Bachelor's Thesis Computing Science

Radboud University Nijmegen

Optimising Lung Cancer Detection by Oversampling Malignant Nodules

---

Author:
Jen Dusseljee
s1003489

First assessor:
Dr. ir. Colin Jacobs

Daily supervisor:
Lars Leijten

Second assessor:
Prof. Tom Heskes

January 18, 2024

Abstract

Lung cancer is the leading cause of cancer-related deaths worldwide because it is often diagnosed when the disease has already progressed to later stages, with poor treatment outcomes. Methods for detecting lung cancer in its early stages involve finding malignant nodules in lung CT scans, either in screenings or as incidental findings in other clinical procedures. Both of these methods rely on the work of radiologists, who have an increasingly high workload. Because of this, recent research has focused on using AI systems to help detect these lung nodules. This research has shown that it can be an effective tool but is not yet at the radiologists' level when it comes to detecting primary lung cancers.

In this thesis, we investigate whether we can improve the performance of an existing system at detecting malignant lung nodules. The specific method we will investigate is the application of oversampling on malignant nodules in the training process. We compared different methods for determining which nodules to oversample and validated our results on two screening and three clinical datasets. Our research indicates that oversampling is effective at improving the system's performance on malignant nodules in one of our screening datasets but is inconclusive about a similar effect in the other screening datasets or in clinical datasets.

Contents

1 Introduction **3**

2 Preliminaries **5**
2.1 Computer Vision 5
2.1.1 Neural Networks 5
2.1.2 Convolutional Neural Networks 6
2.1.3 Training Neural Networks 7
2.1.4 Residual Networks 8
2.1.5 Object Detection using YOLO 9
2.2 Lung Cancer 10
2.2.1 Pulmonary Nodules 11
2.2.2 Detecting Pulminary Nodules 12

3 Related Work **14**
3.1 Automated pulmonary nodule detection 14
3.2 Public challenges 14
3.2.1 LUNA16 14
3.2.2 DSB2017 15
3.3 Automated nodule detection in non-screening CT scans . . . 15

4 Methods **18**
4.1 Retraining 18
4.1.1 YOLOv5 Candidate Detection 18
4.1.2 ResNet50 False Positive Reduction 18
4.2 Oversampling 19
4.2.1 Determining malignancy 19
4.3 Experiments 20
4.4 Validation 20
4.4.1 Datasets 20
4.4.2 Evaluation 22

5 Results **24**
5.1 Retraining without oversampling 24
5.2 Retraining with oversampling 25

5.3 Characteristics of missed nodules 27

6 Discussion **32**
6.1 Future work 33
6.2 Conclusions 34

A Training data distributions **41**

B Detailed baseline results **43**
B.1 Second retraining 43
B.2 Individual raters 43
B.3 MILD: earliest vs latest scan 43

Chapter 1

Introduction

Lung cancer is one of the leading causes of preventable mortality worldwide and the leading cause of cancer-related deaths[1]. One of the biggest reasons for this is that it is often detected in a late stage, where cancer has already progressed and five-year survival rates are low. Because of this, early detection of lung cancer is crucial to improve survival rates. In the first stage of lung cancer, when treatment outcomes are most optimistic[2, 3], it is present in the form of a malignant pulmonary (lung) nodule and does not usually cause symptoms and is thus often overlooked until the disease progresses to become symptomatic. These nodules can be found as incidental findings on CT scans in different clinical routines or in targeted lung cancer screening programs of at-risk populations. Because of the implementation of more of these screenings in several countries and the increasing use of CT scans in other clinical routines, the demand for and workload of radiologists has been increasing over time. This fact, combined with the fact that pulmonary nodules can be very small and thus hard to detect on CT scans, makes for a challenging problem.

In recent years, the rise of deep learning has been a promising new approach to helping address this problem. Several studies have demonstrated the effectiveness of computer-aided detection (CAD) systems that can help radiologists find pulmonary nodules in both screenings and clinical settings. One of these systems is the nodule detection system described in Hendrix et al.[4]. This system has been validated to perform as well or better than radiologists at detecting benign nodules and metastases but underperforms compared to radiologists in the detection of primary cancers. A possible reason behind this is the fact that a large majority of nodules that the system is trained on are benign, while malignant nodules are underrepresented in the training data. We propose to oversample malignant nodules during training to balance the training data to solve this underrepresentation. We hypothesize that this will improve the system's performance on primary cancers.

Our research thus aims to answer the question, "Can oversampling malignant nodules be a useful tool for improving the performance of a pulmonary nodule detection system on primary cancers in clinical datasets?". To answer this question, we will first retrain the system and evaluate its performance on several datasets, including the clinical datasets from the original paper, as well as new screening datasets that contain far more nodules. Next, we will perform three experiments where we oversample nodules in the training data, using different metrics for which nodules to oversample. We will evaluate the performance of our system after these experiments and compare this with the performance of the retrained system without oversampling to find out if oversampling affected our system's performance as a whole and on (primary) cancers specifically.

Previous work already exists that evaluates the effectiveness of oversampling minority classes in classification tasks on pulmonary nodules, as well as research that evaluates oversampling nodules compared to non-nodules for detection tasks on pulmonary nodules[5]. To the best of our knowledge, this research is the first to evaluate the effectiveness of oversampling a specific type of nodule for detection tasks on pulmonary nodules.

Chapter 2

Preliminaries

2.1 Computer Vision

2.1.1 Neural Networks

Artificial Neural Networks[6] (from now on referred to as Neural Networks or NNs) are algorithms that mimic the way the human brain processes information. Where the human brain consists of many neurons feeding information into each other through axons to perform complex decision-making, NNs consist of artificial neurons feeding into each other to perform complex computations. Each neuron takes multiple inputs, multiplies each input x_i with a respective weight w_i, takes the sum of these and a bias b, and performs an activation function f over this result to produce output y. A neuron can thus be described with the function in figure 2.1.

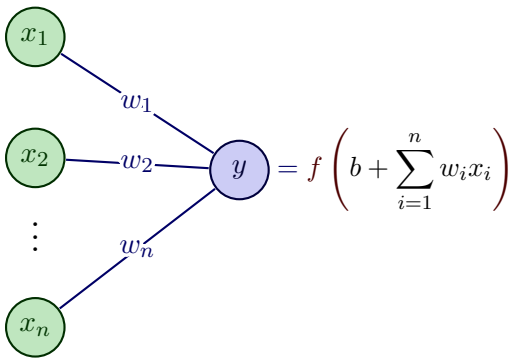


Figure 2.1: A single neuron computation.

These neurons are grouped into layers to form a network, where the outputs of the neurons in layer n are the inputs of the neurons in layer $n + 1$. A complete NN consists of an input layer that can represent the input data, hidden layers that perform most of the complex computations, and an output layer that can represent the output data.

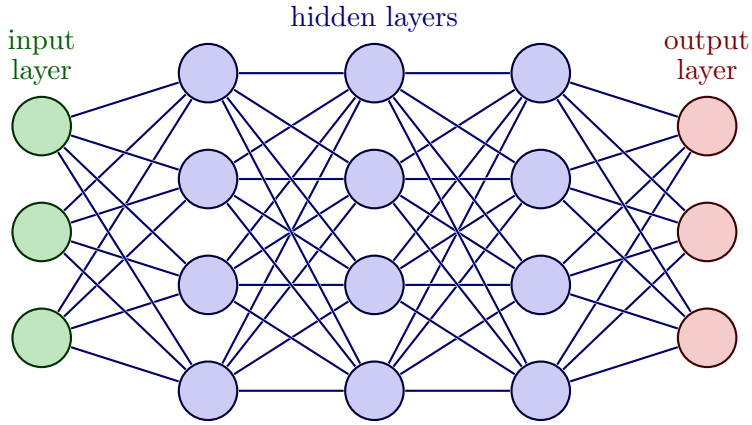


Figure 2.2: A Neural Network with three hidden layers.

2.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a type of Neural Network that is particularly suitable for Computer Vision tasks, as demonstrated by AlexNet in 2012[7]. They generally consist of three kinds of layers: convolutional layers, pooling layers, and fully-connected layers.

Convolutional layers

CNNs use so-called *convolutions* to efficiently detect features (e.g., edges, curves, textures) in images. A 2D convolution is an operation where we apply a filter (or kernel) representing a specific feature to an input image to produce a feature map. Each pixel in this feature map represents how much that area in the input image resembles the feature of the filter. This concept can be translated to apply to data with different dimensionality.

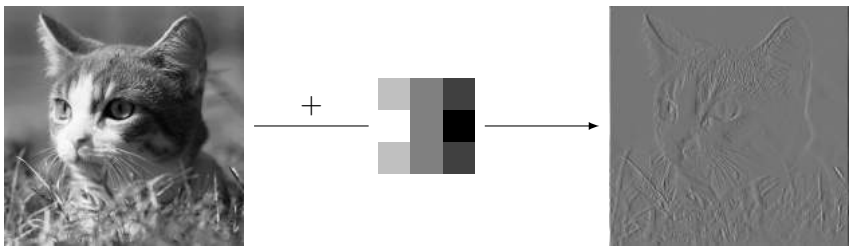


Figure 2.3: A 3x3 convolution.

A 2D convolutional layer contains multiple filters of the same size, each detecting a different feature. The layer outputs a feature map with a channel for each filter. For the first layers, they usually look for simple, abstract features like edges, patterns, or colors. For higher convolutional layers, the

features they look for usually become more complex and concrete.

For example, the first convolutional layer might recognize specific edges, curves, or colors. Deeper convolutional layers could then recognize that these features make up an ear, eye, or nose.

Pooling layers

A pooling layer divides the input to the layer up into small patches and produces a single value for each patch, usually taking either the highest value in the patch or the average of all values in the patch. For example, a three-by-three max-pooling layer divides the input into patches of three by three pixels and outputs the highest value for each patch.

The primary purpose of pooling layers is to reduce the model's computational complexity while preserving spatial hierarchy. It also helps to prevent overfitting, where the model focuses too much on specific properties found in training data instead of generalizing to broader features.

Fully connected layers

Fully connected layers are the layers described in 2.1.1. They are usually the last layers in a CNN, where they use the features extracted in previous layers to make complex decisions.

2.1.3 Training Neural Networks

Besides defining the network architecture, the training process is an essential part of creating neural networks. Training a neural network means finding the optimal configuration of trainable parameters for the network to perform its task as well as possible. In the case of fully connected layers, these trainable parameters are the weights and biases for all the neurons. For convolutional layers, these trainable parameters are the filters the layer uses.

To train a model, we use a training dataset. A set of inputs with an expected output for each input. We then perform the following steps for each input:

1. We perform what is called a "forward pass", feeding the input into the model to get a prediction from the model.

2. We use a loss function to measure the difference between the expected output and the model's prediction. This difference is called the loss. The goal of training is to minimize this loss.

3. Using this loss in a process called backpropagation, we calculate the gradient for each parameter in the network. This gradient tells us in which direction and how much we should change this parameter

to reduce the loss. If we change all parameters according to their gradient, the model's output should be closer to the expected output.

4. Using the calculated gradients, an algorithm called the optimizer calculates new values for all the parameters and updates them. This optimizer ensures that each parameter gradually moves towards a value, such that the whole model generalizes to make accurate predictions on all training data instead of just performing well on the input example presented in this iteration. The speed at which the optimizer updates the parameters based on the gradients is called the *learning rate*.

Performing these steps for all the values in the training dataset is referred to as one *epoch*. During training, as we perform more epochs, the algorithm's performance converges over time. If this process continues for too many epochs, there is a risk that the model will overfit on the training data, where it starts to learn specific features of samples in the training data and starts to perform worse on new samples that do not exist in the training dataset.

The performance of the trained model generally depends on the model's architecture, the quality of the training set, the loss function, and the optimizer used, as well as so-called *hyperparameters* like the learning rate and the number of epochs used.

