Applications of Deep Learning in Medical Image Analysis

Grading of Prostate Cancer and Detection of Coronary Artery Disease

Arvidsson, Ida

2021

Applications of Deep Learning in Medical Image Analysis

Grading of Prostate Cancer and Detection of Coronary Artery Disease

by Ida Arvidsson

LUND
UNIVERSITY

Cover illustration front: Gleason grading performed by AI and two pathologists.
(Credits: Felicia Marginean, Athanasios Simoulis and Ida Arvidsson).
Cover illustration back: Polar map from myocardial perfusion imaging.

Mathematics
Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden

www.maths.lu.se

Doctoral Theses in Mathematical Sciences 2021:4
ISSN: 1404-0034

ISBN: 978-91-7895-798-9 (print)
ISBN: 978-91-7895-797-2 (pdf)
LUTFMA-1072-2021

Printed in Sweden by Media-Tryck, Lund University, Lund 2021

Abstract

A wide range of medical examinations are using analysis of images from different types of equipment. Using artificial intelligence, the assessments could be done automatically. This can have multiple benefits for the healthcare; reduce workload for medical doctors, decrease variations in diagnoses and cut waiting times for the patient as well as improve the performance. The aim of this thesis has been to develop such solutions for two common diseases: prostate cancer and coronary artery disease. The methods used are mainly based on deep learning, where the model teaches itself by training on large datasets.

Prostate cancer is one of the most common cancer diagnoses among men. The diagnosis is most commonly determined by visual assessment of prostate biopsies in a light microscope according to the Gleason scale. Deep learning methods to automatically detect and grade the cancer areas are presented in this thesis. The methods have been adapted to improve the generalisation performance on images from different hospitals, images which have inevitable variations in e.g. stain appearance. The methods include the usage of digital stain normalisation, training with extensive augmentation or using models such as a domain-adversarial neural network. One Gleason grading algorithm was evaluated on a small cohort with biopsies annotated in detail by two pathologists, to compare the performance with pathologists' inter-observer variability. Another cancer detection algorithm was evaluated on a large active surveillance cohort, containing patients with small areas of low-grade cancer. The results are promising towards a future tool to facilitate grading of prostate cancer.

Cardiovascular disease is the leading cause of death world-wide, whereof coronary artery disease is one of the most common diseases. One way to diagnose coronary artery disease is by using myocardial perfusion imaging, where disease in the three main arteries supplying the heart with blood can be detected. Methods based on deep learning to perform the detection automatically are presented in this thesis. Furthermore, an algorithm developed to predict the degree of coronary artery stenosis from myocardial perfusion imaging, by means of quantitative coronary angiography, has also been developed. This assessment is normally done using invasive coronary angiography. Making the prediction automatically from myocardial perfusion imaging could save suffering for patients and free resources within the healthcare system.

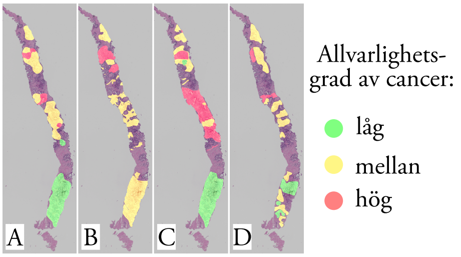
Populärvetenskaplig sammanfattning

Prostatacancer är den vanligaste cancerdiagnosen bland män i flera länder, däribland Sverige. Samtidigt är hjärt- och kärlsjukdom den vanligaste dödsorsaken i världen, varav kranskärlssjukdom är den mest frekventa. Det finns alltså mycket att vinna på att förbättra diagnostiken av dessa sjukdommar.

Under senare år har begreppet artificiell intelligens (AI) letat sig in nästan överallt i samhället, så även inom medicin. Med hjälp av AI skulle en del arbetsuppgifter inom sjukvården kunna underlättas eller helt automatiseras, vilket skulle frigöra tid för sjukvårdspersonal till de mer komplicerade fallen.

Framgången för AI beror till stor del på att datorer har blivit kraftfullare och fler beräkningar kan göras på kort tid. Neurala nätverk är en sorts AI, som bygger på att många enkla beräkningar kombineras i stora nätverk för att hitta avancerade egenskaper och samband. Det speciella med nätverken är att de inte programmeras i detalj, utan i stället får utveckla sig själva genom att träna på stora dataset med många exempel. De får alltså testa sig fram och genom återkoppling från jämförelse med redan kända fall bli bättre på sin uppgift. Den här sortens algoritmer har använts i den här avhandlingen för att göra bedömningar av medicinska bilder.

För diagnostisering av prostatacancer tas små vävnadsprov som färgas och analyseras i ett mikroskop. Där kan cancer ses i form av körtlar som har förlorat sin form och börjat växa okontrollerat. Diagnostiseringen görs med hjälp av en gradering, den så kallade Gleasonskalan, vilken ger en prognos för sjukdomen och används för att avgöra lämplig behandling. Det finns stor variation i vilken bedömning som görs, både mellan men även med samma patolog. AI kan hjälpa till med graderingen, till exempel som en extra bedömmare för att säkerställa rätt diagnos eller för att sortera ut friska fall och därmed underlätta arbetet för det minskande antalet patologer. Exempel på olika bedömningar för samma vävnadsprov visas i figuren nedan.



Variation i bedömning av prostatacancer i samma vävnadsprov. A - AI, B - läkare 1 år 1, C - läkare 1 år 2, D - läkare 2.

Kranskärlssjukdom kan diagnostiseras på flera sätt, varav hjärtscintigrafi är en av de vanligaste. Vid hjärtscintigrafi sprutas en liten mängd av ett radioaktivt ämne in i blodet varefter bilder tas av hjärtat med en gammakamera. Bilderna analyseras för att upptäcka eventuella förträngningar i kärlen som förser hjärtmuskeln med blod. Med hjälp av AI kan bilderna från den ökande användningen av hjärtscintigrafi analyseras automatiskt. Vi har dessutom undersökt om AI kan utläsa mer information än vad det mänskliga ögat kan se.

Projekten i avhandlingen visar flera möjligheter för AI och bildanalys inom medicin med lovande resultat, men diskuterar även de svårigheter som finns. Bara för att en algoritm fungerar bra på ett sjukhus, så är det inte säkert att den gör det på ett annat. Samtidigt så kanske AI kan användas för att göra mer tillförlitliga bedömningar än tidigare.

List of Publications

This thesis is based on the following publications, referred to by their Roman numerals. They are included in this thesis with the permission of the publishers. My contribution to each paper is listed below.

I A. Gummeson, **I. Arvidsson**, M. Ohlsson, N. C. Overgaard, A. Krzyzanowska, A. Heyden, A. Bjartell, K. Åström, "Automatic Gleason Grading of H&E Stained Microscopic Prostate Images using Deep Convolutional Neural Networks", In *Proceedings of the International Society for Optics and Photonics, Medical Imaging: Digital Pathology*, volume 10140, page 101400S, 2017.

This paper is based on the master's thesis by AG, who did the implementations. KÅ and AB had the original idea for the project. The paper was written by me, and revised by the other authors.

II J. Isaksson, **I. Arvidsson**, K. Åström, A. Heyden, "Semantic Segmentation of Microscopic Images of H&E Stained Prostatic Tissue using CNN", In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1252–1256, 2017. ©2017 IEEE

AH came up with the idea and JI did the implementations in his master's thesis. I wrote the paper based on the master's thesis, with input from all authors.

III **I. Arvidsson**, N. C. Overgaard, F. E. Marginean, A. Krzyzanowska, A. Bjartell, K. Åström, A. Heyden, "Generalization of Prostate Cancer Classification for Multiple Sites using Deep Learning", In *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging (ISBI)*, pages 191–194, 2018. ©2018 IEEE

All authors contributed to the idea. I developed the method, did the implementations and wrote the paper, with input from the other authors.

IV K. Tall, **I. Arvidsson**, N. C. Overgaard, K. Åström, A. Heyden, "Automatic Detection of Small Areas of Gleason Grade 5 in Prostate Tissue using CNN", In *Proceedings of the International Society for Optics and Photonics, Medical Imaging: Digital Pathology*, volume 10956, page 109560E, 2019.

AH, NCO, KÅ and I came up with the idea. KT developed the method and did the implementations in his master's thesis, supervised by me and AH. I wrote the paper.

v **I. Arvidsson**, N. C. Overgaard, K. Åström, A. Heyden, "Comparison of Different Augmentation Techniques for Improved Generalization Performance for Gleason Grading", In *Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI)*, pages 923–927, 2019. ©2019 IEEE

All authors contributed to the idea. I developed the ideas, did the implementations and wrote the paper, which was revised by all authors.

vi **I. Arvidsson**, N. C. Overgaard, A. Krzyzanowska, F. E. Marginean, A. Simoulis, A. Bjartell, K. Åström, A. Heyden, "Domain-Adversarial Neural Network for Improved Generalization Performance of Gleason Grade Classification", In *Proceedings of the International Society for Optics and Photonics, Medical Imaging: Digital Pathology*, volume 11320, page 1132016, 2020.

KÅ and I came up with the idea. I developed the idea, did the implementations and wrote the paper, with input from the other authors.

vii F. E. Marginean, **I. Arvidsson**, A. Simoulis, N. C. Overgaard, K. Åström, A. Heyden, A. Bjartell, A. Krzyzanowska, "An Artificial Intelligence-based Support Tool for Automation and Standardisation of Gleason Grading in Prostate Biopsies", In *European Urology Focus*, 2020.

The study was designed by AK, AB, FEM, KÅ, NCO, AH and me. The data was collected and annotated by FEM, AS, AK. I developed and implemented the algorithm. The results were analysed by FEM, AS, AK and me. The paper was mainly written by FEM, AK and me, and revised by all authors.

viii **I. Arvidsson**, N. C. Overgaard, A. Davidsson, J. Frias-Rose, M. Ochoa-Figueroa, K. Åström, A. Heyden, "Prediction of Obstructive Coronary Artery Disease from Myocardial Perfusion Scintigraphy using Deep Neural Networks", In *Proceedings of the IEEE 25th International Conference on Pattern Recognition*, pages 4442–4449, 2021. ©2021 IEEE

MOF and AH came up with the idea for the project. I developed the idea for the method with input from the other authors. I did the implementations. The paper was written by me and revised by all authors.

IX **I. Arvidsson**, N. C. Overgaard, A. Davidsson, J. Frias-Rose, K. Åström, M. Ochoa-Figueroa, A. Heyden, "Detection of Left Bundle Branch Block and Obstructive Coronary Artery Disease from Myocardial Perfusion Scintigraphy using Deep Neural Networks", In *Proceedings of the International Society for Optics and Photonics, Medical Imaging: Computer-Aided Diagnosis*, volume 11597, page 115970N, 2021.

This paper continued the work in Paper VIII, but with additional ideas from MOF. I designed the method, did the implementations and wrote the paper. All authors revised the paper.

X **I. Arvidsson**, A. Davidsson, N. C. Overgaard, C. Pagonis, K. Åström, E. Good, J. Frias-Rose, A. Heyden, M. Ochoa-Figueroa, "Deep Learning Prediction of Quantitative Coronary Angiography using Myocardial Perfusion Images with a Cardiac CZT Camera", *Manuscript*.

MOF came up with the idea for the study. I developed the method and did the implementations. The paper was mainly written by MOF and me, with input from the other authors.

Acknowledgements

This thesis would not have been possible without the help and support from colleagues and friends. I specially want to thank the following:

My supervisors. *Anders Heyden*, *Kalle Åström* and *Niels Christian Overgaard*, for your guidance, support and encouragement. For the many opportunities you have given to me, your interesting ideas and endless enthusiasm. It has been five stimulating and developing years working with you and your combined supervision has been very valuable to me.

All medical collaborators. For collecting data, making annotations and not at least explaining your professions. In particular: *Agnieszka Krzyzanowska*, for great collaboration, your excellent interpretations between the pathologists and me, and your persistence with the prostate project. *Felicia Marginean* and *Athanasios Simoulis*, for all the time you spent on drawing annotations and your patience with the not always so successful algorithm. *Anders Bjartell*, for your guidance and sharing of expertise. *Miguel Ochoa-Figueroa*, for your great ideas and enthusiasm. *Anette Davidsson*, for your input and persistent work.

My colleagues at the Centre for Mathematical Sciences. For good company, valuable input and insightful discussions. Especially: *Gabrielle Flood*, for sharing the last five years with me and for the invaluable teamwork, friendship and support. *Anna Gummeson*, for providing me the best possible start of my PhD and for making the second half of it more fun. *Linn Öström*, for your helpfulness and for being the best office mate possible.

All other collaborators. In particular: *Krister Tham, Linus Mårtensson* and your colleagues at Katam and Eigenvision, for challenging tasks and new perspectives. *Claes Lundström* and *Erik Sjöblom* at Sectra, for support with the prostate project. *Giuseppe Lippolis*, for sharing your images. The master's thesis students working on the prostate project: *Johan Isaksson, Joel Ekelund, Kasper Tall, Elin Olofsson*, for your devotion and accomplishments. All other co-authors: *Mattias Ohlsson, Jeronimo Frias-Rose, Christos Pagonis, Elin Good*, for valuable input on our research.

Family and friends. For support, friendship, caring and inspiration. Especially: *Mamma och Far*, for helping out with everything at home and your unconditional support. *Lovisa*, for explaining all complicated (and not so complicated) medical terms to a novice in the field. *Johannes*, for encouraging me to pursue a PhD. *Elna*, for your generous hospitality and support. *Erik*, for your patience, advice and thoughtfulness.

Svenstorp, 2021-04-15

Funding

The studies and work in this thesis were funded by Vinnova-Swelife and Vinnova-Medtech4Life programs — DOGS: Digital Pathology for Optimized Gleason Score in Prostate Cancer and DOGS-2: Digital Pathology for Optimized Gleason Score-2 — (grant no. 2015–04740 and 2018–02271), Vinnova-Medtech4Life program — AIDA: Analytic Imaging Diagnostics Arena — (grant no. 2017–02447) and strategic research projects eSSENCE and ELLIIT.

Abbreviations

The following table describes the abbreviations and acronyms used throughout the thesis. The page on which each one is first used is also given.

Abbreviation	Meaning	Page
AHA	American Heart Association	11
ANN	Artificial Neural Network	3
AI	Artificial Intelligence	1
AUC	Area Under the ROC Curve	57
BMI	Body Mass Index	26
CAD	Coronary Artery Disease	1
CNN	Convolutional Neural Network	3
CZT	Cadmium-Zinc-Telluride	12
DANN	Domain-Adversarial Neural Network	47
DL	Deep Learning	1
ECG	ElectroCardioGram	14
ESC	European Society of Cardiology	11
FN	False Negative	56
FP	False Positive	56
G3	Gleason grade 3	7
G4	Gleason grade 4	7
G5	Gleason grade 5	7
GAN	Generative Adversarial Network	46
GG	Gleason grade Group	7
GS	Gleason Score	7
H&E	Haematoxylin & Eosin	8
HDI	Human Development Index	6
HSV	Hue Saturation Value	51
ICA	Invasive Coronary Angiography	14
ICC	Intraclass Correlation Coefficient	58
IoU	Intersection-over-Union	58
ISUP	International Society of Urological Pathology	7
LAD	Left Anterior Descending artery	10
LBBB	Left Bundle Branch Block	14
LCx	Left Circumflex artery	10
ML	Machine Learning	2
MPI	Myocardial Perfusion Imaging	1
PCa	Prostate Cancer	1
PRIAS	Prostate Cancer Research International Active Surveillance	23
PSA	Prostate-Specific Antigen	6
QPS	Quantitative Perfusion SPECT	24

continued on next page

Abbreviation	Meaning	Page
RANSAC	RANdom SAmpling Consensus	55
RCA	Right Coronary Artery	10
ReLU	Rectified Linear Unit	34
RGB	Red Green Blue	29
ROC	Receiver Operating Characteristic	57
SIFT	Scale Invariant Feature Transform	54
SPECT	Single-Photon Emission Computed Tomography	12
TN	True Negative	56
TP	True Positive	56
QCA	Quantitative Coronary Angiography	14

Contents

Chapter 1

Introduction

The overall subject of this thesis is automatic assessments of medical images. This by using image analysis and in particular *deep learning* (DL). Setting medical diagnoses with visual assessment is not a novel practice. However, the technology used for medical imaging has evolved and today includes the creation of digital images from different equipment and in huge quantities. This enables the use of *artificial intelligence* (AI) and image analysis for automatic assessments, which can have multiple benefits for healthcare; reduce workload for medical doctors, decrease variations in diagnoses and cut waiting times for the patient as well as improve the sensitivity and the specificity of diagnoses.

The applications studied in this thesis are limited to two fields; grading of *prostate cancer* (PCa) from microscopy images of stained tissue from prostate biopsies and detection of *coronary artery disease* (CAD) using images from *myocardial perfusion imaging* (MPI). The methods, possibilities and limitations are however the same for many other applications, in particular within medicine, and the work presented can hopefully be used as a guide for new applications. An overview of the components of the thesis follows.

1.1 Thesis Overview

The thesis consists of four parts. The first part (Chapters 1 – 3) introduces the subjects, datasets and methods used. In the second part (Chapter 4 – 6) methods for PCa grading are presented and evaluated. The third part (Chapter 7) concerns detection of CAD. Finally, the fourth part (Chapter 8) discusses future applications and presents the conclusions.

Chapter 1 The key concepts of artificial intelligence are introduced. Furthermore, some background information on PCa and CAD is presented.

Chapter 2 The different datasets used are described. These are a key component for the studies and results presented in this thesis.

Chapter 3 The methods used are introduced. These are mainly deep learning methods, but also related aspects such as colour normalisation of microscopy images of stained tissue and registration of images.

Chapter 4 The methods developed for Gleason grading of PCa are described and corresponding results are presented. The focus is to find learning-based methods that generalise well to unseen cases, although the datasets used for training are limited in size and variation.

Chapter 5 Grading of PCa is not a trivial task. In this chapter, two approaches that could facilitate the grading are introduced: semantic segmentation of relevant tissue components and detection of small cancer areas.

Chapter 6 Thorough evaluations of automatic PCa detection and Gleason scoring algorithms are presented. Evaluations are done both by comparing to multiple pathologists and by evaluating on a larger dataset.

Chapter 7 Algorithms for automatic detection of CAD using MPI are described. Furthermore, an algorithm for automatic prediction of the degree of coronary artery stenosis is presented.

Chapter 8 General discussions of the topics in, and related to, this thesis. The conclusions which can be drawn from the presented results and suggestions of topics to be further explored are stated.

1.2 Deep Learning

The term AI seems to be more popular than ever. AI is a very broad term with vague definition. Essentially anything that a machine can do which can be considered "smart" is often included in AI. A sub-field of AI is *machine learning* (ML). The goal of ML is that the machine should teach itself by experience, by training on a dataset or receiving some other sort of feedback. The machine learning algorithm will optimise itself to be as good as possible at the specific task, based on the information in the dataset or received feedback. Ideally, it will also perform well on new data later on.

DL is a way to design the ML algorithm, where simple concepts are built on top of each other and forming a deep structure. It replaces the classical way of designing and extracting hand-made features used for classification with the substantially different strategy of letting the computer itself decide which features are of importance by training on a dataset. While DL is not a novel concept, the availability of large datasets and increased computational

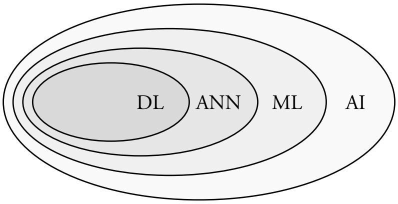


Figure 1.1: An Euler diagram showing the relation between AI, ML, ANN and DL.

power have made DL very successful and popular in recent years. Today DL has outperformed previous state-of-the-art algorithms in several visual recognition tasks and the performance has improved rapidly. The ImageNet Large Scale Visual Recognition Challenge [132] is the largest contest in object recognition. It is held every year. The first time that the winner was a deep neural network was in 2012, when Krizhevsky et al. [91] improved the performance remarkably by lowering the state-of-the-art top-5 error rate from 26.1% to 15.3%. The winners have continued to be DL algorithms ever since.

A DL algorithm consists of a deep *artificial neural network* (ANN). An ANN contains many simple processing units that are combined to form a more complex architecture. When many of these units are grouped in layers and stacked on top of each other, the network is considered a DL model, if it is deep enough. The relation between the terms AI, ML, ANN and DL is illustrated in Figure 1.1. The deep ANN contains a very large number of parameters, whose values are optimised by training on a large dataset. The design of the ANN is inspired by the human brain. The simple units in an ANN corresponds to nerve cells, where input signals are combined and forwarded if strong enough, see Figure 1.2. In the artificial variant these are replaced by a summation and an activation function. These are combined to construct more complex relations, such as the human brain or the deep neural network.

For image analysis, the *convolutional neural network* (CNN) has turned out to be a successful approach. CNNs are sparsely connected neural networks with shared weights, resulting

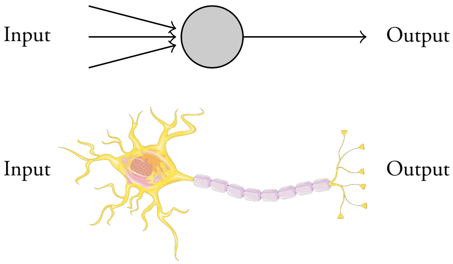


Figure 1.2: Illustration of the similarities between a computational unit in an ANN and a neuron in the human brain. *Image modified and used with permission from Servier Medical Art - Creative Commons Attribution 3.0 Unported License.*

in fewer parameters and therefore easier to train. They are designed to efficiently extract spatial information from images. The goal of image analysis is typically classification or segmentation.

DL is often described as a black box, where decisions are made in a non-understandable way. The initialisation of the neural network and the order of the samples when training are often random. However, when the training is over there is nothing random with a neural network. It consists of well-defined calculations, but its complex and deep structure typically makes it incomprehensible for humans. Therefore, to explain the networks behaviour, one usually does not refer to the trained model but rather to reasons for why it performs as it does. Typically, the trained network can be understood by features in the dataset used for training. A limitation in the dataset will likely create a model that is limited in the same way. For examples outside these boundaries, the network can appear to behave randomly simply since it has not been taught how to handle such examples.

1.3 Medical Image Analysis

DL has gained popularity and success in many fields, not at least within medicine. Large amounts of data are processed in the healthcare system, making it both a field suitable for development of successful DL algorithms and, at the same time, a field which could gain capability using DL. While different types of data are used for medical assessments, images are a very common type of data which makes medical image analysis potentially very useful.

Medical images are created in many different forms using different equipment. For example, they can be created using ultrasound, X-ray, computed tomography, magnetic resonance imaging, microscopes, and scintigraphy. All modalities create very different images, but all images have potential to be analysed automatically using DL. The applications are many, including the two considered in this thesis — detection of cancer and cardiovascular disease.

While the potential for DL within medicine is high, it also carries some difficulties. For example, images from similar machines from different manufacturers or different hospitals can have various appearances and the variations can be hard to forecast. These properties can cause large troubles for DL algorithms and are thus important to be aware of. It is therefore important to evaluate a DL algorithm carefully, e.g. with images from different machines if relevant, to make sure that the algorithm generalises to new data as expected.

How these types of automatic tools should be used in the healthcare is not evident. While they might be used for making the final diagnosis in the future, it is more likely that they will be used as a tool supporting the medical doctors initially. For example, by highlighting potential severe or urgent cases these could be prioritised. Excluding the obvious healthy cases would also reduce the workload for the medical doctors and save time which could instead be spent on the more difficult cases [129]. AI could of course also be used in other parts of the healthcare system, for example to facilitate the workflow and simplify administrative tasks [158].

Setting medical diagnoses is a hard task, requiring many years of study and experience. Assessment of an image is only part of the process to set a diagnosis and replacing that part with an AI system is a much smaller step than letting the AI set the final diagnosis. In a study by Levenson et al. "Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images" from 2015 [95], pigeons were trained with differential food reinforcement to spot cancer in images of breast tissue. As individuals, they did perform worse than the professional pathologists, but as a group they performed equally well and at that point of time far better than AI.

1.4 Prostate Cancer

1.4.1 The Prostate

The prostate is located below the bladder with the urethra passing through it, with size about a walnut, cf. [97] and the references therein. It has an important role in male reproduction, producing the fluid component of the semen. The microanatomy of the prostate tissue consists of glandular structures surrounded by stroma. The glands contain lumen, which is surrounded by an inner layer of epithelial cells, an outer layer with basal cells and a basement membrane surrounding the whole gland, separating it from the stroma. There are also some uncommon cells; neuroendocrine and stem cells. An illustration of the glandular structure is given in Figure 1.3.

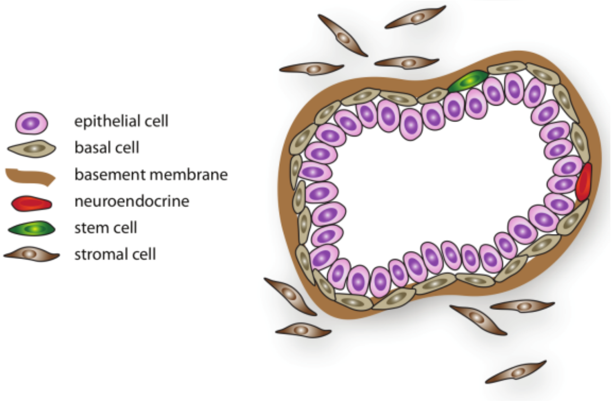


Figure 1.3: Illustration of the normal prostatic glandular structure. The lumen (white) is enclosed by one layer of epithelial cells, one layer with basal cells and a basement membrane. Stroma is surrounding the gland. Image from Lippolis [97], used with permission.

