CLINICAL STUDY

Diagnosing Heart Failure from Chest X-Ray Images Using Deep Learning

Takuya Matsumoto,[1], Satoshi Kodera,[2] MD, Hiroki Shinohara,[2] MD, Hirotaka Ieki,[2] MD, Toshihiro Yamaguchi,[2] MD, Yasutomi Higashikuni,[2] MD, Arihiro Kiyosue,[2] MD, Kaoru Ito,[3] MD, Jiro Ando,[2] MD, Eiki Takimoto,[2] MD, Hiroshi Akazawa,[2] MD, Hiroyuki Morita,[2] MD and Issei Komuro,[2] MD

Summary

The development of deep learning technology has enabled machines to achieve high-level accuracy in interpreting medical images. While many previous studies have examined the detection of pulmonary nodules in chest X-rays using deep learning, the application of this technology to heart failure remains rare. In this paper, we investigated the performance of a deep learning algorithm in terms of diagnosing heart failure using images obtained from chest X-rays. We used 952 chest X-ray images from a labeled database published by the National Institutes of Health. Two cardiologists verified and relabeled a total of 260 "normal" and 378 "heart failure" images, with the remainder being discarded because they had been incorrectly labeled. Data augmentation and transfer learning were used to obtain an accuracy of 82% in diagnosing heart failure using the chest X-ray images. Furthermore, heatmap imaging allowed us to visualize decisions made by the machine. Deep learning can thus help support the diagnosis of heart failure using chest X-ray images.

(Int Heart J 2020; 61: 781-786)

Key words: Artificial intelligence, Transfer learning, CXR

Heart failure is a global pandemic attributed in part to an increasingly-aged population worldwide.[1)] To address it, it is important to build a system of primary care that provides patients with access to general practitioners as well as cardiologists. A chest X-ray is among the most common non-invasive radiological tests providing primary information about a patient's heart condition. However, it is often difficult for general practitioners to make a precise diagnosis of heart failure using chest X-rays. The recent advent of deep learning has improved the accuracy of machines that support the interpretation of medical images.[2,3)] While some previous studies have examined the detection of cardiomegaly on chest X-rays using deep learning,[4)] research on the use of deep learning to diagnose heart failure on chest X-rays is still scarce. The purpose of this study was to verify whether deep learning can help diagnose heart failure using chest X-ray images.

Methods

Dataset: We used 952 images from the chest X-ray database "ChestX-ray8" published by the National Institutes of Health (NIH).[5)] This database contains frontal-view X-ray images with 8 text-mined disease labels. Two cardiologists verified and relabeled the images as "normal" or representative of "heart failure". Some cases were diagnosed differently by the two cardiologists, but they reached a consensus on the label after discussion. The dataset was then randomly split into 3 parts - training, validation, and test sets, in the ratio of approximately 9:1:1, with reference to a previous study.[6)] "Heart failure" was defined as "cardiomegaly or congestion" in this study. According to the most standard criteria, we diagnosed cardiomegaly when the cardiothoracic ratio (CTR) was greater than 50%,[7)] and defined the radiographic presence of pulmonary edema as congestion. All images were resized and trimmed to a size of 1024 × 1024 pixels and a resolution of 72 dpi using Preview version 10.1 (944.6.16.1).

Model: Our model was implemented using Keras version 2.2.5,[8)] which is a highly modular neural network library in Python version 3.6.9. Our model was trained on the Google Colaboratory platform using a graphical processing unit P100.

Figure 1 shows a summary of the model. It consists of two parts, a convolutional part and a classifier. The first

From the [1]School of Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan, [2]Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan and [3]Laboratory for Cardiovascular Genomics and Informatics, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan.

This work was supported by JSPS KAKENHI grant number JP 19K10479.

Address for correspondence: Satoshi Kodera, MD, Department of Cardiovascular Medicine, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. E-mail: kodera@tke.att.ne.jp

Received for publication December 27, 2019. Revised and accepted April 16, 2020.

