

Bioinformatics Program
Technical University of Munich
Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Pneumothorax and Chest Tube Classification on Chest X-rays Using Deep Learning

Anna Charlotte Gerhaer

Bioinformatics Program
Technical University of Munich
Ludwig-Maximilians-Universität München

Bachelor's Thesis in Bioinformatics

Pneumothorax and Chest Tube Classification on Chest X-rays Using Deep Learning

**Klassifikation von Pneumothorax und Thorax
Drainagen auf Thorax Röntgenbildern mit
tiefen neuronalen Netzen**

Author: Anna Charlotte Gerhafer
Supervisor: Alessandro Wollek
1st Advisor: PD Dr. rer. nat. Tobias Lasser
Computational Imaging and Inverse Problems, TUM
Boltzmannstr. 11
85748 Garching
Germany
2nd Advisor: Prof. Dr. rer. nat. Mathias Wilhelm
TUM School of Life Sciences
Alte Akademie 8
85354 Freising
Germany
Submitted: 15.09.2022

Bachelor's Thesis Statement of Originality

I confirm that this bachelor's thesis is my own work and I have documented all sources and material used.

Berlin, 15.09.2022

Place and Date


A black rectangular box redacting the signature, with a handwritten mark above it.

Signature

Abstract

In recent years, automated systems based on machine learning (ML) algorithms have been developed to support radiologists in chest X-ray analysis. However, for classification of the medical condition pneumothorax, currently available state-of-the-art ML models show a significant drop in performance when evaluated on untreated cases. This has been ascribed to the fact that the models base their prediction on the existence of a chest tube, which is the treatment of a pneumothorax, instead of the pneumothorax itself.

This thesis investigates if the performance on untreated pneumothorax cases, i.e. cases that contain a pneumothorax and no chest tube, can be improved by adding the task of chest tube classification to the training procedure. Several model architectures have been trained and tested on a number of chest X-ray datasets, including an auxiliary model to approximate the chest tube ground truth on datasets where such labels were otherwise not available. To determine if the performance on untreated pneumothorax cases improved, detailed subgroup analysis was conducted, in addition to a qualitative analysis based on saliency maps.

In den letzten Jahren wurden automatisierte Systeme basierend auf Algorithmen des maschinellen Lernens (ML) entwickelt, um Radiologen bei der Analyse von Thoraxröntgenaufnahmen zu unterstützen. Bei der Klassifizierung der Krankheit Pneumothorax zeigen derzeit verfügbare Algorithmen jedoch einen deutlichen Leistungsabfall auf unbehandelten Fällen. Dies wird auf die Tatsache zurückgeführt, dass die Modelle ihre Vorhersage auf dem Vorhandensein einer Thorax Drainage basieren, welche die Behandlungsmethode eines Pneumothorax darstellt, und nicht auf dem Pneumothorax selbst.

In dieser Arbeit wird untersucht, ob die Performance bei unbehandelten Pneumothorax-Fällen, also Fällen, die einen Pneumothorax und keine Thorax Drainage enthalten, verbessert werden kann durch das Hinzufügen der Klassifikation von Thorax Drainagen zur Pneumothorax-Klassifikation. Es wurden mehrere Modellarchitekturen trainiert und an einer Reihe von Röntgendatensätzen getestet, einschließlich eines Hilfsmodells zur Annäherung an die Ground Truth der Thorax Drainage bei Datensätzen, für die solche Labels nicht verfügbar waren. Um festzustellen, ob sich die Performance bei unbehandelten Pneumothorax-Fällen verbessert hat, wurde zusätzlich zu einer qualitativen Analyse auf der Grundlage von Saliency Maps eine detaillierte Subgruppenanalyse durchgeführt.

Acknowledgments

I would like to sincerely thank PD Dr. Tobias Lasser for giving me the opportunity to work on this interesting topic.

I would like to extend my sincere thanks to Prof. Dr. Mathias Wilhelm for taking on the role as the second advisor and his valuable advice to arrange enough time for careful evaluation of the results.

In particular, I very much wish to express my gratitude to my supervisor Alessandro Wollek for his support and many enlightening discussions from which I have learned a lot.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Goal	2
1.3	Outline	2
2	Theoretical Background and Related Work	5
2.1	Convolutional Neural Networks	5
2.2	Binary Cross Entropy Loss	6
2.3	Focal Loss	7
2.4	Previous and Related Work	7
2.4.1	Machine Learning in Radiology	7
2.4.2	CheXNet	7
2.4.3	Chest Tubes as Confounding Factors	8
2.4.4	Reducing the Bias Regarding Chest Tubes	9
2.5	Metrics and Evaluation	10
2.5.1	Sensitivity	10
2.5.2	Specificity	10
2.5.3	Classification Threshold and the Trade-off between Sensitivity and Specificity . . .	11
2.5.4	ROC Curve	11
2.5.5	AUROC	11
2.5.6	Sampling Error and Bootstrapping	12
2.5.7	Youden Index	12
2.5.8	Grad-CAM	13
2.6	Chest X-ray Datasets	14
2.6.1	CANDID-PTX	14
2.6.2	PadChest	15
2.6.3	ChestX-ray14	16
2.6.4	CheXpert	17
2.6.5	LMU Chest X-Ray Pneumothorax - Chest Tube Dataset	18
3	Methodology	19
3.1	Tooling	19
3.2	Chest Tube Classification	20
3.2.1	Data Splits	20
3.2.2	Model Training and Hyperparameter Tuning	21
3.3	Chest Tube Prediction	23
3.4	Pneumothorax Classification	23
3.4.1	Data Splits	23
3.4.2	Model Training and Hyperparameter Tuning	24
4	Results	27
4.1	Chest Tube Classification	27
4.2	Pneumothorax Classification	28
4.2.1	Evaluation on LMU Dataset Cohorts	28
4.2.2	Evaluation on CANDID-PTX Test Set Cohorts	30

5 Discussion	33
6 Conclusion	35
7 Outlook	37
Bibliography	39
List of Abbreviations	i
List of Tables	ii
List of Figures	ii
Appendix	v
Supplementary Data	v

1 Introduction

1.1 Motivation

Chest X-ray imaging is one of the most frequently applied medical imaging modalities, with more than 1.4 billion procedures being performed annually worldwide [UIW19]. Chest X-ray analysis and corresponding diagnosis require well-trained and experienced radiologists.

In recent years, automated systems based on machine learning (ML) algorithms have been developed to support radiologists in the diagnosis and decision-making process. Currently available machine learning models show comparable performances to radiologists regarding the classification of common thoracic medical conditions [Raj17].

One of the medical diseases covered by such systems is called pneumothorax (PTX). Pneumothorax is a medical condition where air leaks into the space between the lungs and the chest wall. In severe cases, the lung completely detaches from the chest wall and collapses, which can be life-threatening if not treated. The treatment of a pneumothorax usually involves a chest tube (CT) being inserted between the ribs to remove the air.

Pneumothorax Appearance in X-ray Images

Depending on the size of the pneumothorax and the positioning of the patient (e.g., standing up or laying down) in the X-ray image, a pneumothorax can be very hard to identify. In X-ray images, a purely air-filled space would appear as a black region. Therefore, a pneumothorax typically shows up as a vague shadow, slightly darker than its environment. In severe cases, where the lung is col-

lapsed, the outer edge of the lung can become visible as a line running parallel to the chest wall, as it can be seen in figure 1.1, A. In less severe cases, or when the patient is not standing up, it can be significantly more difficult to recognize a pneumothorax.

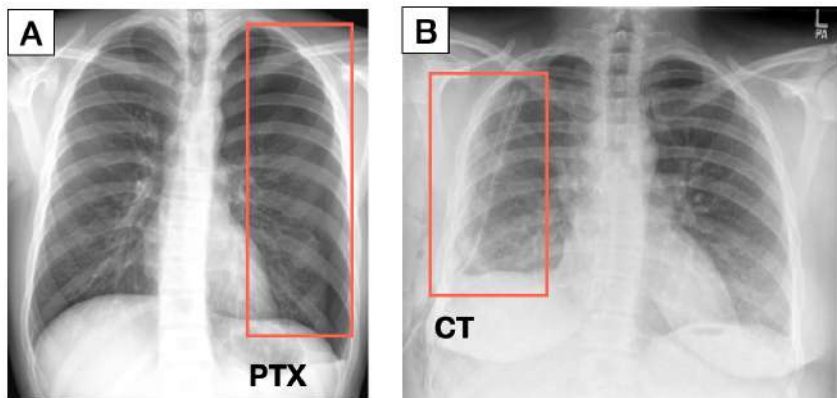


Figure 1.1: A: pneumothorax (PTX), B: chest tube (CT) in X-ray images. The images stem from the CANDID-PTX dataset [Fen21].

The chest tubes, on the other hand, which are inserted to treat pneumothoraces, are typically significantly easier to recognize in an X-ray image than a pneumothorax itself (fig. 1.1, **B**).

Performance Gap and Clinical Utility

Commonly used chest X-ray datasets, such as ChestX-ray14 [Wan17] and CheXpert [Irv19], include pneumothorax cases as well as cases with inserted chest tubes, while only providing labels for pneumothorax. Not surprisingly, one being the treatment for the other, there is a strong correlation between the presence of a pneumothorax and the presence of a chest tube. As a consequence, there is a risk that a machine learning model trained on these datasets learns to recognize pneumothoraces by the presence of a chest tube rather than the appearance of the pneumothorax itself.

Rueckel et al. [Rue20] have discovered that in state-of-the-art ML models there is a considerable gap between the pneumothorax classification performance on cases with and without chest tubes. They found, that the performance was significantly worse on cases without chest tubes. This lead the authors to believe that state-of-the-art models primarily make a pneumothorax prediction based on the existence of a chest tube instead of the pneumothorax itself. In essence, they assumed that the models learn the following shortcut: chest tube \Rightarrow pneumothorax.

In a clinical setting, however, we particularly care about the ability to recognize new pneumothorax cases that have neither been diagnosed nor treated yet, and therefore do not contain chest tubes.

Graf et al. [Gra20] have shown that this bias can be reduced by adding the task of chest tube classification as well as pneumothorax segmentation to the pneumothorax classification task [Gra20]. Inspired by this approach we formed the following hypothesis and investigated it within the course of this project:

The performance on pneumothorax positive cases without a chest tube and the overall pneumothorax performance, respectively, can be improved by solely adding the task of chest tube classification to the pneumothorax classification task.

Thus, we hope the model will learn to distinguish between pneumothoraces and chest tubes and to base its prediction for a pneumothorax on the existence of a pneumothorax instead of a chest tube.

1.2 Goal

The goal of this project is to determine if the performance for pneumothorax positive cases without a chest tube, and the overall pneumothorax performance respectively, can be improved by explicitly introducing chest tube labels and adding the additional task of chest tube classification to the pneumothorax classification task.

1.3 Outline

Publicly available chest X-ray datasets such as ChestX-ray14 [Wan17] do not include chest tube labels. However, there is a smaller chest X-ray dataset called CANDID-PTX

[Fen21], which does include chest tube labels. Therefore, this project was divided into the following four steps:

1. An auxiliary model was trained on the CANDID-PTX dataset to predict chest tubes.
2. The auxiliary model was then used to predict chest tube labels on the larger chest X-ray datasets ChestX-ray14 and CheXpert. Both of those include a higher number of pneumothorax positive images as well as total number of images in comparison to the CANDID-PTX dataset.
3. Pneumothorax classification models were trained on the datasets mentioned in 2. including the noisy chest tube labels as well as excluding those. Additional models were trained on the CANDID-PTX dataset including the ground truth chest tube labels as well as excluding those.
4. In a final step the models' pneumothorax performance was evaluated conducting detailed chest tube and pneumothorax subgroup analysis regarding the established hypothesis.

2 Theoretical Background and Related Work

This chapter summarizes previous and related work as well as foundations regarding machine learning beneficial knowing about for this project.

2.1 Convolutional Neural Networks

A convolutional neural network (CNN) is a special type of an artificial neural network (ANN) that is the most common type or architecture in the area of image processing. An ANN is a machine learning algorithm which is originally inspired by the brain [Hop82].

An ANN consists of a layer of input neurons, followed by an arbitrary number of hidden layers, consisting of an arbitrary number of neurons, followed by an output layer. The neurons of one layer are connected to neurons of the following layer by weights. Those weights determine how much a neuron influences the next layer. Those weights can be changed during backpropagation [GBC16], which is the learning process of the neural network.

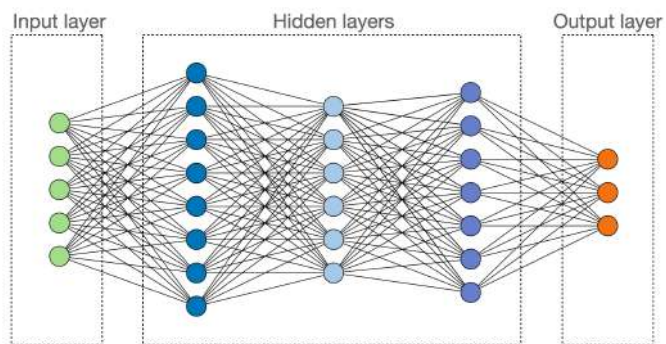


Figure 2.1: Fully Connected Neural Network.

A fully connected neural network is a neural network where for all layer it applies that all neurons of one layer are connected to all neurons of the following layer, as depicted in figure 2.1. One limitation of fully connected neural networks is its struggle regarding computational complexity when it comes to processing large image data.

This is where convolutional neural networks (CNNs) come in. CNNs are characterized by their additional component of a convolutional layer, followed by a pooling layer [LKF10].

A **convolutional layer** can be described as follows; It takes a 3-dimensional array of shape (height, width, n_channels) as input, e.g., an RGB image. The output is the cross-correlation (which is very similar to the convolution) between the input matrix and another matrix, the so-called kernel

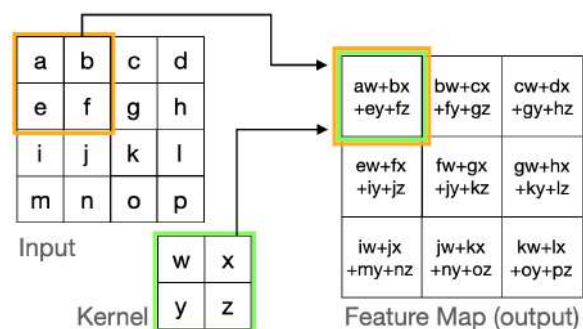


Figure 2.2: Illustration of the cross-correlation.

or filter. The kernel is the set of parameters to learn. The kernel is applied to all sections (restricted by the kernel's size) of the input matrix. This has been illustrated in figure 2.2, with the first step being highlighted. This produces a two-dimensional representation of the image known as a feature map. The kernel is usually of a smaller height and width than the input image while having the same amount of channels. The output of a convolutional layer is a feature map that has as many channels as the convolutional layer has kernels. The kernel's sliding window size is called a *stride*. Padding can be added to the input image.

