The Interpretation of Deep Learning Based Analysis of Medical Images – An Examination of Methodological and Practical Challenges Using Chest X-ray Data

Steinar Valsson, Ognjen Arandjelović[*]
University of St Andrews
North Haugh
School of Computer Science
St Andrews, KY16 9SX
Fife, Scotland
United Kingdom
[*]ognjen.arandjelovic@gmail.com

Abstract

With the increase in availability of annotated X-ray image data, there has been an accompanying and consequent increase in research on machine learning based, and particularly deep learning based, X-ray image analysis. A major problem with this body of work lies in how newly proposed algorithms are evaluated. Usually, comparative analysis is reduced to the presentation of a single metric, often the area under the receiver operating characteristic (AUROC), which does not provide much clinical value or insight, and thus fails to communicate the applicability of proposed models. In the present paper we address this limitation of previous work by presenting a thorough analysis of a state of the art learning approach, and hence illuminate various weaknesses of similar algorithms in the literature, which have not yet been fully acknowledged and appreciated. Our analysis is performed on the ChestX-ray14 dataset which has 14 lung disease labels and metainfo such as patient age, gender, and the relative X-ray direction. We examine the diagnostic significance of different metrics used in the literature including those proposed by the International Medical Device Regulators Forum, and present qualitative assessment of spatial information learnt by the model. We show that models that have very similar AUROCs can exhibit widely differing clinical applicability. As a result, our work demonstrates the importance of detailed reporting and analysis of performance of machine learning approaches in this field, which is crucial both for progress in the field and the adoption of such models in practice.

Introduction

Chest X-ray is one of the most widely available and easy-to-use medical imaging tools in the diagnostics of lung disease. It is relatively inexpensive compared to other imaging techniques (Medizino 2020; Sistrom and McKay 2005). The quality of the acquisition process and the subsequent analysis are of crucial importance as more extensive tests are often only done for acute cases due to cost or lack of availability. A wrongly interpreted X-ray images can lead to a misdiagnosis with severe consequences.

Advances in the field of Machine Learning (ML) have made it possible, in principle, to automate the interpretation of X-ray images or at least assist in the process. Interpreting X-ray images can be quite challenging to do accurately. Junior doctors generally perform rather poorly on the task (Cheung et al. 2018) and even specialists exhibit significant variability between readings (intra-personal variability) or one another (inter-personal variability) (Satia et al. 2013). The difference in contrast between an anomaly and normal tissue can often be minimal and it is often virtually or literally impossible to distinguish between two conditions from an X-ray alone, and further investigation may be needed. The goal here is to emphasise the importance of interpreting model results by training and evaluating the diagnostic capabilities of a model to diagnose and localise 14 disease labels.

Previous work

As noted earlier, the focus of the present work is not on the technical approach itself, but rather on the issues related to the interpretation of the output of machine learning models trained to analyse X-ray imagery. Hence, since all but without exception, previous work suffers from much the same weaknesses (while differing in 'under the bonnet' technicalities), we illustrate this with a representative example – namely the work of Wang et al. (Wang et al. 2017) – without seeking to survey different learning methodologies in detail. The authors describe a data gathering and labelling process using Natural Language Processing (NLP) from radiology reports gathered from institutional Picture Archiving and Communication Systems (PACS), and train a deep CNN model to predict the label corresponding to an input X-ray image. Their experimental corpus includes labelled X-ray images and meta data such as patient ID, age, sex, and the X-ray view position (VP) (antero-posterior or postero-anterior). A total of 14 disease labels are considered: Atelectasis, Cardiomegaly, Consolidation, Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia and Pneumothorax, with the meaning of each being clear from the label itself. Furthermore, for approximately 1000 images the information on the locality of the label (or indeed, the disease) is provided in the form of a bounding box. The promising results reported by the authors have made this work influential, with a number of follow-up methods having been put forward by others, all bearing conceptual and methodological similarity, such as those by Baltruschat et al. (Baltruschat et al. 2019), Rajpurkar et al. (Rajpurkar et al. 2017), Yao et al. (Yao et al. 2017), Li et al. (Li et al. 2018), and Gündel et al. (Gündel et al. 2019).

