Title: Impact of Hybrid Supervision Approaches on the Performance of Artificial Intelligence for the Classification of Chest Radiographs

Authors: Ryan Ellis, BS[2]; Erik Ellestad, BA[1,2]; Brett Elicker, MD[1]; Michael D. Hope, MD[1,2]; Duygu Tosun, PhD[1,2]

(1) Department of Radiology and Biomedical Imaging, University of California – San Francisco
(2) San Francisco Veterans Affairs Medical Center

Summary Statement:
Hybrid supervision approaches for the training of artificial intelligence in the classification of chest radiographs can minimize the number of false negative cases. We found that the most effective approach is extra supervision using a balance of annotated normal and abnormal cases applied to a weakly supervised model.

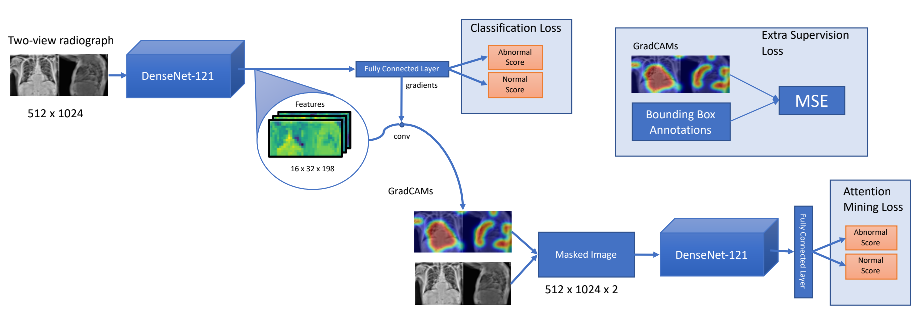
Introduction:
Chest radiography has been the focus of many studies that apply AI in medicine, as it remains one of the most common imaging studies performed. If AI could reliably classify a substantial number of radiographs as normal, and thus free up a radiologist's time to review other studies, the delivery of healthcare could be significantly improved. We evaluated various supervision regimens for the training of AI in the classification of chest radiographs as normal or abnormal.

Methods:
In a retrospective study, 7,000 consecutive two-view chest radiographs obtained from 2012 and 2015 were labeled as normal or abnormal based on clinical reports. The cohort studied is unique in that only two view chest radiographs are included, resulting in a relatively high percentage of normal radiographs (40%). A convolutional neural network (CNN) was trained on this dataset and then evaluated with a subset of 500 radiographs. A subset of 144 radiographs were selected for additional supervision training, where abnormal cases had accompanying bounding boxes locating their pathologies. Each subject's frontal and lateral view radiographs were stitched together side by side, normalized with adaptive histogram equalization, and resized for an input resolution of 512x1024. The CNN architecture was based on DenseNet for feature extraction with an additional attention mechanism to allow for integration of hybrid supervision. Gradient Class Activation Maps (GradCAMs) were estimated by convolving the gradients of the final fully connected layer with the features produced by DenseNet. The GradCAMs produced by the network (16x32 resolution) denoted the salient regions in the image for classification. The performance of CNNs in the automated classification of chest radiographs as normal or abnormal is significantly influenced by the training regimen used. Weak supervision refers to training the model for classification as well as localization by using only the class labels. Extra supervision training was treated as a fine-tuning task, where the weights of the best performing weakly supervised model were used to initialize the weights for training on the annotated subset.

Results:
Training approaches tested were: 1) weak supervision, 2) weak supervision + all annotated cases, 3) weak supervision + only abnormal cases annotated, 4) weak supervision + only normal cases annotated, and 5) weak supervision + balanced normal and abnormal cases annotated. The weakly supervised model achieved an accuracy of 82%, but yielded 75 false negative cases, a sensitivity of 70.0% and a negative predictive value (NPV) of 75.5%. Extra supervision increased NPV at the expense of the false positive rate and overall accuracy. Extra supervision with training using a balance of abnormal and normal radiographs resulted in the greatest increase in NPV (87.2%), improved sensitivity (92.8%), and reduced the number of false negatives by more than fourfold (18 compared to 75 cases).

Conclusion:
Extra supervision using a balance of annotated normal and abnormal cases applied to a weakly supervised model can minimize the number of false negative cases when classifying two-view chest radiographs. Further refinement of such hybrid training approaches for artificial intelligence is warranted to refine models for practical clinical applications.

Figure: Flowchart of the model architecture.



Note: weights are shared for DenseNet-121 and the fully connected layer. MSE = Mean Squared Error.