The **pooling layer** reduces the input to a smaller size by trying to summarize its content [LKF10]. It does this by moving its sliding window with a given stride over the input and summarizing the content of each subsection of the image in one number based on its pooling function. E.g., for max pooling, the maximum of the chosen window is being extracted. This aims at reducing the size of the input, and therefore the following required amount of computation. An example for max pooling as well as average pooling can be seen in figure 2.3.

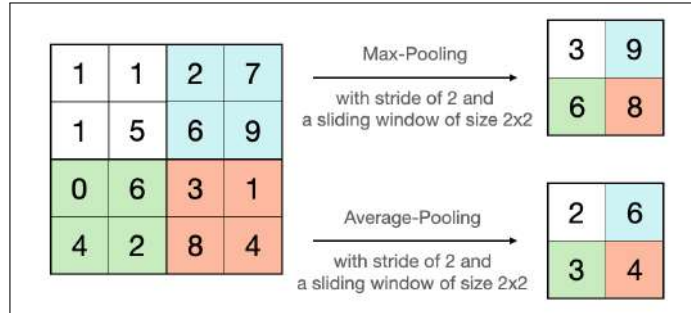


Figure 2.3: Illustration of max pooling and average pooling.

Two well-known and commonly used CNN-based architectures are ResNet18 [He15] and DenseNet121 [Hua16]. The ResNet18 is 18 layers deep and characterized by so-called residual blocks, which allow skip connections within the network. Those are similar to shortcuts by skipping some of the hidden layers. This aims at gaining accuracy by increasingly adding depth [He15].

The DenseNet121 is a CNN consisting of 121 layers and is characterized by its *dense blocks* [Hua16], where each layer receives additional inputs from all preceding layers.

2.2 Binary Cross Entropy Loss

For a neural network to learn by performing backpropagation [GBC16], a loss function needs to be defined. Based on this function, a loss can then be computed for a given input based on the model's output, and the input's ground truth label during the optimization process.

One such function is the binary cross entropy (BCE) loss [Bis06]. For the BCE loss, each prediction is being compared to the ground truth, which for a binary problem is 0 or 1. The loss is then computed based on the difference of the prediction and the corresponding ground truth value, as depicted in equation 2.1, with N = number of samples, y_i = ground truth for sample i and p_i = prediction for sample i .

$$BCELoss := -\frac{1}{N} \sum_{i=0}^{N-1} (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \quad (2.1)$$

By applying the logarithm the punishment for a big difference is significantly higher than for a small difference.

The binary cross entropy loss is a special case of the cross entropy loss, which can be applied to multiclass problems with N samples and M classes:

$$CrossEntropyLoss := -\frac{1}{N} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} y_{ij} \log(p_{ij}) \quad (2.2)$$

2.3 Focal Loss

The focal loss is based on the cross entropy loss, but additionally takes into account class imbalance [Lin17]. By adding the factor $(1 - pt)^\gamma$ it tries to focus learning on strongly misclassified samples [Lin17]. With $\gamma = 0$, the focal loss is equal to the cross entropy loss. By setting γ to greater than zero, it reduces the weight of well classified samples, and consequently increases the weight of strongly misclassified samples. The focal loss is defined as follows:

$$FocalLoss(p_t) = (1 - p_t)^\gamma \log(p_t); \text{ where } p_t = yp + (1 - y)(1 - p) \quad (2.3)$$

2.4 Previous and Related Work

This section provides a summary of related work that has recently been published in the area of artificial intelligence (AI) in thoracic radiology.

2.4.1 Machine Learning in Radiology

Machine learning can be useful in a variety of areas regarding chest X-ray classification.

One option to support radiologists and improve their performance is to have overcautious machine learning models make a preselection.

Another useful application of such models could be in places where there is a shortage of expert radiologists while X-ray imaging machines are available. In this case, the machine learning models could potentially undertake a radiologist’s task to examine a given chest X-ray image.

Another scenario is to integrate machine learning models into the clinical workflow by offering assistance to radiologists, as in an AI assistant, inspired by the 4-eye principle.

2.4.2 CheXNet

CheXNet is a machine learning algorithm that was published by P. Rajpurkar et al. in 2017 in the paper *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning* [Raj17]. It predicts 14 common thoracic diseases in chest X-ray images. The proposed model is a DenseNet121 that has been trained on the ChestX-ray14 dataset [Wan17]. CheXNet can keep up with previous models on all classes and even outperforms them on 4 classes (Emphysema, Mass, Nodule and Pneumonia).

Additionally, a DenseNet121 was trained on the ChestX-ray14 dataset to only predict Pneumonia. This model was evaluated based on the F1-score [TH15] and compared to the performance of four radiologists, separately, as well as averaged. While it outperformed all of them, it was noted that several factors probably negatively influence the

radiologists’ performance in comparison to a clinical setting. Firstly, the radiologists only got to examine frontal X-ray images, whereas in a clinical setting, they usually can take a look at multiple X-ray images taken from different positions, e.g., lateral and frontal. Secondly, the radiologists, as well as the model, were not allowed to consider any background information on the patients, such as sex, age or their medical history, which has been shown to decrease radiologists’ performance when examining X-ray images.

This lead the authors P. Rajpurkar et al. to the assumption that their setup provides a conservative estimate of performance.

2.4.3 Chest Tubes as Confounding Factors

Rueckel et al. examined state-of-the-art chest X-ray classification models for potential biases in the paper *Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs* [Rue20]. They hypothesized and showed that state-of-the-art machine learning models that have been trained on large publicly available chest X-ray datasets do not adequately consider confounding features such as chest tubes[Rue20]. Additionally, it is stated that factors such as the size of a pneumothorax are not sufficiently being considered[Rue20].

They found that the overall performance of state-of-the-art models on predicting pneumothoraces is misleading: It has been shown that the performance drops significantly when the model is evaluated on untreated pneumothorax cases only, thus pneumothorax positive cases without chest tubes.

In their experiment, Rueckel et al. examined two state-of-the-art machine learning models: *AI-1.5* and *AI-CheXNet*. Both were evaluated on an annotated benchmarking cohort. The dataset included detailed labels for pneumothorax, including location and size, as well as chest tubes. This allowed for detailed subgroup analysis, which showed that even though both algorithms perform seemingly well when evaluated on the entire dataset, they actually show significant differences in performance when looking at the subgroup analysis.

The subgroup analysis was conducted on the cohorts depicted in figure 2.4. Here, the gray colored subset is equal to the entire dataset. On the green colored subset, a model which makes a pneumothorax prediction based of the existence on a chest tube instead of a pneumothorax, performs very well. On the red subset, on the other hand, such a model would perform

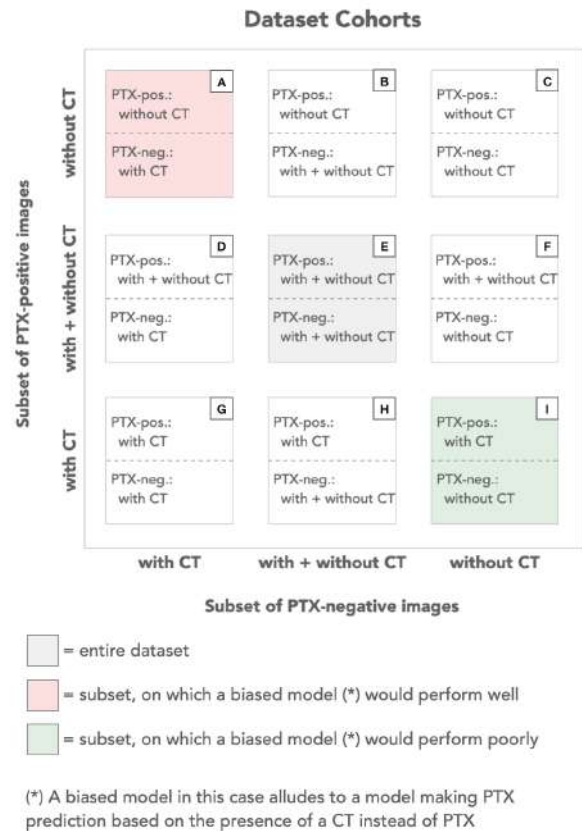


Figure 2.4: Dataset’s cohorts based on pneumothorax (PTX) and chest tube (CT) combinations.

very poorly.

In the detailed subgroup analysis it showed that the algorithms have a 0.205 (*AI-1.5*) and 0.295 (*AI-CheXNet*) AUROC performance gap for the following subsets:

1. On the subset **I** (in figure 2.4) the AUROC was 0.763 (*AI-1.5*) and 0.837 (*AI-CheXNet*).
2. On the subset **A**(in figure 2.4) the AUROC dropped to 0.558 (*AI-1.5*) and 0.542 (*AI-CheXNet*).

As previously mentioned, when evaluated on the entire dataset, their AUROC scores seemed to be fairly high, which the authors ascribed to the fact that a relatively high proportion of pneumothorax positive chest X-ray images in the dataset includes chest tubes. This misleadingly resulted in their overall pneumothorax performance suggesting a higher clinical utility.

Those results lead the authors to believe that chest tubes are relevant confounders that have not been considered in the models’ training set.

In the further course, we will abbreviate the bias just described as the *bias regarding chest tubes*.

2.4.4 Reducing the Bias Regarding Chest Tubes

In *Pneumothorax and chest tube classification on chest x-rays for detection of missed pneumothorax* [Gra20] Graf et al. reduced the bias regarding chest tubes in state-of-the-art machine learning algorithms (described in 2.4.3) by adding the task of classifying chest tubes to pneumothorax classification. Additionally, they added a pneumothorax segmentation model to the pneumothorax classification pipeline.

The authors achieved comparable performance for pneumothorax cases with chest tubes as well as for pneumothorax cases without chest tubes in contrast to previous work.

The proposed classification pipeline consisted of four different models. For chest tube classification, a DenseNet121 architecture was used. For pneumothorax classification, they used an ensemble of three models: First, a DenseNet-121, second, a U-Net to segment the pneumothorax region, and a third method which extracts four patches from specific lung regions, which are most likely to contain a small pneumothorax [Gra20].

The models were trained on a combination of images from a private dataset and the ChestX-ray14 dataset [Wan17].

This work inspired the idea for this project, i.e., reducing the chest tube bias in pneumothorax prediction by adding the task of chest tube classification to the pneumothorax classification problem.

2.5 Metrics and Evaluation

This section provides a short overview of metrics and evaluation methods used in this project to quantify and visualize the statistical performance of the machine learning models.

2.5.1 Sensitivity

The sensitivity of a binary classifier is defined as the rate of truly positive samples that are correctly identified as positives by the classifier:

$$\text{Sensitivity} := \frac{TP}{TP + FN}; \text{ where } TP = \text{True Positives}, FN = \text{False Negatives} \quad (2.4)$$

Sensitivity is also known as *true positive rate* (TPR) as well as *recall* and has the following properties:

- Values lie in the range $[0.0, 1.0]$.
- All other things being equal, a higher value is desirable.
- As an example, a pneumothorax classifier with a sensitivity of 0.85 recognizes 85% of X-ray images with a pneumothorax as such.
- We only need samples belonging to the positive class to measure the sensitivity. Negative samples are not required.
- Negative samples in the test data do not influence the sensitivity. Therefore, the sensitivity is not affected by the class distribution in the test data.

2.5.2 Specificity

Analogous to the sensitivity, the specificity of a binary classifier is defined as the rate of truly negative samples that are correctly identified as negatives by the classifier:

$$\text{Specificity} := \frac{TN}{TN + FP}; \text{ where } TN = \text{True Negatives}, FP = \text{False Positives} \quad (2.5)$$

Specificity is also known as *true negative rate* (TNR) and has the following properties:

- Values lie in the range $[0.0, 1.0]$.
- All other things being equal, a higher value is desirable.
- As an example, a pneumothorax classifier with a specificity of 0.99 recognizes 99% of X-ray images without a pneumothorax as such.
- We only need samples belonging to the negative class to measure the specificity. Positive samples are not required.
- Positive samples in the test data do not influence the specificity. Therefore, just like the sensitivity, the specificity is not affected by the class distribution in the test data.

2.5.3 Classification Threshold and the Trade-off between Sensitivity and Specificity

The raw output of a binary classifier is typically a continuous number that is correlated to the model's confidence that its input belongs to the positive class. In order to derive a binary classification from the continuous output, a threshold needs to be chosen, above which the output is considered to be positive. The choice of this threshold directly influences the model's sensitivity and specificity, and there is a trade-off between the two. A lower threshold corresponds to a higher sensitivity but results in reduced specificity. Picking a higher threshold increases the specificity but reduces the sensitivity, i.e., the model will recognize fewer positive samples as positive.

There is no universal solution to what the best choice of a threshold is; but instead, it depends on the application.

2.5.4 ROC Curve

A receiver operating characteristic (ROC) curve is a plot that visualizes the trade-off between sensitivity (2.4) and specificity (2.5) for a given model and dataset (fig. 2.5) [Faw04]. The procedure to generate such a ROC curve can be described as follows:

- Use the model to compute one continuous number (score) for each sample in the dataset.
- Sort the resulting scores, and append an additional score slightly above the maximum of the existing ones.
- For each of these scores, set the classification threshold equal to the score and compute the resulting sensitivity and specificity values on the dataset.
- Visualize the pairs of (sensitivity, specificity) in a scatter plot. For historical reasons, the horizontal axis is used for 1 - specificity, the false positive rate (FPR). The vertical axis is used for the sensitivity.
- Connect the dots in the scatter plot with straight line segments.

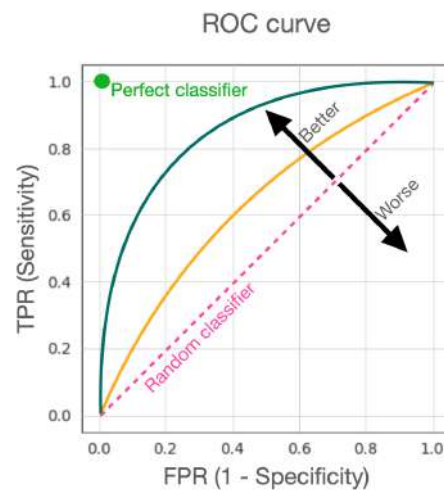


Figure 2.5: ROC curve illustration including a perfect and random classifier.

It is worth noting that the classification thresholds themselves do not appear in the ROC curve, and their values do not influence the shape of the curve.

2.5.5 AUROC

The area under the ROC curve (AUROC, also known as ROC AUC) [Faw04] is a popular metric for binary classifiers. It has the following noteworthy properties:

- The value falls into the range $[0, 1]$. Higher values are preferable.
- Invariance with respect to the classification threshold, as the ROC curve takes into account all possible thresholds.
- Closely related to the previous property, the scale of the values of the predicted scores do not influence the AUROC. For example, the AUROC is the same for a model's probabilities after calibration as well as before.
- It is equivalent to the probability that the model predicts a higher score for a randomly chosen positive sample than for a randomly chosen negative sample [Faw06].
- Consequently, the expected value of the AUROC score of a random classifier is 0.5.
- Different ROC curves can have the same AUROC.
- As a consequence of the sensitivity and specificity not being affected by the class distribution in the test data, the AUROC score shares this property. In particular, this means that strong class imbalances in the test data do not influence the AUROC score.