In none of the aforementioned work, except for that of Baltruschat et al. (Baltruschat et al. 2019), is there a discussion of the shortcomings to any extent. The scores, usually Area Under Curve (AUC) for the Receiver Operating Characteristic (ROC), or the F1-score, are adopted without any consideration of their clinical significance or insight in what is failing in the proposed method when it does (and failure certainly does occur often enough that it ought to have been discussed).

Quantifying performance using a single numerical measure is certainly an attractive proposition: it is usually easily interpretable, quickly absorbed, and provides unambiguous rank ordering of different approaches. While this approach can be appropriate in some problem contexts, it certainly is not in the case of X-ray image analysis, when nuances in what a model is learning or basing its decisions on, can lead to significant clinical differences, yet leave a simple all-encompassing performance measure unaltered (or virtually so). The present paper sheds additional light on this issue and furthers the understanding of the effectiveness Software as a Medical Device (SaMD) may be measured.

Performance quantification

The Food and Drug Administration (FDA), as a part of the IMDRF, has issued guidelines for SaMDs clinical evaluation where they list a number of evaluation functions they'd like to see reported for clinical validation in future SaMDs. These are specificity, sensitivity, accuracy, and the odds ratio (Center for Devices and Radiological Health and Food And Drug Administration 2018). These metrics can all be computed from the values comprising the confusion matrix – a 2×2 matrix containing the empirical True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) ratios measured by applying a model on test data. Sensitivity, or recall, specificity, accuracy, and F1-score are thus defined as follows:

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F1 = \frac{TP}{TP + \frac{1}{2} \times (FP + FN)}. \tag{4}$$

A high sensitivity entails that there are very few false negatives, while high specificity means that there are few false positives. Accuracy describes the proportion of correct diagnoses but has the downside of not accounting for imbalanced data as it is possible to always predict a class with very few samples as another class with more numerous samples and still have high accuracy. Having both sensitivity and specificity included can therefor indicate how well the SaMD performs in a relatively straightforward way. Accuracy can then be looked at with respect to the other metrics.

The Diagnostic Odds Ratio (DOR), is also often used as a single indicator of diagnostic performance. Its value can range from 0 to infinity, with higher values corresponding to better performance. A value of 1 means that the a positive result is as likely to be the result of a true positive or a true negative, and a score below 1 means that there are more negative results for positive examples of a given class. The DOR is independent of sample prevalence, as apposed to accuracy and a 95% confidence interval can be calculated as

$$\ln(DOR) \pm 1.96 \times SE(\ln(DOR)) \tag{5}$$

where

$$SE(\ln(DOR)) = \sqrt{\frac{1}{TP} + \frac{1}{TN} + \frac{1}{FP} + \frac{1}{FN}} \tag{6}$$

A drawback of the DOR is that it is undefined when the confusion matrix contains zero entries (i.e. in practice, if there are no false positives or false negatives). A commonly used *ad hoc* adjustment applied in such cases is to add 0.5 to all values in the matrix.

Model training

As we noted earlier, the method described by Wang et al. (Wang et al. 2017) is an influential and representative of a whole body of work in the area, and hence herein we adopt it as our baseline. We take a pre-trained network and re-training on the task specific data set – that of X-ray images. A key feature of this process is that the entire network is re-trained and not just the classification layer (which is more common in the literature). In particular, we adopt the 121-layer Dense Convolutional Network (DenseNet) (Huang et al. 2017) pre-trained on the ImageNet corpus (Deng et al. 2009), and re-train on the data made by available by Wang et al., using the same training-validation-test split as the original authors and the Binary Cross-Entropy loss function:

$$\ell(x, y) = \frac{1}{N} \sum_{i=1}^{N} l_i \tag{7}$$

where

$$l_n = -[y_n \cdot \log x_n + (1 - y_n) \cdot \log(1 - x_n)] \tag{8}$$

where x and y are respectively the input and the output vectors, and N is the batch size.

