RADBOUD UNIVERSITY

MASTER THESIS

Combining AI with Radiologists: exploring the possibilities in implementation of Computer-Aided Detection

Author:
R.M.W. KLUGE
(S4388267)

Supervisors:
K.G. VAN LEEUWEN
S. SCHALEKAMP
A.P. DE VRIES

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Information Sciences

October, 2020

Radboud University

Abstract

Faculty of Science
Information Sciences

Master of Science

Combining AI with Radiologists: exploring the possibilities in implementation of Computer-Aided Detection

by R.M.W. KLUGE

Since the introduction of deep learning, a new era in radiology has started: the transformation of Computer-Aided Detection (CAD) tools that approaches radiologist-level performance. Compared to traditional CAD, modern deep-learning CAD is applied to various problems in radiology. Less is known about where to implement such tools in the clinical workflow and how this impacts the workflow of radiologists.

As an example use case, we evaluated three (two commercial) CAD systems for pulmonary nodule detection on chest radiographs. We considered three broad strategies: CAD as first reader, CAD as second reader, and CAD concurrently with readers. Even though standalone CAD performs worse than readers, a performance increase of 10% sensitivity or 7% increase in specificity can be achieved depending on the implementation scenario. During CAD as second reader, the possibility of a reader to score an image as 'uncertain' allowed an increase in sensitivity from 69% to 72% using CAD with no further drawbacks. For most other scenarios, the trade-off between specificity - sensitivity versus reading time was observed. Apart from measuring performance, change in workflow and risk were also considered as metrics to capture qualitative results. By comparing scenarios against these metrics, we show the effects of various implementation strategies for CAD on one dataset.

Contents

Chapter 1

Introduction

1.1 AI upcoming – finding its way to the clinic

Since the widespread developments of Artificial Intelligence (AI) in various fields, AI has also arrived at the hospital. Together with advancements in computer vision, the combination of AI and computer vision provide interesting opportunities for the radiology department. Already in 1985, the term computer-aided detection (CAD) for the detection of pulmonary nodules has been introduced (Lampeter, 1985). Current advances in computer vision for medical imaging show that combining AI with radiologists is beneficial (Ardila et al., 2019) (Mayo et al., 2019). AI's future in radiology requires radiologists to acquire new skills and be aware of the pitfalls when introducing AI systems. In previous work, I explored the importance of developing a CAD tool to detect the time-critical symptom *pneumothorax* using deep learning (Kluge, 2020). The next step would be to explore how exactly these tools can be applied in the clinic and what effects these tools have.

According to the model of hierarchical efficacy, Fryback and Thornbury, 1991 shows that technological achievements to improve the diagnostic progress does not necessarily mean that these advancements convert into an improved patient outcome. Thus, a gap exists between the development of tools and increased patient outcome that needs to be filled. The difference between development and implementation of these tools can already be seen during the CAD development life-cycle (Lin and Levary, 1989):

1. Problem definition
2. Design
3. Development
4. Evaluation
5. Implementation (in the clinical setting)
6. Integration (in the workflow of a radiologist)
7. Maintenance

The distinction between development and integration resembles this difference. Nowadays, there has been a lot of research done for the design and development phases of CAD. However, less is known about how to implement such CAD tools in the workflow of the radiologists, and what the difference in consequences are (Nishikawa and Bae, 2018) (Gao et al., 2019) (Liu et al., 2019). Therefore, the focus of this thesis lies in the *integration* phase of the development of CAD tools. Here, the

question remains how we can best fit in the software into a radiologist's workflow. Depending on the integration type, each use case comes with its costs and benefits trade-off. We explore the different implementation use cases of CAD tools and evaluate use cases by comparing them against various metrics.

1.2 Types of integration

Now that we set focus on the integration part, we highlight prior work and explore different implementation strategies. Depending on the type of CAD tool, the clinical workflow implementation differs and might significantly affect the overall performance. Chapter 3.4 demonstrates various factors that influence the performance of CAD tools. It has been shown that when CAD seems beneficial, but at the same time too disruptive to the clinical workflow, the implementation of CAD causes a decrease in adoption rate despite the improved performance (Werth and Ledbetter, 2020). It is important to know beforehand what the goal of the CAD tool would be, before deciding on which use case to consider.

Geras; Mann, and Moy, 2019 and Fujita, 2020 provide a couple of potential use cases for CAD, specifically for mammography:

1. Prevention of missed nodules. Might lead to more false-positives.

2. Classifying benign/malignant nodules. Confirmation of benign/malignant nodules might increase confidence and reduce evaluation time.

3. Concurrent-reader CAD. The prediction of CAD is immediately available to the radiologist. This helps the radiologist in its decision-making process. Another method of concurrent-reading is *interactive* decision support. During *interactive* sessions, radiologists are only shown predictions when clicking on suspected areas. *Interactive* CAD is further explained in Chapter 2.4.

4. Second reader. Here, the radiologist first interprets images without CAD. Then, CAD is used after the reading session as an independent second reader. Upon discrepancies with the radiologist, a second radiologist may be called for. This implementation type increases the sensitivity, but with increased reading time compared to concurrent read (Beyer et al., 2007) (Iussich et al., 2014).

5. First reader. The radiologist is restricted to predictions that CAD did not manage to filter out. This significantly reduces time in, for example, breast cancer screenings. Traditional CAD is shown to be unreliable for the dismissal of normal cases during first read scenarios (Geras; Mann, and Moy, 2019) (Philpotts, 2009). In this case, modern CAD might be beneficial when the sensitivity is shown to be on par with radiologists (Geras; Mann, and Moy, 2019). This does not come without a cost: modern CAD might still filter out nodules otherwise detected by radiologists. Research of Mani et al., 2004 shows no significant increase in performance due to the upper sensitivity limit of CAD, but inter-rater variability decreased.

6. Similar case retrieval. CAD can obtain similar cases, which would give the radiologist more information to support its decision-making process by comparing its decision against decisions from the history (Owais et al., 2019).

As we see, some research regarding implementation strategies exists, but few present direct comparisons of these strategies. Iussich et al., 2014 has shown that, for CT colonography screening, no significant differences between second reader CAD and first reader CAD has been observed. In the case of mammography, the use of concurrent reader CAD significantly increases the sensitivity and specificity of readers versus standalone readers (Kim et al., 2020).

In order to demonstrate what type of CAD implementation affect different outcome metrics, we perform a simulation study where we tackle various scenarios and evaluate them against predefined metrics. As an example, we use three CAD tools (traditional CAD, modern CAD, ensemble CAD) for pulmonary nodule detection on chest radiographs, which we evaluated and compared against the results of radiologists. For radiologists, nodule detection on chest radiographs remains a challenging task, as the number of cancerous nodules initially missed lies between 19-26% (Austin; Romney, and Goldsmith, 1992). We use this research setting as an example task of the use case of CAD, but we expect that the results will generalize towards other CAD tasks as well.

Using this simulation study, we additionally address the issue in which we evaluate whether modern (deep learning) CAD outperforms traditional computer vision CAD systems.

Chapter 2

Background

2.1 Types of CAD

We subdivide CAD tools into various categories:

1. CADe – Computer-aided detection; provides localization of the affected tissue.
2. CADx – Computer-aided diagnosis; provides a classification of the found tissue of CADe.
3. CADq – Computer-aided quantification. Here, CAD quantifies the results (e.g., determining the volume of tumours). These consists of CADe and CADx tools, as when these substages fail, quantification is infeasible.
4. CAST – Computer-aided simple triage. Here, CAD tools are implemented to determine the workload list of the radiologist. This can increase the time of assessing urgent/critical diseases (Goldenberg and Peled, 2011).

As we see, some of these tools need the preceding functionality to achieve the desired results (e.g., quantification requires detection and diagnosis). In this thesis, we focus on the detection of nodules, therefore requiring a CADe tool.

2.2 Traditional CAD vs modern CAD

There are two types of algorithm development approaches seen during the development phase: traditional methods and modern (deep learning) methods. Using traditional methods, experts define rule-based features based on the disease classification, such as the size of found nodules. These traditional methods cause for the introduction of human bias in these algorithms (Gao et al., 2019). The second type are modern (deep learning) CAD methods. These methods do not introduce human feature knowledge, and designs features itself. Due to the rise of *deep learning* methods, computer algorithms and performances changed significantly. In the current state, *deep learning* methods outperform traditional computer vision methods in several fields (Yassin et al., 2018). This significant impact calls for a separation of methods over time. Research has yet to prove that this also the case for diagnostic image analysis (Litjens et al., 2017).

2.2.1 Traditional CAD

In traditional CAD systems, humans are required for feature engineering. Here, features are manually constructed by, i.e. machine learning methods or other image processing techniques, and the algorithm then decides which of these given features are most important.

The effectiveness of using traditional CAD tools is not always apparent. A large study between 90 hospitals shows that traditional CAD tools lead to decreased specificity and do not necessarily improve breast cancer detection rates. At best, traditional CAD tools provide no significant benefit in the clinic (Fenton et al., 2007). Next to that, it is estimated that the use of traditional CAD tools increases the reading time by 20% (Oakden-Rayner, 2019), or even cases where reduced performance leads to an increase in biopsies (Gilbert et al., 2008). Some studies observe a significant difference in usage of a CAD tool to increase the sensitivity, where other studies do not find any difference (Hoop et al., 2010). This could potentially be attributed to the fact that the observers have difficulties in interpreting CAD markings, and neglect true positive cad marks and accept false-positive cad marks.

2.2.2 Modern CAD

Due to recent advances in computation, the popularity in neural networks, more specifically deep learning, kept rising steadily. When *tensorflow* (Abadi et al., 2016) was introduced, the concept of neural networks became accessible to researchers and allowed for fast iterations. Another big difference is the fact that deep learning is task agnostic, meaning that it can learn any specific task as long as the dataset for that task is available. In this way, modern CAD tools are less dependent on human feature engineering and introduce less bias than traditional CAD systems (Gao et al., 2019). The neural networks of modern CAD tools themselves will generate features and evaluate its feature usefulness automatically. The benefit of this is that the model performances are significantly increased, and development time is reduced from years to a few months. Another benefit of deep learning is that new iterations of the algorithm can be developed quicker, as it learns from the input data, and additional input data can lead to better performance. The downside of a deep learning approach is that the exact features are unknown, therefore making this a *black-box* approach. Therefore, with modern CAD tools, we give up explainability for algorithm performance. Explainability helps to generate the trust of a radiologist by providing an insight into the algorithm. It would help the radiologist when the algorithm can explain its prediction of a certain class. For example, when the CAD system classifies benign or malignant nodules, we are interested in *why* exactly this nodule is benign or malignant. Another downside of deep learning approaches is that these methods require large annotated datasets, which are costly to construct. As there is no standardization of datasets, it is hard to compare various methods against each other, as each dataset will contain its own biases and annotation differences. This makes various deep learning methods hard to compare (Kluge, 2020).

In terms of diagnostic imaging, modern CAD tools such as breast cancer detection show a successful false-positive reduction by 69% and decrease reading time by 17% (Mayo et al., 2019). Kim et al., 2020 shows that traditional CAD has problems with distorted or asymmetrical cancers, which modern CAD solves. For lung cancer detection, Ardila et al., 2019 shows a modern CAD system that outperforms all six radiologists and reduces the number of false-positives and false negatives by 11% and 5%, respectively.

2.3 Factors affecting radiologist performance

When we introduce CAD into the workflow of radiologists, there are psychological effects at play. One problem here is that the radiologist makes the final decision regarding the patient's treatment, even though the AI system might perform better than the radiologist. This effect is called the *automation bias* (Geras; Mann, and Moy, 2019). According to Hsu and Hoyt, 2019, five factors influence the impact of CAD tools on radiologist performance:

1. Trust in the CAD system.
2. Human-computer interaction: is the CAD tool intuitive and efficient? E.g. the number of clicks required.
3. Confidence of the radiologist in itself.
4. CAD tool explainability.
5. Previous training and understanding with CAD tools.

These factors come into play when the radiologists directly interact with CAD, i.e. during *CAD as Concurrent Reader* (Chapter 3.3.2)

2.4 Interactive CAD

During *CAD as Concurrent Reader*, we can choose not to present detected nodules to the radiologist immediately like in *prompting* scenarios. This way, we rely on the radiologist's knowledge and only provide additional information given by CAD upon request. In other words, scores remain hidden unless the radiologist asks for a prediction in a particular region. This is called *interactive* CAD (Samulski et al., 2010). Hupse et al., 2013 saw that *interactive* CAD has a significant impact compared to default concurrent-read CAD. It has been shown that *Interactive* CAD was used as a tool to confirm the suspicion of radiologists and as a measure to prevent the radiologist from getting distracted by false-positive markings. However, using this method, reading time increased by an average of 10 seconds per image.

Other research with an interactive approach is the research of Nishikawa and Bae, 2018. This CAD implementation tries to reduce the uncertainty of radiologists by, next to presenting CAD predictions upon request, providing similar cases with known outcomes. This would make CAD more evidence-based and provide radiologists with additional knowledge of historical decisions (Muramatsu et al., 2010). As a final next step, Nishikawa and Bae, 2018 suggests allowing CAD to intervene when radiologists wrongfully miss a nodule. This time, location information would remain hidden, and only the suggestion to re-evaluate the image is given.

The last problem with using CAD as an interactive format are some ethical problems: nodules can go undetected when radiologists do not query the particular location. Geras; Mann, and Moy, 2019 shows that hybrid systems (both prompting and interactive CAD) can be used to increase reader performance significantly. The most confident prediction areas are then given in advance, whereas the rest of the image can be queried interactively.

Chapter 3

Methods

To evaluate our research question, we applied CAD tools for nodule detection to a dataset containing solitary nodules and normal images. Then, we outline the CAD systems used, specify the implementation types and finally describe the metrics used to evaluate the results.

3.1 Study population

The foundation of this study is built on the study of Schalekamp et al., 2014b. We used the same dataset and reader results to compare radiologist performance and scenarios against various CAD systems. The existing dataset consists of 300 posterior-anterior (PA) and lateral chest radiographs from four institutions (111 with solitary nodules, 189 normal images). All subjects were 40 years or older.

For all cases, an expert radiologist and clinical researcher established the reference standard using the chest radiographs and corresponding CT. Nodules needed to be visible on the PA radiographs; were < 30 mm in diameter, and were classified into four categories:

1. well visible (category 1)
2. moderately subtle (category 2)
3. subtle (category 3)
4. very subtle (category 4)

An overview of the distribution of nodule sizes can be seen in Figure 3.1.

Using an annotation platform, such as *grand-challenge* [1], 12 readers without previous clinical experience with CAD scored the chest radiographs for the presence of nodules between 0-100 based on their prediction confidence (0, no nodule; 100, definitely a nodule). Then, the performance of these readers for three scenarios was evaluated: readers standalone as a baseline, readers with CAD, and readers with interactive CAD (iCAD).

Split by nodule size To further investigate the effectiveness of the CADs, we evaluated the performance on subsets of the data split by nodule size. This way, we gained insight into how CAD behaves for small or larger nodules. We split nodule sizes into two categories: small (7-15mm) and large (15-36mm). Each split consisted of all normal images plus a corresponding positive nodule class small/large nodules.

[1] https://grand-challenge.org/

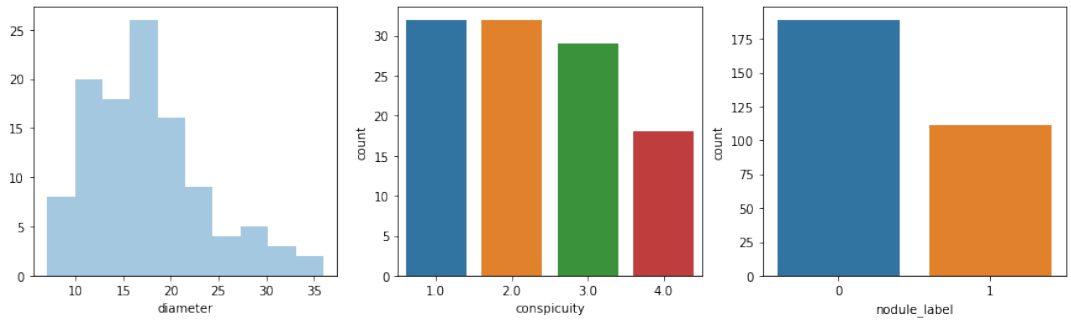


FIGURE 3.1: Overview of distribution of nodule size, conspicuity, and class label.

