Master thesis

Cornelia Hofsäß

Generalizability of deep-learning-based pulmonary embolism detection from CT data

April 28, 2022

supervised by:
Prof. Dr.-Ing. Tobias Knopp
Dr. Jens-Peter M. Zemke

Dr. Hannes Nickisch
Dr. Tanja Lossau
Dr. Tobias Klinder

Hamburg University of Technology
Institute for Biomedical Imaging
Schwarzenbergstraße 95
21073 Hamburg

University Medical Center Hamburg-Eppendorf
Section for Biomedical Imaging
Martinistraße 52
20246 Hamburg

Ich versichere an Eides statt, die vorliegende Arbeit selbstständig und nur unter Benutzung der angegebenen Quellen und Hilfsmittel angefertigt zu haben.

Hamburg, den 28.02.2022

Contents

List of Abbreviations 9

1 Introduction **3**
1.1 Motivation 3
1.2 Research Questions 5
1.3 Thesis Outline 5

2 Theoretical Foundation **7**
2.1 Pulmonary Embolism 7
2.2 Convolutional Neural Networks 11
2.3 CT and DECT 18
2.4 Related Research 27

3 Methods and Materials **31**
3.1 Data set 31
3.2 Evaluation 42
3.3 Methodology Approach 46

4 Experiments and Interpretation **49**
4.1 Input 49
4.2 Training 62
4.3 Evaluation 73

5 Conclusion **87**

List of Tables

2.1 Wells Score 9
2.2 Related Research 29

3.1 Overview of Data Sets 37
3.2 Five-Number Summary 39
3.3 Overview of Peripheral and Proximal Embolisms 41

4.1 Hyperparameter Setting 50
4.2 Data Setup 54
4.3 Evaluation Across Data Sets 58
4.4 Combined Training 59
4.5 Results of Vessel Experiments 61
4.6 Hyperparamter Setups 63
4.7 Rendering Plots on two Monoenergetic Energy Levels 83
4.8 Comparison with the CADPE Challenge Winner 85

List of Figures

2.1 Pulmonary Embolism . . . 7
2.2 Clinical Process . . . 8
2.3 CTA Images with Pulmonary Embolisms . . . 9
2.4 Lung Anatomy . . . 10
2.5 Basic Concept of Supervised Learning . . . 11
2.6 U-Net Structure . . . 14
2.7 Visualization of Cross-Correlation . . . 15
2.8 Visualization of Maxpooling . . . 16
2.9 Conventional CT Setup . . . 18
2.10 CT Measurement . . . 19
2.11 X-Ray Spectrum . . . 21
2.12 Dual-Energy CT Setup . . . 22
2.13 Compton Scattering and Photo Effect . . . 23
2.14 Dual-Energy CT Representations . . . 25
2.15 DECT Material Decomposition . . . 26
2.16 Key Components of Classical PE Detection Approaches . . . 27

3.1 Development of the INHOUSE Data Set . . . 33
3.2 Missing and Corrected Annotation . . . 34
3.3 Inconsistent Labelling Methods . . . 36
3.4 Inconsistent Labels after Dilated Merging Step . . . 36
3.5 Comorbidity Cases . . . 37
3.6 Volume Box Plots and Violin Plots . . . 38
3.7 Smallest Embolisms within the Public Data Sets . . . 39
3.8 Separation of Peripheral and Proximal Embolisms . . . 40
3.9 Histograms of Peripheral and Proximal Embolisms . . . 41
3.10 Example of Different FROC Curves . . . 42
3.11 Example Sketch of Connected Components . . . 43
3.12 Example Scenarios of Ground Truth and Predicted Segmentations . . . 44
3.13 Methodology Approach . . . 47

4.1 Training Curve . . . 50
4.2 FROC Naive Experiment . . . 51
4.3 Lung Mask . . . 52
4.4 Negative Examples of Naive Training . . . 52
4.5 Training Curves Cross-Validation . . . 53
4.6 Performance of Different INHOUSE Setups . . . 54
4.7 Separate Training and Evaluation . . . 57
4.8 CT Scan with and witout Vessel Mask Overlay . . . 60

4.9 Merged Masks 60
4.10 Binary and Multiplied Vessel Mask 61
4.11 Training with Different Patch Sizes 63
4.12 Training with Different Learning Rates 64
4.13 Influence of Rotation Transformations 65
4.14 Influence of Scaling Transformations 66
4.15 Influence of Gaussian White Noise 67
4.16 Influence of Intensity Transformations 68
4.17 Classical Data Augmentation 69
4.18 DECT Data Augmentation 71
4.19 Performance on Healthy Cases 73
4.20 Evaluation on Proximal and Peripheral Embolisms 75
4.21 Histograms of Predictions and Ground Truth Volume 76
4.22 Renderings 77
4.23 Predictions outside the Region of Interest 78
4.24 Possibly Wrong False Positive Prediction 79
4.25 Special Cases of False Positive Predictions 79
4.26 False Negative Predictions 1 80
4.27 False Negative Predictions 2 81
4.28 Low Contrast True Positive Examples 82
4.29 Results of the CADPE Challenge 85

List of Abbreviations

AFP	average false positive
BDM	basic material decomposition
BraTS	brain tumor segmentation
CAD	computer aided detection
CADPE	computer aided detection for pulmonary embolism
CNN	convolutional neural network
CT	computed tomography
CTA	computed tomography angiography
DC	dice score
DECT	dual energy computed tomography
DL	deep learning
DVT	deep vein thrombus
EndoCV	endoscopy computer vision challenge
FN	false negative
FP	false positive
FROC	free-response-receiver-operating-characteristic
FUMPE	Ferdowsi University of Mashhad's PE
ISBI	IEEE international symposium on biomedical imaging
IVC	inferior vena cava
KNN	k-nearest-neighbor
LV	left ventricular
OPI	one-pixel intersection
PA	pulmonary artery
PAE	pulmonary artery embolism
PE	pulmonary embolism
pp	percentage points
Q-score	Qanadli-score
RSNA	Radiological Society of North America
ROC	receiver-operating-characteristic
ROI	region of interest
RV	right ventricular
SECT	single energy computed tomography
SPECT	spectral energy computed tomography
TP	true positive
TPR	true positive rate
UKK	Uniklinik Köln
VCS	vena cava superior
VNC	virtual non-contrast

Acknowledgements

First of all, I would like to acknowledge that I had the opportunity to write my Master's thesis at Philips Medical and thus gain an insight into medical and healthcare technology. There, I would like to express my gratitude to my main supervisors Dr. Tanja Lossau and Dr. Hannes Nickisch from Philips, who supported me during the whole master thesis and Dr. Tobias Klindler for his advice and support. My special thanks go to Dr. Roman Johannes Gertz and Prof. Dr. Alexander C. Bunck from the UKK (Uniklinik Köln) for their assistance with medical background information and for providing data. Last but not least, I wish to show my appreciation to Dr. Jens-Peter M. Zemke and Prof. Dr. Tobias Knopp from the TUHH (Technische Universität Hamburg), who supervised my work at the university.

1 Introduction

1.1 Motivation

Pulmonary embolisms (PEs), more precisely pulmonary artery embolisms (PAEs), are blood clots in the pulmonary arteries that obstruct them and thus prevent blood flow in the lungs. They are life-threatening and must be detected early. The mortality rate is approximately 30%, but with early detection, through proper treatment, this can be reduced to 2%, [1]. The state-of-the-art standard for diagnosing PEs are computed tomography angiography (CTA) scans. This involves injecting a contrast agent according to specific protocols to achieve a contrast in the pulmonary arteries. As a result, the blood clots become visible through contrast recesses within the arteries, [2, 3, 4, 5].

Since the 21st century, different approaches have been made to identify PEs automatically, so-called computer aided detection (CAD) systems for pulmonary embolism detection. First, the focus laid more on classical approaches, [6, 7, 8], where image processing was applied to extract certain features, which were then utilized for the classification of pulmonary embolisms, for example with conventional machine learning techniques, such as decision trees or k-nearest-neighbor (KNN), [8]. Due to the great success of deep learning (DL) methods in other areas [9], the trend has moved towards combining feature extraction and classification in deep end-to-end networks, [10, 11].

However, a major problem in evaluating the different approaches is that they are mostly assessed only on private data sets. This makes an objective comparison impossible. Because of this problem, annotated data sets such as the computer aided detection for pulmonary embolism (CADPE) data set, [3], and the Ferdowsi University of Mashhad's PE (FUMPE) data set, [2], have been made available to the public. This allows a proper comparison of the approaches. For example, Tajbaksh *et al.*, [12], achieved in their publication a sensitivity of 83% at an average false positive (AFP) rate of 2 on their own data set, but only a sensitivity 40% at 2 AFP on the CADPE challenge. Furthermore, performing well on a public data set also does not generally infer a good generalization in clinical applications. This requires a high sensitivity on a wide range of data and a low false positive rate, especially on healthy cases, which is a well-known challenge, due to the high amount of false positives (FPs), from which CADPE systems suffer, [10, 13]. Mueller-Peltzer *et al.* tested their CAD system for PE detection on data, collected over a three-year period in an emergency department. From a total of 3331 predicted emboli, only 258 (8%) were true positives (TPs) and 3073 (92%) were FPs,

demonstrating the limitations of CADPE systems in the clinical usecase and the complexity of this task, [13]. If the performance of CADPE systems can be improved, they could be used in practice as follows. The CADPE system has access to computed tomography (CT) data from various clinics and runs in the background. If a pulmonary embolism is suspected, an alarm is triggered and a team of experts is alerted. They receive the suspicious CT and the associated segmentations, which can be then examined in more detail.

The suspicion of PE often arises as an incidental finding on a CT scan in the context of another diagnosis, as the symptoms of pulmonary embolisms are often nonspecific. Because of the wide range of protocols in CT imaging, the image acquisition was likely not acquired with the correct contrast protocol, so the patient will need to be exposed to further radiation for a more accurate diagnosis. Computer automated tools are not suitable for low contrast images. This is due to the fact that CADPE systems are only trained with CTA scans according to PE protocols and thus do not generalize on different contrasts. Creating a large database with different contrast PE images turns out to be difficult, not only because of the lack of availability of medical data, but also because even for specialists the annotation in low contrast images is hardly manageable, [14]. Therefore, so-called dual energy computed tomography (DECT) data can be used. These are CT data taken simultaneously with low and high x-ray energy spectra. This allows different image representations, including different simulated contrasts. From one CT scan with associated label, multiple images with poorer contrasts can be generated, [15, 16, 17]. Including these in training could enable contrast-independent networks. Thus, the CADPE system would no longer be limited to CTA according PE protocols, but could be applied on all available CT images.

In this work, we use a deep learning approach based on the 3D U-Net structure, [18]. This architecture is state-of-the-art for segmentation tasks and outperforms the prior used convolutional neural networks (CNNs), for example by introducing this architecture, Ronneberger *et al.* won two cell tracking challenges at IEEE international symposium on biomedical imaging (ISBI) 2015, [18]. Since then, many adaptations from this architecture were generated with which various other challenges were won, such as the brain tumor segmentation (BraTS) challenge 2019, where substructures of brain tumours were segmented, [19], or the endoscopy computer vision challenge (EndoCV) 2021 challenge, where colon polyps were detected and segmented, [20]. In addition to detection, we also focus on segmentation, which makes the problem even more complex. In particular, we investigate the generalization ability of the networks, for which we use three different data sets, the public CADPE and FUMPE data set and a private data set, named INHOUSE. In contrast to the public data sets that have been specifically created for benchmarking CADPE algorithms, the INHOUSE data set results from collecting data for three year in a university clinic. The acquired images are not dedicated pulmonary embolism suspects, but are mostly diagnosed retrospectively as comorbidities, making this data set much more challenging. Various experiments for performance improvement are made based on changes in the network input and different training strategies. This is followed by a detailed evaluation for more precise understanding of the network behavior and the responsible factors for the generalizability. Since our INHOUSE data set is a DECT data

set, we are able to generate different contrasts and investigate and improve the generalization ability of the networks.

1.2 Research Questions

As already mentioned in the previous section, we use a network that is based on the 3D U-Net architecture to segment PEs. Here, we investigate *whether this network structure is suitable for segmentation for pulmonary embolisms* and compare the performance of our method with other approaches in the literature. One of the main investigations is the analysis of generalizability, examining how the trained networks perform on other data sets and *what factors play a role in performance in general and in generalizability*. This includes investigating *how conventionally trained networks perform on different contrasts and how this performance can be improved*.

By answering these questions, the basic problems of pulmonary embolism detection and generalization for clinical application will be highlighted. It should provide a basis for further work to improve generalizability, especially with DECT data. There, we put emphasis on the understanding of the network behavior and the influence of changes regarding the input, the training and the evaluation. To sum up, the following questions will be answered in this work:

- To what extent is the 3D U-Net structure suitable for PE segmentation?
- Which factors influence the generalizability of the network?
- How do conventional networks perform on different contrast images?
- How can contrast-independent predictions be realized?

1.3 Thesis Outline

This section gives a short overview of how the thesis is structured. In Chapter 2, theoretical background information is given, which is necessary to understand this work. First, a basic medical knowledge about pulmonary embolism is provided. Second, the general concept of CNNs and the used network structure are introduced. Third, the fundamentals of CT, especially DECT are explained. Finally, a short summary of related research is given. In Chapter 3 the used data sets are investigated in detail. This includes the analysis of the quality of annotations and presenting the pre-processing steps, which are necessary for their utilization for training and evaluation. In addition, it involves the precise investigation of the statistics of the data sets. This is important for further interpretations of the behavior of the networks that are trained with these data sets. Afterwards the metrics for performance evaluation are explained. At the end of this chapter, the methodology approach is introduced. The executed experiments are structured in experiments concerning the input of the network, the training strategy and the evaluation. This should help to understand how different changes in the individual components of the network influence the performance and impact the

generalizability. In Chapter 4 the performed experiments are described in the same structure with simultaneous interpretation of the results, starting with different experiments regarding the input data, such as the influence of the used training sets on the generalizability or the usage of additional information in form of lung or vessel masks on the overall performance. Then, we analyse the effects of different changes in the hyperparameters and trying different data augmentation strategies, first classical strategies and then an extension using DECT data. Finally, an intensive evaluation follows, in which the network behavior is closely examined. In Chapter 5, we summarize the main results of this work and discuss which further analyses can be carried out in future research projects.

2 Theoretical Foundation

In this section, the necessary theoretical foundations for understanding this work are laid. First of all, medical information about PEs in general is given in Section 2.1. Then, the basic concept of CNNs is briefly explained and the network structure we use is shown, see Section 2.2. Here, we assume that the reader already has some basic knowledge in the field of deep learning and neural networks. Section 2.3 contains general information about the generation of CT, in particular DECT data. Finally, we briefly review the current state of research, see Section 2.4.

2.1 Pulmonary Embolism

A PE is a blockage of an artery in the lung, mostly caused by a blood clot, a so-called thrombus, which obstructs the blood flow, shown in Figure 2.1. It usually occurs after vascular injuries in order to close the damaged area in the vessel from inside. It originates either directly in the lung arteries or it originates from different locations. In up to 80% of cases a PE is caused by a deep vein thrombus (DVT), which is a blood clot in the leg that travels through the blood stream from the deep veins of the leg or pelvis and via the inferior vena cava, the right atrium and ventricle of the heart into both pulmonary arteries, [2, 3, 21].

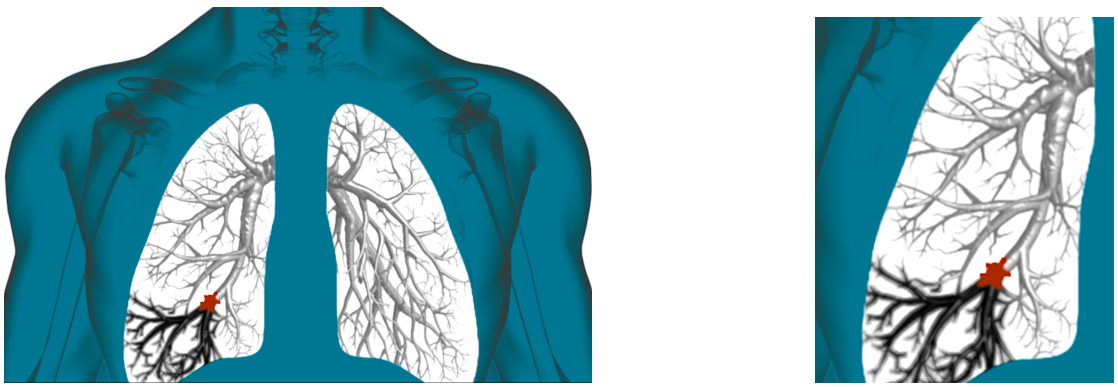


Figure 2.1: Example illustration of a PE. Blood clot (red) obstructs part of the arterial tree, which results in dead space within the lung (dark region).

The thrombus causes congestion in the lung areas and the affected region can no longer receive blood flow, resulting in a dead space. As a consequence, the lungs can no longer fulfill their natural function of oxygenating the blood and releasing CO_2, which leads to a lack of oxygen supply to the organs and thus to the failure of important functions. In addition, there is an increased pressure on the right heart, which can lead from shortness of breath to complete right heart failure, e.g., heart insufficiency, and even death, [4]. The lethality rate for pulmonary embolisms is about 30% but can be reduced to 2% with early detection, [10]. Each year, about 430 000 people in Europe are affected of PEs, in America there are about 300 000 - 600 000, 12 000 - 80 000 of which die from it, [3, 22]. Pulmonary embolism is the third leading cause of cardiovascular death, after chronic ischemic and myocardial infarction, [4].

Using Figure 2.2, we will now briefly present the clinical process, from symptoms, via diagnosis, up to therapy.

Typical symptoms of pulmonary embolism are dyspnea, meaning difficult breathing, or tachypnea, which stands for an increased respiratory rate, as a consequence of the increase in functional dead volume. As a result of concomitant pleurisy or pulmonary infarction, symptoms may include cough, hemoptysis (coughing up blood), and chest pain of a pleuritic nature (worsened by breathing). Tachycardia (increase in heart rate above $100\,\mathrm{min}^{-1}$) may occur due to the right heart strain, [21, 23, 24]. However, as the symptoms are often very nonspecific, pulmonary embolism is one of the most common unexpected findings of an autopsy, [25].

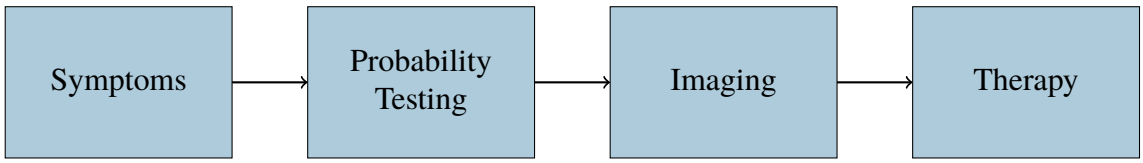


Figure 2.2: Clinical course, starting with various symptoms of the patient, progressing to exclusion procedures of PEs by probability tests, to imaging procedures, and if PE is successfully diagnosed, to various therapy interventions.

After the symptoms have given a first suspicion of a pulmonary embolism, so-called probability tests are first used to exclude a pulmonary embolism. One probability test is the Wells Score, developed by Philip Steven Wells in 1995, which uses clinical criteria to determine the likelihood of a pulmonary embolism, see Table 2.1, [21, 24, 26]. Another method to rule out a PE are D-dimer tests. D-dimers are proteins formed during the degradation of fibrin, a protein in the blood. During the body's own process of degrading a blood clot, the fibrin is split and the resulting D-dimers can then be detected by the test, which is an indication of a blood clot. These tests usually have a high sensitivity but a low specificity, which means that if the patient has a pulmonary embolism, they are likely to be positive. However, this does not mean that if they have a positive test result, the patient has a pulmonary embolism, [21, 23, 24, 26].

symptoms	**points**
clinically suspected DVT	3
alternative diagnosis is less likely than PE	3
tachycardia (heart rate > $\frac{100}{\text{min}}$)	1.5
surgery or immobilization (min. 3 days) within the last month	1.5
history of DVT or PE	1.5
hemoptysis (coughing up blood)	1
malignancy (under therapy or palliative therapy within last 6 months)	1

Table 2.1: The Wells Score predicts the likelihood of PE by considering different symptoms and summing up the corresponding points. If the score > 4, then the risk of suffering from PE is likely, if score ≤ 4, it is unlikely, [26].

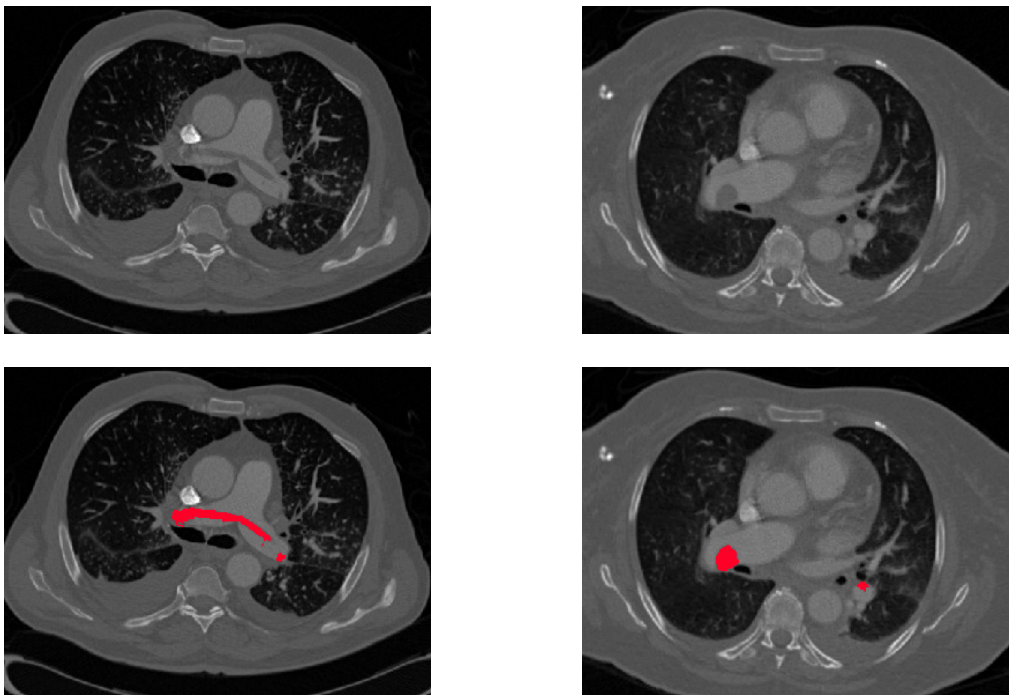


Figure 2.3: Examples of CTA images with a riding thrombus (left) and multiple contrast cavities (right) within the pulmonary arteries. The embolisms are marked in red.

As the first choice of imaging methods for the diagnosis of pulmonary embolisms, a CTA is used. Here, an iodine-containing contrast agent is injected according to specific protocols that specify the dose and time period between injection and image acquisition. Due to the strong absorption property of iodine, the arteries become visible. Pulmonary embolisms can

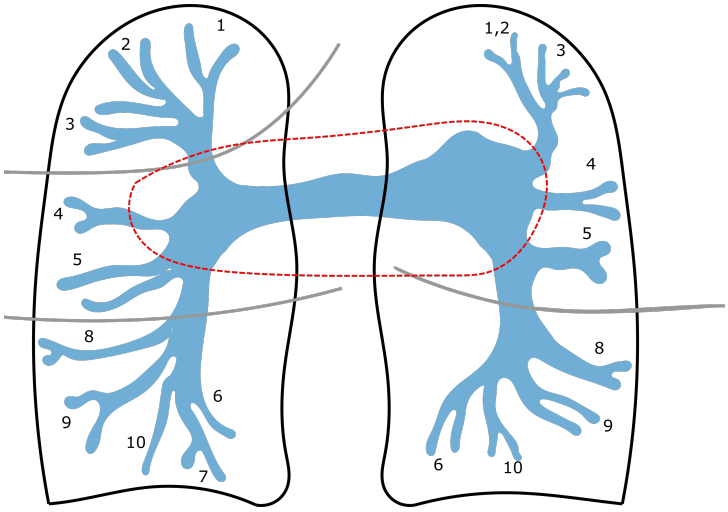


Figure 2.4: Schematic representation of the lung segments and pulmonary arteries, inspired from [2]. The right lung has three lobes and ten segments, while the left lung has only two lobes and eight segments. Segment 1 and segment 2 are fused in the left lung and segment 7 (which is omitted in the literature) is merged with segment 8, [28]. The lobes are separated by gray lines and the segments are numbered.

be observed on a CTA as a filling defect surrounded by a contrast rim in a pulmonary artery, see Figure 2.3, [4, 23, 26].

Figure 2.4 shows a sketch of the general structure of the lung and the arterial tree. The inspection process is straight forward. Starting at the pulmonary trunk, encircled in red, and walking along the arteries, tracing each arterial branch by examining contrast variations. The average radiologist has a sensitivity between 77% and 94%, [27], depending on the experience, the concentration and especially on the time that is taken to analyse a scan, making it difficult to measure. Another imaging method is a ventilation-perfusion scan, also referred to as a lung scintigraphy, in which the distribution of an inhaled radioactive gas is measured. The thrombus can be identified by the observation that some areas are ventilated but not perfused. This method is used, for example, in case of contrast material allergies or pregnancy due to its lower radiation exposure, [4, 21, 23].

After the patient is diagnosed with pulmonary embolism, various treatment methods can be applied, depending on the progression and severity of the embolism. Mostly treatment of pulmonary embolisms relies on anticoagulant medications, such as lysetherapy, which usually involves intravenous injection of heparin to dissolve the thrombi. Another option is the use of so-called inferior vena cava (IVC) filters. These are inserted into the inferior vena cava and prevent thrombi from reaching the lungs travelling from the lower extremities. In severe pulmonary embolisms, surgical removal by embolectomy is also considered. The clot is removed directly from the opened pulmonary arteries. To prevent further embolisms from occurring, prophylactic measures are usually performed, such as administration of low-dose heparin and also compression of the lower extremities, [23, 24].

2.2 Convolutional Neural Networks

In this section, a short introduction to deep learning, in particular CNNs, is given. We explain the basic concepts of training a neural network and introduce the U-Net structure on which our later implemented networks are based.

In this work we focus on neural networks, especially CNNs from the area of supervised learning, which means that for each input data, the correct label is available. Thereby, we consider a segmentation task, where the input data are 3D CTA images and the corresponding labels are the respective segmentation masks.

Our input data set can be described as the set $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1,\dots,N}$, where N is the amount of data, $\mathbf{X}_i \in \mathbb{R}^{h_i \times w_i \times d_i}$ is the CT input image, measured in the so-called Hounsfield scale (HU), and $\mathbf{Y}_i \in \mathbb{N}_0^{h_i \times w_i \times d_i}$ is the corresponding label mask, with height h_i, width w_i and depth d_i.

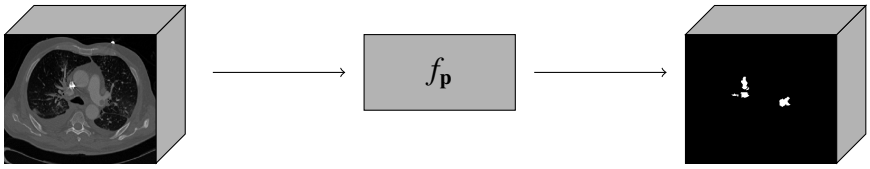


Figure 2.5: The basic concept is to determine the function $f_{\mathbf{p}}$, predicting for the 3D input CTA $\mathbf{X}$ the segmentation mask $f_{\mathbf{p}}(\mathbf{X})$, which is compared to the ground truth segmentation mask $\mathbf{Y}$.

The Hounsfield scale is used to describe the recorded attenuation of X-rays passing tissue during a CT measurement. It is a relative quantitative measure, which is determined by a linear transformation $T_{\text{HU}} : \mathbb{R} \to \mathbb{R}$ of the attenuation coefficient $\mu \in \mathbb{R}$, defined as

$$T_{\text{HU}}(\mu) = 1000 \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}. \tag{2.1}$$

The image with the calculated HU values is then usually displayed as a grayscale image. Water, with an absorption coefficient of $\mu_{\text{water}} = 0.192 \, \frac{1}{\text{cm}}$ always has a CT value of 0 HU and air with $\mu_{\text{air}} \approx 0 \, \frac{1}{\text{cm}}$ corresponds to -1000 HU, regardless of the X-ray spectrum. For all other materials, the HU value depends on the used X-ray spectrum, see Section 2.3. A change of 1 HU corresponds to a change in X-ray attenuation of 0.1% relative to water. In practice, the HU values of different materials are mostly covered by the interval $[-1000, 3095]$, whose 4096 different integer values can be expressed by 12 bits, $2^{12} = 4069$, [29, 30, 31].

The label mask $\mathbf{Y}_i$ of image $\mathbf{X}_i$ consists of $N_{cc,i} \in \mathbb{N}$ different connected components, which can be interpreted as different embolisms. We define the set of all different connected components as

$$\underline{N}_{cc,i} \coloneqq \left\{1, \ldots, N_{cc,i}\right\}. \tag{2.2}$$

For simplification, we omit the index i in the following and consider only one tuple $(\mathbf{X}, \mathbf{Y})$ of our data set. Each entry $y_{\mathbf{i}}$ of the matrix $\mathbf{Y}$, where $\mathbf{i}$ is the multi index, defined as $\mathbf{i} = (i, j, k)$, can take the following values

$$y_{\mathbf{i}} = \begin{cases} c_r, & \text{if } y_{\mathbf{i}} \text{ belongs to the connected component } c_r \in \underline{N}_{cc}, \\ 0, & \text{if } y_{\mathbf{i}} \text{ is background.} \end{cases} \tag{2.3}$$

In contrast to detection problems, we can have several embolisms within one instance. The objective is to detect and to segment them as accurate as possible. In Section 3.2 an overview of different evaluation methods is given.

The objective is to generate a function $f_{\mathbf{p}} : \mathbb{R}^{h \times w \times d} \to \mathbb{R}^{h \times w \times d}$, dependent on the parameters $\mathbf{p} \in \mathbb{R}^{\ell}$ that approximates for each normalized input value $\mathbf{X}$ the exact label mask $\mathbf{Y}$

$$f_{\mathbf{p}}(\mathbf{X}) \approx \mathbf{Y}. \tag{2.4}$$

According to the universal approximation theorem, for each function there exists a neural network that can approximate it arbitrarily well. There are different mathematical proofs about the needed restrictions of the network, as the number of layers, the number of neurons or the properties of the activation function, to make this universality possible. For example Cybenko shows in [32] that with a hidden layer any continuous function on a compactum can be approximated arbitrarily well. But although we know that such a network exists that does not mean that we are able to construct or even recognize such a network, [33].

To evaluate the quality of the network $f_{\mathbf{p}}$, we need a convex cost function C, which should be zero if the prediction approximates the ground truth values for all data.

