University of Colorado Denver

HEALTH ADMINISTRATION
RESEARCH CONSORTIUM
Research Brief

Volume 8 Issue 2
Spring 2023

Generalized Radiographic View Identification with Deep Learning

Authors

Xiang Fang, Leah Harris, Wei Zhou, Donglai Huo

Introduction

Radiography or X-Ray imaging is one of the most frequently performed exams in medical imaging. In 2006 alone, about 377 million diagnostic and interventional radiologic examinations were performed in the US, and over 70% were radiographic studies[1]. Although the technology to conduct Radiography procedures has rapidly developed, basic radiographic views (image as seen by the image receptor) did not change much over the years.

There are usually several vital factors to describe a radiographic view: the anatomy (chest, abdominal, foot, etc.), the laterality (left or right), the projections (Antero-Posterior, lateral, oblique), and body position (supine, erect, flexion, extension, etc.). Clinically, when an X-ray order is prescribed, it may contain one or multiple views with clear instructions. For example, an order of "XR Chest 2 View (PA, LAT)" instructs the technologists to take two chest X-ray views (one posterior-anterior view and one lateral view) for radiologists' reading.

Digital imaging, including computed radiography (CR) and digital radiography (DR), has become the Radiology department's dominant equipment. X-ray images taken are transferred and stored electronically in the picture archiving and communication system (PACS) system with digital imaging and communications in medicine (DICOM) format. The DICOM header contains rich information regarding the patient, exam techniques, and other imaging options. Theoretically, it could also contain information regarding radiographic views, such as anatomy, laterality, and projections. However, such information is only available for specific vendors, and it depends on the technologist to select the proper protocol for the workstation before the exam.

A previous study reported[2] that 15% of the exams missed laterality information in the header. In current clinical practice, x-ray technologists put additional lead markers to mark the laterality and the body position, along with their name initials. These additional markers could also be added later from the acquisition workstation digitally. During the marker-adding process, it is also possible that human errors can happen with wrong markers, leading to the wrong side or wrong body part exams. Adverse events of the wrong side, wrong site, wrong procedure, and wrong patient in radiology are significant issues that must be addressed[3].

Machine learning has been successfully applied in image classification for natural images[4], and researchers have shown great success in applying pre-trained machine learning models with transfer learning in radiology[5]. Specifically, task-specific machine learning models generate high accuracy and efficiency in identifying different X-ray views, including laterality[6] and projection[7] [8]. However, how well machine learning could learn to classify different X-ray views, in general, remains unclear. The difficulty levels for identifying the view difference could be low (Chest PA vs. Chest Lateral, Figure 1a and Figure 1b), medium (Knee Lateral vs. Oblique, Figure 1c and Figure 1d), or high (Foot regular vs. Foot Standing, Figure 1e and Figure 1e).

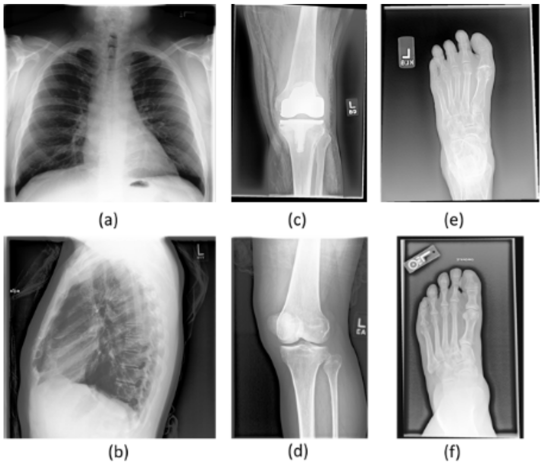

Figure 1: Radiographic View Examples

In this research brief, we explore the possibility of using popular machine learning models with transfer learning to identify radiographs generally. The results will indicate the possibility of building a quality control program based on machine learning models to ensure the correct image content of X-ray views before the patient leaves the exam room. The view identification and model performance are checked at different difficulty levels. Since this is a general view identification and classification task, the model prediction is solely based on the image contents, without training

[1] Mettler, F.A., Jr., et al., Radiologic and nuclear medicine studies in the United States and worldwide: frequency, radiation dose, and comparison with other radiation sources--1950-2007. *Radiology*, 2009. 253(2): p. 520-31.