2.1.4 Residual Networks

In theory, adding more layers to a convolutional neural network should improve the model's performance because the model can perform more complicated calculations. In practice, however, we find that as the number of layers increases past a certain point, the model's performance actually decreases. This is counterintuitive because we would expect a model with more layers to perform at least as well as a model with fewer layers, as the additional layers could theoretically learn the identity function, leaving the outputs of the previous layers unchanged. Because of this performance decrease beyond a certain number of layers, there is a limit to how many layers we can add, which limits how complex a model's calculations and decision-making can be.

Residual Blocks

Residual networks, first introduced in[8], solve this issue by introducing the concept of residual blocks. In a regular neural network, each layer is responsible for learning a mapping $\mathcal{H}(x)$ from the inputs of that layer to the desired outputs of that layer. In a residual network, each residual block is responsible for learning a *residual* mapping $\mathcal{F}(x) = \mathcal{H}(x) - x$, meaning the *difference* between the inputs and desired outputs of the block. This task

is much easier, and mappings like the identity function become very easy to learn since $\mathcal{H}(x)$ simply needs to be 0. The full output of a residual block is the residual mapping over the input added to the input, or $\mathcal{F}(x) + x$. Adding the input of the block to the output like this is also called a *skip connection* or *residual connection*.

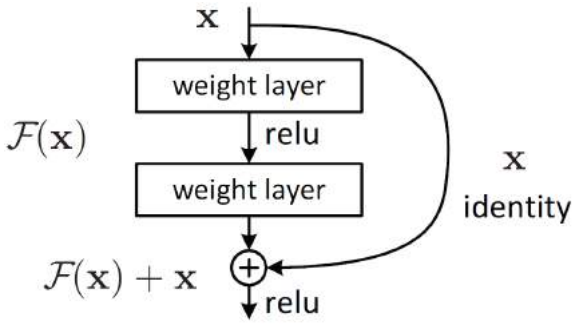


Figure 2.4: A residual block. From [8]

A Residual Network consists of multiple of these residual blocks chained together.

2.1.5 Object Detection using YOLO

"You Only Look Once" (YOLO) is a CNN architecture for real-time object detection created by Redmond et al. in 2016[9]. The model takes in an image and produces a set of bounding boxes for detected objects, with class probabilities for each bounding box. Where previously existing architectures usually consist of separate steps for predicting regions containing objects, classifying those objects, and refining the predictions, YOLO uses a single CNN that simultaneously predicts bounding boxes for objects and calculates probabilities for the classes of the objects detected in these bounding boxes. Because of this single-network approach, YOLO is much faster at detecting objects, even fast enough to work on real-time video footage.

The network used in YOLO consists of multiple blocks of convolutional and max-pooling layers for a total of 24 convolutional layers, followed by two fully connected layers. To detect objects, the input image is first stretched into a square of a fixed size and then broken up into a grid of $S \times S$ patches. For each patch, the model predicts C class probabilities for the contents of the patch and B bounding boxes for the objects that the patch covers. Each bounding box has a center position relative to the patch it belongs to, a width and height relative to the size of the whole image, and a confidence score. The final output of the model is a class probability map that shows the class probabilities for each patch and a set of bounding boxes with, for each bounding box, a confidence score that the box covers an object.

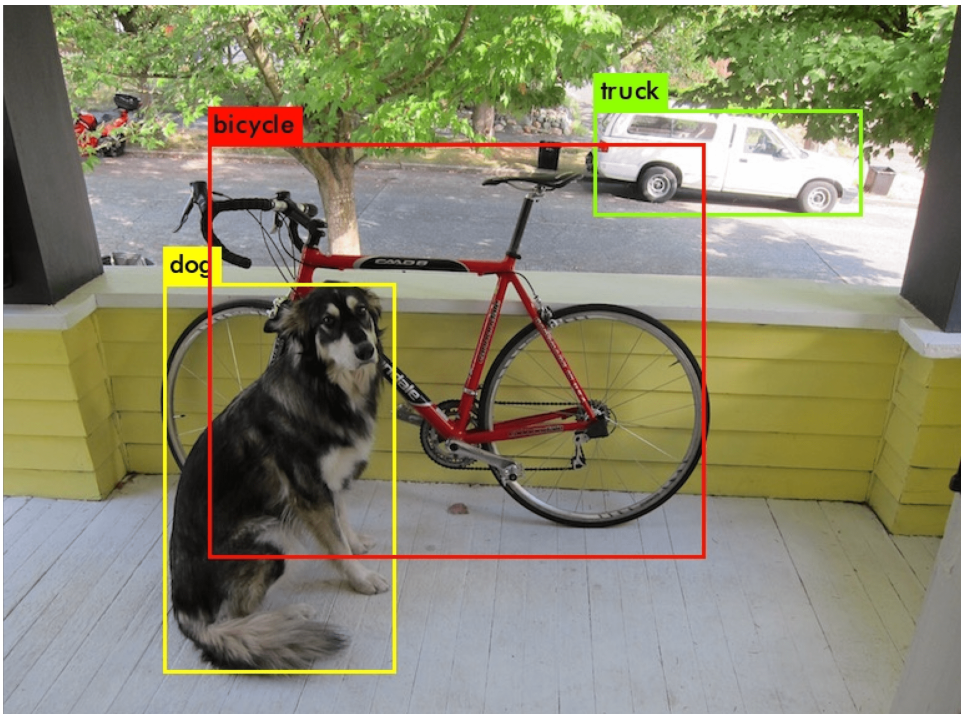


Figure 2.5: YOLO detects objects and predicts their class. From the website of the author[10]
.

During a post-processing stage, overlapping bounding boxes are removed using a process called *Non-Max Surpression*, bounding boxes with a confidence score below a given threshold are removed, and center positions are recalculated to be relative to the original image instead of to the grid cell. Using the class probability map, we determine the most likely object for each bounding box. The final output is a list of bounding boxes, with for each bounding box: the position and size, scaled and adjusted to the original image; the class label; a confidence score that combines the confidence score in the previous step with the class probability.

In the original paper, YOLO was trained for detection on the Pascal VOC 2007 dataset, containing 20 classes for people, animals, vehicles, and a variety of indoor items. The model can be retrained for your own detection tasks, containing different numbers and types of classes.

2.2 Lung Cancer

The term 'cancer' refers to a group of diseases in which genetic mutations cause cells to multiply uncontrollably, taking over the surrounding tissue and spreading to other parts of the body. The specific type of cancer, and

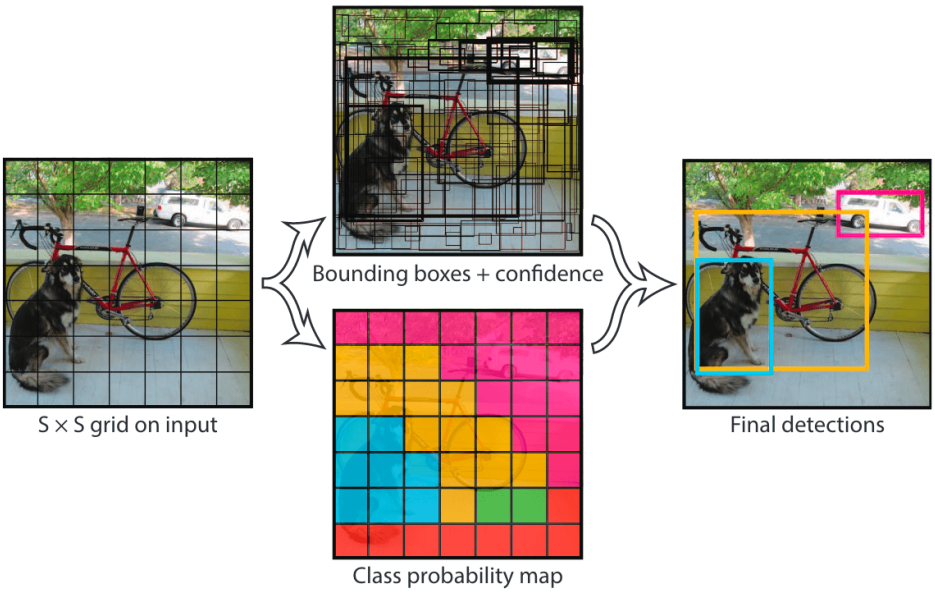


Figure 2.6: YOLO predicts bounding boxes and class probabilities for each patch to determine final detections. From Redmon et al.[9]

with that, the most appropriate treatment, is determined by the location of the cancer and the mutations and type of cells that make up the cancer. Another critical factor in determining the treatment and outcome of the cancer is the stage it is in. The stage ranges from stage 0, where cancerous cells are present but have not spread into the surrounding tissue, to stage 4, where the cancer has widely spread (metastasized) to other body parts.

Lung cancer is cancer that originates in the lungs. Compared to other cancers, it has a very low five-year survival rate[11], largely because the disease causes very few distinct or noticeable symptoms in the early stages[12] and is therefore often detected after it has already progressed to the later stages where treatment outcomes are significantly poorer. It is, therefore, crucial to increase the likelihood of finding the disease early on while it is in a pre-symptomatic stage.

2.2.1 Pulmonary Nodules

One way lung cancer can be detected in a pre-symptomatic stage is in the form of pulmonary (lung) nodules on a CT scan. Pulmonary Nodules are round or semi-round structures commonly found in the lungs on medical images like CT scans. They are usually defined as being between 3 and 30 millimeters [13, 14]. Structures smaller than 3 millimeters are usually called micronodules, while structures bigger than 30 millimeters are usually called masses. Pulmonary nodules can be solid (fully opaque) or sub-solid (at least

partly transparent). Sub-solid nodules can be further classified into part-solid nodules (having an opaque and a transparent component) or non-solid nodules (fully transparent), also called pure ground glass nodules. Additional categories for pulmonary nodules are calcified (containing calcium deposits) and perifissural (located near or attached to a division between lung lobes). Most pulmonary nodules found are benign and do not require

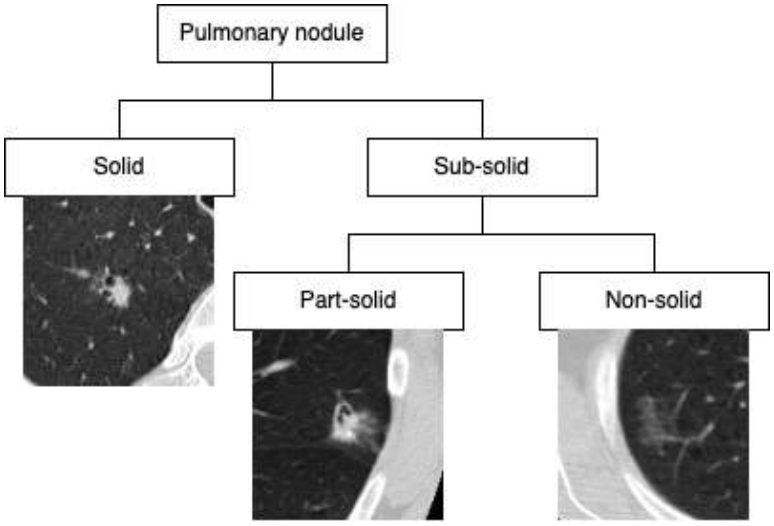


Figure 2.7: Classifying nodules into solid and sub-solid types.