The objective of machine learning is to find the optimal parameters $\mathbf{p} \in \mathbb{R}^{\ell}$ of the function $f_{\mathbf{p}}$, such that the costs $C : \mathbb{R}^{\ell} \to \mathbb{R}$ will be minimized,

$$\min_{\mathbf{p} \in \mathbb{R}^{\ell}} C(\mathbf{p}). \tag{2.5}$$

Therefore, we use the gradient descent algorithm. In each step we want to adapt the parameter vector by $\Delta\mathbf{p}$, such that the cost function decreases. To do so, we assume that $C \in C^1$ and approximate the costs with a first-order Taylor polynomial around $\mathbf{p}$

$$C(\mathbf{p} + \Delta\mathbf{p}) \approx C(\mathbf{p}) + \nabla C(\mathbf{p})^{\mathrm{T}} \Delta\mathbf{p}, \tag{2.6}$$

where $\nabla C(\mathbf{p})$ denotes the gradient of the cost function, defined as

$$\nabla C(\mathbf{p}) \coloneqq \left(\frac{\partial C(\mathbf{p})}{\partial p_1}, \ldots, \frac{\partial C(\mathbf{p})}{\partial p_l} \right)^{\mathrm{T}}. \tag{2.7}$$

To minimize the costs, we want to choose $\Delta\mathbf{p}$ such that $\nabla C(\mathbf{p})^{\mathrm{T}}\Delta\mathbf{p}$ become as negative as possible. According to the Cauchy-Schwarz inequality for any vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, we have

$$-\|\mathbf{u}\|_2\|\mathbf{v}\|_2 \leq \mathbf{u}^{\mathrm{T}}\mathbf{v} \leq \|\mathbf{u}\|_2\|\mathbf{v}\|_2, \tag{2.8}$$

where we can see that $\mathbf{u}$ has to be parallel to $\mathbf{v}$ to become as negative as possible. Thus we choose

$$\Delta\mathbf{p} = -\eta\nabla C(\mathbf{p}), \tag{2.9}$$

with suitable small learning rate $\eta > 0$. In each step of the gradient descent method, the costs are reduced by $\Delta C \approx -\eta \cdot \|\nabla C\|_2$, [34]. The choice of the learning rate plays an important role, if the learning rate is too small, the convergence speed will be too low, if the learning rate is too large, the linearisation from Equation (2.6) will not hold, and the costs might diverge, [33].

We start the gradient descent method with a randomly initialized parameter vector $\mathbf{p}$ and iterate

$$\mathbf{p} \leftarrow \mathbf{p} - \eta\nabla C(\mathbf{p}), \tag{2.10}$$

until we meet a stopping criterion, [34]. There exist many different initialization strategies, that we not discuss in the scope of this work. For more information see [35, 36].

After we have seen that the goal of training is to minimize the cost function and thus the predicted values approach the exact labels, we now want to look at how our function $f_{\mathbf{p}}$, i.e., our neural network, is constructed. The network structure is shown in Figure 2.6. Based on this, the main operations of a CNN are briefly explained below.

In this example we consider only one input image, i.e., the channel size is $c = 1$ and the dimension of our input is $1 \times h \times w \times d$. Normally, in a gradient descent update step, a mini-batch of size mb, i.e., several images are passed through the network at the same time.

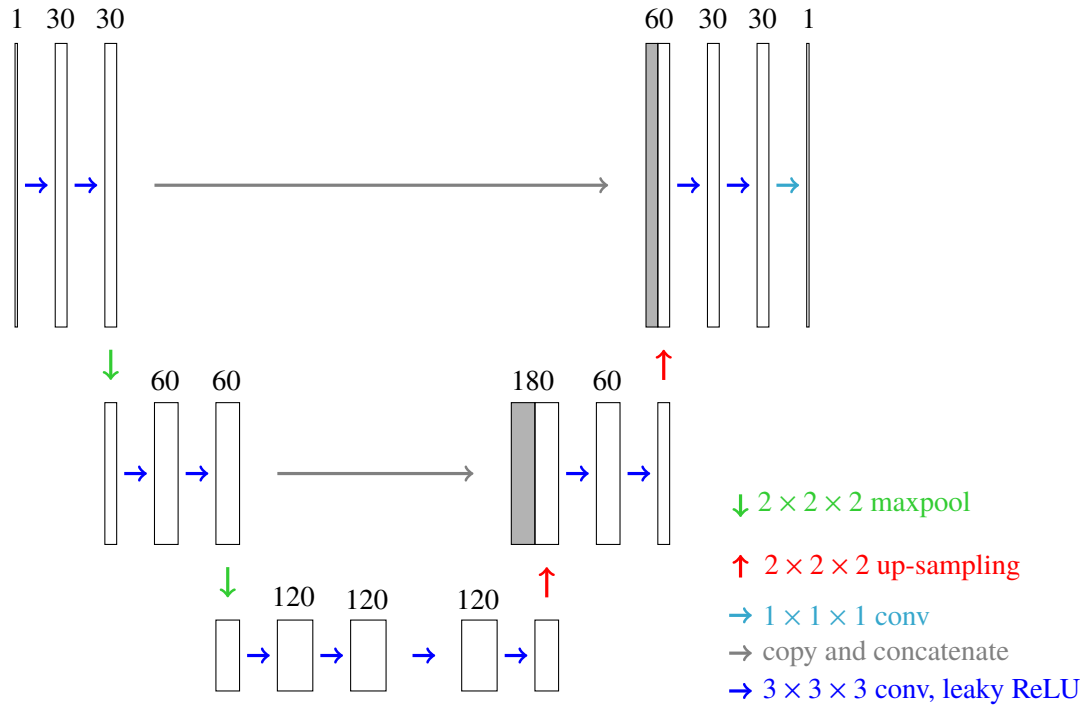


Figure 2.6: The U-Net architecture consists of an encoder, which reduces the dimension of the input space and captures the context of the image, and a decoder, which transforms the dimensionally reduced data back to the original input size. The encoder contains three blocks of two convolutional layers, followed by a maxpooling layer, except in the last block, where no maxpooling layer exists. Above each output the number of channels is stated, which is identical to the number of filters used. In each convolution block the number of filters are doubled, starting with 30 filter in the first block. The decoder contains also blocks with two convolutional layers, but followed by an up-sampling layer instead of maxpooling. In the last layer a $1 \times 1 \times 1$ convolution is used with sigmoid activation. The sigmoid activation function returns the segmentation mask representing the pixel-wise classification. The output of each encoder block has a shortcut connection to the input of the corresponding decoder block.

Convolutions: The main operations which are performed in a CNN are convolutions. These are realized in PyTorch and TensorFlow through cross-correlations which can easily be visualized by multiplying each element of the kernel with the receptive field of the image, displayed in Figure 2.7. The output of the cross-correlation operation is called a feature map. In a CNN in each convolutional layer many different filters are applied, which are excited by different properties of the image. In our implementation, see Figure 2.6, we use filters of size $3 \times 3 \times 3$ and we pad the image with zeros such that the output images always have the same size as the input images. We start by using 30 filters in the first convolutional block and double it for each layer of the encoder. In [37], Zeiler *et al.* visualize the filters of the different layers and by which part of the image they were stimulated. They found out that filters in the first layers detect more coarse features, such as edges or colors. Some filters resemble the Gabor filters known from image processing. The deeper one goes into the network, the more complex are the features detected by the filters. Additionally, for each feature map a bias is added. Note that the weights of the filters and the bias are the trainable parameters $\mathbf{p}$. For more information about the convolution operation in neural networks, see [38].

$$\begin{array}{|ccc|} \hline a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \\ \hline \end{array} \star \begin{array}{|cc|} \hline f_{1,1} & f_{1,2} \\ f_{2,1} & f_{2,2} \\ \hline \end{array} = \begin{array}{|cc|} \hline a_{1,1}f_{1,1} + a_{1,2}f_{1,2} & a_{1,2}f_{1,1} + a_{1,3}f_{1,2} \\ +a_{2,1}f_{2,1} + a_{2,2}f_{2,2} & +a_{2,2}f_{2,1} + a_{2,3}f_{2,2} \\ a_{2,1}f_{1,1} + a_{2,2}f_{1,2} & a_{2,2}f_{1,1} + a_{2,3}f_{1,2} \\ +a_{3,1}f_{2,1} + a_{3,2}f_{2,2} & +a_{3,2}f_{2,1} + a_{3,3}f_{2,2} \\ \hline \end{array}$$

Figure 2.7: Visualization of the cross-correlation operation. The filter is superimposed on the image, marked in blue. This field is called the receptive field. This region is multiplied with the filter weights and summed up, resulting in an entry of the feature map. Then the filter is shifted by the so-called stride $s = 1$ to the right. This process is repeated until the whole image has been sampled.

Activation function: After the convolution operation and the added bias, an activation function is applied. We use the leaky ReLu function $\Phi : \mathbb{R} \to \mathbb{R}$, which is defined as

$$\Phi(x) := \max(x, 0) + \min(0, \alpha x) = \begin{cases} \alpha x, & x \leq 0, \\ x, & x > 0, \end{cases} \tag{2.11}$$

with $\alpha = 0.01$. It should be mentioned that the leaky ReLu function is used although it is not differentiable at $x = 0$, but it is at least piecewise differentiable. The derivative can be expressed by the Heaviside function. The Heaviside function $H : \mathbb{R} \to \mathbb{R}$ is defined as

$$H(x) := \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases} \tag{2.12}$$

The piecewise defined derivative can then be expressed by

$$H(x) + \alpha H(-x), \forall x \in \mathbb{R} \setminus \{0\}\,. \tag{2.13}$$

Fortunately the case that we reach zero on a computer is unlikely, as long as we do not run out of precision points. It is necessary that the activation function is a non-polynomial function to learn more complex scenarios, otherwise all layers of the network would collapse into one layer and only polynomial problems could be approximated, see [39, 40, 41].

Pooling: Another import layer is the pooling layer, which reduces the size of the feature maps. A filter runs over the feature map and pools the data, for example by taking the maximum value as the output. This reduces the size of the output image. Here, we use pooling filters of size $2 \times 2 \times 2$ and strides of the same size, which means that after a pooling layer the size is halved, see Figure 2.8, [38].

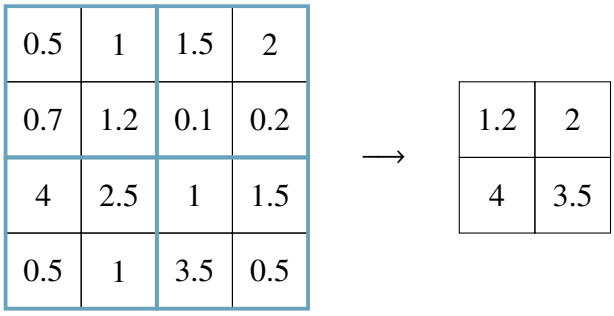


Figure 2.8: Visualization of a 2D maxpooling operation with an activation map of $\mathbf{A} \in \mathbb{R}^{4\times4}$ and a pooling filter $\mathbf{P} \in \mathbb{R}^{2\times2}$, resulting in an output activation map of halved size.

Shortcut connection: It has been observed that in deep networks, the so-called vanishing gradient problem occurs. This means that when adjusting the weights the gradient becomes smaller and smaller with each propagation of the error to the previous layer, leading to a vanishing gradient, which impedes learning, [42]. This can be overcome with so-called shortcut connections. Here, the output of a previous layer is simply forwarded to the input of a back layer by concatenating them, [42]. In the U-Net structure each output of an encoder block is concatenated with the input of the corresponding decoder block, see Figure 2.6.

Up-sampling: Because the output size of the network has the same dimension as the input size of the network, up-sampling techniques are needed to undo the dimension reduction by the maxpooling operation. Thus, these layers increase the resolution of the output, [18].

Backpropagation: The training of the network is done with the well-known backpropagation algorithm. It consists of two steps: forward propagation and back propagation. In the forward propagation, the input data is propagated forward through the network and the prediction of the network is calculated. As already mentioned, the quality, e.g., error of the network is determined by the cost function. In the backpropagation step, the error of the network is propagated back through the network. There, the partial derivative of the cost function according to each parameter of $\mathbf{p}$ is calculated. These can then be adjusted such that the cost decreases, see Equation (2.9). For a more detailed description see [34].

2.3 CT and DECT

In conventional CT, a single X-ray spectrum is used to image the human body based on different attenuation of tissues within the patient. Figure 2.9 shows the main components of a CT set-up. The source emits X-rays that pass through an object, where they are partly absorbed. The attenuated beams are then measured at the detector, [43].

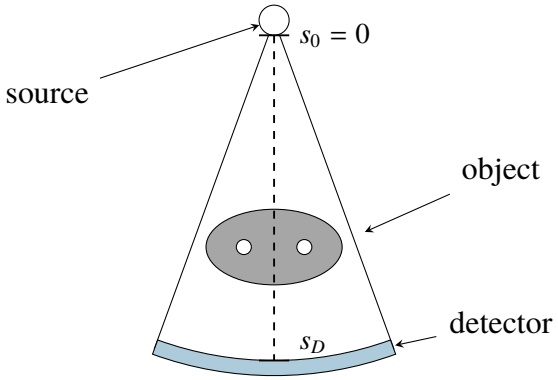


Figure 2.9: A conventional CT setup consists of an X-ray source that irradiates an object (patient), and a detector, measuring the attenuated radiation. The unit consisting of X-ray tube and detector is called gantry and can be rotated around the object. We consider the X-ray attenuation along the displayed line between the source at $s_0 = 0$ and the detector at s_D.

Let $I : \mathbb{R} \to \mathbb{R}_+$ be the intensity of the X-ray. By travelling through an inhomogeneous object along a line between the source and the detector, drawn in Figure 2.9, it is damped by the attenuation μ. We consider the attenuation as a function $\mu : \mathbb{R} \to \mathbb{R}$, where $\mu(s)$ describes the attenuation along our X-ray beam between the source s_0 and the detector s_D. We assume that $\mu(s) = 0$ outside the considered object.

The decrease of the radiation intensity I of the X-ray beam by travelling through an inhomogeneous object can be described by

$$\frac{dI}{I(s)} = -\mu(s)ds, \tag{2.14}$$

resulting in an ordinary differential equation. Using the initial condition $I(0) = I_0$, which is the X-ray intensity at the source, leads to the Beer-Lambert law

$$I_D = I(s_D) = I_0 e^{-\int_0^{s_D} \mu(s)ds}, \tag{2.15}$$

where I_D is the intensity at the detector and s_D is the source detector distance.

Converted, this results in the so-called projection p

$$p = -\ln\left(\frac{I_D}{I_0}\right) = \int_0^{s_D} \mu(s)ds. \tag{2.16}$$

In a CT measurement, we want to determine the attenuation coefficient $\mu : \mathbb{R}^2 \to \mathbb{R}$ at each point (x, y) of our patient coordinate system. For simplification we consider only one slice of our CT measurement, resulting in a 2D image. The representation of the spatial distribution of the absorption coefficient is our CT image, introduced in Section 2.2. As we have already seen this cannot be measured directly. To reconstruct μ we have to measure different projections p. To achieve this, the source and detector are rotated by at least 180° plus an fan angle around the object, resulting in projection data $p(\xi, \gamma)$ for different rotation angles γ at respective detector voxels ξ. Figure 2.10 sketches the rotated gantry. Here, (x, y) is the fixed coordinate system of the patient and (ξ, η) is the rotated coordinate system of the gantry.

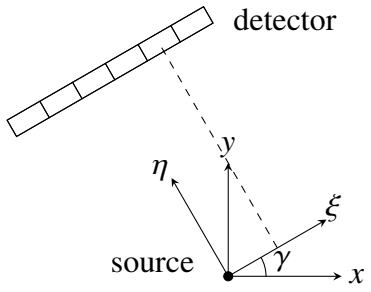


Figure 2.10: The (ξ, η) system is rotated by different angles γ relative to the fixed (x, y) system. The dashed line is one example path of $\delta_{\xi,\gamma}$.

The relation between the projection data p and the attenuation coefficient μ can be expressed by the Radon Transformation. In general the Radon transformation R maps a function on $\mathbb{R}^n$ into the set of its integrals over its hyperplanes of $\mathbb{R}^n$, [44]. Here, we consider the projection along the path $\delta_{\xi,\gamma} : [0, s_D] \to \mathbb{R}^2$, described by its distance ξ to the origin and the rotation angle γ

$$p(\xi, \gamma) = R\left(\mu(x, y)\right) = \int_{\delta_{\xi,\gamma}} \mu(x, y) ds. \tag{2.17}$$

The integration path is defined as follows,

$$\delta_{\xi,\gamma} = \mathbf{R}_\gamma^{\mathrm{T}} \begin{pmatrix} \xi \\ \eta \end{pmatrix}, \tag{2.18}$$

where $\mathbf{R}_\gamma$ is the rotation matrix

$$\mathbf{R}_\gamma = \begin{pmatrix} \cos(\gamma) & -\sin(\gamma) \\ \sin(\gamma) & \cos(\gamma) \end{pmatrix}. \tag{2.19}$$

Equation (2.17) can be rewritten as follows using the transformation theorem

$$\begin{aligned} p(\xi,\gamma) &= R\left(\mu(x,y)\right) \\ &= \int_{\delta_{\xi,\gamma}} \mu(x,y)ds \\ &= \int_0^{s_D} \mu\left(\delta_{\xi,\gamma}(\eta)\right) \|\delta'_{\xi,\gamma}(\eta)\|_2 d\eta \\ &= \int_0^{s_D} \mu(\xi\cos(\gamma) - \eta\sin(\gamma), \xi\sin(\gamma) + \eta\cos(\gamma))d\eta. \end{aligned}$$

Without loss of generality, we can extend our integration limits by assuming that μ is zero outside our considered object. Thus, for every ξ and every γ we get the following relation

$$p(\xi,\gamma) = \int_{-\infty}^{+\infty} \mu\left(\xi\cos(\gamma) - \eta\sin(\gamma), \xi\sin(\gamma) + \eta\cos(\gamma)\right) d\eta. \tag{2.20}$$

This defines a mapping of 2D data $\{\mu(x,y) \mid x,y \in \mathbb{R}\}$ to the 2D data $\{p(\xi,\gamma) \mid \xi,\gamma \in \mathbb{R}\}$. The image $\mu(x,y),\ x,y \in \mathbb{R}$ can be reconstructed with various methods, like the Fourier Based Reconstruction, using the Fourier Slice theorem or the Filtered Backprojection [1, 44, 45, 46].

The aforementioned reconstruction methods are based on the Beer-Lambert law, see Equation (2.15), where no energy dependencies are considered. Thus, we assume that the X-ray spectrum is monochromatic, although in fact the emitted photons have different energies resulting in a polychromatic spectrum, shown in Figure 2.11. The continuous spectrum results mostly from the so-called "Bremsstrahlung" which originates from the deceleration of an electron by the deflection at an atomic nucleus. Thereby, photons with arbitrary energy between 0 and the complete kinetic energy of the electron can be produced. The characteristic peaks result from the interaction of an electron with the K-shell: An electron of the K-shell is ejected and leaves a hole, which is filled by an electron from an outer shell. Thereby, a photon is emitted with exactly this characteristic energy that corresponds to the energy difference between an outer and an inner shell [15, 46, 47].

Thus, the attenuation law of Equation (2.15) has to be adapted for polychromatic radiation by integrating over the energy E

$$I_D = \int_0^{E_{max}} I_0(E) e^{-\int_0^{s_D} \mu(s,E)ds} dE, \tag{2.21}$$

which cannot be easily solved for μ as in the monochromatic case.

Assuming monoenergetic radiation as in traditional CT measurements, the energy information of the attenuation coefficient is lost. In contrast to the so-called single energy computed tomography (SECT), spectral energy computed tomography (SPECT) or DECT take advantage of this energy dependency for material differentiation and tissue characterization. Therefore, they have to measure with more than one spectrum. The theoretical foundations were already

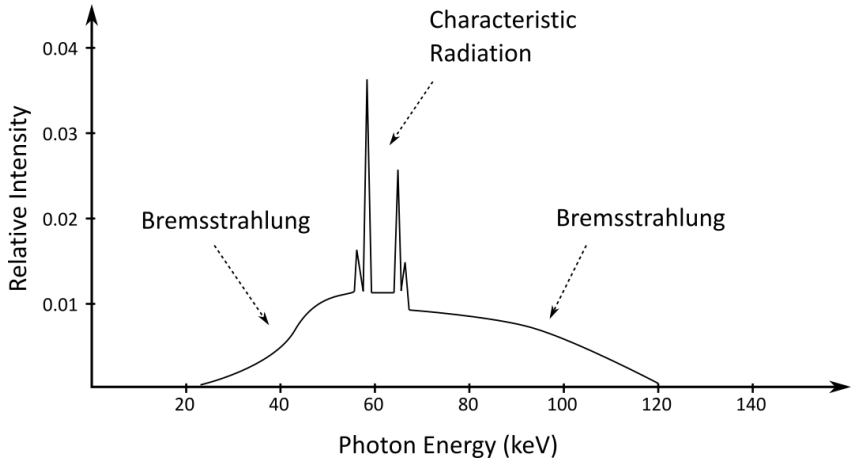
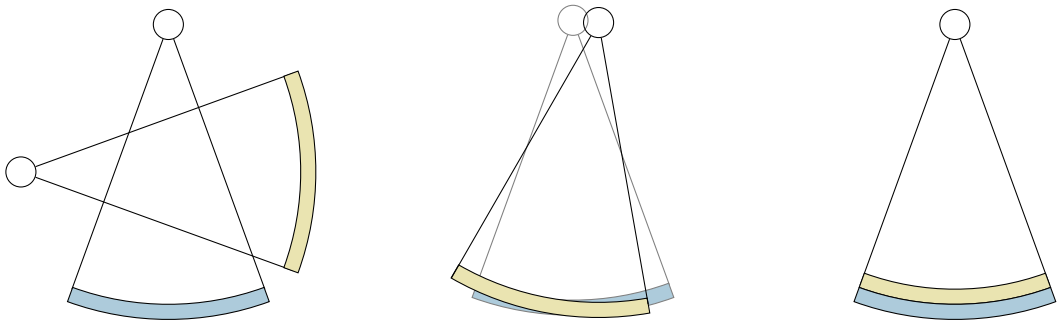


Figure 2.11: Typical X-ray spectrum for tube voltage of 120keV, inspired from [1] and [15].

established in 1979 shortly after the invention of CT, but due to technical limitations at that time, it could not be applied for the clinical use case. The terms DECT and SPECT are often used mistakenly as interchangeable in literature: while SPECT is able to measure with more than two spectra and therefore needs more advanced systems like photon counting detectors, DECT uses only two spectra and is a subset of SPECT, [15, 43].

In this work we use DECT data. As already mentioned it contains projection data of two different spectra, which are combined to get more information about the material. Mostly X-ray tube voltages of 80 kV for the lower spectrum and 140 kV for the higher spectrum are applied. It is desirable to have as little overlap of the spectra as possible, but simultaneously the tube voltage must not be too low because then the radiation will be completely absorbed. If the tube voltage is too high, it will result in an increased radiation exposure and the difference in soft tissue attenuation might be lower, [15]. There are several methods to acquire low and high energy spectra data, three of them are drawn in Figure 2.12 which shows typical DECT configurations. Figure 2.12a shows a dual-source detector configuration, where the same object is scanned by nearly perpendicular high and low energy spectral X-ray sources. One advantage is that the spectra are generated in two different tubes independently from each other, making separation of spectra simple. Disadvantages are limited space by using two sources and detectors resulting in a smaller field of view and cross-scatter effect which interferes with the respective other measurement. In addition it is vulnerable to motion artifacts. Figure 2.12b shows a single source detector configuration which switches fast between low and high energy spectra. The advantages are that this method is really fast and robust against motion artefacts, but the projections of the two measurements are not perfectly aligned, making material decomposition in the projection domain difficult. Figure 2.12c displays a so-called sandwich detector, which consists of a two-layered scintillator detector, where the first layer absorbs lower-energy and the second layer absorbs higher-energy photons. This results in simultaneous measurements of the projection data which is robust to motion artefacts and can be directly decomposed in the projection domain. One disadvantage is that



(a) Dual-Source **(b)** Fast kV Switching **(c)** Sandwich Detector

Figure 2.12: Different DECT configurations, with two perpendicular sources and two detectors (Dual-Source), one source and one detector (fast kV switching) or one source and two detector layers (sandwich detector). The low energy is highlighted in yellow and the high energy in blue.

the spectra can not be separated perfectly [15, 43, 48]. In this work, the images are captured with a Philips IQon spectral CT scanner, which uses sandwich detectors for separating low- and high-energy data.

Now, we take a closer look at how material characterization with two measured spectra is realized. Due to the fact that DECT material characterization relies on the different energy dependencies of attenuation for different materials, we investigate which factors contribute to the attenuation within matter. The total attenuation can be decomposed into five summands resulting from different physical effects

$$\mu_{total} = \mu_{PE} + \mu_C + \mu_R + \mu_{pair} + \mu_{ph,n}, \tag{2.22}$$

where μ_{PE} denotes the attenuation coefficient for photoelectric absorption, μ_C for Compton scattering, μ_R for Rayleigh scattering, μ_{pair} for pair production and $\mu_{ph,n}$ for photon-nuclear reaction. In medical use cases, the predominant factors for attenuation result from the Compton scattering and the photoelectric effect, [15, 46, 48]. Thus, Equation (2.22) simplifies to

$$\mu_{total} \approx \mu_{PE} + \mu_C. \tag{2.23}$$

Figure 2.13 contains the main principle of the Compton scattering and the photoelectric effect. The Compton effect is a phenomenon in which an incoming photon collides with an electron of the outer shell and is scattered. The electron thereby receives part of the energy of the incoming photon and becomes a free electron. The Compton effect does not strongly vary between different materials and is only weakly dependent on the photon energy. In the photoelectric effect, the incoming photon collides with an electron of the K-shell. The photon is completely absorbed and the electron receives the complete energy of the incoming photon, which is sufficient to be ionized. An electron from the outer layer then falls back onto the K-shell, emitting a photon with exactly the energy difference between the outer and inner

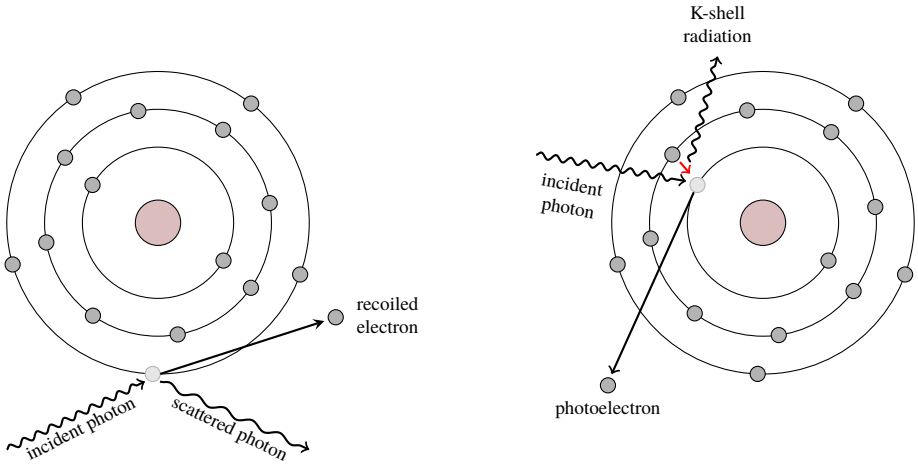


Figure 2.13: Compton scattering (left) and photo effect (right), where the red circle stands for the atomic nucleus and the gray circle represents the electrons in the different shells.

shell. Materials with small atomic numbers exhibit a low photoelectric effect, while materials with large atomic numbers exhibit a very strong photoelectric effect. Moreover, if the incident radiation is close to the K-shell binding energy, the probability of photoelectric activity is highest and therefore the absorption spectrum has local maxima there, [46, 48].

The material decomposition is based on exactly these two effects, as they have different strengths in different materials.

Mainly responsible for the material decomposition is the photoelectric effect, since it depends very strongly on the energy and the atomic number. To distinguish between different materials, they must differ in their atomic numbers. The Compton effect conversely depends mainly on the electron density. For example, materials with higher atomic numbers such as calcium (Z=20) and iodine (Z=53) can be easily distinguished from materials with lower atomic numbers such as hydrogen (Z=1), carbon (Z=6), nitrogen (Z=7) and oxygen (Z=8), [15].

The attenuation can be written as a linear combination of energy dependent basis functions f and energy independent coefficients $a_x,\ x \in \{PE, C\}$

$$\mu(x, y; E) = \underbrace{a_{PE}(x, y) f_{PE}(E)}_{\mu_{PE}} + \underbrace{a_C(x, y) f_C(E)}_{\mu_C}, \tag{2.24}$$

where

$$f_{PE}(E) = \frac{1}{E^3}, \tag{2.25}$$

and $f_C(E)$ is the so-called Klein-Nishina function $f_{KN}(E)$

$$f_{KN}(E) = \frac{1+\alpha}{\alpha^2}\left[\frac{2(1+\alpha)}{1+2\alpha} - \frac{1}{\alpha}\ln(1+2\alpha)\right] + \frac{1}{2\alpha}\ln(1+2\alpha) - \frac{1+3\alpha}{(1+2\alpha)^2}, \tag{2.26}$$

with

$$\alpha = \alpha(E) = \frac{E}{510 \cdot 975}\,\text{keV}. \tag{2.27}$$

The coefficients of Equation (2.24) are approximately

$$a_{PE}(x,y) \approx K_1 \frac{\rho(x,y)}{A} Z^4,\ \forall x,y \tag{2.28}$$

and

$$a_C(x,y) \approx K_2 \frac{\rho(x,y)}{A} Z,\ \forall x,y, \tag{2.29}$$

where K_1 and K_2 are constants, ρ is the mass density and Z is the atomic number, [1, 17, 43, 49]. As we have seen in Equation (2.16) and (2.20), in conventional CT our measured raw data are line integrals of the attenuation coefficient in beam direction. Here we also measured line integrals of the energy-independent coefficients a_{PE} and a_C because the energy dependent basis functions can be drawn out of the integral.

$$\int \mu(x,y;E)ds = A_{PE} f_{PE}(E) + A_C f_C(E), \tag{2.30}$$

where

$$A_{PE} = \int a_{PE}(x,y)ds \quad \text{and} \quad A_C = \int a_C(x,y)ds. \tag{2.31}$$

Inserting Equation (2.30) into Equation (2.21) and exploiting the fact that we have taken two measurements, one with low energy, and one with high energy X-ray spectra, results in the following non-linear system of equations

$$I_{low}(A_{PE}, A_C) = \int I_0(E) e^{A_{PE} f_{PE}(E) + A_C f_C(E)} dE, \tag{2.32}$$

$$I_{high}(A_{PE}, A_C) = \int I_0(E) e^{A_{PE} f_{PE}(E) + A_C f_C(E)} dE. \tag{2.33}$$

This system has to be solved for A_{PE} and A_C; several approaches exists in current research, which would go beyond the scope of this work. Assuming, that we have A_{PE} and A_C, a_{PE} and a_C can be reconstructed similar as in the conventional CT using a Radon transformation, [1, 49].

Figure 2.14 shows some DECT representations in blue, which can be generated from the low- and high-energy raw data (marked in red). In the first step the raw data is decomposed into Compton and photo data and reconstructed to the Compton $a_c(x,y)$ and photo $a_{PE}(x,y)$ images (Basic Decomposition). As shown in Equation (2.24) these can be combined linearly. Evaluating the energy-dependent basis functions at different energy levels, different so-called monoenergetic representations can be computed. Monoenergetic images simulate how the actual image would look like if the data were measured with a monochromatic X-ray beam at that energy and with the intensity corresponding to that energy. Thereby, different contrasts

can be generated. For example iodine has a high atomic number and is highly energy-dependent and illustrates the change within the monoenergetic images well. Due to the fact that iodine has a K-edge of 33.2 keV the attenuation is highest at 40 keV and decreases at higher energies. Thus the lower the energy, the higher the iodine attenuation and the better the contrast, because low-Z materials are less energy-dependent. Usually, energies between 40 keV and 200 keV are considered [15, 16, 17, 48].