3.2 Computer-aided Detection (CAD)

During the evaluation of nodule detection CAD, we considered three CAD systems: CAD-A, CAD-B, and CAD-A+B. CAD-A and CAD-B are both commercially available product, whereas CAD-A+B is derived from the mentioned CADs. Each CAD has its characteristics, which we describe in the paragraphs below.

CAD-A In this study, CAD-A can be classified as a *traditional* CAD system, as this version was released before the introduction of neural networks for medical imaging and makes use of human feature engineering. CAD-A is optimized to detect nodules between 9 - 30mm in diameter (Schalekamp et al., 2014b). Using default clinical threshold of 0.35, CAD-A achieves a 74% sensitivity with 1.0 false-positive per image (Schalekamp et al., 2014a).

CAD-B CAD-B is a modern CAD tool, as it has been developed by using neural networks and has been trained using more than 40.000 chest radiographs. Nam et al., 2019 showed very high specificity of 95.2% and 80.7% sensitivity using this CAD and showed that it could detect 100% of high conspicuous nodules and most large (>3cm) nodules. The limitations of CAD-B lie in the detection of small (<1 cm) and less conspicuous nodules Lee et al., 2020. When using this CAD as a second reader, radiologists saw an increased performance for malignant nodule detection (Nam et al., 2019).

CAD-A+B We reason that two combined CAD systems are better than one. Combined reasoning mimics human nature, as multiple opinions prove for a more robust prediction (Sagi and Rokach, 2018). Therefore, we chose to create an ensemble of the two previously described CADs. We achieved this by calculating the mean prediction of the two algorithms, resulting in CAD-A+B.

3.2.1 Choosing the ROC cut-off threshold

Following the study of Schalekamp et al., 2014b, each reader assigned a score between 0 and 1. From the continuous scores of this study, we chose the cut-off point for a positive prediction at $>= 0.5$. Then, we can calculate and plot the average sensitivity and specificity for each reader. In Figure 3.2, we see that this results in a mean reader sensitivity of 69% (interquartile range 67-73%) and mean specificity of 87% (interquartile range 84-94%) for standalone readers.

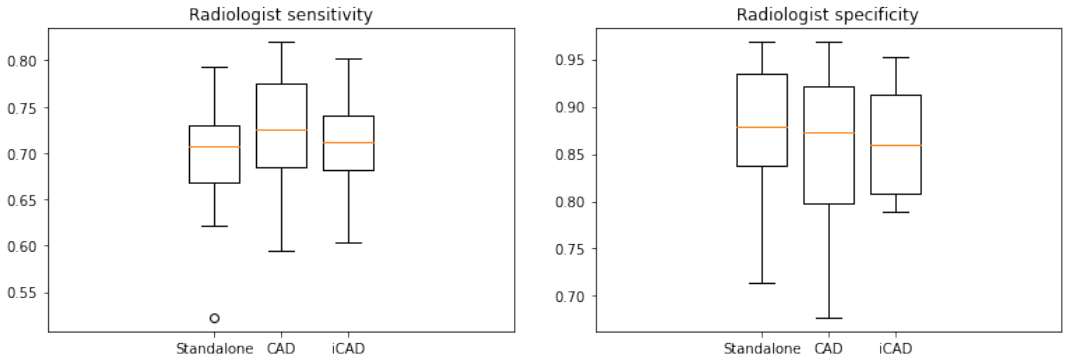
CAD systems are designed for concurrent read and often come with a factory setting threshold. Depending on the purpose of CAD, we can alter the threshold setting. In order to evaluate different use cases, we defined various points of interest on the ROC curve. The question remained that we needed to determine a point on the ROC curve with a certain specificity and sensitivity. There were two options:

1. Choose a point on the CAD ROC based on the performance of average readers (Figure 3.2).
2. Choose a point on the CAD ROC that gives the best trade-off between specificity and sensitivity as possible (most upper-left point on the ROC).

The first option gave CAD an equal error rate as radiologists. The second option rejects the notion of a balanced trade-off between specificity/sensitivity, which results in either a very high specificity and a very low sensitivity, or vice-versa.

We decided that CAD should achieve a sensitivity/specificity equal to the upper quartile of the average radiologists, depending on the use case of CAD. Therefore, for each CAD, we defined both high-sensitivity and high-specificity modes. The upper quartile sensitivity for the high-sensitivity mode was 73%, and for high-specificity mode, we acquired the upper quartile specificity of 94%. The baseline values for each CAD with the different modes are displayed in Table 3.1.

When we inspect the distribution of scores of *CAD-A*, we stumble upon a problem. On the ROC curve, the maximum sensitivity of CAD is limited and does not reach a sensitivity above 80%. This is because there exist no prediction probability thresholds between 0.0 and 0.13. The more continuous these numbers are, the more precise one can define thresholds. In this case, *CAD-B* has an advantage. By inspecting the ROC curve in Figure 4.3, we see that *CAD-B* provides variations on the curve for all false-positive rates. This means that for *CAD-B* we can choose these high sensitivity or specificity thresholds if needed.



(A) Average reader sensitivity

(B) Average reader specificity

FIGURE 3.2: Performance of radiologists on the nodule dataset. Shows clear differences in sensitivity between standalone and the use of CAD-A.

TABLE 3.1: CAD with high specificity/sensitivity modes. Thresholds are based on the upper quartile performance of the average reader.

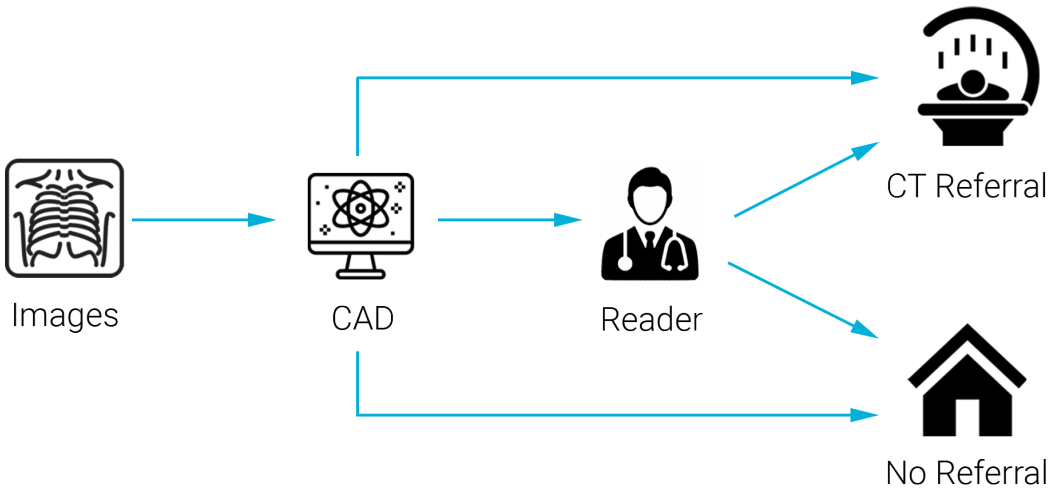
	Mode	Sensitivity	Specificity	Threshold
Readers (upper quartile)	Standalone	73%	94%	0.5
Readers (upper quartile)	with CAD-A	77%	92%	0.5
CAD-A	High Sens	73%	39%	0.36
CAD-A	High Spec	34%	92%	0.71
CAD-B	High Sens	74%	29%	0.0052
CAD-B	High Spec	34%	93%	0.1586
CAD-A+B	High Sens	73%	45%	0.2172
CAD-A+B	High Spec	47%	93%	0.3853

3.3 Implementation types

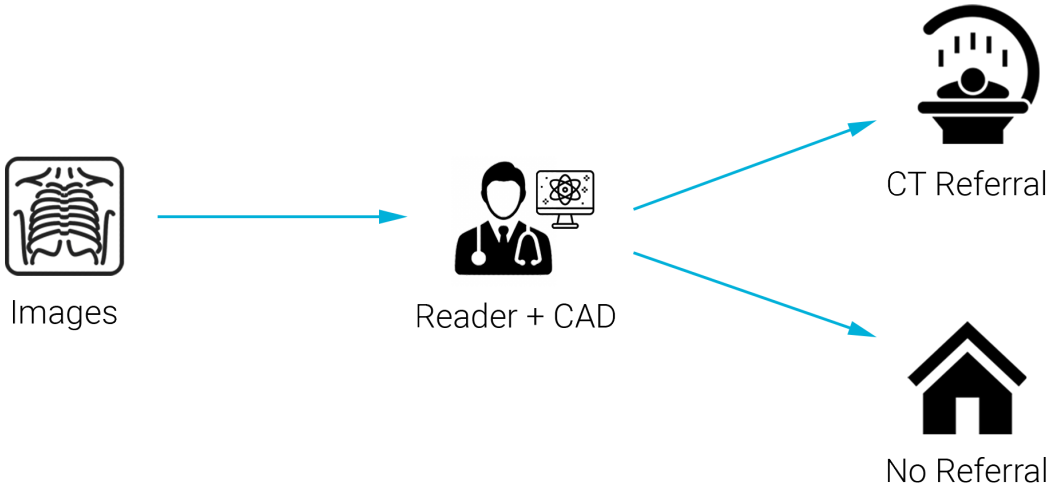
As seen in Chapter 1.2, we can separate the integration types into three broad categories:

1. CAD as first reader.
2. CAD with reader concurrently.
3. CAD as second reader.

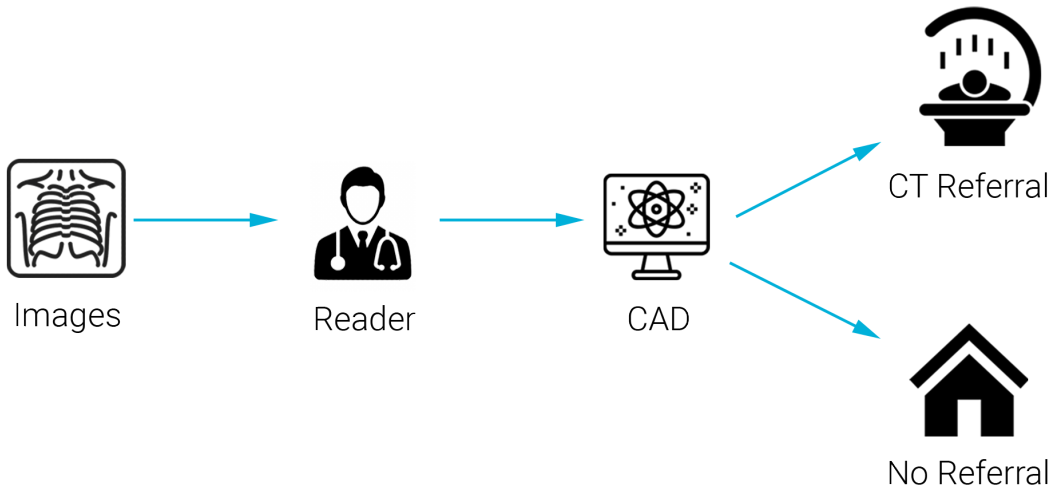
These integration types are further broadly illustrated in Figure 3.3. In the following chapters, we will further explain the three different integration possibilities and use the previously mentioned CAD modes (Table 3.1) when applicable. For analysis, we created the simulations using Python 3.8.



(A) CAD as First Reader.



(B) CAD as Concurrent Reader.



(C) CAD as Second Reader.

FIGURE 3.3: Overview of CAD integration types.

3.3.1 CAD as First Reader

Applying CAD as a first reader allows images to skip the workload queue of radiologists. In these cases, CAD took the responsibility of the decision and processed a subset of the images without the intervention of the radiologists. There are two approaches to filtering images:

1. Filtering *normal* images
2. Filtering *nodule* images

3.3.1.1 Filtering *normal* images

When we applied CAD to filter out *normal* images, we prioritized the detection of true negatives. By choosing such a priority, we simultaneously limited the number of false-negatives (type 2 error) or *false omission rate*. In order to tackle this problem, we needed to acquire an operating point on the ROC curve that gave the highest sensitivity possible, such that the algorithm minimizes the risk of falsely omissing an image containing a nodule.

3.3.1.2 Filtering *nodule* images

Instead of reducing the workload for radiologists by filtering *normal* images, we can choose to filter out *nodule* images using high-specificity CAD. In this approach, the filtered *nodule* cases would automatically be forwarded to CT screening. This reduces the initial reading time for radiologists, and possibly a shorter time to diagnosis. The trade-off here was the possible increased rate of false-positives, which may lead to an unnecessary increase in additional radiation exposure for patients.

3.3.1.3 Redistributing images to radiologist

After applying CAD as a first-read filtering solution to save time, we can look at ways to make use of the saved radiology time. The saved time in itself could be the end-goal, but we can also choose to make use of the saved time to increase detection rates. One way to increase detection rates is by double-reading hard images. For screening scenarios, it is a recommended practice to double-read images to increase recall rates (Ciatto et al., 2005). Therefore, after we have substituted the radiologist read of a subset of images by a standalone CAD read, we redistribute the saved radiology reads to be used for double reading of images that appear harder to CAD. We expect that when the number of double-reads increase, we obtain better performance for these images. But first, how do we determine which images are *hard*?

We specify two theories that determine which images are *hard*, and thus being selected for double-read:

1. Confident CAD. Cases that have a high (or very low) CAD predicted probability of containing a nodule. E.g., for *nodule* filtering, this means a high prediction score, and for *normal* filtering a low prediction score.
2. Uncertain CAD. In this case, the CAD is not sure about its predictions, meaning that it assigns a score that is neither at the extremes of the prediction scale (intermediate probabilities).

Both theories have a different understanding of what *hard* cases entail. For confident CAD, we applied the double read strategy to verify that readers have the same

confidence as CAD, and there is no uncertainty among these images. For uncertain CAD, we reasoned that CAD is unsure about its prediction, and possibly readers would be unsure too.

We applied these scenarios to both the filtering of *nodules* and *normals*, and defined four scenarios as seen in Figure 3.4. In the following paragraphs, we further explain these scenarios.

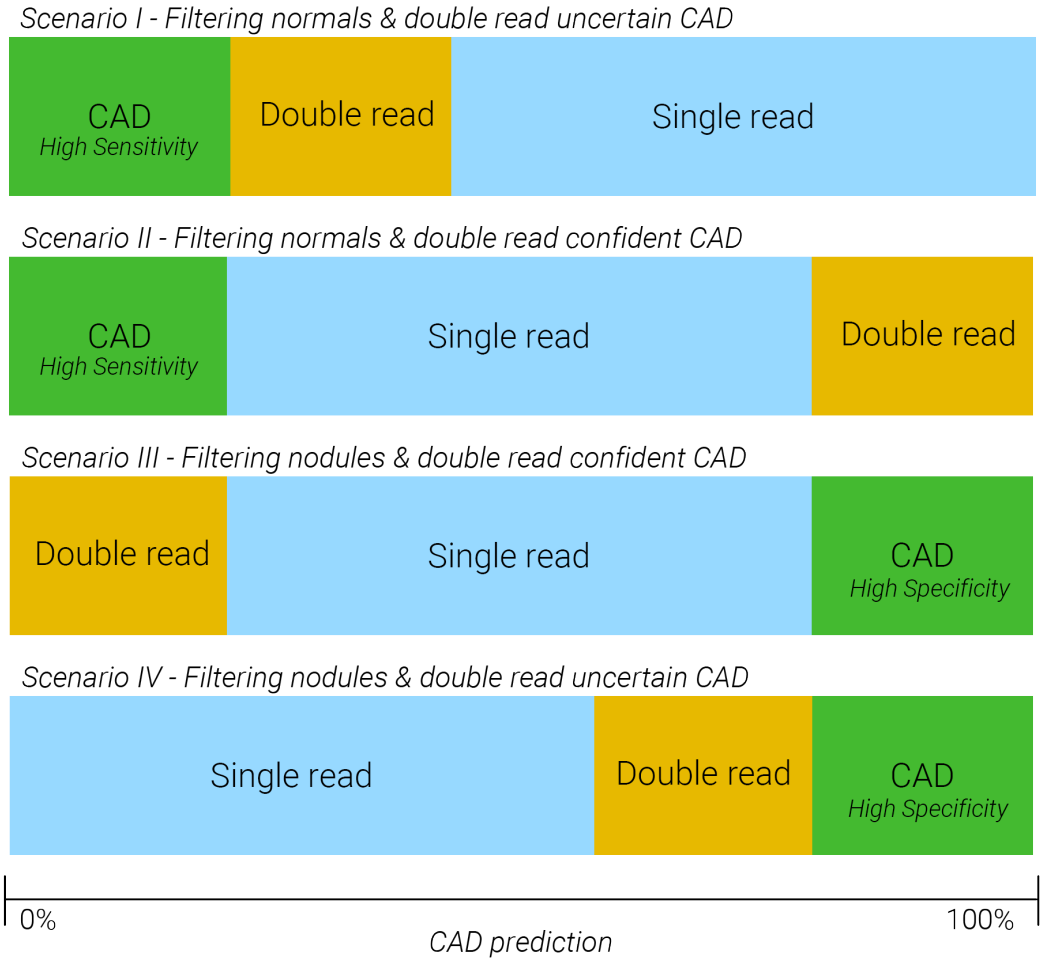


FIGURE 3.4: CAD as First Reader – Double read strategies depending on the implementation use case of CAD.