[2] Filice, R.W. and S.K. Frantz, Effectiveness of Deep Learning Algorithms to Determine Laterality in Radiographs. *J Digit Imaging*, 2019. 32(4): p. 656-664.

[3] Seiden, S.C. and P. Barach, Wrong-side/wrong-site, wrong-procedure, and wrong-patient adverse events: Are they preventable? *Arch Surg,* 2006. 141(9):p. 931-9.

[4] Russakovsky, O., et al., ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 115(3): p. 211-252.

[5] Litjens, G., et al., A survey on deep learning in medical image analysis. *Med Image Anal*, 2017. 42: p. 60-88.

[6] Filice, R.W. and S.K. Frantz, Effectiveness of Deep Learning Algorithms to Determine Laterality in Radiographs. *J Digit Imaging*, 2019. 32(4): p. 656-664.

[7] Yi, P.H., et al., Deep-Learning-Based Semantic Labeling for 2D Mammography and Comparison of Complexity for Machine Learning Tasks. *J Digit Imaging*, 2019. 32(4): p. 565-570.

[8] Rajkomar, A., et al., High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *J Digit Imaging*, 2017. 30(1): p. 95-101.

specifically for the markers or incorporating the DICOM header information.

Materials and Methods

IRB approval was obtained for this HIPAA-compliant retrospective study, and the requirement of written informed consent was waived. An initial database search of the EMR (Electronic medical records) system (Epic Systems Corporation, Verona, WI, USA) in our facility was performed, and all X-ray exams, including both CR and DR exams between 2013/01/01 and 2018/11/01, were included in the initial search. The type of each exam was identified based on our internal exam code. The following exclusion criteria were applied: X-ray exams for children (age < 18); exam types performed less than 200 times. 120 X-ray exam types were identified and included to construct the database. One hundred exams were randomly selected for each exam type, and DICOM images of the selected exams were extracted from our PACS using a customized Python application (Python Software Foundation, https://www.python.org/). Most exams contain more than one X-ray view. For instance, a hand exam may have posterior-anterior, lateral, and oblique views. In total, 15046 images were included in the curated dataset, belonging to 143 different x-ray views.

Classes at Different Levels

Each X-ray image was manually assigned to a "class" or "label" at each of 4 different levels by an experienced board-certified technologist (LH). The labeling results at different levels served as the ground truth to train and validate the proposed machine-learning approach for identifying X-ray views. The labeling convention was defined as given below. An example of this hierarchy for "ankle" is illustrated in Figure 2.

- Level 1: Anatomy Level. Examples: "Abdomen," "Chest," "Foot." In total, 25 classes or labels were assigned at this level.
- Level 2: Laterality Level. For anatomies that have laterality, they were further labeled at this level. Examples: "Foot_L," "Finger_R," "Chest_None." "None" was assigned if there was no appropriate laterality level. In total, 41 classes or labels were assigned at this level.
- Level 3: Projection Level. Examples: "Foot_L_AP," "Head_None_Lat," and "Ankle_R_Lat." Information on projection directions was included in this level. "None" was assigned if there was no appropriate laterality level. In total, 108 classes or labels were assigned at this level.