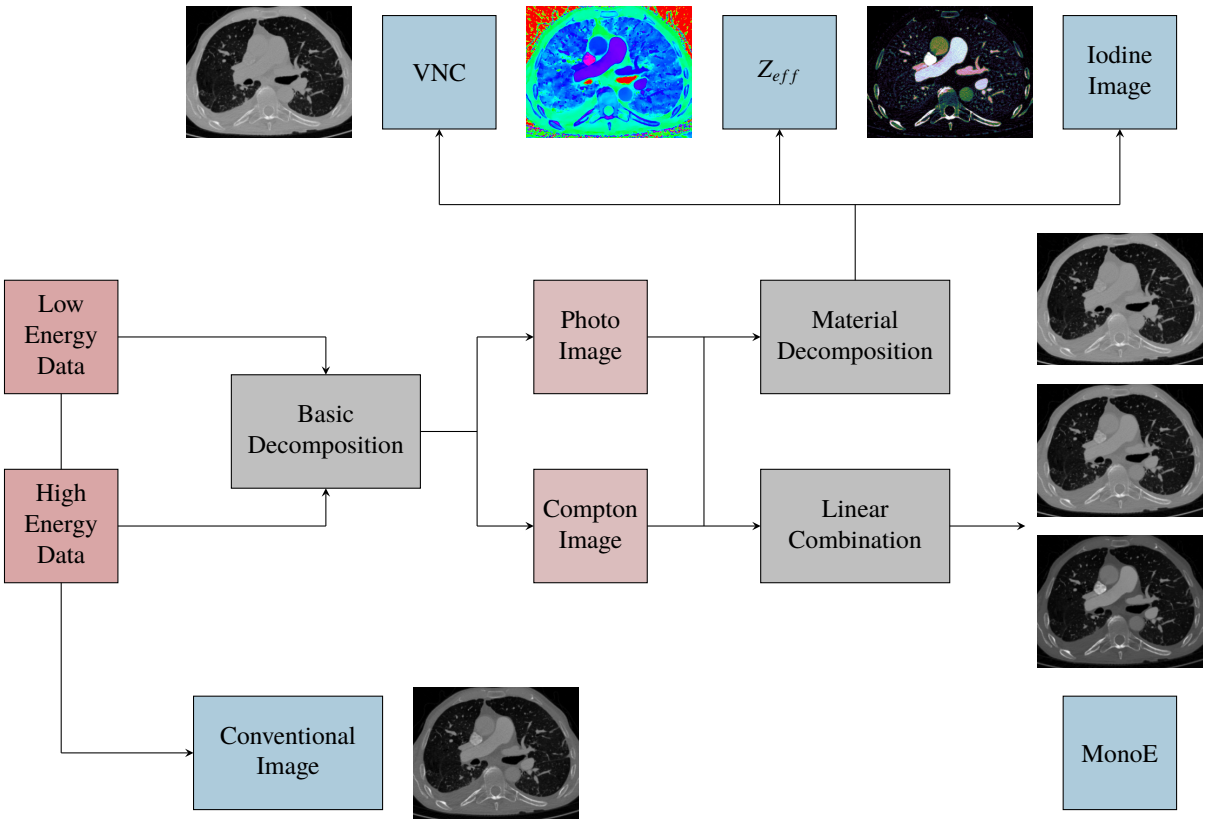


Figure 2.14: Different DECT representations, which can be constructed with material decomposition or linear combination from the low and high energy data.

Unlike with the monoenergetic images, which have different contrast, but the same basic anatomic information as in conventional CT images, different basic material decomposition (BDM) images can also be constructed. The main principle of the material decomposition is that different materials have different contributions to Compton and photo attenuation, see a_{PH}, a_C in Equation (2.24). Figure 2.15 shows a qualitative sketch, how the material decomposition can be performed. For example, an iodine-water material pair can be used. The attenuation of pure water consists of the following contributions of photo and scatter basis functions

$$\mu^{\text{water}}(E) = a_{PH}^{\text{water}} f_{PE}(E) + a_C^{\text{water}} f_C(E). \tag{2.34}$$

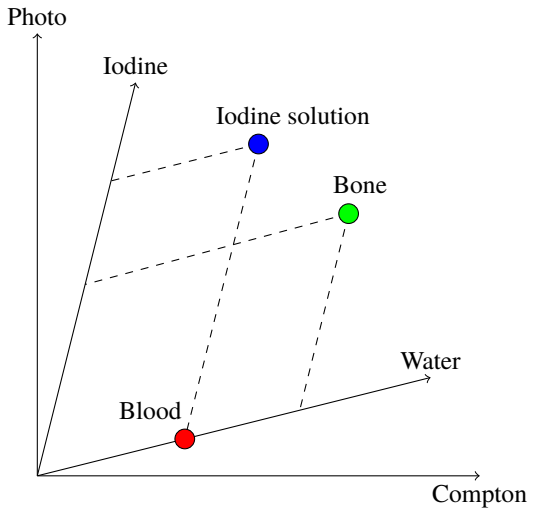


Figure 2.15: Material decomposition with iodine-water as the basis pair, inspired by [16]. Each material can be decomposed by its proportions of iodine and water. Blood (red dot) has no iodine proportion. An iodine solution (blue) has the same water contribution as blood and an additional iodine proportion. Bone (green dot) has a high iodine as well as water proportion.

Thus, the relative contribution of the photoelectric and Compton effect of water can be represented on a line. While in water the Compton effect dominates, iodine has a higher photoelectric effect and can be represented on another line within the Compton photo plot. Now, each measured Compton and photo attenuation can be expressed by the water-iodine basis pair

$$\mu^{\text{Bone}}(E) = a_{PH}^{\text{Bone}} f_{PE}(E) + a_{C}^{\text{Bone}} f_{C}(E) = a_{\text{water}}^{\text{Bone}} \mu_{\text{water}} + a_{\text{I}}^{\text{Bone}} \mu_{\text{I}}. \tag{2.35}$$

The decomposition does not imply that the material physically consists of this proportion of water and iodine, but that the attenuation of the material is the same as that of this combination of water and iodine. In this way, only iodine components can be displayed in the image, or the iodine component can be virtually suppressed to generate virtual non-contrast (VNC) images. Looking at the coefficients in Equation (2.28) and Equation (2.29), one can see that the attenuation can be parameterized by the effective atomic number Z. Thus, Z can be determined as the ratio of the two coefficients.

2.4 Related Research

In this section an overview of related research is presented. Different CAD approaches for the automatic detection of PEs are compared. We differentiate between classical and deep learning approaches.

Various efforts to detect PEs in CT scans have been made since the last decade [3]. While first publications use classical approaches, e.g., [6, 7, 8], current papers put a stronger focus on deep learning approaches, such as [10, 11, 12, 50], due to the great success in other medical imaging areas [9]. Table 2.2 provides an overview of related research papers. In the following, we briefly discuss the basic concepts. Classical approaches consist mostly of the steps displayed in Figure 2.16.

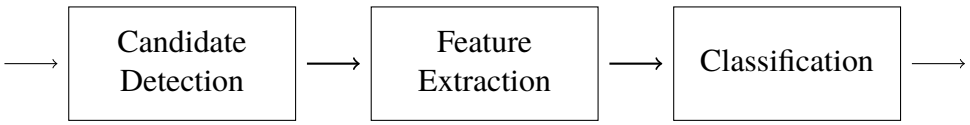


Figure 2.16: Key components of classical PE detection approaches. The candidate detection step makes a preselection of suspicious regions, mostly implemented with classical image processing techniques, on which different features, considering for example the intensity or geometrics, are extracted. The hand-crafted features are then used to detect pulmonary embolisms, mostly using classical machine learning algorithms.

The classical approach is realized for example in [6, 7, 8]. For candidate detection a vessel segmentation is performed in all approaches, but realized differently. While Zhou *et al.* extract the vessels via a clustering approach based on Expectation-Maximum [7], Bouma *et al.* apply intensity and morphological operations [8]. The proposed candidates are determined based on intensity values, the eigenvalue of the hessian matrix and morphology transformations, [8]. From the candidates, different features are extracted. These include geometric features, such as size or shape, intensity features or general location information. These features are utilized for the PE classification, which can be realized with conventional machine learning algorithms, such as KNN or decision trees, [8].

From Table 2.2 it can be seen that the earlier publications focused strongly on the classical approach, while in the later publications these have been partially or completely replaced by deep learning techniques. For example, in Tajbaksh *et al.*, the feature extraction and classification step was replaced by a CNN that combines both steps. Candidate selection is still based on a classical approach. A novelty here is that the candidates are aligned by the vessels, so that the most important 3D information from the cross-sectional and longitudinal 2D image can be included, which is then used for the classification. The vessel alignment approach was also used by Yang *et al.* and Lin *et al.* In addition, it is evident from the table that the trend is moving more and more towards end-to-end approaches. While the work of Yang *et al.* still consists of two separate steps, Lin *et al.* implemented an end-to-end network. Also Huang *et al.* implemented an end-to-end architecture called P-net, which has 3D inputs of the

entire images, but with reduced slices, unlike the 2D CNNs which consider only the cross sections, [11]. The end-to-end architecture eliminates the need for tedious and complicated pre-processing.

Furthermore, it must be mentioned that in earlier papers, the evaluation was mostly done on private data sets. Therefore, sensitivities of 100%, such as in Masutani *et al.* are difficult to compare with other sensitivities. There, no concomitant disease or image artifacts were included in the data set. Zho *et al.* has included such cases in their evaluation, whereby the sensitivity becomes worse. In Bouma *et al.* they achieve a higher sensitivity at that time despite using comorbidities.

Due to the lack of public data sets and poor possibility for comparison, annotated data sets, e.g., in the form of the CADPE challenge or the FUMPE were published. For example, it can be seen that Tajbaksh *et al.* achieves 83% on its own data set, but only 40% on the CADPE data set. Also, the work of Yang *et al.* and Lin *et al.* was evaluated on the CADPE data set. However, it must be added that here the data set was annotated independently, and thus the comparison with the previous results is not fair.

As already mentioned, we use a deep learning approach in this work which is based on the 3D U-Net structure. We train with three different data sets, the CADPE, FUMPE and INHOUSE data set. The INHOUSE data set results from collecting data for three years in a university clinic, where the PEs are diagnosed retrospectively as comorbidities. We investigate the generalizability and make several efforts to improve it.

Authors	Year	Approach	Data set (#cases)	Results (TPR)
Masutani *et al.* [6]	2002	**Classical:** one of the first CAD publications for PEs 1. vessel segmentation, 2. feature extraction, 3. voxel-based initial candidate proposal 4. connected component-based classification	**private** 19 11 with 21 PE wo comorbidities	**private** 100% at 7.7 AFP 85% at 2.6 AFP 100% casewise
Zhou *et al.* [7]	2005	**Classical:** include comorbidities 1. vessel segmentation 2. region extraction 3. feature extraction; 1,2 are based on EM segmentation	**private** 14 with 163 PE	**private** 52% at 11.2 AFP
Bouma *et al.* [8]	2009	**Classical:** better generalization due to comorbidities 1. vessel segmentation 2. candidate detction 4. feature extraction 5. classification	**private** 39 train (202 PE) 19 test (116 PE)	**private** 63% at 4.9 AFP
Gonzales *et al.* [3]	2013 2019	**CADPE challenge**: public data set for benchmarking UA-2.5D: U-Net based segmentation network on five slice input CT images MeVis: classical approach that finds filling defects	**public** 20 train (105 PE) 20 test (130 PE) 40 postchallenge	Winner 2013 MeVis 40% at 1.35 AFP Winner 2019 UA-2.5D 74% at 2 AFP
Tajbaksh *et al.* [12]	2015	**DL:** CNNs for FP reduction 1. lung segmentation 2. candidate detection 3. vessel alignment 4. longitudinal and cross-sectional image as CNN input	**private** 121 (326 PE) **public** CADPE	**private** 83% at 2 AFP **public** CADPE 40% at 2AFP
Yang *et al.* [50]	2019	**DL:** two stage CNN for PE detection 1. 3D candidate proposal network 2. vessel alignment of each cube 3. 2D classification network of cross-sections	**private** 129 (269 PE) **public** CADPE	**private** 84.2% at 2 AFP **public** 75.4% at 2 AFP
Lin *et al.* [10]	2019	**DL:** end-to-end CNN for PE detection 1. 3D candidate proposal network 2. vessel alignment subnet 3. 2D classification network of cross-sections	**private** 129 (269 PE) **public** CADPE	**private** 80.7% at 2 AFP **public** 86.8% at 2 AFP

Table 2.2: Related research of CAD algorithm of PEs. The overview contains next to the main author and the year of publication, a short key point summary of the used approach. There, we mainly differentiate between classical and deep learning (DL) approaches.

3 Methods and Materials

This chapter gives an overview of the methods and materials applied in this work. First of all, in Section 3.1 a detailed description of the existing data sets is provided, followed by the introduction of various evaluation metrics in Section 3.2 and concluded by presenting an overview in Section 3.3 of the methods used and experiments performed in order to answer the research questions of this work, see Section 1.2.

3.1 Data set

3.1.1 Overview

The data sets presented in the following contain CTA scans. The acquisition of these scans usually consists of the following steps: injection of a contrast material and capturing of the images. There exist standardized protocols which define the amount of contrast material and time between contrast agent injection and image acquisition. Usually, for imaging the pulmonary arteries, 20-30 ml of a 370 mg/ml iodine solution is injected as a contrast material. To ensure that the pulmonary arteries are full of contrast material, a strictly defined time interval of 6-13 seconds must pass between contrast material injection and image acquisition [5]. To reduce motion artifacts the images are captured during a single breath-hold.

We examine a total of four different cohorts from three different sources. Two data sets result from public challenges, the so-called FUMPE data set [2] and the CADPE data set [3]. The remaining data sets are in-house data sets that were provided by the Uniklinik Köln (UKK). Table 3.1 gives an overview of these data sets. Before we compare the data sets in more detail, we first describe the data collection and processing steps.

Public data sets

The FUMPE data set, published in 2018, consists of 35 CTAs, while the CADPE data set includes a total of 91 scans. The first subset of CTA scans from the CADPE data set was published in 2013 for the CADPE challenge: 20 CTA scans were provided for training, while 20 were used for evaluation of the challenge. The other 51 scans were first made available in 2019. Both the FUMPE data set and the CADPE data set have the same objectives: to

create publicly available data sets for benchmarks of existing algorithms and to facilitate the development of new algorithms, as well as the detection and segmentation of PEs. Most of the CADPE systems developed earlier were evaluated on private data sets, making an objective comparison between the algorithms difficult, [2, 3].

The FUMPE scans were made using a NeuViz 16 multi-slice helical CT scanner designed by Philips and Neusoft Medical System Co. While some of the scanned patients suffered from typical symptoms of pulmonary embolism such as dyspnea, tachypnea and pleuritic chest pain with haemoptysis, other patients had non-specific symptoms like tachycardia, palpitations, wheezing and cough, [2].

The CADPE data set was collected with SIEMENS Somatom Sensation 40 scanners at six different hospitals belonging to the Unidad Central de Radiodiagnóstico in Madrid in Spain. In addition to PEs there could be other pulmonary diseases within the data set, [3].

Both data sets were labelled with semi-automated software tools. The FUMPE data set was annotated by a radiologist with 5 years of experience who marked the region of interests (ROIs) in each scan, which were then passed to a segmentation software. Afterwards, a radiologist with 18 years of experience approved the annotations. Unlike other published data sets, FUMPE contains additional prognosis information such as right ventricular (RV) and left ventricular (LV) ratio, the reflux of the contrast material into the IVC, the pulmonary artery (PA) diameter and the Qanadli-score (Q-score), computing a weighted sum of the number of clots within the arteries. The higher the Q-score, the higher the mortality and morbidity rates, [2].

The CADPE data set was annotated by three advanced radiologists, each with more than 15, 20, and 19 years of experience, respectively. Each of them marked ROIs of PEs independently, passed to a semi-automated tool for segmentation. At the end the multiple image segmentations were fused with the STAPLE algorithm, [51]. Afterwards, a manual inspection of the combined segmentations was performed, [3].

In-House Data Sets

In 2021, we received a total of 169 CTA scans from the UKK, where 114 patients suffered from PEs, denoted as the INHOUSE data set, and 55 were healthy, called the NORMAL data set. The main difference from the scans of the public challenges is that these scans are DECT scans, i.e., each scan consists of a photo, scatter and noise image in addition to the conventional image, which allows for the calculation of different DECT representations.

Figure 3.1 illustrates the development of the data set from the UKK in a flow chart. There, all steps from image acquisition up to the pre-processing are included.

Firstly, in the image acquisition step the DECT images were recorded with a Philips IQon Spectral CT scanner. Secondly, the images were annotated. Scans without any PEs were grouped together to form the NORMAL data set, and those with PEs to form the INHOUSE data set, respectively.

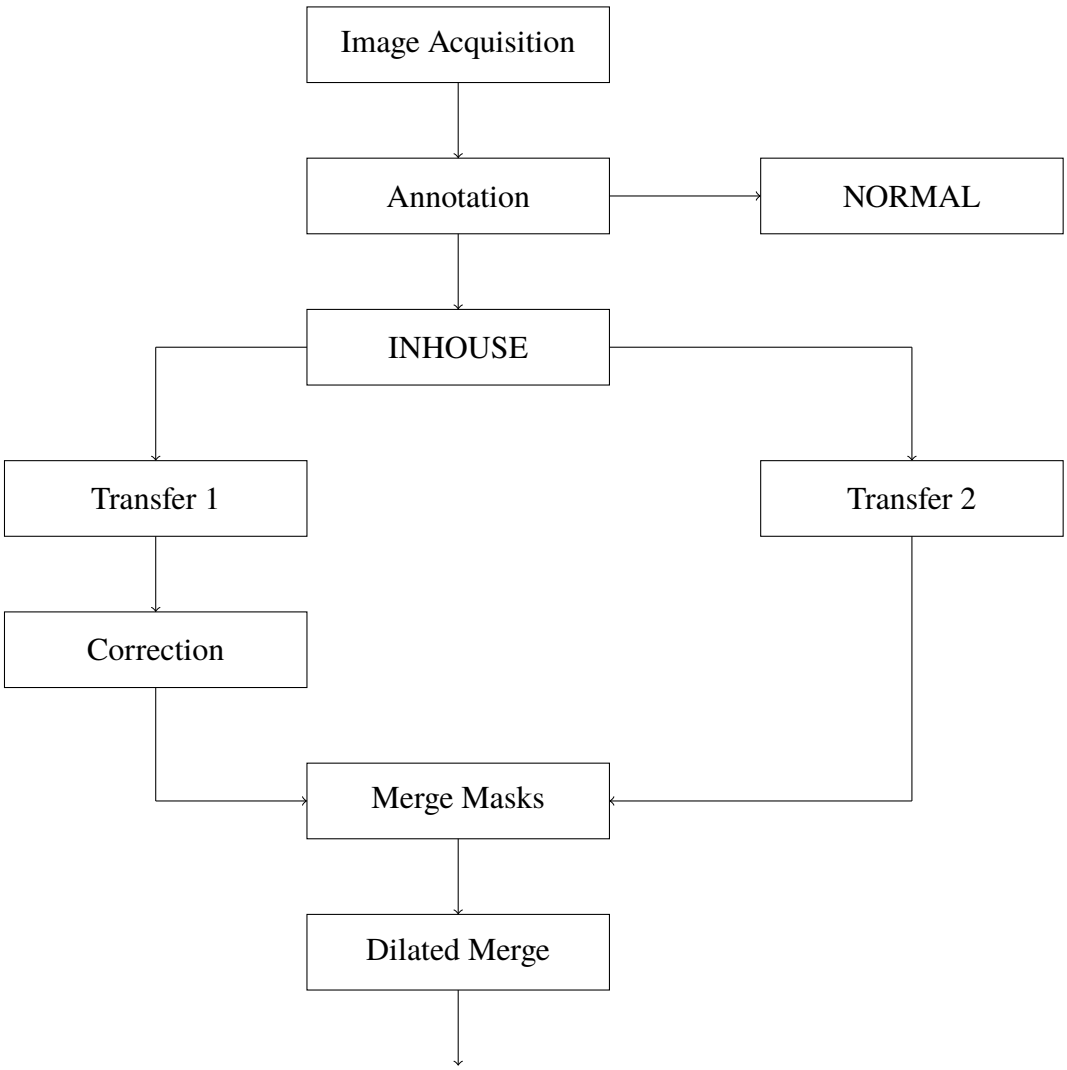


Figure 3.1: The development of the INHOUSE data set in a flow chart. The images are collected. Scans without PEs are grouped together to the NORMAL dataset, scans with PEs are grouped together to the INHOUSE dataset. The INHOUSE dataset are transmitted via two transfers. After correction of the annotations from Transfer 1, the masks from both transfers are merged. At the end a dilated merging is performed to get a consistent clustering of the PEs.

The review of the INHOUSE data set has shown that it was prone to label inconsistencies and missing annotations. The following paragraph provides detailed information on how these were addressed by manual correction and a data re-transfer.

The so-called INHOUSE data set was transmitted via two transfers: Transfer 1 (April 2021) and Transfer 2 (November 2021). The first transfer resulted in 66 scans with corresponding annotations. We noticed that some annotations were missing, presumably due to transmission losses. Therefore, all data were checked and corrected if necessary. For this purpose, the data set was divided among three Philips employees. Each person checked and corrected the annotations in their part. To go through all slices of a scan takes about 30 min. In case of uncertainties, the scans were noted and discussed together. At the end, all notations were gone through a second time and checked by one person. The complete process took a total of one month.

The left image of Figure 3.2 shows an example of a transmitted annotation, where a huge proximal thrombus is missing in the main pulmonary artery. The right image displays the same scan with our corrected annotation.

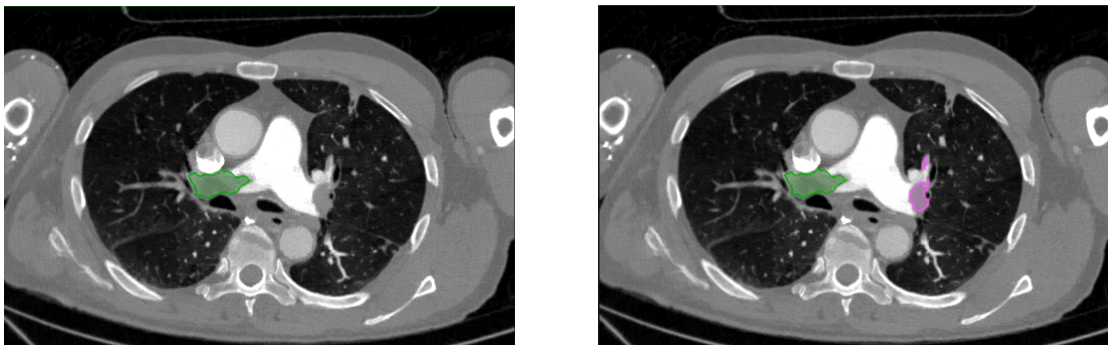


Figure 3.2: Axial view of CT scan and label mask overlay (green) with a missing annotation in the left main pulmonary artery (left image) and correction (pink) of the missing embolus (right image).

The second data transfer consists of 56 scans that had already been transferred during Transfer 1, but with novel annotations and 48 completely new scans. The 56 annotations from Transfer 2 were compared with the already-transmitted and corrected annotations from Transfer 1. Cases in which the annotations differed from each were identified automatically. These were then manually checked and merged (Merge Masks). Although the transmissions of Transfer 2 were loss-free, some PEs added during the correction were still missing in the label masks. However, many of the embolisms added in the correction step were present in the new annotations of the second transmission.

We have noticed that in the new transferred cases, the way of annotations differs from each other, e.g., some cases consists of several embolisms, but each embolism is annotated with the

same label, see Figure 3.3. Such inconsistent labelling methods obstruct a correct evaluation of the network performance. If one embolism is marked with several labels, and the network predicts only a part of the clot, it seems that the network does not detect all embolisms, when it actually does. The other way around, if several embolisms are marked with the same label, it seems that the network detects all existing embolism, although it only detects one of them.

To overcome that, we used a dilated merging procedure as the last post-processing step to generalize the annotations. We applied this step not only to the 48 new cases of Transfer 2, but also to our 56 merged intersections of the corrected data from Transfer 1 and Transfer 2, and to 10 corrected cases of Transfer 1 that were missing in Transfer 2, resulting in our 114 processed cases of the INHOUSE data set, see Figure 3.1.

Let $\mathbf{Y} \in \mathbb{N}^{w \times h \times d}$ be the label mask of the image $\mathbf{X} \in \mathbb{R}^{w \times h \times d}$, $w, h, d \in \mathbb{N}$. For each label $\mathbf{Y}$ of our data set, we compute the connected component label mask $\mathbf{C} \in \mathbb{N}^{w \times h \times d}$ that contains the connected components c_j, $j = 1, \ldots, N_{cc}$, $N_{cc} \in \mathbb{N}$.

For each mask $\mathbf{C}$, we compute the dilation with the structuring element $\mathbf{B}$

$$\mathbf{D} = \mathbf{C} \oplus \mathbf{B} := \bigcup_{b \in \mathbf{B}} \mathbf{C}_b, \tag{3.1}$$

where $\mathbf{C}_b$ is translation of $\mathbf{C}$ by b. Afterwards, the connected component label mask $\hat{\mathbf{C}}$ of the dilated label mask $\mathbf{D}$ is determined, with the new connected components $\hat{c}_k$, $k = 1, \ldots, \hat{N}_{cc} \leq N_{cc}$. In order to determine which components c_j belong to the same embolism, we define the set $S_{\hat{c}_k}$, containing all labels $c_j \in S_{\hat{c}_k}$ that are equivalent to $\hat{c}_k$, meaning they are part of the same embolism. All elements of the same equivalence class $c_j \in S_{\hat{c}_k}$ get the same new label $\hat{c}_k$ assigned.

$$\mathbf{C}^{c_j} \leftarrow \hat{c}_k \quad \forall c_j \in S_{\hat{c}_k}, \tag{3.2}$$

where

$$\mathbf{C}^{c_j} := \begin{cases} 1, & \text{if } c_{\mathbf{i}} = c_j, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3}$$

where $\mathbf{i}$ is the multi index of the array element. Thereby,

$$c_j \in S_{\hat{c}_k} \text{ if } \mathbf{C}^{c_j} \cap \hat{\mathbf{C}}^{\hat{c}_k} \neq \emptyset. \tag{3.4}$$

Figure 3.4 shows the CT images from Figure 3.3 after applying the dilated merging algorithm.

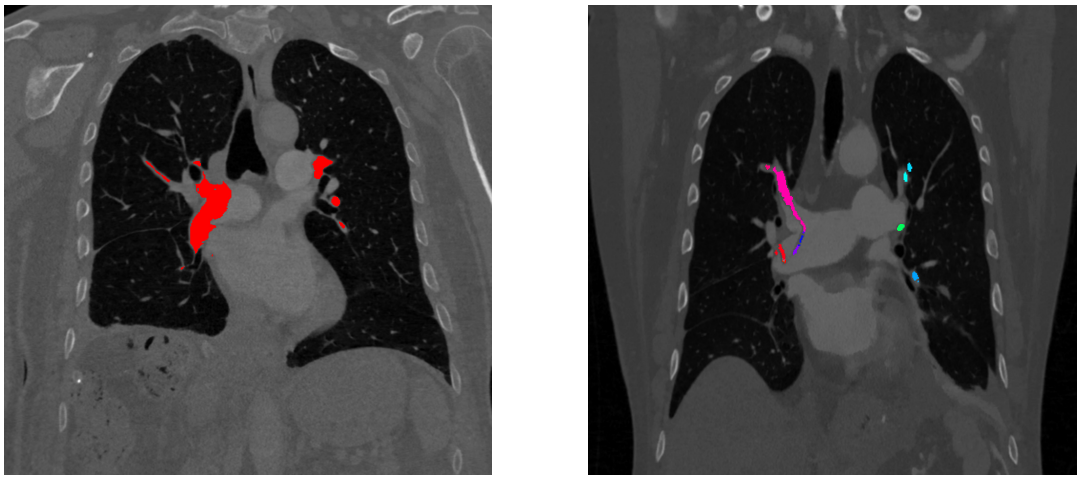


Figure 3.3: One label (red) for all embolisms in case 268 (left) and several labels (coloured overlay) for one embolism in case 288 (right) from the new transmitted data set.

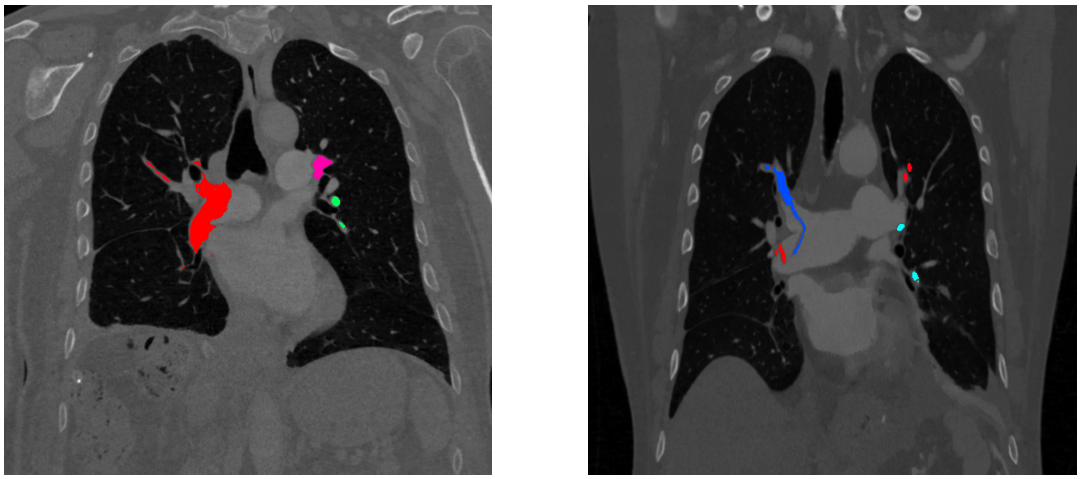


Figure 3.4: Images from Figure 3.3 after dilated merging. Different embolisms that have been marked with the same label were separated (left) and embolisms that had several labels were merged (right).

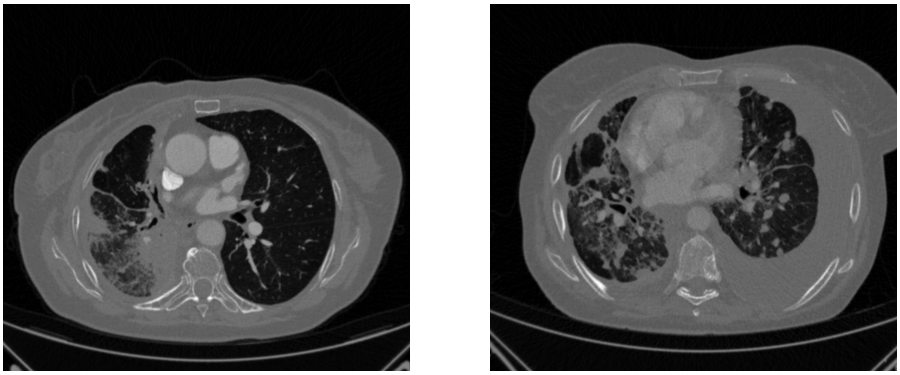


Figure 3.5: Comorbidities, such as cancer in the right lung (left and right) and infiltration (right). Separation of vessels is more challenging, due to less contrast to neighboring artefact tissue. Pulmonary embolisms are hardly visible.

During the data processing steps we noticed that there exist many cases within the INHOUSE data set with comorbidities, which all were marked as comorbidity cases. Two examples are shown in Figure 3.5 where it is obviously difficult to detect embolisms within. These were labeled as comorbidities and account for approximately 30% of the data set and can later be manually excluded from the training and/or evaluation.

Data Set Comparison

data set	year	N	N_{PE}	$\frac{N_{\text{PE}}}{N}$	DECT	$\mu\ [\text{cm}^3]$	$\sigma\ [\text{cm}^3]$	$r_\mu\ [\text{cm}^2 \times \text{cm}]$
INHOUSE	2021	114	552	4.84	yes	1.44	4.55	$0.77^2 \times 0.5$
NORMAL	2021	52	0	0	yes	-	-	$0.77^2 \times 0.5$
FUMPE	2018	35	110	3.14	no	4.24	8.73	$0.64^2 \times 0.91$
CADPE	2013/19	91	317	3.48	no	4.4	9.63	$0.7^2 \times 0.84$

Table 3.1: Overview of the used data sets, where N is the number of cases and N_{PE} the number of embolisms in total, μ the mean and σ the variance of the embolism volume and r_μ the average resolution of the 3D scans.