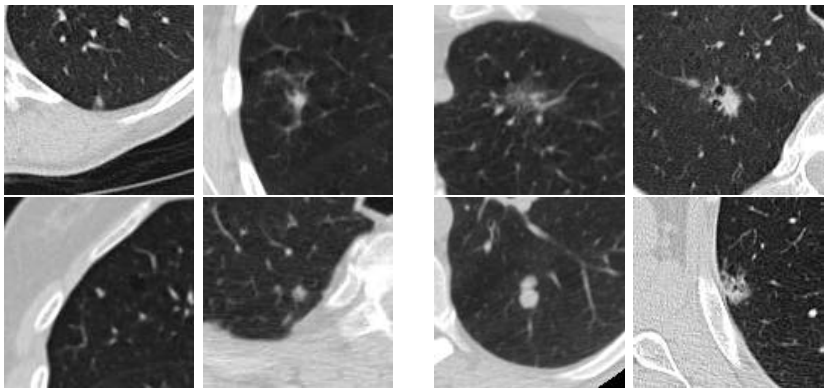
treatment, but occasionally, they can be cases of lung cancer or a different metastasized cancer. To assess the malignancy of a nodule, radiologists look at properties like size, shape, texture, solidness, and spiculation (spikiness around the edges of the nodule). Based on this assessment, further research can be done, e.g., follow-up CT scans to track the nodule's growth, PET CT scans to determine the metabolic activity of the nodule, or more invasive tests like a biopsy of the nodule.

2.2.2 Detecting Pulminary Nodules

To detect these pulmonary nodules, without the existence of specific symptoms, we primarily rely on finding pulmonary nodules as incidental findings or in lung cancer screenings.

Incidental findings

Currently, many of these cases are detected as incidental findings on routine clinical examinations, where pulmonary nodules are found on CT scans made for other clinical purposes[15]. These CT scans vary in resolution, radiation doses, and if contrast is used, based on the protocol in which they were



(a) Benign nodules. (b) Malignant nodules.

Figure 2.8: Examples of benign and malignant nodules.

made. The patients in these scans also tend to be less healthy than patients in screening scans.

Screenings

Another approach to detecting early-stage lung cancer without relying on symptoms is the use of screening programmes. In these screening programmes, members of a high-risk target population undergo low-dose CT scans at regular intervals.

Different screening trials like the National Lung Screening Trial (NLST) in the U.S. and the Nederlands–Leuvens Longkanker Screenings Onderzoek (NELSON) trial in the Netherlands and Belgium have shown effective in detecting lung cancer in earlier stages and improving treatment outcomes for lung cancer patients[16, 17, 18].

These trials call for the implementation of national lung cancer screenings of at-risk populations (usually long-term smokers between 50 and 75 years of age). Several countries, like Croatia, Poland, Italy, and Romania, have already implemented lung cancer screenings[19]. However, it is still not widely adopted across Europe, partly because of the lack of trained radiologists in many European countries[20].

Chapter 3

Related Work

3.1 Automated pulmonary nodule detection

Quite some research has already been done on the detection of pulmonary nodules using deep learning. In 2019, Pehrson et al.[21] published a systemic review evaluating various feature-based and deep learning-based algorithms from 41 different papers. Additionally, Traoré et al. in 2020[22] evaluated various state-of-the-art object detection models for the purpose of pulmonary nodule detection. Both of these papers used the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)[23, 24] dataset for evaluation. LIDC-IDRI is the largest publicly available database of annotated lung CT scans containing both clinical and screening sets. Four radiologists made the annotations in this dataset using a consensus-based approach. The annotations include subjective malignancy ratings made by the radiologists.

3.2 Public challenges

3.2.1 LUNA16

To encourage the engagement of the (scientific) community, the LUng Nodule Analysis challenge in 2016 (LUNA16)[25] challenged participants around the world to come up with algorithms to detect pulmonary nodules. The challenge consisted of a candidate detection track, where a model had to find candidate nodules, and a false positive reduction track, where a model had to eliminate false positives from the candidate nodules found by the first model. Submissions were evaluated using a subset of the LIDC-IDRI dataset, and the most successful submissions were discussed by the organizers in [26].

3.2.2 DSB2017

In 2017, the topic of Kaggle's yearly Data Science Bowl (DSB2017)[27] was automatic lung cancer detection. The goal for participants was to build an algorithm that could, given the CT scans for a given patient, accurately predict whether the patient would be diagnosed with lung cancer within one year. Entries in this competition were evaluated using datasets from three different screening trials. An observer study in 2021 by Jacobs et al.[28] demonstrated that the performance of two of the top three entries in this competition was comparable to that of radiologists.

3.3 Automated nodule detection in non-screening CT scans

The work done within our group by Hendrix et al.[4], which my work continues on, presents a deep learning-based system to detect pulmonary nodules automatically. Where previous work primarily focused on detecting nodules in a screening setting, this system aims to detect nodules in CT scans from clinical routines.

Additionally, where most work thus far has been validated on publicly available datasets like LIDC-IDRI, this work validated the system's performance in a multi-center retrospective study with a reliable reference standard by multiple radiologists.

The system consists of three stages, as shown in Fig. 3.1.

Lung detection (YOLOv5) The first stage uses YOLOv5 (the 5th version of the YOLO framework) to detect the lungs in each layer of the CT scan. These lung detections are combined into a 3D search area for the next stages.

Nodule candidate detection (YOLOv5) The second stage takes five layers of the CT scan (within the search area) at a time and combines them into a single five-channel image. It then uses YOLOv5 to detect possible nodules (candidates) in these images.

False positive reduction (ResNet50) The last step uses a False Positive Reduction (FPR) model based on work from [29] to reduce the number of false positives from step two while maintaining high sensitivity. Where the candidate detection step takes a global look at the whole scan, this step takes a detailed look at individual candidate nodules.

To do this, it takes nine slices at different angles of the 3D patch containing the nodule candidate. It then processes each slice using a ResNet50 model (a residual network with 50 layers) and combines the results using

a convolutional layer to produce a single confidence score for each nodule candidate. These confidence scores are averaged with those produced by YOLO in step 2 to give a final confidence score for each nodule candidate. All nodules with a confidence score below a given threshold are excluded.

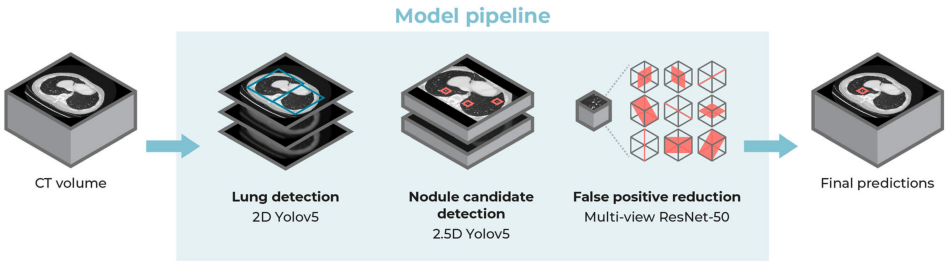


Figure 3.1: The three different stages of the nodule pipeline as presented in Hendriks et al.[4].

The lung detection part of the system was trained on 500 thorax and thorax-abdomen CT scans (500 patients) from Radboudumc. The rest of the system was trained on 602 clinical CT scans (602 patients) from Radboudumc and 888 scans (887 patients) from the LUNA16 dataset.

The system was evaluated on an internal test set containing 100 CT scans (100 patients) from Radboudumc and an external test set containing 100 scans (100 patients) from Jeroen Bosch Ziekenhuis (JBZ). Comparing the performance of the system to the performance of the individual radiologists shows that the system outperformed the radiologists in the detection of benign actionable[1] nodules and metastases, but underperformed compared to radiologists in the detection of primary cancers[2].

My work will build on this research by evaluating the existing system's performance on several screening datasets. Additionally, we will attempt to improve the system's performance at detecting primary cancers. While benign nodules pose a lesser threat and metastasized cancers from other body parts often present complex treatment challenges, catching more primary lung cancers in early stages is a crucial step in improving treatment outcomes for lung cancer. To do this, we will oversample malignant nodules, which are currently underrepresented in the training data. Previous research has shown oversampling to be effective in classification models trained on unbalanced datasets[30]. Previous work also exists that demonstrates the effectiveness of oversampling in pulmonary nodule detection tasks[5, 31]. However, this work focuses on oversampling nodules compared to non-nodules

[1]Benign actionable nodules are nodules that, according to protocol, needed follow-up but turned out to be benign.

[2]Primary cancers are cancers that originated in the lungs and did not metastasize from cancers elsewhere.

and focuses on the false positive reduction part of the system, whereas ours focuses on oversampling an underrepresented type of malignant nodules and retraining the system as a whole.

Chapter 4

Methods

4.1 Retraining

To acquire our baseline measurements and for our oversampling experiments, we will retrain the nodule detection model and the false positive reduction model from the system described in Hendrix et al.[4] using the same training data as was used in the paper, which consists of 602 clinical CT scans from Radboudumc and 888 CT scans from the LUNA16 dataset.

4.1.1 YOLOv5 Candidate Detection

Instead of pre-processing the training data for the YOLOv5 candidate detection ourselves, we use existing, pre-processed training data available within our group that can directly be used to retrain the YOLOv5 model. This is the same data as used in the original paper and consists of CT scan slices of 5 layers each and corresponding labels of the annotated nodule positions and sizes. The training data contains a balanced division of positive slices, which do contain a nodule annotation, and negative slices, which do not contain any nodule annotations. Using this balanced, pre-processed dataset, we train the model for 50 epochs.

4.1.2 ResNet50 False Positive Reduction

To train the false positive reduction part of the network, we first detect all candidate nodules in the training data using the retrained model from the previous step. Multiple pre-processing steps are then taken to extract 3D patches of all these candidate nodules and annotation labels for each patch. These patches and labels are then used to train the model. During training, a balanced dataloader is used to ensure that the model is trained on an approximately equal number of positive patches, which contain a nodule annotation, and negative patches, which do not contain any nodule

annotations. This pre-processed training data is used to train the model for 30 epochs using the balanced dataloader.

4.2 Oversampling

In our oversampling experiments, we will oversample annotated nodules in the training data for both the candidate detection and false positive reduction models. To do this, we make copies of the preprocessed input for each model. This means that for the candidate detection model, we make copies of all slices containing oversampled nodule annotations and corresponding labels. We will also copy a random negative slice for each positive slice we copy. This way, our training data remains balanced, as it is in the original study.

For the training input to the false positive reduction model, we simply make copies of the patches containing oversampled candidate nodules, along with the corresponding annotation label. Since the training process for this model uses a balanced dataloader, we do not need to copy negative patches.

4.2.1 Determining malignancy

To determine which annotated nodules from the training data to oversample, we need to determine their malignancy. We do not have accurate malignancy labels for all annotated nodules in the training data, but we do have some information we can use to decide which ones are worth oversampling.

For the nodule annotations in the Radboudumc data, we have a list of annotated nodules that were found to be cancer. This list was made by Ward Hendrix from our group by cross-referencing nodule annotations with biopsy data and later CT scans from the Integraal Kankercentrum Nederland (IKNL) and should thus serve as a reliable ground truth for nodule malignancy. This list is partially complete, so we will not be able to oversample all malignant Radboudumc nodules. We also do not know which annotated nodules marked as cancer are primary cancers and which are metastases.

The LUNA16 data is based on the LIDC-IDRI dataset and does not contain objective malignancy ratings. Instead, it contains subjective suspiciousness ratings from 1 (Highly Unlikely for Cancer) to 5 (Highly Suspicious for Cancer) from the radiologists who annotated the data. Because these suspiciousness ratings were not validated by later biopsies or other further testing and the interpretation of CT scan findings is subject to a high inter-rater disagreement[32, 33], they do not serve as an objective ground-truth. Still, they do serve as an indication of malignancy. These subjective malignancy ratings are complete for the whole LUNA16 dataset. For our research, we will consider annotated nodules with the highest suspiciousness rating as malignant.

We will compare these methods for determining malignancy by running multiple experiments.

4.3 Experiments

Using all the methods described above, we will run the following experiments:

Baseline Retraining the system on the original data without oversampling.