Scenario I & II - Filtering *normals* In the first two scenarios, we saved images by filtering *normal* images using the high-sensitivity CAD thresholds. By filtering normal images, we saved several images which we could consecutively spend on double-reading another set of images. The scenario was considered a success when the performance was better than the performance of a single reader. We also suspected that the performance would not surpass the mean ROC curve of the double reader, and considered double reader performance as upper-bound for these scenarios.

Scenario III & IV - Filtering *nodules* In the third and fourth scenario, we excluded images from the normal workflow that have a high CAD probability of containing a nodule (*nodule* filtering). When the CAD system detected a nodule, the patient

was immediately forwarded to obtain a CT-scan and skipped assessment by the radiologist. This provided a low risk for the patient in terms of nodule detection, but allowed for the possibility of an increased rate of false-positives and thus extra time and radiation dose for the patient.

3.3.2 CAD as Concurrent Reader

In this section, we present concurrent reading strategies where the radiologist has had direct interaction with CAD-A. In general, there were a multitude of integration strategies:

1. Prompt. When the image was presented, at the same time, the CAD score and localization were shown.

2. Interactive. Only at the radiologist request, the prediction score of a particular region of the image was shown.

3. Hybrid. A combination of prompt and interactive mode. Here, confident CAD predictions were immediately shown, and less confident predictions were available upon request. For CAD-A, this mode was not available, and therefore could not be tested.

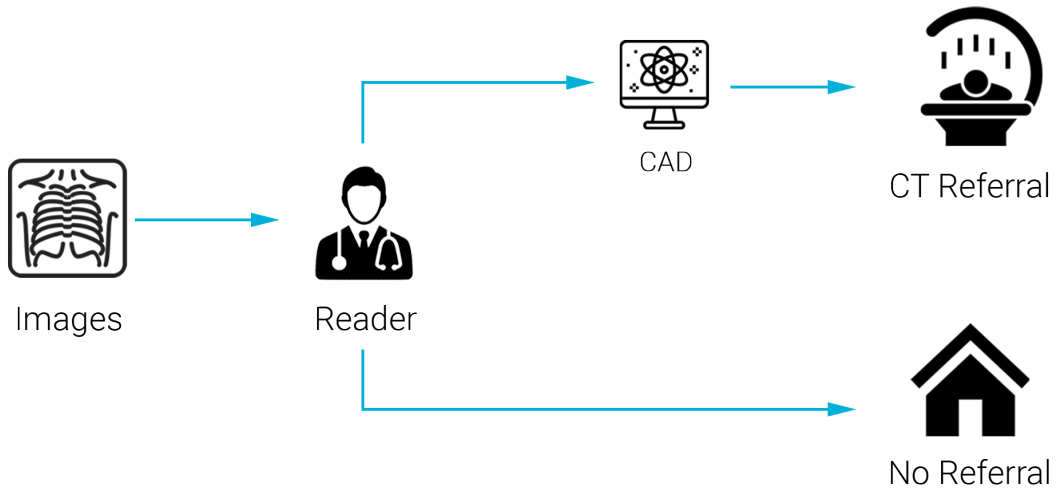
We referred to the previous study of Schalekamp et al., 2014b, where readers used bone suppression images (BSI) and CAD-A concurrently. CAD-B and CAD-A+B were not considered for this scenario, as the interaction between CADs - radiologists were not available. This would mean that we would need to find new readers with an equal amount of experience with CAD tools for each CAD evaluation.

3.3.3 CAD as Second Reader

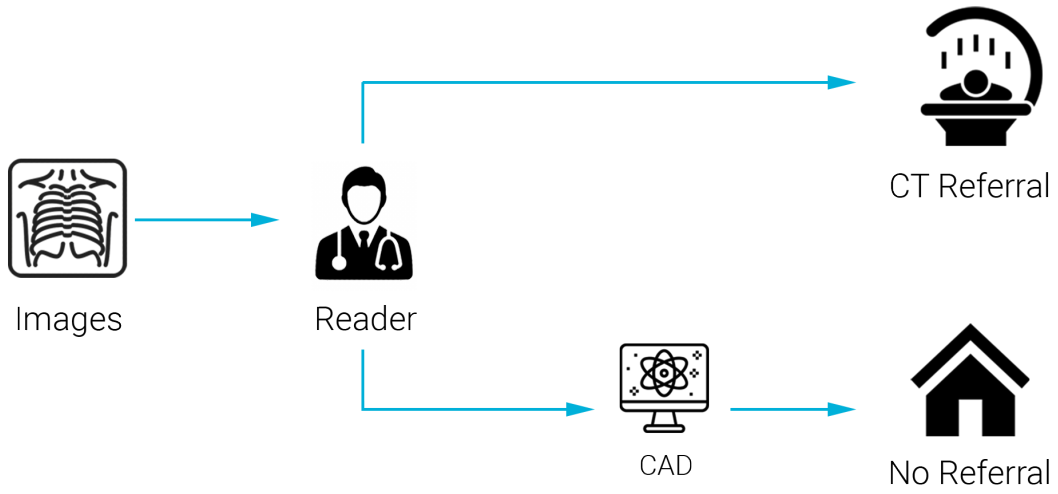
In this section, we explored the possibilities of CAD as a second reader. In general, the task of a second reader was to confirm whether the first reader was correct. Compared to other CADs implementations as second reader (Iussich et al., 2014), our second read strategy does not incorporate a *third* look by the reader. This way, our strategy does not increase reading time but influences the number of CT referrals after the clinical decision. Depending on the use case of CAD, two corrective second read actions could be considered:

1. A positive nodule prediction by the first reader was rectified by the CAD in order to decrease false-positives and unnecessary CT reference (Figure 3.5a).

2. A missed nodule by the first reader was spotted by the CAD as second reader. In this case, this could increase false-positives and reduce the number of false-negatives (Figure 3.5b).



(A) False-positive reduction



(B) False-negative reduction

FIGURE 3.5: CAD as Second Reader, two scenarios for either false-positive or false-negative reduction.

3.3.3.1 CAD for false-positive reduction

For false-positive reduction, CAD acted as a gateway to reduce false-positive predictions. All the positive radiologist predictions ($>= 0.5$) got passed through this gate and are evaluated for possible false-positives. Here, CAD had the 'final' call to decide whether the found nodule was a true positive or false positive. The goal here was to reduce the false-positives as much as possible, while at the same time not losing any true positive predictions.

We chose our model thresholds such that CAD captured as many images containing nodules as possible, which means we chose a high sensitivity mode. High-sensitivity tends to miss few nodules, at the cost of increased false-positives. When CAD does not classify an image as having a nodule, it is most probable that this image did not contain a nodule. Consecutively, when a reader assigned an image as containing a nodule, and CAD was not able to detect a nodule, we could consider this image for second-reading by another reader. We hypothesized that this CAD

pathway might result in fewer unnecessary follow-up CTs and, therefore, reduced healthcare costs.

However, the risk factor was considered high, as CAD could potentially filter out true-positive images, even though readers could detect nodules on these images.

3.3.3.2 CAD for false-negative reduction

The second pathway for CAD as a second reader would be to increase the recall rate of positive nodule patients. This requires a CAD system that is confident about its positive predictions (high specificity mode). The high specificity mode comes with a low recall rate (sensitivity), but in combination with the high sensitivity of the reader, this combination may provide an increase in nodule recall.

This approach gave us a trade-off between additional reads and increased nodule recall. Because readers scored continuous values between 0-100, the confidence of a reader's prediction was captured. We divided false-negative reduction into two scenarios:

1. Reader was confident in its prediction (prediction of 0.0%).
2. Reader was insecure in its prediction (prediction between $> 0\%$ and $< 50\%$).

Confident readers Confident radiologists are readers that score images with a probability of 0.0. Then, the radiologist would be confident that a particular image does not contain a nodule. When CAD was then applied using high-specificity parameters, this approach could increase nodule recall.

Insecure readers The reader is not completely confident about rating an image as negative, and scores images between > 0.0 and 50. We suspected that insecure readers provide a better estimate to obtain additional nodules compared to the confident reader category.

3.4 Metrics

To measure the quality of CAD tools and implementation strategies, we calculated various metrics to compare the strategies. While comparing these metrics, we found that some metrics were negatively correlated with each other. Depending on the scenario, one metric would be more important than others, thus creating a trade-off between one metric versus another. One example would be that in a certain scenario, the specificity increased, but at the same time, the average reader time was increased as well. According to Fryback and Thornbury, 1991 efficacy hierarchy model, we first need to achieve a reliable diagnostic accuracy efficacy before we can achieve benefit higher upon the hierarchy. Diagnostic accuracy efficacy could be achieved by maximizing the CAD tool's performance: AUC, sensitivity, and specificity.

Sensitivity The sensitivity, or recall, represents the percentage of retrieved nodule cases. It is calculated by dividing the number of true positives by the number of positive cases: $\frac{TP}{TP+FN}$ (Parikh et al., 2008). The higher this number, the better. When maximizing this metric, we optimized for a scenario where we were not allowed to miss any nodules.

Specificity The specificity, or true negative rate, represents the percentage of correct negative classified cases as normals. High specificity means a low chance of false-positive predictions. Specificity is calculated by dividing the number of true negatives by the number of true negatives and false positives: $\frac{TN}{TN+FP}$. Mostly, this metric is balanced with *sensitivity*, as ruling out a positive prediction means sacrificing the probability of recall.

Area Under the Curve (AUC) The Area Under the Curve (AUC) score is a collection of the sensitivity and specificity combination. When plotting the sensitivity - specificity combination for each threshold, we acquired an ROC curve. When calculating the surface area underneath this curve, we could obtain the AUC score. The higher this score, the better the combinations of specificity - sensitivity is considering all thresholds. For binary classification such as nodule detection, an AUC score of 0.5 is considered chance level.

The metrics specified above (sensitivity, specificity, and AUC) do not take into account the prevalence of a positive (nodule) class. This is dangerous, as a positive prediction of a rare scenario (i.e., low prevalence situations) have a higher significance to change the sensitivity and specificity compared to high prevalence scenarios (Halligan; Altman, and Mallett, 2015). We overcame this problem automatically due to our high prevalence simulation. Choosing the best model based on AUC score does not necessarily mean a performance translation to clinical scenarios with different prevalence rates. Therefore, the AUC score and ROC curve are not enough for the evaluation of screening tests (Qin et al., 2020). For clinical end-users using a machine learning model, the AUC score is not suitable for guidance (Sendak et al., 2020).

Reading time The reading time determines the interpretation time the radiologist looks at an image. Reading time changes depending on the number of cases to read and how much time is spent on these images. In this metric, we do not address the reading time that follows from the prediction (e.g., taking into account follow-up CT). Reading time can be influenced by multiple factors, such as self-confidence and trust in the CAD tool. More of these factors are further explained in Chapter 2.3.

Risk Oakden-Rayner, 2019 states that we could define risk as to the level of required supervision. If there is no human supervision present in situations where healthcare workers make decisions based on CAD output, we assume this circumstance is a high-risk situation.

Change in workflow For some implementation methods, the workflow of the radiologist would have to disruptively change. For these methods, it is shown that change in workflow significantly decreases the adoption of these tools in the clinic (Werth and Ledbetter, 2020). Therefore, we defined three levels: low, medium and high. A *low* change in workflow means that the basic workflow remained unaffected; CAD performed outside the visible scope of radiologists. For a *medium* change in workflow, the workflow was altered, but no significant training was needed. For a *high* change in workflow, radiologists would require training in order to achieve optimal performance.

Chapter 4

Results

In this chapter, we present the findings of our analysis. First, we establish a baseline for the readers and CADs, and then we dive into the analysis of our experiments. We further dive into the details of the results in the consecutive chapters, but a summary table of all the evaluated use cases can be found in Table 4.5.2.2.

4.1 Baseline – Readers

Readers determined the baseline performance for our scenarios. The mean AUC scores of individual readers (12 readers) without the use of CAD was 0.83. In Figure 4.1, we show a distribution of the predictions that each reader scores. We observed that some readers do not score uniformly from 0% to 100%, but instead round to tens (i.e., *obs 5*). Other readers tend to score only two values: 0 or 50, and some others only score confident positives scores between 50 and 100 (*obs 6* and *obs 12*). Therefore, we conclude that the predicted scores are subjective.

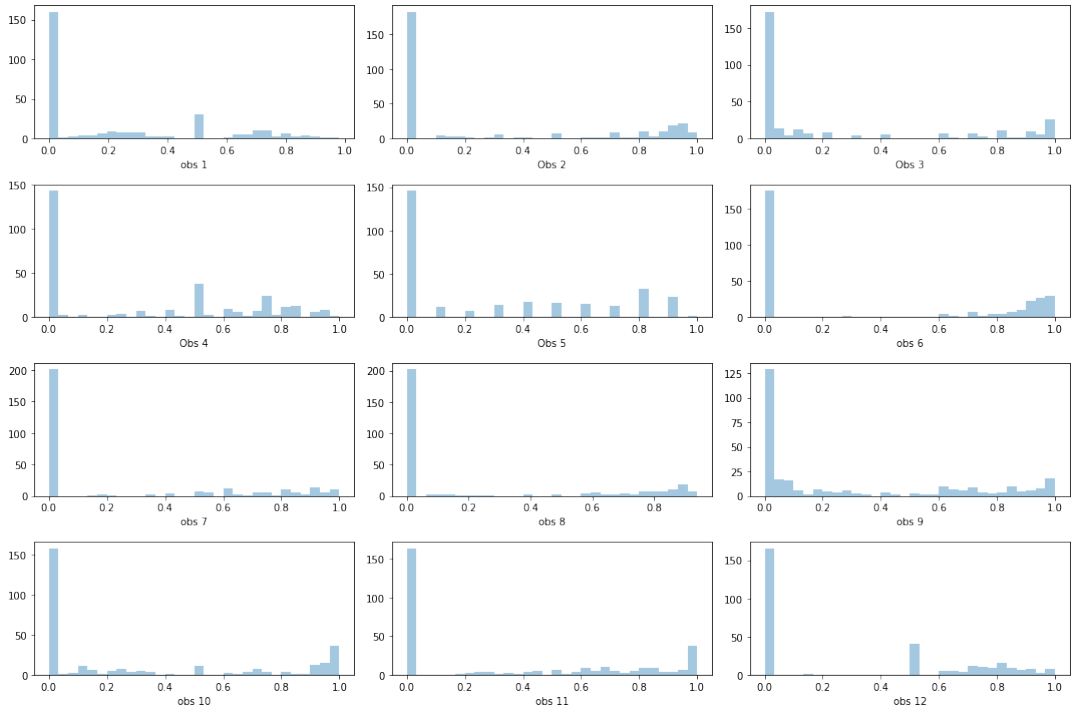


FIGURE 4.1: Individual reader distribution scores. The scores show a different distribution for each reader. The readers are probably not correctly educated to score uniformly and score subjectively based on a threshold of 0.5.

