The Effects of Active Learning on Computer-Aided Diagnosis in Multi-Disease Prediction of Chest X-rays

Djordi Janssen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
d.janssen@student.utwente.nl

ABSTRACT

The manual annotation of medical images in order to improve computer-aided diagnosis requires a profound level of expertise and is highly time-consuming. Active learning strategies have been widely used in cases where manual annotation is burdensome, as these strategies intend to optimize computers models while reducing the amount of necessary training data. In this research, four active learning strategies have been tested against a passive approach by using a convolutional neural network that has been trained on the National Institutes of Health Chest X-Ray dataset, containing 112,120 multi-label X-rays of the thorax. Additionally, we studied whether the performance of the learning strategies was affected by the amount of (available) training data and examined the impact of the strategies on the individual diseases. We found that active learning had a positive effect on both the area under the ROC curve and the ranking metrics when compared to passive learning, and that the optimal strategy was dependent on the size of the training data. The standard deviations from the average AUC of the individual diseases were higher when using active learning compared to passive learning. In conclusion, this study demonstrated that active learning is beneficial for training computer-aided diagnosis models on the prediction of multiple disease of the thorax.

Keywords

Computer-aided diagnosis, Radiography, Deep learning, Active learning, Convolutional neural network

1. INTRODUCTION

Radiography is one of the most frequently used technique for the diagnosis of diseases in modern medicine [1]. The radiologist is responsible for reviewing and interpreting radiographic images and eventually suggests the most likely diagnosis. Since the interpretation of a radiographic image depends on the knowledge and experience of the radiologist, diagnoses are prone to interobserver variability.

In the search for ways to minimize this variability and to save time, intelligent computer models which are able to make a diagnosis on radiographic images have been developed. In this field of computer-aided diagnosis (CAD), artificially intelligent models are trained on a large number of radiographic images and can learn to distinguish between normal and abnormal medical findings or even be trained to predict diseases. However, the adoption of CAD models in clinical institutions requires a very high accuracy. As with any complex classification problem, the model has to be trained on a large amount of data to achieve the desirable performance. Unfortunately, due to the lack of publicly available medical training data that has high-quality annotations, it is still very difficult to obtain a satisfactory accuracy for these models [12].

A common approach to obtain more training data is by having a human annotator manually label the data. A well-known example for this procedure is the spam filter of a mailbox, which is continuously retrained whenever a user flags a mail as spam. While the spam filter gets constantly updated with new training data, we are not sure whether an extra mail will actually improve its performance - perhaps the algorithm could already classify this mail very well. We call this form of tuition *passive learning*: the computer model is being taught something without knowing why and if the information will improve the algorithm [15]. The reason that a passive learning approach works well for spam filters is because new training data is effortlessly generated by the enormous group of mail users. This also means that it is not that important if some mails are less informative: overall the algorithm has been trained on so much data that it will have seen more than enough informative instances.

In contrast to spam filters, the manual annotation of radiographic images requires a profound level of expertise and is highly time-consuming, which poses a serious financial burden on the optimization of CAD models. Moreover, if we would ask a radiologist to annotate a random set of images and it turns out that the model was already able to predict these scans very well, the model will not benefit significantly and a substantial amount of resources would be wasted. Hence, using a passive learning approach can be very inefficient in situations like this, where limited resources are available. The discipline of *active learning* (AL) tries to overcome this impediment by looking for ways by which a human annotator would label only those images that are the most informative to the model. The fundamental notion in active learning is that the learning system can actively query on which instances it wants to be trained, in the attempt to achieve a high accuracy with as few labeled instances as possible [13]. Figure 1 illustrates the concept of active learning, where the computer model queries X-rays from the pool of unlabeled instances U_L, after which the radiologist annotates these images resulting in a labeled set L.

31th Twente Student Conference on IT July 5th, 2019, Enschede, The Netherlands.

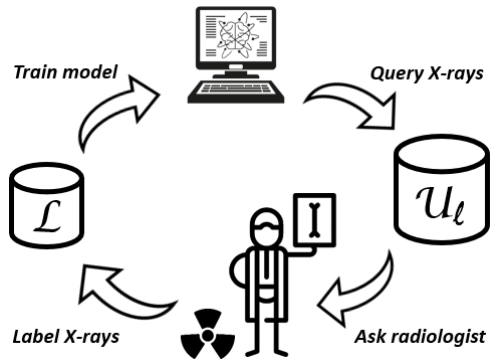

Figure 1. Schematic overview of pool-based active learning in CAD. U_L is the pool of unlabeled instances and L is the set of newly labeled instances.