A Retraining with all nodules marked as cancer in the Radboudumc data oversampled by a factor of 5.

B Retraining with all nodules in the LUNA16 with a malignancy rating of 5 (Highly Suspicious for Cancer) given by any radiologist oversampled by a factor of 5.

C (A+B) Retraining with all nodules marked as cancer in the Radboudumc data **and** all nodules in the LUNA16 data with a malignancy rating of 5 given by any radiologist oversampled by a factor of 5.

The distribution of nodule characteristics in the training data for each experiment can be seen in table 4.1. More detailed information about the exact training inputs for each experiment can be found in appendix A.

4.4 Validation

4.4.1 Datasets

To validate our experiments, we will test the systems on the same clinical datasets as the original paper. We will also run the systems on multiple screening datasets. Statistics about the distribution of nodule characteristics in the datasets used for testing can be seen in table 4.2.

Clinical datasets (Radboudumc and JBZ)

These test sets contain 100 clinical CT scans (100 patients) from Radboudumc and 100 clinical CT scans (100 patients) from Jeroen Bosch Ziekenhuis (JBZ). Both test sets were balanced to contain 25 scans with no findings, 25 scans with benign actionable nodules, 25 scans with metastases, and 25 scans containing primary cancers.

Table 4.1: Characteristics of nodules in the training sets for the different experiments. (A: Oversampling on cancers in Radboudumc. B: Oversampling on suspicious nodules in LUNA16. C: Oversampling on both.)

Dataset	**Radboudumc**	**LUNA16**	**Baseline**	**A**	**B**	**C (A+B)**
Nodules (n)	2489	2281	4770	5322[1]	5398[1]	5950[1]
Scans (n)	602	888	1490	1490	1490	1490
Patients (n)	602	887	1489	1489	1489	1489
Nodules per diameter threshold (n, % of total)						
$\geq$ 4 mm	1622 (65.2)	1896 (83.1)	3518 (73.8)	4030 (75.7)	4141 (76.7)	4660 (78.3)
$\geq$ 5 mm	1076 (43.2)	1326 (58.1)	2402 (50.4)	2883 (54.2)	3025 (56.0)	3508 (59.0)
Diameter (mm)						
Mean	6.1	7.0	6.5	7.2	7.7	8.3
Median	4.7	5.4	5.0	5.3	5.4	5.7
IQR	3.7-6.5	4.3-7.6	4.0-7.0	4.1-8.2	4.1-9.2	4.2-10.3
Volume (mm^3)						
Mean	374.1	510.0	439.1	620.9	785.9	952.2
Median	55.0	82.8	67.6	76.7	82.4	95.8
IQR	26.4-147.2	43.1-235.4	33.2-185.9	35.5-285.8	36.7-407.8	39.0-581.4
Nodules per scan (n)						
Median	3	2	2			
IQR	1-6	1-3	1-4			
Nodules per type (n, % of total)						
Solid	1339 (53.8)	1400 (61.4)	2739 (57.4)	3248 (61.0)	3249 (60.2)	3738 (62.8)
Part-solid	95 (3.8)	354 (15.5)	449 (9.4)	469 (8.8)	542 (10.0)	577 (9.7)
Non-solid	160 (6.4)	309 (13.5)	469 (9.8)	487 (9.2)	488 (9.0)	503 (8.5)
Perifissural	684 (27.5)	0 (0.0)	684 (14.3)	689 (12.9)	684 (12.7)	695 (11.7)
Calcified	211 (8.5)	218 (9.6)	429 (9.0)	429 (8.1)	435 (8.1)	437 (7.3)
Benign versus malignant nodules[2](n, % of total)						
Benign	2351 (94.5)	2124 (93.1)	4475 (93.8)	4475 (84.1)	4475 (82.9)	4475 (75.2)
Malignant	138 (5.5)	157 (6.9)	295 (6.2)	847 (15.9)	923 (17.1)	1475 (24.8)

[1] No new nodules were added, only duplicates of existing nodules.
[2] Malignancy is estimated, as explained in 4.2.1.

Screening datasets (NLST, DLCST, and MILD)

In addition to the clinical datasets, we also use datasets from the National Lung Screening Trial (NLST)[34], the Danish Lung Cancer Screening Trial (DLCST)[35] and the Multicentric Italian Lung Detection (MILD) trial[36]. Since these datasets contain many more nodule annotations than the clinical datasets, we can more accurately assess if any possible performance increases are statistically significant. We have made cancer-enriched subsets from these datasets to test our systems.

These subsets were composed by first selecting the earliest scan for each

participant. Of these scans, we included all scans with annotated malignant nodules and added scans without cancer for a total of at most 1000 per subset.
The annotations of nodules for the NLST and DLCST datasets came from existing data within our group, used in research by Vendkadesh et al.[37]. The annotations for the MILD dataset were collected by the MILD trial researchers in collaboration with Radboudumc researchers, where nodules found in later scans were retroactively added to earlier scans. We have excluded annotations of nodules with a diameter under 3 mm or over 30 mm from all datasets.

4.4.2 Evaluation

To evaluate the performance of our systems on the dataset, we use the same evaluation method as the LUNA16 challenge[1]. In this method, we take the system's predictions, a list of nodule annotations by radiologists it should find, and a list of excluded nodule annotations that should not count as true or false positives. These excluded annotations can be annotations that were made by a minority of radiologists or nodules with a diameter of less than 3 mm or over 30 mm. Using this, we calculate the sensitivity of the system for different numbers of false positives per scan to create a Free-response Receiver Operator Characteristics (FROC) curve.

For these FROC curves, we also calculate 95% confidence intervals using bootstrapping. In this process, we take 1000 random samples (bootstraps) of the data to estimate the distribution of our data and calculate a range for which we can be 95% confident that it contains our true result. Doing this gives us an indication of how much variability we can expect if we were to repeat our experiment under similar conditions.

We also compare our systems's performance after each experiment with the performance of our baseline system. To do this, we calibrate our system in each experiment to 1 false positive per scan for each dataset and compare the sensitivities. We evaluate the statistical significance of any changes in sensitivity between experiments using a two-sided paired permutation test. In this test, we randomly swap individual results of the two systems 1000 times, recalculating the difference in sensitivity each time. We then calculate a p-value by measuring how many of these random swaps resulted in a difference in sensitivity greater than the one measured without any swaps. A p-value of less than 0.05 is considered statistically significant.

[1] `https://luna16.grand-challenge.org/Evaluation/`

Table 4.2: Characteristics of nodules in the test sets.

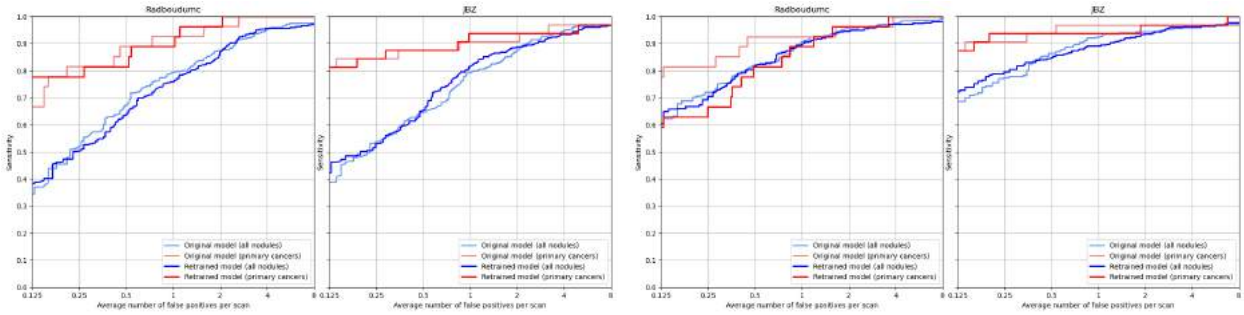
Dataset	**Radboudumc**	**JBZ**	**MILD**	**DLCST**	**NLST**
Nodules (n)	319	303	2788	1283	1182
Scans (n)	100	100	1000	806	1000
Patients (n)	100	100	1000	806	1000
Nodules per diameter threshold (n, % of total)					
$\geq$ 4 mm	250 (78.4)	262 (86.4)	1441 (51.7)	1094 (85.3)	1173 (99.2)
$\geq$ 5 mm	188 (58.9)	215 (71.0)	859 (30.8)	703 (54.8)	1009 (85.4)
Diameter (mm)					
Mean	7.2	9.0	5.0	6.2	11.9
Median	5.5	6.6	4.1	5.2	9.2
IQR	4.1-8.6	4.7-11.9	3.3-5.5	4.4-6.5	5.9-14.5
Volume (mm^3)					
Mean	573.6	1049.9	283.9	520.6	N/A
Median	87.7	160.6	36.6	73.6	N/A
IQR	36-332	55-889	18.3-87.3	44.3-145.1	N/A
Nodules per scan (n)					
Median	1	2	1	2	1
IQR	1-4	1-4	1-2	1-3	1-1
Nodules per type (n, % of total)					
Solid	269 (84.3)	247 (81.5)	1069 (83.3)	1959 (70.3)	797 (67.4)
Part-solid	12 (3.8)	19 (6.3)	163 (12.7)	386 (13.8)	141 (11.9)
Non-solid	8 (2.5)	10 (3.3)	51 (4.0)	199 (7.1)	222 (18.8)
Perifissural	24 (7.5)	17 (5.6)		187 (6.7)	22 (1.9)
Calcified	6 (1.9)	10 (3.3)		57 (2.0)	
Benign versus malignant nodules (n, % of total)					
Benign	127 (39.8)	158 (52.1)	2682 (96.2)	1220 (95.1)	468 (39.6)
Actionable	63 (19.7)	87 (28.7)			
Malignant	192 (60.2)	145 (47.9)	106 (3.8)	63 (4.9)	714 (60.4)
Primary cancer	27 (8.5)	32 (10.6)			
Metastasis	165 (51.7)	113 (37.3)			

Chapter 5

Results

5.1 Retraining without oversampling

We first retrained the system on the original dataset and hyperparameters used in Hendrix et al. Fig. 5.1a shows the performance of our retrained candidate detection model (without false positive reduction) compared with the original model. Our retrained model seems to perform similarly to the original model, as expected. Fig. 5.1b shows the performance of our complete retrained system (including false positive reduction). Our retrained system performs similarly to the original system overall but, interestingly, performs worse in detecting primary cancers in the Radboudumc dataset. It even performs worse on these nodules than the retrained model without false positive reduction. To make sure our training process is correct, we retrained the system again in the exact same way. The results of both retraining compared to the baseline can be found in Appendix B.1 and indicate that our training process is correct but subject to variation on the relatively small clinical test sets.



(a) Without false positive reduction.

(b) With false positive reduction

Figure 5.1: FROC curves of the retrained system (without oversampling) and the original system on the clinical datasets.

We also evaluated our retrained system on subsets of three screening datasets from the MILD trial, NLST, and DLCST. Fig. 5.2 shows the free response operator characteristic (FROC) curves of the system on each clinical and screening dataset. It also shows the 95% confidence interval for each nodule category.

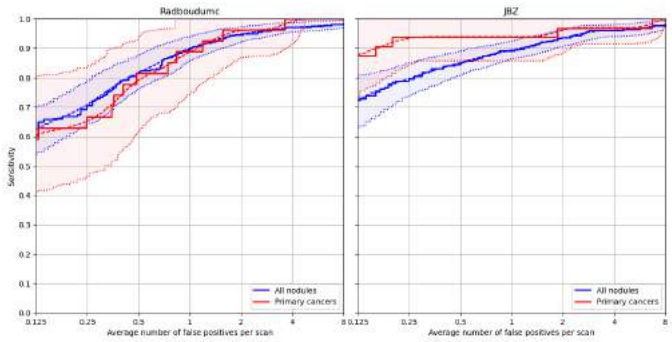
We found that the system performs much worse on screening than clinical datasets, reaching substantially lower sensitivities at the same number of false positives per scan. Part of this is because of the way we count true and false positives. For the clinical datasets, nodules are only counted if they were found by three or more out of five radiologists and ignored otherwise. This means that the nodules that are hardest to detect (and were thus detected by a minority of radiologists) are excluded from the clinical datasets. This is not the case for the screening datasets, which were annotated by just one or two radiologists. Appendix B.2 shows the FROC curves on the clinical datasets if we take the annotations of a single rater as the ground truth.

