THE UNIVERSITY OF CHICAGO

INVESTIGATION OF DEEP LEARNING IN MEDICAL IMAGING FOR ENHANCED WORKFLOW, IMPROVED DIAGNOSIS, AND EXPLANATORY ARTIFICIAL INTELLIGENCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MEDICAL PHYSICS

BY
JENNIE SANDOZ MARIE CROSBY

CHICAGO, ILLINOIS
JUNE 2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The journey from conceptualizing this project to the results presented here included many people. My advisor, Dr. Maryellen Giger, took the time to mentor, support, encourage, and motivate me. I am so honored to have been advised by such a fantastic and well-respected scientist. She helped me immensely to grow as a scientist myself.

The members of my thesis committee were invaluable to the process, providing support, guidance, and scientific expertise to ensure my work was of the highest quality. Dr. Heber MacMahon generously shared his expertise in thoracic radiology, which was crucial to this project. Dr. Samuel Armato, in addition to his strong leadership of the graduate program, contributed his expertise in image processing, also a major component of this work. Dr. Hania Al-Hallaq contributed her knowledge as a radiation oncology physicist, ensuring my work was valuable and understandable to those outside of radiology.

I want to acknowledge Dr. Feng Li, who provided the annotations and interpretations for the radiographs used in this work.

I was incredibly honored to have the opportunity to mentor three fantastic students: Thomas Rhines, Sophia Chen, and Clara Duan. Thomas, a physics undergraduate at the University of Chicago, greatly contributed to this work. He conceived his own ideas under the framework of this project and pursued those ideas himself. He is an extremely capable scientist, who made me think by asking great questions. Sophia Chen, an undergraduate at Vanderbilt University, worked with me for a summer. She was incredibly helpful for getting the visualization portion of this work started. Clara Duan, a high school student at the time, greatly helped this project through her contributions organizing the database.

My fellow Graduate Program in Medical Physics (GPMP) students were so helpful both academically and socially. My classmates, Scott, Adam, and Sam, were supportive and helpful throughout my time in the program. My office mates, Madeleine and Jordan, always listened to me and helped me when I needed it. The other GPMP students were also supportive and I am very fortunate to know them.

The GPMP faculty were supportive of my goals in a variety of ways. A special thank you to the faculty who taught my courses, as well as the faculty who let me shadow them in the clinic. I also want to thank the GPMP support and administrative staff for their help over the years.

This dissertation would not have been possible without the support of the following: the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health (NIH) under grant number T32 EB002103, NIH Grant S10 OD025081 “Protected Radiomics Analysis Commons for Deep Learning in Biomedical Discovery,” University of Chicago Department of Radiology Grant “LDCT lung screening deep learning project,” and the Lawrence H. Lanzl Graduate Fellowship Award.

Last, and certainly not least, I would like to give a huge thank you to my family. I owe everything to my mother and father who never wavered in their love and support. They sacrificed so much to make sure I got the highest-quality education available. I cannot thank them enough for all the opportunities they provided for me. I am so blessed to have an incredible brother, John-John, who was always supportive and willing to chat about my work. This work could not have happened without the support of my husband, Spencer. He provided so much encouragement, love, and pizza throughout the process. I am incredibly fortunate to have so many people who supported me.

ABSTRACT

In the past few years, applications of deep learning have experienced explosive growth due to their role in solving complex problems. Deep learning has recently been gaining attention for use in medical imaging and applications of deep learning are being explored to enhance radiology practice, including for the selection and preparation of images for interpretation, analysis of image quality, and assistance for diagnostic decision-making tasks, among many other clinical applications. For the use of deep learning in medical imaging, it is important to understand physical limitations of medical images as well as techniques with which to augment inputs and forms of output with which to enhance specific task performance. The primary goals of this research are to investigate deep learning in medical imaging through contributions in (i) workflow enhancement, (ii) diagnostic improvements, and (iii) AI output understanding (i.e., explainability) through the specific tasks of detection and visualization of pneumothorax on thoracic radiographs. However, this specific investigation of pneumothorax detection and visualization could yield procedures applicable to other imaging applications.

Pneumothorax, the abnormal presence of air between the lung and chest wall, can be diagnosed using a chest radiograph; visual indications of pneumothorax in a chest radiograph include a fine line at the edge of the lung and a change in texture outside the lung. Due to the overlapping structures within a frontal chest radiograph due to 2D projection radiography, pneumothorax can be difficult for even an experienced radiologist to detect. The detection of pneumothorax within the radiograph is further complicated by the wide variety of sizes and severities pneumothorax can possess.

Deep Learning for enhancing radiology workflow: Once medical images are acquired, the organization of the images is typically accomplished semi-automatically using DICOM header information. However, DICOM header information can be incorrect or inconsistent due to input error or use of equipment from multiple vendors. Deep learning has a role in classifying images in terms of their projection view in order to compensate for the lack of DICOM accuracy. In this work, it was demonstrated that a convolutional neural

network trained from scratch with radiographs manually classified by radiographic view can classify, rapidly and accurately, unseen test images into the various views resulting from a dual-energy chest radiography study. In addition, a convolutional neural network trained from scratch was able to classify between frontal chest radiographs acquired anteroposteriorly and posteroanteriorly with a high performance. Convolutional neural networks trained to classify projection view could be used by a researcher to quickly assemble and organize a dataset of images of a certain view or by a radiologist for application within the clinical workflow; thus, enhancing radiology workflow.

Deep learning for improving medical image diagnosis: After a network is trained to classify projection view and a dataset is organized, the images themselves may be input to a network to address a specific diagnostic question. Unless the task is relatively simple, a large number of images are needed in the dataset to apply deep learning methods. Training from scratch requires thousands to millions of images, depending on how many classification categories are required. For a dataset that is too small to train a network from scratch, transfer learning can be applied. Transfer learning is using a trained network and retraining a portion of it for the new task, known as fine-tuning, or using the output from various network layers in a classifier, known as feature extraction. Deep learning, and more specifically, transfer learning techniques, could be useful for triaging radiographs and if a radiograph has a high probability of having a pneumothorax, it could be prioritized for prompt reading by a radiologist, thus decreasing the time from image acquisition to diagnosis.

The input resolution to convolutional neural networks is limited due to computational resources. Thus, when deep learning algorithms are applied to medical images, the images are often downsampled to enable input to the convolutional neural network. As a result, the input spatial resolution is reduced, possibly obscuring signs of disease. For the specific case of pneumothorax, a fine line at the edge of the lung and a lack of lung texture outside the lung are the primary visual indications of a pneumothorax. However, if the radiograph is downsampled for input to deep learning, some of those visual indications may be obscured.

In this work, the impact of image resolution on the deep learning detection of pneumothorax was investigated, finding that a higher-resolution input to train the convolutional neural network led to an improved classification performance.

Deep learning for understanding AI output: The ability to interpret and understand deep learning algorithms' output is vital to the eventual clinical implementation of the algorithms. Deep learning visualization techniques can be used for the enhancement of AI workflow. Visualization of neural network output provides an explanation of neural network decision-making by indicating the locations within the image that have the greatest influence on the final classification. The visualization of pneumothorax on frontal chest radiographs enables the localization of pneumothorax within the image as well as further investigation of the impact of input image resolution. Visualization tools allow for greater understanding of AI output and enable troubleshooting to determine the causes of incorrect classifications. In this work, two visualization methods were performed and compared for two levels of spatial resolution, finding that the localization ability was superior when visualizing output on the higher-resolution images.

The medical significance of this research is that it could improve both the delivery and effectiveness of patient care by incorporation of deep learning along various stages of the clinical workflow (both within the image preparation and the image interpretation tasks). The application of deep learning to workflow enhancement through radiograph view classification, improved diagnosis through transfer learning techniques, along with the analysis of deep learning output visualization, yields improvements for pneumothorax detection with deep learning, which can be applied to other imaging tasks.

Keywords: chest radiography, pneumothorax, convolutional neural networks, deep learning, machine learning, computer vision, thoracic imaging

CHAPTER 1

INTRODUCTION

1.1 Application of Computers to Medical Imaging Tasks

The benefit of a medical imaging examination to a patient depends on the image quality and the ability of the radiologist interpreting the image. The enhancement of radiologist performance is possible through the application of computers to various clinical tasks.

The use of computers to analyze radiographs has been discussed since the mid-1950s [1]. Substantial efforts were made during the 60s and 70s to investigate and apply computers to the tasks of detecting or classifying abnormalities on medical images [2]. These early efforts were hampered by the lack of digital data and limited computational power. In the mid-1980s, technology had advanced sufficiently for further investigation of the use of computers to detect abnormalities on chest radiographs and mammograms [2]. The first computer-aided detection (CADe) system for mammography was Food and Drug Administration (FDA) approved in 1998 and a commercial CADe system for lung nodule detection on chest radiographs was FDA approved in 2001 [2]. CADe is intended to provide decision support to radiologists to augment their performance and improve efficiency [3].The investigation and application of computers to medical imaging has increased since these times due to the much larger quantity of digital images as well as the increase of computational power.

As early as 1994, convolutional neural networks, computer algorithms applied to analyze images, were being applied to medical images [3, 4]. Deep learning involves convolutional neural networks with many hidden layers and, therefore, successive layers of representation [5]. Deep learning has found diverse applications in medical imaging; deep learning methods have been applied to various imaging modalities for many different tasks, including detection, segmentation, and classification [6]. The applications of deep learning to augment current radiology workflows include: report generation, image quality assessment, triaging acquired images, detecting various conditions/diseases on images from many modalities, and many other clinical tasks [7].

One potential clinical application of deep learning is to expedite the time from acquisition to radiologist interpretation for medical images. Expediting interpretation is particularly important for conditions that require immediate treatment or have a high likelihood of disease. A study by Rachh et al. [8] reported the median time from acquisition of a chest radiography study to radiologist interpretation was 520 minutes in their clinical system. Chest radiography could particularly benefit from deep learning applications since 52 million chest radiographs are ordered per year in the United States, as of 2017 [8]. There are multiple recent publications on the topic of detecting conditions and diseases in chest radiographs using deep learning [9, 10, 11, 12, 13]. One of the challenges of using computerized methods for chest radiographs is its low contrast sensitivity for subtle abnormalities [3]. Most of the studies apply deep learning methods uniformly to various conditions that present differently in the radiographs. Many groups report strong performances, making the detection and diagnosis of disease on chest radiographs using deep learning appear to be a solved problem. However, the detection of pneumothorax on frontal chest radiographs using deep learning requires additional considerations compared to other lung conditions.

1.2 Pneumothorax

Pneumothorax (PTX) is the abnormal presence of air between the lung and the chest wall (Figure 1.1). It can occur spontaneously, due to trauma, or from a medical procedure. Spontaneous PTX can occur for patients with underlying lung disease, such as emphysema; for those patients, even a small PTX will cause extreme breathing difficulties [14]. PTX is caused by procedures such as transbronchial lung biopsy, thoracentesis (draining fluid via a needle inserted between the lungs and chest wall), and central vein/pulmonary artery catheterization [15]. PTX can also be caused by a pressure differential or due to barotrauma, injury caused by increased air pressure [15]. PTX is associated with lung diseases such as chronic obstructive pulmonary disease (COPD) and acute respiratory distress syndrome

(ARDS) [15].

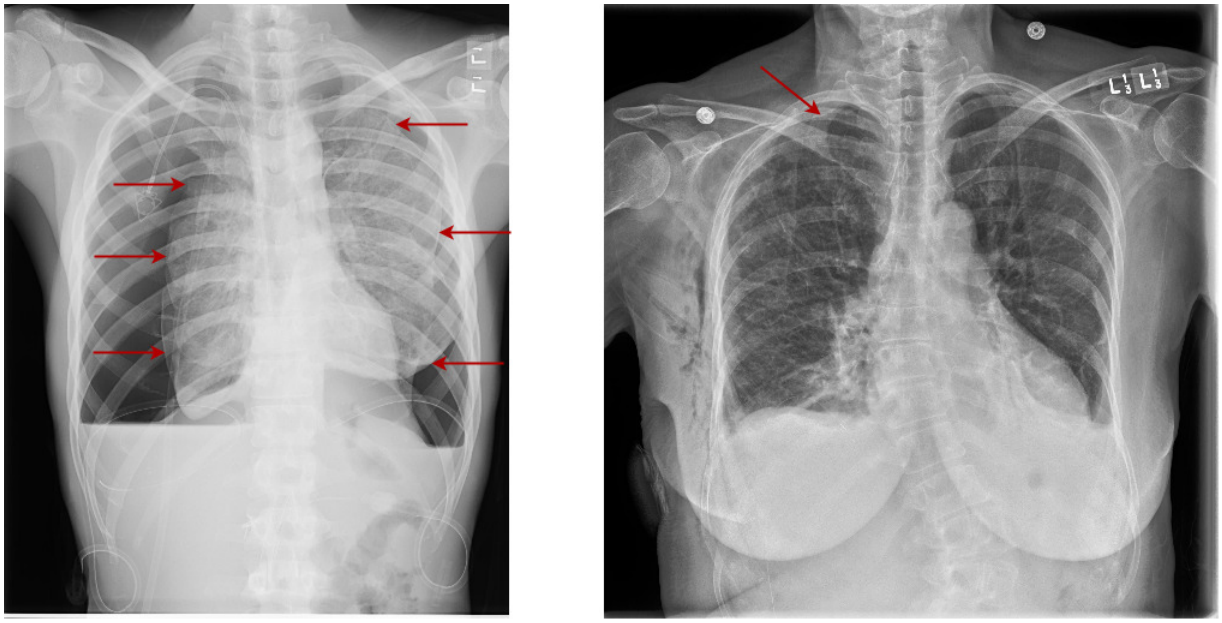


Figure 1.1: Two PTX cases of different sizes. On the left, the patient has large bilateral PTXs. On the right, the patient has a small PTX on the right side. The arrows point to the edge of the lung.

A study conducted by Chen et al. [15] regarding PTX in the ICU at their hospital found that 60 cases (3% of total ICU patients) occurred over 36 months. They found that 58% of the cases (35) were caused by a procedure, the most common was thoracentesis (19 patients, 54%). In addition, 92% of the patients with PTX (52 patients) had an underlying condition, with the most common being a malignancy (18 patients). Of the patients with PTX, 33 (55%) had ARDS, the most common cause of respiratory failure.

After patients undergo a lung biopsy, they are monitored for PTX via chest radiographs, with between 17% and 27% of patients developing a PTX following a lung biopsy [16]. Chen et al. [15] found that the prognosis of patients with PTX due to a procedure was better than the patients whose PTX was caused by barotrauma and hypothesized that early detection by radiography may be able to improve their prognosis.

1.3 Pneumothorax Imaging

PTX can be imaged using many different modalities including: radiography, computed tomography (CT), and ultrasound [17]. Imaging PTX using each of these modalities will be briefly reviewed.

A standard chest radiography examination consists of an erect posteroanterior (PA) radiograph and a left lateral projection acquired during full inspiration [18]. PA radiographs are obtained when the x-ray beam enters the back of the chest (posterior) and exits the front (anterior), where the detector is located. A chest radiography exam can also be performed using a portable x-ray machine, acquiring a radiograph anteroposteriorly (AP). AP radiographs are obtained when the x-ray beam enters the front of the chest (anterior) and exits through the back of the chest (posterior), where the detector is positioned. Portable radiography is recommended for patients too unstable or unable to travel to a radiology department [19]; however, if a standard chest radiography exam is possible, it is preferred due to the superior diagnostic quality and acquisition of multiple projections [20]. Whether the image was acquired AP or PA affects the appearance and measured dimensions of pathologic findings, such as tumors, as well as thoracic structures, such as the heart, due to photon beam divergence [21]. An example of an AP and a PA radiograph for the same patient is shown in Figure 1.2.

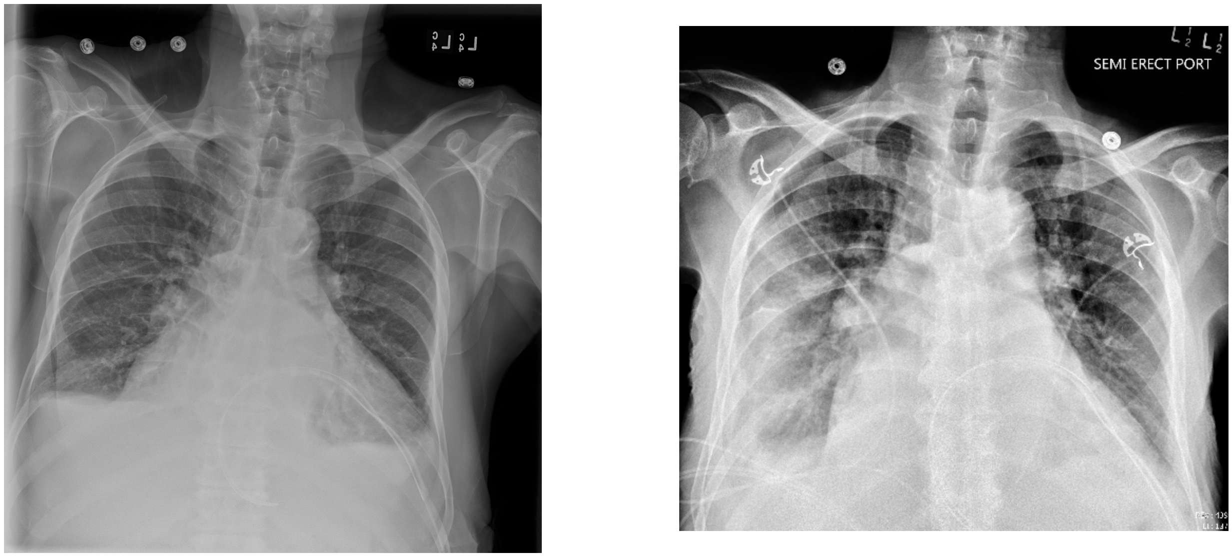


Figure 1.2: Example of a PA (left) and an AP (right) radiograph for the same patient. The AP radiograph was acquired a day after the PA radiograph for this patient. Note that the window/level is different between the two images.

While a standard chest radiography exam is preferable to portable radiography, the American College of Radiology (ACR) appropriateness criteria [19] recommends portable AP radiographs for immediate assessment following an interventional procedure in the chest or abdomen to check for PTX. In the ACR appropriateness criteria for acute respiratory illness in immunocompetent patients, chest radiography is considered sufficient for the diagnosis of most PTXs; a standard chest radiography exam is preferred [22]. Kirkpatrick et al. [23] evaluated the sensitivity of radiography for the detection of PTX and found that for supine AP radiographs of 225 trauma patients, the sensitivity is 20.9%. CT is considered the gold standard for the detection of PTX; however, sometimes patients are too ill to be transported to a CT scanner and the increased radiation dose of a CT scan compared to a chest radiography study should be considered [17]. The ACR appropriateness criteria for acute respiratory illness in immunocompetent patients recommends CT scanning only when the underlying causes of the PTX need to be identified [22].

Ultrasound is another method for imaging PTX. A meta-analysis by Ding et al. [17]

investigated 20 studies regarding the detection of PTX via ultrasound and radiography and found that the pooled sensitivity of ultrasound was 0.88 (0.85-0.91) and the specificity was 0.99 (0.98-0.99). For radiography, the pooled sensitivity was 0.52 (0.49-0.55) and the specificity was 1.00 (1.00-1.00).

Due to the ubiquitous nature of chest radiography and the recommendations from the ACR, radiography for the detection of PTX will be the focus of this work. The visual signs of a PTX on a frontal chest radiograph can be subtle; the most specific visual signs include a fine line at the edge of the lung and a lack of lung texture between the ribcage and lung [24]. PTX can have a wide range of sizes and severities. The detection of PTX is further complicated by the overlapping structures in a chest radiograph, potentially obscuring the subtle visual signs. The detection of PTX yields decision-making tasks relevant to many imaging exams since the PTX is a high spatial frequency structure with fine detail within a complex background of varying normal structure, texture, and non-anatomical devices.

1.4 Research Scope and Objectives

The diagnosis of PTX using chest radiography involves the acquisition of a chest radiography exam and then the interpretation of the acquired radiograph(s) by a radiologist who uses their clinical judgment to determine whether a PTX is present. Deep learning has the potential to improve the efficiency of the process and enhance radiologist performance. Three areas of potential improvement in the process are: the classification of radiographic views for workflow enhancement, the detection of PTX within the frontal radiograph for improved diagnosis, and display of the location of the PTX to the radiologist for explanatory deep learning output. Figure 1.3 shows the overall workflow incorporating the contributions in workflow enhancement, diagnostic improvements, and AI output understanding through the tasks of detection and visualization of pneumothorax on chest radiographs. These three areas have distinct objectives; therefore, deep learning methods applied need to be thoughtfully chosen with the knowledge of each method's limitations and knowledge of the task. The different deep learning techniques for application to medical images, as well as their advantages and disadvantages, will be discussed in Chapter 2. Within those deep learning methods, there are many deep learning architectures that can be applied; four architectures and their innovative components will be discussed in Chapter 2.

The classification of radiographic view using deep learning methods can enhance radiology workflow, which will be the focus of Chapter 3. A standard chest imaging exam consists of a frontal and lateral image; however, some radiology departments use dual-energy imaging, yielding two additional views: soft tissue and bone. As discussed previously, PTX may also be imaged using AP radiographs acquired on a portable x-ray machine. Chapter 3 includes methods and results for the classification of radiograph view resulting from a dual-energy chest examination, as well as the classification of radiographs acquired AP versus PA. The application of deep learning to classification of view can ensure that the correct image, the frontal radiograph, is displayed to the radiologist who is interpreting the exam. In addition, if deep learning methods are being applied after this step, classification of view using deep learning methods can ensure the correct image view is provided to the following deep learning algorithm(s). Figure 1.3 shows the four images resulting from a dual-energy thoracic radiography study and the classification of radiographic view, providing the frontal image to the following step in the process for the detection of PTX.

Due to the subtle visual signs, overlapping structures in a frontal chest radiograph, and wide range of PTX size/severity, the detection of PTX within an image can be a challenging task. Deep learning has a role in the detection of the visual signs of a PTX to expedite radiologist interpretation so the PTX can be treated as soon as possible and/or to assist a radiologist in their interpretation. Chapter 4 will focus on the use of deep learning to improve medical image diagnosis through the detection of PTX on radiographic images, with emphasis on the effect of input image resolution on deep learning performance. Appendix A will further discuss the impact of image resolution on deep learning performance; results from the use of higher-resolution images for the detection and localization of pneumothorax will be presented and discussed. Appendix B provides more information and analysis of the possible confounding variables when applying deep learning for the detection of pneumothorax. Deep learning methods can provide radiologists with an output score indicating the probability that the radiograph has PTX. Figure 1.3 shows images with a probability of PTX output by the deep learning algorithm that exceeds the threshold probability; therefore, those images are used as input for the next step.

In addition to a probability of PTX output by the deep learning algorithm, indicating the position of the detected PTX can assist radiologist interpretation, as well as provide explanation for the probability score given. Interpretable output from deep learning algorithms can direct radiologist attention to the region of the image with PTX. If the radiograph is determined not to have a PTX by the radiologist but has a high probability output by the deep learning network, a visualization of the output can reveal the cause of the misclassification. Chapter 5 will focus on the visualization (i.e., explainability) and human interpretability of deep learning output. Visualization background and methods, as well as results from the visualization of pneumothorax, will be discussed. Figure 1.3 shows visualizations of PTX for two visualization methods, which will be discussed in detail in Chapter 5.

The findings and implications of this work will be summarized in Chapter 6. Overall, this work addresses the application of deep learning to various stages of the process of PTX diagnosis using radiographic images, including workflow enhancement, diagnostic improvements, and visualization for deep learning output understanding. These improvements could yield techniques suitable for other applications of deep learning in medical imaging. Future research directions will be proposed in Chapter 6.

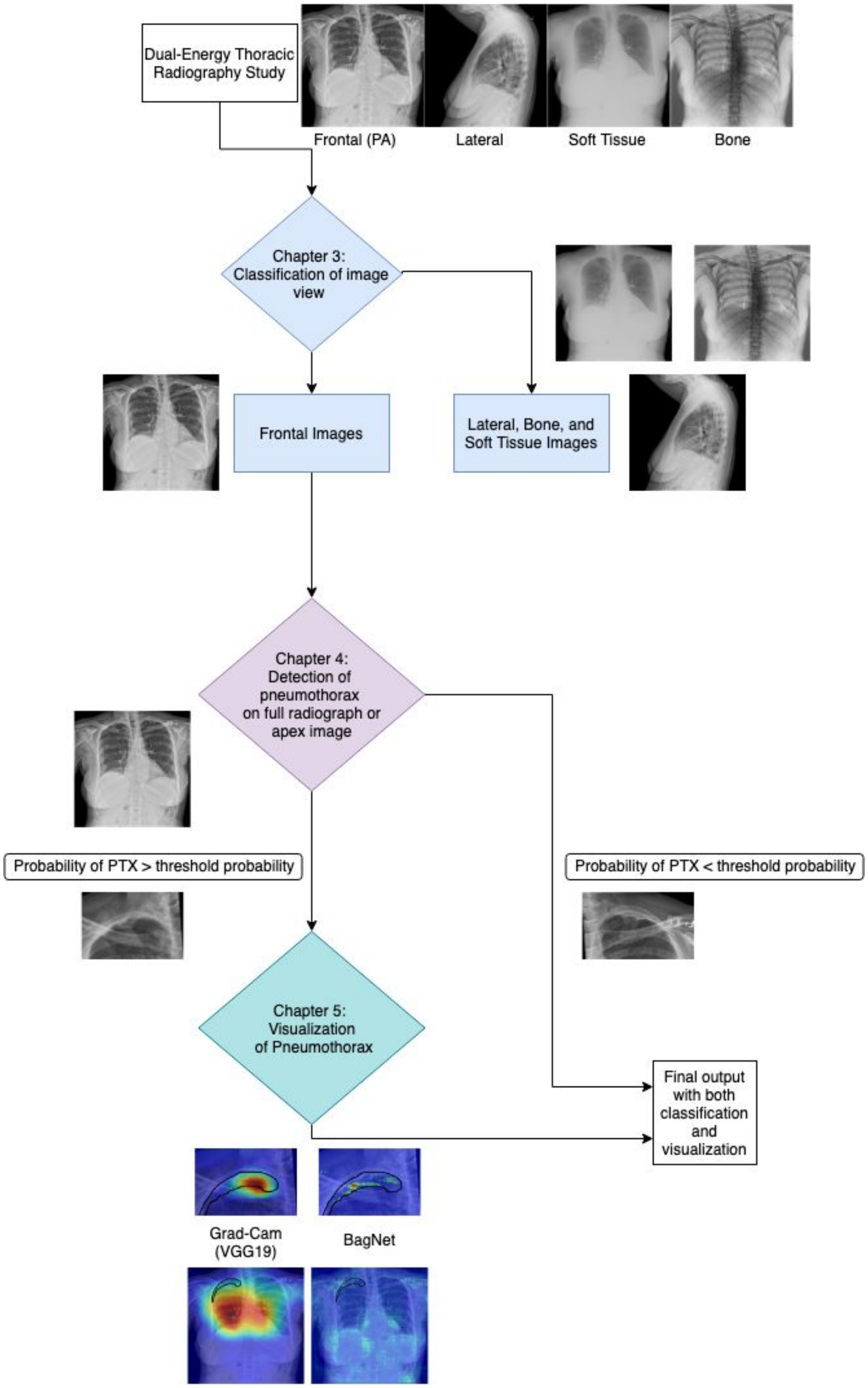


Figure 1.3: Depiction of the overall workflow incorporating the contributions in workflow enhancement, diagnostic improvements, and AI output understanding through the tasks of detection and visualization of PTX on chest radiographs. The start of the workflow is the acquisition of a thoracic radiography study and the output is a visual summary of the detection and localization of PTX.

CHAPTER 2

BACKGROUND OF DEEP LEARNING METHODS APPLIED

The first known general purpose computer, the Analytical Engine, was designed in the 1830s and 1840s. A collaborator of the inventor, Lady Ada Lovelace, commented on the invention saying that "it could do whatever we know how to perform" [5]. Artificial intelligence (AI) pioneer Alan Turing pondered Lovelace's comment and determined that computers are capable of learning, going beyond what we know how to order them to perform [25, 5]. This conclusion inspired a new paradigm of programming. The classical paradigm of programming involves human input of rules and data to apply those rules to and then the program provides the outputs. The new paradigm involves humans inputting data and outputs and the program provides the rules, which can be applied to new data [5].

Machine learning is the ability of an algorithm to learn patterns or associations without being explicitly programmed to do so. Machine learning can be applied through the computerized extraction of features, which are then put into a classifier such as a support vector machine (SVM) [3]. Deep learning, a subset of machine learning, refers to the use of neural networks, which are so named due to their structure being inspired by and reminiscent of the human neuron system. Deep learning specifically refers to neural networks with many hidden layers and, therefore, successive layers of representation [5]. The primary advantage of deep learning methods is that there is no feature engineering or selection required; the network determines and learns the appropriate features for the optimal performance [5]. Deep learning is used due to its flexibility; rules-based programming may fail when exposed to a new problem, whereas deep learning responds based upon learned parameters from the training set [7]. Deep learning is being successfully applied in many fields including natural language processing, autonomous driving, and targeted marketing [6].

As discussed in the previous chapter, computer methods have been applied and investigated for decades to perform many tasks relevant to medical imaging. The recent resurgence of machine learning is due to the advancement of graphical processing units (GPUs) that are capable of training deep neural networks. Computing power prior to GPUs was inadequate for updating the weights properly and using networks deeper than 5 layers [6]. The success of deep neural networks in various tasks outside of medical imaging has encouraged the application of these methods in medical imaging for the enhancement of human healthcare. Deep learning has been compared to the training of a new radiologist; given many examples and expert clinical truth, radiologists learn what to look for within the image [3].

There are three broad categories of machine learning: supervised learning, unsupervised learning, and reinforcement learning [6]. For supervised learning, examples labeled with the truth are input for training so the network can learn how to optimize the performance to achieve the highest accuracy on the validation test [5]. Examples of supervised learning methods are support vector machines, neural networks, and decision trees. In unsupervised learning, unlabeled input data are given to the neural network, and it learns its own classification scheme to divide the dataset into the network's self-determined classes [5]. Unsupervised methods include K-means, hierarchical clustering, and fuzzy C-means. The third category of machine learning is reinforcement learning, where a task is performed and the program receives negative or positive reinforcement [7]. The goal of reinforcement learning is for the network to learn the consequences of actions in an environment; reinforcement learning is used to teach networks how to play video games, for instance [7].

The detection of PTX in chest radiographs is a challenging problem, and the maximum transparency in training/testing is needed; therefore, the remainder of this work will be focused on supervised learning methods via deep learning. Supervised learning is the most common category of machine learning applied to medical images [6].

Before discussing deep learning further, specific terminology that pertains to deep learning will be defined. Input is given to the deep learning model and a prediction is output. Inside the deep learning model are layers, which are a sequence of data transformations. Transformations implemented by a layer are parameterized by the weights. When a deep learning model is trained using supervised methods, truth is provided along with the input data, so a loss function can measure how far the output is from the truth or expected output. After one iteration over all the training data ("epoch"), the loss function outputs a loss score for the validation dataset ("validation loss"), which is then input to an optimizer. There are several types of optimizers, such as stochastic gradient descent (SGD) [26], adaptive gradient (AdaGrad) [27], and Adam [28]; the optimizer updates the weights to minimize the loss. The adjustment of the weights per epoch is given by the learning rate, which can be reduced when the loss plateaus or starts increasing to improve convergence to the optimal solution. A deep learning model is the set of weights learned by the trained network, which can then be applied to an unknown image for prediction. Testing of the deep learning model is performed using an unseen set of examples to evaluate the trained model's performance on new data. Figure 2.1 is a diagram of the components of training a deep learning model.

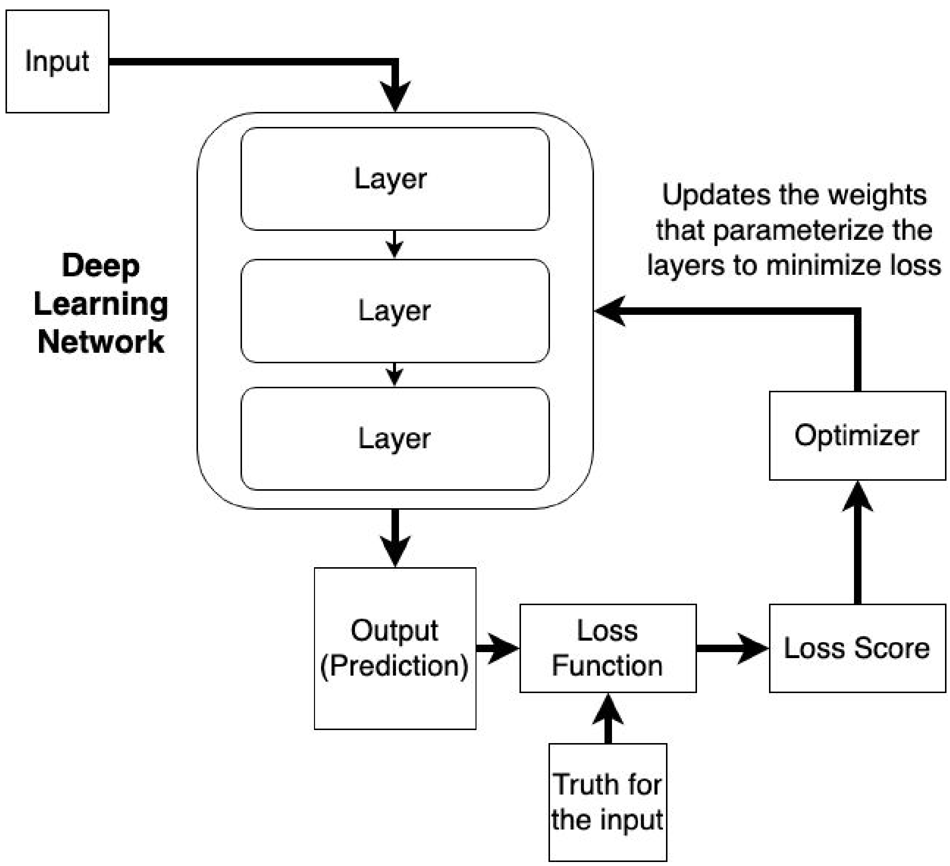


Figure 2.1: Process of training a deep learning model via supervised methods. Input is given, the output from the model is compared to the truth, which is then used to optimize the weights to minimize the loss.

Convolutional neural networks (CNNs) are a type of deep learning algorithm. CNNs assume a geometric relationship in the inputs, such as the rows and columns of the image [6]. One property of CNNs that makes them useful for many computer vision tasks is that the patterns they learn are translation invariant, meaning that once the network learns a pattern, it can detect the same pattern no matter where it is in the input image [5]. Each convolutional layer has filter elements (also called kernels) [7]. The filter element is moved across the image and the output at each location creates the output value. The step size for the movement of the filter element is called the stride; a stride of 1 indicates the filter element shifts 1 pixel at a time. The receptive field is the region of the input space visible to a filter element. In CNNs, specialized layers are used to amplify the important features from the convolutional layers. One such layer is an activation layer, which follows a convolutional layer [6]. The most common activation layer used is the rectified linear unit (ReLU), which has an output of zero for any negative value and keeps the value for any positive value. Another vital layer for CNNs is the max-pooling layer; it takes the outputs from the convolutional kernel and finds the maximum value, therefore rewarding the convolution function that extracts the most important features from the image [6]. Another important component for training a CNN is regularization, which rescales the weights between a pair of layers to a more effective range. A common form of regularization is "dropout," which sets 50% of the weights (typically, the percentage is set by the user) to zero [6]. The weights set to zero are random and vary each training round; with many iterations, the important connections will be kept. The final layer of a CNN is the classification layer; the classification layer outputs scores to quantify the image's likelihood of belonging to each class. Softmax is the classification layer used for a mutually exclusive classification; the score output from a softmax classification layer is a probability of belonging to a given class since the scores sum to 1. There is not a formula to determine the number and types of layers needed for a given problem; it is a trial-and-error process requiring the training of various architectures to determine the optimal configuration for a given problem [5].

Computer analysis of medical images requires customization to the task and imaging modality [3]. Training a CNN for image classification can be accomplished through a variety of techniques, although choosing the correct technique requires knowledge of the data set and the task. One strategy is training from scratch, which will be discussed further in the next section. Another strategy is transfer learning, which can be further divided into two techniques: fine-tuning and feature extraction. Knowledge of the task and the limitations of the techniques assist in making the optimal choice for the needed task.

2.1 Training From Scratch

Training from scratch involves using a network with all weights unknown and randomly initialized at the beginning of training. The training set with its accompanying truth labels is then used to train all the parameters of the network to optimize the performance. The weights are adjusted using the validation dataset after each iteration of the training data. After the training is complete, the test set is used to evaluate the performance of the trained network.

CNNs have a large number of weights that need to be optimized for the task the network is trained to perform. For example, a relatively shallow network, AlexNet [29], has over 60 million parameters. Therefore, training with a small dataset may not lead to strong performance; the number of training examples may not be enough to optimize the parameters for the task. The required dataset size for training from scratch depends on the depth of the network and complexity of the task. Medical imaging datasets are far more limited in number than the natural color images for which neural networks were created, and obtaining an adequate dataset of medical images is a significant obstacle for the medical application of deep learning.

Due to the large number of weights, if there is a small number of training examples, it is likely the CNN will overfit the data. Overfitting is a common issue for deep learning and occurs when the CNN learns the weights to optimize the performance on the training set but cannot generalize to the validation set or an unseen test set [7]. To mitigate the risk of overfitting, the performance on the validation set following each round of the training data can be monitored and training stopped after the performance on the validation set decreases by a certain amount. Another technique to reduce overfitting is the use of dataset augmentation [5]. Dataset augmentation is the use of image transformation techniques to increase the number of example images the network uses for training. Augmentation can consist of shifting an image, zooming (magnifying), rotating, flipping, and adding noise, for example. Figure 2.2 shows two examples of image augmentation techniques.

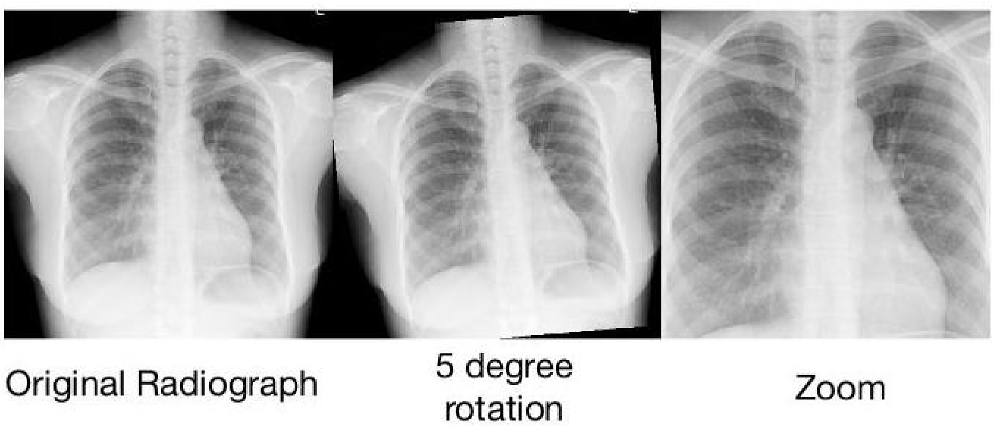


Figure 2.2: Two dataset augmentation techniques, rotation and zooming, applied to a radiograph.

In medical imaging tasks, dataset augmentation should be applied such that the transformations are realistic and clinically useful [30]. For example, the slight rotation of a radiograph is realistic, often patients are slightly slanted in their radiographs. However, rotating an image by a large angle would be unrealistic since the trained CNN would be unlikely to receive a clinically acceptable image with that appearance. Data augmentation techniques should be chosen such that they best capture medical image statistics [30]. Figure 2.3 shows two examples of image augmentation techniques that are unrealistic and do not represent potential images the network would classify.

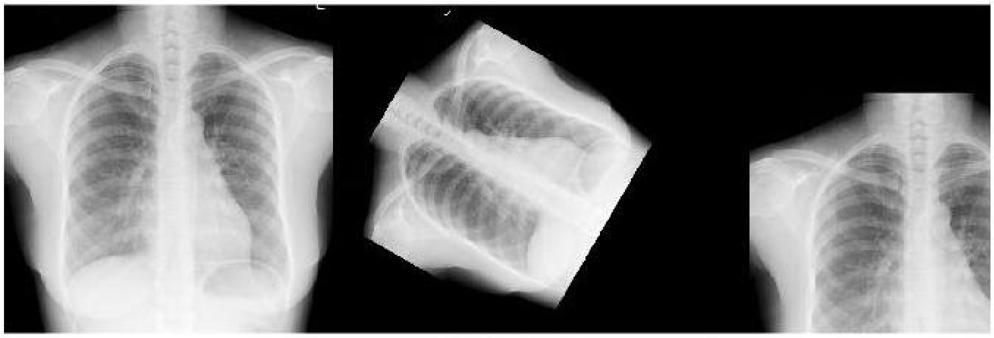


Figure 2.3: Image augmentation techniques applied to a radiograph that do not represent variations in the chest radiographs the network would need to classify in clinical usage.

2.2 Transfer Learning

Transfer learning is the use of a model pre-trained on other data for a new task. The model is often pre-trained on ImageNet [31] images. ImageNet is a collection of 3.2 million labeled color images from 5,247 categories. If the dataset used to pretrain the network is large enough and diverse enough, it can act as a generic model of the visual world and the features can be applicable to other tasks [5]; therefore, ImageNet is commonly used. The two primary types of transfer learning are fine-tuning and feature extraction. The following sections will discuss the concepts for each method.

2.2.1 Fine-Tuning

Fine-tuning, a form of transfer learning, is a widely used and relatively stable method of customizing a CNN for a new task. Fine-tuning begins with a pre-trained network, then chosen layers are unfrozen (i.e., made trainable), and a new classifier is added for the new classes needed for the new task. Then the new data are used to retrain the unfrozen layers and to train the classification layer. The weights of the unfrozen layers are adjusted to optimize the performance for the new task. Figure 2.4 is a schematic of fine-tuning a CNN.

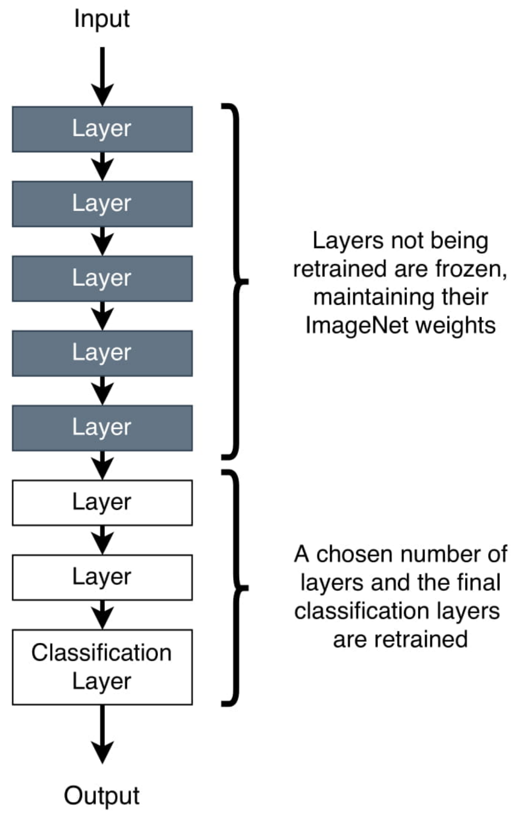


Figure 2.4: Schematic of the general concepts of fine-tuning: some layer weights are frozen, and chosen layers, as well as the final classification layer, are retrained for the new task.

A study from Tajbakhsh et al. [32] investigated the application of fine-tuning to medical imaging tasks and compared the performance of the fine-tuned AlexNet [29] network to one trained from scratch. The investigation was performed for four applications: polyp detection in colonoscopy videos (detection task), image quality assessment in colonoscopy videos (image classification task), pulmonary embolism detection in CT images (detection task), and boundary segmentation in ultrasonographic images (image segmentation task). They found that fine-tuning an AlexNet CNN pre-trained on ImageNet outperformed, or matched (in the worst case), the performance of the AlexNet CNN trained from scratch for the same task. They also found that fine-tuning is more robust to the size of the training set, effectively learning the task with fewer training images [32]. The ability to learn the task with fewer training images is advantageous for medical imaging applications, where there is a lack of large, annotated datasets.

2.2.2 Feature Extraction

Feature extraction uses a pre-trained network and the representations learned by the network to obtain the features corresponding to a new input. Features are extracted from chosen layers of a CNN. The features are arrays of numbers, not meaningful for human interpretation. The features are then put into a new untrained classifier, such as a support vector machine (SVM), to select features and create an accurate classifier. Figure 2.5 is a schematic of feature extraction of a CNN.

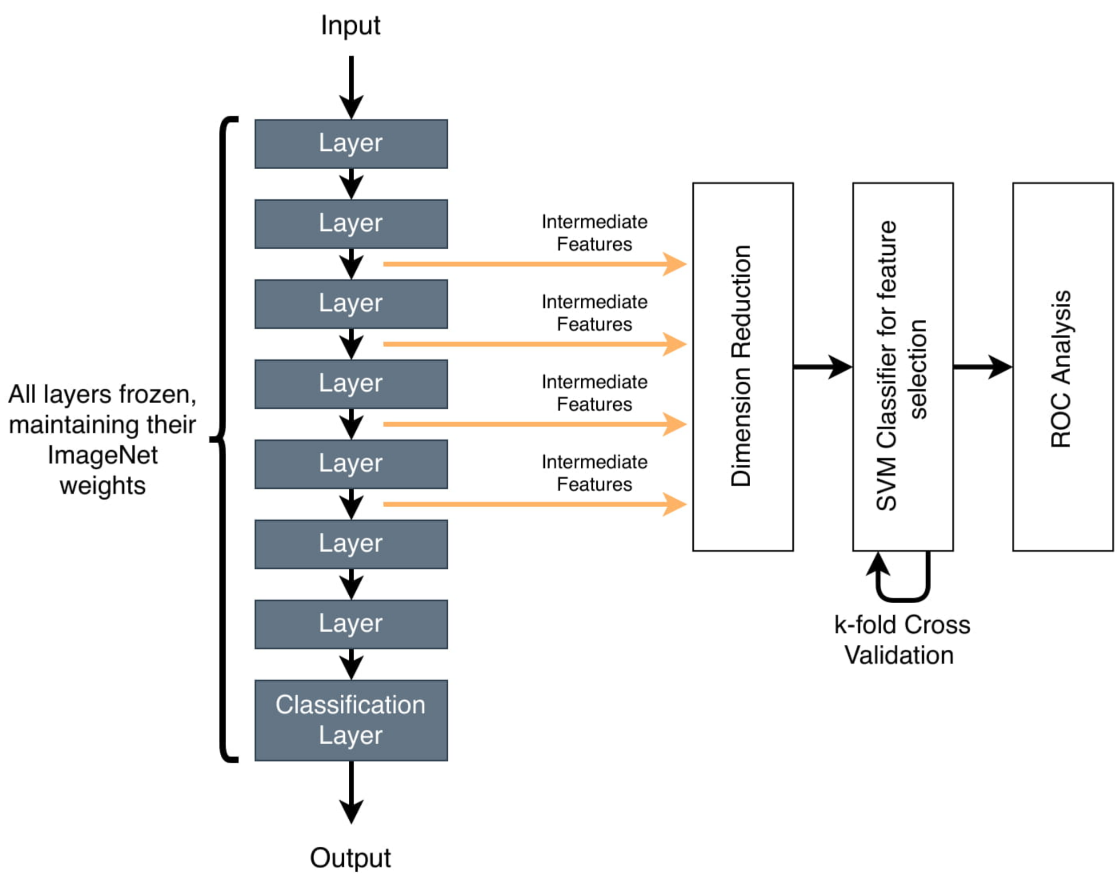


Figure 2.5: Schematic of the general concepts of feature extraction: intermediate values are extracted from chosen layers, dimensionality is reduced, then a classifier is used to select features, and ROC analysis is performed.