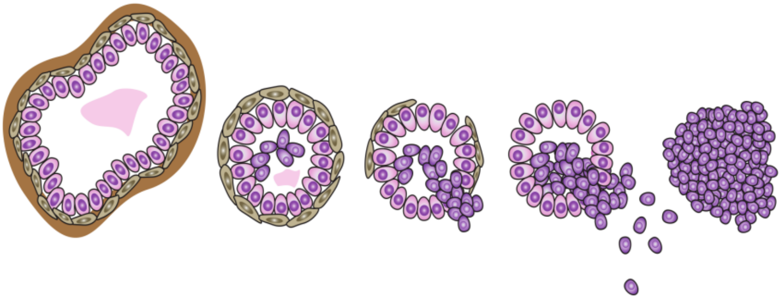


Figure 1.4: Schematic illustration of the transformation from normal gland to cancer area. Missing basal cells is the first indication of cancer. Image from Lippolis [97], used with permission.

The shape of a normal gland is irregular, but with the structure described above. Missing basal cells is a clear indication of cancer. The growing cancer will destroy the glandular structure and cancer cells will appear scattered or in solid clusters. This is illustrated in Figure 1.4.

1.4.2 Prostate Cancer

According to the global cancer statistics 2020 [144], PCa was the most frequently diagnosed cancer among men in many countries, for example in Northern and Western Europe (including Sweden). Worldwide among men it was the second most common cancer diagnosis (14.1%, 1.4 million new cases) after lung cancer (14.3%) and the fifth leading cause of cancer death (6.8%, 375 000 cases). The incidence of PCa is three times higher in high and very high *Human Development Index* (HDI) countries compared to lower HDI countries, but the mortality rate is approximately the same. In Sweden it is the leading cause of cancer death among men. The reason for the varying incidence rates worldwide is likely caused by the differences in diagnostic practices, where *prostate-specific antigen* (PSA) testing has affected the incidence rates over time in many countries. The risk factors for PCa are limited, but include advancing age, family history of PCa and certain genetic mutations.

While examinations such as PSA blood test can give an indication of PCa, biopsy of the prostatic tissue is the best way to get a conclusive diagnosis [135]. In the last years multiparametric magnetic resonance imaging has become more important for detection of prostate cancer, enabling targeted biopsies [85] and preliminary automatic prediction of cancer grade using DL [25]. For correct treatment, a good classification of the severity of the cancer is necessary. The standard procedure is that the diagnosis is determined by pathologists based on ocular inspection of prostate biopsies in order to classify them according to severity and Gleason score, see next subsection. Typically 8-12 biopsy cores are used to set diagnosis. There is a high pressure on the pathologists to always be both efficient and meticulous.

There are several different approaches used to treat PCa, depending on age and general health conditions of the patient but also how severe the cancer is. For patients with tumours that are not expected to grow for several years, one option is to use active surveillance. Under active surveillance, regular tests are conducted until signs of progression are observed. Curative treatments include radical prostatectomy, i.e. removing the whole prostate using surgery, and radiotherapy. For patients with reoccurring or advanced cancer there are few options except chemotherapy or hormone treatment, all with side effects. There are however potential new treatments, such as interstitial photodynamic therapy [148].

1.4.3 Gleason Grading

Gleason grading was named after Donald Gleason, who developed a method for PCa classification in the 1960's. Today this is the standard method, although it has been updated in various ways since then [48, 50]. The Gleason grade ranks cancer by severity based on the growth pattern of the cancer cells. The original scale went from 1 to 5, see Figure 1.5, where 1 is the lowest grade of cancer and 5 is the most severe grade of cancer. However, Gleason grade 1 and 2 are not used in practice for grading of prostate biopsies anymore [50, 135]. The Gleason grades used are thus *Gleason grade 3* (G3), *Gleason grade 4* (G4) and *Gleason grade 5* (G5).

The Gleason grade is a label assigned to individual regions of the prostate tissue and different regions may have different Gleason grades. To describe the whole prostate biopsy the *Gleason score* (GS) is used. The GS is constructed by two Gleason grades; the grade covering the largest area of the biopsy and the highest grade different from the first occurring on the biopsy, e.g. GS $3 + 4 = 7$. To distinguish e.g. GS $3 + 4$ and GS $4 + 3$, both with sum 7, the *Gleason grade group* (GG) was suggested by [50]. There is also another scale, the *International Society of Urological Pathology* (ISUP) grade [46], very similar to the GG [49]. In this thesis the GG will be used. The link between the GS and the GG can be seen in Table 1.1.

G3 and G4 are mainly detected by larger patterns, compared to G5 where single cancer cells sometimes can be seen. Since the more malignant tumours can split up into scattered locations, single cells of G5 can occur intermingled with benign tissue. Therefore, it is of great importance to find even very small areas of the highest grade. A single cancer cell can be detected by its differently shaped nucleus, but also by the nuclear texture and size [26, 135].

Table 1.1: The Gleason scores and corresponding grade groups from [50].

Gleason score (GS)	{3+3, or lower}	{3+4}	{4+3}	{4+4, 3+5, 5+3}	{4+5,5+4,5+5}
Gleason grade group (GG)	1	2	3	4	5

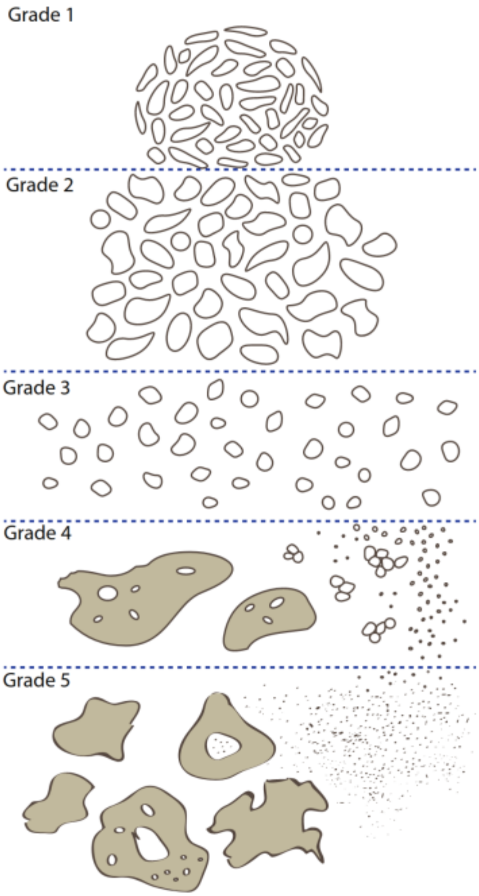


Figure 1.5: Schematic drawing of the Gleason grading system, illustrating how the cancer cells are spread in the tissue for the different grades. Image from Lippolis [97], used with permission.

1.4.4 Staining

Before the pathologists examine the prostatic tissue, it is prepared in a few steps. The tissue is embedded in paraffin, sliced into thin sections and mounted onto glass slides. The slides are stained, to highlight different features. The examination is done using a light microscope, either a traditional or a digital one.

There are many different stains, with different properties. There are immunohistochemical stainings, such as AMACR and p63, which highlight cancerous and benign glands respectively. Prognostic markers, such as Ki-67, have association with aggressive features of PCa [97]. The most common staining for cancer diagnosis is however *haematoxylin and eosin* (H&E), which are the stains that have been used in the studies in this thesis.

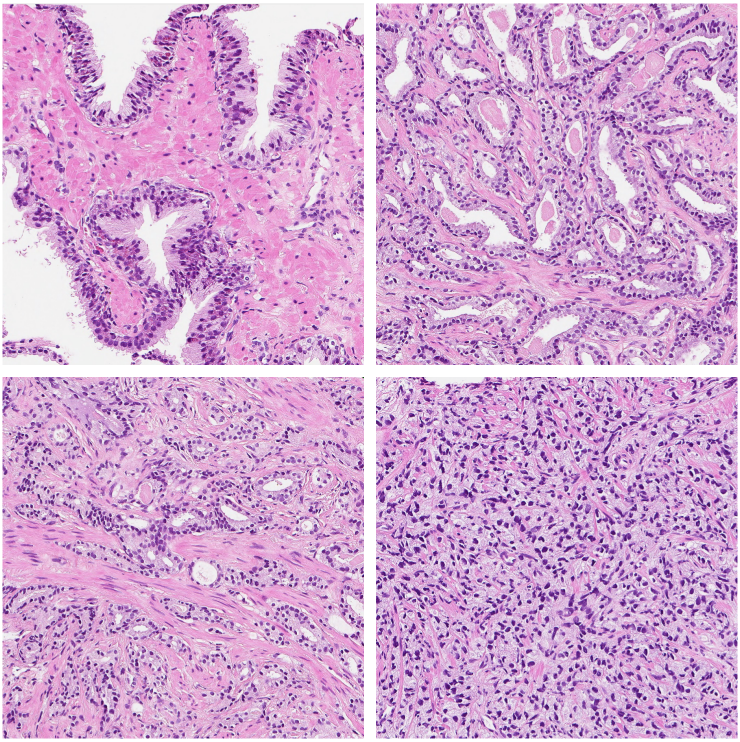


Figure 1.6: Examples of tissue stained with H&E from Skåne University Hospital, Malmö, Sweden, classified as Benign (top left), G3 (top right), G4 (bottom left) and G5 (bottom right).

Haematoxylin is a basic dye, which colours the nuclei purple. Other parts of the tissue are stained into different shades of pink, brighter for stroma and darker for epithelial cytoplasm, by the acidic dye eosin [97]. The stains absorb light and the parts which have not been stained therefore appear white. Examples of H&E stained tissue with different Gleason grades can be seen in Figure 1.6. The H&E staining is an old technique and also the most commonly used for Gleason grading, although immunohistochemical staining sometimes is used to confirm the diagnosis. However, the diagnosis will probably rely on H&E staining for at least some more decades according to [57].

Differences in the preparation of the histology slides, such as stain concentration, staining duration and tissue thickness, result in differences in staining appearances from laboratory to laboratory and over time [67]. Also, storage and handling of the slide will affect the appearance, since the stains can fade when exposed to light [106]. These variations are hard to avoid and are typically larger between different labs although variations occur also within the same lab. While this usually is not a problem for humans using a traditional procedure, it can be problematic for an AI algorithm trained on a dataset with few variations. This problem will be further investigated and discussed in this thesis.

1.4.5 Digital Pathology

Since the mid 1990's, microscope imaging has improved and now allows for whole slides to be digitised. The pathologist's light microscope can thus be replaced with a computer and this allows for *digital pathology*. Multiple studies have compared the usage of light microscopes and digital pathology, see [77] and the references therein, and found that the performance is approximately the same with both approaches. Which approach that is the most efficient one is not clear, but there are other benefits with the digital approach; reduced risk of patient and slide misidentification, reduced risk for tissue loss or damage, improved telepathology consultation (i.e. study slides from a distance) as well as simplified annotation drawings and measurements using different software. The possibility to use automatic image analysis to make assessments of the tissue is another great benefit.

The benefits with automatic analysis of tissues are many. The pathologists' workload is heavy due to the high incidence of PCa resulting in large volumes of histological material. Also, there is a lack of pathologists resulting in long queues for cancer patients and delays in diagnosis. A system that for example automatically recognises all benign cases would drastically reduce the workload, leaving only the hard cases to the pathologists to examine. Moreover, multiple studies have shown that there is a high variability in grading between different pathologists in multiple studies, [2, 47, 110, 123], something that can have considerable impact on which diagnosis and treatment the patient gets. With a second opinion from an automated image analysis algorithm, diagnoses might come closer to consensus, decreasing over- and undertreatment. Hence there is good reason to try to automate the pathological analysis process.

1.5 Coronary Artery Disease

1.5.1 Cardiovascular Physiology

The heart and circulatory system have a central role in the functioning of the body. Oxygen and nutrients needed by the body, as well as waste products such as carbon dioxide, are transported by the blood via the circulatory system. The heart is a muscular organ that functions as a pump of the blood. The coronary arteries are part of the circulatory system, supplying the heart muscle itself with blood. The main three coronary arteries are the *left anterior descending artery* (LAD), *left circumflex artery* (LCx) and *right coronary artery* (RCA), see Figure 1.7.

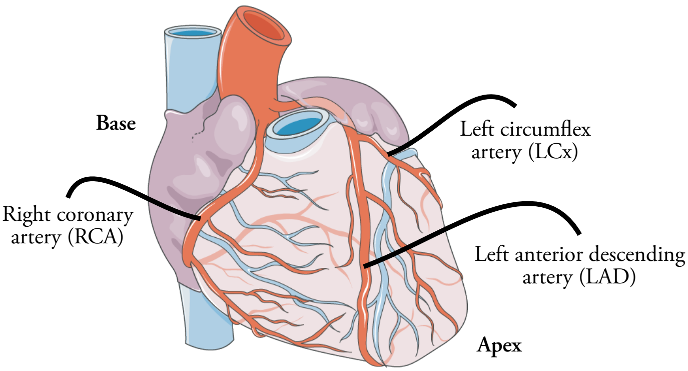

Base
Left circumflex artery (LCx)
Right coronary artery (RCA)
Left anterior descending artery (LAD)
Apex

Figure 1.7: The three main coronary arteries, supplying the heart muscle with blood. *Image modified and used with permission from Servier Medical Art - Creative Commons Attribution 3.0 Unported License.*

1.5.2 Coronary Artery Disease

According to World Health Organization [162], cardiovascular disease is the leading cause of death world-wide, estimated to represent 31% of all global deaths in 2016. Of these death cases, CAD and stroke are estimated to correspond to 82%. These two diseases are usually acute events, where the blood is prevented from flowing to the heart or brain. The blockage that prevents the blood flow is usually caused by build-up of fatty deposits on the inner walls of the blood vessels. There are multiple behavioural risk factors of CAD, such as tobacco use, unhealthy diet, physical inactivity and harmful use of alcohol.

Multiple methods exist to diagnose CAD. Two common methods are described in the following subsections; non-invasive MPI and invasive coronary angiogram. The American College of Cardiology Foundation and *American Heart Association* (AHA) guidelines advocate the use of pre-test probability estimates — derived from the Diamond and Forrester model and Coronary Artery Surgery Study — to guide diagnostic testing for patients with suspected stable CAD. The motivation for this is that it is simple to use and can be implemented at the physician's first encounter with the patient [27, 39, 54]. Additionally, guidelines for the diagnosis and management of chronic coronary syndromes from the *European Society of Cardiology* (ESC) are also used by referring physicians. The ESC pre-test probability estimates from 2013 have been used in this thesis [109]. These tools are of great value not only to the referring physicians (i.e. the physician that establishes medical necessity) but also to nuclear medicine specialists reading MPI images in order to have a clinical scenario of the patient before the test is performed. Both the ESC pre-test probability and AHA pre-test probability are determined based on age, gender and nature of symptoms, as described in [3] and [61] respectively. The different pre-test probabilities are given in Table 1.2 (ESC) and Table 1.3 (AHA).

Table 1.2: Pre-test probabilities according to the ESC scale from 2013.

	Chest pain					
	Typical		Atypical		Non-anginal	
Age	Men	Women	Men	Women	Men	Women
30-39	59%	28%	29%	10%	18%	5%
40-49	69%	37%	38%	14%	25%	8%
50-59	77%	47%	49%	20%	34%	12%
60-69	84%	58%	59%	28%	44%	17%
70-79	89%	68%	69%	37%	54%	24%
80+	93%	76%	78%	47%	65%	32%

Table 1.3: Pre-test probabilities according to the AHA scale. Interm. is short for intermediate.

	Chest pain						Dyspnoea	
	Typical		Atypical		Non-anginal			
Age	Men	Women	Men	Women	Men	Women	Men	Women
30-39	Interm.	Interm.	Interm.	Very low	Low	Very low	Very low	Very low
40-49	High	Interm.	Interm.	Low	Interm.	Very low	Low	Very low
50-59	High	Interm.	Interm.	Interm.	Interm.	Low	Low	Very low
60-69	High	High	Interm.	Interm.	Interm.	Interm.	Low	Low

1.5.3 Myocardial Perfusion Imaging

MPI is one of the most common cardiological examinations performed today for diagnosis and risk assessment in patients with suspected CAD. It is a method using nuclear medicine, where radioisotopes attached to drugs are injected into the bloodstream and travel internally to specific organs or tissue. The emitted gamma radiation is captured by external detectors to form 2-dimensional images, *scintigraphy*, or 3-dimensional images, *single-photon emission computed tomography* (SPECT). MPI provides valuable information on for example ischaemia, myocardial injuries and left ventricular ejection fraction [31, 45, 52, 122]. The technique has seen improvements in recent years with the introduction of *cadmium-zinc-telluride* (CZT) technology, allowing to perform the scan in shorter times [44], use low dose radiotracer protocols [51, 122] and achieve high diagnostic performance at the same time [119].

From the SPECT images, a 2-dimensional polar map is constructed to simplify the analysis. One example is shown in Figure 1.8. Analyses are currently performed visually by physicians based on the AHA standard 17-segment model of the polar maps, see Figure 1.9. The centre represents the cardiac apex, and the three surrounding circles consist of the apical, mid and basal areas, compare with Figure 1.7.

Analysing the MPI studies is very time consuming and the number of nuclear medicine specialist with experience in MPI is limited and steadily decreasing. At the same time, the number of MPI examinations increases each year, for example since there are new recom-

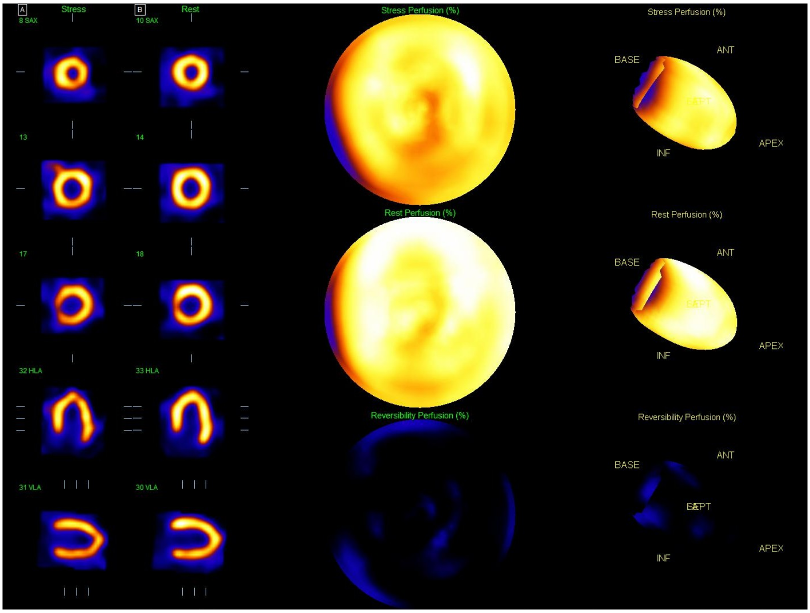


Figure 1.8: The image shows part of a DICOM image, excluding e.g. patient information. It illustrates an example of constructed polar maps in artificial colouring of the intensities from a patient with obstructive CAD in the RCA.

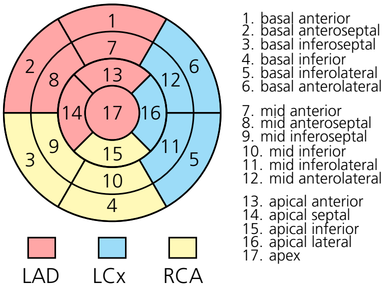


Figure 1.9: The three coloured regions correspond to the regions used in this work: LAD, LCx and RCA. The division is done according to the AHA standard 17-segment model.

mendations for the healthcare to increase medical imaging such as MPI [88]. Furthermore, the qualitative approach is subjective and suffers from inter-observer variability between nuclear medicine specialists [151].

Apart from major observer dependent problems, the MPI technique itself carries additional problems with artefacts which can originate from several sources. First, it is important to carefully position the patient's heart when performing MPI with the CZT camera to avoid image artefacts [70]. Furthermore, a slight movement from the patient during the imaging session, the body habitus from the patient (especially in obesity [53]) and the proper limits of the technique can all result in inconclusive studies. This will prompt further studies and more radiation delivered to the patient, decreasing the quality of life for the individual and increasing the costs for the healthcare system. Artefacts can be caused by the breast or diaphragm, but also *left bundle branch block* (LBBB). These artefacts can be a source of false positive results in CAD detection. LBBB causes delayed activation of the left ventricle, making it contract later than the right ventricle. It is normally detected using *electrocardiogram* (ECG) [145].

1.5.4 Quantitative Coronary Angiography

Patients who are thought to have significant CAD in the MPI study are further examined by means of *invasive coronary angiography* (ICA). The coronary angiogram is constructed using X-ray imaging and an injected dye to see the vessels and possible blockages (Figure 1.10). To avoid inter-observer variability and achieve reproducibility, the degree of coronary artery stenosis can be evaluated by means of *quantitative coronary angiography* (QCA) [60, 146]. QCA has been used since the late 1980s. It measures the diameter of coronary artery stenosis and expresses it in percentage based on the angiogram, see Figure 1.11. The edges of the artery are automatically or semi-automatically detected to provide quantitative estimates from the coronary angiograms.

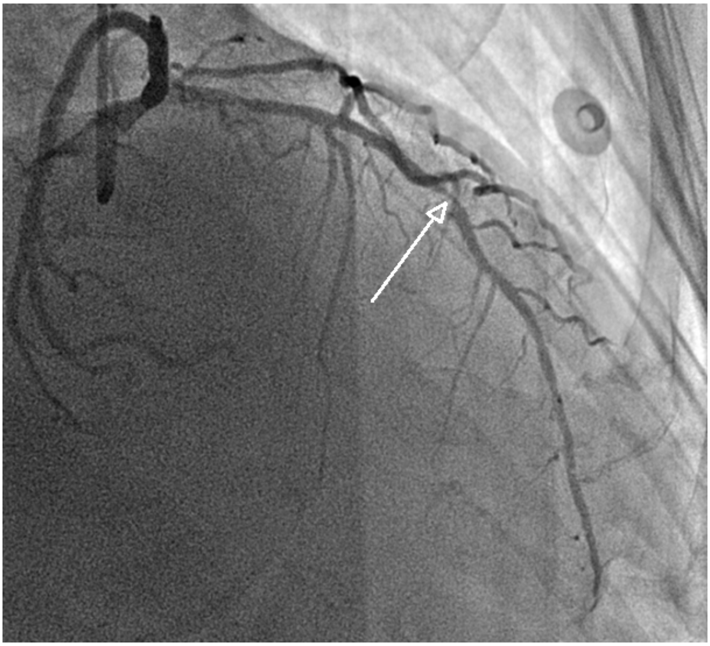
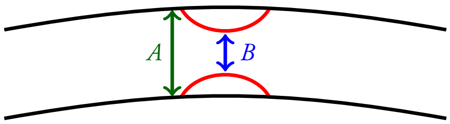


Figure 1.10: An obstruction, see the arrow, confirmed in the ICA with a QCA value of 92%.



$$\text{Diameter stenosis (\%)} = \frac{B}{A} \cdot 100$$

Figure 1.11: Illustration of how the diameter stenosis is determined using QCA, by measuring the minimal lumen diameter B compared with the reference diameter A.

1.5.5 DL Applied to CAD

In the field of cardiovascular image analysis, DL has been applied to most examinations, including MPI [100]. An image analysis algorithm could reduce some of the problems listed above, such as the subjectivity and time consumption [62, 151]. An automatic or even semi-automatic system would improve the healthcare system, reducing subjectivity and potentially extracting more information than the human eye can see, such as predicting QCA from MPI.

Chapter 2

Data

With the introduction of ML, in particular DL, the demand and interest in large datasets have increased rapidly. Apart from the design of suitable neural networks, attention should be given to the construction and usage of the dataset for a successful algorithm. For example, a small or skewed dataset will be a limiting factor. Both the training and evaluation of ML algorithms are often based on a dataset, making it even more crucial that the dataset is handled correctly to avoid biases and incorrect conclusions. How this typically is done is described below. Thereafter follows two sections, introducing the datasets used in this thesis for PCa grading and for CAD detection, respectively.

For ML algorithms, the dataset is typically split into three parts — *training*, *validation* and *test*. The training dataset is used to train the model, e.g. tune the weights in the neural network. During training, the validation dataset is used to measure the performance and compare different settings, such as network architectures or hyperparamters, to choose the optimal design. Finally, the test dataset is used to evaluate the performance to assure reliable results.

The test dataset should represent the expected future use cases as well as possible. For some medical applications it might be desirable to include samples from different hospitals or machines, while this is irrelevant for other applications. For example, when grading microscopy images of PCa, it is desirable to include images with different origins. For these images, there are inevitable variations from e.g. different staining procedures and different microscopes at different hospitals, which also typically varies over time. To claim that an algorithm is useful at more than one hospital it is therefore important that it is tested on images from multiple different hospitals, a desire that sometimes is hard to fulfil due to limited availability of datasets. However, for algorithms developed to be used with e.g. images from a single specific machine the requirements on the dataset could be lower.