We also see that the system performs worse on malignant nodules in MILD. This can be partly attributed to the fact that we included the earliest scan of each patient, and annotations of nodules found in later scans were added to these earliest scans when these malignant nodules were very hard to detect. This effect is demonstrated in Appendix B.3.

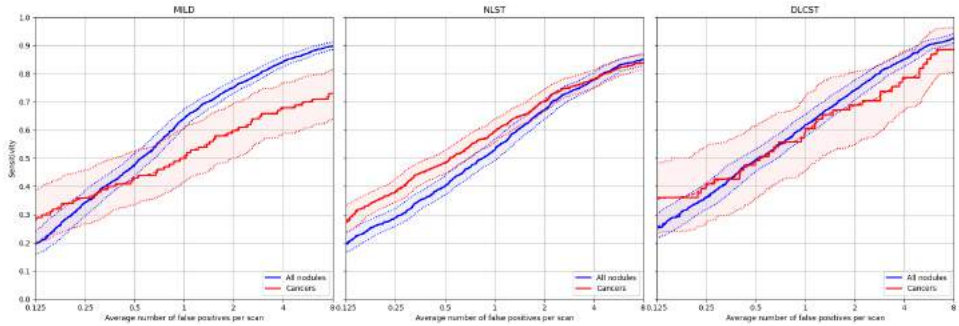
5.2 Retraining with oversampling

The FROC curves of the systems after retraining without oversampling and after our different oversampling experiments, as described in section 4.2, can be seen in Fig. 5.3. These graphs all seem to indicate that oversampling on seemingly malignant nodules improves the performance of the system on malignant nodules without sacrificing a lot of performance on all nodules.

We also measured the sensitivity of each system on the different datasets at an operating point of 1 false positive per scan. The results of this can be seen in Table 5.1. We found no statistically significant difference in performance on the clinical datasets from Radboudumc and JBZ, but we did find that our system reached a significantly higher sensitivity on malignant nodules in our NLST subset after oversampling on cancers in Radboudumc (65.0% vs. 59.6%, $p = 0.04$) and after oversampling on both cancers in Radboudumc and suspicious nodules in LUNA16 (68.4% vs. 59.6%, $p < 0.01$). Oversampling on both cancers in Radboudumc and suspicious nodules in LUNA16 also gave a significantly higher sensitivity on all nodules in our NLST subset (59.2% vs. 53.2%, $p = 0.01$). This is likely because our NLST subset consists primarily of malignant nodules (60.4%). There is no significant difference between oversampling on nodules labeled malignant in the Radboudumc dataset and oversampling on nodules marked suspicious



(a) Performance on clinical datasets.



(b) Performance on screening datasets.

Figure 5.2: FROC curves of the retrained system (without oversampling). The shaded bands represent the 95% confidence intervals per nodule category.

in LUNA16. The most effective oversampling method seems to combine the two and oversample as many likely malignant nodules as possible.

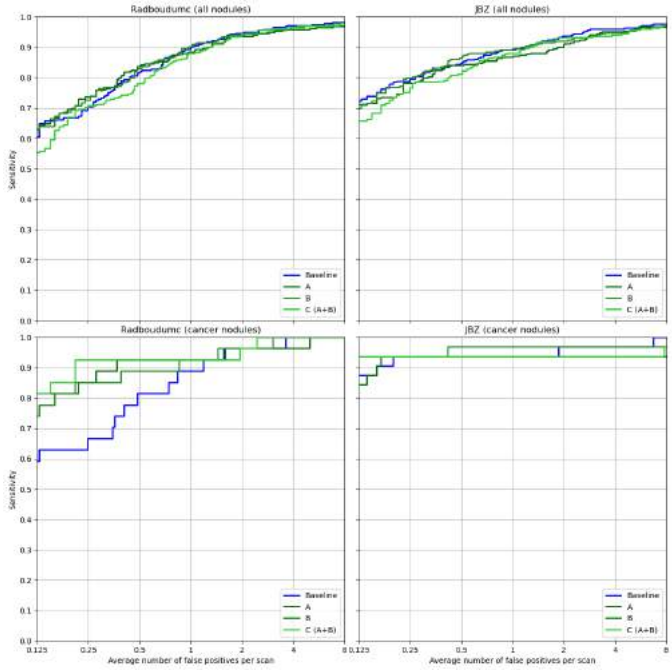
Interestingly, we also found that oversampling on cancers in Radboudumc made the system significantly less sensitive on all nodules in our MILD subset (61.9% vs. 64.9%, $p = 0.03$). This was not the case when oversampling on suspicious nodules in LUNA16 or when oversampling on both cancers in Radboudumc and suspicious nodules in LUNA16.

When looking at the performance of just the YOLOv5 candidate detection model, we found that the model performs significantly worse at detecting all nodules in experiments A, B, and C for both JBZ and our MILD subset ($p < 0.01$), in experiments B and C for Radboudumc ($p = 0.01$) and experiment B for our DLCST subset ($p < 0.01$). The YOLOv5 candidate detection model did not perform significantly better on cancers in any dataset after any experiment. These results indicate that while retraining with oversampling has a positive effect on the ResNet50 false positive reduction model, it has a negative effect on the YOLOv5 candidate detection model.

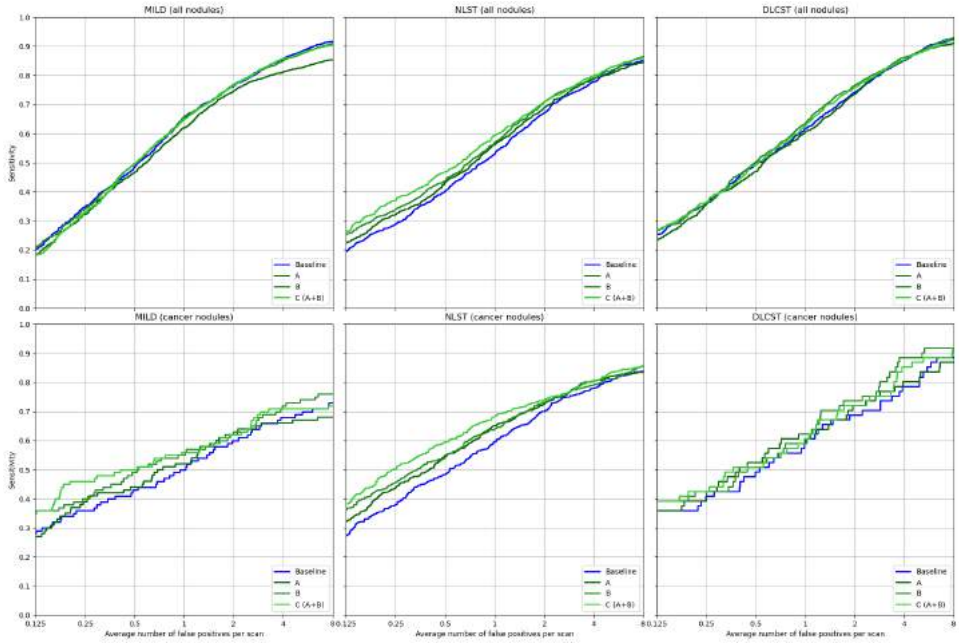
5.3 Characteristics of missed nodules

We evaluated the characteristics of missed nodules in our NLST subset. We focused on this dataset because it contains the most malignant nodules and showed the most significant results in our experiments. Table. 5.2 shows the number of missed nodules in our NLST subset for each experiment and information about the diameter distribution of missed nodules. We calibrated the system for each experiment to an average of 1 false positive per scan. We see that the diameter of our missed nodules decreases for our oversampling experiments. For our baseline experiment, the mean and median diameter of missed nodules are 10.2 and 8.2, respectively, with an IQR of (5.3-13.4). For experiment C, the mean and median diameter of missed nodules are 9.5 and 7.2, with an IQR of (5.1-12.1). This is in line with what we would expect. In Table. 4.1, we see that for our oversampling experiments, the mean and median diameter of our nodules increases compared to the baseline experiment. Because we train on more large nodules, our system also learns to detect more large nodules.

We have also inspected visual examples of nodules missed by the baseline retrained system but found after retraining with oversampling on both Radboudumc and LUNA16. Six examples can be found in Fig. 5.4. Fig. 5.5 shows six nodules that were still missed after retraining with oversampling on Radboudumc and LUNA16. All visual examples are nodules from our NLST subset.



(a) Performance on clinical datasets.



(b) Performance on screening datasets.

Figure 5.3: Performance of the system after our oversampling experiments, compared to baseline. (A: Oversampling on cancers in Radboudumc. B: Oversampling on suspicious nodules in LUNA16. C: Oversampling on both.)

Table 5.1: Comparison of our systems, retrained with oversampling, with our baseline system. All systems were calibrated on each dataset to 1 FP/scan.
(A: Oversampling on cancers in Radboudumc. B: Oversampling on suspicious nodules in LUNA16. C: Oversampling on both.)

(a) Comparison of models without FPR.

	All nodules		(Primary) cancers	
	Sensitivity (%)	p	Sensitivity (%)	p
Radboudumc				
Baseline	76.8 (70.7, 83.4)		91.5 (76.7, 100.0)	
A	76.2 (67.8, 85.5)	0.92	90.2 (75.0, 100.0)	>0.99
B	69.7 (58.4, 79.6)	0.01	99.6 (95.2, 100.0)	0.23
C (A+B)	66.6 (56.9, 76.4)	0.01	96.1 (87.1, 100.0)	0.60
JBZ				
Baseline	81.8 (75.5, 87.3)		92.7 (80.0, 100.0)	
A	69.2 (59.0, 79.6)	<0.01	90.5 (80.0, 100.0)	>0.99
B	62.0 (52.7, 71.2)	<0.01	93.4 (81.1, 100.0)	>0.99
C (A+B)	56.6 (47.8, 65.6)	<0.01	91.2 (78.8, 100.0)	>0.99
MILD				
Baseline	51.9 (48.3, 55.7)		50.4 (40.8, 60.2)	
A	32.2 (29.2, 35.1)	<0.01	41.5 (31.7, 51.3)	0.28
B	41.8 (39.0, 44.8)	<0.01	56.8 (46.7, 67.3)	0.48
C (A+B)	39.2 (36.2, 41.8)	<0.01	56.2 (46.8, 65.6)	0.58
NLST				
Baseline	56.4 (52.8, 59.7)		64.8 (60.9, 68.4)	
A	53.0 (49.8, 56.4)	0.14	64.6 (60.7, 68.4)	0.94
B	54.3 (51.0, 57.7)	0.34	66.0 (62.0, 69.9)	0.76
C (A+B)	54.3 (51.2, 57.6)	0.36	67.0 (63.5, 70.5)	0.44
DLCST				
Baseline	55.7 (51.4, 60.9)		56.9 (43.6, 70.5)	
A	55.3 (50.9, 59.7)	0.86	62.9 (50.0, 74.5)	0.59
B	48.5 (43.6, 52.9)	<0.01	55.5 (41.7, 69.0)	>0.99
C (A+B)	53.0 (48.1, 58.2)	0.22	60.1 (47.8, 72.2)	0.70

(b) Comparison of systems with FPR.