The selection of layers must be done thoughtfully since it is computationally intensive to store the extracted features and process them through a new classifier. The trained convolutional base's earlier layers extract local, generic features, such as textures and edges.

The later layers of the pre-trained model extract features more specific to the task it was trained to perform; for example, an ear of a cat for a model trained to classify between photos of dogs and cats [5]. Therefore, when using feature extraction on a new dataset that is unlike the images in the ImageNet dataset used for pre-training, it is likely that the earlier layers will provide the most useful features for the classification.

2.3 Network Architectures

2.3.1 AlexNet

AlexNet was the winner of the ImageNet challenge in 2012, using a deep convolutional neural network to classify a dataset of 1.2 million natural color images [29]. It achieved an accuracy of 85%, over 10% higher than the second-place entry; the human accuracy on the ImageNet database is 95%. Convolutional neural networks had been introduced and implemented prior to AlexNet in 2012 [33, 4]. The unique aspect of Alex Krizhevsky's CNN implementation was the use of GPUs (graphical processing units), which made deeper networks possible [29]. AlexNet is considered to have begun the latest deep learning boom due to its strong performance on the ImageNet challenge [34].

AlexNet's architecture consists of 5 convolutional layers and 3 fully connected layers for a total of 8 layers [29]. The required input image size to AlexNet is 256 by 256 pixels. The output of the last fully connected layer goes to a final classification layer, which has output nodes for the number of classification categories desired. Figure 2.6 is a schematic of the architecture of AlexNet.

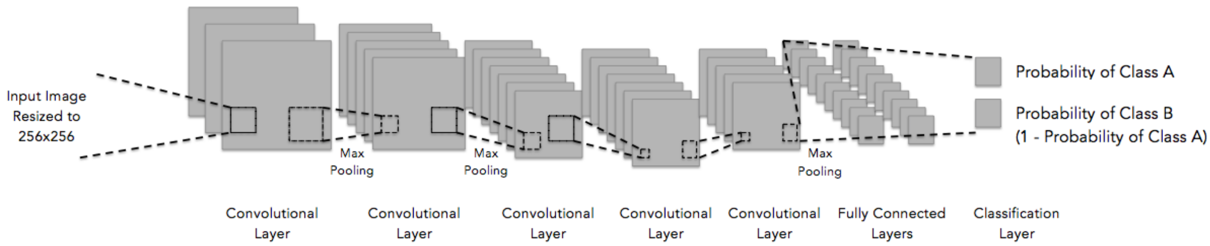


Figure 2.6: The architecture of AlexNet, which takes an input image resized to 256 x 256 pixels and outputs a probability score of belonging to each class. The probabilities of belonging to each class add up to 1 if using a softmax classifier.

AlexNet only has 5 convolutional layers; therefore, it can be used to establish a baseline in performance. Since it has fewer layers than other CNNs, it is less prone to overfitting if training with few images, therefore theoretically able to accurately generalize for high performance on an independent test set. The implementation does not require substantial computational resources relative to other networks due to the CNN's shallowness, therefore it should be simpler to implement in practice.

2.3.2 VGG19

VGG19 was the winner of the localization portion and received second place in the classification portion of the 2014 ImageNet challenge [35]. The primary innovation introduced by Simonyan et al. [35] was the use of 3 x 3 convolutional filters in each layer, as well as increasing the depth compared to prior networks, such as AlexNet. The depth was increased to 16 and 19 weight layers in VGG16 and VGG19, which were both released for public use.

The architecture of VGG19 consists of an input layer (224 x 224 pixel size), 5 convolutional blocks, with max-pooling layers between, for a total of 16 weight layers, plus 3 fully connected layers, and a final classification layer. The convolutional blocks consist of 3 x 3 filters, which is the smallest receptive field to capture the concepts of left/right, center, and up/down. The stride is 1 pixel. The max-pooling layers perform the max-pooling operation over a 2 x 2 pixel window with a stride of 2. The three fully connected layers have 4096, 4096, and 1000 channels, respectively. The last fully connected layer's channel number depends on the number of output classes. The total number of weights is 144 million [35].

VGG19 is a frequently used network for medical imaging tasks [3]. This is likely due to its relative depth; it is not too simplistic. However it is also shallower than other popular networks, meaning that it does not need as many images with which to train. The architecture is an improvement upon AlexNet and is publicly available, therefore it can be implemented without purchase of proprietary algorithms.

2.3.3 ResNet-50

ResNet won the 2015 ImageNet challenge with an accuracy of 95.5%, outperforming human classification performance (95%) on the 1.2 million natural color images in the ImageNet dataset. The innovation introduced by ResNet was the addition of residual blocks using identity mapping [36].

Prior to ResNet, the increasing performance from year to year on the ImageNet challenge was due to the addition of more layers. However, it was found that adding more layers to a CNN eventually leads to performance degradation in the form of higher training error [36]. This is known as the "vanishing gradient problem." To update the weights during training of a CNN, the loss (or error) function is backpropagated through the network. When backpropagating through a network with many hidden layers, each layer has a small derivative, causing the gradient to decrease such that when it gets to the earlier layers, it is too small to update the weights correctly. ResNet is built to preserve the gradient through the introduction of residual blocks (Figure 2.7) [36]. ResNet-50 consists of 16 residual blocks, each with 3 convolutional layers, and 2 pooling layers, in addition to an input (224 x 224 pixels) and output layer.

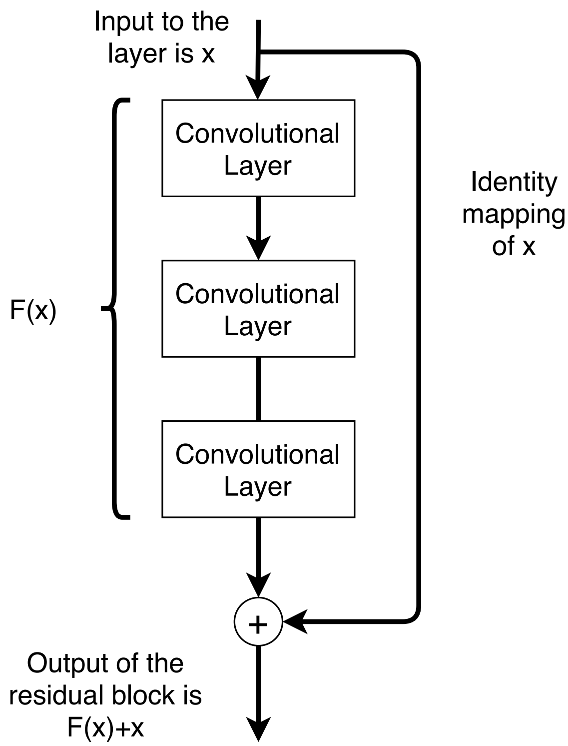


Figure 2.7: A residual block in ResNet-50 includes three convolutional layers and a skip connection that maps the input to the block to the output.

ResNet was found to be easier to optimize than non-residual CNNs and able to gain accuracy with increasing depth [36]. The residual block has an identity function so adding new layers will not negatively impact performance; regularization will skip over layers that do not increase performance. The identity function prevents the loss of information in those layers as well as preserving the gradient when backpropagating through the network to update weights during training. By using the residual blocks, He et al. [36] was able to increase the network depth 8 times deeper than VGG while still having lower complexity. ResNet-50 has 23 million parameters, compared to VGG19's 144 million parameters [36]. ResNet-50 is the ResNet architecture with the fewest parameters, therefore the most appropriate to train with a relatively small dataset, such as one obtained for a medical imaging task.

2.3.4 BagNet

BagNet is a variant of the ResNet-50 architecture that incorporates a unique approach. Brendel et al. [37] developed BagNet to address the lack of interpretability for most CNNs. The role of hidden neurons in CNNs depends on the roles of other neurons, making it difficult to understand the output due to all the interdependencies. Brendel et al. specifically configured their network to be as interpretable as possible [37]. To increase interpretability, BagNet uses the occurrences of small local image features, related to the bag-of-features models used prior to deep learning's tractability. Bag-of-features models do a spatial aggregate of the evidence from patches of an image; however, CNNs non-linearly integrate information from the whole image. BagNet was created to combine the concepts to create a high-performing and flexible CNN with the interpretability of a bag-of-features model.

BagNet does not take into account the spatial ordering of the features. Since it focuses on local features, the analysis of how each part of the image influences the classification is relatively straightforward, leading to an output of class probability score as well as a heatmap showing the class evidence extracted from each part of the image. BagNet is a variant of the ResNet-50 architecture, simply changing the 3x3 convolutions to 1x1 convolutions to limit the receptive field size of the topmost convolutional layer to 33 x 33 pixels. BagNet extracts class activations on each 33 x 33-pixel patch of the image. Each patch has a heatmap that is summed up over all of the patches; accumulating the activations yields the total evidence for the final classification. Through experiments on the ImageNet dataset, Brendel et al. found that most class evidence was around the shapes of objects, and BagNet outperformed AlexNet but not VGG16 [37]. BagNet's ideal use is in applications where it is desirable to trade accuracy for better interpretability.

Interpretability of deep learning output for medical imaging tasks is of great importance for eventual clinical implementation [7]. BagNet's unique structure, allowing it to output heatmaps of class evidence, rather than gradients, allows for a more intuitive understanding of the output. BagNet is suggested for application in medical imaging tasks, even though the authors did not perform experiments on any medical imaging datasets [37].

2.4 Evaluation of Deep Learning Network Performance

There are two primary methods to evaluate the performance of a trained deep neural network: cross-validation and use of an independent test set [5]. Cross-validation refers to the use of a portion of the dataset being used for training and a portion for testing, which rotates to a different part of the dataset, leading to the use of many folds for the cross-validation. The use of an independent test set consists of testing the trained CNN using images independent from the images used for training/validation.

There are many metrics used to quantify the performance of the test set including accuracy, area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, specificity, precision-recall curve, and more. AUC [38] will be used in this work due to its robusticity for unbalanced datasets. In addition, AUC is often the performance metric chosen for medical imaging applications with deep learning so performance can be directly compared to similar studies. Specific performance evaluation and statistical analysis for each method will be provided in each of the following chapters.

2.5 Interpretability of Deep Learning for Radiology

In addition to the network performance evaluation, the ability of the algorithm to explain the result or provide interpretable output is of interest. Deep learning, due to its many hidden layers and its ability to learn without explicit rules being programmed, is often considered to be a “black box”. Deep learning uses high dimensional functions that cannot be characterized with simple terms [7]. Deep learning’s status as a “black box” may be adequate for some applications; however, there is growing concern about being able to explain the output of deep learning to gain clinician confidence in the output results [7]. In addition, explainability of the output could help in identifying biases in the network’s performance and allow the developers to address those issues.

Interpretability of results is important for clinical adoption of deep learning. There are techniques to visualize the most influential aspects of an image for the classification assigned by the trained network. More discussion of interpretability and visualization methods is presented in Chapter 5.

2.6 Discussion & Conclusions

The techniques for training and applying deep learning: training from scratch, fine-tuning, and feature extraction, have their own advantages and disadvantages. The deep learning technique should be chosen to best conform to the needs of a given task. For medical imaging applications of deep learning, issues of dataset size and interpretability are of particular importance. There is generally a lack of large, annotated datasets of medical images, making transfer learning often a more suitable option for application to medical imaging tasks. In addition, for eventual clinical implementation of the deep learning algorithms, explanation of the output is needed to increase clinician confidence and identify biases. The four CNN architectures introduced are summarized in Table 2.1 and were selected to apply a variety of CNN architectures to the tasks. The following chapters will discuss applications of these deep learning methods for workflow enhancement, improved diagnosis, and explanability of deep learning output.

Table 2.1: Summary of the four CNN architectures discussed in the chapter, including the year they were released, number of weights, number of weight layers, notable characteristic(s) and why they were selected to include in this work.

Network Architecture	**Number of Weights**	**Number of Weight Layers**	**Notable Characteristic(s)**	**Motivation for Selection**
AlexNet (2012) [29]	61 million	8	First CNN that used GPU technology	Relatively shallow CNN to establish a baseline in performance
VGG19 (2014) [35]	144 million	19	Use of 3x3 convolutional filters, increased depth	Deeper CNN, strong performance for other medical imaging tasks
ResNet-50 (2015) [36]	23 million	50	Residual block design utilizing identity mapping	Innovative design to address the vanishing gradient problem
BagNet (2019) [37]	18 million	50	Related to bag-of-features models, uses number of occurrences of local image features	Designed for increased interpretability, visualization component of the output

CHAPTER 3

INVESTIGATION OF DEEP LEARNING METHODS IN THE TASK OF CLASSIFYING THORACIC RADIOGRAPHIC VIEWS

3.1 Introduction & Motivation

Chest radiography studies are common, with 68 million ordered annually as of 2018 [8]. A typical chest radiography study consists of a frontal radiograph, as well as a lateral chest radiograph. However, some radiology departments use dual-energy imaging in order to obtain two more images in the study: a bone image and a soft-tissue image. The four views resulting from a dual-energy chest radiography study are shown in Figure 3.1. The type and view of each image in the study is typically specified in the DICOM header in order to enable accurate hanging protocols [39]. The DICOM header may be incomplete or incorrect, leading to incorrect classification of image view. Images acquired in the same study may have identical DICOM headers or inconsistent differences in header fields, making it difficult to identify each image's view in an automated fashion.

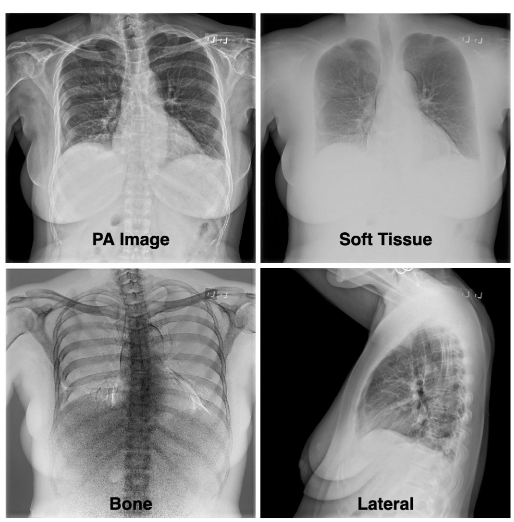


Figure 3.1: The four radiographic views resulting from a dual-energy chest radiography study: frontal, lateral, soft tissue, and bone.

DICOM descriptors are used in clinical practice to transfer images and hang them appropriately to enable efficient interpretation. In addition, the DICOM header information is applied for the display, post-processing, and storage of the acquired images. To ensure smooth clinical workflows and apply appropriate post-processing techniques, accurate view classification of the acquired images is needed.

The frontal images in each chest radiography study can be acquired AP or PA, which affects the appearance and measured dimensions of pathologic findings, such as tumors, as well as thoracic structures, such as the heart [21]. Whether the image was acquired AP or PA is typically specified in the DICOM header. In our institutional dataset, 14% of the images lacked the specification of AP or PA in their DICOM header information. To distinguish between AP and PA images, a radiologist uses visual cues, such as the clavicle and scapula, the heart shadow, or markers placed by the technologist.

A CNN that could classify radiograph view efficiently without disrupting clinical workflow would be a valuable addition to radiology departments. The classification of frontal versus lateral radiographs by a CNN was published by Rajkomar et al. [40]. Their method was to pretrain neural networks and fine tune them for the classification of frontal versus lateral radiographs. However, this work [41] applies a CNN to dual-energy imaging, classifying them between frontal, lateral, soft tissue, and bone. The purpose of this work was to train CNNs from scratch for the task of distinguishing between the four views resulting from a dual-energy chest radiography study, not only between frontal and lateral. Beyond a clinical workflow benefit, CNNs trained to classify chest radiograph view could be a valuable tool for researchers compiling large databases from various sources.

3.2 Methods

3.2.1 Dataset

In order to train the CNNs from scratch to classify radiograph view, three datasets were obtained and formulated. Two of the datasets were formulated using radiographs from University of Chicago Medical Center: (i) a training/validation/test (TVT) set and (ii) an independent testing set referred to as “ITS”. All the radiographs were retrospectively obtained under a HIPAA-compliant, IRB-approved protocol. The radiographs were obtained from existing research datasets, as well as from a teaching set. The radiographs were not selected specifically for the view classification tasks. The third dataset (NIHTS) was comprised of 1,000 radiographs from the publicly released ChestX-ray14 dataset from NIH Clinical Center [11]. The patient characteristics and primary disease states are listed in Table 3.1 for the three datasets used.

Table 3.1: Patient and image characteristics for the thoracic radiographs used for view classification via deep learning methods.

		Train/Validation/Test Set (TVT) (827 imaging studies, 1,910 images, 310 patients)	**Independent Test Set ("ITS") (1,368 imaging studies, 3,757 images, 262 patients)**	**NIH Testing Set (NIHTS), (1,000 imaging studies, 1,000 images, 612 patients)**
Sex	**Female**	131 (42.3%)	107 (40.8%)	236 (38.6%)
	Male	179 (57.7%)	155 (59.2%)	376 (61.4%)
Age	**Mean**	58.5	58	51
	Standard Deviation	17.5	17.6	14.1
Devices (Number of Imaging Studies)	**Leads/Catheter/Tube**	516 (62.4%)	1,496 (79.4%)	621 (62.1%)
	Device	80 (9.7%)	255 (13.5%)	111 (11.1%)
	Surgical Metal	198 (23.9%)	566 (30%)	62 (6.2%)
	None	237 (28.7%)	267 (14.2%)	332 (33.2%)
Primary Disease State (Number of Imaging Studies)	**No Significant Abnormality**	111 (13.4%)	70 (5.1%)	515 (51.5%)
	Pleural Effusion	66 (8%)	55 (4%)	49 (4.9%)
	Emphysema	11 (1.3%)	7 (0.5%)	10 (1%)
	Pneumothorax	575 (69.5%)	1,182 (86.4%)	15 (1.5%)
	Other: Diffuse lung disease, cardiomegaly, consolidation, edema, fibrosis, hernia, infiltration, pleural thickening, pneumonia	9 (1.1%)	0 (0%)	140 (14%)
	Airspace Opacity	22 (2.7%)	16 (1.2%)	0 (0%)
	Lung Nodule	10 (1.2%)	5 (0.4%)	30 (3%)
	Atelectasis	23 (2.8%)	33 (2.4%)	241 (24.1%)
Vendors Identified in DICOM Header (Number of images)	**Fujifilm**	1,099 (57.5%)	1,792 (47.7%)	No vendors identified in publicly released ChestX-ray8 dataset
	Canon	55 (2.9%)	1,035 (27.5%)	
	GE	729 (38.2%)	877 (23.3%)	
	Riverain	27 (1.4%)	19 (0.5%)	
	Philips	0 (0%)	26 (0.7%)	
	Siemens	0 (0%)	8 (0.2%)	
Frontal and Associated Images (Number of Images)	**Frontal (AP/PA)**	818 (42.8%)	1,368 (36.4%)	1,000 (100%)
	Bone, Lateral, and Soft-tissue Images	1,092 (57.2%)	2,389 (63.6%)	0 (0%)

The TVT dataset consisted of 1,910 thoracic radiographs from 827 imaging studies from 266 patients at the University of Chicago Medical Center. The TVT dataset was formulated from a teaching set and two research datasets. The teaching set was generated by selecting PTX cases, in addition to false-positive cases, between February 2006 and March 2017. After PTX cases were selected, radiographs for the same patients either prior to or after the date of the radiograph demonstrating PTX were obtained for inclusion in the dataset. The radiographs in the two research datasets were found by searching radiology reports for chest radiographs with PTX between July 2016 and February 2017, as well as between May 2015 and June 2016. Again, radiographs for the same patients prior to or after the date of the radiograph demonstrating PTX were included in the dataset. Overall, the acquisition dates ranged from February 2006 to February 2017, and the images were from four radiographic imaging equipment manufacturers. The chest radiographs had a variety of primary disease states, such as: PTX, atelectasis, airspace opacity, pleural effusion, emphysema, lung nodules, and diffuse lung disease. Many of the radiographs also included medical devices such as pacemakers, catheters, chest tubes, surgical markers, and metal implants. None of the radiographs were excluded due to the presence or absence of devices or their primary disease state.

The "ITS" dataset used as an independent test set consisted of 3,757 thoracic radiographs from 1,368 imaging studies of 262 patients at the University of Chicago Medical Center. The "ITS" was formulated from a research dataset that was generated by searching radiology reports for chest radiographs with PTX between January 2013 and April 2015. Radiographs for the same patients prior to or after the date of the radiograph demonstrating PTX were included in the dataset if they were available. Note that none of the patients in the "ITS" had images included in the TVT. Overall, the "ITS" consisted of images acquired between August 2007 and February 2017 and from 6 different vendors' equipment.

The third dataset, the NIHTS, was comprised of 1,000 frontal radiographs (500 AP and 500 PA) from 612 patients drawn from the ChestX-ray14 dataset. View information (AP or PA) for the images was released by NIH Clinical Center with the images [11].

3.2.2 Methods for the 4-way Radiograph View Classification

When the DICOM header information was used to classify radiograph view for the 1,910 radiographs in the TVT dataset, 38% were left unclassified. For the development of the CNN approach for view classification, the 1,910 radiographs in the TVT dataset were manually classified into the 4 views: frontal (818 images), lateral (418 images), soft tissue (389 images), and bone (285 images). A radiologist with 18 years of clinical experience and 28 years of research experience verified the manual view classification. After the truth for the radiograph views was established, the TVT dataset was used to train a CNN from scratch.

To train the CNN from scratch, a NVIDIA Deep Learning GPU Training System (DIGITS 6.0.0, NVIDIA Corporation, Santa Clara, CA) was used. A Tensorflow framework (Tensorflow-gpuV1.2.1, Google LLC, Mountain View, CA) was used to train a CNN with AlexNet architecture [29] (Figure 3.2) from scratch on an NVIDIA Titan X Pascal GPU. Training from scratch is a three-step process: training, validation to tune the configuration of the network, and testing after training/validation is complete. The three steps will be referred to as the "training, validation, testing" method. The highest class likelihood output by the trained network was used as the network-assigned view for each test case.

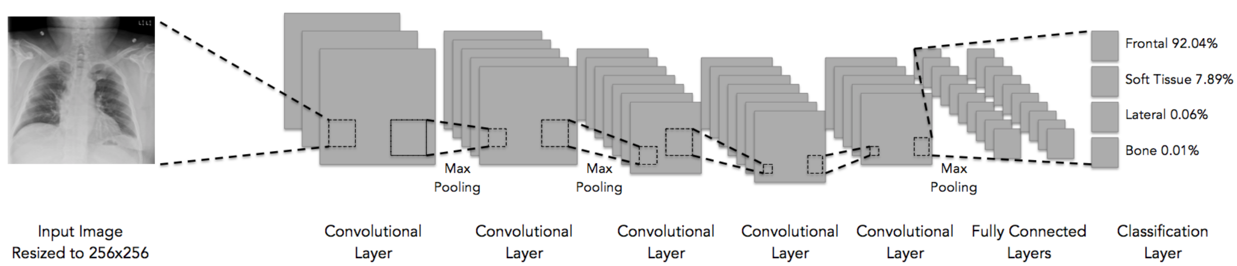


Figure 3.2: Schematic of a thoracic radiograph from the test set input to an AlexNet-architecture CNN trained from scratch with four output nodes corresponding to the four image view types. The highest percent likelihood was used as the assigned view for each test set image. In this case, the test image would be assigned to "frontal" since it has the highest percent likelihood (90.89%).

The TVT dataset was used for training (1,242 radiographs, 65%), validation (382 radiographs, 20%), and testing (286 radiographs, 15%) of the trained network. This dataset was used for training CNNs to classify between frontal vs. all else and frontal vs. lateral (binary classifications), as well as a 4-way classification of image type. Bilinear interpolation was used to downsample the images to 256 x 256 pixels for input to the CNN, and training was performed for 25 epochs, with a validation interval of 1 epoch. The base-learning rate was set to 0.001 and adaptive gradient was used as the solver type, since it was found to have higher performance than other solver types for this task. Data augmentation was performed to double the images used for training the CNN. During training, horizontal flipping was randomly applied to the training images in each epoch.

3.2.3 Methods for the Classification of Radiographs Acquired AP vs. PA

Using only DICOM header information to classify the frontal radiographs into those acquired AP and those acquired PA left 311 of the 2,186 (14%) radiographs unclassified. The 2,186 frontal (AP and PA) radiographs from the TVT and "ITS" datasets were manually classified and verified by a radiologist with 18 years of clinical experience. The 2,186 frontal radiographs (1,475 AP and 711 PA) were used to train a CNN for the classification between frontal radiographs acquired AP and PA. There were too few AP and PA images in the TVT alone to use for training a network from scratch; therefore, the frontal images from the "ITS" were also used for the training dataset.

To train a CNN to classify between AP and PA radiographs, the same network architecture (AlexNet) and training parameters as the 4-way view classification were used. The data augmentation via horizontal flipping was not included in the CNN training for the AP vs. PA classification, since frontal radiographs, both AP and PA, are oriented such that the heart is on the right side of the image; therefore, training the CNN with frontal radiographs with horizontal flip would not represent a potential variation for the classification between AP and PA radiographs. 65% of the images were used for training (1,421 radiographs), 20% for step-wise validation (437 radiographs), and 15% for testing (328 radiographs).

Chest radiographs may be imprinted with a label, such as those acquired on a mobile x-ray machine, which are imprinted with labels such as "port" or "portable." The CNN trained to classify between AP and PA radiographs could possibly learn how to read the labels or use the presence or absence of labels for classification. To address this, an additional study was performed to evaluate the impact of imprinted labels when training a CNN. The images with labels were cropped to remove the label. Then the cropped images and the images without imprinted labels were used for training a CNN with the same architecture and training parameters as the CNN trained with the uncropped radiographs. The dataset consisting of images without imprinted labels and the images cropped to remove the labels will be referred to collectively as "cropped images" for the remainder of this chapter.

3.2.4 Independent Clinical Evaluation & Statistical Analysis

After the "training, validation, and testing" was performed to train the CNNs from scratch, the CNNs were tested using independent test sets. The three datasets and how they were used for each task are shown in Table 3.2. The network trained to classify among the four views resulting from a dual-energy study was independently tested using the 3,757 radiographs in the "ITS." The networks trained to classify between radiographs acquired AP and PA were tested with the 1,000 radiographs in the NIHTS, since the frontal radiographs from the TVT and "ITS" datasets were used for training the CNNs from scratch.

Table 3.2: The three datasets and their usage for each of the tasks reported in Chapter 3.

Task	**Dataset**		
	TVT	**"ITS"**	**NIHTS**
4-way classification	Used for training/validation/testing	Used as an independent test set	Not Used
AP vs. PA	Frontal (AP & PA) radiographs used for training/validation/testing		Used as an independent test set

To evaluate the performance of the CNNs trained from scratch, receiver operating characteristic (ROC) analysis was performed. The area under the ROC curve (AUC) was used as a performance assessment metric [38] and 95% confidence intervals were calculated. DeLong's method [42] was used to compare ROC curves and calculate the p-value to determine whether there was a statistically significant difference in performance. To further evaluate performance, the percentage of images misclassified and the relative frequency distributions of percent likelihood for each view were examined.

3.3 Results

3.3.1 Results of the 4-way Radiograph View Classification

For the binary classification between frontal radiographs and all other views, the trained CNN yielded an AUC of 0.997 (95% CI: 0.996, 0.998) on the "ITS." The trained network yielded an AUC of 0.9998 (95% CI: 0.9993, 0.9999) on the "ITS" for the binary classification between frontal and lateral radiographs. For the frontal vs. lateral classification, 12 radiographs from 11 patients were misclassified of the 2,601 images. Of the 12 misclassified radiographs, 3 were in landscape orientation, one had a metal shoulder implant, and 2 had a large portion of the abdomen in the radiograph, comprising greater than 50% of the image. The lowest AUC when the classifications were assessed pairwise was for the classification between the frontal and soft-tissue radiographs, 0.998 (95% CI: 0.997, 0.999). For the frontal vs. soft tissue classification, 49 images from 39 patients were misclassified of the 2,524 images. 51% of the misclassified radiographs were soft-tissue images that had been generated from poor energy subtraction, resulting in visible bony anatomy. Other misclassified radiographs were due to high image noise (n=8), a large portion of the skull in the image (n=1), and an abdomen image. The binary classification combinations and their performances are shown in Table 3.3.

Table 3.3: Performance of the binary combinations from classifications on the “ITS” (3,757 images) given by AUC.

Task		AUC for “ITS”	95% Confidence Interval
AP/PA	vs. All Else	0.9972	(0.9960, 0.9980)
AP/PA	vs. Lateral	0.9998	(0.9993, 0.9999)
AP/PA	vs. Soft Tissue	0.9979	(0.9967, 0.9987)
AP/PA	vs. Bone	0.9991	(0.9982, 0.9996)
Lateral	vs. Soft Tissue	0.9999	(0.9969, 1.0000)
Lateral	vs. Bone	0.9996	(0.9957, 1.0000)
Bone	vs. Soft Tissue	1.0000	(0.9947, 1.0000)

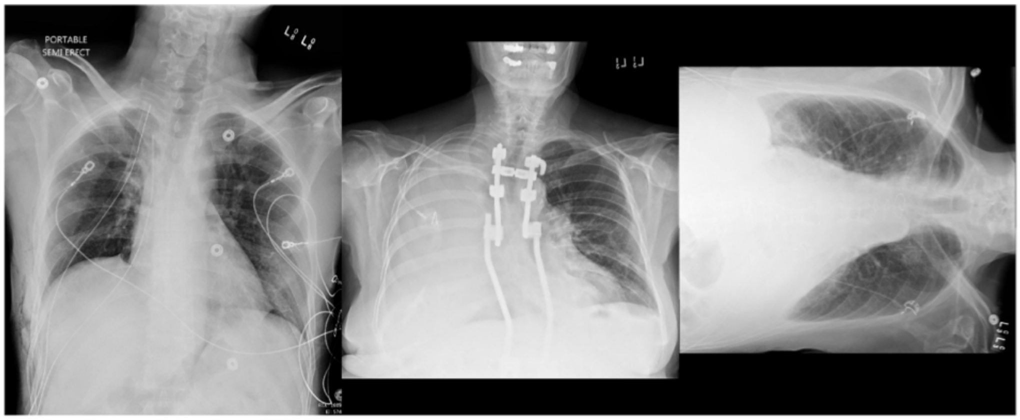
For the 4-way classification, the CNN had a training accuracy of 99.18% when training was completed. The trained CNN reported an accuracy of 98.19% on the “ITS.” The percentage of each “ITS” radiograph view classified into the 4 categories is shown in Table 3.4. For the 4-way classification, the 57 images misclassified of the 3,757 images included 22 frontal radiographs, 5 lateral radiographs, 24 soft-tissue images, and 6 bone images. The prevalence of images with PTX in the “ITS” dataset is higher than the prevalence of other lung conditions (Table 3.1). To verify that the presence of PTX did not influence the classification, the prevalence of misclassified images with PTX was compared to the overall prevalence in the “ITS” dataset; 84% of the misclassified radiographs (48 of 57) had PTX, which is similar to the overall prevalence of PTX in the “ITS” dataset (86%.) Example frontal and lateral radiographs from the “ITS” with their percent likelihoods output by the network are shown in Figure 3.3. Figure 3.4 shows examples of frontal images from the “ITS” and their percent likelihood output by the trained CNN. Figures 3.5, 3.6, and 3.7 show example images from the “ITS” and the percent likelihood assigned by the trained CNN for lateral, soft-tissue, and bone images, respectively. To further evaluate the 4-way classification performance, relative frequency histograms for each of the 4 classes and their percent likelihoods is given in Figure 3.8.

Table 3.4: Results in terms of the percentage of images classified into each image view category by the trained CNN.

		Class Assigned by Trained CNN			
		Frontal	**Lateral**	**Soft Tissue**	**Bone**
Truth	**Frontal**	**98.2%**	0.2%	0.7%	0.9%
	Lateral	0.5%	**99.2%**	0.3%	0%
	Soft Tissue	4.2%	0%	**95.8%**	0%
	Bone	0.5%	0.3%	0.2%	**99%**

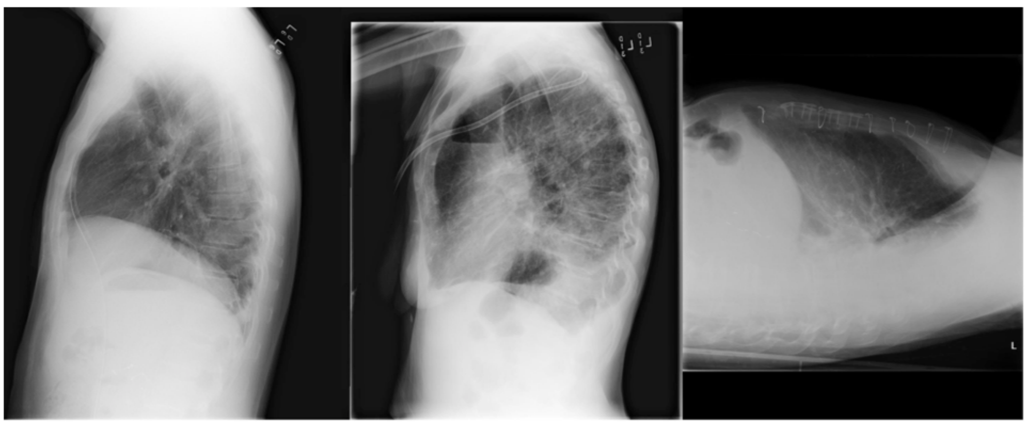
		Percent Likelihood Output from Trained CNN				
Image	*Truth*	*Frontal*	*Lateral*	*Soft Tissue*	*Bone*	*Correctly Classified?*
	Frontal	**100%**	0%	0%	0%	Yes
	Frontal	**54.3%**	0%	45.7%	0%	Yes
	Frontal	39.3%	0.1%	**60.6%**	0%	No
	Lateral	**74.9%**	8.4%	16.5%	0.2%	No

Figure 3.3: Example image views from the “ITS” with output of the percent likelihoods from the trained network.



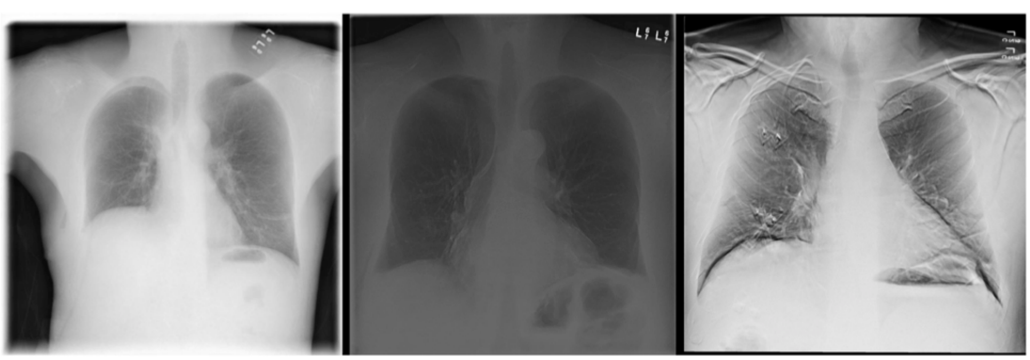
100% 50% 1%

Figure 3.4: Example frontal images with their likelihood of being frontal.



100% 52% 15%

Figure 3.5: Example lateral images with their likelihood of being lateral.



100% 50% 0%

Figure 3.6: Example soft-tissue images with their likelihood of being soft tissue.

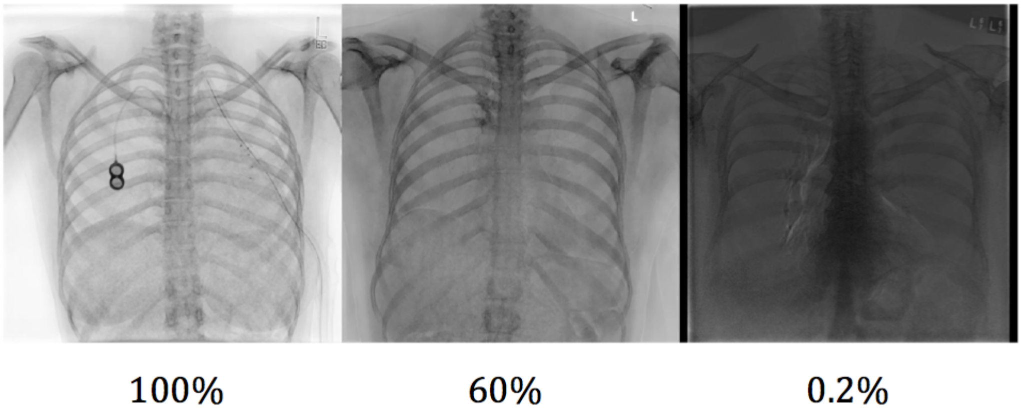


Figure 3.7: Example bone images with their likelihood of being bone.

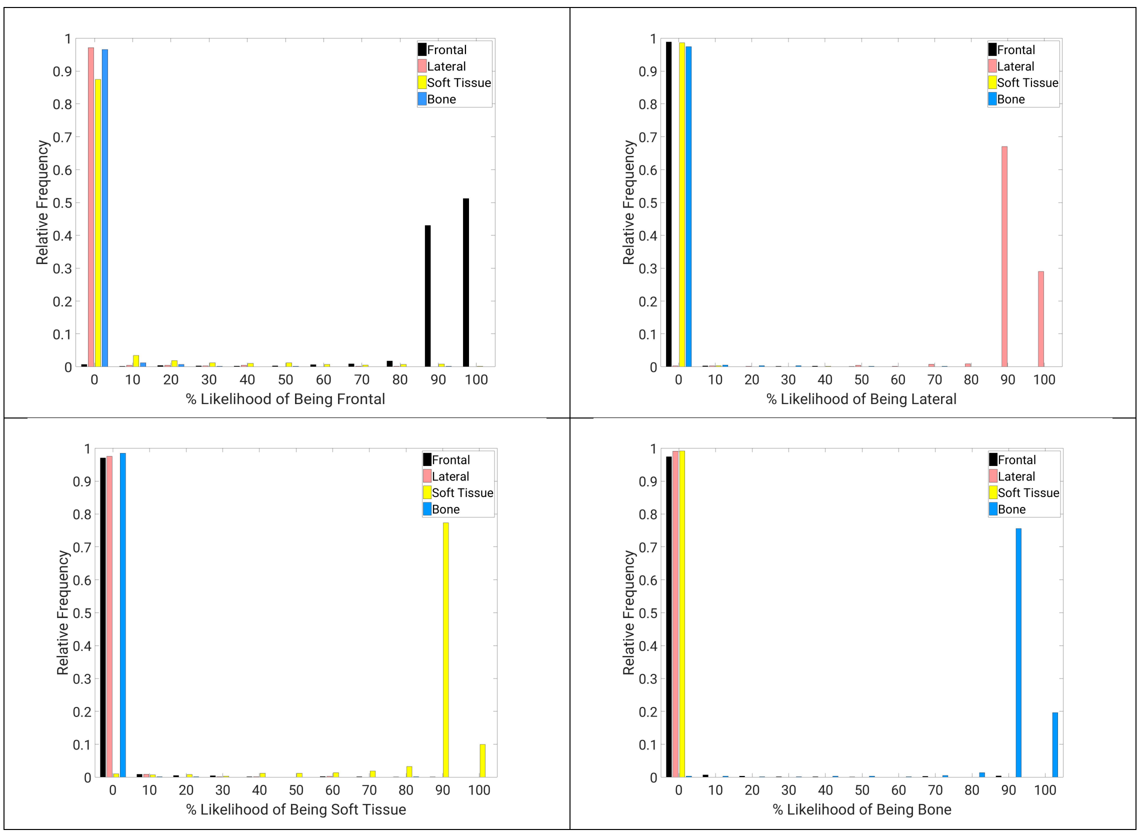


Figure 3.8: Histograms of percent likelihood for the four-way classification.

Since the 1,910 radiographs in the TVT set were manually classified into the four views, the time required to manually classify the 3,757 images of the "ITS" could be estimated. The time to manually classify the "ITS" was estimated to be 11.6 hours, if performed at the same rate as the manual classification of the TVT. For the classification via deep learning, loading the 1,910 radiographs of the TVT for training took 57 seconds, training/validating/testing the CNN took 229 seconds, and classifying the 3,757 images in the "ITS" took 171 seconds.

3.3.2 Results of the AP vs. PA Radiograph Classification

The NIHTS was used as an independent test set to test the classification between AP and PA radiographs, yielding an AUC of 0.973 (95% CI: 0.961, 0.981). For the 1,000 test set images, 92 were misclassified. None of the 92 radiographs had PTX, which is similar to the overall PTX prevalence in the NIHTS (1.5%). Training the CNN from scratch (training, validation, and testing) took 267 seconds, and classifying the NIHTS was performed in 130 seconds.

For the investigation of the impact of imprinted labels, the prevalence of labels and their wording was analyzed. The AP images from the TVT and "ITS" were found to have 12 unique imprinted label wordings, with the two most common being "Semi Erect" (20%) and "Portable" (44%). 1,193 of the 1,475 AP images had imprinted labels (81%). For the PA images from the TVT and "ITS," only 4%, 27 of the 711 images, had imprinted labels. There were 4 unique imprinted label wordings, with "upright" being the most frequent. 119 images of the 500 APs in the NIHTS had imprinted labels, with the most frequent being "Portable" (24%) and "AP" (18%). There were no imprinted labels on any of the 500 PA images in the NIHTS. The distribution of label wording for the AP images in the TVT, "ITS," and NIHTS datasets are shown in Table 3.5.

Table 3.5: Distribution of label wording for APs in the NIHTS, TVT and "ITS."

Label Wording	**APs in the TVT and "ITS" Set (1,475 total)**	**APs in the NIHTS Set (500 total)**
Portable	649 (44%)	119 (24%)
AP	17 (1%)	90 (18%)
Supine	30 (2%)	14 (3%)
Sitting	23 (2%)	2 (0.4%)
Semi-Upright	57 (4%)	1 (0.2%)
Erect	122 (8%)	1 (0.2%)
Semi Erect	301 (20%)	2 (0.4%)
Other	616 (42%)	3 (0.6%)
Total	**1,193 (81%)**	**119 (24%)**

After the CNN was retrained for the task of classifying between AP and PA radiographs with the cropped images (i.e., images with labels removed and images originally without labels), it yielded an AUC of 0.946 (95% CI: 0.931, 0.958) on the NIHTS. The percentage of radiographs classified into each of the categories by the network trained with the original radiographs and the network trained with cropped radiographs is given in Table 3.6 and Table 3.7, respectively. The relative frequency distributions of the percent likelihood assigned by the network trained with unaltered images and the network trained with cropped images are shown in Figure 3.9.

Table 3.6: Percentage of images classified as AP and PA by the CNN trained with original uncropped radiographs. The AUC from the independent evaluation was 0.973 (95% CI: [0.961, 0.981]) in the task of classifying between AP and PA images.

		Class Assigned by CNN Trained with Original Uncropped Radiographs	
		AP	**PA**
Truth	**AP**	**93.6%**	6.4%
	PA	8.2%	**91.8%**

Table 3.7: Percentage of images classified as AP and PA by the CNN trained with cropped radiographs. The AUC from the independent evaluation was 0.946 (95% CI: [0.931, 0.958]) in the task of classifying between AP and PA images.

		Class Assigned by CNN Trained with Cropped Radiographs	
		AP	**PA**
Truth	**AP**	**89.4%**	10.6%
	PA	12.2%	**87.8%**

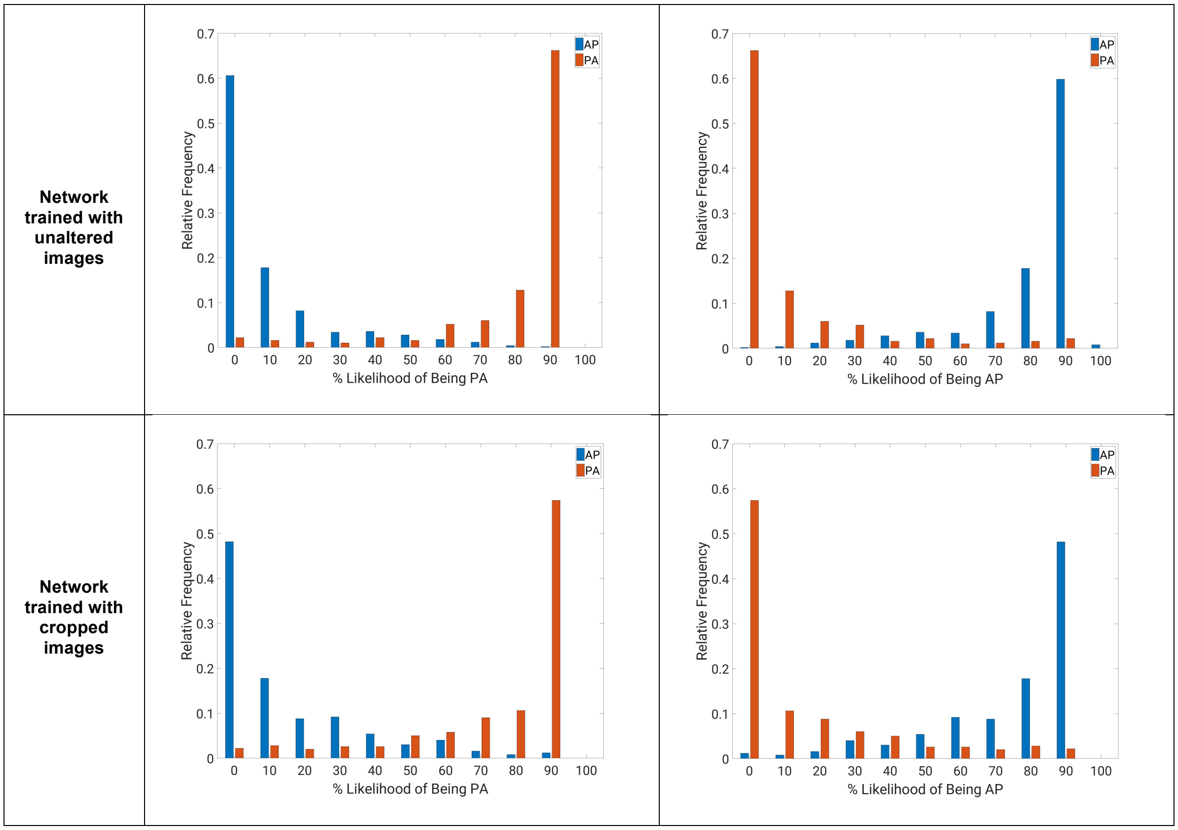


Figure 3.9: Histograms of percent likelihood for the AP vs. PA classification for the network trained with unaltered images and the network trained with cropped images.

The AUCs from the test set from both CNNs, one trained with the original radiographs and one trained with the cropped images, were compared to evaluate the impact of imprinted image labels on the training of a CNN from scratch. DeLong's method was used to calculate the p-value and determine whether there was a statistically significant ($p<0.05$) difference between the AUCs from the two networks. The p-value was calculated to be <0.001, showing a statistically significant decrease in performance for the network trained with the cropped images (0.946 vs. 0.973 with the uncropped radiographs). The ROC curves for both CNNs are shown in Figure 3.10.