For the localisation of salient image region corresponding to the label, we used Gradient-weighted Class Activation Mapping, or Grad-CAM, based on work by Zhou et al. (Zhou et al. 2016) and further improved on by Selvaraju et al. (Selvaraju et al. 2016). Herein we summarize the process for the reader's benefit. Firstly, an input image is run through the model and the activations from the forward pass on the last convolutional layer saved. Then, back-propagation with respect to a given label is performed and the output gradients from the backwards pass on the same convolutional layer also saved. Next, the gradients are pooled together into a single layer and multiplied by the activations saved earlier. An average pooling is applied to the activation, per feature, leaving a $H \times W$ matrix. A ReLU function is then applied to the matrix, removing all negative feature output and the remaining features then normalized around the maximum entry in the array. At this point the Grad-CAM heatmap has been generated and can be overlayed on top of the original image.

In the end, we compared two models. One that just follows the method mentioned above and another one where the network was modified to use metadata by virtue of two additional binary nodes, corresponding to a patient's gender and the X-ray VP, in the last prediction layer. We'll refer to the first model as the standard model and the second one as the modified model.

Analysis

In line with the primary aims of this work, we started by assessing the different methods' performance using the most widely used metric in the literature, namely the AUROC. Under this metric, the standard and the modified models stand on par with one another, the former achieving the AUROC value of 0.800 and the latter the marginally higher value of 0.806. We note that this is consistent with the previous reports in the literature, with the reported AUROC ranging from 0.745 (see Wang et al. (Wang et al. 2017)) to 0.806 using the method proposed by Baltruschat et al. (Baltruschat et al. 2019). The picture painted by comparing the per label AUROC values, shown in Table 1, is similar: on some labels one model performs somewhat better, on others the other. Weighted by the frequencies of the labels, as we saw earlier, the difference all but disappears.

Both the standard and the modified model achieve nearly identical empirical AUROC scores which, as we noted already, are normally used as the metric for ranking different methods in the field. Thus, superficially, this result suggests that the two methods are performing on par. Yet, in clinical terms, which is really what is of ultimate interest, this is far from the case – a closer look shows that the models actually perform rather differently.

Consider a slightly more nuanced comparison of the methods' performances summarized in Table 2. In terms of specificity and accuracy, the standard model can be seen to be superior. This is significant. For example, the difference of 0.023 in specificity means that out of 1000 patients, 23 can be (correctly) not subjected to further investigation and tests, thereby reducing unnecessary discomfort caused and reducing the financial burden on the health care system. On the other hand, the modified model has a higher recall so it is more likely to detect disease present in patients that have it. The difference in recall of 0.025 means correctly diagnoses 25 more patients in a 1000 than the standard model. To contextualize this, patients and healthcare professionals were willing to exchange 2250 FP diagnoses of colorectal cancer for one additional TP diagnosis (Boone et al. 2013). Similarly, 63% of women found > 500 FPs reasonable per one life saved, and 37% would tolerate 10,000 or more (Schwartz 2000).

Table 1: Comparison of the standard and modified models using the standard AUROC score, per label and overall.

Label	Modified model	Standard model
Atelectasis	0.763	**0.768**
Cardiomegaly	0.875	**0.887**
Consolidation	**0.749**	**0.749**
Edema	**0.846**	0.835
Effusion	0.822	**0.830**
Emphysema	**0.895**	0.873
Fibrosis	0.816	**0.818**
Hernia	**0.937**	0.896
Infiltration	0.694	**0.697**
Mass	**0.820**	0.814
Nodule	**0.747**	0.739
Pleural Thickening	**0.763**	0.762
Pneumonia	**0.714**	0.708
Pneumothorax	**0.840**	0.829
Average	**0.806**	0.800

Table 2: Coarse model comparison.

Model	Specificity	Sensitivity	Accuracy	DOR
Standard	0.741	0.726	0.741	9.56
Modified	0.718	0.751	0.720	10.63

Reflecting on these observations, it is neither correct to say that the methods perform comparably, nor that one is superior to the other. Rather, there are significant differences between the two, and the question which is to be preferred in a specific context is one which demands collaborative consultative effort between teams of clinicians who understand the particular operative environment of interest, and, no less importantly, medical ethicists whose role in the process is still inadequately appreciated.