Released in advance online on J-STAGE July 18, 2020.

doi: 10.1536/ihj.19-714

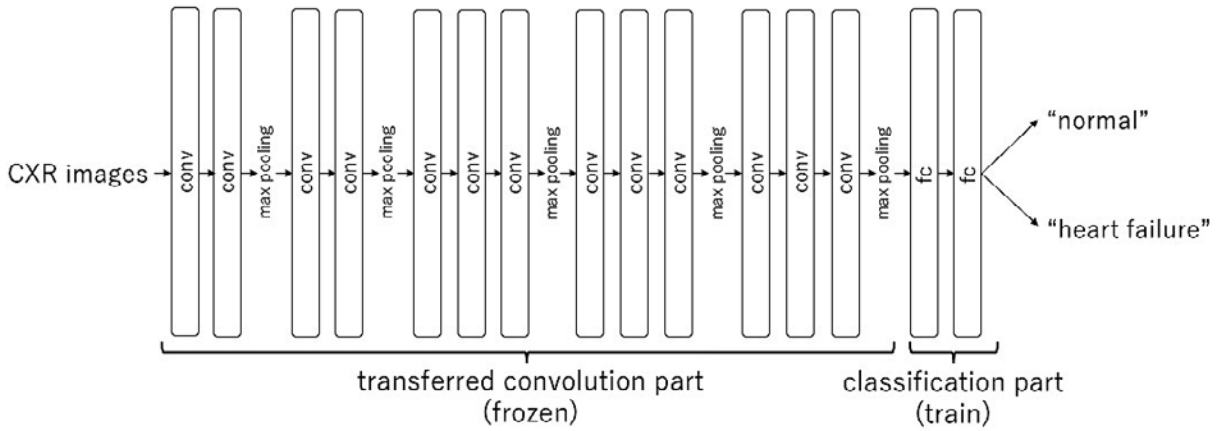
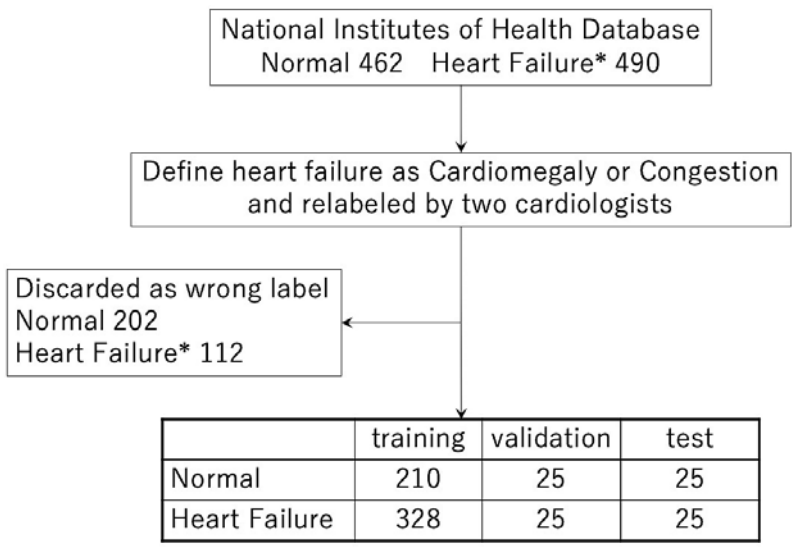


Figure 1. Summary of the model. The first 13 layers, transferred from VGG16, were frozen, and the last 2 layers were trained based on the ChestX-ray8 dataset. conv indicates convolution layers; and fc, fully connected layers.



	training	validation	test
Normal	210	25	25
Heart Failure	328	25	25

Figure 2. Flowchart of labeling and allocation. *In the original ChestX-ray8 dataset, it was labeled cardiomegaly, and not heart failure.