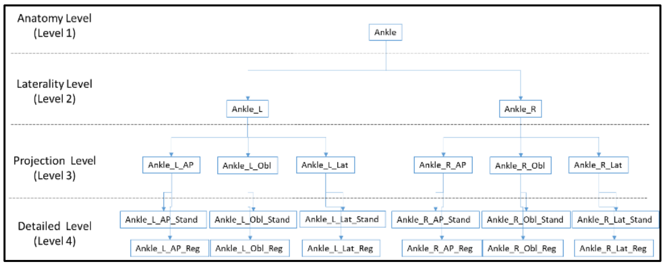

Figure 2: The Hierarchic Tree of The Classification (Label) of Images at Four Different Levels

- Level 4: Detailed Level. This is the most detailed classification level. Examples: "Foot_L_AP_Stand," "Pelvis_None_None_Inlet," CSpine_None_AP_Ext." "None" was assigned if there was no appropriate laterality level. In some clinical scenarios, e.g., "Foot_R_AP_Reg" vs. "Foot_L_AP_Stand," the difference between variable views is so subtle that even experienced technologists have to identify them based on additional information, such as the markers in the image. In total, 143 classes or labels were assigned at this level.

Allowed Labels

Internal ambiguity does exist for the label or class assignment for X-ray views. For example, "wrist" is part of a hand image. Therefore, labeling a "hand" image as "wrist" in certain situations might be acceptable. To account for this issue, a series of "allowed labels" were created and assigned for each label at different levels, and the model performance under "allowed labels" was also evaluated.

Deep Learning Model and Transfer Learning

For GPU acceleration, machine learning models were trained with a Linux-based computer with Keras deep learning library[9] and CUDA 9.1 (Nvdia Corporation, Santa Clara, CA). The computer has an Intel Xeon® processor E5-2660 processor, 16TB hard disk space, 128GB RAM, and 4 NVIDIA GeForce GTX 1080Ti graphics processing units (Nvidia Corporation, Santa Clara, CA).

Inception V3 [10] was selected to perform the classification task in this study. Inception V3 was pre-trained with the ImageNet database[11], and such infrastructure has demonstrated promising image classification capacity in several settings. We used transfer learning to adjust the model parameters to fit into the radiography data. The top layer of the original Inception V3 was removed. A polling layer, a fully connected layer, a dropout layer, and a final activation layer with sigmoid activation were added. Categorical cross-entropy was used as the loss function, and the learning rate was set to 0.0001. The total number of

[9] Chollet, F.C.O. and others, Keras. 2015, \url{https://keras.io}.

[10] Szegedy, C., et al. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[11] Russakovsky, O., et al., ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 115(3): p. 211-252.

Epochs is 40. For each level of the classification task, the same neural network infrastructure was trained and validated, yielding four separate models for four labeling levels.

To construct image datasets for Inception V3, all original images in DICOM format were converted to PNG format. These images were then resized to 299 by 299 pixels, and pixel values were normalized to [0, 1] in the training process. For each class at each level, the images were randomly split into training sets (70%), validation sets (15%), and test sets (15%). Real-time data augmentation was performed by applying the following random image transformations: image rotation (-10 degrees to 10 degrees), image translation (60 pixels in each direction), image shearing (-10 degrees to 10 degrees), and image zooming (0-20%) for each epoch. In addition, the horizontal flip was turned on only for Level 1 classification data augmentation since the image orientation could be an essential feature in classifying the laterality.

Model Output

The model's output on each image is a vector of "scores" corresponding to each class. The "predicted class" was the output class with the highest score. We also record the output classes with the second and third-highest scores for further analysis. We repeated the processing for all four levels, respectively. We generated a "heat map" for each prediction task based on a Gradient-Weighted Class Activation Mapping (Grad-CAM) approach to understand the essential features recognized by the neural network for making classification decisions.

Performance Evaluation

ISLVRC (ImageNet Large Scale Visual Recognition Challenge) classification task[12] using top-5 classification errors has been successfully implanted to evaluate the performance of classification models. In ISLVRC, the "error" is defined as a false negative rate, and "top-5" indicates a certain tolerance level of the prediction errors for multi-label situations. We applied a similar concept to evaluate the performance in this multi-label classification problem, only to report sensitivity (True Positive Rate or 1-False Negative Rate), precision, and accuracy. We report "top-2" and "top-3" sensitivity instead of "top-5" sensitivity as we do not have as many classes as in ISLVRC. We also report "Allowed Label" sensitivity to evaluate the model performance when the "error" is "reasonable" (in the pre-defined class list).