This research studies the effects of active learning strategies on computer-aided diagnosis for the prediction of multiple diseases on chest X-rays.

2. TECHNICAL BACKGROUND

2.1 Dataset

This research uses the National Institutes of Health Chest X-Ray dataset, which is one of the largest publicly accessible radiology datasets published at the time of writing [17]. The dataset contains 112,120 X-ray images of the thorax with disease labels from 30,805 unique patients. The images in this dataset have been labelled according to 15 categories, of which 14 are medical diseases and 1 class is labeled as 'no findings'. Each X-ray can be diagnosed with 0, 1 or multiple diseases (i.e. multi-label). Figure 2 illustrates X-rays for 8 out of the 14 common thorax diseases that have been annotated in the dataset. Moreover, Figure 5 in the Appendix shows the distribution of the diseases in the dataset and depicts the co-occurrence of these diseases. Lastly, a small number of X-rays have been annotated with bounding boxes, which represent the coordinates of the location at which a disease has been identified.

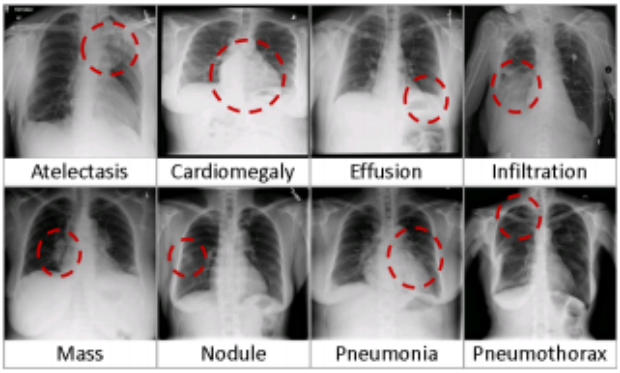

Figure 2. Eight common thoracic diseases observed in chest X-rays [17].

2.2 Convolutional Neural Network

This study focuses on the effects that active learning strategies can have on computer-aided diagnosis. Therefore, the creation of a high performing CAD model is not an objective per se, but should be seen as a positive side-effect. For this reason, the decision was made to use existing work optimized for the dataset instead of attempting to create a model from scratch. The particular model used in this research is a convolutional neural network (CNN) published on Kaggle [10], mainly due to its favorable balance between performance and simplicity. The neural network consists out of 4 hidden layers, uses dropout and pooling, has sigmoid as activation function and has been built using the Keras API [3].

2.3 Active Learning Framework

Several frameworks have been developed that stimulate and facilitate the use of active learning for machine learning applications. This research makes use of a framework called modAL [4], which has been built for Python and is integrated with scikit-learn [11]. While the framework already contains various built-in active learning strategies, its modular design supports extensibility, allowing users to add custom strategies.

3. RESEARCH QUESTIONS

During the course of this research the following questions have been answered:

PRQ. What are the effects of active learning on computer-aided diagnosis for the prediction of multiple diseases in chest radiography?

- **SRQ1.** How does the performance of an active learning approach compare to passive learning?
- **SRQ2.** How are the learning strategies affected by the size of the (available) training data?
- **SRQ3.** How does the effectiveness of the learning strategies vary amongst the individual diseases?

4. RELATED WORK

One of the most prominent works on active learning is the extensive literature survey done by Settles [13]. This paper gives a good introduction to active learning and presents an overview of various well-known query learning frameworks for optimizing machine learning models. Unfortunately, the majority of these strategies do not generalize to multi-label classification problems or are computationally inefficient for the use on larger datasets. As a consequence, these frameworks were not applicable to this research and different strategies have been used, which will be discussed extensively in Section 5.1.

Already various studies have experimented with the use of active learning in the medical domain. A research conducted by Liu [9] examined the effects of active learning applied to gene expression data for cancer classification. The results showed that when using the same amount of training data, active learning could improve the area under the receiver operating characteristics (ROC) curve to 0.81, compared to a value of under 0.50 using passive learning. Moreover, the research concluded that about 82% of the training data could be saved when using active learning. Another study by Hutchison et al. [8] researched the impact of active learning on screening diabetic retinopathy. Similarly, they reported that the amount of training data could be reduced by 80% when using active learning over a passive approach, while retaining an area under the curve of 0.856.