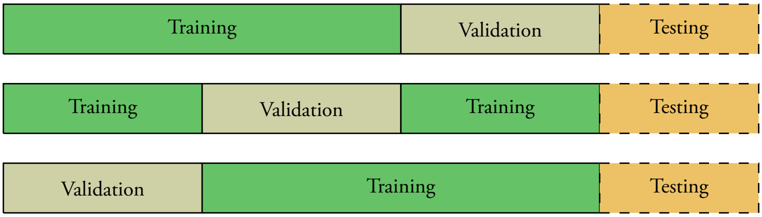


Figure 2.1: The dataset is typically split in three parts; *training* to train the network, *validation* to choose e.g. network design and hyperparameters, and *test* to measure the performance on new data. Cross-validation can be used to utilise all data for both training and validation. The test dataset is sometimes omitted. The figure illustrates 3-fold cross-validation with an external test dataset.

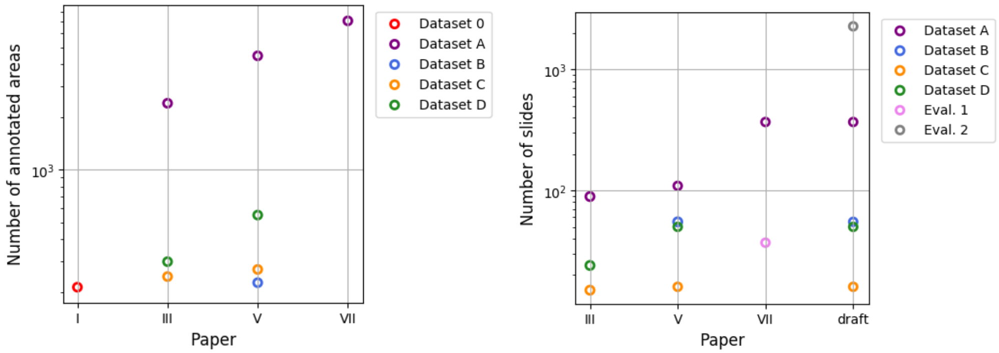
A common problem is that the available datasets are too small, or smaller than desired. The test dataset is then sometimes omitted and only the validation dataset is used for evaluation. To reduce the risk of incorrect conclusions, the dataset can be divided into train and validation multiple times to be able to both train and validate on all data, called *cross validation*. The data is split into n parts, where $(n-1)$ parts are used for training and the remaining one for validation of the result. This is repeated n times, using the different parts for validation. An external test dataset is sometimes but not always used. The procedure is illustrated in Figure 2.1.

2.1 Prostate Cancer Dataset

Several different datasets have been used for the development and evaluation of the Gleason grading algorithms presented in this thesis. Each one of them is introduced below, with details on their sizes, origins and annotations. All of them consist of H&E stained tissue, but with different types of annotations. The tissue referred to as benign consists of all tissue considered to be non-cancerous. Most of the datasets have been collected in parallel with the development of the methods in this thesis. This has resulted in continuous change of the prerequisites for the algorithm development and shift in detail level of the annotations over time due to continuous progression. The sizes of the datasets range between 213 smaller images, used for both training and validation, to more than 2000 whole slide images, used for testing only. The sizes of the different datasets used in the different papers are illustrated in Figure 2.2.

2.1.1 Dataset 0

The smallest dataset, hereafter referred to as *Dataset 0*, was used in Paper I presented in Section 4.1. It has previously been used in [83, 97]. The dataset has two sources; PathXL in Belfast and Beaumont Hospital in Dublin in conjunction with the Prostate Cancer Research



(a) **(b)**

Figure 2.2: Number of (a) annotated regions and (b) available slides in some of the datasets, used in different papers. Dataset A refers to both the original and the extended version. Eval. 1 and 2 mean Evaluation Dataset 1 and 2 respectively. Draft refers to the study presented in Section 6.2.

Consortium. The tissue was scanned on two different scanners in 40X magnification; a Leica SCN400 Digital Slide Scanner and an Aperio Scanscope CS scanner. It was unknown which images originated from which of the two sources. There are some small differences in hue and saturation between the samples, see Figure 2.3, possibly due to their different origins. In total there are 213 images, cropped from whole slide images such that each image is considered to consist of only one Gleason grade or benign tissue. The number of images of each class was 52 benign, 52 G3, 52 G4 and 57 G5.

Due to uncertainty in the annotations of these images, whether the whole image or only

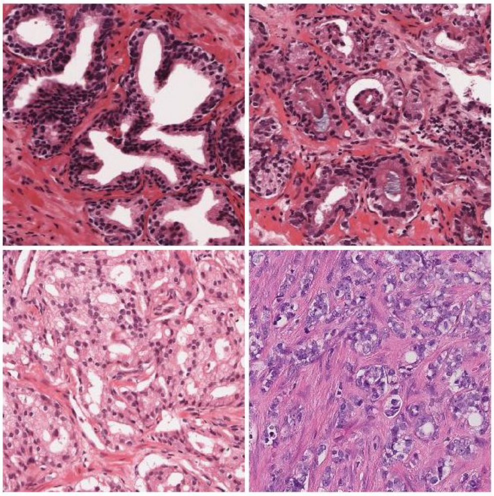


Figure 2.3: Examples of images from Dataset 0 classified as Benign (top left), G3 (top right), G4 (bottom left) and G5 (bottom right).

the majority of the image consisted of tissue with one Gleason grade, the dataset was not used in the further studies. Instead of letting a second pathologist annotate them, more data was instead annotated to the datasets presented in the next paragraph.

2.1.2 Dataset A, B, C and D

For the remaining papers included dealing with PCa, data from *Dataset A, B, C* and *D* was, in part or in full, used. The largest dataset, Dataset A, was used for training and validation in different partitions. The remaining three datasets were only used for testing, except for the study presented in Chapter 6.2. Details of the datasets are given below.

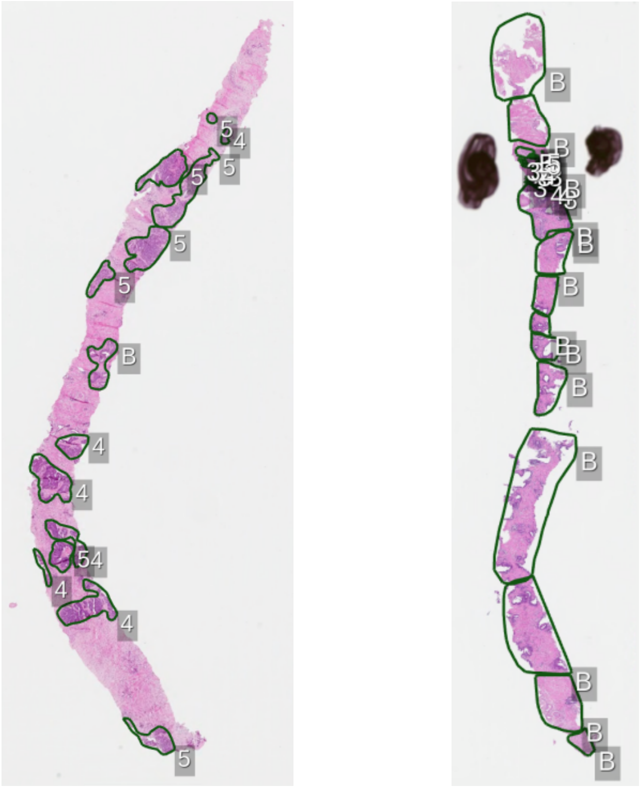
Dataset A originates from Skåne University Hospital in Malmö, Sweden, collected between 2014 and 2018. Pathology reports were available for all cases. The H&E stained slides had been used for standard pathological diagnostics and the clinical diagnosis was available for each patient. The slides were scanned on an Aperio CS2 scanner (Leica, Newcastle, UK) at a resolution of 0.247 µm/pixel (40X magnification).

Dataset B was collected at Helsingborgs lasarett, Sweden, but scanned in Malmö on the same scanner as Dataset A. The conditions for this dataset were very similar as for Dataset A with available pathology reports. The studies which used Dataset A and Dataset B were approved by the Regional Ethics Committee in Lund, Sweden (no. 2005/494 and no. 2018/11).

Dataset C was obtained from Linköping University Hospital, Sweden, and approved by the Regional Ethics Committee in Linköping, Sweden (no. 2013/195-31). It was scanned on an Aperio AT Turbo scanner (Leica, Newcastle, UK) at a resolution of 0.5 µm/pixel (20X magnification). No clinical information was available.

Dataset D originates from Erasmus University Medical Center in Rotterdam, the Netherlands. The slides were scanned on a Hamamatsu HT 2.0 scanner (Hamamatsu Photonics K.K, Tokyo, Japan) with resolution 0.228 µm/pixel (40X magnification). No clinical or patient information was available and the slides were considered remnant material.

The digital images from all four datasets were uploaded to the Sectra IDS7 software (Sectra AB, Linköping, Sweden). Two senior consultant pathologists (P1 and P2, working at the same institute but in different departments at Skåne University Hospital in Malmö, Sweden) annotated the regions with Gleason patterns G3, G4 and G5 or benign within the IDS7 platform. Examples of annotations are given in Figure 2.4. No distinction was made between the different morphologies, such as the cribriform pattern, within the three Gleason patterns. In case the Gleason patterns were intermixed, the most representative pattern was assigned. Consecutive biopsy slides, double stained immunohistochemically against p63 and AMARC [21, 126] as well as the original pathology reports were available



(a)

(b)

Figure 2.4: Annotations of cancer areas drawn by pathologists in the Sectra IDS7 software. The pen marks next to the tissue in (b) are drawn by the pathologist setting the clinical diagnosis, to highlight the cancer region.

for most cases in Dataset A and Dataset B and could be used in the event of ambiguity during annotation.

Slides with folded and fragmented tissue were included in the study, as long as the sections were deemed of good enough quality for pathologists' routine use. The annotations performed by the two pathologists were pooled and the total number of slides and annotated regions can be seen in Table 2.1. The numbers in the table correspond to the full datasets.

Table 2.1: Number of annotated slides and regions in the Dataset A, B, C and D.

Dataset	Nbr of slides	Nbr of annotated regions			
		Benign	G3	G4	G5
A	109	2535	726	849	357
B	55	117	48	52	9
C	16	54	9	189	17
D	50	29	304	209	8

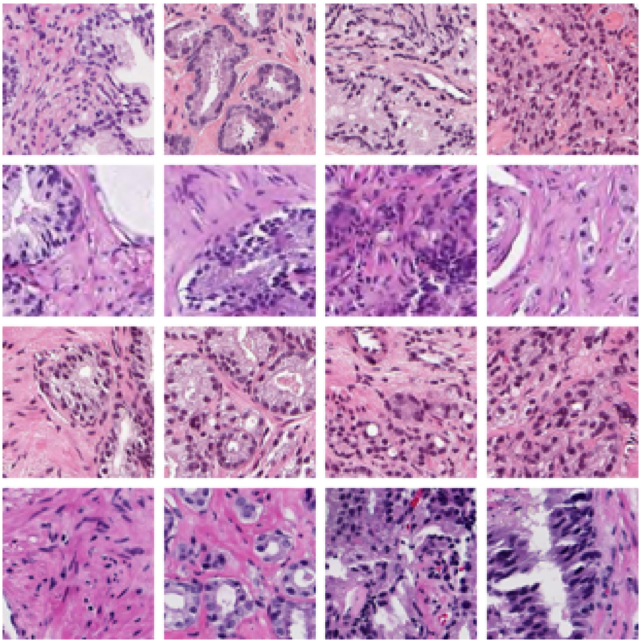


Figure 2.5: Examples of H&E stained tissue with different grades (left to right: Benign, G3, G4, G5) from the different datasets (top to bottom: Dataset A, B, C and D).

Note that in some of the publications only parts of the datasets were used. Example images from the different datasets can be seen in Figure 2.5.

2.1.3 Extended Dataset A

Dataset A was extended in Paper VII, denoted *Extended Dataset A*, using the same ethical permit, with additional annotations drawn from scratch but also annotations done by correcting preliminary annotations from a developed AI algorithm. Details for the algorithm used are given in Section 6.1. These two parts were denoted "Train 1", consisting of 476 annotated biopsy slides from 119 patients, and "Train 2", with an additional subset of 222 slides from a further 55 patients, see Table 2.2. The slide selection and annotation drawing were done in the same way as for the smaller version.

Table 2.2: Numbers of biopsy slides in Extended Dataset A (Train 1 and Train 2) and in Evaluation Dataset 1 (Test). The clinical diagnosis given to each biopsy was extracted from clinical records.

	Benign	3+3	3+4	4+3	3+5	4+4	4+5	5+4	5+5	Biopsies	Patients
Train 1	221	99	40	33	8	19	37	10	8	476	119
Train 2	146	34	16	10	1	9	6	0	0	222	55
Total to train	367	133	56	43	9	28	43	10	8	698	174
Test	13	9	6	3	3	0	3	0	0	37	21

2.1.4 Evaluation Dataset 1

For testing, a separate cohort of 37 biopsy scans from 21 patients, denoted *Evaluation Dataset 1* (“Test” in Table 2.2), with the same origin as Dataset A was collected and used in Paper VII presented in Section 6.1. The study using this cohort was covered by the same ethical permit as Dataset A and Dataset B. The slides were scanned with Aperio CS2 (same as Dataset A), but 36 of the slides were also scanned with a Hamamatsu S60 scanner (Hamamatsu Photonics K.K, Tokyo, Japan; Scanner 2) 24 months later, at a resolution of 0.220 µm/pixel.

With the same procedure and conditions as above, e.g. including slides with folded and fragmented tissue, the slides were annotated by the same pathologists (P1 and P2). However this time, all slides were annotated by both pathologists. Furthermore, P1 repeated the annotations on the same samples after 1-year (first evaluation was named P1-1 and second P1-2). This allowed assessment of intra- and inter-observer differences. The GS given at the time of diagnosis for each example was obtained from electronic medical records.

2.1.5 Evaluation Dataset 2

A second test cohort has been collected with patients included in accordance with the *Prostate Cancer Research International Active Surveillance* (PRIAS) protocol [20, 147]. Active surveillance is becoming an increasingly accepted standard clinical approach to low-risk PCa. The patients included in this cohort all have GS of 3+4 or less with $\leq$ 10% cancer per biopsy and $\leq$ 2 biopsy cores with cancer. Follow-up of the patients is done regularly with for example PSA blood tests and new prostate biopsies. The latter are scheduled 1, 4, 7, 10 years after diagnosis and thereafter every fifth year. Due to the composition of the cohort most biopsies will only contain low-grade cancer or benign tissue.

The cohort originates from three hospitals in Sweden; Skåne University Hospital in Malmö and Lund respectively as well as Centralsjukhuset Kristianstad. The clinical diagnosis was available for each patient, but no detailed annotations of cancer regions. The biopsies, collected between 2007 and 2018, were retrospectively scanned between 2018 and 2021 on an Aperio CS2 scanner (same as Dataset A). In total, data from 180 patients and more than 5000 biopsies were available. However, the slides from 2010 and earlier were faded and considered to be of too poor quality to include. Furthermore, only slides from 88 patients have been scanned and analysed this far. The dataset used for testing consisted of data from 2263 slides whereof 260 slides contained cancer. The cancer length was reported for 255 of the cancer slides, whereof 108 had $\leq$ 1 mm cancer and 149 had $\leq$ 2 mm cancer. The average cancer length was 2.9 mm. The dataset is denoted *Evaluation Dataset 2* in this thesis, used in Section 6.2. Ethical approval was provided by the Regional Ethical Committee in Lund (no. 2008/708).

2.2 Coronary Artery Disease Dataset

One dataset has been used for development and evaluation of automatic detection of CAD in this thesis. However, only part of the dataset was available for Paper VIII and only part of the dataset had QCA values, which was used in Paper X. The dataset as a whole is described in this section. The studies were approved by the Regional Ethics Review Board in Region Östergötland, Sweden (approval number 2019/00097).

Adult subjects referred to MPI at Linköping University Hospital during 1st of June 2014 to the 30th of October 2019 (n=3658) were considered. The referral for stress testing was at the clinical discretion of the referring cardiologist. In total 759 patients were selected from the database, with MPI studies conducted in a dedicated CZT cardio camera [52, 119] (D-SPECT Spectrum Dynamics). Patients met the following requirements: they had undergone ICA maximum 6 months after MPI or had a low or very low pre-test probability of ischaemia according to the AHA. For all patients, information about CAD in the RCA, LAD and LCx was available, as well as information about artefacts in the MPI images from breast, diaphragm or LBBB. Details can be seen in Table 2.3. For 275 patients, the QCA value was also available. Details of this subset is given in Paper X, Section 7.4.

MPI was performed according to the European Association of Nuclear Medicine guidelines [76]. After a stress test, MPI was done on the CZT camera in both upright and supine positions with at least 1 million myocardial counts. All patients performed either a physical bicycle ergometer stress test, a pharmacological stress test with regadenoson, or a combination of both, at the discretion of the nuclear medicine physician. All subjects received prior, routine instructions, sent to their home, to avoid potential regadenoson agonists for at least 24 h before MPI (e.g., coffee, tea, cola drinks, chocolate and cacao). In case of use of a pharmacological stress test, 400 µg (5 ml) of regadenoson were administered intravenously. In the case of a combined protocol, a bicycle ergometer stress test with 30-50 watts was used with regadenoson after 2 minutes of cycling.

Left ventricular myocardial contours were computed using standard Cedars-Sinai Medical Center *Quantitative Perfusion SPECT* (QPS) software version 2012. Left ventricular contours were defined by a technologist with >15 years of experience in nuclear cardiology who was blinded to angiographic and clinical findings. When needed, the technologist corrected the gross initial left ventricular localisation, the left ventricular mask, and the valve plane position. Thus, the polar maps from upright and supine position were registered to each

Table 2.3: The number of cases with artefacts and with CAD in the different arteries. Note that each case can have a positive label for multiple arteries and artefacts.

Total	Normal	LAD	RCA	LCx	LBBB	Breast	Diaphragm
759	387	154	141	123	66	55	80

other when reconstructed, except some minor noise, and the three arteries are depicted at the same position in each image.

MPI was routinely assessed visually by nuclear medicine consultants, each with at least 10 years of experience in MPI, by using the 17-segment model of the left ventricle, see Figure 1.9 on page 13, and a conventional 4-point grading system: 0 – normal uptake, 1 – equivocal uptake, 2 – moderate uptake, and 3 – severe reduction of uptake, according to current guidelines [113, 139]. All MPI images were retrospectively re-evaluated by an experienced nuclear medicine physician. Only segments with an uptake score ≥ 2 at stress were considered to have a definite uptake reduction. Of these, (i) segments with reversible defects (ischaemic defects) were defined as those with at least 1-point decrease in uptake score on the rest acquisition and (ii) the other segments were considered to have a fixed defect (myocardial infarction segments) except for those with a definitely normal contractility on gated SPECT and for which the final diagnosis was attenuation artefact. Segments with attenuation artefact were excluded for the final definition of the stress defects area. Ground truth for LBBB was provided from ECG.

The ground truth for pathological MPI images was obtained by means of ICA conducted routinely according to standard techniques. Coronary angiograms were analysed by a blinded experienced observer. Per cent lumen area reductions due to intracoronary atheromatous plaques were first determined visually on end-diastolic frames and with the help of a quantitative angiography software (General Electric Advantage Workstation, Cardiac X-Ray Applications, Stenosis Analysis v1.6) for stenosis visually assessed to be around the 50% threshold by an experienced angiographer physician. Where applicable, two separate measurements in orthogonal views of the same stenotic segment were obtained and values were averaged to represent an approximate measurement of the per cent vessel area stenosis. Any stenosis $\geq$ 50% was considered significant and regarded as a positive QCA test. Total coronary vessel occlusions were marked as 100% lumen area stenosis. When no visible stenotic lumen was seen on angiography with a marginally patent vessel (with other than normal flow) the stenosis was also regarded as a total occlusion. In the case of "normal" MPI images a follow up period of at least 6 months was carried out.

All patients had data from stress in up-right and supine position. Some patients also had images in rest, but not all of them. The reason for this is mainly that no rest examination was carried out on patients who were considered healthy in the stress examination. To not affect the algorithm by missing data, the rest images were not used. The polar maps were cropped in full resolution which meant that the resulting images had a size of either 296 × 296 pixels or 336 × 336 pixels. Since the majority of these crops had the smaller size, all were resized to have the size 296 × 296 pixels. The polar maps are usually shown in artificial colouring to the medical doctors. However, since the colouring does not add any additional information to the greyscale intensity image the original data was instead used with the developed DL systems. A conversion between colour and greyscale was developed

Table 2.4: Characteristics of the auxiliary parameters.

Parameter	Average	Standard Deviation	Range of Values	Comments
Gender	-	-	$\{0,1\}$	36% women
Age	66	11	$[21,98]$	-
BMI	28	5	$[16,52]$	14 values missing
Angina symp.	-	-	$\{0,1\}$	-
AHA prob.	-	-	$\{0,1,2,3\}$	-
ESC Pre-test prob.	32	39	$[0,100]$	-

in Paper VIII, since only part of the images were collected in greyscale at that time. For description, see Section 7.2. Examples of MPI images for patients with CAD in different arteries and with different artefacts can be seen in Figure 2.6.

Other parameters are also relevant and used by nuclear medicine specialists to set the diagnosis. Therefore those parameters, denoted "auxiliary parameters", were also extracted from the database. The relevant auxiliary parameters available were: *gender*, *age*, *body mass index* (BMI), *angina symptoms* (no or yes) and *pre-test probability* according to ESC and AHA (ranked as *very low* (0), *low* (1), *intermediate* (2) or *high* (3)). Details of the auxiliary parameters can be seen in Table 2.4. In 14 cases the BMI value was unknown.

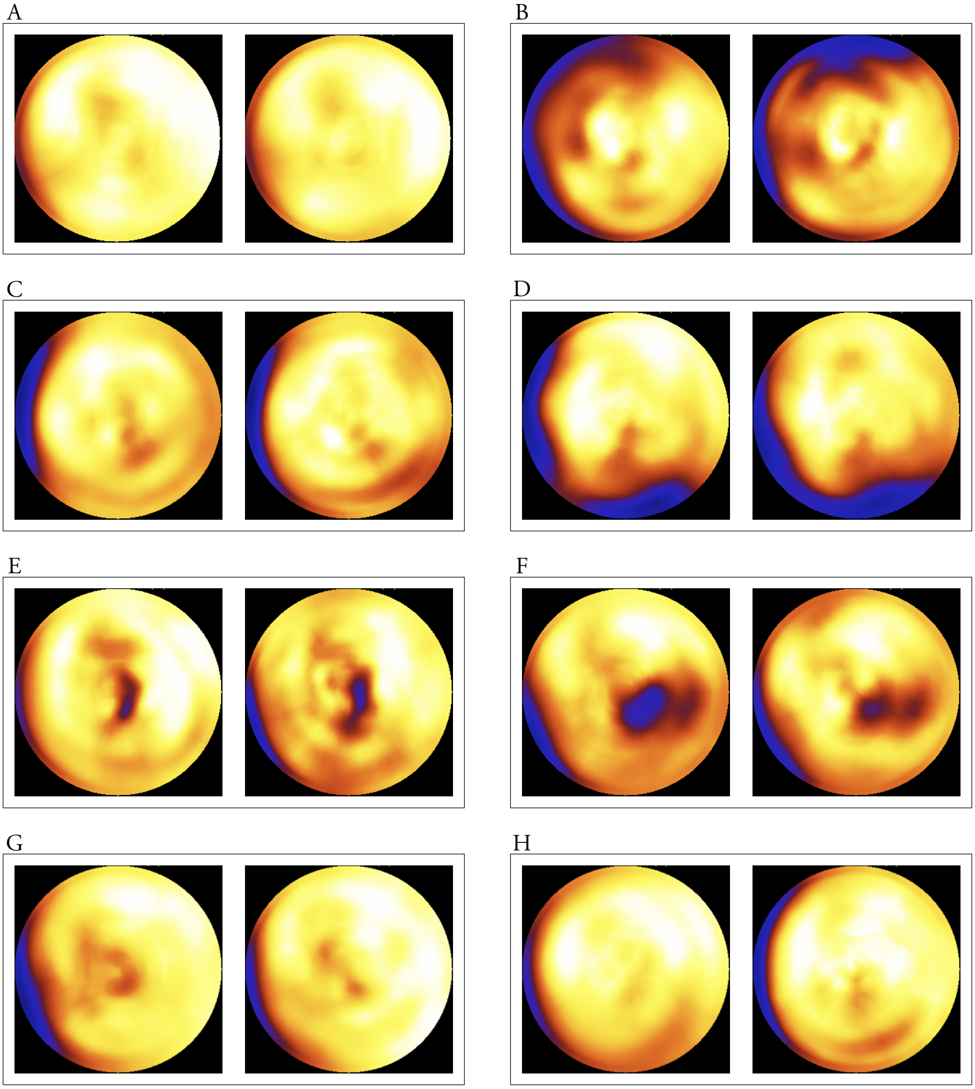


Figure 2.6: Polar maps with artificial colouring from four patients; (A) healthy, myocardial ischaemia in (B) LAD, (C) LCx, (D) RCA, (E) LAD and RCA, (F) LCx and RCA, (G) LBBB in the septum and apex of the left ventricle, territory of the RCA and LAD, and (H) healthy but with a diaphragm artefact from attenuation, in the inferior wall of the left ventricle, RCA/LCx territory. Compare with the segments in Figure 1.9 on page 13. Dark areas indicate poor blood flow.