	All nodules		(Primary) cancers	
	Sensitivity (%)	p	Sensitivity (%)	p
Radboudumc				
Baseline	90.1 (85.8, 93.9)		88.8 (74.3, 100.0)	
A	89.0 (84.1, 93.6)	0.73	93.0 (82.6, 100.0)	>0.99
B	90.1 (84.7, 94.4)	0.89	92.1 (79.3, 100.0)	>0.99
C (A+B)	87.8 (82.7, 92.8)	0.62	92.6 (82.1, 100.0)	>0.99
JBZ				
Baseline	89.4 (86.3, 92.5)		94.0 (85.7, 100.0)	
A	87.0 (84.1, 90.0)	0.47	94.0 (86.5, 100.0)	>0.99
B	89.6 (86.8, 92.2)	>0.99	97.2 (91.9, 100.0)	>0.99
C (A+B)	88.4 (84.8, 91.9)	0.78	94.0 (86.1, 100.0)	>0.99
MILD				
Baseline	64.9 (61.6, 68.1)		50.5 (41.0, 59.8)	
A	61.9 (58.7, 64.9)	0.03	52.1 (42.1, 61.5)	0.87
B	65.5 (62.1, 68.4)	0.59	55.6 (46.2, 64.4)	0.58
C (A+B)	64.7 (61.6, 67.8)	0.92	55.8 (46.5, 65.6)	0.47
NLST				
Baseline	53.2 (49.5, 56.8)		59.6 (55.5, 63.8)	
A	56.2 (52.5, 60.0)	0.14	65.0 (60.9, 68.9)	0.04
B	56.7 (53.2, 60.4)	0.07	64.2 (60.2, 67.8)	0.08
C (A+B)	59.2 (55.7, 62.4)	0.01	68.4 (64.9, 71.9)	<0.01
DLCST				
Baseline	61.5 (57.6, 65.5)		58.9 (45.3, 72.1)	
A	60.3 (55.6, 64.3)	0.51	62.3 (51.1, 73.8)	0.86
B	63.7 (58.3, 68.7)	0.40	60.7 (48.8, 73.7)	>0.99
C (A+B)	62.7 (58.1, 67.4)	0.69	60.4 (48.4, 73.5)	>0.99

Table 5.2: Characteristics of missed nodules in our NLST subset, at an operating point of 1 false positive per scan, for each experiment. (A: Oversampling on cancers in Radboudumc. B: Oversampling on suspicious nodules in LUNA16. C: Oversampling on both.)

(a) Characteristics for all missed nodules.

Experiment	**Baseline**	**A**	**B**	**C (A+B)**
Missed nodules (n, % of total)	555 (47.0)	519 (43.9)	507 (42.9)	480 (40.6)
Missed nodules per diameter threshold (n, % of missed)				
$\geq$ 4 mm	549 (98.9)	513 (98.8)	501 (98.8)	475 (99.0)
$\geq$ 5 mm	445 (80.2)	408 (78.6)	398 (78.5)	362 (75.4)
Diameter of missed nodules (mm)				
Mean	10.2	9.8	9.9	9.5
Median	8.2	7.7	7.7	7.2
IQR	5.3-13.4	5.2-12.8	5.2-12.8	5.1-12.1
Types of missed nodules (n, % of missed)				
Solid	312 (56.2)	285 (54.9)	282 (55.6)	264 (55.0)
Part-solid	70 (12.6)	59 (11.4)	59 (11.6)	51 (10.6)
Non-solid	164 (29.5)	168 (32.4)	158 (31.2)	155 (32.3)
Perifissural	9 (1.6)	7 (1.3)	8 (1.6)	10 (2.1)

(b) Characteristics for cancers.

Experiment	**Baseline**	**A**	**B**	**C (A+B)**
Missed nodules (n, % of total)	289 (40.5)	253 (35.4)	256 (35.9)	228 (31.9)
Missed nodules per diameter threshold (n, % of missed)				
$\geq$ 4 mm	288 (99.7)	252 (99.6)	255 (99.6)	227 (99.6)
$\geq$ 5 mm	253 (87.5)	216 (85.4)	221 (86.3)	192 (84.2)
Diameter of missed nodules (mm)				
Mean	12.1	11.7	11.8	11.4
Median	10.7	10.3	10.3	9.7
IQR	6.6-16.7	6.4-15.7	6.4-15.8	6.1-15.2
Types of missed nodules (n, % of missed)				
Solid	197 (68.2)	169 (66.8)	171 (66.8)	155 (68.0)
Part-solid	44 (15.2)	37 (14.6)	40 (15.6)	31 (13.6)
Non-solid	48 (16.6)	47 (18.6)	45 (17.6)	42 (18.4)
Perifissural	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)

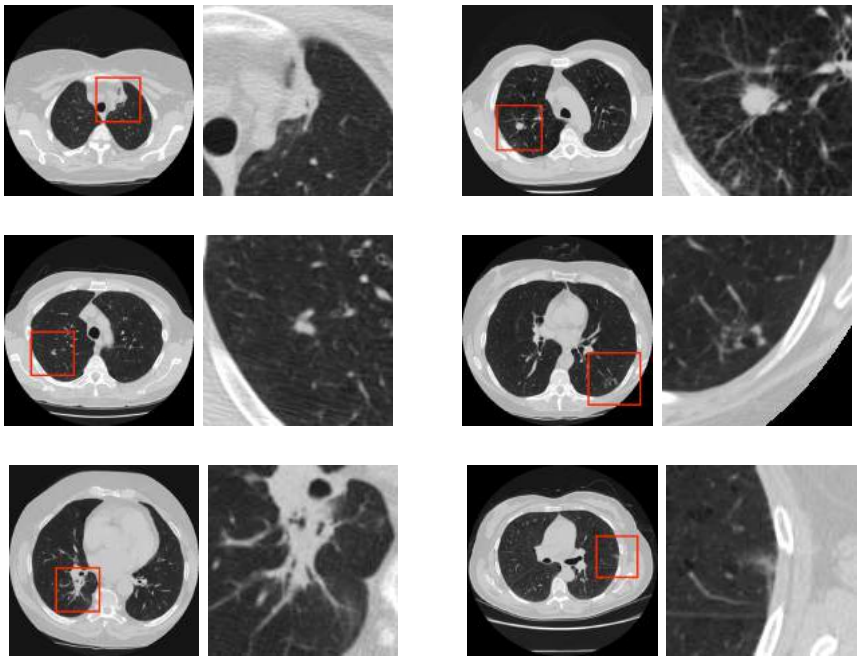


Figure 5.4: Malignant nodules missed by the baseline retrained system but found by the system after retraining with oversampling on Radboudumc and LUNA16.

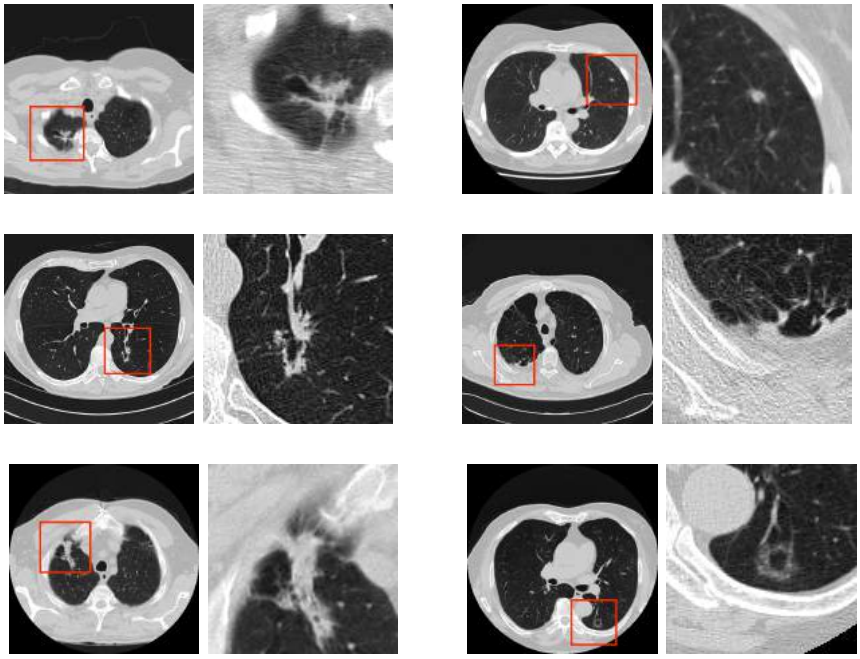


Figure 5.5: Malignant nodules missed by the system after retraining with oversampling on Radboudumc and LUNA16.

Chapter 6

Discussion

In this thesis, our primary focus was to answer the question of whether oversampling can be used to improve the performance of a deep learning-based pulmonary nodule detection system on primary cancers in clinical datasets. To do this, we first retrained the system and evaluated it on clinical test sets from Radboudumc and JBZ, as was done in the original paper by Hendrix et al. Additionally, we evaluated the system on three additional datasets from the lung cancer screening trials: MILD, NLST, and DLCST. These additional test sets contain many more scans (1000, 1000, and 806 scans, respectively, compared with 100 scans per clinical test set), thus making it easier to find statistically significant differences in performance.

In this initial evaluation, we found that our system performed better on the clinical test sets we tested than on the screening test sets. This is in part because of the evaluation method used, where the clinical datasets are evaluated on using a ground truth based on consensus between five radiologists, while the screening datasets are evaluated on using annotations by one or two radiologists. When adjusted for this, the difference is reduced but is still present.

After this initial retraining and evaluation, we retrained the system on training sets where we oversampled on malignant nodules. We ran multiple experiments because we did not have a complete list of malignant nodules for the training data, only a partially complete list for the Radoudumc training data, and a list of subjective ratings for the LUNA16 training data. We retrained our system with oversampling only on all nodules in Radboudumc marked as cancer (experiment A), oversampling only on all nodules in LUNA16 rated highly suspicious of cancer (experiment B), and with both groups of nodules oversampled (experiment C). All oversampled nodules were included five times in the training data. After evaluating all experiments, we found that oversampling can positively affect the system's performance on malignant nodules, although a statistically significant improvement was only seen in one dataset The results of experiment B seem

slightly better than those of experiment A on our subsets of MILD and NLST, although this is not a statistically significant difference. The biggest performance increase was seen in experiment C on our NLST subset, reaching a significantly higher ($p < 0.01$) sensitivity at 1 false positive per scan of 68.4 (64.9, 71.9), compared to the baseline sensitivity of 59.6 (55.5, 63.8). These results indicate that oversampling was effective, with the number of nodules oversampled having a bigger effect than the accuracy of malignancy labels.

We also found that while oversampling as a whole was effective, it significantly decreased the performance of the YOLOv5 candidate detection model in almost all experiments on all datasets when looking for all nodules, without having a significant effect when looking for only cancers. This indicates that oversampling is best used only on the ResNet50-based false positive reduction model.

Even though we have shown that oversampling can be a valuable tool in improving the performance of our nodule detection system, we have only seen statistically significant improvements in the performance on cancers in screening datasets. The question remains whether this improvement also holds for the detection of primary cancers in clinical data. For one, we have seen that our system performs quite differently on screening data than on clinical data, meaning that conclusions we can draw about the performance of our system on screening data can not be directly translated to clinical data. Secondly, the screening datasets only contain general malignancy information for nodules and do not make a distinction between primary cancers and metastases. We can likely assume that most of the malignant nodules in screening data are primary cancers since participants in these screening trials are otherwise healthy and thus unlikely to have pre-existing cancer that has progressed enough to have metastasized, but we do not have concrete statistics to support this assumption.

6.1 Future work

Even though our research serves as a good proof-of-concept for oversampling on a specific class of nodules in the training of a nodule detection system, a lot of research can still be done to expand on this.