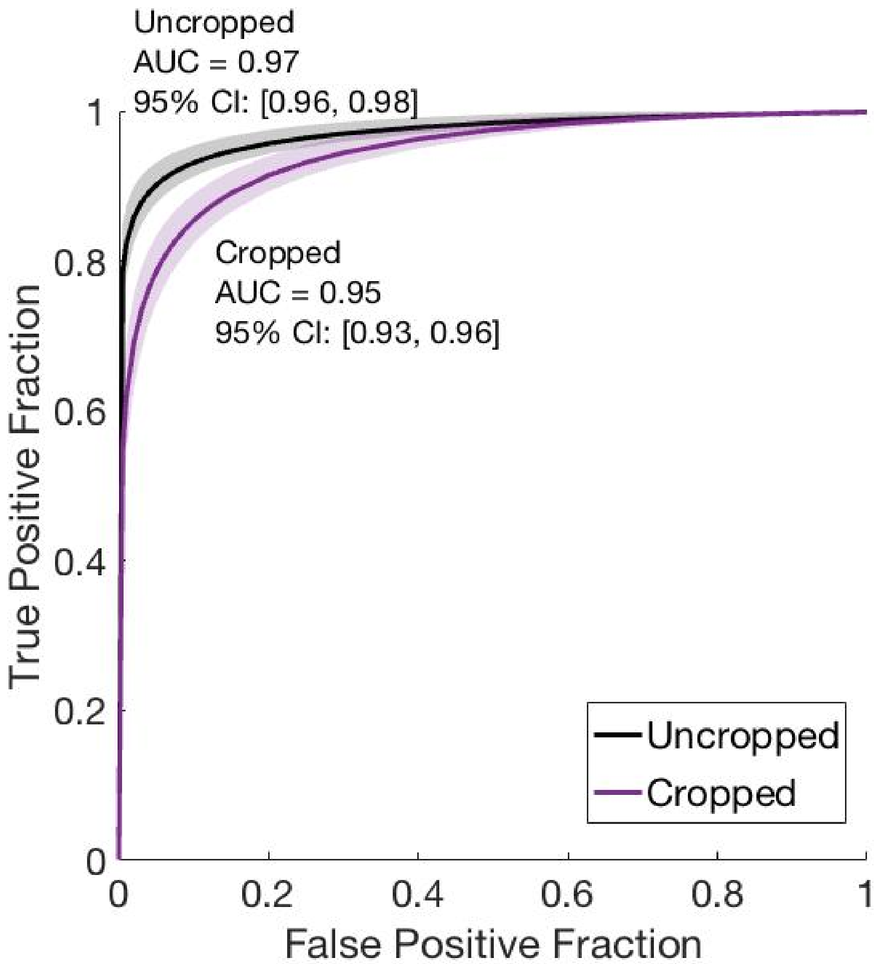


Figure 3.10: ROC curves for convolutional neural networks trained with original uncropped and cropped images in the task of classifying between AP and PA chest radiographs. The shaded area represents the 95% confidence interval for the curve.

To further evaluate the impact of imprinted image labels, ROC analysis was performed separately for the 119 AP images with imprinted labels and the 381 AP images without imprinted labels in the NIHTS. The ROC curves were compared using DeLong's method [42] to calculate the p-value to determine whether there was a statistically significant difference in performance. For the task of distinguishing the 119 AP images with imprinted labels from the 500 PA images, the CNN trained with the unaltered dataset yielded an AUC of 0.967 (95% CI: 0.954, 0.979), and the CNN trained with the cropped dataset yielded an AUC of 0.932 (95% CI: 0.906, 0.953). The p-value was calculated to be <0.001. For the task of distinguishing the 381 AP images without labels from the 500 PA images, the CNN trained with the unaltered dataset yielded an AUC of 0.975 (95% CI: 0.964, 0.982), and the CNN trained with the cropped dataset yielded an AUC of 0.952 (95% CI: 0.937, 0.963). The p-value was calculated to be <0.001. Both groups had a statistically significant difference when tested with the CNN trained with the cropped images. The AUC and p-values for both groups are given in Table 3.8.

Table 3.8: AUC values from ROC analysis in the task of distinguishing between AP and PA images, performed separately for the group of AP images with labels and the group of AP images without labels from the NIHTS. The p-values for the differences between AUCs are reported. The p-values were calculated using DeLong's method to determine whether the two ROC curves were statistically significantly different.

Trained Network	**119 AP with labels vs. 500 PA without labels**	**381 AP without labels vs. 500 PA without labels**	**p-value for the difference between the AUCs per each network (Across row)**
Trained with Unaltered Images	0.967 95% CI: (0.954, 0.979)	0.975 95% CI: (0.964, 0.982)	0.63
Trained With Cropped Images and Images Without Labels	0.932 95% CI: (0.906, 0.953)	0.952 95% CI: (0.937, 0.963)	0.27
p-value for the difference in AUCs between the two networks (down column)	<0.001	<0.001	

When testing with the same CNN, we failed to show a statistically significant difference in performance between the 119 APs with labels and 381 APs without labels. Two images that have similar label wording that were classified differently by the network trained with cropped images are shown in Figure 3.11. Figure 3.12 gives an example image where cropping to remove the imprinted label may have caused the removal of useful anatomical information for the classification.

Image	**Percent Likelihood from Network Trained with Original Uncropped Images**		**Percent Likelihood from Network Trained with Cropped Images**		*Correctly Classified by both networks?*
	AP	*PA*	*AP*	*PA*	
	**90.0%**	10.0%	**90.2%**	9.8%	Yes
	**53.4%**	46.6%	42.1%	**57.9%**	No

Figure 3.11: Two AP images from the NIHTS with similar label wording and location that were classified differently from one another by the network trained with cropped images.

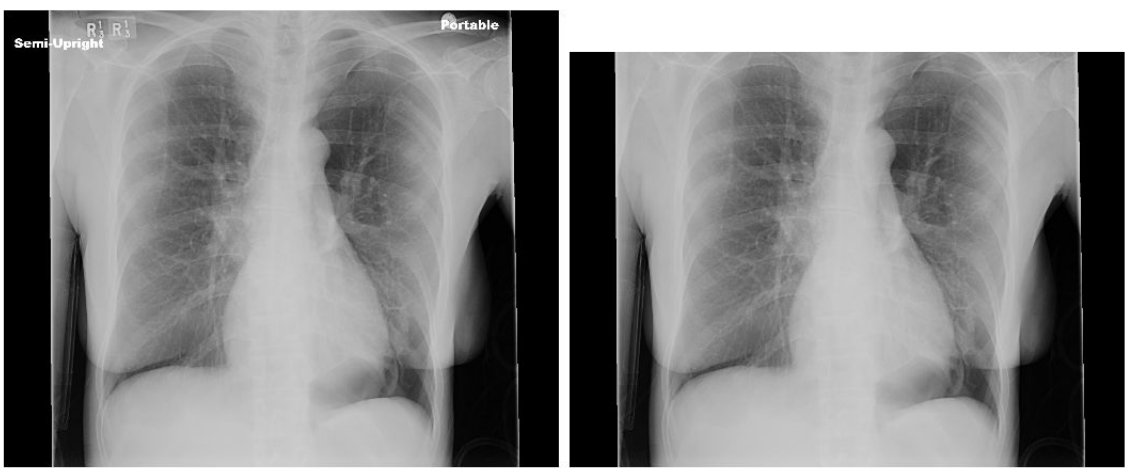


Figure 3.12: Example of a cropped image (right) that may have removed relevant anatomical information from the original uncropped image (left).

3.4 Discussion

The strong performance of the CNNs trained from scratch demonstrates that CNNs can be applied to tasks of classifying thoracic radiograph views in an efficient manner [41]. The radiographs had varying disease states, presence or lack of devices, were from 6 vendors' equipment, and had a wide range of acquisition dates (Table 3.1). The strong performance of the CNNs on the independent clinical test sets shows the robustness of the CNN for various vendors' equipment and across time. The dataset was unbalanced for the disease state, with 69.5% of the TVT and 86.4% of the "ITS" having PTX. To ensure that the presence or absence of PTX did not impact the classification performance, the prevalence of PTX in the misclassified images was compared to the overall prevalence of PTX images in each independent testing dataset ("ITS" and NIHTS). The prevalence of PTX in the misclassified images was similar to the overall prevalence within each testing dataset; therefore, the view classification performed by the CNNs did not depend on the presence or absence of PTX and was robust to varying disease state.

The classification between the four views resulting from a dual-energy study can be useful for displaying, hanging, and storing the radiographs correctly. The binary classification between frontal and lateral radiographs via a trained CNN can also be useful for radiology departments that do not routinely perform dual-energy imaging.

The classification of frontal radiographs acquired AP and PA can also be performed rapidly and accurately by a trained CNN. Classifying these views could assist radiologists in making correct diagnoses, as well as ensuring appropriate post-processing is applied, such as horizontal flipping of an image so the heart is on the correct side of the image. For the investigation of the impact of imprinted labels on the AP vs. PA classification, the presence of devices, artifacts, or imprinted labels did not account for the misclassified images. This was found by visually inspecting the images and not finding a disproportionate number (compared to the overall prevalence in the NIHTS) misclassified due to presence or absence of devices, artifacts, or imprinted labels. There is a statistically significant decrease in performance when the CNN is trained with the cropped images, which may be due to the network's learned recognition of labels when trained with the original radiographs. Some of the imprinted labels were overlaid on anatomy, and removal of the labels via cropping may have removed anatomy that was relevant for the classification. There were many possibilities for how to evaluate the impact of imprinted labels. Ideally, radiographs without imprinted labels would have been used to train the network to avoid training a CNN to detect labels; however, the images without labels were too small of a dataset for training a CNN from scratch. Another method could have involved obscuring the label, putting a black box over the label, for example. However, the network, instead of learning the labels, would likely learn to identify the black box as an indication of an AP image. The label was cropped out of the image to minimize the incorrect correlations that could be learned by the CNN. All images, cropped and uncropped, were resized to 256 x 256 pixels for training the CNN. The network could have learned an altered aspect radio as an indication of an AP image, since most of the AP images were cropped. However, the cropped images varied in their original size and the number of rows cropped out varied widely, since each image was manually cropped and the imprinted labels were at many different positions within the radiograph across the dataset. Due to this, the change in aspect ratio, a possible confounder, varied across the dataset. In addition, we tested on an unaltered independent test set; the aspect ratios of the AP images in the independent test set were unchanged. The decrease in classification performance when trained with the cropped images demonstrates that the potential impact of imprinted labels should be considered when training deep learning models. Despite the decrease in performance, the classification performance was strong and could be incorporated into clinical workflows.

Clinical implementation of these CNNs trained from scratch would allow DICOM metadata to be checked using the percent likelihood scores output by the network. The rapid classification by a trained CNN potentially facilitates integration into the clinical workflow without compromising efficiency. In addition to a potential clinical benefit, the time savings introduced by using these trained CNNs can make it possible for researchers to rapidly classify large datasets, especially when they are from various sources and may have different DICOM header fields.

The CNN architecture used (AlexNet) is a relatively shallow network compared to some other commonly used network architectures [35]. The AlexNet architecture was chosen to determine whether a relatively shallow network could perform these tasks adequately. The use of a CNN trained from scratch to perform radiograph view classification has research and clinical applications; a shallower network can be more widely implemented due to its reduced computational requirements compared to deeper CNNs. The classification performance could likely be improved using a more complex and deeper network; however, the risk of overfitting would be increased, meaning that there are too few training examples and the network learns the training examples too well so it is not able to generalize on an independent test set.

3.5 Conclusions

CNNs trained from scratch can be applied to classify thoracic radiographic views for images with various disease states, devices, and manufacturers, leading to the enhancement of radiology workflow [41]. The classification via trained CNNs is rapid, lending itself to incorporation in clinical workflow with minimal time requirements. In addition, the rapid and accurate classification could be a helpful tool for researchers compiling large datasets from various sources. The trained CNN could help verify DICOM header information to ensure correct post-processing, display, and storage.

The robustness of the CNN trained from scratch is demonstrated through the large acquisition time range and number of equipment vendors. The lowest AUC in this study was for the task of distinguishing frontal images acquired AP and PA when the CNN was trained with cropped images (0.946, 95% CI: 0.931, 0.958). This is still a strong performance that could be used clinically. The use of an AlexNet CNN architecture means that the trained network could be integrated into clinical or research workflows without the dedication of significant computation resources or purchase of proprietary algorithms.

CHAPTER 4

INVESTIGATION OF DEEP LEARNING METHODS IN THE TASK OF DETECTION OF PNEUMOTHORAX IN CHEST RADIOGRAPHS

4.1 Introduction & Motivation

The detection of pneumothorax in a frontal chest radiograph can be a difficult task due to the overlapping structures resulting from 2D projection radiography, as well as the wide range of sizes and severity PTX can have. The computerized detection of PTX had been investigated prior to deep learning's tractability; a study by Sanada et al. [24] in 1992 detected PTX through ROI placement on the upper lung, enhancing edges, and removing rib edges based on the location of posterior ribs as determined separately. Additional steps were taken to reduce noise and then the detection of the PTX pattern was performed using the Hough transform. When the method was applied to 22 radiographs with PTX and 28 normal radiographs, 77% of pneumothoraces were able to be detected. The detected PTX pattern was marked and displayed. While promising initial results, 23% of the PTX images were not detected due to the placement of the ROI. In addition, residual rib edges caused false positives; 15 of 22 false positives were due to rib edges. The radiographs required digitization for input to the algorithm, and the authors found that when the same image was digitized a few times, the results varied [24]. This study showed the feasibility of computer detection of PTX and identified some potential issues with these computerized methods.

Multiple groups have reported strong performances applying deep learning methods for the task of detection of PTX in chest radiographs [12, 9, 11, 10]. Table 4.1 summarizes the methods and results from other PTX detection studies. Some of these studies [11, 9, 10] used frontal chest radiographs from the publicly available ChestX-ray14 dataset [11] and trained neural network architectures using chest radiographs to detect many different lung diseases. A few groups downsampled the full radiographs to an input size of 224 x 224 pixels [12, 9]. At that input image size, the most specific visual signs of a PTX, a fine line at the edge of the lung with a change in texture outside the lung, are typically not visible due to the low spatial resolution. Thus, while the reduced input matrix size may be adequate for detecting other lung diseases or conditions with deep learning, it may be inadequate for PTX and further investigation of the effect of input resolution is needed. A recent study [43] investigated the impact of image resolution on the detection of eight conditions from the ChestX-ray14 dataset, finding the maximum classification performance between 256 x 256 pixels and 448 x 448 pixels for emphysema, cardiomegaly, hernias, edema, effusions, atelectasis, masses, and nodules. However, the study did not investigate the impact of image resolution for the detection of PTX.

Table 4.1: Methods and results for currently published studies using deep learning for the detection of PTX on chest radiographs.

Study	Dataset Source	Neural Network Architecture	Input Image Size	Dataset Augmentation Techniques	AUC from ROC Analysis in the task of detecting pneumothorax
Yao et al., 2018 [1]	ChestX-ray14	Two stage custom model with densely connected image encoder with a recurrent neural network decoder	512 x 512	· 4 directional shift by 25 pixels · Rotation between 15 · Scaled between 80% and 120%	0.841
Rajpurkar et al., 2017 [2]	ChestX-ray14	Custom designed network, CheXNet	224 x 224	· 224 x 224 random crops of 256 x 256 images · Horizontal flipping	0.889
Wang et al., 2017 [3]	ChestX-ray14	ResNet-50 from scratch	1024 x 1024	None	0.799
Cicero et al., 2017 [4]	Institutional data set with 1,299 PTX radiographs	GoogleNet from scratch	224 x 224	· 224 x 224 random crops of 256 x 256 images · Horizontal flipping	0.861

In order to investigate the impact of the effective resolution of the input images on the deep learning detection of PTX, a dataset consisting of frontal chest radiographs with and without PTX was assembled. To construct a dataset with images of a higher effective resolution, the radiographs were cropped into two "apex images" (Section 4.2.2) for input to the neural network. "Effective resolution" refers to the resolution when the image is downsampled to the required input size for the CNN (224 x 224 pixels). The frontal chest radiographs were downsampled to the required input size from the original acquisition resolution. The "apex images," cropped from the original-resolution radiograph, were downsampled to the same required input size; therefore, the apex images had a higher effective resolution since it was a limited field of view but downsampled to the same size as the full radiographs. The purpose of generating these levels of resolution for neural network input was to investigate whether the CNNs are learning to identify the visual signs of PTX for the classification or are learning spurious correlations.

Different techniques for CNN training were performed to evaluate the performance on the two resolution levels. This is due to the varying strengths and weaknesses of deep learning approaches and networks, as discussed in Chapter 2. The application of various techniques and architectures allows for comparison across many different deep learning methods and identification of the method with the strongest performance in the task of classifying between radiographs with and without PTX. The two levels of input image resolution and the deep learning methods investigated for each level that will be presented in this chapter are shown in Figure 4.1.

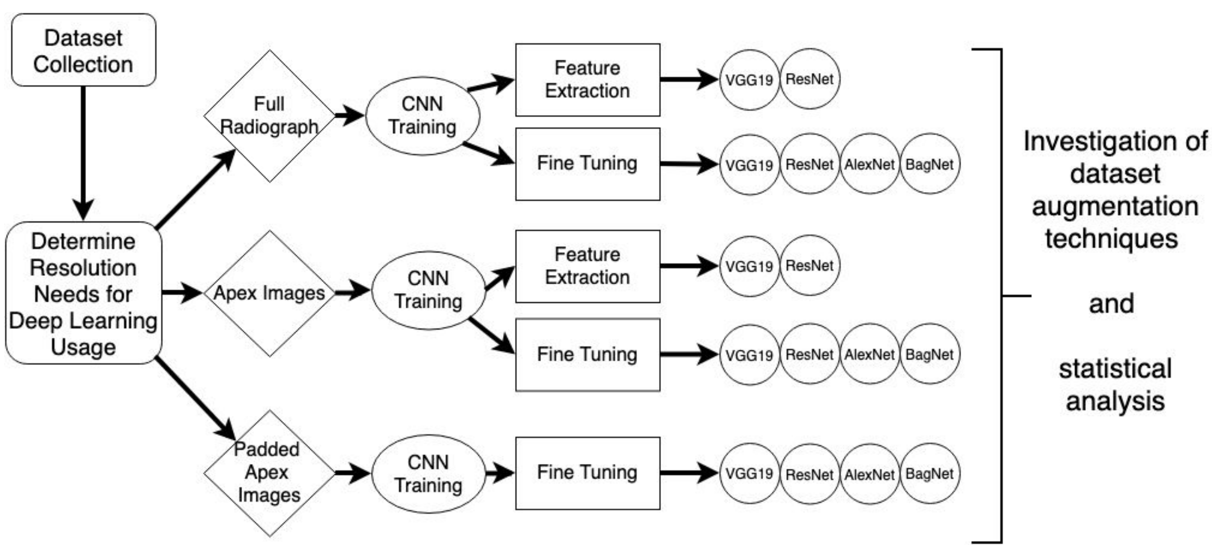


Figure 4.1: Deep learning methods presented and discussed in this chapter.

4.2 Data Acquisition and Preprocessing Methods

4.2.1 Image Data and Verification of Ground Truth

The chest radiographs were obtained from two sources: the University of Chicago Medical Center and the publicly-available ChestX-ray14 dataset from NIH Clinical Center [11]. Supervised training of deep learning models requires the truth for the input data; since the truth is used to adjust the weights to obtain the optimum performance, high-quality, accurate truth is needed.

University of Chicago Report Searching

A total of 1,295 radiographs with PTX in the apex and 668 without PTX were downloaded retrospectively from the PACS system at the University of Chicago Medical Center under a HIPAA-compliant, IRB-approved protocol. The identification and use of chest radiographs with PTX visible in the apex(es) was justified by an initial analysis that showed that 94% of our preliminary PTX radiograph dataset (225 of 240) had PTX in the apex. In addition, the biological justification is that air rises to the top of the pleural cavity when the patient is upright or semiupright; therefore, PTX is commonly visible in the apex [14]. The PTX radiographs had been identified by searching radiology reports for PTX cases from January 2013 to March 2017. A radiologist with 18 years of clinical experience verified the ground truth for each image since clinical radiology reports are known to be flawed for deep learning applications and using them as "truth" should be avoided, if possible, since they are generated for patient care, not research [44].

The PTX radiographs from University of Chicago used for training and validation were found by searching radiology reports for PTX cases between January 2013 and April 2015, as well as from a teaching set. The radiographs without PTX were from the same patients prior to or after the PTX diagnosis as well as from patients who did not have a radiograph in the PTX portion of the dataset. For use as the test set, 225 radiographs with PTX in the apex and 350 radiographs without PTX were collected. The PTX radiographs used for testing were found by searching radiology reports between May 2015 and February 2017, and the radiographs without PTX used for testing were from the patients prior to or after the PTX diagnosis. The patients included in the testing set did not have a radiograph used for training or validation. The PTX and non-PTX radiographs' acquisition dates ranged from February 2006 to March 2017.

ChestX-ray14 Database

The second source of chest radiographs was the ChestX-ray14 dataset from NIH Clinical Center [11]. All of the images labeled as "pneumothorax" in that dataset were downloaded. A radiologist reinterpreted all of the images labeled as "pneumothorax" at two separate instances with the previous interpretation not visible. Since the SIIM/ACR PTX challenge [45] released new truth for a portion of this dataset, cases from the ChestX-ray14 database were only used if both the radiologist truth and the challenge truth agreed. The number of ChestX-ray14 images for which both our radiologist and the challenge truth agreed yielded 1,616 radiographs with PTX in the apex and 1,752 radiographs without PTX. Note that the challenge truth did not encompass the full dataset of "pneumothorax" labeled images from ChestX-ray14 so one cannot conclude that only 1,616 radiographs of the 5,298 radiographs labeled as "pneumothorax" in the ChestX-ray14 dataset are correctly labeled; since images were only used if PTX was visible in the apex while chest radiographs with PTX visible elsewhere were not included in the dataset. The images from ChestX-ray14 were randomly separated into training (70%) and validation (30%), ensuring patients used for training did not have images in the validation set.

Table 4.2 shows the number of radiographs with and without PTX and their sources, as well as patient characteristics and distribution into training, validation, and testing. Images without PTX had either a disease/condition other than PTX or no significant abnormality.

Table 4.2: The chest radiograph dataset was obtained from two different image sources and split into training, validation, and testing sets. UC= University of Chicago Medical Center, NIH= ChestX-ray14 dataset from NIH Clinical Center, DL= Deep learning

		PTX		No PTX	
Image Source		**UC**	**NIH**	**UC**	**NIH**
Number of Chest Radiographs		1,295	1,616	668	1,752
Number of Patients		436	495	598	1,027
Acquisition Date Range		January 2013 to March 2017	No dates provided	February 2006 to March 2017	No dates provided
Manufacturer Identified in DICOM Header	**Canon Inc.**	516	No manufacturer identified	45	No manufacturer identified
	FUJIFILM Corporation	499		471	
	GE Healthcare	257		113	
	Other	23		39	
Age	**Mean**	55.0	47.0	56.7	47.3
	Standard Deviation	18.4	17.7	18.0	15.7
Sex	**Female**	181 (42%)	235 (47%)	276 (46%)	565 (55%)
	Male	255 (58%)	260 (53%)	322 (54%)	462 (45%)
Number of Apex Images		1,388	1,684	2,526	5,052
Full Radiographs Used in Each DL Stage	**Training**	749	1,131	223	1,226
	Validation	321	485	95	526
	Testing	225	0	350	0
Apex Images Used in Each DL Stage	**Training**	808	1,175	1,131	3,538
	Validation	352	509	478	1,514
	Testing	228	0	917	0

4.2.2 Image Preparation for Network Input

Resizing full radiographs

Due to computational limitations, CNNs are typically limited to an input size around 224 x 224 pixels. The full radiographs were downsampled for input to the CNN. The original radiograph sizes varied (Table 4.3); however, they were all downsampled to the same size. The radiographs from ChestX-ray14 were released at a matrix size of 1024 x 1024, although the original acquisition sizes varied (Table 4.3). The training radiographs were downsampled to 256 x 256 pixels via bilinear interpolation, and then 224 x 224-pixel random cropping was applied to augment the training data. The radiographs used for validation and testing were downsampled via bilinear interpolation to 224 x 224 pixels.

Table 4.3: The original radiograph sizes and pixel sizes for the dataset. The “other” category includes the images that had sizes in common with less than 10 images in the dataset. UC= University of Chicago Medical Center, NIH= ChestX-ray14 dataset from NIH Clinical Center

		UC	
		PTX	**No PTX**
Original radiograph size	1760 x 1760	80	178
	1760 x 2140	39	37
	2012 x 2012	17	3
	2021 x 2021	96	40
	2140 x 1760	234	164
	2140 x 2140	58	113
	2800 x 3408	66	7
	3408 x 2800	134	14
	4280 x 3520	88	16
	Other	483	96
Original radiograph pixel size (mm x mm)	0.1 x 0.1	88	16
	0.125 x 0.125	527	46
	0.194222 x 0.194222	77	21
	0.194311 x 0.194311	88	65
	0.194317 x 0.194317	21	8
	0.1988 x 0.1988	45	12
	0.2 x 0.2	411	490
	Other	38	10

		NIH	
		PTX	**No PTX**
Original radiograph size	2020 x 2021	12	8
	2021 x 2020	16	7
	2021 x 2021	166	103
	2048 x 2500	74	89
	2500 x 2048	304	434
	2544 x 3056	27	42
	2992 x 2991	402	443
	2056 x 2544	253	263
	Other	362	363
Original radiograph pixel size (mm x mm)	0.139 x 0.139	305	339
	0.143 x 0.143	707	746
	0.168 x 0.168	306	446
	0.171 x 0.171	72	77
	0.194311 x 0.194311	210	141
	Other	16	3

Portioning into Apex Images

To investigate the effect of image resolution on the deep network performance, each full radiograph was separated into two images of the top portion of the lungs (referred to as "apex images"). The separation into apex images was performed by the placement of points at the apices of the lungs and the costophrenic angles, in addition to the placement of a point at the top of the aortic arch (Figure 4.2). The separation between the right and left lungs was found by connecting the apex points with a line and the costophrenic angle points with a line and then using the line that bisected those two lines (blue line in Figure 4.2). The portion of the lung above the point placed at the top of the aortic arch and below the point placed at the apex was used for the apex image. The costophrenic angle points were used to denote the side boundary of the lung, so the apex image would not include a large portion outside the body.

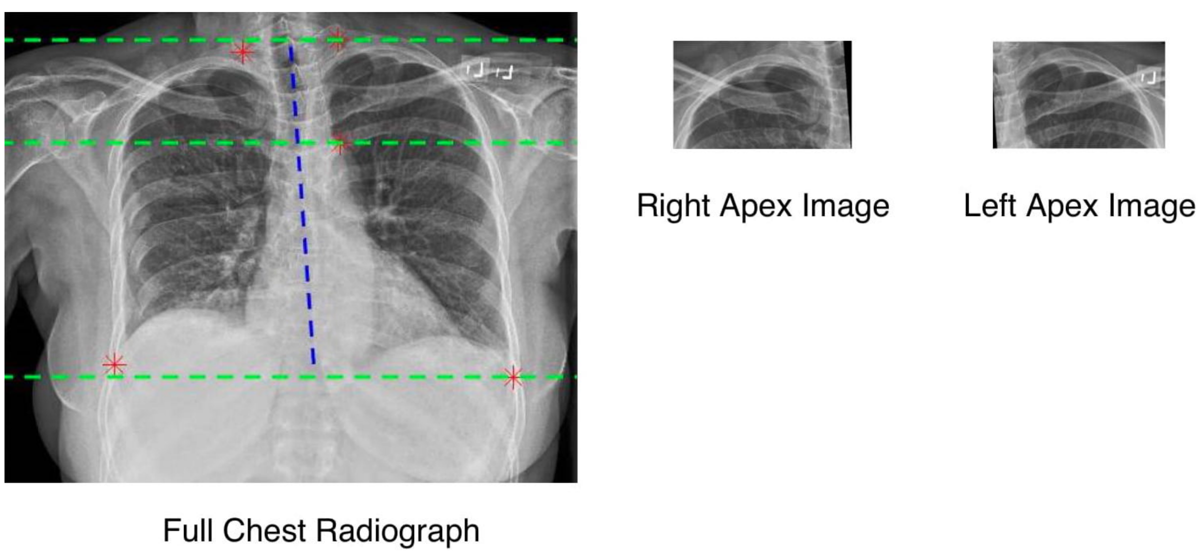


Figure 4.2: Points placed at the apices, aortic arch, and costophrenic angles (red) are used to divide the full radiograph (left) into two apex images (middle, right). The portion of the lungs above the point at the top of the aortic arch is used for the apex images. The midline (blue), used as the line of division between the right and left sides, is calculated as the bisection of the line between the points at the apices and the line between the costophrenic angle points.

If only one apex of a radiograph had PTX, the other apex was used as a non-PTX apex image. For radiographs without PTX, both apexes were used as non-PTX apex images. Table 4.2 gives the number of apex images from each source. Not all extracted apex images were included in the apex image set; apex images were excluded from the training and validation sets if the top of the apex was cut off in the radiograph or the chest was off center such that the edge of the apex was cut off in the radiograph. All apex images from the independent test set radiographs were included.

Padded Apex Images

Apex images were padded with zeros, meaning that new pixels with a value of zero were added around the edges of the apex image to restore it to the original radiograph size. These padded apex images were generated in order to investigate the influence on performance when the CNN's attention is directed to the relevant portion of the image. Comparing the performance of the padded apex images versus the apex images and full radiographs enables the determination of the influence of image resolution.

Comparison of Effective Resolutions for Downsampled Images

The full chest radiographs, apex images, and padded apex images used for training were all downsampled to 256 x 256 for random 224 x 224 pixel cropping for CNN input. Figure 4.3 shows the appearance of these images when resized. The full radiographs and padded apex images had the same effective resolution when downsampled since they were both downsampled to the same matrix size from the original acquisition resolution or, in the case of the ChestX-ray14 images, from their 1024 x 1024 resolution as released by NIH. The apex images had a higher effective resolution since they were cropped from the radiographs at the original acquisition resolution (or from the 1024 x 1024 ChestX-ray14 images) and then that limited field of view was downsampled to the same matrix size as the full radiographs and padded apex images. The effective resolution of the downsampled apex images cropped from the University of Chicago images was, on average, 3.3 times the effective resolution of the full radiographs downsampled to the same matrix size (standard deviation=0.67). For the ChestX-ray14 images, the effective resolution of the downsampled apex images was, on average, 3.2 times the effective resolution of the full radiographs downsampled to the same matrix size (standard deviation=0.43).

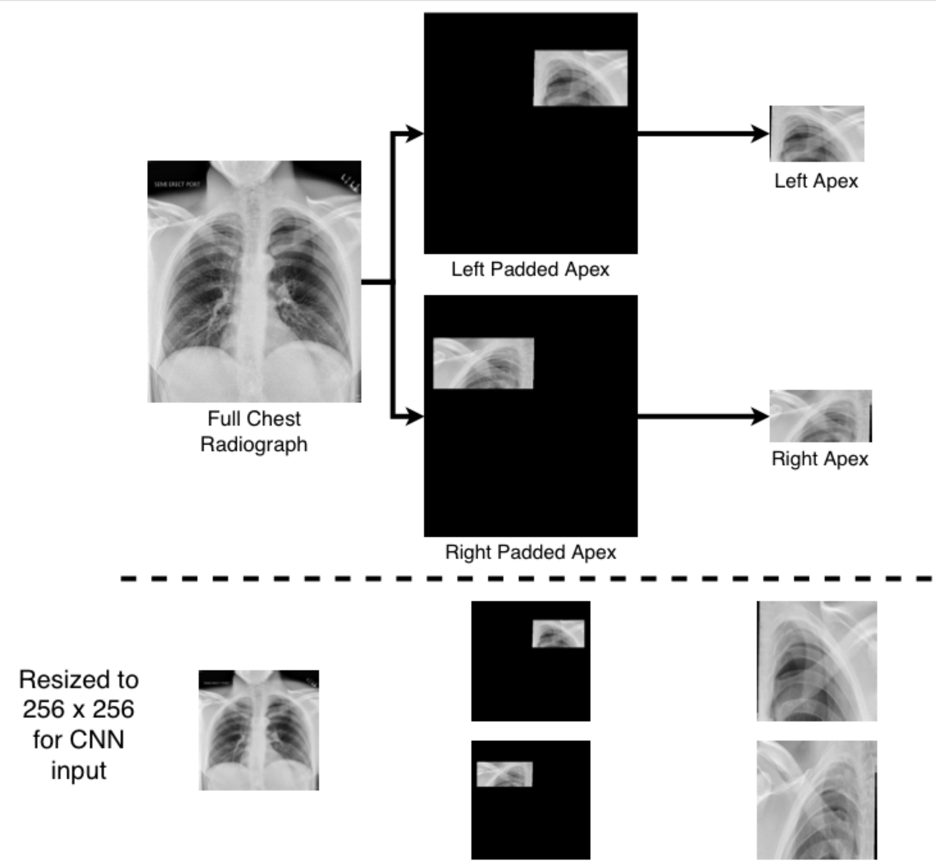


Figure 4.3: The padded apex images are generated by obscuring the radiograph outside of the apex specified using the 5 points (Figure 1). The full chest radiograph, the padded apex images, and the apex images all were resized to 256 x 256 via bilinear interpolation for fine-tuning the CNN.

4.3 Fine-Tuning Methods and Statistical Analysis

As noted in Chapter 2, fine-tuning is a form of transfer learning in which a CNN trained on other data (in this case, ImageNet [31]) has most of its layer weights frozen and selected layers retrained for a new task [5]. Due to the relatively small dataset, the technique of fine-tuning was chosen so the CNN did not need to be completely trained "from scratch", but rather only a few layers (and therefore fewer weights) had to be retrained.

The chest radiographs were used for fine-tuning four different CNN architectures pre-trained on ImageNet: AlexNet [29], VGG19 [35], ResNet50 [36], and BagNet [37], which were introduced in Chapter 2. Fine-tuning was performed using the training dataset and validation dataset. Then, the test set was used to evaluate the performance of the fine-tuned network. The test set images, 225 PTX radiographs and 350 radiographs without PTX from the University of Chicago Medical Center, were also separated into apex images/padded apex images and tested separately. For statistical analysis and computation of the ROC curve, the apex/padded apex with the highest network-assigned probability of PTX was used as the single value for each test set image. The same cases were used in each stage of fine-tuning with either full radiographs or apex images. Table 4.2 provides the number of radiographs and apex images with and without PTX used for each stage of the CNN fine-tuning.

4.3.1 Augmentation Techniques

Dataset augmentation was applied to increase the number of training samples on which the network could learn. There are many potential dataset augmentation techniques; current literature was reviewed to identify effective augmentation techniques for the task. A few of the recently-published studies apply 224 x 224 random cropping of the 256 x 256 pixel image for input to the neural network [9, 12] and apply horizontal flipping [9, 12, 10]. Therefore, for training the network, the full radiographs were resized to 256 x 256, and then 224 x 224 random cropping and horizontal flipping was applied. To further increase the number of training samples, a rotation range of ± 5 degrees was applied, in addition to a height shift and width shift of 12 pixels, as well as a zoom range of 0.2 (magnification between 0.8 and 1.2) to be similar to the methods in Yao, et al. [10]. The validation and test images were unaltered, other than being resized to 224 x 224 for testing with the fine-tuned network (which had a required input size of 224 x 224). Example training images with dataset augmentation techniques applied are shown in Figure 4.4, along with validation and test images that are not altered other than being resized to the required CNN input size.

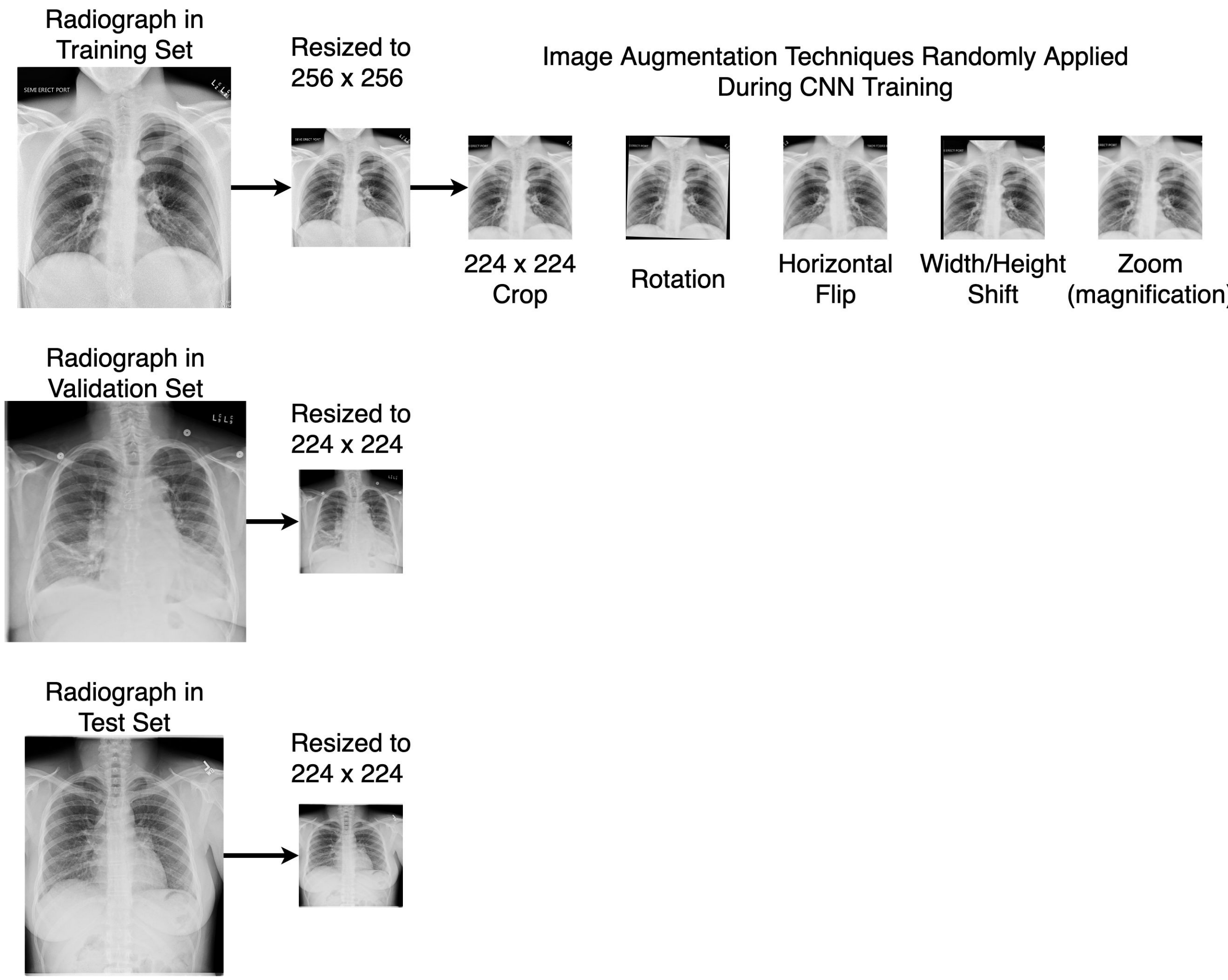


Figure 4.4: Resizing and augmentation techniques for the full radiographs

The apex images were also used in fine-tuning a neural network in a similar fashion as the full images. After being cropped from the radiograph at full resolution, the apex images were resized to 256 x 256 pixels, and then, for dataset augmentation, 224 x 224 pixel random crops were used and horizontal flipping was applied. The same dataset augmentation techniques applied to the full radiographs were also used for the apex images. As with the full images, validation and test set apex images were unaltered, and none of the patients in the test set overlapped with those in the training or validation sets. Example training images with dataset augmentation techniques applied are shown in Figure 4.5, along with validation and test images that are not altered other than being resized to the required CNN input size.

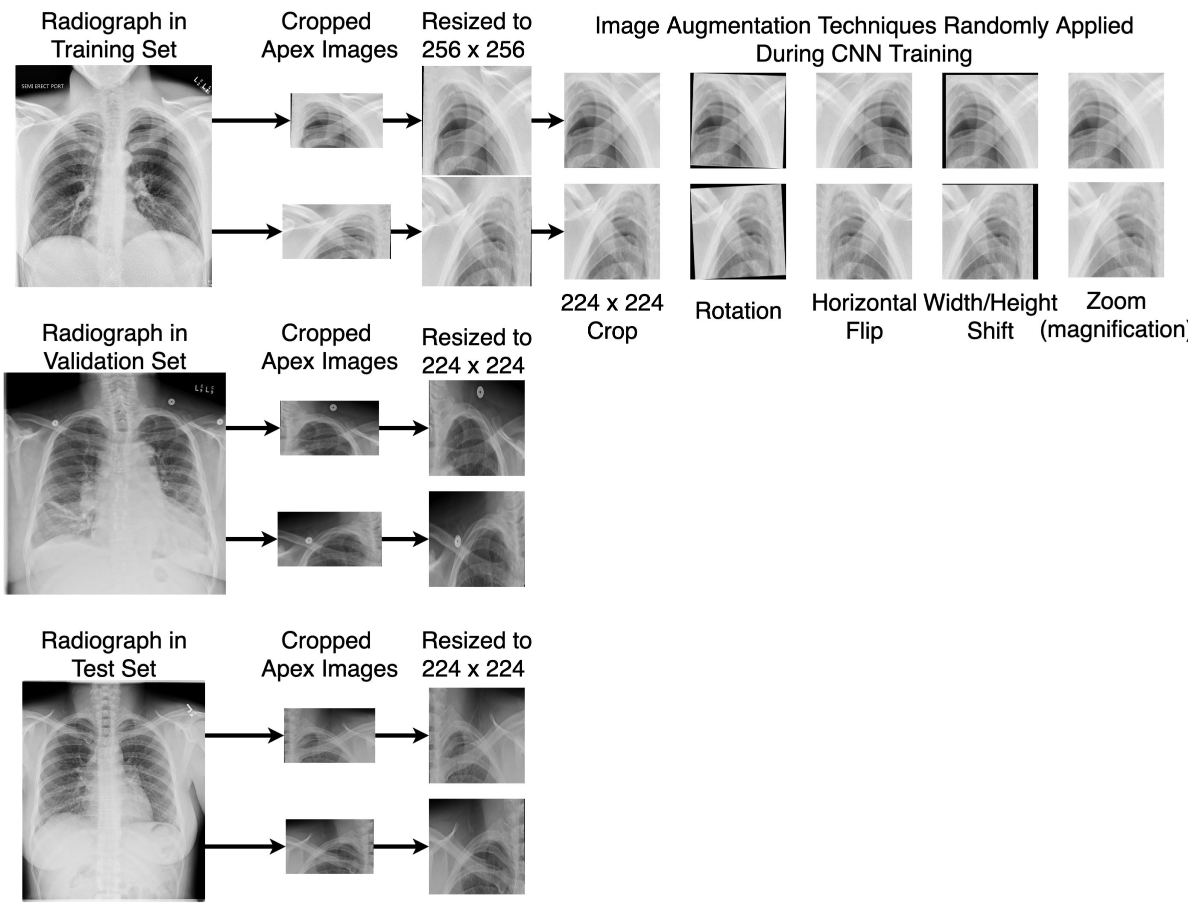


Figure 4.5: Resizing and augmentation techniques for the apex images

The padded apex images were also used for fine-tuning the four CNNs pretrained on ImageNet. The same image augmentation techniques, horizontal flipping and 224 x 224 random cropping, applied for training with the full radiographs and apex images were also applied to the padded apex images. Example training images with dataset augmentation techniques applied are shown in Figure 4.6, along with validation and test images that are not altered other than being resized to the required CNN input size.

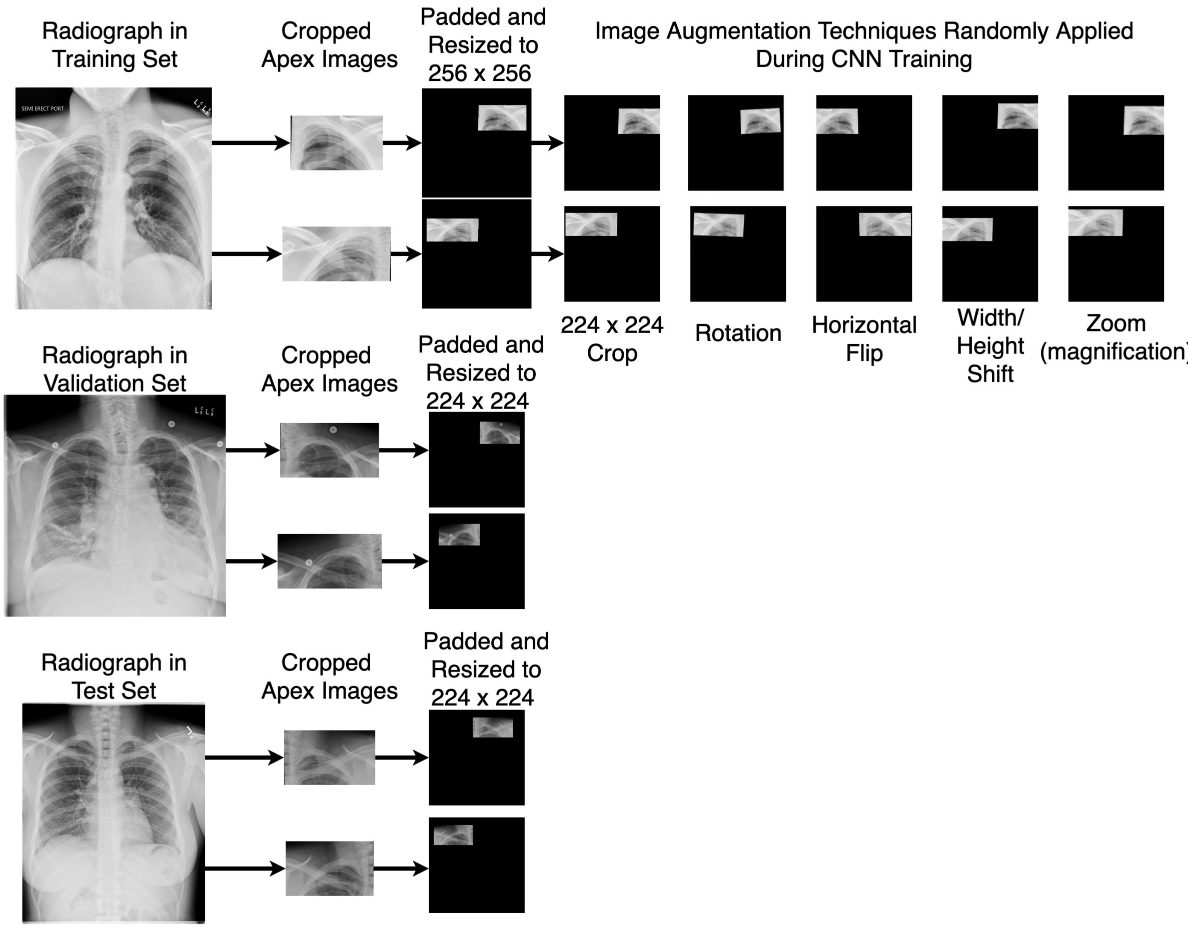


Figure 4.6: Resizing and augmentation techniques for the padded apex images

Each CNN was fine-tuned using the training and validation sets, and after training was complete, the fine-tuned networks were tested on the test set. The same test set was used for the testing of the full radiograph method, the padded apex image method, and the apex image method.

4.3.2 Statistical Evaluation

The performance of the fine-tuned networks for the full radiographs, the padded apex images, and the apex images was evaluated using receiver operating characteristic (ROC) analysis and the area under the ROC curve (AUC) served as the performance metric in the task of distinguishing between PTX and non-PTX cases [38]. Due to statistical variations in the initialization of CNNs, each time a CNN is trained/fine-tuned, the performance will be slightly different. Therefore, to properly represent the results of the fine-tuning, the CNNs were each fine-tuned five times and the CNN with the median AUC was reported as the result and used for the visualization of CNN output (Chapter 5). Fine-tuning each CNN five times was chosen to gauge the variability of network performance while minimizing training time.

The ROC curves of the CNNs fine-tuned with full radiographs, padded apex images, and apex images were compared using DeLong's method [42] to determine whether there was a statistically significant difference in performance in the task of distinguishing between images with and without PTX. To compare ROC curves, the highest probability of PTX between the two apexes/padded apexes from each test image was used as the single value for each test image in the ROC analysis since the detection task was on a per-case basis and not a per-lung basis. Three comparisons were made: the full radiograph ROC curve compared to the padded apex ROC curve, the full radiograph ROC curve compared to the apex ROC curve, and the padded apex ROC curve compared to the apex ROC curve. The level of statistical significance was $p<0.017$ after correction for multiple comparisons using the Bonferroni method [46].

Detection sensitivity and specificity were also calculated. If the probability of the presence of PTX from the fine-tuned network for an independent test image was 0.5 or greater, it was classified as a PTX image. If the probability of PTX was less than 0.5, the image was classified as a non-PTX image. The choice of the threshold is arbitrary; since it is a probability, 0.5 was used as the threshold for this case. The threshold was chosen to be the same for all the CNNs investigated in this work in order to directly compare the results.