Chapter 3

Methods

The different methods used in this thesis are introduced in this chapter. The focus is on DL, but also related methods used, as for example preprocessing, are introduced. The first section describes a method for stain normalisation of histology images. It is followed by introduction of neural networks and explanations of different architectures which are useful for different tasks in medical image analysis. Thereafter data augmentation is introduced, which can be very helpful when training DL algorithms with limited datasets and aiming at good generalisation performance. Finally, registration of images and different metrics to measure performance are described.

3.1 Digital Stain Normalisation

Images of H&E stained tissue vary in appearance, see Figure 3.1. The reasons for this are many: different procedures when staining the tissue, different manufacturers of the stains as well as different scanners with different settings will all affect the final image. Several suggestions for how the images could be modified digitally to appear more similar, independent of their origin, have been published [106, 131, 160]. In general, the idea is to digitally separate the two stains and then normalise them individually to a target image. In this way, variations in hue and intensity of the stains between different sites can be decreased. This can be very useful if this type of images is used by a DL algorithm, since such algorithms have a tendency to only work well on images similar to the images used when training the algorithm.

The images of H&E stained tissue mainly consists of three colours; blue or purple from the haematoxylin, pink from the eosin and white for the background. An example image and its *red-green-blue* (RGB) values are shown in Figure 3.2a and 3.2b. The RGB values can

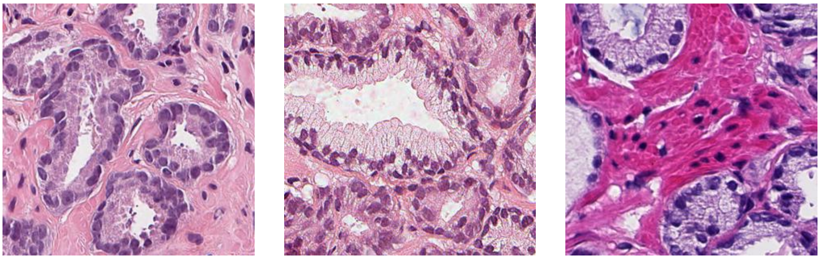


Figure 3.1: Three examples of G3 tissue, with very different staining appearances. All three examples are from Dataset A.

be divided into two parts; one with the haematoxylin pixels and one with the eosin pixels. While these components are not linear in the RGB space, their *optical densities* are linear (see Figure 3.2c). The optical densities O, a $3 \times n$ matrix for an image containing n pixels, are determined using the Beer-Lambert law,

$$I = I_0 \exp(-O)\,, \tag{3.1}$$

where I is the observed intensity and I_0 is the illuminating light, i.e. the intensity of the background pixels. The optical density is given by

$$O = -\log(I/I_0)\,. \tag{3.2}$$

The optical density is determined by two factors; the stain appearances and the stain densities. A *stain vector* describes how the obtained colour is related to the stain concentration. The *stain matrix* S has size 3×2 with each column representing the RGB colour of one stain, i.e. one stain vector. The process of the optical density is modelled as

$$O = SV + \epsilon\,, \tag{3.3}$$

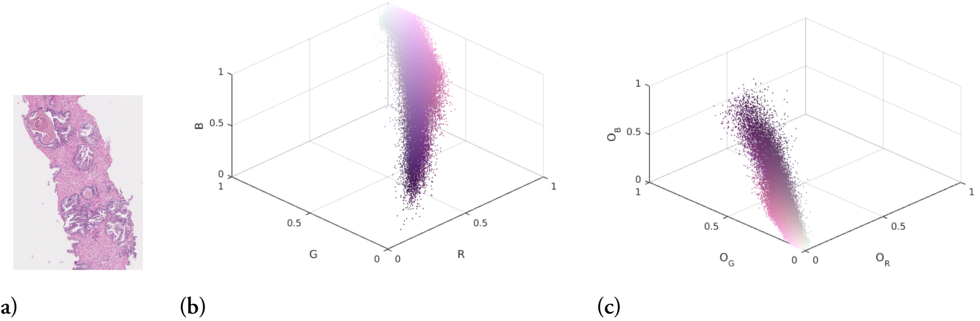


Figure 3.2: Illustration of (a) an image of H&E stained tissue, (b) the RGB values of the pixels in the image and (c) the corresponding optical densities.

where V is the stain concentration (a $2 \times n$ matrix) and ϵ is additional noise, which should be minimised. The factorisation is commonly called colour deconvolution.

A method to separate stains when the stain vectors are given is presented in [131]. This method has the drawback that the stain vectors need to be determined manually and these can differ for different slides even from the same lab. In an automatic setting, this is not an appealing approach. In [106], the stain vectors are automatically found using singular value decomposition of the optical density. The amount of stain in the two images is however assumed to be the same, which could be suboptimal. The method in [160] solves that problem by not modifying the stain concentration, but instead only changing the colour appearance. A very similar approach was presented by [86], but they use a pretrained classifier to classify pixels into the two stain classes to estimate the colour basis. The methods in [86] and in [160] were evaluated in [134], for the task of epithelial-stroma classification using different techniques. Both methods were successful and none was stated to be better than the other. In this thesis the method by Vahadane et al. [160], described in more detail below, has been used.

The idea in [160] is to use the optical density as described above. The estimation of S and V is based on sparse non-negative matrix factorisation [94] and the LASSO algorithm [157]. The motivation for non-negative matrix factorisation is that neither the stain concentration nor the stain matrix can be negative; $S \geq 0$ and $V \geq 0$ (elementwise inequality). Furthermore, since the two stains only bind to different biological structures a sparsity constraint is imposed on the matrix V. This results in the optimisation problem

$$\begin{aligned} \underset{S,V}{\text{minimise}} \quad & \frac{1}{2}||O - SV||_F^2 + \lambda \sum_{i=1}^{2} ||V(i,:)||_1 \,, \\ \text{subject to} \quad & S, V \geq 0 \,, \\ & ||S(:,i)||_2^2 = 1, i = 1, 2 \,. \end{aligned} \tag{3.4}$$

where λ is a regularisation parameter. The last constraint is to remove equivalent solutions of type $(S/\alpha,\, V\alpha)$, $\alpha > 0$. To solve the problem alternating optimisation over S and V is used, keeping the other argument fixed. The default value of the regularisation parameter is $\lambda = 0.1$.

To normalise the staining of one (source) image to another (target) image, the procedure above is performed for both of the images. Thereafter, the dynamic range of the stain concentrations V_s for the source image is normalised to be the same as the stain concentrations V_t for the target image, for each of the stains i;

$$\widehat{V_s}(i,:) = V_s(i,:) \frac{RM(V_t(i,:))}{RM(V_s(i,:))} \,, \tag{3.5}$$

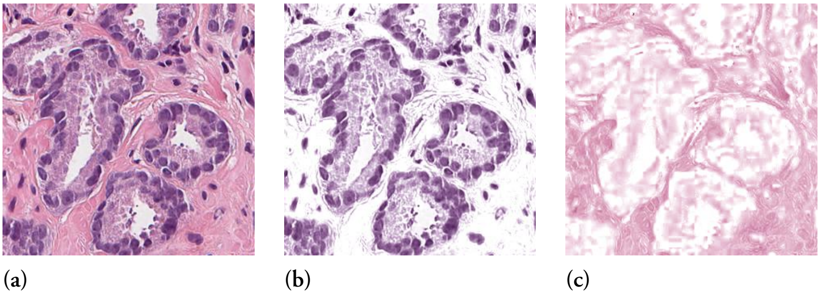


Figure 3.3: Example of digital stain separation using the method in [160]: (a) original and (b, c) the two stain channels.

where $RM(\cdot)$ is a function that computes robust pseudo maximum of each stain at 99%, i.e. the maximum when neglecting the 1% highest values. The new optical density for the source image is determined by the normalised stain concentrations $\widehat{V}_s$ and the stain matrix S_t for the target image according to (3.3); $\widehat{O}_s = S_t\widehat{V}_s$. Finally, the optical density is converted to RGB using (3.1).

An example of separation into the two stain components using the method described above can be seen in Figure 3.3. Examples of normalisation of different images can be seen in Figure 3.4. As can be seen, the results are not ideal.

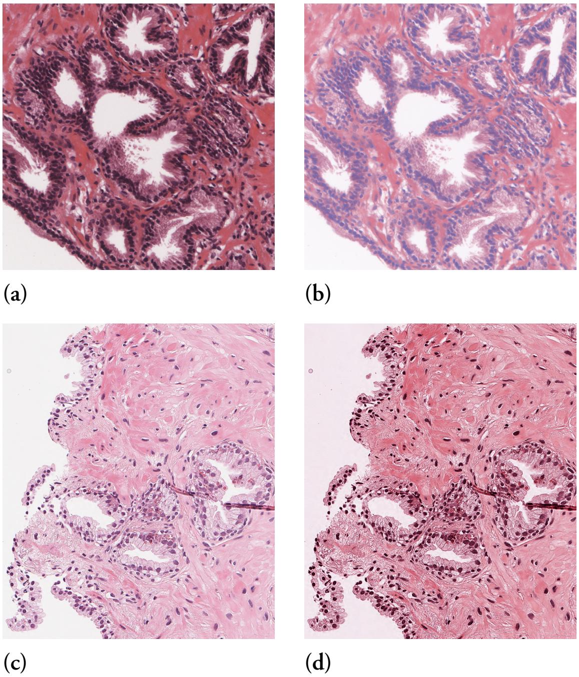


Figure 3.4: Top row: (a) image from Dataset 0 and (b) the same image normalised to have the same stain appearance as (c). Bottom row: (c) image from Dataset A and (d) the same image normalised to have the same stain appearance as (a).

There exist more recent approaches of stain normalisation based on DL, for example [137], presenting superior results to the methods described above. Another approach for stain normalisation is to train cycle-consistent adversarial networks [164], as described in Section 3.6.

3.2 Artificial Neural Networks

The goal with this section is to give an overview of neural networks, introduce the key concepts and some common architectures. The content of the section is limited to what has been used to produce the results in this thesis. More thorough descriptions can for example be found in [63].

3.2.1 The Perceptron

The neural network consists of many, or at least one, *neurons* or *perceptrons*. The neuron produces one output value y from multiple inputs $x_1, ..., x_n$. The output is determined by weights $w_1, ..., w_n$, a bias b and an activation function φ (described in the next subsection) as

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right). \tag{3.6}$$

The bias is often replaced with an additional input node $x_0 = 1$ and a corresponding weight $w_0 = b$, giving the simpler expression

$$y = \varphi\left(\sum_{i=0}^{n} w_i x_i\right). \tag{3.7}$$

3.2.2 Activation Functions

The *activation functions* are used to allow the neural network to learn non-linear relations. There are multiple different ones. Some common ones are the sigmoid function, which for example is used when the output should be limited to the range $(0, 1)$;

$$\varphi(x) = \frac{1}{1 + \mathrm{e}^{-x}}. \tag{3.8}$$

The hyperbolic tangent limits the output to the range $(-1, 1)$;

$$\varphi(x) = \tanh(x) = \frac{\mathrm{e}^{x} - \mathrm{e}^{-x}}{\mathrm{e}^{x} + \mathrm{e}^{-x}}. \tag{3.9}$$

The *rectified linear unit* (ReLU) is another activation function that commonly is used in CNNs;

$$\varphi(x) = \max(0, x)\,. \tag{3.10}$$

Softmax is a special activation function, since it combines all of its inputs $\boldsymbol{x} = (x_1, \ldots, x_n)$. It is typically only used in the last layer of a classification CNN, where it makes sure that the output lies in the range $[0, 1]$ as well as sums up to one and therefore often are interpreted as probabilities. It is defined as

$$\varphi(\boldsymbol{x})_i = \frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_j}}\,. \tag{3.11}$$

Note that the softmax fulfills $0 < \varphi(\boldsymbol{x})_i \leq 1$ and $\sum_{i=1}^{n} \varphi(\boldsymbol{x})_i = 1$.

3.2.3 Multilayer Perceptron

To determine multiple and more complex features from the input, multiple neurons are stacked next to each other in layers and the layers are stacked following each other. The structure is called a *multilayer perceptron*. The connections between the neurons can be done arbitrarily but will be limited to feed-forward networks in this thesis, meaning that no neuron connects to a previous neuron. Except the initial input layer and final output layer, the intermediate layers are called hidden layers. An illustration of an ANN with two hidden layers is illustrated in Figure 3.5.

3.2.4 Convolutional Neural Networks

A CNN consists of a sequence of layers, whereof at least one is a *convolutional* layer. It typically also includes some sort of *pooling* layer, commonly max pooling. When using

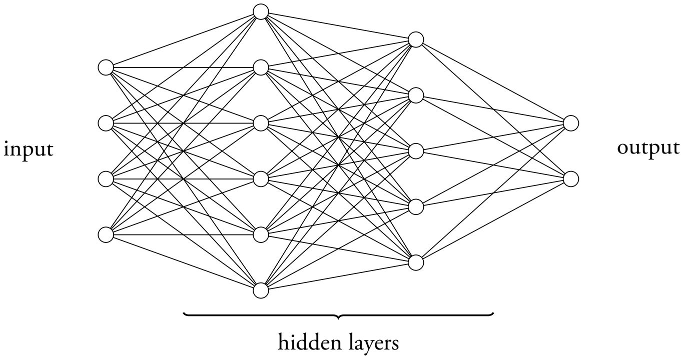


Figure 3.5: A multilayer perceptron with two hidden layers and two output nodes. Each circle represents a neuron, each line a weight and each column of neurons a layer.

images as input, these operators take a three-dimensional data array as input and produce a three-dimensional data array as output. The two first dimensions give the spatial information while the third gives a feature vector for each pixel. For an RGB image, these feature vectors are given by the three colour components. Further down in the network, the number of features is typically increased while the spatial resolution is decreased if the network should be used for classification.

A continuous, one-dimensional convolution of a function x and a kernel w is defined as

$$(x * w)(t) = \int x(a)w(t-a)da\,. \tag{3.12}$$

The corresponding discrete definition is

$$(x * w)(i) = \sum_{a} x(a)w(i-a)\,. \tag{3.13}$$

For two dimensions it is defined as

$$(x * w)(i,j) = \sum_{a}\sum_{b} x(a,b)w(i-a,j-b)\,. \tag{3.14}$$

When images are used as input, the convolutional layers extract spatial information and create features (channels of the image) by letting a fixed size 4-dimensional convolutional kernel convolve spatially over the data array. The kernel consists of weights, but compared to the multilayer perceptron the network is only sparsely connected and the weights are shared. For an input image I with i_1 channels (three for an RGB image) and a convolutional kernel w with size $k_1 \times k_2 \times i_1 \times i_2$ the output will have i_2 channels. The result for pixel (m_2, n_2, j) of a convolution with I and w is given by

$$(I * w)(m_2, n_2, j) = \sum_{m_1=-\lfloor a \rfloor}^{\lceil a \rceil} \sum_{n_1=-\lfloor b \rfloor}^{\lceil b \rceil} \sum_{i=1}^{i_1} I(m_2-m_1, n_2-n_1, i)w(m_1, n_1, i, j)\,, \tag{3.15}$$

where $a = (k_1 - 1)/2$ and $b = (k_2 - 1)/2$, using the ceil and floor operators. The spatial size of the output could be either smaller or the same if the input is padded by for example adding appropriate number of zeros surrounding the input. Illustration of a convolution without padding when $i = j = 1$ is given in Figure 3.6. Convolutions can be used with different strides, i.e. shifts between the convolutions. In this example stride 1 was used.

Pooling is used to reduce the spatial resolution. Max pooling or average pooling are common methods. For a fixed spatial window the pooling layer produces only one output. Max pooling keeps the highest value and average pooling keeps the average of the input. The stride is typically the same as the size of the window.

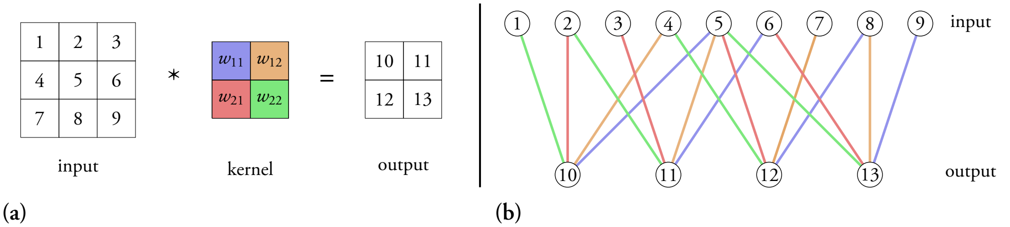


Figure 3.6: Convolution between a 3 × 3 image with one channel and a 2 × 2 filter, with stride one and without any padding, is illustrated in two different ways in (a) and (b). The numbers denote the indices of the nodes and not their pixel values. The coloured lines in (b) correspond to multiplications with the corresponding coloured weights in (a). Note the shared weights and hence fewer tunable parameters compared to a fully connected ANN. Idea to figure from [65] and [82].

An example of a simple CNN architecture is given in Figure 3.7. Each convolutional layer is followed by a ReLU activation function. The output in this case consists of three nodes, which for example could be used to predict which of the three classes the image belongs to.

3.2.5 CNN Architectures

A CNN can be designed in many ways. There are, however, some architectures which have shown very good performances and have become standard CNN models. One is the *AlexNet* [91], which won the ImageNet Large Scale Visual Recognition Challenge [132] in 2012. This was followed by the deeper *VGG* networks which used smaller filter kernels [138]. Some other designs, which have been used in this thesis, are introduced below. An overview of their information is given in Table 3.1.

The *ResNet* was first presented in [68]. It introduced the skip connections, illustrated in Figure 3.8, making it possible to train much deeper networks by giving the network the option to copy the activations from layer to layer (or ResNet block to ResNet block).

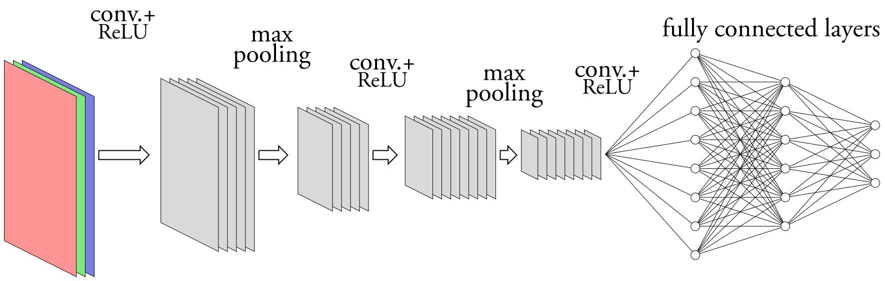


Figure 3.7: A CNN taking an RGB image as input and making classification into three classes.

Table 3.1: The sizes and features of a few different common CNN models for classification.

Network	Year	Feature	Parameters	Depth
VGG16	2014	Deeper	138,357,544	23
ResNet152	2015	Deeper, shortcut connections	60,419,944	152
ResNet50	2015	Deeper, shortcut connections	25,636,712	50
Inception V3	2015	Wider	23,851,784	159
Xception	2016	Depthwise separable convolutions	22,910,480	126
DenseNet121	2017	Shortcut concatenations	8,062,504	121

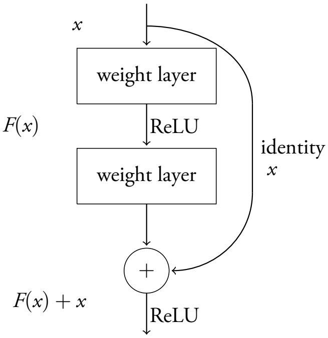


Figure 3.8: The ResNet building block with a skip connection, adding a previous activation to a later one.

Another popular design is the *inception* network. The original design, inception V1, was called GoogLeNet [149]. It consists of inception modules, illustrated in Figure 3.9a, which use multiple different filter sizes and concatenate the results. The 1×1 convolutions only use features from one pixel to construct new features. The inception design has been updated and the version that has been used in this thesis is inception V3, which for example has factorised the large convolutional layers into smaller ones and uses asymmetric convolutions, see Figure 3.9b.

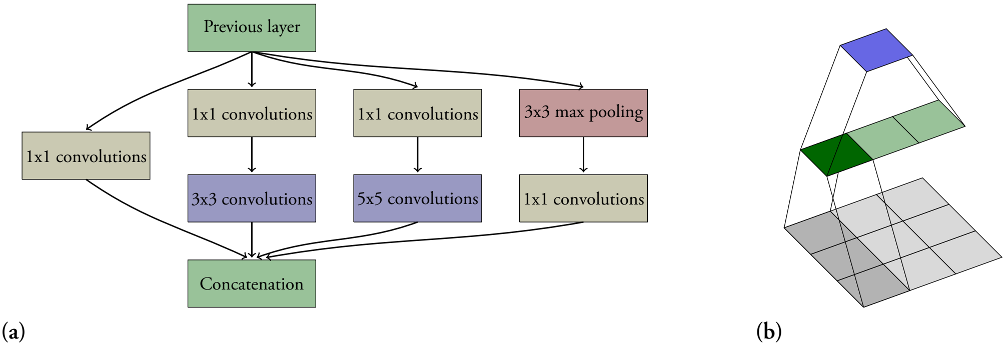


Figure 3.9: (a) The inception module and (b) the factorisation of a 3×3 convolution as 3×1 and 1×3 convolutions.

A variant of the inception is the *xception* model [29], where the inception modules have been replaced with depthwise separable convolutions, i.e. a spatial convolution performed independently over each channel followed by a pointwise (depthwise) convolution. It also adds residual connections. It has approximately the same number of parameters, but outperformed the inception V3 on several datasets.

The *DenseNet* [74] resembles the ResNet, but replaces the addition in the residual blocks with concatenations. It has fewer parameters and is thus suitable for smaller datasets. It exists in different sizes, where the smaller DenseNet121 design has been used in this thesis.

3.2.6 From Classification to Segmentation

The CNNs described above are designed for classification. Another common goal when applying CNNs to images is segmentation, dividing the image into regions belonging to different classes. Some CNN designs for this task will be introduced in Section 3.5. Another approach is to use the classification CNN for segmentation. By letting the CNN predict the class for multiple small areas from a sliding window over the image, a coarse segmentation can be obtained. Thanks to the CNN's design, a classification CNN can often be modified to output this coarse segmentation of a large image instead of only a class per smaller image.

3.3 Training ANNs

An essential part of using DL is the training of the algorithms. While the first step is to choose a network design with a suitable size to the task at hand, the network will most likely not perform well without proper training. The previous section only described different network designs, but to be useful the weights which usually are initialised randomly need to be optimised. In this section various aspects related to training of the ANN will be described. The parameters related to the training of a specific network are commonly called hyperparameters, which needs to be chosen and can be tuned for optimal training. Further descriptions and details for training procedure and optimisation can be found in [63].

The training is done by presenting the training data for the network and comparing the output it produces with the desired output. The error is quantified by the *loss function*, introduced in subsection 3.3.1, and is used to determine how the network should be optimised. The loss is a function of the weights in the network. The optimisation is done using a method called *backpropagation*, where the partial derivatives of the loss function with respect the weights are computed. The achieved gradients are used to update the weights using an optimisation scheme such as *stochastic gradient descent* or *Adam* [87], introduced in Subsection 3.3.2. The training procedure is iterated over all the training data a certain number of times. One such iteration is called an *epoch*. The training data is not necessarily processed one by one, but instead commonly presented in *mini-batches* of multiple examples, for which the losses are summed up. The training can either be performed for a fixed number of epochs or *early stopping* can be used, where the stopping is determined based on the performance on the validation data. While training, the loss is observed for both the training and validation data. If the loss for the validation data starts to increase, the model is *overfitting* or *overtraining*. This means that the performance of the model on the training data do not generalise to the validation data and probably not to other datasets either.

3.3.1 Loss Function

The loss function or objective function defines the error for the model and is thus what should be minimised while training. It connects the output $\hat{y}$ from the ANN to the true classes y of the examples. For a normal classification problem $y_i = 1$ for exactly one value of i and $y_i = 0$ for all other values of i. *Cross-entropy* is a common loss function when training classification networks, defined as

$$L(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)\,. \tag{3.16}$$

For a correct classification $\hat{y}_i \approx y_i$ and the loss $L \approx 0$, while for an incorrect classification the loss is higher.

When the output is not discrete classes but instead continuous variables, the mean squared error for the n values could for example be used as loss function instead:

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,. \tag{3.17}$$

The total loss for a batch of examples is given by the sum of the loss for each example. When training ANNs where the number of examples of each class is not the same, the loss is typically weighted to give each class the same importance regardless. This is achieved by multiplying the loss for each example by a weight, which for example could be inversely proportional to the total number of examples of that class.

3.3.2 Optimiser

Stochastic gradient descent and its variants are the most used optimisation algorithms for DL. Gradient descent takes a step with length μ, known as the learning rate for ANNs, in the direction of the greatest descent for each weight w_{ij} as

$$w'_{ij} = w_{ij} + \Delta w_{ij}, \text{ where } \Delta w_{ij} = -\mu \frac{\partial L}{\partial w_{ij}} \,. \tag{3.18}$$

In stochastic gradient descent the update is done on a random mini-batch consisting of n of the total N samples simultaneously;

$$\Delta w_{ij} = -\mu \frac{1}{n} \sum_{n \in N} \frac{\partial L}{\partial w_{ij}} \,. \tag{3.19}$$

The *learning rate* is an important parameter for optimal training. Often is the learning rate initialised at a higher value and then gradually decreased while training. To accelerate the learning and avoid noisy gradients, a second *momentum* term can be added. It accumulates an exponentially decaying moving average of past gradients;

$$\Delta w_{ij}(t+1) = -\mu \frac{1}{n} \sum_{n \in N} \frac{\partial L}{\partial w_{ij}} + \alpha \Delta w_{ij}(t) \,. \tag{3.20}$$

The parameter α determines how much impact the previous gradients should have.