Table 3.1 shows that the public data sets have approximately the same number of embolisms per scan with 3.14 embolisms per case in the FUMPE data set, 3.48 embolisms per case in the CADPE data set, except for the INHOUSE data set, where the average number of embolisms is much larger, with 4.84 embolisms per case. Another metric by which we can estimate the complexity of the data set is the mean volume of the embolisms. One can easily imagine that if the data set consists of a few large proximal embolisms they will be more easily detected than many small peripheral embolisms. Related to that, the CADPE data set has the largest

thrombi with a mean volume of 4.4 cm^3, closely followed by FUMPE with 4.24 cm^3. Both are roughly 3 times larger than the mean PE volume of the INHOUSE data set with 1.44 cm^3.

In order to get a better understanding of the appearance of PEs within the three data sets, we created box plots, overlayed with violin plots of the volume distribution in Figure 3.6. Here, the distribution of the range between the minimum and the third quartile without outliers is displayed. Considering the distribution of the data sets one can observe that the INHOUSE data set consists of many more small embolisms than the public data sets. Although the CADPE data set has the largest mean value, it seems that the FUMPE data set has more larger embolisms between 3 cm^3 and 7 cm^3. In all data sets the volume range of the embolisms is widespread, e.g., INHOUSE, where 4.54 mm^3 is the smallest and 65 235.14 $mm^3 \approx 65.24$ cm^3 is the largest embolism, see Table 3.2. This large range of the search space illustrates the complexity of the problem, because CAD systems have to detect different appearances of embolisms. By comparing embolisms with smallest volume, it becomes noticeable that the minimum of the public data sets is much smaller than the minimum volume of the INHOUSE data set, although the majority of the embolisms have a smaller volume than those of the public data sets, which becomes evident through comparison of the first quartile, the median and the third quartile. FUMPE has the largest embolisms between the first and third quartiles, but the outliers are much smaller than those in the CADPE and INHOUSE data sets.

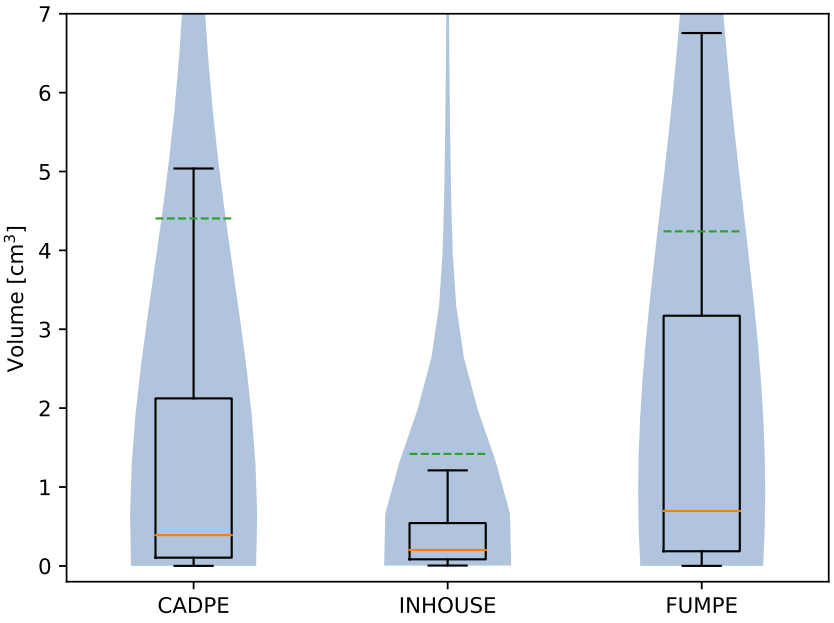
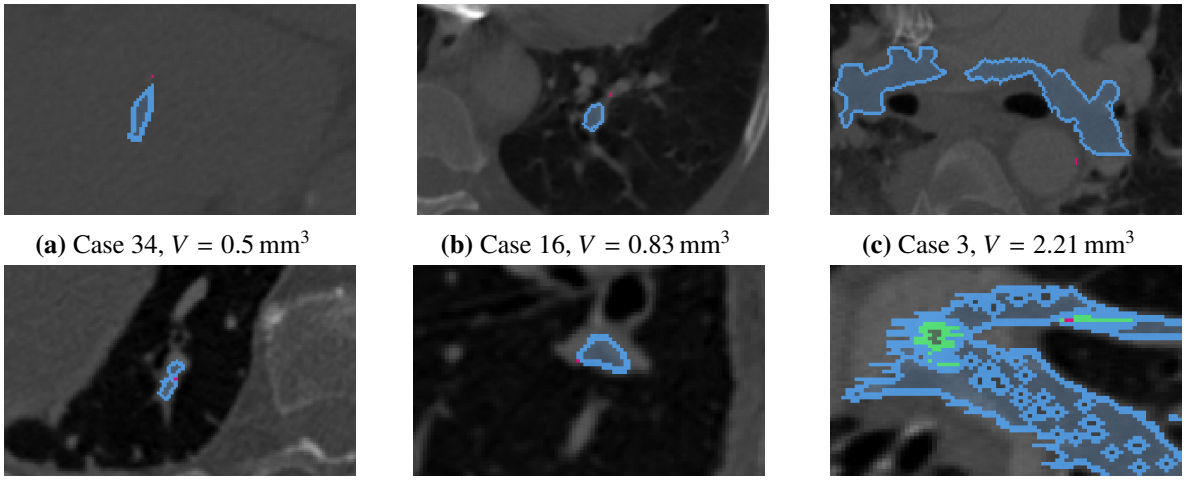


Figure 3.6: Volume box plots and violin plots of the used data sets, where only the range between the minimum and third quartile is shown. The orange line shows the median value and the dotted green line shows the mean value for the respective data set.

data set	min $[\mathrm{mm}^3]$	Q_1 $[\mathrm{mm}^3]$	median $[\mathrm{mm}^3]$	Q_3 $[\mathrm{mm}^3]$	max $[\mathrm{mm}^3]$
INHOUSE	4.54	82.61	190.2	536.87	65235.14
FUMPE	0.5	185.86	697.03	3171.75	46271.03
CADPE	0.34	104.55	390.36	2123.84	63645.34

Table 3.2: The five-number summary contains both extrema of the data set (min, max), the first quartile (Q_1), the median and the third quartile (Q_3).

We have examined the embolisms with the smallest volumes from the public data sets in more detail, and found that these mostly result from poor annotation. Figure 3.7 demonstrates some examples where embolisms with smallest volumes are marked in red. Obviously, these result from bad labelling and do not indicate an additional embolism. To give an example in Figure 3.7a and 3.7c some embolisms are not even located inside an artery. While both embolisms of Figure 3.7a lie within the pleural cavity below the lung, the red label of Figure 3.7c lies within the aorta descendens. Both are obviously not embolisms. In all other cases it seems that the red marked label belongs to the adjacent embolism, marked in blue or green. Because the label masks are from public challenges, we decided to keep them for benchmarking with other methods.



(a) Case 34, $V = 0.5\,\mathrm{mm}^3$ **(b)** Case 16, $V = 0.83\,\mathrm{mm}^3$ **(c)** Case 3, $V = 2.21\,\mathrm{mm}^3$

(d) Case 69, $V = 0.34\,\mathrm{mm}^3$ **(e)** Case 49, $V = 0.38\,\mathrm{mm}^3$ **(f)** Case 41, $V = 0.96\,\mathrm{mm}^3$

Figure 3.7: Embolisms with smallest volume V of the FUMPE data set (first row) and of the CADPE data set (second row) are marked in red. Adjacent annotations within the image cutout are marked in blue and green.

3.1.2 Peripheral and Proximal

To get a better understanding of the complexity of the different data sets we investigated the apperance of proximal and peripheral embolisms within the data sets. Table 3.3 gives an overview of the percentage, mean volume, and variance of the peripheral and proximal embolisms within the INHOUSE, FUMPE and CADPE data set. Due to the fact that peripheral embolisms are smaller, we can say as a rule of thumb that the more peripheral embolisms are within a data set, the more complex is the data set. But it should be noted that besides the volume, the contrast is another indication for the complexity. In this work we did not categorize embolisms in proximal and peripheral manually, but automatically. We generated additional lung masks for each image and determined the proportion of each thrombus lying inside and outside of the lung mask. If the majority of voxels was located outside the lung mask, we labeled the embolus as proximal; if a larger volume was located inside the lung mask, we labeled it as peripheral, see Figure 3.8.

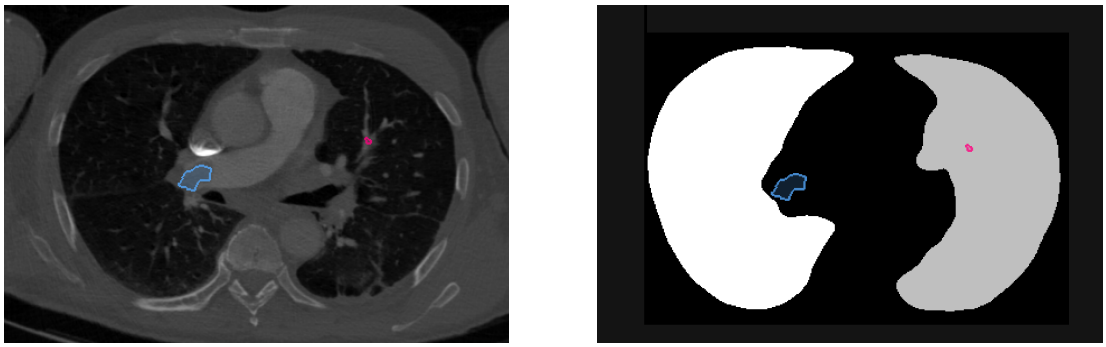


Figure 3.8: CT image (left) and lung mask (right) with label overlay. The blue embolism lies within the main pulmonary artery, which is not captured by the lung mask, which we use as an indicator for separating proximal and peripheral embolism, while the red embolism lies within the lung mask, specifically within a peripheral artery and is therefore a peripheral embolism.

Table 3.3 contains an overview of the peripheral and proximal embolisms in the data sets. One can immediately see that the INHOUSE data set has the highest percentage of peripheral embolisms, followed by the CADPE data set. In contrast, the FUMPE data set contains more proximal embolisms. The INHOUSE data set has the smallest mean volume of proximal and peripheral emboli, therefore it is the most challenging data set. In contrast to the FUMPE data set, the CADPE data set has few proximal embolisms, but these have a significantly larger volume, although we have already observed that this is mainly due to some large outliers. The FUMPE data set, on the other hand, has many more proximal embolisms, but these have a slightly smaller volume than those of the CADPE data set.

Figure 3.9 contains the histograms of the volume distribution of peripheral and proximal embolisms of all three data sets. Here it can be identified that in all three data sets the distribution of the peripheral embolisms in a logarithmic scale looks normally distributed.

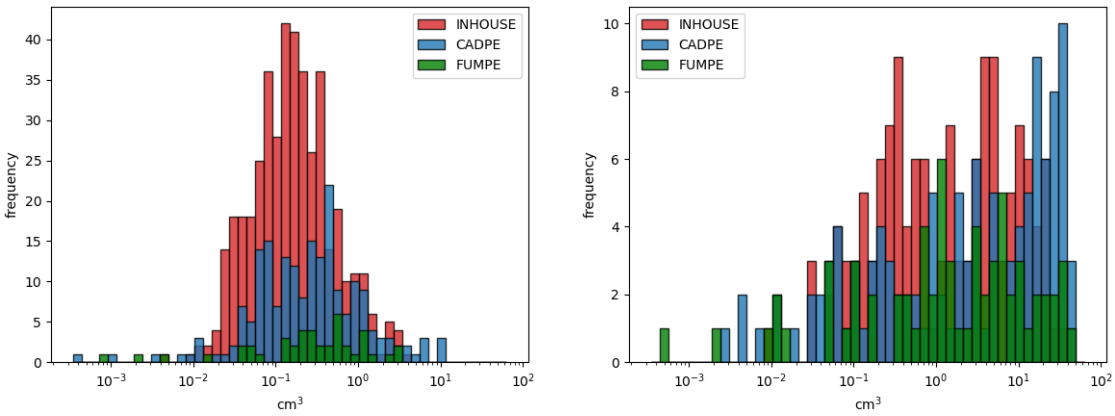


Figure 3.9: Volume distribution of the peripheral embolism (left) and the proximal embolism (right) of the INHOUSE (red), FUMPE(green) and CADPE (blue) data set using a logarithmic scaled x-axis.

The proximal embolisms are more distributed over the larger volumes; however, there are also some proximal embolisms that have a smaller volume and also some peripheral embolisms with a larger volume.

data set	N_{Proxi}	N_{Peri}	$\frac{N_{\text{Proxi}}}{N_{\text{PE}}}$	$\frac{N_{\text{Peri}}}{N_{\text{PE}}}$	μ_{Proxi}	μ_{Peri}	σ_{Proxi}	σ_{Peri}
unit			[%]	[%]	$[\text{cm}^3]$	$[\text{cm}^3]$	$[\text{cm}^3]$	$[\text{cm}^3]$
INHOUSE	131	421	23.73	76.27	5.07	0.31	8.34	0.47
FUMPE	67	43	60.91	39.09	6.59	0.58	10.55	0.76
CADPE	117	200	36.91	63.09	10.72	0.71	13.59	1.55

Table 3.3: N_i denotes the number of proximal and peripheral embolisms within the data set, μ_i the mean volume and σ_i the variance, $i \in \{\text{Proxi}, \text{Peri}\}$.

In summary, due to the variable appearance of pulmonary embolisms, detection and segmentation of pulmonary embolisms is a challenging task. In addition, we have determined that the annotations require several processing steps before they can be used for training. After a closer analysis of the data sets, it seems that due to the large number of peripheral embolisms, the small volume and other difficulties, such as the comorbidities, the INHOUSE data set is the most challenging.

3.2 Evaluation

As already mentioned, various evaluation metrics will be used to quantify the network performance. We will use free-response-receiver-operating-characteristic (FROC) curves to characterize the performance of the networks at different possible decision thresholds τ. In a FROC curve the sensitivity is plotted against the AFP rate, unlike the conventional receiver-operating-characteristic (ROC) curve where the sensitivity is plotted against the FP rate. While the ROC curve is suitable when there is a binary classification (diseased and not diseased), the FROC curve is suitable for problems in which several findings can be present in one object, i.e., a detection task.

The sensitivity or true positive rate (TPR) is defined as the number of true positives divided by the total amount of positive values on a per embolus basis

$$TPR = \frac{TP}{TP + FN}. \tag{3.5}$$

The AFP rate is defined as the number of false positives divided by the number of considered cases N:

$$AFP = \frac{FP}{N}. \tag{3.6}$$

The FROC is determined by computing the pair (TPR, AFP) depending on a free parameter τ for the threshold. At the lowest threshold, we accept all proposed predictions. This results in the highest number of TP, but also in the highest amount of FP (upper right point in FROC). The highest threshold for τ rejects more proposal predictions which results in the lowest number of TPs, but also less FP (lower left point in FROC). Here we select for the lower threshold $\tau = 0.5$ and for the upper threshold $\tau = 1$, e.g., $\tau \in [0.5, 1]$. Figure 3.10 shows several examples of FROC curves. In the ideal case we will have a sensitivity of 100% at an AFP of zero.

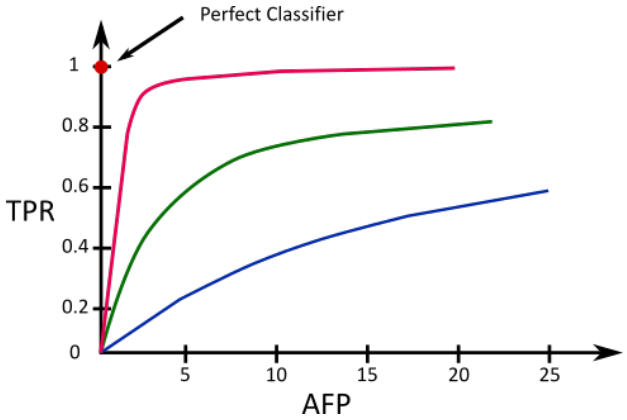


Figure 3.10: In a FROC curve the TPR is plotted against the AFP. The objective is to reach the left upper corner, having the highest possible TPR with the smallest amount of AFP.

In the following we explain how the pair (TPR, AFP) is determined for a given limit τ. Let $\mathbf{Y} \in \mathbb{N}_0^{h \times w \times d}$ be our label mask with connected components $c_r \in \underline{N}_{cc}$, with $\underline{N}_{cc} = \{1, \ldots, N_{cc}\}$, introduced in Section 2.2. Analogously, we define the network segmentation mask $\tilde{\mathbf{Y}} \in \mathbb{N}_0^{h \times w \times d}$ with $\tilde{N}_{cc}$ different connected components $\tilde{c}_s \in \underline{\tilde{N}}_{cc}$ with $\underline{\tilde{N}}_{cc} = \left\{1, \ldots \tilde{N}_{cc}\right\}$ as the set of predicted connected components.

For every connected component $c_r \in \underline{N}_{cc}$, we define an index set $\mathbf{I}_{c_r}$ for which holds

$$\mathbf{I}_{c_r} = \left\{ \mathbf{i} \;\middle|\; y_\mathbf{i} = c_r,\ y_\mathbf{i} \in \mathbf{Y} \right\} \tag{3.7}$$

and analogously the index set for each $\tilde{c}_s \in \underline{\tilde{N}}_{cc}$,

$$\mathbf{I}_{\tilde{c}_s} = \left\{ \mathbf{i} \;\middle|\; \tilde{y}_\mathbf{i} = \tilde{c}_s,\ \tilde{y}_\mathbf{i} \in \tilde{\mathbf{Y}} \right\}. \tag{3.8}$$

The network output is actually $\mathbf{P} \in [0, 1]^{h,w,d}$ which values can be interpreted as probabilities, denoting the network confidence. In our FROC curve, we use the average probability of each connected component as a decision criterion for including it in the calculation of the (TPR, AFP) pair. For each predicted connected component $\tilde{c}_s \in \underline{\tilde{N}}_{cc}$ we compute the average probability

$$p_{\tilde{c}_s} = \sum_{\mathbf{i} \in \mathbf{I}_{\tilde{c}_s}} \frac{p_\mathbf{i}}{\left|\mathbf{I}_{\tilde{c}_s}\right|} \tag{3.9}$$

If $p_{\tilde{c}_s} > \tau$, the connected component $\tilde{c}_s$ will be considered by computing the pair (TPR, AFP) belonging to this threshold τ.

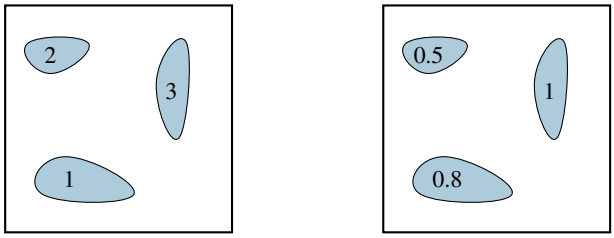


Figure 3.11: Connected component mask with $\tilde{c}_1 = 1,\ \tilde{c}_2 = 2,\ \tilde{c}_3 = 3$ and average probabilities $p_1 = 0.8,\ p_2 = 0.8,\ p_3 = 1$.

Figure 3.11 shows a sketch of a mask with three connected components and the respective average network probability. With a threshold of $\tau = 0.7$, only the components c_1 and c_3 would be considered as network predictions and included in the calculation.

To compute the TPR, we have to count how many of our connected components of our label mask $\mathbf{Y}$ are detected from our network. To count $c_r \in \underline{N}_{cc}$ as a TP, there has to be at least one match with a connected component $\tilde{c}_s \in \underline{\tilde{N}}_{cc}$ of our predicted label mask $\tilde{\mathbf{Y}}$. Therefore, we define a matcher function

$$M_{\underline{\tilde{N}}_{cc}}(c_r) = \begin{cases} 1, & \text{if } \sum_{\tilde{c}_s \in \underline{\tilde{N}}_{cc}} m(c_r, \tilde{c}_s) > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{3.10}$$

where $m(c_r, \tilde{c}_s)$ defines the matching criterion. We have a match between c_r and $\tilde{c}_s$ if a matching criterion is fulfilled, meaning that the corresponding matching function f outputs a greater value than a given matching threshold θ

$$m(c_r, \tilde{c}_s) = \begin{cases} 1, & \text{if } f(c_r, \tilde{c}_s) > \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{3.11}$$

In this work, we will use the one-pixel intersection (OPI) as a matching criterion, meaning that we have a match if there is at least one pixel overlap between c_r and $\tilde{c}_s$. That is fulfilled if the intersection of the index sets contains at least one element

$$f_{\text{OPI}}(c_r, \tilde{c}_s) = \left|\mathbf{I}_{c_r} \cap \mathbf{I}_{\tilde{c}_s}\right| > 0 = \theta. \tag{3.12}$$

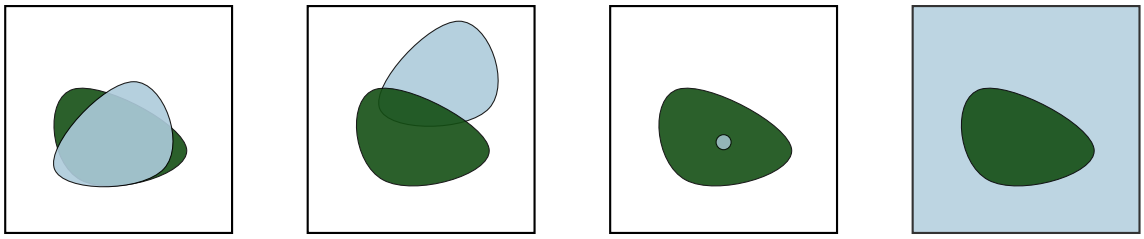


Figure 3.12: Four different scenarios where the OPI counts a match between the ground truth connected component (green) and the network predicted connected component (blue). In the most left image the prediction and reference label fits well, while in the other scenarios there is a spatial shift between them, or the network prediction is much smaller or larger.

Using the OPI all cases from Figure 3.12 are counted as a hit. Another more rigorous criterion could be the dice score (DC), which is generally defined as

$$DC = \frac{2|X \cap Y|}{|X| + |Y|} S \tag{3.13}$$

considering the sets X, Y. Using a lower threshold for the DC to determine if there is a match between prediction and ground truth, results in the following matching function

$$f_{\text{DC}}(c_r, \tilde{c}_s) = 2\frac{\left|\mathbf{I}_{c_r} \cap \mathbf{I}_{\tilde{c}_s}\right|}{\left|\mathbf{I}_{c_r}\right| + \left|\mathbf{I}_{\tilde{c}_s}\right|} > \theta. \tag{3.14}$$

When using the dice coefficient we use $\theta = 0.2$ as a lower bound. Here only the most left images would be interpreted as a match, making the dice coefficient a much stricter criterion. Depending on whether we focus more on segmentation or detection, either the OPI or the DC metric can be used.

In conclusion, the TPR can be computed by counting all matches of all images $i = 1, \ldots, N$ and dividing by the number of connected components of the reference images

$$TPR = \frac{\sum_i \sum_r M_{\underline{\tilde{N}}_{cc,i}}(c_r)}{\sum_i N_{cc,i}}. \tag{3.15}$$

The false positive rate can be computed by counting all connected components of the network prediction which have no match with a reference label and dividing by the number N of considered cases

$$FPR = \frac{\sum_i \left(\tilde{N}_{cc,i} - \sum_s M_{\underline{N}_{cc,i}}(\tilde{c}_s)\right)}{N}. \tag{3.16}$$

3.3 Methodology Approach

In this section the methodology approach is explained using the illustration in Figure 3.13. To answer the research questions structurally, the experiments are divided into three different parts. First, investigations where we focus on the input of the network, see Section 4.1, second, experiments where we intervene directly in the training, see Section 4.2, and third, different evaluations to assess the results in more detail are made, see Section 4.3.

First of all, in Section 4.1.1 a naive experiment is presented. This should provide a reference for comparison with all subsequent experiments and introduce the complexity of the problem. Afterwards in Section 4.1.2, the influence of the input data shall be examined in more detail by assessing the impact of the revision of the INHOUSE data set from Section 3.1.1 on the network performance. Furthermore in Section 4.1.3, we want to analyse how the choice of the data set determines the behaviour of the network and how well networks trained with a certain data set generalize to other data sets, followed by the usage of a combined data set for training, see Section 4.1.4. In addition, lung masks are used during trainings to reduce the number of false positives outside the region of interest. To further improve the performance, we also investigate the extent to which vessel masks can be used in this context, see Section 4.1.5. All these experiments are based on the input of the network, by passing different data sets, annotations and additional masks to the network.

In order for the networks to achieve good performance on their own data, we first tune the hyperparameters, see Subsection 4.2.1. We process in such a way that we change only one parameter at a time based on the initial values and examine its influence on the performance. Additionally in Section 4.2.2, classical data augmentation is used to create more variance in the data. Different geometric and intensity transformations, filters and the addition of noise are investigated. To analyse how the networks perform on images with different contrasts, we evaluate the networks on different monoenergetic energy levels, see Section 4.2.3. Afterwards, these are additionally introduced into the training to obtain a contrast-independent prediction.

The last part focuses more precisely on the evaluation. In addition to the metrics already used, such as the OPI and the DC, the center of mass for each predicted connected component is determined in order to compare our approach with the approaches of the CADPE challenge, see Section 4.3.6. Further experiments are the evaluation of the network on healthy images without pulmonary embolisms, see Section 4.3.1, and the differentiation between peripheral and proximal embolisms within the evaluation, see Section 4.3.2 . To understand the network behaviour in more detail, we take a closer look at the cases in which the network fails, i.e., how the false positive predictions of the network look like. We also take a closer look at embolisms, where the network has difficulties to detect them, see Section 4.3.4. Finally, we look at some examples where the networks fail without the DECT data augmentation, but the non-contrast networks succeed, see Section 4.3.5.

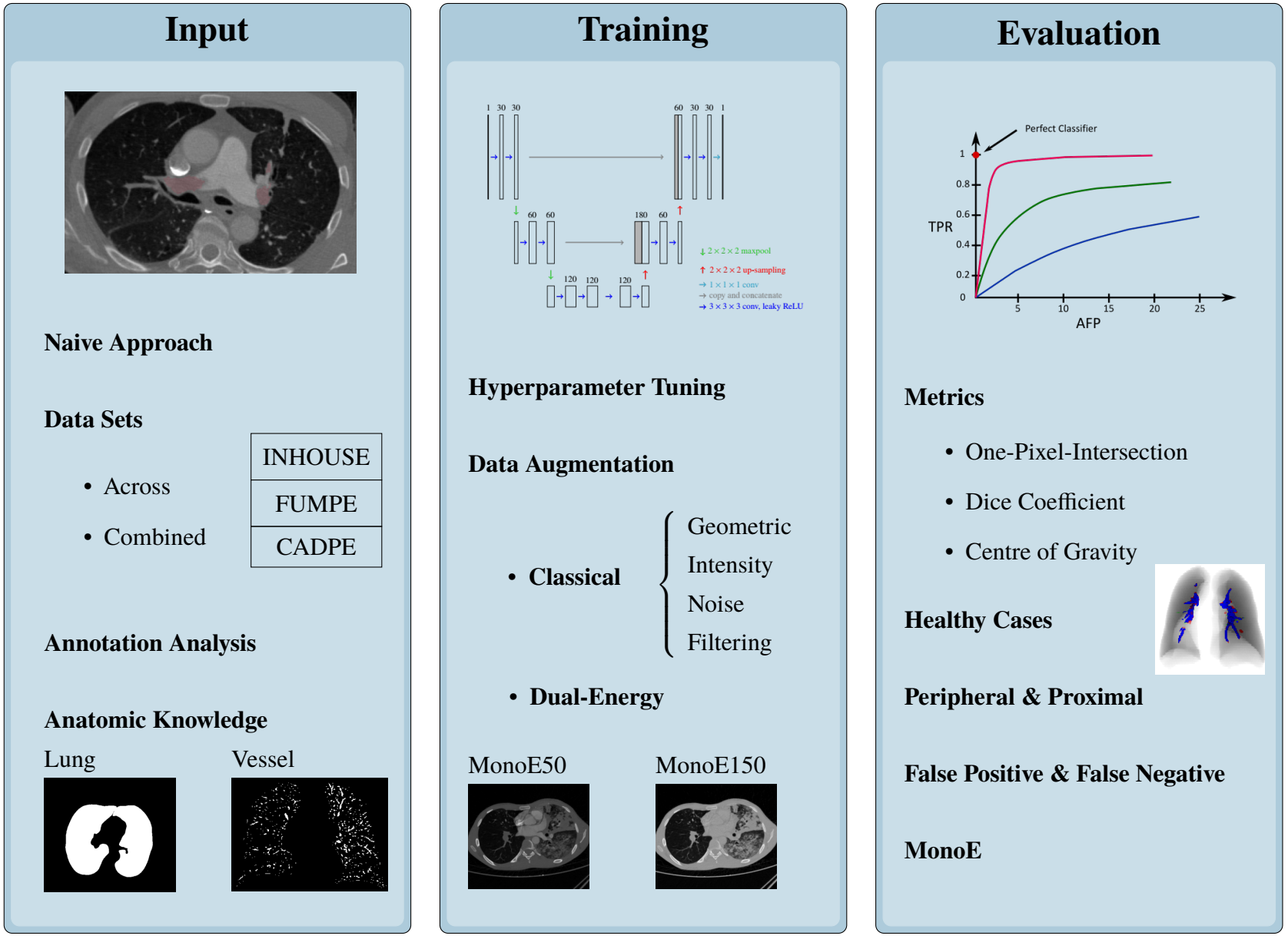


Figure 3.13: Methodology approach of the performed experiments to answer the research questions.

4

Experiments and Interpretation

In this chapter, the experiments performed are presented and the results are interpreted. We use Figure 3.13 from Section 3.3 as a guideline to answer the research questions step by step. Section 4.1 contains a naive approach and a detailed analysis of the influence of changes concerning the input of the network. This is followed by the investigation of different training strategies, such as testing different hyperparameters, classical data augmentation approaches and DECT augmentation, see Section 4.2. We conclude this chapter with a detailed evaluation of the performance of our networks in Section 4.3 to get a better understanding of their behaviour.

4.1 Input

In this section, we perform various experiments with different approaches regarding the input data. First, we present the experimental setup, having a closer look on a naive approach. Extending on this basis, we do further experiments to investigate and improve the performance and generalizability.

4.1.1 Naive Approach

To get an understanding of the problem and its complexity, the results of the first naive approach are presented. As a first experiment, we trained the network with the obtained data from Transfer 1 without our correction, see Figure 3.1. Additionally, this training ran with our first choice of hyperparameters, listed in Table 4.1. During this work, we adapt the hyperparameters step by step, see Section 4.2.1.

In order to understand the network behaviour, we have a closer look at the training behaviour and the FROC, which are displayed in Figure 4.1 and Figure 4.2. As we can see in the training plot, Figure 4.1, the validation loss decreases while the validation metric increases, meaning that the network is learning. Nevertheless, the performance of the network is bad. First of all, we can see in Figure 4.2 that both training and evaluation FROC curve are overall quite low. Moreover, the network generalizes very poorly, which can be seen from the large gap between them. Considering the sensitivity, which is defined in Section 3.2, at an AFP of 5 the OPI is 48.73% for training and 22.92% for testing.

Hyperparameter		**Setting**
Epochs	ep	1000
Mini-batch Size	mbs	8
Learning Rate	lr	8
Patch Size	ps	$72 \times 72 \times 72$
Loss	C	CrossEntropy + DiceLoss
Optimizer		AdaDelta
Augmentation Parameters	mixture	Flip x-axis Rotate Logarithmic Scale

Table 4.1: Setting of the hyperparameters in the naive approach. In all experiments, we use the sum of CrossEntropy and DiceLoss as the cost function and AdaDelta for optimization, which is an extension of the gradient descent algorithm. A detailed investigation of different hyperparameters and different data augmentation strategies can be seen in Section 4.2.1 and Section 4.2.2.

