Development of Model Observers for Quantitative Assessment of Mammography Image Quality

vorgelegt von
M. Sc.
Tobias Kretz

von der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
-Dr.-rer.-nat.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Benjamin Blankertz
Gutachter:	Prof. Dr. Klaus-Robert Müller
Gutachter:	Prof. Dr. Ioannis Sechopoulos
Gutachter:	Dr. Clemens Elster

Tag der wissenschaftlichen Aussprache: 01. September 2020

Berlin 2020

Abstract

Assurance of image quality has become a basic need in our society as images play a crucial role in this era of social media and digitization. Applications range from surveillance to medical imaging. Image quality is defined as the degree of clarity of its elements. Images undergo a chain of processes before being perceived by the viewer, starting from acquisition to digitization, compression, pre-processing and final presentation. Image quality expresses the extent of distortion of the original image information by various degradation processes.

In mammography, image quality specifically defines the perceptibility of clinically relevant structures, like abnormal masses and suspicious micro-calcifications, that allows one to reliably detect breast cancer at an early stage. The image quality in mammography can be assessed by analyzing recordings of technical phantoms that contain artificial lesions with various diameters and contrast levels. This way, a contrast-detail information can be determined, which expresses the ability of an observer to detect the individual lesions with prescribed accuracy. Automatic procedures are favored for objectively determining this contrast-detail information.

In this thesis, several data analysis methods for assessing image quality in mammography are developed, tested and compared to the current state-of-the-art procedure which is recommended by a European organization. The first contribution of this thesis is the implementation of a virtual mammography for the simulation of realistic images of a technical phantom. Simulated images of a known virtual specimen allow different image quality assessment procedures to be assessed and compared. In addition, virtual mammography makes it possible to set up large data sets at low cost that are required for the application of data-intensive techniques, such as deep learning.

Automatic procedures for image quality assessment in mammography usually comprise expert pre-processing and the utilization of mathematical model observers. Our second contribution is the development of a method that utilizes a recently developed parametric model observer. This way, the contrast-detail information is based on established measures of image quality. Furthermore, an uncertainty of the prediction can be estimated. Compared to the current practice, our proposed approach reduces the workload for quality assurance measurements. Even though our method needs fewer images, it still requires expert pre-processing.

Deep learning is able to avoid such expert pre-processing and to derive representative features automatically from the images. Our third contribution is the exploration of the applicability of deep learning methods for image quality assessment in mammography. Specifically, a deep learning observer is developed that estimates the contrast-detail information directly from a single image. No cumbersome pre-processing of the images is required.

Our last contribution is the explanation of predictions of the developed deep learning observer as well as the estimation of the uncertainty of its prediction. The former gives interesting insights into the learned strategy, whereas the latter expresses the confidence about our proposed model.

Zusammenfassung

Das Sichern der Bildqualität ist von enormer Bedeutung für unsere Gesellschaft, da Bilder diese Ära der sozialen Medien und Digitalisierung prägen; die Anwendungen reichen von Überwachung bis zu medizinischer Bildgebung. Die Qualität eines Bildes ist definiert als Grad der Klarheit seiner Bestandteile. Bevor Bilder von einem Beobachter wahrgenommen werden, durchlaufen sie eine Prozesskette, angefangen von der Aufnahme zu Digitalisierung, Bildkompression, Bildbearbeitung und schließlich Darstellung. Die Bildqualität beschreibt dann in welchem Ausmaß das ursprüngliche Bild durch verschiedene Artefakte gestört wird.

In der Mammographie beschreibt Bildqualität die Wahrnehmbarkeit von klinisch relevanten Strukturen wie abnormen Massen und verdächtigem Mikrokalk, deren Nachweis ermöglicht, Brustkrebs in einem frühen Stadium zu erkennen. Die Bildqualität kann durch das Analysieren von Aufnahmen technischer Phantome, die künstliche Läsionen unterschiedlicher Größe und Kontrastniveaus enthalten, bestimmt werden. Dadurch wird eine Kontrast-Detail Information ermittelt, die angibt, ob ein Beobachter die Läsionen mit ausreichender Genauigkeit erkennen kann. Favorisiert werden dabei automatische Prozeduren, die Objektivität gewährleisten.

In dieser Arbeit werden verschiedene Analysemethoden zur Bestimmung der Bildqualität in der Mammographie entwickelt, getestet und mit dem aktuellen Standard verglichen, der von einer europäischen Organisation empfohlen wird. Unser erster Beitrag ist die Implementierung einer virtuellen Mammographie, um realistische Bilder eines technischen Phanomts simulieren zu können. Simulierte Bilder eines bekannten, virtuellen Testobjekts gewährleisten Vergleichbarkeit und ermöglichen die Validierung verschiedener Verfahren. Des Weiteren können mittels virtueller Mammographie große Datensätze mit geringem Aufwand erstellt werden, die notwendig sind um datenintensive Methoden, wie *deep learning*, anzuwenden.

Typischerweise setzen sich automatische Prozeduren zur Bestimmung der Bildqualität in der Mammographie zusammen aus aufwendiger Bildbearbeitung und Anwendung eines modellbasierten Beobachters. Unser zweiter Beitrag ist die Entwicklung einer Methode, die auf einem kürzlich entwickelten parametrischen Modell basiert. Dadurch lässt sich die Kontrast-Detail Information auf etablierte Maße für die Bildqualität zurückführen. Außerdem kann eine Unsicherheit der Vorhersage angegeben werden. Verglichen mit dem gängigen Standard reduziert unser Verfahren den Aufwand von Qualitätskontrollmessungen. Auch wenn unser Verfahren weniger Bilder braucht, müssen diese zunächst aufwendig bearbeitet werden.

Deep learning ermöglicht es aufwendige Bildbearbeitung zu umgehen, indem repräsentative Merkmale direkt aus den Bildern bestimmt werden. In unserem dritten Beitrag versuchen wir herauszufinden, ob sich *deep learning* für die Bestimmung der Bildqualität in der Mammographie eignet. Ein *deep learning observer* wurde entwickelt, der die Kontrast-Detail Information direkt aus einem Einzelbild bestimmen kann. Die Bilder müssen nun nicht mehr aufwändig bearbeitet werden. Unser letzter Beitrag behandelt das Nachvollziehen spezifischer Vorhersagen unseres *deep learning observers* sowie eine Abschätzung der Unsicherheit derselben. Ersteres liefert interessante Einblicke in die erlernte Strategie, wohingegen letzteres ermöglicht abzuschätzen wie sicher die Vorhersagen unseres Modells sind.

This dissertation is dedicated to my wife Varidhi, who encouraged me to pursue my goals and kept me on track.

-

Thank you Vari for all your support, your patience and motivation.

Acknowledgement

My interest in the field of data analysis emerged because of the guidance that received during my master thesis at Justus-Liebig Universität Gießen (JLU Gießen). Recalling this, I want to thank my former supervisor Dr. Christopher Dietz, who encouraged me to focus on machine learning in my master thesis.

Three years ago, the decision to start my PhD at the Physikalisch-Technische Bundesanstalt (PTB) in Berlin was largely influenced by the same. This initiated a very exciting period in my life. During my PhD, I was extremely blessed by highly qualified, helpful and encouraging environment at PTB. I also got immense and continuous support from my professor Dr. Klaus-Robert Müller and his research group at the Technische Universität Berlin (TU Berlin). It was always motivating and relieving talking to him.

I am particularly grateful for the direction given by my main supervisor Dr. Clemens Elster from PTB. I highly appreciate his extraordinary guidance and effort invested in my work. I wish to acknowledge the assistance provided by Dr. Mathias Anton. I am thankful for many fruitful discussions with him and his innumerable remarkable advice. I would like to express my deep gratitude to Dr. Tobias Schaeffter for giving me the opportunity of doing the PhD at PTB Berlin.

My special thanks goes to my fellow researchers- Gerd, Manuel, Franko, Kathy, Jörg, Steffen, Selma, Lara, Manuel- with whom I had a great time while working. Without a doubt, I want to express my love and thanks to my family who universally supported me all along the way. Finally, I thank Varidhi, my loving wife, for her patience, her support and her vast encouragements. Without her, I would not be where I am today.

Berlin, 2020
Tobias Kretz

"Isn't it funny how day by day nothing changes but
when you look back everything is different..."
C. S. Lewis

Table of Contents

List of Figures

List of Tables

Abbreviations

Mo	Molybdenum
Rh	Rhodium
W	Tungsten
ANN	artificial neural network
AUC	area under curve
CDMAM	contrast-detail phantom for mammography
CNN	convolutional neural network
DLO	deep learning observer
ELBO	evidence lower bound
EUREF	European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services
KL	Kullback-Leibler
LRP	layerwise relevance propagation
MSOpi	modified simple observer for pooled images
PCA	principal component analysis
ReLU	rectified linear units
ROC	receiver operating characteristic
SNR	signal-to-noise-ratio
SPR	scatter-to-primary-ratio
SSO	semantic segmentation observer

Symbols

M	number of samples in a mini-batch
N	total number of training samples
$\boldsymbol{W}^k$	matrix of weights in layer k with $\boldsymbol{W}^k \in \mathbb{R}^{n_k \times n_{k-1}}$
$\boldsymbol{\alpha}^k$	output of a fully-connected layer k with $\boldsymbol{\alpha}^k \in \mathbb{R}^{n_k \times 1}$ and $\boldsymbol{\alpha}^k = \sigma\left(\sum \boldsymbol{W}^k \times \boldsymbol{\alpha}^{k-1} + \boldsymbol{b}^k\right)$
$\boldsymbol{b}^k$	vector of biases in layer k with $\boldsymbol{b}^k \in \mathbb{R}^{n_k \times 1}$
$\hat{y}_i$	neural network prediction for input x_i
$\mathbb{E}_q\left[f(x)\right]$	expectation of a function $f(x)$ with respect to the distribution $q(x)$, given as $\mathbb{E}_q\left[f(x)\right] = \int q(x) f(x) dx$
$\mathcal{D}$	data consisting of input output pairs.
$\mathcal{H}(x)$	Heavyside function with $\mathcal{H}(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$
$\mathcal{L}(y, \hat{y})$	loss function that measures the deviation of the output y and the predicted output $\hat{y} = \psi(x, \theta)$
$\phi(x)$	cumulative distribution of standard normal distribution
$\psi(x, \theta)$	model with input x and parameters θ
$\sigma(x)$	(non-)linear activation function
θ	variable to describe all learnable parameters of a model
n_k	number of neurons in a fully connected layer k
x_i	input sample, e.g. image
y_i	(ground truth) label for input x_i

1
Introduction

1.1 Motivation

According to the Word Health Organization (WHO), breast cancer is the leading cancer among women worldwide (WHO, 2014). In 2018, worldwide 54.4 breast cancer cases were reported per 100, 000 women in highly developed countries while less developed countries were shown to have 31.3 breast cancer cases per 100, 000 women (Bray et al., 2018). The overall mortality rate in highly developed countries in 2018 was 21% (Bray et al., 2018). In less developed countries breast cancer is usually detected in an advanced stage (Anderson et al., 2003) leading to a higher mortality of around 47.6% in 2018 (Bray et al., 2018).

Obviously, the survival rate depends on the stage in which cancer is diagnosed (Richards, 2009). Cancer that is diagnosed in early stages is more likely to be treated successfully (Hiom, 2015). For this purpose, early detection of breast cancer in asymptomatic women plays an important role in reducing breast cancer mortality (Kösters and Gøtzsche, 2003). Once an abnormality is detected, it can be accurately diagnosed via biopsy, where a part of the abnormal structure is extracted and examined in a laboratory (Veronesi et al., 1997).

One major criticism of early detection is its tendency of overdiagnosis. Overdiagnosis in general describes the diagnosis of a disease that would not cause symptoms or cause deaths (Welch and Black, 2010). In breast cancer, overdiagnosis can lead to unnecessary surgeries and psychological stress of patients that would not develop symptoms in their lifetime or die from breast cancer due to a rather slowly progressing cancer type (Paci and Duffy, 2005). Nevertheless the WHO recommends the use of early detection (WHO, 2014). Especially mammography, where the breast is imaged with low energy X-Rays (ICRU, 2009), has shown evidence for early detection (Oeffinger et al., 2015). For this purpose, mammography screening programs are implemented in many countries all over the world (Althuis et al., 2005). As a consequence, efficiency studies, like the Swedish randomized trials (Nyström et al., 2002), report a reduction of breast cancer mortality of around 21% for woman. However, some other studies such as the Canadian randomized trial report no significant decrease of mortality by mammography screening (Miller et al., 2014).

Another objection that may be raised is that by screening a predominantly healthy population the disadvantages outweigh the benefits of mammography screening (Gøtzsche and Olsen, 2000; Gøtzsche and Jørgensen, 2013). Not only is the use of X-Ray radiation connected with the risk of inducing cancer (Colin et al., 2011; Yaffe and Mainprize, 2011), especially when the patients belong to a high risk group for breast cancer (Nadine et al., 2006), but falsely positive diagnosed cases result in unnecessary operations and stress to the patients (Jørgensen and Gøtzsche, 2009; Biller-Andorno and Jüni, 2014). The latter may be addressed by reducing the frequency of medical examinations from annual to biennial while conserving the benefits from significant mortality reduction (Mandelblatt et al., 2009). The examinations employing X-Ray radiation are always connected with the risk of inducing cancer. The absorbed dose is taken as a metric to quantify this risk. A higher radiation dose increases this risk but at the same time increases the signal-to-noise-ratio (SNR), making it more likely to detect the abnormalities (Boone et al., 2001). Thereby, a higher radiation dose improves the quality of the image.

Consequently, mammography poses the problem of trading off radiation dose with the achieved image quality. The ultimate goal is to keep the radiation dose low while maintaining sufficient image quality.

In mammography the term image quality is usually defined as the ability of a device to depict radiologically significant details (ICRU, 2009). In other words, image quality evaluates how likely breast cancer can be detected in the presented image. A higher image quality indicates a higher success rate for cancer detection. More general, image quality describes the visual effect of an image (Wang et al., 2002). Image acquisition, processing and representation introduce image distortions that decrease the obtained visual performance measured by observers (Liu et al., 2019). Accurate image quality assessment methods can be divided into methods where a reference image is available and methods where no reference image is available (Bosse et al., 2017). In both cases, a metric has to be chosen to quantify image quality. Full-reference image quality assessment commonly measures the deviation to the reference image (Bosse et al., 2017). The more an image deviates from a reference, the lower is the image quality. No-reference image quality assessment aims to determine subjective quality measures like the mean opinion score (Wang et al., 2002) or objective image quality metrics based on modelling parameters of visual perception (Liu et al., 2019).

For mammography, no reference images are available but image quality can be technically assessed by recording images of technical phantoms (Perry et al., 2006). One type of such technical phantoms comprise several details of different size and contrast embedded in a uniform background. A suitable measure of image quality is an observer's performance to successfully detect these lesions in a recorded image. These observers can be humans as well as mathematical model observers. Employing humans to measure image quality is time consuming and yields highly subjective scores that vary significantly between different observers (Karssemeijer and Thijssen, 1996). Mathematical model observers, on the other hand, enable fast and objective measurement of image quality. Such model observers define a metric that enables to distinguish features relevant for certain tasks, e.g. lesion detection.

It has been shown that the assessment of image quality using technical phantoms is linked to the detection of calcifications in clinical images (Warren et al., 2014; Mackenzie et al., 2016). Here, the targets that serve as a surrogate for the calcifications, are embedded in the anthropomorphic structure of the breast.

Technical estimations of image quality depend on the method to analyze images as well as on the manufacturing performance of the technical phantoms (Fabiszewska et al., 2016). It has been reported that the reproducibility of the targets in technical phantoms can vary up to 10% (Borowski et al., 2012) leading to a variance of the determined image quality of the same device with different phantoms of the same type (Fabiszewska et al., 2016). For the purpose of standardized image quality measurement, such a behavior is undesired and procedures to robustly estimate the image quality are required. Current procedures to automatically estimate image quality rely on the exact knowledge of the signature induced by these targets. In addition, sophisticated pre-processing is required to derive image quality estimates and many images are needed for a reliable determination (Mackenzie et al., 2017).

Deep learning could be an alternative for robust estimation of image quality in mammography without further pre-processing. Deep learning comprises methods where large models are trained to solve a specific task. It has been shown that deep learning can be used as a tool for computer aided detection of lesions in mammography with a performance comparable to expert radiologists (e.g. Kooi et al. (2017)). Furthermore, deep learning has been proposed for quantifying the quality of visual perception of images in the context of image quality assessment (e.g. Hou et al. (2014); Bosse et al. (2017)). Even though these studies prove the benefit of using deep learning for image quality assessment, it has not been used to assess mammography image quality so far.

In this thesis, different methods for mammography image quality analysis will be developed that comprise conventional approaches as well as deep learning. A conventional method utilizing a recent parametric mathematical model observer determines contrast-detail with fewer images than the current practice. Furthermore, the derived quantities are accompanied with an estimation of the uncertainty, which is crucial for reliable conformity assessment. In addition, a deep learning oberserver shall be developed and its potential benefit shall be explored. Virtual mammography is implemented and used to create a data base for training a deep learning observer as well as test data bases for assessing its performance in comparison with the current practice. Finally, methods of interpretability and uncertainty for deep learning are applied to explain the deep learning observer's predictions and to estimate the accompanied uncertainty of these predictions.

1.2 Objective and Aims

The goal of this work is to develop precise and reproducible procedures to improve and ease the estimation of image quality in mammography from images of a technical phantom. Specifically, different methods are developed and assessed in terms of their performance and uncertainty. The main objectives of this thesis can be summarized as follows:

- **Objective 1:** ***Implementation of a virtual mammography to simulate radiographic images.*** Virtual mammography is a tool for simulating realistic mammography images and can be used for method assessment and to develop data-intensive analysis procedures. We aim to explore state-of-the-art procedures to implement such a virtual mammography. This simulation tool shall then be used to simulate mammography images of a technical phantom to develop and asses methods that determine the image quality. Furthermore, the virtual mammography will be used to set up a large data base to explore the benefits of the application of deep learning, as well as of other methods, for image quality assessment.

- **Objective 2:** ***Application of a parametric model observer for mammography image quality assessment.*** Mammography image quality can be assessed by measuring the performance of an observer to detect structure details in images of a technical phantom. Mathematical observers assess image quality objectively but current procedures do not include estimations of the uncertainties of predictions. We aim to develop a procedure for mammography image quality assessment that utilizes a parametric model observer. For a reliable decision whether a device produces images of sufficient quality, an uncertainty has to be determined. Therefore, we also aim to quantify the uncertainty that is accompanied with the results of our method. The virtual mammography will be used to assess our proposed approach and to compare it with a reference method that is suggested in the EUREF guideline for mammography quality assurance.

- **Objective 3:** ***Development of deep learning methods for mammography image quality assessment.*** Common automatic techniques to determine the image quality in mammography require error prone pre-processing of recorded images and depend critically on the manufacturing accuracy of technical phantoms. In this thesis, different approaches to apply deep learning for mammography image quality assessment shall be explored. The virtual mammography will be used to simulate large data sets to facilitate the training of data-intensive deep learning models. In particular, it shall be investigated, if deep learning can be applied such that the contrast-detail information is determined directly from the images of the CDMAM phantom. This way, image quality can be determined without pre-processing. Our proposed method will be compared to the current automatic procedure recommended in the EUREF guideline.

- **Objective 4:** ***Estimation of the uncertainty and analysis of the explainability of the developed deep learning observer.*** Neural networks are usually designed to predict point estimates of some values of interest and can be viewed as black box solutions for specific problems. Recent progress in the field of the interpretability of deep learning models makes robust explanation of deep learning predictions possible, especially for classification problems. Such explanations give insights why the model ends up with a particular prediction. We aim to explore recent explanation methods and apply them to our developed deep learning procedure.

  In addition to interpretability, growing interest towards methods that aim for an assessment of the predictive uncertainty of deep learning initiated a vast amount of research resulting in a variety of different methods. Here, we will review different methods for the estimation of the predictive uncertainty and apply them to our deep learning method. A benchmark shall be proposed to compare different uncertainty methods in the context of multivariate regression.

1.3 Thesis Structure and Overview

Chapter 2 (Fundamentals) provides a brief overview of the background information for the topics mammography and machine learning. The fundamental imaging physics are briefly reviewed that are needed to set up virtual mammography for realistic image simulation. The importance of mammography image quality assessment is explained along with a description of the protocol for image quality assurance in mammography according to the current European guideline. Furthermore, machine learning is introduced as a tool to automatically derive methods that learn from data to solve a certain task. Specifically, neural networks are described along with an algorithm how they can be trained efficiently. The probabilistic interpretation, that can be used to model predictive uncertainties, concludes the brief review of machine learning.

Chapter 3 (Virtual mammography) is dedicated to the description of our implemented virtual mammography as a framework to simulate realistic mammography images. In addition, an alternative method for image quality assessment in mammography is developed that employs a recently published parametric model observer. Virtual mammography is used to assess our proposed approach and to compare it with the current practice. The limitations and benefits of our novel approach are discussed. Furthermore, virtual mammography is an essential tool for the development of deep learning approaches for image quality estimation. It is used to develop a large data set to facilitate the training of deep learning methods.

Chapter 4 (Mammography Image Quality Assurance Using Deep Learning) describes the developments of methods to assess image quality in mammography using deep learning. Two methods that follow two different strategies are described. As a result, a semantic segmentation observer (SSO) and a deep learning observer (DLO) are presented. The SSO estimates a global figure of merit for image quality and is therefore difficult to compare with the current practice that uses local contrast information. Our DLO, on the other hand, can extract the local contrast information and is therefore compared with the current practice. A sparse sampling approach further ensures that our method is robust against variations of the phantom due to deviations in the manufacturing process of technical phantoms.

Chapter 5 (Explainability and Uncertainty of the deep learning observer) is dedicated to the interpretability as well as to the estimation of uncertainty of our newly developed DLO. Interpretability aims to open the black box nature of deep learning by providing an explanation why a model yielded its prediction for a given input. Different explanation methods are reviewed and applied to our developed DLO. This leads to interesting insights in how the observer determines image quality. Growing interest in estimating the predictive uncertainty resulted in the development of different models. For a simplified example, it is described, how state of the art methods can be easily modified to predict point estimates along with an associated uncertainty. Different approaches are explained and the corresponding results are compared. A benchmark test is proposed that allows one to recommend a method and a strategy to estimate the uncertainty for mammography image quality assessment using deep learning. These findings are finally applied to determine the uncertainty of contrast-detail curves derived by our DLO.

1.4 List of Publications

At this point, I also want to thank my co-authors for the permission to use the content of our publications for my thesis.
During my PhD studies, the following work has been published.

Peer-reviewed journals

[1] Kretz, T., Müller, K.R., Schaeffter, T. and Elster, C., (2020). Mammography Image Quality Assurance Using Deep Learning. *IEEE Transactions on Biomedical Engineering*, https://doi.org/10.1109/TBME.2020.2983539.

[2] Kretz, T., Anton, M., Schaeffter, T. and Elster, C., (2019). Determination of contrast-detail curves in mammography image quality assessment by a parametric model observer. *Physica Medica*, 62, pp.120-128.

[3] Anton, M., Khanin, A., Kretz, T., Reginatto, M. and Elster, C., (2018). A simple parametric model observer for quality assurance in computer tomography. *Physics in Medicine & Biology*, 63(7), p.075011.

[4] Dietz, C., Kretz, T., Thoma, M. H., (2017). Machine-learning approach for local classification of crystalline structures in multiphase systems. *Physical Review E* 96(1), p.011301.

Conferences/Workshops

[5] Kretz, T., Anton, M., Elster, C., (2018). Image quality and quality assurance in mammography using mathematical model observers. In: *49th Annual Conference of the German society for medical physics (DGMP).*

[6] Kretz, T., Anton, M., Khanin, A., Reginatto, M., Elster, C., (2018). Virtual Mammography. In: *VirtMess 2018: Metrology for virtual measuring devices.*

Further publications

[7] Kretz, T., (2019). Virtual mammography. *PTB-News 3.2019.*

[8] Kretz, T., Elster, C., (2019). Using machine learning for quality assurance in mammography. *PTB Annual Report 2019 - News of the Year.*

2 Fundamentals

This chapter provides relevant fundamentals before immersing into the detailed description of our contributions. For better understanding, this chapter is split into two parts starting with the fundamentals of mammography, followed by machine learning. More specifically, an introduction to mammography and its imaging physics is given in the beginning. A definition of the term image quality in the context of mammography is provided, followed by a description of the current practice to estimate image quality that will be used as a reference method in this work.

Image quality estimation methods can be assessed in general by virtual mammography, that realistically simulates real mammography image recordings. Realistic simulated images are also useful to generate large data sets and hence virtual mammography supports the development of deep learning approaches. For a realistic simulation of mammography images the image formation process needs to be modeled. Thus, a brief review of the image formation process is provided in the mammography section. A detailed review is out of scope for this work and the interested reader is referred to the dedicated literature (e.g. Russo (2017); Barrett and Myers (2013)).

The second part of this chapter focuses on a brief description of machine learning. Artificial neural networks (ANN) are reviewed as a powerful tool in machine learning to solve a variety of tasks. Such complex models can be trained with the back-propagation algorithm by using reference data. Again, the interested reader is referred to the dedicated reference literature for more detailed information (e.g. Bishop (2006); Mitchell (1997); Orr and Müller (1998)).

The fundamental descriptions of mammography and image quality assessment are mainly important for chapter 3 whereas the basics of machine learning are used in chapters 4 and 5 to develop and assess deep learning applications for mammography image quality assessment.

2.1 Mammography

The term mammography describes a radiography technique where the breast tissue is exposed to X-Ray radiation to look for abnormalities in the imaged breast. Abnormalities in the breast tissue include large, low-contrast (cancerous) masses and small high-contrast micro-calcifications (ICRU, 2009). The goal of mammography is to record an image of the patient breast such that these abnormalities can be diagnosed by a radiologist. In this section a brief overview of the underlying imaging physics of mammography shall be given.

The mammography imaging process can be roughly split into three different stages that are the production, the interaction with matter and the detection of transmitted X-Ray photons. A detailed description can be found in e.g. Russo (2017); Barrett and Myers (2013); ICRU (2009), since the description of the underlying physics is reduced to those processes that are necessary for the development of a computer model of mammography that produces realistic simulated images.

Production of X-Rays

X-Rays are a form of high energy electromagnetic radiation, that can be found naturally in our universe and can be produced artificially by X-Ray tubes (Barrett and Myers, 2013). The X-Ray tube consists of a vacuum tube in which electrons are emitted from a cathode and accelerated in an electric field. The resulting high energy electrons are then used to bombard a metal anode. As part of the interaction of the electrons with the atoms of the metal anode X-Rays are created (Russo, 2017).

The interaction of high energy electrons with the atoms of the target material generates two kinds of X-Rays: Bremsstrahlung and characteristic X-Rays (Barrett and Myers, 2013). Bremsstrahlung occurs when the negatively charged electrons are decelerated in the electric field of the positively charged protons in the nucleus of the target atoms (Koch and Motz, 1959). Bremsstrahlung exhibits a continuous spectrum (Reed, 2005) which means that the electrons can lose between 0% and 100% of their initial kinetic energy. The energy loss is then emitted as X-Ray photons with corresponding photon energy. If the impinging electron has sufficient energy, it can hit an inner orbital electron of the atoms of the target material causing its emission. The resulting vacancy then causes electrons from higher energy levels to be transferred into the inner shells. If an electron jumps from a outer shell to an inner shell, the difference of the corresponding energy levels is emitted as a photon (Archer and Wagner, 1988). This effect produces an characteristic spectrum with many X-Ray photons of few discrete energies. The spectrum produced by an X-Ray tube combines characteristic radiation and Bremsstrahlung into a single spectrum. This depends on the target material, the tube voltage that accelerates the electrons and the number of free electrons, which is proportional to the time current product.

For mammography, soft X-Rays with low energies between 15 keV and 30 keV achieve a high contrast between normal breast tissue and abnormal masses (Russo, 2017). Molybdenum (*Mo*), Rhodium (*Rh*) or Tungsten (W) are commonly used as target materials for tube anodes in mammography (ICRU, 2009). Using *Mo* as an anode material produces a spectrum with characteristic X-Rays at 17.9 keV and 19.5 keV which is optimal for small breasts (Russo, 2017). *Rh*, on the other hand, produces characteristic X-Rays at 20.2 keV and 22.7 keV which are more eligible for the imaging of dense breasts, due to the higher energy (Russo, 2017). Incorporating W as an anode material yields X-Rays at very low energies between 8 keV and 10 keV.

Table 2.1: Typical anode material/filter combinations to generate X-Ray spectra with different beam qualities to optimally image different breast sizes in digital mammography.

Anode material	Filter material	Filter thickness [μm]
Mo	Mo	0.030
Mo	Rh	0.025
Rh	Rh	0.025
W	Rh	0.050

Especially for Mo and Rh anodes, the additional continuous spectrum generated by the Bremsstrahlung is typically undesired since it adds unnecessary radiation dose while reducing the contrast between healthy and malignous tissue (Russo, 2017). Filters are frequently used to minimize the contribution of the continuous spectrum . For Mo and Rh anodes, the filters are typically made of the same materials as the anode, while for W anodes, Rh filters are commonly used in order to suppress the contribution of the low energetic characteristic peak in the W spectrum. Table 2.1 summarizes typical anode material and filter combinations used in digital mammography.

Figure 2.1 shows a typical spectrum from a W anode (blue) and a Mo anode (orange) for a tube voltage of 28 kV. The unfiltered spectra shown in Figure 2.1a comprise characteristic radiation and the continuous Bremsstrahlung. The undesired continuous radiation can be filtered out of the spectrum as shown in Figure 2.1b using typical anode material filter combinations which are listed in Table 2.1. The design of the X-Ray tube influences the image quality or, more precisely, the spatial resolution (Huda et al., 2003). This occurs because the initial electron beam that targets the anode is spatially distributed. Hence, different electrons hit the anode in a slightly different position. Therefore, the different X-Ray photons are not generated at the exact same position. The area of the anode that is targeted by the incoming electron beam is often referred to as the focal spot (Curry et al., 1990). This focal spot is the area from which

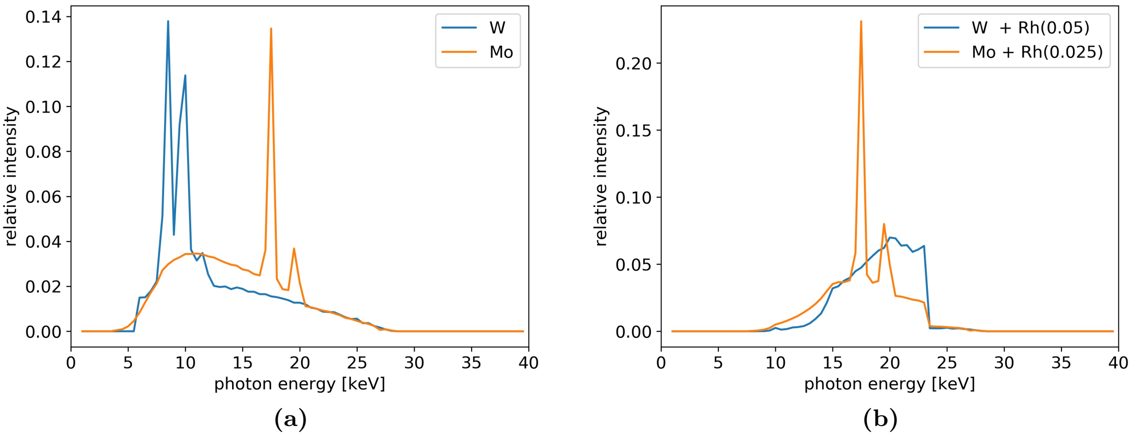


Figure 2.1: Typical normalized X-Ray spectra for breast cancer detection in mammography for a tube voltage of 28 kV. (a) Unfiltered spectra for a Tungsten (W) and a Molybdenum (Mo) anode comprise characteristic radiation (peaks) and a continuous part. (b) Achieved X-Ray spectra for W and Mo after filtering the undesired continuous part using typical anode filter combinations (cf. Table 2.1).

X-Ray photons are emitted. The smaller the focal spot size, the better the resulting spatial resolution of the system, which contributes to the image quality that can be achieved with the corresponding mammography device (Curry et al., 1990).

Interaction of Photons with Matter

When photons pass through an object, they can either penetrate the object without any interaction or interact with the atoms of the target by depositing all or a part of their kinetic energy. The interaction of photons with the atoms of the target can be divided into photoelectric absorption, Compton scattering and coherent scattering (Tipler and Mosca, 2007). Theoretically, a photon that is near the nucleus of an atom, can also interact via pair production if its kinetic energy exceeds 1.02 MeV (Ashok and Thomas, 2003). However, typical photon energies of X-Ray radiation fall in the range 100 eV to 200 keV (Seibert, 2004) where photoelectric absorption and Compton scattering are the predominant processes (Seibert and Boone, 2005).

Rather than understanding the interaction of single photons with matter, one is more interested how a beam that comprises many X-Ray photons interacts with matter. Some of the photons in the beam will penetrate the target without interaction, while others will interact and thus be removed from the beam via interaction. Hence, the initial beam is attenuated by the target. The Beer-Lambert law describes an exponential decrease of the intensity if the beam passes through a homogeneous material:

$$N(E) = N_0(E)e^{-\int_0^x \mu(E,x)\cdot dx}\,, \tag{2.1}$$

where $N_0(E)$ describes the initial intensity, x denotes the thickness of the material and $\mu(E,x)$ denotes the energy dependent attenuation coefficient.

The human breast is composed of skin, adipose and glandular tissue. Micro-calcifications consist of denser matter, whereas masses consist of soft materials with attenuation properties similar to glandular tissue (Russo, 2017). The differences of the linear attenuation coefficients of the soft composites of the breast tissue tend to increase for decreasing energy. For very low energies, photoelectric absorption is the predominant process of photon interaction. The Compton scattering is more likely to occur with increasing energy. At energies above 25 keV, the Compton scattering is more likely than the photoelectric absorption. This incoherent scattering reduces the energy of X-Ray photons and changes their path such that they reach the detector in a slightly different angle or even at a different position. This undesired effect reduces the contrast of the image. Therefore, mammography primarily uses anode materials that yield many low energetic X-Ray photons like Mo, Rh and W.

Detection of photons

The goal of mammography is to translate the locally different attenuation properties of an object into an image such that these differences can be perceived as contrasts. After penetrating the target, the X-Ray radiation has to be transformed into an image. This can be done by either transforming the incoming X-Ray radiation into light or by direct transformation to an electric signal (Barrett and Myers, 2013).

The detection of X-Ray photons is more difficult as they have higher energies than visible light, so they are more likely to penetrate through thin materials. However, fluorescent materials such as cesium iodide (CsI) can be used to transform the incoming X-Ray radiation into photons with lower energy that can be detected either with film-screen methods or with CMOS cameras (Noel and Thibault, 2004). The number of low-energy photons induced by an incoming X-Ray beam depends on the energy distribution of the X-Ray photons in the beam and on the number of photons per energy.

Film-screen detectors use a chemical reaction of silver atoms to locally record the photon intensity and the developed film can then be presented on an illuminated screen (Del Guerra, 2004). In digital mammography, fluorescent layers are directly combined with CMOS cameras that translate the intensity of optical photons into an electric signal that can be further digitized into pixel intensities (Spahn, 2005). Direct digital mammography allows to convert the incoming X-Ray radiation directly into an electric signal. Such direct conversion detectors consist of photo-conductive layers (e.g. amorphous Selenium) where the incoming X-Ray photons generate electron-hole pairs (Russo, 2017). The number of produced primary electrons is determined by the spectrum of the incoming X-Rays. For any conversion procedure, the efficiency of the detection depends on the dimension of the reactive layer as well as on the material properties and can be expressed as the detector quantum efficiency (Yaffe and Rowlands, 1997)

$$\eta\left(E\right) = 1 - e^{-\int_0^x \mu(E,x)\cdot dx} , \tag{2.2}$$

where μ denotes the attenuation coefficient of the conversion material and x denotes the thickness of the conversion layer.

In the detection stage, image quality is mainly influenced by noise. The total variance of the pixel values in the raw image contains contributions of quantum noise, electronic noise and additional spatial noise (Burgess, 2004). Electronic noise, such as shot noise, is independent of the incoming radiation while quantum noise and additional spatial noise are directly correlated with the incoming exposure (Burgess, 2004). At this point, we do not aim for an extensive analysis of noise and its characteristics but refer the interested reader to dedicated literature (e.g. Pisano and Yaffe (2005); Whitman and Haygood (2012)).

In summary, the underlying principle physics in mammography can be divided into the three stages production, interaction and detection of X-Ray photons. In the same way, it is necessary to model these stages for realistic simulation of mammography images. Each of the three stages contributes substantially to the limitations of the resulting image quality. The main contribution during the X-Ray production comes from blurring that occurs if the focal spot is not a point source. In the interaction stage, scattering becomes the dominant process for photons with energies higher than 25 keV. In x-ray projection imaging, including mammography, scattering adds to the quantum noise and thus reduces the resulting contrast to noise ratio and hence image quality. Finally, in the detection stage detector noise and spatial resolution substantially degrades the image quality of the recorded image.

Besides mammography, other imaging modalities can also be used to diagnose breast cancer at an early stage. Mammography as a 2D planar imaging modality can be extended to breast tomosynthesis by combining several 2D planar images from different angles to produce a pseudo-3D image (Niklason et al., 1997). Besides that, breast Magnetic Resonance

Imaging (MRI) and breast Computer Tomography (CT) are modalities especially designed to image the human breast for early diagnosis of breast cancer (ICRU, 2009). Also infrared (IR) spectroscopy (Obrig et al., 2000) was successfully used to diagnose breast cancer (Pogue et al., 2001). Breast MRI, breast tomosynthesis, IR spectroscopy, and breast CT can improve the accuracy of the detection of abnormalities, especially in dense breasts, but a nation wide screening using these modalities would be too expensive (e.g. Gilbert et al. (2016); Plevritis et al. (2006)). So far, mammography is the only imaging modality with proven evidence to affect the mortality by early detection of breast cancers. A thorough image quality assurance is necessary to ensure that mammography devices produce images of sufficient quality. This has to be done so that the low contrast lesions and small micro-calcification can be perceived and diagnosed reliably in recorded images.

2.1.1 Image Quality Assessment

In general, the term image quality describes the quality of human perception of images (Wang et al., 2002). Before an image is perceived by a viewer, it undergoes several steps starting from acquisition to digitization or compression and finally presentation. Each of these procedures can induce distortions that later affect the visual perception of the image. Methods that assess image quality can be divided in two groups, depending on the availability of a reference image (Bosse et al., 2017). Full-reference image quality assessment allows one to determine the effects of image processing, such as compression, by comparing the final image with a prior reference image. The closer the final image resembles the reference image, the better the image quality. Common measures in full reference image quality assessment are the mean standard error (MSE) or the structural similarity index (SSIM) (Bosse et al., 2017). In contrast, no-reference image quality assessment aims to assess image quality without further requirement of reference images. Here, image quality can be assessed in terms of subjective scores, such as the mean opinion score of viewers (Liu et al., 2019), or objective measures like the model parameters of visual perception from image statistics (Wang et al., 2002).

In mammography, the term image quality comprises the performance of a device to depict clinically relevant details in a recorded image (Haus and Yaffe, 2000). Such details include micro-calcifications and masses. A micro-calcification is usually shown as a small detail with higher contrast, whereas abnormal masses are larger in dimension, but lower in contrast (ICRU, 2009). Image quality in mammography can be estimated in a no-reference manner by recording images of technical phantoms and reporting an observer's performance on a specific task, such as lesion detection (Perry et al., 2006). Such observers can be either humans or mathematical model observers. Mathematical model observers make a fast and objective assessment of image quality possible, whereas human observer studies are highly subjective and time consuming. Model observers estimate image statistics that are specific to a certain task (Barrett et al., 1993). For image quality estimation in mammography this task is lesion detection. To be more specific, the observer determines whether a lesion is present in an image or not.

In general, mammography image quality is related to a set of certain technical parameters. In practice, radiologists classify images as inadequate quality if the image has low measured values of radiographic contrast, signal-to-noise-ratio (SNR) and spatial resolution as well as high presence of artifacts in the image (Haus and Yaffe, 2000). A clear understanding of how the individual parameters contribute to the image quality is still missing(ICRU, 2009). Furthermore, these technical parameters depend on the acquisition parameters and of course they are task dependent: a set of acquisition parameters well suited to depict a small, high contrast micro-calcification might not be optimal to visualize low contrast masses in the image (Huda et al., 2003). Besides, determining image quality by measuring these technical parameters, such as the modulation transfer function (MTF) or the noise power spectrum (NPS) requires the validity of the assumption of the linear system theory (Chen, 1998) (cf. Anton et al. (2018)). However, nonlinear image processing violates these assumptions. Therefore, task-based image quality assessment with model observers are usually preferred (Anton et al., 2018).

Recalling that mammography image quality is interpreted as the ability of an observer to perceive clinical relevant details in the image, the receiver operating characteristic (ROC) curve methodology can be used as a tool to infer the observer performance with the mammography device (Metz, 1979).

The ROC analysis provides a measure of the performance for a binary classification task, such as the presence or absence of abnormalities in a medical image(Obuchowski, 2003). For each image, a numeric test statistic is computed. This test statistic varies for different images. It follows that different images with and without lesions (i.e. the signature of a target) can be described by the two distributions of the numeric test statistic for both classes. Figure 2.2 visualizes an example ROC analysis. The Figure shows the distributions of the test statistic of the two classes for a binary task in two different cases. As it can be seen, the overlap between the two class distributions is reduced in case 2. Figure 2.2 shows the corresponding ROC curves for the two cases along with the ideal and the random performance. Ideally, both class distributions do not overlap, which means that the observer can perfectly distinguish between both classes. The random performance defines the scenario where the distribution of the test

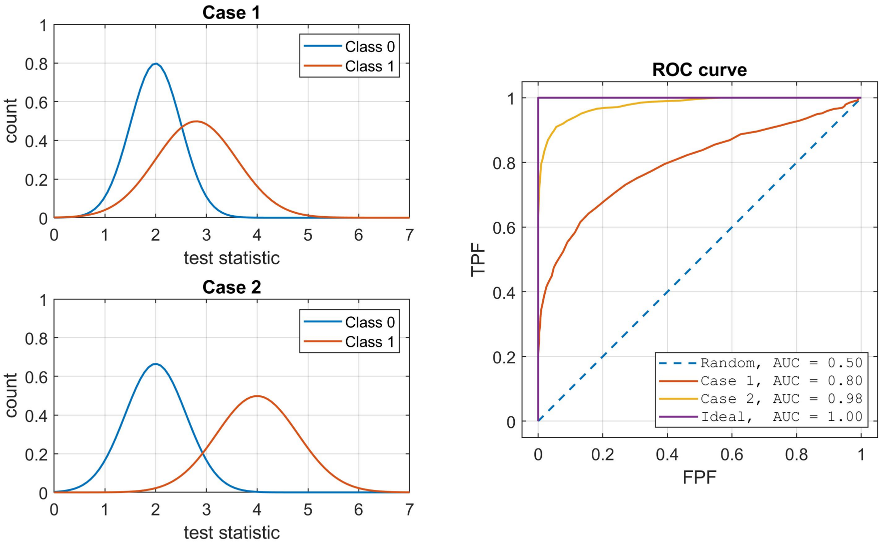


Figure 2.2: Example ROC study for a binary task where the class samples are distributed with respect to a test statistic as shown on the left for two different cases. On the right the corresponding ROC curves are shown together with the ideal scenario without overlap of the different class distributions and the random case, where the class distributions for both classes are equal.

statistic of both classes overlaps completely, which means that the observer cannot distinguish between the different classes. The ROC curve plots the true positive fraction (TPF) against the false positive fraction (FPF). Both parameters, TPF and FPF, are calculated as a function of a threshold for the test statistic. TPF and FPF are the part of the distribution of class 1 and class 0 that are greater than the chosen threshold. The area under curve (AUC) (Hanley and McNeil, 1983) is the integrated area under the ROC curve and is useful as a metric for the image quality. The AUC is always in the range between the random performance ($AUC = 0.5$) and the ideal performance ($AUC = 1$).