13 layers form the convolutional part, which extracts features of the image, and the last two layers constitute the classifier, which classifies images as "normal" or "heart failure." Given the small dataset used for this study, there was concern over whether the model would be able to extract a sufficient number of features from the images. We thus used the general machine learning approach called transfer learning.[9] The concept of transfer learning is to transfer and use knowledge learned in one task to improve the learning of a different task. In this study, the convolutional part of the model was transferred from the well-known network model VGG16,[10] which was pre-trained on the ImageNet dataset containing 14 197 122 images classified into 1000 categories.[11] The transferred convolutional part was not updated during the training (also referred to as "frozen"), and only the classifier part was trained on the ChestX-ray8 dataset. Enlargement, reduction, translation, and rotation were randomly performed on the images for data augmentation. However, left-right reversal and shear deformation, which are usually performed in deep learning, were not performed in this study, because chest X-rays are asymmetrical, and the inclination of each part is important for interpretation. We applied binary cross-entropy as the loss function, stochastic gradient descent (SGD) as the optimizer, and trained our model for 150 epochs at a learning rate of 0.0001. We used accuracy and log loss of the validation dataset as evaluation metrics, and calculated sensitivity and specificity.

Visualization: Once we had trained the model, we generated gradient-class activation maps (grad-CAMs)[12] using the test set. The Grad-CAMs produced a heatmap that highlighted regions in the image that were important for classification. This method provided an insight into the "black box" nature of the model and helped us to better understand how the model made decisions.

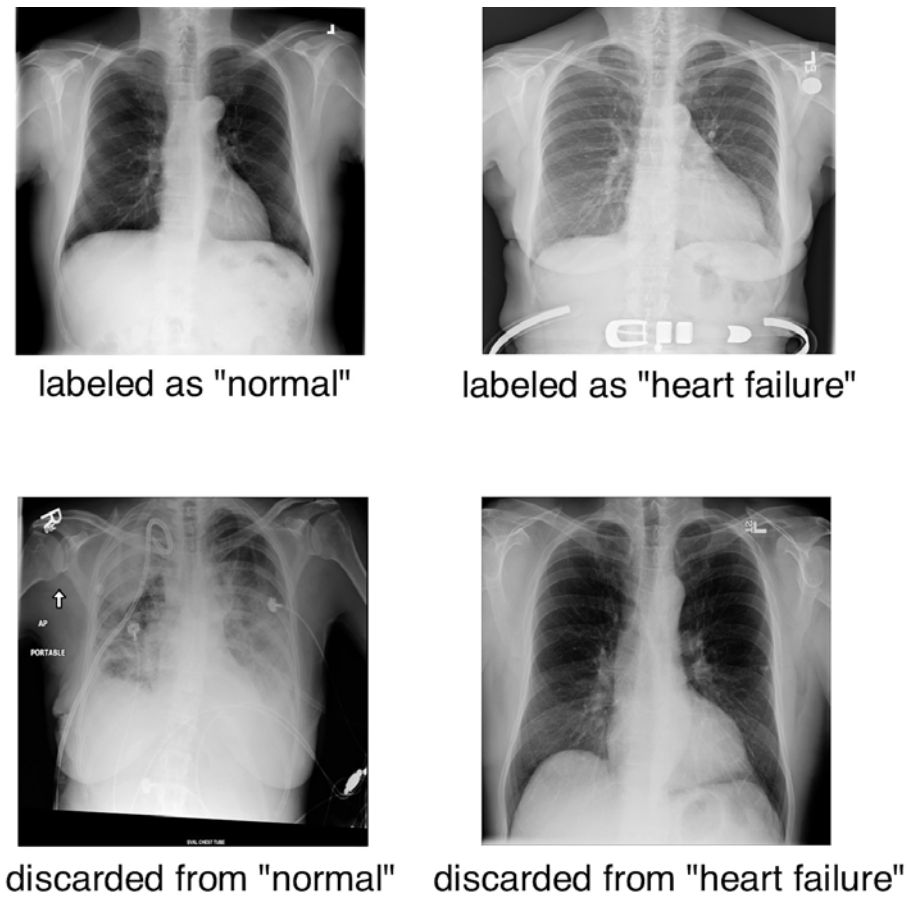


Figure 3. Examples of actual labeling.