Many of these properties are desirable in some situations and inappropriate in others. For example, the AUROC, on its own, is not a suitable metric in use cases where well-calibrated probabilities are required. Nor is it suitable in situations where the specificity is more important than the sensitivity, or vice versa. In contrast, if all we need is a summary of a classifier's overall performance for different thresholds, the AUROC can be suitable.

2.5.6 Sampling Error and Bootstrapping

In all practical machine learning applications, the available test data is a limited subset of the real distribution of data that a model is expected to be confronted with once deployed. As a consequence, every statistic we compute on the test data is associated with a sampling error. Bootstrapping [IBM22] is a popular method to estimate the magnitude of such a sampling error. The available dataset is resampled N times with replacement, and for each resampled dataset (having the same size as the original dataset), the quantity of interest (e.g., the AUROC score) is computed. The sampling error can now be estimated by calculating the standard deviation of these N values. If the values are normally distributed, the 95% confidence interval can be derived by multiplying the standard deviation by 1.96 [YalC].

2.5.7 Youden Index

One common option is to choose a threshold based on getting the highest value of the youden index [SCP11]. The youden index is defined as follows:

$$\begin{aligned} \text{YoudenIndex} &:= \text{Sensitivity} - (1 - \text{Specificity}) \\ &= \text{Sensitivity} + \text{Specificity} - 1 \end{aligned} \tag{2.6}$$

Choosing the threshold where the youden index is highest aims at providing a good trade-off between sensitivity and specificity by maximizing the sum of sensitivity and specificity.

2.5.8 Grad-CAM

Grad-CAM [Sel19] is short for *Gradient-weighted Class Activation Mapping* and is a technique to visualize which region of an image a CNN-based classification model pays attention to when making a prediction. This aims at making convolutional neural networks more transparent and explainable.

Often, there is a trade-off between performance and interpretability: Rule-based models, for example, can be easily followed and understood in their decision-making but are outperformed by less interpretable neural networks in many tasks. Those, on the other hand, are often called *black boxes* [BMA19], as in the magic happens behind closed doors, because it is hard to understand what makes them come to a certain decision. However, nowadays, in many common machine learning tasks, including image classification, neural networks outperform rule-based algorithms.

Researchers have come up with various methods to make neural networks, and image classifiers in particular, more interpretable. Grad-CAM is one of them, as it can be used to visualize how much influence different regions in an image have on the model's output.

It is class-discriminative, meaning that a heatmap localizes the observed class in the image. An example is shown in figure 2.6. Here we have four input images, with all four including cats. In the according Grad-CAM visualizations it can be seen that the cats' location is being distinctively highlighted.

Making ML models more explainable is an important factor, especially when trying to incorporate them in a clinical workflow. If a doctor does not understand why a model makes a certain decision, it will be harder to integrate it, and it probably won't be used as much.

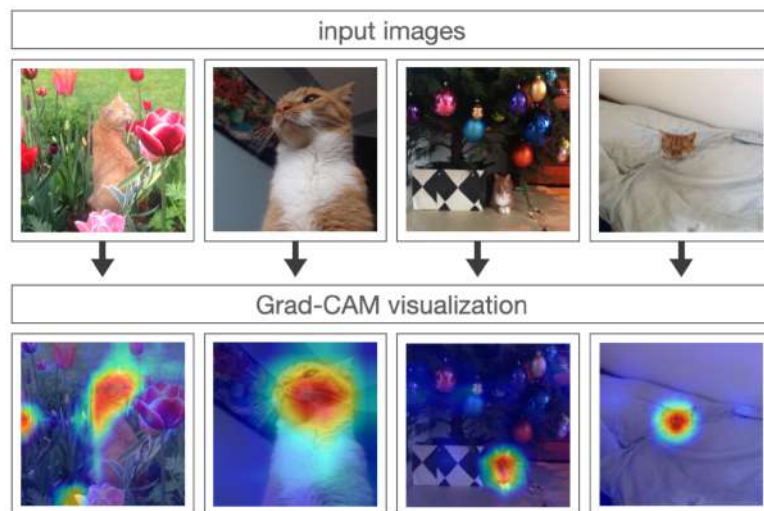


Figure 2.6: Grad-CAM visualizations for ResNet18, pre-trained on ImageNet [Den09], for the class *tabby cat*.

Operating principle

Grad-CAM is based on a technique called *Class Activation Mapping* (CAM) [Zho15]. CAM is only applicable to CNN-based models that have a *Global Average Pooling* (GAP) layer included as a second-last step.

CAM assigns meaning to each position in the final convolutional layer by computing the linear combination of activations, which are being weighted by the according output

weights for the considered class. The resulting heatmap is then being mapped onto the input image. Due to the architecture of a CNN it is known that the activation in a given region, is directly linked that region of the according input image.

In comparison to CAM, Grad-CAM can be applied to all kinds of CNNs. It generalizes CAM by including gradient information: The heatmap is generated by using the gradients that arrive at the last convolutional layer of a given model. More precisely, the gradients of the final layer are used as weights. The map is then created by computing the weighted sum of the final feature maps while using those weights.

Areas of application

Grad-CAM can be applied to various kinds of machine learning problems, such as image classification, image captioning, and visual question answering models. Grad-CAM can be a useful tool at different stages of a deep learning project.

During the model development, it can be useful for error diagnosis to potentially identify why a model performs poorly or if it misunderstands a task. For example, it would be suspicious if a chest X-ray model's output were determined almost exclusively by the pixels in the upper right corner of an X-ray image, where usually no part of the human body is depicted, but instead only the background.

In a deployed setting, Grad-CAM can be used for the user to gain trust in the model so that the model is being used regularly and finds its place in everyday usage. This scenario is very applicable to current state-of-the-art models in chest X-ray examination. We have well-performing models and now we need to establish radiologists' trust in them so that they are actually being used instead of being forgotten. We neither want experienced radiologists to lose trust in a well-performing model and to consequently abandon the model only because of incorrect predictions which might seem obvious to an expert radiologist nor do we want to irritate or confuse inexperienced radiologists by incorrect predictions.

2.6 Chest X-ray Datasets

This section provides a brief overview of chest X-ray datasets that were used in this project. More detailed information about the way these datasets were used will be provided in section 3.

2.6.1 CANDID-PTX

The CANDID-PTX dataset [Fen21], short for *Chest x-ray Anonymised New Zealand Dataset in Dunedin-Pneumothorax*, includes 19,237 frontal chest X-ray images. Those were collected from 13,745 patients at Dunedin Hospital in New Zealand. 10,278 out of the 19,237 images were obtained from male patients and 8,929 images were obtained from female patients. The patients' age ranges from 16 to 101 years. The mean age of the patients is 60.1 years with a standard deviation of 20.1 years.

The images are saved as DICOMs, short for *Digital Imaging and Communications in Medicine*, and are of size 1024 x 1024 pixels.

The dataset includes the following three class labels: chest tube, pneumothorax and rib fracture. Those are not mutually exclusive, as it can be seen in figure 2.7. About 1 in 6 images has at least one of the labels, which results in a 1:5 positive-to-negative ratio. There are 3,196 pneumothorax positive images obtained from 1,036 patients, 335 rib fracture-positive images from 217 patients and 1,423 chest tube positive images from 583 patients. Ground truth labels were assigned for each image after being reviewed by multiple physicians and/or radiologists. The dataset includes bounding box annotations for rib fracture cases, and segmentation annotations for chest tube and pneumothorax cases.

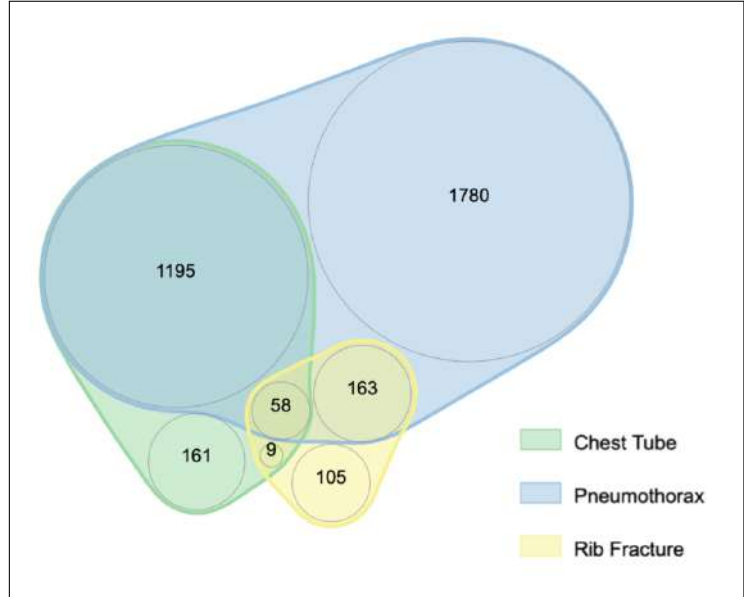


Figure 2.7: Label distribution in CANDID-PTX dataset, venn diagram created using nVennR [PAQ18].

The dataset is publicly available and has been published in 2021.

2.6.2 PadChest

The PadChest dataset [Bus20], short for *Pathology Detection in Chest radiographs*, includes 160,861 images obtained from 69,882 patients. The images were acquired from 2009 to 2017 and annotated by radiologists from San Juan Hospital in Alicante, Spain. The dataset includes images from a total of six different position views, such as frontal and lateral.

The reports are labeled with 174 different radiographic findings, 19 differential diagnoses and 104 anatomic locations. The labels include but are not limited to pneumothorax and several different kinds of tubes, such as nasogastric tubes and chest drain tubes.

27% of the acquired medical reports were manually reviewed by trained physicians. The remaining reports were labeled by a recurrent neural network with attention mecha-

Subset sizes	CANDID-PTX	PadChest	ChestX-ray14	CheXpert	LMU dataset
#total imgs	19,237	160,861	112,120	224,316	6,434
#PTX-pos.	3,196	411	5,302	17,313	1,652
#CT-pos.	1,423	612	-	-	1,865

Table 2.1: Dataset overview: Listed are the total number of images included in each dataset as well as the absolute number of images regarding pneumothorax (PTX) and chest tube (CT) positive images. If a label is not included in a dataset, the according cell is marked with a dash (-).

nisms. The image sizes range from 1500 to 4280 pixels in height and from 1760 to 3520 pixels in width. They are saved as PNG files.

The dataset is publicly available and has been published in 2019.

2.6.3 ChestX-ray14

The ChestX-ray14 dataset [Wan17] includes 112,120 frontal chest X-ray images from 30,805 patients. The images were collected from the clinical PACS database at the National Institutes of Health Clinical Center. The acquired images stem from studies that were performed between 1992 and 2015. The dataset includes class labels for 14 common chest pathologies, as listed in 2.2. Labels were extracted from the associated medical reports using natural language processing (NLP). The labels are not mutually exclusive, and 51,708 images contain at least one of the pathologies, leaving 60,412 images without any pathologies. The dataset includes 5,302 pneumothorax positive images. In addition to the classification labels for all images, bounding box annotations are provided for approximately 100 images per class.

For this dataset, a train-test split has been provided by the authors [Wan17]. The given training set includes 86,524 images, with 2,637 being pneumothorax positive. The test set includes the remaining 25,596 images, with 2,665 being pneumothorax positive. Therefore the pneumothorax positive-to-negative ratio in the training set is 1:33, wherein the test set provides a pneumothorax positive-to-negative ratio of 1:10. A similar pattern can be seen for the other classes represented. There is no further information on how this split was produced or what the intentions were when making this split.

It has been ensured that all images from a patient are either part of the training set or all images from a patient are part of the test set.

One of the advantages of this dataset is the large amount of images included. Another benefit is that bounding box annotations are provided for a small part of the dataset.

A limitation, on the other hand, is that the image labels are extracted by NLP. Its labeling accuracy has been estimated to be higher than 90%.

The dataset is publicly available and has been published in 2017.

Table 2.2: ChestX-ray14 label distribution: Number (#) of positive (pos.) images for each class.

Class	# pos. images
Atelectasis	11,535
Cardiomegaly	2,772
Consolidation	4,667
Edema	2,303
Effusion	13,307
Emphysema	2,516
Fibrosis	1,686
Hernia	227
Infiltration	19,871
Mass	5,746
Nodule	6,323
Pleural Thickening	3,385
Pneumonia	1,353
Pneumothorax	5,298
No Finding	60,412

2.6.4 CheXpert

The CheXpert dataset [Irv19], short for *Chest eXpert*, consists of 224,316 frontal and lateral chest X-ray images. The images were collected from Stanford Hospital along with according medical reports. They were acquired from studies that were performed between October 2002 and July 2017 in inpatient centers as well as outpatient centers. Labels were extracted from medical reports for 13 medical diseases, as listed in 2.3, based on medical relevance and frequent occurrence in the reports. The additional label *No Finding* was introduced for the images that do not show any of the pathologies listed above. Only 8.86% of the images have the label *No Finding*. The dataset captures uncertainties from the medical reports by introducing the label *uncertain* alongside *positive* and *negative*.

The labels were extracted from the free text medical reports by an automated labeler that extracted observations based on provided rules. The labeler worked on three different levels. First, it checked if given words or sentences occurred in a given medical report. Next, it classified all mentions as positive, negative or uncertain. In a last step, all the classified mentions have been condensed to a final label for the 14 observations.

Additionally, two sets have been provided, that were annotated by board-certified radiologists: First, a validation set consisting of 200 images from 200 different patients, which has been manually annotated by 3 radiologists, and secondly, a test set including 500 images from 500 different patients, which have been manually annotated by five radiologists.

Having those small additional manually annotated sets is a benefit compared to other available chest X-ray datasets, which presumably have a higher number of incorrectly labeled cases due to NLP labeling.

The dataset is publicly available and has been published in 2019 [Irv19]. Two versions of this dataset are available:

- CheXpert-v1.0 (440 GB): contains the original images sized at approximately (3000, 3000) pixels.
- CheXpert-v1.0-small (11 GB): contains downscaled images of approximate size of (390, 320) pixels.

Table 2.3: CheXpert label distribution: Number (#) of positive (pos.) images for each class.