4.1.1 Multiple readers

We evaluated the performance of multiple readers to assess maximum achievable performance. We considered the performance of double, triple, quadruple, and quintuple reading and averaged their predictions to calculate a mean prediction per image. The results are shown in Figure 4.2. Overall, we see that the AUC score increases when the number of readers increases. The difference between quadruple (four) reading versus quintuple (five) reading seems minimal.

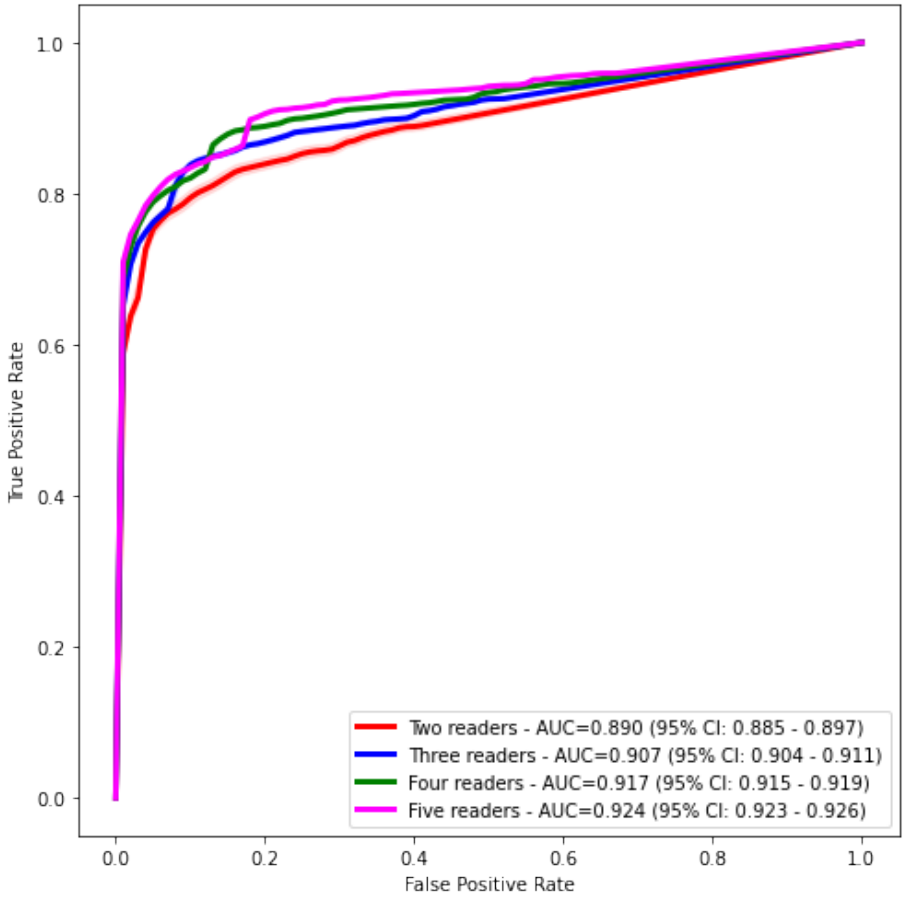


FIGURE 4.2: ROC curves of the combination of ensemble readers. More readers result in an increase in AUC score.

4.2 Baseline – CAD

An overview figure containing all CADs versus baseline readers is shown in Figure 4.3. From this figure, we can conclude that CAD is not on par with radiologist performance, as these curves do not intersect with readers' sensitivity - specificity combination.

In Figure 4.4, we display the distribution of scores per CAD. CAD-A shows a normalized distribution, but no scores exist between 0 and 0.13. This discrepancy is seen in the ROC curve as well, as CAD-A is unable to achieve higher sensitivity than 0.80 due to this score gap. CAD-B tends to assign most images a low score. This might make this CAD not able to perform well in high sensitivity scenarios, as setting a specific threshold that generalizes across other datasets might be hard. The

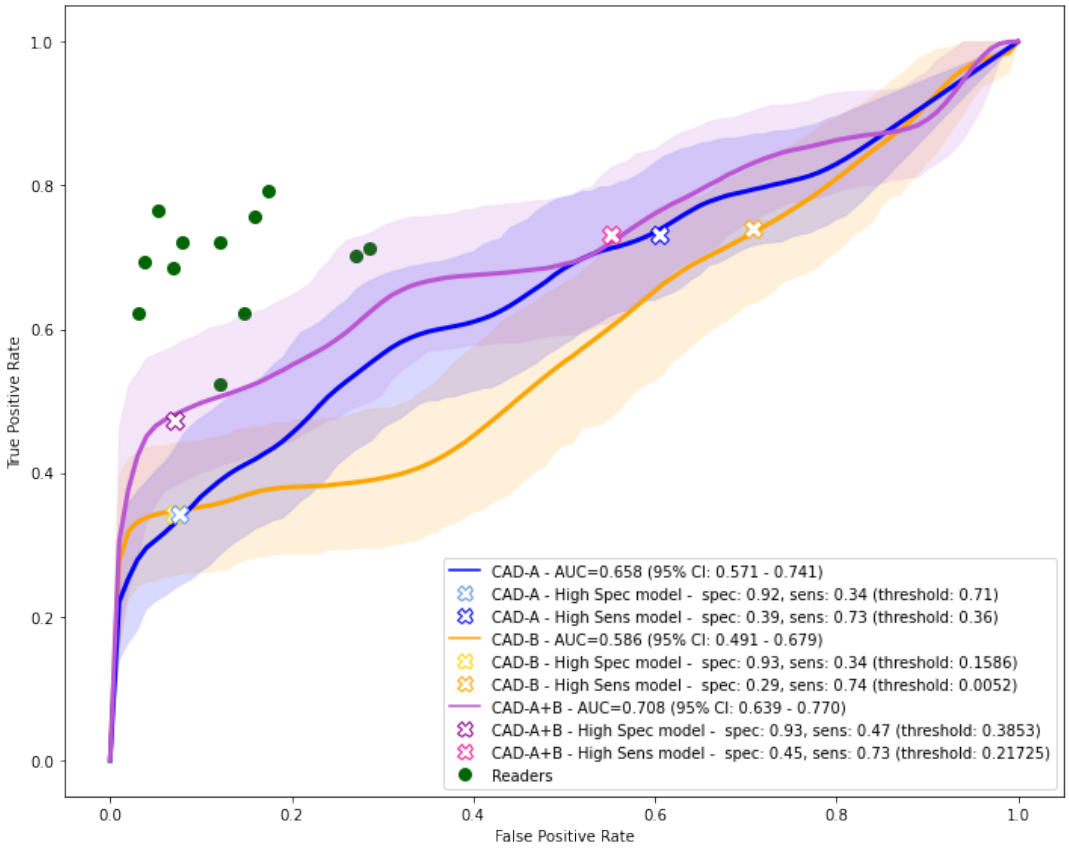


FIGURE 4.3: ROC curves of CAD and ROC points of individual readers at threshold 0.5.

ensemble of CAD-A and CAD-B (CAD-A+B) restores this normalized distribution in some sense, allowing for a better threshold selection. Here, we get the normalized distribution of CAD-A, while having the continuity of CAD-B.

CAD-A Traditional CAD tool CAD-A achieved a performance of 0.656 AUC. Using the default clinical threshold of 0.35, we obtained a specificity of 38% and 74% sensitivity.

CAD-B CAD-B achieved a performance of 0.586 AUC, which is lower than CAD-A, without overlap in 95% confidence intervals for some points on the ROC. Using the threshold of 0.30, different results on our dataset were reported: 97% specificity and 32% sensitivity compared to a specificity of 94.6 - 100% and 70.3 - 91.1% sensitivity as reported by Nam et al., 2019.

CAD-A+B The ensemble CAD-A+B, combining the predictions of CAD-A and B, resulted in an AUC score of 0.707. Figure 4.3 shows that the ensemble performs better than each CAD individually, especially in the high specificity region (low false-positive rates).

4.2.1 Split by nodule size

The results in Figure 4.5 show that for large nodules (15-36mm), CAD-A, and CAD-A+B (AUC of 0.646 and 0.682) significantly outperform CAD-B (AUC of 0.513).

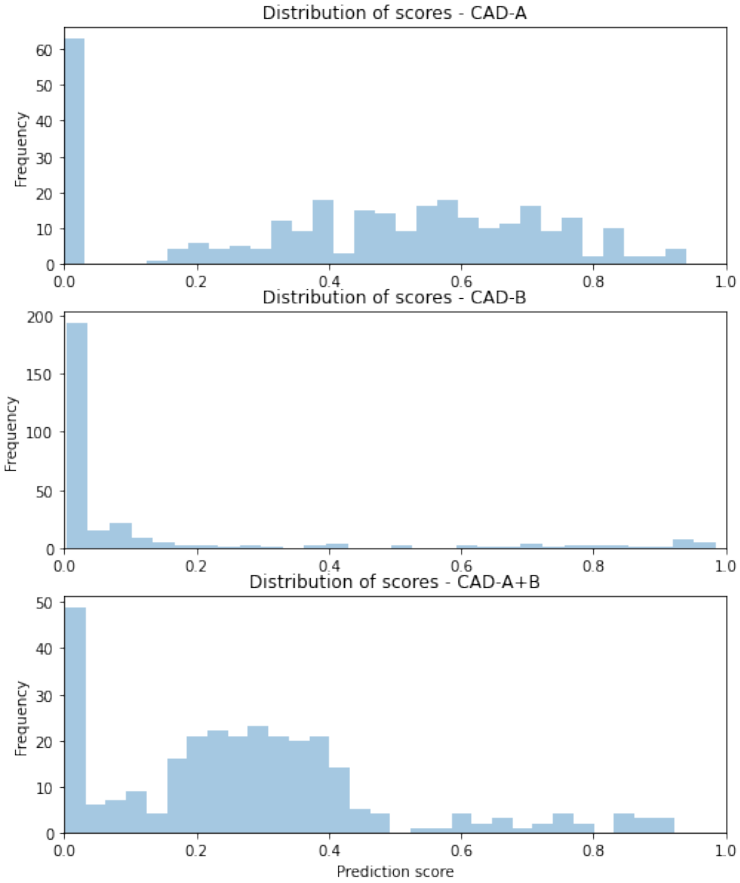
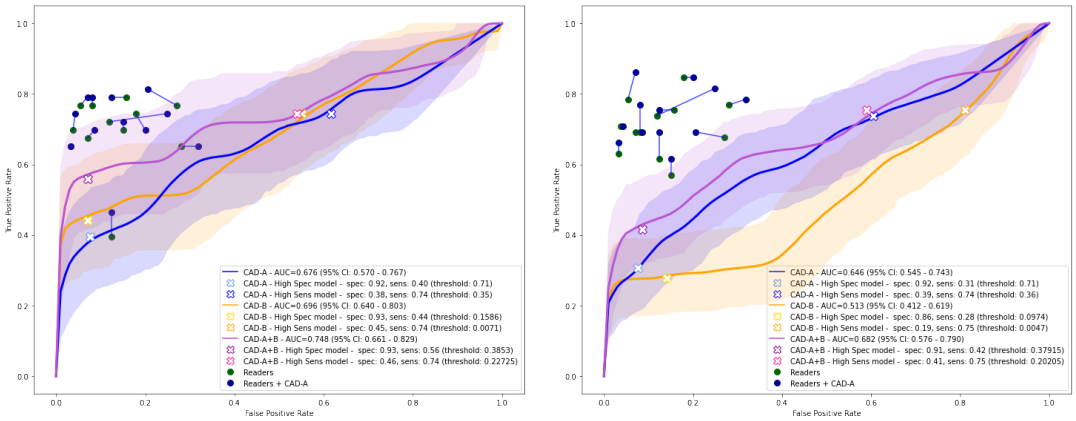


FIGURE 4.4: CAD score distribution.

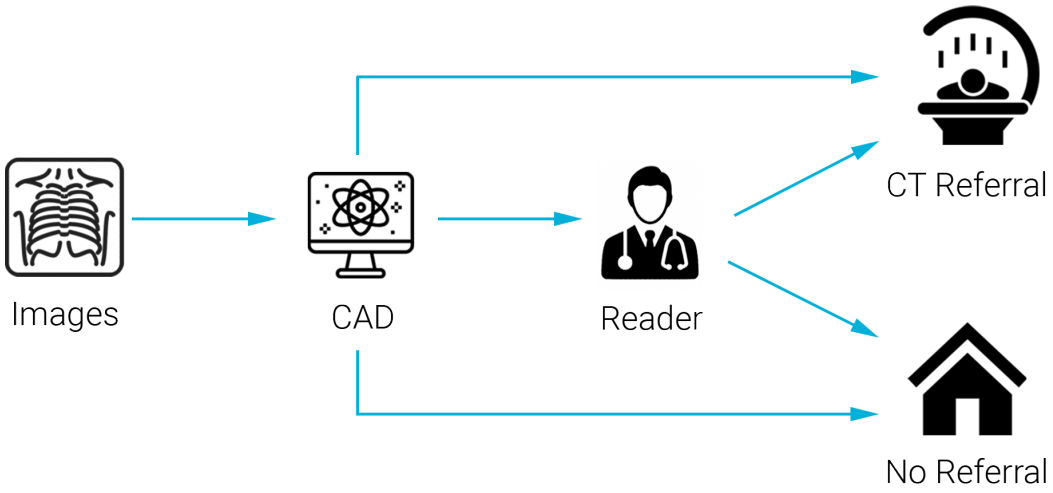
CAD-B performed well for high-specificity, but beyond the high-specificity region, CAD-B performed at chance level.



(A) ROC of small nodules (7-15mm) (B) ROC of large nodules (15-35mm)

FIGURE 4.5: ROC of CAD and readers split by nodule size: small and large. CAD-B performs significantly worse with large nodules.

4.3 Experiment – CAD as First Reader



In this section, we explore the possibilities of having CAD as a first reader. We make certain decisions and see how that affects the results of our dataset. During first read, the responsibility of CAD was to reduce the number of scans that the radiologist needs to assess.

4.3.1 Filtering *normal* images

Acquiring a ROC cutoff point at readers' upper-quartile sensitivity level, CAD-A had a sensitivity of 73% and specificity of 39%. Consecutively, when applying the CAD towards image filtering in order to reduce the number of *normal* images to be read, CAD removed 104 / 293 (35%) images, where 30 false-negatives were observed. This means that 28% of the filtered images were not normal but contain a nodule.

4.3.2 Filtering *nodule* images

Figure 4.6 shows the number of filtered *nodule* images per CAD system. Here, we see that on average, 19% of these filtered *nodule* images were false-positives. As a reference comparison, we show the mean performance of the readers on the right figure. Although CAD had a better specificity compared to the mean reader specificity, we see that for these images of this set, readers tend to classify fewer images as false-positives (3%). However, reader reference also introduced false-negatives (9.7%), meaning that readers missed images containing a nodule that would otherwise have been detected by CAD. Still, for all the CAD filtered images, the average error rate over all the readers (single-read) was 13%[1].

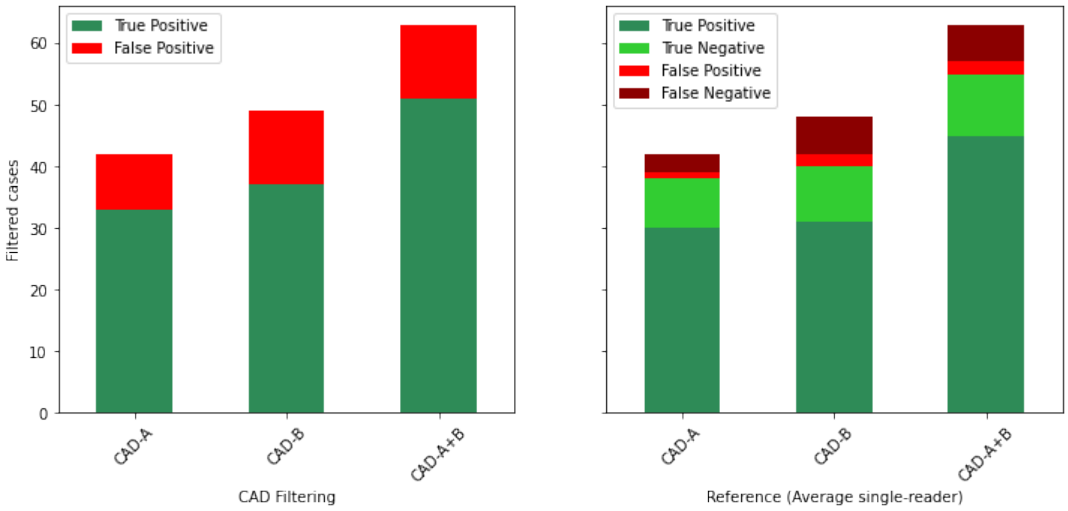


FIGURE 4.6: True/false positives for the CAD filtering operation (left) versus the average single reader on these filtered cases (right).