Overall, the following metrics were reported to evaluate the performance of each class:

$$\text{Sensitivity} = TP/(TP+FN)$$

$$\text{Precision} = TP/(TP+FP)$$

For each level, the overall performance is evaluated by:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

where: TP = True positive; FP = False positive; TN = True negative; FN = False negative

In addition, the following performance evaluation metrics were reported for each class:

- Top2_Sensitivity: If the correct label is in the top two prediction choices of the model output, then the prediction is assumed correct.
- Top3_Sensitivity: If the correct label is in the top three prediction choices of the model output, then the prediction is assumed correct.
- AllowedLabel_Sensitivity: If the prediction is one of the "Allowed Labels" for that class, then the prediction is assumed correct.

Results

The overall accuracy for each level, before and after "Allowed Labels" are counted, is shown in Figure 3.

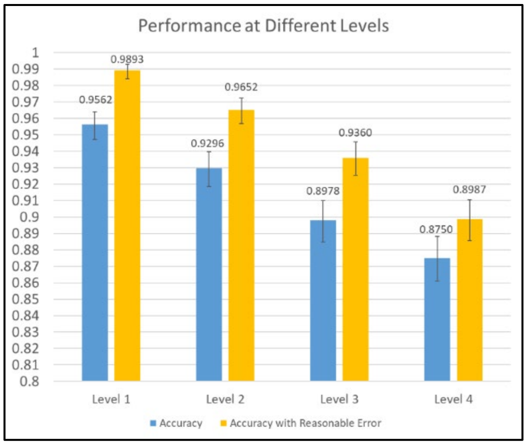

Figure 3: The Overall Accuracy and The Accuracy Allowing "Reasonable Errors" for All Levels
(A 95% CI Accuracy Range Was Reported in The Error Bars in The Images)

As expected, the overall performance of classification models decreases as the classification level increases.

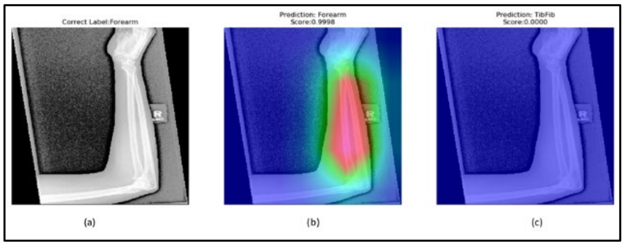

Figure 4: A Typical Level 1 Case With Correct Prediction

[12] Russakovsky, O., et al., ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 115(3): p. 211-252.

Figure 4 shows a representative level 1 case with the original image (Figure 4a) with the label "Forearm," the correct prediction of "Forearm" (Figure 4b) with the highest score of 0.9998 and the prediction (Figure 4c) of "Tibfib" with the second highest score of almost 0. The successful level 1 classification results also corroborated with the peak intensity regions in the heat map overlapped on the original X-ray image (figure 4b).

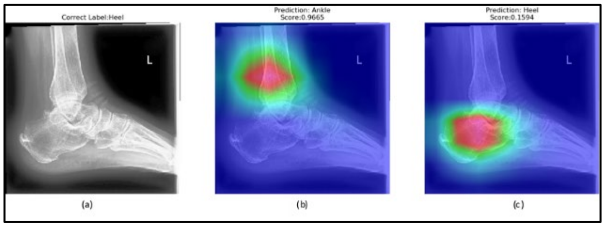

Figure 5: A Typical Level 1 Case With the Wrong Prediction

Figure 5 shows another level 1 case with the original image (Figure 5a) labeled as "Heel." The prediction (Figure 5b) with the highest score of 0.9665 was incorrect, with the label "Ankle." However, the prediction with the second highest score of 0.1594 (Figure 5c) was correct, with the label "Heel." In this case, since "ankle" is in the list of "Allowed Labels" for "heel," when calculating "AllowedLabel_Sensitivity," the prediction is deemed as correct. Similarly, since the correct label "heel" is within the top 2 predictions, it is deemed as correct when calculating "Top2_Sensitivity" and "Top3_Sensitivity".