Several other directions have been researched that try to cope with the lack of medically annotated data. One successful example is the extraction of annotations by text

Table 1. Data subset composition for every iteration

	Iteration						
		1	2	3	4	5	6
Data Partition	Total set	5,000	10,000	20,000	50,000	80,000	112,120
	Test set	1,000	2,000	4,000	10,000	16,000	22,424
	Validation set	400	800	1,600	4,000	6,400	8,970
	Initial train set	180	360	720	1,800	2,880	4,036
	X-Pool	3,420	6,840	13,680	34,200	54,720	76,690
	Final trained set	1,206	2,412	4,824	12,060	19,296	27,043

mining of radiology reports. Zech et al. [19] have evaluated several natural language processing techniques that can generate disease labels by extracting information from radiology reports. They reported that the best-performing model had an average sensitivity and specificity across all findings of 90.25% and 91.72%, respectively. Despite these eminent results, it should be noted that this approach diminishes the quality of annotation due to computer intervention.

4.1 Contribution

Whereas earlier studies have examined the impact of active learning on the prediction of a single disease, so far no research had been done on the multi-label classification variant. Hence, this research contributes to the field by studying the effects of active learning on computer-aided diagnosis for the prediction of *multiple* diseases. Additionally, not earlier has the effect of active learning been studied on the NIH Chest X-ray dataset, or any other radiographic dataset with comparable size.

5. METHODS

For efficient and objective results, the concept of active learning has been applied in a slightly different manner. Whereas the conventional active learning approach requests a human annotator to label instances, our dataset already contains the labels for all of the X-rays. Therefore, we pretend that for a portion of the images we do not have the label and we will call this set the X-pool. Correspondingly, querying any number of instances from the X-pool will reveal their labels as if a human annotator was consulted.

5.1 Learning Strategy Selection

The strategies that are used in this research proceed from a multi-label strategy framework proposed by Esuli & Sebastiani [6]. Whilst this framework defines a total of 12 active learning strategies, not every strategy could be tested due to the long training time of the neural network. Accordingly, a selection of 4 strategies has been made, mainly based on the performance presented by [6]. Each of those strategies, as well as random sampling, will be described in detail in the following sections. Before continuing it should be stated that although the neural network conventionally predicts a value on the interval $[0, 1]$, for two of the strategies (i.e. minimal confidence and average confidence) it was necessary to transform this interval to $[-1, 1]$. On this interval -1 denotes that the neural network is certain that a disease is not present (i.e. disease negative) and 1 is used to indicate certainty for disease present (i.e. disease positive). Finally, we will refer to $x^\star_S$ as the most informative image given a strategy S.

5.1.1 Minimal Confidence

The strategy of minimal confidence (MC) regards the most informative image as being the image on which it is most unsure about how to label. For every X-ray on which the model makes a prediction, it identifies the disease which has the lowest absolute probability. Afterwards, this value is compared amongst the entire set of X-rays, and the X-ray with the lowest confidence is selected. Note that we use the absolute value, since it is not important whether the model predicts -0.8 (80% sure of disease negative) or 0.8 (80% sure of disease positive): in both cases the neural network is 80% confident. The mathematical notation for MC is:

$$x^\star_{MC} = \underset{x}{^{argmin}} \mid \psi(P(y^*|x) \mid \tag{1}$$

, where $y^* = \underset{y}{^{argmin}} \mid \psi(P(y|x)) \mid$ stands for the least confident disease on image x and $\psi(z) = 2z - 1$ is the transformation to the [-1, 1] confidence interval.

5.1.2 Average Confidence

Similar to the strategy of minimal confidence, the informativeness measure of average confidence (AC) is based on the uncertainty with which the model predicts diseases. The drawback of MC however, is that this strategy is only concerned with the disease it is most unsure about, thus discarding all information about the other diseases. AC addresses this shortcoming by first taking the average probability of all the disease labels from an image and then compares this value across the entire set of X-rays, regarding the image with lowest value as the most informative instance. This approach has been mathematically formulated in equation 2:

$$x^\star_{AC} = \underset{x}{^{argmin}} \frac{1}{N} \sum_{i=1}^{N} |\psi(P(y_i|x))| \tag{2}$$

, where N symbolizes the total amount of diseases, $\sum_{i=1}^{N} y_i$ ranges over the complete set of disease labels Y such that $y_i \subseteq Y$ and $\psi(z) = 2z - 1$ is again the transformation to the [-1, 1] confidence interval.

5.1.3 Maximum Score

According to strategy of maximum score (MS), the most informative instance is the image for which the model has the highest confidence that (at least) one disease is positive. This strategy relies on the assumption that in supervised learning tasks, it is generally the positive instances rather than the negative ones which are the most useful [6]. The score of a label is equal to its probability on the closed interval [0,1], with 1 having the highest score and 0 the lowest score. For every X-ray, MS identifies the label with the highest score and then compares these scores for all the X-rays, again selecting the image with the best score. Equation 3 expresses the most informative instance according to MS:

$$x^\star_{MS} = \underset{x}{^{argmax}} P(y^*|x) \tag{3}$$

, where $y^* = \underset{y}{argmax} P(y^*|x)$ denotes the label with the highest score for image x.