Understanding data & findings interpretation

A major concern of relevance to the efforts in the development of medical applications of machine learning concerns data used for training and testing algorithms. Notable problems include quality control (both of data itself as well as of its labelling), the clinical relevance and appropriateness of any associated annotations, data balance, and numerous others. Indeed, concerns regarding the ChestX-ray14 cor-

Table 3: Mean model loss dependency on the number of labels per image.

N	0	1	2	3	4	5	6	7	8
Loss	0.055	0.206	0.346	0.491	0.647	0.827	0.956	1.134	1.353
Count	9861	7992	5021	1958	572	152	31	8	1

pus have been raised too. Indeed, their nature mirrors the aforementioned pervasive ones: labelling accuracy (quality control), confounding information (quality control), clinical meaning of labels (quality control & clinical significance), and the usefulness of the labels (clinical significance and appropriateness) (Oakden-Rayner 2017). Consider the following quality control concern: since some pneumothorax images are of patients that have already been treated and who hence have a chest drain, a machine learning algorithm can learn to detect the presence of a drain and thus to correctly label the image, rather than than learning to detect directly the condition itself (a similar issue in an anatomically different context was noted by Tun et al. (William, Arandjelovic, and Caie 2018)). This is illustrated in Figure 1 which shows on the left the original image, with the drain tube indicated, and on the right the learnt class (pneumothorax) activation map.

Another important observation is that an image can have more than one class label associated with it, (e.g. both 'Pneumonia' and 'Infiltration' labels can be associated with the same X-ray image). Using the same loss function used to train the network, we can compute the mean model loss as a function of the number of labels, N, associated with an image (n.b. N ranges from 0 for healthy lungs and goes up to 8, which is the maximum number of labels in this corpus). The loss increases at a linear rate with each additional label (see Table 3), suggesting that the number of labels does not effect the per label accuracy.

Looking at all instances of images with a single label and examining the mean activations across classes reveals a clear bias. An example is illustrated in Table 4. The mean activation for the correct, ground truth label 'Consolidation' is only 0.0842 whereas the mean activation for 'Infiltration' is 0.2724 – a 3.2-fold difference.

This observation is corroborated further by the plot in Figure 2 which shows counts of the number of times each class exhibits among the three highest mean activations for single label images across all classes. 'Infiltration' is the most frequent class in the corpus and for six out the fourteen ground truth labels it exhibits the highest activation mean. In seven cases it is the second most activated class, and in one the third. In other words, it is *always* amongst the top three most activated output classes , regardless of what the true, target label is. The same can be seen for the three other most common classes, namely 'Atelectasis', 'Effusion', and 'Mass'. The frequency of high activations is highly affected by the number of class instances in the corpus.