4.4 Network-Specific Methods

4.4.1 AlexNet-Specific Methods

AlexNet had a high performance on the radiograph classification task (Chapter 3). Due to the high performance of AlexNet on the radiograph classification task and its prominence as the CNN that began the latest wave of interest in CNNs, an AlexNet architecture was fine-tuned for the task of classifying between images with and without PTX. The AlexNet architecture was pretrained on ImageNet, with the final convolutional layer and classification layer being retrained for the new task. Since the AlexNet CNN pretrained on ImageNet has a 1000-class output to correspond to the ImageNet classes, the output and classification layer had to be modified to correspond to 2 classes, PTX and no PTX. The CNN was fine-tuned for 20 epochs, with a stochastic gradient descent (SGD) solver. The initial learning rate was 0.001 and it was reduced by a factor of 0.1 every 7 epochs. The software used for fine-tuning the network was PyTorch (version 1.1.0) implemented in Python (version 3.6.8). The fine-tuning was performed on a nVidia Tesla V100 GPU.

4.4.2 VGG19-Specific Methods

VGG19 was selected due to its high performance in many medical imaging tasks [47]. For VGG19, the last two convolutional layers and final classification layer were re-trained for the new task of distinguishing between radiographs with and without PTX using the assembled dataset of frontal chest radiographs.

A VGG19 architecture network pretrained on ImageNet was used, and the last two convolutional layers and the final classification layer were retrained for the task of distinguishing between apex images with and without PTX. The final classification layer was retrained for two epochs before the fine-tuning took place in order to improve the initialization. The optimizer used was SGD with an initial learning rate of 0.001, which was reduced by a factor of 0.1 three epochs after the validation loss plateaued.

To address overfitting, in addition to dataset augmentation and the use of fine-tuning, the validation loss was monitored and training ceased five epochs after the validation loss started increasing. A relatively high dropout (0.6) was also used to reduce overfitting. The software used for fine-tuning the network was Keras (version 2.2.2) implemented in Python (version 3.5.5) with a Tensorflow (Tensorflow-gpu, version 1.10.0) backend. The fine-tuning was performed on a nVidia Tesla V100 GPU.

4.4.3 ResNet50-Specific Methods

To fine-tune a ResNet50 architecture for classifying between images with and without PTX, a ResNet50 CNN pretrained on ImageNet was loaded. All the layers were frozen other than the last two convolutional layers and the final classification layer. Like VGG19, the newly added classification layer was retrained for 2 epochs and then the final two convolutional layers and the final classification layer were fine-tuned. An Adam optimizer was used, with an initial learning rate of 0.001, which was reduced by a factor of 0.1 three epochs after the validation loss plateaued. The maximum number of epochs was set to 50; however, training ceased five epochs after the validation loss started decreasing. The dropout was set to be 0.5 to help reduce overfitting, in addition to a modified learning rate and early stopping of the fine-tuning. The software used for fine-tuning the network was Keras (version 2.2.2) implemented in Python (version 3.5.5) with a Tensorflow (Tensorflow-gpu, version 1.10.0) backend. The fine-tuning was performed on a nVidia Tesla V100 GPU.

4.4.4 BagNet33-Specific Methods

The BagNet33 architecture network was pretrained on ImageNet. As the authors of the BagNet paper [37] suggested for fine-tuning for a new task, the final classification layer was changed to correspond to a 2-class classification task. Then the pre-trained network was initialized with the ImageNet weights and each layer was fine-tuned for the new task of classifying between images with and without PTX. This training was performed for 20 epochs and the optimizer used was SGD with an initial learning rate of 0.001. The learning rate was reduced by a factor of 0.1 every 7 epochs. The software used for fine-tuning the network was PyTorch (version 1.1.0) implemented in Python (version 3.6.8). The fine-tuning was performed on a nVidia Tesla V100 GPU.

4.5 AlexNet Results for the Full Radiographs, Apex Images, and Padded Apex Images

Table 4.4 provides the results from the performance evaluation of the fine-tuned AlexNet CNNs using the test set of 225 radiographs with PTX and 350 radiographs without PTX. When the ROC curves were compared using DeLong's method [42] a statistically significant difference in performance ($p<0.017$, corrected for the three comparisons using the Bonferroni correction) was seen between the CNN fine-tuned with the full radiographs (AUC=0.702) and the CNN fine-tuned with the apex images (AUC=0.862) with a p-value of <0.001. Figure 4.7 shows the probability distributions from the AlexNet CNNs fine-tuned with full radiographs and apex images for the test set images with and without PTX. Figure 4.8 consists of two scatterplots showing the probability of PTX for each test case as output from analysis of its full radiograph and its corresponding apex images. Statistically significant differences in performance ($p<0.017$) were seen between the CNN fine-tuned with the padded apex images (AUC=0.823) and the CNN fine-tuned with the apex images (AUC=0.862) with a p-value of 0.016, as well as between the CNN fine-tuned with the full radiographs (AUC=0.702) and the CNN fine-tuned with the padded apex images (AUC=0.823), with a p-value of <0.001. Figure 4.9 gives the probability distributions for the test set images from the CNNs fine-tuned with full radiographs, apex images, and padded apex images. The ROC curves for the fine-tuned AlexNet CNNs are shown in Figure 4.10.

Table 4.4: Summary of the results from testing the AlexNet CNNs fine-tuned with the full radiographs, apex images, and padded apex images.

AlexNet CNN fine-tuned with:	AUC for the task of classifying between images with and without PTX (95% CI)	False Negatives	False Positives	Average Probability of PTX output by the fine-tuned CNN for cases with PTX (standard deviation)	Average Probability of PTX output by the fine-tuned CNN for cases without PTX (standard deviation)	Detection Sensitivity	Specificity
Full Radiographs	0.702 (0.657, 0.745)	75 (33%)	126 (36%)	0.67 (0.22)	0.53 (0.17)	67%	64%
Apex Images	0.862 (0.831, 0.890)	99 (44%)	39 (11%)	0.58 (0.31)	0.18 (0.23)	56%	89%
Padded Apex Images	0.823 (0.787, 0.854)	123 (55%)	37 (11%)	0.45 (0.22)	0.18 (0.21)	45%	89%

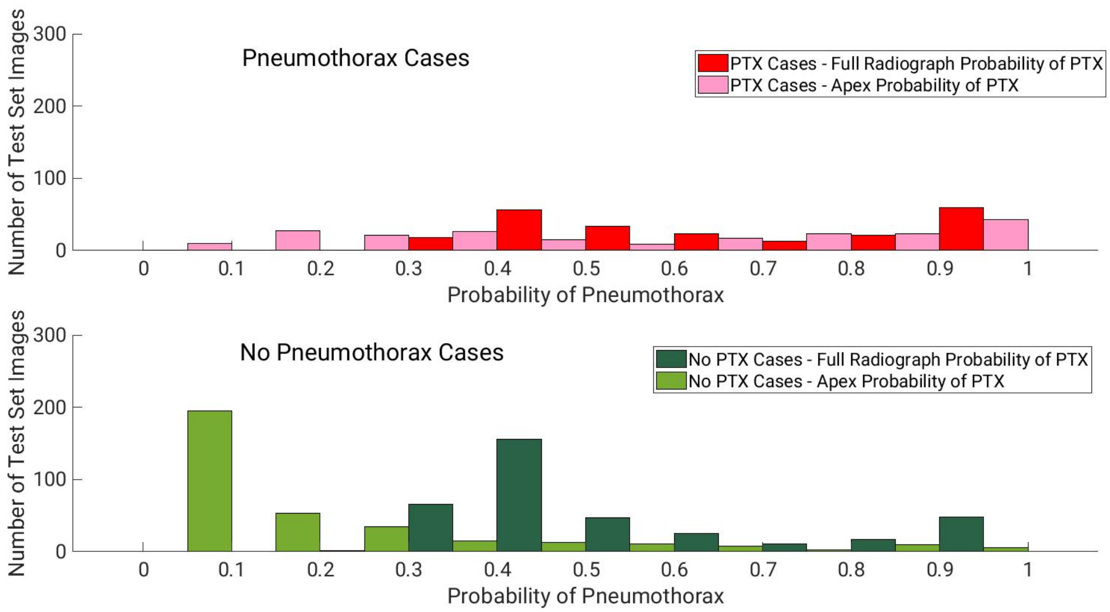
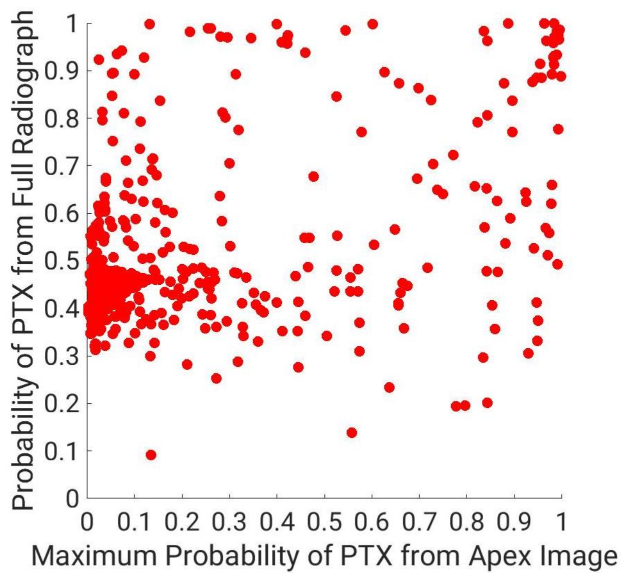
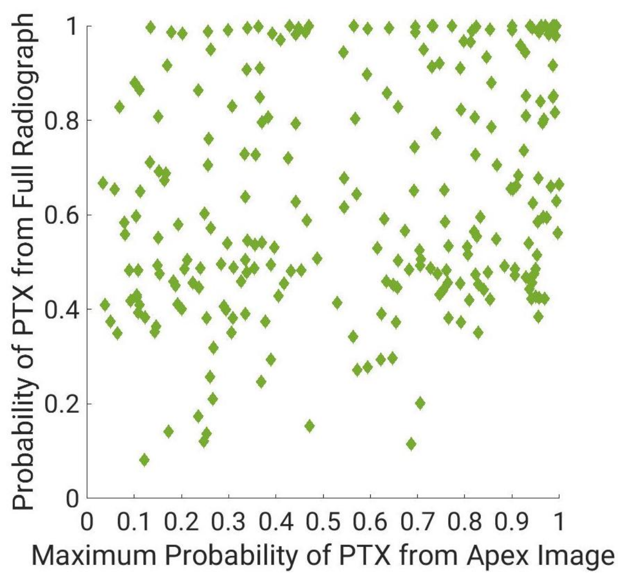


Figure 4.7: Distribution of probability of PTX from the fine-tuned AlexNet CNNs for the independent test set images plotted separately for the PTX and non-PTX cases. For the PTX cases, the probability of PTX should equal 1. For the no PTX cases, the probability of PTX should equal 0.



(a) Scatter plot of the probability of PTX from the **fine-tuned AlexNet CNNs** for the independent test set **radiographs with PTX**. Ideally, points should be in the top right corner, corresponding to a 100% probability of PTX.



(b) Scatter plot of the probability of PTX from the **fine-tuned AlexNet CNNs** for the independent test set **radiographs without PTX**. Ideally, points should be in the bottom left corner, corresponding to a 0% probability of PTX for the test set images without PTX.

Figure 4.8: Scatter plots of probability of PTX output by the fine-tuned AlexNet CNNs

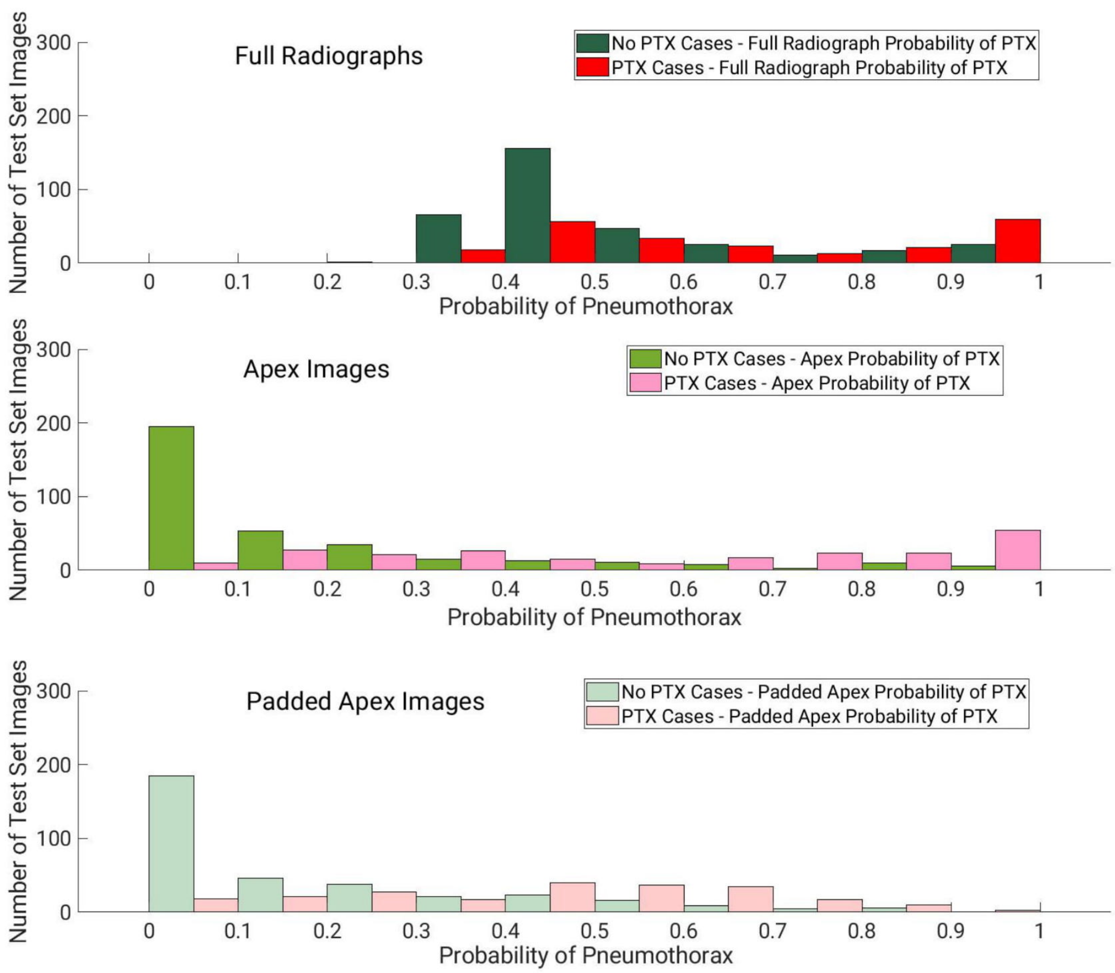


Figure 4.9: Distribution of probability of PTX output by the three fine-tuned AlexNet networks for the independent test set radiographs. An ideal distribution would have the probability of PTX equal to 1 for the PTX cases and a probability of PTX equal to 0 for the cases without PTX.

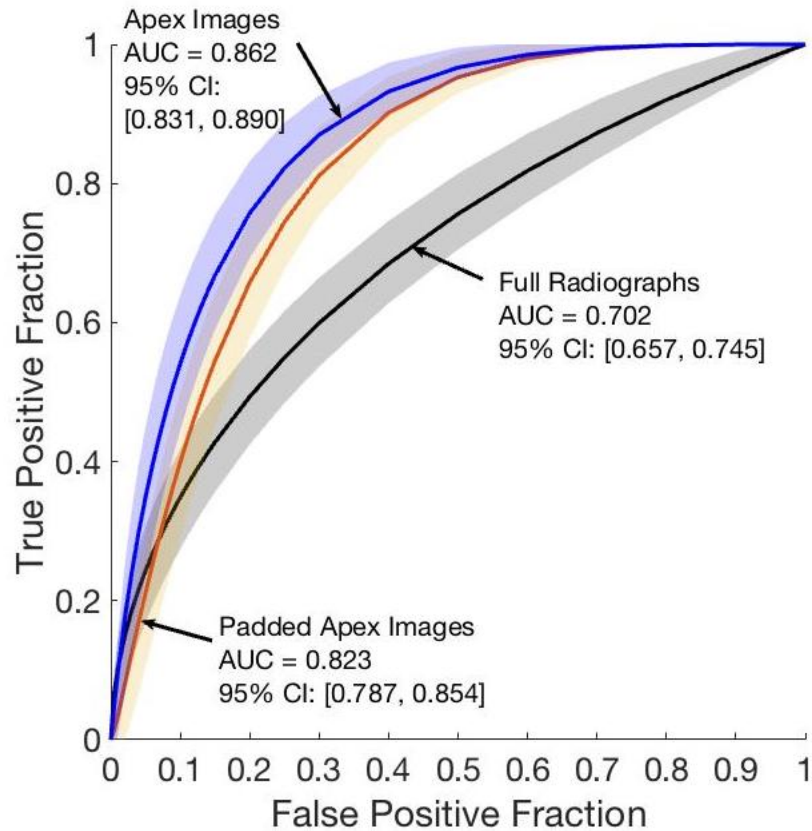


Figure 4.10: ROC curves for the three fine-tuning approaches for AlexNet showing their performances on the independent test set in the task of distinguishing between radiographs with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

4.5.1 Discussion and Conclusions from AlexNet Fine-tuning

The performances of the three networks, when quantified through ROC analysis, all show a statistically significant difference ($p<0.017$) from one another. The performance of the AlexNet CNN fine-tuned with the apex images has the highest performance with an AUC of 0.862 (95% CI: 0.831, 0.890). However, the CNN fine-tuned with the full radiographs has the highest detection sensitivity of 67%, compared to 56% for the CNN fine-tuned with the apex images. For the specificity, the CNN fine-tuned with the full radiographs has the lowest specificity (64%) and the CNNs fine-tuned with the apex images and padded apex images have the same specificity (89%). Due to a high classification AUC in the task of classifying between images with and without PTX, as well as a high specificity, the CNN fine-tuned with the apex images has the best performance of the fine-tuned AlexNet CNNs.

4.6 VGG19 Results for the Full Radiographs, Apex Images, and Padded Apex Images

Table 4.5 provides the results from the performance evaluation of the fine-tuned VGG19 CNNs using the test set of 225 radiographs with PTX and 350 radiographs without PTX. When the ROC curves were compared using DeLong's method [42] a statistically significant difference in performance ($p<0.017$) was seen between the CNN fine-tuned with the full radiographs (AUC=0.792) and the CNN fine-tuned with the apex images (AUC=0.898) with a p-value of <0.001. Figure 4.11 shows the probability distributions from the VGG19 CNNs fine-tuned with full radiographs and apex images for the test set images with and without PTX. Figure 4.12 consists of two scatterplots showing the probability of PTX for each test case as output from analysis of its full radiograph and its corresponding apex images. A statistically significant difference in performance was also seen between the CNN fine-tuned with the padded apex images (AUC=0.824) and the CNN fine-tuned with the apex images (AUC=0.898) with a p-value of <0.001. We failed to show a statistically significant difference in performance between the CNN fine-tuned with the full radiographs (AUC=0.792) and the CNN fine-tuned with the padded apex images (AUC=0.824), with a p-value of 0.08. Figure 4.13 gives the probability distributions for the test set images for the CNNs fine-tuned with full radiographs, apex images, and padded apex images. The ROC curves for the fine-tuned VGG19 CNNs are shown in Figure 4.14.

Table 4.5: Summary of the results from testing the VGG19 CNNs fine-tuned with the full radiographs, apex images, and padded apex images.

VGG19 CNN fine-tuned with:	AUC for the task of classifying between images with and without PTX (95% CI)	False Negatives	False Positives	Average Probability of PTX output by the fine-tuned CNN for cases with PTX (standard deviation)	Average Probability of PTX output by the fine-tuned CNN for cases without PTX (standard deviation)	Detection Sensitivity	Specificity
Full Radiographs	0.792 (0.753, 0.827)	50 (22%)	108 (31%)	0.76 (0.27)	0.32 (0.28)	78%	69%
Apex Images	0.898 (0.871, 0.921)	91 (40%)	27 (8%)	0.55 (0.33)	0.12 (0.20)	62%	89%
Padded Apex Images	0.824 (0.789, 0.855)	109 (48%)	42 (12%)	0.48 (0.18)	0.23 (0.19)	52%	88%

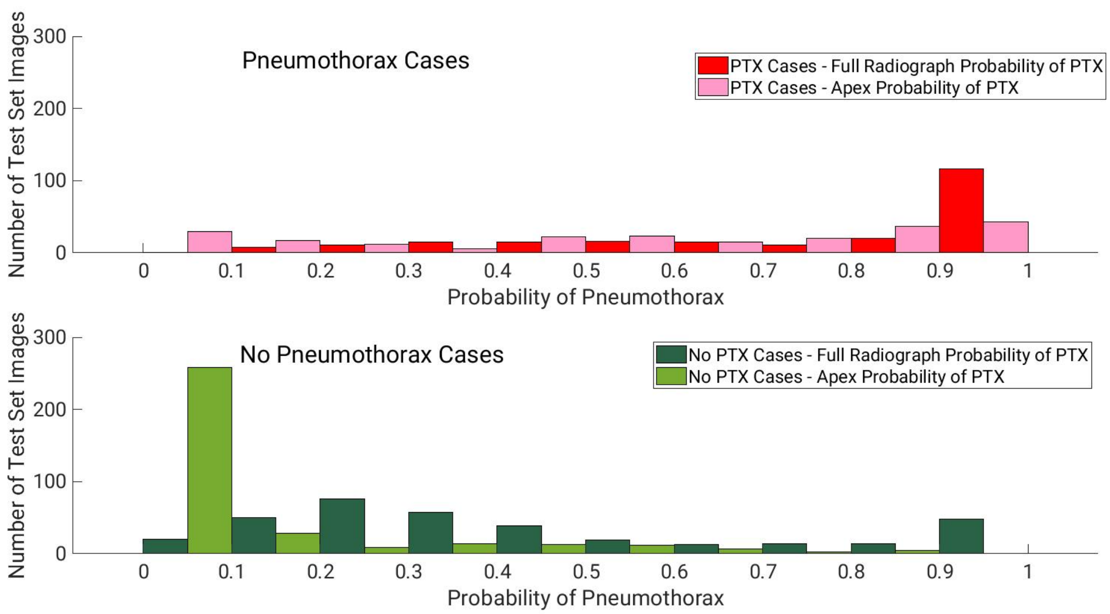
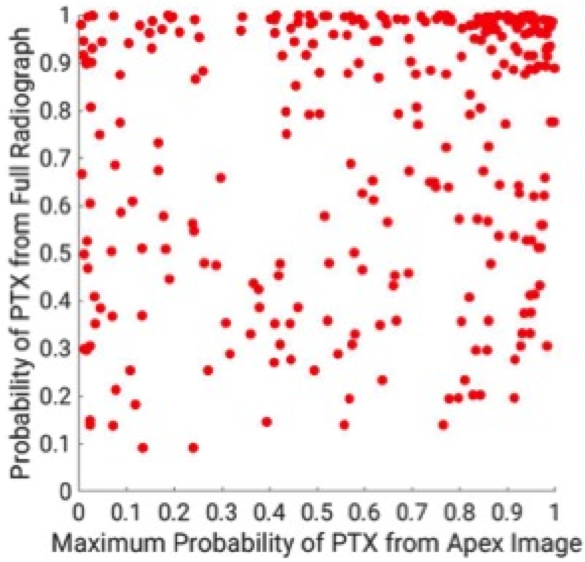
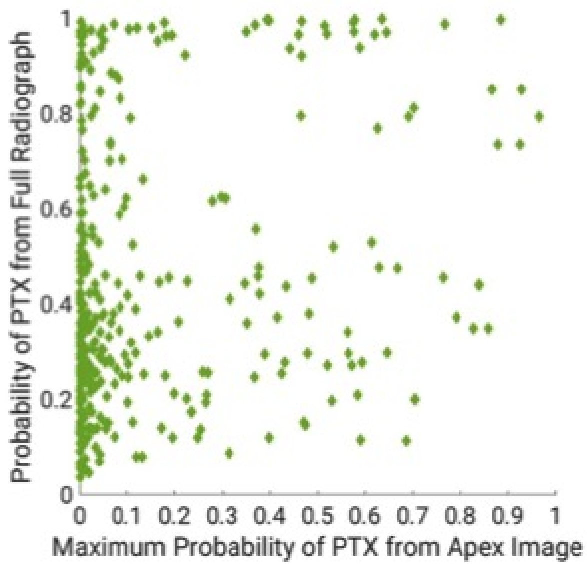


Figure 4.11: Distribution of probability of PTX from the fine-tuned VGG19 CNNs for the independent test set images plotted separately for the PTX and non-PTX cases. For the PTX cases, the probability of PTX should equal 1. For the no PTX cases, the probability of PTX should equal 0.



(a) Scatter plot of the probability of PTX from the **fine-tuned VGG19 CNNs** for the independent test set **radiographs with PTX**. Ideally, points should be in the top right corner, corresponding to a 100% probability of PTX.



(b) Scatter plot of the probability of PTX from the **fine-tuned VGG19 CNNs** for the independent test set **radiographs without PTX**. Ideally, points should be in the bottom left corner, corresponding to a 0% probability of PTX for the test set images without PTX.

Figure 4.12: Scatter plots of probability of PTX output by the fine-tuned VGG19 CNNs

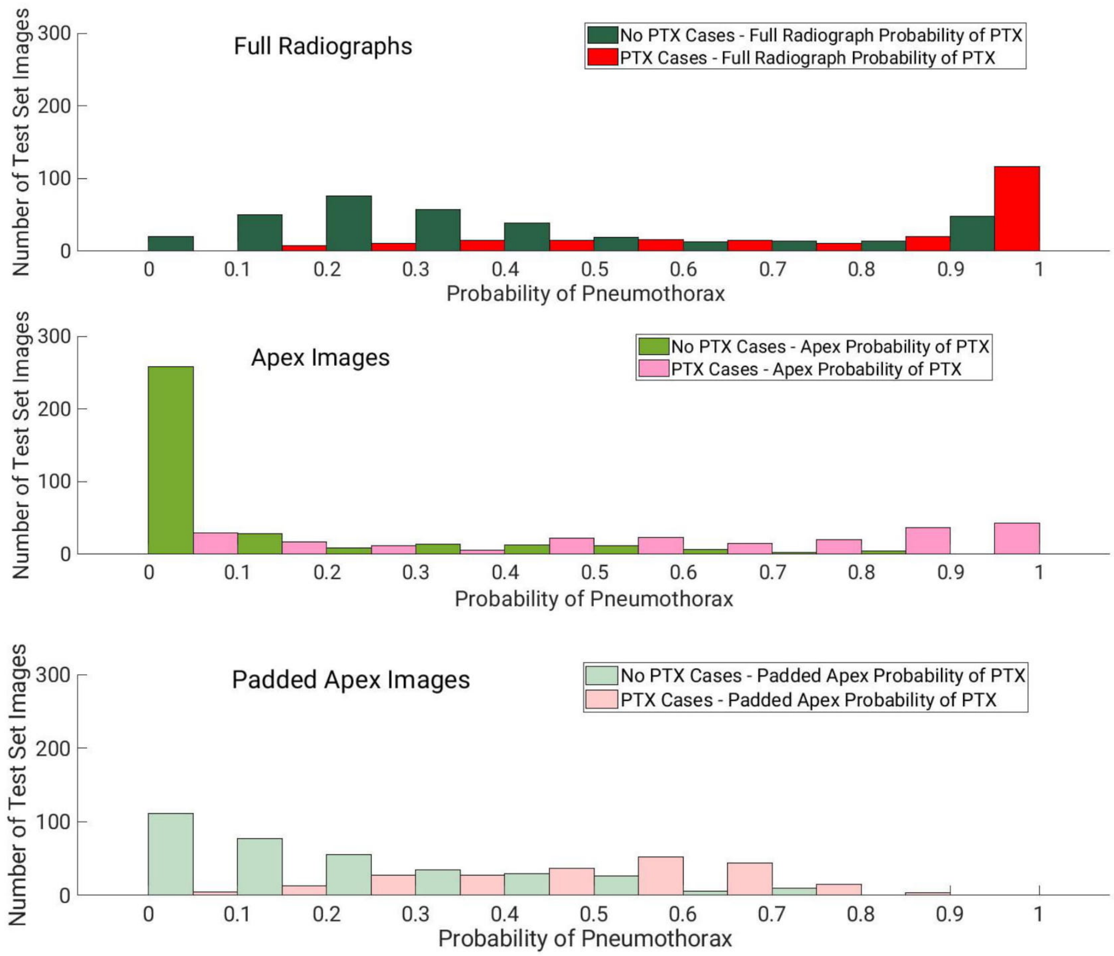


Figure 4.13: Distribution of probability of PTX output by the three fine-tuned VGG19 networks for the independent test set radiographs.

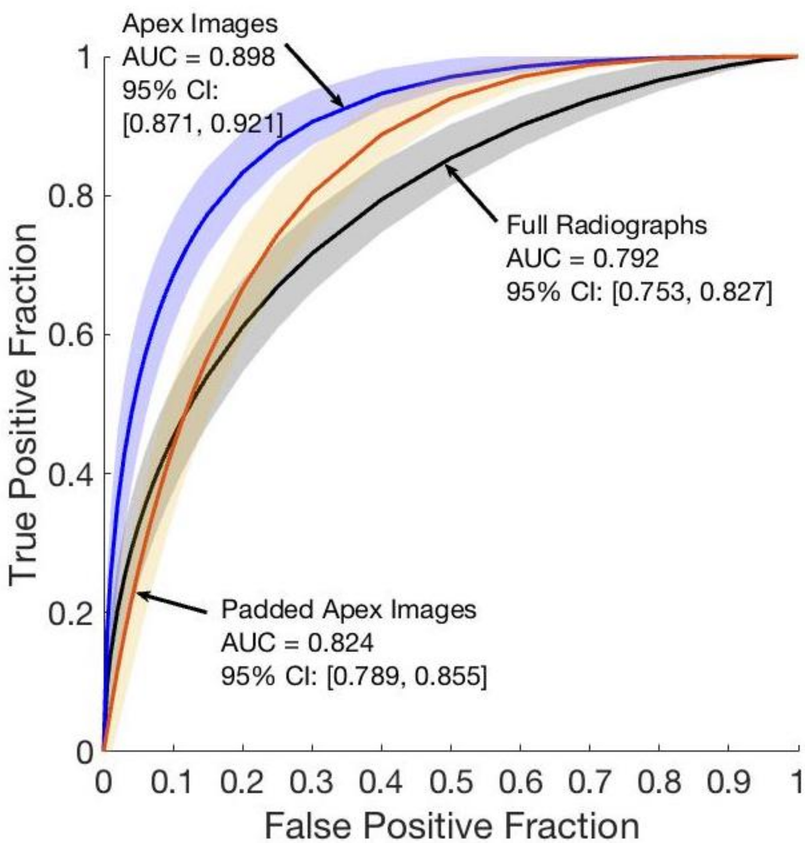


Figure 4.14: ROC curves for the three fine-tuning approaches for VGG19 showing their performances on the independent test set in the task of distinguishing between radiographs with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

4.6.1 Discussion and Conclusions from VGG19 Fine-tuning

The comparison of AUCs shows that fine-tuning a CNN with apex images rather than full radiographs leads to a statistically significant improvement in performance on the same test set ($p<0.001$) of cases with and without PTX.

From the distributions shown in Figure 4.13, and analysis of the quantity of false positives/negatives, the network fine-tuned with the full radiographs had a superior ability to correctly classify PTX cases; however, there were many false positives (non-PTX cases incorrectly classified as having PTX). The network fine-tuned with the apex images had a superior ability to correctly classify non-PTX images. The highest detection sensitivity was from the CNN fine-tuned with the full radiographs (78%) compared to the sensitivity of the CNN fine-tuned with the apex images (62%). For the specificity, the CNN fine-tuned with the apex images is the highest (89% vs. 69% for the CNN fine-tuned with the full radiographs). Due to the high percentage of false positives for the network fine-tuned with the full radiograph, the overall performance of the network fine-tuned with apex images was superior when quantified using ROC analysis in the task of detecting images with and without PTX.

4.7 ResNet-50 Results for the Full Radiographs, Apex Images, and Padded Apex Images

Table 4.6 provides the results from the performance evaluation of the fine-tuned ResNet-50 CNNs using the test set of 225 radiographs with PTX and 350 radiographs without PTX. When the ROC curves were compared using DeLong's method [42], a statistically significant difference in performance ($p<0.017$) was seen between the CNN fine-tuned with the full radiographs (AUC=0.669) and the CNN fine-tuned with the apex images (AUC=0.765) with a p-value of 0.0022. Figure 4.15 shows the probability distributions from the ResNet CNNs fine-tuned with full radiographs and apex images for the test set images with and without PTX. Figure 4.16 consists of two scatterplots showing the probability of PTX for each test case as output from analysis of its full radiograph and its corresponding apex images. A statistically significant difference in performance was also seen between the CNN fine-tuned with the padded apex images (AUC=0.657) and the CNN fine-tuned with the apex images (AUC=0.765) with a p-value of <0.001. We failed to show a statistically significant difference in performance between the CNN fine-tuned with the full radiographs (AUC=0.669) and the CNN fine-tuned with the padded apex images (AUC=0.657), with a p-value of 0.69. Figure 4.17 gives the probability distributions for the test set images for the CNNs fine-tuned with full radiographs, apex images, and padded apex images. The ROC curves for the fine-tuned ResNet CNNs are shown in Figure 4.18.

Table 4.6: Summary of the results from testing the ResNet-50 CNNs fine-tuned with the full radiographs, apex images, and padded apex images.

ResNet-50 CNN fine-tuned with:	AUC for the task of classifying between images with and without PTX (95% CI)	False Negatives	False Positives	Average Probability of PTX output by the fine-tuned CNN for cases with PTX (standard deviation)	Average Probability of PTX output by the fine-tuned CNN for cases without PTX (standard deviation)	Detection Sensitivity	Specificity
Full Radiographs	0.669 (0.622, 0.714)	3 (1%)	343 (98%)	0.82 (0.20)	0.71 (0.21)	99%	2%
Apex Images	0.765 (0.717, 0.807)	225 (100%)	0 (0%)	0.001 (0.003)	0.0003 (0.001)	0%	100%
Padded Apex Images	0.657 (0.611, 0.702)	225 (100%)	0 (0%)	0.13 (0.03)	0.12 (0.02)	0%	100%

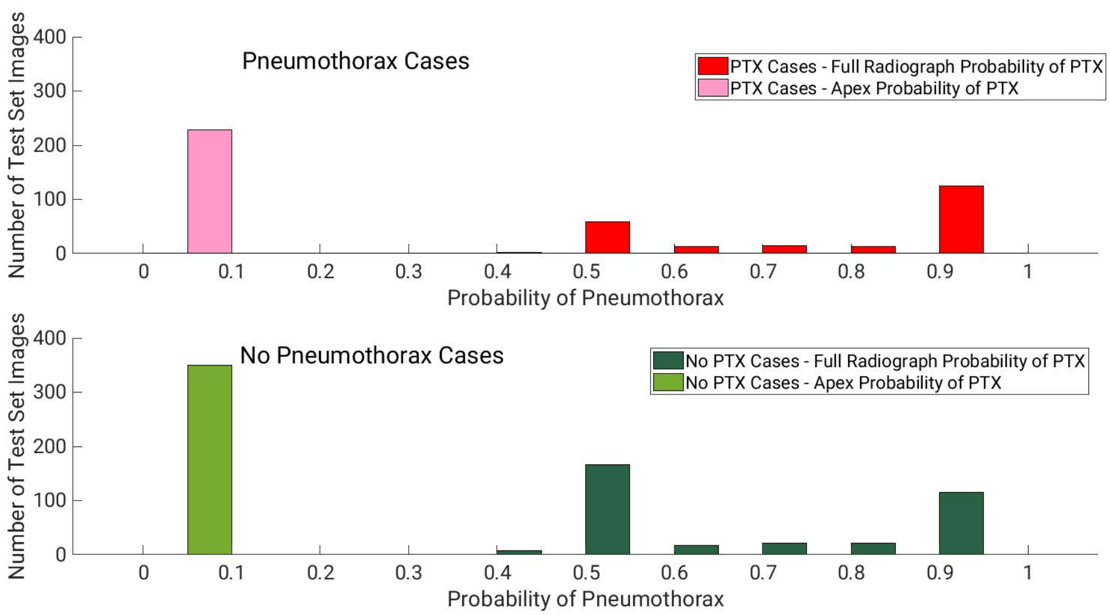
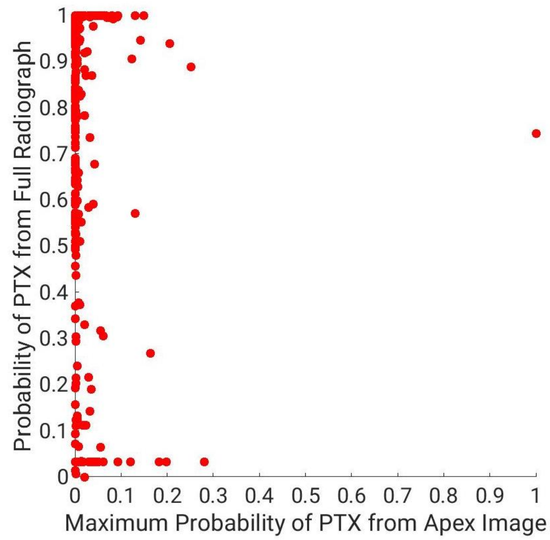
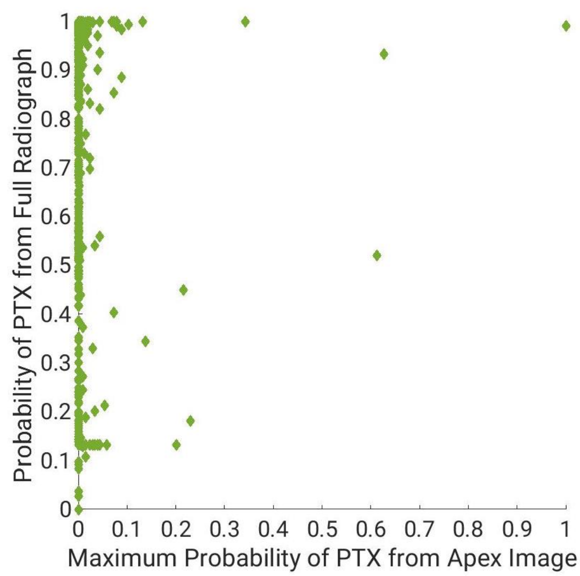


Figure 4.15: Distribution of probability of PTX from the fine-tuned ResNet CNNs for the independent test set images plotted separately for the PTX and non-PTX cases. For the PTX cases, the probability of PTX should equal 1. For the no PTX cases, the probability of PTX should equal 0.



(a) Scatter plot of the normalized probability of PTX from the **fine-tuned ResNet50 CNNs** for the independent test set **radiographs with PTX**. Ideally, points should be in the top right corner, corresponding to a 100% probability of PTX.



(b) Scatter plot of the normalized probability of PTX from the **fine-tuned ResNet50 CNNs** for the independent test set **radiographs without PTX**. Ideally, points should be in the bottom left corner, corresponding to a 0% probability of PTX for the test set images without PTX.

Figure 4.16: Scatter plots of normalized probability of PTX output by the fine-tuned ResNet50 CNNs.

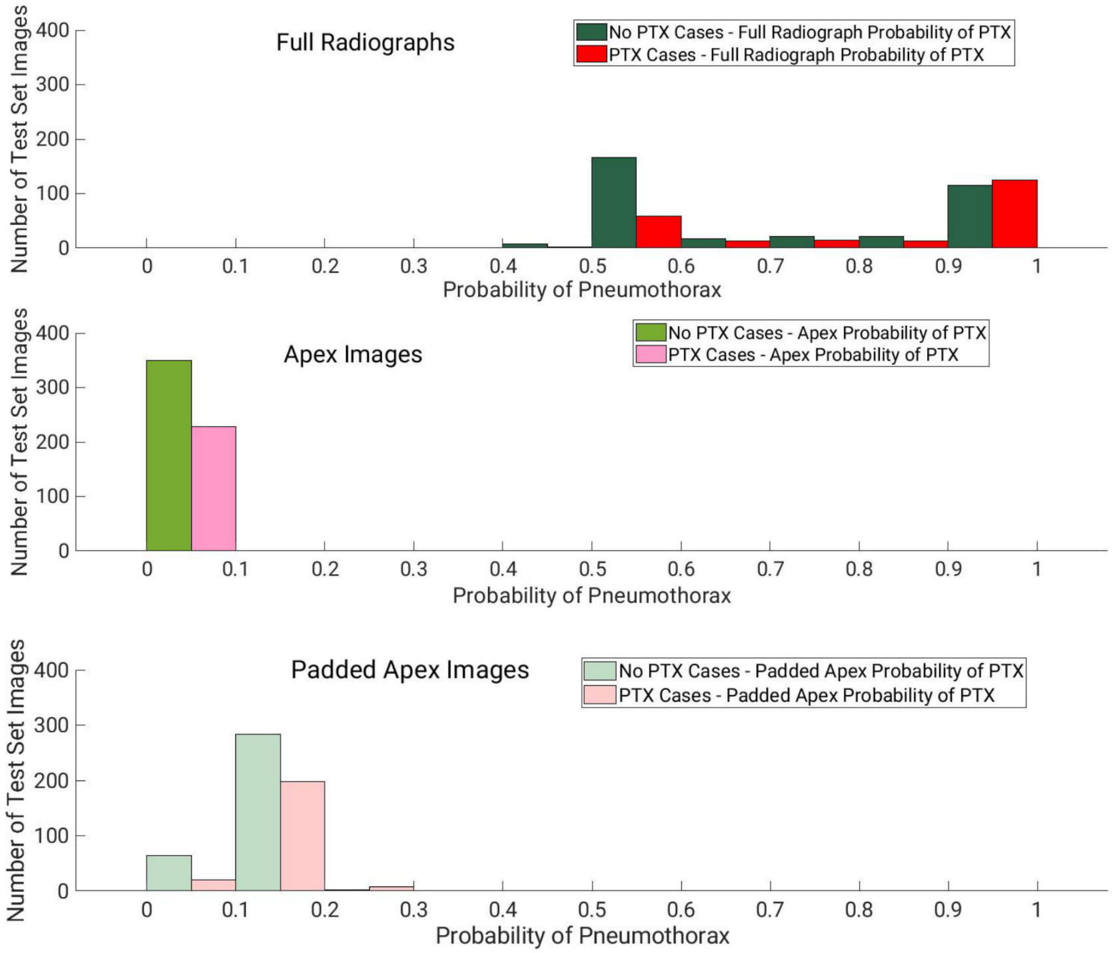


Figure 4.17: Distribution of probability of PTX output by the three fine-tuned ResNet networks for the independent test set radiographs. An ideal distribution would have the probability of PTX equal to 1 for the PTX cases and a probability of PTX equal to 0 for the cases without PTX.

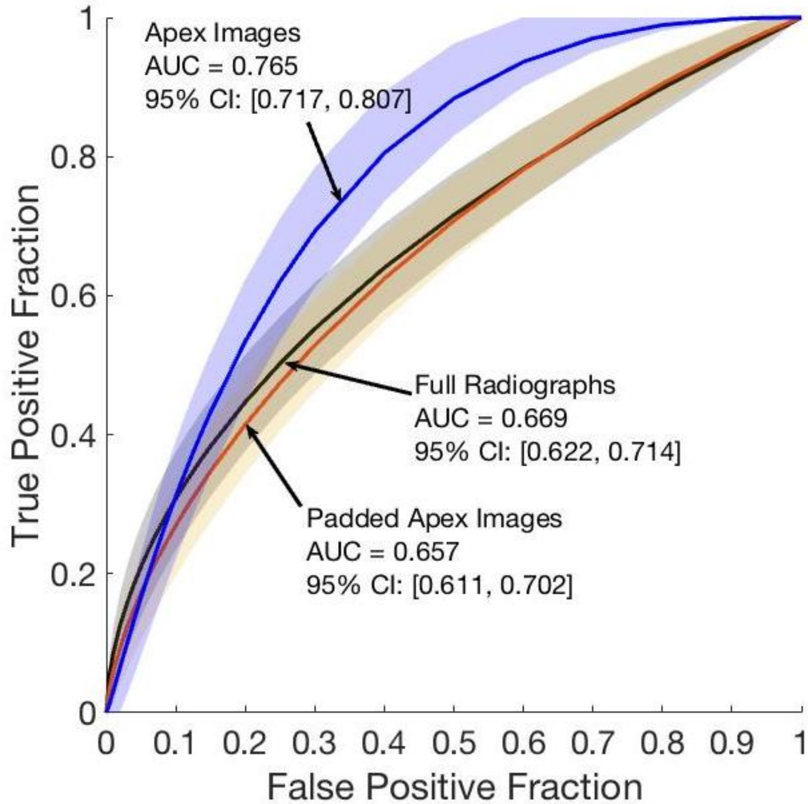


Figure 4.18: ROC curves for the three fine-tuning approaches for ResNet50 showing their performances on the independent test set in the task of distinguishing between radiographs with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

4.7.1 Discussion and Conclusions from ResNet50 Fine-tuning

The fine-tuned ResNet50 CNN output was distinct from the other fine-tuned networks. For the network fine-tuned with the full radiographs, almost all the test set images were assigned a probability of PTX over 0.5 (99% detection sensitivity and 2% specificity). However, for the CNNs fine tuned with the apex images and padded apex images, all test set images were assigned a probability of PTX less than 0.5, leading to a detection sensitivity of 0%. A potential cause for this unusual behavior is underfitting; retraining only the final two convolutional layers and the final classification layer for the new task may not have been sufficient to adequately learn the underlying structure of the training data.

However, the probability threshold for classifying an image as having PTX or not is arbitrary. When ROC analysis was performed, the performance was stronger than the sensitivity/specificity would suggest; the CNN fine-tuned with apex images yielded an AUC of 0.765 (95% CI: 0.717, 0.807) in the task of distinguishing between images with and without PTX. Despite almost all of the test set images being classified as having PTX by the CNN fine-tuned with full radiographs, a statistically significant difference in performance was not observed between that CNN and the CNN fine-tuned with padded apex images ($p=0.685$), which classified all test set images as not having PTX.

4.8 BagNet-33 Results for the Full Radiographs, Apex Images, and Padded Apex Images

Table 4.7 provides the results from the performance evaluation of the fine-tuned BagNet-33 CNNs using the test set of 225 radiographs with PTX and 350 radiographs without PTX. When the ROC curves were compared using DeLong's method [42] a statistically significant difference in performance ($p<0.017$) was seen between the CNN fine-tuned with the full radiographs (AUC=0.784) and the CNN fine-tuned with the apex images (AUC=0.922) with a p-value of <0.001. Figure 4.19 shows the probability distributions from the BagNet CNNs fine-tuned with full radiographs and apex images for the test set images with and without PTX. Figure 4.20 consists of two scatterplots showing the probability of PTX for each test case as output from analysis of its full radiograph and its corresponding apex images. Statistically significant differences in performance were seen between the CNN fine-tuned with the padded apex images (AUC=0.890) and the CNN fine-tuned with the apex images (AUC=0.922) with a p-value of 0.012, as well as between the CNN fine-tuned with the full radiographs (AUC=0.784) and the CNN fine-tuned with the padded apex images (AUC=0.890), with a p-value <0.001. Figure 4.21 gives the probability distributions for the test set images for the CNNs fine-tuned with full radiographs, apex images, and padded apex images. The ROC curves for the fine-tuned BagNet CNNs are shown in Figure 4.22.

Table 4.7: Summary of the results from testing the BagNet-33 CNNs fine-tuned with the full radiographs, apex images, and padded apex images.