The learning rate is a crucial parameter which should be high enough to avoid getting stuck in local minima or too slow training, but also low enough to achieve convergence. There exists a number of algorithms that adapt the learning rates for each model parameter. In this thesis, the Adam optimiser [87] has been used. It includes estimates of both first and second moments of the gradients when computing the individual adaptive learning rates for different parameters.

3.3.3 Regularisation

Regularisation is a group of strategies which reduce the risk of overfitting. There are multiple different techniques, such as *augmentation* which will be further discussed in Section 3.8. Early stopping mentioned in the introduction to this chapter can also be considered a regularisation technique, limiting the training performance but optimising the generalisation performance. A CNN is in itself regularised thanks to the shared weights in the convolutional layers. Other strategies are parameter regularisation or *weight decay*, where a term driving the weights w closer to zero is added to the objective function. The term is commonly the L_2 regularisation $\frac{\lambda}{2}||w||_2^2$, where λ is a hyperparameter determining the impact of the term. Training the network for multiple tasks can also have a regularising effect; a variant of this will be described in Section 3.7. Combining several models and let them vote on the final output, known as *bagging* or *ensemble*, can also reduce the generalisation error. Two other common techniques are described below.

Dropout is a powerful and computationally inexpensive regularisation technique. Dropout is applied to one or multiple layers in the network while training, where a predefined ratio of the connections to the following layer are dropped, meaning that some weights are put to zero. The connections to drop are randomly selected for each example. This prevents the network from relying on only a few features and instead forces it to take multiple features into consideration. Dropout is commonly applied to some of the last layers in the ANN.

Batch normalisation is motivated by the difficulty of training very deep networks. While training, each weight is updated using the gradient, based on the assumption that no other layer is changed. However, all layers are updated simultaneously. High-order terms in the updates can both be very small or exponentially large, making it hard to choose an appropriate learning rate. Batch normalisation is a method of adaptive reparametrisation which reduces these problems. While training, the activations H from a batch of examples in a layer are normalised as $H' = \frac{H-\mu}{\sigma}$, where μ is a vector with the mean of each unit and σ the corresponding standard deviation of H. When utilising the trained network, these parameters are replaced with the averages of the mean and standard deviation respectively from the training data.

3.3.4 Transfer Learning

For faster and potentially better training of ANNs, *transfer learning* can be used. The network is then pre-trained on a large dataset, before it is fine-tuned on the relevant dataset. There are multiple benefits with this approach. For example, for the standard CNN designs there already exists pre-trained weights from training on the ImageNet dataset [38]. Using those, the training can be faster than if training from scratch. Furthermore, for small datasets the risk of overfitting can be reduced by freezing some of the initial layers, meaning that these layers will not be optimised when training. However, if the datasets used for pre-training and fine-tuning are very different, the usefulness of pre-training is limited. This is often the case for medical images, where the image appearance is quite different from the images in ImageNet, and [125] found that transfer learning offers little benefit to the performance.

3.3.5 Implementation

The networks used to produce the results presented in this thesis have been implemented in either MATLAB (The MathWorks, Inc., Natick, Massachusetts, United States) or Python (Python Software Foundation, https://www.python.org/). In MATLAB the CNN library MatConvNet [161] was used, which provided all standard CNN components and allowed for GPU processing. The Python implementations were done using the Keras library [30] with the Tensorflow backend [1].

3.4 Autoencoder

The *autoencoder* is a neural network consisting of two consecutive parts; the encoder and the decoder. The network is trained to reconstruct its input, which might not seem very useful. That is however typically not the goal with the autoencoder. By restricting the network such that it cannot copy the input directly, e.g. by having latent layers that are smaller than the input and output layers, a prefect reconstruction is prevented. This way the network is forced to prioritise which features of the input that are important to represent the dataset. For a trained autoencoder the output from the encoder will thus be a compressed version of the input. Since no labelled data is needed to train the network, it is an unsupervised learning method.

As described above, the autoencoder can be used for dimensionality reduction. This is illustrated in Figure 3.10, showing an encoder f, decoder g and latent layer (encoding) smaller than the input and output. The goal when training is to minimise the difference between the input and output, $L(x, g(f(x)))$, where L is a loss penalising differences between the two variables. In particular, a linear autoencoder trained with the mean squared error as loss will learn to span the same subspace as principal component analysis [63]. More sophisticated autoencoders will however be able to learn more complex dimensionality reductions. Often are tied weights, i.e. shared weights between the encoder and decoder, used for the autoencoder. This has, however, not been used in the autoencoders in this thesis.

There are multiple other use cases for autoencoders, such as denoising autoencoders, learning manifolds and generative modelling. This section will be restricted to using autoencoders for dimensionality reduction.

The typical autoencoder is symmetric, where the decoder is a mirrored version of the en-

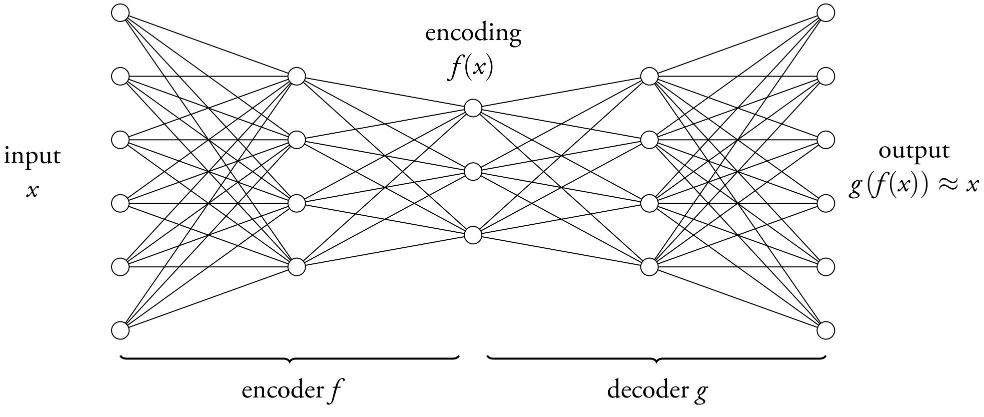


Figure 3.10: Illustration of an autoencoder for dimensionality reduction, with two fully connected layers in the encoder and decoder respectively. The autoencoder is trained to reconstruct its input.

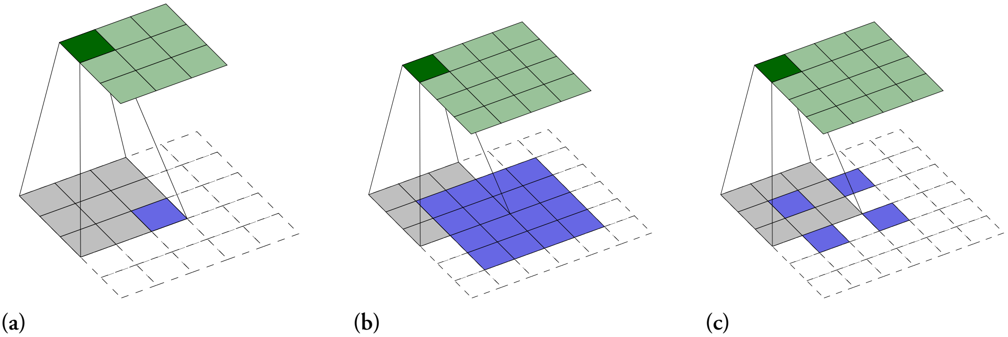


Figure 3.11: Transposed convolution for a convolution with unit stride and kernel size 3, when using (a) no zero padding and (b) zero padding 1. In (c), the transposed convolution for a convolution with stride 2, kernel size 3 and zero padding 1 is shown. All of the transposed convolutions have stride 1. The blue pixels are the input, white pixels are zeros added by the padding and green pixels are the results after the transposed convolutions.

coder. For the simple fully connected neural network, this mirroring is obvious. However, the decoder design is less apparent for a CNN. A max pooling layer in the encoder is replaced with upsampling in the decoder. The convolutional layers in the encoder are in the decoder replaced with *transposed convolutions*, sometimes called *deconvolutions* (however not to be confused with the inverse of a convolution, which is the mathematical definition). The transposed convolution is described in detail in [43]. It is computed as a convolution, but with adapted zero padding and stride. The relation between a convolution and the corresponding transposed convolution depends on the zero padding and stride of the convolution. There are many different variants, whereof three examples are given here.

For a convolutional layer with unit stride, no zero padding and a convolutional kernel size k, the transposed convolution is given by a convolution with the same stride and kernel size, but with padding $k - 1$. Such an example, for an input with size 3×3 and $k = 3$ is illustrated in Figure 3.11a. To compare, the transpose convolution corresponding to a convolution with unit stride, odd filter size $k = 2n+1, n \in \mathbb{N}$ and padding n (i.e. the input and output from the convolution will have the same size), is identical to the convolution. This is illustrated in Figure 3.11b for $k = 3, n = 1$. The transposed convolution of a convolution with non-unit stride is more complicated. The kernel size is still the same, but for the transposed convolution the stride is 1 and zeros are added both as padding but also in between each input unit. An example is given in Figure 3.11c.

For efficient downsampling of microscopy images of prostatic tissue, an autoencoder could be used. An example of this is shown in Section 4.2, downsampling images from 20X magnification to 5X magnification.

3.5 Semantic Segmentation

The aim of the neural networks discussed thus far has been classification, i.e. for each image the desired output is a class. However, if the aim is not to determine the class of the image, but rather the precise localisation of one or multiple objects in the image, the design of the neural networks has to be modified. The goal is then to do *semantic segmentation*, as illustrated in Figure 3.12.

One way to obtain a segmentation is to perform multiple classifications on overlapping patches of a larger image, in a sliding-window setup as described in Subsection 3.2.6. There are however other approaches, suitable when high resolution segmentations are desirable. The *fully convolutional network* was introduced in [102]. The idea with these networks is to extend the usual contracting classification network with successive upsampling layers, increasing the resolution of the output. Features from the contracting path are concatenated with the upsampled features, to give an output based on high resolution features as well as deep features. The concept can seem similar to the autoencoder, but with the fundamental difference that the desired output is different. The architecture was further developed in [130], introducing the *u-net*, and the rest of this section will focus on this architecture.

The u-net performs semantic segmentation and outperformed previous state-of-the-art methods on several semantic segmentation challenges. It has been widely used and proved successful in many different applications, in particular for medical images [42], and has also been modified to e.g. handle 3D volumes [32]. It consists of a contracting and a symmetric expanding part. An illustration can be seen in Figure 3.13.

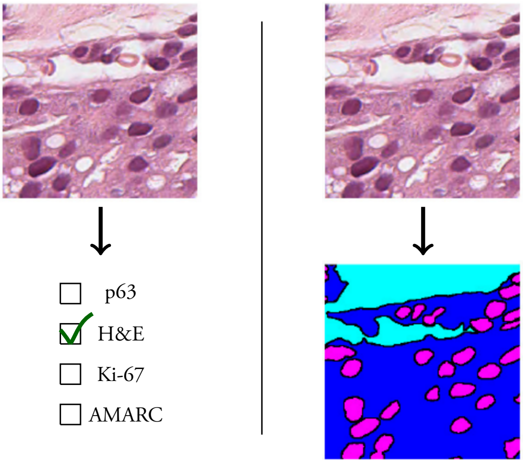


Figure 3.12: There can be different goals when applying DL to images. Two examples are illustrated here: (left) classification of which staining is used on a prostatic tissue sample and (right) semantic segmentation of the same image into relevant tissue components.

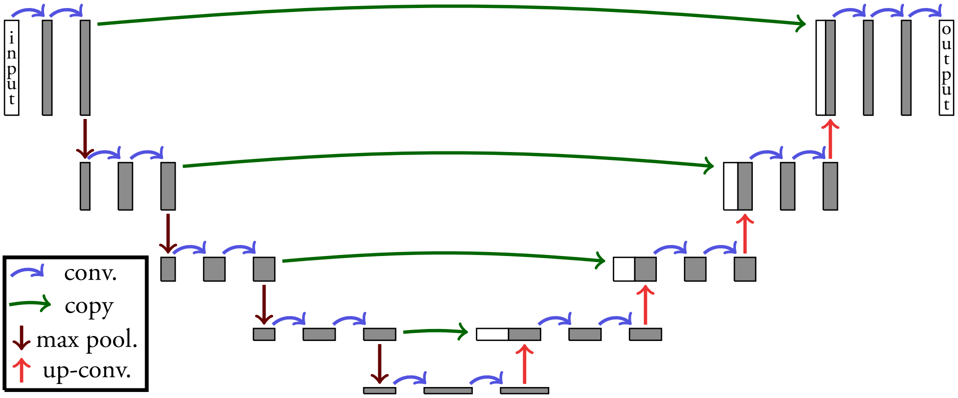


Figure 3.13: Illustration of the u-net, used for semantic segmentation. The copied feature maps (white boxes) are concatenated with the corresponding feature maps in the expanding part. Image adapted from [130].

The contracting part of the network consists of convolutional layers and activation functions (originally 3 × 3 convolutions and ReLU activations) as well as downsampling using max pooling layers (originally 2 × 2 with stride 2). The expanding part consists of *up-convolutions*, described below, convolutional layers and activation functions symmetrical to the contracting part, as well as concatenation with cropped feature maps from the contracting part at the same resolution. Cropping is needed since border pixels are lost in the convolutional layers. This is also the reason for why the output is slightly smaller than the input.

The up-convolution used in the original u-net is performed by upsampling of the feature map by a factor 2 followed by a 2 × 2 convolution. Note the difference between the up-convolution and the transposed convolution described in the previous section about autoencoders.

To train the u-net annotations on pixel level are needed. The output and the true segmentation are compared when training, calculating the loss by averaging over the individual losses (e.g. the cross-entropy loss) for all the pixels.

After the u-net was presented, multiple other neural network architectures for semantic segmentation have been developed. The networks often have a trade-off of giving high precision and model size. Two efficient designs are the DeepLab model [28] and the Pyramid Scene Parsing Network [163].

3.6 CycleGAN

The *generative adversarial network* (GAN) was first introduced in [64]. The idea is to have two models; one generative model G that generates data and one discriminative model D that predicts whether an example comes from the training data or is generated by G. Both models are trained simultaneously; G to be as good as possible at generating data that fools D and D to discriminate between the generated and the training examples.

A continuation of the idea was presented in [164]; "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks", known as *CycleGANs*. Instead of generating new data from scratch as with the GAN, the idea here is to translate images from one domain X to another domain Y. For example; images of zebras are translated such that they appear to be of horses and vice versa.

The CycleGAN consists of four parts; two generators, G_X and G_Y, and two discriminators, D_X and D_Y. The generator G_X takes an image from domain X as input and outputs an image appearing to be from domain Y. The discriminator D_Y aims at distinguish between images y_i from domain Y and images x_i from domain X that have been translated to appear to be from domain Y; $G_X(x_i)$. Furthermore, the aim is to be cycle-consistent, i.e. $G_Y(G_X(x_i))$ for $x_i \in X$ should be as close to x_i as possible and vice versa for $y_i \in Y$ and $G_X(G_Y(y_i))$. This is all illustrated in Figure 3.14. The discriminators and generators can be designed in different ways. However, only the original design used in [164] has been used in this thesis.

To train the CycleGAN, images from the two domains are needed. There are however no restriction to have paired images; e.g. images from the same scene or of a similar object. In this thesis CycleGANs have been used to translate microscopy images of prostate tissue from different sites. These CycleGANs were trained without any annotations from pathologists. However, care must be taken when applying this method; when training a CycleGAN to translate between Dataset A and Dataset C, only benign tissue was included from Dataset C. This resulted not only in a CycleGAN able to alternate the stain appearance as intended, but a CycleGAN which also translated all cancer examples from Dataset A such that they appeared benign by adding nuclei as can be seen in Figure 3.15. This problem has been

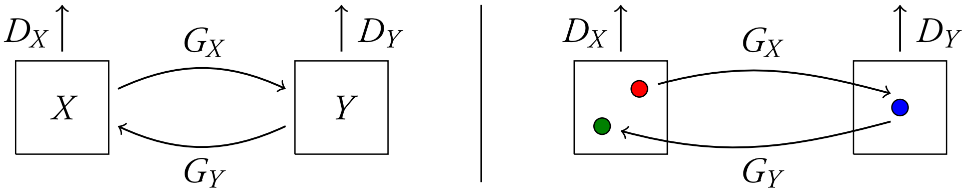


Figure 3.14: Illustration of a CycleGAN architecture with two generators, G_X and G_Y, and two discriminators, D_X and D_Y. Right: the red dot represents an example x_i, the blue dot is $G_X(x_i)$ and the green dot is $G_Y(G_X(x_i))$. To be cycle-consistent the distance between the red and green dot should be as small as possible. Image adapted from [164].

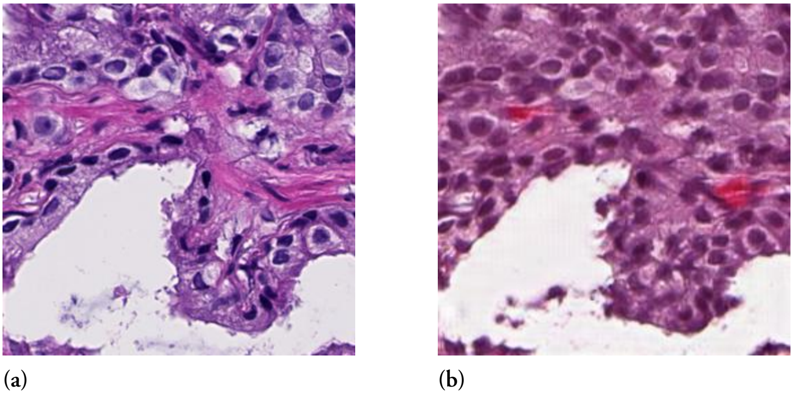


Figure 3.15: Example of an (a) image of tissue with G3 from Dataset A which in (b) is transformed to appear as it is from Dataset C. Note the nuclei added by the CycleGAN.

considered and solved by using images of both benign and malignant tissue from each site, see Section 4.3.

The example above illustrates that CycleGANs are powerful but sometimes learn unwanted behaviour. In [128] it was found that principal component analysis was better than CycleGANs if the only desired modifications were spatial shift, rotation or colour shift. For translation between the prostate sites should also small details in e.g. the stain or structure be modified, wherefore the CycleGAN was assumed to be superior. The experiments and results are presented in Section 4.3.

GANs can be used to boost training sets, by generating artificial training data. This was for example done in [133], using conditional GANs (i.e. GANs for which the class is defined when generating the data) to generate immunohistochemically stained Ki-67 images. The usage of the CycleGANs is (at least) twofold for classification of histology samples; they can be used to normalise images from different sites to look similar, a method competing with the method described in Section 3.1, and they can be used as augmentation of the training data by generating multiple variations of each example, see Section 3.8.

3.7 Domain-Adversarial Neural Network

When training a DL model, the dataset used is of great importance to make sure that the model learns relevant features of the data and that it will be able to generalise to new data. However, it is typically difficult to produce a dataset without bias towards some feature. To overcome this problem a *domain-adversarial neural network* (DANN), which was introduced in [59], can be used if the undesirable feature, the *domain*, is known.

The idea with this network design is to force the trained network to not bias certain features, the domain, which in reality should be irrelevant. This means that the goal is that some latent space of the class predictor, and all layers below that, should be independent of the domain. To achieve this, the network has two outputs; the *class y* that should be predicted as well as possible and the *domain d* which ideally is not better than random guessing when the network is fully optimised.

The DANN idea described in [59] consists of three parts: the feature extractor $G_f(\cdot, \theta_f)$ with parameters θ_f and output f; the class predictor $G_y(\cdot, \theta_y)$ with parameters θ_y and output y; and the domain predictor $G_d(\cdot, \theta_d)$ with parameters θ_d and output d. An example of such a design connecting the three components is illustrated in Figure 3.16. It has two outputs, y and d. The corresponding losses, L_y and L_d, are determined using e.g. the cross-entropy loss as

$$L_y^i(\theta_f, \theta_y) = L_y(G_y(G_f(x_i, \theta_f), \theta_y), y_i)\,, \tag{3.21}$$

$$L_d^i(\theta_f, \theta_d) = L_d(G_d(G_f(x_i, \theta_f), \theta_d), d_i)\,, \tag{3.22}$$

for an example x_i with true class y_i and domain d_i. For n examples with known classes and $\hat{n}$ examples (the same or different ones) with known domains, the training of the DANN is done by optimising

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f, \theta_y) - \alpha\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} L_d^i(\theta_f, \theta_d)\,, \tag{3.23}$$

as both a minimisation and maximisation problem. The hyperparameter α controls the the trade-off between the two losses. The solution is given by a saddle point $(\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_d)$;

$$(\hat{\theta}_f, \hat{\theta}_y) = \underset{\theta_f, \theta_y}{\arg\min}\, E(\theta_f, \theta_y, \hat{\theta}_d)\,, \tag{3.24}$$

$$\hat{\theta}_d = \underset{\theta_d}{\arg\max}\, E(\hat{\theta}_f, \hat{\theta}_y, \theta_d)\,. \tag{3.25}$$

Such a saddle point can be found by using the following gradient updates;

$$\theta_f \leftarrow \theta_f - \mu\left(\frac{\partial L_y}{\partial \theta_f} - \lambda\alpha\frac{\partial L_d}{\partial \theta_f}\right)\,, \tag{3.26}$$

$$\theta_y \leftarrow \theta_y - \mu\frac{\partial L_y}{\partial \theta_y}\,, \tag{3.27}$$

$$\theta_d \leftarrow \theta_d - \mu\alpha\frac{\partial L_d}{\partial \theta_d}\,, \tag{3.28}$$

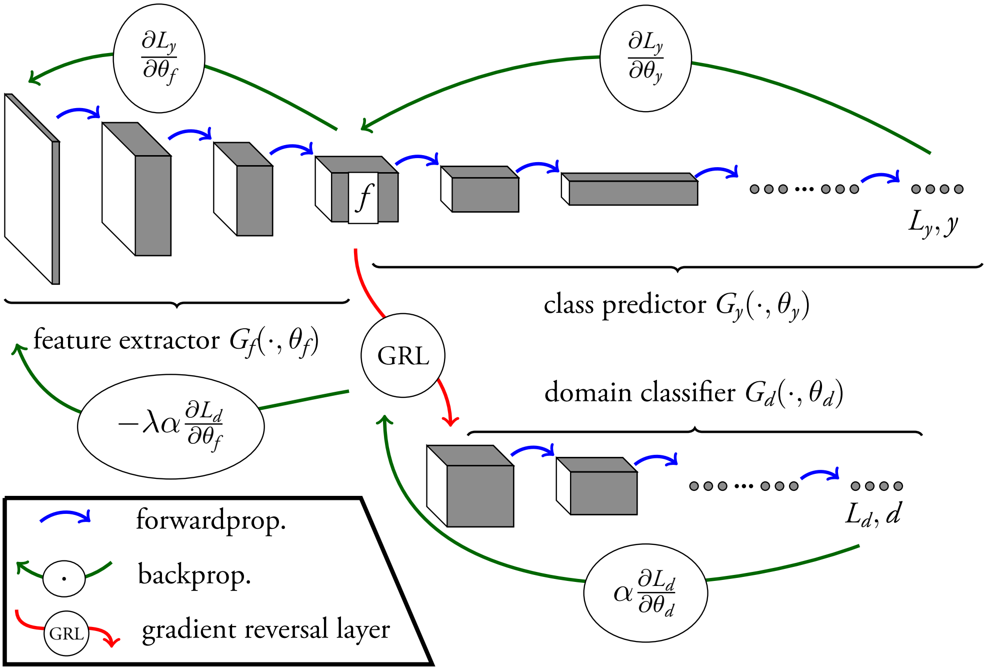

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
f
L_y, y
class predictor $G_y(\cdot, \theta_y)$
feature extractor $G_f(\cdot, \theta_f)$
GRL
$-\lambda\alpha\frac{\partial L_d}{\partial \theta_f}$
domain classifier $G_d(\cdot, \theta_d)$
L_d, d
$\alpha\frac{\partial L_d}{\partial \theta_d}$
forwardprop.
backprop.
GRL
gradient reversal layer

Figure 3.16: Illustration of a DANN architecture with one input layer, one gradient reversal layer and two output layers. Both the class y and the domain d are predicted, and the total loss is a weighted sum of L_y and L_d. Image adapted from [59].

where μ is the learning rate and λ controls the strength of the adversarial component in the feature extractor. The hyperparameter λ was not used in the first publication [59], but the combination of α and λ has previously been used in e.g. [92].

By the design of the DANN, the different weights are optimised for different tasks. The weights before the gradient reversal layer, θ_f, are optimised to perform well on the grade classification while performing badly on the domain prediction. The weights between the gradient reversal layer and the grade classification, θ_y, are only optimised for the main classification, while the weights between the gradient reversal layer and the domain classifier, θ_d, are optimised for the domain prediction. This is all done simultaneously, in an adversarial way.

The DANN approach has previously been used successfully for mitosis detection in breast cancer histopathology images [92] and magnetic resonance images of traumatic brain injuries [84]. In this thesis the approach has been used to train a network to perform Gleason grading while forcing it to not be able to determine which site the images are from. The work and results are described in Section 4.4.