5.1.4 Average Score

Just as with AC, the strategy of average score (AS) accounts for every disease label instead of only considering the best label for the metric. With AS, the average score is computed for every X-ray and subsequently the image with the highest overall score is selected. Equation 4 shows the most informative instance according to AS, where the symbols used have the same meaning as with AC.

$$x^{\star}_{AS} = \underset{x}{argmax} \frac{1}{N} \sum_{i=1}^{N} P(y_i|x) \tag{4}$$

5.1.5 Random Sampling (RS)

With random sampling, images are drawn from the X-pool using a pseudo-random number generator. This 'strategy' is the default for training of artificially intelligent models.

5.2 Data Transformation

Due to the high quality of the dataset, only small changes had to be made to prepare the data. Firstly, the disease labels of the patients in the dataset were presented in a comma separated format. To use these labels as the target output of our neural network, this format had to be modified into a one-hot encoded vector. The length of this vector equalled to the number of unique diseases, where for every label the number '1' indicates disease positive and '0' indicates disease negative. Subsequently, the distribution of diseases had been plot showing that 'Hernia' is highly underrepresented in the population. Only about 0.2% of the images had been marked with 'Hernia', which is significantly lower than the second lowest diagnosed disease with a positive rate of 1.3%. Consequently, the decision was taken to remove 'Hernia' from the target vector so that the overall results would not be biased. Ultimately, the X-rays were converted into numerical data with a shape of 128x128 pixels using the image data generator from Tensorflow [2].

5.3 Data Partitioning

The next step in the process is to divide the dataset into smaller subsets, each serving a different purpose. The initial dataset has been split according to the Pareto principle into a 80% train- and 20% test division. Hereafter, 10% of the training data has been set apart as validation set. Finally, 5% of the remaining training set was used to initially fit the neural network, while the other 95% constitutes the X-pool from which data will be actively queried depending on the learning strategy. Note that we want to test the performance of our learning strategies on different data set sizes, meaning that N images have been randomly sampled from the complete dataset to form an initial dataset, where N is dependent on the iteration. The exact composition of data samples for each iteration can be found in Table 1, with the full dataset used in iteration 6. The final trained set is the total number of images that the model had been trained on at the end of an iteration.

5.4 Learning Process

Functioning as base model, a neural network was created which has subsequently been fitted on the initial training set belonging to a certain iteration. Afterwards, a copy of the model was made for every learning strategy to ensure that each strategy had the exact same starting point. From this moment onwards, for every strategy the neural network was repetitively fitted onto new batches of images. Each strategy was allowed to query from the X-pool for a total of 20 times. The amount of images requested per query was equal to 1.5% the size of the original X-pool, so that at the end of the learning process 30% of the images were drawn from the X-pool. After a query had been made the images that were used were removed from the X-pool, so that the model could not be fit more than once on the same image. In order to reduce variability, the aforementioned process has been repeated a total of 5 times for each strategy per iteration. The pseudocode of the active learning process is illustrated in Algorithm 1, where the parameters m, s and x represent the model, strategy and X-pool, respectively.

Algorithm 1 Actively teaching the neural network

```
procedure ActiveLearning(m, s, x)
    queries ← 20
    n_images ← int(length(x) × 0.3 ÷ queries)
    i ← 1
    while i ≤ queries do
        images ← s.query(n_images)
        m.train(images)
        x.remove(images)
        i ← i + 1
    end while
end procedure
```

5.5 Evaluation Metrics

In order to evaluate the performance of the neural network given some strategy, several evaluation metrics have been considered. Firstly, a very popular metric that plays a central role in the evaluation of diagnostic tests in the medical world is the receiver operating characteristic (ROC) curve [7]. The ROC curve depicts the relationship between the sensitivity and false positive rate at alternating thresholds. The area under the curve (AUC) of the ROC indicates how good the model is in predicting diseases that are positive while giving as few falls alarms as possible. Additionally, various other multi-label evaluation metrics were used such as Label Ranking Average Precision, Ranking Loss [5] and Coverage Error [16]. These metrics are based on a notion of ranking, where each disease label has been ranked according to its predicted probability. In essence, these metrics measure how good the model is in distinguishing the stochastic order amongst the diseases.