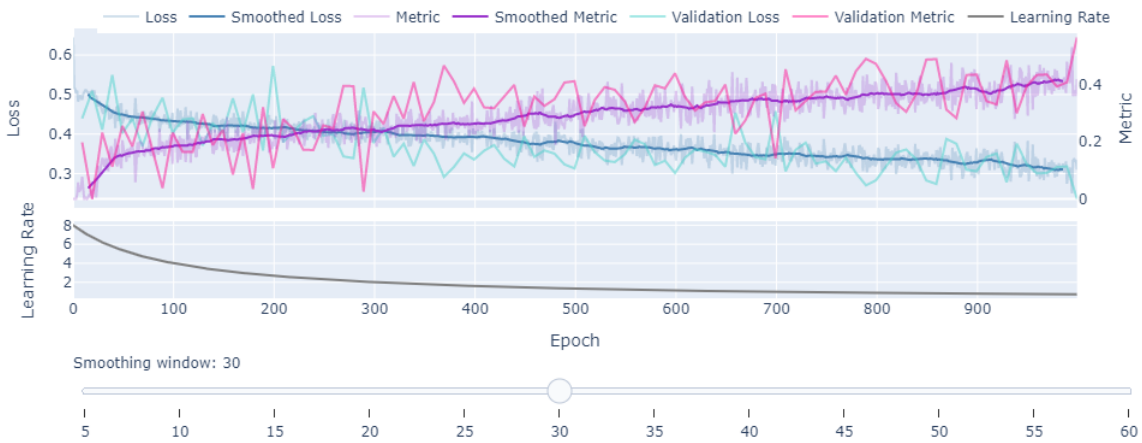


Figure 4.1: Training behaviour of the naive approach with real (low opacity) and smoothed (high opacity) training loss in blue, real validation loss in cyan, real (low opacity) and smoothed (high opacity) training metric in purple and real validation metric in pink. The training and validation metrics increase while the training and validation losses decrease.

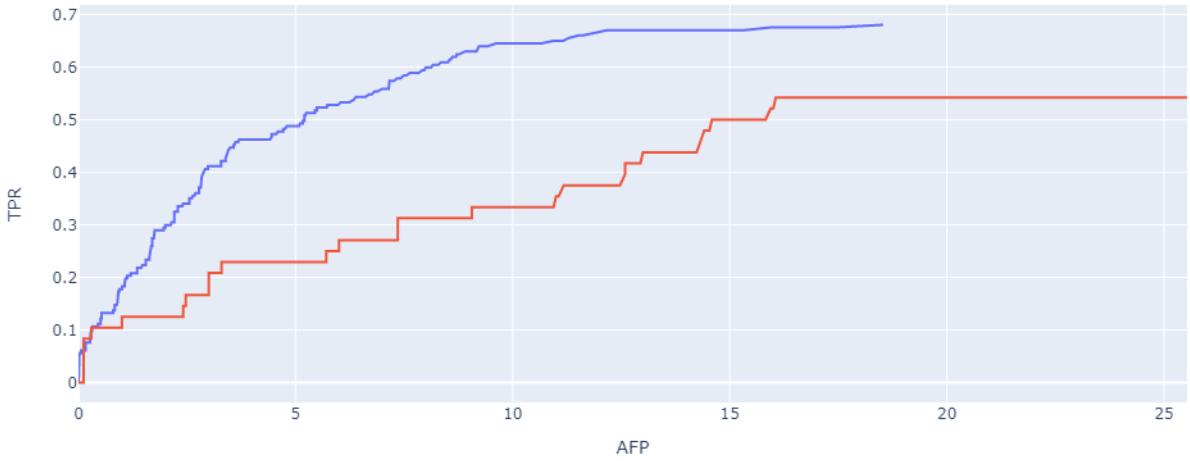


Figure 4.2: FROC curves of the naive approach. Training curve (blue) and validation curve (red) are still too low. The huge gap between them indicates overfitting.

This can have several reasons. First, poor choice of hyperparameters can negatively affect training, so one approach would be to examine these. Another approach would be to examine the data, as poorly labelled or complex cases that are already difficult for humans to recognize can impede training.

To further investigate the network behaviour, we have selected some cases where the network shows poor performance, see Figure 4.4. There, two effects stand out. One of them is that the network tends to detect sharp edges outside the ROI. For example, the edges on the ribs were detected (displayed in the first row) or parts of the CT device, see second row. However, FPs lie not only outside, but also inside the ROI, which is displayed in the last row. This shows that the network has not yet learned the right parameters for detecting PEs, but rather realizes a kind of edge detection.

This naive approach shows that the detection of PEs is not a trivial task and gives a guideline for improving network performance. Firstly, to omit FPs outside the ROI, for each image, the respective lung mask is generated and the image is cropped based on it. Figure 4.3 displays an example of an original scan, its corresponding lung mask and the cropped image.

Secondly, the performance has to be increased in general and generalization has to be improved. Thus as the next step, the annotations will be investigated and the hyperparameters will be tuned.

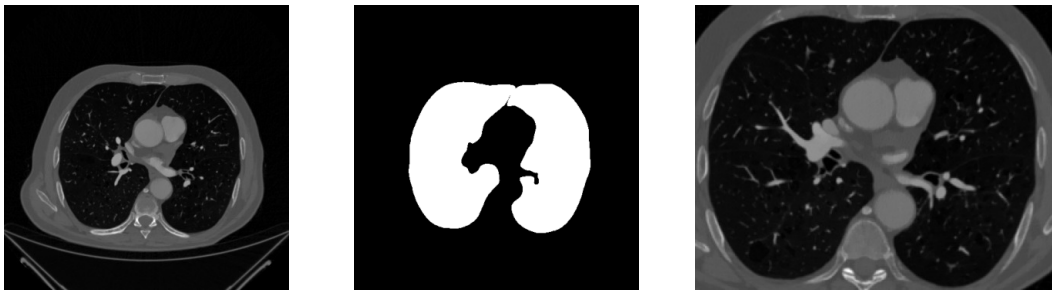
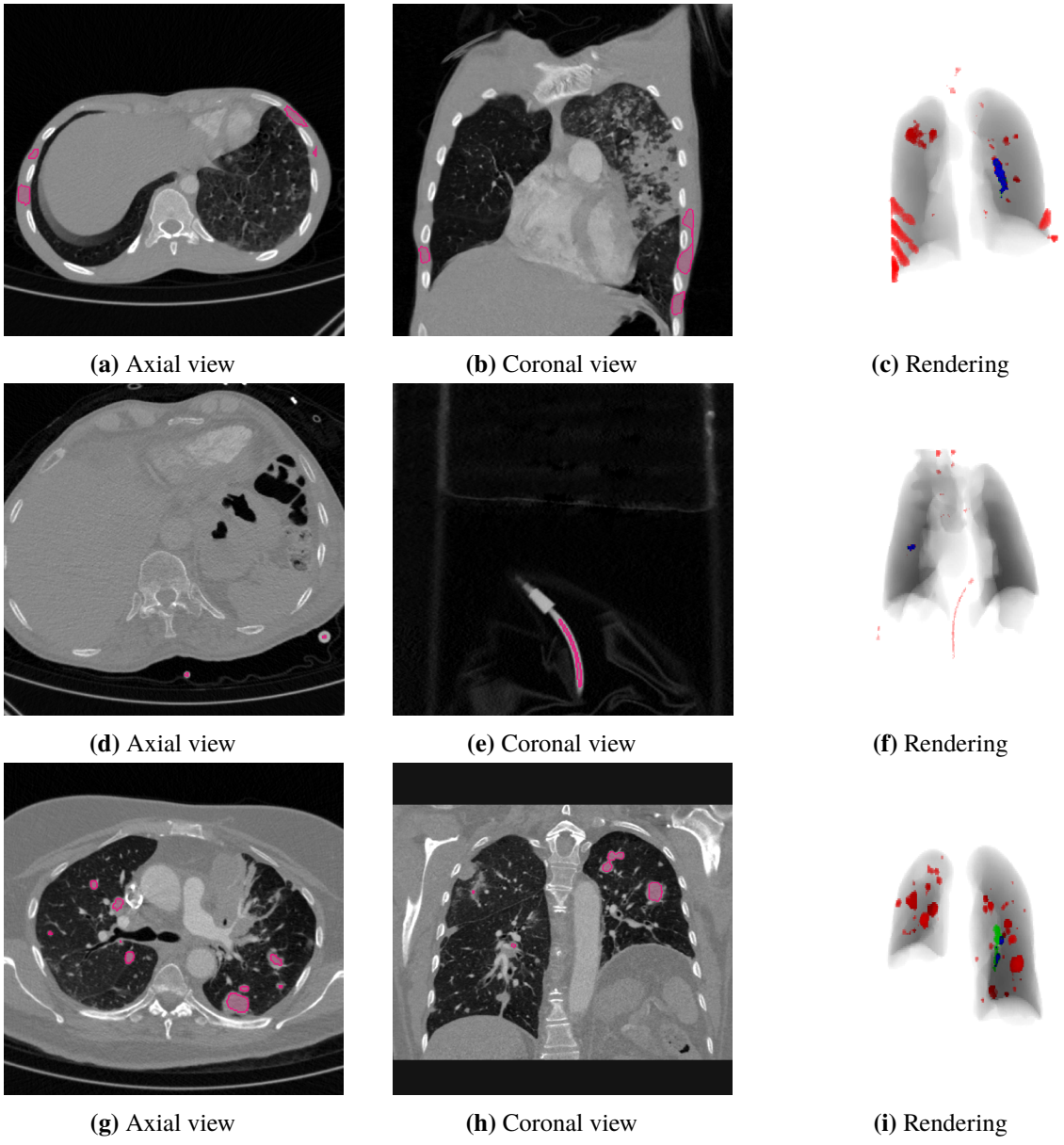


Figure 4.3: Original image, lung mask and cropped image of case 290 (axial view). Based on the segmentation mask, a bounding box is determined such that the complete lung mask fits in it. Using this bounding box, the image was then cropped without margins.



(a) Axial view **(b)** Coronal view **(c)** Rendering

(d) Axial view **(e)** Coronal view **(f)** Rendering

(g) Axial view **(h)** Coronal view **(i)** Rendering

Figure 4.4: Axial and coronal views of case 1, 257 and 99 with network predictions as overlay (a),(b) and renderings of the lung with false negatives (FNs) in green, FPs in red and TPs in blue (c). The rendering shows a 2D representation of the 3D image, the opacity indicated the density of the object.

4.1.2 Annotation Analysis

In this subsection, we perform experiments regarding the INHOUSE annotations. As mentioned in Section 3.1.1, there were two batches of the INHOUSE data set. Additionally, the data were processed in different steps: the labels were corrected, comorbidity cases were marked and annotations were standardized. To investigate the influence of these steps and the influence of the quality of the data in general on the stability of training and performance, the following experiments are conducted.

First of all, we use 4-fold cross-validation to investigate the performance difference between various splits. Here, we have used the annotations from Batch 1 before we correct them. The data set is split into four equal parts, where one part is used for testing and the remaining three parts are used for training. Each of the four parts was used once for testing, resulting in four trained networks. Cross-validation is usually used to examine how the method performs in general on different test sets. We use it to study how the data influences the training. Figure 4.5 shows the four different training curves of the splits. One can observe that there is a large difference in performance between the splits, especially between split 1 and split 2. While in the first split, the training is successful, in the second split the validation loss does not decrease properly and also the validation metric does not seem to increase. The OPI of these two splits is pictured in Figure 4.6 (a). While small variations between different splits are normal, with an OPI of 60% in split 1 and only 28.33% in split 2, the variance is quite high.

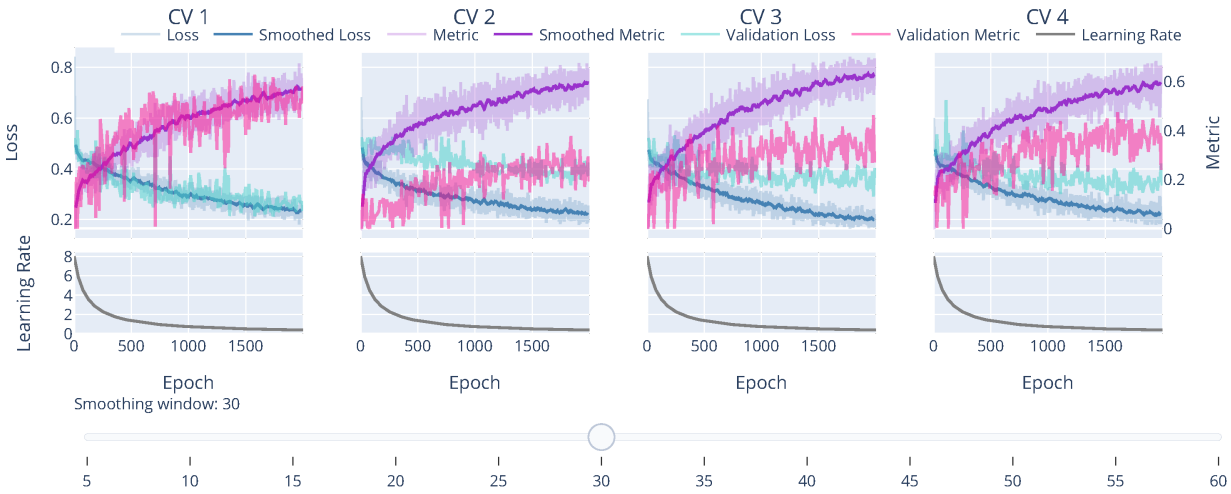


Figure 4.5: Training curves of the splits by using 4-fold cross-validation with real (low opacity) and smoothed (high opacity) training loss in blue, real validation loss in cyan, real (low opacity) and smoothed (high opacity) training metric in purple and real validation metric in pink.

This large variance could be due to the fact that the nature of the test data from split 2 is poorly conditioned. However, if the training is reproduced with the same splits, there will be a large

variance between the individual trainings. This means that the training is quite unstable. To investigate which factors in training lead to this instability, we analysed all random operations.

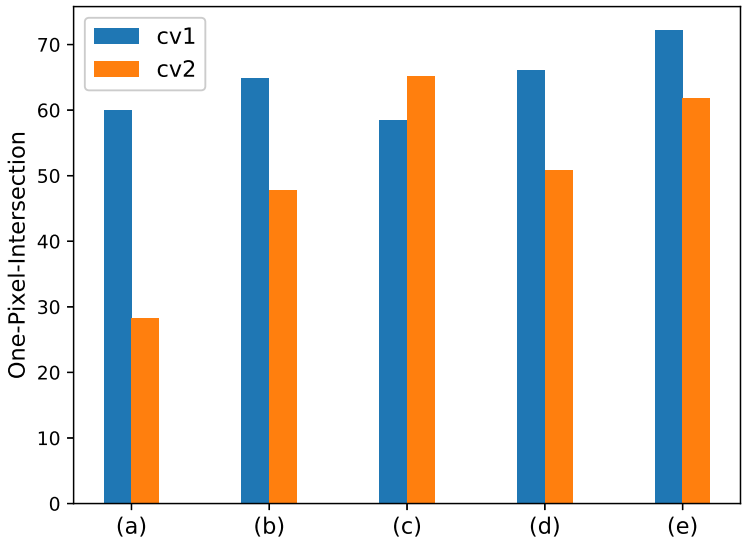


Figure 4.6: OPI of different trainings with different representations of the data from Transfer 1 and Transfer 2 (split 1 in blue and split 2 in orange)

Experiment	(a)	(b)	(c)	(d)	(e)
Annotation	T1	T1 (corrected)	T1 (corrected)	T1 (corrected)	T2
Training	×	×	×	–	×
Evaluation	×	×	–	–	×
$\overline{\text{OPI}}$	44.17	56.35	61.83	58.49	67.01

Table 4.2: Data setup and corresponding mean OPI over the two splits of the experiments from Figure 4.6. The annotation row indicates whether the data are from Transfer 1 with or without corrected labels or Transfer 2. If comorbidity cases were used for training and/ evaluation it is marked with a cross.

In training, there are three factors that introduce randomness: network initialization, patch selection and data augmentation. In the case of poor network initialization, it should be possible to stabilize the training by adjusting the choice of hyperparameters, especially the learning rate. We can exclude data augmentation as the disturbing factor, because the same behavior occurs also without data augmentation. It is much more likely that the choice of patches strongly affects the stability of the training. First, we have seen that the appearance of

pulmonary embolisms is highly variable. If patches are randomly selected in which embolisms are poorly visible even to humans, this may have a negative effect on the training. In addition, if some annotations are missing, i.e., the network receives a patch with a pulmonary embolism but it is not labeled, performance drops sharply. This makes the training very unstable and hugely dependent on the quality of the patches. With a large amount of data, such outliers in the annotations may not have much impact, but here the amount of data is not enough to compensate.

As mentioned in Section 3.1.1 we had gone through all cases of batch 1 and corrected if necessary. Afterwards, we repeated the training with our new annotations which is shown in Figure 4.6 (b). As we can see, the performance increases on both splits. Note that the training and test cases in experiments (a) and (b) are exactly the same, only the annotations differ.

In contrast, the data in the two splits in experiments (c), (d) and (e) differ from each other and also from those in (a) and (b). In experiment (c), we included the comorbidity cases only in the training data and excluded them from the evaluation data, due to the fact that pulmonary embolisms are very difficult to detect here. In experiment (d), these cases were excluded from both training and evaluation. At last, we trained again the same network with our complete processed INHOUSE data set from Transfer 1 und Transfer 2, see Figure 3.1. Because more data are available here, comorbidity cases were not treated differently. Table 4.2 shows the data used for the different experiments from Figure 4.6 and the mean OPI of the two splits.

It can be seen that the correction of the annotations has greatly increased the performance. Also, the evaluation of the two splits no longer shows such a large difference. The exclusion of the comorbidity cases is also recommended, since these already have a large influence on the evaluation with small amounts of data. After preserving the data from the second transfer, we have enough data to compensate for outliers.

Overall, these experiments have shown that the conditioning of the data has an enormous impact on the performance and stability of the training. By processing the data in different ways, Figure 3.1, we were able to increase the mean OPI by more than 20 percentage points (pp). Despite improvement of the data, there still remains a variance between repeated trainings, due to the strong influence of the chosen patches. In order to be able to compare the experiments in the following, seeds have to be set for all random operations. Otherwise, no precise statements can be made as to whether the improvement was achieved by the chosen methods or by the random choice of the patches.

4.1.3 Across Data Sets

In order to investigate the generalizability of the individual data sets, we trained the same network separately with the three different data sets: INHOUSE, FUMPE and CADPE. After training, we evaluated each network on each data set separately. To also analyse the stability of the data sets, we always trained two splits. In addition, we trained the same experiment with three different seeds ($s \in \{0, 13, 42\}$) to investigate the scatter between different replications,

resulting in $3 \cdot 2 \cdot 3 = 18$ different networks. Each network is evaluated on its own test data set but also on the test sets of the other data sets .

Each column of Figure 4.7 contains networks that have been trained with one of the three data sets. The rows indicate the data sets on which the networks have been evaluated. The results of all networks trained with the INHOUSE data set are shown with red bars, FUMPE with green bars and CADPE with blue bars. Additionally, the first black horizontal line in each bar indicates the value of the DC, the second one indicates the OPI.

Table 4.3 contains the mean OPI metric and the mean DC of all splits and seeds. It makes sense to investigate both metrics, because as we have seen in Section 3.2, evaluating the dice score gives more information about the segmentation and evaluating the OPI gives more information about the detection ability. To ensure that the networks do not tend to predict too large regions only to get a high OPI both metrics should be considered.

Considering the individual plots, for example the evaluation on the respective own test data set (see plots on the diagonal), it is noticeable that not only with the INHOUSE data set the performance varies depending on the split and different seed. In the INHOUSE data set, the largest difference of the OPI is between training with seed 13 and seed 42 in split 1, which is around 30 pp. But also in FUMPE we have a similar strong scatter considering for example the performance difference of split 1 trained with seed 0 and seed 13. In contrast, the results when training with the CADPE data set are less variable. Here, the difference between the best and worst results is only around 10 pp. The training seems to be more stable with the CADPE data set.

The plots outside the diagonals show how well the individual networks generalize, i.e., how well they perform on the other data sets. It can be clearly seen that there is good generalization for most of the networks. For example, the INHOUSE and CADPE networks perform even better on the FUMPE data than on their own data. FUMPE on the other hand, generalizes the worst of all the networks. This is not surprising, since already in the data analysis in Section 3.1.1 we saw that FUMPE is the less variable data set, for example, it has much less peripheral embolisms than INHOUSE and CADPE. This can also be deduced from the fact that all networks perform best on the FUMPE data set. Consequently, this data set appears to be the easiest of all three data sets, but is not suitable for training as a result. In contrast, the mean performance of all three data sets is worst on the INHOUSE data set. Again, this is not surprising, as we have seen that this data set is the most challenging with the smallest embolisms on average and additional comorbidity cases. Interestingly, CADPE performs even better on the INHOUSE data set than INHOUSE itself. CADPE seems to be the most suitable for training among all data sets.

Comparing the OPI and DC metric, it is noticeable that the DC is always slightly smaller than the OPI, as the latter is a more strict metric. In general, however, the DC seems to be proportional to the OPI in most cases. For example, when training and evaluating using INHOUSE (first row first column), it is easy to see that when the OPI is large, there is also a larger DC. Some FUMPE networks have a larger difference between OPI and DC. For example, for training and evaluation with FUMPE (middle plot), for split 1 (left bar, dark

green), seed $s = 0$ and seed $s = 42$ the OPI is much larger than the DC. This is because the FUMPE networks tend to have larger predictions most of the time. We have a closer look on this in Section 4.3.3.

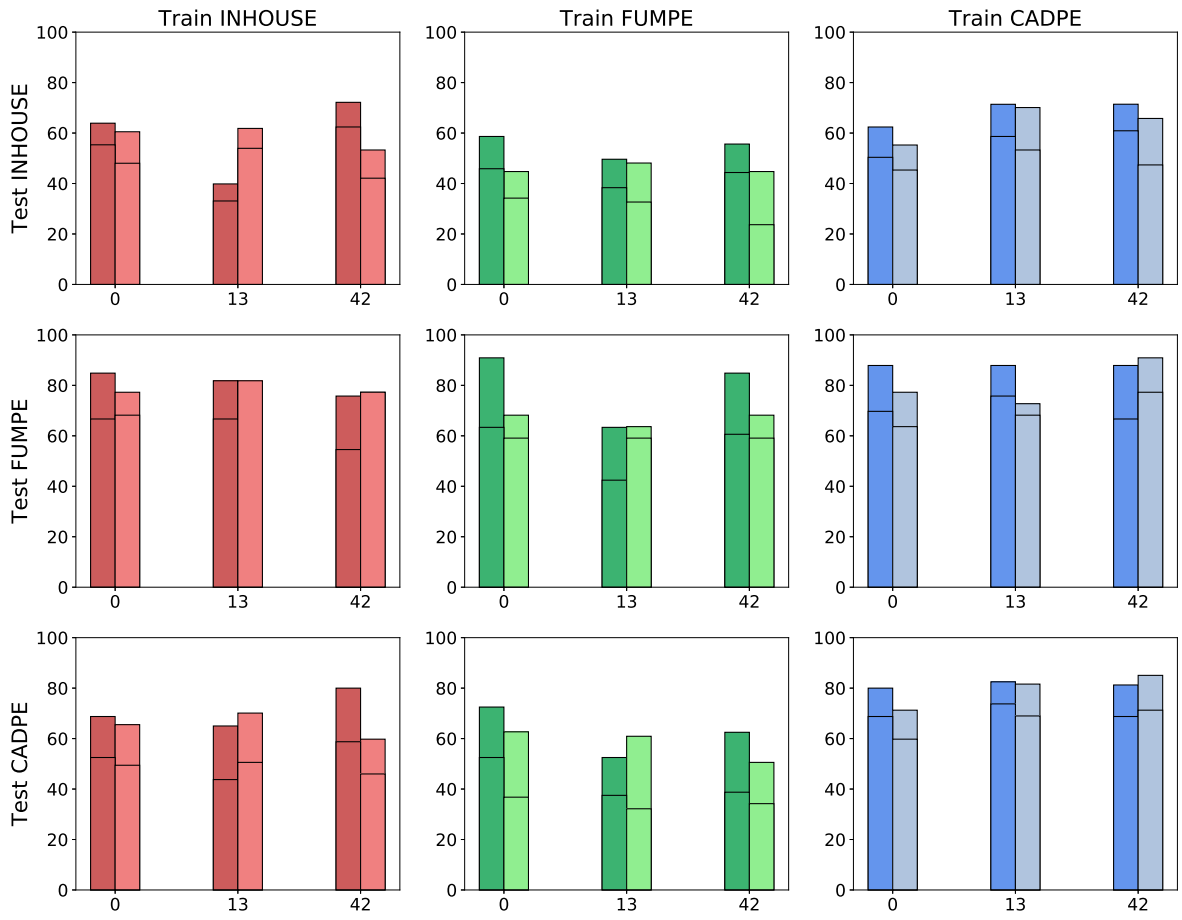


Figure 4.7: Separate training and evaluation with the INHOUSE, CADPE and FUMPE data sets. In each column of the 3×3 plots the networks were trained with the same data set, red indicated training with INHOUSE, green FUMPE and blue CADPE. The rows denote the data set on which the networks were evaluated. Note that in each row, the same test data set is used for all columns. With each data set the training was repeated with three different seeds $s \in \{0, 13, 42\}$ each for two different splits on the data set, marked with a darker color for split 1 and a lighter color for split 2. The first bar indicates the value for the DC and the second for the OPI.

data set	INHOUSE		FUMPE		CADPE	
	OPI	DC	OPI	DC	OPI	DC
INHOUSE	58.60	49.15	50.25	36.52	66.07	52.65
FUMPE	79.79	69.19	73.19	57.28	84.09	70.20
CADPE	68.19	50.16	60.28	38.65	80.28	68.55

Table 4.3: Mean OPI and DC over all seeds and both splits of each plot from Figure 4.7. The rows indicates on which data set the networks were evaluated, the columns indicates with which data set the networks were trained. On all data sets, networks trained with the CADPE data set have the highest OPI and DC on average.

In summary, generalizability depends on the nature of the data. While the networks that have been trained with the FUMPE data set with the least number of cases and also the least peripheral embolisms generalizes the worst, networks that have been trained with the CADPE data set performs the best on its own and on all other data sets. In addition, we have seen that performance is also highly dependent on the complexity of the data sets. While all networks perform well on FUMPE, the detection of embolisms is most challenging on the INHOUSE data set.

4.1.4 Combined Data Set

In this section we trained the network with a combined data set to investigate the influence of the amount of data on the performance and generalizability. Due to the incorrect annotations within the INHOUSE data set at this time, we only use the FUMPE and CADPE data sets for the combined training.

Because of the different number of images in the data sets, we need to consider whether to balance between data sets or between the individual cases. Balancing between data sets means that when the mini-batches are created, the probability of selecting cases from both data sets is equal. Thus, with 91 cases from CADPE and 35 cases from FUMPE, each case from the first data set has a weight of $w_C = \frac{91}{91+35} = 0.28$ and from the second data set $w_F = \frac{35}{91+35} = 0.72$. Another approach would be to balance between cases, meaning that each scan is equally likely and thus a CADPE case is used more often than a FUMPE case for training. In this experiment, each case would have a weight of $w = w_F = w_C = 0.5$.

It can be observed that combined training improves performance on both data sets. However, an equal distribution between the individual cases is better suited than between the data sets, because as we have already seen in the last section, the FUMPE data set does not generalize quite as well and should therefore not be weighted so heavily. In general the more data is available the better it is for training, but the weighting should be according to the quality of the data set.

w_F	1	0.72	0.64	0.57	0.5	0
w_C	0	0.28	0.36	0.43	0.5	1
CADPE	60.92	77.50	77.50	80.00	80.00	71.25
FUMPE	66.67	73.81	66.67	73.81	73.81	72.73

Table 4.4: OPI evaluation on FUMPE and CADPE test data of combined training with different weights w_F for training cases from the FUMPE data set and w_C for training cases from the CADPE data set.

4.1.5 Vessel Mask

In order to improve the sensitivity and to reduce the average number of FPs, we create the vessel masks and include them into training. The different approaches are listed in Table 4.5 and described below in detail.

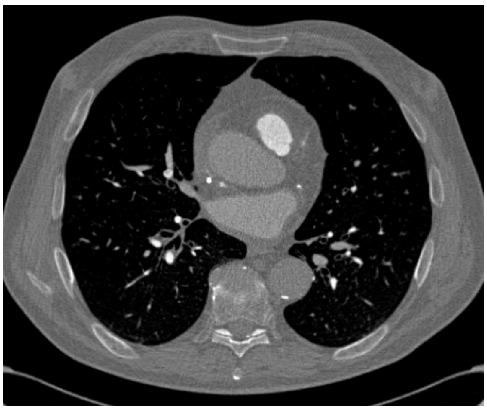
The initial idea behind this was to reduce the search space by searching for pulmonary embolisms only within the vessel mask. However, the vessel mask mainly includes only the peripheral embolisms and many proximal embolisms are not included.

Instead, we performed two other experiments using the vessel mask. One was to reduce the number of false positive predictions within the veins, to use the vein mask as a second label so that the network not only learns what an embolism appears like, but also what a healthy vein looks like. Therefore we merged the vessel mask and the embolism mask. Figure 4.9 shows an example of the resulting label mask, where the vessel has the label 2 and the embolism has the label value 1.

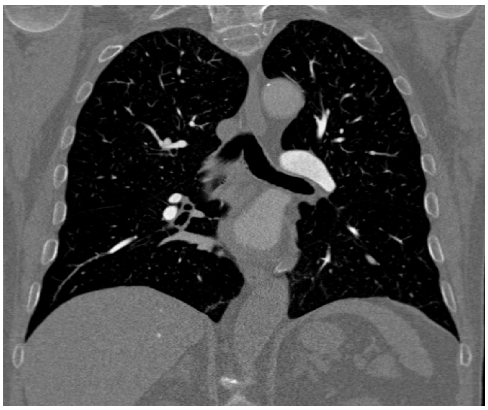
As a last strategy we add the vessel mask as an additional channel to the input images. Due to the fact that many PEs lie within the vessel mask, this should help the network to locate the blood clots. We used two approaches to do this. The first was to add the vessel mask as a binary image as a second channel. The second approach was to multiply the input image with the vessel mask. Figure 4.10 displays the binary and multiplied vessel masks used as a second input channel in the network.

To compare the different approaches, we again define seeds in all experiments so that all random processes behave the same. For comparison, we also show the same training without using the vessel mask. Table 4.5 contains the OPI of the different experiments run and evaluated once with the CADPE data set and once with the FUMPE data set.

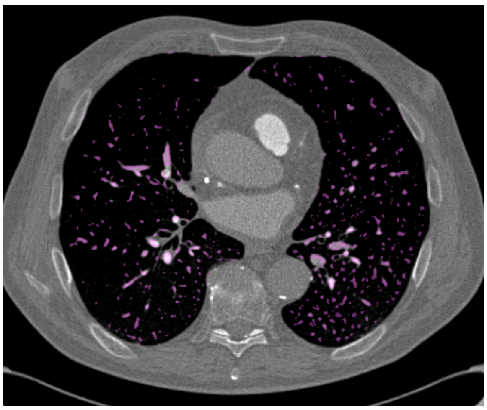
It can be inferred that the approach of taking the vessel mask as the second label slightly degrades the performance on both data sets. The network is possibly too focused on learning what a healthy vein looks like. More experiments with more complex network structures would need to be performed to gain further insights.