3.8 Augmentation

The idea with augmentation is to artificially create more data from existing data. For DL this can be very useful, since the increased dataset and increased diversity in the dataset reduce the risk of overfitting when training. By modifying the data slightly, the DL algorithm is encouraged to not rely too much on details in the data but instead base its decision on the combination of multiple features.

Augmentation can be done in different ways; by defining random transformations manually or by training models to transform the data based on existing variations. This section will start by describing some simple augmentation methods for scalar data, thereafter continue with augmentation methods used to augment images and end by mentioning a few more complex methods which can be used for augmentation.

3.8.1 Augmentation of Scalars

Examples of suitable augmentation techniques for different types of scalars are given in the following paragraphs. Other types of scalars and augmentation techniques do of course also exist but have not been used in this thesis. The augmentation methods described here were used to augment the auxiliary parameters used to detect CAD.

Augmentation of a binary parameter, e.g. the gender, can be obtained by randomly switching the value of the parameter with some probability p %. Thus, in average the parameter has its true value in $(100 - p)$% of the occasions, and the other option at the remaining p %. This way the model cannot rely on only this parameter when making its prediction, but the model can still learn that the parameter is of importance.

For discrete, ordered parameters, e.g. values in the set $x \in \{0, 1, 2, 3\}$ such as the AHA pre-test probabilities, it is reasonable that the augmentation should take the ordering into account. Thus, the probability that $x = 0$ is altered to 3 is lower than that it is altered to 1. For example, one could limit the possible alternations to only one step or use different probabilities for different alternatives.

Finally, for continuous scalar variables the augmentation can for example be achieved by randomly adding a value in a given range. Thus, for a variable x and a chosen range $[-t, t]$, generate samples as $x' = x + t'$, where t' is drawn uniformly at random from the range $[-t, t]$. If the range of the original variable is limited, the generated value can be thresholded to not end up outside this range.

The reason for using randomised values above and not use exactly every ith iteration, is that different combinations of different augmentations should be used each time a new augmented example is generated.

3.8.2 Augmentation of Images

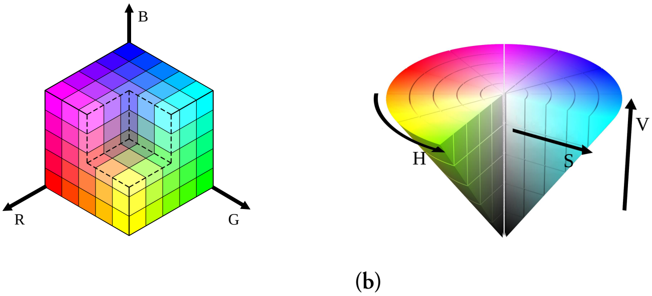
There are a number of augmentation techniques for images that can be useful depending on what the image depicts. For example, for images of cats and dogs flipping the image in vertical direction is undesirable, while for an microscopy image of tissue the orientation of the image is irrelevant. In the following paragraphs techniques for geometric, colour and other types of appearance augmentations are described.

Geometric augmentations include e.g. translation and rotation of the images. For the PCa dataset used, this has been achieved by cropping patches at random positions and with random orientations within annotated areas. Flipping the images in the horizontal and vertical direction has also been used on these datasets. Other types of geometric augmentation include zooming in or out slightly, or distortions. This has not been used, since the magnification was known and was fixed for these datasets. For the MPI dataset only rotation by a small random angle was used, since the polar maps have a fixed position and orientation in the images.

Colour augmentation of the images of H&E stained tissue was designed to mimic the variations in staining that exists, which has been shown to be useful [101]. Thus, the red, green and blue channels were not altered independently to avoid generating unrealistic examples (it is however possible that also unrealistic variations could be useful). The *hue-saturation-value* (HSV) colour space was used instead, see Figure 3.17. The conversion between RGB and HSV is defined as:

$$
\begin{aligned}
\mathrm{H} &= 60^{\mathrm{o}} \cdot \begin{cases} \text{undefined} & \text{if } \max(\mathrm{R},\mathrm{G},\mathrm{B}) = 0\,, \\ \frac{\mathrm{G}-\mathrm{B}}{\max(\mathrm{R},\mathrm{G},\mathrm{B})-\min(\mathrm{R},\mathrm{G},\mathrm{B})} & \text{if } \mathrm{R} = \max(\mathrm{R},\mathrm{G},\mathrm{B})\,, \\ 2+\frac{\mathrm{B}-\mathrm{R}}{\max(\mathrm{R},\mathrm{G},\mathrm{B})-\min(\mathrm{R},\mathrm{G},\mathrm{B})} & \text{if } \mathrm{G} = \max(\mathrm{R},\mathrm{G},\mathrm{B})\,, \\ 4+\frac{\mathrm{R}-\mathrm{G}}{\max(\mathrm{R},\mathrm{G},\mathrm{B})-\min(\mathrm{R},\mathrm{G},\mathrm{B})} & \text{if } \mathrm{B} = \max(\mathrm{R},\mathrm{G},\mathrm{B})\,, \end{cases} \\
\mathrm{S} &= \begin{cases} \frac{\max(\mathrm{R},\mathrm{G},\mathrm{B})-\min(\mathrm{R},\mathrm{G},\mathrm{B})}{\max(\mathrm{R},\mathrm{G},\mathrm{B})} & \text{if } \max(\mathrm{R},\mathrm{G},\mathrm{B}) \neq 0\,, \\ 0 & \text{otherwise}\,, \end{cases} \\
\mathrm{V} &= \max(\mathrm{R},\mathrm{G},\mathrm{B})\,.
\end{aligned}
\tag{3.29}
$$

The colour augmentation of an image RGB was generated using three parameters, h, s, v, by randomly and independently changing the H, S and V values as described in Algorithm 1. An example of how the image appearance is changed is given in Figure 3.18.



(a) (b)

Figure 3.17: (a) The RGB colour space as a cube. The two corners that are not displayed are black (origin) and white (R,G,B all max). (b) HSV colour space as a cone. Images adapted from [71] and [72].

Algorithm 1: Colour augmentation of RGB image.

```
Input: image RGB, scalars $h, s, v$
Output: image R′G′B′
RGB $\rightarrow$ HSV                        // defined by eq. (3.29)
$h' \in [1/h, h]$                            // drawn uniformly at random
$s' \in [1/s, s]$                            // drawn uniformly at random
$v' \in [1/v, v]$                            // drawn uniformly at random
$H' = H \cdot h'$
$S' = S \cdot s'$
$V' = V \cdot v'$
$H'S'V'[H'S'V' > 1] = 1$                     // to avoid undefined HSV values
$H'S'V' \rightarrow R'G'B'$                  // inverse of eq. (3.29)
```

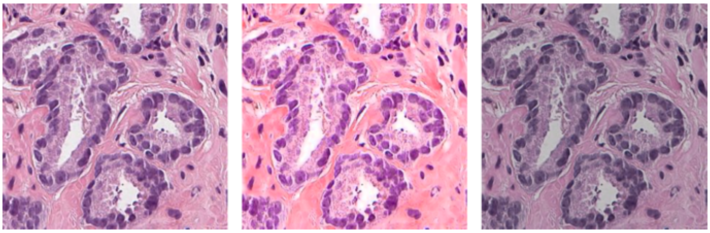


Figure 3.18: Illustration of (left) a patch and (right) two of its most extreme colour augmentations using $h = 1.04, s = 1.25, v = 1.25$.

Except augmenting the colour, the appearance of an image can be modified by blurring, clipping the intensity and adding noise. Suggestions for how this could be done is given by Algorithm 2, 3 and 4 respectively. Illustrations of some random augmentations, created by colour augmenting, blurring, intensity clipping and adding noise, can be seen in Figure 3.19.

Algorithm 2: Augmentation by blurring the image.

```
Input: image $im$, scalar $b$
Output: image $im'$
$b' \in [0, b]$                                    // drawn uniformly at random
$G_{b'}$ = 2D Gaussian filter with standard deviation $b'$
$im' = \text{conv}(im, G_{b'})$                              // convolution
$im'[im' > 1] = 1$
```

Algorithm 3: Augmentation by clipping the intensity.

```
Input: image $im$, scalars $t_1, t_2$
Output: image $im'$
$t'_1 \in [0, t_1]$                                 // drawn uniformly at random
$t'_2 \in [t_2, 1]$                                 // drawn uniformly at random
$im[im < t'_1] = t_1$
$im[im > t'_2] = t_2$
$im' = (im - t'_1)/(t'_2 - t'_1)$
```

Algorithm 4: Augmentation by adding noise.

```
Input: image $im$, scalar $n$
Output: image $im'$
$N'$ = random matrix with same size as $im$        // random values from $\mathcal{N}(0, n)$
$im' = im + N'$
$im'[im' < 0] = 0$
$im'[im' > 1] = 1$
```

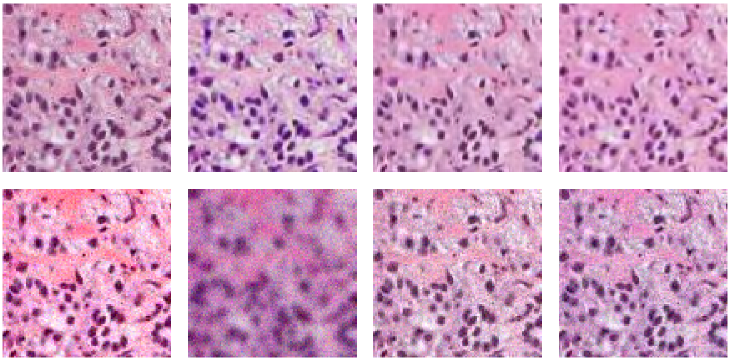


Figure 3.19: Examples of random augmentations of the top left image, using colour augmentation, blurring, intensity clipping and adding noise.

Augmentation can also be achieved by using GANs or CycleGANs, which were described in Section 3.6, to generate more data. The aim of these models is to generate images as realistic as possible. However, to expand the data outside the range of the existing training data, the methods described above might be better. It should also be pointed out that it is not obvious whether a larger dataset with unrealistic variations or a smaller dataset with realistic examples is preferable when training ANNs. However, over-extensive augmentation is expected to result in worse performance, since the neural network will be heavily biased towards characteristics that have not been augmented [136].

3.9 Rigid Registration

It is sometimes useful to digitally align pairs of microscopy images of tissue. For example, consecutive slides can be stained with different stains and by digitally aligning them the annotations from one slide, created manually or automatically such as in [155], can be transferred to the other. Another use case is for evaluation on pixel-level for the same slide but scanned with two different scanners, which has been done in this thesis (see Section 6.1). The registration is based on that corresponding key-points are found in both images, which are aligned to each other. The method, presented in Lippolis et al. [98], is described in this section.

The key-points are found using *scale invariant feature transform* (SIFT) [103]. The SIFT key-points are found as local extrema in space and scale, and are stable under certain changes in e.g. illumination. Each key-point has a descriptor, similar to a fingerprint, associated with it. The descriptors from the two images are preliminary matched in the following way. Let $\boldsymbol{x}^i = (x_1^i, x_2^i), i = 1, ..., N_1$ and $\boldsymbol{y}^i = (y_1^i, y_2^i), i = 1, ..., N_2$ be the found key-points in the two images. Construct the $N_1 \times N_2$ distance matrix $D = [d_{ij}]$, where d_{ij} is the Euclidean distance between $\boldsymbol{x}^i$ and $\boldsymbol{y}^j$. The preliminary match for each index i is then given by the points $\boldsymbol{x}^i$ and $\boldsymbol{y}^{j^*}$, where $j^* = \arg\min_j d_{ij}$, if Lowe's criterion holds:

$$\frac{\min_j d_{ij}}{\min_{j \neq j^*} d_{ij}} < 0.77 \,. \tag{3.30}$$

Thus, only matches where the nearest neighbour of $\boldsymbol{x}^i$ in the set of all $\boldsymbol{y}^j$ is much closer than the next-nearest neighbour are kept.

The images are aligned based on the matching key-points using Procrustes analysis, i.e. including translation, rotation and scaling. The transformation between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by a rigid transformation $\boldsymbol{y} = T(\boldsymbol{x})$, where

$$T(\boldsymbol{x}) = \begin{pmatrix} a & -b \\ b & a \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} + \begin{pmatrix} t_1 \\ t_2 \end{pmatrix} \,. \tag{3.31}$$

For the N matches $\{(\boldsymbol{x}^i,\boldsymbol{y}^i)\}_{i=1}^N$ the transformation is given by $\boldsymbol{y}^i = T(\boldsymbol{x}^i) + \epsilon^i, i = 1, ..., N$, where ϵ^i are small errors. The optimal transformation T is found by minimising the sum of the squared errors; $\min_T \frac{1}{2}\sum_{i=1}^N ||\epsilon^i||^2$. The minimisation is done in the following way: Collect all the parameters in a vector $\boldsymbol{z} = (a, b, t_1, t_2)$ and introduce the matrix $B(\boldsymbol{x})$:

$$B(\boldsymbol{x}) = \begin{pmatrix} \boldsymbol{x}_1 & -\boldsymbol{x}_2 & 1 & 0 \\ \boldsymbol{x}_2 & \boldsymbol{x}_1 & 0 & 1 \end{pmatrix} . \tag{3.32}$$

Then $T(\boldsymbol{x}) = B(\boldsymbol{x})\boldsymbol{z}$ and the error is given by $\epsilon^i = \boldsymbol{y}^i - B(\boldsymbol{x}^i)\boldsymbol{z}$, linear in $\boldsymbol{z}$. Introducing the variables $\boldsymbol{Y}^T = [(\boldsymbol{y}^1)^T, ..., (\boldsymbol{y}^N)^T]$ and $\boldsymbol{B}^T = [B(\boldsymbol{x}^1)^T, ..., B(\boldsymbol{x}^N)^T]$, the error-minimisation becomes a least square problem with respect to $\boldsymbol{z}$:

$$\min_{\boldsymbol{z}} \frac{1}{2}||\boldsymbol{Y} - \boldsymbol{Bz}||^2 , \tag{3.33}$$

where $||\cdot||$ is the norm in $\boldsymbol{R}^{2\times N}$. The desired $\boldsymbol{z}$ is given by the solution to $\boldsymbol{B}^T\boldsymbol{Bz} = \boldsymbol{B}^T\boldsymbol{Y}$, giving the optimal transformation. To be solvable at least two corresponding non-degenerate point pairs are needed, but the problem is more well-conditioned with at least four point correspondences.

Some of the preliminary matches found above can be *false matches*, or *outliers*. In [120], an approach to find the largest possible set of inliers was presented. However, the *RANdom SAmpling Consensus* (RANSAC) algorithm invented by Fischler and Bolles [55] was used in this thesis to avoid including the outliers when aligning the images. The idea with the RANSAC method is to estimate the model parameters using as little data as possible (randomly drawn) and evaluate the obtained parameters using the remaining data, for a fixed number of iterations. The algorithm is described in Algorithm 5, where k is the number of iterations, t is a threshold to determine if a data point fits well with the model (if it is an *inlier*) and n is the number of data points used to estimate the model parameters (in this case four were used). The optimal model parameters found using RANSAC and Procrustes analysis are used to align the images.

Algorithm 5: RANSAC.

```
while iterations < k do
    Randomly select n data points
    Estimate the model parameters using the selected data points
    Using the threshold t, count how many of the other data points that are inliers
    if number of inliers is higher than in all previous iterations then
        Save the inliers
    end
end
Optimise the model using the saved inliers.
```

3.10 Metrics

To measure the performance multiple different metrics can be used. They have different advantages and are suitable for different use cases. Some metrics, like the *accuracy*, are intuitive and easy to understand, while others are more abstract. In this section the different metrics used in this thesis are described, where all except the *agreement index* are well-known methods.

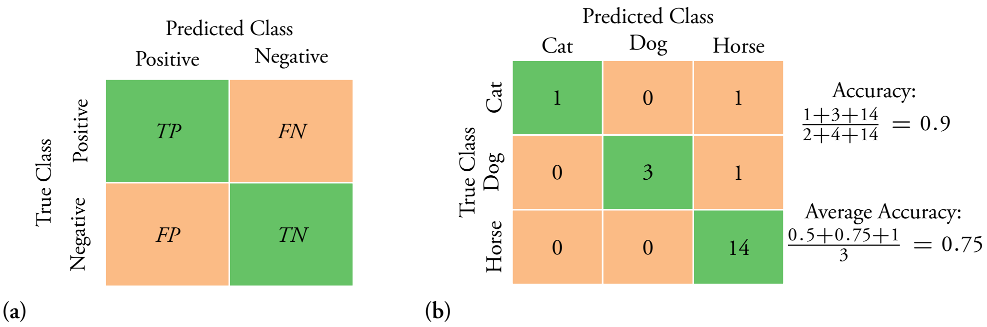
For classification, the result can be presented in a *confusion matrix*. The correct classifications will end up on the diagonal and all off-diagonal results are incorrect. For binary classification, the confusion matrix consists of *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN), as can be seen in Figure 3.20a. Based on these parameters, the *sensitivity* and *specificity* are defined as

$$\text{sensitivity} \equiv \frac{TP}{TP + FN}, \tag{3.34}$$

$$\text{specificity} \equiv \frac{TN}{TN + FP}. \tag{3.35}$$

The accuracy is given by the sum of the correct classifications divided by the total number of examples. However, if the number of examples of different classes not are the same, an *average accuracy* is preferable. It is given by the average of the accuracy for each class. An example of such a confusion matrix and corresponding accuracy are given in Figure 3.20b.

The accuracy only measures if the result is correct or not, and does not consider how certain the model was of its prediction. For binary classification with true label $y_t \in \{0, 1\}$ and an output $y_p \in [0, 1]$, the classification is typically determined by whether y_p is larger than some threshold t or not, typically $y_p > 0.5$. The closer to 0.5, the less certain is the result.



(a)

		Predicted Class	
		Positive	Negative
True Class	Positive	*TP*	*FN*
	Negative	*FP*	*TN*

(b)

		Predicted Class		
		Cat	Dog	Horse
True Class	Cat	1	0	1
	Dog	0	3	1
	Horse	0	0	14

Accuracy:
$\frac{1+3+14}{2+4+14} = 0.9$

Average Accuracy:
$\frac{0.5+0.75+1}{3} = 0.75$

Figure 3.20: (a) The confusion matrix for a binary classification problem and (b) example of a confusion matrix for three classes, with corresponding accuracy and average accuracy.

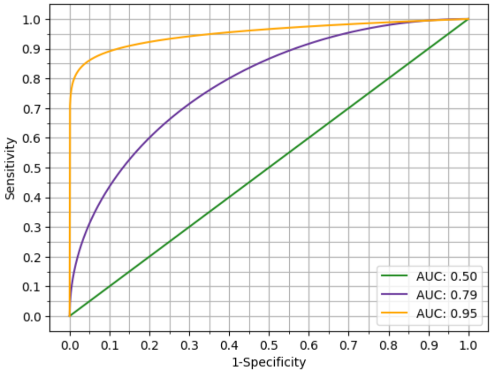


Figure 3.21: Illustration of three different ROC curves and their corresponding AUCs. The green curve is equivalent to random guessing.

To illustrate this, a *receiver operating characteristic* (ROC) curve is useful. It is constructed by varying the threshold and checking $y_p > t$, to measure the sensitivity and specificity for different thresholds. The curve is given by the sensitivity as function of one minus the specificity (1-specificity). Such a curve is illustrated in Figure 3.21. The *area under the ROC curve* (AUC) measures the results, also considering the uncertainty. A perfect ROC curve gives AUC $= 1$ and a ROC curve based on random guessing will in average give AUC $= 0.5$.

One benefit with the AUC instead of accuracy as a metric, is that the threshold does not need to be chosen. Choosing the threshold is a trade-off between false positives and false negatives, and for medical applications it is an ethical question. With a low threshold, there will be more false positive results, resulting in possibly overdiagnosis and increased work for the medical doctors. On the other hand, a high threshold will instead increase the risk that a patient with disease is overlooked.

The metrics described above are based on that a ground truth, or at least a gold standard, exists for each example. However, this is not always the case and the agreement between different raters should instead sometimes be measured. *Cohen's kappa coefficient* κ is such a metric, introduced by Cohen [34]. Instead of only measuring the agreement as an accuracy, it also takes the possibility of the agreement occurring by chance into account. The probability of chance agreement is defined as

$$p_c = \frac{1}{N^2} \sum_k n_{k1} n_{k2} \,, \tag{3.36}$$

where the number of categories is k, observations is N and times rater i predicts category k is n_{ki}. If the accuracy is p_o, Cohen's kappa is given by

$$\kappa \equiv \frac{p_o - p_c}{1 - p_c} \,. \tag{3.37}$$

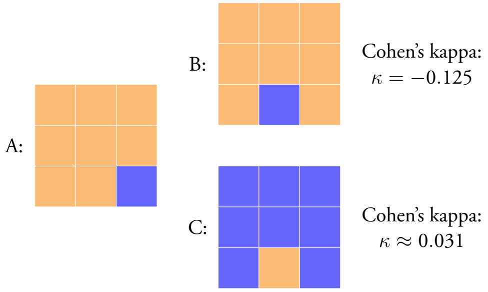


Figure 3.22: An example where the agreement between two raters according to Cohen's kappa is unexpected. For each pixel the raters (A, B, C) could choose blue or orange. Comparing A and B gives $\kappa = -0.125$, comparing A and C gives $\kappa \approx 0.031$.

The higher the value of Cohen's kappa, the better is the agreement according to this metric. There are however a few drawbacks with Cohen's kappa [124, 159]. One example is that the metric is hard to interpret. Furthermore, agreement occurring by chance is calculated based on the assumption that the rater simply guesses when uncertain, which often not is the case. An example to illustrate the unintuitive and confusing results kappa can give is shown in Figure 3.22.

Another metric to measure agreement between raters based on multiple subjects is the *intraclass correlation coefficient* (ICC). The definition depends on if e.g. the raters for each subject are chosen randomly or if the same raters are used for all subjects. Further details can be found in [108] and [89]. The complex definition gives a precise metric which however is hard to interpret.

To measure the performance for a segmentation task, other metrics are needed. One option is to use pixel-wise metrics, i.e. consider each pixel as an example and use e.g. accuracy, average accuracy or Cohen's kappa. Another option is *Intersection-over-Union* (IoU), or *Jaccard index* first introduced as coefficient of community in [81]. For segmentations X and Y, it is defined as

$$\mathrm{IoU} \equiv \frac{\text{area of overlap}}{\text{area of union}} = \frac{|X \cap Y|}{|X \cup Y|} . \tag{3.38}$$

For multiple classes, the IoU is given by the average of the IoU of each class. A third option is the *Dice coefficient* [40], defined by

$$\mathrm{D} \equiv \frac{2|X \cap Y|}{|X| + |Y|} . \tag{3.39}$$

The Dice coefficient is very similar to IoU, with the relation

$$\mathrm{D} = \frac{2\mathrm{IoU}}{1 + \mathrm{IoU}} . \tag{3.40}$$

One difference is however that the Dice coefficient is not a proper metric, since it does not satisfy the triangle inequality. Both IoU and the Dice coefficient provide a balanced measure of the accuracy of each class, independent of their prevalence in the data — but they do lack intuitive understanding.

One problem with the metrics mentioned above, is that the ordering of classes is not considered, which sometimes is desired. For ordered classes, it could be desirable that the size of the error is reflected by the metric, by e.g. weighting different errors differently. For the IoU and Dice coefficient there is no straight forward way for such weighting. For Cohen's kappa, the errors could be weighted but it has other limitations as discussed above. Therefore, a new metric was constructed, the *agreement index* introduced in Paper VII [107], to measure the agreement between different raters for Gleason grading. This metric was used pixel-wise for segmentation as described below.

Each pixel was checked for overlap in the result from the two raters and for a correct overlap a score of 3 was given. A penalty of -1, -2 or -3 is given to each incorrect pixel and each incorrect class (B vs G3 vs G4 vs G5). For example, if a pixel was classified as G3 by one rater and G4 by another rater, the penalty would be –1, but a classification of G3 vs G5 would generate a penalty of –2. The agreement index is computed for each individual slide as the sum of the pixel scores and normalised such that a perfect overlap gives an index of 1. Thus, let n denote the number of pixels with tissue, x_i the first rater's prediction (from an algorithm or pathologist) for pixel i and y_i the corresponding label by the second rater, where the prediction is coded as benign=0, G3=1, G4=2, G5=3. The agreement index is then given by

$$d = \sum_i \frac{3 - |x_i - y_i|}{3n}, \tag{3.41}$$

where a perfect overlap gives $d = 1$ and no overlap gives $d = 0$.

Chapter 4

Generalisation of Gleason Grading ANNs

In this chapter, methods for automatic detection of PCa and prediction of Gleason grade using DL are described. The automatic approach has the potential to increase the accuracy and efficiency of PCa diagnosis, which would be of uttermost use. The chapter is based on Paper I, III, V and VI.

With the introduction of digital pathology, image analysis can be used to facilitate pathologists in their daily work. As for many other image analysis tasks, DL and especially CNNs have turned out to be very successful for detection of cancer from digitised prostate biopsy samples. In the last years multiple papers have presented promising suggestions towards this goal. For example, [99] successfully separated benign and malignant tissue and was able to automatically exclude slides containing only benign tissue. In [101] the false negative rate was reduced to a quarter of a pathologist's for the task of detecting tumours in breast cancer. For Gleason grading, promising results were early presented in [66] and [37].