Implementing meaningful ROC studies, however, requires many images to be analyzed. Only with a sufficient number of images the distribution of the test statistic can be inferred reliably. ROC studies are complex and time consuming and hence are often not very practical (ICRU, 2009). Instead, image quality can be expressed by system measurements or observer properties. Objective measures of the system properties determine technical parameters like noise, spatial resolution, contrast and artifacts and are shown to correlate with observer-dependent contrast-detail measurements (Marshall, 2006). Contrast-detail measurements employ test patterns that contain different structures that vary in their dimension (e.g. diameter of circular discs) and their radiation contrast, given by the structure thickness (ICRU, 2009). Images of this test phantom are then viewed by observers to decide which structures are visible. The observer decision can then be translated into contrast-detail curves that display the corresponding minimum structure thickness for each structure diameter such that an employed observer can detect this structure with prescribed accuracy (Loo et al., 1983).

In Europe, the European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services (EUREF) developed a guideline on how to assess the image quality in mammography (Perry et al., 2006). Rather than measuring the system parameters directly, the image quality is evaluated in terms of contrast-detail measurements. This can either be done by employing a radiologist or by an automatic procedure. By annually assessing the image quality of mammography devices in terms of contrast-detail measurements, their depreciation can be tracked and a recommendation of when certain units have to be replaced can be given. Devices from different vendors can be compared and new devices can be checked whether they produce images of sufficient quality for breast cancer diagnosis.

2.1.2 CDMAM Phantom

Image quality assurance in mammography is important to ensure that (new) devices produce images with a sufficient contrast resolution in order to reliably depict abnormalities in patient images (acceptance test). Furthermore, image quality measurements are carried out to test whether devices that were already approved once, still reach the sufficient image quality prerequisites (constancy test). Therefore, image quality measurements need to be reproducible and consistent. For this purpose, usually technical phantoms are developed to be used for the assessment of the diagnostic image quality. Guidelines were developed by the American College of Radiology as well as by a European commission to standardize quality assurance. They particularly describe which phantoms should be used and how image quality can be assessed (Li et al., 2010).

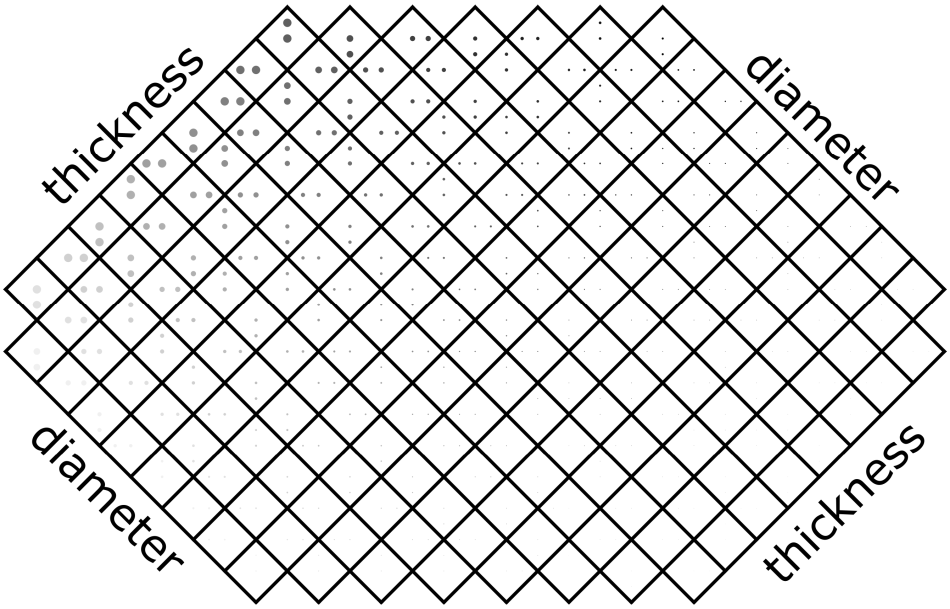


Figure 2.3: Schematic representation of the CDMAM phantom for mammography image quality assessment according to European guidelines. The different thicknesses are indicated as different gray levels. Figure from Kretz et al. (2019).

In this work, the focus is on the European guideline that recommends to use the contrast-detail phantom for mammography (CDMAM) phantom (version 3.4, Artinis) to evaluate the imaging performance of a device. The CDMAM phantom comprises 205 individual square cells arranged in a regular grid pattern as illustrated in Figure 2.3. The individual cells are separated from each other by solid thick grid lines. The choice of different materials with different attenuation properties cause different transmission of X-Ray radiation and hence, varying intensities in the recorded image. The CDMAM phantom consists of a 16 $cm \times 24$ cm aluminium plate with a thickness of $0,5$ mm, pure gold discs and a polymethyl methacrylate (PMMA) cover (Thijssen et al., 2007).

Within each cell, two gold disc are located: one in the center of the cell and one randomly located in one of the four corners. The two gold discs in one cell have the same diameter and thickness. These two properties vary in each row and column, respectively, as it can be seen in Figure 2.3. In total, sixteen different diameters uniformly separated on a logarithmic scale from $0,06$ mm to $2,00$ mm and sixteen different thicknesses uniformly separated on a logarithmic scale from $0,03$ μm to $2,00$ μm are present in the CDMAM phantom. A lower thickness and a lower diameter reduce the amount of X-Ray photons that interact with the gold atoms and thus more radiation can pass, resulting in a transmission profile that is similar to that of the surrounding background. As a consequence, the SNR of image regions from targets with small thickness is more similar to the SNR of the background regions. Therefore, these artificial lesions are harder to detect in the resulting image.

The CDMAM phantom was designed in such a way that an observer was supposed to determine the location of the second gold disc in one of the four corners. Comparing the estimated position with the actual known position enables to decide whether the lesion was correctly detected or not. This procedure is repeated for a series of images for every cell of the CDMAM phantom. This way, the threshold gold thickness can be determined for every diameter such that the proportion of correctly detected discs crosses a pre-defined threshold.

This information is then used to evaluate the image quality. Solving this task with human observers is a slow and tedious task (Yip et al., 2009) and hence automatic procedures were developed for an objective and fast evaluation of mammography image quality in terms of contrast-detail curves.

2.1.3 EUREF Guideline Automatic Procedure

A standardized guideline to evaluate mammography image quality is important to guarantee consistent and robust quality assurance protocols that are comparable for different imaging devices. In Europe, the European Reference Organization for Quality Assured Breast Cancer Screening and Diagnostic Services (EUREF) proposed a guideline that defines the protocol for quality assurance in mammography (Perry et al., 2006), that will be referred to as EUREF guideline in the following.

Among other things, this guideline defines how to evaluate the contrast resolution by means of recordings of the CDMAM phantom. According to the EUREF guideline, images of the CDMAM phantom can be evaluated by either human observers or automatic procedures. Besides long evaluation times, the employment of humans may introduce a bias since human performance is highly subjective. Computer algorithms, on the other hand, enable an objective and fast assessment of image quality and the EUREF guideline further describes how to relate automatic procedures with a corresponding expectation from a human observer (Perry et al., 2006).

Evaluating images of the CDMAM phantom to assess image quality in mammography is a multistage process: the image needs to be pre-processed before an observer carries out a scoring that can be translated into a metric that expresses image quality. This metric is the so called contrast-detail curve, that links the diameter of gold discs with a thickness, such that an observer can localize the position of the disc in a series of images with a prescribed accuracy (Perry and Puthaar, 2006). For the automatic CDMAM readout, a minimum of 16 recorded images is required according to the EUREF guideline (Perry et al., 2006). The contrast-detail curve summarizes local information: in each column of the CDMAM phantom the gold discs have a fixed diameter, whereas their thickness varies in each row. Thus, the CDMAM scoring procedure aims to identify that row of one specific column, where the observer cannot detect the position of the second gold disc with sufficient accuracy.

2.1.3.1 Pre-processing

The contrast-detail curve summarizes local detectability information in the phantom. To compute this local information, many procedures require an error-prone pre-processing of the image of the CDMAM phantom. A decision has to be carried out for the individual cells of the CDMAM phantom. Therefore, the full image should be segmented into its individual cells, which can be done by aligning a mask of the known grid pattern with the image. Then, in each cell the actual position of the lesion has to be determined, which again can be done by comparing the cell with a mask that is known a priori. Different specimen of the CDMAM phantom, however, show deviations of up to 10% in diameter and $< 10\%$ in thickness of the gold discs, which cause a significant deviation of the estimated contrast-detail curves for the same imaging device but different phantoms (Borowski et al., 2012; Fabiszewska et al., 2016).

This introduces an uncertainty of the known ground truth mask and hence the positioning of the individual lesions has to be refined to achieve a reliable contrast-detail curve estimation.

All in all, a combination of pre-processings are applied to an image of the CDMAM phantom prior to the CDMAM scoring. First of all, the CDMAM phantom with its dimensions (16 *cm* × 24 *cm*) does not cover the whole area of the detector (typically 24 *cm* × 30 *cm*). Hence, the images contain bright areas outside the actual phantom image. These margins have to be cropped before further evaluation. Since photons not passing the phantom reach the detector more or less without any interaction on their path, the margins appear with higher pixel intensity. Therefore, they can be excluded simply by filtering out those pixels with gray values above a certain threshold. In addition, the CDMAM phantom image has to be aligned with a template mask to map the ground truth positions of the gold discs to the local pixels in the image in order to determine the position of each artificial lesion. Doing this for each image individually, further accounts for shifts in the positioning of the phantom in the acquisition procedure. An efficient alignment can be done by determining the position of the grid lines. The grid is presented in the phantom as thick straight lines and thus the Hough transform can be applied to detect them (Karssemeijer and Thijssen, 1996). Once the grid lines are detected, their intersections mark the four corners of each cell in the CDMAM phantom. This way, the whole phantom image can be segmented into its individual cell. Knowing the exact positions of the artificial lesions in each cell, makes it possible to check whether lesions are detected correctly, to compute the scoring of an observer.

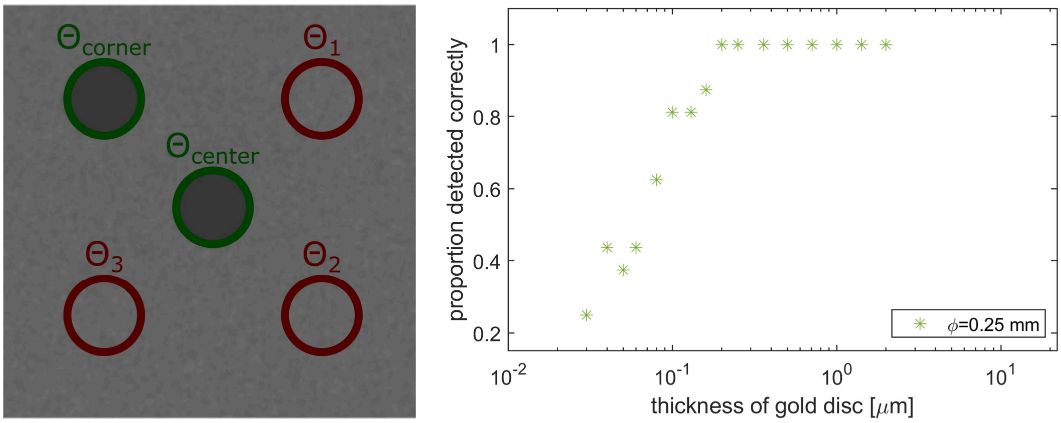
2.1.3.2 Contrast-detail Scoring

After successfully segmenting the image of the whole CDMAM into its individual cells, a mathematical observer is employed to localize the position of the second gold disc. This disc is positioned in one of the four corners of the square cell in a specified distance to the disc located in the center of the cell. In the current EUREF guideline, this observer simply calculates the intensity mean of the four potential corner disc positions and chooses the one out of the four possibilities with the minimum mean intensity. Note that gold has a higher X-Ray attenuation, thus, less radiation is transmitted which results in a lower intensity of the image regions that are in the shadow of the gold discs.

In practice, n_{images} of the CDMAM phantom are recorded, and for each cell a proportion of correctly detected gold discs is reported (Perry et al., 2006). Figure 2.4a shows an example image of one individual CDMAM cell, along with the true disc positions (green) and the three potential corner positions (red) that do not contain the gold disc. The values θ_k denote the mean (lesion) intensities in the corresponding region.

The mean lesion intensity of each true location is then compared with the mean intensity in the three potential red corner positions. For a cell with gold disc diameter d and thickness t, the proportion of correctly detected discs is obtained as

$$p_d(t) = \frac{n_{correct}}{2n_{images}}, \tag{2.3}$$



(a) (b)

Figure 2.4: (a) Illustration of the disc localization in a CDMAM cell. Green locations show the true position in the center and the true corner position, while red positions mark the potential corner positions that do not contain the gold disc. (b) Scoring result for the localisation of discs with diameter 0.25 mm for 16 images of all relevant thicknesses of the CDMAM phantom.

where $n_{correct}$ denotes the number of correctly located lesions. This can be calculated as

$$n_{correct} = \sum_{i=1}^{n_{images}} \mathcal{H}\left(min(\theta_1^i, \theta_2^i, \theta_3^i) - \theta_{corner}^i\right) + \mathcal{H}\left(min(\theta_1^i, \theta_2^i, \theta_3^i) - \theta_{center}^i\right) ,$$

where $\mathcal{H}(x)$ denotes the Heavyside function.

This means that for each of the n_{images} images the observer reports two scorings by comparing first the average intensity in the center lesion region θ_{center} with the three potential positions $\theta_1, \theta_2, \theta_3$ and repeating the same with the true corner position θ_{corner}. The combined result for a series of n_{images} images for one cell is then computed via equation 2.3. According to the EUREF guideline at least 16 images have to be analyzed leading to a total of 32 localizations per CDMAM cell (Perry et al., 2006). For gold discs with high thickness the difference between average intensities in the lesion and background region increases, thus, the observer correctly detects the corresponding lesion more often than lesions with low thickness. For a given diameter, the proportion of corrected thicknesses for all available thicknesses is collected as demonstrated in Figure 2.4b. For small thicknesses the curve tends to 0.25 which is the expected proportion of correctly detected lesions by chance for this one out of four task. For higher thicknesses, the curve tends to unity, which means that the observer correctly detects (almost) all lesions. The raw data is then fitted by assuming an underlying psychometric curve (Karssemeijer and Thijssen, 1996) given as

$$p_d(t) = \frac{0.75}{1 + \exp\left[-\zeta(C(t) - C_T)\right]} + 0.25 ,$$

where t denotes the thickness, ζ and C_T denote fit parameters that are estimated using a nonlinear least-square procedure and $C(t)$ denotes the contrast at thickness t given as

$$C(t) = ln\left(1 - e^{-\mu_{gold}t}\right) .$$

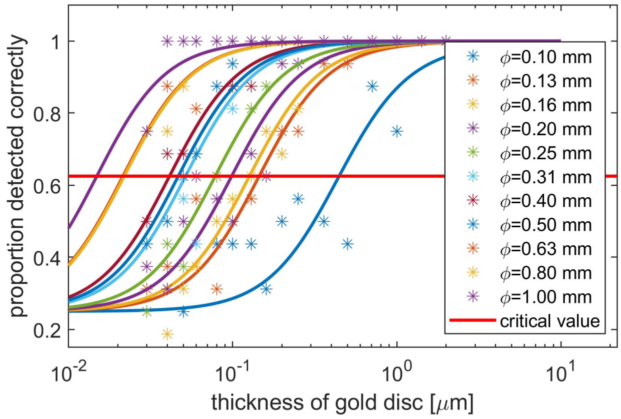


Figure 2.5: Psychometric curves for the localization performance of the automatic readout procedure of the EUREF guideline using 16 simulated images of the CDMAM phantom. The red horizontal line marks the minimum prescribed accuracy. The Figure was taken from Kretz et al. (2019) and modified by adding the critical value.

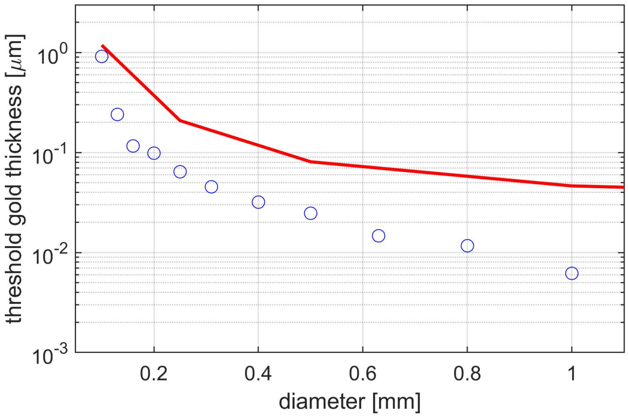


Figure 2.6: Example contrast-detail curve obtained by the EUREF guideline procedure for 16 simulated images. The red line illustrates the acceptable limit curve. The contrast-detail curve must be below the limit curve to meet the acceptance criterion. Figure from Kretz et al. (2019).

For a photon energy of 31 keV, the mass attenuation coefficient of gold is $\mu_{gold} = 0.047\ \mu m$.

The psychometric fit is carried out for all diameters, which is visualized in Figure 2.5 for a scoring of 16 simulated images of the CDMAM phantom.

The bigger the diameter of the gold discs, the more the corresponding psychometric curve shifts to the left, meaning that high detection accuracy is achieved for lower threshold gold thicknesses. For the determination of contrast-detail curves, the EUREF guideline prescribes a minimum accuracy of 0.625 which marks the center of the interval between pure guessing (0.25) and certainty (1) (Perry et al., 2006). This limit value is illustrated as the red horizontal line in Figure 2.5. Calculating the intersection of the psychometric curves with this limit value results in a predicted threshold gold thickness for each diameter that can be plotted as a contrast-detail curve, see Figure 2.6.

The contrast-detail curve displays the minimum gold thickness for each diameter, such that an observer can locate the position with at least 62.5% accuracy.

2.2 Machine Learning

The term machine learning frames a broad field within the area of Artificial Intelligence (AI). All you need for machine learning is data, a task and a performance measure of how good you solve this given task. Then, machine learning describes the process how an algorithm improves its performance to solve a given task by learning an optimal solving strategy from the data. In more technical terms, Mitchell (1997) defines machine learning as follows:

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Thus, after defining the performance measure P for a specific task T, the task of machine learning is to optimize the parameters of a selected model to accurately solve this task on a given set of data.

A remarkable benefit of machine learning is that the strategy of solving the task T is not explicitly programmed, but an algorithm is trained to solve this task, given a set of data (Bishop, 2006).

The learning process itself can be categorized into different stages: representation, evaluation and optimization (Domingos, 2012).

Representation addresses the challenge to represent the model in such a way that a computer can handle. Practically, this is done by selecting the kind of model that should be trained. Examples for such model representations are artificial neural networks (LeCun et al., 2015; Schmidhuber, 2015; Hassoun et al., 1995), support vector machines (Cortes and Vapnik, 1995; Müller et al., 2001; Suykens and Vandewalle, 1999; Schölkopf et al., 2002), random forests (Breiman, 2001; Ho, 1995) and many more. Besides the representation of the model, also the question of how to represent the data needs to be considered. For a given task T, it can be profitable to use a set of engineered features from the data rather than using the raw data directly (Zheng and Casari, 2018).

Evaluation is needed to separate good models and bad models or to ensure that a chosen model successfully improves over time. It is done by evaluating the performance measure P which is commonly referred to as the objective function or loss function.

Finally, **optimization** describes the (fine)-tuning of the parameters of the best model to achieve the highest scoring as measured by the performance measure P. Optimization thus defines an algorithm to update the model parameters to optimally solve the task T. Ideally, the scoring, measured by the performance measure P, is high on an particular training set as well as an independent test set that was not used for training. In this situation, the method successfully learned to generalize solving the task even on data that was never seen before.

The field of machine learning can be categorized according to different kinds of learning: supervised, semi-supervised, unsupervised and reinforcement learning (Lison, 2015). They all try to solve a task T by improving a performance measure P but utilize different concepts to reach that goal. In supervised learning, the data contains samples along with known labels for each sample. The set of known labels is often referred to as the ground truth. A model can then be used to make a prediction on the data, and the performance measure compares the prediction with the actual label and forces the model to predict the labels accurately.

In contrast, unsupervised learning learns to solve a task T, such as feature detection (Le, 2013), solely by using the data without further knowledge of labels. Semi-supervised learning is a hybrid between supervised and unsupervised learning: the data contains samples that are labeled as well as samples without labels. Reinforcement Learning (Sutton and Barto, 2018), on the other hand, describes a totally different approach. Here, an agent is trained to solve task T. Every good decision of this agent is rewarded and every bad decision punished, respectively. The agent is trained to avoid punishments, but collect many rewards instead. All different approaches brought tremendous success in their relevant fields. This work focuses only on supervised learning.

The supervised learning problem can be technically formulated as:

$$\begin{array}{ll} \text{Given} & D : \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\} \\ \text{find} & \psi(x, \theta) \\ \text{that best approximates} & x_i \mapsto y_i \forall i \end{array}$$

where D denotes a data set containing N training examples x_i with corresponding labels y_i, and $\psi(x, \theta)$ is the model function with model parameters θ.
Machine learning can be interpreted as a method for function approximation. Commonly, it is assumed, that the data y is generated by an unknown function f

$$y_i = f(x_i) + \epsilon_i \,,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ denotes some additional noise with variance σ_i^2. This function f can be either continuous, in case of a regression task, or discrete, for classification tasks. Machine learning aims to optimize the parameters θ of a selected model $\psi(x, \theta)$ to optimally approximate

$$x_i \mapsto y_i \forall i \,.$$

Theoretically, machine learning benefits from the *universal approximation theorem* (Cybenko, 1989) which states that any continuous function can be approximated arbitrarily accurately with neural networks. Later, an algorithm is introduced that allows one to optimize these parameters for several different models to closely approximate the ground truth labels.

Besides a correct prediction of the data in the training set, an optimal model should be able to generalize well, such that future unseen data can be predicted as well. One common challenge is here to avoid overfitting during the training procedure (see e.g. (Hastie et al., 2009)). Overfitting is when the model accurately predicts the labels of the data in the training set, but fails when it comes to the prediction of data that was not represented in the training set. This means that the model starts to fit the noise in the data, which leads to bad parameters estimates. To avoid overfitting, the training needs to be regularized. Common regularization strategies comprise data augmentation (Hernández-García and König, 2018), dropout (Hinton et al., 2012), early stopping (Orr and Müller, 1998) and regularization costs in the objective function (Wang and Klabjan, 2017).

Among other potential machine learning models, the artificial neural network (ANN) is one strong driver to open up one additional field of machine learning: the so-called deep learning. Deep learning usually employs big models with many layers of individual units, each described by a set of parameters. As mentioned earlier, representation is one important component of the great success of machine learning algorithms. If the data can be represented by certain features, a machine learning algorithm can successfully solve a task on them. Identifying a suitable set of features, however, requires a lot of expert knowledge. In many cases it is unknown which features should be extracted to represent the data for optimally solving a certain task. Deep learning bypasses the problem of feature engineering by discovering a suitable representation of the data by itself (Goodfellow et al., 2016). Deep models learn simple representations in the initial structure, and the deeper the model gets, the more complex the representations become (Sun et al., 2014). The initial elements, for example will only identify the colors of a pixel whereas a deep network can learn complex representations, for example, object shapes in its late elements. Thus, increasing the complexity of the model enables a better representation, whereas on the other hand, the number of learnable parameters increases significantly. Therefore, (high) computational power is a good friend of deep learning (LeCun, 2019). The absence of large data sets as well as lacking high performance computer hardware caused a decreasing popularity of deep learning, despite its introduction in the 1980's (LeCun, 2019). Finally, in 2012, Krizhevsky et al. (2012) implemented deep learning for image classification on the large ImageNet (Berg et al., 2010) data set, achieving better results than every state-of-the-art algorithm so far. They report a highly optimized implementation using graphics processing units (GPUs) to successfully train a deep model with around 60 million parameters. Since then, deep learning was successfully applied in various fields including speech-recognition, object detection and natural language processing (see e.g. LeCun et al. (2015)).

For more in-depth information on the subject, we refer the interested reader to the dedicated literature (Bishop, 2006; Orr and Müller, 1998; Lison, 2015; Mitchell, 1997; Goodfellow et al., 2016). In the following, we focus on a detailed description of the ANN and how to train a model.

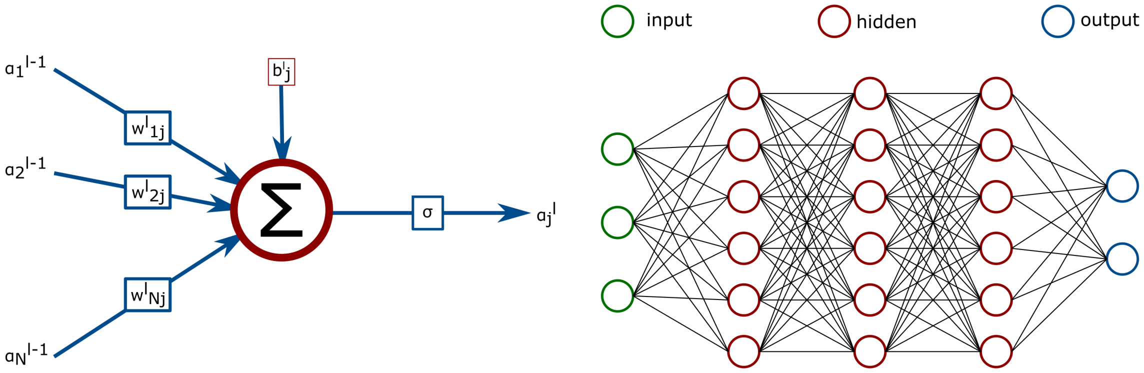
2.2.1 Artificial Neural Networks

In neurophysiology a neuron describes a biological cell for information processing. A neuron consists of a cell body, dendrites and the so called axon (Jefferys and Cooper, 2007). A neuron receives information from other neurons and, depending on whether the information reaches a certain threshold, the neuron transmits a signal through its axon. Motivated by its biological counterpart, an endeavour to develop such neurons artificially was started in the early 1940s by McCulloch and Pitts (see McCulloch and Pitts (1943)). Based on their pioneering work, Rosenblatt introduced the perceptron in 1958 as a unit that linearly combines its inputs. It gives 1 as an output if the sum of all inputs is greater than 0 and 0 otherwise (Rosenblatt, 1958). Figure 2.7a depicts the Rosenblatt perceptron.

Mathematically, an artificial neuron j multiplies its inputs α_i^{l-1} with weight factors w_{ij}^l and outputs a sum of all weighted inputs and a bias b_j after applying an nonlinear activation function σ

$$\alpha_j^l = \sigma\left(\sum_i w_{ij}^l * \alpha_i^{l-1} + b_j^l\right).$$

A feed-forward ANN combines several artificial neurons and arranges them in layers as shown in Figure 2.7b. The first layer is referred to as input layer and the last layer is called output



(a) **(b)**

Figure 2.7: (a) Schematic visualization of an artificial neuron according to the McCulloch-Pitts model. A neuron outputs the weighted sum of its inputs after applying a activation function σ. (b) A feedforward ANN combines multiple artificial neurons in layers and stacks the several layers to a network structure. Each neuron is connected to every neuron in the following layer.

layer. An ANN combines these with (several) hidden layers in between. Especially in deep learning, the number of hidden layers is high. In a fully connected layer, every neuron is connected with every neuron of the next following layer. The parameters θ of such an ANN comprise all the weights and biases of every neuron and need to be optimized for an accurate prediction of the underlying ground truth.

The output can be mathematically expressed as a nested evaluation of all the layers outputs. Assuming there are n_k neurons in layer k of a fully connected network, the weights can be expressed as a matrix $\boldsymbol{W}^k \in \mathbb{R}^{n_k \times n_{k-1}}$. The inputs for layer k, are the n_{k-1} outputs $\boldsymbol{\alpha}^{k-1} \in \mathbb{R}^{n_{k-1} \times 1}$ of the previous layer. With the vector of all biases $\boldsymbol{b}^k \in \mathbb{R}^{n_k \times 1}$ in layer k, the output of layer can be calculated using the matrix-vector multiplication:

$$h_k\left(\boldsymbol{\alpha}^{k-1}\right) = \sigma_k\left(\boldsymbol{W}^k \cdot \boldsymbol{\alpha}^{k-1} + \boldsymbol{b}^k\right), \tag{2.4}$$

where σ_k denotes the activation function in layer k. In this notation, it is assumed that the input $\boldsymbol{x}$ can be expressed as a one dimensional vector and is denoted by $\boldsymbol{\alpha}^0$.

For a fully connected neural network with four layers, as illustrated in Figure 2.7b, the output $\hat{y}$ can be computed as:

$$\hat{y} = \psi\left(x, \theta\right) = h4(h3(h2(h1(x))))$$

ANNs were introduced in the 1960s where they were computationally difficult to train (Goodfellow et al., 2016). In the late 1980s, ANNs gained new popularity with the development of the back-propagation algorithm (Rumelhart et al., 1986), which will be explained in more detail in the next section.

Since their introduction ANNs were further developed. New kinds of layers were introduced to ease the training and consider correlations in the inputs. In 1998, LeCun et al. (1998) introduced a neural network architecture which achieved remarkable results in hand written digit recognition. Here, the input are images. Adjacent pixels in images are usually highly correlated and not independent. A fully connected layer does not consider such correlations, which led to the development of convolutional layers that are able to account for correlations in the local neighborhood of a pixel. Such convolutional neural network (CNN) usually combine convolutional layers and pooling layers together with batch-normalization and nonlinear activation functions, which are briefly reviewed in the following.

Convolutional Layers: An image usually consists of several hundreds or thousands of pixels. Representing each pixel individually as an input neuron for a fully connected network would lead to a huge number of parameters that is computationally difficult to handle. Convolutional layers act as filters and use parameter sharing to reduce the number of learnable parameters. Such convolutional layers learn the entries of a convolution kernel that is referred to as filter. By defining a filter size, one can shift the filter over the pixels of the image to calculate a feature map that expresses, how present the feature of the filter is in parts of the image. The development of such convolutional layers was motivated by local receptive fields as described in (LeCun et al., 1998). Instead of learning a weight for every pixel, only the filter weights are optimized, leading to a great reduction of the number of trainable parameters.

Pooling Layers: Once a feature is detected by a convolutional filter, its exact position is less relevant (LeCun et al., 1998). Position invariance, which is an essential requirement in many applications, can be achieved by sub-sampling the feature map. This further reduces the dimension of the feature map, which can additionally reduce the number of learnable parameters. Commonly, max-pooling only keeps the maximum activation in a 2×2 receptive field in the feature map. Sub-sampling of the feature maps reduces the dimensions and thus the complexity. Additionally, it also reduces the output's sensitivity on shifts and distortions. For many tasks (e.g. image classification), the position of the object in the image is less relevant and shift invariance is a desired property of the trained neural net.

Activation Layers: Without the application of an activation function the layer output as described in Equation 2.4 and thus the neural network output, reduces to a linear function. Thus, to account for nonlinearities, nonlinear activation functions are a key feature. Sigmoid functions were initially used to model the biological neuron's behavior that only fires a pulse if the excitation reaches a certain threshold (Rumelhart et al., 1986). Similarly, sigmoid logistic functions are zero until the input reaches a threshold, and one for values significantly larger than the threshold. However, sigmoid activation functions exhibit bad convergence due to the vanishing gradient problem (Pascanu et al., 2013). Therefore, Nair and Hinton (2010) introduced rectified linear units (ReLU) to avoid this problem.

$$ReLU(x) = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{else.} \end{cases}$$

This function was found to be computationally stable and is therefore widely used in neural networks.

Normalization Layers: It was found that the convergence of neural network training is in general much faster if the average over each input over the training set is close to zero (LeCun et al., 2012). Therefore, a normalization of the input over the training set can be used to stabilize the training process. Similarly, Ioffe and Szegedy (2015) state in their work that the same phenomena occur in the distribution of each layer's input. These distributions change during the training, and Ioffe and Szegedy (2015) introduce batch-normalization layers to

Table 2.2: Overview of different loss functions. M can be either the total number of examples in the whole training set or the size of a mini-batch.

Name	$\mathcal{L}\left(y_i, \psi\left(x_i, \theta\right)\right)$	setting
Mean squared error	$\left(y_i - \psi\left(x_i, \theta\right)\right)^2$	regression
Mean absolute error	$\lvert y_i - \psi\left(x_i, \theta\right)\rvert$	regression
Huber loss	$\begin{cases} \frac{1}{2}\left(y_i - \psi\left(x_i, \theta\right)\right)^2 & \text{if } \lvert y_i - \psi\left(x_i, \theta\right)\rvert < \delta \\ \delta\lvert y_i - \psi\left(x_i, \theta\right)\rvert - \frac{1}{2}\delta^2 & \text{else.} \end{cases}$	regression
Log loss	$y_i \log\left(\psi\left(x_i, \theta\right)\right) + \left(1 - y_i\right) \log\left(1 - \psi\left(x_i, \theta\right)\right),$ $\psi\left(x_i, \theta\right) \in (0, 1)$	binary classification
Cross entropy	$\sum_{j=1}^{c} y_i^{(j)} \log\left(\psi^{(j)}\left(x_i, \theta\right)\right),$ $\psi^{(j)}\left(x_i, \theta\right) \in (0, 1)\, \forall j$	classification

normalize the input of each layer for every training batch. This way, they achieved a more robust training procedure.

2.2.2 Learning and Back-propagation

Until now, the general concept of machine learning was discussed and ANNs were introduced as one possible kind of models among many others. ANNs can be described by their model parameters θ, which are the weights and biases of each neuron of the net, respectively. The idea is to optimize the model parameters θ of our model ψ such that the resulting predictions fit the underlying label y.

As mentioned in the definition of machine learning a performance measure P is needed to determine the deviation between model output and ground truth. This performance measure is more commonly referred to as the loss function, here denoted as $\mathcal{L}$. The loss function penalizes large deviations of the ground truth y and the model output $\psi\left(x, \theta\right)$, and therefore it should be minimized during the training

$$\arg\min_{\theta} \sum_{i=1}^{M} \mathcal{L}\left(y_i, \psi\left(x_i, \theta\right)\right) \ ,$$

where M denotes either the total number of samples or the size of a mini-batch, which will be explained later. Table 2.2 provides an overview of commonly used loss functions. Note that the log loss and the cross entropy are special loss functions, requiring that $\psi(x_i, \theta)$ is non-negative and normalized. In classification problems this is usually ensured by computing the softmax function (Murphy, 2012).

Once a loss function is defined, the question arises, how this function can be used to minimize the deviation between the model and the ground truth. Among other examples, the gradient descent algorithm (Cauchy, 1847) describes how the gradients of the loss function can be used to minimize its values stepwise. In a small neighborhood $\boldsymbol{\Delta x}$, any differentiable function f can be approximated by using its gradient:

$$f(\boldsymbol{x} + \boldsymbol{\Delta x}) \approx f(\boldsymbol{x}) + \boldsymbol{\Delta x} \cdot \boldsymbol{\nabla} f(\boldsymbol{x}).$$

Following this concept, the function f can be reduced by taking small steps $\eta > 0$ in the direction of the negative gradient.

$$f(\boldsymbol{x} - \eta \boldsymbol{\nabla} f(\boldsymbol{x})) \approx f(\boldsymbol{x}) - \eta \boldsymbol{\nabla} f(\boldsymbol{x}) \cdot \boldsymbol{\nabla} f(\boldsymbol{x}) < f(\boldsymbol{x})$$

The parameters to optimize are the model parameters θ. To reduce the loss function with respect to θ, the gradient descent algorithm determines the updates of the model parameters by moving θ a small step along the direction of the gradient of the loss function.

$$\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}\left(y, \psi\left(x, \theta\right)\right)}{\partial \theta} \tag{2.5}$$

The step size η is called learning rate and defines how much the parameters θ change with each update. The choice of η is crucial. Choosing η too small results in slow convergence or converge into a local minimum. On the other hand, if η is chosen too big, then the optimizer can step out of the minimum and diverge to infinity (Zeiler, 2012). Besides, the assumption of linearity is no longer valid for large values of η.

To estimate the parameters and their updates for all samples in the training set can be very time and memory consuming. Therefore, *stochastic gradient descent* methods are commonly used that approximate the true gradient by considering only a single sample or a small sub-sample of the whole training set in one parameter update (Bottou, 2010). The stochastic nature of this process is introduced through the fact that these sub-samples, which are referred to as batches, are drawn randomly from the data. Stochastic gradient descent further acts as a regularization, avoiding that the optimization gets stuck in a local minimum (Ruder, 2016).

However, the approximation of exact gradients of the loss function introduces some noise, so that the optimizer will not always go into the optimal direction. As a consequence, *stochastic gradient decent with momentum (SGDM)* was introduced to avoid this. SGDM updates its parameters based on the gradient as well as on the previous gradient (Qian, 1999). The sequence of gradients V_i is then computed by

$$V_i = \beta V_{i-1} + \eta \frac{\partial \mathcal{L}\left(y, \psi\left(x, \theta\right)\right)}{\partial \theta} \,,$$

with a momentum factor β. The parameter updates are then given by

$$\theta \leftarrow \theta - V_i \,.$$

Stochastic gradient descent with momentum resembles conjugate gradient methods (Fletcher and Reeves, 1964). Such conjugate gradient methods avoid taking multiple steps in the same direction of the gradient and hence ensure a faster convergence (Qian, 1999). In fact, conjugate gradient methods terminates provably in the minimum after n line-searches for a positive-definite n-dimensional quadratic form (Qian, 1999).

In order to determine the updates of the model parameters θ, their gradients of the loss function need to be calculated. For a hierarchical model, such as an ANN, the gradients of the model parameters can be obtained by applying the chain rule. To demonstrate this, we assume a simple model of one input neuron $x = \alpha_0$, one single hidden neuron α_1 and the output neuron $\hat{y} = \psi(x, \theta)$, as depicted in Figure 2.8.

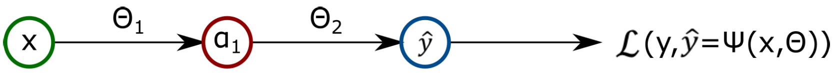


Figure 2.8: Schematic to demonstrate the back-propagation algorithm for a hierarchical model with one input, one hidden, and one output neuron.

The model is ψ is described by its parameters $\theta = \{\theta_1, \theta_2\}$. For given x, θ_1, θ_2 we can evaluate the model and calculate the loss function $\mathcal{L}(y, \hat{y})$. The question is how to update the model parameters such that the model approximates the true output closely. As mentioned, the (stochastic) gradient descent method allows the model parameters to be updated once the gradients are determined. The gradients to be determined are:

$$\frac{\partial \mathcal{L}(y, \psi(x, \theta))}{\partial \theta_1} \text{ and } \frac{\partial \mathcal{L}(y, \psi(x, \theta))}{\partial \theta_2}.$$

The loss function $\mathcal{L}(y, \hat{y})$ depends only implicitly on the model parameters θ. However, $\hat{y}$ depends explicit on the model parameters θ_2 in our case. Thus, we can apply the chain rule to determine the gradient with respect to θ_2:

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial \theta_2} = \frac{\partial \mathcal{L}(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}(\alpha_1, \theta_2)}{\partial \theta_2}.$$

For differentiable loss and activation functions, these gradients can be expressed analytically. In the same way, the gradient with respect to θ_1 can be determined. The loss function depends explicitly on $\hat{y}$, but $\hat{y}$ does depend only implicitly on θ_1. However, $\alpha_1 = \alpha_1(x, \theta_1)$ depends explicitly on θ_1 and thus applying the chain rule once more yields

$$\frac{\partial \mathcal{L}(y, \hat{y})}{\partial \theta_1} = \frac{\partial \mathcal{L}(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}(\alpha_1, \theta_2)}{\partial \alpha_1} \cdot \frac{\partial \alpha_1(x, \theta_1)}{\partial \theta_1}.$$

By applying the chain rule, the gradient of the loss function can be traced back to the influence of the individual model parameters. With the knowledge of the gradients, the updates of the individual parameters can then be performed according to Equation 2.5, in order to reduce the loss function and thus better approximate the ground truth values. The improved efficiency of the back-propagation algorithm relies on the fact that all the gradients are computed simultaneously, requiring only the evaluation of one forward pass, followed by a single backward pass (Rumelhart et al., 1986).

2.2.3 Probabilistic Interpretation

So far, machine learning was introduced in general and ANNs were discussed as one type of model to approximate functions. Models can be trained using the back-propagation algorithm. In addition, there is a probabilistic interpretation of deep learning which aims for the determination of probability distributions to enable an understanding of the model's confidence and uncertainty.

If we assume that the data x of our data set D follows an underlying probability distribution, we are looking for the parameters θ of our model ψ that are most likely to generate the ground truth y. By starting with an initial prior distribution $p(\theta)$ we are interested in how to update the parameters after seeing the data, which is denoted as the posterior distribution $p(\theta \mid D)$.

The Bayes rule (Bayes, 1763) enables the determination of the posterior

$$p(\theta \mid D) = \frac{p(y \mid x, \theta)p(\theta)}{p(D)} \tag{2.6}$$

by multiplying the prior $p(\theta)$, with the likelihood $p(y \mid x, \theta)$ which equals the distribution of the output y viewed as a function of the model parameters θ. The expression

$$p(D) = \int p(y \mid x, \theta)p(\theta)d\theta$$

is called model evidence and acts as a normalization of the posterior in equation 2.6.

For simple models, as in linear regression, the marginalisation can be expressed analytically, whereas for most applications, the marginalisation and thus the posterior $p(\theta \mid D)$ needs to be approximated.

After finding distributions that describe the parameters θ to optimally represent the training set D, the ultimate goal is to find a way to determine a prediction for any new datapoint x^*. This can be done by invoking Bayesian inference

$$p(y^* \mid x^*, D) = \int p(y^* \mid x^*, \theta)p(\theta \mid D)d\theta$$

where the posterior predictive distribution $p(y^* \mid x^*, D)$ is the prediction according to the assumed statistical model averaged over the posterior for all model parameters θ.