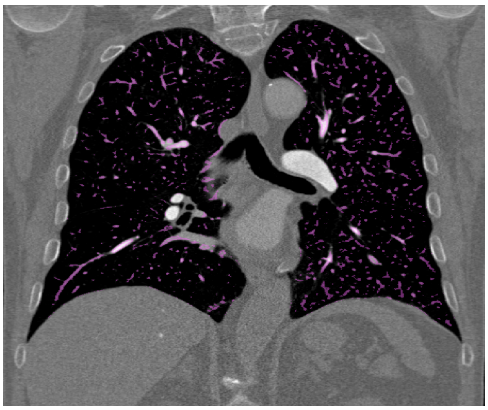
(a) Axial view



(b) Coronal view

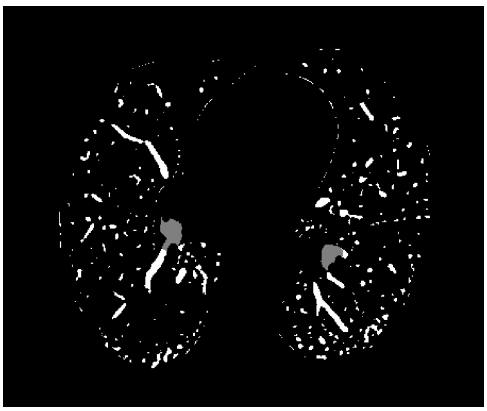


(c) Axial view with vessel mask overlay



(d) Coronal view with vessel mask overlay

Figure 4.8: CTA window (ww = 900HU, wl = 100HU) on case 4 from CADPE data set (a), (b) and with vessel mask overlay (c), (d) in red.



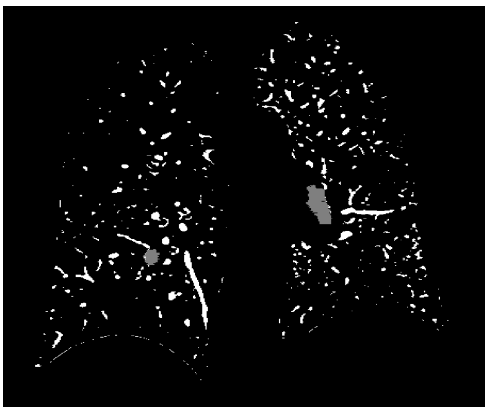


Figure 4.9: Merged label and vessel mask, axial view (left) and coronal view (right).

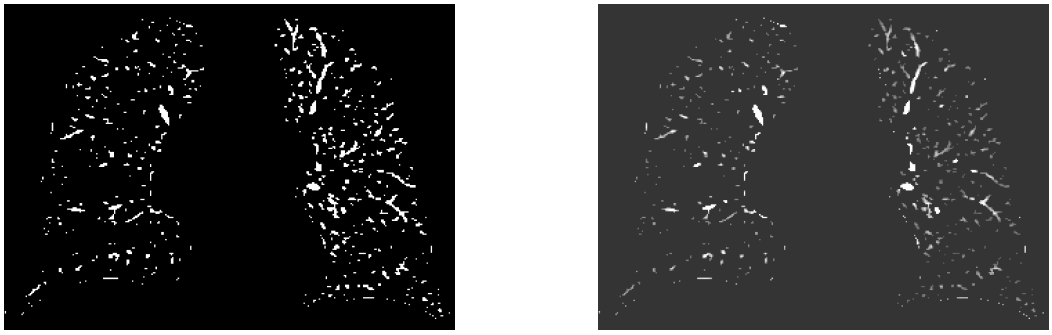


Figure 4.10: Binary (left) and multiplied (right) vessel masks, used as a second input channel of the network, coronal view

Experiment	Data set	
	CADPE	**FUMPE**
Separate training without vessel mask	71.25%	66.67%
Vessel mask as a second label	69.81%	61.91%
Binary vessel mask as additional input channel	80.00%	71.43%
Multiplied vessel mask as additional input channel	81.25%	61.90%

Table 4.5: Results of different experiments with vessel mask compared to the training without vessel mask, performed with the CADPE and FUMPE data set. As evaluation metric the OPI is used.

When using the vein mask as an additional input channel, the performance increases by 10 pp when training with the CADPE data set. On the FUMPE data set, the performance only increases when using the vessel mask as a binary image.

In summary, this section has shown that using the vein mask as a binary additional input channel can improve performance. This helps the network to localize the embolisms more accurately.

4.2 Training

In this section, different experiments regarding the training strategies are presented. At first, hyperparameters are tuned to stabilize the training and to obtain the best possible performance. After that, different data augmentation techniques are applied to increase performance and the generalization ability. First, we use classical methods like geometric or intensity methods, then we train the network with DECT representations with the objective to make the network more robust against contrast variations.

4.2.1 Hyperparameter Tuning

At the beginning of this work, different experiments in order to find the best hyperparameters were performed. As already mentioned in Section 4.1.1, wrong hyperparameters destabilize training. In addition, we have seen that the hyperparameters were not yet optimal in the first experiment, and thus the performance suffers. The right choice of hyperparameters is essential for the network to learn properly.

We have seen in Figure 4.1 that the loss has not yet gone into saturation. To make the loss converge, we increase the number of epochs from 1000 to 2000. As a result, the OPI increased by around 10 pp. As already mentioned, we cropped the images within the lung region, which highly improves the network performance too.

In order to make the training less stochastic, we increase the mini-batch size from 8 to 10, due to memory limits, the mini-batch size cannot be enlarged further when the patch size stays constant.

To investigate the impact of the hyperparameters, in all random operations the seed will be set fixed at $s = 0$. As we have already observed, the selected patches have an enormous influence on the performance which leads to a huge variance in the result. With the seed fixed, we can trace the performance change back to the modification of the hyperparameters.

As a starting point, we train two experiments with different hyperparameter settings. One with a large learning rate, a small mini-batch size, and a moderate patch size, the other with a small learning rate, a large mini-batch size and a much smaller patch size, see Table 4.6. All other remaining hyperparameters are set like in Table 4.1. The OPI from Setting 1 is about 12 pp higher than in Setting 2.

To investigate how the individual hyperparameters influence the training, we repeat the training several times by changing only one parameter. We concentrate mostly on the patch size ps and learning rate lr. The mini-batch size is set to $mbs = 10$ and the number of epochs to $ep = 2000$, as in Setting 1, see Table 4.6.

Hyperparameter		**Setting 1**	**Setting 2**	Naive Approach
Epochs	ep	2000	2000	1000
Mini-batch size	mbs	10	32	8
Learning rate	lr	10	1	8
Patch size	ps	$64 \times 64 \times 64$	$40 \times 40 \times 40$	$72 \times 72 \times 72$
OPI at 5 AFP		70.96%	58.04%	22.92%

Table 4.6: Comparison of two different hyperparameter settings and the setting from the naive approach, which differ in mini-batch size, learning rate and patch size. Note that the naive approach had been trained before we corrected the annotations of batch 1.

Figure 4.11 shows the impact of the patch size on the performance. There, the same training with different patch sizes ($ps = 32^3, 40^3, 50^3, 64^3$) was realized. This was performed twice, once with a learning rate of $lr = 4$ and once with a learning rate of $lr = 10$. In both cases, the best results were achieved with a patch size of $ps = 64^3$, showing that for detection of PEs, information about the environment is necessary.

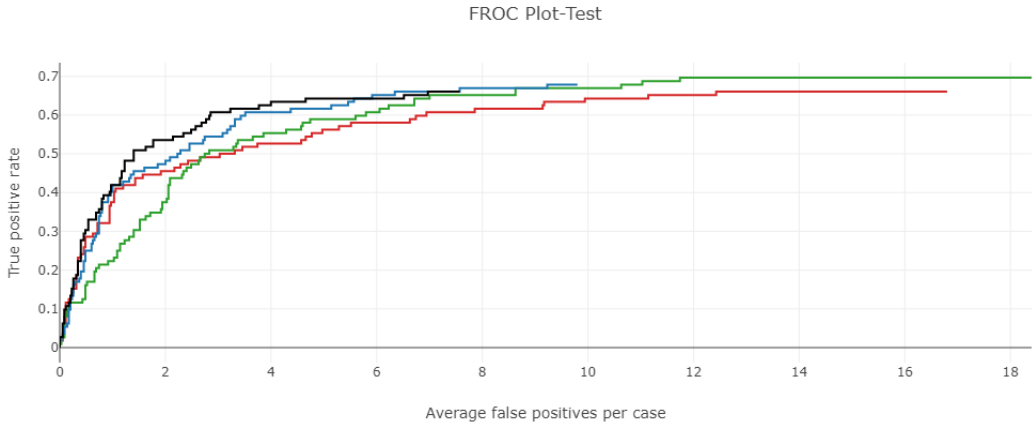


Figure 4.11: Comparison of different patch sizes, $ps = 64^3$ (black), $ps = 50^3$ (blue), $ps = 40^3$ (red), $ps = 32^3$ (green) trained with a learning rate of $lr = 4$.

While with a learning rate of $lr = 4$ at an average of 2 FP, it seems that the performance increases with larger patch size, this relation does not hold for all AFPs per case, e.g., at 4 AFP we have a higher performance with $ps = 32^3$ as with $ps = 40^3$, see Figure 4.11. At a threshold of $\tau = 0.5$ (upper left point in the FROC curve) all networks achieve a similar performance, but it is noticeable that with decreasing patch size, the AFP rate increases strongly. The same behaviour can be observed with a learning rate of 10.

In both cases at an AFP rate of 5 with patch size $ps = 64^3$ we get roughly an at least 10 pp better result than with the worst patch size. Thus, in the following experiment we will always use $ps = 64^3$ for the patch size.

Considering the average resolution of the INHOUSE data set in Table 3.1, a patch size of $ps = 32^3$ would result in a volume of $V = 32^3 \cdot 0.77^2 \cdot 0.5\,\text{mm}^3 = 9.71\,\text{cm}^3$, which corresponds to a cube with a side length of $a = 2.12$cm. That seems to be too small to get enough context information, e.g., regarding the mean volume of proximal embolisms of $5.07\,\text{cm}^3$, see Table 3.3 .

In another experiment, the influence of the learning rate is investigated. Figure 4.12 shows several FROC curves where each network is trained with a different learning rate. Considering the learning rates $lr = 1, 4, 8, 10$ it becomes clear that the performance increases with increasing learning rates. However, if the learning rate is set too high, for example to $lr = 20$, the performance drops again.

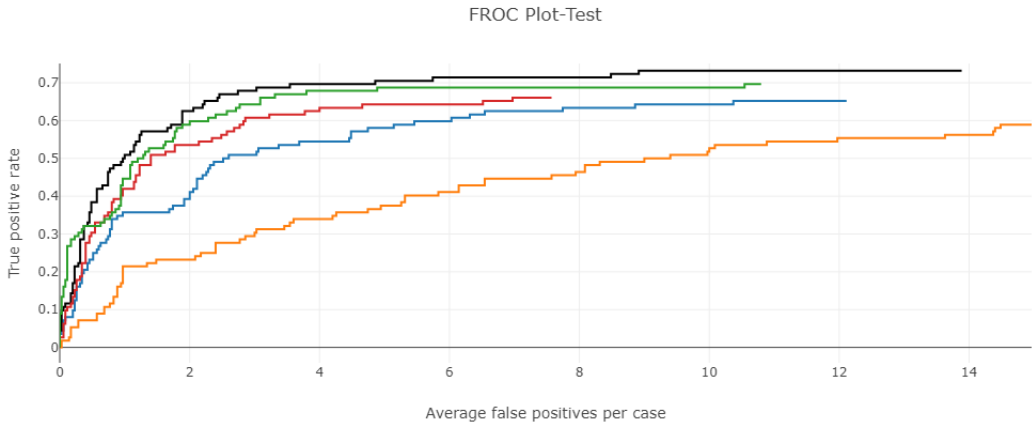


Figure 4.12: Comparison of different learning rates, $lr = 10$ (black), $lr = 8$ (green), $lr = 6$ (red), $lr = 1$ (blue), $lr = 20$ (orange).

In summary, we have seen that the choice of hyperparameters influences the network performance strongly. Several investigations show that we achieve best results with Setting 1 of Table 4.6. Thus, in the following experiments an epoch number of $ep = 2000$, a mini-batch size of $mb = 10$, a learning rate of 10 and a patch size of $ps = 64^3$ is always used. Having a good performance on the test data set is necessary before generalization ability can be investigated.

4.2.2 Classical Data Augmentation

To get more variance in the data set and thus increase generalizability, we use classical data augmentation techniques. These include geometric transformations, such as scaling, translations, rotations and reflections, intensity operations, for example gamma correction or intensity shifts. Other augmentation techniques are noise injections, like adding white noise to the images, or filtering methods, like sharpening or blurring. Figure 4.17 shows an example image once in its original form and after applying different data augmentation techniques.

We tried some of these data augmentation techniques on the CADPE and the corrected INHOUSE data set of batch 1 and investigate their influence on the training.

Figure 4.13 shows the influence of different rotation angles on the network performance. We trained once with the CADPE data set (blue lines) and once with the INHOUSE data set (red lines). For comparison, we trained completely without data augmentation techniques (see dashed lines). Afterwards, during training the images were randomly rotated with a probability of $p = 50\%$. The rotation angle followed a normal distribution, where we changed the standard deviation in different experiments. Figure 4.13 shows the OPI for four different trainings with different rotation angles $\sigma_\alpha = 5°, 10°, 20°, 30°$.

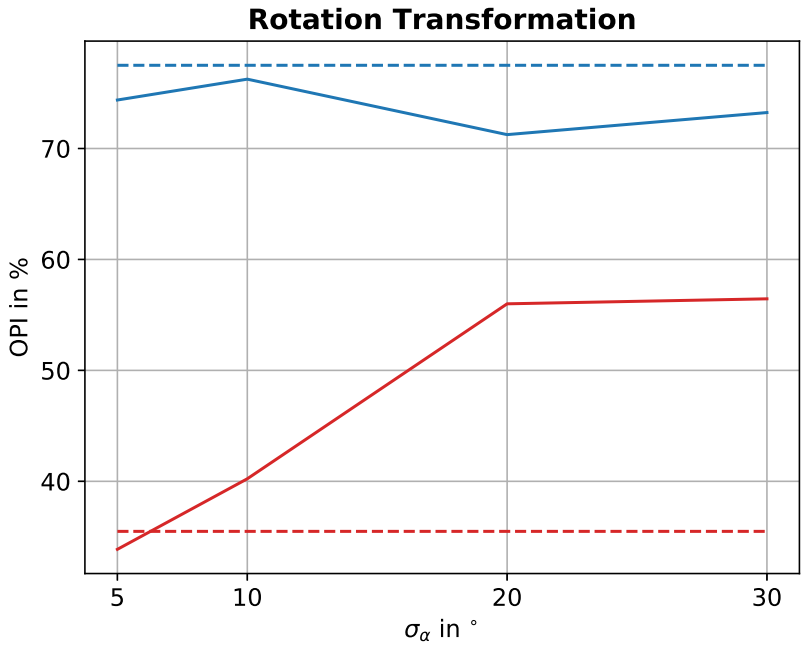


Figure 4.13: Influence of different random rotation angle on the performance (OPI), trained with the CADPE data set (blue) and INHOUSE data set (red), compared to the performance without data augmentation (dotted lines).

It can be observed that rotation shows no improvement when training with the CADPE data set. The performance here is even slightly worse, compared to training without data augmentation.

However, this may be due to the fact that the training was performed without seeds and may be related to the variability in the training. In contrast, rotation during training leads to a significant improvement in the INHOUSE data set. While with a standard deviation of $\sigma_\alpha = 5°$ the performance with an OPI of 33.87% is still very similar to the one without data augmentation (OPI = 35.49%), it increases very strongly with larger rotation angles. With a standard deviation of $\sigma_\alpha = 20°$ the performance increases by 23 pp to an OPI of 56%.

The large difference in the influence of rotation between the two data sets is probably due to the fact that the CADPE data set is with its 91 cases much larger than the INHOUSE data set. In addition, the comorbidity cases were removed, so that only 45 images were available, which corresponds to only half of the CADPE data.

Figure 4.14 shows the influence of different scaling factors on the network performance, where the scaling factor $\sigma_S = 1.25, 1.5, 1.75$ defines the standard deviation for normal sampling. When training with the CADPE data set as well as with the INHOUSE data set, small scaling factors $\sigma_S < 1.5$ do not have a great impact, ignoring small variations within the training, whereas with scaling factors $\sigma_S \geq 1.5$, this influences the training negatively. With a scaling standard deviation of $\sigma_S = 1.75$, performance drops by 8 pp from 38.71% to 30.65% for INHOUSE and 10 pp from 75% to 65% for the CADPE data set, compared to a scaling factor of $\sigma_S = 1.25$.

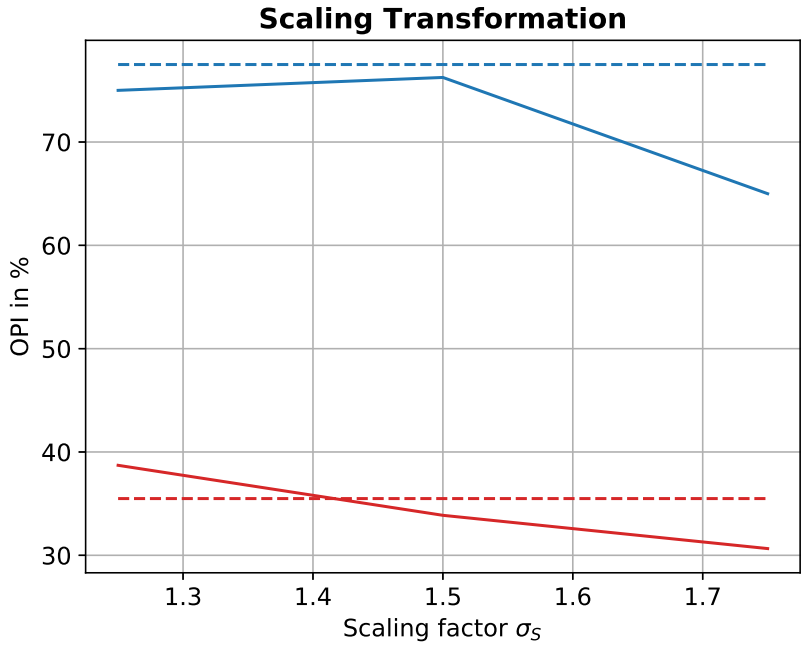


Figure 4.14: Influence of scaling on the performance, trained with the CADPE data set (blue) and with the INHOUSE data set (red), compared to the performance without data augmentation (dotted lines).

The influence of Gaussian white noise is investigated on the CADPE data set. Three networks are trained with different standard deviations $\sigma_N = 0.1, 0.2, 0.3$, which is displayed in Figure 4.15. It is clearly visible that the noise has a negative impact on the performance. Even with a standard deviation of $\sigma_N = 0.1$, the performance is 7.5 pp worse than when training without noise. At $\sigma_N = 0.3$, the performance is even 30 pp worse. In contrast to other problems, where the noise makes the networks more robust, the detection of pulmonary embolisms is very sensitive to noise.

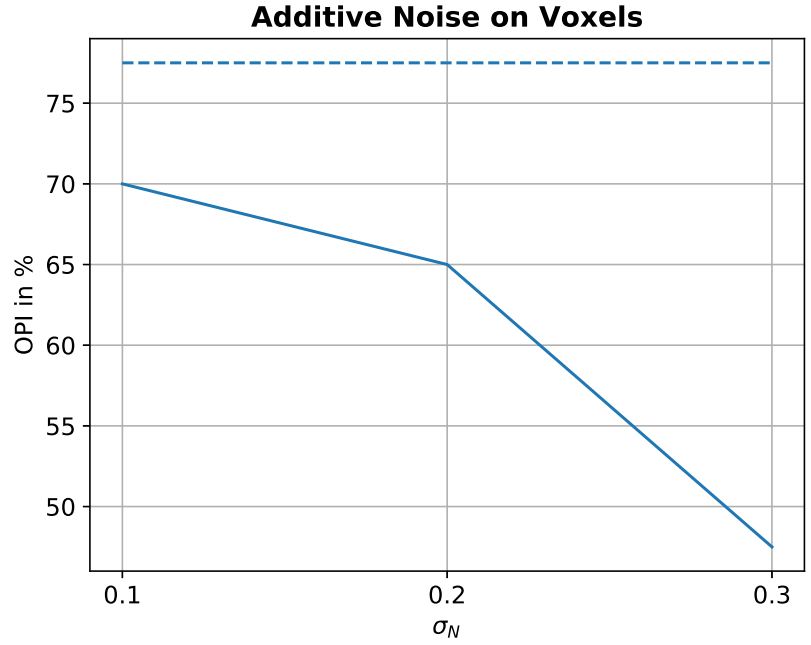


Figure 4.15: Influence of gaussian white noise on the performance, trained with CADPE, compared to the performance without data augmentation (dotted line).

As a last experiment, we investigated different intensity transformations. Figure 4.16 shows the impact of gamma distribution (blue) and intensity shifts (black) on the performance. Firstly, we trained the network where the intensities are shifted by a normally-distributed random value. Here, we tried the three different standard deviations $\sigma_I = 0.1, 0.2, 0.3$. Small intensity shifts like $\sigma_I = 0.1$ do not worsen the performance, but also do not improve it. With increasing σ_I, the performance decreases from 77.5% to 68.8%. Secondly, the intensity is changed via a gamma transformation, defined as

$$I_{out} = I_{in}^{\gamma}, \tag{4.1}$$

where the input values are raised to the power γ. With $\gamma < 1$, dark regions are strongly lightened, the entire image looks brighter, but with less contrast. With $\gamma > 1$, brighter areas are darkened. We applied the gamma transformation to the image with different values

$\gamma = 0.75, 0.95, 1.05, 1.25$. As you can see, the performance for $\gamma < 1$ decreases much more than for $\gamma > 1$. For gamma close to 1, the performance is approximately equal to the performance without data augmentation. Even for $\gamma = 1.25$ the performance decreases only slightly. However, no improvement can be achieved this way.

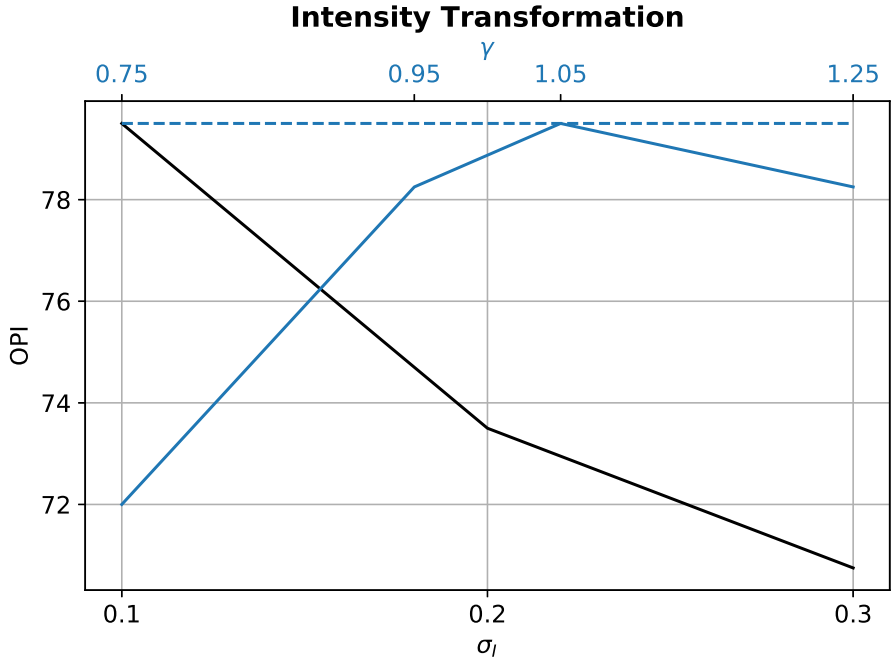
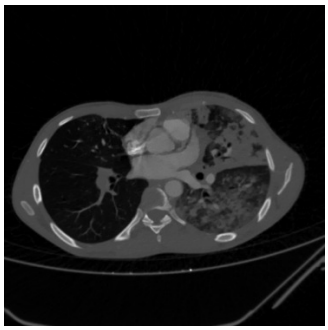


Figure 4.16: Influence of different intensity transformations, like intensity shifts (black line) and gamma transformation (blue line) on network performance, compared to training without data augmentation (dotted line) for the CADPE data set.

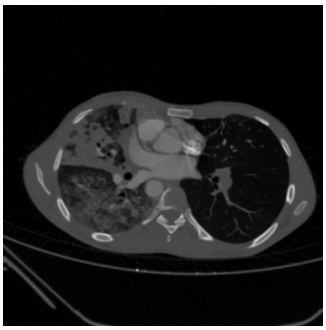
In addition, the extent to which the reflection of the images influences the training was also investigated. Here we have the same performance for both data sets as without the data augmentation. Accordingly, this does not affect the performance on the test data.

We also trained two networks with the CADPE data set, where we randomly blurred the images with a gaussian filter with standard deviation $\sigma_B = 0.1, 0.2$. This also did not have an impact on performance.

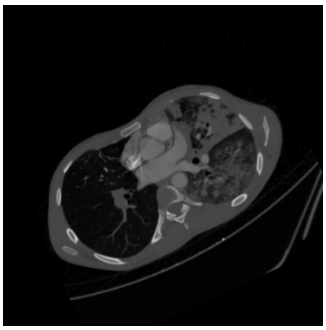
In summary, different data augmentation techniques were investigated. By examining the two different data sets, it became very clear that data augmentation is particularly useful when there is few data available. When there is already a lot of data available, as with the CADPE data set, data augmentation does not necessarily harm performance, but also does not provide a significant improvement. In conclusion, the following data augmentation methods are recommended: Rotation with standard deviation $\sigma_\alpha \leq 30°$, scaling with factors $\sigma_S < 1.5$, small intensity transformations like intensity shifts with $\sigma_I = 0.1$ or gamma transformations with gamma close to 1, flipping and Gaussian blurring with $\sigma_B < 0.2$.



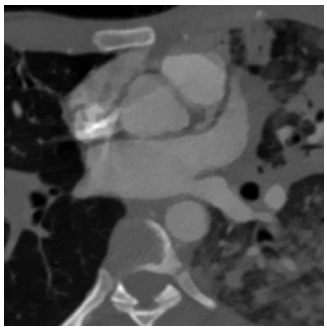
(a) Original image



(b) Flipping

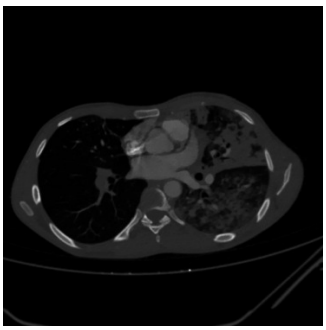


(c) Rotation



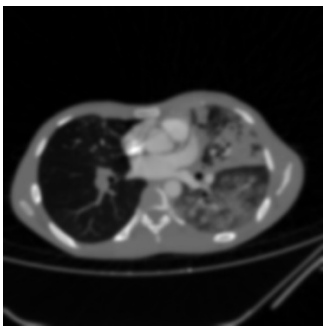
(d) Scaling

(e) Intensity Shift



(f) Gamma $\gamma = 1.5$

(g) Gamma $\gamma = 0.5$



(h) Blurring

(i) Sharpening

(j) Gaussian Noise

Figure 4.17: Original image and transformed images after applying different classical data augmentation techniques.

4.2.3 Dual-Energy CT Augmentation

In this section, the generalization of the already trained networks to images with different contrasts is investigated. For this purpose, we create different contrasts using the DECT representations by generating monoenergetic images with different energy levels. As described in Section 2.3, the higher the energy level, the lower the contrast.

Figure 4.18 shows conventional CT image and monoenergetic images with energy levels $e \in \{50\,\text{keV}, 70\,\text{keV}, 100\,\text{keV}, 150\,\text{keV}, 200\,\text{keV}\}$.

To improve generalizability, we randomly generated monoenergetic images at different energy levels on the fly during training, thus realizing online data augmentation with DECT data. These were generated uniformly distributed between different energy levels within the interval $[e_{min}, e_{max}]$. Here we used the following intervals $I_1 = [60\,\text{keV}, 80\,\text{keV}]$, $I_2 = [50\,\text{keV}, 100\,\text{keV}]$, $I_3 = [50\,\text{keV}, 150\,\text{keV}]$ and $I_4 = [50\,\text{keV}, 200\,\text{keV}]$. While training with interval I_1 produces images similar to the conventional ones, intervals I_2, I_3 and I_4 produce increasingly low-contrast images. All networks were trained once without and once with data augmentation. As our classical data augmentation technique, we used rotation with a normally distributed rotation angle with standard deviation $\sigma_\alpha = 10°$ in x and z direction, flipping along the x-axis, and a random scaling following a normal distribution with standard deviation $\sigma_S = 1.25$.

To allow for comparison of the networks, the seed of all random operations was set to zero for all networks trained with the INHOUSE data set. Thus, each networks receives exactly the same patches, only the contrast differs. In addition, we also tested how well the networks trained with the FUMPE and CADPE data set perform on the monoenergetic data. For this we used the best networks from Section 4.1.3.

Considering the networks trained without DECT data augmentation, see Figure 4.18 CADPE (blue), FUMPE (green) and INHOUSE (red), one can see that the performance of CADPE and INHOUSE is very good on the conventional data. Since the monoenergetic images with 70 keV are of the same contrast as the conventional images, the performance here is equivalent. The performance on the monoE50 images is mostly quite similar. On the one hand, one would assume that the performance is improved, since the contrast is better, but also the noise in the image is increased, which can in turn have a negative effect. CADPE and INHOUSE get slightly worse here, FUMPE increases a bit. FUMPE has poor generalization capability, which we have already seen in Section 4.1.3.

As the energy level increases, the performance drops very sharply. For the monoE200 images, the performance of INHOUSE and CADPE drops by two thirds compared to the monoE70 images from 55.34% and 59.33% to 20% and 14.67%.

This shows that the networks perform only on similar contrasts and do not work on CTs acquired without a contrast agent.

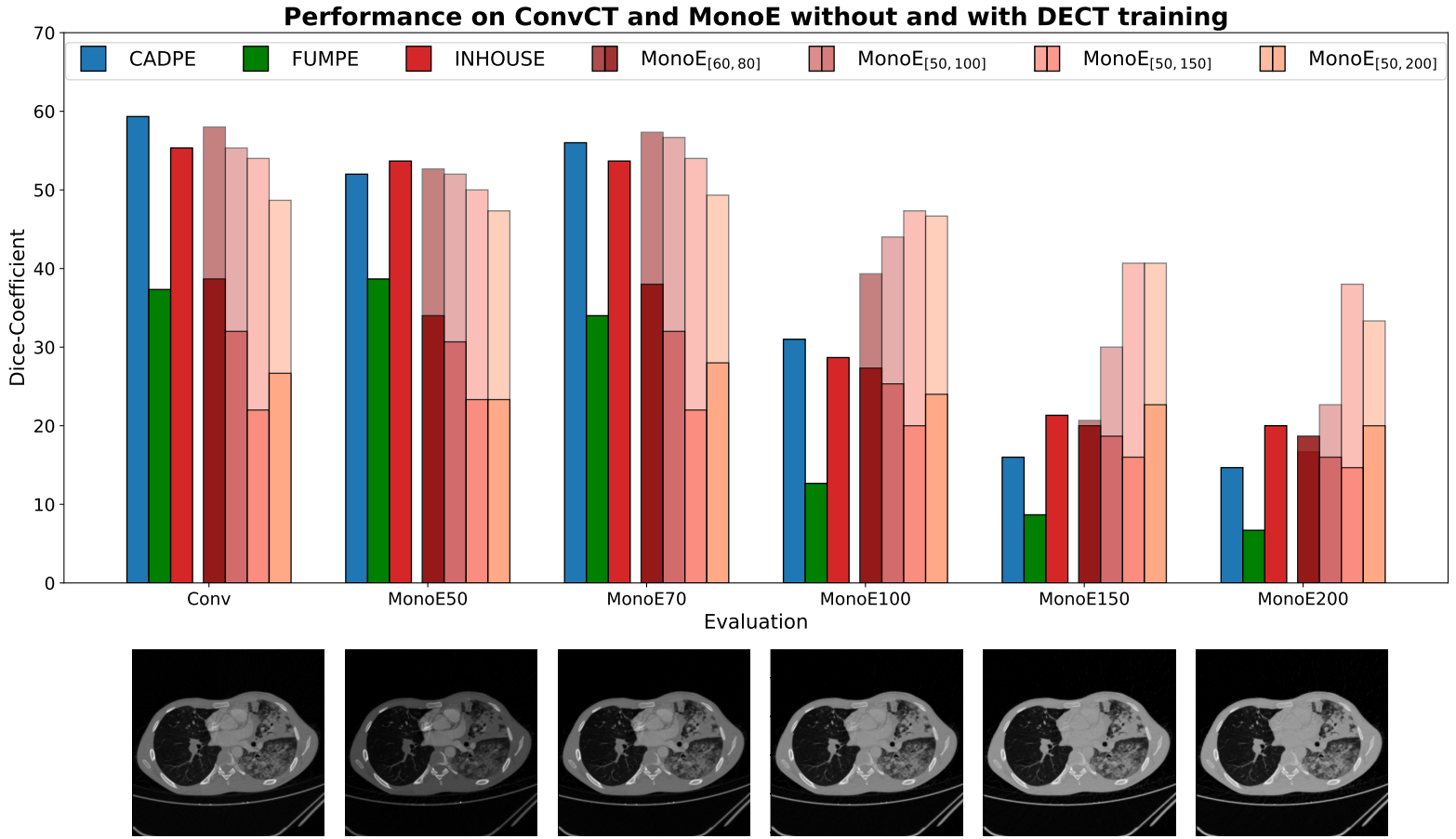


Figure 4.18: Performance comparison on conventional and monoenergetic images with networks trained with and without DECT data augmentation. The DC is determined for a network trained separately with the CADPE (blue), FUMPE (green) and INHOUSE (red) data set without DECT, but with classical data augmentation. Furthermore, networks with DECT data augmentation are trained with (brighter bar) and without (darker bar) additional classical data augmentation. The energy range increases (red color gradient), from low to high contrast variations.