BagNet-33 CNN fine-tuned with:	AUC for the task of classifying between images with and without PTX (95% CI)	False Negatives	False Positives	Average Probability of PTX output by the fine-tuned CNN for cases with PTX (standard deviation)	Average Probability of PTX output by the fine-tuned CNN for cases without PTX (standard deviation)	Detection Sensitivity	Specificity
Full Radiographs	0.784 (0.743, 0.821)	29 (13%)	196 (56%)	0.75 (0.17)	0.59 (0.16)	87%	44%
Apex Images	0.922 (0.898, 0.942)	85 (38%)	15 (4%)	0.60 (0.36)	0.08 (0.16)	62%	96%
Padded Apex Images	0.890 (0.861, 0.914)	127 (56%)	17 (5%)	0.47 (0.31)	0.10 (0.16)	44%	95%

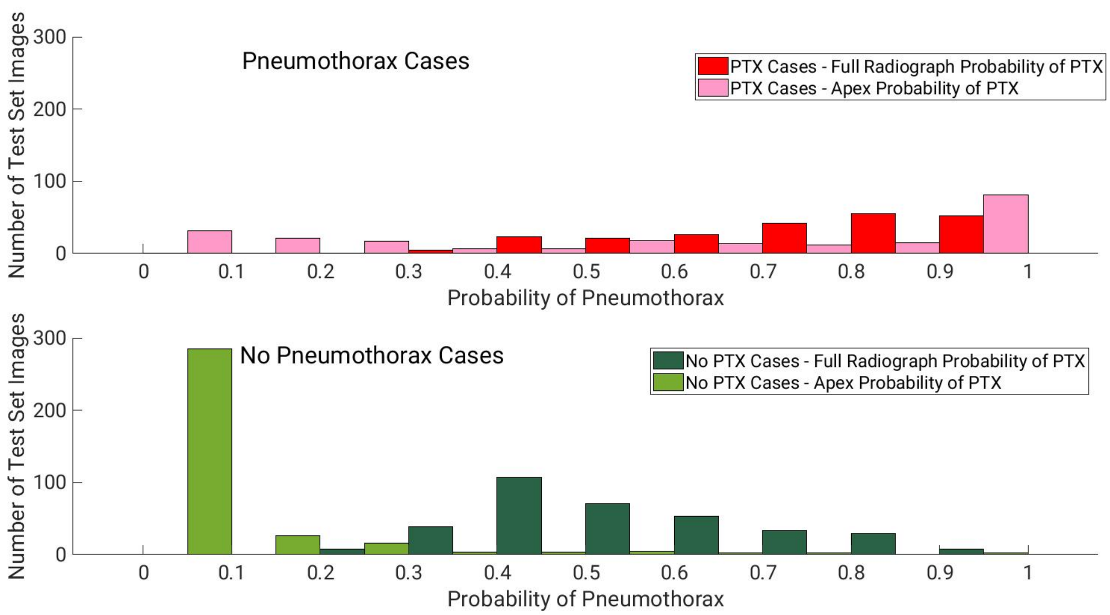
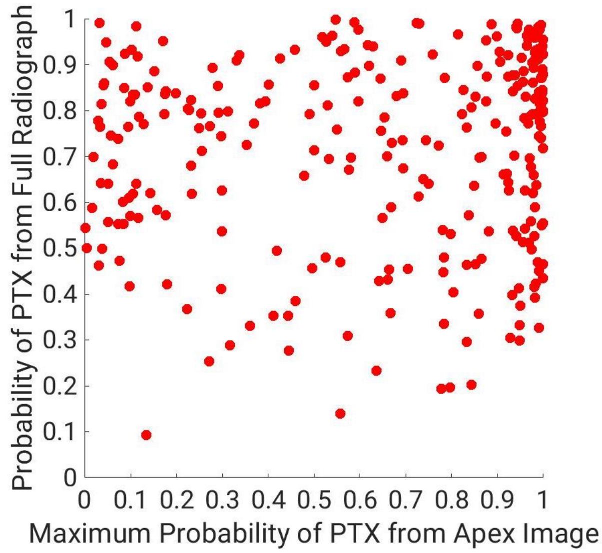
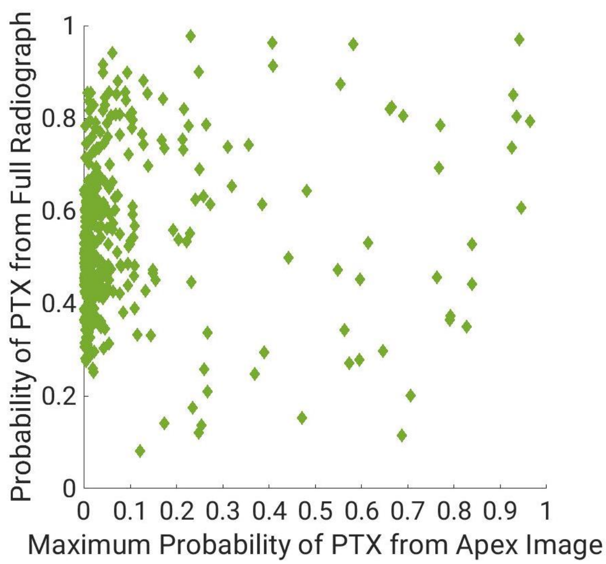


Figure 4.19: Distribution of probability of PTX from the fine-tuned BagNet CNNs for the independent test set images plotted separately for the PTX and non-PTX cases. For the PTX cases, the probability of PTX should equal 1. For the no PTX cases, the probability of PTX should equal 0.



(a) Scatter plot of the probability of PTX from the **fine-tuned BagNet CNNs** for the independent test set **radiographs with PTX**. Ideally, points should be in the top right corner, corresponding to a 100% probability of PTX.



(b) Scatter plot of the probability of PTX from the **fine-tuned BagNet CNNs** for the independent test set **radiographs without PTX**. Ideally, points should be in the bottom left corner, corresponding to a 0% probability of PTX for the test set images without PTX.

Figure 4.20: Scatter plots of probability of PTX output by the fine-tuned BagNet CNNs

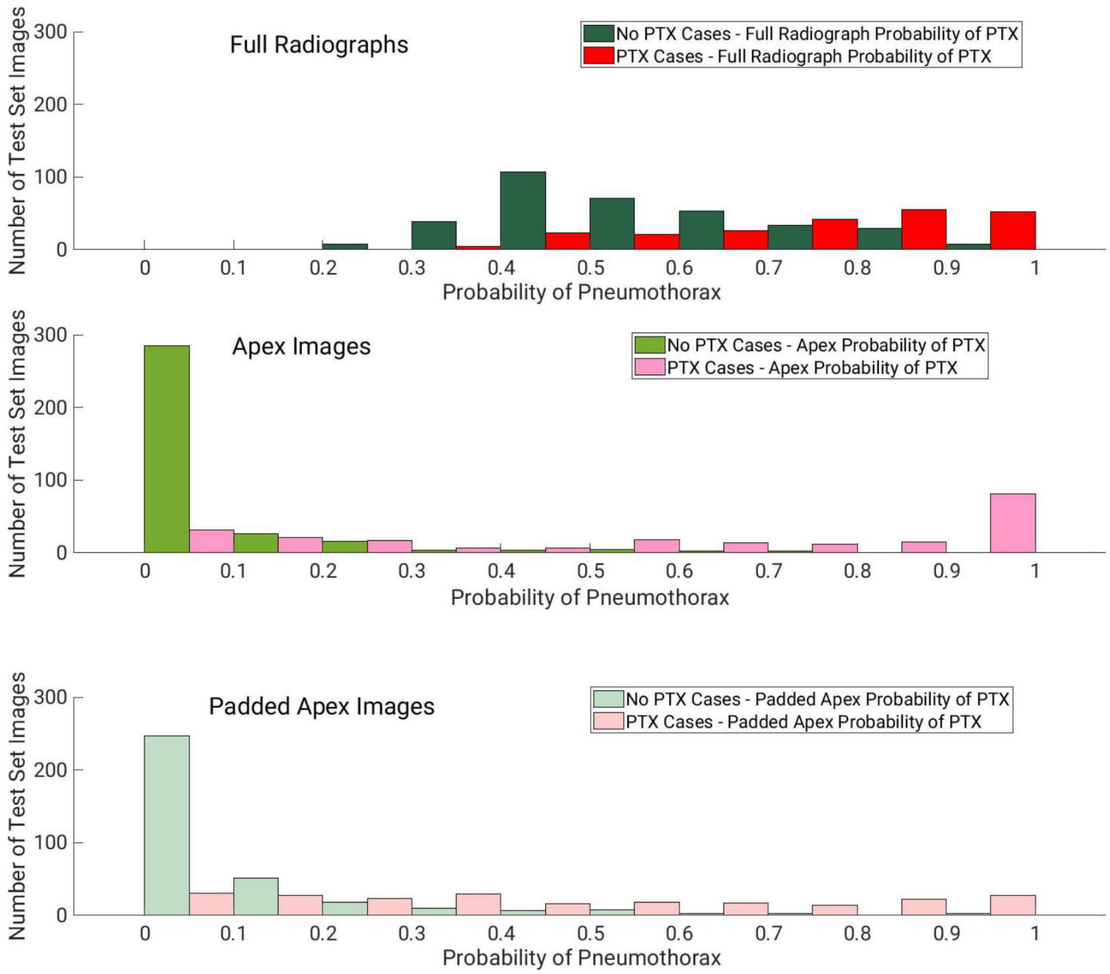


Figure 4.21: Distribution of probability of PTX output by the three fine-tuned BagNet networks for the independent test set radiographs. An ideal distribution would have the probability of PTX equal to 1 for the PTX cases and a probability of PTX equal to 0 for the cases without PTX.

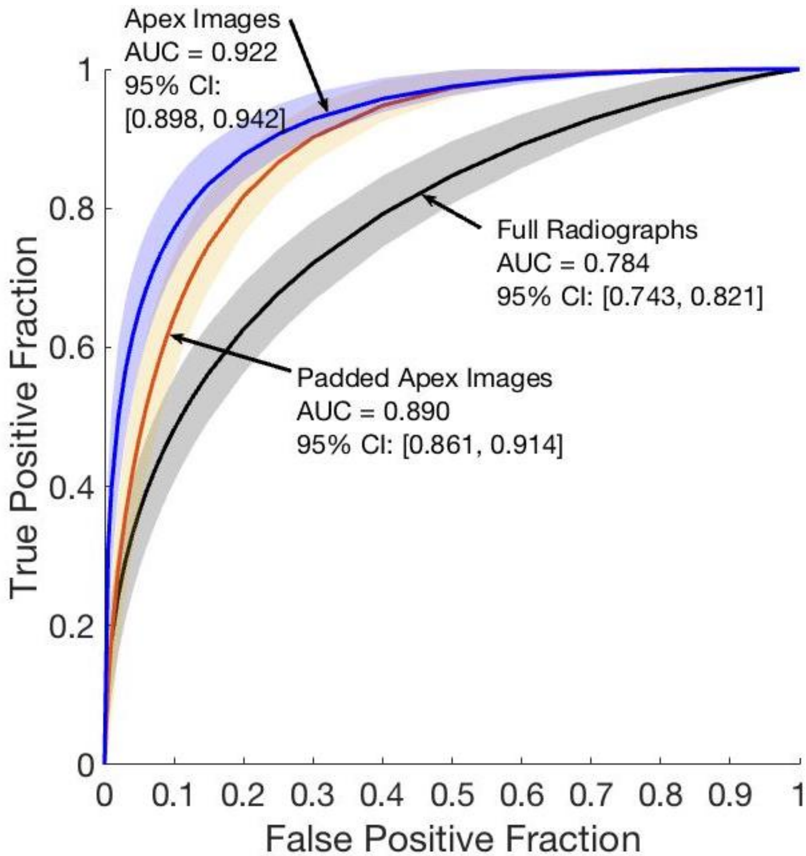


Figure 4.22: ROC curves for the three fine-tuning approaches for BagNet showing their performances on the independent test set in the task of distinguishing between radiographs with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

4.8.1 Discussion and Conclusions from BagNet Fine-tuning

The BagNet network is a variant of the ResNet network; however, the fine-tuned BagNet CNN's performance was not similar to that of the fine-tuned ResNet. All of the performances of the fine-tuned BagNet networks were statistically significantly different from one another.

The detection sensitivity was highest for the CNN fine tuned with the full radiographs (87% versus 62% for the apex images and 44% for the padded apex images). However, the specificity is highest for the CNN fine-tuned with the apex images (96% versus 44% for the full radiographs and 95% for the padded apex images).

4.9 Comparison of the Fine-Tuned CNNs: AlexNet, VGG19, ResNet50, and BagNet33

The purpose of this work was to use deep learning to improve medical image diagnosis, with emphasis on the effect of image resolution on deep learning performance. Each of the four networks had varying performance on the test set, which are summarized in Table 4.8. Figure 4.23 plots the AUC values for each of the fine-tuned CNNs in the task of classifying between images with and without PTX.

Table 4.8: Performance of the four fine-tuned networks for the full radiographs, apex images, and padded apex images.

		AUC	Detection Sensitivity	Detection Specificity
AlexNet	**Full Radiographs**	0.702 (95% CI: 0.657, 0.745)	67%	64%
	Apex Images	0.862 (95% CI: 0.831, 0.890)	56%	89%
	Padded Apex Images	0.823 (95% CI: 0.787, 0.854)	45%	89%
VGG19	**Full Radiographs**	0.792 (95% CI: 0.753, 0.827)	78%	69%
	Apex Images	0.898 (95% CI: 0.871, 0.921)	62%	89%
	Padded Apex Images	0.824 (95% CI: 0.789, 0.855)	52%	88%
ResNet50	**Full Radiographs**	0.669 (95% CI: 0.622, 0.714)	99%	2%
	Apex Images	0.765 (95% CI: 0.717, 0.807)	0%	100%
	Padded Apex Images	0.657 (95% CI: 0.611, 0.702)	0%	100%
BagNet33	**Full Radiographs**	0.784 (95% CI: 0.743, 0.821)	87%	44%
	Apex Images	0.922 (95% CI: 0.898, 0.942)	62%	96%
	Padded Apex Images	0.890 (95% CI: 0.861, 0.914)	44%	95%

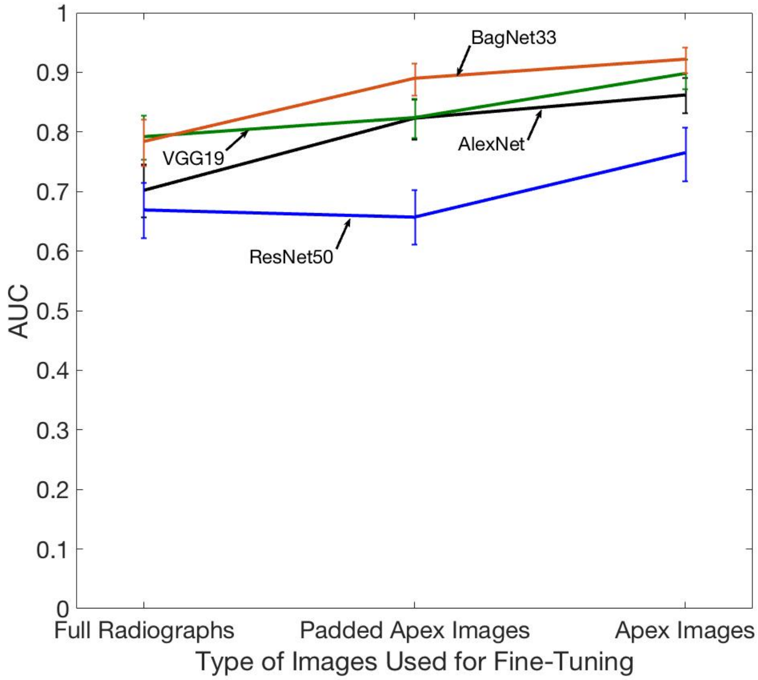


Figure 4.23: AUC values for each fine-tuned network architecture (AlexNet, VGG19, ResNet50, BagNet33) for the task of classifying between images with and without PTX for the three image types (full radiographs, padded apex images, and apex images). The error bars show the 95% confidence interval for the AUC.

To compare the fine-tuned CNNs directly to one another when they are fine-tuned with the same type of image, the ROC curves were compared and the p-values calculated using DeLong's method [42]. Figure 4.24 shows the ROC curves for the four network architectures fine-tuned for classification of the full radiograph. Table 4.9 shows the p-values for the comparison of performance between the networks. Comparing the performance of the four networks to one another leads to 6 comparisons. After correcting the level of statistical significance for multiple comparisons using the Bonferroni method [46], the level of statistical significance was 0.0083.

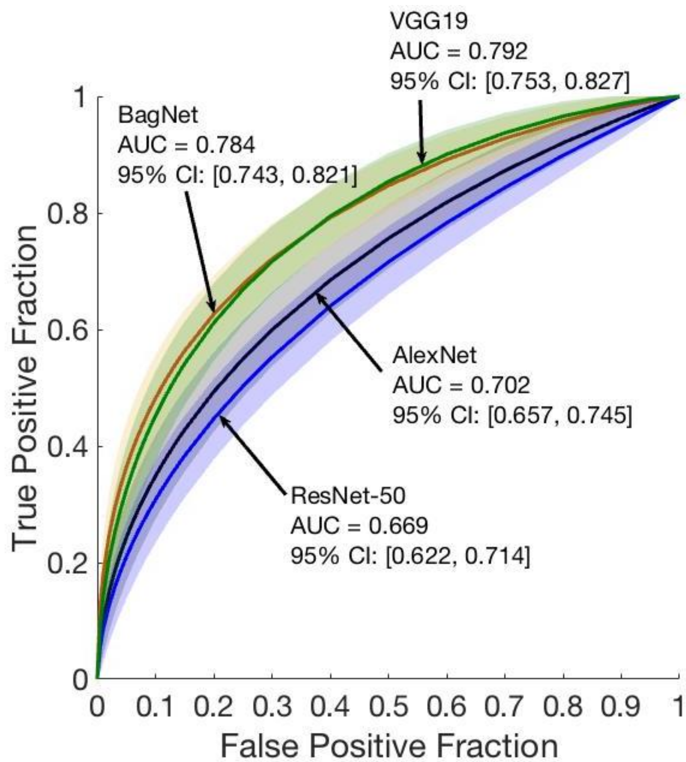


Figure 4.24: ROC curves for the four fine-tuned networks in the task of classifying between full radiographs with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve. The curve colors are: black (AlexNet), blue (ResNet-50), green (VGG19), and orange (BagNet).

For the full radiograph, the performance of the fine-tuned BagNet CNN and the fine-tuned VGG19 were similar, failing to show a statistically significant difference in performance (p=0.627). The performance of the fine-tuned AlexNet network and the ResNet50 network also failed to show a statistically significant difference (p=0.318). However, the remainder of the comparisons showed statistically significant differences in performance (Table 4.9).

Table 4.9: p-values from comparisons of the ROC curves for the four fine-tuned networks in the task of classifying full radiographs with and without PTX. p-values were calculated using DeLong's method to compare ROC curves.

	BagNet33	**ResNet50**	**VGG19**
AlexNet	0.0053	0.318	0.0012
BagNet33		<0.001	0.627
ResNet50			<0.001

Figure 4.25 shows the ROC curves for the four network architectures fine-tuned for classification of the apex images. The p-values resulting from comparing the ROC curves for the CNNs fine-tuned with apex images to one another are shown in Table 4.10.

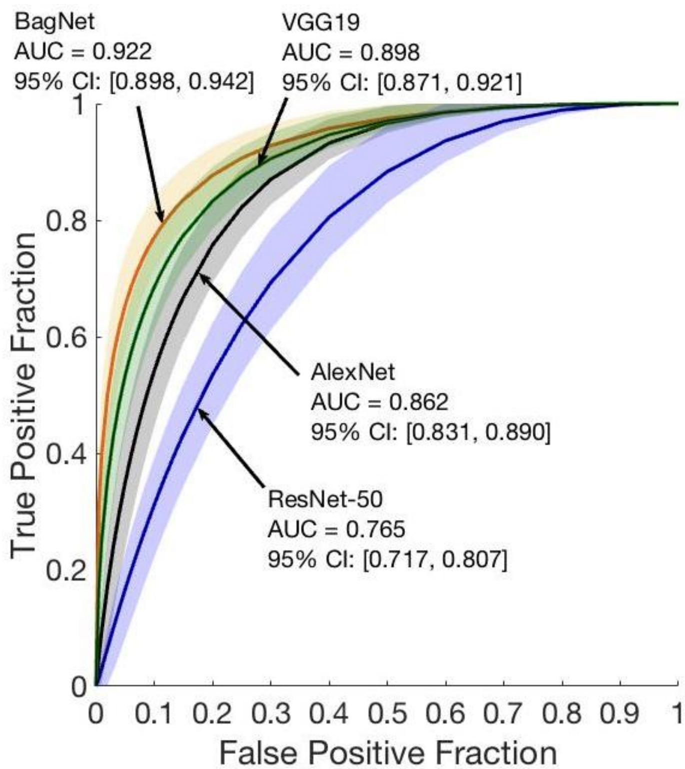


Figure 4.25: ROC curves for the four fine-tuned networks in the task of classifying between apex images with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

For the apex images, the only comparison between network performance that failed to show statistical significance was between the fine-tuned BagNet network and the fine-tuned VGG19 network (p=0.086).

Table 4.10: p-values from comparisons of the ROC curves for the four fine-tuned networks in the task of classifying apex images with and without PTX. p-values were calculated using DeLong's method to compare ROC curves.

	BagNet33	**ResNet50**	**VGG19**
AlexNet	<0.001	<0.001	0.004
BagNet33		<0.001	0.086
ResNet50			<0.001

Figure 4.26 shows the ROC curves for the four network architectures fine-tuned for classification of the padded apex images. The p-values resulting from comparing the four CNNs fine-tuned with the padded apex images to one another are shown in Table 4.11.

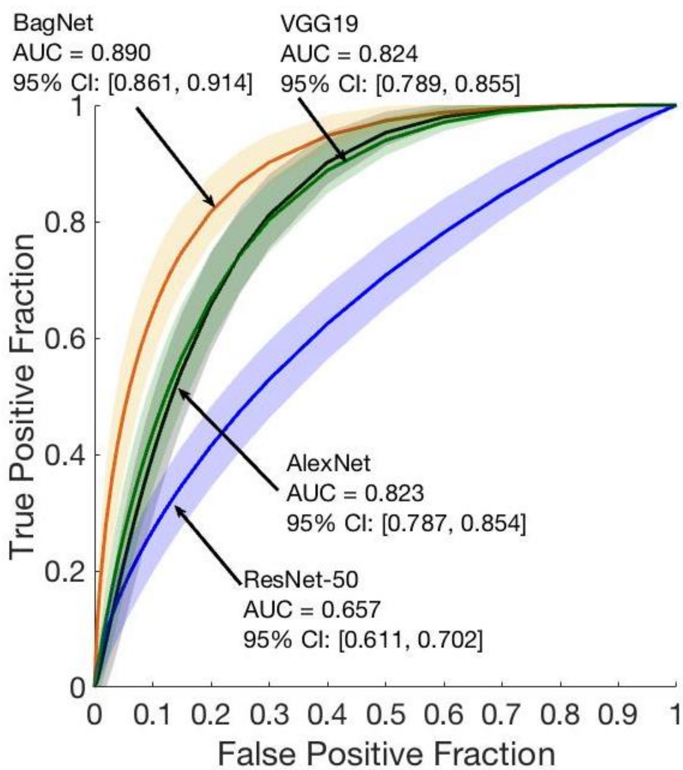


Figure 4.26: ROC curves for the four fine-tuned networks in the task of classifying between padded apex images with and without PTX. The shaded region around each line denotes the 95% confidence interval for the curve.

For the CNNs fine-tuned with the padded apex images, the only statistically significant difference in performance was seen between the performance of the fine-tuned AlexNet network and the VGG19 network (p=0.804).

Table 4.11: p-values from comparisons of the ROC curves for the four fine-tuned networks in the task of classifying padded apex images with and without PTX. p-values were calculated using DeLong's method to compare ROC curves.

	BagNet33	**ResNet50**	**VGG19**
AlexNet	<0.001	<0.001	0.804
BagNet33		<0.001	<0.001
ResNet50			<0.001

4.10 Feature Extraction

Feature extraction, the use of intermediate features from a pre-trained network that are then used in a classifier, was also performed for the classification of full radiographs and apex images with and without PTX. The software used for extracting the intermediate features from the pre-trained network was Keras (version 2.2.2) implemented in Python (version 3.5.5) with a Tensorflow (Tensorflow-gpu, version 1.10.0) backend on a nVidia Tesla V100 GPU. The features were extracted from the chosen layers and then pooled to reduce the dimensions. The features were then normalized and concatenated into a text file.

After the intermediate features were extracted, the text file of features along with the truth for each image were loaded into Matlab (version R2017b). Principal component analysis (PCA) was performed to reduce the dimensionality to have a feature space of 50 features. Then a support vector machine (SVM) was trained, with 5-fold cross validation, to construct predictions.

For fine-tuning, the training, validation, and testing sets were distinct; however, for feature extraction all the images were grouped together, since 5-fold cross validation was used with the SVM. Cross validation is an alternative to the use of an independent test set for testing the performance of CNNs. The 5-fold cross-validation was performed over 50 runs and the average AUC is reported. The standard deviation of the AUC was also calculated.

4.10.1 Network Specific Methods and Feature Extraction Results

VGG19

The VGG19 CNN was pretrained on ImageNet. The intermediate features were extracted from each of the five max-pooling layers of the CNN. The intermediate features were then downsampled using average-pooling, which summarizes the average presence of a feature, to reduce the dimensions. For the full radiograph, the average AUC over the 50 runs was 0.735 with a standard deviation of 0.037. For the apex images, the average AUC over the 50 runs was 0.776 with a standard deviation of 0.026.

ResNet50

The ResNet50 CNN was pretrained on ImageNet and the intermediate features were extracted from the max-pooling layer just before the final classification layer. For VGG19, there were 5 max-pooling layers between blocks of convolutional layers; however, in the ResNet50 architecture, there is only one max-pooling layer, just before the final classification layer.

For the full radiograph, 5-fold cross validation over 50 runs yielded an average AUC of 0.503 with a standard deviation of 0.023. For the apex images, the average AUC was 0.501 with a standard deviation of 0.015.

Comparison of Feature Extraction Results

Feature extraction requires the extraction of features from intermediate layers of a CNN, creating large feature vectors. The dimensionality is reduced through PCA and feature selection; however, the processing and storage of the large feature vectors in the meantime is computationally expensive. If the performance was better than that of fine-tuning, the computational cost may be worthwhile; however, for this task, fine-tuning yields better performance.

In addition, since the feature vectors extracted are not human-interpretable, selecting which layers to extract from becomes a trial-and-error process. As will be discussed further in the next chapter, explainability and interpretability is vital for clinical implementation of deep learning. Since the feature vectors extracted are simply arrays of numbers, it is difficult to determine how the pre-trained CNN is interpreting the new input image. The additional steps of feature value pooling, feature reduction, and SVM further dilute the explainability of the output.

4.11 Discussion & Conclusions

Published studies have reported strong performance for the detection of various lung diseases on chest radiographs when using deep learning algorithms. In this chapter, CNNs were used for the detection of PTX to determine the effect of image resolution on performance when using deep learning to improve medical image diagnosis.

Fine-tuning yielded a stronger performance compared to feature extraction. Four different architectures were fine-tuned in order to more fully evaluate the impact of image resolution to observe general trends, rather than investigating the impact of image resolution on a single architecture. For all four of the fine-tuned networks, the performance of the CNNs fine-tuned with apex images was the highest and statistically significantly different than the performance with any other image type (full radiograph or padded apex image). The detection sensitivity was highest for the full radiographs for all four networks; however, the specificity was highest for the networks fine-tuned with the apex images or equivalent to the specificity for the padded apex images.

For the specific case of PTX, deep learning could be applied to triage acquired radiographs and prioritize them for radiologist interpretation based upon the likelihood of a condition requiring immediate care. In that case, it would likely be more helpful to have a high sensitivity so patients with PTX are not overlooked. In order to balance a high classification performance with a high sensitivity, future work could investigate the combination of output from a CNN fine-tuned with full radiographs and a CNN fine-tuned with apex images.

These results showing the effect of image resolution of the input to deep learning algorithms demonstrates the importance of neural networks trained for specific tasks, which should be considered when training a single network to classify many diseases on a chest radiograph. Some thoracic diseases may have visual signs that are visible on downsampled radiographs; however, more visually subtle diseases, such as PTX, may require different methods to address image resolution when training a network for detection.

To further investigate the disparity in performance between the full radiograph and the apex images, as well as to address the human interpretability of deep learning output, visualization methods will be presented in Chapter 5. The visualization will allow for further analysis and determination of what aspects of the radiograph are most influential for the classification. Further investigation of the impact of increased input image resolution, in the form of patches of the apex images, is presented as related work in Appendix A. Appendix B provides more analysis of the results for different subgroups within this dataset for the fine-tuned VGG19 CNNs.

CHAPTER 5

INVESTIGATION OF DEEP LEARNING METHODS FOR THE VISUALIZATION, AND SUBSEQUENT EXPLANATORY AI, OF PTX IN CHEST RADIOGRAPHS

5.1 Introduction & Motivation

There are many potential applications of deep learning for the improvement of radiology practice, in addition to the application of deep learning to improve diagnosis from medical images as discussed in the previous chapter. Some of these applications include postprocessing, image quality analytics, radiation dose estimation, and radiology reporting, along with many others [7]. Deep learning has the potential to greatly impact radiology practice; however, several challenges have been identified. One such challenge is deciphering the "black box" of deep learning. Other challenges include the lack of large annotated medical image datasets, lack of standards, lack of regulations, workflow integration, radiologist perspectives, and medicolegal issues [7]. Being able to decipher the "black box" would help address some of the other challenges, including the need for large datasets, regulatory issues, and radiologist perspectives.

Interpretability, being able to decipher the "black box", is the degree to which a human can understand the cause of a decision or the degree to which a human can consistently predict the model's results [48]. Interpretability can help to address the dataset size requirement for deep learning. Interpretability helps detect bias in the model; biases are learned from the training data, so interpretability can help determine whether the training data are large enough and diverse enough. In addition, interpretability helps to check causality, ensuring that the training data are sufficient for the network to learn causal relationships. Interpretability can also help with regulatory challenges; interpretability can help verify that predictions are not discriminatory (explicitly or implicitly) against protected groups, in addition to helping to ensure that privacy and sensitive information in the data is protected. Interpretability can also assist radiologists in understanding the output of the deep learning algorithms; radiologists need to be able to trust the algorithms in order to apply them to clinical tasks where they affect patient care. A recent study that surveyed clinicians to determine the type of explanations they want from clinical systems found that there is a need for CNN models that reflect a similar analytic process to established methodology of evidence-based decision making in the clinic [49].

Explanations of CNN output can be provided by visualizations of what CNNs have learned on images. There are three primary methods for visualizing the output of a trained CNN on images: (i) visualizing intermediate activations, (ii) visualizing layer filters, and (iii) visualizing heatmaps of class activation in an image [5]. Visualizing intermediate activations provides a view into how an input is decomposed into the different filters learned by the CNN. Visualizing filters displays the visual pattern each filter is meant to respond to, which is abstract and difficult to apply for general interpretability for clinical tasks. Heatmaps of class activation are helpful for understanding which aspects of an image were identified as belonging to a given class. The class activation heatmap is a two-dimensional grid of scores associated with a specific output class [5]. The heatmap indicates how important each location within the image is with respect to the class under consideration; each image has a heatmap corresponding to each class the network can classify. Class activation maps can help to localize the object within the image, as well as to investigate the causes of misclassification.

In one of the comparison studies mentioned in the previous chapter, Rajpurkar et al. [9], generated class activation maps as part of their study and published one with PTX; however, they did not analyze the accuracy of the visualization over their full dataset. In another comparison study, Wang et al. [11], performed class activation mapping and analyzed the overlap between the influential region of the heatmap and the bounding box ground truth. Various thresholds were applied to identify the influential region of the heatmap, and the overlap was reported as an intersection over union. The accuracy ranged from 8% when the heatmap was thresholded such that only the top 10% of values denoted the influential region to 46% when the heatmap was thresholded such that the top 90% of values denoted the influential region.

As discussed in the previous chapter, the CNNs fine-tuned with apex images had the strongest performance in the task of classifying between images with and without PTX, particularly for the fine-tuned VGG19 network (AUC= 0.898) and the fine-tuned BagNet network (AUC= 0.922). Visualization can assist in comparing the performance of the different CNN architectures, as well as comparing the performance of the full radiographs and the apex images. For this work, class activation mapping was performed with a couple of visualization tools. Visualization was performed to determine whether the fine-tuned CNNs are actually being trained to identify the most specific visual signs of PTX that radiologists use for detection of PTX, including the fine line at the edge of the lung and absence of lung texture outside the lung.

5.1.1 Overview of Visualization Tools

Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to visualize network output [50]. Grad-CAM is a visualization tool, not a network architecture, and can be applied to any neural network architecture that contains fully connected layers.

Grad-CAM uses the gradient information going into the final convolutional layer of a trained network to determine the neurons that are the most important for the classification decision. Grad-CAM outputs a "heatmap," which is a 2-D grid of scores associated with an output class, indicating how important each location in the input image is for the classification [5]. The heatmap can also be considered to be a spatial map of "how intensely the input image activates the class" [5].

Grad-CAM is a common visualization technique used in both of the comparison studies previously mentioned [9, 11]. It is applied after training is complete, therefore it does not require any modification of CNN architecture or retraining. There must be a balance between making a model interpretable while still truthfully representing the model's output; Selvaraju, et al. [50] compared the performance of Grad-CAM to occlusion mapping, which is currently considered to be the most faithful representation of CNNs. They found that Grad-CAM was more interpretable by humans than occlusion mapping and was more faithful to the model than previous state-of-the-art visualization methods, Class Activation Mapping (CAM) and contrastive Marginal Winning Probability (c-MWP).

The limitations of Grad-CAM include the inability to localize multiple occurrences of the same class and its failure to localize the entire region of the object [51]. Since comparison studies use Grad-CAM and it is considered more faithful/human interpretable than other visualization methods, Grad-CAM was applied for visualization of the fine-tuned networks' output.

BagNet

As discussed in Chapter 2, BagNet was created in order to combine the interpretability of bag-of-features models with the high performance of CNNs. Class evidence from patches of the image is summed to create a full image classification as well as a heatmap of the class evidence. Since the class evidence is evaluated on a patch-wise basis, the heatmap is very specific, identifying specific pixels, rather than a general region as Grad-CAM does [37].

BagNet is an architecture of its own, which sums up the number of local image features to determine the final classification. The heatmap for each test set image is a component of the network's output; it is not a visualization tool separate from the CNN as Grad-CAM is. Since interpretability of CNNs is vital for clinical implementation of CNNs in medical imaging, BagNet was used to generate a different type of heatmap to compare to the Grad-CAM heatmaps implemented for the other three CNN architectures (AlexNet, VGG19, and

ResNet50).

5.2 Methods

5.2.1 Dataset

The dataset used for visualization was the same dataset used for the detection of pneumothorax via fine-tuning and feature extraction (Table 4.2). The fine-tuned networks were used for the visualizations presented in this chapter. Visualization of the fine-tuned CNNs' output on the independent test set was performed for the full radiographs and the apex images.

Once the ground truth was established for the dataset, labeled as having PTX or not, a radiologist with 18 years of clinical experience drew contours on 225 radiographs with PTX in the apex of the lung. The radiologist was instructed to draw a closed contour around the free air in the pleural space in order to encompass the full spatial extent of the PTX. The contouring was performed in a darkened room and the window/level could be changed for each image. The radiologist could not zoom the image. The contours were used for the quantitative evaluation of the visualization.

5.2.2 Performance Evaluation and Statistical Analysis

Fuzzy c-means clustering was applied to separate the PTX-class heatmap into an influential region and a background region. Fuzzy c-means was chosen to avoid choosing an arbitrary threshold, since the threshold greatly impacts the results. After the fuzzy c-means clustering was performed, the Dice similarity coefficient ("Dice score") between the radiologist contour and the influential region was calculated for the full radiographs and apex images. The Dice similarity coefficient quantifies the overlap using the following equation: $DSC = \frac{2|x \cap y|}{|x|+|y|}$, where x and y are the radiologist contoured area and the influential region of the heatmap, respectively. The Dice score can range between 0 and 1, with a perfect overlap corresponding to a Dice score of 1. The Dice score was used to quantify the performance on a per image basis; a histogram of the Dice score was used to characterize the performance of the population. The median and mean Dice scores were calculated, along with the standard deviation.

In order to quantify the distribution of Dice score, the skewness was calculated. Skewness is a measure of the asymmetry of the distribution around the mean; a skewness of zero means the distribution is symmetric about the mean. If the skewness is positive, the data are spread out more to the right of the mean; if the skewness is negative, the data are spread out more to the left of the mean. Skewness is calculated using the equation: $S = \frac{E(x-\mu)^3}{\sigma^3}$, where x is the Dice score, μ is the mean Dice score, σ is the standard deviation of the Dice score, and $E(x-\mu)$ is the expected value of $x-\mu$. The skewness was calculated for the distribution of Dice score for the full radiographs, as well as for the apex images. In this case, the skewness should be negative, with the data spread out to the left of the mean, meaning that the mean Dice score is closer to 1 (perfect overlap) than if the skewness is positive.

5.3 AlexNet - Grad-CAM Specific Methods and Results for the Full Radiographs and Apex Images

5.3.1 AlexNet - Grad-CAM Specific Methods

For AlexNet, Grad-CAM was implemented using PyTorch (version 1.1.0) with Python (version 3.6.8). A pre-trained AlexNet architecture network was loaded from the PyTorch model library. The final classification layer was changed to a 2-class output, as was performed for fine-tuning. Then the model weights saved from fine-tuning using the full radiograph were loaded into the model, and Grad-CAM was performed to output two heatmaps for each image, one for the PTX class and the no PTX class. After the heatmaps were generated for the full radiographs, the weights saved from fine-tuning with the apex images were loaded, and the heatmap generation process was repeated for the apex images. To quantify the performance of Grad-CAM, 2-category fuzzy c-means clustering was applied to separate the PTX-class heatmap into an influential region and a background region. Then the performance was evaluated with Dice score and skewness.

5.3.2 AlexNet - Grad-CAM Visualization Results for the Full Radiographs

As presented in Chapter 4, the AlexNet CNN fine-tuned on the full radiographs yielded an AUC of 0.702 (95% CI: 0.657, 0.745) in the task of distinguishing between radiographs with and without PTX. The detection sensitivity was 67% and the specificity was 64%. Figure 5.1 shows PTX-class heatmaps for test images correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.2 shows examples of true positives, false negatives, true negatives, and false positives.

When the radiologist truth was compared to the influential regions of the PTX-class heatmaps determined by 2-category fuzzy c-means, the average Dice overlap score was calculated to be 0.03 with a standard deviation of 0.06. The median Dice score was 0; 186 test set PTX images had Dice scores less than 0.01 and 126 of those had a Dice score of 0.

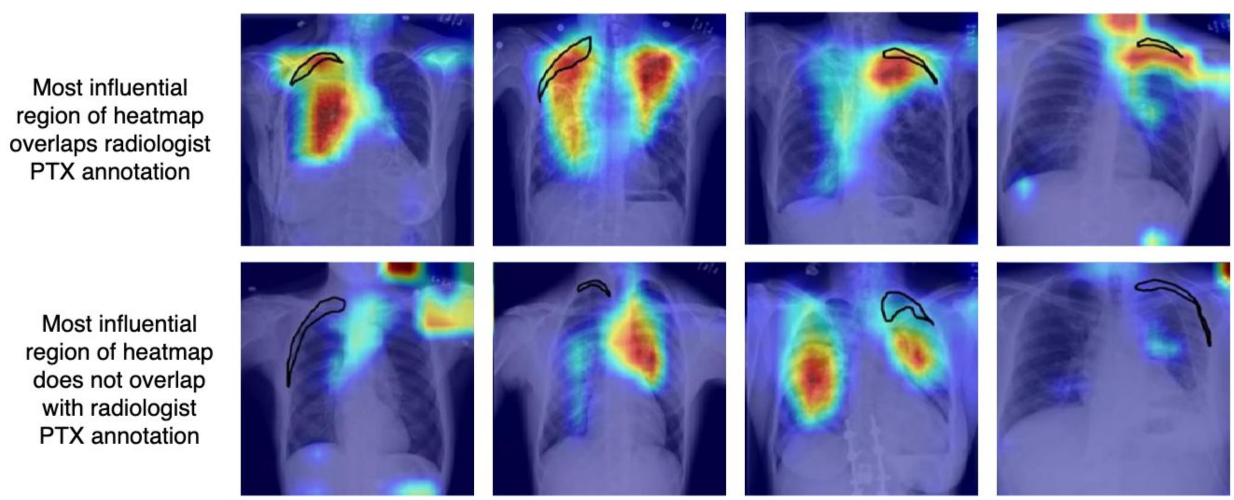


Figure 5.1: Examples of **PTX-class** Grad-CAM heatmaps for true PTX cases correctly classified by the fine-tuned AlexNet network (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the influential regions identified in the heatmap (orange to dark red) overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

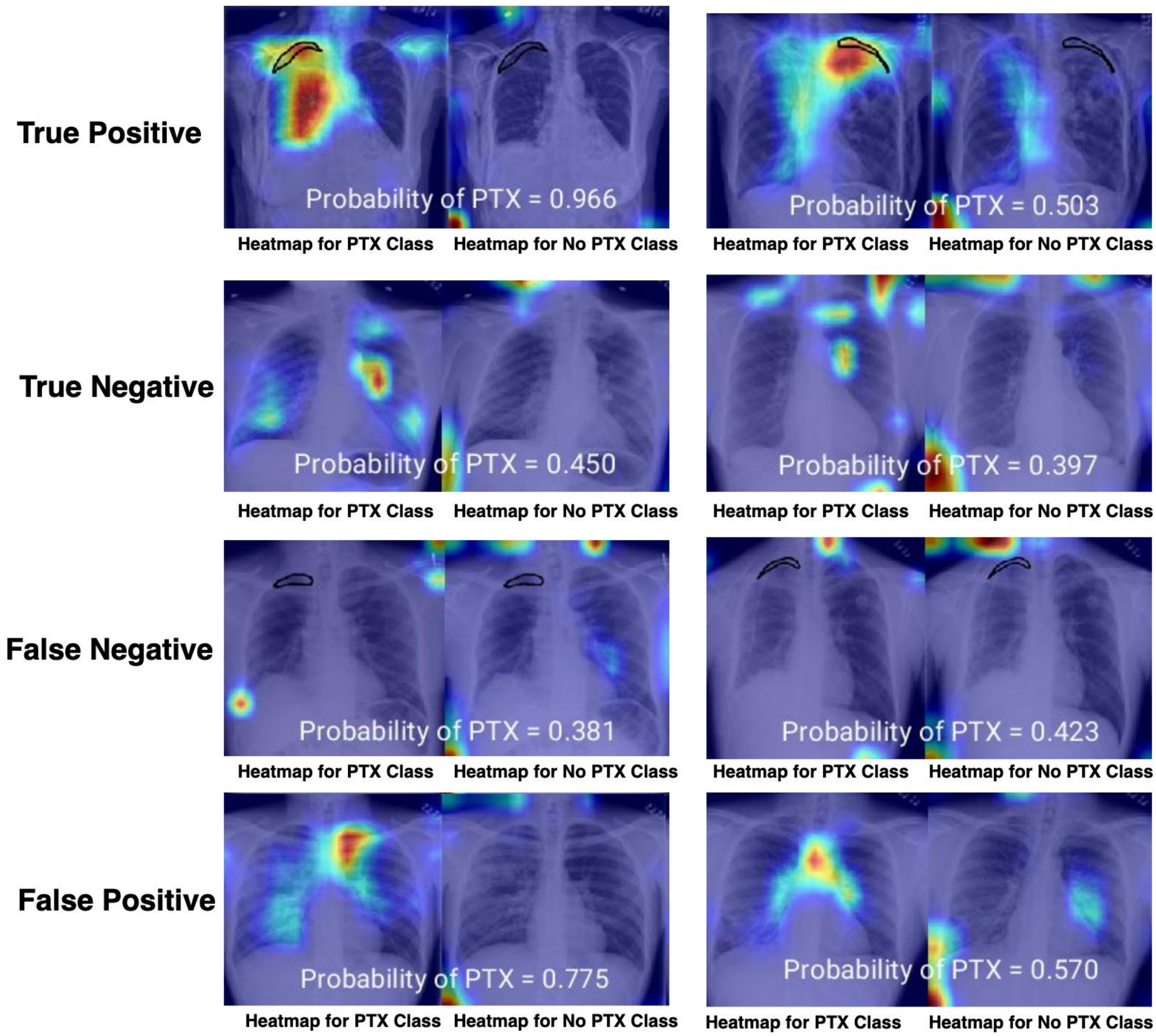


Figure 5.2: Heatmaps for true-positive, true-negative, false-negative, and false-positive cases from the AlexNet CNN fine-tuned with the full radiographs. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap for the PTX class and a heatmap for the No PTX class.

5.3.3 AlexNet - Grad-CAM Visualization Results for the Apex Images

For the test set of 225 PTX images and 350 images without PTX, ROC analysis yielded an AUC of 0.862 (95% CI: 0.831, 0.890) in the task of distinguishing between images with and without PTX. The detection sensitivity was 56% and the specificity was 89%. Figure 5.3 shows examples of cases that were correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.4 shows examples of true positives, true negatives, false negatives and false positives, and their heatmaps corresponding to each of the two classes the network is able to classify.

When the radiologist truth was compared to the influential regions of the PTX-class heatmaps determined by 2-category fuzzy c-means, the average Dice overlap score was calculated to be 0.22 with a standard deviation of 0.19. The median Dice score was 0.19; 64 test set PTX images had Dice scores less than 0.01 and 39 of those had a Dice score of 0.