6. RESULTS

Figure 3 illustrates the average AUC per strategy for each iteration, including standard deviations.

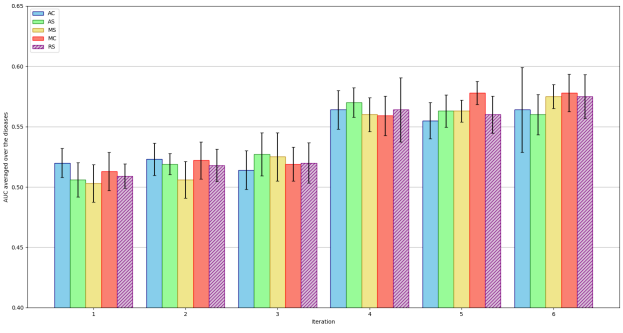

Figure 3. A bar chart showing the AUC of the ROC per learning strategy averaged over all diseases. Error bars represent standard deviations.

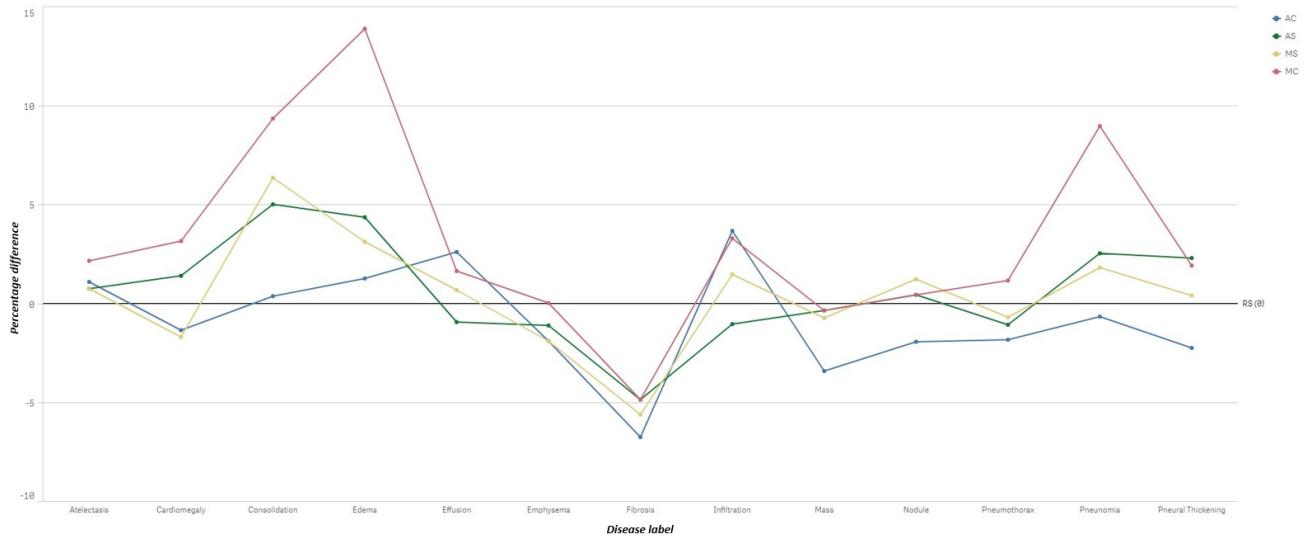

Figure 4. A line graph displaying the relative percentage difference in AUC per disease label compared to the reference (random sampling). The data in this graph originates from iteration 5.

This figure shows that the average AUC ranged from 0.503 in the first iteration to 0.578 in the last iteration using the strategies MS and MC, respectively. Although the differences in AUCs across the various techniques were relatively small, an active learning strategy outperformed the random sampling technique in each iteration. More specifically, AC achieved the highest average AUC during the first two iterations, AS for iterations 3 and 4 and MC during the last two iterations. MS was the only strategy to perform worse than random sampling for the majority of iterations (4 out of 6). The standard deviations were considerably lower for random sampling compared to other strategies during most of the iterations (4 out of 6). Figure 3 shows that maximal AUC has been reached in the fifth iteration using MC (AUC=0.578). Even by adding 7,747 (40,1%) images to the training dataset from the fifth to the sixth iteration, the AUC of the MC in the fifth iteration is still superior when compared to the AUC of the RS in the sixth iteration (AUC=0.575). Although not formally tested, our data suggest that during these iterations 40,1% of data could be saved when using an active learning approach instead of random sampling.