2.2.3.1 Variational inference

As stated above, for most problems it is not tractable to compute the true posterior. Thus, rather than calculating the true posterior directly, a good approximation $q_\phi(\theta)$ of the posterior $p(\theta \mid D)$ has to be determined which can be expressed analytically and represents the true posterior sufficiently, with respect to parameters ϕ. Variational inference (Jordan et al., 1999) replaces the true posterior in Equation 2.6 with an optimization of the parameters ϕ of an approximate distribution $q_\phi(\theta)$ such that the approximate distribution is as close as possible to the true posterior $p(\theta \mid D)$. Commonly the Kullback-Leibler (KL) divergence is used to measure the similarity between to distributions (Kullback and Leibler, 1951). The KL is given as

$$KL\left[q_\phi(\theta)||p(\theta \mid D)\right] = \int q_\phi(\theta) \log \frac{q_\phi(\theta)}{p(\theta \mid D)} d\theta = \mathbb{E}_{q_\phi(\theta)}\left[\log(q_\phi(\theta))\right] - \mathbb{E}_{q_\phi(\theta)}\left[\log(p(\theta \mid D))\right] .$$

The optimal approximation $q^*_\phi(\theta)$ minimizes the KL divergence

$$q^*_\phi(\theta) = \arg\min_\phi KL\left[q_\phi(\theta)||p(\theta \mid D)\right] .$$

With this approximation of the true posterior in equation 2.6, variational inference defines the posterior predictive distribution as

$$p(y^* \mid x^*, D) = \int p(y^* \mid x^*, \theta)q^*_\phi(\theta)d\theta\,.$$

The integral to calculate the KL divergence between the approximate distribution and the true posterior cannot be computed directly, but minimizing the KL divergence is equivalent to maximizing the so called evidence lower bound (ELBO) (Blei et al., 2017), which is given as

$$ELBO = \mathbb{E}_{q_\phi(\theta)}\left[\log(p(y \mid x, \theta))\right] - KL\left[q_\phi(\theta)||p(\theta)\right]. \tag{2.7}$$

Maximizing the ELBO is equivalent to maximizing the first term in Equation 2.7, whereas the KL divergence between $q_\phi(\theta)$ and the prior $p(\theta)$ is minimized. The first term is the expected log-likelihood, and maximizing it forces $q_\phi(\theta)$ to explain the data well, while the KL divergence punishes large deviations of the approximate distribution from the prior distribution of the parameters. In this way, the KL divergence can be interpreted as a stochastic regularization technique (Gal, 2016).

In practice, a Monte Carlo approximation of the negative *ELBO* is commonly used as learning objective (Gal and Ghahramani, 2016)

$$\mathcal{L}_{MC}\left(\theta\right) = \frac{1}{M}\sum_{m=1}^{M} - \log\left(p\left(y_i \mid x_i, \theta\right)\right) + \frac{1}{N}KL\left[q_w(\theta)\|p(\theta)\right]\ , \tag{2.8}$$

where M is the size of a mini-batch and N is the total number of training samples. The second term, the KL divergence, can often be expressed analytically, and therefore this loss function is commonly used as learning objective.

3 Virtual Mammography

Simulations are widely used to understand and study physical processes. In medical imaging, simulations can be used to conduct large virtual clinical trials, rather than time and money consuming real data clinical trials. This way, simulations provide a tool to gain understanding of the underlying imaging physics and thus provide useful information for developers of systems as well as their end users.

In this chapter we will *implement a virtual mammography to simulate radiographic images* of the CDMAM phantom. Virtual mammography guarantees that the ground truth of the virtual specimen is known. A known ground truth along with the ease of generating large image data sets makes virtual mammography a useful tool for the development of data-intensive methods, such as deep learning. Furthermore, the known ground truth allows an inference to be judged based on the simulated images, which makes virtual mammography a suitable tool for the assessment of methods that quantify the image quality.

As outlined in chapter 2, the image formation process in mammography can be divided into three sub-processes, namely generation, interaction and detection of X-Ray radiation. In order to simulate realistic mammography images, each of these has to be modeled subsequently. First, an X-Ray source is simulated to generate an (initial) X-Ray spectrum that can be pre-filtered to obtain the desired properties. The fundamental principle of imaging in mammography is to record the X-Ray transmission profile of an object in the beam path. Such images can then be used to detect breast cancer. Modeling the transmission profile requires the simulation of the interaction of X-Ray photons with the matter of the object. Every material has its own, unique X-Ray absorption properties. Therefore, an object like the CDMAM phantom that is composed of different materials has to be represented by a digitized phantom. This digitized phantom needs to summarize local X-Ray attenuation properties such that a transmission profile can be computed. After computing the transmission profile for every beam, the detection of the X-Ray's has to be modeled to translate the incoming radiation into pixel intensities to yield the final image.

Additional to the implementation of virtual mammography, the *application of a parametric model observers for mammography image quality assessment* is proposed in this chapter. Mathematical model observers are commonly used to develop fast and objective automatic procedures for image quality assessment. Such model observers determine a test statistic from a set of training data which is suitable to solve a given task, e.g. lesion detection. The automatic procedure recommended in the EUREF guideline for mammography quality assurance employs a basic model observer to determine the image quality in terms of contrast-detail curves. This procedure, introduced previously in section 2.1.3, is used as a reference to discuss the advantages as well as the limitations of the proposed alternative approach.

Most of the work presented in this chapter is based on Kretz et al. (2019).

3.1 Related Work

Simulations have been widely used to model realistic radiographic images.
For example, Elangovan et al. (2014) developed a tool to model 2D mammography as well as tomosynthesis, which can be used to report useful information to different vendors. Their simulation pipeline starts with the definition of a mathematical phantom. Then the primary transmission through this phantom is calculated and subsequently degraded by adding noise and scattering. The images simulated this way show good agreement with real images with respect to their contrast-to-noise ratio and image blurring of test objects.

Similarly, Gong et al. (2006) set up a simulation of digital mammography, breast tomosynthesis and cone-beam computer tomography to analyze the lesion detectability for these different breast imaging modalities. Again, the image is simulated by deriving a primary transmission through a digital phantom. This primary image is subsequently degraded by various effects in the image receptor such as electronic noise.

Also, Mackenzie et al. (2017) use virtual mammography to study the uncertainty related to the determination of contrast-detail curves in mammography quality assurance. As a result, the authors demonstrate the use of 16 images only, which is recommended in the EUREF guideline, is insufficient since the uncertainty of the determined parameters is too high.

Furthermore, simulations have been designed to study the influence of several acquisition parameters in digital tomosynthesis (Malliori et al., 2013) as well as in mammography (Bakic et al., 2002), where the latter work focuses more on a realistic modeling of breast tissue and compressing dynamics.

This list of simulation tools is not complete, cf. also Lazos et al. (2000); Freud et al. (2004, 2005) for further approaches. Altogether, the current state-of-the-art simulations favor the simulation of a primary image that is degraded considering various noise effects in a second step.

3.2 Simulation Tool for Mammography

Following the current practice in radiography simulations, a tool is implemented to simulate mammography images to create a data base of images for the assessment and development of methods. The virtual mammography was designed and optimised for the simulation of images of the CDMAM phantom that is used to determine the image quality in terms of a contrast-detail curve. The basic workflow of our simulation is visualized in Figure 3.1. A primary image is derived that is then phenomenologically degraded by adding scattering, noise

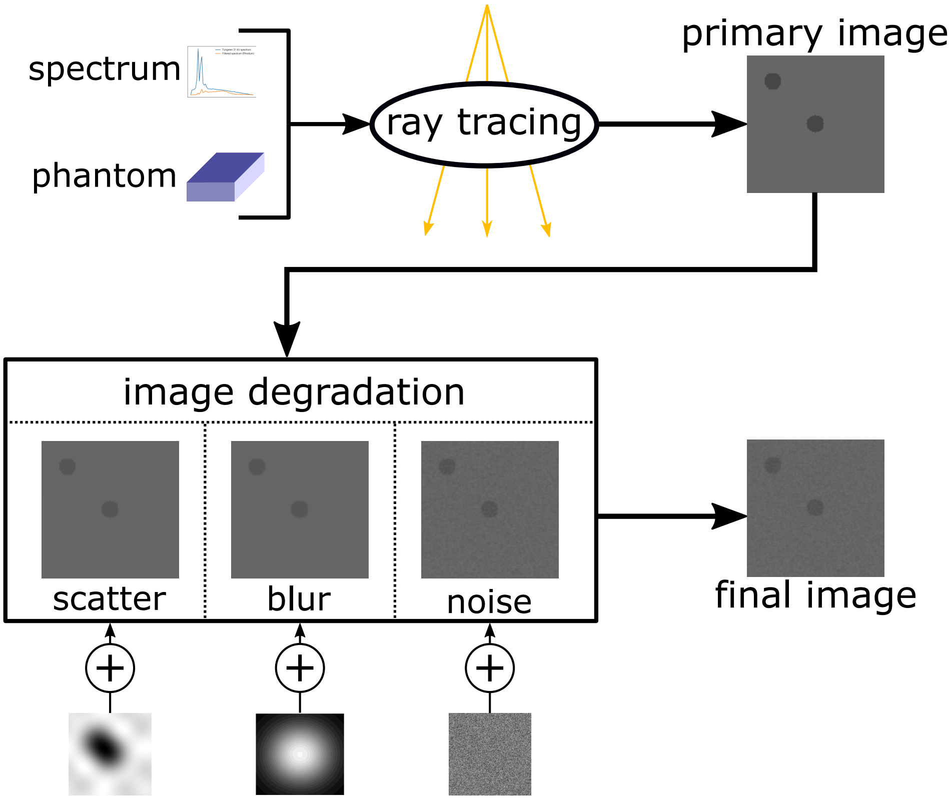


Figure 3.1: Schematic of the simulation pipeline for the virtual mammography. A backward ray casting procedure is used to simulate the transmission of an initial spectrum through a mathematical object. The transmitted radiation is recorded in a primary image and then further degraded by adding scatter, blurring and noise. Figure from Kretz et al. (2019).

and blurring to the primary image.

For the production of the primary image, the object to be imaged has to be represented mathematically, the X-Ray spectrum has to be modeled and the transmission profile for each individual beam path has to be calculated.

3.2.1 Primary Image Simulation

In order to simulate a radiographic image of an object, a digital phantom that describes the object is required. For X-Ray radiography, the material in terms of its X-Ray attenuation parameters and its thickness define the contrast in the resulting image. We implemented a digital CDMAM phantom as a 3D voxel object. The voxels are defined such that each voxel contains only one material. By further specifying the dimensions of the voxels, both parameters, material composition and thickness, are known such that the transmission of incoming radiation can be determined. The dimension of the voxels has to be chosen small enough, so that the discretization error can be neglected.

The CDMAM phantom (see section 2.1.2) is composed of background material, grid material and gold for the discs. The CDMAM phantom comprises various cells each containing gold discs of varying diameter and thickness. To model gold discs of different thicknesses, multiple layers of gold voxels are stacked until the desired thickness is reached. Every voxel that does not belong to the grid region or the gold disc region is modeled as a voxel containing the background material. This way, a model of the entire CDMAM phantom was built, as well as models of the individual cells of the phantom.

As a next step, X-Ray radiation, its absorption and detection need to be modeled. As described in chapter 2, in mammography usually Molybdenum (*Mo*), Rhodium (*Rh*) or Tungsten (*W*) anodes are used to generate the X-Ray spectra. The methodology of interpolating polynomials, developed by Boone et al. (1997), is employed to simulate an initial virtual spectrum. They provide data of measured spectra for all three relevant anode materials and developed a method to calculate spectra for various tube voltages. The spectra modeled this way are normalized to the time-current product that was used in the original measurements and comprise Bremsstrahlung as well as characteristic radiation. Usually the initial spectrum is pre-filtered to reduce the effect of the undesired Bremsstrahlung that adds to the applied radiation dose and reduces the contrast. The pre-filtering can be modeled by applying the Beer-Lambert law (cf. 2.1) for a given initial spectrum and a filter of specified material and thickness.

A real X-Ray tube emits radiation as an area source, whereas it is modeled as a point source in our simulation. This means that all rays originate from the same position. Different starting positions would result in a blurring in the simulated image (Elangovan et al., 2014) but increase the computational complexity of the simulation by increasing the number of potential beam paths. Therefore, blurring will be added during the degradation stage phenomenologically. As a consequence, the geometry of the X-Ray beams is only determined by the position of the source and the target point at the detector. The center of the individual detector pixels is taken as target points. Hence, the resulting number of beams equals the number of pixels in the detector.

With a known initial spectrum, a digital phantom and the radiological beam path, the absorption profile of each individual beam can be determined. The path of the X-Rays from the source through the object to the detector pixel is calculated by using Siddon's algorithm (Siddon, 1985). For each detector pixel, the X-Ray's path through the phantom is a little bit different, resulting in a slightly varying distance that the X-Rays travel through different materials. Siddon's algorithm calculates for each detector pixel ij the distances that the X-Rays travel through each of the three materials $\delta x^{ij}_{background}$, δx^{ij}_{grid} and δx^{ij}_{gold} as indicated in Figure 3.2.

The resulting radiation $N^{ij}(E)$ reaching the detector pixel ij can then be calculated by applying the Beer-Lambert law (cf. equation 2.1)

$$N^{ij}(E) = N_f(E)\exp^{-(\mu_{bg}(E)\delta x^{ij}_{bg}+\mu_{grid}(E)\delta x^{ij}_{grid}+\mu_{gold}(E)\delta x^{ij}_{gold})}, \tag{3.1}$$

where $N_f(E)$ denotes the pre-filtered initial spectrum and μ denote the attenuation coefficient for background (bg), grid and gold, respectively.

The transmitted radiation of each beam targets one single pixel of the detector without further attenuation. To translate this information into an image, the detector needs to be modeled. Each detector pixel is hit by an attenuated X-Ray spectrum and transforms the incoming photon fluence into a pixel intensity by integrating over all energies. The characteristic curve of the detector is assumed to be linear which means that an increase in the incoming radiation intensity linearly increases the pixel intensity. The resulting pixel intensity PV^{ij} for

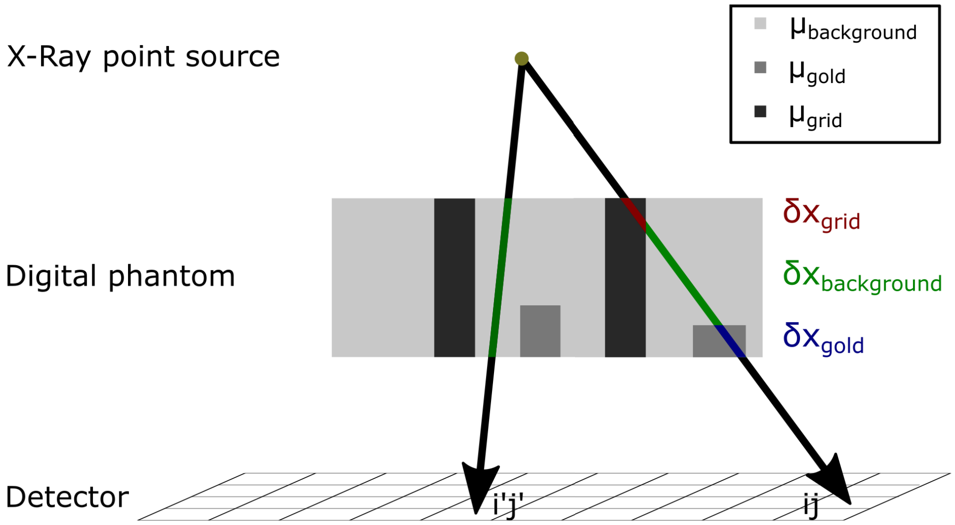


Figure 3.2: Schematic of the detection of transmitted radiation created by an X-Ray point source. The ray targeting detector pixel ij travels through the phantom in a unique way resulting in different distances in different materials that are calculated by Siddon's algorithm to determine the transmitted radiation after the phantom.

pixel ij can be expressed as

$$PV^{ij} = \int \eta(E) N^{ij}(E) E dE \,,$$

where $N^{ij}(E)E$ denotes the incoming exposure given in equation 3.1 and $\eta(E)$ denotes the energy dependent detector efficiency (cf. chapter 2).
Following these steps a primary image can be simulated.

3.2.2 Image Degradation

The primary image is simulated without considering any degradation effects. In reality, however, images are degraded by several effects like detector noise, scattered radiation and limited spatial resolution. Recalling that each sub-process of the image formation procedure in mammography has an effect on the resulting image quality, these contributions have to be modeled if we want to infer image quality from the simulated images.

During the generation of the X-Rays, blurring is the major effect that limits the spatial resolution and hence the resulting image quality (ICRU, 2009). Photon scattering adds quantum noise in digital mammography and reduces the contrast-to-noise-ratio in the final image (ICRU, 2009). Finally, image quality is also limited by the detector noise (ICRU, 2009).

In our virtual mammography, these three image degradation processes are considered phenomenologically which produces images that are visually very similar to real images as will be shown. Blurring occurs when the focal spot of the X-Ray source is not a single spot. Then the rays targeting the pixel ij originate from different positions in the source and the corresponding beams take slightly different paths through the phantom, resulting in a different attenuation profile. As a result, the sharp edges that occur when considering a point source are softened for realistic focal spots. Besides this focal spot blur, blurring also occurs in the scintillator and as an effect of the pixel size. In our simulation tool, blurring is simulated by filtering the primary image with a Gaussian filter with standard deviation 0.7 *Pixels* (cf. Spyrou et al. (2002)). The Gaussian filter softens the edges of the imaged objects. The degree of how much the edges are softened is determined by the standard deviation of the Gaussian filter.

Scattering is known to significantly reduce the contrast in film/screen imaging (Ducote and Molloi, 2010). In digital imaging, scattering furthermore adds quantum noise and thus reduces the contrast-to-noise-ratio. To realistically model scattering, one would have to perform time consuming Monte-Carlo simulations. We use a phenomenological approach by applying a convolution technique, as suggested in (Love and Kruger, 1987). A scattered image is calculated by a convolution of the primary image with a point spread function. This convolution can be performed in the Fourier domain: the Fourier transform of the primary image is multiplied with the Fourier transform of an ideal low-pass filter with $D_0 = 2\ Pixels$. By applying the inverse Fourier transform to this product, the final scattering image results.

The amount of scattering is controlled by the scatter-to-primary-ratio (SPR) (Boone et al., 2000). Setting the SPR allows the pixel intensities of the scattering image to be scaled to similar values as those in the primary image. This way, scattering becomes a noticeable effect that reduces the contrast-to-noise-ratio in the resulting image. In Boone et al. (2000), the authors describe how to choose the SPR realistically for an equivalent breast thickness of 6 cm. The scattering image is then added to the blurred primary image.

The noise characteristics of a recorded mammography image are strongly influenced by the type of detector that is used as image receptor (Ravaglia et al., 2009). Thus, rather than trying to perfectly model one specific type of detector, the detector noise is modeled phenomenologically. Noise in mammography images is spatially correlated. Therefore, we use a combination of Gaussian white noise with standard deviation σ_{wn} and a spatial Gaussian filter with standard deviation $\sigma_{sc} = 0.55\ Pixels$, to simulate such correlated noise.

The strength of noise is controlled by the signal-to-noise-ratio (SNR) which is a free parameter in our simulation. The SNR combines the pixel intensity of the signature of a lesion with the strength of the noise. The total standard deviation of the noise image is determined through the standard deviation of the Gaussian white noise σ_{wn} and the standard deviation of the Gaussian filter σ_{sc}. The obtained noise image is then added to the primary image degraded by blurring and scattering.

The simulation parameters of our virtual mammography are listed in Table 3.1. The main parameters that influence the resulting image quality are the tube voltage ($18 - 40\ kV$), the time-current product ($100 - 140\ mAs$) and the SNR ($\approx 3 - 10$). To avoid discretization errors, the object should be shifted in x- and y-direction relative to the detector for different realizations. This is implemented by including random shifts in the objects x- and y-positions that follow random normal distributions with zero mean and a standard deviation of 1 $Pixel$.

3.2.3 Results

The goal of virtual mammography is to simulate images that are similar to real images. As a reference, a set of 16 images acquired at the German reference center in Münster by a Hologic®

Table 3.1: Free parameters that are used to control the virtual mammography to simulate radiographic images of digital objects.

Parameter	Description	Value
Anode material	Target material of the X-Ray tube anode	Mo, W
Tube voltage	Tube voltage to accelerate free electrons for X-Ray production.	$18 - 40\ kV$
Time-current product	Exposure produced by the X-Ray tube	$100 - 140\ mAs$
Filter	Material and thickness to filter initial spectrum for specific beam quality	Rh ($0.025\ \mu m$) Rh ($0.05\ \mu m$) Mo ($0.03\ \mu m$)
Source to imager distance	Distance between detector and X-Ray tube	$\approx 700\ mm$
Source to object distance	Position of the object	$\approx 600\ mm$
N_x	Number of pixels of the detector in x dimension	e.g. $2323\ px$
N_y	Number of pixels of the detector in y dimension	e.g. $3654\ px$
Pixel pitch	Size and distance of neighbored pixels	$0.04 - 0.07\ mm/px$
SNR	Noise level to control image quality	$3 - 10$
SPR	Scatter level to control contrast to noise ratio	≈ 0.18 equiv. thickness $6\ cm$
Number of images	Total number of images	e.g. 200
Number of random positions	Number of applied random shifts	e.g. 10

Selenia® Dimensions® mammography system using a tube voltage of 31 kV for a tungsten anode with a 0.05 μm Rh filter and a time-current product of 139 mAs[1].

The digital models of the individual cells of the CDMAM phantom were taken to simulate images. The simulation parameters were chosen in accordance with the acquisition parameters from the real images and are listed in Table 3.2. The noise parameters were chosen such that the resulting simulated images matched the real images visually closely.

One of the resulting simulated CDMAM cell images is displayed in Figure 3.3 together with the corresponding cell from a real image. The visual similarity is obvious. The plotted pixel intensities in the middle of the image demonstrate that the difference between lesion and background is the same for simulated and real images, whereas the mean pixel values of background and lesion of the real image are slightly higher than in the simulated image. This

[1] At this point I want to thank Mr. A. Sommer from the German reference center in Münster for providing us these reference images.

Table 3.2: Parameters of simulated images that are visually in good agreement with a set of real images.

Parameter	Value
Anode Material	W
Tube voltage	31 kV
Time-current product	139 mAs
Filter	Rh (0.05 μm)
Source to imager distance	700 mm
Source to object distance	629 mm
Pixel pitch	0.07 mm/px
SNR	5.7
SPR	0.18
Number of images	500
Number of random positions	10

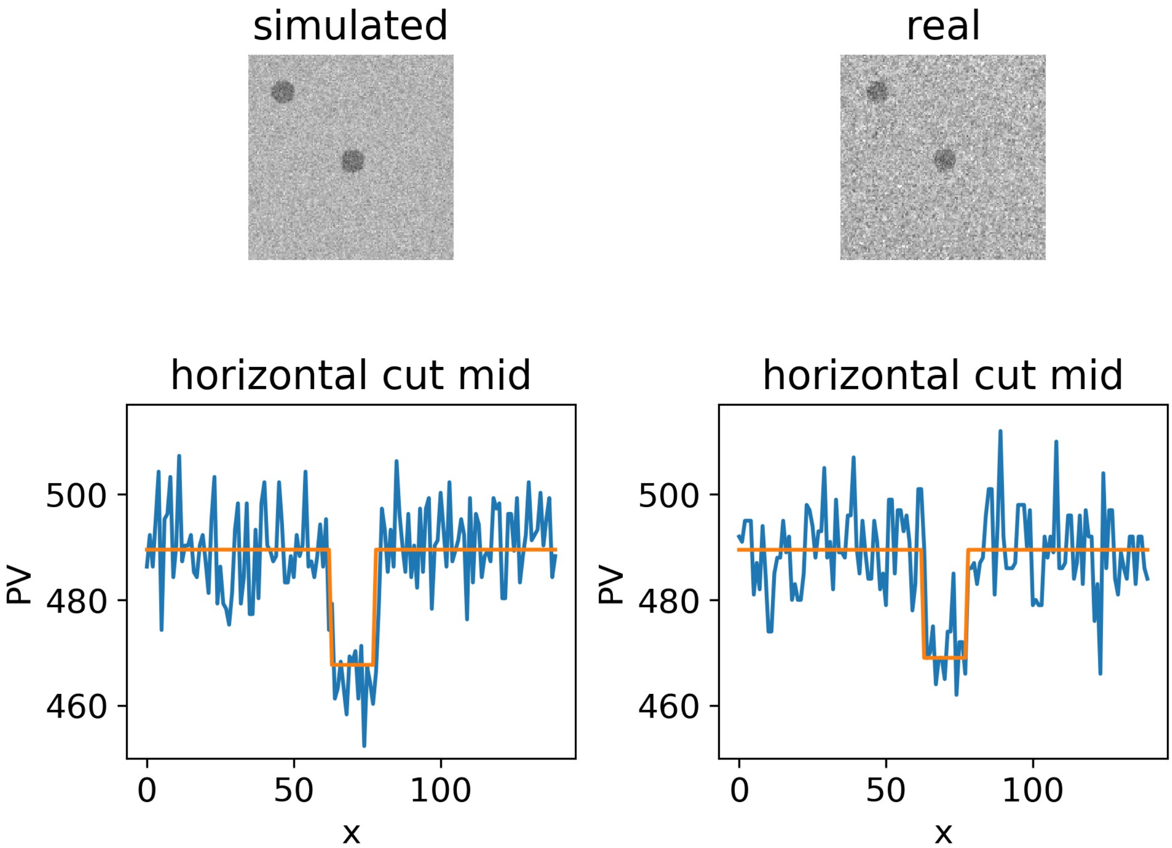


Figure 3.3: Comparison of an exemplarily simulated image with the corresponding cell of a real image of a CDMAM phantom The images are shown in the top with the simulated cell on the left and the real cell on the right. The plots below show the pixel intensities (blue) for a horizontal cut through the middle of the cells. The red curve indicates the mean of the lesion and the background regions, respectively for both the simulated and the real image.

might be caused by the choice of the SPR and the normalization of the detector. Also, the discs appear a little bigger in diameter in the simulated images. A reason for the observed discrepancy may be that a real CDMAM phantom is manufactured with a production error and thus the diameters can vary significantly from the true value (Borowski et al., 2012). For

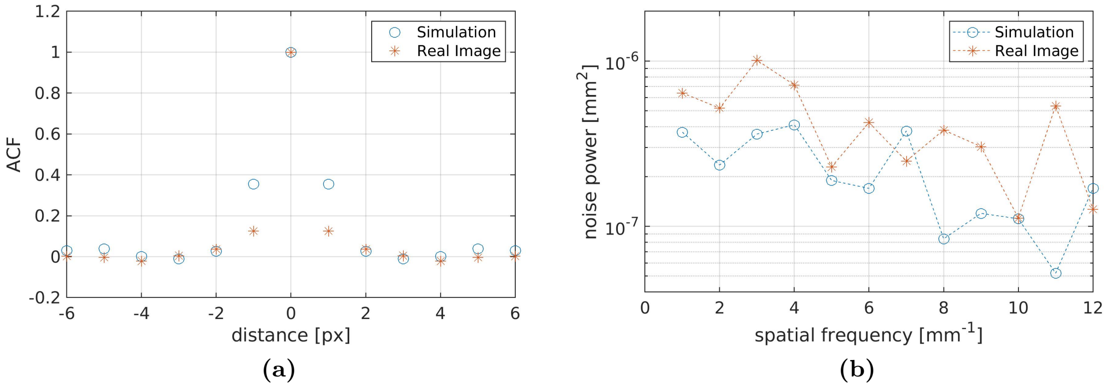


Figure 3.4: (a) Noise auto correlation function for simulated images (blue) and real images (red). (b) Normalized noise power spectrum for the simulated images (blue) and the real images (red). Figure from Kretz et al. (2019).

our digital voxel phantom, the actual diameters according to the specification were considered. Note that the diameter difference between the real images and our simulations is irrelevant if we compare different methods applied on the latter. However, this difference influences the results obtained by the application of the same method on real and simulated images.

Especially the noise characteristics are important to achieve a satisfactory similarity between simulated and real images. Image noise can be characterized by different properties. The simplest parameter is the standard deviation over all image pixels or a sub-region of the image, respectively. The standard deviation of the simulated images was chosen such that it matched the standard deviation of the real images. Besides this, the spatial structure of the noise and hence its correlations are also an important indicator.

The autocorrelation function (ACF) expresses the correlation of the noise between different pixels of the image. It can be applied as a measure of the correlation as a function of the distance between pixels. For an image B, the autocorrelation function can be calculated by shifting the image B by d pixels and calculating the 2-dimensional correlation ACF for the image B and the shifted image B^d

$$ACF(d) = \frac{\sum_m \sum_n \left(B_{mn} - \overline{B}\right)\left(B^d_{mn} - \overline{B^d}\right)}{\sqrt{\left(\sum_m \sum_n \left(B_{mn} - \overline{B}\right)^2\right)\left(\sum_m \sum_n \left(B^d_{mn} - \overline{B^d}\right)^2\right)}}$$

with mean images $\overline{B}$ and $\overline{B^d}$.

Figure 3.4a shows the auto-correlation for a 65×65 *Pixels* region in a simulated image (blue) and a real image (red), respectively. The correlation length in both cases is about 1 *Pixel*, where the simulated images show a slightly higher correlation length. Pixels with a distance of more than two Pixels appear to be uncorrelated.

The spatial structure of the noise can be further described by the noise-power spectrum (Bendat and Piersol, 2011). In a finite region $\tilde{b}$ of the image and a discrete set of frequencies, a sample j of the noise-power spectrum can be obtained as

$$NPS^{(j)}(k,l) = \frac{x_0 y_0}{N_x N_y} \left| \sum_{m,n} \tilde{b}^{(j)}(m,n) e^{-2\pi i (k \cdot m/N_x + l \cdot n/N_y)} \right|^2 ,$$

with x_0 and y_0 as characteristic length scales and N_x and N_y defining the dimensions of the region $\tilde{b}$ in x- and y-direction, respectively (ICRU, 2009). This formula describes the square of the absolute Fourier transformation of the image region $\tilde{b}$ multiplied by a length factor. An estimate of the noise-power spectrum can be calculated by averaging over N samples (ICRU, 2009)

$$\overline{NPS}(k,l) = \frac{1}{N} \sum_{j=1}^{N} NPS^{(j)}(k,l).$$

Figure 3.4b visualizes the normalized noise-power spectrum for a 25×25 *Pixels* sub-region of a real image (red) and a simulated image (blue), for a characteristic length of $x_0 = y_0 = 0.06\ \mu m$. It can be seen that the normalized noise-power spectrum of simulated and real images appear similar. This shows that the simulated images exhibit a sufficiently similar noise in terms of their auto correlation and their noise-power spectrum, even though noise effects were treated only phenomenologically. Therefore, our virtual mammography appears to produce realistic simulations of images of the CDMAM phantom.

3.2.4 Discussion

Virtual mammography was implemented and optimized to simulate realistic mammography images of the CDMAM phantom. Our results show that the simulated images have similar noise properties as a set of real reference images. The images were simulated by subsequently degrading primary images that were modeled following state-of-the-art virtual mammography (Elangovan et al., 2014; Mackenzie et al., 2017). Simulations in general have to find a compromise between the closeness of the simulation to the reality and the computational cost (see e.g. (Elangovan et al., 2014)). The degradation was modeled only phenomenologically to efficiently simulate mammography images. Blurring, scattering and detector noise were considered as degradation effects since they largely determine the resulting image quality (ICRU, 2009). Blurring occurs because the X-Rays do not originate from a point source but from multiple spots from a source area. To capture this effect correctly, one has to model a ray for every detector pixel and every origin and calculate their transmission profile through the phantom. The number of rays increases linearly with the number of origin points, thereby increasing the computational cost. Instead, in our approach blurring was modeled by applying a Gaussian filter to the primary image. This operation is much faster and captures the reduction of the spatial resolution closely. Scattering and detector noise can be simulated more correctly by using time consuming Monte Carlo experiments (Dance and Day, 1984). Again, the approximation of the scattering effects via convolutional filtering and the phenomenological treatment of noise save time while maintaining a close similarity between simulated images and real images. Basic noise characteristics of the simulated images and a set of reference real images matched in terms of noise standard deviation, auto correlation function and normalized noise power spectrum. This indicate that the phenomenological virtual mammography presented here is a powerful tool to simulate images that are similar to real mammography images.

Conducting real clinical trials is both time and money consuming. Virtual mammography, on the other hand, is a good alternative to simulate realistic images to conduct large virtual trials at low cost. For image quality measurements in mammography, several images of technical phantoms such as the CDMAM phantom are recorded (Perry et al., 2006). It has been shown that the CDMAM phantom is produced with a manufacturing tolerance which is high enough to influence image quality measurements (Fabiszewska et al., 2016). Such an uncertainty is critical for a reliable assessment and the development of automatic procedures to determine image quality. The virtual specimen, on the other, is known exactly. Furthermore, virtual mammography can be used to create large data sets, which allow data-intensive methods, such as deep learning, to be applied.

3.3 Parametric Model Observer for Image Quality Estimation

In this section the application of a recently published parametric model observer for mammography image quality assessment is proposed. Virtual mammography, as described above, is used to assess the proposed alternative procedure and compare it with a reference method. In chapter 2.1.3, the current automatic procedure, referred to as the EUREF guideline procedure, was introduced which estimates mammography image quality in terms of the contrast-detail curve. Contrast-detail curves summarize local contrast information by showing the minimum thickness of the object to be detected with a specified probability as a function of the object's diameter. In order to obtain this curve, the guideline procedure recommends that multiple (at least 16) images of the CDMAM phantom are acquired and pre-processed by segmenting the phantom into its individual cells and aligning a phantom mask to determine the artificial lesion positions in each cell. The contrast-detail curve is determined by applying an observer to localize these lesions in each cell individually. By doing this for different images of the same cell, a proportion of correctly detected lesions for each combination of diameter and thickness can be recorded, which is then recast into a contrast-detail curve by choosing a minimum threshold for the probability of correctly detecting the lesion position. The figure of merit of this four-alternative-forced-choice (4AFC) procedure is the proportion of correctly detected discs in a series of image. The 4AFC approach was originally designed for a human observer in order to reduce the probability of just guessing the correct position. The contrast-detail curve obtained by the EUREF guideline procedure relies on an observer's ability to correctly localize a disc in one out of four corners. The uncertainty of the curve determined by this approach is derived assuming a binomial distribution and thus strongly depends on the number of images that are used. Therefore, the EUREF guideline procedure requires the acquisition of more than one image and its uncertainty depends strongly on the number of images (Mackenzie et al., 2017).

An advantage of mathematical model observers is their objectivity. The use of a mathematical observer prevents a possible bias of the results due to learning effects which may be present with human observers. However, many established methods for objective task-based images quality assessment incorporate mathematical model observers to solve a binary task, such as lesion detection (Barrett and Myers, 2013; Barrett et al., 2006). As described earlier, the AUC is a meaningful figure of merit to quantitatively measure a models predictive performance (Bradley, 1997) and is frequently used as a measure for image quality (Barrett et al., 1993, 1998) (cf. section 2.1.1).

Recently, a parametric model observer was developed and applied to images obtained using computer tomography (Anton et al., 2018). This parametric observer assumes a simple geometry of the image and estimates the area under curve (AUC) (cf. Chapter 2) by an analytically derived expression. Here, an approach is explored where this parametric observer is employed to determine contrast-detail curves that are derived from the AUC as an established measure for image quality. The parametric model observer derives the AUC for every cell of the phantom separately. Similar to the EUREF guideline procedures, the proposed method requires the pre-processing of images. To be more specific, the whole phantom image has to be segmented into its individual cells.

Our study is restricted solely to the effect of employing different observers. Using the virtual mammography, the cells of the CDMAM phantom are simulated individually to avoid the need of implementing an error-prone pre-processing. For all simulations, the position of the discs is known exactly.

3.3.1 Parametric Model Observer

The individual cells of the CDMAM phantoms are of a relatively simple structure. Two discs are placed on a flat background. The decision whether a lesion is present in an image or not can be taken by model observers. Model observers can be divided into two groups, namely, non-parametric and parametric model observer. Non-parametric observers, such as the Channelized Hotelling Observer (CHO) (Myers and Barrett, 1987) or the nonprewhitening matched filter (NPW) (Barrett et al., 1993) assume underlying statistical distributions in the images without parametrizing the distribution of lesion and background regions. Based on a set of training images, a test statistic can be calculated for test images. A suitable observer allows a specific task to be solved according to this test statistic. For the task of lesion detection (classification) this test statistic should make it possible to distinguish between images with lesions and images without a lesion present. For such a binary task, the performance of an observer can be assessed via ROC analysis, as described in chapter 2. Parametric approaches assume in addition specific parametrizations of the lesion and background regions as well as the image noise. Since more prior information is thus used, parametric model observers for image quality assessment require fewer images than non-parametric observers (Anton et al., 2018).

In Anton et al. (2018) a parametric model observer was developed which assumes that a single image consists of a flat background with mean μ and the signature of a lesion with mean θ. The noise in the image is assumed to be independent Gaussian noise with constant standard deviation σ. Figure 3.5a visualizes the basic assumptions of the parametric observer.

The background region x_0 can be described by a vector of length n_0, while x_1 denotes a vector of length n_1 that includes all pixels from the lesion region

$$x_0 \sim \mathcal{N}\left(\mu \cdot 1_{n_0}, \sigma^2 I_{n_0}\right),\ x_1 \sim \mathcal{N}\left(\theta \cdot 1_{n_1}, \sigma^2 I_{n_1}\right),$$

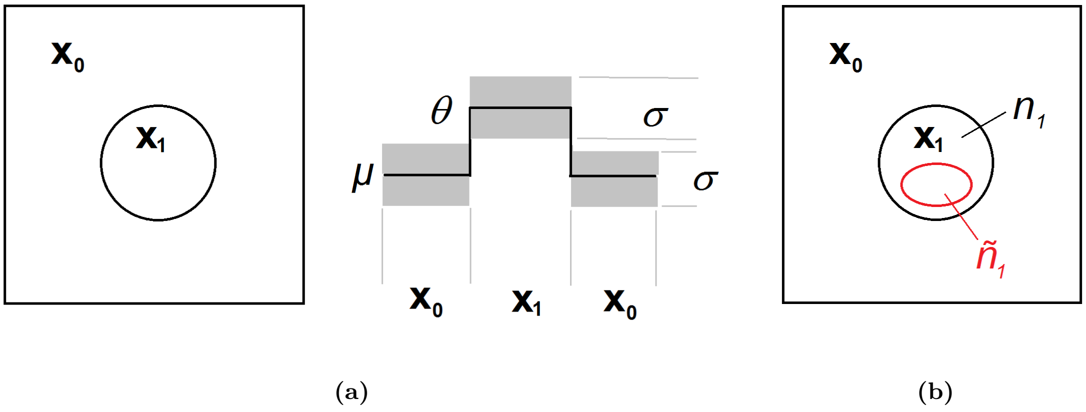


Figure 3.5: (a) The image comprises a background region x_0 and a lesion region x_1. The background and lesion regions have mean pixel intensities of μ and θ, respectively. The noise in the image is assumed to be the same for background and lesion regions with a standard deviation σ. (b) Only a subset of $\tilde{n}_1$ of all pixels in the lesion region are used for testing . Figure from Anton et al. (2018).

where 1_n stands for an $n \times 1$ all-ones matrix and I_n denotes the identity matrix of dimension n. Then an analytical expression for the AUC can be expressed as

$$AUC = \phi\left(\frac{1}{\sqrt{2}}\frac{|\mu - \theta|}{\sigma/\sqrt{n_1}}\right),$$

where ϕ denotes the standard cumulative distribution function of the standard normal distribution (Anton et al., 2018). The analytical expression of the AUC allows us to determine its value directly, by estimating the mean lesion intensity θ, the mean background intensity μ and the standard deviation of the noise σ. The EUREF guideline procedure requires multiple images, whereas the AUC can - at least in principle - be estimated for single images.

The AUC depends on the number of pixels n_1 in the lesion region. It can be tuned by using only a subset $\tilde{n}_1$ of the lesion pixels as indicated in Figure 3.5b. A smaller fraction $\tilde{n}_1 = f \cdot n_1$ with f between 0 and 1 decreases the uncertainty of AUC (Anton et al., 2018). The estimate of the AUC can then be determined as follows

$$A\hat{U}C = \phi\left(\frac{1}{\sqrt{2}}\frac{|\hat{\mu} - \hat{\theta}|}{\hat{\sigma}/\sqrt{\tilde{n}_1}}\right) = \phi\left(\frac{1}{\sqrt{2}}\frac{|\hat{\mu} - \hat{\theta}|}{\hat{\sigma}/\sqrt{f \cdot n_1}}\right),$$

with the estimates for the mean of the signature of a lesion $\hat{\theta}$ and the mean background $\hat{\mu}$ and a pooled estimate of the noise $\hat{\sigma}$ as

$$\hat{\sigma} = \sqrt{\frac{1}{n_0 + n_1 - 2}\left((n_0 - 1)\,\sigma_0^2 + (n_1 - 1)\,\sigma_1^2\right)},$$

where σ_1 and σ_0 are the standard deviations in the lesion and background region, respectively. Note that while the AUC is estimated with the sub-region $\tilde{n}_1$, the estimates $\hat{\mu}, \hat{\theta}, \hat{\sigma}$ are obtained on the complete background and lesion regions with n_0 and n_1 pixels, respectively.

A drawback of the procedure is that the number of pixels might be too low. This can be avoided by pooling p images with the same parameters. This pooling is done by concatenating the vectors x_0 and x_1 of p images. Pooling allows the estimations of μ, θ and σ to be obtained more robustly. The described parametric model observer is introduced as the modified simple observer for pooled images (MSOpi) in (Anton et al., 2018) and will be further employed for the assessment of image quality in mammography.

3.3.2 Modified Mammography Image Quality Assessment

In this section a procedure is developed that utilizes the parametric model observer MSOpi presented above to determine contrast-detail curves from images of single cells of the CDMAM phantom.

The MSOpi parametrizes each CDMAM cell image in the same way as the EUREF observer and makes use of the simple geometry in images of CDMAM cells. In contrast to the EUREF guideline procedure, the MSOpi determines the AUC as a figure of merit. This transforms the 4AFC disc localization in each cell into a two-alternative-forced choice (2AFC) approach of detection a lesion in each cell. The AUC as figure of merit is an established measure in task-based image quality assessment (Barrett and Myers, 2013). Under the assumption of independent Gaussian noise, an analytical expression allows the AUC to be calculated by estimates pixel intensity in the lesion and background region, an estimate of the standard deviation of the noise and the number of pixels that form the lesion. The AUC is calculated for every cell in the CDMAM phantom and measures how accurate the observer can distinguish between lesion and background. The AUC increases with increasing difference of the pixel intensity in the lesion and background region. This image contrast, on the other hand, is defined by the thickness of the gold disc. Therefore, the AUC depends on the thickness of the gold disc and choosing a threshold for the AUC allows to determine the thickness, for which the observer can detect the lesion with sufficient accuracy. Computing this thickness, where the AUC reaches the limit value, for every diameter individually directly results in the contrast-detail curves.