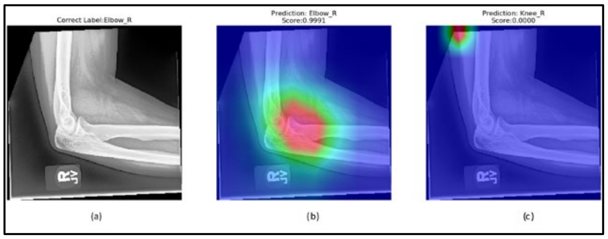

Figure 6: A Typical Level 2 Case With Correct Prediction

Figure 6 shows a level 2 case with the original image (Figure 6a) labeled "Elbow_R." The prediction was correctly made with the highest score of 0.9991 (Figure 6b). Note that the generated heat map (peak intensity region, Figure 6b) correctly pointed to the laterality-related anatomy, indicating that the model's correct judgment (Right Elbow) was based on the anatomical features instead of the marker placed by the technologist.

Discussions

Deep learning has been proven to perform image classification tasks at a comparable performance level with human beings. In the ImageNet challenge, where natural pictures are used for classification, the error rate is around 6%[13]. Without other benchmarks, we would expect a similar performance of classification models in radiography view identification.

In this paper, we investigated the performance of a state-of-the-art machine-learning model in performing classification tasks of X-ray images. The performance was evaluated at different levels, from simple anatomic to subtle, challenging views. We found that the overall classification accuracy decreases with the levels ranging from 0.9562 in level 1 to 0.875 in level 4.

Some data variations inside each class affect the model performance. Some variations come from natural anatomic differences, such as height, weight, sex, and age. Some variations come from the physiological changes due to the symptoms, such as broken bones, pulmonary edema in the chest, or implants in the pelvis. Some other variations come from the definition of the view itself. For example, "Finger_L_Obl_None" contains images for all different fingers. It will always be beneficial to have a larger dataset to improve the generality of the model performance and detect subtle differences between similar classes.

The ambiguity of the X-ray image, or the multi-label issue, obviously affects the classification accuracy. As shown in Figure 5, the "heel" image was classified as "ankle." Since "ankle" is part of the image, it is a "reasonable error." Therefore, we introduced the concept of "allowed labels" to further investigate the impact of the complexity of X-ray images that multi-anatomical structures may be present. If such "reasonable errors" were allowed, the overall accuracy was improved to 0.9893 in Level 1, 0.9652 in Level 2, 0.9360 in Level 3, and 0.8987 in Level 4, as shown in Figure 3. An alternative approach to inspect labeling errors is to examine if the correct label is among the top predictions, even if it is not the first choice. We get a "top3_Sensitivity" rate ranging from 0.9953 in Level 1 to 0.9695 in Level 4, indicating that although the model's first prediction is wrong in some scenarios, the chances of finding the correct label in the first three predictions are still considerably high. Please note that the top choice may be wrong (not "reasonable error").

There are some limitations in this study. Although we tried to construct comprehensive X-ray image datasets, we could not include much data for practical reasons. We did not include X-ray images from children (age < 18) due to limitations from IRB protocol. Some unpopular X-ray views were not included in this study due to the limited number of data points or clinical practice choices. For example, many X-ray views specific to orthopedic surgeons are not included. The classification task could be done differently for this

[13] Russakovsky, O., et al., ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015. 115(3): p. 211-252.

hierarchy classification problem. We train one general model for each level to identify all the views. The downside of this method is that the error may come from different levels. For example, an "Elbow Left AP" image may be wrongly identified as an "Elbow Right AP" view in the Level 3 task. Although we try to classify the image at Level 3, a lower-level error may be carried over. An alternative solution would be "cascade" classification, or to train different models at different levels for specific tasks; for example, after we get the Level 1 classification to know it is a Knee image, we train specific models to different "Knee Left" or "Knee Right" in Level 2. We expect the performance would improve since the number of classes reduced significantly for each classification task. However, the number of models needed for classification will be increased significantly, and the task has to be specified individually in this situation.

