACL tear detection on MRI by 5-fold cross-validation. d. Comparison of AUROC for SARS-CoV-2 detection on chest CT. e. Comparison of AUROC for hemorrhage detection on head CT. The error bar in a indicates the mean absolute deviation. The error bars in b,d, and e indicate the 95% confidence intervals for AUROC scores. The error bars in c indicate the standard deviation among 5-fold cross-validation by
each model. Two-sided P values were calculated by comparing the paired models between RadImageNet and ImageNet. AUC comparisons were evaluated by the DeLong test. The 95% confidence intervals of AUROC were calculated by the DeLong method.

Figure 4

Visualizations of gradient class activation maps from accurately predicted images by all eight pre-trained models. a. A male hand and wrist x-ray image with a validation true bone age of 13 years and 6 months. The RadImageNet models highlighted the distal radial and ulnar, metacarpal, and phalangeal physis. b. A chest radiograph of an ICU patient with pneumonia. The RadImageNet models highlighted the pulmonary infiltrate in the right lung. c. A sagittal T2 fat saturation image from a knee MRI demonstrating an ACL tear. The RadImageNet models highlight the region of the ACL. d. A SARS-COVID-2 positive chest CT image. The RadImageNet models highlight the infiltrates bilaterally. e. A hemorrhage-positive head CT image. The RadImageNet models highlight the areas of intracranial hemorrhage.

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

- supptable1.jpg
- supptable2.jpg
- supptable3a.jpg
- supptable3b.jpg
- supptable4.jpg
- supptable5.jpg
- suppfig1.jpg
- suppfig2.jpg
- suppfig3.jpg
- suppfig4.jpg
- nrreportingsummary.pdf