Table 2 displays the performance of each strategy on the ranking metrics averaged over all iterations. The highest values are indicated in bold whereas the lowest values have been underlined. MC scored the best on LRAP, whereas AS scored highest on both LRL and CE. AS and MS score relatively well on all ranking metrics, with only small deviations amongst the two strategies. It is shown that for all the ranking metrics random sampling performs significantly worse than an active learning approach, independent of the strategy chosen. Moreover, from these results we can conclude that the differences between the active learning strategies are relatively small when compared to the differences between the active learning strategies and RS.

Table 2. The scores for LRAP, LRL and CE for all strategies averaged over the iterations.

	MC	AC	MS	AS	RS
LRAP	**0.736**	0.728	0.734	0.732	0.711
LRL	0.181	0.178	0.176	**0.174**	0.221
CE	2.472	2.421	2.376	**2.365**	2.825

To zoom in on the individual diseases, we showed the AUC of the active learning approaches for the individual diseases during the fifth iteration in Figure 4. To compare the active learning strategies with the random sampling method, we set RS as the reference, and thus show the relative AUCs of the four active learning strategies. Overall, the active learning strategies demonstrate consensus on which diseases it can predict better or worse than random sampling. That is, for the disease labels 'Atelectasis', 'Consolidation' and 'Edema' all four strategies significantly outperformed random sampling, whereas in the case of 'Emphysema', 'Fibrosis' and 'Mass' the strategies performed worse or equal to RS. The same trend, although to a lesser extent, also appears in other iterations. The highest relative difference in AUC was observed for the prediction of 'Edema', for which MC scored 13.8% higher than random sampling. Random sampling was between 4.9% and 6.8% better at the prediction of 'Fibrosis' than the active learning strategies.

Table 3. The standard deviations of the average AUC between individual diseases

	MC	AC	MS	AS	RS
1	0.030	0.041	0.042	**0.049**	0.042
2	**0.060**	0.050	0.036	0.053	0.044
3	0.046	0.042	**0.048**	0.042	0.035
4	0.061	0.061	**0.071**	0.064	0.057
5	**0.072**	0.058	0.060	0.057	0.050
6	**0.073**	0.062	0.062	0.057	0.054
Overall	**0.057**	0.052	0.053	0.054	0.047

From Figure 4 we observed that the performance of the learning strategies was dependent on the disease. To get a better insight into the variation between the AUCs of the individual diseases, we computed the standard deviation from the average AUC for each strategy per iteration (Table 3). The highest and lowest values per iteration are indicated in bold and underlined, respectively. From iteration three onwards, RS shows the lowest standard deviations from the average AUC across the individual diseases. In addition, the overall standard deviation is lowest using RS (SD=0,047). On the other hand, MC and MS combined showed the highest standard deviation for the

majority of the iterations (i.e. 5 out of 6).

7. DISCUSSION

We found that the optimal learning strategy was dependent on the size of the (available) training data. Differences in the average AUC of the ROC were relatively small across the different learning strategies. However, in every iteration at least two active strategies were superior to random sampling in terms of AUC. In addition, our study demonstrated that active learning strategies performed better on the multi-label ranking metrics when compared to RS. Interestingly, for some specific diseases active learning strategies outperformed RS in terms of AUC, whereas for other diseases RS scored better on this metric. The standard deviations from the average AUC for the individual diseases were considerably lower when compared to the active learning strategies, indicating lower differences between the predictions of the individual diseases.

The AUCs of the different learning strategies were dependent on the data size on which the neural network had been trained. AC and AS combined had the highest AUC up to the fourth iteration (i.e. 5,000 - 50.000 X-rays), whereas MC had the highest AUC during the last two iterations (i.e. 80.000 - 112,120 X-rays). These observations may be, at least to some extent, explained by the goals of the different learning strategies. To elaborate on this, both AC and AS consider the average informativeness of *all* diseases. When the neural network has not been exposed to a lot of training data yet, the model is uncertain about its predictions on any of the diseases. At this point, AC and AS teach the model highly varying information about a lot of different diseases in order to broaden the knowledge of the model. Once the neural network has been trained on a larger amount of data, the model is more confident about its predictions of the diseases. Hence, it could be more effective for the model to only learn about those particular diseases it is least confident about in order to deepen the knowledge of the model, which is the goal of MC since it only considers the highest uncertainty on a *single* disease.

The active learning strategies scored remarkably higher than random sampling on all of the evaluated ranking metrics. This means that the neural network had a much better idea on the stochastic order of the diseases when trained with active learning strategies compared to using a passive approach. Interestingly, we observed that the differences between active learning and RS were more obvious in the ranking metrics than in the AUC of the ROC curve. An explanation for this is that although the model was not highly confident on its predictions yet as shown by the sub-optimal AUC, it was able to predict relatively well which diseases were more likely than others in comparison to RS as shown by the ranking metrics.