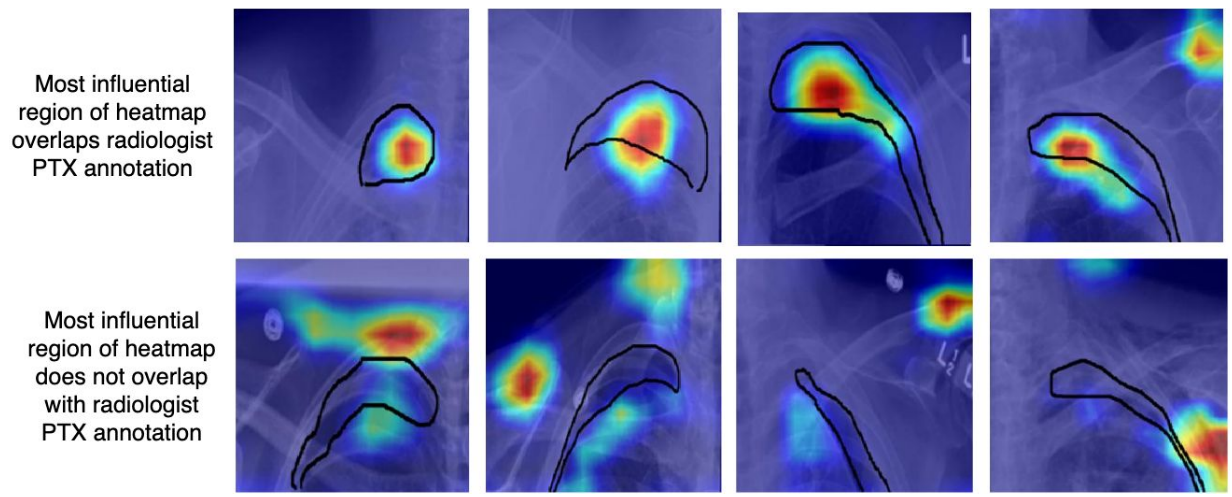


Figure 5.3: Examples of **PTX-class** Grad-CAM heatmaps for true PTX cases correctly classified by the AlexNet network fine-tuned with apex images (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the influential regions identified in the Grad-CAM visualization overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

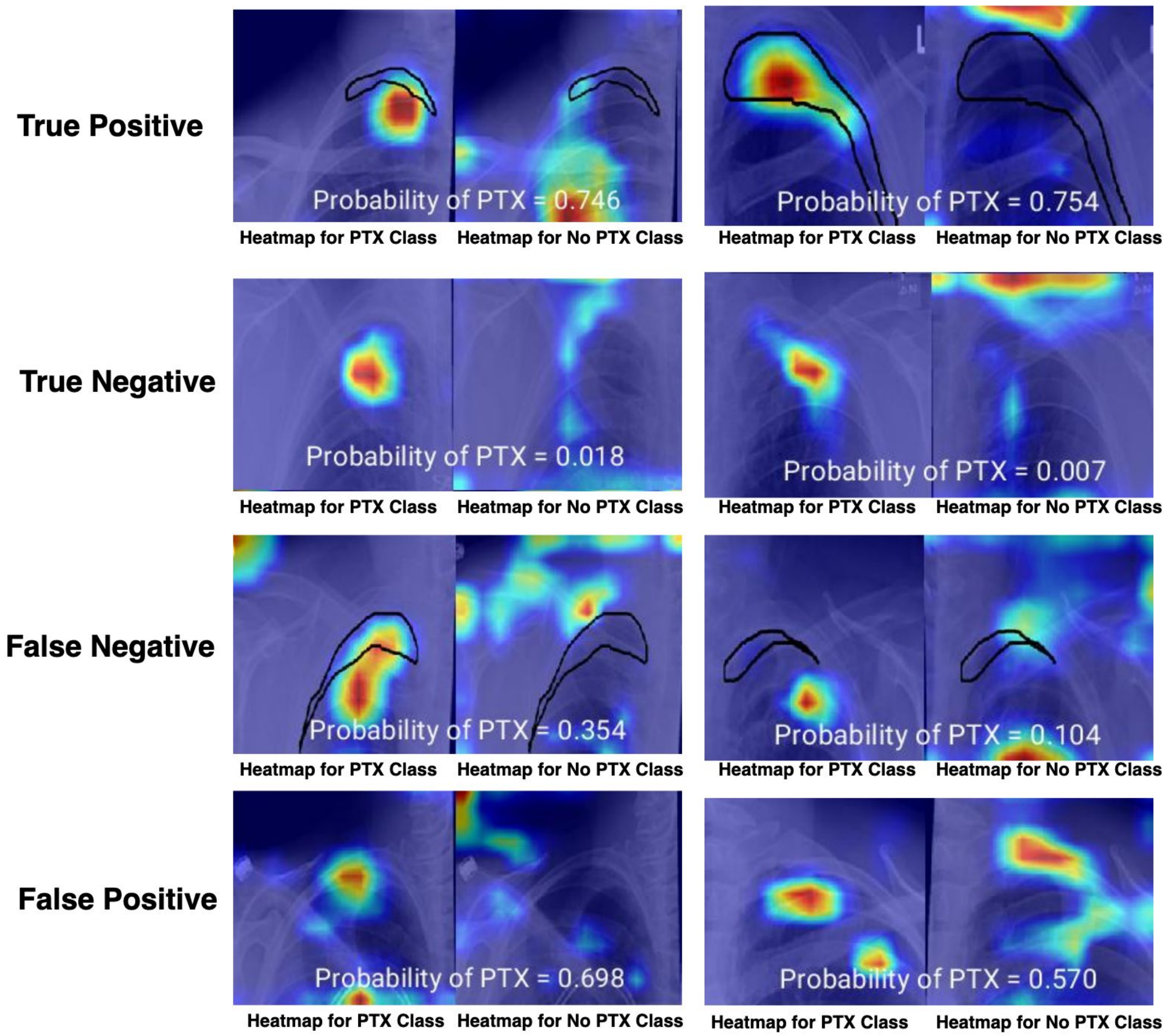


Figure 5.4: Apex heatmaps from the fine-tuned AlexNet CNN for true-positive, true-negative, false-negative, and false-positive cases. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap showing the most influential regions for the PTX class and a heatmap showing the most influential regions for the No PTX class.

5.3.4 Comparison between AlexNet - Grad-CAM visualization results for the full radiographs and apex images

Figure 5.5 is a histogram of the Dice score for the overlap between the influential region of the Grad-CAM PTX-class heatmap and the radiologist annotated location of the PTX. The distributions of Dice score for the full radiographs and the apex images are shown with the median, average, and standard deviation of Dice score, along with the skewness.

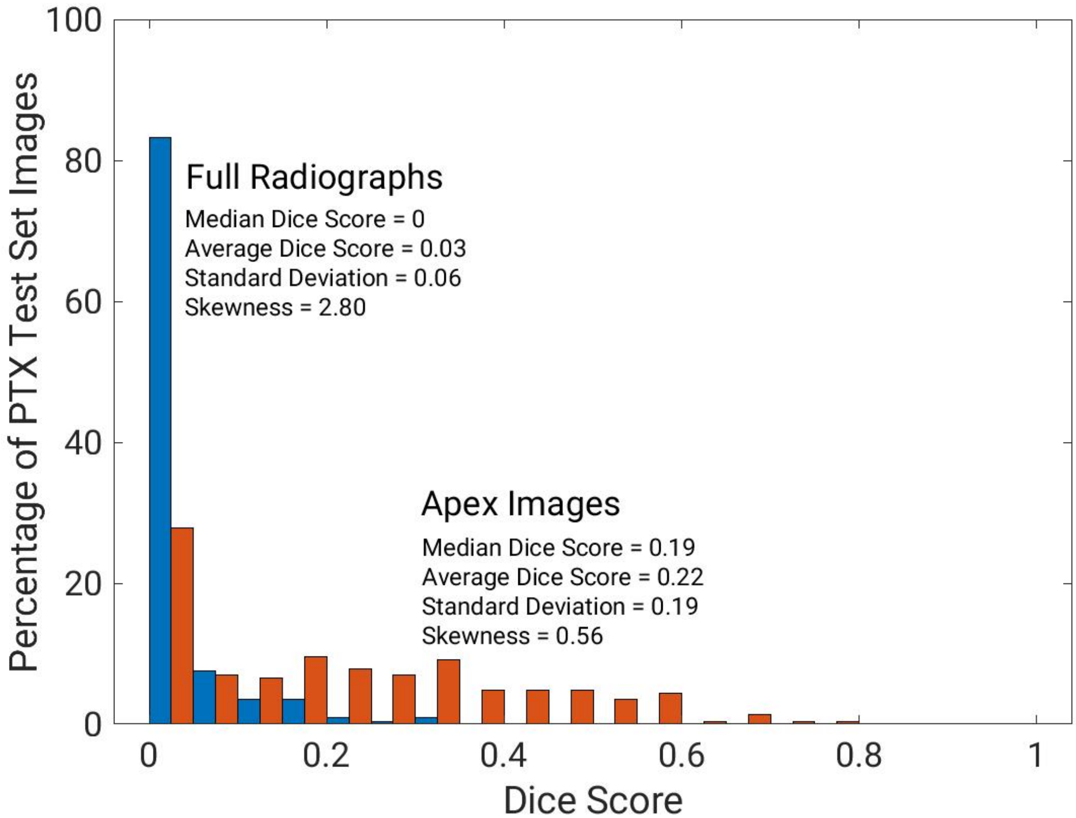


Figure 5.5: Histogram of Dice score for the full radiographs and the apex images, calculated using the radiologist-annotated location of PTX and the region of influence (from 2-category fuzzy c-means) of the PTX-class heatmap generated by Grad-CAM on the fine-tuned AlexNet CNN.

Since the Dice score was calculated using the PTX-class heatmap, another metric of interest is the probability of PTX along with its corresponding Dice score. Figure 5.6 is a scatter plot showing the probability of PTX for each test image, as output by the AlexNet CNNs fine-tuned with the full radiographs and apex images, as well as its Dice score when quantifying overlap between its PTX-class heatmap and the radiologist annotation. Ideally, points should be in the top right corner, corresponding to a high Dice score and high probability of PTX.

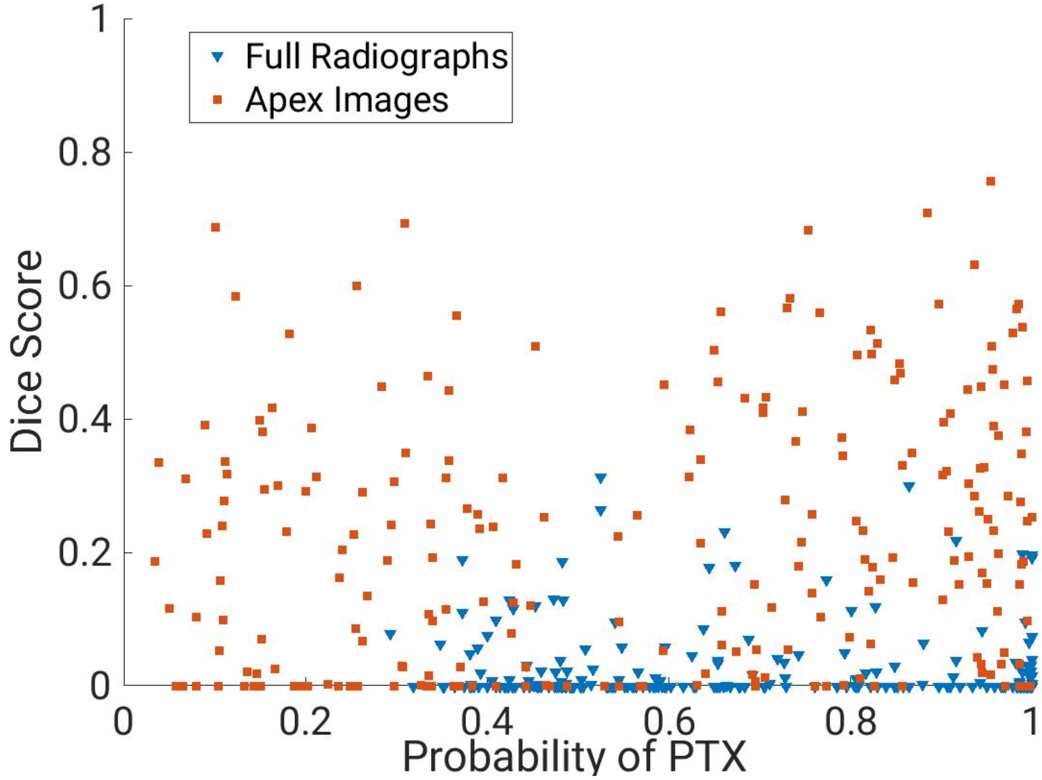


Figure 5.6: Scatter plot of Dice score and probability for the full radiographs and the apex images with PTX for the fine-tuned AlexNet - Grad-CAM visualizations of the activations for the PTX class. Ideally, points should be in the upper right corner, corresponding to a high probability of PTX, as well as a high Dice score.

5.3.5 AlexNet - Grad-CAM Discussion and Conclusions

The Dice score quantifying the overlap of the influential region of the heatmap with the radiologist-annotated location of PTX was higher on average for the AlexNet CNN fine-tuned with the apex images than the CNN fine-tuned with the full radiographs. In addition, the distribution of Dice score for the full radiographs is quite skewed to the right of the mean (skewness=2.80) compared to the distribution of Dice score for the apex images (skewness=0.56). In the scatter plot of Dice score and probability of PTX (Figure 5.6), neither of the fine-tuned AlexNet CNNs have points clustered in the upper right corner, corresponding to a high Dice score and high probability of PTX.

5.4 VGG19 - Grad-CAM Specific Methods and Results for the Full Radiographs and Apex Images

5.4.1 VGG19 - Grad-CAM Specific Methods

For VGG19, Grad-CAM was implemented using Keras (version 2.2.2) with Python (version 3.5.5). The model architecture (modified for a 2-class classification) and weights saved from fine-tuning with the full radiographs were loaded. Grad-CAM was then performed to generate the heatmaps for each class. The process was repeated, loading the architecture and weights from fine-tuning with the apex images, then performing Grad-CAM for heatmap generation. To quantify the heatmaps and their overlap with the radiologist annotation of the PTX, 2-category fuzzy c-means clustering was applied to separate the heatmap into an influential region and a background region. Then the performance was evaluated with Dice score and skewness.

5.4.2 VGG19 - Grad-CAM Visualization Results for the Full Radiographs

As presented in Chapter 4, the VGG19 network fine-tuned on the full radiographs yielded an AUC of 0.792 (95% CI: 0.753, 0.827) in the task of distinguishing between radiographs with and without PTX. The detection sensitivity was 78% and the specificity was 69%. Figure 5.7 shows examples of cases that were correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.8 shows examples of true positives, false negatives, true negatives, and false positives.

When the radiologist truth was compared to the influential regions of the Grad-CAM heatmaps determined by 2-category fuzzy c-means, the average Dice overlap score was calculated to be 0.04 with a standard deviation of 0.07. The median Dice score was 0.01; 111 test set PTX images had Dice scores less than 0.01 and 70 of those had a Dice score of 0.

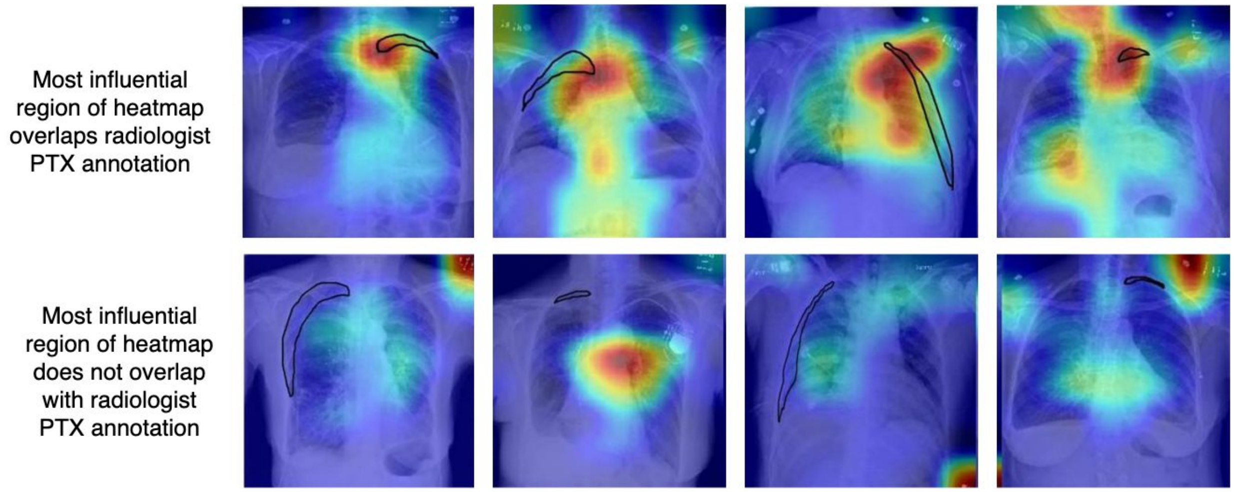


Figure 5.7: Examples of Grad-CAM **PTX-class** heatmaps for true PTX cases correctly classified by the fine-tuned VGG19 network (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the influential regions identified in the heatmap (orange to dark red) overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

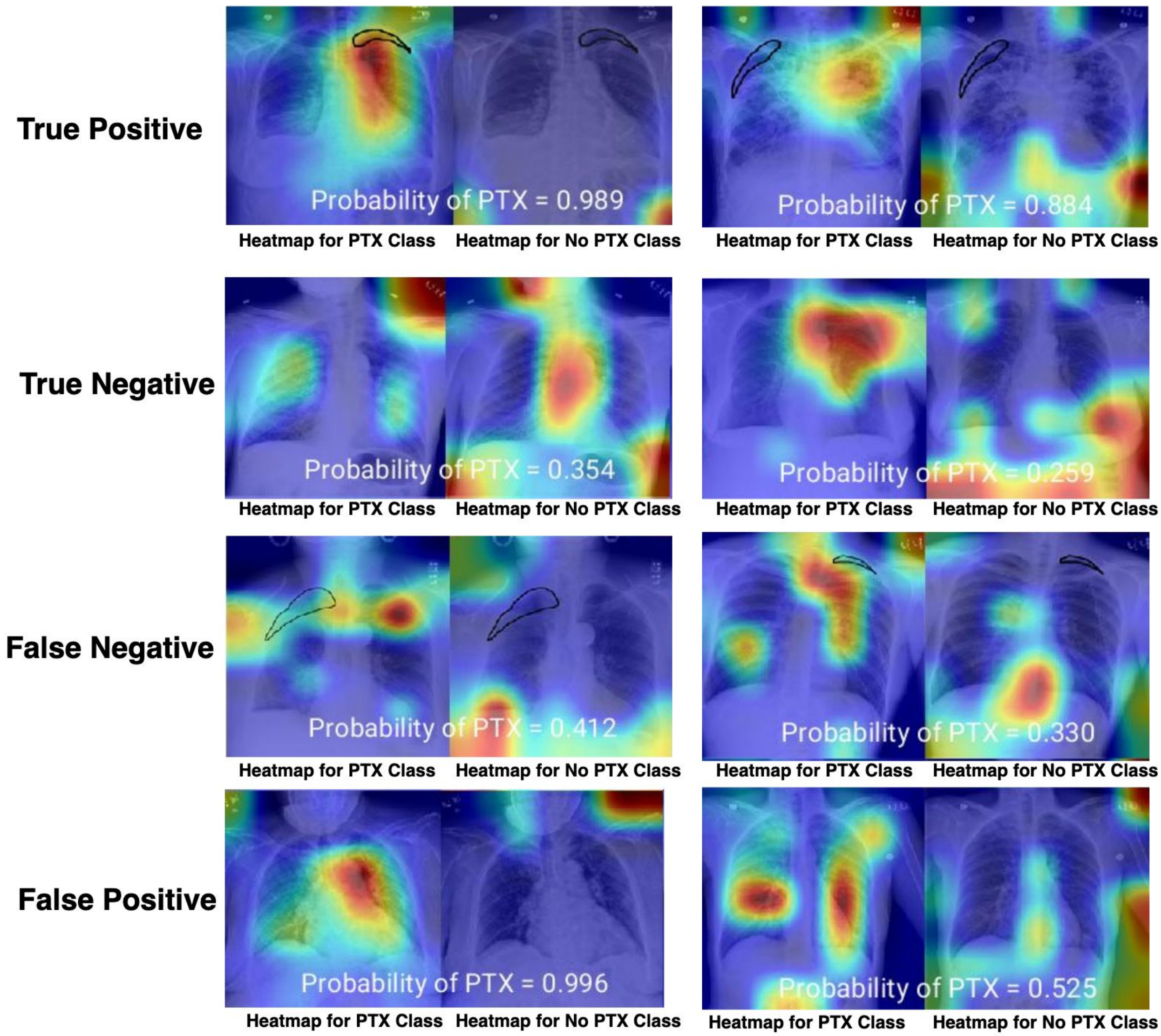


Figure 5.8: Heatmaps for true-positive, true-negative, false-negative, and false-positive cases from the fine-tuned VGG19 network. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap for the PTX class and a heatmap for the No PTX class.

5.4.3 VGG19 - Grad-CAM Visualization Results for the Apex Images

For the test set of 225 PTX images and 350 images without PTX, ROC analysis yielded an AUC of 0.898 (95% CI: 0.871, 0.921) in the task of distinguishing between images with and without PTX. The detection sensitivity was 62% and the specificity was 89%. Figure 5.9 shows example heatmaps of test set PTX images correctly classified (true positives) by the fine-tuned VGG19 CNN and the radiologist annotation of the PTX. Figure 5.10 shows examples of true positives, true negatives, false negatives and false positives.

For the apex images, the median Dice score was 0.36 and the average Dice score was 0.35 with a standard deviation of 0.22. Of the test set PTX images, 24 had Dice scores less than 0.01 and 17 of those had a Dice score of 0.

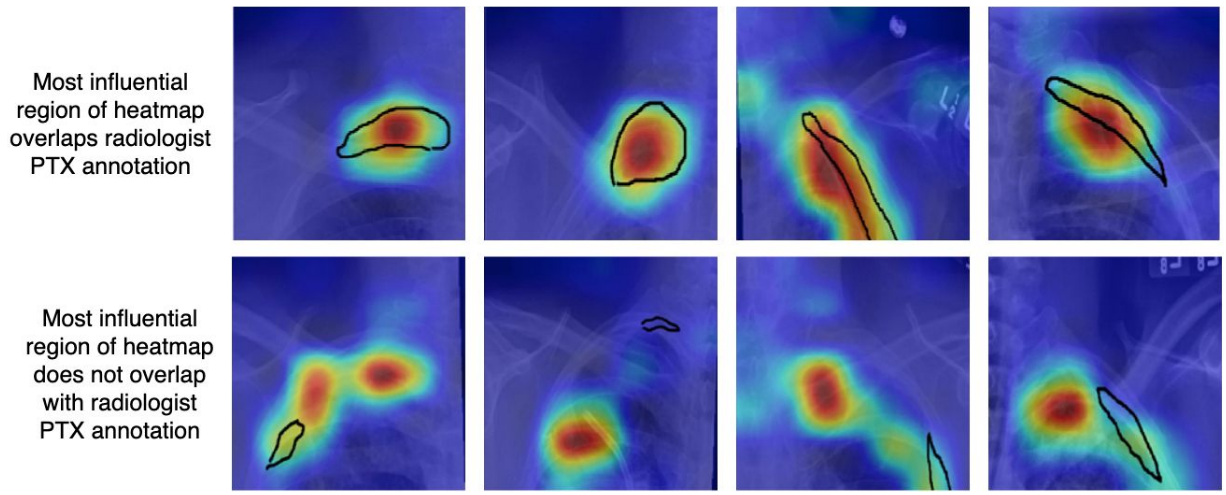


Figure 5.9: Examples of Grad-CAM **PTX-class** heatmaps for true PTX cases correctly classified by the VGG19 CNN fine-tuned with apex images (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the influential regions identified in the Grad-CAM visualization overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

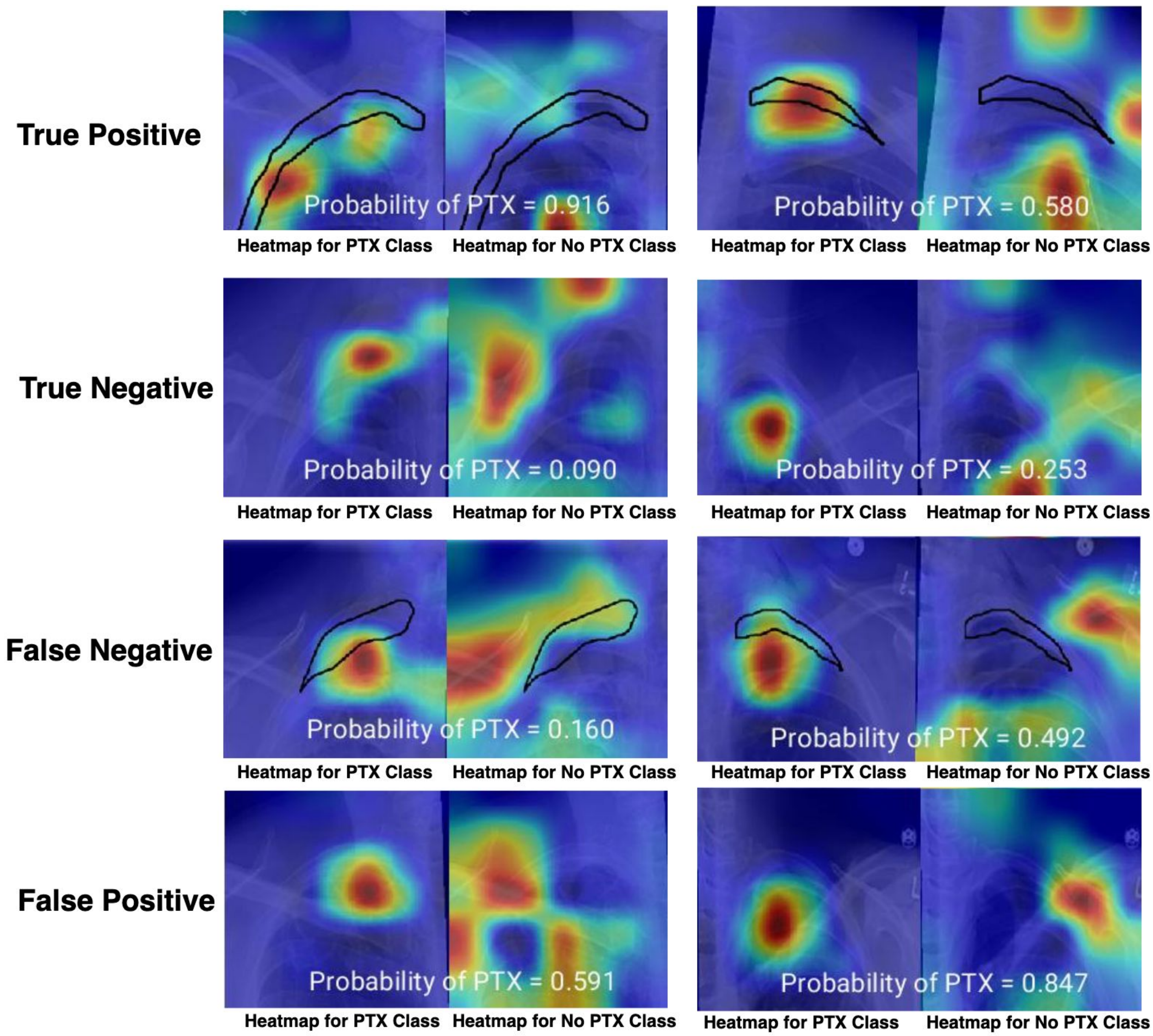


Figure 5.10: Apex heatmaps for true-positive, true-negative, false-negative, and false-positive cases from the fine-tuned VGG19 CNN. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap showing the most influential regions for the PTX class and a heatmap showing the most influential regions for the No PTX class.

5.4.4 Comparison between VGG19 - Grad-CAM visualization results for the full radiographs and apex images

Figure 5.11 is a histogram of the Dice score for the overlap between the influential region of the Grad-CAM PTX-class heatmap and the radiologist annotated location of the PTX. The distributions of Dice score for the full radiographs and the apex images are shown with the median, average, and standard deviation of Dice score, along with the skewness.

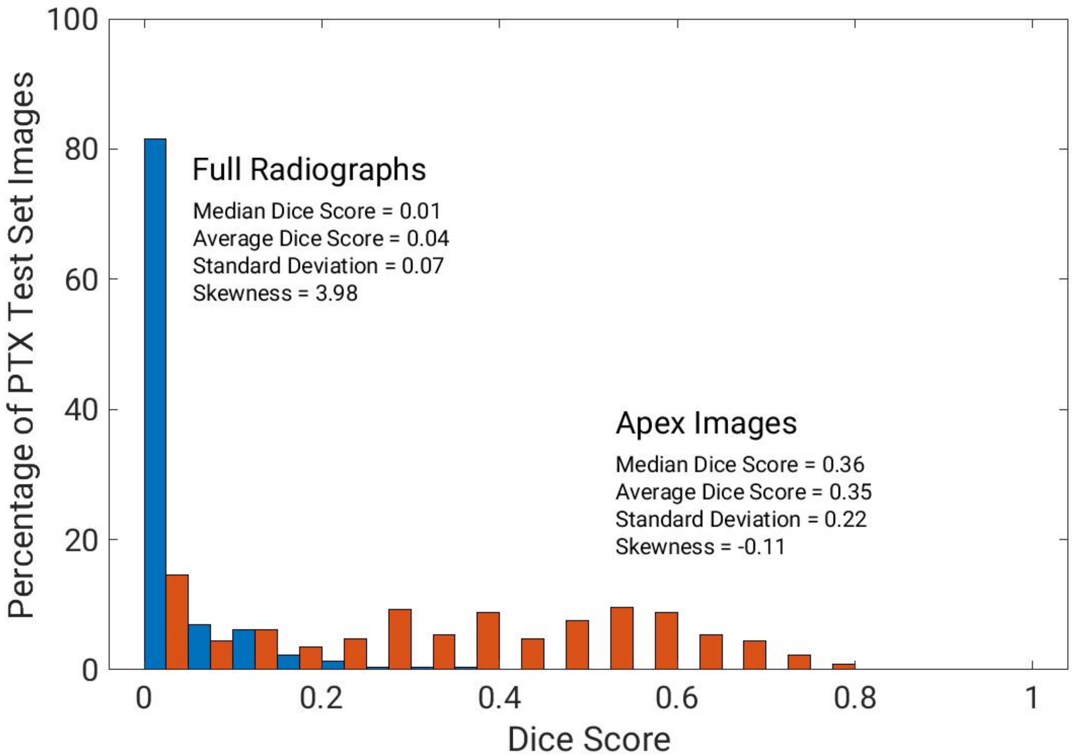


Figure 5.11: Histogram of Dice score from VGG19-GradCAM heatmaps for the full radiographs and the apex images, calculated using the radiologist-annotated location of PTX and the 2-category fuzzy c-means-determined region of influence of the heatmap for the PTX class.

Since the Dice score was calculated using the PTX-class heatmap, another metric of interest is the probability of PTX along with its corresponding Dice score. Figure 5.12 is a scatter plot showing the probability of PTX for each test image, as output by the VGG19 CNNs fine-tuned with the full radiographs and apex images, as well as its Dice score when quantifying overlap between its PTX-class heatmap and the radiologist annotation. Ideally, points should be in the top right corner, corresponding to a high Dice score and high probability of PTX.

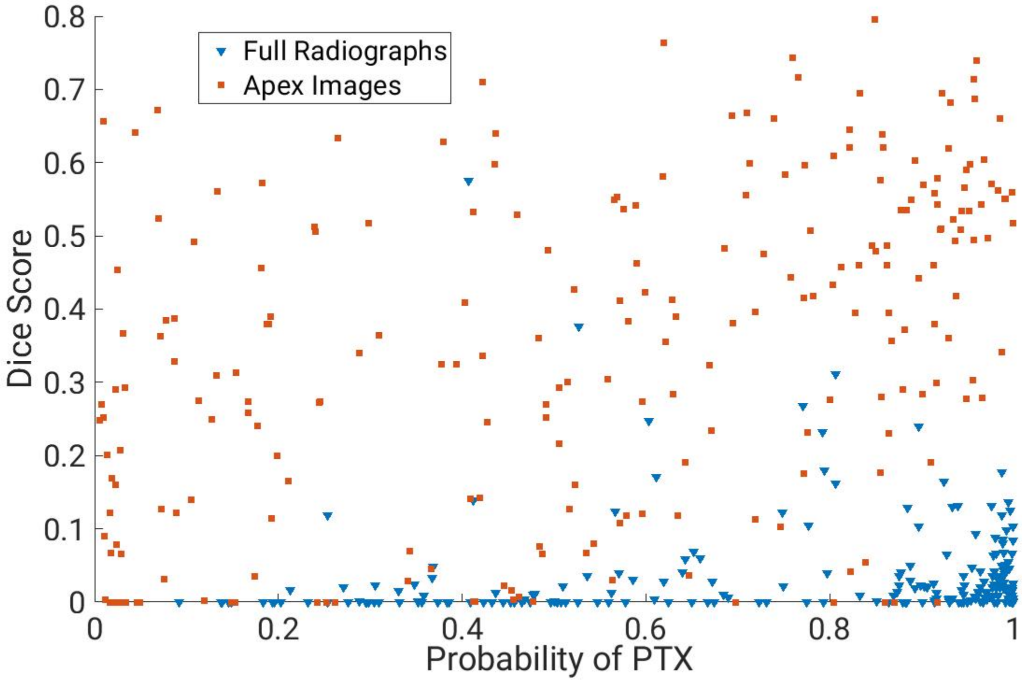


Figure 5.12: Scatter plot of Dice score and probabilities from VGG19 for the full radiographs and the apex images with PTX. Ideally, points should be in the upper right corner, corresponding to a high probability of PTX, as well as a high Dice score.

5.4.5 VGG19 - Grad-CAM Discussion and Conclusions

In addition to a high classification performance, the Grad-CAM heatmaps from the VGG19 CNN fine-tuned with apex images yielded an improved localization ability compared to the VGG19 CNN fine-tuned with full radiographs. The average Dice score is higher for the apex heatmaps than the full radiograph heatmaps. The distribution of Dice score for the full radiographs was highly skewed to the right of the mean (skewness=3.98); the distribution of Dice score for the apex images was negative, therefore spread to the left of the mean (skewness=-0.11). The scatter plot of Dice score and probability of PTX for each test set image (Figure 5.12) shows that the VGG19 CNN fine-tuned with the apex images has a higher concentration of points near the upper right corner compared to the CNN fine-tuned with the full radiographs.

5.5 ResNet50 - Grad-CAM Specific Methods and Results for the Full Radiographs and Apex Images

5.5.1 ResNet50 - Grad-CAM Specific Methods

To generate heatmaps with the fine-tuned ResNet50 CNN, Grad-CAM was implemented using Keras (version 2.2.2) with Python (version 3.5.5). The CNN architecture and weights saved from fine-tuning with the full radiographs were loaded; then Grad-CAM was applied to generate the heatmaps for the PTX and No PTX classes. After those heatmaps were generated for the test set images, the architecture and weights saved from fine-tuning with the apex images were loaded. Then the process was repeated to generate a Grad-CAM heatmap for each of the two classes on all the test images. 2-category fuzzy c-means clustering was applied to separate the PTX-class heatmap into an influential region and a background region in order to quantify the heatmap overlap with the radiologist annotation of the PTX. Then the performance was evaluated with Dice score and skewness.

5.5.2 ResNet50 - Grad-CAM Visualization Results for the Full Radiographs

As presented in Chapter 4, the ResNet50 CNN fine-tuned on the full radiographs yielded an AUC of 0.669 (95% CI: 0.622, 0.714) in the task of distinguishing between radiographs with and without PTX. The detection sensitivity was 99% and the specificity was 2%. Figure 5.13 shows PTX-class heatmaps for test images correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.14 shows examples of true positives, false negatives, true negatives, and false positives.

When the radiologist truth was compared to the influential regions of the Grad-CAM heatmaps determined by 2-category fuzzy c-means, the average Dice overlap score was calculated to be 0.03 with a standard deviation of 0.05. The median Dice score was 0; 141 test set PTX images had Dice scores less than 0.01 and 100 of those had a Dice score of 0.

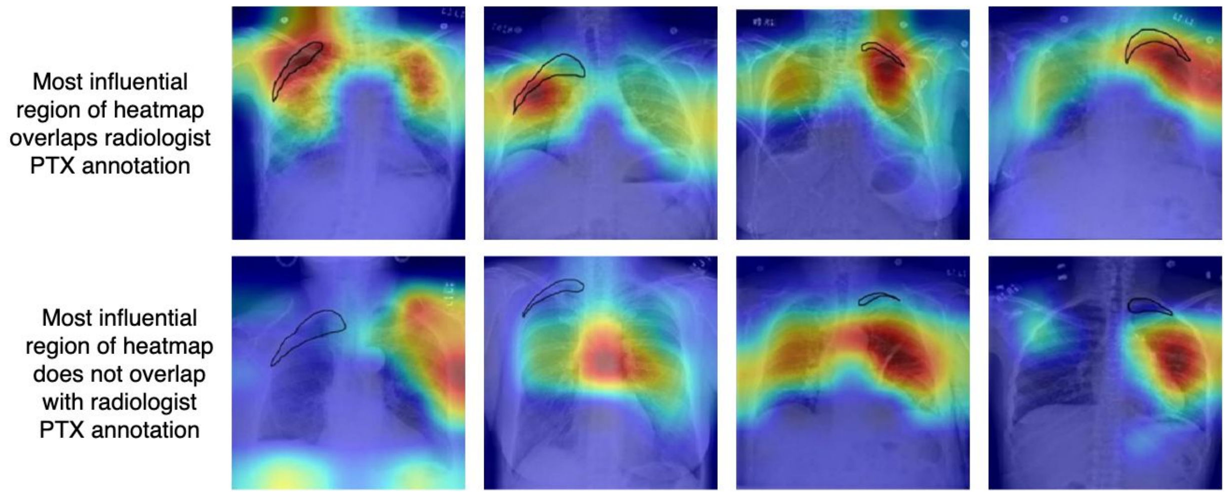


Figure 5.13: Examples of Grad-CAM **PTX-class** heatmaps for true PTX cases correctly classified by the fine-tuned ResNet50 network (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the influential regions identified in the heatmap (orange to dark red) overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

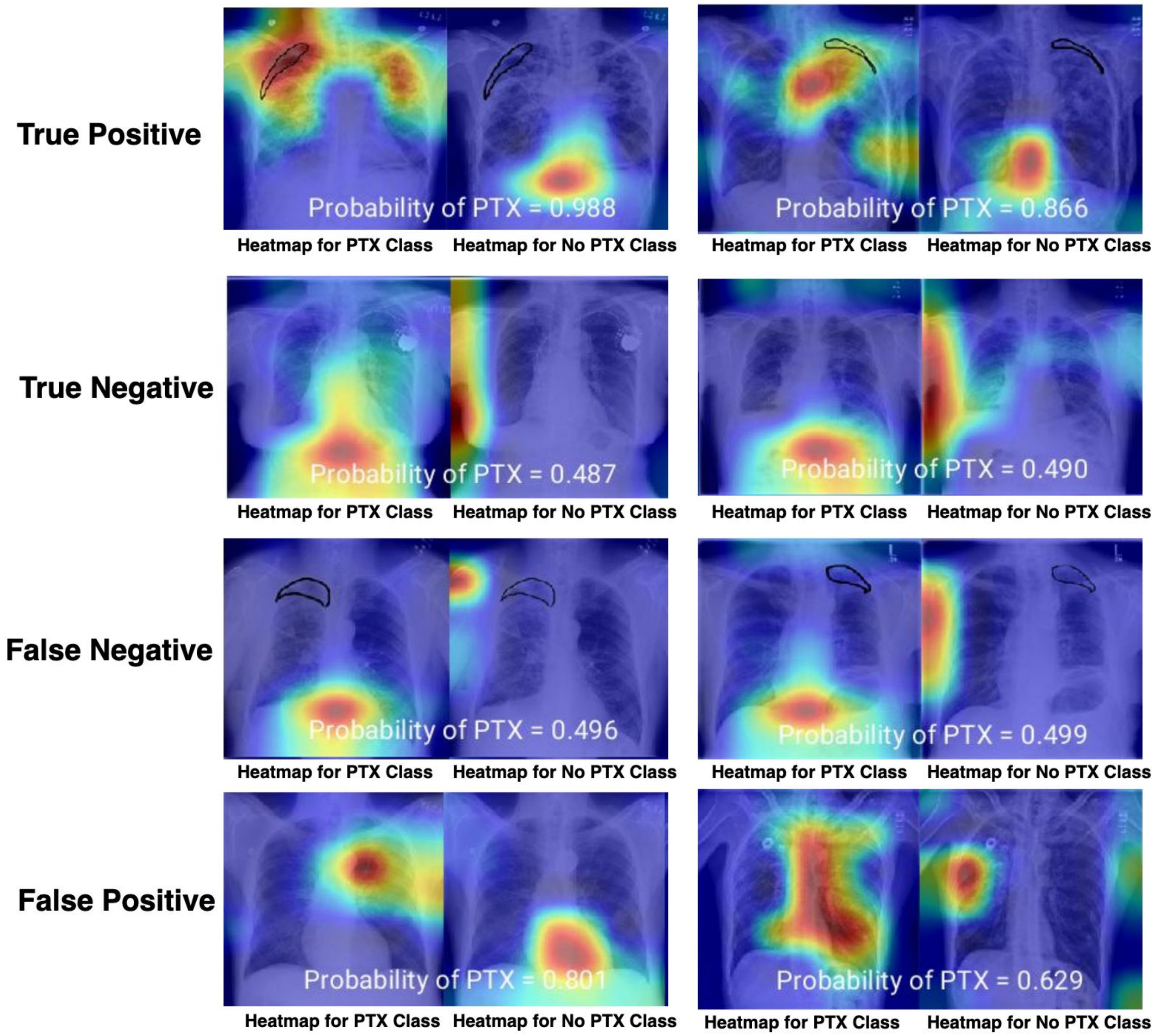


Figure 5.14: Heatmaps from ResNet50 for true-positive, true-negative, false-negative, and false-positive cases. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap for the PTX class and a heatmap for the No PTX class.

5.5.3 ResNet50 - Grad-CAM Visualization Results for the Apex Images

For the test set of 225 PTX images and 350 images without PTX, ROC analysis yielded an AUC of 0.765 (95% CI: 0.717, 0.807) in the task of distinguishing between images with and without PTX. The detection sensitivity was 0% and the specificity was 100%. Figure 5.15 shows example heatmaps for two true-negative cases and two false-negative cases.

When the radiologist truth was compared to the influential regions of the Grad-CAM heatmaps determined by 2-category fuzzy c-means, the average Dice overlap score was calculated to be 0.01 with a standard deviation of 0.06. The median Dice score was 0; 212 test set PTX images had Dice scores less than 0.01 and 209 of those had a Dice score of 0. None of the test images with PTX were correctly classified (probability of PTX >0.5) by the ResNet50 network fine-tuned with the apex images. All of the test set images had a probability of PTX less than 0.5.

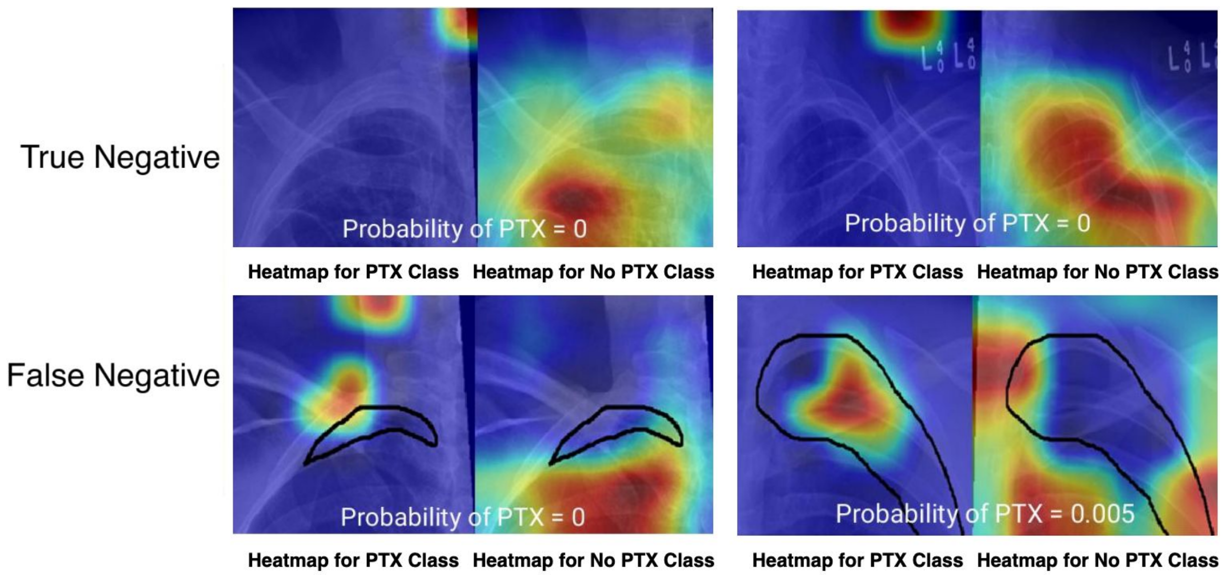


Figure 5.15: Example **PTX-class** apex heatmaps for true negatives and false negatives from the fine-tuned ResNet50 CNN. The false-negative cases have the true location of the PTX annotated as a black contour line.

5.5.4 Comparison between ResNet50 - Grad-CAM visualization results for the full radiographs and apex images

Figure 5.16 is a histogram of the Dice score for the overlap between the influential region of the Grad-CAM PTX-class heatmap and the radiologist annotated location of the PTX. The distributions of Dice score for the full radiographs and the apex images are shown with the median, average, and standard deviation of Dice score, along with the skewness.

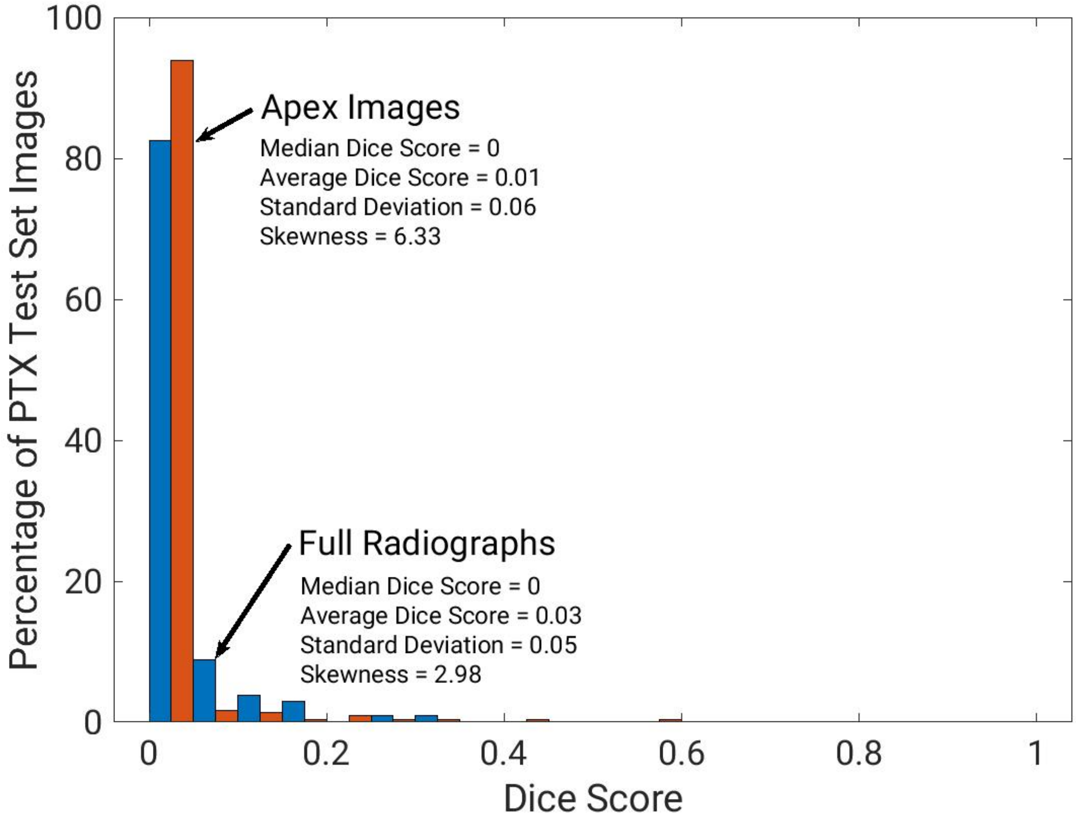


Figure 5.16: Histogram of Dice score from ResNet50-GradCam heatmaps for the full radiographs and the apex images, calculated using the radiologist-annotated location of PTX and the 2-category fuzzy c-means-determined region of influence of the PTX-class heatmap.

Figure 5.17 is a scatter plot showing the probability of PTX for each test image, as output by the ResNet50 CNNs fine-tuned with the full radiographs and apex images, as well as its Dice score when quantifying overlap between its PTX class heatmap and the radiologist annotation. Ideally, points should be in the top right corner, corresponding to a high Dice score and high probability of PTX.

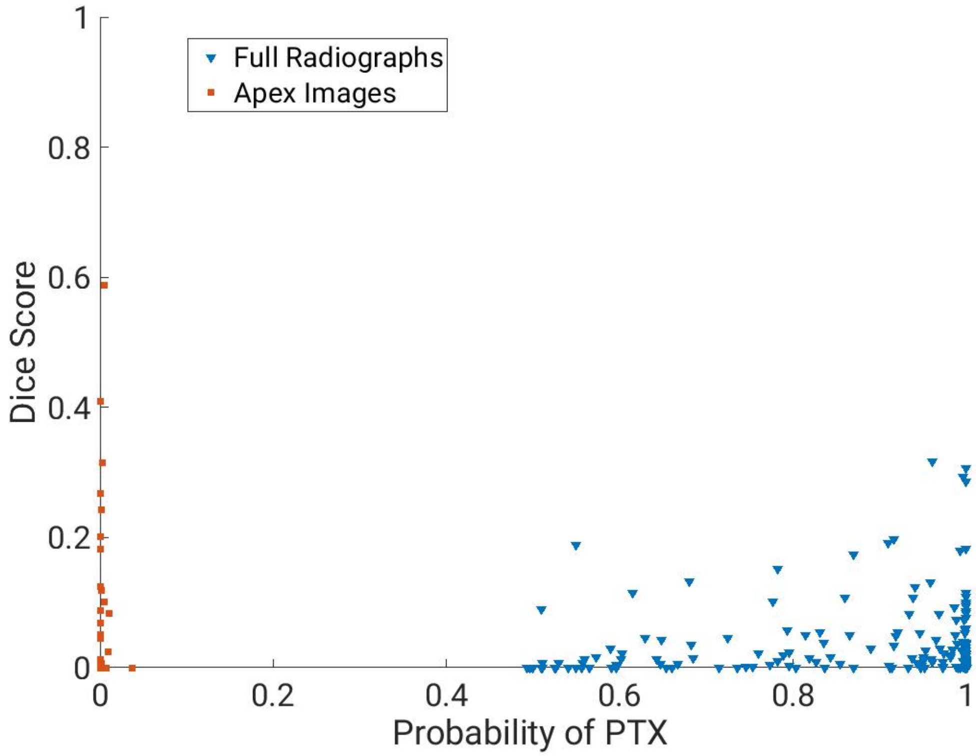


Figure 5.17: Scatter plots of Dice score and probabilities from ResNet50 for the full radiographs and the apex images with PTX. Ideally, points should be in the upper right corner, corresponding to a high probability of PTX, as well as a high Dice score. Ideally, points should be in the upper right corner, corresponding to a high probability of PTX, as well as a high Dice score.