Class	# pos. images
Atelectasis	29,333
Cardiomegaly	23,002
Consolidation	12,730
Edema	48,905
Enlarged Cardiom.	9,020
Fracture	7,270
Lung Lesion	6,856
Lung Opacity	92,669
Pleural Effusion	75,696
Pleural Other	2,441
Pneumonia	4,576
Pneumothorax	17,313
Support Devices	105,831
No Finding	16,627

2.6.5 LMU Chest X-Ray Pneumothorax - Chest Tube Dataset

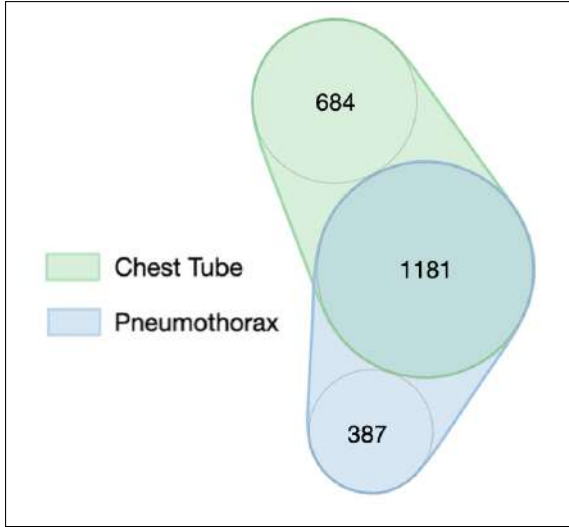


Figure 2.8: Label distribution in the LMU dataset, venn diagram created using nVennR [PAQ18]

The *LMU Chest X-Ray Pneumothorax - Chest Tube Dataset* [Rue20] (*LMU dataset*) includes 6,434 supine frontal chest X-ray images from a total of 4,688 patients. The images were taken between 2010 and 2018 and were acquired from Ludwig-Maximilian-Universität (LMU) Hospital, Munich. The X-ray images are radiologically annotated for pneumothorax size ($<1\text{cm}$, $1\text{-}2\text{cm}$, $>2\text{cm}$) and location (left or right) as well as inserted chest tubes.

The dataset includes 1,652 pneumothorax positive images from 1,173 patients and 4,782 pneumothorax negative cases from 3,515 patients. 1,865 out of the 6,434 images are labeled positive for a chest tube. The age of the patients is between 35 to 83. 41.4% of the images are from female patients. The images are saved as DICOMs

along with a CSV file including the labels.

The images were reviewed and labeled by two well-trained 4th-year medical students. They labeled the first 50 images with direct supervision of a radiology resident. The remaining images, they labeled independently. Uncertain cases were left for review by a radiology resident with three years of experience in thoracic imaging.

This dataset was gathered with the intention of creating a challenging benchmark dataset that would allow for detailed subgroup analysis. Moreover, this dataset was created with the intent to quantify a potential bias regarding chest tubes [Rue20]. Rueckel et al. stated that only supine chest X-rays were included since, among other things, those cases have a higher prevalence of chest tubes due to a high proportion of intensive care unit (ICU) patients. Additionally, ICU patients are more likely to have a concomitant disease, which can negatively influence a model's pneumothorax detection performance [Rue20].

This dataset, unlike the others previously mentioned, is not publicly available.

3 Methodology

The source code for the following experiments is available at <https://github.com/anna-charlotte/ptx-and-ct-classification>. For access to the final chest tube classification model and the final pneumothorax classification models, please contact the author.

3.1 Tooling

The classification models were implemented in Python [VD09] using the deep learning frameworks PyTorch [Pas19] and PyTorch Lightning [FT19].

The framework Ray Tune [Lia18] was used to control the hyperparameter search space and manage the experiments.

Besides monitoring the models' performance (AUROC and loss) on TensorBoard [Aba15], Grad-CAM was additionally used to help identify potential bugs. An example of models' suspicious behavior that was observed during training can be seen in figure 3.1. Here, the Grad-CAM visualizations indicate that the models mostly focused on the pixels in the upper right corner as well as the pixels located at the bottom of the X-ray images. Neither of those locations are areas where we would expect a model to check for pneumothorax as well as chest tubes. This does not tell where the error occurred, but it indicates suspicious behavior.

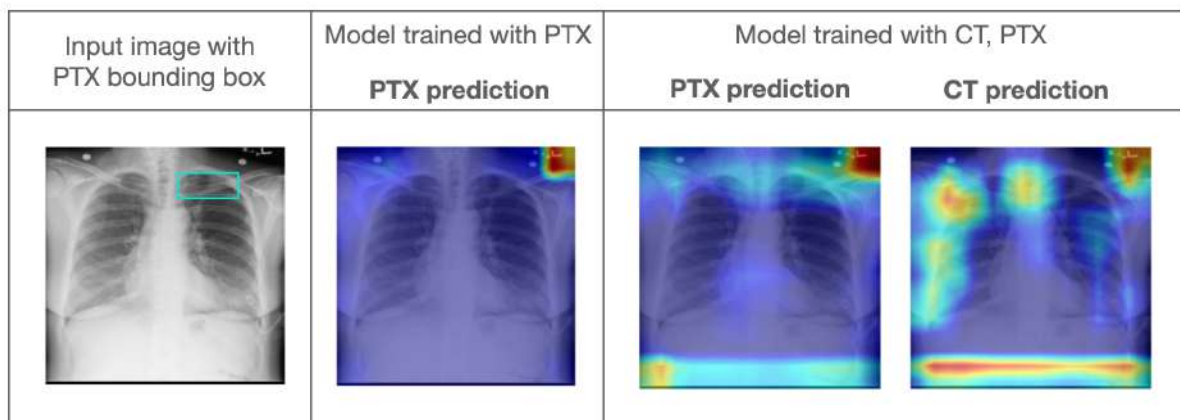


Figure 3.1: Grad-CAM used for error analysis: Using Grad-CAM during training assisted with identifying suspicious behavior: Models' focus on upper right corner, as well as the bottom line of some images. The image stems from the ChestX-ray14 dataset [Wan17].

3.2 Chest Tube Classification

3.2.1 Data Splits

This section will give a brief overview of the datasets and data splits used for training, validation and testing of the auxiliary chest tube classification model.

CANDID-PTX

For chest tube classification, the CANDID-PTX dataset [Fen21] was used for training, validation and testing. One image was removed due to being assigned to two different patients, which is not possible and therefore must be erroneous¹. This resulted in a total of 19,236 images.

The dataset was split according to a 70-10-20 train-val-test split, which resulted in 13,465 training images (70%), 1,925 validation images (10%) and 3,846 test images (20%). The label distribution was kept similar for all splits, as it can be observed in figure 3.2. The splits were made while taking into account the patient ids to ensure that all images from each patient only occur in one split, not within multiple splits. This applies to all splits being made in the further course of this project.

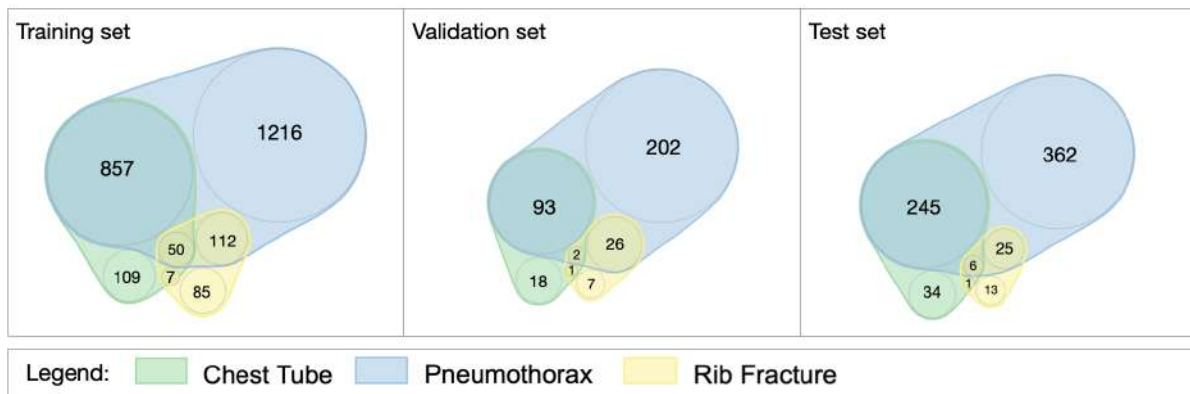


Figure 3.2: Label distribution for CANDID-PTX training, validation and test sets.

PadChest

The PadChest dataset [Bus20] was used for additional validation and testing of the chest tube classification model that was trained on the CANDID-PTX dataset. This way it was ensured, that the model was additionally validated and tested on X-ray images from a distribution that is different from the training set. This was particularly important since the chest tube classification model, later on, would be used to predict chest tube labels on a dataset that would also be from a different distribution.

The provided labels include seven different kinds of tubes. According to LMU radiologists, the class *chest drain tube* is the subset of tubes included in PadChest, which is the equivalent to the *chest tube* in CANDID-PTX, as illustrated in figure 3.3.

¹Image ID of removed image: 7.6.23.312237.44.944525357435.4597.1820932216019, assigned to patient ID's 10463 and 10464

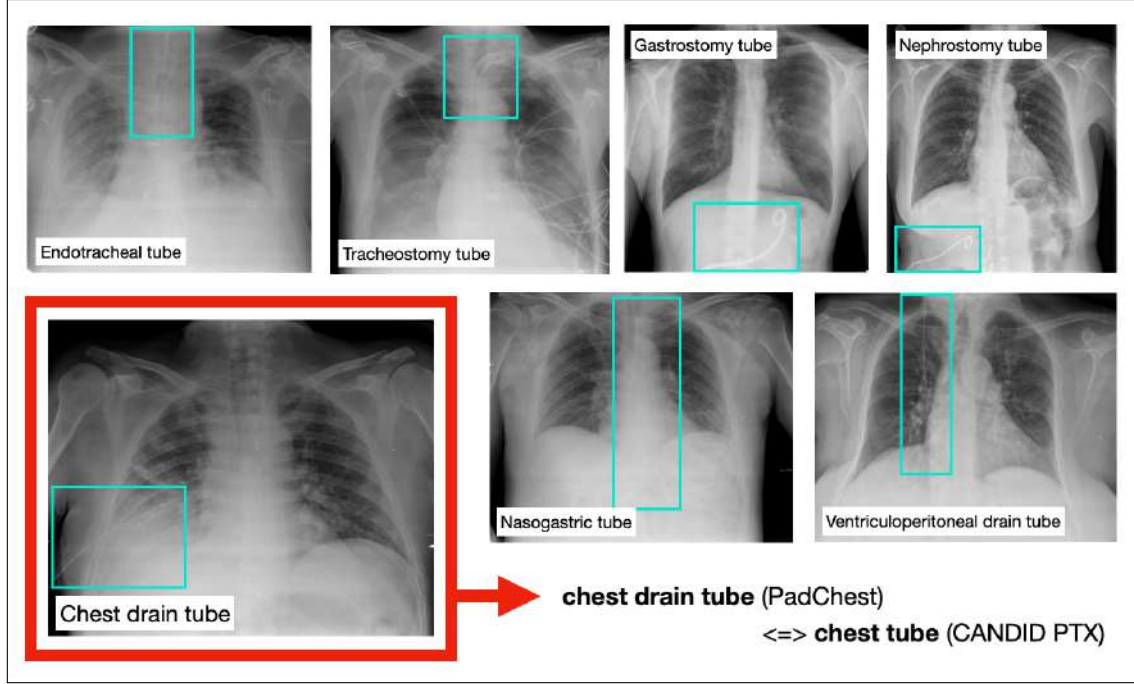


Figure 3.3: Classes including tubes in PadChest: The PadChest dataset includes seven different tubes. PadChest’s class *chest drain tube* is equivalent to the class *chest tube* in the CANDID-PTX dataset. The images stem from the PadChest dataset [Bus20].

In the course of this project, only part of the PadChest dataset was used. First, 103 images were removed due to their label being NaN, which resulted in a total of 160,758 images, with 612 being chest drain tube positive. Filtering for frontal chest X-rays only resulted in 110,551 images, with 494 being chest drain tube positive with a positive-to-negative ratio of 1:223. All positive images and twice the amount of negative images were extracted, which resulted in a total of 1482 images. Of those, one-third was used for validation (494 images) and the remaining 988 images for testing. When the splits were made it was ensured that all images from each patient only occurred in one split, not within multiple splits.

3.2.2 Model Training and Hyperparameter Tuning

Four different network architectures were part of the hyperparameter search space: EfficientNet B02, MobileNetV2, ResNet18 and VGG11. The weights were initialized with weights from a model that was pre-trained with images from ImageNet [Den09] or random weights.

The images were downsampled to different sizes. Due to GPU memory constraints, smaller batch sizes had to be used for larger image sizes; batch size 8 was used for size (1024, 1024), 16 for size (512, 512), and 32 for size (256, 256). Afterwards the images were normalized based on the mean and standard deviation of the ImageNet [Den09] training set images.

Different learning rates were tried out, ranging from 1e-02 to 1e-07, with and without reducing the learning rate by a factor ranging from 2 to 10 when reaching a plateau on the validation loss (ReduceOnPlateau). For optimizers, SGD and Adam optimizers with

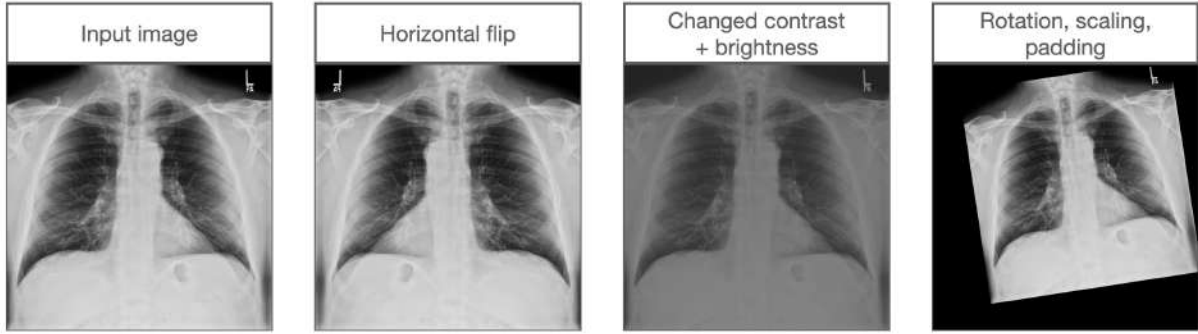


Figure 3.4: Examples of the data augmentation transformations used during training. The image stems from the CANDID-PTX dataset [Fen21].

$\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used.

For the loss function, BCE loss was used as well as focal loss. To prevent the model from overfitting, weight decay was used with a value of $1e-04$.

To increase the models’ robustness and generalizability, various forms of data augmentation were used as depicted in figure 3.4: Horizontal flipping, increasing and decreasing brightness and contrast, as well as rotating, scaling and padding. Using data augmentation seemed to be especially important due to the fact that the model would eventually be used on images coming from a distribution different from the one from the training set. Depending on the hospital that the images were retrieved from or the X-ray imaging machines that were used to take an X-ray image, the images can differ quite a lot throughout different datasets. This can be seen when looking at random samples from CANDID-PTX, PadChest as well as CheXpert, as depicted in figure 3.5.

Using data augmentation during training can help make the model generalize better and, therefore, potentially perform better on new datasets [SK19]. Different probabilities for an image to be augmented were tried, ranging from 0.0 to 1.0.

Also included in the hyperparameter search space were the target classes, that influence the loss. The target classes are the labels that are provided in the CANDID-PTX dataset: chest tube, pneumothorax and rib fracture. Models were trained including all three classes, only chest tube and pneumothorax, as well as only on chest tube.