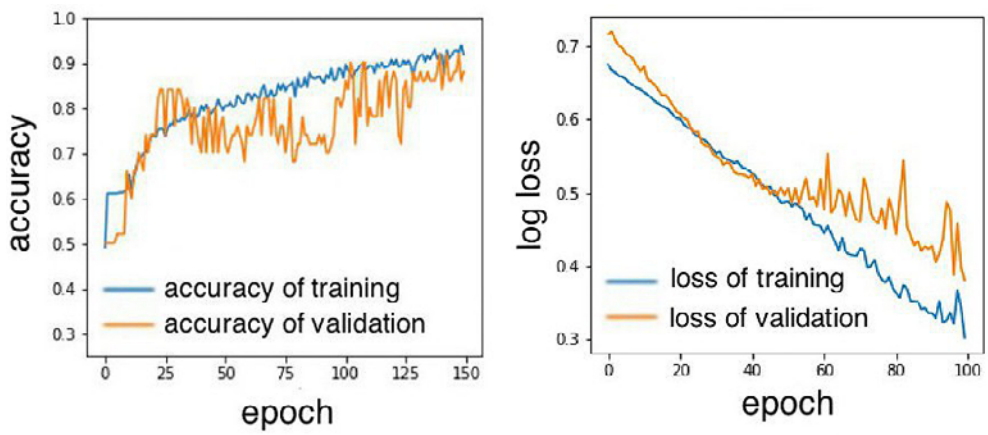


Figure 4. Learning curve of the model. The blue graph represents the training set and the orange graph the validation set. The horizontal axis represents the number of instances of learning and the vertical axis represents accuracy or log loss.

Results

A flowchart of the labeling and allocation is shown in Figure 2. The two cardiologists evaluated the 952 images, and relabeled 260 as "normal" and 378 as "heart failure." For each label, 25 different images were randomly allocated to the validation set and 25 to the test set, with the remaining images allocated to the training set. Examples of the labeling are shown in Figure 3.

The entire training process took 33 minutes. The learning curve of the model is shown in Figure 4. As learning progressed, accuracies on the training set and the