Since the AUC is estimated per cell, our modified procedure requires the same pre-processing steps as the EUREF guideline procedure. Also the assumption of knowing the lesion region exactly is required. The pre-processing is avoided by simulating the images of single cells. This enables us to solely test the observer performance and compare our method against the current practice. A CDMAM cell comprises two artificial lesions: one in the center and the other one in one of the four corners. Both of these regions are merged together for the estimation of $\hat{\theta}$. All remaining pixels are used for the estimate $\hat{\mu}$ of the background area.

The AUC estimate of the MSOpi depends on the number of pixels and hence on the diameter of the gold discs in the particular cell. To express the dependence of the AUC on the number of pixels by its dependence on the true disc diameter d, we modify the fraction f of pixels as

$$f = f(d) = \gamma \frac{d^2}{d_{min}^2} \,,$$

where d_{min} stands for the minimum diameter of interest, which is $d_{min} = 0.1\ mm$ in our case, in accordance with the EUREF guideline. Hence, the fraction f scales with d^2 which is proportional to the number of pixels, but does not change when several pictures of the same cell are pooled. The factor γ is a free factor and can be used to shift the resulting contrast-detail curves. Hence, an AUC estimate is expressed by

$$A\hat{U}C = \phi\left(\frac{1}{\sqrt{2}}\frac{|\hat{\mu}-\hat{\theta}|}{\hat{\sigma}^2}\frac{\sqrt{\gamma}d}{d_{min}}\right) . \tag{3.2}$$

For each cell of the CDMAM phantom, the AUC estimate is calculated. For a fixed diameter d, the AUC estimate depends on the difference between lesion and background and hence on the thickness of the gold disc. In accordance with the EUREF guideline procedure the dependence of the AUC for a given diameter d on thickness t is modeled as a psychometric curve

$$AUC(t) = \frac{0.5}{1+\exp\left[-\zeta\left(C(t)-C_T\right)\right]} + 0.5\,,$$

where $C(t)$ denotes the contrast of a gold disc with thickness t. ζ and C_T stand for free parameters that are determined using a nonlinear least square fit procedure. The AUC varies between 0.5 for very small thicknesses and 1 for large thicknesses. We define a critical AUC value of 0.75 which marks the center between pure guessing and certainty for a binary task. By estimating the free parameters ζ and C_T we can determine the thickness t_{cd} for which $AUC(t_{cd}) = 0.75$ for each diameter. Plotting theses thicknesses for each diameter yields the contrast-detail curve.

3.3.3 Data

The virtual mammography was used to simulate different sets of images of individual cells of the CDMAM phantom. Each cell was simulated individually with a tube voltage of 31 kV, a time-current product of 139 mAs and a noise standard deviation $\sigma_0 = 5.7$. A detailed list of the simulation parameters was given earlier in Table 3.2. For each cell, 500 images were simulated. This enables the estimation of the variability of our suggested approach in comparison with the EUREF guideline procedure. For the 205 cells that are present in the CDMAM phantom this results in a total of $205 \cdot 500 = 102,500$ simulated images. Additionally, images with different noise properties were simulated by varying the SNR. Two additional noise levels $\sigma_l = \frac{1}{2}\sigma_0$ and $\sigma_h = \frac{3}{2}\sigma_0$ were chosen to simulate higher and lower image qualities in order to ensure that our proposed approach is suitable to correctly detect the change in image quality.

3.3.4 Results

The simulated images with noise level σ_0 were found to be visually very similar to real images acquired under comparable conditions. The suggested approach that employs the MSOpi for contrast-detail curve estimation can be applied to single images, whereas pooling images ensures a more accurate estimation of μ, θ and σ. Figure 3.6a shows the logistic regressions that are used to model the dependence of the AUC on the thickness for each diameter for 5 pooled images that were all simulated using the noise level σ_0.

The corresponding contrast-detail curve can be derived from the psychometric cuves in this case by choosing a minimum threshold for the AUC (0.75). Figure 3.6b shows the corresponding contrast-detail curves for the psychometric curves in Figure 3.6a.

As mentioned in section 2.1.3, the psychometric curves can be stabilized by smoothing the

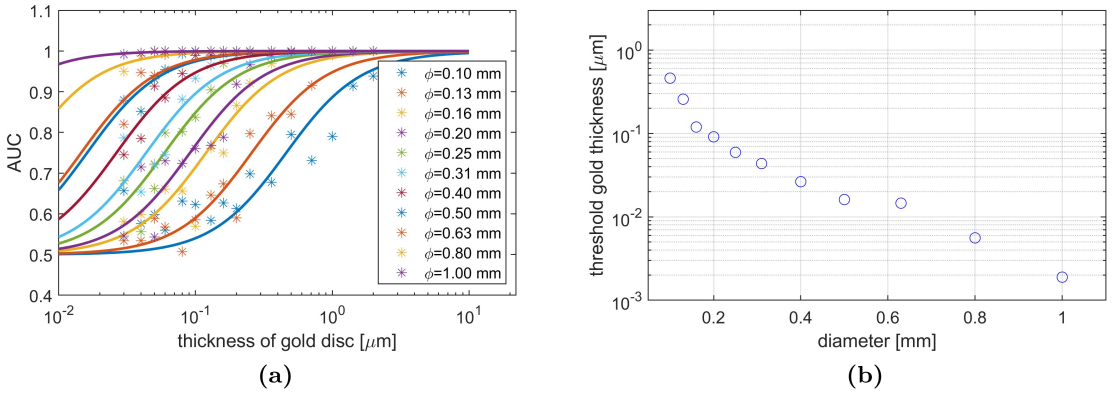


Figure 3.6: (a) Example of the logistic regressions for the AUC of the MSOpi. For each diameter the dependence of the AUC on the thickness is plotted from 5 pooled simulated CDMAM images. (b) Corresponding contrast-detail curve for a threshold of 0.75 for the AUC. Figure from Kretz et al. (2019).

data of the proportion of correctly detected lesions. However, this smoothing operation can be done for the AUC in the same way. It was found that this smoothing operation equally influences the logistic curves obtained by both the procedures. The difference of the methods is not altered by the smoothing operation. Therefore, we report results without smoothing in the following.

In order to compare the proposed alternative with the EUREF guideline procedure, both methods were applied repeatedly using the same simulated images. For the EUREF guideline procedure 16 images were randomly drawn from the set of 500 images with noise level σ_0 and for the MSOpi procedure 5 and 16 images were randomly drawn and used to determine the contrast-detail curve, respectively. This procedure was repeated 5000 times, each time drawing different images. Only for a few repetitions (< 10), outlying results were obtained from both approaches because of a bad termination of the nonlinear fit. These outliers were removed from the results. Table 3.3 lists the quantitative results for the mean and 95% coverage intervals for the repeated applications after outlier removal.

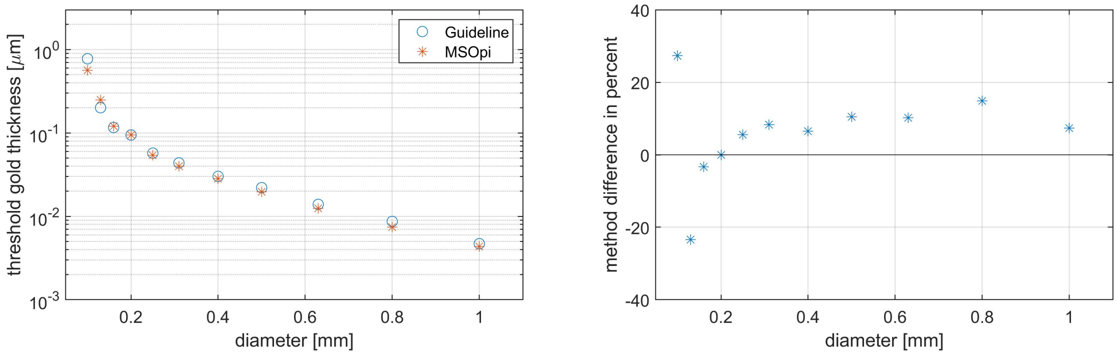
Figure 3.7a displays the mean contrast-detail curve approach for the EUREF guideline procedure with 16 images (blue) and the MSOpi procedure with 5 images (blue) along with a plot of their difference in Figure 3.7b.

The mean contrast-detail curves appear very similar, but they are not identical. The difference plot 3.7b highlights that both methods mainly deviate for the two smallest diameters. Note that even though the difference between the methods appears to be rather high it is within the limits of uncertainty, cf. Table 3.3. For all remaining points of the contrast-detail curve, the method difference is significantly smaller than the corresponding uncertainties.

Depending on the noise, the signatures of the artificial lesions can be detected better or worse in the image. Increasing noise should decrease the image quality, which in terms of the contrast-detail curve would result in a shift on the y-axis to higher values. In order to check whether the proposed approach captures this behavior correctly, two different noise levels σ_l and σ_h were considered in the simulation in addition to σ_0. Again, the two different methods were repeatedly applied to 16 random sample images in case of the EUREF guideline procedure and 5 random image samples for the MSOpi procedure. The mean contrast-detail

Table 3.3: Mean threshold gold thickness for the EUREF guideline approach from 16 images and for the proposed approach from both 5 and 16 images. Uncertainties indicate the intervals containing (pointwise) 95 % of all contrast-detail curves. Table from Kretz et al. (2019).

	mean threshold gold thickness [μm]		
diameter [mm]	EUREF 16 images	MSOpi 5 images	MSOpi 16 images
0.10	0.780 ± 0.210	0.560 ± 0.150	0.568 ± 0.089
0.13	0.201 ± 0.038	0.248 ± 0.042	0.248 ± 0.025
0.16	0.116 ± 0.022	0.120 ± 0.023	0.120 ± 0.014
0.20	0.094 ± 0.018	0.095 ± 0.019	0.094 ± 0.010
0.25	0.057 ± 0.011	0.054 ± 0.011	0.054 ± 0.006
0.31	0.044 ± 0.008	0.040 ± 0.009	0.040 ± 0.005
0.40	0.030 ± 0.007	0.028 ± 0.007	0.028 ± 0.004
0.50	0.022 ± 0.006	0.020 ± 0.005	0.020 ± 0.003
0.63	0.014 ± 0.005	0.013 ± 0.004	0.013 ± 0.002
0.80	0.009 ± 0.005	0.007 ± 0.003	0.007 ± 0.002
1.00	0.005 ± 0.005	0.004 ± 0.003	0.004 ± 0.001

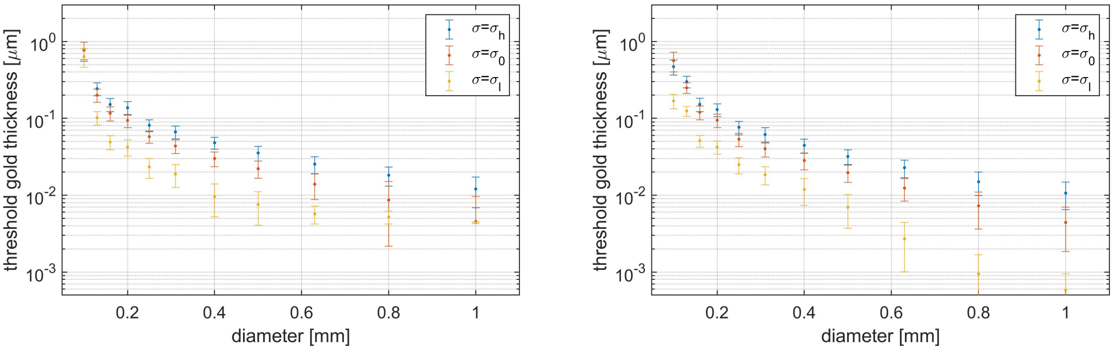


(a) **(b)**

Figure 3.7: (a) Mean contrast-detail curve as obtained by repeatedly applying the EUREF Guidelines procedure to a random sample of 16 images (blue) and the MSOpi procedure to a random sample of 5 images. Images were simulated using the virtual mammography with a noise level σ_0. (b) Difference of the contrast-detail curve relative to the EUREF Guidelines curve. Figure from Kretz et al. (2019).

curves, along with an error bar that represents the intervals containing pointwise 95% of all contrast-detail curves, are shown in Figure 3.8a for the three different noise levels for the EUREF guideline procedure and in Figure 3.8b for the MSOpi procedure.

In both cases the results show the expected behavior: the contrast-detail curve indeed shifts to higher thicknesses with increasing noise level. The contrast-detail curves for both approaches appear very similar for noise levels σ_h (blue) and σ_0 (red), whereas the contrast-detail curve for the lowest noise (yellow) shows large deviations for the smallest diameter $d = 0.1$ mm and diameters beyond 0.5 mm. Except for the smallest diameter, the contrast-detail curves of the EUREF Guidelines procedure in Figure 3.8a stay clearly beyond the contrast-detail curves for the higher noise levels, which raises some doubt as to the determination of the results for the smallest diameter. For the MSOpi procedure, however, the contrast-detail curve for low noise levels clearly stays below the results for higher noise levels, which appears to be reasonable.



(a) **(b)**

Figure 3.8: (a) Mean contrast-detail curves for repeated application of the EUREF Guidelines procedure for 16 images simulated with three different noise levels. (b) Same results for the MSOpi procedure using 5 images. Error bars indicate intervals containing (pointwise) 95% of the contrast-detail curves. Figure from Kretz et al. (2019).

3.3.5 Bayesian Estimation

So far, the proposed procedure determines point estimates of the AUC that are used to compute a contrast-detail curve. Repeated applications of our method as well as the EUREF guideline procedure yield slightly different results. This can be explained by statistical fluctuations in the data. To get an idea of the fluctuation of a result that is achieved with a specific measurement its uncertainty has to be determined. A measure of the uncertainty can be derived using Bayesian statistics (Gelman et al., 2013). The Bayesian approach assumes underlying distributions for every parameter. Also for the figure of merit, a distribution is assumed that initially is pre-determined by a prior distribution that summarizes all prior knowledge about this value (D'Agostini, 2003). Furthermore, a statistical model is introduced for the data. The statistical model describes the sampling distribution, and the observed data are viewed as one realization of this sampling distribution. The probability for the observed data viewed as a function of the unknown parameters is known as the likelihood function (Berger et al., 1988). In a Bayesian inference the prior distributions for all unknowns are updated in view of the observed data, resulting in the posterior distribution (Gelman et al., 2013). According to Bayes theorem (cf. chapter 2), the posterior is proportional to prior times the likelihood. The posterior distribution encodes one's state of knowledge about the figure of merit and can be viewed as the most comprehensive uncertainty characterization of the result (Gelman et al., 2013) which allows one to carry out probabilistic statements. Derived quantities such as the posterior mean and posterior standard deviation can be taken as point estimate and its associated standard uncertainty.

In case of our proposed approach, the Bayesian interpretation implies that all parameters μ, θ, σ and AUC are described by a distribution. The quantities $\hat{\mu}, \hat{\theta}, A\hat{U}C$ and $\hat{\sigma}$ can be derived as point estimates from the corresponding posterior distribution.

In Khanin et al. (2018), a simple Monte Carlo algorithm is provided for an efficient generation of samples from the posterior distribution $p(AUC|\mathcal{D})$ given the data $\mathcal{D}$. In the case of the CDMAM phantom, this enables us to compute the posterior function $p(AUC|t, d)$ for each cell with disc diameter d and lesion thickness t. For very small thicknesses and hence low contrasts, this posterior distribution will have a peak at 0.5 while for bigger contrasts, the distribution will shift towards a value of 1. Figure 3.9 shows this behavior for the cells with a disc diameter of 0.2 mm.

As it can be seen, the posterior distribution shifts towards 1 the higher the thickness of the gold disc in the cell. The higher the AUC, the better the detectability of the lesion in the cell. The conditional probability $p(t \mid AUC, d)$ describes how the thickness t of the discs varies for different values of AUC. Contrast-detail curves can then be expressed by the distribution $p(t \mid AUC = \tau, d)$, where τ denotes a limit value for AUC above which the decision "lesion detectable" is assigned to a cell of the phantom. This distribution $p(t \mid AUC = \tau, d)$ is the distribution of the critical thickness for a given diameter, i.e. a distribution along the standard contrast-detail curve in the y-direction. The mean of this distribution can be viewed as the value of the contrast-detail curve at diameter d, and its variance as the associated standard uncertainty. The Bayes theorem (cf. chapter 2) states that the sought conditional distribution is given by

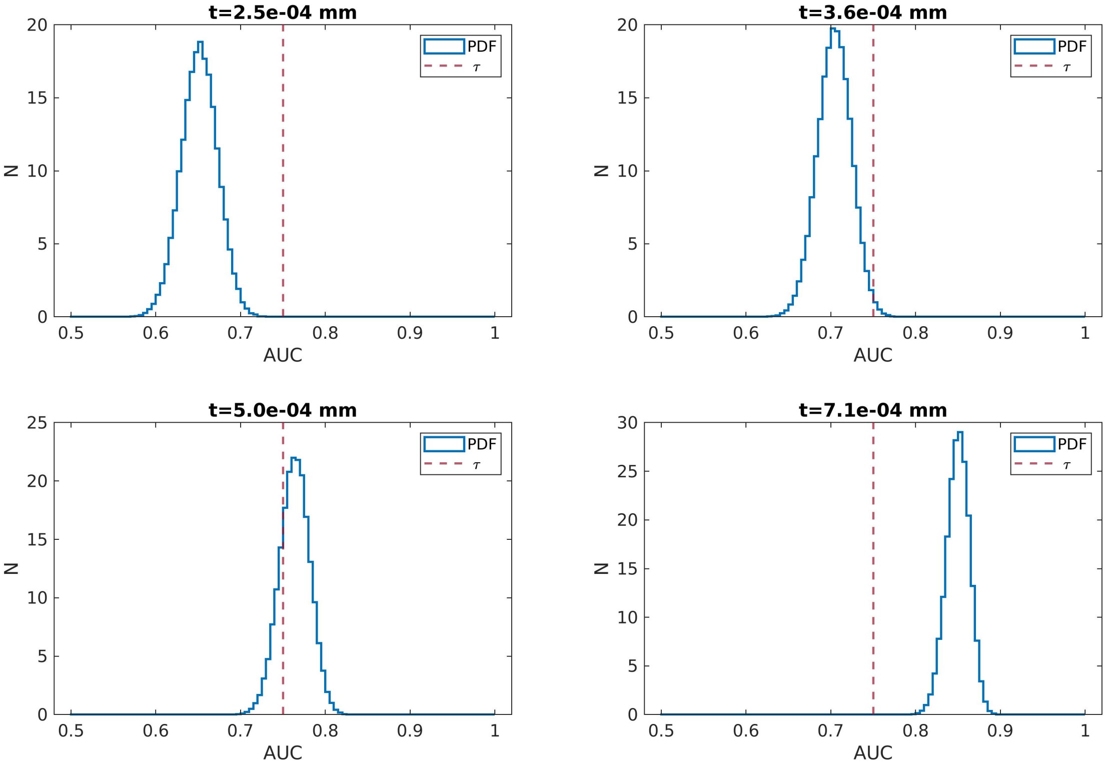


Figure 3.9: Four examplaric histograms of the Bayesian posterior of the AUC for a disc diameter of 0.2 mm. The red line visualizes the limit value τ.

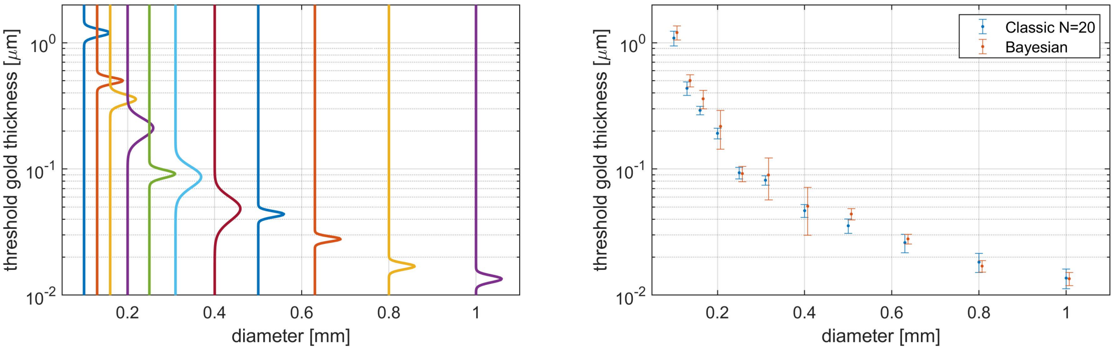
$$p(t \mid AUC = \tau, d) = \frac{p(AUC = \tau \mid t, d)p(t \mid d)}{p(AUC = \tau \mid d)} \propto p(AUC = \tau \mid t, d)p(t)\,,$$

where $p(t \mid d) = p(t)$ is assumed. Choosing a non-informative prior $p(t) \propto 1$ then enables us to determine the contrast-detail curve via the distribution $p(t \mid AUC = \tau, d)$ by recording $p(AUC = \tau \mid t, d)$ for every cell. The limit value τ is indicated in Figure 3.9 as a red line. For small t, the values $p(AUC = \tau | t, d)$ will be close to zero and peak at a certain thickness before they decrease again for higher values of the thickness. For each diameter, $p(t \mid AUC = \tau, d)$ is evaluated for all (≈ 12) existing values of the thickness in the phantom resulting in a sparse representation of $p(t \mid AUC = \tau, d)$. A Gamma distribution

$$f(x, a, b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{\frac{-x}{b}} , \tag{3.3}$$

with a shape parameter a and a scale parameter b is fitted to model $p(t \mid AUC = \tau, d)$ for each diameter individually. As mentioned above, this can be viewed as an approximate posterior distribution function for one point of the contrast-detail curve. The unknown fit parameters are estimated using a nonlinear least-squares procedure. Figure 3.10a shows the resulting waterfall plot of the posterior distribution functions for each point of the contrast-detail curve.

The Bayesian approach directly yields a point estimate of the contrast-detail curve with an associated uncertainty. Mean and variance of the Gamma distribution as in Equation 3.3 are given as $\mu = ab$ and $\sigma^2 = ab^2$. On the other hand, an estimate of the uncertainty for the "classical" MSOpi approach can be determined in terms of the standard deviation for repeated experiments, as described in section 3.3.2. Figure 3.10b depicts the results of the classic approach with 20 repetitions (blue) and the Bayesian approach (red). In both cases, 5 images were pooled to yield contrast-detail curves of comparable accuracy as those determined by the EUREF guideline procedure.



(a) (b)

Figure 3.10: Results of the Bayesian estimation of contrast-detail curves with pooling of 5 images. (a) Posterior distribution functions for each point of the contrast-detail curve. (b) Comparison of the contrast-detail curves derived by the classic MSOpi approach with 20 repetitions (blue) and the Bayesian approach (red). Error bars show extended (k=1.96) standard deviations.

Overall, the Bayesian approach accurately determines the contrast-detail curve. In three cases ($d = 0.16\ mm, d = 0.2\ mm, d = 0.5\ mm$), the mean threshold gold thickness as predicted by the Bayesian approach lies significantly higher than the threshold thickness determined using the standard MSOpi procedure. For diameters of $d = 0.2\ mm, d = 0.31\ mm$ and $d = 0.4\ mm$, the Bayesian approach yields significantly larger uncertainties. Note that due to the sparse sampling of the gamma distribution the fit is sometimes determined by only very few points (see Figures in Appendix A.2).

Nevertheless, the Bayesian approach yields accurate results and enables the determination of uncertainties on the basis of single images without the need of repeated analysis.

3.3.6 Discussion

Virtual mammography was used to assess a developed alternative procedure for contrast-detail estimation in mammography image quality assurance. The current EUREF guideline procedure was taken as a reference. This automatic procedure comprises several pre-processing steps before the actual determination of the contrast-detail curve (Karssemeijer and Thijssen, 1996).

Our proposed approach requires the same pre-processing of the CDMAM as the current automatic procedure. The 4AFC task, tackled by the current EUREF guideline procedure, is transformed into the binary task of lesion detection, following well established procedures of objective task-based image quality assessment (Barrett and Myers, 2013; Pepe et al., 2003). As a result, the contrast-detail curve is determined based on the AUC as a meaningful figure of merit which is frequently used to measure image quality (cf. Barrett and Myers (2013); Pepe et al. (2003); Barrett et al. (2015)). The AUC is determined by applying the parametric observer MSOpi to every cell of the CDMAM phantom individually. Since the EUREF guideline procedure uses a proportion of correctly detected lesions to determine the contrast-detail curve, it necessarily requires multiple images of the phantom to be taken for the analysis. Our novel approach, on the other hand can assess the AUC and hence the contrast-detail curve from - at least in principle - single images only. Using the virtual mammography, we could show that the novel approach with 5 images works as precise as the EUREF guideline procedure with 16 images in terms of the variability of the methods under repeated application. Thus, the workload for the quality assurance measurements can be reduced by the novel approach.

The EUREF guideline procedure estimates contrast-detail curves with large variabilities. Especially for curves close to the limit curve, the decision whether the quality assurance criteria are fulfilled becomes very unreliable. Mackenzie et al. (2017) demonstrate that for 16 images the variability of the contrast-detail curve for small diameters seems too high, led to the recommendation to record more than 16 images. A drawback of the current practice is that uncertainties cannot be estimated directly from 16 images but have to be obtained by repeated experiments from a large data set. The parametric model observer MSOpi, on the other hand, can be modified into a Bayesian estimation of the AUC, and the resulting posterior distribution can be used to characterize the uncertainty. A Bayesian procedure was introduced that determines an approximate posterior function of each point of the contrast-detail curve. Hence, an estimate of the uncertainty can be obtained from single applications of our method, rather than assessing the uncertainty in terms of the variance of repeated experiments. However, to approximate posterior probability distribution functions for each point of the contrast-detail curve requires that a gamma distribution is fitted to the corresponding data. Since only a limited number of thickness values for each diameter are present in the phantom, the actual posterior distribution function might be under-sampled. This can lead to an error-prone fitting procedure. Best results where achieved when the fit routine is initialized with parameters close to the actual value such that the fit routine converges rapidly. Nevertheless, contrast-detail curves can be determined accompanied with an uncertainty of each point. Especially for contrast-detail curves close to a limit curve, the uncertainty plays a critical role in reliably deciding whether the image quality is sufficient or not.

Our contribution can be seen mainly as a proof of principle that the determination of contrast-detail curves can be improved with the proposed alternative method. Furthermore, the uncertainty can be estimated directly along with the prediction of contrast-detail curves. However, to be used in a clinical routine, further validation and testing has to be done. For real images, the whole image pre-processing framework has to be implemented, such that the image is aligned correctly and thus can be segmented into the individual cells prior to the application of the parametric observer. Also, the lesions in the cell need to be aligned properly to reliable estimate the parameters for the calculation of the AUC. Even though the proposed procedure seems promising it still requires cumbersome pre-processing. Rather than implementing the novel procedure end-to-end for clinical applications and thus serve as an alternative software for automatic image quality assessment, a different method was explored that can assess image quality without any pre-processing of the phantom images.

Despite of the promising results presented in this chapter, several limitations are still inherent in our proposed approach. Similar to the automatic procedure of the EUREF guideline, our alternative method for determination of mammography image quality requires expert pre-processing of the images before the contrast-detail curve is determined. In addition, the uncertainty of the contrast-detail curve can be directly determined, but requires a nonlinear fitting routine, which can fail in cases where the data is too sparse, or when the starting values of the fitting routine are poor.

3.4 Summary

We developed a virtual mammography as a framework to simulate radiographic images of any object, with a known digital representation of its X-Ray attenuation properties. Following state-of-the-art simulation techniques a primary image is retrieved via backward ray casting and transformed into a noise free image. A phenomenological degradation was included to degenerate the primary, noise free images to make them visually comparable to real images with the same acquisition parameters. It is shown that even with a phenomenological treatment of image noise the simulated images closely resemble real images in terms of their noise characteristics.

Virtual mammography was then used to simulate images of the individual CDMAM phantom cells to develop and assess novel approaches for the determination of image quality in mammography. An alternative method was developed that utilizes a recently published parametric model observer to determine contrast-detail curves . The novel approach traces the contrast-detail curve back to the AUC as an established figure of merit for image quality assessment, which simplifies the assessment of image quality to a binary task. Extensive tests on simulated images demonstrated that the proposed alternative procedure requires fewer images than the current automatic procedure in the EUREF guideline. This reduces the workload for quality assurance measurements for mammography screening. One further benefit of the novel procedure is that the contrast-detail curve can be estimated along with an associated uncertainty. Especially when compared to a threshold, the uncertainty is a crucial parameter to decide whether a mammography unit has sufficient image quality for mammography screening or not.

Our results demonstrate the benefits of the proposed methods, while still expert pre-processing has to be done before the contrast-detail curve is determined. This can be interpreted as feature engineering in classical machine learning. Deep learning bypasses such feature engineering and derives representative features directly from the data. Therefore, deep learning could be useful for mammography image a ´quality assessment. Cumbersome image pre-processing can be avoided such that the contrast-detail curve is determined directly from the images. This will be explored in the next chapter.

4

Mammography Image Quality Assurance Using Deep Learning

Image quality can be estimated from images of technical phantoms, such as the CDMAM phantom. Local contrast information is expressed in contrast-detail curves by computing the minimum thickness of a lesion for a specific structure diameter, such that its signature in the image can be detected with a specified probability. For this purpose, the EUREF guideline employs an observer to local image patches to carry out the localization of a target. These local image patches are the individual cells of the CDMAM phantom. Therefore, the image of the whole phantom has to be pre-processed. Pre-processing an image and employing an observer to derive a test statistic can be interpreted as feature extraction in machine learning. Classical machine learning employs expert knowledge to engineer characteristic features to solve a certain task.

In this chapter our contribution is the *development of deep learning methods for mammography image quality assessment.* Using the virtual mammography, introduced in chapter 3, large data sets of realistic phantom images with known ground truth can be simulated. This facilitates the training of deep learning methods to assess image quality in mammography. Deep learning bypasses expert feature extraction as described earlier and is able to learn features that are representative to solve a certain task. Such methods could become very handy to avoid error prone pre-processing in mammography image quality assurance.

The benefits of the developed methods are compared with the automatic procedure that is currently suggested in the EUREF guideline for mammography quality assurance.

After a short review of the history of deep learning in medical imaging, two different newly developed observers will be introduced that utilize deep learning. One uses segmentation while the other one aims to directly derive the contrast-detail curve from the entire image. An extensive and thorough comparison of the trained deep learning observer with the EUREF guideline procedure is provided.
Most of the work presented in this chapter is published in (Kretz et al., 2020).

4.1 Deep Learning in Medical Imaging

Machine learning, especially its sub-field deep learning, has been applied in medical imaging for a long time, (cf. Litjens et al. (2017); Yassin et al. (2018); Sahiner et al. (2019)). In the 1960's, algorithms were employed to analyze radiographic images and help to interpret them (Meyers et al., 1964; Becker et al., 1964). With the increasing performance of algorithms and computer hardware, machine learning could be applied in the 1980's to detect and diagnose cancer in chest radiography (Giger et al., 1988) and mammograms (Chan et al., 1987). The success of these early machine learning applications, like in traditional machine learning methods in general, is based on the quality of some features extracted by an expert. After finding a suitable set of features, algorithms were trained to automatically solve a given task. Therefore, these methods require expert knowledge before they can be successfully implemented. Deep learning, on the other hand, is designed to engineer such features on its own. Such algorithms are applied directly to the data and automatically determine a suitable set of features during the training process. This approach is substantially different from traditional machine learning. Early researches that report the first usage of deep learning in medical imaging are the application of a CNN to classify regions in mammograms as either normal or malignous tissue (Sahiner et al., 1996) and the development of a massively trained artificial neural network (MTANN) to reduce the number of false positives in the detection of lung nodules (Suzuki et al., 2003). However, due to restrictions in computational power, these neural nets could not be designed to be very deep.

Over the last decade, improvements in computational power paved the way for the great success of deep learning applications especially in image analysis. At the same time, the number of applications in medical imaging has increased significantly. Deep learning is used for image segmentation to detect anatomical structures (Hu et al., 2016) or for lesion segmentation (Kline et al., 2017) and many more. Neural networks are used to classify lesions in various imaging modalities and body regions. To some extent, algorithms were already successfully trained to achieve expert performance, as in the case of skin lesion classification in (Esteva et al., 2017). Besides these applications, data driven methods are explored for image reconstruction, e.g. (Schlemper et al., 2017) and artifact reduction, e.g. (Jin et al., 2017).

Deep learning has also been applied for quality assessment in medical imaging. For example, a CNN has been trained to assess the goodness of the representation of some key structures in fetal ultrasound images (Wu et al., 2017). Their algorithm achieved comparable performance to the subjective image quality estimation of three medical doctors. In (Lee et al., 2018), a method is described where images of the lung are classified by a neural network according to their diagnostic image quality as rated by a radiologist. To the best of our knowledge, deep learning has not yet been applied to perform image quality assessment in mammography. According to the EUREF guideline, image quality is evaluated in terms of a contrast-detail curve. Hence, the underlying task is a multivariate regression where the method takes images as input and predicts a multivariate contrast-detail curve.

Additional to the application for classical imaging tasks, deep learning also influenced other areas like treatment planning in radiotherapy. Tseng et al. (2017) used deep reinforcement to develop automated radiation adaption protocols for lung cancer treatment. Other research focuses on the determination of the response to the treatment, where deep learning is used to predict the response after combined Chemo-Radiotherapy to identify patients that can benefit from these treatments (Bibault et al., 2018).

Deep learning is widely used in medical imaging. Several studies report impressive results for the application of deep learning in medical imaging, whereas most of these tasks are far from being solved (Summers, 2019). Developing a deep learning method that is robust and has a high repeatability is a challenging task. Especially in the medical imaging community, methods are often trained on data sets that are not publicly available and hence different methods are difficult to evaluate and to compare.

In this chapter, the usefulness of deep learning on the determination of image quality in mammography is explored. Deep learning can potentially bypass error-prone pre-processing leading to a more robust and accurate determination of image quality.

4.2 Semantic Segmentation Observer

As described above, deep learning can be applied in medical imaging in many different ways. Accordingly, deep learning can be used in different ways for mammography image quality assessment. The task as performed by a radiologist is to detect several signatures of artificial lesions of a technical phantom, which are very similar to the detection of lesions in radiographic images. Detecting lesions in radiographic images can be done by so-called semantic segmentation, where each pixel of an input image is classified into a pre-defined set of classes. In this section a method is proposed where image quality is determined via semantic segmentation.

Image quality assessment in mammography takes place in a no-reference setting, where the image quality is expressed in terms of a contrast-detail curve. No-reference or blind image quality assessment can be done for natural images as well as for medical images by utilizing deep learning (Bosse et al., 2017; Hou et al., 2014). Most of those algorithms learn to map images to a corresponding subjective quality score (Bianco et al., 2018). To this end, few articles focus on a pixel-wise approach. Bhattacharya et al. (2010) incorporate deep learning for photo-quality assessment. They use semantic segmentation to elaborate visual aesthetic features. These features are further mapped to a human rating of the quality of an image by support vector regression.

Also, Kim and Lee (2017) utilize deep learning to learn a visual sensitivity map that represents the human visual perception. Inspired by these approaches, we explore the usefulness of semantic segmentation for image quality assessment in mammography. Each pixel of the CDMAM image is labeled to belong to either a background, a grid or a lesion class. Image quality assessment in mammography tries to identify how accurate these lesions can be detected. Hence, knowing the ground truth mask allows one to determine typical metrics for the quality of the performed semantic segmentation. These metrics express how successful the semantic segmentation approach can identify the lesions in the image. This requires that the positions of each lesion is known exactly. For our virtual mammography the virtual specimen, and hence the ground truth segmentation mask, is known exactly. For real images this segmentation mask has to be computed. Since the structure of the CDMAM phantom is known, the ground truth mask can be calculated by creating a mask from the known pixel spacing of the detector and a subsequent cross-correlation of this mask with the image. The cross-correlation makes it possible to align the mask with the grid structure in the image and in turn, the desired ground truth segmentation mask is computed. Despite the limitation of requiring a ground truth mask, image quality measurements via semantic segmentation could also be used as a benchmark for methods that actually utilize semantic segmentation for computer-aided detection. The hope could be, that an algorithm that reliably identifies suspicious regions in real patient images is also good in detecting artificial lesions in a technical phantom.

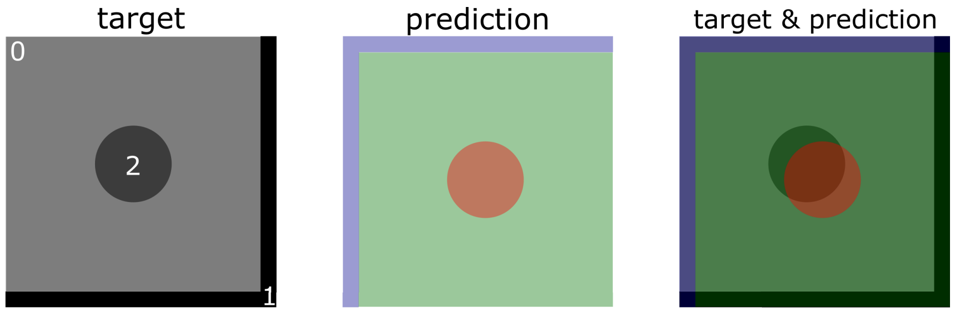
Semantic segmentation describes the task of assigning a unique label to each pixel that marks its membership to an object or a region it belongs to (Long et al., 2015). This way, semantic segmentation comprises "scene understanding" by answering the questions what is visible in the image and where. Deep learning was found to be a powerful tool for pixel-wise semantic segmentation and marks the state of the art in this field of research (Noh et al., 2015; Lin et al., 2017; Chen et al., 2017). Originally, CNN's were introduced to make use of spatial correlations of pixels in their neighborhood. However, this spatial information is lost in the final fully connected layers of a typical CNN. In semantic segmentation, this spatial information is crucial and thus Fully Convolutional Networks (FCN) were developed by Long et al. (2015) to overcome this limitation. FCN's replace the final fully connected layers of conventional CNN's with up-sampling layers that transform the downsized feature map into a pixel-wise output or segmentation map. Thus, the down-sampling path of conventional CNN's to extract features to interpret the context or semantic information is combined with an up-sampling path to recover the spatial information (Long et al., 2015).

The U-net, introduced for biomedical image segmentation (Ronneberger et al., 2015), modifies such FCN's by an up-sampling path with many feature channels so that the up-sampling portion is symmetric to the down-sampling part. Further, skip connections are used to combine context information and spatial information (Drozdzal et al., 2016). The U-net was shown to achieve good performance, even when only few annotated images are available (Ronneberger et al., 2015) and it can be applied to many different tasks.

According to the EUREF guideline, mammography image quality assessment is carried out by analyzing images of the CDMAM phantom and determining a contrast-detail curve. This curve determines the detectability of small structures by visualizing the minimum thickness for each diameter such that an observer can detect the lesion with prescribed accuracy. Contrast-detail measurements were found to relate to the detection of micro-calcifications (Mackenzie et al., 2016). However, they are not the only method to assess image quality. As an example image quality can be determined in terms of the ROC-AUC as described in chapter 2.

Here, a method is developed that utilizes the U-net to segment an image of the CDMAM phantom into background, grid and lesion pixels. Traditional performance measures for semantic segmentation require a known ground truth mask. It will be shown that such performance measures correlate with image quality. Commonly used metrics in semantic segmentation are the overall accuracy (Long et al., 2015), which is measured by

$$\widehat{PA} = \frac{\sum_i n_{ii}}{\sum_i t_i},$$



$$\text{IoU} = \frac{1}{3}\left(\frac{n_{00}}{t_0 + n_{10} + n_{20}} + \frac{n_{11}}{t_1 + n_{01}} + \frac{n_{22}}{t_2 + n_{02}}\right)$$

Figure 4.1: Illustration of the Intersection over Union (IoU). The ground truth target is displayed on the left with the prediction in the center. On the right the combined target and prediction is used to demonstrate how to calculate the IoU. t_i is the number of pixels in class i in the ground truth mask. n_{ij} denotes the number of pixels belonging to class i while they are predicted as class j.

and the mean intersection over union (Long et al., 2015), which can be calculated as

$$\widehat{IoU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_{k \neq i} n_{ki}} ,$$

where t_i denotes the total numbers of pixel in class i, n_{ij} denotes the number of pixels of class i predicted to belong to class j and n_{cl} stands for the number of classes. Figure 4.1 illustrates the meaning of the Intersection over Union (IoU) for a simple example. The ground truth mask consists of 3 classes. In the prediction, class 1 is classified on the opposite site, leading to a shift of the lesion. From the overlay it can be seen how the IoU can be calculated.

Besides the overall measures, we report the individual accuracy and intersection over union for the three classes separately. The overall accuracy depicts how many pixels are correctly classified. The IoU is a measure of similarity of two data sets. If two data sets are equal, then their IoU is unity since intersection and union are the same in this case. The more the two data sets deviate from each other, the less they have in common, which means that the intersection between the two sets decreases. The union on the other hand increases since the two sets contribute more samples that are only present in one of the sets.

4.2.1 Data

The virtual mammography delivers mammography images from digitized phantoms. As described in chapter 3, a digital CDMAM phantom can be used to simulate images for the purpose of image quality assessment. The digitized phantom is a voxel object and the virtual specimen, specifically the position of the artificial lesions is known exactly for all simulated images of this phantom.

Four different values of the SNR were simulated with a tube voltage of 31 kV and a time-current of 100 mAs. For each noise level, 200 images were simulated and randomly split into training data, test data and validation data. The remaining simulation parameters were chosen as listed in Table 3.2. To further verify the accuracy of our method, a set of real images was included as test data in addition.

To enable an efficient training procedure to be carried out on a single graphics processing unit (GPU), the images were down-sampled patchwise. From each entire image of the CDMAM image, sub-images with dimension of 400 × 400 *Pixel* were cut out with minimum overlap of 2 *pixels*. This way, each image yields 60 sub-images that are used as input for the neural network. Each training sample was labelled with the extracted ground truth segmentation mask, cut out at the exact same position as the sample image. The images are transformed into 8 bit grayscale images with pixel intensities between 0 and 255 and normalized by subtracting the median and scaling the intensities to be in the interval between 0 and 1.

4.2.1.1 Data Augmentation

The U-net was shown to be able to outperform state-of-the-art segmentation even in situations where only a few annotated training samples were available (Ronneberger et al., 2015). This is achieved by using extensive data augmentation.

Similarly, data augmentation is utilized here to improve the performance of training a U-net to segment CDMAM images into background, grid and lesion pixels. The goal of data augmentation is to include perturbed image samples to learn a more robust and generalized neural net.

The augmentations can be added to the original data set as done in section 4.3 to construct one whole big data set. Here, a different strategy was chosen where the possible augmentations are applied randomly whenever the training algorithm draws a data sample from the original data set (cf. Shorten and Khoshgoftaar (2019)). This way, the same image is represented in a different variation in each training epoch (cf. chapter 2). As an advantage, this methodology reduces the amount of data that has to be stored physically. Table 4.1 gives an overview of the included augmentations along with a short explanation and the corresponding parameter range of the applied transformations.

Especially blurred samples and images with a gradient in the pixel intensity are important to account for images of different quality levels. Real images were found to have a gradient in y-direction in the background level. As a consequence, the background level is dependent on the position in the image. Therefore, the training data must contain such images that contain a gradient in the pixel intensities in order to train a neural network robustly. The virtual mammography simulates images without gradients. Thus, a gradient image of given strength is added to the image in the augmentation procedure.