Model were trained with early stopping after 8 epochs based on the AUROC score (sec. 2.5.5).

They were evaluated on the CANDID-PTX validation set and PadChest validation set separately.

Hyperparameter	Value
Model	ResNet18
Pre-trained on ImageNet	True
Resized to	(512, 512)
Target classes	chest tube
Learning rate	$5e-05$
ReduceOnPlateau	False
Loss	focal loss
Optimizer	Adam
Weight decay	$1e-04$
Data augmentation probability	0.75

Table 3.1: Hyperparameter values of final chest tube classification model.

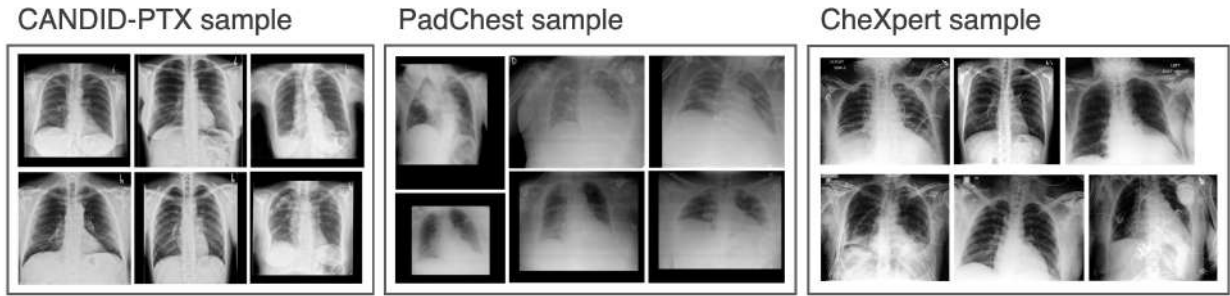


Figure 3.5: Comparison of randomly sampled chest X-ray images from CANDID-PTX, PadChest and CheXpert dataset. The images stem from the given datasets [Fen21], [Bus20], [Irv19].

The hyperparameters for the final chest tube classification model were chosen based on the highest AUROC score for the class chest tube on the PadChest validation set (sec. 3.2.1).

3.3 Chest Tube Prediction

As a next step, the final chest tube classification model described in section 3.2 was used to predict chest tube labels for the ChestX-ray14 dataset and CheXpert dataset. The predictions were then saved to CSV files. In the further course of this project, those predicted probabilities were used as training ground truth labels in the form of soft labels, ranging from 0.0 to 1.0 (instead of converting them into one-hot encoded vectors, so-called hard labels).

For the evaluation with AUROC, however, no continuous format is supported, meaning that the probabilities have to be transformed to one-hot encoded labels, wherein 0 corresponds to *no chest tube*, and 1 corresponds to *chest tube*.

To do so, the threshold was set to 0.173 based on the maximal youden index (sec. 2.5.7). Therefore, all probabilities higher than 0.173 were set to 1, and all probabilities less or equal to 0.173 were set to 0.

3.4 Pneumothorax Classification

This section deals with the pneumothorax classification models and the according training and evaluation process.

3.4.1 Data Splits

This section will give a brief overview of which datasets and data splits were used for training, validation and testing.

CANDID-PTX

The CANDID-PTX dataset was again used for pneumothorax classification. The splits used are equivalent to the ones described in section 3.2.1.

ChestX-ray14

The ChestX-ray14 dataset was used for training, validation and testing. As mentioned in section 2.6.3 a train-test-split has been provided along with the dataset. The given split seemed rather unusual compared to a commonly used 70-10-20 split where the label distribution is being kept fairly balanced between the sets. The given training set’s class distribution is quite imbalanced, with a positive-to-negative ratio of 1:223. However, since the goal for this project was primarily to see if a performance difference could be observed, instead of only focusing on improving the performance, the given split was used in the project. Additionally, the provided training set was split into a new training set, including 74,523 of the images (85%), and a validation set, including 12,001 images (15%). Again, it was assured that all images of each patient occur only in one set.

CheXpert

The CheXpert dataset was used for training, validation and testing. As mentioned in section 2.6.4 a validation and test have been provided. Since those are very small, 200 and 500 samples, other splits were used for this project, following a 70-10-20 split. This resulted in a training set consisting of 134,000 images, a validation set of 19,001 images and a test set of 38,229 images. Again, it was assured that all images of each patient occur only in one set.

LMU Chest X-Ray Pneumothorax - Chest Tube Dataset

The LMU dataset was used for testing. This dataset allows for detailed subgroup analysis based on all combinations of pneumothoraces and chest tubes, as describes in figure 2.4.

3.4.2 Model Training and Hyperparameter Tuning

The hyperparameter search space for the pneumothorax classification models is similar to the one described in section 3.2.2.

For pneumothorax classification, three different network architectures were part of the hyperparameter search space: ResNet18, ResNet50 and DenseNet121.

For DenseNet121, even smaller batch sizes had to be used then previously introduced; batch size 2 was used for size (1024, 1024), 8 for size (512, 512), and 16 for size (256, 256).

Due to storage constraints, models that were trained on the CheXpert dataset were only trained with images of size (256, 256) pixels, stemming from the version *CheXpert-v1.0-small*.

Additionally, to address the class imbalance in some of the training sets used, over-sampling was introduced. By oversampling the training set, images are randomly sampled with replacement with a given probability for their class [Bro20]. More precisely, all classes are being given a weight w_c . The probability that a randomly sampled image belongs to

a certain class can be computed as follows:

$$p = \frac{w_c}{\sum_{i=0}^n w_i}$$

For the class weights, the following weights were tried: (1.0, 5.0), (1.0, 10.0), (1.0, 20.0), (1.0, 40.0). The first weight corresponds to pneumothorax negative images and the second one to pneumothorax positive images.

Models were trained with target classes chest tube and pneumothorax (CT, PTX), as well as only pneumothorax (PTX).

On the CANDID-PTX training set, two models were trained: one with ground truth chest tube labels and one without chest tube labels. Additionally, on the ChestX-ray14 training set and the CheXpert training set two models were trained each; one with noisy chest tube labels, created as described in section 3.3, and one without chest tube labels.

A total of six final pneumothorax classification models were chosen based on their AUROC scores on their corresponding validation sets (CANDID-PTX, ChestX-ray14, CheXpert validation set). The final hyperparameter values are listed in table 3.2.

	Hyperparameter values for ...					
Hyperparameters	Model trained on CANDID-PTX without CT labels	Model trained on CANDID-PTX with CT labels	Model trained on ChestX-ray14 without CT labels	Model trained on ChestX-ray14 with CT labels	Model trained on CheXpert without CT labels	Model trained on CheXpert with CT labels
Model	ResNet18	ResNet18	DenseNet121	DenseNet121	DenseNet121	DenseNet121
Pre-trained on ImageNet	True	True	True	True	True	True
Resized to	(512, 512)	(512, 512)	(512, 512)	(512, 512)	(256, 256)	(256, 256)
Batch size	32	32	7	7	16	16
Target classes	PTX	CT, PTX	PTX	CT, PTX	PTX	CT, PTX
Learning rate	5e-05	5e-05	1e-04	1e-04	1e-04	1e-04
ReduceOnPlateau factor	None	None	0.1	0.1	0.5	0.5
Loss	BCE	BCE	BCE	BCE	BCE	BCE
Oversampling	None	None	based on PTX with weights: neg. cases: 1.0 pos. cases: 10.0	based on PTX with weights: neg. cases: 1.0 pos. cases: 20.0	based on PTX with weights: neg. cases: 1.0 pos. cases: 5.0	based on PTX with weights: neg. cases: 1.0 pos. cases: 5.0
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam
Weight decay	1e-04	1e-04	1e-04	1e-04	1e-04	1e-04
Data augmentation probability	0.75	0.75	0.25	0.75	0.5	0.25

Table 3.2: Hyperparameter values of the final pneumothorax classification models.

4 Results

4.1 Chest Tube Classification

The final chest tube classification model was evaluated on the CANDID-PTX test set as well as on the PadChest test set. On the CANDID-PTX test set it has an AUROC score of 0.984, with the ROC curve being depicted in orange (fig. 4.1). On the PadChest test set it has an AUROC score of 0.936, with the ROC curve being depicted in blue.

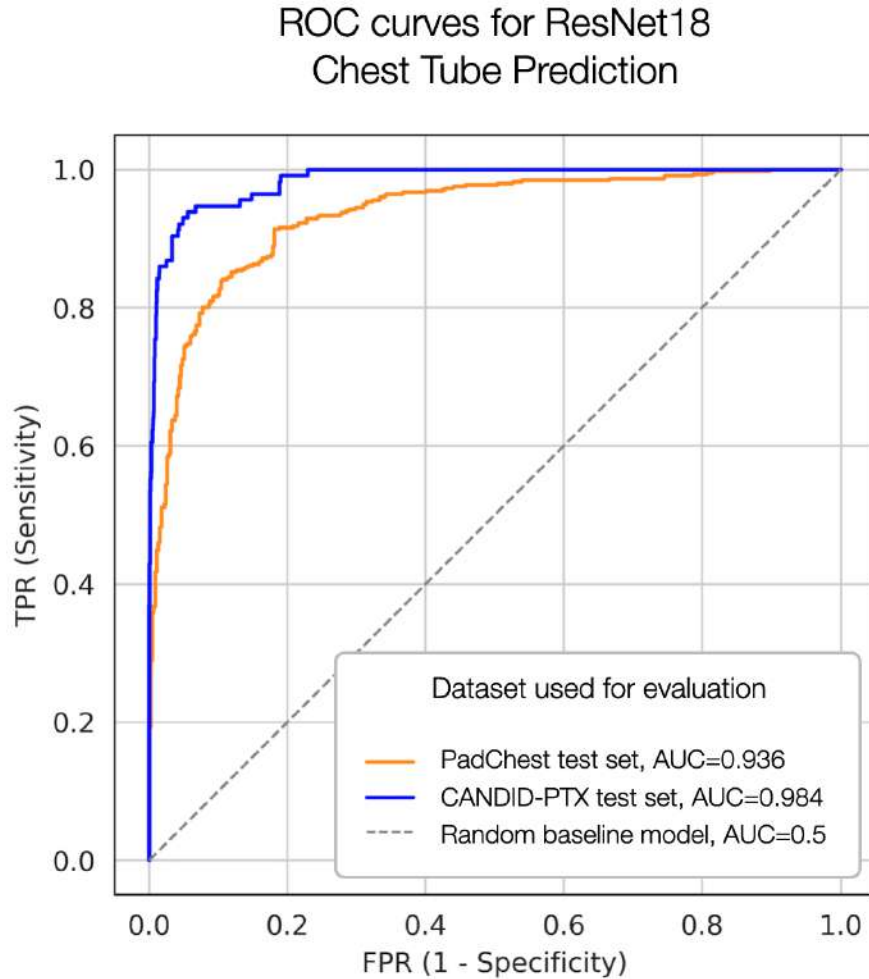


Figure 4.1: ROC curves for the final chest tube classification model, evaluated on the PadChest test set as well as the CANDID-PTX test set.

4.2 Pneumothorax Classification

For pneumothorax classification, all six final models, as described in table 3.2, were evaluated on cohorts of the LMU dataset (fig. 4.2) as well as on cohorts of the CANDID-PTX dataset (fig. 4.3). A discussion of the results follows in section 5.

As a quick recap, the cohorts were created by extracting different subsets of a given dataset by filtering for all possible combinations of PTX and CT cases, as described in figure 2.4. More precisely, the subsets were created by filtering all PTX positive images, as well as all PTX negative images, for either only with CT, only without CT, or with and without CTs. This resulted in the 9 cohorts depicted in figures 4.2 and 4.3.

With regards to the problem explored in this project, the subplots **A**, **E** and **I** are particularly interesting:

- **E** represents the evaluation on the entire dataset.
- In **A** we have the evaluation on a subset on which a model with a bias regarding CT would perform very poorly; all PTX positive images do not contain a tube, whereupon a biased model would incorrectly classify those as pneumothorax negative. All PTX negative images, on the other hand, do contain CTs, whereupon a biased model would mistakenly classify those as pneumothorax positive.
- For the subset used in **I**, it is the other way around; on this subset, a biased model would perform very well.

It may be noted that a non-biased model would perform equally well or poorly on all given subsets.

In both evaluation figures (fig. 4.2, fig. 4.3), the color of the ROC curves corresponds to the dataset that the model was trained on; orange corresponds to CANDID-PTX, blue to CheXpert and green to ChestX-ray14. The dotted lines correspond to models that were trained without chest tube labels, whereas the models represented by the solid line were trained with CT labels (one-hot encoded for CANDID-PTX, soft labels for ChestX-ray14 and CheXpert).

For ease of reference, in the further course, the pneumothorax classification models will be dubbed by a composition of the dataset that they were trained on and if they were trained with or without CT labels. To give an example, *ChestX-ray14 model with CT* represents the model that was trained on the ChestX-ray14 dataset with CT labels. Furthermore, the models trained on the same datasets are called *partner models* in the following, e.g. *ChestX-ray14 model with CT* is the partner model of *ChestX-ray14 model without CT* and the other way around.

4.2.1 Evaluation on LMU Dataset Cohorts

For the overall PTX performance on the entire LMU dataset (fig. 4.2, **E**), it can be seen that the AUROC score is higher for all models trained with chest tube, compared to their corresponding partner model trained without CT labels: for the ChestX-ray14 and CheXpert models the performance difference is 0.02, for the CANDID-PTX models it is 0.05. The best performing model is the CANDID-PTX model with CT with an AUROC score of 0.75. The worst performing model on this subset is the ChestX-ray14 model without CT, with an AUROC score of 0.65.

LMU dataset cohorts						
AUROC score	CANDID-PTX model		ChestX-ray14 model		CheXpert model	
	with CT	w/o CT	with CT	w/o CT	with CT	w/o CT
on subset I	0.83	0.72	0.69	0.67	0.81	0.8
on subset A	0.39	0.59	0.57	0.56	0.37	0.32
AUROC (I) - AUROC (A)	0.44	0.13	0.12	0.11	0.44	0.48

CANDID-PTX test set cohorts						
AUROC score	CANDID-PTX model		ChestX-ray14 model		CheXpert model	
	with CT	w/o CT	with CT	w/o CT	with CT	w/o CT
on subset I	0.97	0.92	0.74	0.72	0.94	0.94
on subset A	0.23	0.75	0.37	0.38	0.22	0.25
AUROC (I) - AUROC (A)	0.74	0.17	0.37	0.34	0.72	0.69

Table 4.1: AUROC scores and performance gaps: AUROC scores for all pneumothorax classification models on subset **I** as well as subset **I** on both dataset cohorts, LMU dataset and CANDID-PTX test set. Additionally, the performance gaps in AUROC scores between the subsets **I** and **A** are given. The abbreviation *w/o* stands for *without*.

On all other subsets, the CheXpert model with CT and the ChestX-ray14 model with CT perform slightly better than the corresponding partner model, by 0.01 to 0.05.