4.3.3 Redistributing images to radiologists

4.3.3.1 Scenario I & II - Filtering *normals*

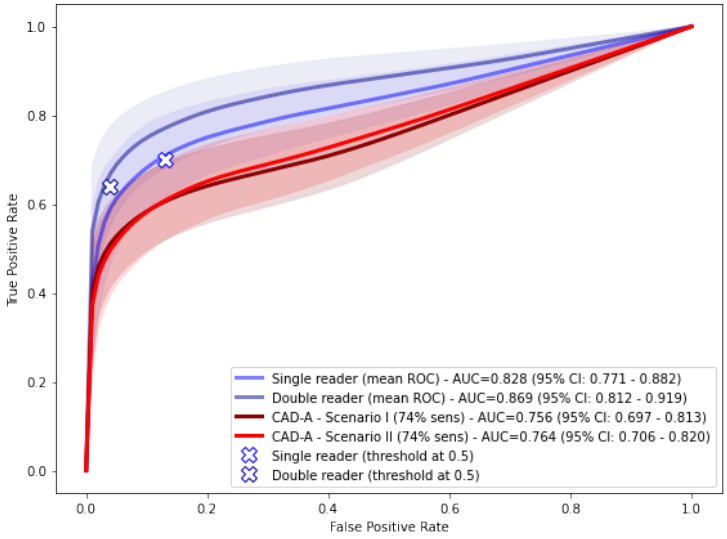
Figure 4.7 shows the performance for the reader distribution for scenarios I & II. For all CADs, overall performance was found worse compared to baseline single reader mean ROC (0.76 vs 0.83 AUC). CAD-B and CAD-A+B (Figure 4.7b and 4.7c) show a higher score (0.79 and 0.78 AUC) compared to CAD-A (0.76 AUC, Figure 4.7a). We also see that for CAD-B and CAD-A+B with a higher sensitivity threshold of 96%, the number of filtered images decreased (from 84 to 21 images, and from 110 to 16 images for CAD-B and CAD-A+B, respectively). Thus, the ROC approximated the performance of a single reader and showed no improvements.

4.3.3.2 Scenario III & IV - Filtering *nodules*

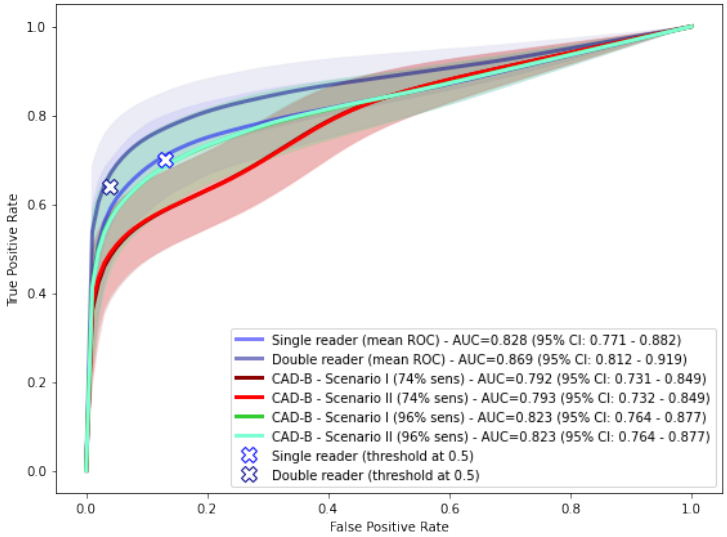
In Figure 4.8 we show the results for these scenarios. We conclude that all CADs benefit positively from this approach, but, due to the spread of the confidence intervals, there seemed to be no significant effect between CADs.

[1] Mean error rate for this set for all CADs is calculated by $\frac{\text{False Nodule+False Normal}}{\text{\# images}}$

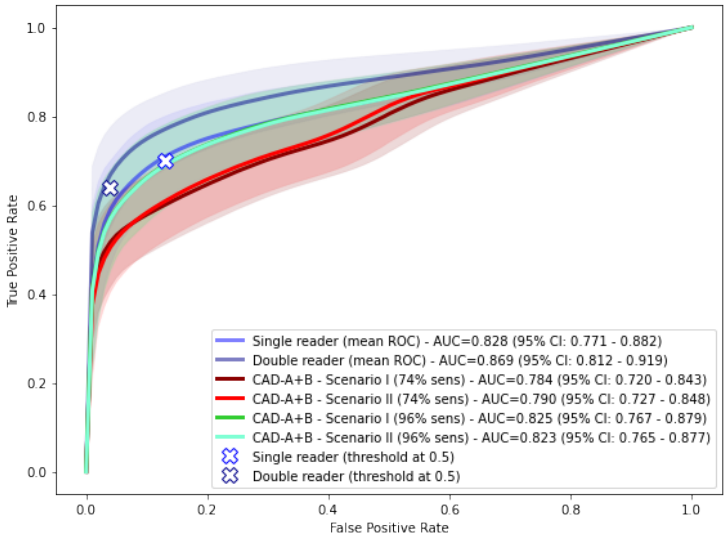
Figures 4.9 and 4.10 show the double-read results for Scenario III and IV, respectively. Scenario III showed fewer false-positives compared to single-read. However, the number of false-negatives during double-reading were similar to the reference single-reader (average error of 24% and 23%). For scenario IV, the double-read strategy caused a decrease in average error from 23% to 18% compared to the single-reader reference. Thus, scenario IV was more beneficial for this particular dataset.



(A) CAD-A *normals* filtering.

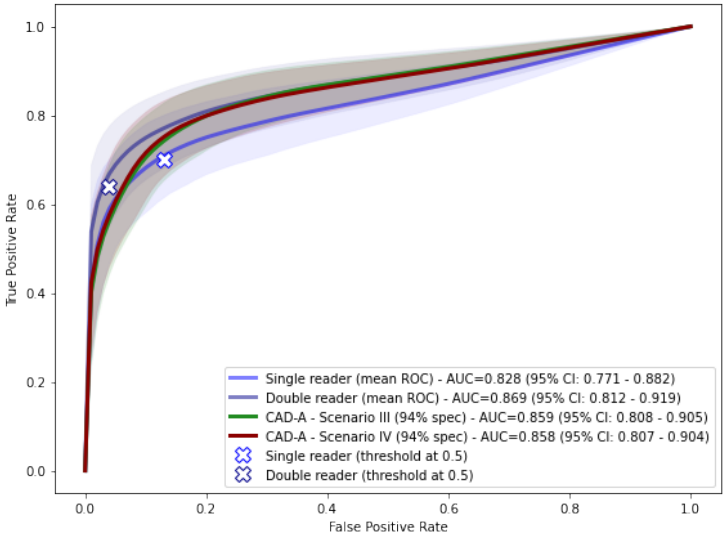


(B) CAD-B *normals* filtering.

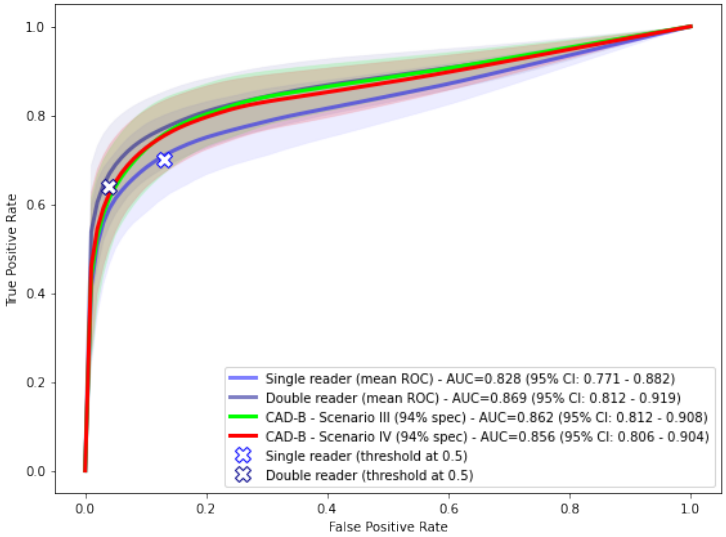


(C) CAD-A+B *normals* filtering.

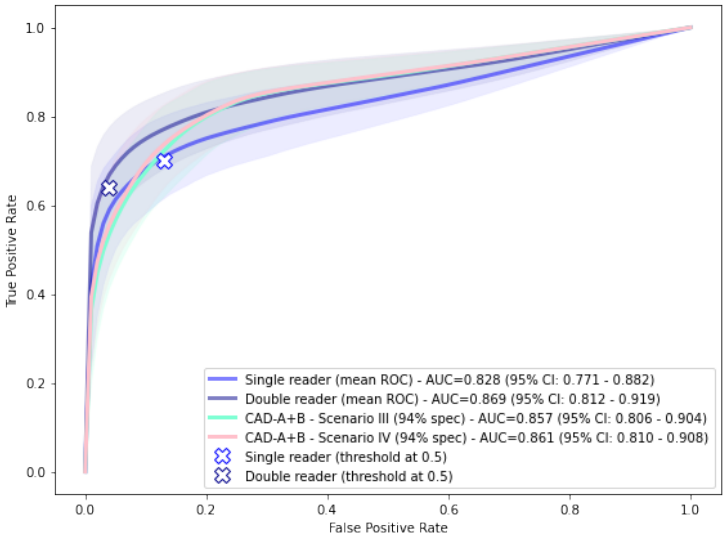
FIGURE 4.7: ROC curves of scenarios I & II. All CADs performed worse than the baseline single reader.



(A) CAD-A *nodules* filtering.



(B) CAD-B *nodules* filtering.



(C) CAD-A+B *nodules* filtering.

FIGURE 4.8: ROC curves of scenarios III & IV. Each CAD perform better than baseline first-reader.

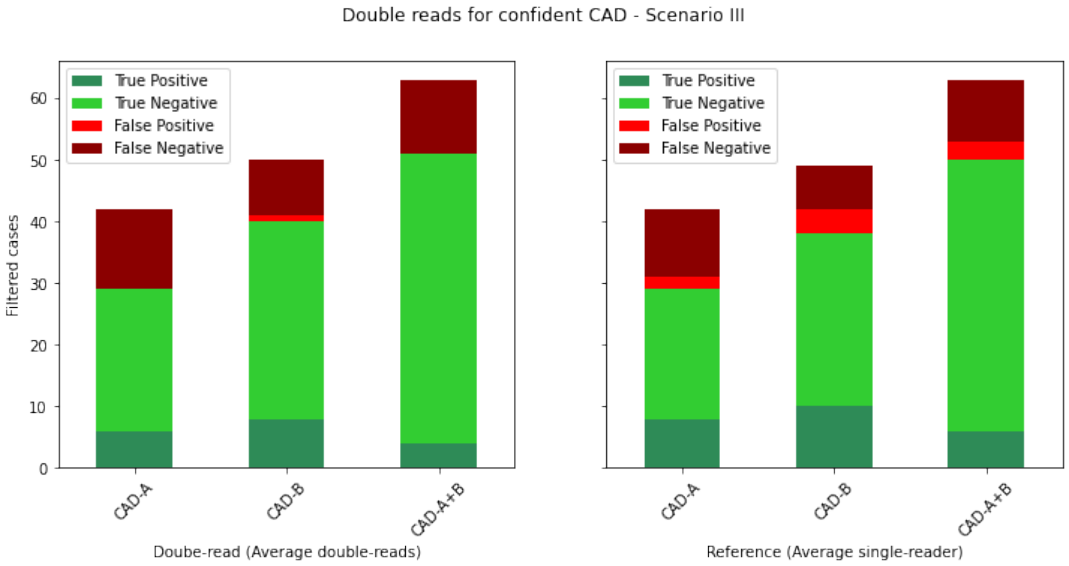


FIGURE 4.9: CAD filtering vs reference for scenario III.

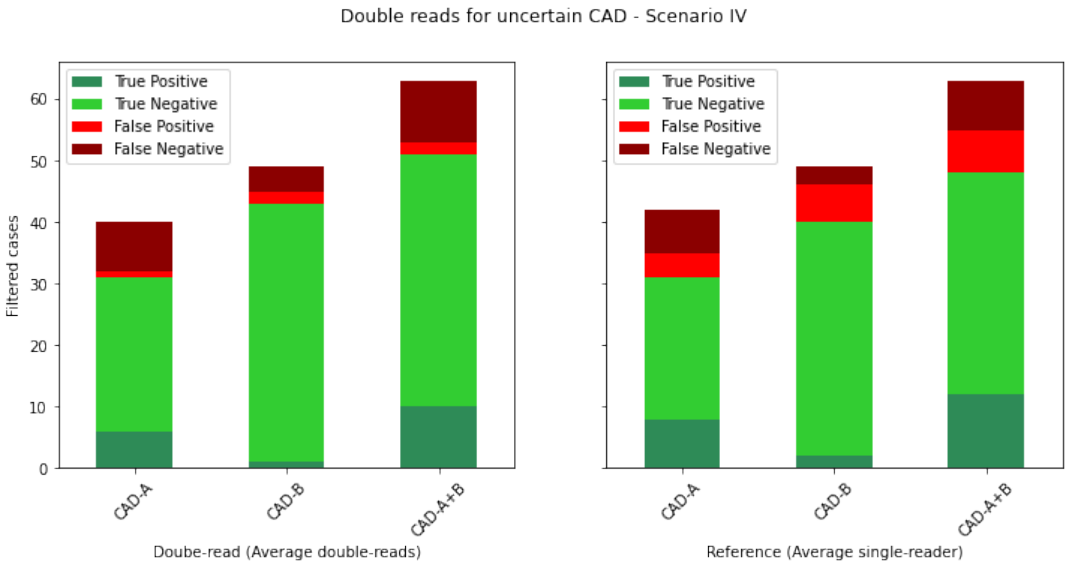
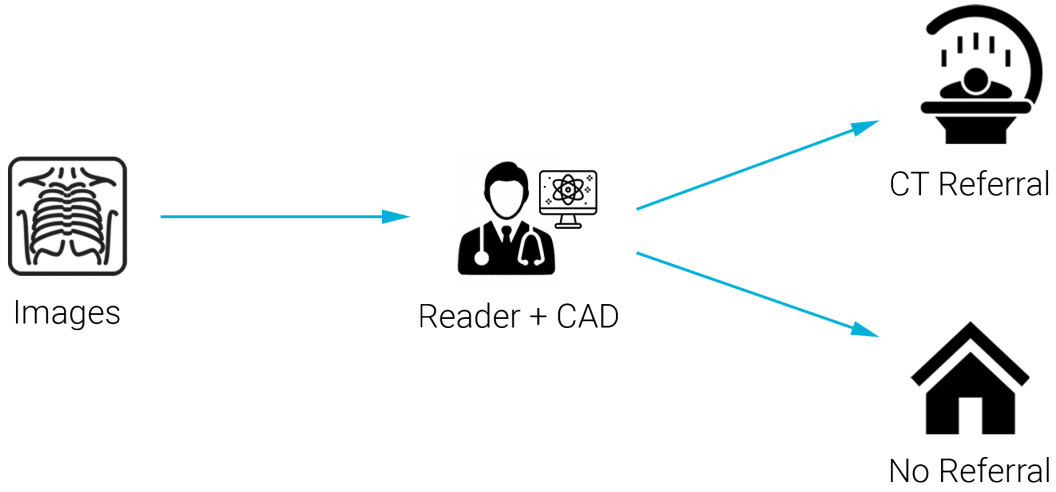


FIGURE 4.10: CAD filtering vs reference for scenario IV.

4.4 Experiment – Concurrent reading (Reader + CAD)



It was found that reader performance increased by the use of CAD as a prompting solution, at the cost of increased reading time. Figure 4.11 shows the ROC of all CADs including prompting and interactive CAD on reader performance.