To improve generalizability, we trained networks with monoenergetic data. These were trained once without additional classical data augmentation (darker bar) and once with classical data augmentation (lighter bar), see Figure 4.18. Without classical and only with DECT data augmentation, the performance is already very poor on the conventional images. This could be due to the fact that the network concentrates more on detecting the different contrasts and thus the filters generally become simpler, i.e., contrast-independent edge detectors are formed which generalize poorly.

Beside the generally poor performance, two effects can be seen. First, the network trained with the interval I_1, i.e., with a lower contrast deviation from the conventional images, has the highest DC on the conventional data. On the other hand, the performance decreases the most with decreasing energy level, with a reduction factor of 2. In contrast, the DC for the training with the largest contrast variation is 12 pp lower on the conventional data, but decreases only by a factor of 1.35.

For DECT training with additional classical data augmentation, the performance increases strongly. This shows that DECT data augmentation is only promising in combination with classical data augmentation. In order for the network to focus not only on the contrast variation, but also on the complexity of the actual problem, a large variance in the data is essential.

Considering the network trained with the energy interval I_1, the performance on the conventional data is even increased by 4 pp, compared to the performance of the network trained only with conventional images.

Additionally, we can see that the performance on the conventional, monoE50 and monoE70 data decreases slightly with increasing contrast variation within the training. However, evaluating on images with an energy level of 100 keV onwards, this effect seems to be reversed and the performance increases with a larger energy range in training. At 100 keV, training with the interval I_3 results in a 10 pp higher DC compared to the network trained with the interval I_1, at 150 keV and 200 keV the performance is even 20 pp larger. It is noticeable that the performance with the network trained with I_3 was slightly better on the monoE200 data, even though it was extended only to 150 keV, in contrast to the performance with I_4, whose interval went to 200 keV. It is probably more difficult to extract information from the images with high energy levels, since the pulmonary embolisms are difficult to see at such poor contrasts.

Overall, two statements can be made. DECT Augmentation with slight contrast variation improves the performance on the conventional data with a small improvement on different contrasts. DECT augmentation with stronger contrast variation gives very good generalization on different contrasts. Thus, DECT data augmentation is highly recommended to obtain a contrast-independent prediction.

4.3 Evaluation

Although in the previous sections the networks have already been evaluated and compared, in this section we will examine some evaluations in more detail and perform additional analysis to gain an even better understanding of the characteristics of the trained networks.

4.3.1 Healthy Cases

First and foremost, we want to investigate how our trained networks respond to healthy cases. It would be desirable that they do not segment false positives on healthy CT images at all. However, since the networks have some false positive predictions and have not yet been trained with the healthy cases, the first expectation would be that the number of false positives would be much lower on the healthy cases than on the diseased cases.

To examine the relationship between the average false positives on the healthy cases and the average false positives on the diseased cases, we plotted the former as a function of the latter in Figure 4.19. For each value of the respective number of false positive predictions on the sick cases, the corresponding threshold τ was taken and for this threshold the amount of false positives on the healthy cases was calculated and plotted.

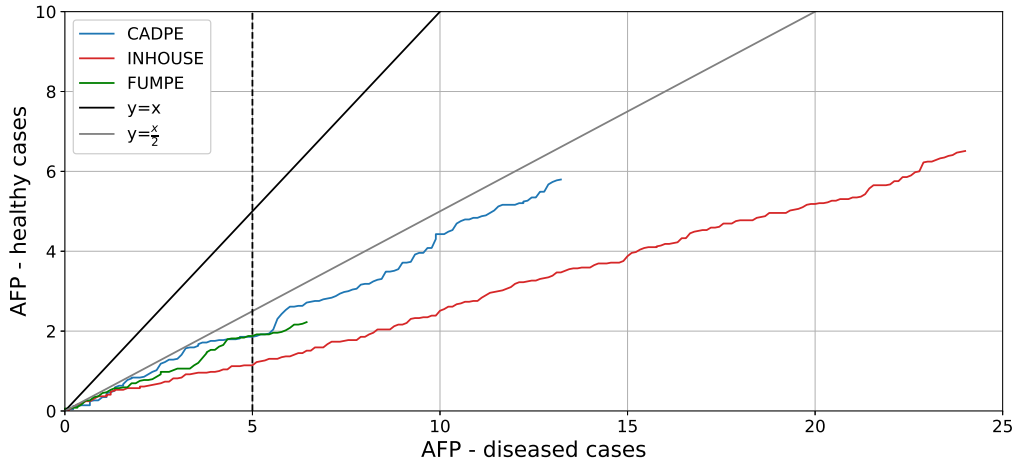


Figure 4.19: AFP of healthy cases against AFP of diseased cases.

It can be seen that for all three networks, the number of false positives on the healthy images is significantly lower than on the diseased images. With a mean false positive rate of AFP = 5 (dotted line) on the diseased cases, CADPE has 1.86, FUMPE 1.88 and INHOUSE only 1.14 false positives on average. Thus, all curves are clearly below the bisector.

Furthermore, the amount of healthy cases was determined, where the network predicts no segmentations at all. These make up 34.69% of the healthy cases for FUMPE, 36.73% for

CADPE and 51.02% for INHOUSE. This means that although we have much less false positive predictions on the healthy cases overall, we still have little cases where not a single embolus is detected at all. This is an undesirable result, since all cases where at least one embolus is segmented would have to be manually checked. For clinical usage, the number of casewise FP predictions has to be further reduced.

In summary, the fact that the networks have much less false positives on the healthy cases than on the diseased images is a positive, but the number of false positives is still too high. To reduce this further, the healthy cases could also be included in the training. Another suggestion would be to retrain the network and to sample more often at the locations of the false positives.

4.3.2 Evaluation on Peripheral and Proximal Embolisms

We have only seen the performance on the whole test data set. Another interesting point would be to investigate to what extent the performance differs on the peripheral and the proximal emboli. Therefore, we evaluated these separately, see Figure 4.20.

Surprisingly, all FUMPE and CADPE networks show similar performance on both proximal and peripheral emboli. With different seeds, the overall performance on both types of thrombi varies. For the INHOUSE data set, however, the difference in performance between the peripheral and proximal embolisms varies strongly. While training with seed 42 achieves good performance on both types of emboli, training with seed 13 performs three times worse on the peripheral data than on the proximal data. Only the performance on the peripheral embolisms varies significantly, on the proximal thrombi the OPI is quite similar in all cases.

This means that the strong variance in training with the INHOUSE data set is related to the detection of the peripheral emboli. This is strongly dependent on the patches used for training.

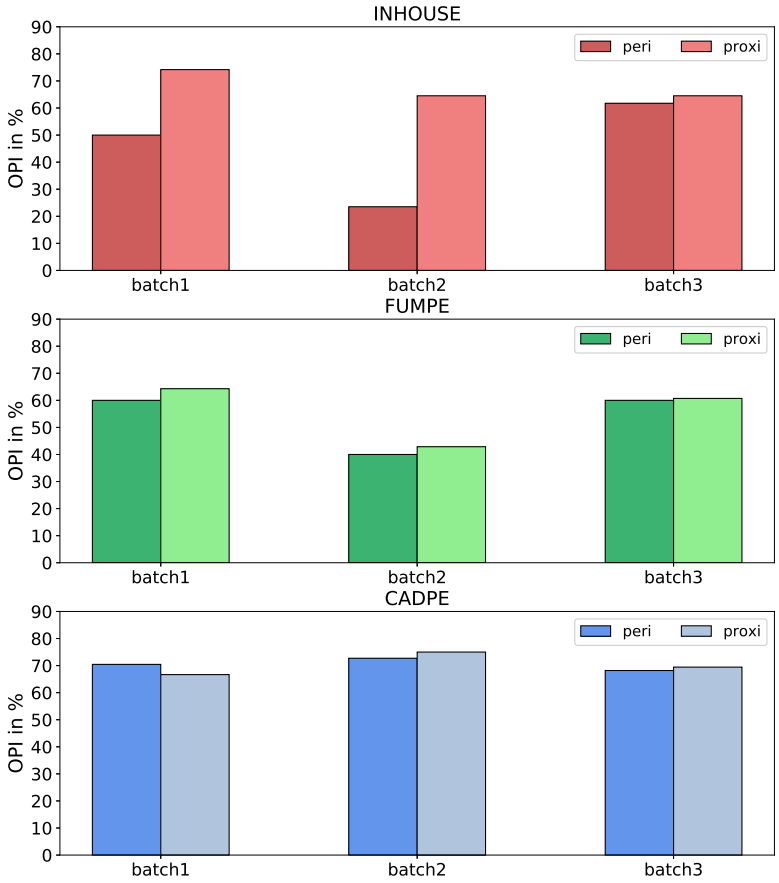
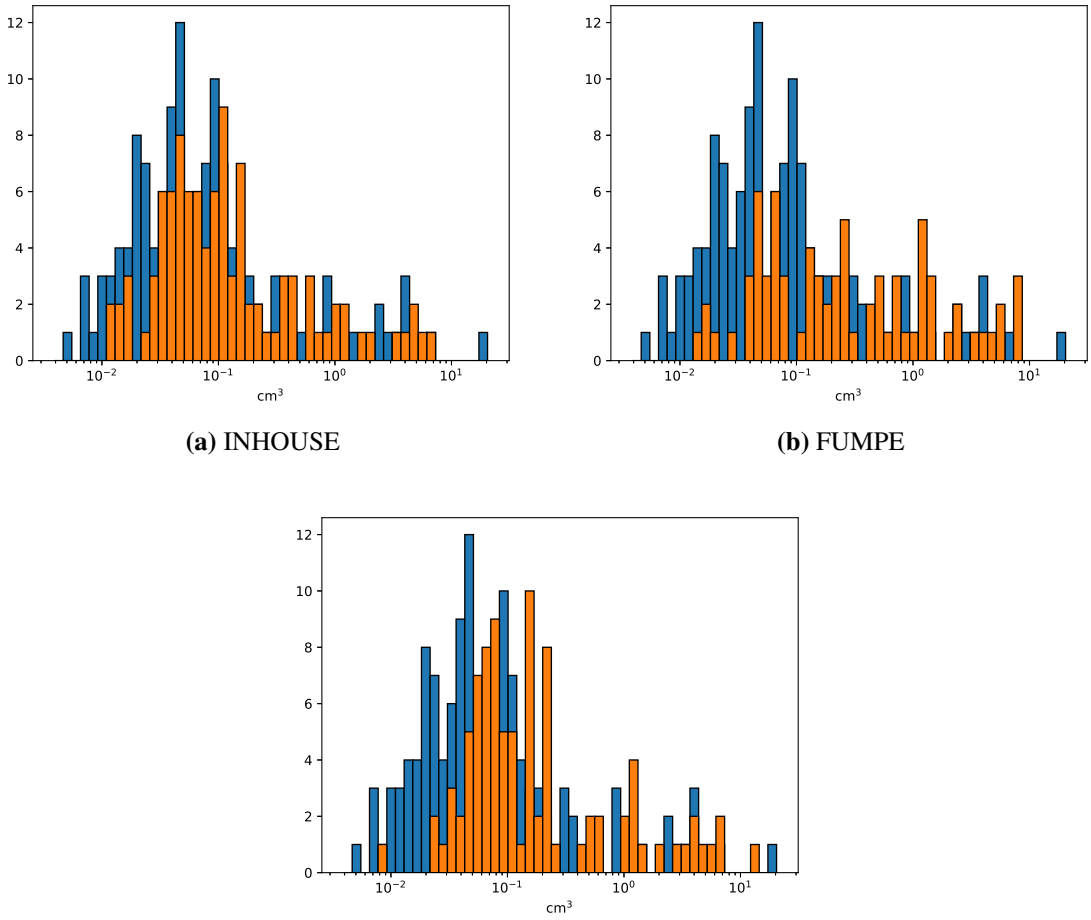


Figure 4.20: Separate evaluation on peripheral and proximal embolisms. Evaluation of the networks from Figure 4.7, trained with cross-validation split 1, and different batches on the respective test data set. Here, the evaluation on peripheral embolisms (darker left bar) and on proximal embolisms (brighter, right bar) is considered individually. The red bars represent training with the INHOUSE data set, green with FUMPE and blue with CADPE.

4.3.3 Segmentations of Separate Trained Networks

In this subsection we take a closer look at the segmentations of the separately trained networks from Section 4.1.3. First we will investigate the segmentations of the networks trained with the different data sets on the INHOUSE data set. Then we will show the segmentations of the individual networks on their own data set with render plots.

The histograms in Figure 4.21 show the volume of the ground truth embolism in blue and the respective volume of the segmented connected components of the networks trained separately with the INHOUSE, FUMPE and CADPE network in orange. It can be seen that all networks have more difficulties to detect smaller embolisms with a volume below 0.03cm^3. This would correspond to a cube-shaped embolus with an edge length of 3mm. The predicted volumes of INHOUSE and CADPE correlate better with the ground truth values than those of FUMPE.



(a) INHOUSE

(b) FUMPE

(c) CADPE

Figure 4.21: Histograms of ground truth embolus volume of the INHOUSE data set (blue) and connected component embolus volume (orange) in cm^3 predicted by a network at an AFP of 5, trained separately with the INHOUSE (upper left), FUMPE (upper right) and CADPE (lower) data sets.

Figure 4.22 shows renderings to visualize the segmentations of the networks. We selected four cases from the own test data set with the best performance. These give a visual insight of the segmentation characteristics of the networks. It can be observed that the connected components of the CADPE networks match accurately with the ground truth data, while

for example the FUMPE network exhibits a higher false positive rate. It is noticeable that the embolisms of the FUMPE network are mostly larger than the reference components. This could be due to the fact that FUMPE has the most proximal embolisms and tends to overestimates the area of the embolus.

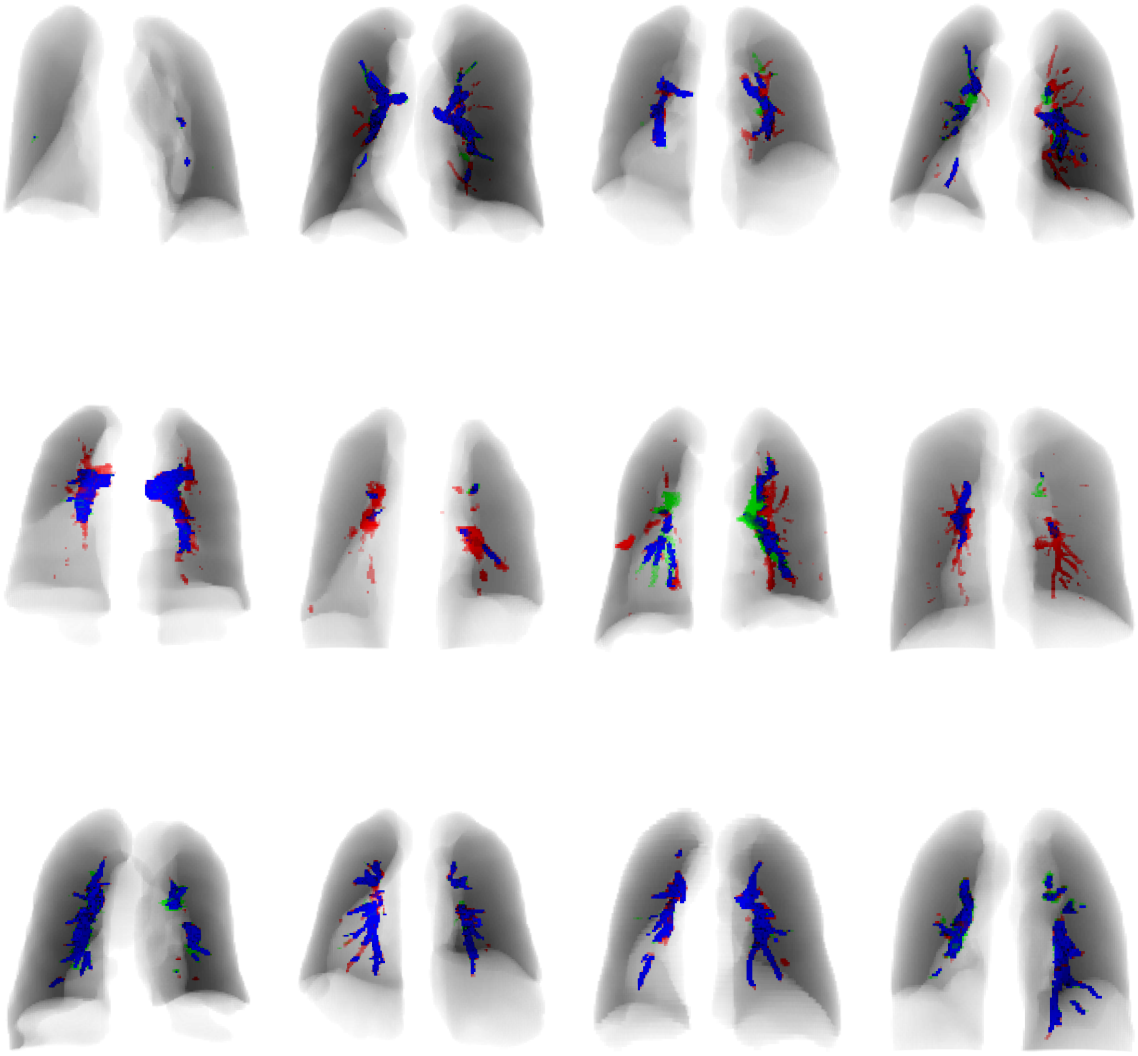


Figure 4.22: Renderings with lung mask (gray values), FN pixels (green), TP pixels (blue) and FP pixels (red) of the INHOUSE (first row), FUMPE (second row) and CADPE (third row) network.

4.3.4 FP and FN Examples

In this subsection, we take a closer look at where the network fails. That means, on the one hand, where the network makes predictions that are actually not embolisms (FP) and, on the other hand, which embolisms the network does not detect (FN).

To do this, we manually examine these cases in the test data. Considering the false positives, three different types stand out. First there exist cases where the predictions make no sense because they do not even lie within the ROI. Figure 4.23 displays some examples where pulmonary embolisms are predicted that are not located within the pulmonary trunk (left and right images). In the left image, the prediction seems to be located within the vena cava superior (VCS) and in the right image, it lies between the esophagus, trachea and left main pulmonary artery. In the middle image, the prediction lies outside the body. These examples show that the context information in the single patches is not sufficient to avoid predictions in such nonsensical areas. The network would either have to receive more global information in order to learn in which areas the embolisms are located or would have to sample even more often over the individual patches where the false positive predictions are located so that it learns the local difference even more precisely.

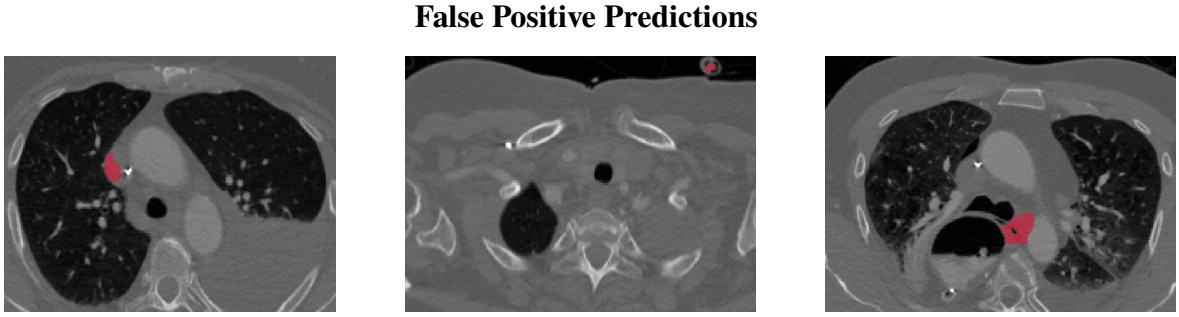


Figure 4.23: Example cases where the predictions does not lie inside of the ROI. FP marked in red lie in VCS (left image), outside the human body (middle image) or between the esophagus, trachea and left main pulmonary artery.

Fortunately, the network only suffers from a few of these predictions outside the ROI. It is much more frequently the case that the network segments within the pulmonary arteries, mainly in the small peripheral arteries. On closer examination, we noticed that many of the FP predictions are not false positives at all, but overlooked pulmonary emboli. Figure 4.24 shows three examples of false positives from our test data set which we would mark as embolisms.

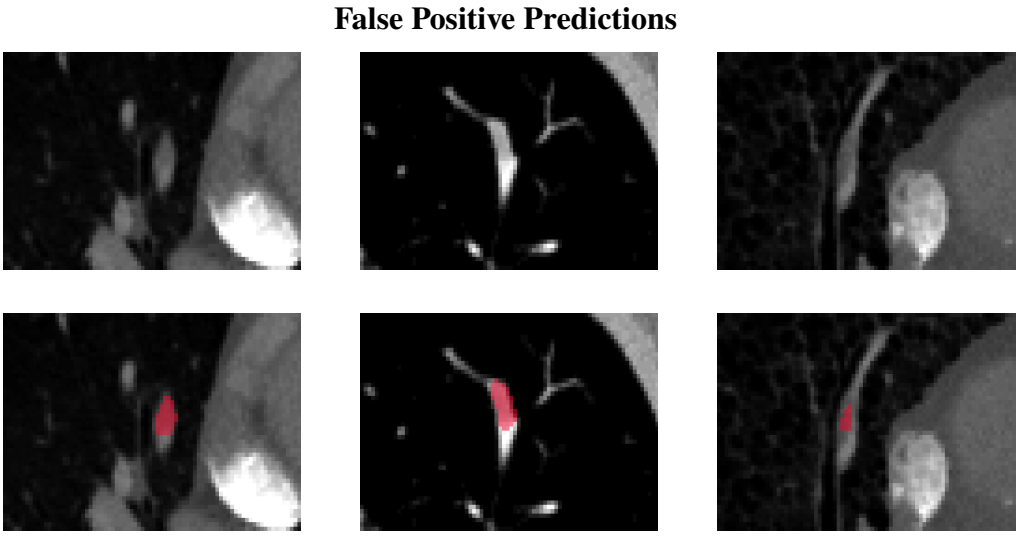
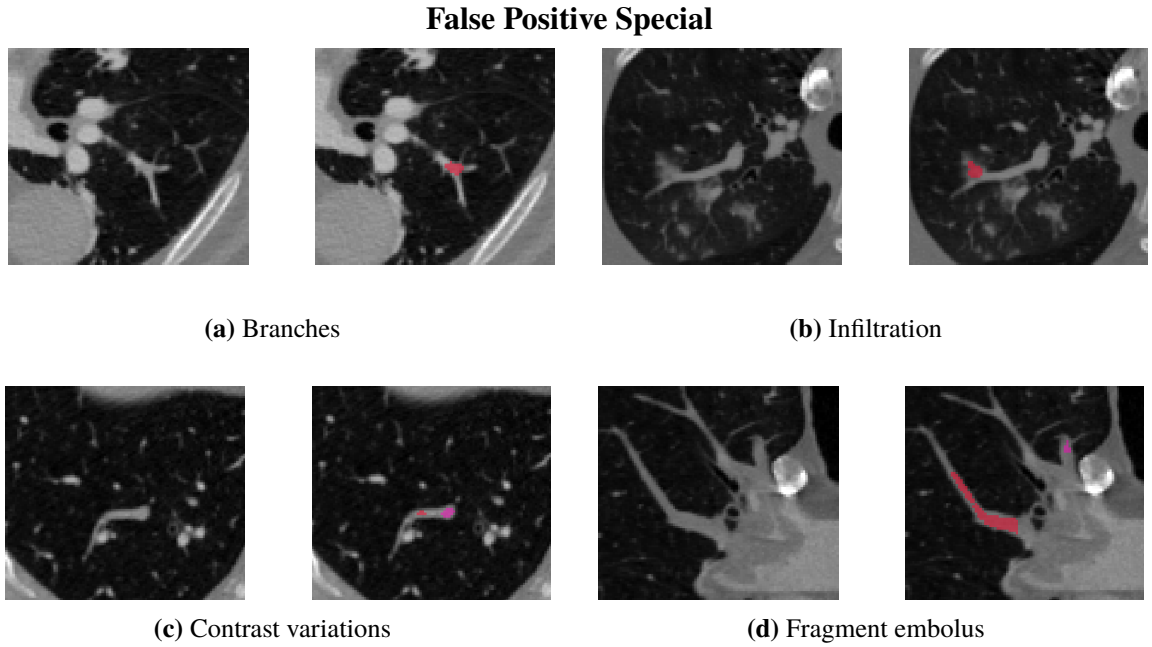


Figure 4.24: Examples without (upper row) and with (lower row) network prediction overlay (red), where the detected regions count as FP segmentations, but seem to be real embolisms.



(a) Branches **(b)** Infiltration

(c) Contrast variations **(d)** Fragment embolus

Figure 4.25: Examples without (left) and with (right) network predictions overlay (red) that contain special cases of FP predictions, like segmentations of branches (a) where a small intensity drop is visible, of surrounding comorbidities like infiltration (b), small contrast variation within peripheral veins (c) and fragments of an embolus which are not included in the annotations mask (d).

Figure 4.25 shows different special cases of FP segmentation. Common mistakes include segmentation of branches (4.25a), surrounding infiltrates (4.25b), or small contrast variations in peripheral arteries (4.25c) that are all no emboli.

One special case that can easily be omitted is displayed in Figure 4.25d. Here the two red regions are counted as FPs, but obviously they are fragments of the huge proximal embolism whose label was not drawn broadly enough. This can easily be avoided by a subsequent post-processing.

Last but not least, we examine in more detail the embolisms that the network did not detect, i.e., the FN predictions. We compared the best INHOUSE network with the best CADPE network, where INHOUSE failed to detect a total of 25 embolisms and CADPE had 26 false negatives. The intersection of both false negatives consists of 18 emboli, which we analysed manually in more detail. Of these 18, we found only 6 that were clearly recognizable to us as embolism. We would not recognize the remaining 12.

Figures 4.26 and 4.27 contain several examples of FN examples. The left image of Figure 4.26 shows one peripheral embolism, which is a borderline case because the artery is so thin that it is difficult to make a statement here. Right next to it, a proximal thrombus was not detected, but it looks like a false label because it lies between the arterial branches. The two right images contain clearly recognizable embolisms, which are not detected, where the rightmost embolism seems to be a fragment of the embolism below.

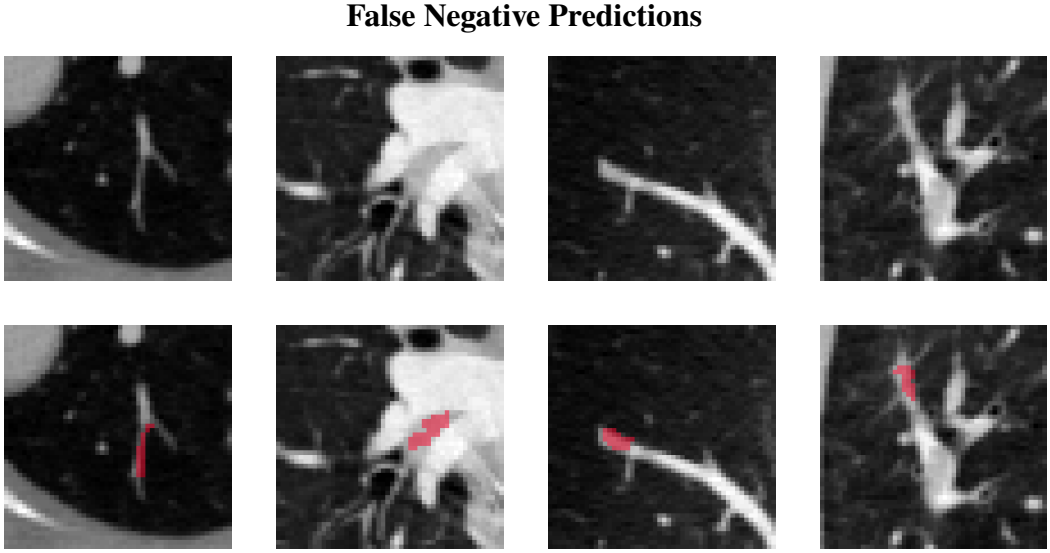


Figure 4.26: Example images without (upper row) and with (lower row) ground truth overlay, which were not detected by the network, like border cases in small peripheral arteries (most left), regions that were wrongly annotated (left), embolisms with low contrasts (right) or not detected fragments of other embolisms (most right).

Figure 4.27 contains FN examples that we would not label as embolism. The two leftmost images show comorbidity cases, where in the leftmost image, not even the artery is visible. But also in the two rightmost images, a thrombus is not visible.

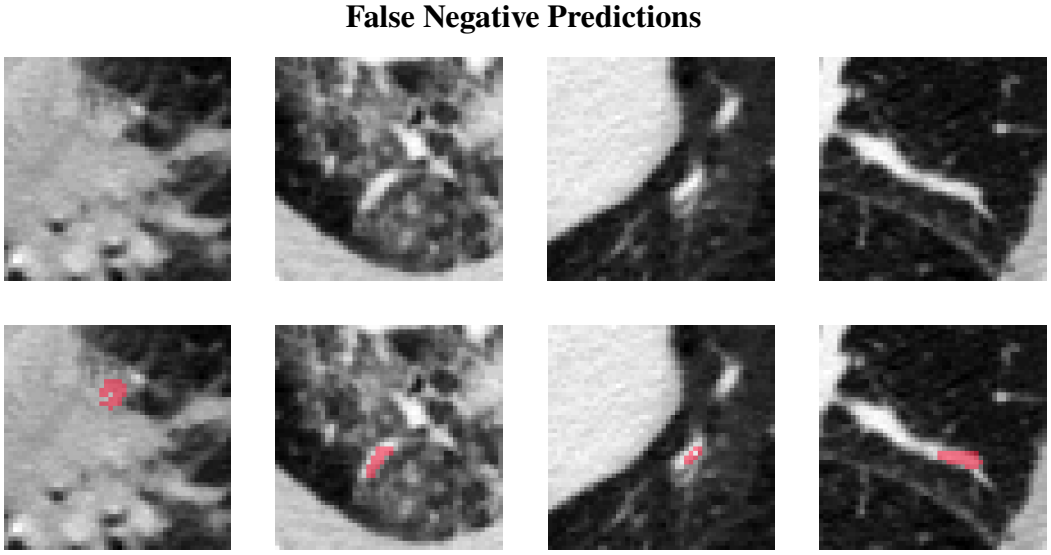


Figure 4.27: Example images without (upper row) and with (lower row) ground truth overlay that were not detected by the network that we would not annotate. Both in the two leftmost images which are strongly affected by comorbidities and in the better visible rightmost images, no embolisms are recognizable to us, despite the annotations.