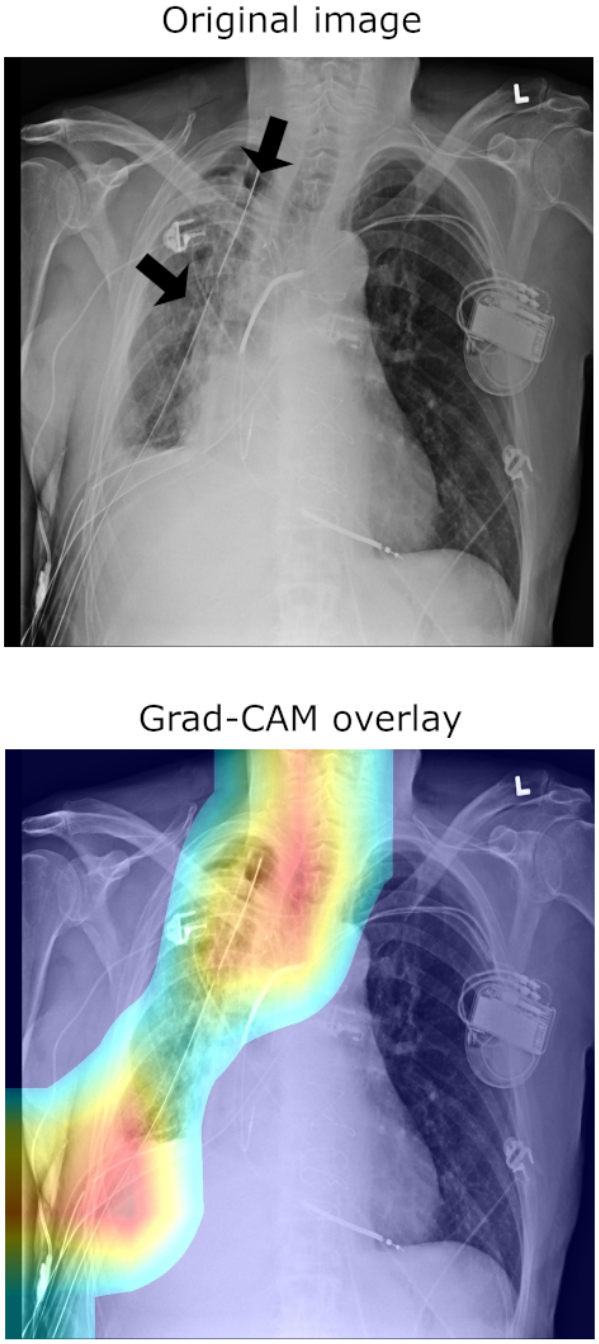

Figure 1: Image labelled as 'Pneumothorax' after it has been treated by a drain tube.

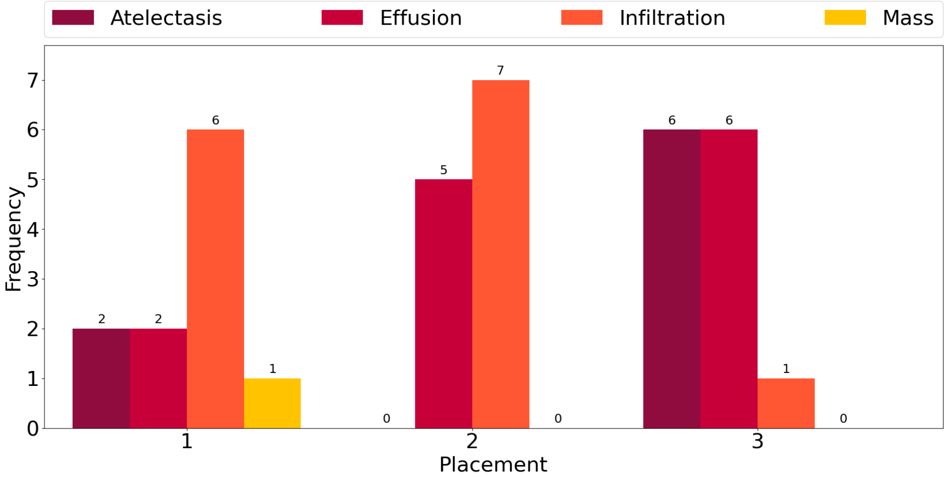

Figure 2: Frequency of highest activation mean.

Table 4: Mean activation of 'Consolidation' for single label images, across different ground truth target labels.

Class	Mean activation
Atelectasis	0.134
Cardiomegaly	0.023
Consolidation	0.084
Edema	0.075
Effusion	0.244
Emphysema	0.011
Fibrosis	0.006
Hernia	< 0.001
Infiltration	0.272
Mass	0.061
Nodule	0.050
Pleural Thickening	0.024
Pneumonia	0.025
Pneumothorax	0.019