The applied random flips effect both image and mask in the same way. All remaining data augmentation methods leave the mask unaltered.

4.2.2 Neural Network

Inspired by (Ronneberger et al., 2015), we set up a U-net architecture that keeps the dimensions of the output same as the input image. The basic architecture is shown in Figure 4.2.

In the down-sampling path, each two consecutive convolutions are followed by a max pooling operation. Each convolution is carried out together with a batch-normalization and a ReLU operation. This procedure is subsequently used to down-sample the initial 400 × 400 *Pixel* image into 512 25 × 25 *Pixel* feature maps. In the up-sampling path, up-convolutions are employed to up-sample the feature maps to the original image size. In each step, the feature

Table 4.1: Overview of the different transformations used to augment the image patches of the CDMAM phantom to train a neural network to segment the image into background, grid and lesion pixels. The third column specifies the parameter range of the applied transformations. In each epoch these parameters are randomly and uniformly drawn from the corresponding intervals.

Operation	Explanation	Values
Flip	With probabilities p_v and p_h, the input image is flipped vertically and horizontally.	$p_v = p_h = 0.5$
Brightness	The brightness of the image is adjusted by a factor b. A factor of $b = 0$ results in a black image, whereas $b = 1$ gives the original image.	$b \in [0.1, 1.9]$
Contrast	The contrast of the image is controlled by a contrast factor c where a value of $c = 1$ conserves the original image and a factor smaller or greater than one reduces or increases the contrast in the image, respectively.	$c \in [0.1, 1.9]$
Gamma	A gamma correction is applied to the image. $\gamma, 1$ leaves the original image unaltered and values greater than 1 make shadows in the image darker.	$\gamma \in [1, 1.8]$
Blur	A Gaussian filter with standard deviation σ is applied to account for blurring in the images.	$\sigma \in [0, 5]$
Gradient	With a probability of p_{grad} a gradient is added to the image. The pixel range lies in $[0, 255]$. A gradient image is created such that the pixel intensities vary linearly between $m \cdot 255$ for $y = 0$ and 255 for $y = 255$. The gradient image is then blended into the the original image.	$p_{grad} = 0.2, m = 0.8$

map is concatenated to the corresponding map in the down-sampling path and fed into two consecutive convolution blocks. The final 64 feature maps of size 400×400 *Pixel* is transformed into the output segmentation map via 1×1-convolution. The segmentation maps consists of 3 channels that represent the 3 classes (background, grid and lesion) in our case. The network was implemented in PyTorch (Paszke et al., 2019). The network was trained for 20 epochs with a batch size of 32 images using a stochastic gradient descent solver with momentum 0.9 and a learning rate of $5 \cdot 10^{-4}$. Cross entropy loss was used as a cost function with a weight regularization factor of 0.1. During the training procedure the model that achieved the best overall accuracy on a random set of validation images that were not used for training was kept for further evaluation.

Splitting the image of a CDMAM phantom into pixels of background, grid and lesion results in a large imbalance between the different classes since there are more pixels belonging to the background than to the grid and lesion class. If this is not considered in the training phase, the neural network will concentrate on correctly classifying the background class since this is the major contribution to the corresponding loss function. Therefore, we use weighted cross entropy as a loss function, where each class is multiplied with a weight that corresponds to the inverse of the relative frequency of class pixels in the ground truth images.

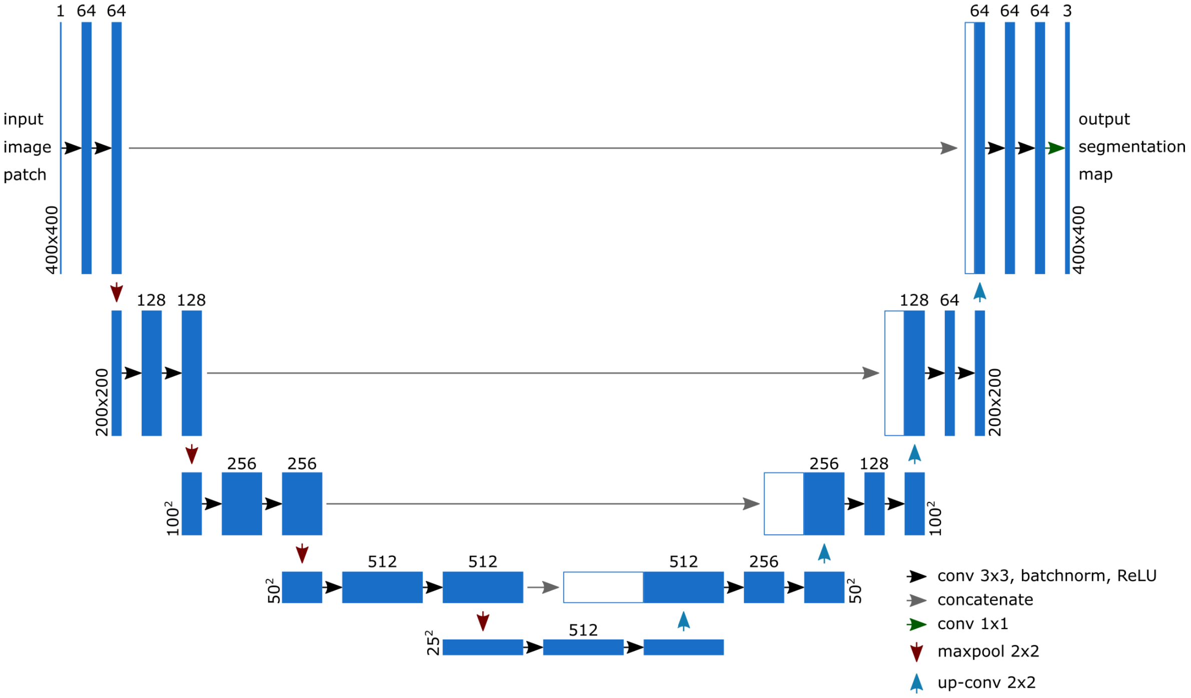


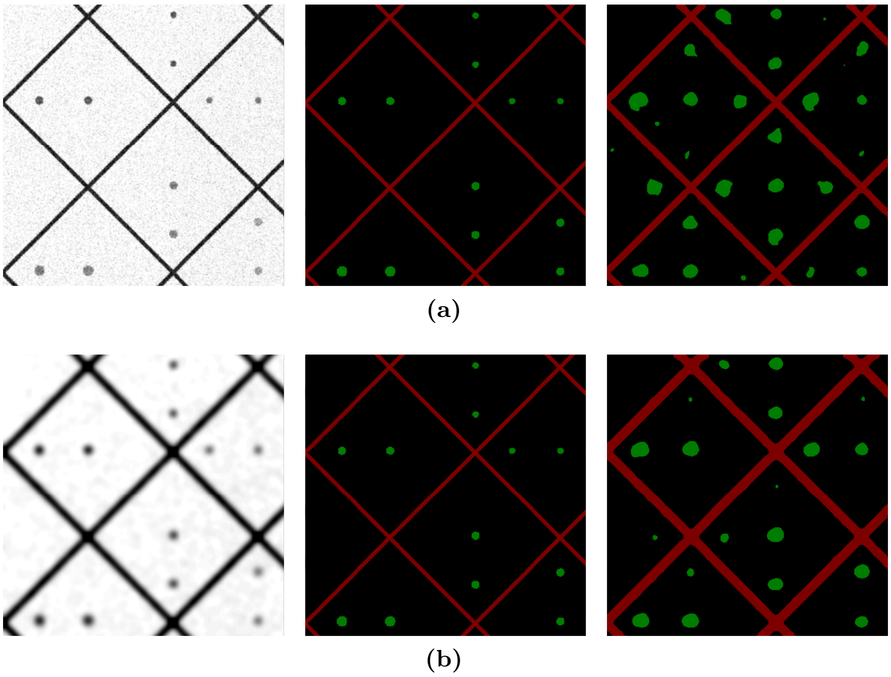
Figure 4.2: Architecture of the U-net that reads-in an image patch of 400×400 *Pixels* image patch and outputs a segmentation map of the same size with three classes (background, grid and pixel). Blue boxes visualize feature maps with the number of channels on top of the box. The arrows denote different operations.

4.2.3 Results

The U-net was trained with the parameters reported in the previous section for 48 hours on a Nvidia®Tesla®K80 GPU and tested on images that were not used for training the net. However, the test images were not truly independent, since test images were created with the same simulation parameters as the training images. Test images as well as training images were simulated with four different noise parameters and thus four different image quality levels. We will refer to these as four different SNR values in the following.

Figure 4.3 shows the result of the trained U-net for an example image of the highest SNR class. The input image, which is a 400×400 *Pixel* patch of the whole image, is shown on the left with the corresponding ground truth in the center and the network output on the right. In Figure 4.3a the raw image was used as an input image, whereas Figure 4.3b shows the resulting segmentation, when a Gaussian filter with standard deviation $\sigma = 4$ is applied to the input image. The Gaussian filter leads to a blurring in the grid and lesion regions respectively, at the same time the fluctuation in the background. As a result, the trained U-net classifies less pixels as lesion pixels that actually belong to the background. For a more detailed evaluation, Table 4.2 lists the quality scores for example test images for the four different image quality levels $SNR\ 1-4$ and different standard deviations σ of the Gaussian filter. The quality of the semantic segmentation is measured for test images of the whole CDMAM phantom in terms of the pixel accuracy and the Intersection over Union. Next to the overall results, the individual results for the three different classes are presented.

It can be seen that for all different SNRs the overall pixel accuracy $\widehat{PA}$ as well as the mean intersection over union $\widehat{IoU}$ decreases with increasing standard deviation of the Gaussian filter.



(a)

(b)

Figure 4.3: Results of the segmentation of a patch of the CDMAM phantom into three different classes: background (black), grid (red) and lesion (green). The left side shows the input patch with the corresponding ground truth in the center and the network output segmentation mask on the right. (a) Results for the raw image and (b) results when the same image is blurred with a Gaussian filter with standard deviation $\sigma = 4$ before it is fed into the neural net.

The Gaussian filter blurs the grid lines in the image significantly (cf. Figure 4.3). The U-net correctly reacts on blurred grid lines and classifies them accurately. However, the ground truth is not affected by the blurring and thus the grid lines in the ground truth segmentation map remain the same. As a result, more pixels of the background class are (correctly) classified as grid pixels by the U-Net. This ground truth, however, was not altered for the blurred images resulting in a decrease of the overall accuracy and IoU for increasing standard deviation of the Gaussian filter (cf. Table 4.2).

For image quality assessment, the lesion class plays the crucial role. While the pixel accuracy of the lesion class also decreases for increasing blurring, the Intersection over Union increases. This is in accordance with the finding from Figure 4.3, where for a higher blurring less background pixels are wrongly classified as lesion pixels. The goal of image quality assessment is to find out whether the device yields a sufficient resolution of small contrasts to ensure adequate diagnostic quality. Thus, image quality assessment should be able to distinguish different quality levels. In Table 4.2 the Intersection over Union of the segmentation map and the ground truth without applying a Gaussian filter ($\sigma = 0$) are highlighted as bold numbers for each quality level. With increasing index $SNR\ i$, the images were simulated with a larger amount of noise, resulting in lower image quality. Indeed, the numbers decrease for decreasing image quality and thus could be used as an indicator for image quality evaluation.

To investigate this effect in more detail, the confusion matrix (Stehman, 1997) is taken into account. The confusion matrix or error matrix is a map that reports the observed frequencies of classes in the ground truth as well as the prediction of a classifier. Besides the total number of class members in the ground truth as well as in the prediction, a binary confusion matrix presents the true positives, the false positives, false negatives and the true negatives. For a

Table 4.2: Quality measures for the performed semantic segmentation for test images for four different quality levels SNR 1 – 4 after applying a Gaussian filter with standard deviation σ. The SNR decreases monotonically from SNR 1 to SNR 4 Presented are the overall pixel accuracy $\widehat{PA}$ and the mean intersection over union $\widehat{IoU}$ as well as the individual class scores for classes background (bg), grid and lesion.

SNR	Gaussian filter	$\hat{PA}$	$\hat{IoU}$	$\hat{PA}$ bg	$\hat{PA}$ grid	$\hat{PA}$ lesion	$\hat{IoU}$ bg	$\hat{IoU}$ grid	$\hat{IoU}$ lesion
SNR 1	$\sigma = 0$	0.9348	0.5550	0.9339	0.9944	0.6513	0.9310	0.5940	**0.1400**
	$\sigma = 1$	0.9346	0.5513	0.9337	0.9971	0.6274	0.9308	0.5790	0.1440
	$\sigma = 2$	0.9337	0.5445	0.9331	0.9990	0.5864	0.9299	0.5491	0.1545
	$\sigma = 3$	0.9298	0.5366	0.9290	0.9997	0.5702	0.9257	0.4984	0.1857
	$\sigma = 4$	0.9248	0.5342	0.9240	0.9999	0.5236	0.9204	0.4500	0.2321
SNR 2	$\sigma = 0$	0.9349	0.5560	0.9338	0.9980	0.6504	0.9311	0.5992	**0.1378**
	$\sigma = 1$	0.9353	0.5532	0.9344	0.9987	0.6263	0.9315	0.5843	0.1438
	$\sigma = 2$	0.9338	0.5455	0.9330	0.9994	0.5974	0.9300	0.5488	0.1577
	$\sigma = 3$	0.9304	0.5364	0.9297	0.9998	0.5563	0.9263	0.5018	0.1812
	$\sigma = 4$	0.9250	0.5353	0.9242	0.9999	0.5251	0.9207	0.4502	0.2350
SNR 3	$\sigma = 0$	0.9312	0.5521	0.9298	0.9969	0.6709	0.9272	0.6015	**0.1276**
	$\sigma = 1$	0.9333	0.5512	0.9321	0.9985	0.6481	0.9294	0.5845	0.1396
	$\sigma = 2$	0.9338	0.5455	0.9330	0.9994	0.5993	0.9299	0.5494	0.1573
	$\sigma = 3$	0.9300	0.5377	0.9292	0.9998	0.5776	0.9260	0.4999	0.1873
	$\sigma = 4$	0.9251	0.5342	0.9244	0.9999	0.5189	0.9208	0.4513	0.2305
SNR 4	$\sigma = 0$	0.9284	0.5508	0.9266	0.9989	0.6928	0.9242	0.6076	**0.1206**
	$\sigma = 1$	0.9333	0.5516	0.9321	0.9993	0.6515	0.9294	0.5856	0.1398
	$\sigma = 2$	0.9339	0.5473	0.9330	0.9996	0.6178	0.9301	0.5511	0.1608
	$\sigma = 3$	0.9302	0.5392	0.9292	0.9999	0.5901	0.9261	0.4999	0.1915
	$\sigma = 4$	0.9252	0.5359	0.9244	0.1000	0.5316	0.9208	0.4514	0.2356

Table 4.3: Confusion Matrix for "lesion" against "no lesion" classification for an example test image simulated with the SNR 1, $\sigma = 0$ setup. The values indicate the total number of pixels or the relative fractions with respect to the corresponding total amount. Gray cells highlight the true positives, false positives, false negatives and true negatives. The precision (red) and the recall (green) are used to calculate the F_1-score (see text).

			ground truth			
			lesion	$\overline{lesion}$		
		total	65,447	8,422,795		
prediction	*lesion*	281,732	42,625	239,107	15.13%	84.87%
	$\overline{lesion}$	228,206,510	22,822	8,183,688	2.78%	99.72%
			65.13%	2.84%		
			34.87%	97.16%		

multi-class classification a binary confusion matrix can be determined where the actual class is evaluated against all other classes.

True positives are members that were correctly identified by the classifier. False positives are members that actually belong to class i but are classified as something different. On the other hand, false negatives are all samples that are classified into class i while they actually do not belong to class i. True negatives finally are samples that are correctly classified of not belonging to a certain class i. Several numbers can be extracted from the confusion matrix. Two very important figures are the so-called precision and the recall. The precision denotes the fraction of true positives and the total number of predicted members in class i, whereas the recall divides the true positives by the total number of members in the ground truth class i. With these two values, the F_1 score

$$F_1 = 2\frac{Recall * Precision}{Recall + Precision}$$

can be taken as performance measure to evaluate the semantic segmentation.

Table 4.3 shows an example of a confusion matrix for the semantic segmentation result of a test image of the whole CDMAM phantom, simulated with SNR 1 without applying a Gaussian filter. The gray cells denote the four values of interest and report the total number of pixels that fulfil the condition in each cell. The precision is given in the red cell, while the recall is presented in the green cell of the matrix. The Recall equals the overall pixel accuracy $\widehat{PA}$ from Table 4.2. For example test images of the entire CDMAM phantom for the four different image quality levels SNR $1-4$ the corresponding performance measures of the lesion class are presented in Table 4.4.

With decreasing image quality, the trained U-net classifies more pixels as lesion. This increases the true positives and thus the recall since the amount of lesion pixels in the ground truth mask remains constant. However, the increase in the number of pixels classified as lesions is higher than the increase of the true positive fraction, resulting in a decrease in the precision of the U-net. Thus, the F_1-score decreases with decreasing image quality as well. Precision as well as F_1-score show the expected behavior that the corresponding values decrease with decreasing image quality. In case the position of the lesion - and hence the ground truth for our semantic segmentation - is known exactly, which is a common assumption in mammography image quality assessment, precision or F_1-score could be used to assess the image quality via semantic segmentation.

Note that the results presented are a proof of principle of how deep learning might be useful for mammography image quality assessment. The current EUREF guideline procedure measures image quality in terms of the contrast-detail curve that depicts local contrast-detail information. The results presented in this section yield a global measure of image quality in

Table 4.4: Performance measures of the lesion class for the segmentation of the entire CDMAM phantom image. The four different SNR i denote four image sets of different image quality, where the image quality decreases for increasing index.

	predicted lesion pixels	true positives	precision	recall	F_1
SNR 1	281,732	42,625	15.13%	65.13%	24.55%
SNR 2	285,945	42,568	14.89%	65.04%	24.23%
SNR 3	322,622	43,909	13.61%	67.09%	22.63%
SNR 4	355,820	45,342	12.74%	69.28%	21.52%

terms of the F_1 score or the precision. To transform this into a contrast-detail curve, the segmentation result needs to be evaluated for the individual cells of the CDMAM phantoms. Thus, an error-prone post-processing of the segmentation result needs to be carried out in order to determine a contrast-detail curve. This way, our method could be modified to distinguish between cells where the signature of the lesion can be detected with sufficient probability and cells where this is not the case, by choosing a threshold for the derived metric. This can be recast into contrast-detail curves as described in chapter 3. In other words, the application of the semantic segmentation observer (SSO) alone is not sufficient to estimate the contrast-detail curve without cumbersome post-processing.

Nevertheless the developed SSO is able to precisely classify a filtered image into background, grid and lesion pixels. With known lesion positions, global metrics (e.g. the F_1-score) can be calculated that measure the performance of the semantic segmentation which correlates with image quality.

Rather than fine tuning and extensive testing of the employed U-net on a bigger data set, we suggest a completely different approach in the next section where deep learning is directly applied to images of the CDMAM phantom to predict the contrast-detail curve. In contrast to the semantic segmentation this approach is directly comparable with the results of the current EUREF guideline procedure. This will help to see whether a deep learning approach is beneficial for mammography image quality assurance.

4.2.4 Discussion

Virtual mammography as described in chapter 3 can be used to simulate large image data sets of the CDMAM phantom. This helps to explore the usefulness of deep learning in the context of mammography image quality assurance. The goal of image quality assurance in mammography is to ensure that a device is able to depict small contrasts in the image such that a reliable diagnostic can be carried out. For quality assurance measurements a technical phantom is used to determine the contrast resolution that can be achieved by an observer under standard acquisition parameters.

Over the past decade, deep learning has demonstrated its potential for many different tasks, especially image analysis. Traditional machine learning algorithms utilize a set of features, extracted by an expert, that are especially useful to solve the given task. For mammography image quality assessment, classical data analysis methods or human observers require that the image of the entire phantom is segmented into single patches that are further analyzed by comparing average pixel intensities and localizing the signature of a lesion. This analysis pipeline is especially suitable for the human observer. The question remains whether this is also the optimal strategy for an algorithm. One great advantage of deep learning is that during the training phase, the network develops a set of features by itself, according to which the given task can be solved.

In this section, semantic segmentation was used to train a U-net to classify the individual pixels into three different classes: background, grid and lesion. Our results show that the U-net can be successfully learned to distinguish between lesion pixels, grid pixels and background pixels. The F_1-score for the lesion class can be used as a global measure for image quality since it correctly correlates with the quality of different images. While this is beneficial on its own, the current EUREF guideline procedure expresses image quality in terms of a contrast-detail curve. Rather than being a global measure of image quality, the contrast-detail curve summarizes local contrast information. To recast our global F_1-score into a contrast-detail curve, the segmentation result needs to be evaluated individually for each cell of the CDMAM phantom. By choosing a minimum threshold for the F_1-score, a contrast-detail curve can be extracted, similarly to the EUREF method. Therefore, the segmentation result has to be separated into the individual cells which represents an error-prone post-processing.

The good initial results of the SSO demonstrate its potential for image quality assessment as a proof of principle. However, a contrast-detail curve cannot be directly obtained without cumbersome post-processing. Quality assurance measurements, following the EUREF guideline, compare the determined contrast-detail curve to a limit curve and carry out a binary classification: sufficient image quality or not. Future research could explore whether this classification can be done utilizing our proposed segmentation approach. To do this, it has to be analyzed, if for a set of test images that, according to the EUREF guideline approach, are classified as sufficient image quality or not there exists a threshold such that the assessment via our SSO comes to almost the same judgment? However, this is beyond the scope of this work and therefore, the SSO was not refined any further or compared to the current practice. Instead, a method shall be developed that directly regresses a contrast-detail curve from images of the CDMAM phantom via a CNN. The application of semantic segmentation for mammography quality assurance could be especially beneficial in cases where computer aided diagnostics or detection is used to analyze images. Using semantic segmentation for quality assurance could then allow the effect of different acquisition parameters and different mammography devices on the detectability of lesions in a technical phantom to be studied in a controlled way. This could potentially lead to an increased performance of computer aided detection algorithms for breast cancer diagnostics or at least serve as a benchmark for methods that utilize semantic segmentation. However, this is beyond the scope of this work. Our results indicate that further research to explore this approach could be worthwhile.

The semantic segmentation observer (SSO) determines a global measure of image quality, but requires the knowledge of the ground truth segmentation mask. This mask is known for our virtual specimen, but has to be computed for real images. To determine a contrast-detail curve, sophisticated post-processing has to be done to translate the global metric into local information.

4.3 Deep Learning Observer

Existing methods to determine image quality in mammography in terms of contrast-detail curves require the exact positions of the lesions to be known as well as error prone pre- or post-processing. Deep learning, however, is able to learn representative features itself from big data sets such that a specific task can be solved. In this chapter, a deep learning observer (DLO) is developed that can be applied directly to images of the phantom to determine contrast-detail curves. Therefore, a database of labeled training images is required. Such a data base is created by combining simulated images from the developed virtual mammography, cf. section 3.2, with a set of real images. In supervised learning each image requires a label. These labels are the contrast-detail curves determined with the EUREF guideline software, cf. section 2.1.3. For each set of simulated images the EUREF guideline procedure is applied and the determined contrast-detail curve is then assigned to each image individually.

4.3.1 Data

High quality representative data plays a critical role for the successful, data intensive training of deep neural networks (Abu-Mostafa et al., 2012; Gudivada et al., 2017). To train a neural network for the precise prediction of contrast-detail curves from images of the CDMAM phantom a data set containing images of several different quality levels and different mammography devices is needed. The goal is to cover a wide range of different acquisition parameters that is representative such that the trained neural network generalizes well and thus allows the regression of contrast-detail curves in most clinically relevant application settings.

Creating such a representative data set can be done by virtual mammography, as described previously in section 3.2. The three main simulation parameters tube voltage, the time-current product and the SNR control the modeling of different acquisition settings and devices. Especially the time-current product and the SNR influence the ability of the device to depict small contrasts in the resulting image and hence mainly influence the contrast-detail curves. A higher time-current product equals a higher radiation dose and ultimately leads to a higher SNR and hence shifts the corresponding contrast-detail curve downwards, which represents a better image quality. On the other hand, the SNR emulates the quality of the imaging detector and higher values result in a better image quality as determined in terms of contrast-detail curves.

In total, sixteen sets of 200 images each were simulated and split into a training set, a test set and one independent test set to check the generalization ability of the trained network as illustrated in Figure 4.4.

The corresponding simulation parameters for each set are given in Table 4.5. The 200 images from each set differ in the random noise and in randomly applied shifts of the position of the phantom. These shifts are normally distributed with zero mean and a 1 *Pixel* standard deviation in x- and y-direction individually.

In order to construct a representative training set, 11 simulated sets were combined such that the corresponding contrast-detail curves span a range, indicated as the blue area in Figure 4.5. The parameters of the 4 simulated test sets are chosen differently from the parameters that were used to construct the training set such that they fall in the range that is

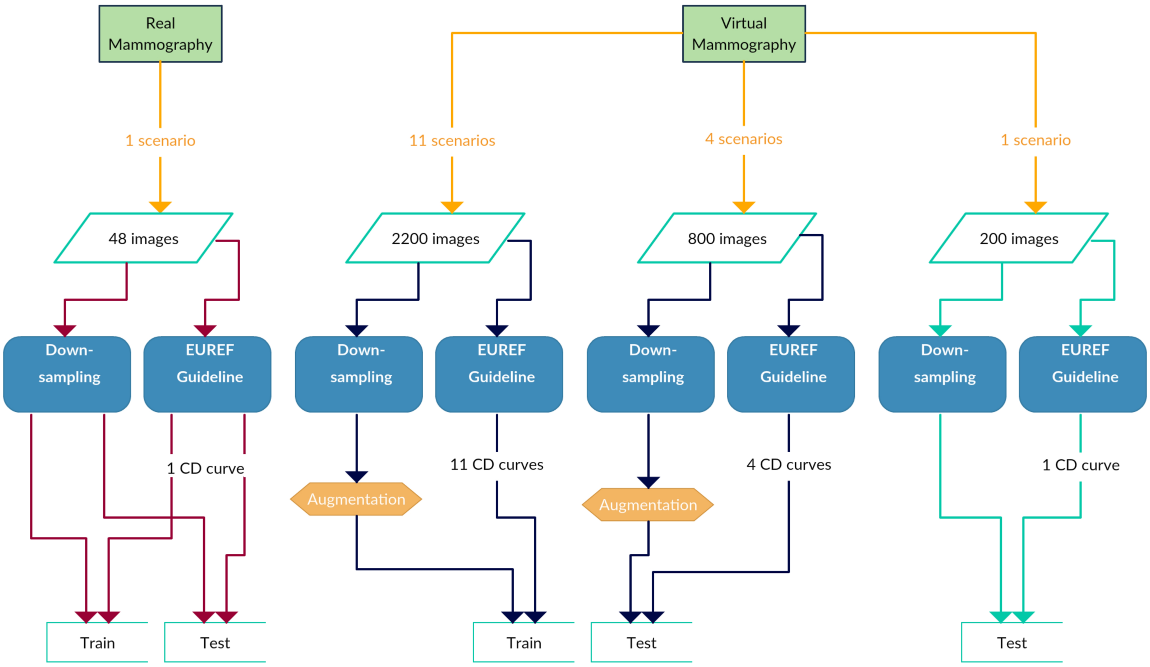


Figure 4.4: Schematic of the image data set consisting of 48 real images and 16 sets of 200 simulated images, respectively. Real images are split into train and test set but were acquired using the same acquisition parameters. Simulated images are split into train and test sets, such that the acquisition parameters are different which ensures the independence of training and test sets. Figure from Kretz et al. (2020).

Table 4.5: Simulation parameters for the 16 different simulation scenarios used to train and test a neural network to estimate a contrast-detail curve from a CDMAM phantom image. Table from Kretz et al. (2020).

Set	No.	Voltage $[kVp]$	Exposure $[mAs]$	SNR
Training	1	25	110	4.5
	2	25	110	4.9
	3	31	100	5.8
	4	25	110	5.9
	5	28	105	6.7
	6	31	100	6.7
	7	25	110	7.2
	8	27	100	7.3
	9	31	100	7.6
	10	25	100	8.5
	11	31	100	8.5
Test	1	28	115	4.9
	2	25	110	6.7
	3	25	110	9.7
	4	25	110	11.7
Generalization Test	1	23	105	3.1

spanned by the training set, as it can be seen as red curves in Figure 4.5. One further set is simulated with parameters such that the corresponding contrast-detail curve lies outside the training range, see the orange curve in Figure 4.5.

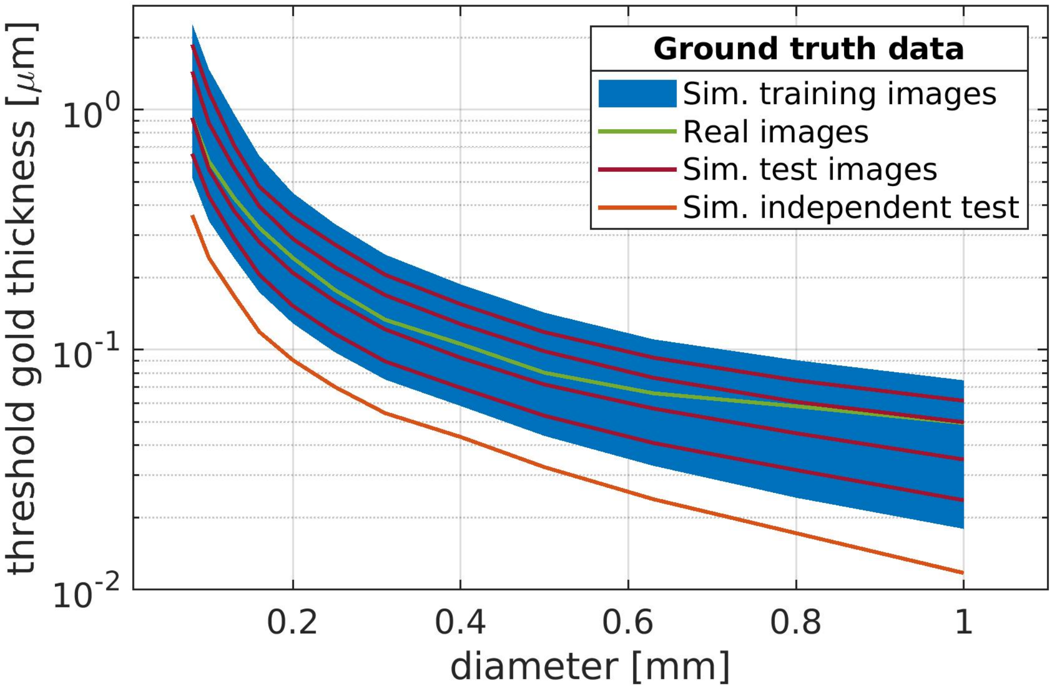


Figure 4.5: Overview of the different contrast-detail curves. 11 simulation scenarios form a range of image quality levels (blue). Four different simulation settings are used to construct independent test sets that lie within the training range (red). The set is enhanced with real images (green). One independent test set (orange) falls outside the training range and is used to check the generalization ability of the trained net. Figure from Kretz et al. (2020).

Simulations are an easy and fast way to construct large data sets with easily controllable image qualities. However, a trained model should be applied to real images in a clinical quality assurance procedure and thus, the training set should further include samples of real images. Therefore, 48 images of the CDMAM phantom were taken that were acquired on a Siemens® Mammomat Inspiration with an exposure of 100 mAs and a tube voltage of 30 kVp. Since the acquisition parameters of all 48 images are the same, the EUREF guideline procedure is applied to all 48 images to determine a contrast-detail curve that is then assigned as a label to every image individually. Henceforth, all 48 real images have the same label. The 48 images are then split into two sets of 24 images for training and 24 images for testing. Thus, even though the real image sets are not truly independent, since they were generated using the same acquisition parameters, none of the real images used for training are included in the test set.

4.3.1.1 Downsampling

Mammography uses high resolution images (approx. $3,600 \times 2,300$ $pixels$) with typical pixel distance $\approx 0.05 - 0.07$ mm. To achieve a reasonable performance when training a deep neural network on a huge database, the CDMAM images are compressed before they are fed into the neural net. Since the CDMAM phantom itself comprises a highly redundant structure, a sparse sampling approach seems to be feasible to reduce the dimensionality of the input images. However, dimension reduction algorithms as, for example, the principal component analysis (PCA) (Wold et al., 1987) or shearlet transformations (Easley et al., 2008) reduce the dimensionality with respect to a corresponding basis. To avoid the assumption of low dimensionality in a specified basis, random down-sampling is chosen to reduce the image dimensions. Random down-sampling is achieved by choosing random pixel indices in x- and y-direction individually with respect to their original order. Choosing a high sample rate preserves more of the original image, while the performance of the training process will decrease. On the other hand, low sample rates ease the training process, whereas a lot of image information is lost. In case of the CDMAM phantom, an image dimension reduction to 250×250 *pixels* was found to ensure an efficient performance while the achieved accuracy does not suffer too much from the high down-sampling rate. Higher downsampling rates were found to leave out too much information such that the contrast-detail curves could not be estimated with sufficient accuracy. Keeping more information, on the other hand, significantly slowed down the training process without providing a significantly better accuracy. Using random downsampling to reduce the original images with a size of 3654×2323 *pixels* (cf. section 3.2) to a size of 250×250 *pixels* means omitting 99% of the original image area. This number may seem high but note that the CDMAM phantom contains a lot of redundant information that is not needed by a neural network to estimate a contrast-detail curve.

4.3.1.2 Data Augmentation and Normalization

Even though the orientation of the CDMAM phantom is strictly regulated in the quality assurance protocol for mammography screening, the trained network should be invariant under rotation and scaling of the recorded image. For this reason, the training data set is further augmented. First of all, the imaging area is usually not fully covered by the phantom. This means that the final image consists of the phantom region and a margin area that has a much higher intensity since the incoming radiation is not absorbed by the phantom. To enable the estimation of the contrast-detail curve without prior segmentation of the phantom area, the training data has to include image samples that contain a margin. In the simulations, a margin is added to the image simply by placing the image in a bigger image before down-sampling. Phantom pixels usually have an intensity between 300 and 600, whereas the intensity for the margin pixels for each image is chosen randomly between values of $2,600$ and $10,000$. The pixel intensity in the margin region is much higher because the X-Rays in this area are not attenuated at all. To produce a more realistic margin, a small random perturbation of 2‰ of the pixel intensity was added to each pixel of the margin area. Each simulated training scenario is augmented by including versions with margins between 0 and 100 *pixels*. Note that by adding a margin of 100 *pixels* the margin area covers over 60% of the original image. This is far from clinical scenarios where the margin area usually covers between 25% to 50%. Going beyond the clinical relevant range led to greater robustness and performance of this approach. Introducing images with different margins for the same ground truth is supposed to lead the trained neural network to be robust for images of different devices where the size of the detector determines the margin in the actual image. After adding six different margins, mirroring is included as data augmentation. The different sets are mirrored vertically, horizontally and both vertically & horizontally. Thus, the 200 images of each training scenario are multiplied to $200 \cdot 6 \cdot 4 = 4,800$ images. As mentioned earlier, the training set is enriched by including 24 real images. Since this is strongly under-represented compared to the $4,800$ images of one

Table 4.6: Final composition of the training set comprising augmented simulated images and real images and the final test set of 4 simulation sets and real images that were not used for training plus 1 independent simulation test set. Factors denote the multiplication of images through data augmentation. Table from Kretz et al. (2020).

Set	Type	No. images	No. scenarios
Training	simulation	$2200 \cdot 6 \cdot 4$	11
	real	$24 \cdot 200$	1
Test	simulation	$800 \cdot 6$	4
	real	24	1
	simulation	200	1

simulation scenario, each real training image is copied 200 times, without alteration, to have a fair representation in the final training data set.

Data augmentation is usually only applied for the training images. Here, we use the different margins also in the test set, to ensure that the network successfully learned to ignore the margin pixels which would prove that no initial data processing is needed.

The final sizes of training and test data sets are given in Table 4.6. The training set after data augmentation consists of $57,600$ images. Testing of the trained network is performed on $4,800$ simulated test images, 24 real images and a further simulated generalization test set of 200 images.

4.3.2 Neural Network Architecture and Training Strategy

After setting up the database to train a neural network to estimate contrast-detail curves from single images of the CDMAM phantom, a network architecture needs to be chosen. A contrast-detail curve is estimated only on a finite number of abscissa values. For each of a discrete set of diameters, the corresponding thickness is computed such that the signature of this lesion can be detected with prescribed accuracy. This can be understood as simultaneously solving single regression problems at each abscissa value. As a reference for the regression, the contrast-detail curve determined with the current EUREF guideline procedure is used. Thus, the network should learn the regression of the contrast-detail curve and be able to generalize even beyond the range defined by the training data.

Figure 4.6 shows a schematic overview of the workflow to determine contrast-detail curves with a neural net. The incoming image is randomly down-sampled and fed into a convolutional neural network (CNN). CNN's showed superior performance in several computer vision tasks, where inputs are commonly images [refs]. Typical elements like convolution, activation, normalization and pooling are organized into layers and stacked onto each other. These structures can be interpreted as feature extractors with the derived features fed into one or more fully connected layers to finally perform the given task. In our case, this is the regression of the 12 points that form the contrast-detail curve.

Inspired by the image regression example in the Matlab® deep learning Toolbox (MATLAB, 2019), a regression CNN was set up with an architecture as listed in Table 4.7. Stages of convolutional layers followed by ReLU activation functions and batch-normalization layers are combined by using 2×2 pooling layers.

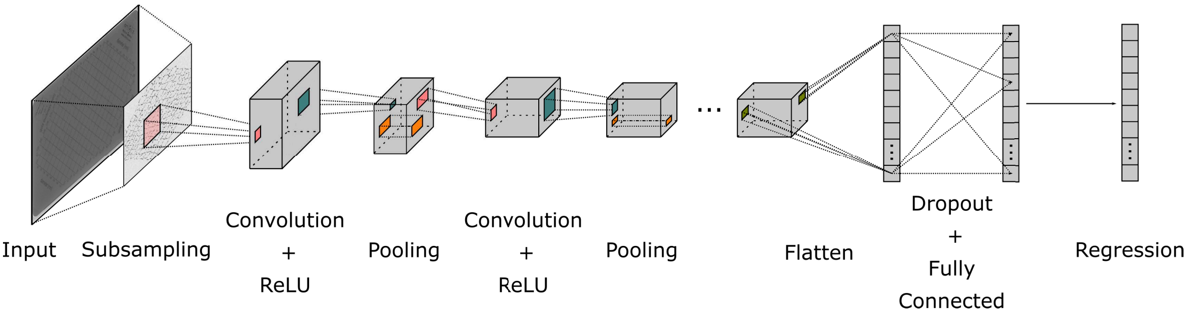


Figure 4.6: Schematic overview of the processing pipeline. The image is randomly downsampled and fed into a convolutional neural net. The latter comprises several convolution, pooling, normalization and fully connected layers to regress a contrast-detail curve. A detailed overview on the network architecture is given in Table 4.7.

Table 4.7: Architecture of the CNN for contrast-detail curve estimation from CDMAM images. Table from Kretz et al. (2020).

Layer	Type	Filter dimensions	Output dimensions
1	image input		$250 \times 250 \times 1$
2	convolutional	$3 \times 3 \times 1 \times 8$	$250 \times 250 \times 8$
3	batchnorm + relu		$250 \times 250 \times 8$
5	maxpool	2×2	$125 \times 125 \times 8$
6	convolutional	$3 \times 3 \times 8 \times 16$	$125 \times 125 \times 16$
7	batchnorm + relu		$125 \times 125 \times 16$
9	maxpool	2×2	$62 \times 62 \times 16$
10	convolutional	$3 \times 3 \times 16 \times 32$	$62 \times 62 \times 32$
11	batchnorm + relu		$62 \times 62 \times 32$
13	fully connected	123008×512	$1 \times 1 \times 512$
14	dropout (0.2)		$1 \times 1 \times 512$
15	fully connected	512×512	$1 \times 1 \times 512$
16	fully connected	512×12	$1 \times 1 \times 12$
17	regression layer		1×12

After three convolution stages, a fully connected layer follows. For better training performance and generalization, a dropout layer with a dropout probability of 0.2 is added, which transfers into the next fully connected layer. The final layer is a fully connected layer with 12 neurons that represent the twelve points of the contrast-detail curve. The mean squared loss (see also chapter 2) was utilized as the corresponding loss function.

4.3.2.1 Incremental Learning

Rather than training the full network architecture directly on all of the data, incremental learning was applied which makes use of transfer learning. Not only the data was varied in the training process but also the architecture itself was developed into its final form as presented in Table 4.7 and learned incrementally via transfer learning (Kretz et al., 2020).

Incremental learning in general describes the procedure when either the network architecture (Wang and Kuh, 1992; Zhang, 1994) is refined throughout the training process or the training set is altered over time by adding more or different data, as e.g. in (Engelbrecht and Brits, 2001). Incremental learning aims to transfer already acquired knowledge to either a refinement of the network architecture or to the correct prediction of new data. Restarting the training of the network with random network parameters would cause the network to forget all the information that was collected in earlier training epochs and is referred to as "catastrophic forgetting" in (French, 1999; Kirkpatrick et al., 2017) , while transfer learning uses the parameters as learned so far as a new starting point for further, refined training.

For the presented method, the network was previously trained without the second fully connected layer on a subset of the training data. Real images as well as all simulated images augmented with margins or mirrored images were removed from the training set. Then the simulated margins were included in the training set, together with the mirrored images. Transfer learning was used to refine the neural network on the enhanced data set.

After successfully learning on the enhanced training set, the second fully connected layer was added and trained with dropout of 0.2. Training a deeper network with incremental learning increased the achieved performance of this approach. If the deep network was trained directly from scratch without the incremental procedure, the achieved accuracy drops significantly. This is why an incremental learning procedure was used to achieve a deeper network gradually.

In a final step, the real data was included in the training data and the network was again trained on the whole data set using transfer learning. In each step the neural network was trained using stochastic gradient descent with momentum (cf. Chapter 2) and an initial learning rate of $lr = 10^{-3}$. The network was trained in batches of 128 images from the training data set for 150 epochs. At every 30th epoch the learning rate was decreased by multiplying it with 0.1.

4.3.3 Results

The trained network was applied to the different test scenarios that were previously described, and compared with the current practice of the EUREF guideline. None of the test images were used to train the neural net. However, the real test images are not truly independent, since they were recorded under the same conditions as the real images included in the training set.

Figure 4.7 shows the prediction of the trained CNN of 4 different simulated test scenarios along with the corresponding ground truth displayed as lines. Each simulated test scenario consists of 200 images, labeled with the same ground truth. Figure 4.7a shows the mean values for all single predictions of the corresponding set for the 200 images. Error bars indicate the corresponding standard deviations. One can see that the trained network is able to correctly predict the contrast-detail curve of a given image. It is also obvious that the variability of the network predictions increases with the increase in predicted threshold thickness values. Higher curves represent devices with a low SNR and hence a low image quality.