Prompt During *prompt* concurrent reading, Figure 4.12 shows that in general, every reader benefited from a prompting CAD implementation. The use of CAD significantly improved reader performance (0.824 to 0.841 AUC). For very subtle lesions (low conspicuity), CAD helped readers spot nodules more often (39% recall vs 22%). However, the average reading time increased from 23 seconds to 30 seconds per case.

Interactive The benefit for *interactive* concurrent reading was debatable. Not every reader seemed to benefit from this approach. As seen in Figure 4.12, interactive CAD shows an ROC curve in between standalone and prompting, which means that CAD as an interactive tool might be better than standalone readers, but not as good as prompting CAD. Also, the increase in reading time was less compared to *prompting*: average reading time increased from 23 seconds to 26 seconds.

4.4.1 Split by nodule size

When we split the nodules per size, Figure 4.13 shows that for small nodules, CAD did not necessarily improve performance. For larger nodules, the AUC scores of standalone and prompting CAD were further apart, meaning that CAD improved performance.

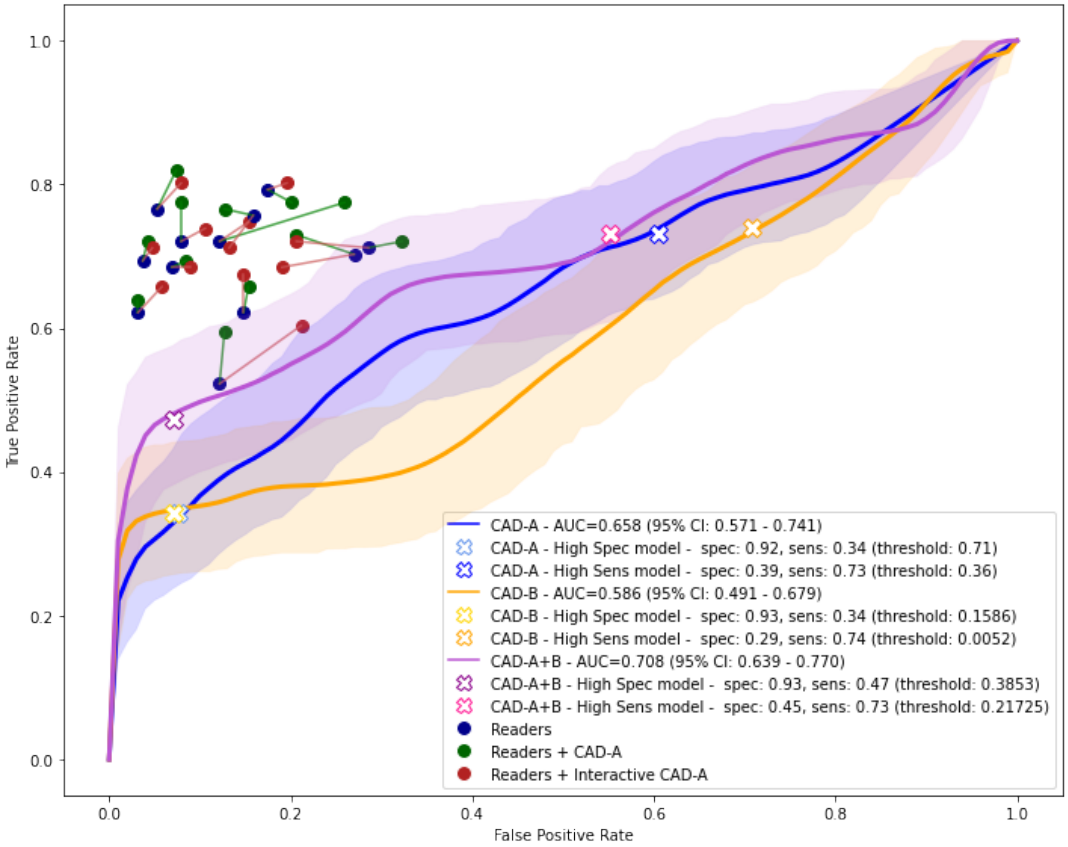


FIGURE 4.11: ROC curves of CAD and ROC points of individual readers with CAD and interactive CAD.

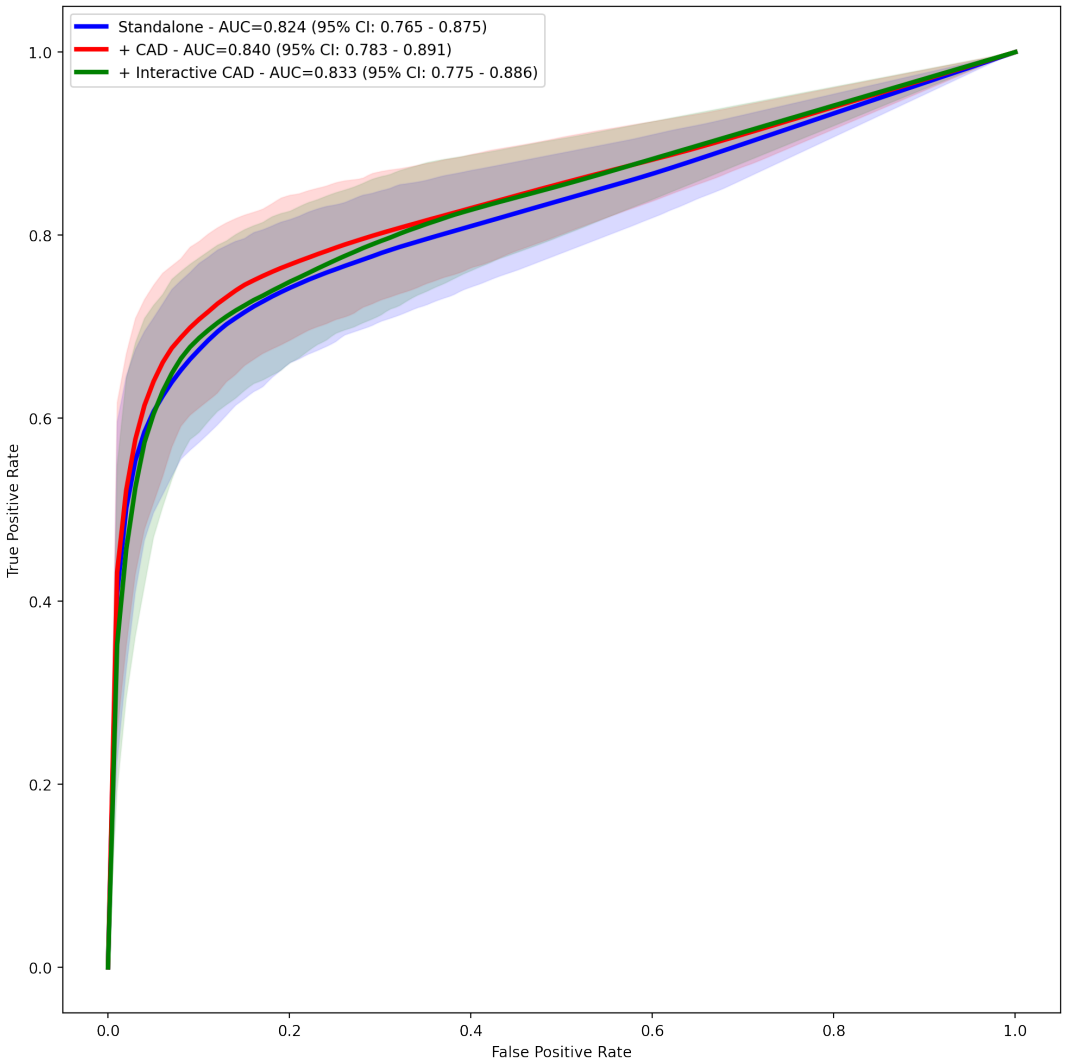
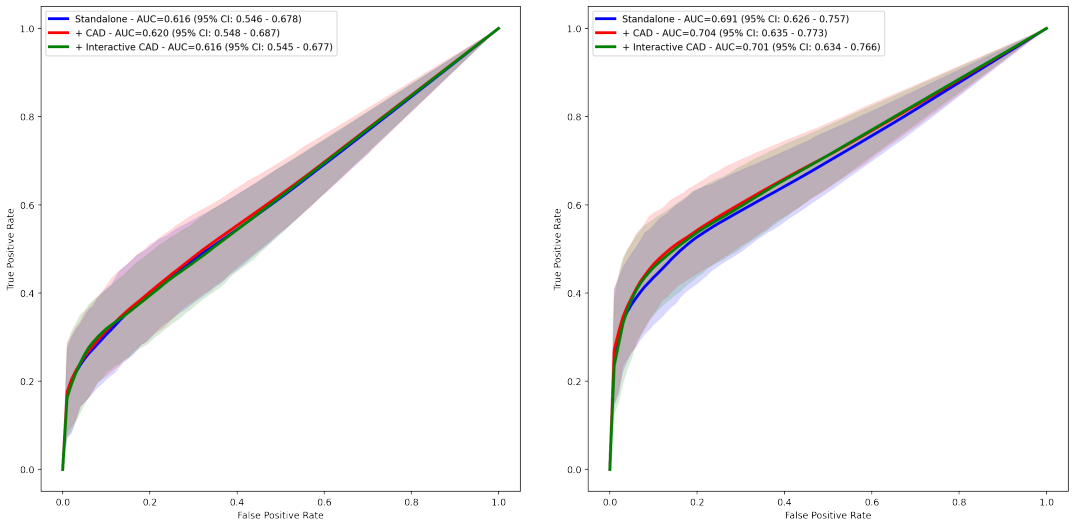


FIGURE 4.12: ROC of readers without CAD vs readers with CAD vs readers with interactive CAD. In general, the concurrent use of CAD tools helped to achieve better performance.

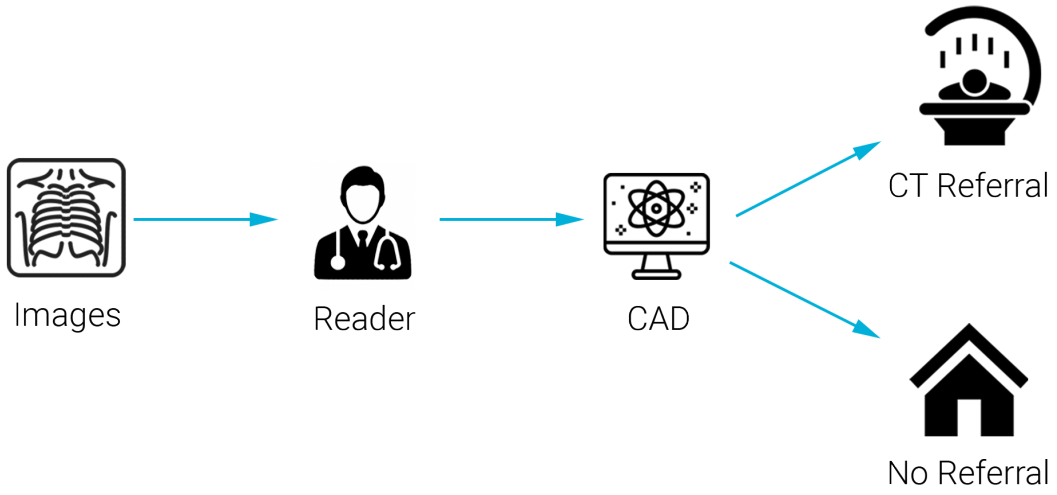


(A) Mean concurrent reader ROC of small nodules (7-15mm)

(B) Mean concurrent reader ROC of large nodules (15-35mm)

FIGURE 4.13: Mean readers ROC with various CAD implementations split by nodule size: small and large.

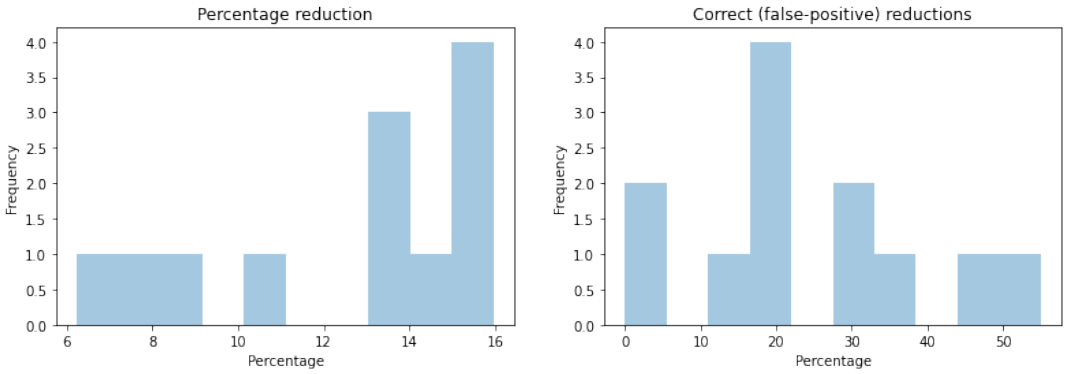
4.5 Experiment – CAD as Second Reader



4.5.1 CAD for false-positive reduction

On average, readers scored 100 cases as nodule images. From these 100 images in our set, 76 contained a nodule. In other words, readers scored 24 images as false-positive. In this scenario, we try to minimize this number while keeping the true-positive predictions of the readers. When applying CAD in this scenario, we reduced the 24 false-positives, but also (unwillingly) removed true-positive images. In Table 4.1 the percentage of correct false-positive reductions is represented in *% fp reductions*. This number needs to be maximized in order to be effective.

CAD-A's high-sensitivity mode with a threshold at 0.21 reduced an average of 13 images per reader. Figure 4.14 shows the distribution of the percentage reduced per reader. The average reduction percentage was 13%. Of these reduced images, 24% of the images contained false-positives, at the cost of 76% of the images being rejected containing nodules. We concluded that for any of our CADs, applying false-positive reduction in this scenario is infeasible.



(A) Total reduction percentage (B) Percentage of reductions that are false-positive

FIGURE 4.14: CAD as second reader, false-positive reductions after positive predictions by radiologists. The percentage of correct reductions need to be as high as possible.

TABLE 4.1: Results for various CADs for second read false-positive reduction. Here, our goal was to reduce false-positives when readers were certain about a positive prediction ($\geq$0.5).

CAD	Avg. reductions	% fp reductions	% total reduction
CAD-A (81% sens, 28% spec)	13 / 100	24%	13%
CAD-B (91% sens, 10% spec)	9 / 100	17%	9%
CAD-B (95% sens, 6% spec)	5 / 100	22%	5%
CAD-A+B (90% sens, 9% spec)	5 / 100	16%	5%
CAD-A+B (95% sens, 6% spec)	3 / 100	22%	3%

4.5.2 CAD for false-negative reduction

For false-negative reduction, we looked at all images that readers classified as negative. Then, CAD evaluated these negatives images using a high-specificity setting. On average, readers scored a total of 193 images as negative, where 32 of these images were falsely rejected and contained a nodule (17%). The task of CAD was to reduce this percentage even further so that nodules were never discarded and therefore lowered the risk. We show that CAD-B performs better than CAD-A+B for this scenario.

4.5.2.1 *Confident* readers

For *confident* readers, CADs using default high-specificity mode (representing upper-quartile reader specificity), readers on average predicted 160 images as 0.0 (no chance of having a nodule), where 24 of these images were false-negatives (15%). Table 4.2 shows for each CAD the average additional image reads with the percentage of these that are actual true-positives. We demonstrated that for higher specificity thresholds such as 99% spec for CAD-B, on average, two additional images were recalled for extra read, where 81% were true-positives.

4.5.2.2 *Insecure* readers

For CAD as second-reader for insecure predictions (reader prediction greater than 0, but lower than 0.5), the average reader scored 33 images in this range. 8 of the 33 images (24%) were incorrectly predicted as negatives. We display the results for each CAD setting in Table 4.3. In this table, we observe that higher CAD specificity modes resulted in a higher percentage of correct increases. By varying the specificity modes, we trade-off accuracy for recall. CAD-B has had a higher percentage of correct increases compared to CAD-A, while CAD-A+B outperformed CAD-B.

TABLE 4.2: Results for various CADs for second read false-negative reduction for certain readers. Here, CAD intervened to increase true-positives based on reader probabilities of 0.0. On average, 160 cases were scored using this probability.