One drawback with this technique is that the algorithms tend to overfit to the training data and not generalise to data from other sources, for example different hospitals, due to for example inevitable stain variations. To overcome this problem, many different approaches have been suggested such as using digital stain separation and normalisation, using CycleGANs for normalisation, using CycleGANs for augmentation or training with colour augmentation. While the idea with the first two approaches is to achieve normalisation of the data, the idea with the last ones instead is to force the algorithm to adapt to more variations, possibly outside the range of the training data. Another approach is to use a DANN to force the network to neglect variations between hospitals. All these methods have been evaluated and are presented in the following sections.

4.1 First Gleason Grading Algorithm

Digital pathology has grown during the last years. This work, published 2017 (Paper I, [66]) based on the master's thesis by Gummeson [65], was one of the first publications presenting a DL solution for automatic prediction of Gleason grade in PCa. Previous studies which investigate automatic PCa classification include [83, 97, 99].

It is hard to train a network to recognise a pattern it has not seen before. Consequently, it is crucial that a large dataset covering as many variations as possible is available to achieve a well-trained network. Due to the difficulties of obtaining data and the need for at least one expert pathologist to annotate the data, the dataset used in this study was rather limited.

Two of the previous studies, [83, 97], are especially relevant since they used the same dataset as this work and it is thus possible to compare their result with this study's result in a fair way. The method proposed by Ref. [97] uses SIFT as descriptor of local texture, which are then clustered using a bag of words approach. The histogram of the frequency of the clusters are used as a descriptor for classifying the whole image. In Ref. [83] a pre-trained CNN is used for feature extraction and either support vector machines or random forest are used for classification.

Ref. [99] presents a classifier for benign versus malignant tissue. They have trained a CNN from scratch, similar to the one presented in this work but with significantly more data. For each pixel they get the likelihood that the pixel is either benign or malignant. Using these results for a whole slide they were able to determine which slides that contained only benign tissue and thus not had to be diagnosed by a pathologist. They were able to automatically exclude 30-40% of the slides which only contained benign tissue. Compared to their study, this work aimed for classification into four classes — Benign, G3, G4 and G5 — but has not investigated the results for whole slides.

4.1.1 Dataset

Dataset 0 was used in this study. For DL the dataset is rather small. Since each image is of one class only, and the orientation of the image is irrelevant, multiple patches were created from each image to increase the amount of training data. In total more than 12 000 patches were created from each class. The same number of patches from each class was used, to not promote any class more than the others.

The images were divided into four parts and 4-fold cross-validation was used. The images differed substantially in size and five images were too small to go through the CNN and were thus rejected. To not prefer any class during training the same number of patches of each class was created, even though the number of images of each class varied as well as the size of each image. The images were first low-pass filtered and down-sampled in both dimensions, from 40X to 6X magnification. The down-sampling was done to cover a larger spatial area, which is beneficial since the Gleason grade cannot be determined from too small areas, without needing a too large CNN which would increase the requirements of the computer. The chosen resolution was considered good enough to still be able to determine the Gleason grade.

4.1.2 Method

The CNN architecture is illustrated in Figure 4.1. The input consisted of an RGB image with three channels and size 106 × 106 pixels. The sizes of the convolutional kernels were either 3 × 3 or 4 × 4, followed by ReLU activation functions. The max pooling layers were used with 2 × 2 filters and stride 2. The last two layers were fully connected, where no spatial information was left. The output of the network used the softmax function and consisted of four channels, one each for the four classes (Benign, G3, G4, G5). Dropout was as regularisation used before the last two layers. The network was trained from scratch using stochastic gradient descent with momentum and the cross-entropy loss. The learning rate was decreased for each iteration in the training, in an attempt to reach a good local minimum. The networks were trained for 100 epochs, and the average results of the last 25 epochs were used. The CNN was implemented using MatConvNet [161]. Training the four networks needed to do 4-fold cross-validation with 100 epochs took 58 hours.

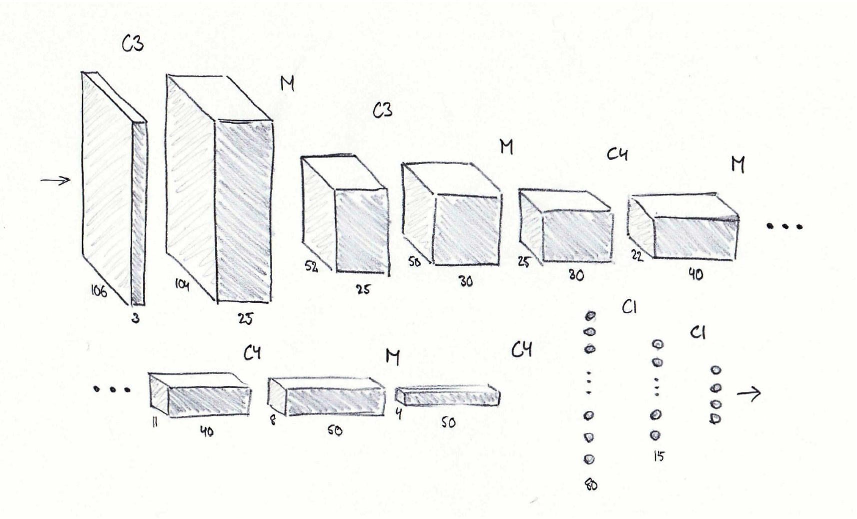


Figure 4.1: The design of the convolutional network which is trained for Gleason grading. It takes an RGB image of size 106 × 106 × 3 as input and returns a 1 × 1 × 4 vector as output. C3 means a convolutional layer with a 3 × 3 kernel, M is a max pooling layer with 2 × 2 filters and stride 2. The last two layers are fully connected and use the dropout regularisation technique.

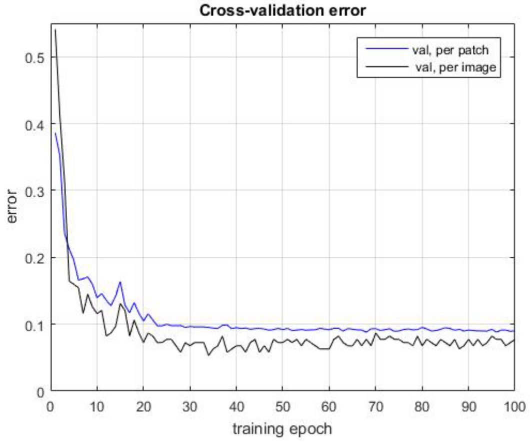


Figure 4.2: Validation error as function of epoch iteration. The final result is 92.7% correct on per image level, when averaging over the last 25 epochs.

4.1.3 Results

Figure 4.2 shows the mean of the validation error as function of iteration, both per patch in blue and per image in black. To estimate the total per image error, the mean of the last 25 epochs was used. This gave a final error rate of 7.3%. The confusion matrix, see Figure 4.3, shows the classification errors that occur when looking at the classification of each whole image.

In Figure 4.4 an example of a misclassified image is given, together with segmentation maps showing the classifications using the four different CNNs. The left result is achieved when the image was used as validation image, while it was used as training image in the three other cases. The image's true label is G4, which is illustrated by green. Red means G3.

		Predicted class			
		B	G3	G4	G5
True class	B	51	0	0	0
	G3	3	46	3	1
	G4	0	6	43	0
	G5	0	3	0	52

Figure 4.3: Confusion matrix, illustrating the classification errors. Benign is denoted B and the Gleason grades are denoted G3, G4 and G5.

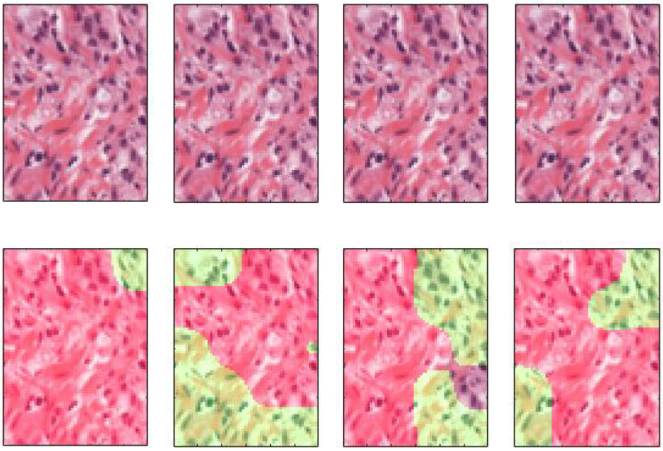


Figure 4.4: Example of an image (upper row) and segmentation maps (lower row) showing the classifications from the four different nets from the different cross-validation simulations. The left result is obtained when the image is used as validation image, while it is used as training image for the others. The true label is G4, which means that the classification ought to be green. The resolution of the images is the same as used when training the CNN.

4.1.4 Discussion

The error rate of 7.3% is an improvement to previous state-of-the-art methods using the same dataset. Ref. [97] achieved an error rate of 12.7% and Ref. [83] achieved an error rate of 10.2%.

Based on the confusion matrix, see Figure 4.3, it can be seen that all images with the true label Benign are correctly classified, while G3 seems to be hardest to classify, with misclassification to all other classes. It is probable that some of the misclassifications are worse than others, e.g. it is worse to classify G5 as Benign than as G4. During training these misclassifications have been equally penalised, but to reduce the risk of these large errors it might be beneficial with different penalisation for different errors.

Different parts of the misclassified image shown in Figure 4.4 are misclassified with different CNNs. This indicates that a larger dataset for training would make the classification more robust.

The input size, 106 × 106 pixels, was chosen since this was considered large enough to see the context which is needed to determine the Gleason grade. However, some cancerous images in the dataset were smaller than this and thus could not be classified with the CNN. Due to the large variations in e.g. size and structure of different cancerous areas, it is hard to design a CNN which is optimal for all types. Since only the images were available and not the surrounding tissue, it was not possible to determine whether the CNN was able to detect the smaller areas too.

4.2 Generalisation using Normalisation and Autoencoders

The issue that algorithms trained on images from only one site might not generalise to other sites, due to for example inevitable stain variations and differences in the equipment, was highlighted in this work (Paper III, [6]). Neglecting this aspect can potentially lead to overestimated results and incorrect conclusions.

The purpose of this work was to investigate the performance and ability of generalisation of different preprocessing and classification techniques. Two different CNNs were used for classification, using images in 20X and 5X magnification respectively. For the latter, three different preprocessing approaches were compared; (i) regular downsampling, (ii) regular downsampling, digital stain separation and normalisation, and (iii) using the encoding part of an autoencoder for downsampling. The aim of using an autoencoder was to get a more efficient downsampling and extracting more information than the regular one does. Finally, it was also investigated to what extent colour augmentation of the training data improved the performance and ability to generalise. The evaluation was done using three datasets from different origins, which have been annotated by the same two pathologists to avoid inter-observer variations.

4.2.1 Dataset

Part of Dataset A, C and D were used in this work, cf. Subsection 2.1.2. An overview of the amount of data is given in Table 4.1. Since the networks used required a minimum size of 312×312 pixels in 20X magnification, only the annotated areas which were at least this large were included in the datasets. Dataset A, the largest of the three, was used for training and validation of the networks, where three quarters of the annotated areas of each class were used for training and the rest for validation. Dataset C and D were used for testing and were thus not used to tune the networks during training.

Patches were created from the training images to train the networks. Spatial data augmentation was used to expand the amount of data, using rotation and flipping. The same number of patches was extracted for each class, to not promote any of them. The same patches were used to train all the different networks, although colour augmentation was performed for some of the trainings as described in Section 4.2.2.

The validation and testing were performed on 312×312 pixel patches, i.e. for each patch the networks suggested one class only. Since the classification should be invariant to rotation and flipping each patch was tested 8 times individually; once for each 90° rotation and for their mirrored versions. The same patches were used for validation and testing of all different methods.

Table 4.1: Details of the datasets used for training and validation (A) and testing (C and D). (©2018 IEEE)

Dataset	Number of slides	Nbr of annotated areas				Nbr of patches per class for training	Nbr of patches per class for validation or test
		Benign	G3	G4	G5		
A	88	1527	343	411	113	76320	590
C	15	48	9	171	17	-	600
D	24	24	165	106	3	-	49

4.2.2 Method

Six different classification CNNs were used, using different resolutions and preprocessing techniques. Descriptions of the different components follow.

As preprocessing, normalisation of the stains was used. To normalise the two stains individually, the method and implementation provided by [160] and described in Section 3.1, which is based on sparse non-negative matrix factorisation, was used. As input to the classifier the two stain channels were used, see Figure 3.3 on page 32.

An autoencoder was trained for efficient downsampling of the images instead of using regular downsampling. Since the images should be downsampled from 20X to 5X, an autoencoder with two max pooling layers and zero padding for the convolutions was used. The design is shown in Figure 4.5. All convolutions were followed by ReLU activation functions, except the C1 convolution in the bottom of the network where softmax was used and the last C3 convolution where the hyperbolic tangent was used. The softmax was chosen to try forcing the network to choose one of the channels and thus ideally getting the different channels to represent different types of the tissue. Since the input images were normalised to $[-1, 1]$ the hyperbolic tangent was used as the last activation function. The first half of the autoencoder, including the C1 layer, was used to downsample the data as preprocessing. Illustration of a three channel encoded result as an RGB image is shown in Figure 4.6.

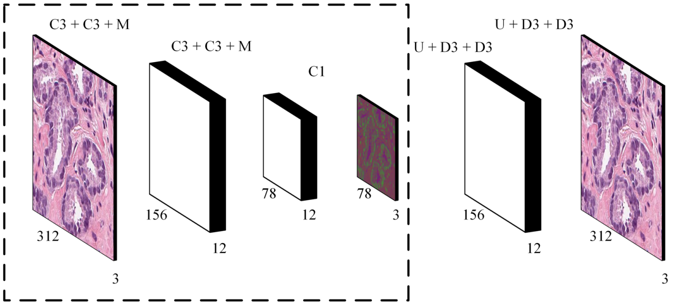


Figure 4.5: The design of the autoencoder, where the part in the dashed rectangle is used for downsampling of the data. C3 and C1 means convolution with square filters with sizes 3 and 1, M and U means 2 × 2 max pooling and up-sampling, and D3 means transposed convolution [43] using filters with size 3. (©2018 IEEE)

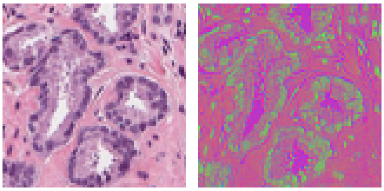


Figure 4.6: Example of an input image and corresponding result when the image is encoded with the autoencoder. The intensity of the result is increased, for a more clear visualisation. (©2018 IEEE)

The two different CNNs used for classification, illustrated in Figure 4.7, have the difference that the 20X classifier has two extra layers of convolutions (using zero padding) and max pooling layers. The 5X network was inspired by the networks presented in [66, 99], using a similar structure and input resolution. ReLU was used as activation function after each convolution except the last where softmax was used.

The networks were implemented in Keras [30] and trained using the Adam optimiser with default parameters [87]. The cross-entropy loss was used for the classifier and mean squared error for the autoencoder. They were both trained with a batch size of 32. When training the classifiers, 50% dropout was used between the three final fully connected layers. All input images were, after the preprocessing, normalised to $[-1, 1]$ by multiplying by 2 and subtracting 1.

In addition to spatial augmentation the colours of the images were also augmented as described in Section 3.8, with parameters $h = 1.04, s = 1.25, v = 1.25$, see Figure 3.18 on page 52. The autoencoder was trained using colour augmentation, while only some of the classifiers were trained with colour augmentation.

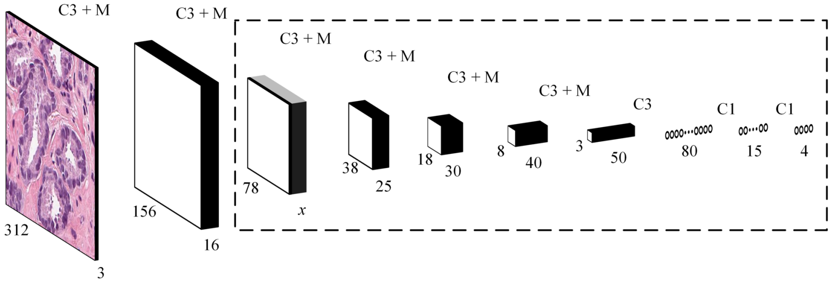


Figure 4.7: The design of the CNN classifiers, where inputs in 20X uses the whole network and inputs in 5X uses the part in the dashed rectangle. C3 and C1 means 3×3 and 1×1 convolution, M means 2×2 max pooling with stride 2. The number of features x in the third layer depends on the input type; $x = 16$ for 20X, $x = 3$ for 5X using regular downsampling or the encoder and $x = 2$ for digital stain separation. (©2018 IEEE)

Table 4.2: Results on the validation and test datasets, where the first percentage gives the accuracy for Gleason grading and the one in parenthesis the accuracy for benign versus malignant. (©2018 IEEE)

Input size &	Validation accuracy (%)	Test accuracy (%)	
preprocessing	Dataset A	Dataset C	Dataset D
20X – colour augmentation	**81 (95)**	46 (76)	50 (92)
5X	73 (94)	46 (80)	45 (88)
5X – colour augmentation	79 (95)	42 (81)	49 (92)
5X – stain normalisation	78 (94)	48 (79)	**53** (89)
5X – autoencoder	75 (92)	50 (83)	47 (90)
5X – autoencoder, colour augmentation	75 (93)	**53 (83)**	50 **(94)**

4.2.3 Results

An overview of the accuracy for the different methods as well as the different validation and test datasets is given in Table 4.2. Accuracy for both Gleason prediction as well as prediction of benign versus malignant are given.

4.2.4 Discussion

All methods generalised quite well for benign versus malignant, cf. Table 4.2, while the results for Gleason grading were significantly worse for the test datasets. Since the datasets are annotated by the same two pathologists, there ought to be no difference due to inconsistency between pathologists. However, it could be due to inconsistency within these pathologists.

Approximately the same accuracy was obtained with the different methods. The highest for the validation set was achieved with the network using the highest input resolution, suggesting that some important information was lost when downsampling to 5X. The best results for the test datasets were obtained when the autoencoder was used for downsampling and the classifier was trained using colour augmentation. Thus the autoencoder seems to make a more clever downsampling than the regular downsampling. The use of colour augmentation increased the CNN's ability to generalise, even when regular downsampling was used. The improvement when using the digital stain separation and normalisation was similar to when colour augmentation was used.

4.3 Normalisation and Augmentation using CycleGANs

As described in the previous section, limited generalisation performance is a fundamental problem when using DL applied to digital pathology. In this work, more approaches for improving the generalisation performance of a CNN trained for Gleason grading were evaluated. The main ideas were the same — to either normalise data from different sites to each other or to create artificial data to train on — but the techniques were different. This work was originally published in Paper V [7].

The first idea, to normalise data from different sites, has been used multiple times, for example in [6, 14]. Using this approach, the generalisation performance has been slightly improved but better results are needed. In this paper another technique is investigated; training a CycleGAN to translate images from one site to another.

For the second approach, to ensure sufficiently large variation in the training data, the most obvious way would be to include data from as many sites as possible. But this is seldom possible in practice since data, especially annotated data, is not available. Artificial approaches to expand the existing dataset are consequently preferable. The use of augmentation techniques is standard when training neural networks, but which techniques that are used vary. In this work rotation, flipping, colour augmentation, addition of noise, intensity clipping and blurring were performed. That these different augmentation techniques improve the generalisation performance is shown in [154], where they found that the more augmentation used the better performance although colour augmentation seemed to be the most important one. They performed the colour augmentation on each stain individually after a colour deconvolution, an approach which was not used in this work. The reason for this was to avoid the colour deconvolution which not always was successful, probably due to weak stains and the fact that both stains had approximately the same colour. Instead, the augmentation was carried out on each channel individually after transforming to the HSV colour space. Using CycleGANs to augment the existing training data was also tested.

4.3.1 Dataset

In this work, the entire Datasets A, B, C and D were used, cf. Subsection 2.1.2. Dataset A was used for training and validation, and the remaining datasets were used for testing. To train the CycleGANs, all datasets were used but only the images and not the labels. Therefore, the datasets still could be used as test sets. From the annotated areas, 256×256 pixel patches in 10X magnification were extracted. The number of patches for each of the datasets can be seen in Table 4.3, and examples of tissue from the different sites can be seen in Figure 2.5 on page 22. For the training dataset, rotations and flipping were performed before the patches were generated, thus more patches were extracted from each annotation compared to the validation and test sets where this type of augmentation was not performed.

Table 4.3: Details of the datasets used for training and validation (A) and testing (B, C and D). (©2019 IEEE)

Dataset	Nbr of slides	Nbr of annotated areas				Nbr of extracted patches			
		Benign	G3	G4	G5	Benign	G3	G4	G5
A train	109	1902	545	637	268	348504	224808	315312	186672
A val.		633	181	212	89	9679	1396	2249	1115
B test	55	117	48	52	9	3628	879	487	42
C test	16	54	9	189	17	1635	296	2087	335
D test	50	29	304	209	8	977	1909	1103	47

4.3.2 Method

The goal was to construct an algorithm which generalises well to different sites. Therefore, only Dataset A was used for training and validation, and the three other datasets were used for testing. The method described here was thus designed to optimise the performance on the test datasets, while the classification network only was trained and optimised using Dataset A.

CycleGANs were trained to translate between Dataset A and each test dataset as described in Section 3.6. Since no annotations are needed to train the networks, other datasets can relatively easy be added in the future. Furthermore, compared with other types of digital stain normalisation techniques often used, this approach has the advantage of not only being able to change the colour, but also smaller structures in the images such as the stroma texture thanks to that spatial information also is considered.

Images from all Gleason grades were included when training the CycleGANs, to make sure that the translation did not modify this feature of an image. Examples of translations between Dataset A and Dataset B can be seen in Figure 4.8.

The CycleGANs were used in two different ways. To start with, they were used to normalise the test images to Dataset A, i.e. the classifier was applied to the test images after they had been transformed to appear as Dataset A. Secondly, the CycleGANs were used in the opposite direction to increase the Dataset A by a factor four. Thus for each image from Dataset A, three other images were constructed by transforming the image such that it appeared to belong to Dataset B, C and D respectively. An example is shown in the top row in Figure 4.9. This extended dataset was used as an augmented expansion of Dataset A.

Except the standard augmentation techniques used, such as rotating and flipping the images, some additional techniques were also used in this work; namely blurring the images, clipping the intensity, adding noise, augmenting the colours and, as stated above, transforming the images using the trained CycleGANs. The details of the augmentation techniques are given in Section 3.8. The following parameters were used: $b = 2, t_1 = 0.1, t_2 = 0.9, n = 0.05, h = 1.04, s = 1.25, v = 1.25$. Examples of augmented versions can be seen in Figure 4.9.

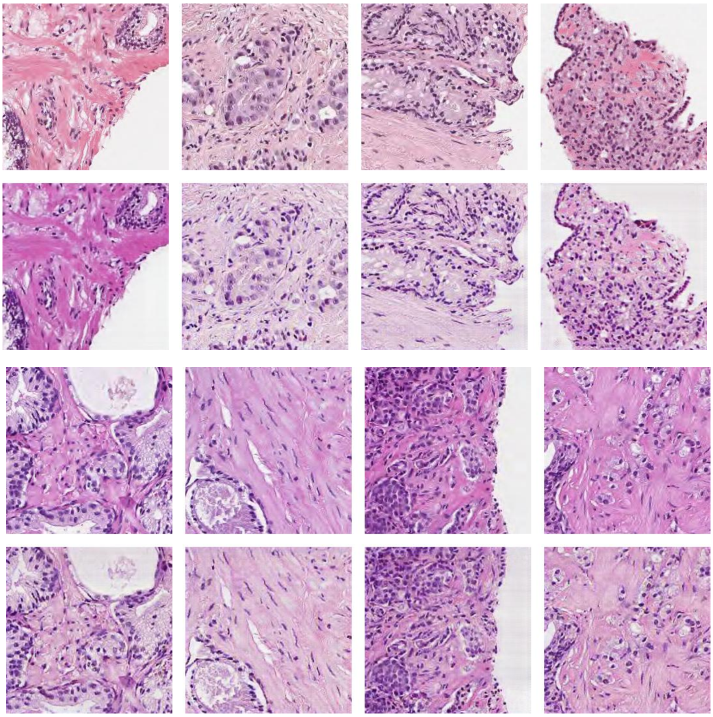


Figure 4.8: Images from Dataset A transformed to Dataset B (top rows) and images from Dataset B transformed to Dataset A (bottom rows), using a CycleGAN. (©2019 IEEE)

Each image was randomly augmented for each epoch, thus new variations were generated while training the network. To avoid that the augmentation techniques removed the effects of each other, for example that the blurring removed the added noise, the augmentations were applied in the following order: blurring, intensity clipping, noise addition, colour augmentation.

The CNN's design was inspired by the neural networks presented in [6, 66, 99]. The design can be seen in Figure 4.10, where each convolution is followed by a ReLU activation function except the last one where softmax was used. Dropout with 50% was used in the two C1 layers. The networks were implemented in Keras [30], trained to minimise the cross-entropy loss using the Adam optimiser [87] and the batch size 64. The loss for the classifier was weighted to give each class the same importance, despite the imbalanced classes.