In the area of validation, a lot of work can still be done in the collection and accurate annotation of clinical datasets to be used for testing. The work by Hendrix et al. provided us with two high-quality datasets, but with the small size of these datasets (100 CT scans each), it is difficult to find if small changes in improvement are statistically significant. Alternatively, more research could be done on the difference in performance of detection algorithms on clinical versus screening datasets to find whether conclusions

about one can be translated to the other.

On oversampling, many further experiments can be run that we did not have time for during this thesis. One obvious example would be to only retrain the ResNet50 false positive reduction model with oversampling, as our research seems to indicate that this would lead to better results. Additionally, one could experiment with the oversampling parameters we used. It would be interesting to see what the effect is at different oversampling rates and at what rates the improvements would start to diminish. One could also be more specific in which nodules to oversample, only oversampling nodules that fall within diameter/volume ranges or other characteristics that are currently hard for the system to recognize.

Lastly, it would be interesting to find ways to add augmented or new data for the system to train on. One could apply data augmentation techniques to the data added in oversampling, like rotating or translating the slices, or more modern techniques like Generative Adversarial Networks[38], Variational Auto-Encoders[39], or deepSMOTE[40]. Alternatively, one could add malignant samples from other datasets to the training data. This way, the system is presented with new data during training, whereas our oversampling method merely duplicates existing data.

6.2 Conclusions

We can draw two main conclusions from our research.

First, we can conclude that our system performs better on the clinical test sets we tested (Radboudumc and JBZ) than on our subsets the screening test sets (MILD, NLST, DLCST), both before and after oversampling. This holds even after adjusting for the differences in establishing ground truths from annotations.

Secondly, we have found indications that oversampling on a specific type of pulmonary nodule improves the performance of our system on that type of nodule. This improvement is visible for the system as a whole but not for the YOLOv5 candidate detection model, which performed significantly worse on all nodules. We found that this holds for malignant nodules in our subset of NLST, but we could not determine if this holds for clinical datasets as well. We also found that the number of nodules we oversample on seems to have a bigger effect than the metric we use to determine which nodules to oversample, with a combination of objective cancer ratings and subjective suspiciousness ratings being the most effective. In most experiments, oversampling did not have a significant adverse effect on the performance of all nodules. We only saw a significant adverse effect when we oversampled on nodules marked as cancer from the Radboudumc data when evaluated on all nodules in our subset of MILD.

Bibliography

[1] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians*, 71(3):209–249, May 2021.

[2] Marta Soares, Luís Antunes, Patrícia Redondo, Marina Borges, Ruben Hermans, Dony Patel, Fiona Grimson, Robin Munro, Carlos Chaib, Laure Lacoin, Melinda Daumont, John R Penrod, John C O'Donnell, Maria José Bento, and Francisco Rocha Gonçalves. Treatment and outcomes for early non-small-cell lung cancer: A retrospective analysis of a Portuguese hospital database. *Lung Cancer Management*, 10(2):LMT46.

[3] S. S. Birring and M. D. Peake. Symptoms and the early diagnosis of lung cancer. *Thorax*, 60(4):268–269, April 2005.

[4] Ward Hendrix, Nils Hendrix, Ernst T. Scholten, Mariëlle Mourits, Joline Trap-de Jong, Steven Schalekamp, Mike Korst, Maarten van Leuken, Bram van Ginneken, Mathias Prokop, Matthieu Rutten, and Colin Jacobs. Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans. *Communications Medicine*, 3(1):1–12, October 2023.

[5] Shrikant A. Mehre, Sudipta Mukhopadhyay, Anirvan Dutta, Nagam Chaithan Harsha, Ashis Kumar Dhara, and Niranjan Khandelwal. An automated lung nodule detection system for CT images using synthetic minority oversampling. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, pages 120–127. SPIE, March 2016.

[6] Simon S. Haykin. *Neural Networks: A Comprehensive Foundation.* Prentice Hall, Upper Saddle River, NJ, 2. ed., [nachdr.] edition, 1999. Literaturverz. S. 796 - 836.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. Comment: Tech report.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[10] YOLO: Real-Time Object Detection. https://pjreddie.com/darknet/yolo/.

[11] Cancer survival in England - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddis

[12] Leroy Hyde and Charles I. Hyde. Clinical Manifestations of Lung Cancer. *Chest*, 65(3):299–306, March 1974.

[13] Pulmonary Nodules — British Thoracic Society — Better lung health for all. https://www.brit-thoracic.org.uk/quality-improvement/guidelines/pulmonary-nodules/.

[14] Alexander A. Bankier, Heber MacMahon, Jin Mo Goo, Geoffrey D. Rubin, Cornelia M. Schaefer-Prokop, and David P. Naidich. Recommendations for Measuring Pulmonary Nodules at CT: A Statement from the Fleischner Society. *Radiology*, 285(2):584–600, November 2017.

[15] Michael K. Gould, Tania Tang, In-Lu Amy Liu, Janet Lee, Chengyi Zheng, Kim N. Danforth, Anne E. Kosco, Jamie L. Di Fiore, and David E. Suh. Recent Trends in the Identification of Incidental Pulmonary Nodules. *American Journal of Respiratory and Critical Care Medicine*, 192(10):1208–1214, November 2015.

[16] National Lung Screening Trial Research Team, Denise R. Aberle, Amanda M. Adams, Christine D. Berg, William C. Black, Jonathan D. Clapp, Richard M. Fagerstrom, Ilana F. Gareen, Constantine Gatsonis, Pamela M. Marcus, and JoRean D. Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *The New England Journal of Medicine*, 365(5):395–409, August 2011.

[17] Harry J. de Koning, Carlijn M. van der Aalst, Pim A. de Jong, Ernst T. Scholten, Kristiaan Nackaerts, Marjolein A. Heuvelmans, Jan-Willem J. Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, Susan van 't Westeinde, Mathias Prokop, Willem P. Mali, Firdaus A.A. Mohamed Hoesein, Peter M.A. van Ooijen, Joachim G.J.V. Aerts, Michael A. den Bakker, Erik Thunnissen, Johny Verschakelen, Rozemarijn Vliegenthart, Joan E. Walter, Kevin ten Haaf, Harry J.M.

Groen, and Matthijs Oudkerk. Reduced Lung-Cancer Mortality with Volume CT Screening in a Randomized Trial. *New England Journal of Medicine*, 382(6):503–513, February 2020.

[18] Scott J Adams, Emily Stone, David R Baldwin, Rozemarijn Vliegenthart, Pyng Lee, and Florian J Fintelmann. Lung cancer screening. *The Lancet*, 401(10374):390–408, February 2023.

[19] Suzanne Wait, Arturo Alvarez-Rosete, Tasnime Osama, Dani Bancroft, Robin Cornelissen, Ante Marušić, Pilar Garrido, Mariusz Adamek, Jan van Meerbeeck, Annemiek Snoeckx, Olivier Leleu, Ebba Hallersjö Hult, Sébastien Couraud, and David R. Baldwin. Implementing Lung Cancer Screening in Europe: Taking a Systems Approach. *JTO Clinical and Research Reports*, 3(5), May 2022.

[20] Jan P. Van Meerbeeck, Emma O'Dowd, Brian Ward, Paul Van Schil, and Annemiek Snoeckx. Lung Cancer Screening: New Perspective and Challenges in Europe. *Cancers*, 14(9):2343, May 2022.

[21] Lea Marie Pehrson, Michael Bachmann Nielsen, and Carsten Ammitzbøl Lauridsen. Automatic Pulmonary Nodule Detection Applying Deep Learning or Machine Learning Algorithms to the LIDC-IDRI Database: A Systematic Review. *Diagnostics*, 9(1):29, March 2019.

[22] Abdarahmane Traoré, Abdoulaye O. Ly, and Moulay A. Akhloufi. Evaluating Deep Learning Algorithms in Pulmonary Nodule Detection. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1335–1338, July 2020.

[23] Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J. R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.-Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. The Lung Image Database Consortium (LIDC) and Image Database Resource Ini-

tiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Medical Physics*, 38(2):915–931, February 2011.

[24] Samuel G. Armato III, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J.R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. Data From LIDC-IDRI, 2015.

[25] LUNA16 - Grand Challenge. https://luna16.grand-challenge.org/.

[26] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S. N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M. J. de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T. M. C. Manders, Alexander Sóñora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M Schaefer-Prokop, Ernst T. Scholten, Luuk Scholten, Miranda M. Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C. A. Zuidhof, Bram van Ginneken, and Colin Jacobs. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Medical Image Analysis*, 42:1–13, December 2017.

[27] AJ_Buckeye, Jacob Kriss, Josette_BoozAllen, Josh Sullivan, Meghan O'Connell, Nilofer, and Will Cukierski. Data Science Bowl 2017. https://kaggle.com/competitions/data-science-bowl-2017.

[28] Colin Jacobs, Arnaud A. A. Setio, Ernst T. Scholten, Paul K. Gerke, Haimasree Bhattacharya, Firdaus A. M Hoesein, Monique Brink, Erik Ranschaert, Pim A. de Jong, Mario Silva, Bram Geurts, Kaman Chung, Steven Schalekamp, Joke Meersschaert, Anand Devaraj, Paul F. Pinsky, Stephen C. Lam, Bram van Ginneken, and Keyvan Farahani. Deep Learning for Lung Cancer Detection on Screening CT Scans: Results of a Large-Scale Public Competition and an Observer Study with 11

Radiologists. *Radiology. Artificial Intelligence*, 3(6):e210027, November 2021.

[29] Kiran Vaidhya Venkadesh, Arnaud A. A. Setio, Anton Schreuder, Ernst T. Scholten, Kaman Chung, Mathilde M. W. Wille, Zaigham Saghir, Bram van Ginneken, Mathias Prokop, and Colin Jacobs. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT. *Radiology*, 300(2):438–447, August 2021.

[30] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 243–248, April 2020.

[31] Peng Cao, Xiaoli Liu, Jinzhu Yang, Dazhe Zhao, Wei Li, Min Huang, and Osmar Zaiane. A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recognition*, 64:327–346, April 2017.

[32] David S. Gierada, Thomas K. Pilgram, Melissa Ford, Richard M. Fagerstrom, Timothy R. Church, Hrudaya Nath, Kavita Garg, and Diane C. Strollo. Lung Cancer: Interobserver Agreement on Interpretation of Pulmonary Findings at Low-Dose CT Screening. *Radiology*, 246(1):265–272, January 2008.

[33] Joseph K. Leader, Thomas E. Warfel, Carl R. Fuhrman, Sara K. Golla, Joel L. Weissfeld, Ricardo S. Avila, Wesly D. Turner, and Bin Zheng. Pulmonary Nodule Detection with Low-Dose CT of the Lung: Agreement Among Radiologists. *American Journal of Roentgenology*, 185(4):973–978, October 2005.

[34] National Lung Screening Trial (NLST) - NCI. https://www.cancer.gov/types/lung/research/nlst, 09/08/2014 - 08:00.

[35] Danish Lung Cancer Group. Screening for Lung Cancer. A Randomised Controlled Trial of Low-Dose CT-Scanning. Clinical Trial Registration NCT00496977, clinicaltrials.gov, July 2007.

[36] U. Pastorino, M. Silva, S. Sestini, F. Sabia, M. Boeri, A. Cantarutti, N. Sverzellati, G. Sozzi, G. Corrao, and A. Marchianò. Prolonged lung cancer screening reduced 10-year mortality in the MILD trial: New confirmation of lung cancer screening efficacy. *Annals of Oncology: Official Journal of the European Society for Medical Oncology*, 30(7):1162–1169, July 2019.

[37] Kiran Vaidhya Venkadesh, Arnaud A. A. Setio, Anton Schreuder, Ernst T. Scholten, Kaman Chung, Mathilde M. W Wille, Zaigham Saghir, Bram van Ginneken, Mathias Prokop, and Colin Jacobs. Deep Learning for Malignancy Risk Estimation of Pulmonary Nodules Detected at Low-Dose Screening CT. *Radiology*, 300(2):438–447, August 2021.