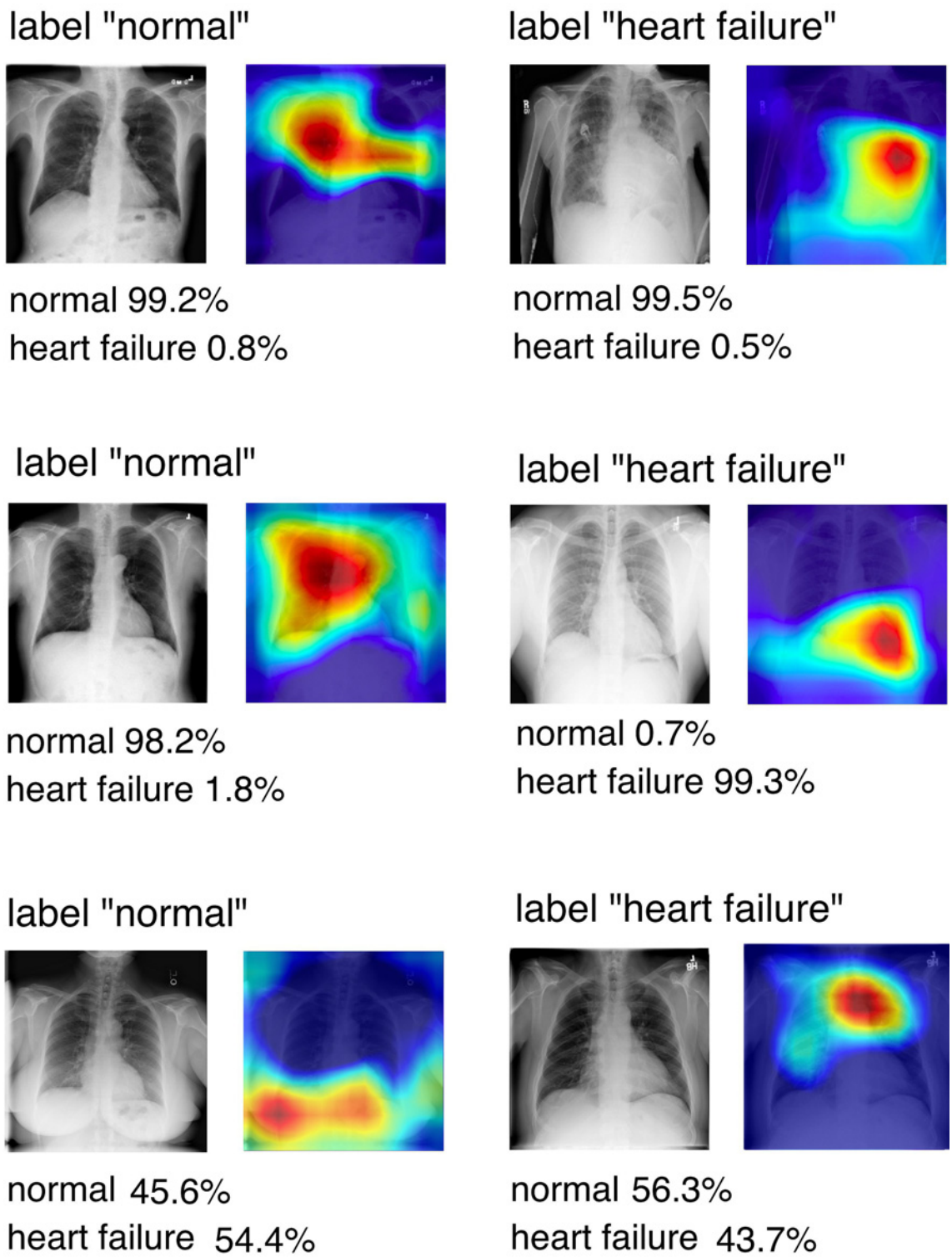


Figure 5. The heatmaps and probabilities of prediction. The original image is on the left and its heatmap is on the right, with its prediction probability written below. The red areas on the heatmaps show important regions, according to which the machine determined the classification.

validation set increased to 93.9% and 92%, respectively, with the log loss on the sets decreasing to 17.6% and 23.8%, respectively. The trained model was able to classify the test data with an accuracy of 82%. The sensitivity and specificity for detection of heart failure on chest-X rays were 75% and 94.4% respectively. Figure 5 shows

randomly selected examples of the prediction probabilities and heatmaps of the chest X-rays from the test set. The red areas on the heatmaps show important regions used by the machine to determine the classification. The prediction probabilities suggest most of the images were diagnosed correctly, and most of the heatmaps focused on the lung field and heart.

Discussion

In this study, we established a model to detect heart failure on chest X-rays by applying deep learning, and obtained an accuracy of 82%. Our model is useful for the diagnosis of heart failure, especially for general practitioners in primary care.

In the initial set-up stage, we used a chest X-ray database open to the public and developed by the NIH, without relabeling it, to obtain an accuracy of diagnosis of 74%. To improve accuracy, we verified the images in the dataset and discarded 202/462 (43.7%) images labeled "normal" and 112/490 images (22.9%) labeled "heart failure" (Figures 2, 3). It is important in deep learning to remove noisy or incomplete images to improve the quality of the learning data. The data relabeling process is a major reason why our method achieved a higher accuracy level than those reported in previous research.[13)]

In addition to the relabeling of the dataset, we used two techniques, data augmentation and transfer learning, to further improve accuracy. The dataset used was small, and this can lead to two major problems in the context of deep learning. One is the risk of "overfitting," which means that the trained network does not fit other and new data. In the case of overfitting, there should be a discrepancy between the learning curves of the training set and the validation set. However, in our case, both learning curves were similar, which suggests that we avoided overfitting through data augmentation and transfer learning. The second major problem arising from the small size of the dataset is that the machine cannot extract a sufficient number of features. To verify that it could detect features needed for the diagnosis of heart failure using chest X-rays, we prepared heatmaps to visualize the process of classification.[12)] The heatmaps showed that the machine focused on the lung field and heart in the images, which means that it extracted features and classified images in the same manner as doctors. The success of our method can thus be attributed to transfer learning.