In Figure 4.7b averages of 16 contrast detail curves, determined from a single image each, are shown. As can be seen, this drastically reduces the variability of the resulting contrast-detail curves. These Figures demonstrate that the trained network is able to correctly predict the contrast-detail curve.

As mentioned earlier, the simulated images were augmented to include margin areas that do not contain information for the contrast detectability and hence the trained network should be able to estimate the contrast-detail curve robustly for different margins. Figure 4.8 displays

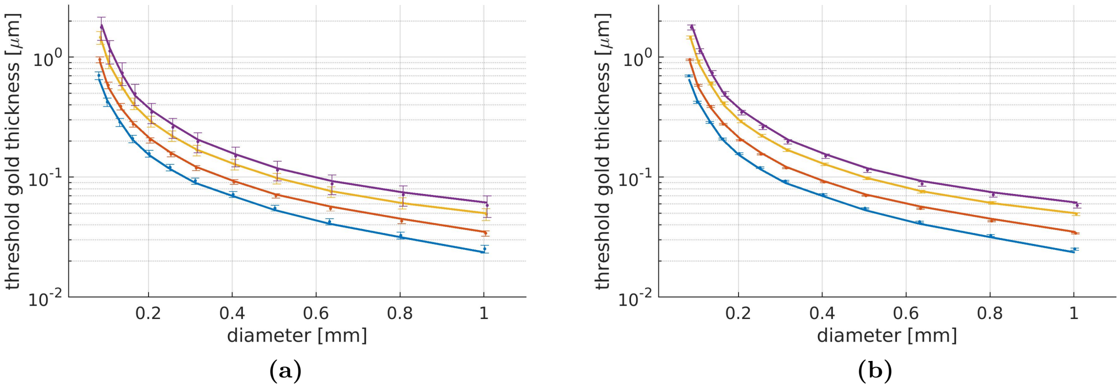


Figure 4.7: (a) Ground truth contrast-detail curves along with the predictions of the trained neural network for 200 single images for each of the 4 different simulated test scenarios. Error bars indicate standard deviations. (b) Averages of 16 predictions of the trained neural net. Figure from Kretz et al. (2020).

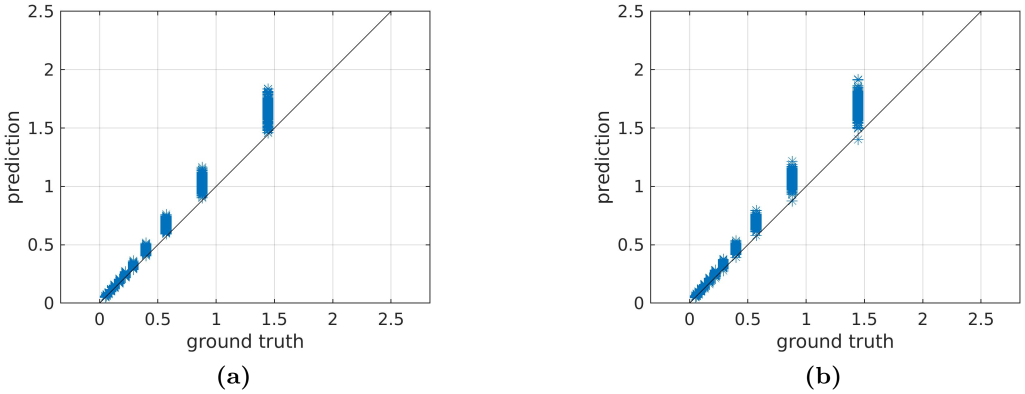


Figure 4.8: Single predictions of the threshold gold thickness (prediction) vs. ground truth values obtained by applying the EUREF guideline procedure for the simulated test images with (a) no additional margin and (b) with additional margin. Figure from Kretz et al. (2020).

the single predictions of the threshold gold thickness against the ground truth as obtained by applying the EUREF guideline procedure for one of the simulated test sets.

Each test image is labeled with the same ground truth, whereas the predictions of single CDMAM phantom images by the trained network vary. The variation increases with the higher values and also with the higher threshold gold thickness. A high threshold gold thickness corresponds to a low diameter. Figure 4.8a shows the results for test images without additional margins, while the results displayed in Figure 4.8b are determined for test images with an additional margin. As can be seen, the mean prediction is the same in both cases, while the results from the test images with additional margin seem to have a bigger variation. Both Figures show that the prediction clearly follows the ground truth.

Figure 4.9 visualizes similar results for the real test images. The predictions vary but are distributed close to the ground truth indicated as black line.

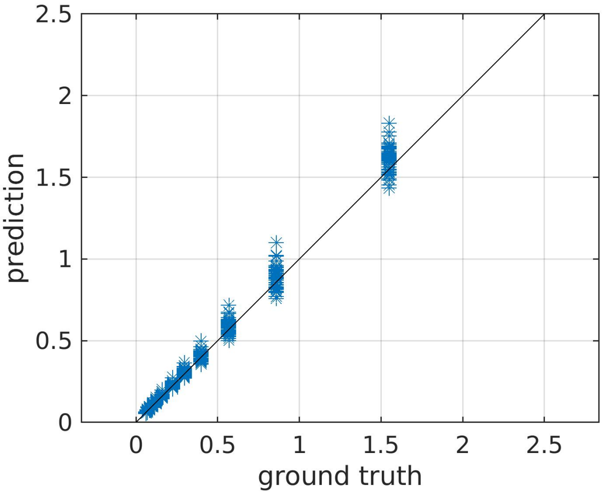


Figure 4.9: Single predictions of the threshold gold thickness (prediction) vs. ground truth values obtained by applying the EUREF guideline procedure for the real test images. Figure from Kretz et al. (2020).

4.3.4 Generalization

Rather than classifying an image as belonging to a predefined quality range, the suggested approach was designed to regress the contrast-detail curve. If correctly learned, the trained network should be able to correctly estimate a contrast-detail curve outside the range that was contained in the training set. For this purpose, an independent generalization test set was constructed with an exposure and a SNR chosen such that the corresponding ground truth curve lies outside the training labels, see Figure 4.5. Figure 4.10 shows the single prediction of the trained network along with the corresponding ground truth. The prediction is highly accurate, which suggests that the trained network generalizes well. Note that the y-axis is presented on a logarithmic scale when comparing the standard deviations with previously shown results.

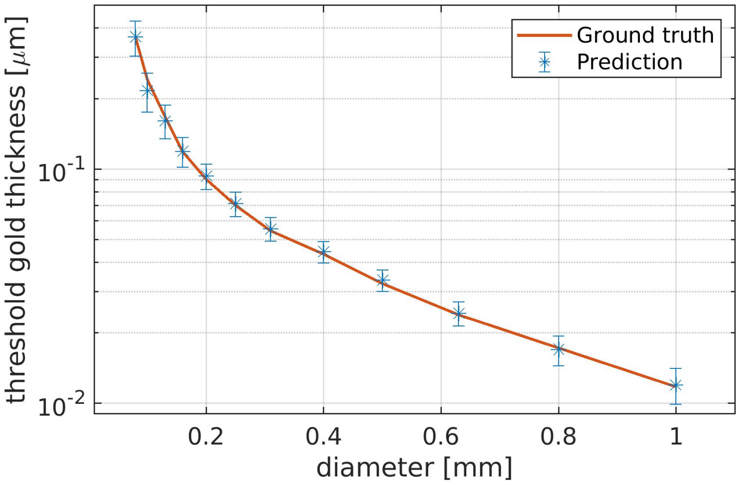


Figure 4.10: Predicted contrast-detail curve for an independent generalization test set. Ground truth (red) and predictions by the trained network for 200 single images. Error bars indicate standard errors. Figure from Kretz et al. (2020).

4.3.5 Variability

So far, the EUREF guideline procedure was applied once for all images of one specific scenario and then assigned to each image as the ground truth label individually. However, due to different noise artifacts in each image, the contrast-detail curve as determined by the EUREF guideline procedure shows a significant variability for images that are recorded under the same conditions. By repeatedly drawing only 16 images out of 48 real images or 200 simulated images, respectively, and applying the EUREF guideline procedure, the resulting contrast-detail curves show variations as well. This means that there is a relevant uncertainty present in the ground truth data. Figure 4.11 demonstrates the effect on the ground truth (red) for the 48 real images. In blue, the single predictions of the trained neural network for the 24 real images of the test set are shown that were not included in the training.

It can be seen that the variability of the single predictions of the trained network are similar to the variability of the ground truth data, when determined by 16 images. In case of the smallest diameter, our approach even shows a smaller variability than the reference method. This highlights the high performance of the neural net which is able to estimate the contrast-detail curve from single images of CDMAM phantoms with a precision similar to the current practice which uses 16 images.

4.3.6 Discussion

Our virtual mammography was taken to simulate realistic images of the CDMAM phantom under different simulation parameters. For these simulated images the the EUREF guideline procedure was applied to generate ground truth contrast-detail curves. The images labeled with the ground truth were then taken to train a DLO to directly estimate contrast-detail curves.

Our results show that a CNN can be successfully trained to assess the image quality in mammography by analyzing images of the CDMAM phantom. The presented approach was

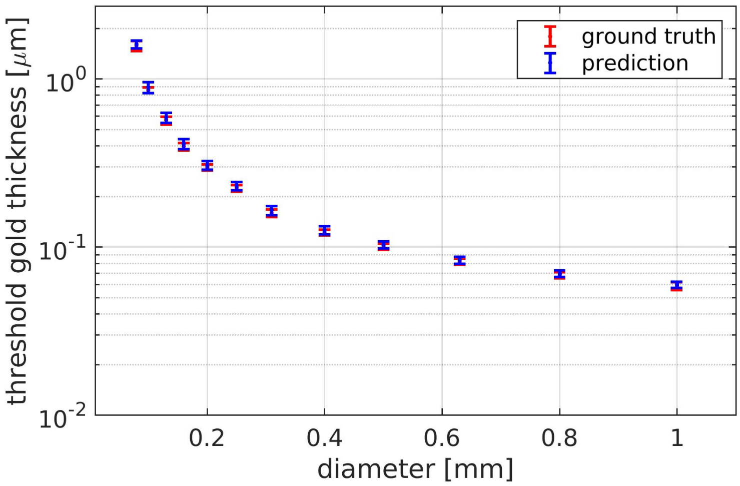


Figure 4.11: Repeated application of the EUREF guidelines procedure for 16 randomly drawn images from the set of 48 real images (red) with common ground truth and predictions of the neural network for single images from the test set of 24 real images. Error bars indicate standard uncertainties. Figure from Kretz et al. (2020).

compared to the current practice in automatic image quality assessment, as outlined in the EUREF guideline. Our trained neural network is able to correctly predict the contrast-detail curve from images of the CDMAM phantom after application of random down-sampling. The accuracy of our method was - for single images - as high as the accuracy of the current practice with 16 images. Therefore, our method can significantly reduce the workload for quality control measurements in mammography.

A high down-sampling rate as utilized in our approach tremendously reduces the information that is present in the image. It was shown that by omitting 99% of the original image, a CNN can still be trained successfully to correctly estimate a contrast-detail curve. Neither a human observer, nor the methods that actually perform a detection task of single lesions in the image are able to work with such a highly down-sampled set of images.

The contrast-detail curve visualizes the thickness for each diameters where the signature of the corresponding lesion cannot be localized with prescribed probability. This means, that the image is divided into cells where the prescribed accuracy is reached and cells where this is not the case. Due to the regular structure of the CDMAM phantom the separation boundary between detectable and non-detectable lesions should be a line or a curve. The relative position of this boundary to the phantom would then be re-cast into the contrast-detail curve. Different image quality levels that lead to different contrast-detail curves would alter the position of the boundary curve in the image. This way, the results indicate, that contrast-detail curves can be estimated in a simpler way by regressing a boundary to an image, rather than detecting individual lesions in the image. Ideally, the trained network should be able to understand this relation.

The EUREF guideline procedure as well as simple mathematical observers and even our introduced segmentation observer require error-prone pre- or post-processing to estimate image quality in terms of contrast-detail curves. First, the image needs to be segmented into individual cells. Then, the positions of the lesions need to be known exactly and localized precisely. The CNN, on the other hand, was applied directly to the down-sampled raw data and learned to pre-process the images by itself, by accounting for mis-alignments or margins in the image. No further, error-prone pre- or post-processing of the data has to be done. Solely a random down-sampling was carried out to reduce the dimension of the input images. The variability of contrast-detail curves obtained by our method from single images matched those variations obtained by the EUREF guideline procedure using 16 images. This shows, that we can derive contrast-detail curves with fewer images than the current guideline procedure. That the trained neural network is able to estimate the contrast-detail curve with highly down-sampled images further demonstrates, that the high resolution CDMAM phantom images contain redundant information. Thus, one could think about designing a simpler phantom to assess image quality from recorded images.

From the three developed methods, the DLO is the only one that allows one to determine contrast-detail curves without further, error-prone pre- or post-processing. The parametric approach, discussed in section 3.3.2, was found to reduce the workload for quality assurance measurements because only 5 images are needed to determine contrast-detail curves with the same accuracy as the EUREF guideline procedure using 16 images. The DLO can even reduce the number of images that are required even further. As visualized in Figure 4.11, the DLO is able to predict contrast-detail curves from single images as accurate as the guideline procedure using 16 images. In addition no further pre-processing or knowledge of the lesion positions is required. On the other hand, our parametric approach allows us to directly estimate an uncertainty of the prediction, which is not possible with our DLO, so far. The semantic segmentation observer determines an overall quality score of the segmentation performance, which could be used as a measure for image quality as well. The promising result of our segmentation approach indicates its potential usefulness for conformity assessment in mammography quality assurance if the question is whether a device produces sufficient image quality or not. However, the computation of these scores requires an existing ground truth mask that has to be constructed for real images.

Summing up, our DLO provides a proof of principle that deep learning can be successfully applied for mammography image quality assessment. To implement this approach in a clinical environment, the network should be rigorously trained and evaluated on a large set of real images from different vendors and for many imaging conditions. This is beyond the scope of this work but seems to be a worthwhile endeavor considering the good performance demonstrated in this work.

The limitations observed were that the deep learning observer (DLO) was majorly trained on simulated data and has to be rigorously tested on real data, before it can be applied in clinical practice. The trained DLO predicts the contrast-detail curve from data not used for training, such as data from a unknown device, accurate, but with a high variation. This uncertainty can be reduced by including similar images in the training data and retraining the model. Retraining the model on different training data will slightly effect its prediction behavior. However, this might challenge the use of the DLO for constancy tests that measures the performance of specific devices over time.

4.4 Summary

In this chapter two different methods are presented to tackle Objective 4 of this thesis. Both of them follow different strategies. First, deep learning is used to segment the image into different classes. By doing so, the network can learn to differentiate between lesion pixels and other pixels. After smoothing the image with a Gaussian kernel, this works with sufficient success. However, the result obtained in this way cannot be translated into an assessment of image quality in a direct way. First of all, if image quality is assessed in terms of contrast-detail curves, the application of post-processing is required. Scores for each single cell have to be computed to identify those cells, where the trained network cannot detect the lesion pixels with sufficient accuracy. Besides, the computation of semantic segmentation scores requires the knowledge of a ground truth mask. It follows that the uncertainty associated with the production of the CDMAM phantom directly influences the accuracy of the result.

To circumvent this, an approach was presented where the images of the CDMAM phantom were down-sampled and directly mapped to the corresponding contrast-detail curves. Incremental learning was applied to fine-tune a regression CNN and to extend the training data to better generalize the net. The trained network was found to estimate the image quality, based on single images only, as accurate as the reference method, the EUREF guideline procedure. Further, no cumbersome pre-processing is needed and the translation into the contrast-detail curve can be done without further knowledge of the phantom. As a result, the network does not perform a detection of the individual lesions, but separates the phantom image into a region where lesions can be detected and a region where lesions cannot be detected. Therefore, the high resolution images of the CDMAM phantom contain a lot of redundant information. Therefore, a sparse sampling approach as the random down-sampling that was used here can be used without losing too much information.

The presented approach was mainly trained and extensively tested on simulated images of the CDMAM phantom. Real images were included in the training set. For testing on real images a subset of the real images that were not taken for training was used. However, they were recorded with the same acquisition parameters on the same device as the real training images. Yet, the success of training a neural network to estimate the contrast-detail curve directly from single images is a proof of principle that using neural nets is advantageous for the task of image quality assessment in mammography. To apply this procedure in a clinical environment, a huge database of real images from different vendors and recorded with many different acquisition parameters would be required. Our results suggest that this would be a worth-wile endeavor which, however, is outside the scope of this work.

Despite the great success of the prediction of contrast-detail curves with deep learning, the results presented in this chapter provide point estimates for the values of interest. The variability of predictions was assessed by predicting a set of images and calculating for the mean and the variance. This gives a first rough estimate for the underlying uncertainty, while it would be a great advantage if the contrast-detail curve could be predicted along with an estimation of the associated uncertainty. This will be discussed in the next chapter. Without uncertainty estimation, conformity assessment cannot be carried out and meaningful comparisons of different results are not possible.

5 Explainability and Uncertainty of the Deep Learning Observer

Machine learning is a powerful tool, which in combination with high performance hardware allows one to develop algorithms that learn to solve specific tasks given large data sets. However, in critical applications or when human interaction is required it is often essential to understand why specific predictions are made and how trustworthy the model predictions are. Understanding model predictions requires an opening of their black-box nature, which can be done with interpretability techniques, such as explainability. Explainability methods aim to provide an explanation why the model ended up with the specific result. On the other hand, assessing the model confidence can be achieved by an analysis of the model uncertainty. The model confidence then expresses how trustworthy the predictions are, which is essential for conformity assessment. Therefore, this chapter is dedicated to these two important fields: explainability and uncertainty. In particular our contribution is to *estimate the uncertainty and to analyze the explainability of the developed deep learning observer* introduced in chapter 4.

Nowadays, machine learning applications perform as good as humans or better, in very different domains like object detection, speech recognition, cancer classification and many more. In many fields, artificial expert systems can be designed that work optimal in one specific task. The power of humans, on the other hand, is their ability to generalize their knowledge to various tasks instead of only one. So far, an optimal universal artificial intelligence, generalizing as good as the human brain, is not existing. However, if humans take a decision they are more or less incapable of objectively quantifying the underlying uncertainty. The confidence into a human decision is mainly judged based on their experience and their instinct. Mathematical models, on the other hand, allow an objective calculation or at least approximation of the underlying uncertainty of the decisions.

Machine learning models, in particular deep learning models, are often developed as black-box models. Given an input, they yield an output, without stating what particularly caused the model to end up with this result. This makes it challenging to understand model predictions and to identify their learning strategy. Lately, increasing interest initiated research focusing on interpretability of machine learning models. In addition, assessing a model's uncertainty allows one to determine how trustworthy its results are. Making decisions that can be interpreted along with quantifying the uncertainty of the decision is the key for the humans to understand and trust in machine learning applications. Knowing why the model predicts an output for a given input can be further used to resolve the learning strategy and to demonstrate successful training. This enables one to discover novel learning strategies to optimally solve certain tasks from a trained model.

5.1 Interpretability and Explainability

Artificial intelligence is more and more transforming from fiction to reality. The availability of data and computational power allows one to train algorithms to act as expert systems in many fields and applications (Amodei et al., 2016; He et al., 2016; Collobert and Weston, 2008). However, several questions arise on how artificial intelligence should interact with the human. Should artificial intelligence be designed and applied as stand-alone method or should all decisions be carried out with a human-in-the-loop, especially for safety critical applications? This further initiates ethical discussions regarding the concept of responsibility: who is responsible for a machine decision failure? For human-in-the-loop procedures, which by definition integrates human interaction, it is necessary to make interpretable predictions and assess how much the humans can rely on them.

To sum up, there is a vast demand for transparent machine learning models, which initiated the development of concepts to specify the interpretability of machine learning. A strict mathematical definition for interpretability does not exist. It can be described as the degree to which a human can understand the cause of a decision (Miller, 2019). The black box nature of machine learning models is replaced with a trail of transparent decisions, allowing a human user to understand how predictions and decisions are made (Samek, 2019).

The literature divides interpretability into a generalized description of how a model yields a decision and explainability (Molnar, 2019). Explainability aims to give a plausible explanation for one specific input and the corresponding model prediction (Samek et al., 2020). In this way, an explanation is the answer to a why-question: why did the model predict this given this input (Molnar, 2019)?

Interpretability opens the black box nature of machine learning models by allowing one to inspect the model and its predictions. Its techniques were successfully used to gain new insights into complex chemical processes (Schütt et al., 2017). The same can also be used to learn from machines. Interpretability and explainability are growing fields of research in machine learning. In this work, we focus on explainability. Several state-of-the-art methods will be reviewed briefly, which aim at explaining model predictions. The findings of this section are applied to the deep learning observer (DLO), that was introduced in chapter 4, to understand its behavior. There, a CNN was trained on a large set of simulated CDMAM phantom images as well as on a number of real images to estimate a contrast-detail curve. Images were sub-sampled before they were fed into the neural net. The DLO was found to accurately predict the contrast-detail curve even for images whose ground truth contrast-detail curve lies out of the range of curves used for training. In this section, it is investigated why our proposed DLO yields a corresponding contrast-detail curve for a certain input image. In particular, our aim is to identify whether there are pixels in the input image that influence the estimated contrast-detail curves significantly more than others.

5.1.1 Related Work

As previously described, several methods exist to open the black box machine learning model in order to understand the network's behavior. They can be roughly divided into two different approaches: explanation of the internal information of a trained network and explanation of the processing of an external input (Gilpin et al., 2018). An internal explanation aims to demonstrate the representation of the task in the trained net. For example, if a neural network was trained to classify images of dogs and cats into the corresponding class, the internal explanation aims to illustrate what the network "thinks" is characteristic for the class cat or dog. On the other hand, the external explanation would visualize why the network classifies a given input as cat or dog. Both approaches differ fundamentally and the interested reader is referred to the dedicated literature for further information (e.g. Montavon et al. (2018); Samek (2019); Lapuschkin et al. (2019); Bojarski et al. (2017); Taylor (2006)). Explainability can be interpreted as sensitivity analysis that identifies salient features (Saltelli et al., 2000; Baehrens et al., 2010). These salient features can be traced back to the input space and thus giving insight about the image features that are most relevant for the predicted output. In the literature, the sensitivity analysis for deep learning is carried out using different approaches. A salience map for the input image can be obtained by omitting parts of the image (occlusion sensitivity (Zeiler and Fergus, 2014)) and record the effect on the output. A further approach estimates the gradients with respect to the input, for determining salient features in the image (gradient sensitivity (Simonyan et al., 2013; Baehrens et al., 2010)). Lately, another approach, known as layerwise relevance propagation (LRP) (Bach et al., 2015), was developed that determines a relevance score for each feature and defines a set of rules to propagate these features back to the input. In the following section, the three approaches occlusion sensitivity, gradient sensitivity and LRP will be discussed in detail.

5.1.1.1 Occlusion Sensitivity

The basic idea of occlusion sensitivity is to record the influence of different parts of the image on the output of the model (Zeiler and Fergus, 2014). To do so, some parts of the image are modified (set to 0) and the resulting image is fed into the model. The deviation of the output of the modified picture and the output of the full image is then taken as the salience of the modified region on the output. Occluding different regions of the image and combining the determined saliences into a single map results in a so-called heatmap that visualizes how important individual regions are for the determination of the input. The size of the occluded regions is not predetermined so that even the influence of single pixels on the predicted output can be determined.

Occlusion sensitivity was successfully used to localize abnormalities in chest computer tomography images after classifying an image as abnormal (Islam et al., 2017).

5.1.1.2 Gradient Sensitivity

Occlusion sensitivity omits parts of the image to determine their influence on the output. In contrast, Simonyan et al. (2013); Baehrens et al. (2010) suggest a gradient method to explain neural network decisions by calculating the derivative of the class output with respect to the input. This approach derives either an internal or an external explanation, i.e the characteristic image that maximizes the corresponding class score or the regions of the input image that are most relevant for the output, respectively.

Given a class c and a scoring for this class $S_c(X)$ for an input X_0 they aim to rank the elements of X according to their relevance. With the assumption that the scoring function can be approximated by its first order Taylor expansion

$$S_c(X) \approx S_c(X_0) + \left.\frac{\partial S_c(X)}{\partial X}\right|_{X_0}^{T} \times (X - X_0)$$

the change of the scoring function can be expressed in terms of the gradient of the scoring function with respect to the input at the point X_0. Note, that X can either denote any feature within the network or the input image. The corresponding gradients can be determined using standard back-propagation. For a gray-scale image with dimensions $m \times n$ the corresponding saliency map $\mathcal{M} \in \mathbb{R}^{m \times n}$ can be determined as

$$\mathcal{M}_{ij} = \left| \left(\left.\frac{\partial S_c(X)}{\partial X}\right|_{X_0}^{T} \right)_{h(i,j)} \right|$$

where $h(i,j)$ is a mapping from images pixels i and j to the corresponding index of the gradient vector.

5.1.1.3 Layerwise Relevance Propagation

Another way of determining the influence of each pixel on the output is to use pixel-wise decomposition. Following this general idea, the layerwise relevance propagation (LRP) was developed by Bach et al. (2015) and evaluated originally for classification problems. Again, it is assumed that a classifier learns a real-valued mapping from the input image space to the output $S_c : \mathbb{R}^{n \times m} \mapsto \mathbb{R}^1$, where the index c denotes one particular class. In this scenario $S_c(x) > 0$ denotes the confidence of the classifier that the input image x belongs to class c. The network decision $S_c(x)$ for the input x can then be decomposed as

$$S_c(x) \approx \sum_{i,j} R_{h(i,j)},$$

where $R_{h(i,j)}$ denotes the relevance of each pixel x_{ij} given by the mapping $h(i,j)$ that maps the matrix component of the image with the vector index of the relevance scores. $R_{h(i,j)} < 0$ can be interpreted as the evidence against the presence of a structure, while $R_{h(i,j)} > 0$ indicates a high relevance of the image region for the classification output.

More precisely, Bach et al. (2015) define the LRP as a layer-wise decomposition of any hierarchical architecture. The output of each layer l is interpreted as a feature map that can be represented as a feature vector. The LRP then determines a relevance $R^{(l)}$ for the corresponding feature vector in layer l and by doing so propagates the relevance scores back to the input. The basic assumption is that the sum of all relevances in each layer is conserved through the architecture:

$$S_c(x) = \cdots = \sum_k R_k^{(l+1)} = \sum_k R_k^{(l)} = \cdots = \sum_{i,j} R_{h(i,j)}^{(1)} ,$$

where $R_{h(i,j)}^{(1)}$ denotes the relevance vector of the pixels in the input image.

This way, the class score is back-propagated through the architecture to compute a relevance score for the input. This result can then be translated into a heatmap that visualizes which parts of the image contribute most for particular class score $S_c(x)$ of an input image x.

Once the propagation rule is defined, a set of rules is necessary of how to compute the relevance vector of a corresponding feature vector. Assuming a fully connected layer, the output of neuron k in layer l can be obtained as (cf. chapter 2)

$$\alpha_k^l = \sigma\left(\sum_j w_{jk}^l \alpha_j^{l-1} + b_k^l\right) ,$$

where α_j^{l-1} denotes the output of neuron j of the previous layer $l-1$ and w_{jk}^l and b_k^l denote the weights and bias corresponding to the neuron k, respectively. For this case, one now wants to find out the contribution R_j^{l-1} of the neuron j for the relevance scores R_k^l in layer l. One possible rule that has shown good performance in practice is the so called $\alpha\beta$-rule (Bach et al., 2015):

$$R_j^{l-1} = \sum_k \left(\alpha \frac{\alpha_j^{l-1}\left(w_{jk}^l\right)^+}{\sum_j \alpha_j^{l-1}\left(w_{jk}^l\right)^+} - \beta \frac{\alpha_j^{l-1}\left(w_{jk}^l\right)^-}{\sum_j \alpha_j^{l-1}\left(w_{jk}^l\right)^-} \right) R_k^l ,$$

where $()^+$ and $()^-$ denote the positive and negative weights respectively and α and β are hyper-parameters with the constraint $\alpha - \beta = 1$ and $\beta \geq 0$ (Bach et al., 2015).

The relevance R_j^{l-1} is computed by weighting all relevances R_k^l with the weight w_{jk}^l between neuron j in layer $l-1$ and neuron k in layer l relative to the sum of all weights.

Proper LRP propagation rules have been defined for convolution layers, pooling layers and activation functions (Bach et al., 2015) and shown great success to determine robust explanations. Exemplary applications comprise the explanation of classification results for classic image classification (e.g. Bach et al. (2015); Samek et al. (2016); Lapuschkin et al. (2019)), in medical imaging (Sturm et al., 2016), text classification (Arras et al., 2017) and many more. Current research extends LRP to other layers, such as batch-normalization (Montavon et al., 2018; Samek, 2019; Binder et al., 2016).

5.1.2 Proposed Approach

It was shown that salient regions in the input image can be extracted for a classification with different approaches. The LRP calculates a relevance score for each feature and back-propagates them through the model to the input space. This method requires the definition of propagation rules for the different layer types of a neural network that are available for many common layer types including convolutional, pooling and fully connected layers. However, they have not fully defined proper propagation rules for other layers, such as batch-normalization. As mentioned in chapter 2, local re-normalization layers are recommended to re-normalize the features of each layer in the inner network structure.

The gradient method, on the other hand, requires the computation of the gradients with respect to the input features. In other words, the Jacobian of the network parameters needs to be determined. This can be done using the standard back-propagation algorithm explained in chapter 2. However, computing every parameter gradient with respect to the input becomes computational challenging for neural networks with a high number of parameters. If the image dimension is significantly smaller than the number of learnable parameters of the network, it can be more efficient to change individual pixels and compute the forward pass with the perturbed images. Changing each pixel of our down-sampled images of size 250 × 250 *pixels* requires the network evaluation of $62,500$ images, which was found to be faster than deriving the gradients of all ≈ 63 million parameters with respect to the input image. Therefore, occlusion sensitivity as well as a novel proposed approach are favorable in this particular scenario.

Instead of calculating the true gradients, their approximation with a difference quotient is used to determine sensitivity maps. Similar to occlusion sensitivity, this novel approach only requires to change the intensity of individual Pixels and to report the corresponding change on the contrast-detail curve. Different from occlusion sensitivity the relative change of the pixel value is considered in our approach.

In contrast to classification problems, the determination of a contrast-detail curve is a regression of twelve points $\psi(x, \theta) : \mathbb{R}^{n \times m} \mapsto \mathbb{R}^k$ with $k = 12$. Rather than being interested in the relative change of one class score, the goal is to determine the contribution to the output as a whole. For small changes of single pixels in the input the true gradient can be approximated as

$$\frac{\partial \psi(x, \theta)}{\partial x_{ij}} \approx \frac{\frac{1}{k} \sum_{i=1}^{k} \left(\psi^{(i)}(x + h, \theta) - \psi^{(i)}(x, \theta) \right)^2}{\|h\|},$$

where $x \in \mathbb{R}^{n \times m}$ denotes the multidimensional input, x_{ij} denotes the pixel with indices ij and $h \in \mathbb{R}^{n \times m}$ is a matrix that contains one small change h_{ij} at indices ij and zero otherwise. The salience map can then be determined in full resolution of the input image by changing each single pixel individually and combining the approximate gradients on the corresponding position in the salience matrix.

5.1.3 Results

In chapter 4, the DLO was introduced as a tool to estimate contrast-detail curves from single images of the CDMAM phantom, without error-prone pre- or post-processing. Evaluating the accuracy in terms of the root-mean-squared error and the generalization performance for images, whose ground truth contrast-detail curves lie outside all ground truth samples that were used to train the neural network, demonstrates its usability for the assessment of image quality in mammography. However, the model remained a black box. Here, it is explored how the black box nature can be opened to understand why this observer predicts a certain contrast-detail curve for a specific input image. The input image is part of the real test data set and shown in Figure 5.1.

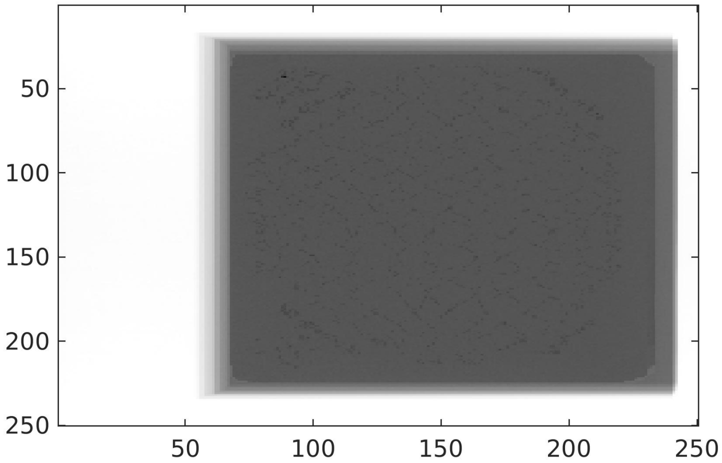


Figure 5.1: Real image of a CDMAM phantom on a logarithmic scale after sub-sampling. The white area marks the margin, where the pixel intensity is much higher than in the actual phantom region.

The image consists of a region that shows the actual phantom and one that shows the margin. This margin appears naturally when the detector size is larger than the actual phantom size. As a consequence, the x-rays in the margin region reach the detector without attenuation and hence, the pixel intensities are much higher here.

Different explainability methods were applied to this input image. The DLO consist of around 63 million network parameters (cf. chapter 4), which made it computationally inefficient to utilize the gradient sensitivity. The LRP is a promising tool that produces robust explanations in image classification problems. However, the DLO utilizes batch-normalization layers in the convolutional blocks, as shown in Table 4.7. These batch-normalization layers pose difficulties to the LRP. In private communication with the developers[1], LRP was applied ignoring the batch-normalization layers as a first approach. This results in more or less uniform heatmaps, as shown in Figure 5.3a. Batch-normalization performs a local re-scaling of features. Ignoring these re-normalizations seems to introduce an averaging with the consequence that all relevances become equal. In addition a LRP propagation rule for batch-normalization layers developed by Binder et al. (2016) was used. However, also in this case the LRP did not produce the expected result.

On the other hand, the batch-normalization layers were an essential part of the DLO, ensuring a stable convergence. Removing the batch-normalization layers from the original architecture in Table 4.7 and retraining on the training data caused a significant drop of the accuracy. It follows that the trained network estimates contrast-detail curves that are not useful, as shown in Figure 5.2.

Figure 5.2a shows an example of the contrast-detail curve estimation of the full DLO together with the corresponding ground truth. As it can be seen, the prediction follows the ground truth closely. Figure 5.2b shows the predicted contrast-detail curve after removing the batch-normalization layers from the original architecture and retraining the resulting network on the training data. Applying the LRP to this network, on the other hand, produced heatmaps that

[1] G. Montavon, Technical University Berlin

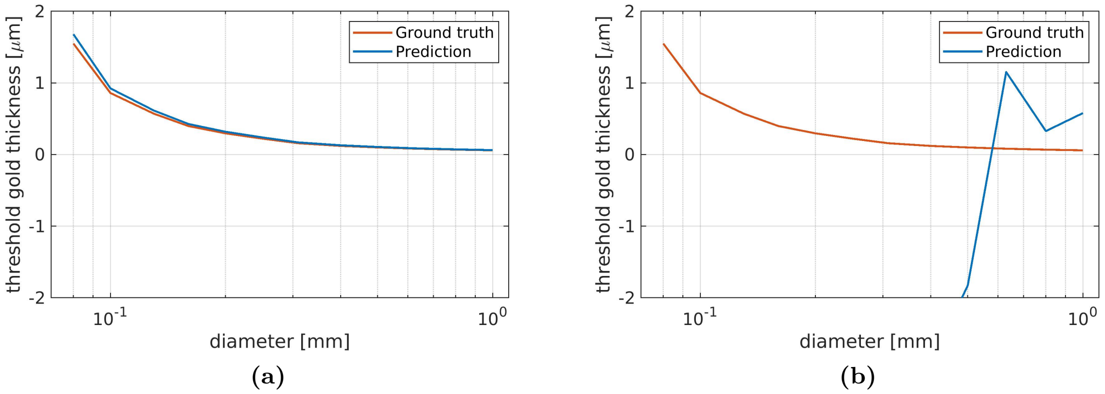


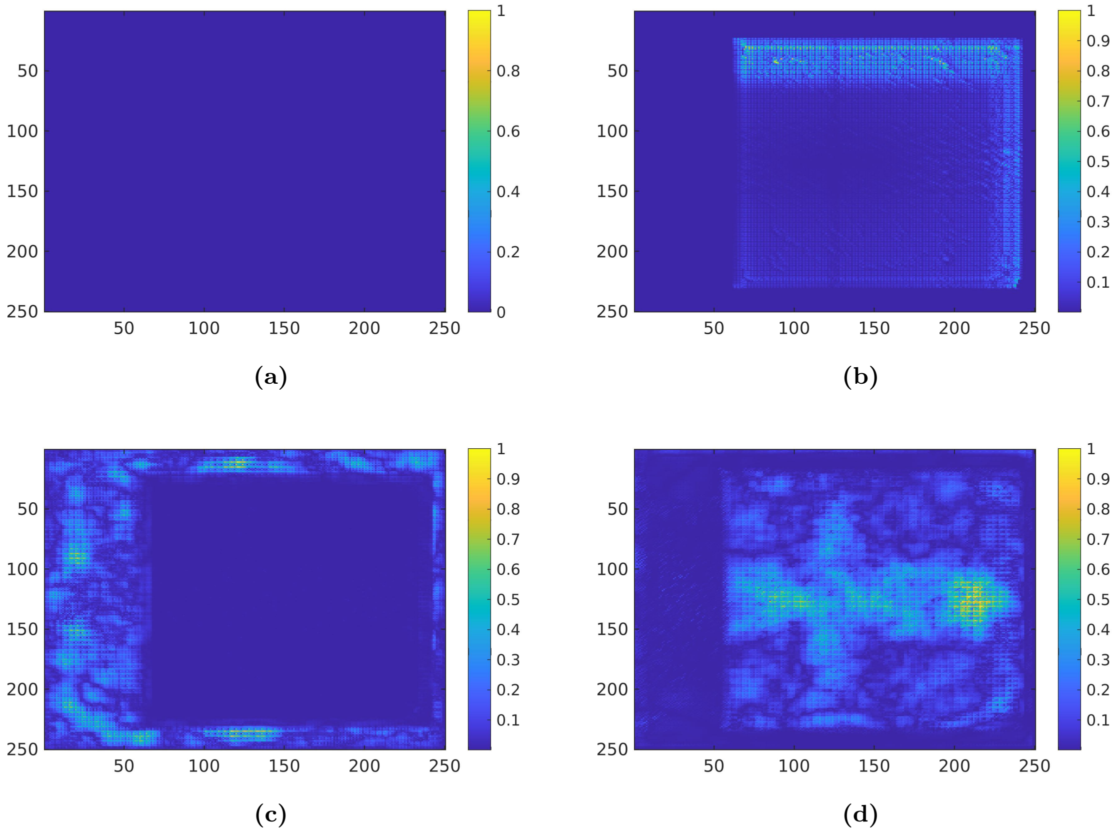
Figure 5.2: Ground truth contrast-detail curves (red) and estimated contrast-detail curves (blue) obtained by (a) the full DLO and (b) the DLO after removing the batch-normalization layers.

were no longer uniform, as visualized in Figure 5.3b. The yellow regions indicate pixels that were highly relevant for the prediction of the neural net. As it can be seen, the LRP sensitivity now highlights the actual area of the phantom, especially the top part of the phantom. So, removing the batch-normalizations resulted in better explanations obtained by the LRP, but the resulting model is not suitable to accurately predict contrast-detail curves.

Neither occlusion sensitivity, nor our proposed approach depend on the architecture of the network and hence, can be used for the original architecture of the DLO. Both methods are applied by perturbing the input image and recording the change in the output. Occlusion sensitivity perturbs an image by zeroing individual pixels or image regions. The sensitivity is then computed as the absolute deviation of the resulting prediction compared to the prediction of the original image. The amount of pixel change is not considered in this approach. Applying the occlusion sensitivity for an input test image (cf. Figure 5.1) results in the heatmap shown in Figure 5.3c. The occlusion sensitivity actually highlights the pixels in the margin area, which should not be the case, since the margin area does not contain relevant information for the image quality.

Our proposed explanation approach, on the other hand, perturbs an image by only slightly changing its intensity. Furthermore, for each perturbed pixel, the absolute deviation of the corresponding prediction to the original prediction is normalized to the degree of perturbation. Figure 5.3d shows the resulting sensitivity image for the real test image from Figure 5.1. The yellow regions, which indicate pixels of high relevance are concentrated in the region of the phantom, especially in the center of the phantom. The white areas in the input image that correspond to the margin are not relevant to the determination of contrast-detail curves.

This indicates, that the trained DLO successfully learned to ignore the margin of the image and estimates the contrast-detail curves based on the actual phantom image. Furthermore, Figure 5.3d gives the impression that not all pixels in the phantom are of the same relevance for the determination of the contrast-detail curve. Especially a horizontal band in the middle of the phantom image and a spot in the right side of the phantom image were responsible for the determination of contrast-detail curves. This horizontal band frames the diameter and thickness combinations of the gold discs in the CDMAM phantom and hence those cells, that actually represent the information that is given in the contrast-detail curve of this particular



(a)

(b)

(c)

(d)

Figure 5.3: Typical normalized sensitivity maps for the input image from Figure 5.1 determined with (a) LRP with batch-normalization in the architecture, (b) LRP after removing all batch-normalization layers from the DLO, (c) occlusion sensitivity and (d) our proposed approach. Yellow regions mark pixels with higher relevance for the neural network's prediction.

image.

5.1.4 Discussion

In chapter 4 our DLO was trained to estimate contrast-detail curves from single images of the CDMAM phantom. The results show that this approach determines contrast-detail curves estimated from single images of comparable accuracy as those obtained by the current recommended software in the EUREF guideline using 16 images. Classical automatic procedures require error prone pre-processing of the raw images. The DLO, on the other hand, utilized solely random sub-sampling of the images before a contrast-detail curve is estimated.

Several explanation methods were employed to understand how the observer can predict accurate contrast-detail curves, even when only 1% of the original image is used (cf. section 4.3.1.1). Explanation methods are used to derive heatmaps that highlight regions in the image which contributed most to the prediction of the neural net. The LRP approach, that demonstrated its applicability in many fields (e.g. Samek et al. (2016); Sturm et al. (2016); Arras et al. (2017)), failed to determine meaningful heatmaps in our case since the architecture of our neural network contained batch-normalization layers, which pose difficulties to the application of LRP. Neither ignoring the batch-normalization layers nor applying a propagation rule suggested by Binder et al. (2016) produced meaningful heatmaps for our DLO. The batch-normalization layers are essential for the performance of our deep leaning observer. Removing them from the architecture and re-training the resulting net on the training data caused a significant drop of the accuracy. Applying the LRP to the architecture without batch-normalization layers results in sensitivity maps that actually highlight the phantom region and hence, produces reasonable results.