The effect of active learning on the area under the ROC curve seemed to be dependent on the individual diseases. We observed that for many diseases, the active learning strategies either *all* performed considerably better or considerably worse than RS on the AUC. This observation is supported by the standard deviations from the AUC of individual diseases, which were substantially higher for active learning strategies than for RS throughout a majority of the iterations. That is, for the active learning strategies there was a relatively high divergence in the performance amongst the individual diseases when compared to RS. Two possible explanations for this divergence are that the performance on the individual diseases is related with (1) the prevalence of the diseases in the sample population and (2) the number of X-rays selected by the learning strategies that contain these diseases. In order to assess to which extent these explanations are associated with the differences in standard deviations, further research needs to be done.

While active learning has demonstrated to positively affect computer-aided diagnosis both in terms of AUC and ranking metrics, the differences observed with random sampling were lower than was expected from the literature [8], [9]. This discrepancy can be attributed to the limitations of this research, which may have suppressed the effectiveness of the active learning strategies. The three most important limitations have been described in Section 7.1.

7.1 Limitations

1. Due to long computational times of up to eight hours per round only 20 queries could be made for every strategy. This means that the X-rays were queried in large batches, with the size of the batches increasing in every iteration. However, the effectiveness of active learning is highest when as few images as possible are queried at the same time. Even when only two X-rays would be queried at once, it could still turn out that these images are very similar and that the model only learns effectively from one of the images. This problem has been recognized by Settles [14], who also emphasizes that serial querying (one at a time) is even a difficult challenge in practice. It would be highly impractical if not infeasible to ask a radiologist to annotate one X-ray, then retrain the model with this labeled image and go back to the radiologist to annotate the next X-ray. Nonetheless, more queries could be made (i.e. reduce the batch size) to better highlight the advantages of active learning.

2. A second limitation of this research was that, due to time constrains, not all variability in the test results could be taken out. Despite the fact that each strategy has been measured for a total of five rounds per iteration and that the standard deviations between the rounds were not extremely high [8], even more rounds are needed to further reduce this variability.

3. The NIH Chest X-ray dataset that has been used contains impurities regarding the reliability of the target labels. The disease labels of the X-rays have not been directly annotated by a radiologist, but are extracted from the radiology reports using natural language processing. More precisely, the labeling accuracy is estimated to be >90%, which almost certainly means that part of the data has been incorrectly labelled, introducing some bias in the results.

8. CONCLUSION

This research examined the effects of active learning on computer-aided diagnosis for multi-disease prediction on chest X-rays. We found that active learning strategies were superior to a passive approach (i.e. random sampling) in terms of average AUC of the diseases as well as on the evaluated ranking metrics during each of the experiments conducted. Additionally, we concluded that the optimal learning strategy depended on the amount of data on which the neural network had been trained. The divergence in AUC between the predictions of individual diseases was much higher by using an active learning strategy compared to a passive approach, indicating

a more pronounced difference between the individual diseases. In conclusion, the use of active learning indicated to have promising prospects for multi-disease prediction in computer-aided diagnosis of diseases based on chest X-rays. Our results underscore the need of additional studies to assess the optimal effect of active learning strategies in computer-aided diagnosis in clinical practice.

9. FUTURE WORK

There are various aspects to this study which showed that further research on active learning for the prediction of multiple diseases is required. Firstly, the active learning strategies that have been used in this research do not cover all the potential multi-label strategies. Several other strategies have been proposed in the literature and further inquiry would be needed to explore the effects of those strategies on computer-aided diagnosis. Secondly, to further study the effects of active learning on CAD it might be interesting to look into how a distinct computer model can influence these effects. For the scope of this research, only one fixed neural network has been used. However, it could very well be the case that using another algorithm or a different set of hyperparameters would have lead to different results. Yang [18] supports this theory by indicating that correctly tuning the hyperparameters can have a high impact on the performance of the active learners. Thirdly, in order to better understand the performance of active learning in CAD it would be valuable to study the co-occurrence of diseases in the population. For instance, when two diseases D_1 and D_2 frequently occur together in a population, then a learning strategy which actively tries to identify D_1 implicitly also trains the model regularly on D_2. Hence, studying these correlations would bring us a deeper understanding of why active learning strategies work well for some diseases and less well for others. Ultimately, whereas this study has researched the effects of active learning on a CAD model that tries to predict a set of 13 diseases, it would be interesting to see how these strategies perform on a model which predicts a smaller set of diseases. Theoretically seen, the expected benefits of active learning are higher when the probability with which an informative instance can be selected using random sampling is lower. Hence, when the model is trained on a smaller set of diseases the chance is lower that we can find a disease by random selection, so the benefits of active learning might be larger in these situations.