It is important to have an effective quality control program in medical imaging. In the current clinical setting, technologists manually assess image quality before the exams are sent to the PACS system. Diagnostic radiography's most common medical error is caused by incorrect patient position, exam type, or laterality. Although manual check works well in many settings, such a process is prone to errors due to variable viewing conditions and depends on personal experience. If any X-ray is inappropriately performed without post-identification of errors, patients will suffer severe consequences, such as misdiagnosis or inaccurate evaluation of treatment[14]. We believe machine learning techniques could be applied to mitigate these medical errors to enable automatic "per exam" quality control. Our results demonstrated that the emerging machine learning approach could automatically check the image contents (X-ray view) purely on the anatomical information. The detection results can be subsequently validated with the reference information from EMR orders to ensure the completion and accuracy of X-ray exams.

Based on the results of this paper, we believe that a clinical quality control system is feasible, which will reliably identify the anatomy, the laterality, and even the projection of the X-ray views. The model performance could be significantly improved if combined with other information. For example, when the marker information is included in detecting the laterality[15], the accuracy is improved to 99%, assuming the marker information is correct. Other information, such as the order information from EMR and DICOM header information from PACS, could also be used to help improve or cross-validate the results from the model. More specific tasks will improve the model performance as well. For example, it was reported[16] to have 100% accuracy if the machine learning algorithms are only trained to differentiate two views (Chest PA or Lateral) or AUC = 1 in differentiating CC vs. MLO view (Mammo views) [17].

If possible, the next step of this project will be to include more data in the datasets, especially images from other institutions. We will also expect a platform that could provide near real-time (a few minutes after PACS upload) feedback to the technologists to reduce the possible wrong side, wrong exam errors in Radiology.

Conclusions

Machine learning methods were developed and applied to classify the X-ray images at Level 1 (Anatomy Level), Level 2 (Laterality Level), Level 3 (Projection Level), and Level 4 (Detailed Level) individually on a comprehensive X-ray image dataset consisting of 15046 different images. Model performance was reported for strict definition and allowing "reasonable errors." Reasonable performance is observed when "reasonable errors" are allowed, indicating the possibility of building a machine learning-based X-ray quality control system.

Acknowledgment

This research brief is adapted from the authors' paper: Fang, X, L Harris, W Zhou, D Huo "Generalized radiographic view identification with deep learning" *Journal of Digital Imaging* 34 (1), 66-74.

This ongoing research project is made possible by the work of faculty at the University of Colorado Business School and the participation of experts from other universities, with the affiliations as follows:

- Xiang Fang Ph.D., Business School, University of Colorado Denver
- Leah Harris, Department of Radiology, University of Colorado Hospital
- Wei Zhou Ph.D., Department of Radiology, University of Colorado School of Medicine, Aurora CO
- Donglai Huo Ph.D., Department of Radiology, University of Colorado School of Medicine, Aurora CO

Cite This Research Brief As:

Fang, X, L Harris, W Zhou, D Huo (2023) "Generalized Radiographic View Identification With Deep Learning" Health Administration Research Consortium Research Brief, University of Colorado Denver, Vol. 8, Issue 2, pp. 1-5.

[14] Seiden, S.C. and P. Barach, Wrong-side/wrong-site, wrong-procedure, and wrong-patient adverse events: Are they preventable? *Arch Surg,* 2006. 141(9): p. 931-9.

[15] Filice, R.W. and S.K. Frantz, Effectiveness of Deep Learning Algorithms to Determine Laterality in Radiographs. *J Digit Imaging*, 2019. 32(4): p. 656-664.

[16] Yi, P.H., et al., Deep-Learning-Based Semantic Labeling for 2D Mammography and Comparison of Complexity for Machine Learning Tasks. *J Digit Imaging*, 2019. 32(4): p. 565-570.

[17] Filice, R.W. and S.K. Frantz, Effectiveness of Deep Learning Algorithms to Determine Laterality in Radiographs. *J Digit Imaging*, 2019. 32(4): p. 656-664.