On subset **I**, the CheXpert models and ChestX-ray14 models perform much better than random. On subset **A**, on the other hand, the ChestX-ray14 models are only slightly than random, with AUROC scores of 0.57 (with CT) and 0.56 (without CT). The CheXpert models perform worse than random, with AUROC scores of 0.37 (with CT) and 0.32 (without CT).

For the CANDID-PTX models, it can be observed that on subset **A** the model without CT outperforms the one with CT, with a performance difference of 0.2; the model with CT has an AUROC score of 0.39, worse than random, whereas the model without CT has an AUROC score of 0.59, better than random. For subset **I**, however, it is the other way around; the model with CT outperforms the model without CT. The performance difference is 0.11. Also, for this subset, both models are better than random, with AUROC scores of 0.83 (with CT) and 0.72 (without CT).

As depicted in table 4.1, all models perform better on subset **I** than on subset **A**. Especially for the CANDID-PTX models, it can be observed that there is a big difference between the performance gaps in the AUROC scores of subsets **A** and **I**; the model with CT has a performance gap of 0.44 in AUROC scores with regards to subsets **I** and **A**, whereas the model trained without CT has a performance gap of only 0.12.

This trend is even stronger on the CANDID-PTX test set cohorts, with performance gaps of 0.74 and 0.17, respectively. For the ChestX-ray14 models and the CheXpert models, the performance gaps regarding subsets **A** and **I** are much more similar, as it can

be seen in table 4.1.

4.2.2 Evaluation on CANDID-PTX Test Set Cohorts

Similar trends can be seen for the evaluation on the CANDID-PTX test set cohorts in figure 4.3. On the entire dataset (**E**) for the CheXpert as well as the ChestX-ray14 models, again, the models with CT labels outperform their partner model by 0.02. Also, the CheXpert models' performance is approximately 0.10 higher compared to the ChestX-ray14 models. The CANDID-PTX models' AUROC scores are even higher by 0.05 to 0.1, with the model with CT having an AUROC score of 0.85 and the model without CT having an AUROC score of 0.9.

When looking at subset **I**, it can be observed that all models trained with CT perform equally well (CheXpert models) or better than their partner model. The best performing model here is the CANDID-PTX model with CT with an AUROC of 0.97, whereas its partner model has an AUROC score of 0.92. In-between those two models, we can observe both CheXpert models with an AUROC score of 0.94 for both models. The ChestX-ray14 models have a lower AUROC score of 0.74 (with CT) and 0.72 (without CT).

On subset **A**, on the other hand, we can see that all models perform worse than the random model, except the CANDID-PTX model without CT labels with an AUROC score of 0.75. The CANDID-PTX model with CT labels has an AUROC score of 0.23, resulting in a performance difference of 0.52. The worst performing model is the CheXpert model, with CT with an AUROC score of 0.22. Its partner model has an AUROC score of 0.25. The ChestX-ray14 models have AUROC scores that are slightly higher but still below random; 0.37 (with CT) and 0.38 (without CT).

ROC Curves for Pneumothorax Classification on the LMU Dataset Cohorts

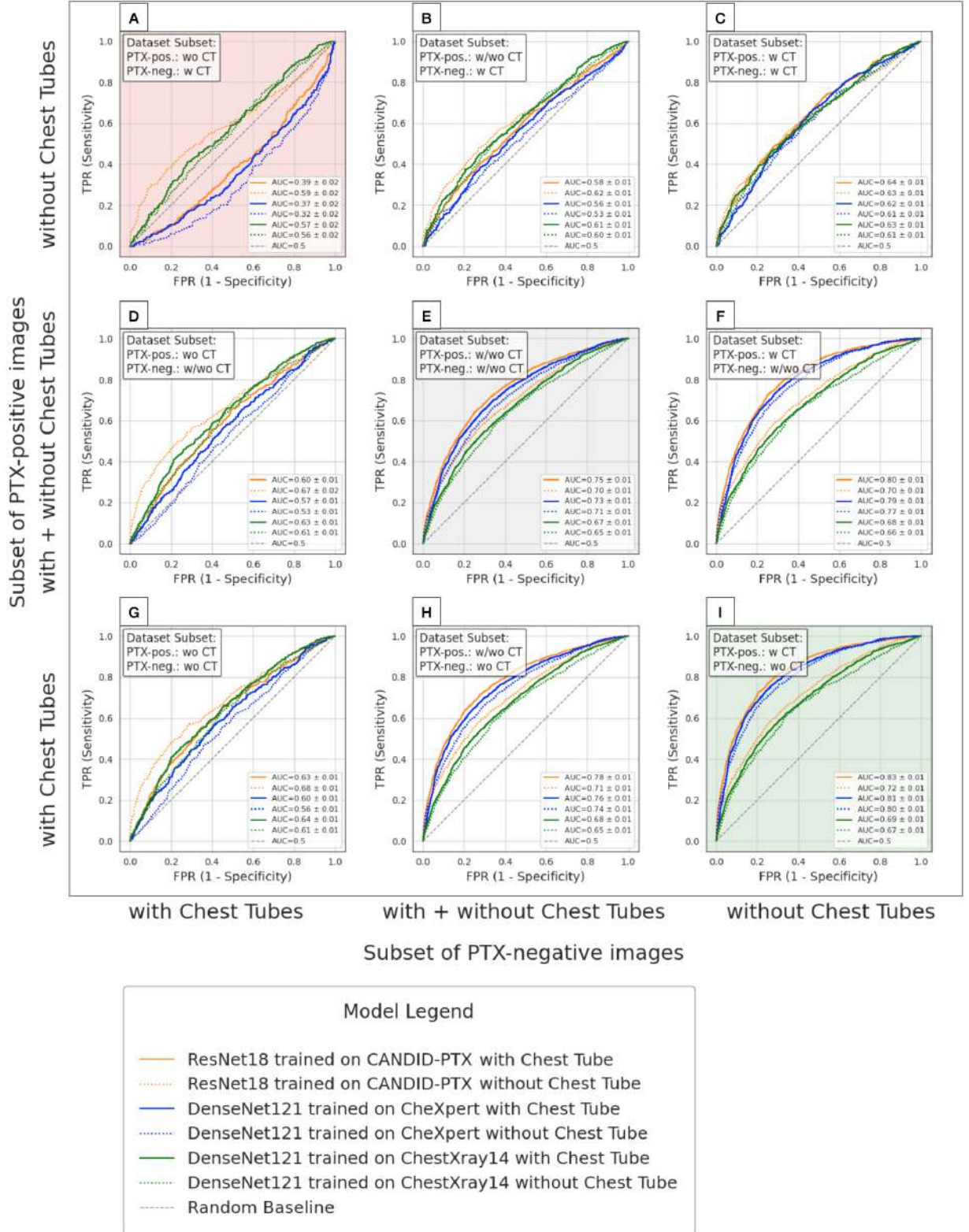


Figure 4.2: ROC curves for pneumothorax classification models on the cohorts of the LMU dataset. The gray plot (E) shows ROC curves on the entire dataset. Subset I is composed of all PTX-pos.-CT-pos. cases and PTX-neg.-CT-neg. cases. Subset A is composed of PTX-pos.-CT-neg. cases and PTX-neg.-CT-pos. cases.

ROC Curves for Pneumothorax Classification on the CANDID-PTX test set

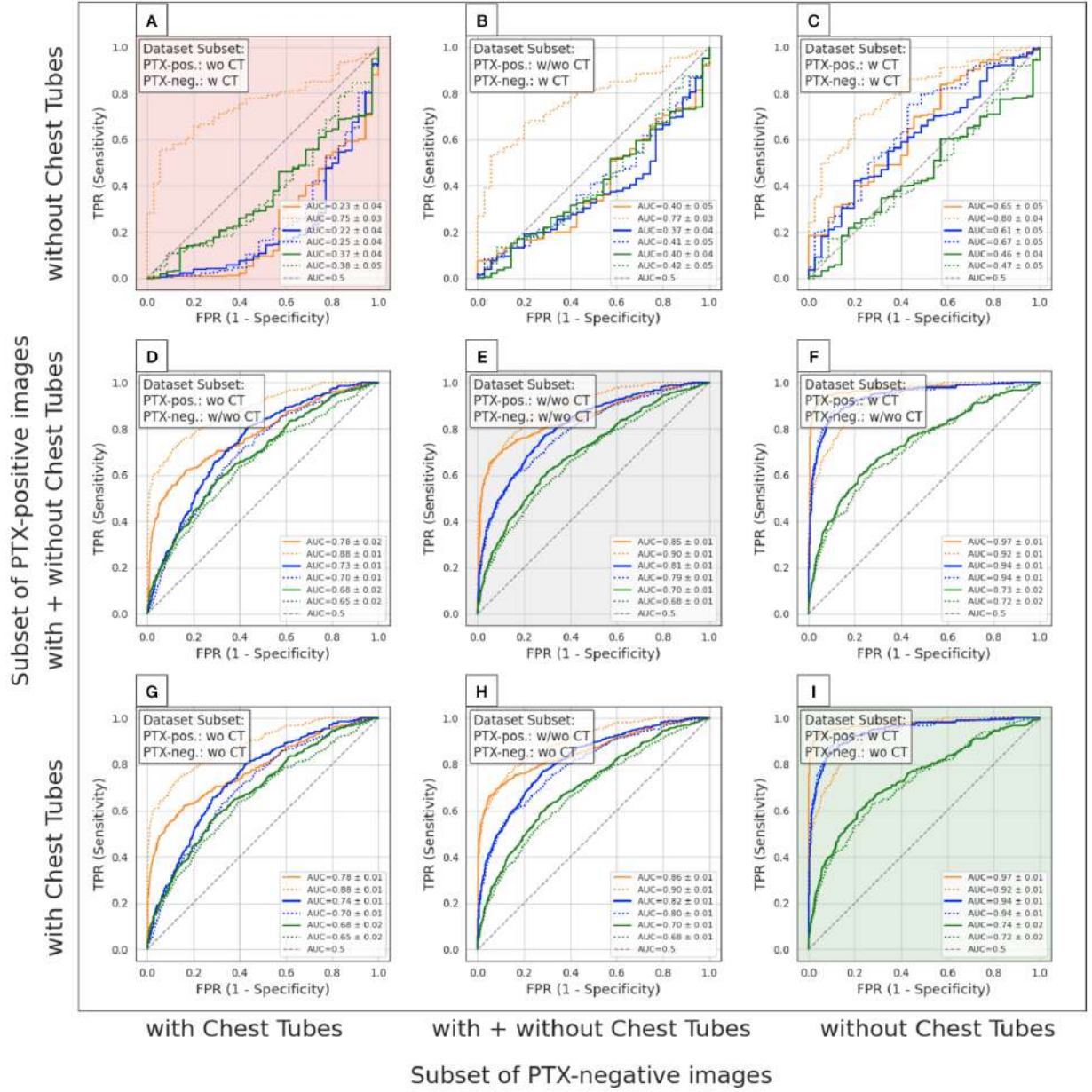


Figure 4.3: ROC curves for pneumothorax classification models on the cohorts of the CANDID-PTX test set. The gray plot (E) shows ROC curves on the entire dataset. Subset I is composed of all PTX-pos.-CT-pos. cases and PTX-neg.-CT-neg. cases. Subset A is composed of PTX-pos.-CT-neg. cases and PTX-neg.-CT-pos. cases.

5 Discussion

As stated in the hypothesis for this project, we hoped that by adding the task of CT classification to the PTX classification procedure, the model would learn to better recognize a pneumothorax by its visual features instead of focussing on CTs. In other words, we hoped the performance on subset **A**, composed of PTX-pos.-CT-neg. and PTX-neg.-CT-pos. cases, would improve.

ChestX-ray14 and CheXpert models

With regards to the results, it is noticeable that the trends for the ChestX-ray14 and CheXpert models are similar: For all models, a clear bias can be seen, i.e., better performance on subset **I** compared to the performance on subset **A**. Throughout all subsets of the LMU cohorts, the models trained with CTs perform a little bit better than their partner model that was trained without CTs.

Additionally, it is noticeable that on subset **A** of the CANDID-PTX test set, all four models perform worse than random, which, in combination with the better performance on subset **I**, indicates a clear bias.

Ultimately, it seems like training with or without CT labels did not make much of a difference for the models trained on ChestX-ray14 and CheXpert. The dataset on which they were trained appears to make a bigger difference, i.e., the CheXpert models have a stronger bias.

CANDID-PTX models

Regarding the CANDID-PTX models, on the other hand, it is especially striking that the performance gaps between the two models, trained with and without CT labels, are much bigger on most subsets compared to the performance gaps within the ChestX-ray14 models and CheXpert models.

Especially noticeable are the big performance gaps between the CANDID-PTX model with CT and the model without CT on subsets **A** and **I** on both test sets. It is striking that the model trained *with* CTs has an even stronger bias. Also, it stands out that the CANDID-PTX model without CTs performs better than random on subset **A** and the one with CTs worse than random. It appears that the best performing model on subset **A** was turned into one of the worst by adding the CT classification task, which presumably shifted its attention towards the chest tubes.

These observations suggest that the bias has not been reduced, and especially for the CANDID-PTX models has even been reinforced when the chest tube classification task was added.

Grad-CAM visualizations
for the pneumothorax classification models

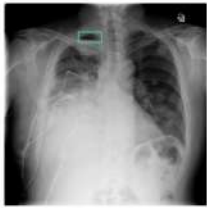
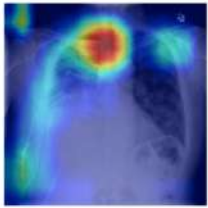
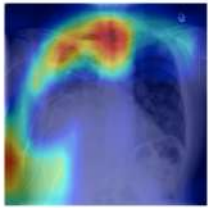
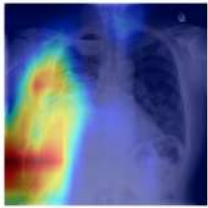

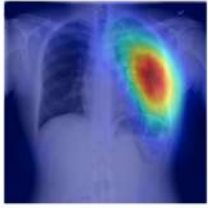
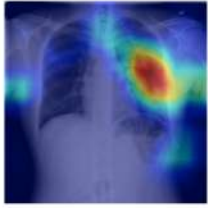
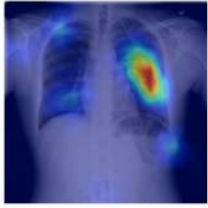

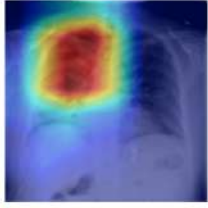
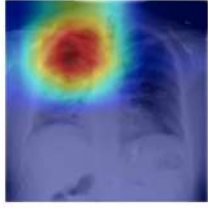
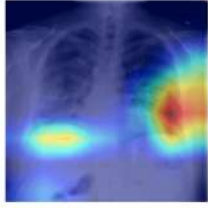
	Input image with PTX bounding box	Model trained with PTX PTX prediction	Model trained with CT, PTX PTX prediction CT prediction	
Models trained on CANDID-PTX:				
Models trained on ChestX-ray14:				
Models trained on CheXpert:				

Figure 5.1: Grad-CAM visualization samples for all PTX classifications models.