Summary and conclusions

Computer software already plays an instrumental role in medicine today and will, without a doubt, play an increasingly important part in future. This observation make it imperative that the evaluation of such software is done rigorously and in a manner which is coherent with its intended clinical application. Indeed, serious concerns have already been raised about the real-world performance of medical software which has previously been reported as successful in at the research stage (Morley, Floridi, and Goldacre 2020). In this paper we looked at this issue in some depth, in the realm of X-ray analysis. We found that in most cases, the analysis of performance reported in research papers is rather poor. In particular, there is an over-reliance on a single, or a few, metric. Worse yet, the clinical significance of these metrics is questionable. Thus, we presented a thorough analysis of a pair of leading machine learning methods for X-ray image based diagnosis. We showed that the widely used standards for performance assessment are overly coarse and often misleading, and that seemingly similarly performing methods can in clinical practice exhibit major differences. Our analysis highlights the subtleties involved in the comprehensive analysis of a machine learning method in this field, potential biases which emerge, as well as often difficult to notice confounding factors. In summary, our work calls for more nuanced evaluation of newly proposed methods and a more thorough reporting of the associated findings, and presents a blueprint for future research efforts.

References

Baltruschat, I. M.; Nickisch, H.; Grass, M.; Knopp, T.; and Saalbach, A. 2019. Comparison of deep learning approaches for multi-label chest X-ray classification. *Scientific Reports*, 9(1): 1–10.

Boone, D.; Mallett, S.; Zhu, S.; Yao, G. L.; Bell, N.; Ghanouni, A.; von Wagner, C.; Taylor, S. A.; Altman, D. G.; Lilford, R.; and Halligan, S. 2013. Patients' & healthcare professionals' values regarding true- & false-positive diagnosis when colorectal cancer screening by CT colonography: discrete choice experiment. *PLoS ONE*, 8(12): e80767.

Center for Devices and Radiological Health; and Food And Drug Administration. 2018. Software as a Medical Device (SAMD): Clinical Evaluation. Technical report, FDA, Center for Devices and Radiological Health.

Cheung, T.; Harianto, H.; Spanger, M.; Young, A.; and Wadhwa, V. 2018. Low accuracy and confidence in chest radiograph interpretation amongst junior doctors and medical students. *Internal Medicine Journal*, 48(7): 864–868.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: a large-scale hierarchical image database. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 248–255.

Gündel, S.; Grbic, S.; Georgescu, B.; Liu, S.; Maier, A.; and Comaniciu, D. 2019. Learning to recognize abnormalities in chest X-rays with location-aware dense networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11401 LNCS: 757–765.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2261–2269.

Li, Z.; Wang, C.; Han, M.; Xue, Y.; Wei, W.; Li, L. J.; and Fei-Fei, L. 2018. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8290–8299.

Medizino. 2020. Buying a new X-ray machine – advice and offers.

Morley, J.; Floridi, L.; and Goldacre, B. 2020. The poor performance of apps assessing skin cancer risk.

Oakden-Rayner, L. 2017. Exploring the ChestXray14 dataset: problems.

Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M. P.; and Ng, A. Y. 2017. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv preprint arXiv:1711.05225*.

Satia, I.; Bashagha, S.; Bibi, A.; Ahmed, R.; Mellor, S.; and Zaman, F. 2013. Assessing the accuracy and certainty in interpreting chest X-rays in the medical division. *Clinical Medicine*, 13(4): 349–352.

Schwartz, L. M. 2000. US women's attitudes to false positive mammography results and detection of ductal carcinoma in situ: cross sectional survey. *The British Medical Journal*, 320(7250): 1635–1640.

Selvaraju, R. R.; Das, A.; Vedantam, R.; Cogswell, M.; Parikh, D.; and Batra, D. 2016. Grad-CAM: Why did you say that? *arXiv preprint arXiv:1611.07450*.

Sistrom, C. L.; and McKay, N. L. 2005. Costs, charges, and revenues for hospital diagnostic imaging procedures: differences by modality and hospital characteristics. *Journal of the American College of Radiology*, 2(6): 511–519.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and Localization of common thorax diseases. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 3462–3471. IEEE.

William, T.; Arandjelovic, O.; and Caie, P. D. 2018. Using machine learning and urine cytology for bladder cancer prescreening and patient stratification. In *Proceedings of the Workshops at the AAAI*, 2–7.

Yao, L.; Poblenz, E.; Dagunts, D.; Covington, B.; Bernard, D.; and Lyman, K. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2921–2929.