5.5.5 ResNet50 - Grad-CAM Discussion and Conclusions

For both of the fine-tuned ResNet50 networks, the output probability distributions were unlike any of the other CNNs fine-tuned. As discussed in Chapter 4, with the full radiographs, the fine-tuned CNN output a probability of PTX around 1 for each test image, with a range from 0.5 to 1; the CNN fine-tuned with the apex images output a probability of PTX around 0 (Figure 5.17). This unusual performance is reflected in the histogram of the Dice score. Both the full radiographs and the apex images have highly skewed distributions of Dice score with low median and average Dice scores.

5.6 BagNet Visualization Specific Methods and Results for the Full Radiographs and Apex Images

5.6.1 BagNet Visualization Specific Methods

Since BagNet's heatmap generation is a component of the output from the trained CNN, the visualizations were saved following the fine-tuning. After fine-tuning was complete and the trained model was being applied to the test set to evaluate the performance, heatmaps for each class were output along with each test image's probability of PTX. To perform the fine-tuning and generation of heatmaps, PyTorch (version 1.1.0) was used within Python (version 3.6.8).

To quantify the performance of the BagNet heatmap visualization technique and its overlap with the radiologist annotation, 3-category fuzzy c-means clustering was applied to separate the PTX-class heatmap into a highly influential region, an influential region, and a background region. The Grad-CAM heatmaps had 2-category fuzzy c-means applied; 3-category fuzzy c-means was applied to the BagNet heatmaps since the BagNet heatmaps do not indicate a general region of influence as Grad-CAM does and as a result there are patches of varying influence throughout the image. When 2-category fuzzy c-means was applied to the BagNet heatmaps, the Dice score did not truly represent the overlap since the "influential region" was scattered throughout the image, even if the region of highest influence was at the PTX location. Changing the fuzzy c-means clustering to 3 categories improved the identification of the highly influential region and the Dice score was more indicative of the actual overlap.

5.6.2 BagNet Visualization Results for the Full Radiographs

As presented in Chapter 4, the BagNet CNN fine-tuned on the full radiographs yielded an AUC of 0.784 (95% CI: 0.743, 0.821) in the task of distinguishing between radiographs with and without PTX. The detection sensitivity was 87% and the specificity was 44%. Figure

5.18 shows PTX-class heatmaps for test images correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.19 shows examples of true positives, false negatives, true negatives, and false positives.

When the radiologist truth was compared to the highly influential regions of the BagNet heatmaps determined by 3-category fuzzy c-means, the average Dice overlap score was calculated to be 0.03 with a standard deviation of 0.04. The median Dice score was 0.02; 85 test set PTX images had Dice scores less than 0.01 and 7 of those had a Dice score of 0.

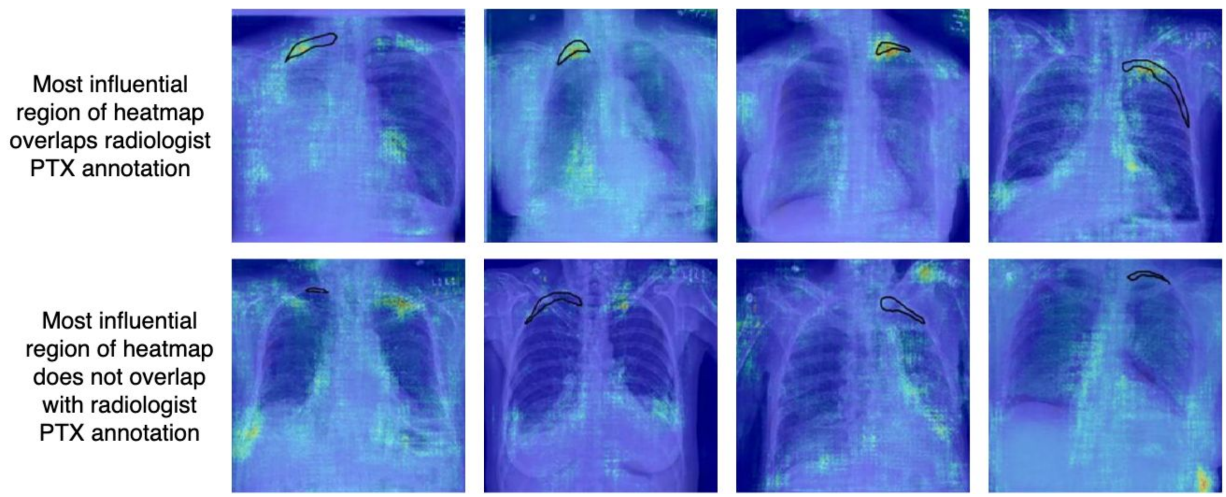


Figure 5.18: Examples of BagNet **PTX-class** heatmaps for true PTX cases correctly classified (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the highly influential regions identified in the heatmap (orange to dark red) overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

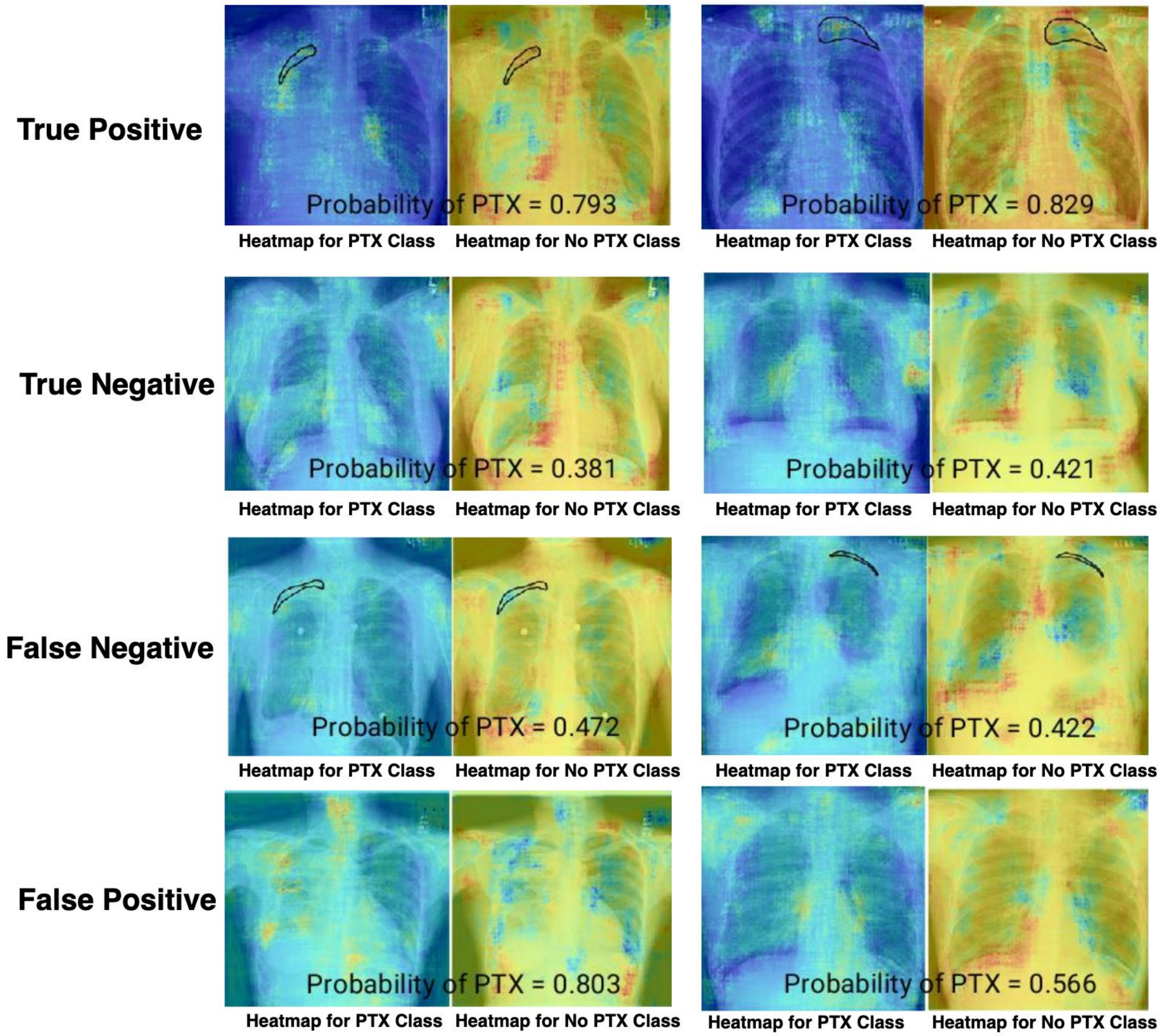


Figure 5.19: Heatmaps for true-positive, true-negative, false-negative, and false-positive cases from the BagNet CNN. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. Each image has a heatmap for the PTX class and a heatmap for the No PTX class. The heatmaps for the No PTX class are opposite to the PTX class heatmap, the most influential region of the PTX class heatmap is the least influential on the No PTX class heatmap. This is due to the classification being binary. If there were more classes, the heatmaps would not be opposites of one another.

5.6.3 BagNet Visualization Results for the Apex Images

For the test set of 225 PTX images and 350 images without PTX, ROC analysis yielded an AUC of 0.922 (95% CI: 0.898, 0.942) in the task of distinguishing between images with and without PTX. The detection sensitivity was 62% and the specificity was 96%. Figure 5.20 shows examples of cases that were correctly identified as having PTX (true positives) with the radiologist annotation of PTX location. Figure 5.21 shows examples of true positives, true negatives, false negatives and false positives, and their heatmaps corresponding to each of the two classes.

When the radiologist truth was compared to the highly influential regions of the BagNet heatmaps determined by 3-category fuzzy c-means, the average Dice overlap score was calculated to be 0.14 with a standard deviation of 0.10. The median Dice score was 0.12; 18 test set PTX images had Dice scores less than 0.01 and 12 of those had a Dice score of 0.

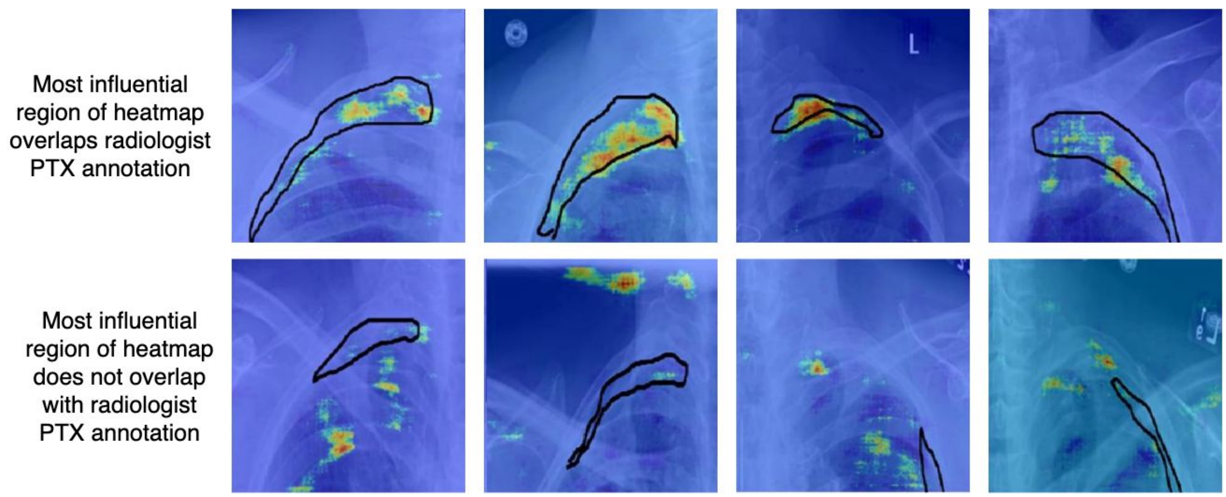


Figure 5.20: Examples of **PTX-class** heatmaps for true PTX cases correctly classified by the BagNet network fine-tuned with apex images (true positives). The annotated location of PTX is denoted by the black outline. The top row shows cases in which the highly influential regions identified in the heatmap overlap, at least partially, with the PTX as annotated by a radiologist. The bottom row shows cases for which highly influential regions did not overlap with the location of PTX.

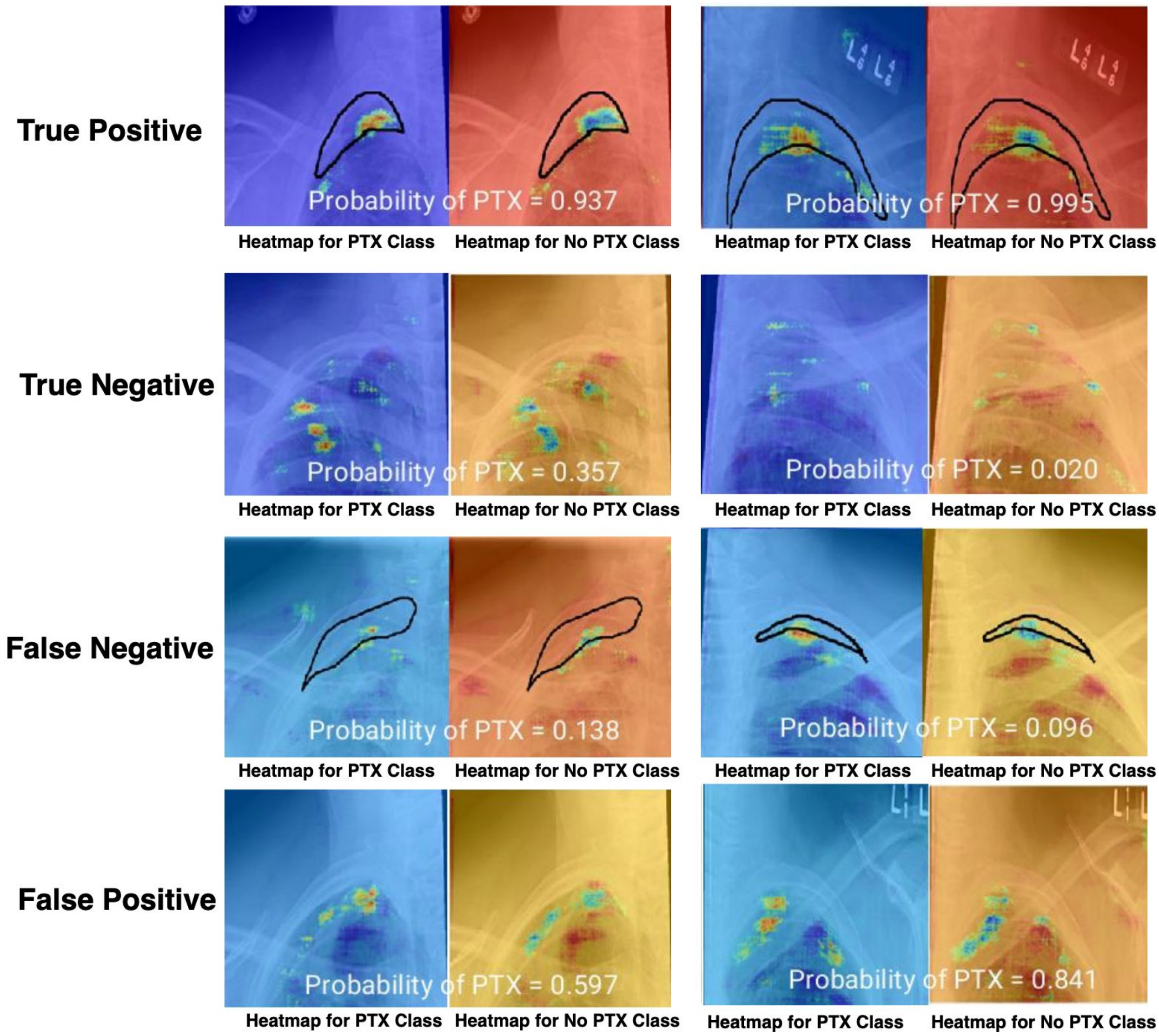


Figure 5.21: Apex heatmaps from BagNet for true-positive, true-negative, false-negative, and false-positive cases. The true-positive and false-negative cases have the true location of the PTX annotated as a black contour line. The heatmaps for the No PTX class are opposite to the PTX class heatmap, the most influential region of the PTX class heatmap is the least influential on the No PTX class heatmap. This is due to the classification being binary. If there were more classes, the heatmaps would not be opposites of one another.

5.6.4 Comparison between BagNet visualization results for the full radiographs and apex images

Figure 5.22 is a histogram of the Dice score for the overlap between the highly influential region of the BagNet PTX-class heatmap and the radiologist annotated location of the PTX.

The distributions of Dice score for the full radiographs and the apex images are shown with the median, average, and standard deviation of Dice score, along with the skewness.

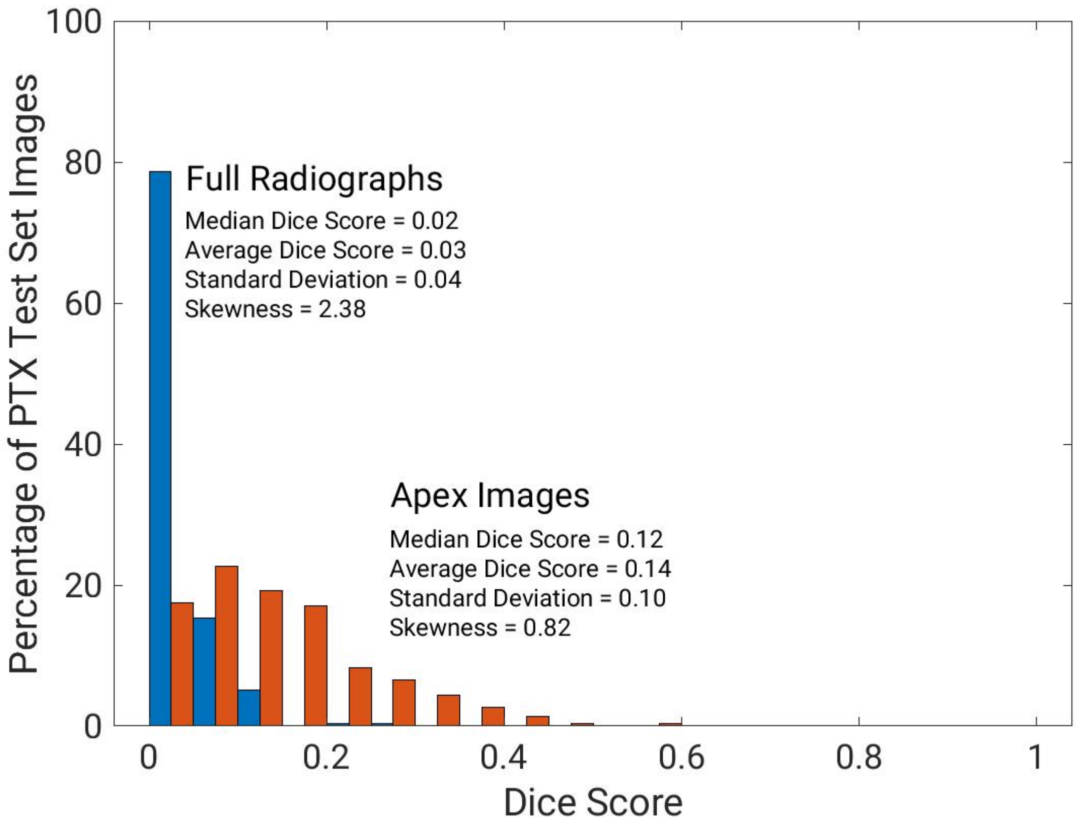


Figure 5.22: Histogram of Dice score for the full radiographs and the apex images, calculated using the radiologist-annotated location of PTX and the 3-category fuzzy c-means-determined region of influence of the BagNet PTX-class heatmap.

Figure 5.23 is a scatter plot showing the probability of PTX for each test image, as output by the BagNet CNNs fine-tuned with the full radiographs and apex images, as well as its Dice score when quantifying overlap between its PTX class heatmap and the radiologist annotation. Ideally, points should be in the top right corner, corresponding to a high Dice score and high probability of PTX.

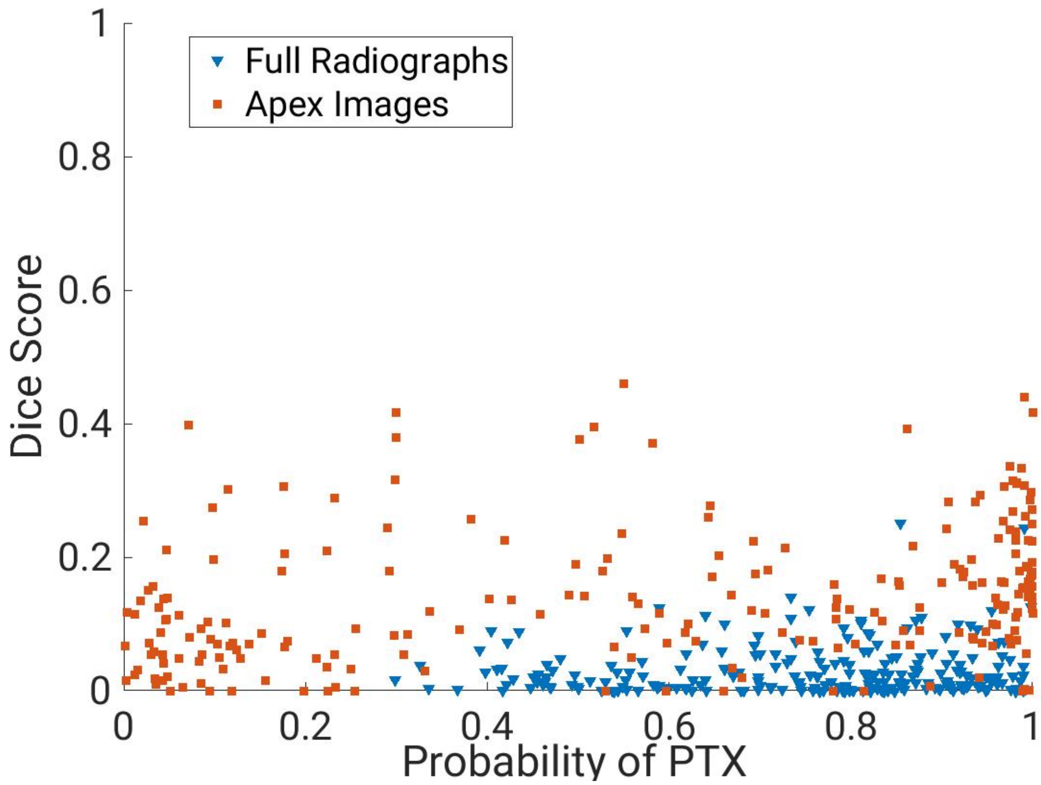


Figure 5.23: Scatter plots of Dice score and probabilities from BagNet for the full radiographs and the apex images with PTX. Ideally, points should be in the upper right corner, corresponding to a high probability of PTX, as well as a high Dice score.

5.6.5 BagNet Visualization Discussion and Conclusions

Analyzing the BagNet heatmaps and calculating the Dice score showed that, on average, the Dice score was higher for the CNN fine-tuned with the apex images (0.14 for apex images vs. 0.03 for full radiographs). The distribution of Dice score was more skewed for the full radiographs (skewness=2.38) than the apex images, with 85 heatmaps having a Dice score <0.01.

The BagNet heatmap has a different appearance than the Grad-CAM heatmaps. The class evidence is summed up across patches to generate a full heatmap. The effect of that is a heatmap with scattered patches of the image that have varying influence on the classification. Qualitatively, many of the apex images had the region of highest influence (dark red) on or within the PTX contour; however, when 2-category fuzzy c-means clustering was applied, the influential region was scattered throughout the image. Therefore, the Dice score did not truly indicate whether the most influential region was overlapped with the PTX. To remove some of the less influential parts of the heatmap from the Dice score calculation and obtain a region of highest influence, 3-category fuzzy c-means was applied. However, even with the 3-category fuzzy c-means clustering, the region of highest influence was scattered, making the Dice score low. Based upon these observations, fuzzy c-means clustering with more categories may more correctly characterize the overlap of the contour with the region of the heatmap that has the maximum influence.

5.7 Discussion & Conclusions for the Four Networks' Visualizations

The median, average, and standard deviation of the Dice score, as well as skewness of the Dice score distribution, for each of the fine-tuned network architectures are summarized in Table 5.1. A Dice score of 1 indicates a perfect overlap between the region of the heatmap with the highest influence and the radiologist annotation. A negative skewness indicates that the data is spread to the left of the mean Dice score. The apex images had higher average Dice scores and the skewness values were less than those of the full radiographs for all four fine-tuned CNN architectures.

Table 5.1: Summary of Quantitative Results for the Visualization Methods

Fine-Tuned CNN Architecture	**CNN fine-tuned with:**	**Median Dice Score**	**Average Dice Score**	**Standard Deviation of Dice Score**	**Skewness**
AlexNet	**Full Radiographs**	0	0.03	0.06	2.80
	Apex Images	0.19	0.22	0.19	0.56
VGG19	**Full Radiographs**	0.01	0.04	0.07	3.98
	Apex Images	0.36	0.35	0.22	-0.11
ResNet50	**Full Radiographs**	0	0.01	0.06	6.33
	Apex Images	0	0.03	0.05	2.98
BagNet	**Full Radiographs**	0.02	0.03	0.04	2.38
	Apex Images	0.12	0.14	0.01	0.82

Visualizing network output for networks trained with full chest radiographs downsampled to 224 x 224 pixels indicates that networks are not identifying the true signs of a PTX and are instead learning incorrect correlations. When fine-tuned with the full radiographs, the highest average Dice score was 0.04 (VGG19). When the VGG19 Grad-CAM heatmaps for the full radiographs were examined, 33% had their region of highest influence (dark red) outside the body.

For apex images, in addition to a higher classification performance, the visualization techniques more consistently denoted the true location of the PTX as a region of high influence on the classification. The highest average Dice score for the apex images was from the fine-tuned VGG19 network, with a Dice score of 0.35. The lowest average Dice score from the four fine-tuned CNNs was 0.03 from the fine-tuned ResNet50, which was the CNN that also produced the lowest average Dice score for the full radiographs (0.01).

Quantification of the quality or utility of deep learning output visualizations is a complex problem [48]. The thresholding of the Grad-CAM heatmap impacts the resulting Dice scores since the use of different thresholds changes the region of the heatmap considered to be influential. For example, the highest average Dice score (calculated for threshold increments of 10%) for the VGG19 CNN fine-tuned with full radiographs is 0.04 (stdev=0.06) at a 70%

threshold (meaning the top 70% of the heatmap values indicate the influential region) and the lowest is 0.01 (stdev=0.04) at a 10% threshold. For the apex images, the highest average Dice score is 0.38 (stdev=0.22) at a 80% threshold and the lowest is 0.06 (stdev=0.07) at a 10% threshold. In this work, the heatmap was separated into a background region and an influential region to try to avoid choosing an arbitrary threshold, and then the Dice similarilty coefficient was calculated to evaluate how much the influential region overlapped with the radiologist annotated PTX. The PTX annotation was an area, which was the primary motivation for choosing the Dice score as the performance metric, since it quantifies the overlap of two areas.

The most specific visual signs of a PTX include a fine line at the edge of the lung and a change in texture outside the lung; therefore, a heatmap that has an area of activation just along the line at the edge of the lung still would qualitatively be considered accurate, but may have a low Dice score if the region of PTX is large. While the Dice scores are small even for the apex images, the full extent of the PTX does not necessarily need to be identified; the line at the edge of the lung would identify the presence of PTX. If performing a segmentation task, identification of the full extent of the PTX would be of greater importance.

As deep learning continues to push towards clinical application, having an interpretable output for radiologists would further increase the utility of deep learning in the clinic. The network output visualization heatmap for each class could be displayed to a radiologist along with the probability of PTX as determined by the fine-tuned network.

CHAPTER 6

SUMMARY AND FUTURE DIRECTIONS

The major contributions of this work to the field of deep learning for medical image analysis are workflow enhancement, diagnostic improvements, and AI output understanding through the tasks of detection and visualization of pneumothorax on thoracic radiographs. This work could improve the delivery and effectiveness of patient care by incorporation of deep learning along various stages of the clinical workflow.

Once medical images are acquired, the organization of the images is usually accomplished using DICOM header information; however, DICOM header information can be incorrect or inconsistent. The results from Chapter 3 demonstrated that CNNs trained from scratch can rapidly and accurately classify between the four radiographic views resulting from a dual-energy study, in addition to classifying between frontal radiographs acquired AP and PA. These results demonstrate the ability of deep learning to be applied for enhancing the radiology workflow.

The detection of pneumothorax within the radiograph is complicated by the wide variety of sizes and severities with which they can present, in addition to the many overlapping structures in a chest radiograph. Chapter 4 investigated the use of transfer learning techniques, specifically fine-tuning and feature extraction, for the classification between images with and without PTX. In addition, the impact of input image resolution on the transfer learning results was investigated. Fine-tuning was performed with AlexNet, VGG19, ResNet50, and BagNet for full frontal chest radiographs, apex images derived from those radiographs, and apex images padded with zeros to the original radiograph size. It was found that the performance for the apex images was higher than the performance of the full radiographs, with a statistically significant difference in performance for all of the fine-tuned CNNs. The specific task of applying deep learning to pneumothorax detection, in addition to the investigation of input image resolution on transfer learning performance, yields diagnostic improvements potentially applicable to other imaging tasks.

The visualization of deep learning output is important for clinical use of deep learning algorithms and the lack of interpretability has been identified as one of the major barriers to clinical adoption. Visualization enables the determination of whether CNNs trained/fine-tuned to detect PTX are actually detecting the visual signs used by radiologists, the fine line at the edge of the lung and the absence of texture outside the lung. Chapter 5 presented visualizations for the 4 fine-tuned CNNs (from Chapter 4) for the full radiographs in addition to the apex images. The performance and localization ability of the heatmaps was quantified through fuzzy c-means clustering and then the Dice score was used as the performance metric for each heatmap. It was found that the average Dice score was higher for heatmaps generated for the apex images than the heatmaps generated for the full radiographs. The distribution of Dice score was quantified by calculating skewness. For all four CNNs, the distribution of Dice score was less skewed for the apex images than for the full radiographs, meaning that the mean Dice score for the apex images was further from zero than the mean Dice score of the full radiographs. The generation of visualizations and evaluation of their performance demonstrated the potential for interpretability of deep learning CNN output. The increased average Dice score and reduced skewness for the apex images compared to the full radiographs further demonstrates the benefits of using a more limited field of view for input to deep learning.

Having completed these studies, several limitations are identified that could be improved in future research. The limitations and suggested future work are presented here. One limitation is the image truth data. The training of deep learning algorithms relies on complete and accurate ground truth for the images in order to update the weights and optimize the performance. For the classification of radiograph view in Chapter 3, there was no ambiguity of truth for the images for the four-way view classification. The projection view was clearly discernible by simply viewing the radiograph. The truth for the AP versus PA classification is a little more difficult to discern; however, a trained radiologist can identify whether a frontal radiograph is acquired AP or PA. Assigning ground truth to the PTX images (Chapter 4) is a more complex process; radiology reports are generated for patient care and should not be used for deep learning truth if avoidable [44]. Therefore, after images at the University of Chicago Medical Center were identified by searching radiology reports and downloaded, a radiologist verified the presence of PTX. Since only one radiologist reviewed these images, some false positives could have been included in the PTX dataset and some of the more subtle cases could have been excluded. For the images from ChestX-ray14, Wang et al. [11] claimed at least a 90% accuracy of the labels; to verify the label truth of images from ChestX-ray14, a radiologist reinterpreted all the images labeled as PTX. After interpreting all of the radiographs at two separate instances with the previous interpretation unknown, the truth was compared to the truth released for a portion of the ChestX-ray14 dataset by the SIIM/ACR Pneumothorax Challenge [45] from another radiologist's interpretations. Radiographs were used only when the challenge truth's radiologist and our radiologist agreed, which, again, could have lead to some false positives being included or subtle cases missed. Potential future work involves having another radiologist, one who specializes in thoracic radiology, view the radiographs again to verify the ground truth.

The test images were from the University of Chicago Medical Center. Even though they were not from any of the same patients as in the training or validation sets, testing the fine-tuned networks on a new external test set from another institution not used for training or validation would be beneficial to evaluate the robustness of the fine-tuned CNNs and to determine whether they are generalizable to images from other institutions.

Another limitation is that the evaluation of PTX detection using deep learning was performed only for PTX visible in the apex. The apex was selected for the investigation of the impact of higher-resolution input images on deep learning performance since PTX is often visible in the apex. Future work could involve using other portions of the radiographs. In addition, the CNNs fine-tuned with full radiographs could be tested on cases in which the PTX is not visible in the apex but visible elsewhere in order to evaluate the generalizability of the fine-tuned CNNs.

Four architectures and two visualization techniques were selected for these studies. A limitation of this work is that there are many more techniques for training CNNs, potential CNN architectures, and visualization techniques. Future work could compare these results to results from other visualization techniques and network architectures. Visualization techniques that perturb different regions of the image to evaluate the importance of each region for the classification, such as occlusion mapping [52], would be valuable to compare with the results presented in Chapter 5. In addition, for calculation of the detection sensitivity and specificity, a probability of PTX output by the CNN of 0.5 or greater caused an image to be considered as a PTX. The threshold could be varied in order to more fully characterize the performance of the CNNs. They may have a poor sensitivity or specificity at the 0.5 threshold; however, at another chosen threshold, the performance may be stronger.

View classification between AP and PA radiographs (Chapter 3) had a high performance; however, a statistically significant decrease in performance was seen when the labels generally associated with AP radiographs were removed. Further investigation of the causes of this performance change could be performed through the application of visualization techniques.

The accuracy of the heatmaps was quantified using the Dice score to calculate the overlap between the radiologist annotation and the influential region of the heatmap. In addition to the limitations and potential future work presented at the end of Chapter 5 (use of other performance metrics and different thresholding techniques), one limitation was the radiologist-annotated PTX contours. The contours were drawn using a standard mouse and the workstation did not allow for zooming the image while drawing the contour. Therefore, due to these hardware and software limitations, the contours may not be precisely denoting the location of the PTX. Potential future work involves drawing the contour again on a high-definition monitor with a high-definition mouse and the ability to zoom in. The analysis of the performance of the heatmaps could be performed again with these improved contours and the performance compared to this work. In addition, since the contour was drawn by one radiologist, having another radiologist draw the contour could be of interest to evaluate the interobserver variability and repeat the analysis.

This work demonstrates the potential applications for deep learning to medical imaging tasks including the enhancement of radiology workflow, improvement of medical image diagnosis, and explanatory output from deep learning algorithms. Workflow enhancement can be achieved through the use of a deep learning model for the classification of radiographic views from a dual-energy, a standard, or a portable chest radiography study, reducing the reliance on DICOM headers for proper display and storage. Deep learning can improve diagnosis through the detection of PTX on frontal chest radiographs; this work demonstrated the impact of input image resolution on deep learning performance, indicating the importance of deep learning algorithms customized for the task being performed. Human-interpretable and explanatory output from deep learning algorithms is needed for clinical implementation. This work showed visualizations of PTX detection on the images and quantified the performance of the visualizations. Overall, this work demonstrates the potential of deep learning to be applied in radiology practice to enhance workflow, improve and enhance diagnosis of medical images, as well as provide human-interpretable explanations of the output.

REFERENCES

[1] L.B. Lusted. Medical electronics. *New England Journal of Medicine*, 252:580–585, 1955.

[2] M.L. Giger, H.P. Chan, and J. Boone. Anniversary paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM. *Medical Physics*, 35(12):5799–5820, 2008.

[3] M.L. Giger. Machine learning in medical imaging. *Journal of the American College of Radiology*, 15(3):512–520, 2018.

[4] W. Zhang, K. Doi, M.L. Giger, Y. Wu, R.M. Nishikawa, and R.A. Schmidt. Computerized detection of clustered microcalcifications in digital mammograpms using a shift-invariant artificial neural network. *Medical Physics*, 21(4), 1994.

[5] F Chollet. *Deep Learning with Python*. Manning Publications, 2018. 152-177.

[6] B.J. Erickson, P. Korfiatis, Z Akkus, and T.L. Kline. Machine learning for medical imaging. *RadioGraphics*, 37(2):505–515, 2017.

[7] G Choy, O Khalilzadeh, M Michalski, S Do, AE Samir, OS Pianykh, JR Geis, PV Pandharipande, JA Brink, and KJ Dreyer. Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318–328, 2018.

[8] P Rachh, AO Levey, A Lemmon, A Marinescu, WF Auffermann, D Haycook, and EA Berkowitz. Reducing STAT portable chest radiograph turnaround times: A pilot study. *Current Problems in Diagnostic Radiology*, 47:156–160, 2018.

[9] P Rajpurkar, J Irvin, K Zhu, B Yang, H Mehta, T Duan, D Ding, A Bagul, C Langlotz, K Shpanskaya, MP Lungren, and AY Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv: 1711.05225*, 2017.

[10] L Yao, E Poblenz, D Dagunts, B Covington, D Bernard, and K Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv: 1710.10501v2*, 2018.

[11] X Wang, Y Peng, L Lu, Z Lu, M Bagheri, and RM Summers. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit*. IEEE, 2017. arXiv 1705.02315.

[12] M Cicero, A Bilbily, E Colak, T Dowdell, B Gray, K Perampaladas, and J Barfett. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Investigative Radiology*, 52:281–287, 2017.

[13] Y. Feng, H.S. Teh, and Y. Cai. Deep learning for chest radiology: A review. *Current Radiology Reports*, 7(24), 2019.

[14] MK Harkness, A Hashim, and D Spence. The "hidden" pneumothorax. *Emergency Medicine Journal*, 21(3):386–387, 2004.

[15] K Chen, JS Jerng, WY Liao, LW Ding, LC Kuo, JY Wang, and PC Yang. Pneumothorax in the ICU: Patient outcomes and prognostic factors. *Chest*, 122:678–683, 2002.

[16] RS Winokur, BB Pua, BW Sullivan, and DC Madoff. Percutaneous lung biopsy. *Semin Intervent Radiol*, 30:121–127, 2013.

[17] W. Ding, Y. Shen, J. Yang, X. He, and M. Zhang. Diagnosis of pneumothorax by radiography and ultrasonography: A meta-analysis. *Chest*, 140:837–839, 2011.

[18] ACR–SPR–STR practice parameter for the performance of chest radiography. *American College of Radiology*, 2017.

[19] ACR–SPR–STR practice parameter for the performance of portable (mobile unit) chest radiography. *American College of Radiology*, 2017.

[20] A.Z. Rasheed, L. Latypov, M. Shahzadi, V.P. Vasudevan, L.N. Gerolemou, and F. Arjomand. Inappropriate utilization of portable chest radiography in diagnosis and treatment of cardiopulmonary disease. *American Thoracic Society Conference 2018.*

[21] J Lemos and J.S. Klein. Methods of examination, normal anatomy, and radiographic findings of chest disease. In William E Brant and Clyde A Helms, editors, *Fundamentals of Diagnostic Radiology*, pages 324–366. Wolters Kluwer/Lippincott Williams Wilkins, Philadelphia, 2012.

[22] ACR appropriateness criteria: Acute respiratory illness in immunocompetent patients. *American College of Radiology*, 2018.

[23] A.W. Kirkpatrick, M. Sirois, and K.B. Laupland. Hand-held thoracic sonography for detecting post-traumatic pneumothoraces: the extended focused assessment with sonography for trauma (EFAST). *J Trauma*, 57(2):288–295, 2004.

[24] H. MacMahon S. Sanada, K. Doi. Image feature analysis and computer-aided diagnosis in digital radiography: automated detection of pneumothorax in chest images. *Medical Physics*, 19(5):1153–1160, 1992.

[25] A.M. Turing. Computing machinery and intelligence. *Mind*, LIX(236):433–460, 1950.

[26] L Bottou and O Bousquet. Stochastic gradient descent. In S Sra, S Nowozin, and SJ Wright, editors, *Optimization for Machine Learning*, pages 351–368. MIT Press, Cambridge, Massachusetts, USA, 2012.

[27] J Duchi, E Hazan, and Y Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[29] A Krizhevsky, I Sutskever, and GE Hinton. ImageNet classification with deep convolutional neural networks. In *Adv Neural Inf Process Syst*, 2012. 1098-1105.

[30] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin. Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annu Symp Proc.*, pages 979–984, 2017.

[31] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[32] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, 35(5):1299–1312, 2016.

[33] Y LeCun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[34] From not working to neural networking. *The Economist*, 25 June 2016.

[35] K Simonyan and A Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[36] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[37] Wieland Brendel and Matthias Bethge. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. *International Conference on Learning Representations (ICLR)*, 2019.

[38] CE Metz. Basic principles of ROC analysis. *Semin Nucl Med*, 8(4):283–298, 1978.

[39] D.A. Clunie. DICOM implementations for digital radiography. *Advances in Digital Radiography: RSNA Categorical Course in Diagnostic Radiology Physics*, pages 163–172, 2003.

[40] A Rajkomar, S Lingam, AG Taylor, M Blum, and J Mongan. High-throughput classification of radiographs using deep convolutional neural networks. *Journal of Digital Imaging*, 30(1):95–101, 2017.

[41] J. Crosby, T. Rhines, F. Li, H. MacMahon, and M.L. Giger. Deep convolutional neural networks in the classification of dual-energy thoracic radiographic views for enhanced workflow. *Journal of Medical Imaging*, 7(1), 2020.

[42] ER DeLong, DM DeLong, and DL Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[43] C.F. Sabottke and B.M. Spieler. The effect of image resolution on deep learning in radiography. *Radiology: AI*, 2(1), 2020.

[44] David A. Bluemke, Linda Moy, Miriam A. Bredella, Birgit B. Ertl-Wagner, Kathryn J. Fowler, Vicky J. Goh, Elkan F. Halpern, Christopher P. Hess, Mark L. Schiebler, Clifford R. Weiss, and et al. Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers—from the Radiology editorial board. *Radiology*, 2019.

[45] SIIM-ACR. Pneumothorax segmentation challenge. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/overview/description, 2019.

[46] C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[47] B Sahiner, A Pezeshk, LM Hadjiiski, X Wang, K Drukker, KH Cha, RM Summers, and ML Giger. Deep learning in medical imaging and radiation therapy. *Medical Physics*, 46(1):e1–e36, 2019.

[48] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[49] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. *CoRR*, abs/1905.05134, 2019.

[50] RR Selvaraju, A Das, R Vedantam, M Cogswell, D Parikh, and D Batra. Grad-CAM: Why did you say that? Visual explanations from deep networks via gradient-based localization. *arXiv: 1610.02391*, 2016.

[51] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 10 2017.

[52] M.D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *ECCV 2014: Lecture Notes in Computer Science*, 2004.

[53] O Ronneberger, P Fischer, and T Brox. U-Net: Convolutional networks for biomedical image segmentation. *arXiv:1505.04597*, 2015.

[54] P. Zarogoulidis, I. Kioumis, G. Pitsiou, K. Porpodis, S. Lampaki, A Papaiwannou, N Katsikogiannis, B Zaric, P Branislav, N Secen, G Dryllis, N. Machairiotis, A Rapti, and K Zarogoulidis. Pneumothorax: From definition to diagnosis and treatment. *Journal of Thoracic Disease*, 6(4):S372–S376, 2014.

[55] Y Dodge. *The Concise Encyclopedia of Statistics*. Springer, 2008. 283-286.

APPENDIX A

FURTHER INVESTIGATION OF DEEP LEARNING METHODS IN THE TASK OF DETECTION OF PTX IN CHEST RADIOGRAPHS THROUGH THE USE OF HIGHER-RESOLUTION PATCHES

In Chapter 4, the results showed that the classification performance was highest for the CNNs fine-tuned with the apex images. To further investigate the impact of image resolution on CNN performance, the apex images were divided into smaller patches and used for fine-tuning. In addition, the patches were used with the U-Net segmentation network to investigate the ability of the U-Net CNN in the task of detecting and segmenting the edge of the lung.

A.1 Fine-Tuning with Higher-Resolution Patches of the Apex Images

To train a network to localize PTX, the location of the PTX must be annotated; however, radiologist annotation is an expensive and time-consuming process. Therefore, VGG19 CNNs were fine-tuned with a small number of annotated radiographs and tested on non-annotated radiographs. The annotated radiographs used for testing the CNNs in Chapter 4 and for quantitatively evaluating the visualizations in Chapter 5 were used to train/validate the VGG19 CNNs. The radiographs used in Chapters 4 and 5 for training and validation were used as a testing set in this extra experiment since the location of the PTX had not been annotated; therefore, they did not have the ground truth necessary to be used for training/validation in this experiment.

A.1.1 Methods

A VGG19 network pre-trained on ImageNet had the last two convolutional layers and the final classification layer retrained for the new task of classifying images with and without PTX. The training set consisted of 225 radiographs with PTX and 350 radiographs without PTX (used for testing in Chapter 4). Data augmentation was applied, including rotations (up to 10 degrees) and small vertical/horizontal shifts (25 pixels). One VGG19 CNN was fine-tuned with the apex images (resized to 256 x 256 via bilinear interpolation) and another VGG19 CNN was fine-tuned with 256 x 256 high-resolution patches extracted from within the apex images. The apex images were resized such that each image would have 9 patches. A roaming region of interest (ROI) of size 256 x 256 pixels was used to generate the patch images such that 128 pixels overlapped between patches in each dimension. The truth for the patches was determined using the radiologist contour; if any portion of the contour was contained within the patch, it was considered to be a patch with PTX. Figure A.1 shows the division of a PTX apex image and an apex image without PTX, and their corresponding radiologist annotation.

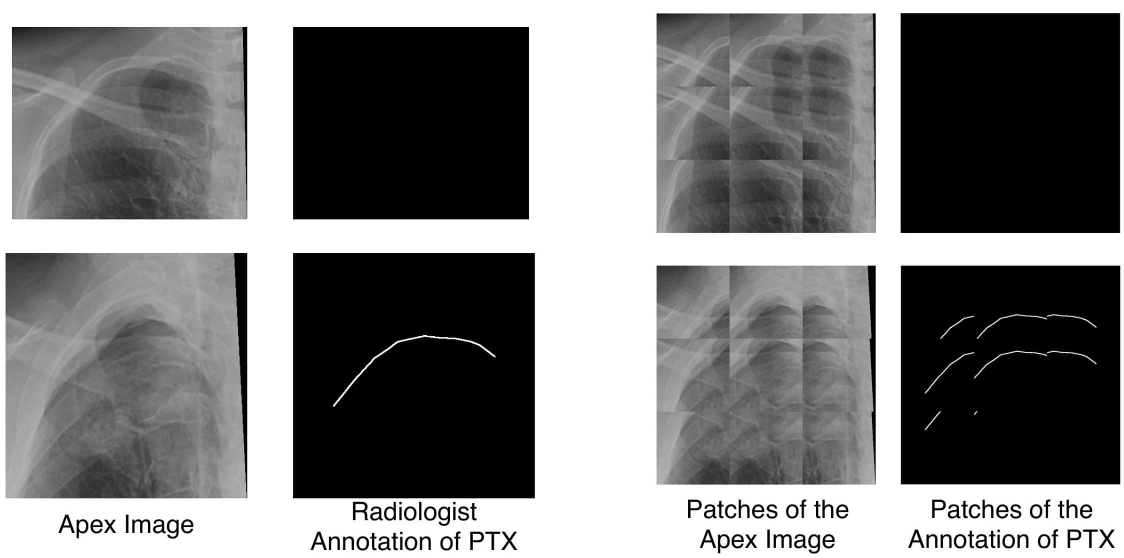


Figure A.1: Apex images and their corresponding radiologist annotation, patches, and patches of the radiologist annotation. The top row is the apex without PTX and the bottom row is the apex with PTX.