A previous study using deep learning to detect cardiomegaly on chest X-rays achieved 87.3% accuracy.[4)] Considering the small sample size and the difficulty of detecting heart failure compared to cardiomegaly, this study can be said to have achieved a similar level of accuracy as the previous one. Furthermore, this study is novel, in that it is the first attempt to apply deep learning to the detection of heart failure on chest X-rays.

Limitations: As described above, the first limitation of this study is the small size of the dataset, given that deep learning usually requires thousands of data items to yield high accuracy. Indeed, images with ambiguous radiolucency were prone to being misdiagnosed by the model. However, given the results of the learning curve and the heatmaps, we think we can overcome this limitation to some extent through data augmentation and transfer learning. The second limitation of this study is that we did not differentiate between non-cardiac diseases, such as pneumonia and heart failure, where a differential diagnosis of these diseases is clinically important. In future work, we will try to create a program that can differentiate between a variety of diseases, including non-cardiac diseases. Third, the ChestX-ray8 dataset that we used offers only limited information on the patients, such as gender and age. Because the clinical background of the patients is uncertain, we had to define heart failure as "cardiomegaly or congestion," which can be judged only from the chest X-ray images. Therefore, there is a possibility of including cardiomegaly resulting from causes other than heart failure, such as hemodialysis. In future work, we will try to establish a program with higher clinical utility that can classify images into more detailed categories such as "heart failure with preserved ejection fraction" (HFpEF) or "heart failure with reduced ejection fraction" (HFrEF). This will be done by adding more clinical information, such as data from echocardiograms or laboratory data.

Conclusion

Deep learning is useful for the diagnosis of heart failure on chest X-ray images. Further research is needed to build larger high-quality datasets to create more clinically useful models.

Acknowledgment

We would like to thank Saad Anis, PhD, of Edanz Group for editing a draft of this manuscript.

Disclosure

Conflicts of interest: None.

References

1. Ambrosy AP, Fonarow GC, Butler J, *et al*. The global health and economic burden of hospitalizations for heart failure: Lessons learned from hospitalized heart failure registries. J Am Coll Cardiol 2014; 63: 1123-33.
2. Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. 2017. Available at: https://arxiv.org/abs/1702.05747. Accessed January 30, 2020.
3. Shen D, Wu G, Suk HI. Deep Learning in Medical Image Analysis. Annu Rev Biomed Eng 2017; 19: 221-48.
4. Que Q, Tang Z, Wang R, *et al*. CardioXNet: Automated detection for cardiomegaly based on deep learning. Conf Proc IEEE Eng Med Biol Soc 2018; 2018: 612-5.
5. Wang X, Peng Y, Lu L, *et al*. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. 2017. Available at: https://arxiv.org/abs/1705.02315v4. Accessed December 15, 2019.
6. Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Healthcare. Ann Thorac Surg 2020; 109: 1323-9.
7. Danzer C. The cardiothoracic ratio: an index of cardiac enlargement. The American Journal of the Medical Sciences 1919;

157: 513-54.
8. Homepage of Keras. Available at: https://keras.io/ja/. Accessed February 14, 2020.
9. Pan SJ, Yang Q. A survey on transfer learning. IEEE Transaction on Knowledge and Data Engineering 2010; 22: 1345-59.
10. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. Available at: http://arxiv.org/abs/1409.1556. Accessed December 15, 2019.
11. Homepage of ImageNet. Summary and Statistics. Available at: http://www.image-net.org/about-stats. Accessed December 15, 2019.
12. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradientbased Localization. arXiv preprint 2016; arXiv: 1610.02391.
13. Zhou S, Zhang X, Zhang R. Identifying cardiomegaly in ChestX-ray8 using transfer learning. Stud Health Technol Inform 2019; 264: 482-6.