10.1148/radiol.2018181422

Appendix E1

Details of Prospective Labels, Data Selection Procedure, and Annotation Procedure

Table E1 details the prospective label categories obtained from our institution. For our study, we defined examinations with code "1" ("No [Clinically] Significant Abnormality") as normal and examinations with either "4" ("Abnormal, Not Previously Reported") or "9" ("Critical Finding") as "abnormal." We exclude nondiagnostic (e.g., research) images in category "5" because they are not clinical in nature, and exclude images in categories "2" ("Abnormal, Previously Reported") and "3" ("Significant Interval Change") because they represent images wherein a known abnormality already exists, and are thus not particularly appropriate as inputs to a triage task. Further, it is possible that these images with previously identified abnormalities contain various types of inherent bias in the imaging protocol or oversample particular conditions (e.g., due to the fact that some pathologies require repeated imaging more than others).

We originally obtained 313,719 images where the DICOM study description was recorded as an x-ray of the chest with a prospective label code of "1" ("no [clinically] significant abnormality," considered normal), "4" ("abnormal, not previously reported," considered abnormal), or "9" ("critical finding," considered abnormal). Frontal images were then selected by isolating examinations containing 'AP' or 'PA' in their DICOM study description. Additionally, only studies containing '1 V' (ie, one view) in their DICOM study description were kept to minimize the chance of including examinations for which the frontal view is normal, but the study is summarized as abnormal. Finally, images with duplicate accession numbers were dropped to ensure that no studies were repeated. After these three filtering procedures, 216,431 images remained, of which a random sample of 200,000 was created for model training and validation.

Given practical limits on available radiologist labeling time, we randomly selected a balanced set of 1,000 images (500 normal, 500 abnormal) on which to obtain expert labels. Each of the two expert radiologists was provided a spreadsheet containing image filenames, and recorded their labels in a spreadsheet. A third party then compared the two radiologist-provided spreadsheets, and identified 72 conflicting studies (7.2% of the studies labeled by hand) on which adjudication was required. The two radiologists were then provided with a spreadsheet of the 72 conflicting filenames (without access to any other information), and together arrived at a consensus triage label for each. After removing test set images with incorrect metadata (e.g., lateral scans incorrectly recorded as 'PA') and removing true negatives to arrive at the same 79% abnormal balance observed in the training set, 533 images remained for use in model evaluation.

Note that our more well-defined labeling protocols for the purposes of this study likely factored into the difference between the expert panel and prospective labels; differences among the expert panel centered mainly around the differences in interpreting the definition of support devices. The 43 differences between the expert panel and prospective labels on the 533-image test set can be broadly categorized into labeling errors in the prospective labels on cases that were normal (44%), differences in interpretation of support devices (23%), and the presence of

miscellaneous pathologies such as scarring or atelectasis missed either by the expert panel or the prospective labels (33%).

Details of CNN Training

All CNN models, whether using pretrained or random initialization, were trained on a single Tesla P100 GPU (16 GB) using the Adam optimizer (31) with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$), the default initial learning rate of 0.001, batch size of 72 (the maximum that would fit on the GPU), learning rate decay rate of $\sqrt{0.1}$, the binary cross-entropy loss function, and dropout rate of 0.1 after each dense layer for DenseNet architectures. To standardize the training procedure over different training set sizes, each model was trained using 50,000 batches (equivalent to twenty epochs for the 200,000-sample dataset) and cross-validation on the development set, while learning rate decay was applied after the development accuracy had not improved for 5,000 batches (equivalent to two epochs for the 200,000-sample dataset). We note here that further optimization could have been performed with respect to both architecture and hyperparameter search; while such fine-tuning tasks can and should play an important role in translating academic results into potential deployment, they are not a focus of this study.

In Figure E1, we present sample output from the training procedure for one of our experimental trials using the DenseNet-121 architecture. We observe that both accuracy and loss for train and development sets track each other closely, implying that the training data has not been overfit. Importantly, though training each trial for 20 epochs requires extensive computational resources, we do observe continuing improvements in train and development accuracy well into our training procedure.

Detailed Results from Bag-Of-Visual-Words+Kernelized SVM (BOVW+KSVM) Comparison

We provide details of the computation procedure for our BOVW+KSVM model here. Image descriptors are first computed using the dense Scale Invariant Feature Transform (SIFT) with spatial subdivisions of 2 × 2 and 4 × 4, 200 visual words are defined by k-means clustering using 300 iterations of the Elkan algorithm, and spatial histograms are created for each image in terms of these visual words using a KD-Tree (23). To leverage the power of model nonlinearity, histograms are explicitly pre-transformed using the linear approximation of the nonlinear χ^2 homogeneous kernel with three dimensions as suggested by (22). A support vector machine (SVM) is then trained on the output of this kernel map using five random seeds on the same training and development data that support training of the neural networks described in the bulk of this manuscript. As with the CNN-based models, coarse hyperparameter search was performed over the homogeneity parameter γ and the SVM slack parameter C. The fact that our experiments show that $\gamma = 0.5$ yields best results is in line with the findings of (22).