Hard vs. Soft CT Labels

It is noteworthy that the CANDID-PTX model with CTs was trained with hard CT labels (0.0 or 1.0) whereas the CheXpert and ChestX-ray14 models were trained with soft CT labels. This could explain why the CANDID-PTX model was so strongly affected by the additional prediction task compared to the other models.

Grad-CAM visualizations

These impressions are also supported by the Grad-CAM visualizations. Indeed, we can observe that the CANDID-PTX model trained *with* CTs tends to pay *more* attention to the CTs during PTX prediction than the corresponding counterpart trained without CTs.

In figure 5.1, an example for all models can be seen. More images with Grad-CAM overlays can be found in the appendix. The majority of examples support the interpretation that when the CANDID-PTX model is trained with CTs, its focus shifts even more towards the chest tubes.

For the ChestX-ray14 and CheXpert models, we cannot observe the same tendency, which is in line with the presumption that the soft labels limit the influence of the CT labels on the training loss.

6 Conclusion

To summarize, in the course of this project, we introduced chest tube labels to several chest X-ray datasets and added the task of chest tube classification to the pneumothorax classification task. By doing so, we hoped the models would learn that pneumothoraces and chest tubes are separate concepts, leading to improved performance on pneumothorax cases, and in particular the untreated ones without chest tubes.

However, the results of our experiments do not support this hypothesis. Instead, we observed negligible performance differences for some of the models (the ones trained on CheXpert and ChestX-ray14 with soft CT labels from the auxiliary model), and for the model trained on CANDID-PTX (with reliable hard labels for CT), we even observed a deterioration of the pneumothorax classification reliability on cases without chest tubes. In other words, the observations indicate that our intervention had the opposite effect of what it aimed to achieve. Instead of shifting focus away from the chest tubes towards the pneumothoraces, the focus on chest tubes seems to have been reinforced. These conclusions are supported by quantitative as well as qualitative observations.

In particular, the Grad-CAM visualizations for the CANDID-PTX model trained with chest tubes support the interpretation that the model does not learn to better focus on the visual features of a pneumothorax. Instead, it seems that the model takes advantage of the increased ability to recognize chest tubes by paying even more attention to chest tubes during PTX prediction, and concluding the presence of a pneumothorax from it.

Even though the pneumothorax AUROC on the entire test sets (subset **E**) has improved marginally for some of the models, the performance on cases that are particularly relevant from a clinical perspective has either been decreased or been unaffected by our intervention. Therefore, the small improvement in the overall PTX performance is unlikely to have a clinical value.

7 Outlook

Before state-of-the-art chest X-ray classification models can be useful with regards to pneumothorax classification within a clinical setting, it will be inevitable to first improve the models' performance for untreated pneumothorax cases. In this chapter we will briefly introduce and discuss further ideas for next steps to get closer to the goal of deployable pneumothorax classification models.

Removing All Chest Tube Positive Cases From Training Set

By removing all chest tube cases from the training set, regardless of their pneumothorax label, we can assure, that a model trained on such a dataset, will not be able to use the chest tubes as a short cut for pneumothorax prediction. There are two potential caveats worth mentioning to this approach. First, since a lot of pneumothorax cases do contain chest tubes, we would loose a lot of pneumothorax positive cases. Secondly, we do not know how a model would react to a never before seen chest tube. This, however, can be handled by leaving all chest tube positive cases in the validation and test set.

Adding Pneumothorax Segmentation to the Pneumothorax Classification Pipeline

While we only introduced chest tube labels and added the chest tube classification task to the pneumothorax classification, Graf et al. additionally extended the pneumothorax classification pipeline among others by adding a segmentation model that detects pneumothoraces based on segmentation annotations. This is one of the main differences between our approach and the one followed by Graf et al. [Gra20]. Most likely, this is an essential step to make the model actually focus on the pneumothorax itself, instead of a chest tube.

Therefore, replicating their approach while using the CANDID-PTX dataset which also provides pneumothorax segmentation annotations seems like a promising next step.

Bibliography

- [Aba15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: <https://www.tensorflow.org/>.
- [Bis06] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006, p. 209.
- [BMA19] Vanessa Buhrmester, David Münch, and Michael Arens. *Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey*. 2019. DOI: 10.48550/ARXIV.1911.12116. URL: <https://arxiv.org/abs/1911.12116>.
- [Bro20] Jason Brownlee. *Random Oversampling and Undersampling for Imbalanced Classification*. Jan. 15, 2020. URL: <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/> (visited on 09/13/2022).
- [Bus20] Aurelia Bustos et al. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”. In: *Medical Image Analysis* 66 (Dec. 2020), p. 101797. DOI: 10.1016/j.media.2020.101797. URL: <https://doi.org/10.1016%2Fj.media.2020.101797>.
- [Den09] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.
- [Faw04] Tom Fawcett. *Pattern Recognition Letters*. 31, Nr. 8. 2004. Chap. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers, pp. 1–38.
- [Faw06] Tom Fawcett. *Pattern Recognition Letters*, 27. 2006. Chap. An introduction to ROC analysis, 861–874.
- [Fen21] Sijing Feng et al. “Curation of the CANDID-PTX Dataset with Free-Text Reports”. In: *Radiology. Artificial intelligence*, 3(6), e210136 (2021). DOI: 10.1148/ryai.2021210136.. URL: <https://doi.org/10.1148/ryai.2021210136>.
- [FT19] William Falcon and The PyTorch Lightning team. *PyTorch Lightning*. Version 1.4. Mar. 2019. DOI: 10.5281/zenodo.3828935. URL: <https://github.com/PyTorchLightning/pytorch-lightning>.
- [GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. <http://www.deeplearningbook.org>. MIT Press, 2016, pp. 200–222.
- [Gra20] Benedikt Graf et al. *Pneumothorax and chest tube classification on chest x-rays for detection of missed pneumothorax*. 2020. DOI: 10.48550/ARXIV.2011.07353. URL: <https://arxiv.org/abs/2011.07353>.

- [He15] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: <https://arxiv.org/abs/1512.03385>.
- [Hop82] John Hopfield. “Neural Networks and Physical Systems with Emergent Collective Computational Abilities”. In: *Proceedings of the National Academy of Sciences of the United States of America* 79 (May 1982), pp. 2554–8. DOI: 10.1073/pnas.79.8.2554.
- [Hua16] Gao Huang et al. *Densely Connected Convolutional Networks*. 2016. DOI: 10.48550/ARXIV.1608.06993. URL: <https://arxiv.org/abs/1608.06993>.
- [IBM22] *IBM SPSS Statistics Bootstrapping*. URL: <https://www.ibm.com/docs/de/spss-statistics/saas?topic=bootstrapping-> (visited on 09/08/2022).
- [Irv19] Jeremy Irvin et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. 2019. DOI: 10.48550/ARXIV.1901.07031. URL: <https://arxiv.org/abs/1901.07031>.
- [Lia18] Richard Liaw et al. “Tune: A Research Platform for Distributed Model Selection and Training”. In: *arXiv preprint arXiv:1807.05118* (2018).
- [Lin17] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. 2017. DOI: 10.48550/ARXIV.1708.02002. URL: <https://arxiv.org/abs/1708.02002>.
- [LKF10] Yann Lecun, Koray Kavukcuoglu, and Clement Farabet. “Convolutional Networks and Applications in Vision”. In: May 2010, pp. 253–256. DOI: 10.1109/ISCAS.2010.5537907.
- [PAQ18] Jose G. Perez-Silva, Miguel Araujo-Voces, and Victor Quesada. “nVenn: generalized, quasi-proportional Venn and Euler diagrams”. In: *Bioinformatics* 34.13 (2018), 2322–2324. URL: <https://doi.org/10.1093/bioinformatics/bty109>.
- [Pas19] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems* 32. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [Raj17] Pranav Rajpurkar et al. *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*. 2017. DOI: 10.48550/ARXIV.1711.05225. URL: <https://arxiv.org/abs/1711.05225>.
- [Rue20] Johannes Rueckel et al. “Impact of Confounding Thoracic Tubes and Pleural Dehiscence Extent on Artificial Intelligence Pneumothorax Detection in Chest Radiographs”. In: *Investigative radiology*, 55(12), 792–798 (2020). DOI: 10.1097/RLI.0000000000000707. URL: <https://doi.org/10.1097/RLI.0000000000000707>.
- [SCP11] E. W. Steyerberg, B. Van Calster, and M. J. Pencina. *Performance measures for prediction models and markers: evaluation of predictions and classifications*. Revista Espanola de Cardiologia (English Edition), 2011, pp. 788–794.
- [Sel19] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. DOI: 10.1007/s11263-019-01228-7. URL: <https://doi.org/10.1007/s11263-019-01228-7>.

- [SK19] Connor Shorten and Taghi Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6 (July 2019). DOI: 10.1186/s40537-019-0197-0.
- [TH15] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool”. In: *BMC Medical Imaging* (2015). DOI: 10.1186/s12880-015-0068-x. URL: <https://bmcmmedimaging.biomedcentral.com/track/pdf/10.1186/s12880-015-0068-x.pdf>.
- [UIW19] Cornell University, INSEAD, and WIPO. *The Global Innovation Index 2019: Creating Healthy Livesâ€”The Future of Medical Innovation*, Ithaca, Fontainebleau, and Geneva. https://www.wipo.int/edocs/pubdocs/en/wipo_pub_gii_2019.pdf. 2019, pp. 129–130.
- [VD09] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.
- [Wan17] Xiaosong Wang et al. “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. DOI: 10.1109/cvpr.2017.369. URL: <https://doi.org/10.1109%2Fcvpr.2017.369>.
- [YalC] *Confidence Intervals*. URL: <http://www.stat.yale.edu/Courses/1997-98/101/confint.htm>.
- [Zho15] Bolei Zhou et al. *Learning Deep Features for Discriminative Localization*. 2015. DOI: 10.48550/ARXIV.1512.04150. URL: <https://arxiv.org/abs/1512.04150>.

List of Abbreviations

AI artificial intelligence

ANN artificial neural network

BCE binary cross entropy

CNN convolutional neural network

CT chest tube

FPR false positive rate

GAP *Global Average Pooling*

ICU intensive care unit

NLP natural language processing

LMU Ludwig-Maximilian-Universität

ML machine learning

PTX pneumothorax

ROC receiver operating characteristic

TPR *true positive rate*

TNR *true negative rate*

List of Tables

2.1	Dataset overview: Listed are the total number of images included in each dataset as well as the absolute number of images regarding pneumothorax (PTX) and chest tube (CT) positive images. If a label is not included in a dataset, the according cell is marked with a dash (-).	15
2.2	ChestX-ray14 label distribution: Number (#) of positive (pos.) images for each class.	16
2.3	CheXpert label distribution: Number (#) of positive (pos.) images for each class.	17
3.1	Hyperparameter values of final chest tube classification model.	22
3.2	Hyperparameter values of the final pneumothorax classification models. . .	26
4.1	AUROC scores and performance gaps: AUROC scores for all pneumothorax classification models on subset I as well as subset I on both dataset cohorts, LMU dataset and CANDID-PTX test set. Additionally, the performance gaps in AUROC scores between the subsets I and A are given. The abbreviation <i>w/o</i> stands for <i>without</i>	29

List of Figures

1.1	A: pneumothorax (PTX), B: chest tube (CT) in X-ray images. The images stem from the CANDID-PTX dataset [Fen21].	1
2.1	Fully Connected Neural Network.	5
2.2	Illustration of the cross-correlation.	5
2.3	Illustration of max pooling and average pooling.	6
2.4	Dataset's cohorts based on pneumothorax (PTX) and chest tube (CT) combinations.	8
2.5	ROC curve illustration including a perfect and random classifier.	11
2.6	Grad-CAM visualizations for ResNet18, pre-trained on ImageNet [Den09], for the class <i>tabby cat</i>	13
2.7	Label distribution in CANDID-PTX dataset, venn diagram created using nVennR [PAQ18].	15
2.8	Label distribution in the LMU dataset, venn diagram created using nVennR [PAQ18]	18
3.1	Grad-CAM used for error analysis: Using Grad-CAM during training assisted with identifying suspicious behavior: Models' focus on upper right corner, as well as the bottom line of some images. The image stems from the ChestX-ray14 dataset [Wan17].	19
3.2	Label distribution for CANDID-PTX training, validation and test sets. . .	20
3.3	Classes including tubes in PadChest: The PadChest dataset includes seven different tubes. PadChest's class <i>chest drain tube</i> is equivalent to the class <i>chest tube</i> in the CANDID-PTX dataset. The images stem from the PadChest dataset [Bus20].	21
3.4	Examples of the data augmentation transformations used during training. The image stems from the CANDID-PTX dataset [Fen21].	22
3.5	Comparison of randomly sampled chest X-ray images from CANDID-PTX, PadChest and CheXpert dataset. The images stem from the given datasets [Fen21], [Bus20], [Irv19].	23
4.1	ROC curves for the final chest tube classification model, evaluated on the PadChest test set as well as the CANDID-PTX test set.	27
4.2	ROC curves for pneumothorax classification models on the cohorts of the LMU dataset. The gray plot (E) shows ROC curves on the entire dataset. Subset I is composed of al PTX-pos.-CT-pos. cases and PTX-neg.-CT-neg. cases. Subset A is composed of PTX-pos.-CT-neg. cases and PTX-neg.-CT-pos. cases.	31

4.3	ROC curves for pneumothorax classification models on the cohorts of the CANDID-PTX test set. The gray plot (E) shows ROC curves on the entire dataset. Subset I is composed of al PTX-pos.-CT-pos. cases and PTX-neg.-CT-neg. cases. Subset A is composed of PTX-pos.-CT-neg. cases and PTX-neg.-CT-pos. cases.	32
5.1	Grad-CAM visualization samples for all PTX classifications models.	34
7.1	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CANDID-PTX dataset.	vi
7.2	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CANDID-PTX dataset.	vii
7.3	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the ChestX-ray14 dataset.	viii
7.4	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the ChestX-ray14 dataset.	ix
7.5	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CheXpert dataset.	x
7.6	Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CheXpert dataset.	xi

Appendix

Supplementary Data

Following, more Grad-CAM visualizations for predictions of the final pneumothorax classification models are provided. The input images stem from the ChestX-ray14 dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on CANDID-PTX


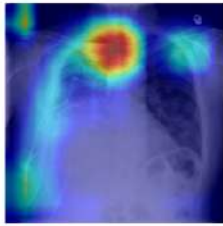
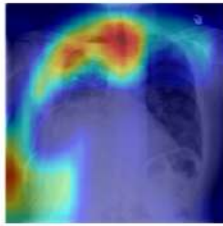
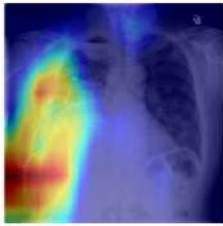