10. ACKNOWLEDGEMENTS

Special thanks go to Meike Nauta and Maurice van Keulen for the active supervision during this research. Furthermore, I would like to thank Doina Bucur and Elena Mocanu for the guidance on academic research and Lando Janssen for proofreading.

11. REFERENCES

[1] Reasons why radiology is so crucial to medical care. https://cmescience.com/reasons-why-radiology-is-so-crucial-to-medical-care, Aug. 2018.

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] F. Chollet et al. Keras. https://keras.io, 2015.

[4] T. Danka and P. Horvath. modAL: A modular active learning framework for python. *Cornell University*, May 2018.

[5] A. Elisseeff and J. Weston. *A kernel method for multi-labelled classification*. MIT Press, Dec 2001.

[6] A. Esuli and F. Sebastiani. Active learning strategies for multi-label text classification. In *Proceedings of the 31st European Conference on Information Retrieval (ECIR 2009)*, 2009.

[7] K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2), 2013.

[8] D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, and et al. *Active Learning for an Efficient Training Strategy of Computer-Aided Diagnosis Systems: Application to Diabetic Retinopathy Screening*, volume 6363. Springer Berlin Heidelberg, 2010.

[9] Y. Liu. Active learning with support vector machine applied to gene expression data for cancer classification. *Journal of Chemical Information and Computer Sciences*, 44(6), Nov 2004.

[10] K. Mader. Chest X-Ray Convolutional Neural Network. https://kaggle.com/kmader/chest-x-ray-cnn, June 2018.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[12] M. I. Razzak, S. Naz, and A. Zaib. Deep learning for medical image processing: Overview, challenges and future. April 2017.

[13] B. Settles. Active Learning Literature Survey. *University of Wisconsin-Madison*, Jan 2010.

[14] B. Settles. From theories to queries: Active learning in practice. In I. Guyon, G. Cawley, G. Dror, V. Lemaire, and A. Statnikov, editors, *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 1–18, Sardinia, Italy, 16 May 2011. PMLR.

[15] S. Thrun. Exploration in active learning. In *Handbook of Brain and Cognitive Science*. MIT Press, Jan 1995.

[16] G. Tsoumakas, I. Katakis, and I. Vlahavas. *Mining Multi-label Data*, pages 667–685. Springer US, 2009.

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[18] Y. Yang. *Towards Practical Active Learning for Classification*. Nov 2018.

[19] J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, and E. K. Oermann. Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*, 287(2):570–580, Jan. 2018.

APPENDIX

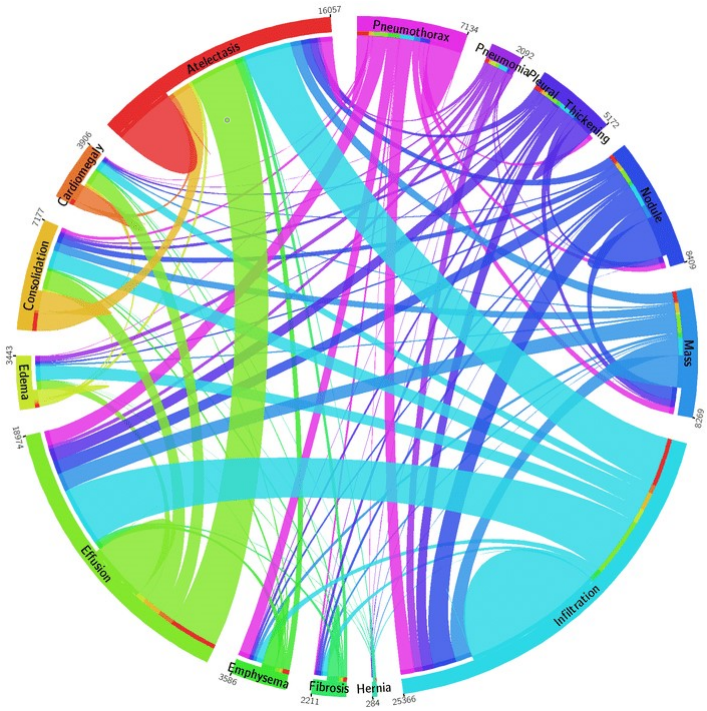

Figure 5. A circular diagram showing the proportions of images with labels from each of the 14 pathology classes and the co-occurrence statistics of the labels [17].