CAD	Follow-up Baseline	Follow-up + CAD	Difference follow-up (%)	% CAD true-positives
CAD-A (90% spec, 36% sens)	160	176	10%	17%
CAD-A (94% spec, 31% sens)	160	170	6%	15%
CAD-A (98% spec, 24% sens)	**160**	**165**	**3%**	**26%**
CAD-B (86% spec, 35% sens)	160	182	14%	21%
CAD-B (93% spec, 34% sens)	160	172	8%	28%
CAD-B (97% spec, 32% sens)	160	167	4%	47%
CAD-B (99% spec, 29% sens)	**160**	**164**	**2%**	**64%**
CAD-A+B (88% spec, 50% sens)	160	179	12%	20%
CAD-A+B (93% spec, 47% sens)	160	173	8%	30%
CAD-A+B (98% spec, 36% sens)	160	164	3%	49%
CAD-A+B (99% spec, 31% sens)	160	163	2%	67%
CAD-A+B (99.5% spec, 22% sens)	**160**	**162**	**1%**	**80%**

TABLE 4.3: Results for various CADs for second read false-negative reduction for insecure readers. Here, CAD intervened to increase true-positives based on reader probabilities between >0.0 and <0.5. On average, 33 cases were scored using this probability interval.

CAD	Follow-up Baseline	Follow-up + CAD	Difference follow-up (%)	% CAD true-positives
CAD-A (90% spec, 36% sens)	33	37	12%	33%
CAD-A (94% spec, 31% sens)	33	35	6%	44%
CAD-A (98% spec, 24% sens)	**33**	**34**	**3%**	**75%**
CAD-B (86% spec, 35% sens)	33	39	18%	33%
CAD-B (93% spec, 34% sens)	33	37	13%	50%
CAD-B (97% spec, 32% sens)	33	36	9%	66%
CAD-B (99% spec, 29% sens)	**33**	**35**	**6%**	**81%**
CAD-A+B (88% spec, 50% sens)	33	39	18%	36%
CAD-A+B (93% spec, 47% sens)	33	37	12%	52%
CAD-A+B (98% spec, 36% sens)	33	36	8%	60%
CAD-A+B (99% spec, 31% sens)	33	35	6%	67%
CAD-A+B (99.5% spec, 22% sens)	**33**	**34**	**3%**	**69%**

Integration	Sensitivity	Specificity	Reading time	Change in workflow	Risk	Summary
Baseline						
Radiologist standalone	*69%*	*87%*	*23 seconds / case*		*Low*	
CAD standalone	*69%*	*37-50%*	*0 seconds*		*High*	
CAD as First Reader						
Filtering - normals	55-58%	91%	↓ 29-35%	None	High	Increase in missed nodules, but high risk.
Filtering - nodules	73-75%	82-83%	↓ 15-22%	None	Low	More nodule cases retrieved, but more false-positives forwarded.
Redistribute images						
Scenario I	53-54%	92-94%	100%	Low, distributes cases among radiologists.	Medium	Increase in missed nodules, but double read to optimize performance.
Scenario II	53%	92-93%	100%	Low, distributes cases among radiologists.	Medium	Increase in missed nodules, but double read to optimize performance.
Scenario III	76-79%	83-84%	100%	Low, distributes cases among radiologists.	Medium	More nodule cases retrieved, but double read should help reduce false-positives.
Scenario IV	76-78%	84%	100%	Low, distributes cases among radiologists.	Medium	More nodule cases retrieved, but double read should help reduce false-positives.
CAD as Concurrent Reader						
Prompt (CAD-A)	72%	86%	30 seconds / case	High. Can cause for significant increase in false-positives.	Low	Increase in nodule findings, but increased reading time

Interactive (CAD-A)	71%	*87%*	26 seconds / case	Medium	Low	Smaller increase in nodule findings, but slight increased reading time.
CAD as Second Reader						
False-positive reduction	55-58%	90-91%	↓ 5% (follow-up CT)	None	Medium	Reduces follow-up CT, but risk of reducing nodule cases.
False-negative reduction						
Confident scenario	71-74%	*82%*	↑ 8-11% (follow-up CT)	None	Low	Increased nodule recall, but also increased unnecessary follow-up CT .
Insecure scenario	71-72%	*86%*	↑ 6-12% (follow-up CT)	Medium, readers can score images as 'uncertain'.	Low	Acts on the indication of doubt of the readers. Increased nodule recall, but also increased unnecessary follow-up CT.

Chapter 5

Discussion

In this thesis, we show various implementation strategies for CAD by evaluating numerous metrics against the task of pulmonary nodule detection on chest radiographs. As a baseline, we used the radiologists mean specificity and sensitivity and compared the results for each scenario against the performance of radiologists without any use of CAD. All evaluated CADs were not directly able to achieve performance similar to the readers. The points on the ROC curve of standalone CAD do not approximate the points on the ROC curve that readers achieve. However, CAD tools can be tuned to be equally or more specific/sensitive towards nodule detection, whereas for readers, this is impossible.

For the evaluated combination scenarios of CAD plus readers, we observed a trade-off between specificity and sensitivity, or a trade-off between specificity - sensitivity and reading time. Depending on the use case, CAD as first reader can be more beneficial compared to CAD as a second reader. To determine which scenarios are best suitable, it is necessary to look at the current hospital scenario, as each hospital setting is different and requires a personalized approach. Important questions need to be asked: what additional reading time would you allow the radiologists in order to achieve additional performance gains? Do we need to save on reading time? Or do we need to reduce the number of false-positives? As an example answer, in some countries, the resources for radiologists are scarce, such as Kenya (Van't Hoog et al., 2011). Saving reading time would be more beneficial compared to optimizing the performance. It has been shown that CAD tools can aid the detection of diseases in such scenarios (Melendez et al., 2016), and improve the patient outcome for a larger population.

Next to implementation strategies, we hypothesized to achieve better performance for modern deep-learning-based CAD-B versus traditional computer vision algorithm CAD-A. However, although Nam et al., 2019 reports an AUC score of 0.92, such AUC scores were not seen in our research. Due to the low standalone AUC score of CAD-B, we conclude that, on this dataset for nodule detection, modern CAD was not able to achieve performance that is on par with radiologists yet. Also, for most scenarios, CAD-B showed no significant improvements over CAD-A. By combining both models, resulting in model ensemble CAD-A+B, AUC scores increase from 0.658 and 0.586 to 0.708, respectively. The use of two CAD systems simultaneously results in better performance, which is in line with the reasoning of the difference in performance between single and double reading.

5.1 Scenarios

5.1.1 CAD as First Reader

Filtering of *normals* comes with a high risk due to the reduced sensitivity, and therefore missing nodules. However, an increased specificity is observed for all CADs. In return, a significantly reduced reading time is observed (∼34%). This scenario is useful when reading time is costly, and when the cost of misdiagnosing is more expensive than missing a diagnosis (i.e., unnecessary expensive/dangerous surgery).

Filtering of *nodules* for CAD as first reader causes a significant increase in nodule recall, but specificity decreases. It is considered low risk because compared to the baseline, more patients get forwarded for follow-up CT. Reading time is reduced by ∼18%. It can be argued that a high false-negative rate introduced by the reference readers is worse compared to the number of false-positives that CAD filtering introduces, as the consequences of these classification differences deteriorate patient outcomes.

5.1.1.1 Image redistribution

The first two scenarios (I and II), where *normals* are filtered, are not feasible due to the poor sensitivity performance of the CADs. Both scenarios show lower AUC scores compared to the baseline single reader (0.76 vs 0.83 AUC). Also, the observed sensitivity is lower compared to the regular *normals* filtering approach without the redistribution of image reads (55-58% versus 53-54%).

For scenarios III and IV, where *nodules* are filtered, redistribution of image reads is beneficial while reading time remains equal to the baseline scenario. Here, the double-reading strategy causes a significant increase in sensitivity (69% to 77%), while the cost of a lower specificity in return is minimal (87% to 84%). There seems to be no difference in performance between scenario III and IV, meaning that applying a double-read strategy to any particular image would already show a benefit.

5.1.2 CAD as Concurrent Reader

CAD as concurrent reader positively affects the reader in its decision, as the sensitivity increases from 69% to 72%. The specificity slightly deteriorates: from 87% to 86%. The reading time increases from 23 to 30 seconds per case for *prompting* mode, whereas *interactive* mode has slightly less increase in reading time: 26 seconds per case. This can be attributed to the fact that readers now need to double-check the proposal location of CAD that readers might have missed in the first case, and judge whether the detection is a true-positive. For the *interactive* mode, the CAD shows fewer false-positives that could distract the reader. The risk factor for this scenario is lowest, as the reader has the final call regarding the clinical decision (Oakden-Rayner, 2019). The level of change in workflow is high, as the readers require training in order to understand the limitations of CAD. Consecutively, adoption rates for this scenario might decrease (Werth and Ledbetter, 2020).

Automation bias For CAD as concurrent reader, the readers were affected by automation bias. Because CAD-A showed poor specificity, one could assume that the absence of a CAD prediction might be more informative to the readers than the presence. As a result, readers were less likely to recall a nodule when CAD did not detect a nodule. This effect was also observed in related work (Alberdi et al., 2004).

A reader could also get distracted with false-positive nodule marks, and would be more prone to miss nodules in other areas of the image (Philpotts, 2009). In order for readers to better understand automation bias, it is important to understand the limitations of CAD. Future work could address the effects of automation bias by, e.g. evaluating the opinion of readers on CAD using a qualitative measure such as a questionnaire afterwards.

5.1.3 CAD as Second Reader

CAD as second reader is only beneficial in some cases. When applying second reader CAD for false-positive reduction, the same performance is achieved compared to CAD as first reader, where CAD is responsible for filtering *normals*. Then, compared to first reader CAD, this scenario provides no other benefits such as additional savings in reader time. Thus, if we prefer a scenario where filtering of *nodules* has priority, we should apply the CAD as first reader. However, recent research suggests that CAD as second reader to reduce false-positives can be beneficial ($\sim$42% reduction where 85% of the filtered cases were successful reductions) (Zelst et al., 2020).

For false-negative reduction and confident readers, we observe a higher sensitivity (69% to 71-74%) but decreased specificity (87% to 82%). We further divided the reader group into an additional category: *insecure* readers. The results of the *insecure* reader group were promising, as the same increase in sensitivity was observed (71-72%), while at the same time the specificity was on par with standalone readers (87%). This shows that when we provide readers with an option to be unsure about an image, CAD can help to resolve this doubt successfully with no additional drawbacks.

Another point to consider is the desired specificity setting for the CAD. When using a higher specificity, the relative percentage of true-positives of all the CAD filtered images using this specificity increases at the same time. However, the trade-off between quality and quantity of filtering shows here as well. We can argue how many additional reads we allow for follow-up CT, with the possibility that the referred patient does not, in fact, has a nodule.

During the simulations, we saw that if we allow for 12% extra CT referrals, $\sim$50% of these referrals contain a true nodule (for CAD-B and CAD-A+B). Translated to a generic setting: we trade-off 12% extra work for a 50% increased detection in these patients.

5.2 Limitations

5.2.1 Retrospective study

As this study was done retrospectively, we cannot confirm the performance in a true clinical setting. A prospective study repeating the approaches is therefore needed to validate the workings in clinical setting (Iussich et al., 2014). This would provide a robust test for the scenarios in a clinical setting and having the possibility to qualitatively study the radiologist for its opinion regarding the change in the clinical workflow.

5.2.2 Model ensemble - CAD-A+B

When we ensembled CAD-A and CAD-B in order to come up with the ensemble CAD-A+B, we needed to combine the scores in some way. Because the distribution

of scores of both algorithms did not line up correctly (Figure 4.4), directly taking the mean of the scores was not ideal. For two similar algorithms, we expected the distribution of scores to resemble two gaussians, one at each tail: one gaussian for positive predictions at the upper prediction scores, one gaussian for negative predictions around lower prediction scores. In reality, this was not the case. For future work, We might need to look into methods in order to align these distributions (i.e., forms of standardization).

5.2.3 Generalizability

It remains a challenge whether the results of the current study will generalize towards other datasets and CAD solutions. At least two factors affect this: prevalence and generalizability of CAD. A different prevalence setting causes a significantly different performance due to the available number of positives. The generalizability of CAD in itself can already be attributed to the development phase, and not necessarily the implementation phase. When the training dataset of CAD does not include the complete population, we expect CAD to perform differently on hospital data that was not captured in this population.

5.2.4 Prevalence

Determining the prevalence rate of nodules is challenging. For our scenario, we accumulated samples with a prevalence rate of 37%. Generally, a lower prevalence ($\sim$0.1%) is seen during clinical or screening situations (Lee et al., 2020). At the same time, the prevalence rate of malignant versus benign nodules is even lower, and nodule features tend to differ from screening situations and clinical situations (Wahidi et al., 2007) (Wang et al., 2014). Due to high variations of prevalence over different studies, there seems to be no rule regarding which prevalence rate to use. As seen in Dembrower et al., 2020, one way to achieve a lower prevalence rate would be to resample additional negative nodule patients.

We suspect that when we repeat the experiments using lower prevalence rates, the scenarios responsible for filtering *normals* will become more effective compared to the filtering of *nodules* due to the abundance of *normal* images. Also, the performance of readers can be taken into question, as readers tend to miss nodules when prevalence rates are lower (Evans; Birdwell, and Wolfe, 2013). Upon the start of the reader study, readers were informed of the high prevalence setting, and that half of the cases contained a nodule.

5.2.5 Dichotomized study

During our experiments, we evaluated nodule detection performance by comparing normal chest radiographs against normal chest radiographs containing a nodule (dichotomization). During real-world clinical setting, various comorbidities can be present (both where a nodule might be present/absent), allowing for external factors that can distract the readers and miss the detection of nodules. Therefore, the reading time metric did not capture the detection of additional comorbidities.

5.2.6 Visibility of nodules

The current requirements of the study design focus on the importance of visible nodules on chest radiographs. However, in a real-world setting, some nodules can be observed on CT but not on chest radiographs. It would be interesting to include

these chest radiographs as positives in our dataset, and observe whether CAD is able to obtain these.

5.2.7 CAD-B

AUFROC scores A metric that takes into account not only whether or not a nodule is detected but also includes the location prediction is called the AUFROC score. In order to apply the AUFROC metric, we need to match the location of the detection with the reference standard. For CAD-A, the location is known, but for CAD-B, the location of the detection was not indicated. Thus, it could mean that the found nodule represents an area that is not the actual location of a nodule. Therefore, we stick to the AUCROC score instead.

Generalizability of CAD-B Using the threshold of 0.30 as reported by Nam et al., 2019, CAD-B achieves 95.2% specificity and 80.7% sensitivity, and a 100% recall for larger nodules on their dataset. However, the results on our dataset using this threshold states differently: 97% specificity and 32% sensitivity, and a 66% recall for large nodules using the high-sensitivity threshold. Although these results are significantly different, we still do not have an explanation of why this is the case.

5.3 Future work

For future work, it would be of interest to evaluate another (better) CAD tool that shows performance on par with readers. We hypothesize that it would magnify the shown performance gains for each scenario. Next, we should consider extending our evaluation dataset and reduce the prevalence to a realistic clinical setting, and test these implementations in a true hospital scenario. Further, we can choose to test these implementations in a true hospital scenario in a routine setting. Finally, a cost-effectiveness study can help gain insights in the cost/benefit trade-off; a metric that we have not yet evaluated.