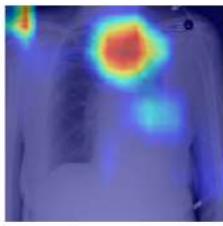
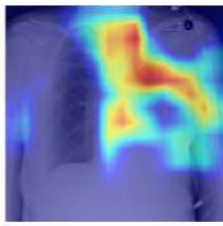
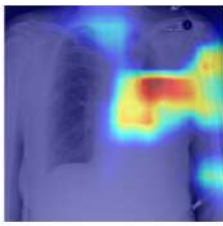

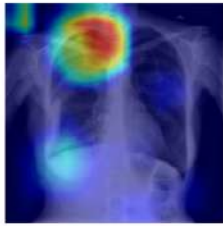
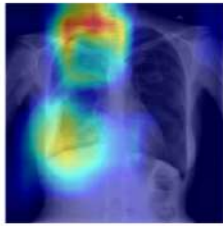
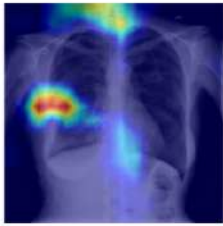

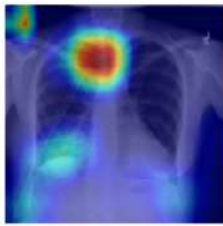
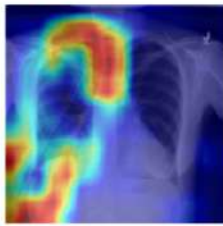
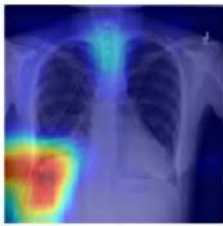

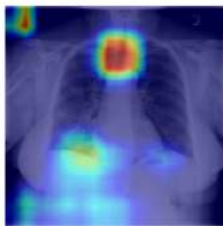
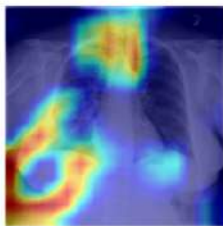
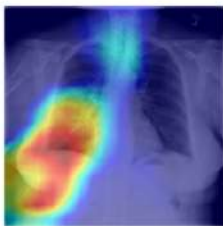
Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.1: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CANDID-PTX dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on CANDID-PTX


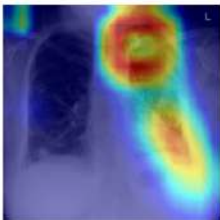
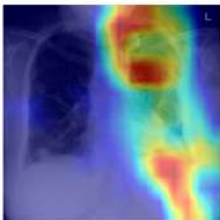
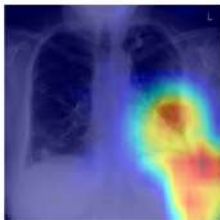

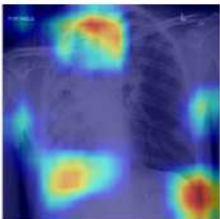
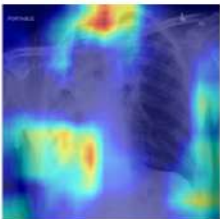
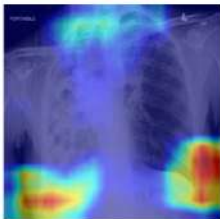
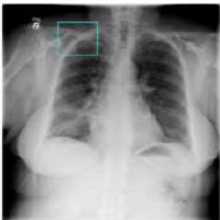
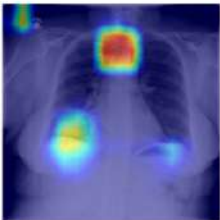
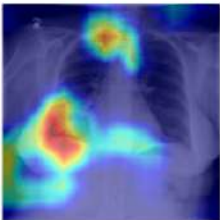
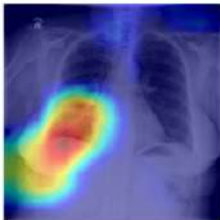
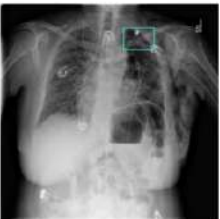
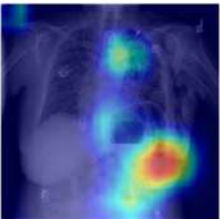
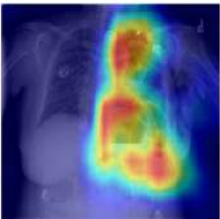
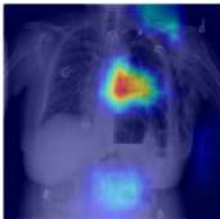

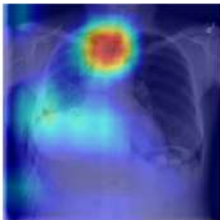
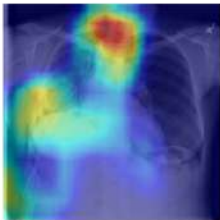
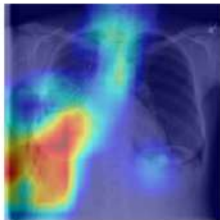
Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.2: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CANDID-PTX dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on ChestX-ray14


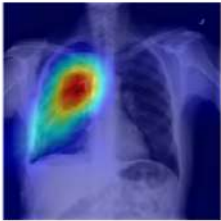
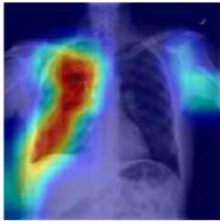
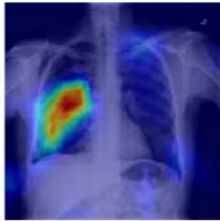

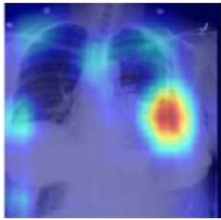
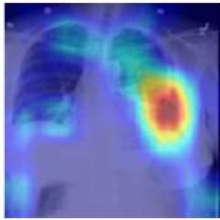
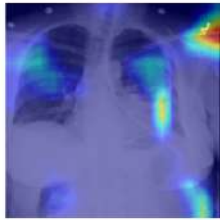

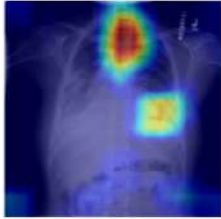
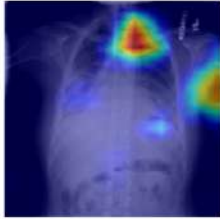
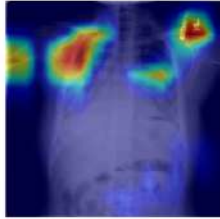

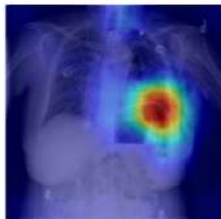
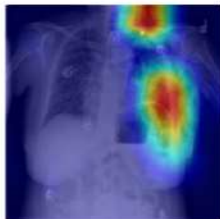
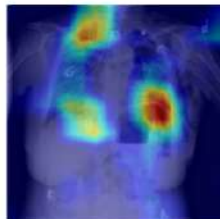

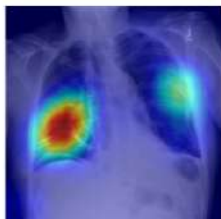
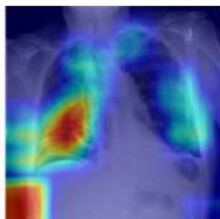
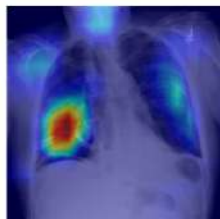
Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.3: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the ChestX-ray14 dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on ChestX-ray14


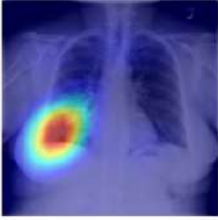
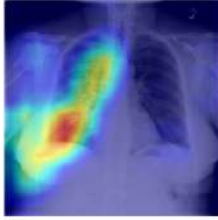
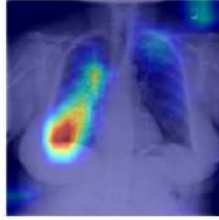
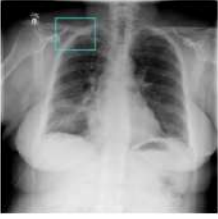
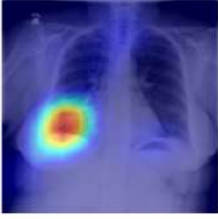
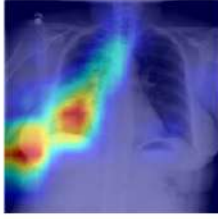
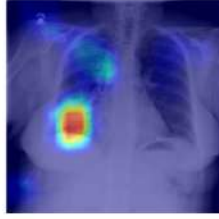



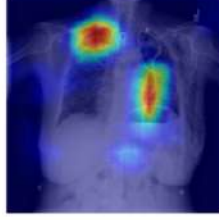

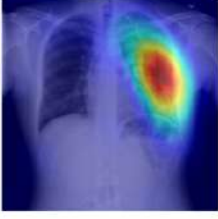
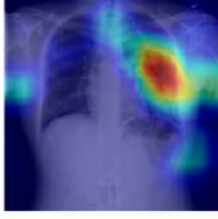
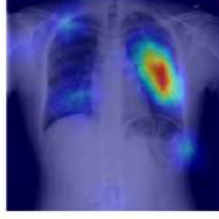

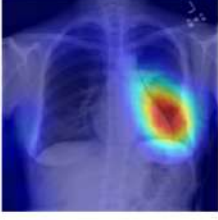
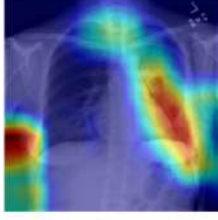

Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.4: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the ChestX-ray14 dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on CheXpert


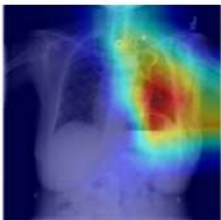
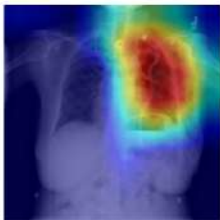
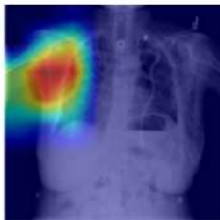
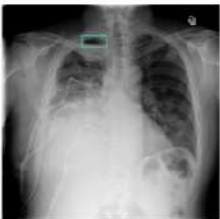
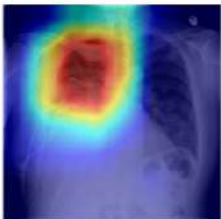
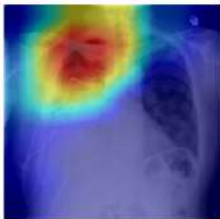
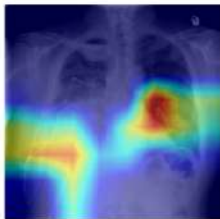
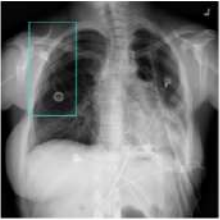
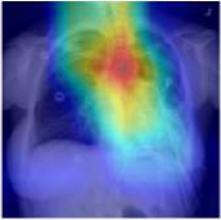
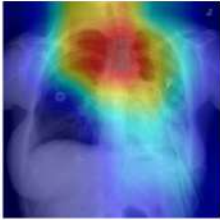
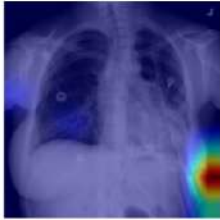

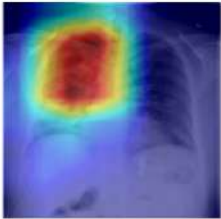
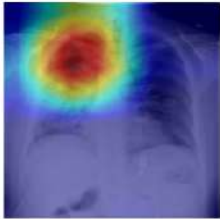
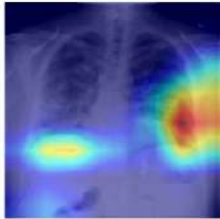

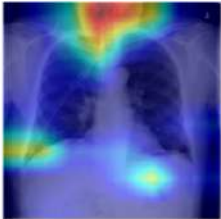

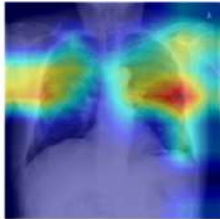
Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.5: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CheXpert dataset.

Grad-CAM visualizations
for the pneumothorax classification models
trained on CheXpert


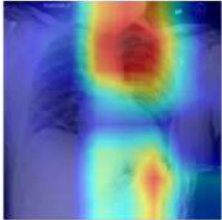
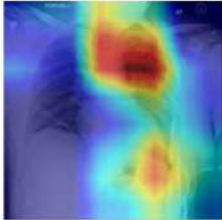
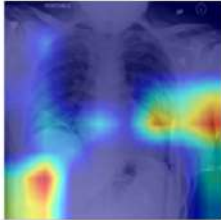

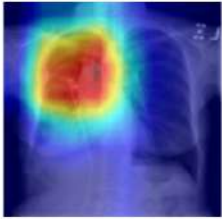
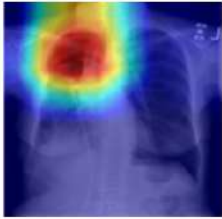
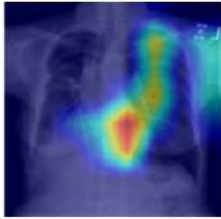

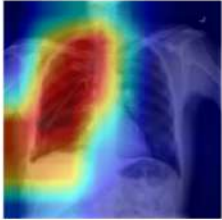
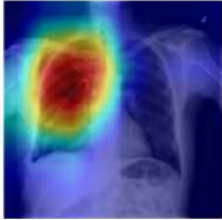
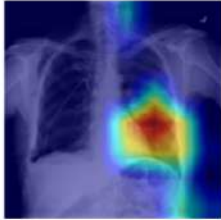

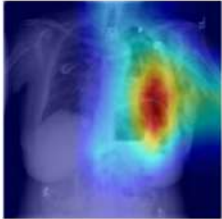
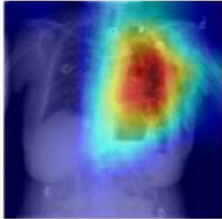
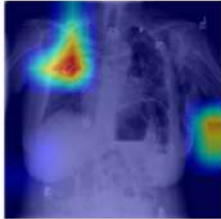

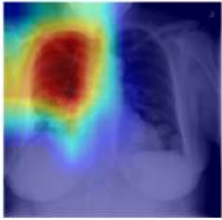
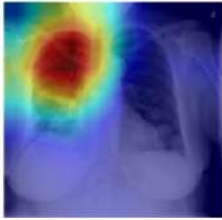
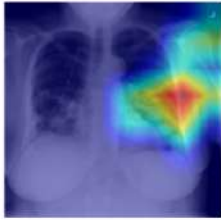
Input image with PTX bounding box	Model trained with PTX	Model trained with CT, PTX	
	PTX prediction	PTX prediction	CT prediction
			
			
			
			
			

Figure 7.6: Grad-CAM visualizations for the pneumothorax classification models that have been trained on the CheXpert dataset.