In Table E2, we report BOVW+KSVM performance averaged over five different random seeds for different training and development set sizes. Note that ROC-AUC values are computed using a score that is the difference between the positive class score and the negative class score, as the maximum of these is used to determine the model prediction. Interestingly, we observe that the relative performance of the CNN and BOVW+KSVM models changes markedly for

different scales of data. For sizes of $O(10^3)$, the BOVW+KSVM outperforms the CNN by 3%, while the representation learning inherent to the CNN results in 7% gain over the BOVW+KSVM with $O(10^4)$ samples. Interestingly, while CNN performance saturates around $O(10^4)$ samples, BOVW+KSVM performance improves substantially up to $O(10^5)$ samples, closing the ROC-AUC gap with the CNN to only 3% at this scale. Improvements in kernel performance with increasing sample size appear to result mostly from improved negative class precision, suggesting that the additional negative examples available as the dataset is scaled are important in driving improved BOVW+KSVM performance. Generally, these findings demonstrate the utility of CNN-based representation learning for achieving high levels of performance on this binary triage task, but also suggest that kernel-based methods may provide utility in certain data regimes, particularly if finer-grained hyperparameter search and higher-dimensional kernel approximations were to be used.

Additional Results

To further elucidate the findings of this study, we have provided examples of additional cases for each classification type (true-positive, true-negative, false-positive, and false-negative) in Figures E2-E5. In Table E3, we also present a breakdown of detection sensitivity by broad pathology class in the test set.

Glossary

To ensure that metrics from machine learning used in this study are accessible to the reader, we define each here for clarity, using common clinical metrics for context when appropriate.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} = Positive\ Predictive\ Value\ (PPV)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} = Sensitivity$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Table E1: Description of Prospective Label Codes

Prospective Label Code	Code Interpretation	Studies Obtained
1	No Significant Abnormality	44,925/216,431 (20.76%)
2	Abnormal, Previously Reported	Not Obtained
3	Significant Interval Change	Not Obtained
4	Abnormal, Not Previously Reported	171,199/216,431 (79.10%)
5	Non-Diagnostic Examination	Not Obtained
9	Critical Finding	305/216,431 (0.14%)

Description of prospective label codes used by our institution. These were assigned by the attending subspecialist radiologist at the time of interpretation.

Table E2: Detailed Performance Metrics for BOVW+KSVM

Train Size	Development Size	Test Accuracy	Precision, ±	Recall, ±	F1, ±	ROC-AUC Score	Gap vs CNN
180,000	20,000	0.88	0.89/0.86	0.98/0.52	0.93/0.65	0.93 (0.90, 0.95)	–0.03
18,000	2,000	0.85	0.88/0.70	0.95/0.48	0.91/0.57	0.87 (0.84, 90)	-0.07
1,800	200	0.85	0.86/0.73	0.96/0.42	0.91/0.53	0.87 (0.84, 0.90)	0.03

Comparison of performance metrics for machine learning baseline using bag-of-visual words features with χ^2 kernel SVM. All reported values are averaged over five random seeds. ± indicates abnormal/normal. Gap versus CNN represents CNN ROC-AUC performance subtracted from BOVW+KSVM performance. Key descriptive statistics are total samples (533), true positives/abnormals (423), and true negatives/normals (110). Bold indicates best observed values (in terms of the BOVW+KSVM model). Best model parameters in each case were $\gamma = 0.5$, C = 1. ROC-AUC refers to area under the receiver operating characteristic curve.

Table E3: Detection Recall (Sensitivity) by Pathology Class

Pathology	Frequency	BOVW+KSVM	AlexNet	ResNet-18	DenseNet-121
Cardiomegaly	10.5%	0.98	0.96	0.96	0.91
Edema	18.2%	1.00	1.00	1.00	0.98
Effusion	30.0%	0.99	0.99	0.99	0.97
Fracture	3.9%	0.90	0.90	0.90	0.71
Opacity	52.5%	0.98	0.97	0.97	0.90
Pneumothorax	5.1%	1.00	1.00	1.00	1.00
Support Device	46.7%	0.96	0.94	0.94	0.89
Normal	20.6%	0.48	0.55	0.55	0.86

Detection recall (sensitivity) by broad pathology class on the test set (533 samples), at the default classifier cutoff value of 0.5. Note that the major difference between the DenseNet-121 and other models is in their relative ability to detect normal cases. Note also that this is a “multi-label” situation wherein each abnormal image can have multiple pathologies; thus, the frequencies do not add up to 100%. BOVW+KSVM indicates bag-of-visual-words + kernelized support vector machine, and other model classes are standard names for neural networks in the literature.