Occlusion sensitivity and the proposed approach are independent of the network architecture. Sensitivity maps are obtained by perturbing an input image, applying the network and comparing the resulting prediction with the baseline prediction. As described in chapter 4, an image normalization is carried out that ultimately shifts all pixel intensities. Images are normalized by subtracting the pixel intensity of the background to the phantom region. As a consequence, a pixel intensity of zero is linked to the background of the phantom. Occlusion sensitivity perturbs an image by zeroing individual pixels or regions (Zeiler and Fergus, 2014). This means, that if a pixel in the margin are is zeroed, it then appears with the same intensity as the background in the phantom region. This affects the prediction of the neural net. Furthermore, the pixels in the margin area have much higher intensities. So, if they are set to zero, their intensity changes a lot more than pixels that belong to the phantom. This change is not considered in the sensitivity maps. As a result, the occlusion sensitivity maps actually highlight the margin area.

Our proposed approach aims to approximate the gradients of the output with respect to the input pixels through the difference quotient. In contrast to gradient sensitivity (Simonyan et al., 2013), we determine a sensitivity map without computing the gradients for every network parameter. Each pixel intensity is only slightly perturbed. The resulting sensitivity maps further consider the relative intensity change of each pixel. In contrast to standard occlusion sensitivity analysis, margin pixels are not altered so much that their pixel intensity suddenly appears to be the same as the background in the phantom region. The resulting heatmaps demonstrate that the trained DLO determines contrast-detail curves based on those pixels that belong to the actual phantom. Pixels in the margin area which contain no relevant information for the determination of the contrast-detail curve are ignored successfully. These margin areas were included in the images to yield a more robust and generalized data set as described in section 4.3.1.2. They appear naturally in real images when the detector area is larger than the phantom area. To be more representative, our simulated images were therefore augmented with margin areas of different sizes, including margins that are beyond the clinical practice. For all test images (simulated and real), our sensitivity maps show no influence of the margin pixels for the predicted contrast-detail curves. This indicates, the network successfully learned that the relevant information is contained solely in the pixels of the actual phantom.

Furthermore, our results show a horizontal band in our sensitivity maps that represents the part of the image that highly influences the resulting contrast-detail curve. This horizontal band is located in the middle of the CDMAM phantom. Exactly this area includes those gold discs that define the corresponding contrast-detail curve. Especially the small diameters (right part of the image) are of greater importance. It appears that the deviation of the predicted contrast-detail curve is bigger for smaller diameters and thus, their contribution to the root-mean-squared error is higher than those of the bigger diameters. Therefore, a change in the region of the small diameters results in higher deviations of the estimated contrast-detail curve, leading to higher relevance scores in our computed sensitivity map. A higher variability of the contrast-detail curves for smaller diameters is confirmed by Mackenzie et al. (2017).

The resulting sensitivity maps further confirm the interpretation of contrast-detail curve estimation in the context of mammography image quality assessment as mentioned in the discussion of chapter 4. The contrast-detail curve itself visualizes the boundary where an employed observer does not reach the required accuracy to detect the gold disc in the corresponding CDMAM cell. In other words, it separates the CDMAM phantom into cells where the observer succeeds in the location task and cells where it does not reach the required accuracy. Due to the regular structure of the CDMAM phantom, the separation between these cells is expected to be a line or a low order curve. Different image qualities would then result in a vertical shift of this separation boundary. Assigning such a separation boundary to an image is a totally different task than to locally check whether lesions can be detected with prescribed accuracy. Especially for the assignment of a separation boundary, the CDMAM phantom contains a lot of redundant information This might explain why the DLO can still accurately assess the contrast-detail curve even though the image was highly down-sampled.

The explanation results that were shown in this chapter emphasize this interpretation of contrast-detail curve determination by visualizing the horizontal band as the region that includes exactly this separation boundary. This is important to understand the neural network's behavior and it further confirms that a sparse sampling of the full image does not only save computational power but can be very helpful for successfully training the neural network and focusing on the regression of such boundaries. This way, our DLO "understood" that for the derivation of contrast-detail curves it does not have to carry out a localization task for individual lesions but can solve the task with the same accuracy by regressing a separation boundary to the image that is then re-cast into a contrast-detail curve.

5.2 Uncertainty

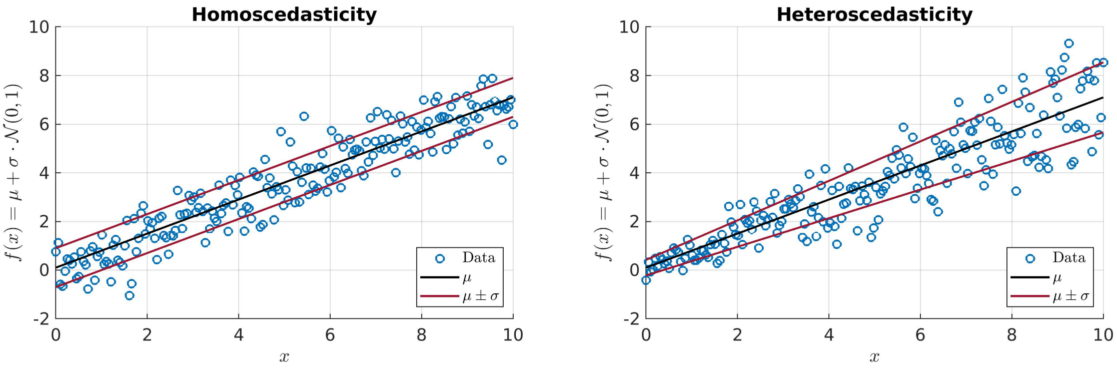
An explanation helps to understand particular predictions of a neural network, but does not allow us to assess how reliable the model predictions are. The uncertainty of a prediction is a crucial parameter to measure such a model reliability that once determined, quantifies how trustworthy the model decision is. An estimation of the associated uncertainty is especially important to make reliable decisions for conformity assessment (Pendrill, 2014; Loftus and Giudice, 2014). Approving mammography devices for screening purpose requires to check whether they produce sufficient image quality. Indeed, this can be seen as a conformity assessment. Therefore, this section is dedicated to the assessment of the measurement uncertainty for our developed DLO.

Broadly, deep learning is assumed to yield accurate predictions. However, quantifying the uncertainty of a model output is highly desired especially in safety critical applications (Kendall and Gal, 2017). Two notorious examples demonstrate the fatality of ignoring uncertainties of the prediction and assuming neural network predictions as accurate. In 2015, the Google photo app classified two African Americans as gorillas, raising a discussion about racial discrimination in machine learning (Guynn, 2015). In May 2016, the autopilot of a Tesla model S falsely classified the trailer of a truck as bright sky, resulting in a deadly crash (NHTSA, 2017). If those algorithms would have considered uncertainties of the prediction along with their predictions, the systems may have been able to yield better decisions.

Bayesian deep learning (cf. section 2.2.3) lends itself to model predictions along with associated uncertainties (Neal, 1993). In Bayesian deep learning, which was already proposed in the early 1990's (e.g. (MacKay, 1992)), the parameters of the neural network are modeled as random variables. In the forward propagation of an input, each network parameter is sampled from a posterior distribution. As mentioned in section 2.2.3, in many scenarios the true posterior for the model parameters cannot be computed and needs to be approximated. However, once a solid approximation is found, Bayesian inference can be used to model predictions along with their uncertainty.

But what exactly are uncertainties? In model-based decisions, system uncertainty occurs naturally because of imperfect or incomplete data, as well as because of stochastic fluctuations in the data (Walker et al., 2003). Depending on their properties, uncertainties are commonly split into reducible (epistemic) and irreducible (aleatoric) uncertainties (Der Kiureghian and Ditlevsen, 2009). Aleatoric uncertainties capture noise that is inherent in the data (Hüllermeier and Waegeman, 2019). They can originate from measurement imprecision of sensors or motion noise. Therefore, aleatoric noise cannot be reduced by including more data. Epistemic uncertainty, on the other hand, describes reducible noise that originates from a lack of knowledge, such as imperfect choice of a model class or insufficient amount of data (Hüllermeier and Waegeman, 2019). Accordingly, the epistemic uncertainty can be reduced by adding more information about the correct class of model and by adding more training data.

Errors can be further divided into homoscedastic and heteroscedastic errors. Homoscedasticity describes the scenario where every input sample contains a constant variance, whereas, in heteroscedasticity the error depends on the input, different inputs occur with different variance. Figure 5.4 demonstrates these two different types of errors for a simple linear function. The



(a) **(b)**

Figure 5.4: Example to demonstrate the different types of errors for a linear function. In the homoscedastic case (a) the data samples are distributed around the mean with constant variance, while in the heteroscedatic case (b) the variance increases for increasing values x.

homoscedastic case in Figure 5.4a shows a constant variance for all values x, whereas the variance in the heteroscedastic case, depicted in Figure 5.4b, increases with increasing values of x.

In a typical setting, it is assumed that the data consists of combinations of inputs and corresponding labels $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}$. Furthermore, there is an underlying function f that maps every training input x_i to the corresponding output y_i:

$$y_i = f(x_i) + \epsilon_i \,, \tag{5.1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ describes the observation noise with variance σ_i^2.

Deep learning now tries to learn a model $\psi(x, \theta)$ to closely approximate the data generating function $f(x)$ from the training data (cf. equation (5.1)). However, if the model class is chosen imperfectly, even the best trained model $\psi^{(opt)}(x, \theta)$ is not exact. The difference between the best model and the actual function $f(x)$ is often referred to as model error (Hüllermeier and Waegeman, 2019). Furthermore, because of insufficient data, the model that is actually learned is not the optimal model. An approximation error remains. Figure 5.5 demonstrates these uncertainties for an ensemble example.

There remains a model error between the best model $\psi^{(opt)}(x, \theta)$ and the actual function $f(x)$. Often, an improved model can be estimated as the mean of several individual models (Hüllermeier and Waegeman, 2019).

Another way to categorize errors are bias and variance. A bias occurs from wrong assumptions made about the model, or from systematic errors in its parameters caused, e.g., by the algorithm used for training. The variance, on the other hand, reflects the uncertainty in the estimated model due to the propagation of random errors in the training data to the estimated model parameters. The bias error usually decreases with increasing model complexity, while at the same time the variance increases. This is also known as bias-variance dilemma (Geman et al., 1992).

Commonly, models are designed that aim to approximate $f(x_i)$ and to estimate the variance σ_i^2 of the observation noise ϵ_i in equation (5.1) simultaneously (Kendall and Gal,

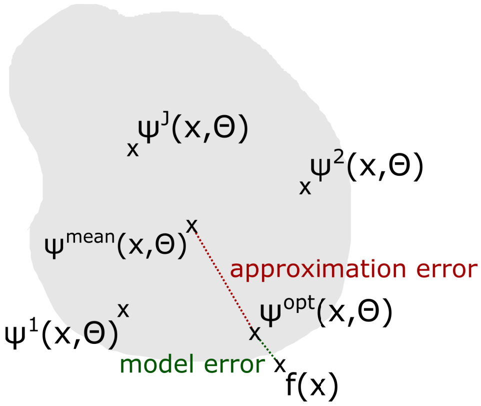


Figure 5.5: Schematic for the demonstration of approximation and model error for an ensemble approach. A model is supposed to approximate a data generating function $f(x)$. Even for the best model, an error remains. In addition, it cannot be perfectly approximated, thus, inducing the approximation error. The model error can be reduced by more information, whereas, the approximation error can be reduced by more training data.

2017). Therefore, the model output combines a predictive mean and a predictive variance

$$\psi\left(x_i,\theta\right)=\begin{pmatrix}\hat{\mu}_i\\ \hat{s}_i\end{pmatrix}\,, \tag{5.2}$$

where $\hat{\mu}_i$ approximates $f(x)$ and $\hat{s}_i=\log(\hat{\sigma}_i^2)$ is the logarithm of the predictive variance. Note that since for the estimated variance $\hat{\sigma}_i^2>0$ has to be true, one commonly models the logarithmic variance $\hat{s}_i$ with the neural network (Gal, 2016).
The prediction is then given as

$$\hat{y}_i=\hat{\mu}_i\,.$$

In the probabilistic interpretation, a proper learning objective is given as

$$\mathcal{L}_{MC}\left(\theta\right)=\frac{1}{M}\sum_{m=1}^{M}-\log\left(p\left(y_i\mid x_i,\theta\right)\right)+\frac{1}{N}KL\left[q_w(\theta)\|p(\theta)\right]\,,$$

where M denotes the size of a mini-batch, as demonstrated in equation 2.8 in section 2.2.3. Assuming a Gaussian likelihood, the negative log-likelihood for the neural network is given as:

$$-\log\left(p\left(y_i\mid x_i\theta\right)\right)=\frac{1}{2}\log\left(2\pi\right)+\frac{1}{2}\hat{s}_i+\frac{1}{2}e^{-\hat{s}_i}\left(y_i-\mu_i\right)^2\,. \tag{5.3}$$

This procedure, introduced by Kendall and Gal (2017) determines the predictive variance $\hat{s}_i$ as loss attenuation. Along with an expression of the Kullback-Leibler divergence in the training objective, it enables one to train inference models that approximate the posterior distribution of the model parameters and hence, completely depict the model uncertainty.

Mammography image quality assessment is done by regressing a contrast-detail curve that depicts the threshold gold thickness for 12 gold disc diameters of the CDMAM phantom, such that an employed observer can detect the lesion with prescribed accuracy. The estimated contrast-detail curve is compared with a limit curve. If the estimated curve is below the limit curve, the device is approved to produce images of sufficient quality. For such a conformity assessment, the estimation of measurement uncertainty is crucial (Pendrill, 2014). The following section provides a short overview of existing methods to estimate the uncertainties of the prediction for deep learning models.

5.2.1 Related Work

A robust estimate of the uncertainty of the prediction combines the uncertainty that is inherent in the data with the uncertainty of the estimation of the model parameters. One way to achieve uncertainties of the prediction is by applying Bayesian neural networks. As mentioned in the fundamentals section 2.2.3, the goal of Bayesian deep learning is to determine the posterior distributions of the model parameters to determine uncertainties of predictions. As discussed there, a direct calculation of the true posterior distribution is intractable in most applications, leading to the application of approximation methods. Several methods have been proposed in the literature, including Laplace approximations (MacKay, 1992) and Hamiltonian methods (Springenberg et al., 2016). In practice, however, Bayesian neural networks are harder to train, and the quality of the uncertainty of the prediction depends critically on the choice of prior distributions for the parameters as well as on approximation errors (Rasmussen and Quinonero-Candela, 2005). Solutions that modify existing models slightly and still enable a robust determination of uncertainties of the prediction are highly desired in practice. Recent developments show that dropout has an interpretation as sampling from the predictive distribution of a model and hence, as an approximate inference method (Gal and Ghahramani, 2016). Dropout was originally introduced to prevent overfitting by randomly dropping units from the neural network in the training procedure (Hinton et al., 2012; Srivastava et al., 2014). In its simplest form, dropout samples a mask z drawn from a Bernoulli distribution and multiplies this mask to the hidden activations. If p denotes the probability that a weight is dropped, then the corresponding Bernoulli distribution produces 0 with probability p and 1 with probability $1 - p$. A sample from the Bernoulli distribution has an expectation value of $1 - p$. Therefore the dropout mask is re-scaled by $1/(1 - p)$.

Using dropout further in the prediction phase can be interpreted as individual samples from an approximate variational posterior predictive distribution and hence, can be taken to determine the uncertainties of the prediction (Gal and Ghahramani, 2016). The author of this study further demonstrates that, as a consequence, for dropout distributions the Kullback-Leibler divergence in the Bayesian deep learning objective can then be approximated as:

$$\frac{1}{N} KL\left[q_{\Phi}\left(\theta\right) \| p\left(\theta\right)\right] = \frac{1}{N} \sum_{l=1}^{L} KL\left[q_{R_l}\left(W_l\right) \| p\left(W_l\right)\right] \tag{5.4}$$

$$KL\left[q_{R_l}\left(W_l\right) \| p\left(W_l\right)\right] \propto \frac{\delta^2}{2} \|R\|^2 - \mathcal{K} H(p)\,, \tag{5.5}$$

where R denotes the product of weight matrix W and dropout mask z, L denotes the number of Layers, $\mathcal{K}$ is the dimension of the layer, δ denotes a typical length-scale and $H(p)$ is the entropy of the random variable given as

$$H(p) = -p \log p - (1 - p) \log(1 - p)\,.$$

The Kullback-Leibler divergence acts as a regularization and combines standard L_2 regularization (Krogh and Hertz, 1992) and dropout regularization (Gal and Ghahramani, 2016). The interpretation of dropout as variational inference is not limited to Bernoulli dropout but generalizes to other forms of dropout masks.

It was found that using Bernoulli dropout slows down the training procedure (Wang and Manning, 2013). They introduced a Gaussian approximation that uses the benefits of dropout training without the need of sampling, which in return accelerates the training process. This way, Gaussian dropout is modeled as multiplicative Gaussian noise and the mask is directly sampled from a normal distribution

$$z \sim \mathcal{N}\left(1, \tfrac{p}{1-p}\right) ,$$

with mean 1 and variance $\frac{p}{1-p}$.

Similarly, concrete dropout was introduced as a possible relaxation of discrete Bernoulli dropout (Gal et al., 2017). The dropout mask is generated as

$$z \sim sigmoid\left(\frac{1}{t}\left(\log\ p - \log\left(1-p\right) + \log\ u - \log\left(1-u\right)\right)\right) ,$$

where $u \sim \mathrm{Unif}\left(0, 1\right)$ denotes a sample from a uniform distribution and t denotes a temperature (Gal et al., 2017). As a further advance, the dropout probability p is no longer set as a hyperparameter but adapted in the training procedure.

For Monte Carlo Variational Inference, the problem of high variances occurs that leads to poor convergence. The local re-parameterization trick (Kingma et al., 2015) addresses this issue. Furthermore, Kingma et al. (2015) discuss that their findings can be interpreted as variational dropout as generalization of Gaussian dropout. Their approach also determines an optimal learning rate adaptively as a parameter that is tuned via back-propagation during the training procedure. In contrast, they show that for their particular choice of posterior the negative Kullback-Leibler divergence in the Bayesian deep learning objective can be expressed as

$$-KL\left[q_\Phi\left(\theta\right) \| p\left(\theta\right)\right] \approx constant + 0.5 \log\left(\alpha\right) + c_1\alpha + c_2\alpha^2 + c_3\alpha^3 ,$$

where $\alpha = \frac{p}{1-p}$ denotes the noise and c_1, c_2 and c_3 are parameters given in (Kingma et al., 2015).

As an alternative, ensemble methods can be used as an approximation of the predictive posterior distribution (Harris, 1989). Bagging predictors to produce a combined prediction is not only shown to achieve robust and improved results (Breiman, 1996), but can also be used to quantify the uncertainty of model parameters (Fushiki et al., 2005). The idea is to train several models of the same type on different bootstrap samples of the whole training data (Breiman, 1996). However, deep learning is usually very data intensive. Training the neural network only on a bootstrap sample can therefore lead to reduced accuracy (Lakshminarayanan et al., 2017). In a recent study, Lakshminarayanan et al. (2017) replaced the bootstrapping of training data by including adversarial samples in the training set. Each ensemble that has a different subset of training images is augmented using adversarial examples.

Furthermore, Wu et al. (2018) report an approach that utilizes deterministic Variational Inference to derive robust uncertainties of the prediction.

5.2.2 Proposed Approach

Our contribution in this area is first of all to apply several of the described uncertainty methods for regression tasks on images. Potential procedures are primarily applied to a toy example, to choose a subset of optimal procedures among all the potential methods. A detailed description of the toy data is given in Appendix A.3. Every image is labeled with a curve that summarizes the local SNR information of various lesions in the image. This curve is similar to the contrast-detail curve and shall be estimated by a neural network. Utilizing a benchmark metric that is described in the next section, enables us to decide which methods work best on the toy data. These selected methods will then be applied to mammography image quality assessment.

The DLO with its architecture described in chapter 4 was incrementally trained to produce accurate contrast-detail curves. This, together with the large number of parameters, makes the DLO inefficient for testing several methods that require training from scratch. Therefore, we propose to replace the architecture of the original DLO with a ResNet architecture (He et al., 2016). The ResNet50 architecture is a very deep convolutional neural network. It consist of more than 40 convolutional blocks and only one fully-connected layer to map the features to the output. Such a ResNet50 architecture was taken and modified. The last fully-connected layer was replaced by two parallel fully-connected layers. The first of these layers was taken to compute $\hat{\mu}_i$. The other layer yields $\hat{\sigma}_i$. In cases where dropout was used, a corresponding dropout layer was included before the fully-connected layers.

Ensemble methods as well as dropout methods result in one estimate of $\hat{\mu}_i$ and $\hat{\sigma}_i$ per ensemble or dropout sample, respectively. Following Kendall and Gal (2017), the averages

$$\begin{aligned}\tilde{\mu}_i &= \frac{1}{J}\sum_{i=1}^{J}\hat{\mu}_i,\\ \tilde{\sigma}_i^2 &= \frac{1}{J}\sum_{i=1}^{J}\hat{\sigma}_i^2 + \frac{1}{J}\sum_{i=1}^{J}\hat{\mu}_i^2 - \left(\frac{1}{J}\sum_{i=1}^{J}\hat{\mu}_i\right)^2 ,\end{aligned}$$

can be taken, where $\tilde{\mu}_i$ is the final predictive mean and $\tilde{\sigma}_i$ is the uncertainty of the prediction.

5.2.2.1 Benchmark metric

In this section, a criterion is suggested that allows us to quantitatively compare different probabilistic models. Our criterion combines the accuracy of the regression with a measure of the quality of uncertainty. A robust and frequently used measure for accuracy of a regression problem is the mean percentage error (MPE) (Bowerman et al., 2005; De Myttenaere et al., 2016; Kim and Kim, 2016). The MPE is defined as:

$$\mathrm{MPE}_i = \frac{100\%}{P}\sum_{p}\frac{\left|y_i^{(p)} - \tilde{\mu}_i^{(p)}\right|}{\left|y_i^{(p)}\right| + \zeta} ,$$

where y_i denotes the multivariate ground truth for input x_i, μ_i denotes the corresponding predicted mean of the neural network ψ, and ζ denotes a numerical stabilizer. The MPE quantifies the deviation of the prediction and the corresponding ground truth.

A fair benchmark should consider the regression accuracy along with the predicted uncertainty. The uncertainty of the prediction $\tilde{\sigma}_i$ for each input is given as the standard deviation of the different output samples. In general, lower estimates of uncertainty are preferred over high uncertainties. However, the actual value of the ground truth should fall into the range of the predicted value and its (extended) uncertainty. The neural networks output multivariate curves. The accuracy was expressed univariately in terms of the MPE. In the following, a framework is proposed that computes a univariate measure of the goodness of the model uncertainty. More information of the quality of the determined uncertainties is conserved studying the point-wise coverage probabilities (Berger, 2013) for each point of the estimated contrast-detail curve. The coverage probability is estimated by the proportion of samples where the interval estimates around the prediction contains the ground truth.

However, a univariate measure is still helpful to roughly compare different methods. Hence, a measure is derived that expresses the quality of the uncertainty prediction. First, for a multivariate output the number of outliers ($y_i \notin [\mu_i - k \cdot \sigma_i, \mu_i + k \cdot \sigma_i]$) is counted as

$$\mathrm{OUT}_i = \frac{1}{P} \sum_p \mathcal{H} \left(\frac{\left| y_i^{(p)} - \tilde{\mu}_i^{(p)} \right|}{k \cdot \tilde{\sigma}_i(p)} - 1 \right) ,$$

where $\tilde{\sigma}_i$ is the model uncertainty, $\mathcal{H}$ denotes the Heavyside function and k is the factor for the expanded measurement uncertainty with $k = 1.96$ for the 95% interval under the normality assumption. This part counts how often the ground truth value lies out of the prediction plus or minus the expanded standard deviation. It varies between 0 if all ground truth values lie within these intervals and 1 if the predicted range does not cover the actual value at all. Moreover, a best model is the one with minimum uncertainty such that (almost) all ground truth values can be explained via the predicted intervals. Hence, the size of σ itself should be considered in our combined uncertainty measure. For a fair combination, this contribution should also vary between 0 and 1 such that both contributions remain comparable. For a series of measurements, the probability of the uncertainty of a single measurement to fall in the interval $[p, q]$ can be expressed via the error function (Glaisher, 1871). Therefore, we take the error function to ensure that the contribution of the size of $\tilde{\sigma}$ is in $[0, 1]$. Specifically, we suggest

$$\mathrm{SIG}_i = \frac{1}{P} \sum_p erf \left(a \cdot \frac{k \cdot \tilde{\sigma}_i^{(p)}}{\left| y_i^{(p)} \right| + \zeta} \right) ,$$

where $erf()$ stands for the error function to ensure that the output varies between 0 and 1 and a denotes a scaling constant that can be chosen such that the convergence towards 1 is faster or slower. A value of $a = 4$ was chosen such that for $\sigma_i / \left| \mu_i^{(k)} \right| = 0.5$ the output was close to 1.

The final measures are averaged for different samples i of the same ground truth

$$\overline{\mathrm{MPE}} = MEDIAN\left(\mathrm{MPE}_i\right), \ \overline{\mathrm{SIG}} = MEDIAN\left(\mathrm{SIG}_i\right), \ \overline{\mathrm{OUT}} = MEDIAN\left(\mathrm{OUT}_i\right) .$$

The final uncertainty quality metric is then determined by a weighted sum of $\overline{\text{SIG}}$ and $\overline{\text{OUT}}$:

$$\overline{\text{UNC}} = w \cdot \overline{\text{SIG}} + (1 - w) \cdot \overline{\text{OUT}}\,.$$

where w decides the ratio of importance for the size of the uncertainties against the number of outliers.

5.2.3 Results

First, the modified ResNet50 architecture was trained on the toy data explained in Appendix A.3. For testing, two independent data sets were created. The performance of the introduced methods for uncertainty assessment was measured with our proposed benchmark metric. The result of the benchmark test is shown in Figure 5.6.

Our benchmark with the toy data set revealed that the ensemble method, Bernoulli dropout, and concrete dropout performed best.

Therefore, these methods were applied for the actual task of contrast-detail determination for mammography. The training data, explained in chapter 4, was taken to train the modified ResNet50 architecture. First, six models were trained as an ensemble. To do so, incremental learning was used by pre-training the neural net on the simulated images without any margins and then fine-tune six models on the whole data set. For the simulated test images, the performance of the network is shown in Figure 5.7a. Figure 5.7a shows the estimated contrast-detail curve for a single image along with the uncertainty of the prediction.

The contrast-detail curves are approximated very accurately by the neural network. Each of the two test sets contains 200 images. Figure 5.7b visualizes the coverage probability for the two contrast-detail curves for the 200 images. Especially the higher contrast-detail curve that is linked to the lower image quality is explained very well with the estimated uncertainty. Accordingly, Figure 5.8 shows the estimated contrast-detail curve along with uncertainties and coverage probabilities for the independent simulated test data set.

As it can be seen, the independent simulated test data can also be approximated accurately.

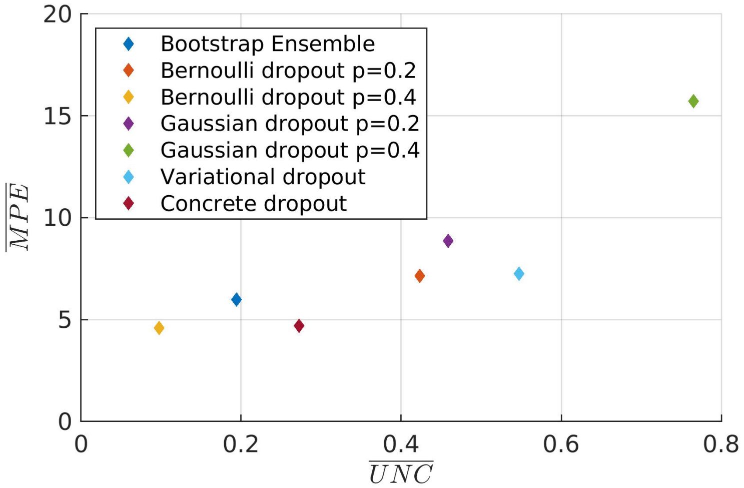
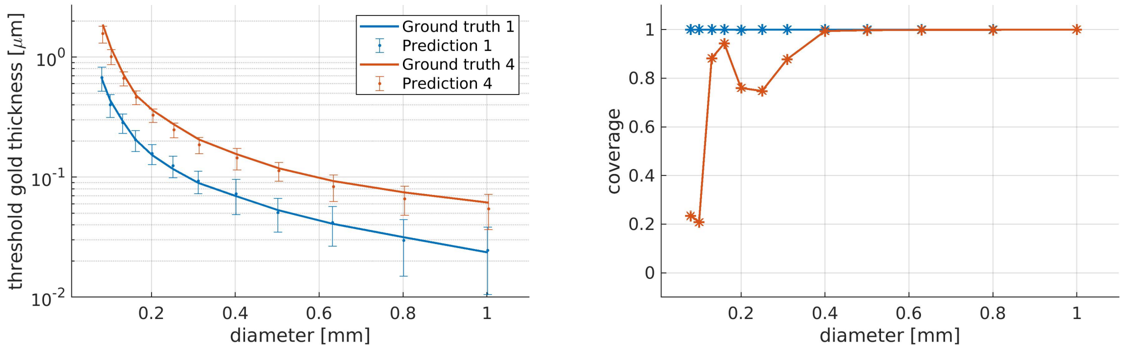
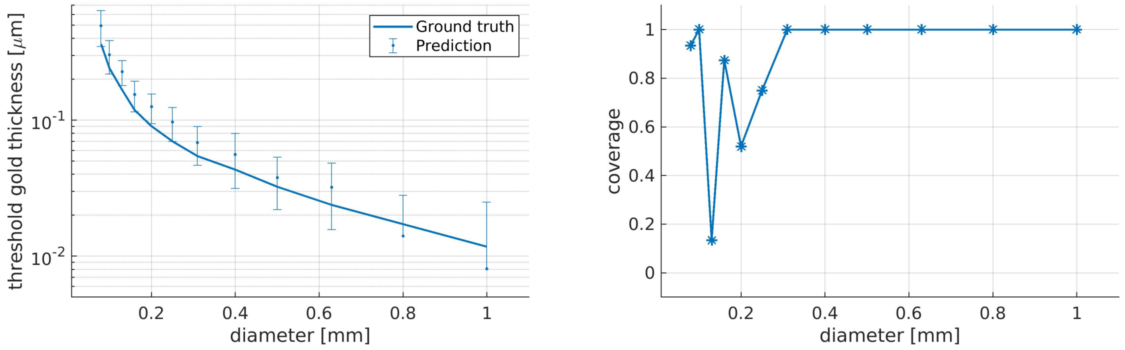


Figure 5.6: Results of the benchmark test on the toy data set. The y-axis shows the mean percentage error $\overline{\text{MPE}}$ and the x-axis displays the proposed uncertainty quality metric $\overline{\text{UNC}}$.



(a) **(b)**

Figure 5.7: (a) Estimated contrast-detail curve and (b) coverage probabilities for the simulated test data. Error bars show 95% intervals.



(a) **(b)**

Figure 5.8: (a) Estimated contrast-detail curve and (b) coverage probabilities for the independent simulated test data. Error bars show 95% intervals.

However, the estimated uncertainties for the independent test data are much higher. In chapter 4, the variability of the EUREF guideline procedure using 16 images of the real test data was compared with the variability of the DLO for single images. Figure 5.9 shows the variability of the EUREF guideline procedure (red) and the uncertainty of the prediction of the trained ResNet50 for a single image.

Especially for the larger diameters, the uncertainties of the prediction are estimated higher than the variability of the EUREF guideline procedure. For the smallest diameter, the estimated uncertainty is in the same range. Table 5.1 lists the calculated predictive variances for the different test sets. The variances are normalized to the ground truth and averaged over the values for one image. The displayed values present the median value for the different images of each set.

It can be seen that the predictive variance for the independent test set is significantly higher than those for all other test sets.

5.2.4 Discussion

Knowing the uncertainty is essential for a reliable conformity assessment (Pendrill, 2014). Hence, it also should be considered for mammography image quality assessment, where

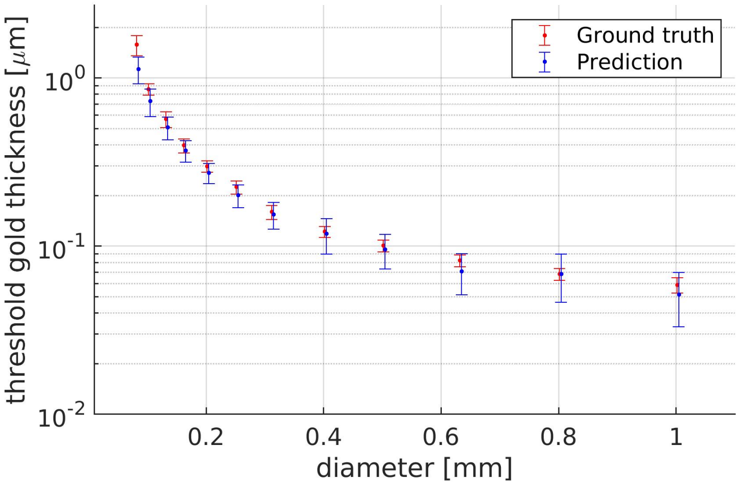


Figure 5.9: (red) Variability of the EUREF guideline procedure with 16 images (cf. chapter 4) and (blue) estimated contrast-detail curve plus uncertainty of the DLO from a single image. Error bars show 95% intervals.

Table 5.1: Uncertainties of the prediction normalized to the corresponding ground truth and averaged over the twelve points for each sample of the corresponding data set. The presented scores are the median of all values for all images in the corresponding set.

Set	$\tilde{\sigma}/y$
Simulated test set 1	0.1175
Simulated test set 2	0.0930
Simulated test set 3	0.0748
Simulated test set 4	0.0649
Independent simulated test set	0.1875
Real test set	0.0713

estimated contrast-detail curves are compared with a limit curve. Mackenzie et al. (2017) used uncertainties to demonstrate that the determination of contrast-detail curves obtained by the EUREF guideline procedure with 16 images is not accurate enough. They found that especially for the small diameters the variability of the determined threshold gold thickness is too high for a reasonable comparison to a limit value.

Our DLO, introduced in chapter 4, determines contrast-detail curves from single images as accurate as those obtained by the EUREF guideline procedure using 16 images. Recently, several methods have been developed, that assess the uncertainty of the prediction of deep learning models. These methods usually predict an observation noise along with an approximation uncertainty. Observation noise is inherent in the training data and and can be learnt as loss attenuation in the training procedure (Kendall and Gal, 2017). Besides, for an imperfect class of models, the data generating function cannot be modeled exactly. Furthermore, finite training samples cause an approximation uncertainty (Hüllermeier and Waegeman, 2019).

Most of the methods that aim to assess the uncertainty of the prediction require many additional training steps. The DLO architecture comprises more than 63 million parameters. Only by incorporating an incremental learning procedure, it was possible to predict contrast-detail curves accurately. However, the many parameters and the complex training procedure impede the application of uncertainty methods. Therefore, the original architecture of the DLO was replaced by a ResNet50 architecture (He et al., 2016). A ResNet50 comprises 11 million parameters. Less parameters reduce the risk of overfitting and speed up the training time. The accuracy of the ResNet50 is comparable to that of the original DLO, if the mean-squared error is taken as the loss function. However, when considering a predictive variance, which is learned by modifying the loss function, the accuracy drops slightly.

The network was learned using a log-likelihood loss function. Treating the observation noise as loss attenuation allows it to be estimated directly in the training procedure, as introduced by Kendall and Gal (2017).

Promising methods to assess the uncertainty of the prediction comprise ensemble methods and dropout. Ensemble methods and dropout can be interpreted as sampling from the posterior predictive distribution in probabilistic deep learning (Lakshminarayanan et al., 2017; Gal and Ghahramani, 2016). A sample of the predictive mean and predictive variance can be obtained by model averaging. Our results demonstrate that these methods which prove their benefit in simple regression tasks and image classification tasks are also applicable for the complex regression of a contrast-detail curve for image quality assessment in mammography.

Based on a simple toy example, a developed benchmark test showed that the ensemble method, Bernoulli dropout, and concrete dropout produce the best estimates of uncertainty, such that the ground truth values are covered by the 95% interval. Applying these procedures further on the simulated mammography images to estimate contrast-detail curves revealed that dropout methods are less suitable than ensemble methods in this context. They tend to overestimate the observation noise and the predictions are not very accurate. Ensemble methods, on the other hand, produce robust estimates of contrast-detail curves and compute uncertainties that cover the ground truth. As a result, an ensemble of 6 individual models estimates a contrast-detail curve accompanied with an uncertainty for a single image.

Our results show that the estimated uncertainties for a single image are similar to the variability of repeated applications of the EUREF guideline procedure using 16 images, especially for the smallest diameter. Table 5.1 further demonstrates that the predictive variance is much higher for the independent data set. This can be explained since the independent test set is outside of the distribution of the contrast-detail curves within the training set. The elevated uncertainty for this case correctly identifies this test set as "out of distribution" (cf. Kendall and Gal (2017); Gal (2016)).

So far, the dropout methods were not suitable for our mammography image quality data set. Including any kind of dropout layer into the ResNet50 architecture caused the training accuracy to decrease significantly. We conjecture that this occurs because dropout is applied before the output layer, which gives the model no chance to learn compensating the noise introduced by dropout. Therefore, the noise in the estimates of the final layer is higher, leading to worse results than those obtained by the ensemble approach (cf. Hinton et al. (2012)). However, the ensemble method produced very reasonable estimates of the uncertainty of the predictions.

The determination of the uncertainty of the predicted contrast-detail curves obtained by our DLO shows reasonable results, while there are still several limitations in the approach applied. The underlying statistical model, and hence the employed loss function, do not account for errors in the input x. This can lead to biased estimates in regression tasks Fuller (2009), and such a bias has not been considered in our uncertainty estimation. Furthermore, our final benchmark metric weights the average size of the uncertainty intervals with the total number of outliers. The choice of these weight factors influences the produced results, which means that a model that is optimal for one choice of weight factors is not necessarily the optimal method, when choosing different weight factors.

5.3 Summary

This chapter was dedicated to the explanation of deep learning and the estimation of uncertainty of the prediction. Explanations are important to understand a neural network's behavior. Robust estimates of the uncertainty of the prediction reflect their reliability. Hence, they allow a human to decide how much to trust the predictions of a neural network.

A difference quotient method was used to determine sensitivity heatmaps that highlight which pixels of an image dominantly contribute to the estimated contrast-detail curve. This procedure does not require the computation of gradients. Therefore, it is also applicable in situations where the user has no access to the neural network parameters directly but can only collect the output for a given input. Our results show that the DLO estimates a contrast-detail curve based mainly on a horizontal band in the CDMAM phantom image. Carrying out the explanation analysis allowed us to understand that the trained network separates the image into detectable and non-detectable discs. This is a different task than performing the local detections of the lesions in the individual cells. The different learning strategy of the DLO further revealed that the random down-sampling is beneficial for this regression task.

Furthermore, several methods for determining the uncertainties of predictions of deep learning were tested on a toy data set. Ensemble method, Bernoulli dropout and concrete dropout worked best on this data. However, for the mammography image quality data dropout methods did not perform well. The ensemble method, on the other hand, produces robust estimates of the uncertainty of the prediction. This way, accurate estimates of the contrast-detail curve can be made from single images, along with a robust estimation of the uncertainty.

6 Conclusion and Outlook

The goal of this thesis was the development of automatic procedures in the context of mammography image quality assessment. Current state-of-the-art methods comprise expert pre-processing with the application of model observers to a estimate contrast-detail curve which visualizes the ability of a device to depict small structures. So far, these methods require the recording of multiple images, expert pre-processing and exact knowledge of the lesion positions in the image. Three alternative methods are developed and discussed in this thesis which improve the current method and reduce the workload for quality assurance measurements. Specifically, four objectives were defined in the beginning of this work that are addresses subsequently.

In chapter 3, the first two objectives are discussed. The first objective of this thesis is driven by the utility that simulation tools give for method development and assessment. Specifically, our first contribution is the implementation of a virtual mammography that follows state-of-the-art methods for mammography image simulation. This tool allows us to simulate realistic mammography images, optimized for simulations of the CDMAM phantom. It is shown, that with our virtual mammography realistic mammography images of the CDMAM phantom can be simulated even though image degradation is modeled purely phenomenologically. Furthermore, the simulated images were used to assess the performance of methods for contrast-detail determination. Our second contribution is the development of an alternative method to determine the contrast-detail curve from images of the CDMAM phantom. The approach is based on a recently developed parametric observer. Specifically, a routine is developed that applies the modified parametric observer on single cells of the CDMAM phantom to determine a contrast-detail curve. This way, the contrast-detail curve is determined based on the AUC as an established figure of merit in task-based image quality assessment. Furthermore, a Bayesian approach is proposed to determine contrast-detail curves accompanied with an uncertainty for each point, which is not possible with the current practice. The novel approach improves the current practice by reducing the amount of pictures that are required to determine accurate contrast-detail curves. However, expert pre-processing is still required. An advantage of our method is that estimates of the uncertainty of the prediction can be made. These are necessary to take meaningful decisions in cases where the resulting contrast-detail curve is close to the prescribed limit curve. For future perspective, an end-to-end implementation has to be provided that combines the pre-segmentation of the images with the application of our proposed approach. Furthermore, the contrast-detail curves determined by our approach have to be calibrated carefully, which can be done by choosing the free factor γ. This calibration should ensure that the results obtained by our automatic procedures correlate well with those obtained by a human observer. To this end, the novel approach was extensively tested for simulated images, where it demonstrated its potential benefit. Before it can be applied in clinical practice it has to be thoroughly tested on real images, especially on images of devices from different manufacturers.

Expert pre-processing of images for contrast-detail estimation can be seen as feature extraction in classical machine learning. Deep learning is able to learn meaningful features directly from the data. Chapter 4 presents our third contribution, which is that it was explored whether it is possible to utilize deep learning to estimate image quality directly from the image without further pre-processing. Specifically, two different methods are developed that follow different strategies. First, semantic segmentation was utilized to train a semantic segmentation observer (SSO) that classifies each pixel of the image as lesion, background or grid pixel. Classical semantic segmentation metrics were found to correlate well with image quality, but require the knowledge of a ground truth mask. Such a mask is easily available for simulations, whereas it is expensive to determine for real images. In addition, a deep learning observer (DLO) is developed that estimates contrast-detail curves directly from images of the CDMAM phantom, without cumbersome pre-processing. The developed DLO is able to accurately estimate contrast-detail curves from a single image that is strongly down-sampled. Summarizing, deep learning can improve image quality assessment since it avoids error-prone pre-processing. The DLO estimates contrast-detail curves from a single image with comparable accuracy as the current state-of-the-art method which requires 16 images. Future work could modify the semantic segmentation observer that demonstrated its potential benefits, such that it can be utilized to determine contrast-detail curves. In addition, this procedure can potentially be adapted to serve as pre-processing for methods such as those developed in chapter 3. Image quality assurance in mammography determines the contrast-detail curve and classifies a corresponding device as approved if the curve falls below a certain threshold curve. Simplified, this can be viewed as a binary classification. With a data set that contains images that emulate the two classes sufficient and insufficient image quality, one could check whether the SSO can distinguish between these two classes. Besides that, one could think of designing a binary classifier, that classifies an image as sufficient or insufficient image quality, respectively. This classifier might be easier to learn since this classification task is a simpler task than the regression of the contrast-detail curve. However, the contrast-detail curve provides more information about the imaging performance, such as which details can be accurately perceived in the image. Furthermore, the concept of determining the image quality by semantic segmentation of a technical phantom could serve as a benchmark for different computer-aided diagnostics algorithms that utilize semantic segmentation to detect lesions in real mammograms. The structure of these phantoms is well known, therefore a ground truth mask can be determined, which can be used, to quantify the performance of different algorithms in this case. The DLO, on the other hand, can be applied directly to images to determine a contrast-detail curve. Before it can be applied in clinical practice, it has to be tested on a large set of real clinical images from devices of different manufactures. Furthermore, the method would benefit from retraining the model on such an extended data set. For real world application a procedure hast to be defined when the DLO should be retrained on new data and how this retraining effects constancy testing.