In summary, we see that the network still makes many false positive segmentations. However, only a portion of these are predictions that do not make sense because they are outside the ROI. This should be corrected in future work. However, many other FP predictions have detected further emboli. Borderline cases, such as detected tissue or branching, are common errors that are very difficult to correct.

In addition, we have also seen that many labels are still not correct, or represent borderline cases. Segmentations that are not visible to humans cannot be recognized by the network either, as is the situation with comorbidity cases, for example. Regarding the comorbidity cases, it should be noted that INHOUSE segmented far fewer false positives on these. CADPE detected 40 false positives on one case, whereas INHOUSE segmented only 9. This is probably because there are more of these cases included in training with the INHOUSE data set and apparently not in the CADPE data set.

4.3.5 Evaluation on Monoenergetic Images

We have identified that the performance without DECT data augmentation decreases rapidly with increasing energy level on the monoenergetic data, i.e., as soon as the contrast decreases. We will now investigate in more detail what the networks segment on the different monoenergetic images.

Table 4.7 contains rendering images of different cases, once evaluated on monoE70 and once on monoE200 images. Thereby, networks were examined, which were trained without DECT data augmentation and additionally a network, where during the training monoenergetic data between 50 keV and 150 keV were also generated.

All networks perform well despite a few false positive predictions on the monoE70 images. It should be mentioned that when creating the render plots, the limit for detecting an embolism was set to $\tau = 0.5$, which yields a larger number of false positives in the image.

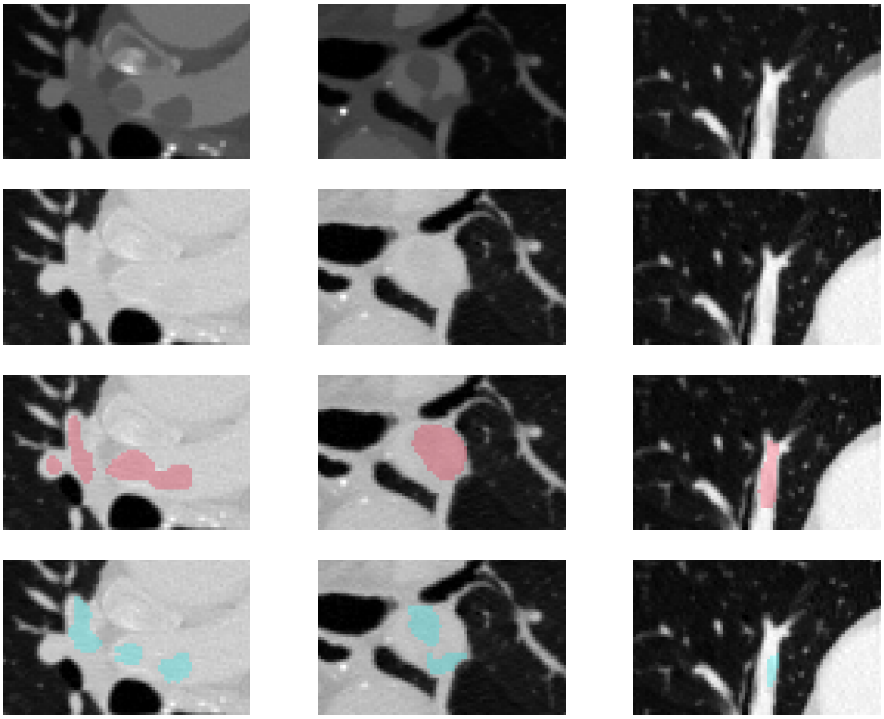


Figure 4.28: Example images with embolisms, which were not detected by networks trained with conventional data on the low contrast monoE150 images (second row), but by networks trained with DECT data. Predictions of the DECT networks are marked in red (third row) and ground truth in the last row. In addition the monoE50 images are displayed in the first row, where the contrast variation due to the embolisms is quite well visible.

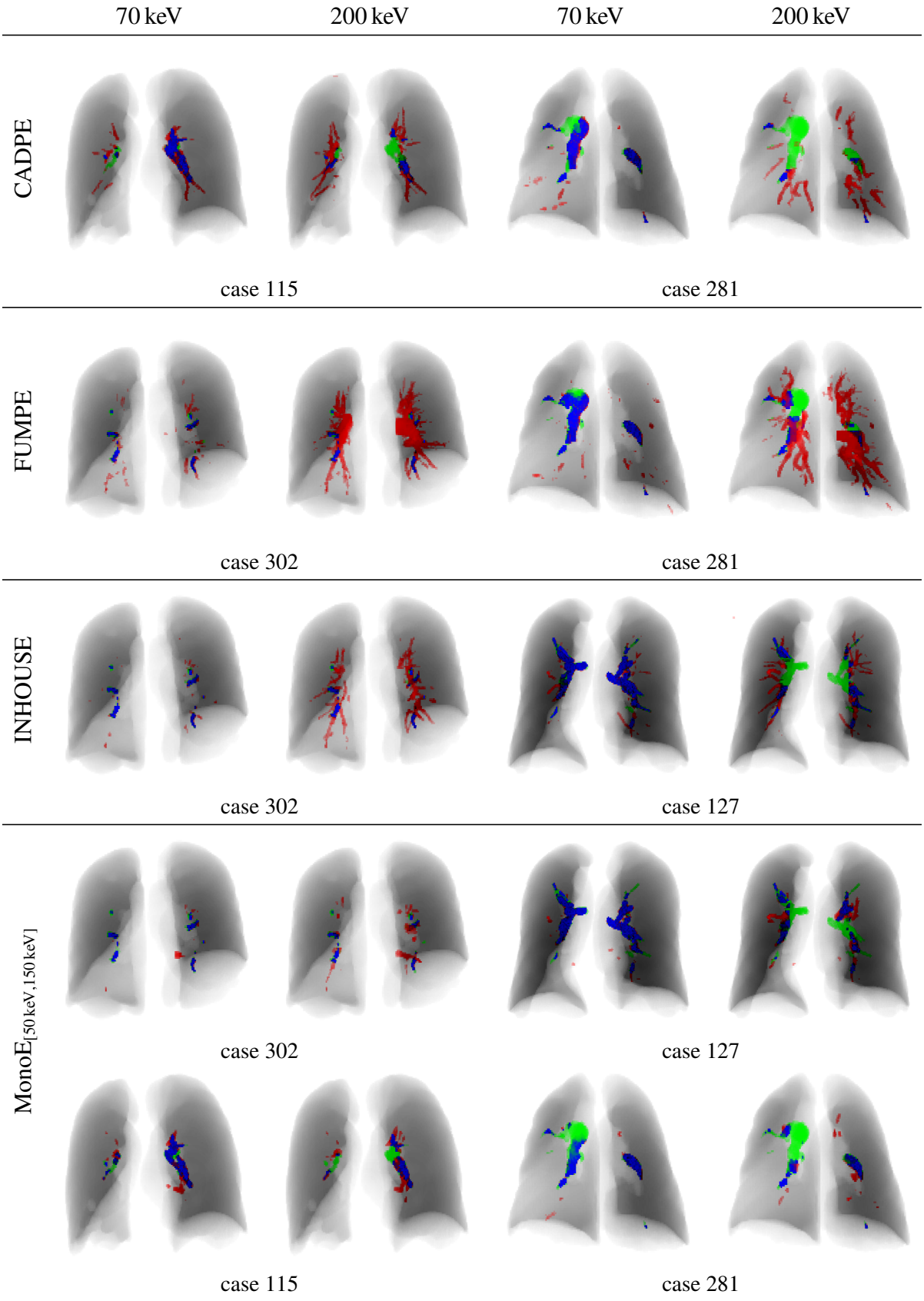


Table 4.7: Render plots of different cases on two monoenergetic energy level (70 keV, 200 keV), evaluated once with networks trained separately on CADPE, FUMPE and INHOUSE without and once on INHOUSE with DECT data augmentation. TP voxel are marked in blue, FN voxel in green and FP voxel in red.

When evaluating the networks trained without DECT data augmentation on the monoE200 images, two phenomena occur. First, a mayor part of the embolisms that were detected on the 70 keV images (blue region) are no longer detected on the monoE200 images (green region). Second, the networks tend to segment much of the vessel tree. It appears that the network no longer detects contrast variations within arteries, but is only excited by edges of a complete artery.

Considering the network trained on monoenergetic data, the segmentation on the monoE200 data differs not so much from the segmentations on the monoE70 data. Although the embolus is no longer segmented as accurately as on the monoE70 data, see case 115, case 281 and case 127, the intersection of prediction and ground truth has decreased. However, the embolisms are still detected for the most part and the network does not tend to segment the complete vessels here.

In addition, Figure 4.28 shows three examples, where the conventional trained INHOUSE network does not detect the ground truth values at an AFP rate of 5, but the monoenergetic network does. In the first row, the corresponding section of the monoE50 image is displayed, because here the embolisms are best visible. Below, the same section from the monoE150 image is visible. Here, the embolism is very difficult to recognize. In the last two rows, the prediction of the monoenergetic network is shown in red and the reference label in blue. The INHOUSE network does not segment anything in these areas. Although the segmentation is somewhat coarser, the monoenergetic network still recognizes all three emboli.

4.3.6 Center of Gravity - Comparison with CADPE Challenge

Additionally, we want to compare our general approach for detecting pulmonary embolisms with other methods. For this purpose, we compared our networks trained only with the CADPE data set with the methods of the CADPE challenge of [3].

In the CADPE challenge, FROC curves, in which the sensitivity was plotted against the average false positive rate, were also used. However, no segmentation of embolisms was performed, but detection instead: the networks predict only individual key points. If a key point lies within the embolism, this counts as a true positive. If multiple key points lie within an embolism, the key point with the greatest confidence is taken. All key points that lie outside the reference embolism are counted as false positives.

Figure 4.29 shows the FROC curves of the different methods that have been benchmarked in the CADPE challenge on 20 test images (left) and two of our networks, trained with the CADPE data set evaluated on two different splits (right). Of course the evaluation will vary slightly on different test images.

The performance of the three best networks UA-2D, UA-2.5D and UA-3D, all from [27], is listed in Table 4.8. Figure 4.29 from [3] shows all participants of the challenge and later submitted approaches, except the UA-3D network, which was later published in [27].

AFP	cv1	cv2	2D	2.5D	3D
1	63.8	65.5	47.5	69	68
2	68.8	62.8	52.5	74	72
4	72.5	73.6	62.5	75	75

Table 4.8: Sensitivity in % at different AFP rates of our networks (cv1, cv2) compared to the UA-2D,2.5D and 3D networks, which were developed by the winner of the CADPE challenge.

The networks of the challenge winner, UA-2D, UA-2.5D and UA-3D, all are based on the U-Net structure [18]. In the UA-2D network axial slices of the data are used as input while the output is a 2D embolism segmentation mask. The 2.5D network, on the other hand, is based on the 2D network, but uses a composition of five input slices, namely the target slice, two slices below and above it. Again, the single slice segmentation mask was used as the output. The 3D network uses the same structure as the 2.5D network, but with slices from the axial, coronal and sagittal plane. For each target scan, three predictions with different input slices from different planes are made and the three predictions are merged, taking the maximum of each pixel.

Although our approach is also based on a U-Net architecture, one main difference is that we use 3D patches as input and 3D segmentation masks as output. In this case, we only consider a small area at a time, but have much more information from the near environment. While in the UA approach, they consider one complete axial slice, but have less information about the environment in the other dimensions.

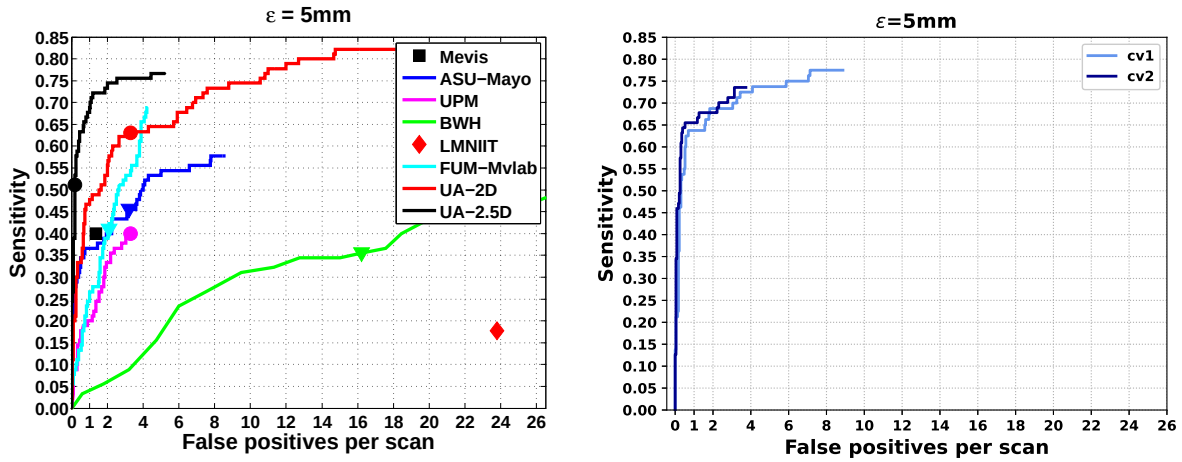


Figure 4.29: Comparison of different methods from the CADPE challenge (left),[27], and our networks trained and evaluated with two different splits of the CADPE data set (right).

One can see that our networks are close behind the UA-3D and UA-2.5D approaches and ahead of the UA-2D approach. An important difference is the way of evaluation: while we calculated the center of mass for each connected component and determined its probability

as the average over the network's output of all pixels within the connected component, the UA approaches used the following post-processing steps. First of all, a closing operation was applied to the binary segmentation mask before the connected components were determined. Instead of determining only the center of mass, the shortest distance to the perimeter was also determined for each pixel. From the points with the largest distance to the perimeter, those five were determined which were closest to the center of mass. This means that, in contrast to our approach, several candidates for each component are available here and thus the probability of a hit could be greater. It would be interesting to see if the same post-processing would improve our evaluation.

5 Conclusion

In this section, the results of the previous experiments are summarized and discussed, with which the research questions from Section 1.2 are answered. Furthermore, we want to discuss which future research questions have arisen from this work.

First of all, in our naive approach in Subsection 4.1.1, we have seen that the segmentation of pulmonary embolisms is a complex problem due to the large variability of location, size, shape and appearance and that a naive training without a more detailed study of the data and the right choice of training strategy is not sufficient to get a good performance.

In Subsection 4.1.2, it was found that the nature of the data set has a major impact on performance and generalization. At one point, the INHOUSE data set was considered separately, as it is the most challenging due to missing annotations and the presence of the large amount of comorbidities. We saw that by correcting the data and using the comorbities only in training, we can can increase performance by 17.66 pp, see Table 4.2. In addition, the effect of the size of training data set on performance became apparent when increasing the size of the data set from 66 cases to 114 cases, caused a 10 pp performance improvement.

In the separated training and evaluation, in Subsection 4.1.3, it was found that the CADPE data set generalizes best of all data sets, followed by the INHOUSE data set and FUMPE data set, which generalizes the worst. This is probably due to the small number of images in the data set, the large number of proximal embolisms, and although FUMPE has fewer outliers, the distribution is denser for the large embolisms than for the other data sets, see Subsection 3.1.1. By this characteristic, all networks have a high sensitivity on FUMPE. On the INHOUSE data set instead, all networks show the worst performance. This implies that the FUMPE data set is the easiest and that the INHOUSE data set is the most challenging. The INHOUSE data set with 112 cases and 552 embolisms is larger than the CADPE data set with its 91 cases and 317 data (see Subsection 3.1.1), the latter generalizes better. Although both data sets cover a similar range of volumes in the range, INHOUSE contains many more peripheral embolisms and has a smaller mean volume. Combined training also improves performance, but data sets should be weighted according to their quality to achieve the best possible result.

With these experimental results we are able to answer the research question, *what factors influence the generalizability.* First, we need a sufficient amount of data, but the nature of the data is much more important for generalization. The number of comorbidities, the size of the

embolisms, the distribution of peripheral and proximal embolisms in the data set, and the distribution of volumina all are important factors influencing the generalizability. In addition, a proper and standardized annotation is very important and an intensive pre-processing is necessary.

In the training section, a performance increase of 13 pp was achieved by examining the hyperparameters, see Subsection 4.2.1. Here it turned out that in particular a larger learning rate and a larger patch size is advantageous. For the latter, a negative correlation with the number of average false positives was seen. It seems that the network benefits from more regional information. This leads to the question whether approaches like in P-NET, [11], where a wider image range, but a lower amount of slices ($199 \times 199 \times 24$) are used as input or UA-2.5D where several cross sections are generated, could extract due to the higher size in two dimensions more regional information compared of using 3D cubes with same sizes in all dimensions.

In a direct comparison with the UA-2.5 network in Subsection 4.3.6, the winner of the CADPE challenge, our approach came very close to their result, although it must be said that our result might still be improved by a similar post-processing. Although the evaluation on two different test data sets makes a good statement about the general behavior, the same test data have to be used for a fair comparison.

This answers the research question, *that the 3D U-Net architecture is suitable for the detection and segmentation of pulmonary embolisms*, but there is still a demand for improvement. This is especially visible in the evaluation section. Although on the healthy data, our networks predict significantly fewer false positives on average, the number of casewise false positive predictions is still too high for clinical applications. Also, closer examination of the false positives has shown that in some cases the location is outside the ROI. This is probably due to the lack of spatial information due to the use of the 3D patches. Again, it would be interesting to try other network structures to overcome this. For example using an approach similar to Lin *et al.*, [10], in which a region proposal network first proposes candidate, after which our 3D network could then be applied for segmentation. It would also be interesting to see whether a vessel alignment in combination with a 2.5D network getting the cross sections as input or our 3D network where 3D cubes are used as the input would perform better.

Classical data augmentation, which was investigated in Subsection 4.2.2, had little effect on performance. On a smaller data set, rotation was found to provide a significant improvement. Other operations had little or no impact.

The analysis on the monoenergetic data in Subsection 4.2.3 has shown that the networks trained with conventional data perform very poorly on different contrasts. A more detailed analysis of the segmentations showed that they tend to segment the entire vessel structure due to the poor contrast, see Subsection 4.3.5. This answers the research question *how conventionally trained networks perform on different contrasts.*

By online data augmentation with the monoenergetic data it stood out that this performs very poorly without additional classical data augmentation, even though the augmented images in the range of [60 keV, 80 keV] are not very different from the conventional images. This

suggests that spectral training focuses more on the different contrasts during learning and not on the original problem, thus generating very simple but contrast-independent filters. Nevertheless, in combination with classical data augmentation, the performance could be significantly improved, which answers the last research question *how a contrast-independent prediction can be realized*. Two conclusions can be derived from this experiment. A small contrast variation in training increases the performance on the conventional data slightly. Training with exactly the same patches, only with a higher contrast variation, increased the performance on the conventional data by 3 pp, with little improvement on the monoenergetic data with contrasts close to the augmentation range. The greater the contrast variation in training, the worse the results on conventional data (and near conventional data), but the better the result on low contrasts. With spectral training in the range of [50 keV, 150 keV], the performance compared to the training without DECT augmentation on the conventional data decreases only by 1.34 pp; however, compared to the DECT training with the interval [60 keV, 80 keV] by 4 pp, the performance on the higher energy levels doubles from 20% to 40%, see Subsection 4.2.3.

Depending on the application, either improvement on conventional can be realized or generalization on different contrasts.

In summary, a comprehensive analysis regarding generalizability in pulmonary embolism detection was performed in this work. We have seen that the problem is very complex and the training is very unstable. The quality and statistics of the data sets have a crucial impact on the training. Using DECT data augmentation, we have managed to make the network more robust to different contrasts. However, many other research questions have arisen in the process. Besides the investigation of different network structures, training with further data sets is of interest. One famous data set is known from the Radiological Society of North America (RSNA) pulmonary embolism detection challenge, [52], which is the largest public annotated data set with more than 12 000 annotated cases, but it is only suitable for detection tasks. This data set could be used to first train a casewise detection network, followed by a segmentation network. Furthermore, we have shown that DECT data augmentation is suitable for performance enhancement on the conventional data as well as for generalization. Here, more training strategies could be pursued with other DECT representations, such as the VNC images. Furthermore, the networks could be trained using only the photo and scatter images, allowing the optimal transformations to be learned by the network independently. Again, it would be of interest to train with a larger DECT data set. Unfortunately, these are to our knowledge not yet publicly available.

We conclude that if some improvements are made, such as training with more conventional data, but also with more DECT data for data augmentation, and reducing the false positive rate through various approaches, such as combining a region proposal network or a casewise detection network with our segmentation network, a PE segmentation network can be used for clinical applications. The CADPE system should have access to CT data from different clinics. It runs in the background and triggers an alarm when a pulmonary embolism is detected and segmented. This information can then be passed on to a team of experts, which can then confirm the diagnosis and enable early treatment.

Bibliography

[1] Oliver Taubmann, Martin Berger, Marco Boegel, Yan Xia, Michael Balda, and Andreas Maier. *Computed Tomography: An Introductory Guide*. Springer, 08 2018.

[2] Mojtaba Masoudi, Hamid Pourreza, Mahdi Saadatmand-Tarzjan, Noushin Eftekhari, Fateme Zargar, and Masoud Pezeshki Rad. A new dataset of computed-tomography angiography images for computer-aided detection of pulmonary embolism. *Scientific Data*, 5:180180, 09 2018.

[3] Germán González, Daniel Jimenez-Carretero, Sara Rodríguez-López, Carlos Cano-Espinosa, Miguel Cazorla, Tanya Agarwal, Vinit Agarwal, Nima Tajbakhsh, Michael B. Gotway, Jianming Liang, Mojtaba Masoudi, Noushin Eftekhari, Mahdi Saadatmand, Hamid-Reza Pourreza, Patricia Fraga-Rivas, Eduardo Fraile, Frank J. Rybicki, Ara Kassarjian, Raúl San José Estépar, and Maria J. Ledesma-Carbayo. Computer aided detection for pulmonary embolism challenge (CAD-PE), 2020.

[4] Thomas Henzler, J. Michael Barraza, John W. Nance, Philip Costello, Radko Krissak, Christian Fink, and U. Joseph Schoepf. CT imaging of acute pulmonary embolism. *Journal of Cardiovascular Computed Tomography*, 5(1):3–11, 2011.

[5] T. Akagawa, T. Gotanda, T. Katsuda, and R. Gotanda. Preoperative 3D CT Pulmonary Angiography Images Using 64 Multidetector Row Computed Tomography for Cancer Patients. In Ákos Jobbágy, editor, *5th European Conference of the International Federation for Medical and Biological Engineering*, pages 583–586, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[6] Yoshitaka Masutani, Heber Macmahon, and Kunio Doi. Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis. *Medical Imaging, IEEE Transactions on*, 21:1517 – 1523, 01 2003.

[7] Chuan Zhou, Heang-Ping Chan, Smita Patel, Philip Cascade, Berkman Sahiner, Lubomir Hadjiiski, and Ella Kazerooni. Preliminary Investigation of Computer-aided Detection of Pulmonary Embolism in Three-dimensional Computed Tomography Pulmonary Angiography Images1. *Academic radiology*, 12:782–92, 06 2005.

[8] Henri Bouma, Jeroen Sonnemans, Anna Vilanova, and Frans Gerritsen. Automatic Detection of Pulmonary Embolism in CTA Images. *IEEE transactions on medical imaging*, 28:1223–30, 03 2009.

[9] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019. Special Issue: Deep Learning in Medical Physics.

[10] Yi Lin, Jianchao Su, Xiang Wang, Xiang Li, Jingen Liu, Kwang-Ting Cheng, and Xin Yang. Automated Pulmonary Embolism Detection from CTPA Images Using an End-to-End Convolutional Neural Network. In *Medical Image Computing and*

Computer Assisted Intervention – MICCAI 2019, pages 280–288, Cham, 2019. Springer International Publishing.

[11] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn Ball, Norah Borus, Andrew Huang, Bhavik Patel, Pranav Rajpurkar, Jeremy Irvin, Jared Dunnmon, Joseph Bledsoe, Katie Shpanskaya, Abhay Dhaliwal, Roham Zamanian, Andrew Ng, and Matthew Lungren. PENet - a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. *npj Digital Medicine*, 3, 12 2020.

[12] Nima Tajbakhsh, Michael B. Gotway, and Jianming Liang. Computer-Aided Pulmonary Embolism Detection Using a Novel Vessel-Aligned Multi-planar Image Representation and Convolutional Neural Networks. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 62–69, Cham, 2015. Springer International Publishing.

[13] Katharina Müller-Peltzer, Alexander Crispin, Robert Stahl, Fabian Bamberg, and Christoph Gregor Trumm. Grenzen künstlicher Intelligenz in der Notfallbefundung. *RoFo : Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin*, 2021.

[14] Pierre-Jean Lartaud, Aymeric Rouchaud, Jean-Michel Rouet, Olivier Nempont, and Loic Boussel. Spectral ct based training dataset generation and augmentation for conventional ct vascular segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 768–775. Springer, 2019.

[15] Reza Forghani, Bruno De Man, and Rajiv Gupta. Dual-Energy Computed Tomography: Physical Principles, Approaches to Scanning, Usage, and Implementation: Part 1. *Neuroimaging Clinics of North America*, 27:371–384, 08 2017.

[16] Reza Forghani, Bruno De Man, and Rajiv Gupta. Dual-Energy Computed Tomography: Physical Principles, Approaches to Scanning, Usage, and Implementation: Part 2. *Neuroimaging Clinics of North America*, 27:385–400, 08 2017.

[17] Negin Rassouli, Maryam Etesami, Amar Dhanantwari, and Prabhakar Rajiah. Detector-based spectral CT with a novel dual-layer technology: principles and applications. *Insights into Imaging*, 8, 10 2017.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[19] Zeyu Jiang, Changxing Ding, Minfeng Liu, and Dacheng Tao. Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task. In *International MICCAI brainlesion workshop*, pages 231–241. Springer, 2019.

[20] Vajira Thambawita, Steven A. Hicks, Pål Halvorsen, and Michael A. Riegler. DivergentNets: Medical Image Segmentation by Network Ensemble. In *EndoCV@ISBI*,

2021.

[21] Jan Bělohlávek, Vladimír Dytrych, and Aleš Linhart. Pulmonary embolism, part I: Epidemiology, risk factors and risk stratification, pathophysiology, clinical presentation, diagnosis and nonthrombotic pulmonary embolism. *Experimental and clinical cardiology*, 18:129–38, 06 2013.

[22] Galinier Michel Elenizi Khaled, Alharthi Rasha. Pulmonary embolism originating from germ cell tumor causes severe left ventricular dysfunction in a healthy young adult with full recovery: a case report. *BMC Cardiovascular Disorders*, 05 2021.

[23] Steven E. Weinberger, Barbara A. Cockrill, and Jess Mandel. 13 - pulmonary embolism. In Steven E. Weinberger, Barbara A. Cockrill, and Jess Mandel, editors, *Principles of Pulmonary Medicine (Sixth Edition)*, pages 179–188. W.B. Saunders, Philadelphia, sixth edition edition, 2014.

[24] Stavros V. Konstantinides. *Management of Acute Pulmonary Embolism*. Springer, 2007.

[25] Meredith L. Turetz, Andrew Sideris, Oren A. Friedman, Nidhi Triphathi, and James M. Horowitz. Epidemiology, pathophysiology, and natural history of pulmonary embolism. *Seminars in interventional radiology*, 35 2:92–98, 2018.

[26] Alastair Moore, Jason Wachsmann, Murthy Chamarthy, Lloyd Panjikaran, Yuki Tanabe, and Prabhakar Rajiah. Imaging of acute pulmonary embolism: An update. *Cardiovascular Drugs and Therapy*, 8, 12 2017.

[27] Carlos Cano-Espinosa, Miguel Cazorla, and Germán González. Computer Aided Detection of Pulmonary Embolism Using Multi-Slice Multi-Axial Segmentation. *Applied Sciences*, 10(8), 2020.

[28] Lechner G. Breitenseher M., Pokieser P. *Lehrbuch der radiologisch-klinischen Diagnostik*. Breitenseher Publisher, 2012.

[29] Johanna DenOtter, Tami D. anf Schubert. Hounsfield Unit. *StatPearls*, 2021.

[30] T.M. Buzug. *Einführung in die Computertomographie: Mathematisch-physikalische Grundlagen der Bildrekonstruktion*. Springer Berlin Heidelberg, 2011.

[31] D R Dance, Stelios Christofides, Andrew D A Maidment, I D McLean, and Kwan-Hoong Ng. *Diagnostic Radiology Physics*. Non-serial Publications. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2014.

[32] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, December 1989.

[33] Michael A. Nielsen. *Neural Networks and Deep Learning*. Determination press, 2015.

[34] Catherine F. Higham and Desmond J. Higham. Deep Learning: An Introduction for Applied Mathematicians. *ArXiv*, abs/1801.05894, 2019.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015.

[36] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[37] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

[38] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning, 2016.

[39] Tomasz Szandala. Review and comparison of commonly used activation functions for deep neural networks. *CoRR*, abs/2010.09458, 2020.

[40] Arun Kumar Dubey and Vanita Jain. Comparative Study of Convolution Neural Network's Relu and Leaky-Relu Activation Functions. In Sukumar Mishra, Yog Raj Sood, and Anuradha Tomar, editors, *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*, pages 873–880, Singapore, 2019. Springer Singapore.

[41] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer Feedforward Networks with a Non-Polynomial Activation Function Can Approximate Any Function. *New York University Stern School of Business Research Paper Series*, 1993.

[42] Humaidi A.J. et al. Alzubaidi L., Zhang J. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 2021.

[43] Richard Garnett. A comprehensive review of dual-energy and multi-spectral computed tomography. *Clinical Imaging*, 67, 08 2020.

[44] Frank Natterer. *The Mathematics of Computerized Tomography*. Society for Industrial and Applied Mathematics, USA, 2001.

[45] Harrison H Barrett. Iii the radon transform and its applications. In *Progress in optics*, volume 21, pages 217–286. Elsevier, 1984.

[46] Martin Berger, Qiao Yang, and Andreas Maier. *X-ray Imaging: An Introductory Guide*. Springer, 08 2018.

[47] Carsten Schirra, Bernhard Brendel, Mark Anastasio, and Ewald Roessl. Spectral ct: A technology primer for contrast agent development. *Contrast media molecular imaging*, 9:62–70, 01 2014.

[48] Anushri Parakh, Manuel Patino, and Dushyant V Sahani. Spectral ct/dual-energy ct. In *Multislice CT*, pages 59–79. Springer, 2017.

[49] Robert Alvarez and A Macovski. Energy-selective reconstructions in x-ray computerized tomography. *Physics in medicine and biology*, 21:733–44, 10 1976.

[50] Xin Yang, Yi Lin, Jianchao Su, Xiang Wang, Xiang Li, Jingen Lin, and Kwang-Ting Cheng. A Two-Stage Convolutional Neural Network for Pulmonary Embolism Detection From CTPA Images. *IEEE Access*, 06 2019.

[51] S.K. Warfield, K.H. Zou, and W.M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[52] Errol Colak, Felipe Kitamura, Stephen Hobbs, Carol wu, Matthew Lungren, Luciano Prevedello, Jayashree Kalpathy-Cramer, Robyn Ball, George Shih, Anouk Stein, Safwan Halabi, Emre Altinmakas, Meng Law, Parveen Kumar, Karam Manzalawi, Dennis Rubio, Jacob Sechrist, Pauline Germaine, Eva Lopez, and John Mongan. The RSNA Pulmonary Embolism CT (RSPECT) Dataset. *Radiology: Artificial Intelligence*, 3, 01 2021.