The testing set consisted of the remaining 2,686 PTX radiographs and 2,070 radiographs without PTX, which were the radiographs used as the training and validation sets when fine-tuning the CNNs in Chapter 4. Note that images from the same case were completely in either the training set or the test set. The network probabilities of belonging to the PTX class or non-PTX class were evaluated separately for the apex images and for the patches derived from the apex images. The apex with the larger probability of PTX was used as the single probability for each test case. For the classification of the patches, the patch with the highest probability of PTX was used as a single probability of PTX for each test case. Then the results were merged using soft voting. ROC analysis was performed and the area under the ROC curve (AUC) used as a performance metric. After the ROC curves were generated for the apex images, patches of the apex images, and their merged results via soft voting, the curves were compared using DeLong's method [42] to determine whether there was a statistically significant difference in performance. Since there were 3 comparisons, the significance level (α) was corrected for multiple comparisons using the Bonferroni correction [46], $\alpha = 0.017$. Figure A.2 gives the overall workflow from acquisition of a frontal radiograph to its merged apex and patch probability of PTX.

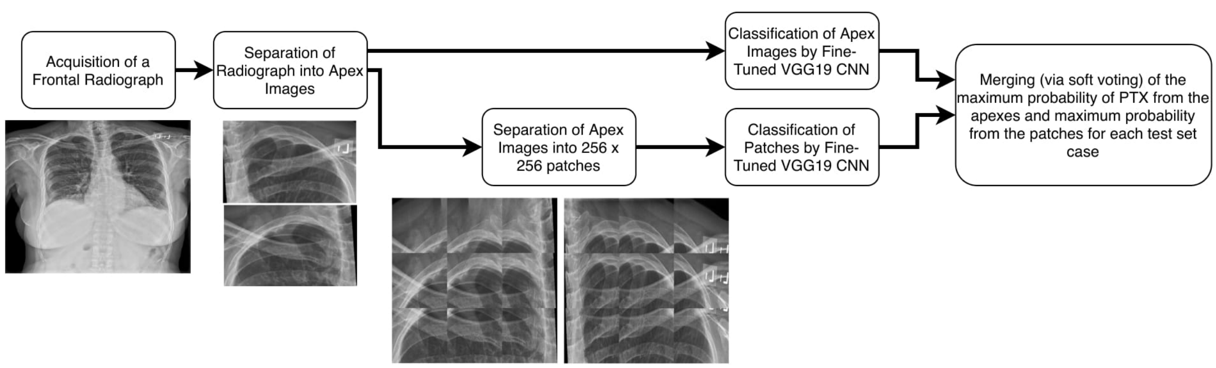


Figure A.2: Overview of the workflow that incorporates portioning the radiograph into apex images and then into patches of the apex image, and their classification.

A.1.2 Results for Patches of the Apex Images

The apex-based network was fine-tuned with apex images and tested on apex images of the independent test set. ROC analysis was performed, yielding an AUC of 0.799 (95% CI: 0.787, 0.809) in the task of distinguishing between apex images with and without PTX. For the patch-based network, which was fine-tuned with 256 x 256 patches of the apex images, the AUC was 0.726 (95% CI: 0.713, 0.738) in the task of distinguishing between apex images containing and not containing PTX for the independent test set, using a single likelihood value for each test set case from the patch with the highest likelihood of PTX. The use of the patch with the highest likelihood allows for localization of the PTX. Other patches with high likelihood value could also be considered as part of the PTX for further localization ability. After merging the output from the two deep learning methods via soft voting, the AUC was 0.826 (95% CI: 0.816, 0.836) in the task of distinguishing between images with and without PTX. This merging of the outputs yielded a statistically significant improvement in performance compared to either network alone ($p < 0.001$). Figure A.3 shows the ROC curves for each network and their merged output.

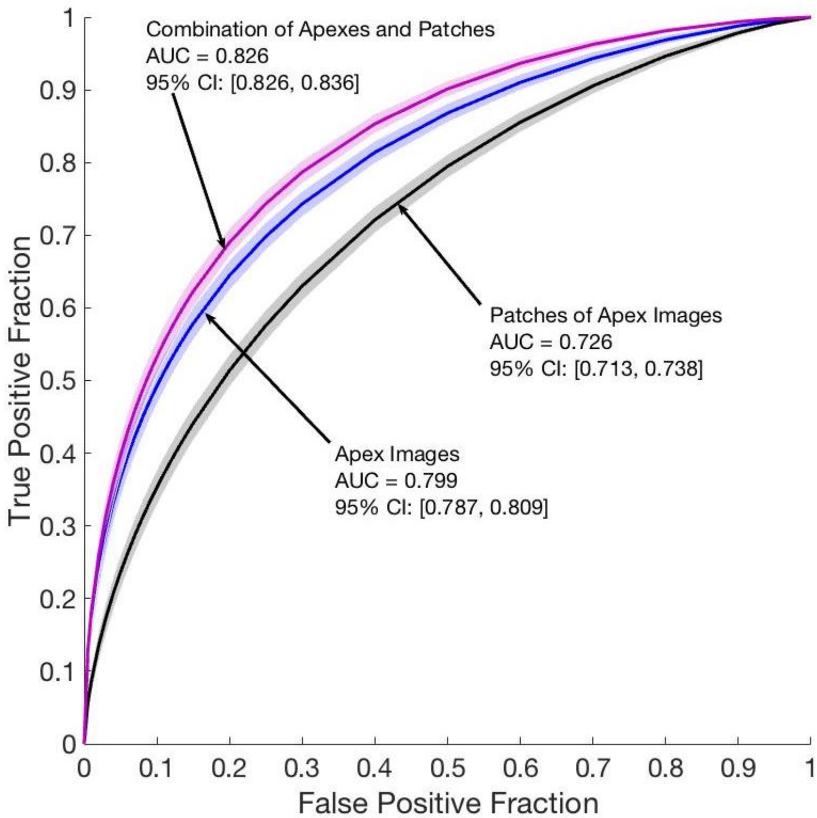


Figure A.3: ROC curves for each method individually and then combined. Each ROC curve is statistically significantly different from one another with p-values <0.001. The shaded area around each curve is the 95% confidence interval of the ROC curve.

A.1.3 Discussion & Conclusions

These results demonstrate the usefulness of deep learning methods for both image-based classification and abnormality-based localization. The VGG19 networks were fine-tuned with a small number of annotated images and still showed a strong performance on the independent test set for detection of the PTX within the image as well as in the indication of the patch with the highest likelihood of containing PTX. The merging of the outputs led to a statistically significant increase in performance, showing that the combination of two levels of image resolution may be synergistic and could be used for other deep learning applications where image resolution is of concern.

A limitation was the use of a small number of annotated radiographs for training; when testing, the patch with the highest probability of PTX was used as the value for the full image; however, it is unknown whether that was a patch that actually contained PTX. Future work could include using more annotated images for training, as well as for testing. In addition, in this study, if any of the contour was contained in the patch, it was considered to be a PTX image. Future work could involve adjustment of the number of pixels of the contour required to consider the patch a PTX patch, excluding those with fewer pixels from training. This could improve network performance due to providing only PTX patches with sizable annotation within them for training.

Another limitation and area of future study is the division of the apex images into the patches. The use of 9 patches per apex image was arbitrarily chosen; however, any number of patches could be chosen. Future work could further investigate which methods for generating the patches would lead to the strongest performance.

A.2 Training a U-Net CNN from Scratch with Higher-Resolution Patches of the Apex Images

In addition to fine-tuning with the higher-resolution patches of the apex image, U-Net, a segmentation network, was investigated to determine whether it could localize the PTX when trained with the higher-resolution patches of the radiographs and their corresponding radiologist annotations.

A.2.1 U-Net Background

About U-Net

U-Net [53] is a segmentation network specifically developed for segmentation of medical images. It is unique in its architecture, which captures context through a contracting path and enables localization through a symmetric expanding path. The four networks discussed in Chapter 2 (AlexNet, VGG19, ResNet50, and BagNet) give a probability score of belonging to each class on a full image basis. U-Net gives a per-pixel class, therefore, it is a segmentation network and can localize the subject of interest.

For most medical imaging tasks, there is a scarcity of high-quality, labeled training data. For the U-Net, data augmentation is performed through elastic deformations of the training images. The rationale for this is that the network is able to learn invariance to deformations, which are common in tissue [53].

U-Net was selected to evaluate whether it could localize the PTX when trained with high-resolution patches of the radiographs and their corresponding radiologist annotation. It is a common segmentation network, specifically designed for medical images, so it is appropriate for the task.

A.2.2 Methods

The database included 240 chest radiographs with PTX from the University of Chicago. Using the five-points placed on anatomical landmarks denoted by a radiologist, the left and right lungs were cropped to yield two apex images from each CXR (480 images). To serve as localization "truth," a radiologist drew contour lines along each PTX. Binary truth masks, identical in size to the cropped images, were made by widening these contour lines to four pixels in width. The 480-image dataset was randomly split by case into three groups: 65% (312) used for training the network, 20% (96) for validation, and 15% (72) for testing the performance of the network. A U-Net neural network was trained from scratch with step-wise validation. Random jitter augmentation enhanced the training, which involved 33 epochs of 100 steps using an Adam optimizer, an adaptive learning rate, 0.05 dropout, and per-layer batch normalization. Subsequently, a scanning ROI method was used to localize the U-Net indicated PTX.

Results on Patches

During evaluation, the total pixel value sum of each scanning ROI was calculated, assigning a per-image image score as the largest ROI pixel sum in that image. The distribution of pixel sum is shown in Figure A.4.

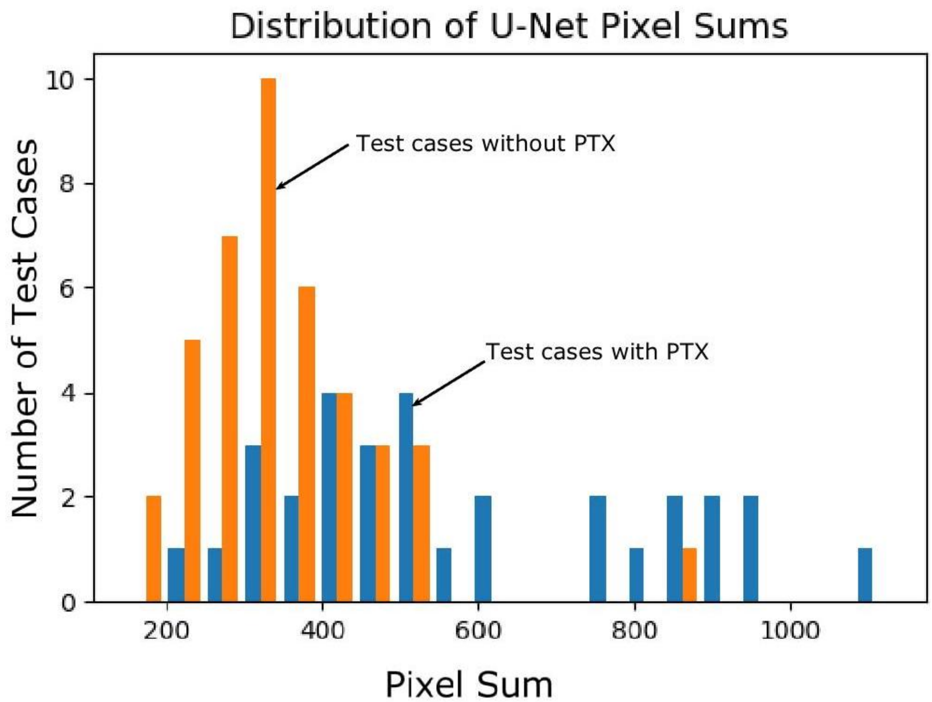


Figure A.4: Distribution of U-Net pixel sums for the test cases with and without PTX.

ROC analysis using the largest ROI pixel sums yielded an area under the curve of 0.793 (95% confidence interval: [0.673, 0.882]). Figure A.5 shows the ROC curve for the U-Net results and Figure A.6 gives examples of test image patches, their annotation truth, and the U-Net output.

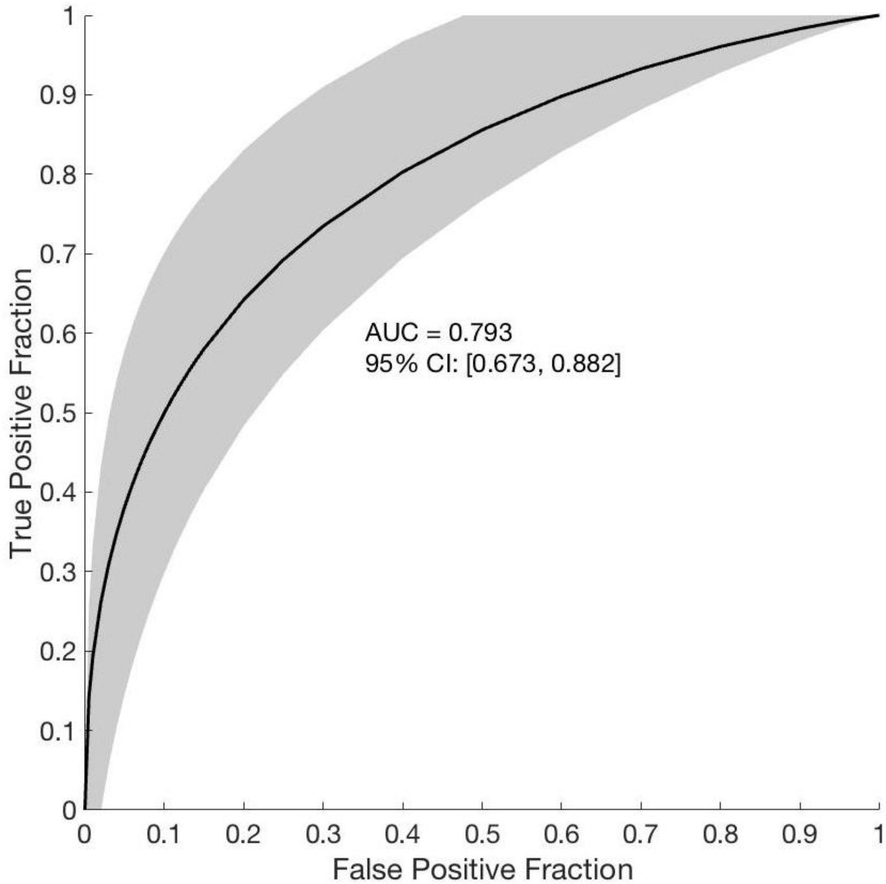


Figure A.5: ROC Curve for the U-Net trained to identify the line on the edge of the lung. The total pixel value sum was used as a score for each image for ROC analysis. The shaded area is the 95% confidence interval of the ROC curve.

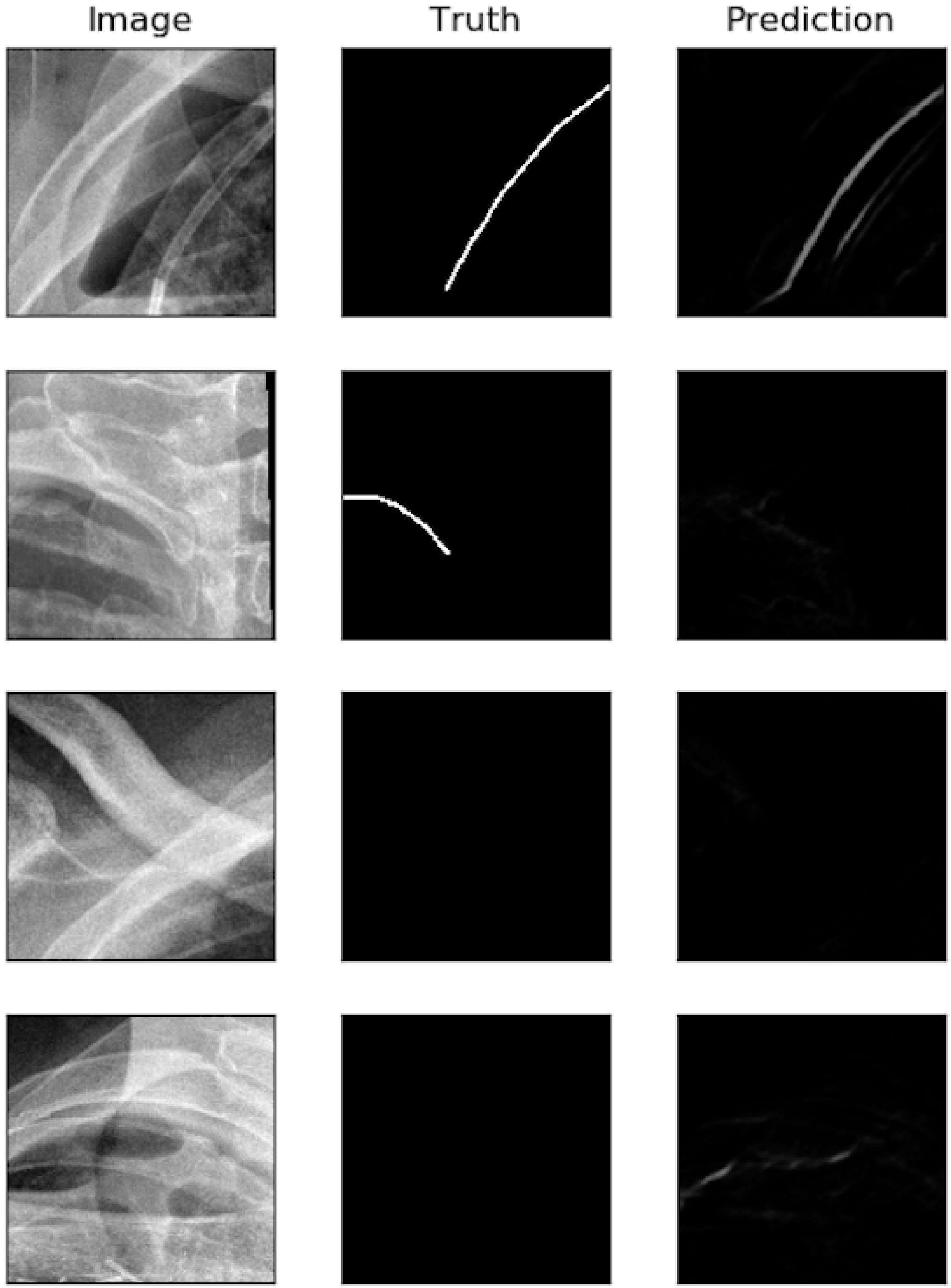


Figure A.6: Four examples of U-Net output: the top two rows show PTX cases, one of which (the top) had the prediction closely match the radiologist annotated truth. The third and fourth row show test cases without PTX; the last row shows an image with scattered segmentation predictions.

A.2.3 Discussion & Conclusions

The implementation of a U-Net architecture combined with a scanning ROI for localization demonstrated the possibility of localization-based identification of PTX in patient CXRs. This was an initial analysis with a small portion of the dataset to determine whether the use of a U-Net was beneficial for PTX localization. The purpose of detection of PTX via deep learning methods, as presented in Chapter 4, was to investigate the impact of input image resolution when using deep learning methods. The U-Net results demonstrated weaker classification performance than the fine-tuning methods discussed in Chapter 4. Future work could apply the U-Net methods to the full dataset at various input resolutions to evaluate the performance and compare to the results presented in Chapter 4.

APPENDIX B
INVESTIGATION OF THE IMPACT OF OTHER LUNG DISEASES AND DEVICES ON THE DETECTION OF PTX VIA DEEP LEARNING METHODS

Patients with PTX may have other lung conditions, such as diffuse lung disease or emphysema. In addition, when patients are diagnosed with a PTX, a chest tube is often inserted to draw out air [54]. There are also other medical devices that often appear in the radiographs of PTX patients, such as leads and catheters. Using images with other lung conditions or devices may train the CNN to detect them or associate them with the presence of PTX. To investigate this potential pitfall of applying deep learning methods to the detection of PTX, test set radiographs were interpreted and the presence/absence of a lung disease or device was recorded. The results from the fine-tuned VGG19 CNN were compared for the images with and without other diseases and devices.

B.1 Assigning Truth to Test Set Radiographs

In order to correctly evaluate the performance of the trained model, the test set radiographs need truth assigned by a radiologist. After the presence of PTX was confirmed, each test set image had 5 points placed by a radiologist: two points placed at the apices, two points placed at the costophrenic angles, and a point on top of the aortic arch. These points split the lungs within the radiograph into six sections, each of which could have individual truth assigned (Figure B.1). The purpose of this was to localize the PTX, as well as to identify other notable characteristics of each section. Five questions were asked of the radiologist for each portion of the radiograph: (1) Does the area have PTX?, (2) Is there diffuse lung disease adjacent to the PTX?, (3) Is there emphysema present?, (4) Are there chest tubes present?, and (5) Are there catheters present? There was also a text box for each portion of the radiograph so the radiologist could provide comments, such as noting the presence of surgical markers or metal implants.

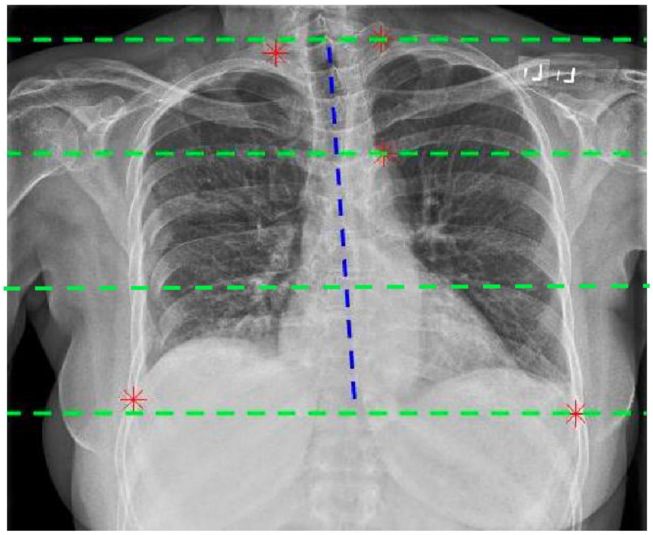


Figure B.1: Points placed at the apices, costophrenic angles, and the top of the aortic arch (red) are used to portion the lungs within the chest radiograph into six sections for truth to be assigned separately.

For the PTX cases, a contour was drawn to denote the location of PTX. Since PTX can be subtle, a "subtlety score" between 1 and 10 was assigned to each PTX, by the same radiologist, with 1 meaning the PTX was very subtle and 10 meaning the PTX was very obvious. Figure B.2 shows the area of contour (mm^2) versus the subtlety score. As the subtlety score increases, the average contour area generally increases. This matches what is expected, since larger PTXs are more likely to be obvious; however, the error bars (the standard deviation) show there is a wide range of contour area assigned to each subtlety score.

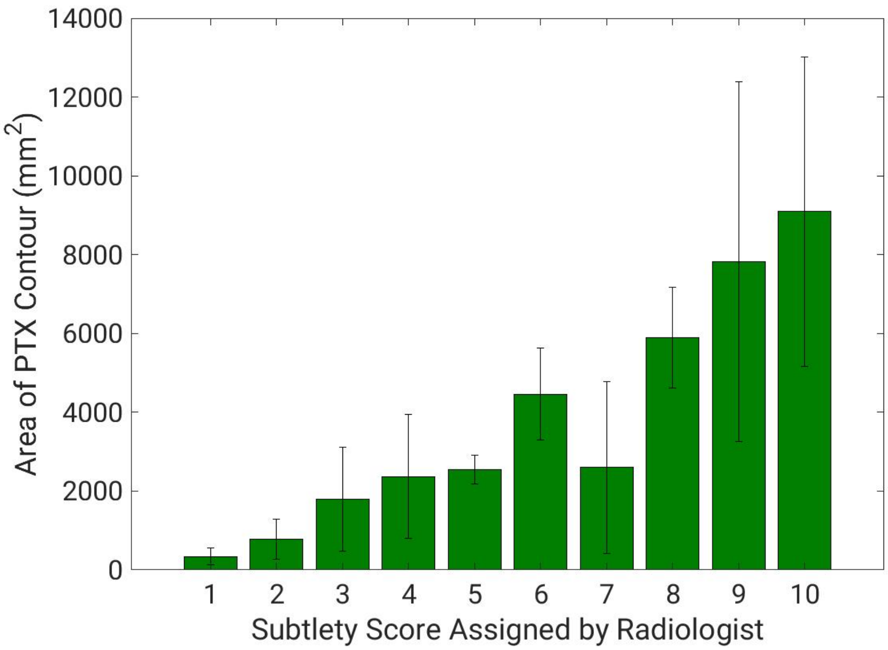


Figure B.2: The average contour area (in mm^2) for each subtlety score, with error bars representing the standard deviation.

B.2 Probability Results from the Fine-Tuned VGG19 CNNs

It may be expected that a fine-tuned CNN would be able to more effectively detect PTX for large and/or obvious cases. Figure B.3 shows the probability of PTX from the fine-tuned VGG19 CNNs for each subtlety score. There is no obvious trend showing the increase in probability of PTX assigned by the fine-tuned CNN as the subtlety score increases. Figure B.4 is a scatter plot of the assigned probability from the VGG19 CNNs versus the area of the PTX contour (in mm^2). Again, there is no obvious trend of having a higher probability of PTX assigned as the area increases.

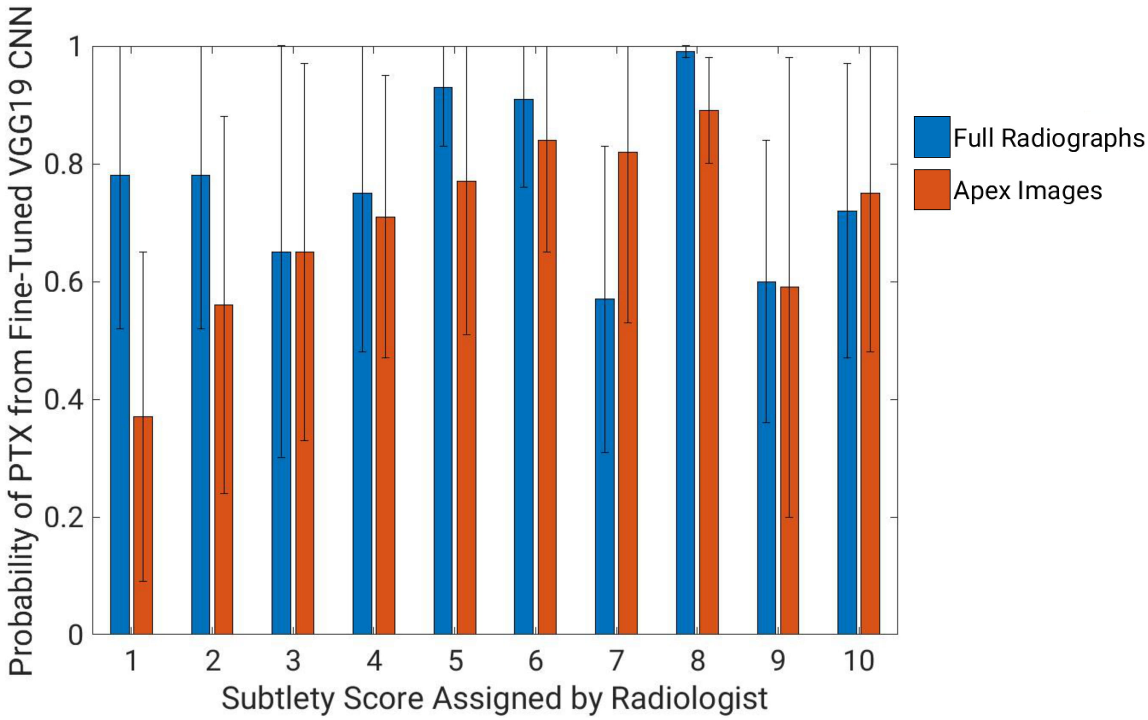


Figure B.3: The average probability of PTX for each subtlety score from the fine-tuned VGG19 CNNs. The blue represents the results for the VGG19 CNN fine tuned with the full radiographs and the orange represents the results from the VGG19 CNN fine-tuned with the apex images. The error bars represent the standard deviation.

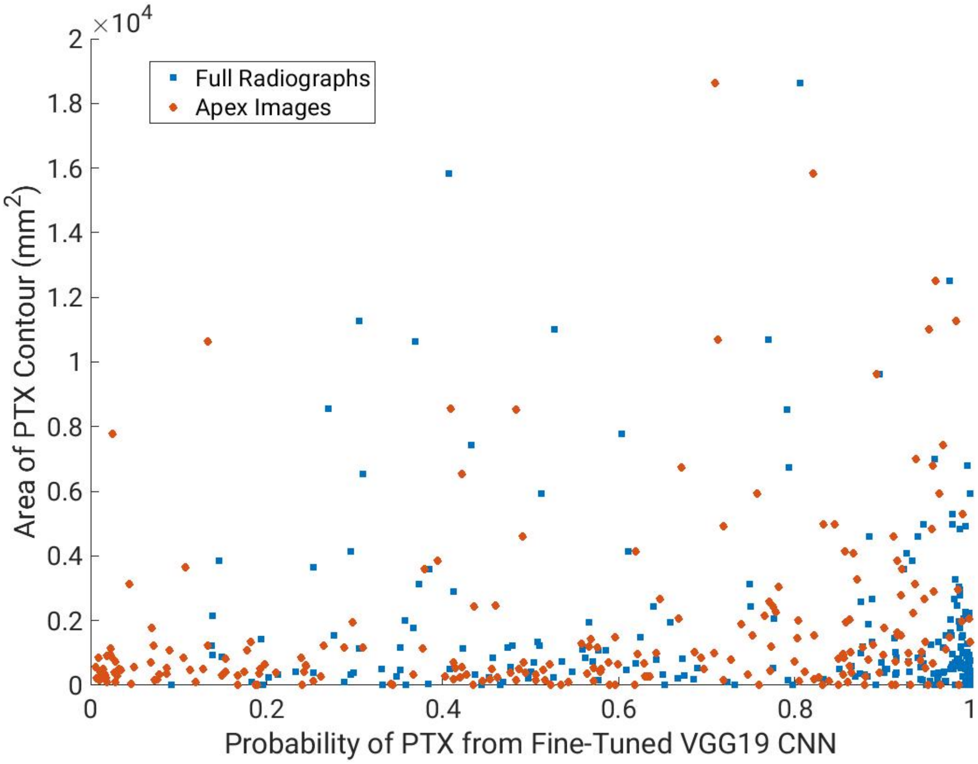


Figure B.4: Scatter plot of the probability of PTX assigned to images with PTX by each of the fine-tuned VGG19 CNNs versus the area of the contour in mm^2.

The test dataset was not evenly split between right and left PTX cases. There were 85 images with PTX on the left, 137 images with PTX on the right, and 3 images with PTX in both lungs. Figure B.5 shows the average probability of PTX output by the fine-tuned VGG19 CNNs for each side. Using the two-sample Kolmogorov-Smirnov test [55] to compare the distributions, the p-value for the comparison between right and left was 0.31 for the VGG19 CNN fine-tuned with the full radiographs and 0.14 for the VGG19 CNN fine-tuned with the apex images, therefore failing to show a statistically significant difference in distributions of PTX probability between the right and left sides.

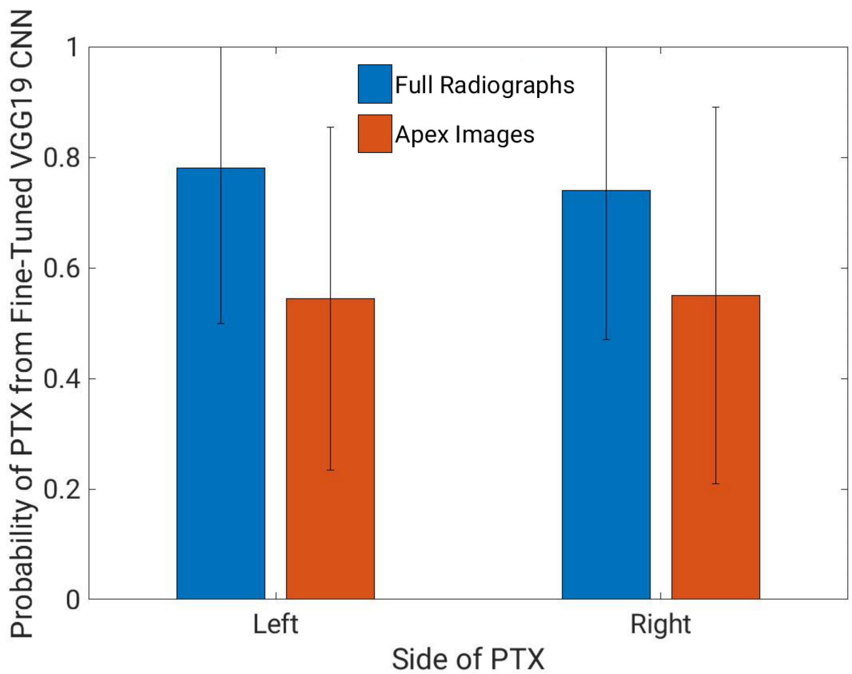


Figure B.5: The average probability of PTX for the right and left sides (excluding the three bilateral PTX cases). The error bars show the standard deviation.

If PTX was present in the image, five questions were answered by the radiologist interpreting the image for each of the six lung regions (Figure B.1): (1) Does the area have PTX?, (2) Is there diffuse lung disease adjacent to the PTX?, (3) Is there emphysema present?, (4) Are there chest tubes present?, and (5) Are there catheters present? To investigate the impact of diffuse lung disease, emphysema, chest tubes, and catheters, the average probabilities of PTX from the fine-tuned VGG19 CNNs were calculated for the groups of images that had "yes" answers and "no" answers for each of the questions. Table B.1 shows the number of "yes" and "no" answers in the apex and anywhere within the radiograph.

Table B.1: Summary of "Yes" and "No" answers in the apex and anywhere in the radiograph

	# In Apex		**# Anywhere in Radiograph**	
Question	Yes	No	Yes	No
2. Diffuse lung disease?	38 (17%)	187 (83%)	153 (68%)	72 (32%)
3. Emphysema?	21 (9%)	204 (91%)	29 (13%)	196 (87%)
4. Chest tube(s)?	20 (9%)	205 (91%)	45 (20%)	180 (80%)
5. Catheter(s)?	29 (13%)	196 (87%)	70 (31%)	155 (69%)

Figure B.6 shows the average probabilities from the VGG19 CNN fine-tuned with apex images for images with "yes" in the apex and images with "no" in the apex for questions 2 through 5. The means were compared using two-sample Kolmogorov-Smirnov tests and all failed to show a statistically significant difference in distribution of PTX probability between the "yes" and "no" groups.

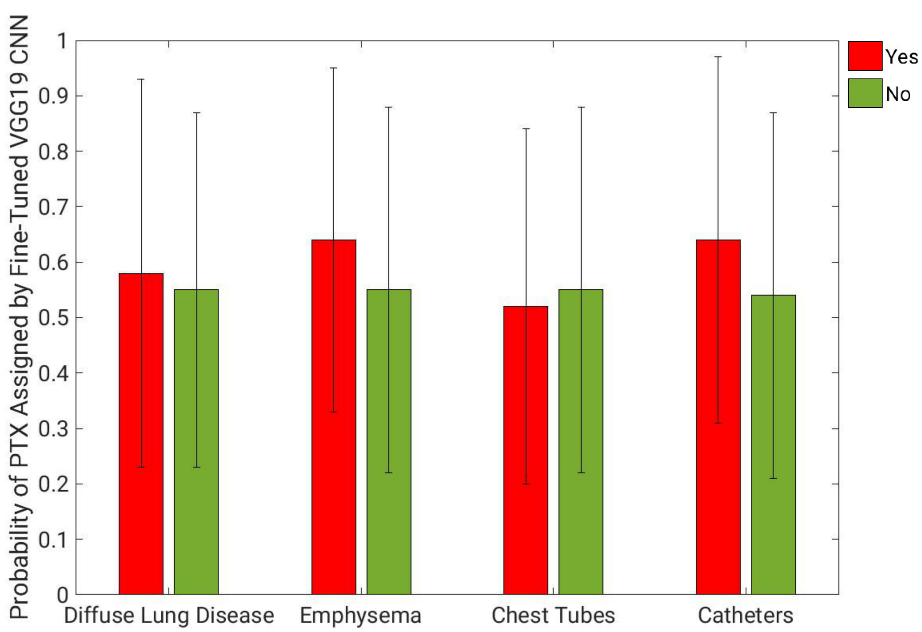


Figure B.6: The probability of PTX from the VGG19 CNN fine-tuned with apex images for cases with and without diffuse lung disease, emphysema, chest tube(s), and catheter(s) in the apex. The error bars show the standard deviation.

Figure B.7 shows the average probabilities from the VGG19 CNN fine-tuned with full radiographs for images with "yes" in the apex and images with "no" in the apex for questions 2 through 5. The distributions of PTX probability were compared using two-sample Kolmogorov-Smirnov tests and all failed to show a statistically significant difference between the "yes" and "no" groups.

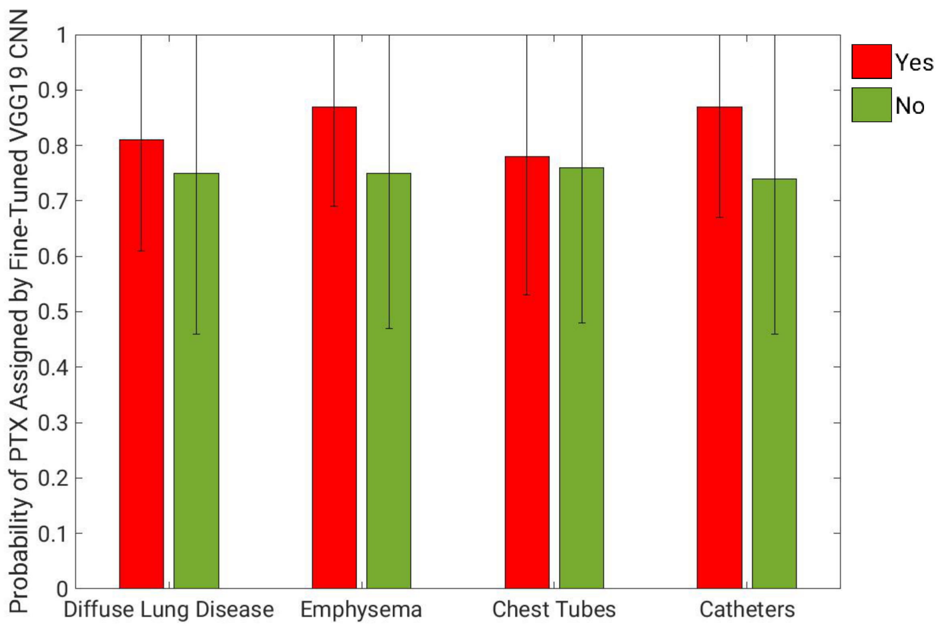


Figure B.7: The probability of PTX from the VGG19 CNN fine-tuned with full radiographs for cases with and without diffuse lung disease, emphysema, chest tube(s), and catheter(s) in the apex. The error bars show the standard deviation.

Since the VGG19 CNN fine-tuned with full radiographs could have had the performance impacted by these conditions/devices other than in the apex, the average probabilities for images with "yes" anywhere in the radiograph (any of the six sections) and "no" are shown in Figure B.8. The distributions of PTX probability were compared using two-sample Kolmogorov-Smirnov tests with a significance level of 0.0125 after correcting for multiple comparisons using the Bonferroni method. A statistically significant difference in distributions was seen for the diffuse lung disease group versus the no diffuse lung disease group (p=0.0078).

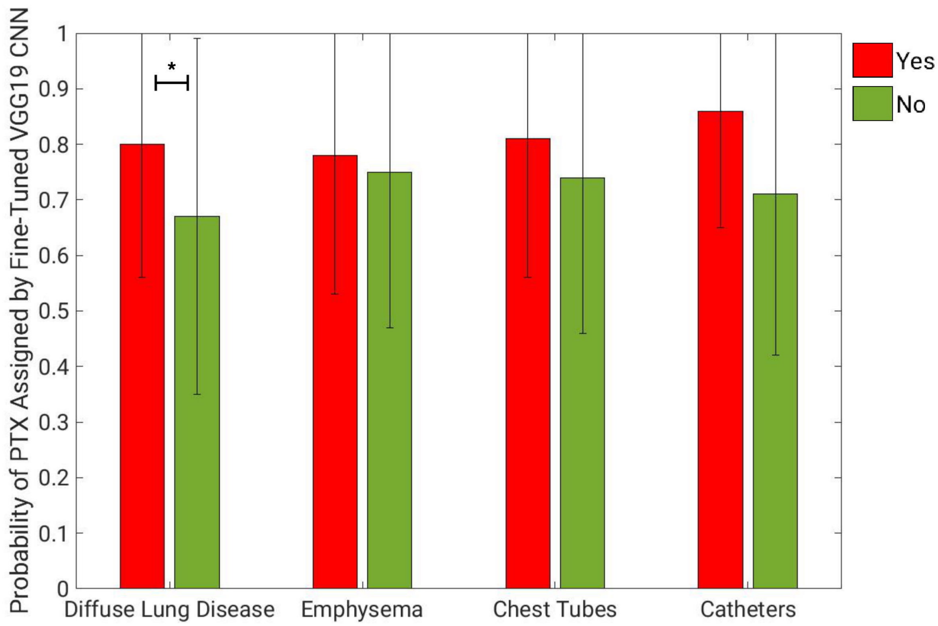


Figure B.8: The probability of PTX from the VGG19 CNN fine-tuned with full radiographs for cases with and without diffuse lung disease, emphysema, chest tube(s), and catheter(s) anywhere in the radiograph. The error bars show the standard deviation. * denotes a statistically significant difference in the distributions of PTX probability between the two groups.

The results demonstrate that other lung conditions, specifically diffuse lung disease, may impact the performance of deep learning for the detection of PTX. The apex images failed to show a statistically significant difference in PTX probability distribution for any of the groups, indicating it may be more robust to confounding diseases within the image. The annotated dataset was limited; future work could involve gathering a larger annotated dataset and repeating the analysis.

B.3 Investigation of the Impact of Resizing an Apex Image to a Square

An extra experiment was conducted where the apex images were padded with zeros; new pixels with a value of zero were added around the edge of the apex image to make them square to avoid the severe change of the aspect ratio when resizing apex images to the square matrix size of 256 x 256 (Figure 4.3). A VGG19 CNN was fine-tuned with all the same augmentation and training parameters as the apex images with the aspect ratio changed. Testing this CNN yielded an AUC of 0.864 (95% CI: 0.832, 0.892) in the task of classifying between images with and without PTX, which was found to be a statistically significant decrease in performance from the CNN fine-tuned with the apex images with their aspect ratio altered as in Figure 4.3 (0.898, 95% CI: (0.871, 0.921)). Therefore, the apex images with the aspect ratio altered were used for fine-tuning all the CNN architectures presented in Chapter 4 and were used for the comparison of performance with the full radiographs and the padded apex images.

LIST OF PUBLICATIONS AND PRESENTATIONS

Peer-Reviewed Publications

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Deep Convolutional Neural Networks in the Classification of Dual-Energy Thoracic Radiographic Views for Enhanced Workflow. J. Med. Imag. 7(1), 016501 (2020), doi: 10.1117/1.JMI.7.1.016501.

A Xiao*, **J Crosby***, H Kang, M Malin, Y Hasan, S Chumura, H Al-Hallaq. Single-institution report of setup margins of voluntary deep-inspiration breath-hold (DIBH) whole breast radiotherapy implemented with real-time surface imaging. J Appl Clin Med Phys 19(4) pg. 205-213 (2018). *Authors contributed equally to the manuscript

Proceedings Papers

J Crosby, T Rhines, F Li, H MacMahon, M Giger, "Deep learning for pneumothorax detection and localization using networks fine-tuned with multiple institutional datasets," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 113140C (16 March 2020); https://doi.org/10.1117/12.2549709

J Crosby, S Chen, F Li, H MacMahon, M Giger, "Network output visualization to uncover limitations of deep learning detection of pneumothorax," Proc. SPIE 11316, Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, 113160O (16 March 2020); https://doi.org/10.1117/12.2550066

J Crosby, T Rhines, C Duan, F Li, H MacMahon, M Giger, "Impact of imprinted labels on deep learning classification of AP and PA thoracic radiographs ," Proc. SPIE 10954, Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications, 109540E (15 March 2019); https://doi.org/10.1117/12.2513026

JD Fuhrman, **J Crosby**, R Yip, CI Henschke, DF Yankelevitz, ML Giger, "Detection and classification of coronary artery calcifications in low dose thoracic CT using deep learning," Proc. SPIE 10950, Medical Imaging 2019: Computer-Aided Diagnosis, 1095039 (13 March 2019); https://doi.org/10.1117/12.2513134

Oral Presentations

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Deep Learning for Pneumothorax Detection and Localization Using Networks Fine-tuned with Multiple Institutional Datasets. Oral presentation at SPIE Medical Imaging 2020.

J Crosby, S Chen, F Li, H MacMahon, M Giger. Network Output Visualization to Uncover Limitations of Deep Learning Detection of Pneumothorax. Oral presentation at SPIE Medical Imaging 2020.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Thoracic Radiograph Workflow to Support Machine Learning Methods for Pneumothorax Detection and Localization. Oral presentation at SIIM Conference on Machine Intelligence in Medical Imaging 2019.

T Rhines, **J Crosby**, F Li, H MacMahon, M Giger. Detection of Pneumothorax Using a U-Net Architecture on Frontal Chest Radiographs. Oral presentation at AAPM Annual Meeting 2019.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Impact of Imprinted Labels on Deep Learning Classification of AP and PA Thoracic Radiographs. Oral presentation at SPIE Medical Imaging 2019.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Multi-Class Deep Learning for Classification of Thoracic Radiographs to Enable Accurate and Efficient Workflow. Oral presentation at RSNA Annual Meeting 2018.

T Rhines, **J Crosby**, F Li, H MacMahon, M Giger. Multi-institutional Deep Network for High Performance Sorting of Over 3000 AP and PA Chest Radiographs. Oral presentation at RSNA Annual Meeting 2018.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Deep Learning in the Classification of Thoracic Radiographic Views to Enable Accurate and Efficient Clinical Workflows. Oral presentation at AAPM Annual Meeting 2018.

J Crosby, H Kang, M Malin, Y Hasan, S Chumura, H Al-Hallaq. Calculation of PTV Margins for Whole Breast DIBH Radiotherapy Using Real-Time Surface Imaging Data. Oral presentation at AAPM Annual Meeting 2017.

J Crosby, T Miller, H Li, L Lan, D Ginat, M Giger. Robustness of Radiomics on Head and Neck CTs: Intra- and Inter-Observer Contouring Effects. Oral presentation at AAPM Annual Meeting 2017.

Poster Presentations

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Application of Convolutional Neural Networks to Tasks Involving Medical Images. Invited poster presentation at Chicagoland Radiology Expo 2019.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Deep Learning for Pneumothorax Detection Using Networks Fine-Tuned with Chest Radiographs From Institutional and Publically Available Datasets. Poster presentation at AAPM Annual Meeting 2019.

J Fuhrman, **J Crosby**, C Henschke, D Yankelevitz, M Giger. Detection and Classification of Coronary Artery Calcifications in Low Dose Thoracic CT Using Deep Learning. Poster presentation at SPIE Medical Imaging 2019.

J Crosby, T Rhines, F Li, H MacMahon, M Giger. Application of Convolutional Neural Networks to Tasks Involving Medical Images. Poster Presentation at NIH NIBIB Training Grantees Meeting 2018.