Appendix A

ROCs

A.1 ROC of individual readers + CADs

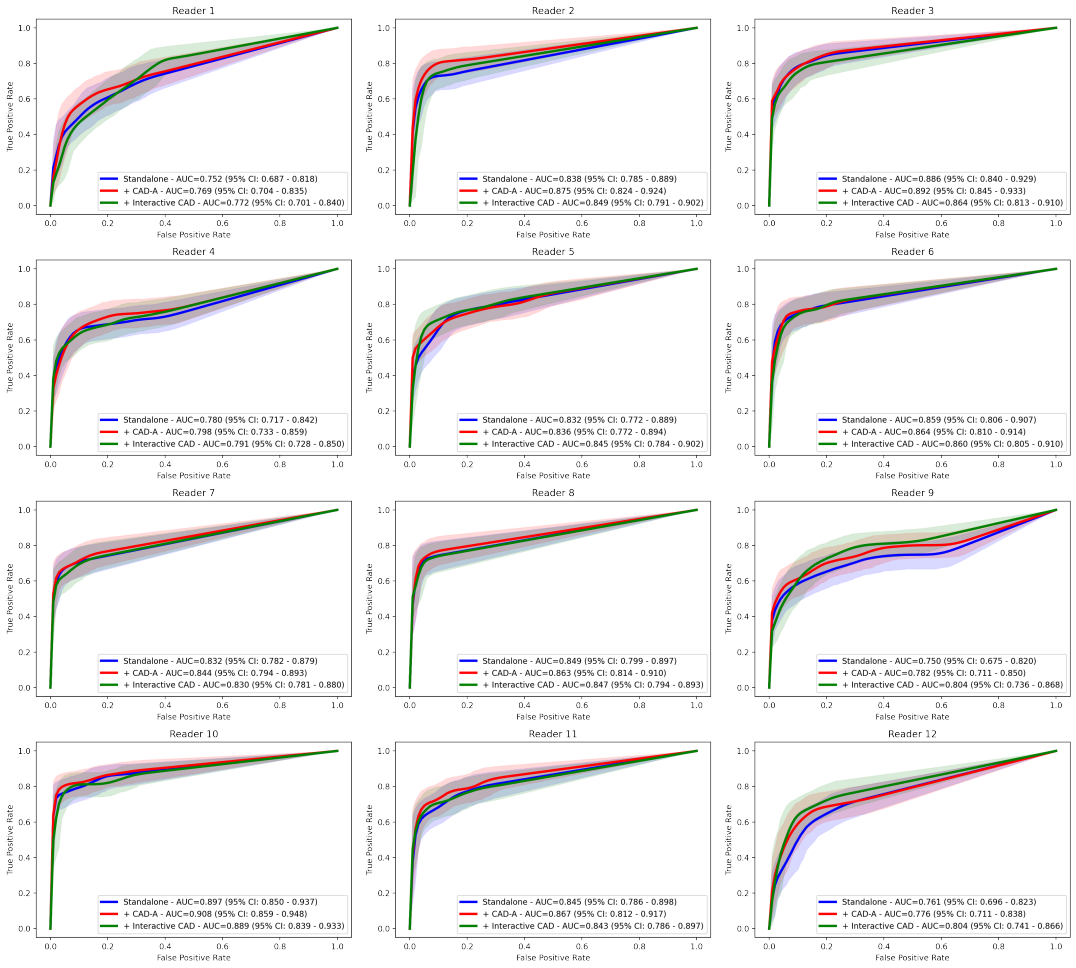


FIGURE A.1: ROC of standalone vs standalone + CAD vs standalone + interactive CAD. In general, the concurrent use of CAD tools help to increase performance.

A.2 Averaging bootstrapped ROC curves

It is not wise to choose a model based on a single ROC curve or AUC score. This would be the same as choosing the model with the best performance on one particular dataset, which introduces the possibility of choosing the model that overfits

most. Variance should be introduced by bootstrapping or cross-validation of the dataset to be tested (Fawcett, 2006). This way, one can determine which model is best based on multiple variations of a dataset where we prefer generalizability.

For some scenarios, it was necessary to average the ROC curves for each CAD in order to visualize the differences in scenarios better. This required an approach that averages the ROC curves while taking into account the bootstrapped confidence interval scores. Averaging ROC curves in itself comes with a challenge, as the thresholds per ROC differ together with the retrieved false-positives. Thus, we need to interpolate to obtain new points on the ROC. There are generally two ways into averaging ROC curves (Fawcett, 2006):

1. Vertical averaging. By fixing the false-positive rate, we average over the true positive rates. This is generally applied when a one-dimensional measure of variation is desired.

2. Threshold averaging. This relies on the underlying threshold that causes the points on the ROC (i.e., false-positive and true-positive rate).

We implemented and added the vertical averaging approach to the *evalutils* package, which can be found here.

Appendix B

Results table

We show the complete results table covering all (hypothesized) scenarios and individual CAD performance below.

Integration	Sensitivity	Specificity	Reading time	Change in workflow	Risk	Summary
Baseline						
Radiologist standalone	*69%*	*87%*	*23 seconds / case*		*Low*	
CAD standalone						
(CAD-A)	*69%*	*50%*				
(CAD-B)	*69%*	*37%*	*0 seconds*		*High*	
(CAD-A+B)	*69%*	*47%*				
CAD as First Reader						
Filtering - normals						
(CAD-A 73% sens)	58%	91%	↓ 35%			
(CAD-B 73% sens)	55%	91%	↓ 29%	None	High	Increase in missed nodules, but high risk.
(CAD-A+B 73% sens)	58%	91%	↓ 39%			
Filtering - nodules						
(CAD-A 94% spec)	73%	83%	↓ 15%			
(CAD-B 94% spec)	75%	82%	↓ 17%	None	Low	More nodule cases retrieved, but more false-positives forwarded.
(CAD-A+B 94% spec)	75%	82%	↓ 22%			
Redistribute images						
Scenario I						
(CAD-A 74% sens)	54%	93%				
(CAD-B 74% sens)	53%	92%	100%	Low, distributes cases among radiologists.	Medium	Increase in missed nodules, but double read to optimize performance.
(CAD-A+B 74% sens)	54%	94%				
Scenario II						
(CAD-A 74% sens)	53%	93%				
(CAD-B 74% sens)	53%	92%	100%	Low, distributes cases among radiologists.	Medium	Increase in missed nodules, but double read to optimize performance.
(CAD-A+B 74% sens)	53%	93%				

Scenario III (CAD-A 93% spec) (CAD-B 93% spec) (CAD-A+B 93% spec)	76% 76% 79%	84% 83% 83%	100%	Low, distributes cases among radiologists.	Medium	Mode nodule cases retrieved, but double read should help reduce false-positives.
Scenario IV (CAD-A 93% spec) (CAD-B 93% spec) (CAD-A+B 93% spec)	76% 77% 78%	84% 84% 84%	100%	Low, distributes cases among radiologists.	Medium	Mode nodule cases retrieved, but double read should help reduce false-positives.
Triage			100%	Low, estimated high-risk cases first.	Low	Faster diagnosis and possibly time-to-treatment.
CAD as Concurrent Reader						
Prompt (CAD-A)	72%	86%	30 seconds / case	High. Can cause for significant increase in false-positives.	Low	Increase in nodule findings, but increased reading time.
Interactive (CAD-A)	71%	*87%*	26 seconds / case	Medium	Low	Smaller increase in nodule findings, but slight increased reading time.
Hybrid (show prompts for confident predictions)	*Increase?*	*Increase?*	*Mix between prompt and interactive?*	*Medium-high. Increases FP rate, but less than prompt.*	*Low*	
Similar case retrieval	*Equal?*	*Increase?*	*Increase?*	*High*	*Low*	
CAD as Second Reader						
False-positive reduction (CAD-A 74% sens) (CAD-B 74% sens) (CAD-A+B 74% sens)	58% 55% 58%	90% 91% 91%	↓ 5%	None	Medium	Reduces follow-up CT, but risk of incorrectly reducing nodule cases.

False-negative reduction						
Confident scenario			↑ +8%			
(CAD-A 94% spec)	71%	*82%*	↑ +10%	None	Low	Increased nodule recall, but also increased unnecessary follow-up CT.
(CAD-B 94% spec)	73%	*82%*	↑ +11%			
(CAD-A+B 94% spec)	74%	*82%*	(follow-up CT)			
Insecure scenario			↑ +6%			
(CAD-A 94% spec)	71%	*86%*	↑ +12%	Medium, readers can score images as 'uncertain'.	Low	Acts on the indication of doubt of the readers. Increased nodule recall, but also increased unnecessary follow-up CT.
(CAD-B 94% spec)	72%	*86%*	↑ +12%			
(CAD-A+B 94% spec)	72%	*86%*	(follow-up CT)			
CAD as 'final call' when uncertainty among radiologists	*Increase?*	*Increase?*	*Increase*	*Medium, third reader resolves conflict*		*Third read by a radiologist, possibly increased performance but also reading time.*

Bibliography

Abadi, Martín et al. (2016). "Tensorflow: A system for large-scale machine learning". *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283.

Alberdi, Eugenio et al. (2004). "Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography". *Academic radiology* 11.8, pp. 909–918.

Ardila, Diego et al. (2019). "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography". *Nature medicine* 25.6, pp. 954–961.

Austin, JH; Romney, BM, and Goldsmith, LS (1992). "Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect." *Radiology* 182.1, pp. 115–122.

Beyer, F et al. (2007). "Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader". *European radiology* 17.11, pp. 2941–2947.

Ciatto, S et al. (2005). "Second reading of screening mammograms increases cancer detection and recall rates. Results in the Florence screening programme". *Journal of medical screening* 12.2, pp. 103–106.

Dembrower, Karin et al. (2020). "Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study". *The Lancet Digital Health* 2.9, e468–e474.

Evans, Karla K; Birdwell, Robyn L, and Wolfe, Jeremy M (2013). "If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening". *PloS one* 8.5, e64366.

Fawcett, Tom (2006). "An introduction to ROC analysis". *Pattern recognition letters* 27.8, pp. 861–874.

Fenton, Joshua J et al. (2007). "Influence of computer-aided detection on performance of screening mammography". *New England Journal of Medicine* 356.14, pp. 1399–1409.

Fryback, Dennis G and Thornbury, John R (1991). "The efficacy of diagnostic imaging". *Medical decision making* 11.2, pp. 88–94.

Fujita, Hiroshi (2020). "AI-based computer-aided diagnosis (AI-CAD): the latest review to read first". *Radiological Physics and Technology* 13.1, pp. 6–19.

Gao, Yiming et al. (2019). "New frontiers: an update on computer-aided diagnosis for breast imaging in the age of artificial intelligence". *American Journal of Roentgenology* 212.2, pp. 300–307.

Geras, Krzysztof J; Mann, Ritse M, and Moy, Linda (2019). "Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives". *Radiology* 293.2, pp. 246–259.

Gilbert, Fiona J et al. (2008). "Single reading with computer-aided detection for screening mammography". *New England Journal of Medicine* 359.16, pp. 1675–1684.

Goldenberg, Roman and Peled, Nathan (2011). "Computer-aided simple triage". *International journal of computer assisted radiology and surgery* 6.5, p. 705.

Halligan, Steve; Altman, Douglas G, and Mallett, Susan (2015). "Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: a discussion and proposal for an alternative approach". *European radiology* 25.4, pp. 932–939.

Hoop, Bartjan de et al. (2010). "Computer-aided detection of lung cancer on chest radiographs: effect on observer performance". *Radiology* 257.2, pp. 532–540.

Hsu, William and Hoyt, Anne C (2019). "Using Time as a Measure of Impact for AI Systems: Implications in Breast Screening". *Radiology: Artificial Intelligence* 1.4, e190107.

Hupse, Rianne et al. (2013). "Computer-aided detection of masses at mammography: interactive decision support versus prompts". *Radiology* 266.1, pp. 123–129.

Iussich, Gabriella et al. (2014). "Computer-aided detection for computed tomographic colonography screening: a prospective comparison of a double-reading paradigm with first-reader computer-aided detection against second-reader computer-aided detection". *Investigative radiology* 49.3, pp. 173–182.

Kim, Hyo-Eun et al. (2020). "Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study". *The Lancet Digital Health* 2.3, e138–e148.

Kluge, Ruben MW (2020). "Pneumothorax Detection On Chest Radiographs: A Comparative Analysis Of Public Datasets, Deep Learning Architectures, And Domain Adaptation Via Iterative Self-Training". MA thesis.

Lampeter, William A (1985). "Computer-Aided Detection of Pulmonary Nodules". *Computer Assisted Radiology/Computergestützte Radiologie*. Springer, pp. 502–506.

Lee, Jong Hyuk et al. (2020). "Performance of a Deep Learning Algorithm Compared with Radiologic Interpretation for Lung Cancer Detection on Chest Radiographs in a Health Screening Population". *Radiology*, p. 201240.

Lin, Chi Y and Levary, Reuven R. (1989). "Computer-aided software development process design". *IEEE Transactions on Software Engineering* 15.9, pp. 1025–1037.

Litjens, Geert et al. (2017). "A survey on deep learning in medical image analysis". *Medical image analysis* 42, pp. 60–88.

Liu, Xiaoxuan et al. (2019). "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis". *The lancet digital health* 1.6, e271–e297.

Mani, Aravind et al. (2004). "Computed tomography colonography: feasibility of computer-aided polyp detection in a "first reader" paradigm". *Journal of computer assisted tomography* 28.3, pp. 318–326.

Mayo, Ray Cody et al. (2019). "Reduction of false-positive markings on mammograms: a retrospective comparison study using an artificial intelligence-based CAD". *Journal of Digital Imaging* 32.4, pp. 618–624.

Melendez, Jaime et al. (2016). "An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information". *Scientific reports* 6, p. 25265.

Muramatsu, Chisako et al. (2010). "Presentation of similar images as a reference for distinction between benign and malignant masses on mammograms: analysis of initial observer study". *Journal of digital imaging* 23.5, pp. 592–602.

Nam, Ju Gang et al. (2019). "Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs". *Radiology* 290.1, pp. 218–228.

Nishikawa, Robert M and Bae, Kyongtae T (2018). "Importance of better human-computer interaction in the era of deep learning: mammography computer-aided

diagnosis as a use case". *Journal of the American College of Radiology* 15.1, pp. 49–52.

Oakden-Rayner, Luke (2019). *The rebirth of CAD: how is modern AI different from the CAD we know?*

Owais, Muhammad et al. (2019). "Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence". *Journal of clinical medicine* 8.4, p. 462.

Parikh, Rajul et al. (2008). "Understanding and using sensitivity, specificity and predictive values". *Indian journal of ophthalmology* 56.1, p. 45.

Philpotts, Liane E (2009). "Can computer-aided detection be detrimental to mammographic interpretation?" *Radiology* 253.1, pp. 17–22.

Qin, Zhi Zhen et al. (2020). "Can artificial intelligence (AI) be used to accurately detect tuberculosis (TB) from chest x-ray? A multiplatform evaluation of five AI products used for TB screening in a high TB-burden setting". *arXiv preprint arXiv:2006.05509*.

Sagi, Omer and Rokach, Lior (2018). "Ensemble learning: A survey". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8.4, e1249.

Samulski, Maurice et al. (2010). "Using computer-aided detection in mammography as a decision support". *European radiology* 20.10, pp. 2323–2330.

Schalekamp, S et al. (2014a). "New methods for using computer-aided detection information for the detection of lung nodules on chest radiographs". *The British journal of radiology* 87.1036, p. 20140015.

Schalekamp, Steven et al. (2014b). "Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images". *Radiology* 272.1, pp. 252–261.

Sendak, Mark P et al. (2020). "Presenting machine learning model information to clinical end users with model facts labels". *NPJ Digital Medicine* 3.1, pp. 1–4.

Van't Hoog, AH et al. (2011). "High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey". *The International journal of tuberculosis and lung disease* 15.10, pp. 1308–1314.

Wahidi, Momen M et al. (2007). "Evidence for the treatment of patients with pulmonary nodules: when is it lung cancer?: ACCP evidence-based clinical practice guidelines". *Chest* 132.3, 94S–107S.

Wang, Yi-Xiang J et al. (2014). "Evidence based imaging strategies for solitary pulmonary nodule". *Journal of thoracic disease* 6.7, p. 872.

Werth, Kyle and Ledbetter, Luke (2020). "Artificial Intelligence in Head and Neck Imaging: A Glimpse into the Future". *Neuroimaging Clinics of North America* 30.3, pp. 359–368.

Yassin, Nisreen IR et al. (2018). "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review". *Computer methods and programs in biomedicine* 156, pp. 25–45.

Zelst, Jan CM van et al. (2020). "Validation of radiologists' findings by computer-aided detection (CAD) software in breast cancer detection with automated 3D breast ultrasound: a concept study in implementation of artificial intelligence software". *Acta Radiologica* 61.3, pp. 312–320.

ADDENDUM

17 November 2020

After the thesis was written, we found in collaboration with the vendor of the *CAD-B* software that the data anonymization process performed by the research group caused the software to malfunction for a subset of the data.

The results of *CAD-B* have consequently been improved beyond the performance reported in the thesis. *CAD-B* results in this thesis should therefore be interpreted as the results of a hypothetical CAD system, making the overall simulation results remain valid.