[38] Yongfeng Dong, Huaxin Xiao, and Yao Dong. SA-CGAN: An oversampling method based on single attribute guided conditional GAN for multi-class imbalanced learning. *Neurocomputing*, 472:326–337, February 2022.

[39] Zhiqiang Wan, Yazhou Zhang, and Haibo He. Variational autoencoder based synthetic data generation for imbalanced learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, November 2017.

[40] Damien Dablain, Bartosz Krawczyk, and Nitesh V. Chawla. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data, May 2021. Comment: 14 pages, 9 figures.

Appendix A

Training data distributions

Table. A.1 shows the number of positive[1] and negative slices that we to train the YOLOv5 candidate detection model in each experiment. Note that the number of slices does not necessarily scale with the number of annotated nodules in each experiment. This is because larger nodules fall into more slices than smaller nodules.

Table. A.2 shows the number of positive[1] and negative patches we used to train the ResNet50 false positive reduction model in each experiment. The number of patches depends on the number of nodule candidates found by the YOLOv5 candidate detection model.

Table A.1: Input data into the YOLOv5 nodule candidate detection in each experiment.

Dataset	**Original**	**A**	**B**	**C (A+B)**
Total slices	60416	76976	104660	121220
Train	48007	61543	81823	95359
Val	12409	15433	22837	25861
Positive slices (%)	30221 (50.0)	38501 (50.0)	52297 (50.0)	60577 (50.0)
Train	23791 (49.6)	30559 (49.7)	40699 (49.7)	47467 (49.8)
Val	6430 (51.8)	7942 (51.5)	11598 (50.8)	13110 (50.7)

[1] "Positive" in this context does not refer to malignancy but means that the patch/slice contains a nodule.

Table A.2: Input data into the ResNet50 false positive reduction model in each experiment.

Dataset	**Original**	**A**	**B**	**C (A+B)**
Total patches	57623	44990	48244	53569
Train	46359	36430	39207	43629
Val	11264	8560	9037	9940
Positive patches (%)	4811 (8.35)	5055 (11.24)	6446 (13.36)	6908 (12.90)
Train	3784 (8.16)	4084 (11.21)	5453 (13.91)	5909 (13.54)
Val	1027 (9.12)	971 (11.34)	993 (10.99)	999 (10.05)

Appendix B

Detailed baseline results

B.1 Second retraining

Fig. B.1 shows FROC curves of the original and two retrained systems (without oversampling). We do not see major differences, suggesting that our training process is correct. The most notable difference we see is when only looking at primary cancers in Radboudumc, where our first retrained system performs notably worse.

B.2 Individual raters

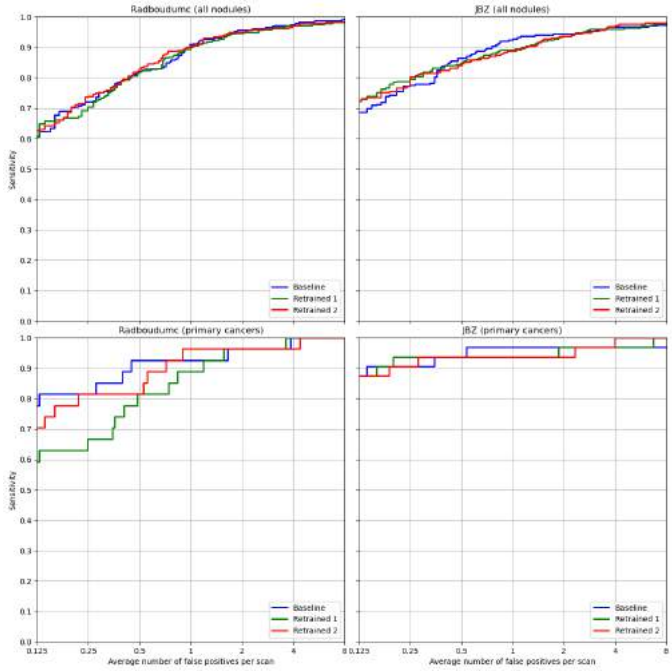
In Fig. B.2, we see the FROC curves of our retrained system when evaluated against a consensus-based ground truth that only counts nodules annotated by a majority of five radiologists and excludes the rest, compared to the same system when evaluated using the annotations of a single radiologist as ground truth.

The system scores better when evaluated using the consensus-based ground truth.

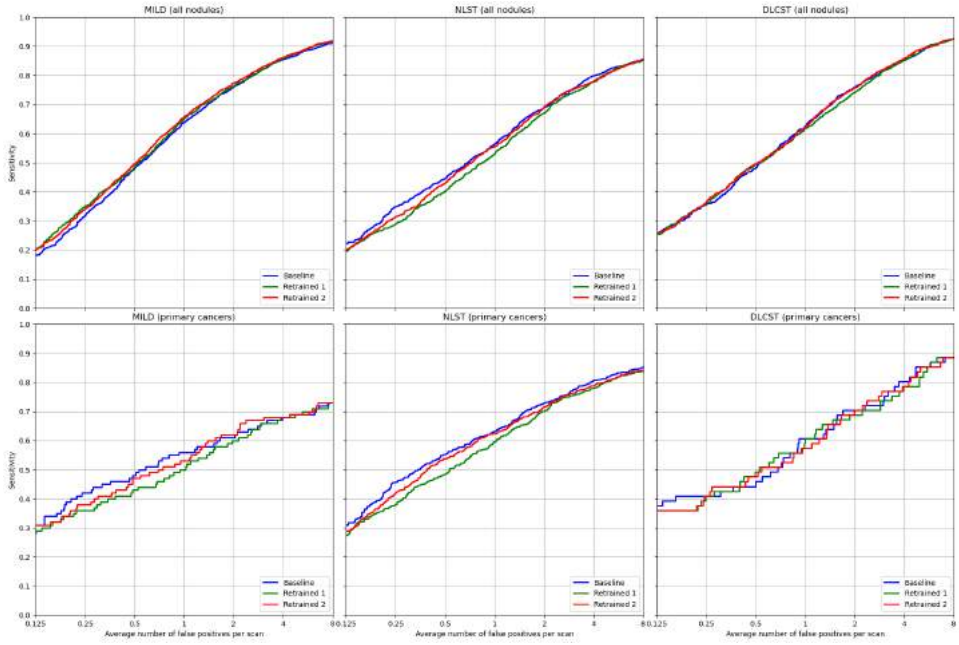
B.3 MILD: earliest vs latest scan

Fig. B.3 shows the performance of our retrained model without oversampling on two different subsets of MILD. The first subset selects the earliest scan of each participant in the trial, while the second subset selects the latest scan of each participant. For both subsets, we then included all scans where cancer was found and included scans without cancer for a total of 1000 scans per subset.

We can see that the system performs better at detecting cancer if we take the latest scan per participant than when we take the first scan.



(a) Performance on clinical datasets.



(b) Performance on screening datasets.

Figure B.1: Performance of original system and two retrained systems (without oversampling).

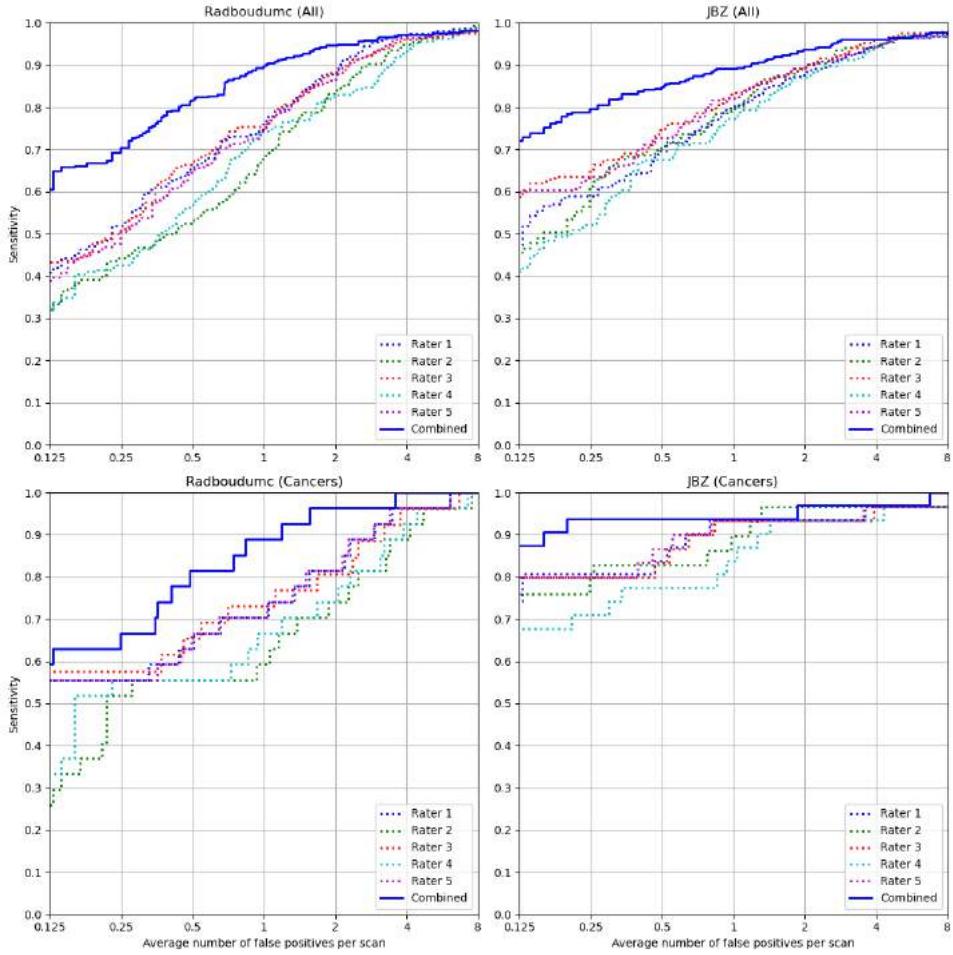


Figure B.2: Performance of retrained models (without oversampling) on clinical datasets, when taking the annotations of a single rater as the ground truth.

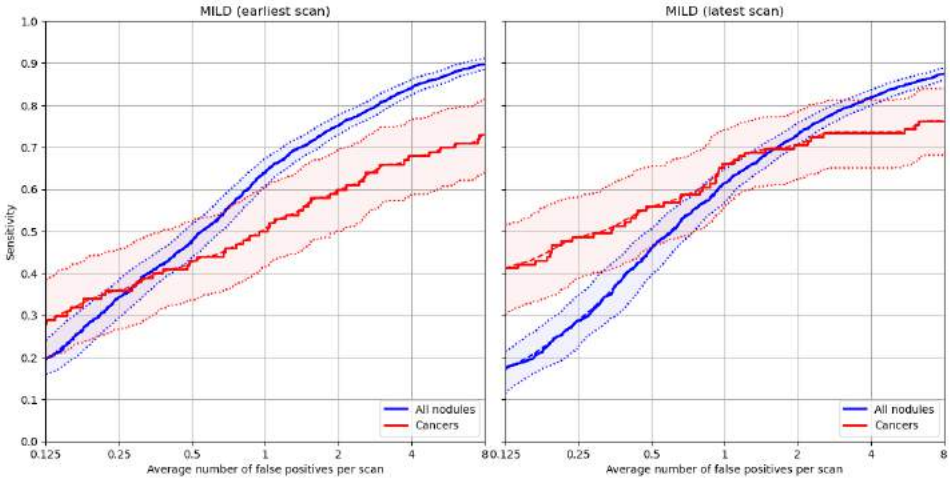


Figure B.3: Performance of retrained models (without oversampling) on our MILD subset when we take the earliest scan per patient, compared to when we take the latest scan per patient.