Deep learning has proven to be beneficial in many tasks and applications. Also in this work, it is shown that mammography image quality assessment may benefit from the employment of deep learning. Error-prone pre-processing is no longer required and our proposed deep learning observer does not depend on the exact knowledge of the positions of the lesions in the phantom. Despite the accurate estimation of contrast-detail curves, it was not clear how the DLO was able to work with highly down-sampled images. In our forth and last objective that is discussed in chapter 5, we investigate the reliability and explainability of our deep learning approach for image quality assessment in mammography. Specifically, we contribute by explaining the predictions made by our trained DLO and by estimating the uncertainty that is accompanied with the determined contrast-detail curve. The application of robust explanation methods revealed that the DLO successfully learned to ignore irrelevant margin areas. Contrast-detail curves are solely determined by the image pixels that actually belong to the phantom. To be more specific, a small horizontal band in the center of the phantom was found to dominantly influence the resulting contrast-detail curves. In conclusion, the proposed deep learning observer learned to split the image into a region where discs can be accurately detected and regions where this is not the case. This information was then recast into contrast-detail curves. Besides an explanation of the network predictions, the estimation of an uncertainty of the prediction is required to take a reliable decision in conformity assessment. For deep learning, this uncertainty can be estimated with different methods. In a simplified example, ensemble methods and concrete dropout were found to produce optimal estimates of the uncertainty of the prediction in terms of coverage probabilities and a proposed uni-variate benchmark metric. In the context of mammography image quality assessment, only ensemble methods produced suitable results. Summing up, with a slight modification of the model architecture, neural networks can estimate an uncertainty of a prediction, without significant loss of accuracy. Especially ensemble methods were found to produce robust estimations of uncertainties of the predictions of our trained DLO. The results presented in this chapter give insights of the learning strategy and quantify the uncertainty accompanied with the prediction. Looking ahead, the explanation analysis revealed that the DLO learned to map the contrast-detail curve to highly down-sampled single images of the CDMAM phantom. This shows that the CDMAM phantom contains redundant information which is not required for the determination of contrast-detail curves with our DLO. Therefore, the DLO can potentially infer meaningful image quality from simpler phantoms than the CDMAM phantom. Furthermore, it could potentially be possible to determine contrast-detail curves directly from real patient images for online image quality assessment. In order to develop such an application, a large data set of different patient images from various devices along with the corresponding contrast-detail curve from each device would be needed. Also, the initial findings regarding the estimation of uncertainty have to be analyzed further. Especially, the different types of uncertainties have to be treated more carefully. The aleatoric uncertainty that is present in the data suggests to use an errors-in-variables model. To do so, a modified learning objective has to be defined. Uncertainty quantification for image regression is relevant also in other contexts and our proposed benchmark could be further developed to provide means for validating uncertainty estimations in such cases. Ideally, this should be supported by benchmark test sets and software that automatically assesses uncertainty estimation procedures in terms of their results for these test sets.

In general, our proposed methods for image quality assessment in mammography are thoroughly tested and evaluated on simulated images. Before they can be applied in clinical practice our methods should be tested on large sets of real images. In addition, it could be worthwhile exploring whether our methods allow us to further reduce the radiation dose that is required for accurate detection of details in the images of technical phantoms. The contributions in this thesis focus on mammography as imaging modality. Our results show that image quality assessment in mammography benefits from the proposed methods. In addition to mammography, image quality is a critical parameter also for other imaging modalities such as tomosynthesis, computer tomography, sonography, and magnetic resonance imaging. It would be intriguing to investigate if these methods can be further adapted such that they can also be successfully applied for other imaging modalities.

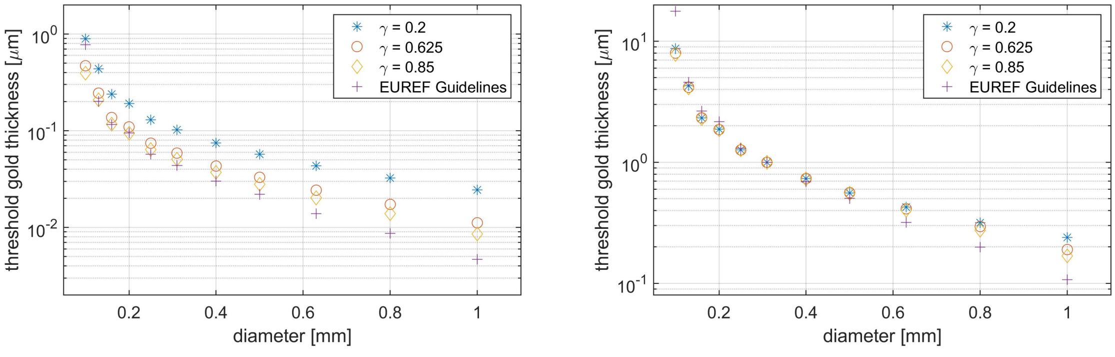
A

Supplemental results

A.1 Calibrating the Parametric Model Observer

In chapter 3 an alternative determination of contrast-detail curves was proposed. A method is suggested that utilizes a parametric observer to locally calculate the AUC for every cell of the CDMAM phantom. For the calculation of the AUC, the proposed approach allows one to choose a free parameter γ, see equation (3.2). Figure A.1 visualizes how the choice of γ affects the resulting mean contrast-detail curve.

The choice of γ scales the argument of equation (3.2). As a consequence, the resulting logistic curves (cf. Figure 3.6a) are shifted either to the right or to the left. This, in turn, leads to the change of the resulting contrast-detail curve indicated in Figure A.1a. However, changing the value of γ only shifts the entire contrast-detail curve with respect to the threshold gold thickness. Figure A.1b shows the normalized contrast-detail curve. This normalization was done by dividing the contrast-detail curves by their corresponding threshold gold thickness



(a) **(b)**

Figure A.1: (a) Variation of the obtained contrast-detail with our proposed approach, depending on the choice of the free factor γ. (b) Contrast-detail curves normalized to the corresponding threshold gold thickness for a diameter $d = 0.25\ mm$. Figure from Kretz et al. (2019).

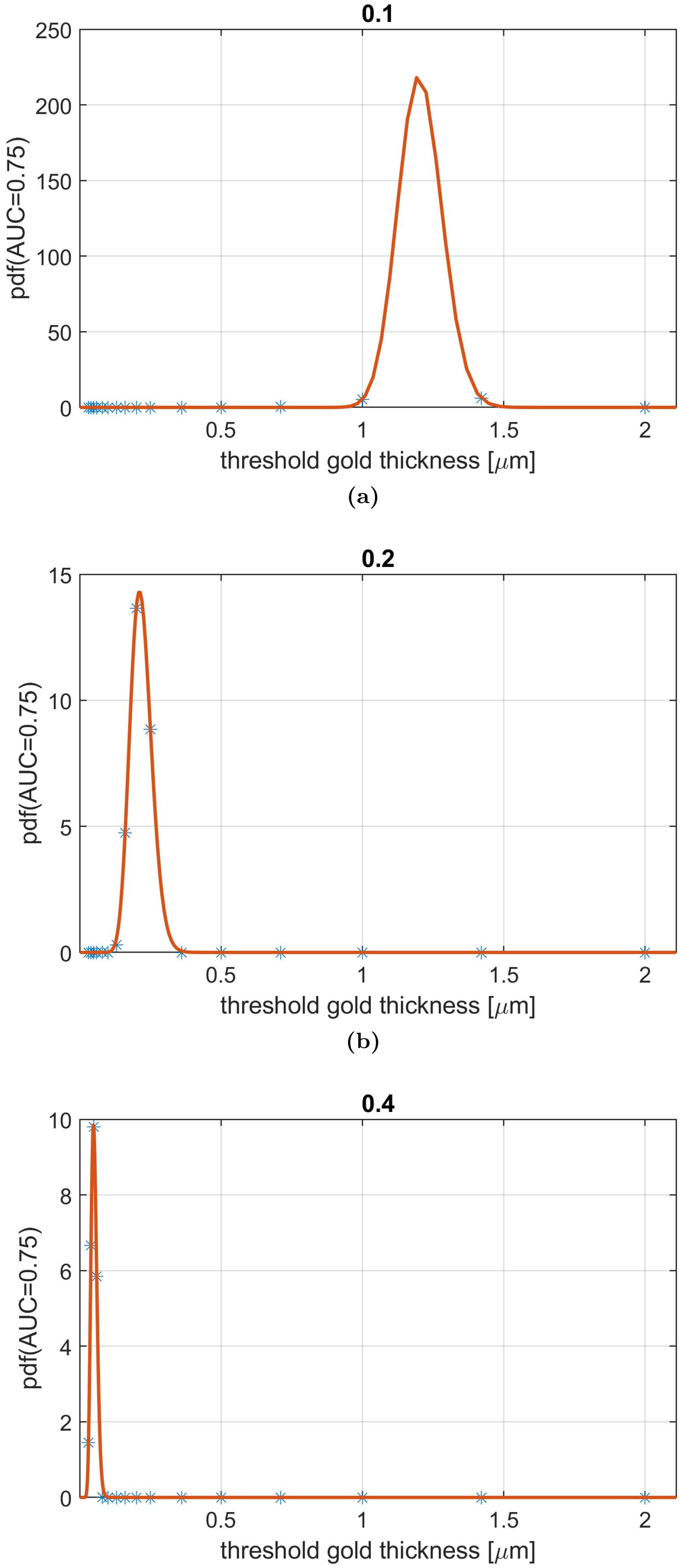
at $d = 0.25\ mm$. Different values of γ do not affect the shape of the resulting contrast-detail curves. A value of $\gamma = 0.625$ was chosen such that the magnitude of the obtained contrast-detail curves matches that of the EUREF guideline procedure.

The free parameter γ can be used to calibrate the results of our proposed approach. For our setting, its value is selected to minimize the difference between the contrast-detail curves obtained by the EUREF guideline procedure and the outcome of our method. Furthermore, it is encouraging that the proposed approach and the current practice both yield curves of similar shape. In equation (3.2) it is shown that with the definition of $f(d)$ the AUC depends on the diameter d but is independent of the total number of pixels in the lesion region. In this way, the same value of γ can be taken without changing the AUC when taking another pixel pitch into account. The diameter of the gold discs is constant, while the number of pixels of the signature evoked by this disc depends on the resolution of the detector. However, different types of detectors typically vary in their noise characteristics. Especially if the noise correlation changes significantly, the choice of γ needs to be adapted.

A.2 Bayesian Estimation

In chapter 3 a method was introduced that utilizes a recently developed parametric model observer to determine contrast-detail curves for mammography image quality assessment. In addition, a Bayesian procedure allows one to determine the posterior function $p(t \mid AUC = \tau, d)$ by reporting the recording $p(AUC = \tau \mid t, d)$ for every cell of the CDMAM phantom with thickness t and diameter d. The parameter τ denotes the critical value of AUC that is used to mark the limit for successful detection. Since there are only few discrete thicknesses for every diameter, the posterior of the contrast-detail curve is represented very sparsely. Therefore, a Gamma distribution (3.3) was fitted to the reported values as an approximate posterior distribution function for one point of the contrast-detail curve.

Figure A.2 shows three typical examples of the approximated posterior distribution for three different diameters in the contrast-detail curve determination. For diameters $d = 0.2\ mm$ (Figure A.2b) and $d = 0.4\ mm$ (Figure A.2c), the Gamma distribution can be fitted accurately because the peak is covered by the available thickness values. For a diameter of $d = 0.1\ mm$ (Figure A.2a), on the other hand, the fit is very uncertain, since there are no values that define the peak shape.



(a)

(b)

(c)

Figure A.2: Fitting a Gamma distribution as an approximate posterior distribution for three points of the contrast-detail curve. For each diameter (a)-(c), a Gamma function is fitted to the points.

A.3 Uncertainty for Deep Learning

A toy data set was created to test different methods to assess an uncertainty of the prediction obtained by a neural network. The toy data consists of images, each labeled with a corresponding curve. The 250 × 250 *Pixel* images have zero background. Nine square lesions of varying size with intensity one are embedded in this background. In a next step, Gaussian white noise with standard deviation σ is added to the image. The corresponding labels summarize local information of each lesion. Figure A.3 shows a typical input image of the toydata set.

The individual points of the label curve are calculated as

$$y_i = \frac{1}{\sigma_i \cdot \sqrt{n_{pixel}}} ,$$

where σ_i is the strength of the noise in the image and n_{pixel} is the number of pixels of the corresponding square lesion. This way, the label curve comprises nine points that vary because of the number of pixels of the corresponding lesions.

Eight different values of σ_i, with 200 images each, were combined to a training set. For testing, two independent values of σ_i were used to simulate 100 images for each noise strength. Table A.1 lists the parameters that were used in the different toy data sets.

The training data was used to train the different ensemble or dropout neural networks, respectively. After successful training, the neural networks were applied on independent test data. The suggested benchmark metric (see chapter 5) was calculated for each trained net for both of the test data sets individually. The numbers presented in Figure 5.6 show the mean benchmark score for both test sets for the different methods. As explained there, Bernoulli

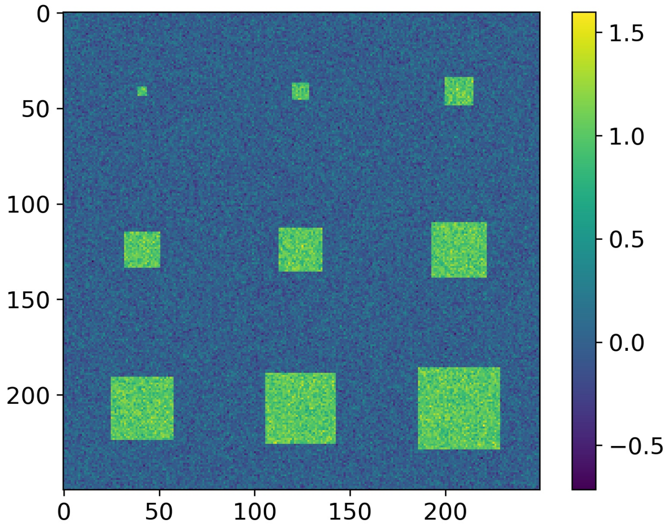


Figure A.3: Typical input image of the toydata set. Nine square lesions with intensity one are embedded in zero background. White noise was then added to the image.

Table A.1: Parameters used to create the eight toy data sets for training and two independent sets for testing of the neural network.

Set	Number	σ	N_{images}
Train	1	0.010	200
	2	0.021	200
	3	0.045	200
	4	0.097	200
	5	0.206	200
	6	0.439	200
	7	0.936	200
	8	1.995	200
Test	1	0.018	100
	2	0.150	100

dropout ($p = 0.4$), concrete dropout and ensemble learning achieved accurate approximations in terms of the mean percentage error (MPE) with reasonable estimations of the uncertainty of the prediction.

References

Abu-Mostafa, Y. S., Magdon-Ismail, M., and Lin, H.-T. (2012). *Learning from data*, volume 4. AMLBook New York, NY, USA:.

Althuis, M. D., Dozier, J. M., Anderson, W. F., Devesa, S. S., and Brinton, L. A. (2005). Global trends in breast cancer incidence and mortality 1973–1997. *International journal of epidemiology*, 34(2):405–412.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182.

Anderson, B. O., Braun, S., Lim, S., Smith, R. A., Taplin, S., and Thomas, D. B. (2003). Early detection of breast cancer in countries with limited resources. *The breast journal*, 9:S51–S59.

Anton, M., Khanin, A., Kretz, T., Reginatto, M., and Elster, C. (2018). A simple parametric model observer for quality assurance in computer tomography. *Phys. Med. Biol.*, 63(7):75011 (16pp).

Archer, B. R. and Wagner, L. K. (1988). Determination of diagnostic x-ray spectra with characteristic radiation using attenuation analysis. *Medical physics*, 15(4):637–641.

Arras, L., Horn, F., Montavon, G., Müller, K.-R., and Samek, W. (2017). " what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142.

Ashok, D. and Thomas, F. (2003). *Introduction to nuclear and particle physics.* World Scientific.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and MÃžller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831.

Bakic, P. R., Albert, M., Brzakovic, D., and Maidment, A. D. (2002). Mammogram synthesis using a 3d simulation. i. breast tissue model and image acquisition simulation. *Medical physics*, 29(9):2131–2139.

Barrett, H. H., Abbey, C. K., and Clarkson, E. (1998). Objective assessment of image quality. iii. roc metrics, ideal observers, and likelihood-generating functions. *JOSA A*, 15(6):1520–1535.

Barrett, H. H. and Myers, K. J. (2013). *Foundations of image science.* John Wiley & Sons.

Barrett, H. H., Myers, K. J., Devaney, N., and Dainty, C. (2006). Objective assessment of image quality. iv. application to adaptive optics. *JOSA A*, 23(12):3080–3105.

Barrett, H. H., Myers, K. J., Hoeschen, C., Kupinski, M. A., and Little, M. P. (2015). Task-based measures of image quality and their relation to radiation dose and patient risk. *Physics in Medicine & Biology*, 60(2):R1.

Barrett, H. H., Yao, J., Rolland, J. P., and Myers, K. J. (1993). Model observers for assessment of image quality. *Proceedings of the National Academy of Sciences*, 90(21):9758–9765.

Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, 53:370–418.

Becker, H., Nettleton, W., Meyers, P., Sweeney, J., and Nice, C. (1964). Digital computer determination of a medical diagnostic index directly from chest x-ray images. *IEEE Transactions on Biomedical Engineering*, BME-11(3):67–72.

Bendat, J. S. and Piersol, A. G. (2011). *Random data: analysis and measurement procedures*, volume 729. John Wiley & Sons.

Berg, A., Deng, J., and Fei-Fei, L. (2010). Large scale visual recognition challenge 2010.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berger, J. O., Wolpert, R. L., Bayarri, M., DeGroot, M., Hill, B. M., Lane, D. A., and LeCam, L. (1988). The likelihood principle. *Lecture Notes-Monograph Series*, 6:iii–199.

Bhattacharya, S., Sukthankar, R., and Shah, M. (2010). A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 271–280.

Bianco, S., Celona, L., Napoletano, P., and Schettini, R. (2018). On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing*, 12(2):355–362.

Bibault, J.-E., Giraud, P., Housset, M., Durdux, C., Taieb, J., Berger, A., Coriat, R., Chaussade, S., Dousset, B., Nordlinger, B., et al. (2018). Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Scientific reports*, 8(1):1–8.

Biller-Andorno, N. and Jüni, P. (2014). Abolishing mammography screening programs? a view from the swiss medical board. *Obstetrical & Gynecological Survey*, 69(8):474–475.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pages 63–71. Springer.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., and Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*.

Boone, J. M., Fewell, T. R., and Jennings, R. J. (1997). Molybdenum, rhodium, and tungsten anode spectral models using interpolating polynomials with application to mammography. *Medical physics*, 24(12):1863–1874.

Boone, J. M., Lindfors, K. K., Cooper III, V. N., and Seibert, J. A. (2000). Scatter/primary in mammography: comprehensive results. *Medical physics*, 27(10):2408–2416.

Boone, J. M., Nelson, T. R., Lindfors, K. K., and Seibert, J. A. (2001). Dedicated breast ct: radiation dose and image quality evaluation. *Radiology*, 221(3):657–667.

Borowski, M., Wrede, S., Neuschaefer-Rube, U., Büermann, L., Danzebrink, H.-U., Krumrey, M., Goebbels, J., Noetel, J., Onel, Y., Feldmann, J., et al. (2012). Development of procedures for nondestructive quality control of phantoms that are used in quality assurance tests according to §16, paragraphs 2 and 3 of the german x-ray ordinance at x-ray systems used for examination of humans vorhaben 3608s20001.

Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and Samek, W. (2017). Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer.

Bowerman, B. L., O'Connell, R. T., and Koehler, A. B. (2005). Forecasting, time series, and regression: an applied approach.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Burgess, A. (2004). On the noise variance of a digital mammography system. *Medical physics*, 31(7):1987–1995.

Cauchy, A. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538.

Chan, H.-P., Doi, K., Galhotra, S., Vyborny, C. J., MacMahon, H., and Jokich, P. M. (1987). Image feature analysis and computer-aided diagnosis in digital radiography. i. automated detection of microcalcifications in mammography. *Medical physics*, 14(4):538–548.

Chen, C.-T. (1998). *Linear system theory and design*. Oxford University Press, Inc.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Colin, C., Devic, C., Noël, A., Rabilloud, M., Zabot, M.-T., Pinet-Isaac, S., Giraud, S., Riche, B., Valette, P.-J., Rodriguez-Lafrasse, C., et al. (2011). Dna double-strand breaks induced by mammographic screening procedures in human mammary epithelial cells. *International journal of radiation biology*, 87(11):1103–1112.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Curry, T. S., Dowdey, J. E., and Murry, R. C. (1990). *Christensen's physics of diagnostic radiology*. Lippincott Williams & Wilkins.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

D'Agostini, G. (2003). Bayesian inference in processing experimental data: principles and basic applications. *Reports on Progress in Physics*, 66(9):1383.

Dance, D. and Day, G. (1984). The computation of scatter in mammography by monte carlo methods. *Physics in Medicine & Biology*, 29(3):237.

De Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48.

Del Guerra, A. (2004). *Ionizing radiation detectors for medical imaging*. World scientific.

Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., and Pal, C. (2016). The importance of skip connections in biomedical image segmentation. In *Deep Learning and Data Labeling for Medical Applications*, pages 179–187. Springer.

Ducote, J. and Molloi, S. (2010). Scatter correction in digital mammography based on image deconvolution. *Physics in Medicine & Biology*, 55(5):1295.

Easley, G. R., Labate, D., and Colonna, F. (2008). Shearlet-based total variation diffusion for denoising. *IEEE Transactions on Image processing*, 18(2):260–268.

Elangovan, P., Warren, L. M., Mackenzie, A., Rashidnasab, A., Diaz, O., Dance, D. R., Young, K. C., Bosmans, H., Strudley, C. J., and Wells, K. (2014). Development and validation of a modelling framework for simulating 2d-mammography and breast tomosynthesis images. *Physics in Medicine & Biology*, 59(15):4275.

Engelbrecht, A. and Brits, R. (2001). A clustering approach to incremental learning for feedforward neural networks. In *IJCNN 01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222)*, volume 3, pages 2019–2024. IEEE.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118.

Fabiszewska, E., Grabska, I., and Pasicz, K. (2016). The threshold contrast thickness evaluated with different CDMAM phantoms and software. *Nukleonika*, 61(1):53–59.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Freud, N., Duvauchelle, P., Pistrui-Maximean, S., Létang, J.-M., and Babot, D. (2004). Deterministic simulation of first-order scattering in virtual x-ray imaging. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 222(1-2):285–300.

Freud, N., Letang, J.-M., and Babot, D. (2005). A hybrid approach to simulate x-ray imaging techniques, combining monte carlo and deterministic algorithms. *IEEE transactions on nuclear science*, 52(5):1329–1334.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Fushiki, T., Komaki, F., Aihara, K., et al. (2005). Nonparametric bootstrap prediction. *Bernoulli*, 11(2):293–307.

Gal, Y. (2016). Uncertainty in deep learning. *University of Cambridge*, 1:3.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

Gal, Y., Hron, J., and Kendall, A. (2017). Concrete dropout. In *Advances in neural information processing systems*, pages 3581–3590.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Giger, M. L., Doi, K., and MacMahon, H. (1988). Image feature analysis and computer-aided diagnosis in digital radiography. 3. automated detection of nodules in peripheral lung fields. *Medical Physics*, 15(2):158–166.

Gilbert, F. J., Tucker, L., and Young, K. C. (2016). Digital breast tomosynthesis (dbt): a review of the evidence for use as a screening tool. *Clinical radiology*, 71(2):141–150.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Glaisher, J. (1871). Liv. on a class of definite integrals.—part ii. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42(282):421–436.

Gong, X., Glick, S. J., Liu, B., Vedula, A. A., and Thacker, S. (2006). A computer simulation study comparing lesion detection accuracy with digital mammography, breast tomosynthesis, and cone-beam CT breast imaging. *Medical physics*, 33(4):1041–1052.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gøtzsche, P. C. and Jørgensen, K. J. (2013). Screening for breast cancer with mammography. *Cochrane database of systematic reviews*, (6).

Gøtzsche, P. C. and Olsen, O. (2000). Is screening for breast cancer with mammography justifiable? *The Lancet*, 355(9198):129–134.

Gudivada, V., Apon, A., and Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1):1–20.

Guynn, J. (2015). Google photos labeled black people'gorillas'. *USA Today*, 1.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

Harris, I. R. (1989). Predictive fit for natural exponential families. *Biometrika*, 76(4):675–684.

Hassoun, M. H. et al. (1995). *Fundamentals of artificial neural networks*. MIT press.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Haus, A. G. and Yaffe, M. J. (2000). Screen-film and digital mammography: image quality and radiation dose considerations. *Radiologic Clinics of North America*, 38(4):871–898.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hernández-García, A. and König, P. (2018). Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hiom, S. (2015). Diagnosing cancer earlier: reviewing the evidence for improving cancer survival.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Hou, W., Gao, X., Tao, D., and Li, X. (2014). Blind image quality assessment via deep learning. *IEEE transactions on neural networks and learning systems*, 26(6):1275–1286.

Hu, P., Wu, F., Peng, J., Liang, P., and Kong, D. (2016). Automatic 3d liver segmentation based on deep learning and globally optimized surface evolution. *Physics in Medicine & Biology*, 61(24):8676.

Huda, W., Sajewicz, A. M., Ogden, K. M., and Dance, D. R. (2003). Experimental investigation of the dose and image quality characteristics of a digital mammography imaging system. *Medical physics*, 30(3):442–448.

Hüllermeier, E. and Waegeman, W. (2019). Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *arXiv preprint arXiv:1910.09457*.

ICRU (2009). ICRU Report No. 82: mammography-assessment of image quality. *Journal of ICRU*, 6.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*.

Jefferys, J. and Cooper, A. (2007). Brain basics. *The Human Brain and Its Disorders*.

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Jørgensen, K. J. and Gøtzsche, P. C. (2009). Overdiagnosis in publicly organised mammography screening programmes: systematic review of incidence trends. *Bmj*, 339:b2587.

Karssemeijer, N. and Thijssen, M. A. O. (1996). Determination of contrast-detail curves of mammography systems by automated image analysis. *Digit. Mammogr.*, 96:155–160.

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584.

Khanin, A., Anton, M., Reginatto, M., and Elster, C. (2018). Assessment of CT image quality using a bayesian framework. *IEEE Transactions on Medical Imaging*, 37(12):2687–2694.

Kim, J. and Lee, S. (2017). Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684.

Kim, S. and Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Kingma, D. P., Salimans, T., and Welling, M. (2015). Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, pages 2575–2583.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Kline, T. L., Korfiatis, P., Edwards, M. E., Blais, J. D., Czerwiec, F. S., Harris, P. C., King, B. F., Torres, V. E., and Erickson, B. J. (2017). Performance of an artificial multi-observer deep neural network for fully automated segmentation of polycystic kidneys. *Journal of digital imaging*, 30(4):442–448.

Koch, H. and Motz, J. (1959). Bremsstrahlung cross-section formulas and related data. *Reviews of modern physics*, 31(4):920.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., den Heeten, A., and Karssemeijer, N. (2017). Large scale deep learning for computer aided detection of mammographic lesions. *Medical Image Analysis*, 35:303–312.

Kösters, J. P. and Gøtzsche, P. C. (2003). Regular self-examination or clinical examination for early detection of breast cancer. *Cochrane Database of Systematic Reviews*, (2):CD003373.

Kretz, T., Anton, M., Schaeffter, T., and Elster, C. (2019). Determination of contrast-detail curves in mammography image quality assessment by a parametric model observer. *Physica Medica*, 62:120–128.

Kretz, T., Müller, K.-R., Schaeffter, T., and Elster, C. (2020). Mammography image quality assurance using deep learning. *IEEE Transactions on Biomedical Engineering*, –:10.1109/TBME.2020.2983539.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Lazos, D., Kolisti, Z., and Pallikarakis, N. (2000). A software data generator for radiographic imaging investigations. *IEEE Transactions on Information Technology in Biomedicine*, 4(1):76–79.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 8595–8598. IEEE.

LeCun, Y. (2019). Deep learning hardware: Past, present, and future. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, pages 12–19.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

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

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Lee, J. H., Grant, B. R., Chung, J. H., Reiser, I., and Giger, M. (2018). Assessment of diagnostic image quality of computed tomography (ct) images of the lung using deep learning. In *Medical Imaging 2018: Physics of Medical Imaging*, volume 10573, page 105731M. International Society for Optics and Photonics.

Li, Y., Poulos, A., McLean, D., and Rickard, M. (2010). A review of methods of clinical image quality evaluation in mammography. *European journal of radiology*, 74(3):e122–e131.

Lin, G., Milan, A., Shen, C., and Reid, I. (2017). Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934.

Lison, P. (2015). An introduction to machine learning. *Language Technology Group: Edinburgh, UK.*

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

Liu, Y.-H., Yang, K.-F., and Yan, H.-M. (2019). No-reference image quality assessment method based on visual parameters. *Journal of Electronic Science and Technology*, 17(2):171–184.

Loftus, P. and Giudice, S. (2014). Relevance of methods and standards for the assessment of measurement system performance in a high-value manufacturing industry. *Metrologia*, 51(4):S219.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Loo, L.-N. D., Doi, K., Ishida, M., Metz, C. E., Chan, H.-P., Higashida, Y., and Kodera, Y. (1983). An empirical investigation of variability in contrast-detail diagram measurements. In *Application of Optical Instrumentation in Medicine XI*, volume 419, pages 68–76. International Society for Optics and Photonics.

Love, L. A. and Kruger, R. A. (1987). Scatter estimation for a digital radiographic system using convolution filtering. *Medical physics*, 14(2):178–185.

MacKay, D. J. (1992). A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

Mackenzie, A., Eales, T. D., Dunn, H. L., Braidley, M. Y., Dance, D. R., and Young, K. C. (2017). Simulation of images of CDMAM phantom and the estimation of measurement uncertainties of threshold gold thickness. *Physica Medica*, 39:137–146.

Mackenzie, A., Warren, L. M., Wallis, M. G., Given-Wilson, R. M., Cooke, J., Dance, D. R., Chakraborty, D. P., Halling-Brown, M. D., Looney, P. T., and Young, K. C. (2016). The relationship between cancer detection in mammography and image quality measurements. *Phys. Medica*, 32(4):568–574.

Malliori, A., Bliznakova, K., Dermitzakis, A., and Pallikarakis, N. (2013). Evaluation of the effect of acquisition parameters on image quality in digital breast tomosynthesis: Simulation studies. In *World Congress on Medical Physics and Biomedical Engineering May 26-31, 2012, Beijing, China*, pages 2211–2214. Springer.

Mandelblatt, J. S., Cronin, K. A., Bailey, S., Berry, D. A., De Koning, H. J., Draisma, G., Huang, H., Lee, S. J., Munsell, M., Plevritis, S. K., et al. (2009). Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms. *Annals of internal medicine*, 151(10):738–747.

Marshall, N. (2006). A comparison between objective and subjective image quality measurements for a full field digital mammography system. *Physics in Medicine & Biology*, 51(10):2441.

MATLAB (2019). *Deep Learning Toolbox version 12.1 (R2019a).* The MathWorks Inc., Natick, Massachusetts.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Metz, C. E. (1979). Applications of roc analysis in diagnostic image evaluation. Technical report, Chicago Univ.

Meyers, P. H., Nice Jr, C. M., Becker, H. C., Nettleton Jr, W. J., Sweeney, J. W., and Meckstroth, G. R. (1964). Automated computer analysis of radiographic images. *Radiology*, 83(6):1029–1034.

Miller, A. B., Wall, C., Baines, C. J., Sun, P., To, T., and Narod, S. A. (2014). Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *Bmj*, 348:g366.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Mitchell, T. M. (1997). *Machine Learning*, volume 1. McGraw-Hill Science/Engineering/Math.

Molnar, C. (2019). *Interpretable Machine Learning.* Lulu. com. `https://christophm.github.io/interpretable-ml-book/`.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

Müller, K.-R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2):181–201.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Myers, K. J. and Barrett, H. H. (1987). Addition of a channel mechanism to the ideal-observer model. *JOSA A*, 4(12):2447–2457.

Nadine, A., Easton, D., Chang-Claude, J., Rookus, M., Brohet, R., Cardis, E., Antoniou, A., Wagner, T., Simard, J., Evans, G., et al. (2006). Effect of chest x-rays on the risk of breast cancer among brca1/2 mutation carriers in the international brca1/2 carrier cohort study: a report from the embrace, genepso, geo-hebon, and ibccs collaborators' group. *J Clin Oncol*, 24(21):3361–6.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Neal, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems*, pages 475–482.

NHTSA (2017). Tesla crash preliminary evaluation report. Technical report, U.S. Department of Transportation National Highway Traffic Safety Administration.

Niklason, L. T., Christian, B. T., Niklason, L. E., Kopans, D. B., Castleberry, D. E., Opsahl-Ong, B., Landberg, C. E., Slanetz, P. J., Giardino, A. A., Moore, R., et al. (1997). Digital tomosynthesis in breast imaging. *Radiology*, 205(2):399–406.

Noel, A. and Thibault, F. (2004). Digital detectors for mammography: the technical challenges. *European radiology*, 14(11):1990–1998.

Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

Nyström, L., Andersson, I., Bjurstam, N., Frisell, J., Nordenskjöld, B., and Rutqvist, L. E. (2002). Long-term effects of mammography screening: updated overview of the swedish randomised trials. *The Lancet*, 359(9310):909–919.

Obrig, H., Wenzel, R., Kohl, M., Horst, S., Wobst, P., Steinbrink, J., Thomas, F., and Villringer, A. (2000). Near-infrared spectroscopy: does it function in functional activation studies of the adult brain? *International Journal of Psychophysiology*, 35(2-3):125–142.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology*, 229(1):3–8.

Oeffinger, K. C., Fontham, E. T., Etzioni, R., Herzig, A., Michaelson, J. S., Shih, Y.-C. T., Walter, L. C., Church, T. R., Flowers, C. R., LaMonte, S. J., et al. (2015). Breast cancer screening for women at average risk: 2015 guideline update from the american cancer society. *Jama*, 314(15):1599–1614.

Orr, G. B. and Müller, K.-R. (1998). *Neural networks: Tricks of the Trade*. Springer.

Paci, E. and Duffy, S. (2005). Overdiagnosis and overtreatment of breast cancer: overdiagnosis and overtreatment in service screening. *Breast Cancer Research*, 7(6):266.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Pendrill, L. R. (2014). Using measurement uncertainty in decision-making and conformity assessment. *Metrologia*, 51(4):S206.

Pepe, M. S. et al. (2003). *The statistical evaluation of medical tests for classification and prediction.* Medicine.

Perry, N., Broeders, M., de Wolf, C., Törnberg, S., Holland, R., von Karsa, L., and Puthaar, E. (2006). *European guidelines for quality assurance in breast cancer screening and diagnosis.* Office for Official Publications of the European Communities, Luxembourg, 4th edition.

Perry, N. and Puthaar, E. (2006). *European guidelines for quality assurance in breast cancer screening and diagnosis.* European Communities.

Pisano, E. D. and Yaffe, M. J. (2005). Digital mammography. *Radiology*, 234(2):353–362.

Plevritis, S. K., Kurian, A. W., Sigal, B. M., Daniel, B. L., Ikeda, D. M., Stockdale, F. E., and Garber, A. M. (2006). Cost-effectiveness of screening brca1/2 mutation carriers with breast magnetic resonance imaging. *Jama*, 295(20):2374–2384.

Pogue, B. W., Poplack, S. P., McBride, T. O., Wells, W. A., Osterman, K. S., Osterberg, U. L., and Paulsen, K. D. (2001). Quantitative hemoglobin tomography with diffuse near-infrared spectroscopy: pilot results in the breast. *Radiology*, 218(1):261–266.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151.

Rasmussen, C. E. and Quinonero-Candela, J. (2005). Healing the relevance vector machine through augmentation. In *Proceedings of the 22nd international conference on Machine learning*, pages 689–696.

Ravaglia, V., Bouwman, R., Young, K., Van Engen, R., and Lazzari, B. (2009). Noise analysis of full field digital mammography systems. In *Medical Imaging 2009: Physics of Medical Imaging*, volume 7258, page 72581B. International Society for Optics and Photonics.

Reed, S. J. B. (2005). *Electron microprobe analysis and scanning electron microscopy in geology.* Cambridge university press.

Richards, M. (2009). The national awareness and early diagnosis initiative in england: assembling the evidence. *British journal of cancer*, 101(S2):S1.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747.*

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Russo, P. (2017). *Handbook of X-ray Imaging: Physics and Technology.* CRC Press.

Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M. A., Adler, D. D., and Goodsitt, M. M. (1996). Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. *IEEE transactions on Medical Imaging*, 15(5):598–610.

Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., Summers, R. M., and Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. *Medical physics*, 46(1):e1–e36.

Saltelli, A., Tarantola, S., Campolongo, F., et al. (2000). Sensitivity anaysis as an ingredient of modeling. *Statistical Science*, 15(4):377–395.

Samek, W. (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K.-R. (2020). Toward interpretable machine learning: Transparent deep neural networks and beyond. *arXiv preprint arXiv:2003.07631*.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., and Rueckert, D. (2017). A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Schölkopf, B., Smola, A. J., Bach, F., et al. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., and Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8(1):13890.

Seibert, J. A. (2004). X-ray imaging physics for nuclear medicine technologists. part 1: Basic principles of x-ray production. *Journal of nuclear medicine technology*, 32(3):139–147.

Seibert, J. A. and Boone, J. M. (2005). X-ray imaging physics for nuclear medicine technologists. part 2: X-ray interactions and image formation. *Journal of nuclear medicine technology*, 33(1):3–18.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60.

Siddon, R. L. (1985). Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical physics*, 12(2):252–255.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Spahn, M. (2005). Flat detectors and their clinical applications. *European radiology*, 15(9):1934–1947.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. (2016). Bayesian optimization with robust bayesian neural networks. In *Advances in neural information processing systems*, pages 4134–4142.

Spyrou, G., Tzanakos, G., Nikiforides, G., and Panayiotakis, G. (2002). A monte carlo simulation model of mammographic imaging with x-ray sources of finite dimensions. *Physics in Medicine & Biology*, 47(6):917.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89.

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R. (2016). Interpretable deep neural networks for single-trial eeg classification. *Journal of neuroscience methods*, 274:141–145.

Summers, R. M. (2019). Are we at a crossroads or a plateau? radiomics and machine learning in abdominal oncology imaging. *Abdominal Radiology*, 44(6):1985–1989.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1891–1898.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Suzuki, K., Armato III, S. G., Li, F., Sone, S., and Doi, K. (2003). Massive training artificial neural network (mtann) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography. *Medical physics*, 30(7):1602–1617.

Taylor, B. J. (2006). *Methods and procedures for the verification and validation of artificial neural networks*. Springer Science & Business Media.

Thijssen, M., Bijkerk, K., and Van der Burght, R. (2007). Manual contrast-detail phantom artinis CDMAM type 3.4. *University Medical Center Nijmegen, Department of Radiology, The Netherlands, Utilisation Manual2006*.

Tipler, P. A. and Mosca, G. (2007). *Physics for scientists and engineers*. Macmillan.

Tseng, H.-H., Luo, Y., Cui, S., Chien, J.-T., Ten Haken, R. K., and El Naqa, I. (2017). Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical physics*, 44(12):6690–6705.

Veronesi, U., Paganelli, G., Galimberti, V., Viale, G., Zurrida, S., Bedoni, M., Costa, A., De Cicco, C., Geraghty, J. G., Luini, A., et al. (1997). Sentinel-node biopsy to avoid axillary dissection in breast cancer with clinically negative lymph-nodes. *The Lancet*, 349(9069):1864–1867.

Walker, W. E., Harremoës, P., Rotmans, J., Van Der Sluijs, J. P., Van Asselt, M. B., Janssen, P., and Krayer von Krauss, M. P. (2003). Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integrated assessment*, 4(1):5–17.

Wang, B. and Klabjan, D. (2017). Regularization for unsupervised deep neural nets. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Wang, E. H. and Kuh, A. (1992). A smart algorithm for incremental learning. In *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, volume 3, pages 121–126 vol.3.

Wang, S. and Manning, C. (2013). Fast dropout training. In *international conference on machine learning*, pages 118–126.

Wang, Z., Bovik, A. C., and Lu, L. (2002). Why is image quality assessment so difficult? In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3313. IEEE.

Warren, L. M., Given-Wilson, R. M., Wallis, M. G., Cooke, J., Halling-Brown, M. D., Mackenzie, A., Chakraborty, D. P., Bosmans, H., Dance, D. R., and Young, K. C. (2014). The effect of image processing on the detection of cancers in digital mammography. *Am. J. Roentgenol.*, 203(2):387–393.

Welch, H. G. and Black, W. C. (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613.

Whitman, G. J. and Haygood, T. M. (2012). *Digital mammography: a practical approach*. Cambridge University Press.

WHO (2014). *WHO position paper on mammography screening*. World Health Organization.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Wu, A., Nowozin, S., Meeds, E., Turner, R. E., Hernández-Lobato, J. M., and Gaunt, A. L. (2018). Deterministic variational inference for robust bayesian neural networks. *arXiv preprint arXiv:1810.03958*.

Wu, L., Cheng, J.-Z., Li, S., Lei, B., Wang, T., and Ni, D. (2017). FUIQA: fetal ultrasound image quality assessment with deep convolutional networks. *IEEE transactions on cybernetics*, 47(5):1336–1349.

Yaffe, M. and Rowlands, J. (1997). X-ray detectors for digital radiography. *Physics in Medicine & Biology*, 42(1):1.

Yaffe, M. J. and Mainprize, J. G. (2011). Risk of radiation-induced breast cancer from mammographic screening. *Radiology*, 258(1):98–105.

Yassin, N. I., Omran, S., El Houby, E. M., and Allam, H. (2018). Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer methods and programs in biomedicine*, 156:25–45.

Yip, M., Chukwu, W., Kottis, E., Lewis, E., Oduko, J., Gundogdu, O., Young, K., and Wells, K. (2009). Automated scoring method for the CDMAM phantom. In *Medical Imaging 2009: Image Perception, Observer Performance, and Technology Assessment*, volume 7263, page 72631A. International Society for Optics and Photonics.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, B.-T. (1994). An incremental learning algorithm that optimizes network size and sample size in one trial. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 215–220. IEEE.

Zheng, A. and Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. " O'Reilly Media, Inc.".