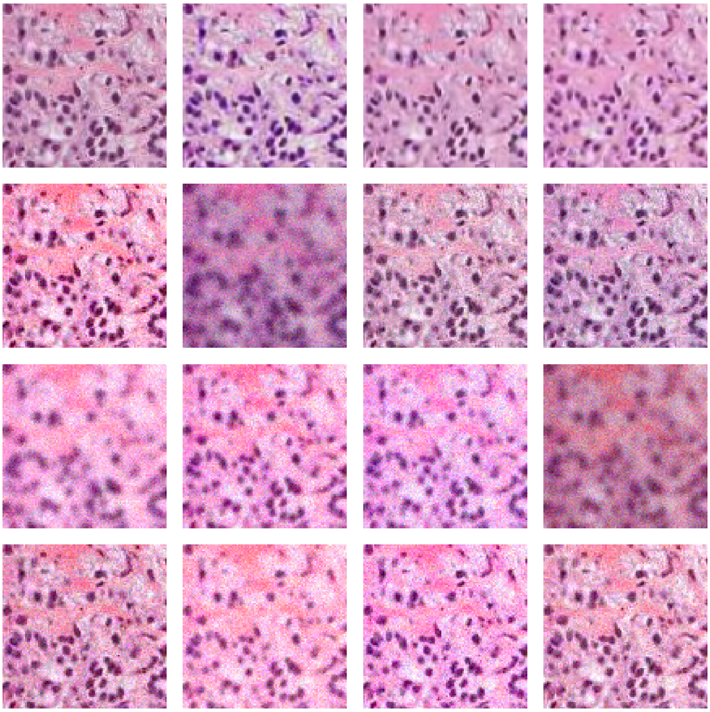


Figure 4.9: Top row: original image and the three images generated by transforming to site B, C and D by the CycleGANs. Remaining rows: examples of generated images combining all the augmentation techniques. (©2019 IEEE)

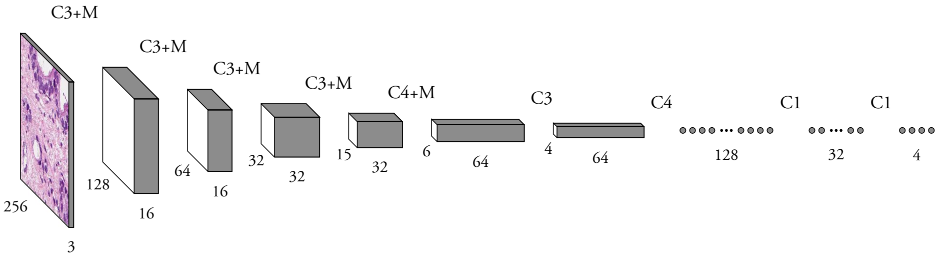


Figure 4.10: The design of the CNN classifier. The spatial size and the number of channels for each layer can be seen in the figure. Here CX denotes a convolution with an $X \times X$ filter, followed by a ReLU activation function, and M denotes a 2×2 max pooling layer with stride 2. After the final C1 layer (i.e. fully connected), softmax was used.

4.3.3 Results

The performances obtained when using the different augmentation techniques individually and combined can be seen in Table 4.4, giving the percentage of correctly classified patches both when no normalisation between the sites was used as well as when CycleGANs were used to normalise the validation datasets to the training site. The augmentation techniques blurring, intensity clipping and addition of noise were only used together. For all of the approaches, approximately the same performance was obtained for the test datasets.

Table 4.4: Average accuracy for Gleason grading of small patches when using different augmentation techniques, where aug includes both blurring, intensity clipping and addition of noise. Colour means that colour augmentation was used and CycleGAN means that the training dataset was increased a factor four by transforming the training data to the test sites. Results in parentheses are the result when first transforming the test images to Dataset A using CycleGANs. (©2019 IEEE)

Augmentation	Validation avg. acc. (%)	Test avg. acc. (%)		
technique	Dataset A	Dataset B	Dataset C	Dataset D
Nothing	77	46 (48)	57 (58)	50 (55)
Colour	77	55 (51)	57 (**62**)	**59** (56)
Colour, aug	77	**59** (57)	**59** (61)	56 (**60**)
CycleGAN	74	44 (44)	52 (49)	48 (51)
CycleGAN, colour	76	55 (53)	57 (55)	50 (56)
CycleGAN, colour, aug	74	42 (50)	43 (53)	47 (51)

4.3.4 Discussion

It is unclear whether normalisation improved the generalisation performance or not. Although the appearances of the training and validation images were more similar after the transformation, the classifier did not always perform better. Using CycleGANs as augmentation did not improve the generalisation performance either and furthermore lowered the accuracy for the validation dataset. On the other hand, colour augmentation improved the performance. Adding the other augmentation techniques, the performance was increased even more. Thus, using less realistic augmentation techniques actually performed better than if more realistic variations were generated by a CycleGAN.

Due to the differences between datasets, it is hard to compare the results to state of the art. For example, in [116] a higher accuracy was achieved, but for the simpler task of only distinguish between low- and high-grade cancer in larger areas. Compared to Paper III [6], where subsets of the datasets used in this study were used, the new results are superior.

One limitation with this study is that the images in the test datasets were used to train the CycleGANs. This could potentially result in CycleGANs which are overfitted to these images. Ideally, the test datasets should have been divided into one part used to train the CycleGANs and one to test the classifier. However, due to the limited sizes of these datasets this was not an appealing approach.

4.4 DANN for Improved Generalisation

Continuing the work in the previous two sections, a new approach to how generalisation performance for Gleason grading can be improved is presented in this section. The method is based on DANNs, and was published in Paper VI [8]. The idea is to train the classification CNN in such a way that it is encouraged to generalise well.

It is in general difficult to produce a dataset without some bias towards any specific feature, and when training DL algorithms the biases in the dataset will be transferred into the network. DL models used in histopathology tend to overfit to the stain appearance of the training data — if the model is trained on data from one lab only, it will usually not be able to generalise to data from other labs. The standard techniques to overcome this problem are to either use colour augmentation of the training data which, artificially, generates more variations for the network to learn, or to normalise the data to appear more similar before the classification is done. A DANN was instead used in this work. It is designed to prevent the model from being biased towards features that in reality are irrelevant such as the origin of an image. To test the technique, four datasets from different hospitals ("domains") for Gleason grading of PCa were used. While data from different sites were needed, there was no need for all of them to be annotated with Gleason grade belonging to train the network.

4.4.1 Dataset

The data used was the same as in the previous section, cf. Subsection 2.1.2, consisting of Dataset A, B, C and D from four different hospitals. The division into training, validation and test was done in the same way. However, the test datasets were also used when training, but only the images and not the labels. Therefore, the datasets could still be used for testing of the classifier. The numbers of whole slides, annotations and extracted patches are given in Table 4.3 and examples of tissue from the different sites can be seen in Figure 2.5 on page 22.

4.4.2 Method

The design of the CNN can be seen in Figure 4.11, with one output for Gleason grading and one for domain prediction. The outputs and cross-entropy losses were denoted y and L_y for the Gleason grade classification and d and L_d for the domain part. For the first three convolutional layers, zero padding was used to keep the spatial size while for the remaining layers this was not used. The number of layers before the gradient reversal layer was varied between one and three, where the design with two layers is illustrated in Figure 4.11. When the classifiers only had one layer in common, one more block with a 3×3 convolutional layer with eight channels and a max pooling layer was added in the domain classifier, while one layer instead was removed when three common layers were used.

A hyperparameter λ determined how much of the gradient that was backpropagated through the gradient reversal layer. The loss for the network was a weighted sum, $L_y + \alpha L_d$, of the losses for the two tasks. These parameters were varied in the range $\lambda \in [0.05, 0.1]$ and $\alpha \in [0.4, 0.6]$.

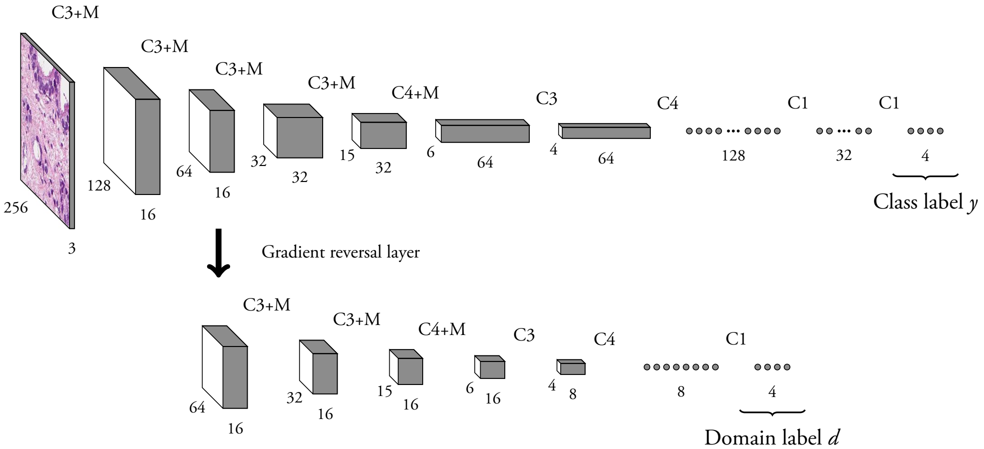


Figure 4.11: The design of the DANN, trained with two outputs to optimise the domain adaption of the Gleason grade classifier. The spatial size and the number of channels for each layer can be seen in the figure. Here CX denotes a convolution with an $X \times X$ filter, followed by a ReLU activation function, and M denotes a 2 × 2 max pooling layer with stride 2. After the final C1 layer (i.e. fully connected), softmax was used.

The network was trained with data from all four datasets. However for the data from the test sites the Gleason grades were not used and the loss L_y for these examples was put to zero. Thus, these examples did only affect the domain classifier and the common layers for the two classifiers. To check that the generalisation performance was not only improved for the specific datasets used when training, even if they were not used to train the Gleason classifier, the DANN was also trained including only two of the three test datasets. One would expect that the more datasets that are included when training the DANN, the better will the network generalise to new datasets.

The CNN was implemented in Keras and trained for 15 epochs, batch size 32, with the Adam optimiser with default parameters [30, 87]. Dropout with 50% was used in the three C1 layers. As augmentation, blurring, intensity clipping, adding noise and colour augmentation was randomly used as described in Section 3.8, with parameters $b = 2, t_1 = 0.1, t_2 = 0.9, n = 0.05, h = 1.04, s = 1.25, v = 1.25$. The network was trained four times from scratch for each set of hyperparameters, to get an estimate of the standard deviation of the performance. Also, the classification part of the CNN was trained without the domain part for comparison, both with and without colour augmentation. When training, the contribution to the loss of each example was weighted against how many examples of that particular class and domain that existed, to not bias any class or domain when training, in the following way:

Let N_A, N_B, N_C, N_D be the number of examples from each site and let $N_A^1, ..., N_A^4$ denote the number of examples from each class from site A ($N_A = \sum_{x=1}^{4} N_A^x$). For each example from site A with grade x, assign a weight for the Gleason classification

$$w_x = \frac{N_A^1}{N_A^x} \,. \tag{4.1}$$

Examples from the other sites will have weight 0, since they are not used to train the Gleason classifier. The total weight for all examples from grade x is $\sum_{i=1}^{N_A^x} w_x = N_A^1$ and the total grade weight for all grades and examples is

$$\sum_{x=1}^{4} \sum_{i=1}^{N_A^x} w_x = 4N_A^1 \,. \tag{4.2}$$

The domain weights for examples from domain d are given by

$$\hat{w}_d = \frac{N_A^1}{N_d} \,. \tag{4.3}$$

The total domain weight for all examples from domain d is $\sum_{i=1}^{N_d} \hat{w}_d = N_A^1$ and the total weight for all examples is

$$\sum_{d \in \{A,B,C,D\}} \sum_{i=1}^{N_d} \hat{w}_d = 4N_A^1 \,, \tag{4.4}$$

giving each Gleason grade and each domain the same importance while training. Further weighting to give the Gleason and domain classifiers different importance is achieved by varying the α and λ parameters.

4.4.3 Results

Multiple designs and hyperparameters for the DANN were tested to find the best configuration, see Table 4.5. The overall best performances for the test datasets were achieved with two common layers for the Gleason and domain adaption classifiers, where the loss for the Gleason classifier was weighted twice as high as the loss for the domain classifier (i.e. $\alpha = 0.5$). In the gradient reversal layer $\lambda = 0.05$ of the gradient was backpropagated.

The best performing DANN was compared to when the classifier was trained without any domain adaption as well as when it was trained with colour augmentation of the training data. These results can be seen in Table 4.6. For two of the three test datasets, the best performance was achieved when using the optimised DANN, while the third test dataset achieved its best result when the network was trained with colour augmentation. For the validation dataset, from the same site as the training data, the best performance was achieved for another set of hyperparameters for the DANN. With these hyperparameters the test accuracy however was sub-optimal.

Table 4.5: Average accuracy results for validation (A) and testing (B, C and D) using different configurations of the DANN. The standard deviation is given in parenthesis, achieved by training each network four times.

	Val. avg. acc. (%)	Test avg. acc. (%)		
Approach	Dataset A	Dataset B	Dataset C	Dataset D
2 layers, $\lambda = 0.1$, $\alpha = 0.5$, colour aug	71.8 (1.8)	50.9 (1.6)	57.4 (0.7)	54.0 (1.6)
2 layers, $\lambda = 0.1$, $\alpha = 0.5$	73.0 (1.4)	52.3 (2.3)	55.9 (0.9)	53.7 (1.3)
2 layers, $\lambda = 0.075$, $\alpha = 0.5$	73.2 (1.0)	52.4 (3.3)	57.8 (1.9)	**55.1 (1.5)**
2 layers, $\lambda = 0.06$, $\alpha = 0.5$	74.2 (0.9)	55.1 (2.9)	58.0 (3.3)	54.7 (1.4)
2 layers, $\lambda = 0.05$, $\alpha = 0.5$	74.8 (1.4)	**55.8 (3.1)**	**58.3 (1.2)**	54.2 (2.7)
2 layers, $\lambda = 0.025$, $\alpha = 0.5$	**77.8 (1.4)**	49.0 (1.6)	56.4 (2.8)	49.9 (2.5)
2 layers, $\lambda = 0.05$, $\alpha = 0.6$	74.1 (1.3)	54.3 (1.8)	57.2 (0.9)	54.8 (1.8)
2 layers, $\lambda = 0.05$, $\alpha = 0.4$	76.1 (1.2)	49.6 (3.8)	55.1 (2.9)	51.3 (3.9)
3 layers, $\lambda = 0.05$, $\alpha = 0.5$	75.5 (0.8)	53.7 (1.1)	55.4 (4.7)	52.5 (1.2)
1 layer, $\lambda = 0.05$, $\alpha = 0.5$	75.8 (1.6)	49.2 (3.2)	56.3 (2.0)	50.7 (1.5)

Table 4.6: Average accuracy results for validation (A) and testing (B, C and D) when using different, and none, techniques for improved generalisation performance. The standard deviation is given in parenthesis, achieved by training each network four times.

	Val. avg. acc. (%)	Test avg. acc. (%)		
Approach	Dataset A	Dataset B	Dataset C	Dataset D
Nothing	**76.2 (0.9)**	45.7 (1.1)	55.5 (2.5)	45.5 (2.4)
colour augmentation	73.7 (0.9)	54.8 (2.3)	57.6 (2.4)	**57.0 (1.5)**
DANN 2 layers, $\lambda = 0.05$, $\alpha = 0.5$	74.8 (1.4)	**55.8 (3.1)**	**58.3 (1.2)**	54.2 (2.7)

Table 4.7: Average accuracy results for validation (A) and testing (B, C and D) using the DANN with two layers, $\lambda = 0.05$, $\alpha = 0.5$, when only including two of the three test datasets when training the DANN. The standard deviation is given in parenthesis, achieved by training each network four times.

Datasets used when training	Val. avg. acc. (%)	Test avg. acc. (%)		
the domain part of the DANN	Dataset A	Dataset B	Dataset C	Dataset D
B, C, D	74.8 (1.4)	**55.8 (3.1)**	**58.3 (1.2)**	**54.2 (2.7)**
B, C	76.3 (2.0)	49.3 (0.7)	55.5 (2.1)	47.7 (0.5)
B, D	75.5 (2.8)	44.3 (4.2)	55.3 (3.0)	49.8 (3.7)
C, D	76.5 (2.1)	46.4 (5.8)	55.3 (4.4)	49.0 (4.5)

To check that the DANN's performance was not only improved for the datasets used when training, it was also trained including only two of the test datasets. These results can be seen in Table 4.7, excluding each of the test datasets once.

4.4.4 Discussion

The second-best accuracy for the validation dataset was achieved when training the Gleason grading classifier without trying to improve the generalisation performance. This is not surprising, since there indeed is no need for generalisation for this dataset. However, the results for the test datasets were the worst with this setting, showing the need for some domain adaption technique in order to train a neural network that generalises well for this type of task.

Excluding each of the test datasets once, see Table 4.7, gives an estimate on how well the network could be expected to work for datasets not included in the training. As can be seen, the results are worsened for all test datasets, independent of which one is excluded. However, most of the results are still better than when no technique is used to improve the generalisation performance, see Table 4.6. Thus, independent of which datasets are used for training the domain part of the DANN, the result is improved for all test datasets and one could expect this is true for other datasets as well.

As intended, the domain predictions of the trained DANN were quite random. Often it got stuck to only predicting one or two of the four domains. Since the goal was that the network could not predict the domain belonging due to the gradient reversal layer, these results were pleasing.

4.5 Generalisation Discussion

There is a need for an automated classification algorithm for Gleason grading due to the high occurrence of PCa and the large variations in Gleason grading between different pathologists. A feasible first step would be an algorithm suggesting Gleason grade to use as a second opinion. In this chapter, new results on automatic Gleason grade classification of H&E stained prostatic tissue were presented. The methods are based on DL, where CNNs were trained from scratch. The accuracy varies between the papers and the different datasets make it hard to compare the results. The evaluation was only performed on patches, i.e. small tissue areas. It is also relevant to evaluate the algorithms on whole slides to see whether they can predict GS too. This will be further explored in Chapter 6.

The first algorithm, Paper I, achieved an accuracy of 92.7% for whole image classification. This was an improvement to previous state of the art of 89.8% for the dataset. However, this dataset is small and the generalisation performance is unsatisfactory (accuracy is 32% when tested on part of dataset A. These results have however not been published). When investigating the misclassified images in Figure 4.4, one can see that different parts of the images are misclassified with different networks. With a larger dataset, these errors probably could be reduced and the performance would have been improved.

In the second part, Paper III, three different datasets were used to evaluate the generalisation performance for several different methods. While the methods did not generalise that well for the task of Gleason grading, they still managed to separate benign and malignant tissue in slides from different sites. The best generalisation was obtained when an autoencoder was used for downsampling, while the highest accuracy on images from the training site was achieved with the classifier using the highest resolution of the input images. Furthermore, both colour augmentation of the training data as well as the use of digital stain separation and normalisation increased the ability to generalise.

The third algorithm, Paper V, continued the work in Paper III but with larger and more datasets. CycleGANs were trained for translation between the different sites and used for normalisation of the test data as well as augmentation of the training data. However, none of these approaches improved the results and the best results were instead found using other augmentation techniques; blurring, intensity clipping, addition of noise and colour modification. Thus, it was not preferable to only generate realistic variations when training the CNN. Instead, the potentially less realistic variations generated by simpler augmentation techniques improved the generalisation performance the most. These results are surprising, since the CycleGANs were trained using the test dataset and the augmented versions created using those should create training data similar to the test datasets. This method does however have some other drawbacks; when using the CycleGANs for augmentation there is a risk that the trained CNN will overfit to these sites and if using the CycleGANs for normalisation training of new CycleGANs is necessary to apply the CNN to images from other sites.

The last method using DANNs, from Paper VI, used the same dataset as Paper V. The achieved results when using a DANN were slightly better than when colour augmentation was used. Also, the results showed that improved performance was achieved for all datasets and not only the ones used when training the domain part of the DANN. However, the algorithm still performed significantly better on data from the same site as the training data.

While there are limitations with all of the presented studies, the main one might be that no estimate of the standard deviation of the results were done in Paper I, III and V. This makes it hard to draw any conclusions about whether some results actually were significantly better than the others.

Another limitation is the limited and different datasets. The different datasets make it hard to draw fair conclusions when comparing all the different methods. A future study experimenting with all techniques on the same datasets would be interesting. Furthermore, the used datasets, especially for testing, are rather small and there is thus a risk of incorrect conclusions. However, no extensive hyperparameter tuning was done and methods such as early stopping were not used, reducing the risk of overfitting to the validation data.

An important aspect when evaluating the methods is the implementation. Some of the methods, e.g. CycleGANs for normalisation but possibly also other normalisation techniques, need new training and tuning when new datasets should be used. In a real-world clinical setting it might also be relevant to have an algorithm which computationally is not too heavy. Most of the proposed methods are similar in this aspect, but the CycleGAN for normalisation has a potential drawback of requiring more computational power. Furthermore, if continuous retraining of the model is relevant, a model that is fairly easy to train is important. Using augmentation is probably the easiest to implement. Using the autoencoder requires two consecutive trainings, making it slightly more complex.

Using either stain normalisation or colour augmentation are standard approaches in computational pathology. Examples include; augmenting the colours [5, 154], using DANN [92], normalisation using CycleGANs [23] and other normalisation techniques [134]. They are all popular techniques which are used and developed continuously. In [156], they state that the combination of colour augmentation and normalisation achieves the best performance, but that colour normalisation can be skipped to save computational resources at a negligible performance cost. Another approach to reduce the problems with generalisation was presented in [11]. They suggest the usage of Picrosirius red and haematoxylin instead of H&E, which more clearly delineates the stromal boundaries, and presents successful results for automatic Gleason grading. The importance of standardisation is highlighted in [56]. A suggestion for how to measure the model performance on novel unlabelled datasets is to use the “representation shift” introduced in [141]. It can be used to detect new data that a model will have problem generalising to.

Chapter 5

Related Aspects of Gleason Grading

In the previous chapter, different techniques for automatic Gleason grading using DL were evaluated. In this chapter, two related studies which could be useful for a complete automatic Gleason scoring algorithm are presented. The first study investigates semantic segmentation of H&E stained tissue into relevant components. This could be a useful preprocessing step before Gleason grading, where the grading could be based on detailed segmentation of the glands and nuclei. The second study focuses on the detection of small G5 areas, which is important to set the correct GS and diagnosis. The patch sizes used in the methods in the previous chapter are larger than the smallest annotated G5 areas, wherefore a risk exists that these areas are overlooked by the algorithm. The studies were published in Paper II and IV.

5.1 Semantic Segmentation of H&E Stained Prostate Tissue

Semantic segmentation of relevant components of H&E stained tissue could be useful for further analysis of the tissue. For example, a Gleason grading algorithm might benefit from having access to individually segmented, pathologically relevant objects from the images. In this work an algorithm for semantic segmentation of microscopy images of H&E stained prostate tissue has been developed. The semantic segmentation is done into the four components: Background, Stroma, Epithelial Cytoplasm and Nuclei. These components are the most relevant parts for PCa diagnosis, since the malignant structures are given by the glands, which are formed by epithelial cytoplasm and nuclei. Thus, if these structures were given, the Gleason grades could more easily be determined by an algorithm. The semantic segmentation algorithm used is based on the u-net, described in Section 3.5, but with some modifications. The most obvious one is that segmentation into more than two classes

Table 5.1: Gold standard statistics. (©2017 IEEE)

	Background	Stroma	Epithelial Cytoplasm	Nuclei
Percentage of pixels	11	20	50	11

was performed. The work is based on the master's thesis by Isaksson [79], published in Paper II [80].

Overall semantic segmentation using CNNs was a young field when this work was conducted — its utilisation towards histopathology even more so — and there was a lack of viable datasets. No available dataset was found providing annotations for more than nuclei segmentation.

5.1.1 Dataset

The data used is part of Dataset A. The data consisted of both benign and malignant areas from one slide. The images were downsampled to 20X, resulting in a pixel size of about 0.5 µm.

To be able to perform supervised training and automatic evaluation of the performance of the algorithm, a gold standard segmentation was needed. It was achieved by manually annotating each pixel into one of the following four classes: Background, Stroma, Epithelial Cytoplasm or Nuclei. A two-pixel-wide border around each individual object was added: one pixel on the classified objects' edges and one pixel just outside the edges. The pixels on the borders were not assigned any class and thus could neither be correctly nor incorrectly classified. This in order to account for inconsistency in the gold standard segmentation, since it is hard to find the exact border. To keep the annotation somewhat simple, the lumen in the stroma and epithelial cytoplasm was not segmented out into a separate class. An example of segmentation is given in Figure 5.1. In total approximately 6.5 million pixels were annotated, all in the same image. Gold standard statistics are given in Table 5.1. The remaining area was covered by the borders.

The data was divided into non-overlapping patches of size 250 × 250 pixels. One third of the patches were randomly chosen and kept as validation data. Due to the limited amount of annotated data, the training data was expanded artificially by using data augmentation. More specifically flipping, rotation and spatial distortion of the data was used. The spatial distortion was obtained by applying a displacement map, consisting of two matrices with the same size as the patches where each element corresponded to how far a given pixel in the original data was to be moved in the x- and y-direction respectively. The matrices were produced by repeating a sine wave vector along the x- and y-axis respectively. The sine wave had a period of 250 pixels (same as the patch size), while its amplitude was set to 6.5% of the patch size. This was considered to be small enough to give reasonable results

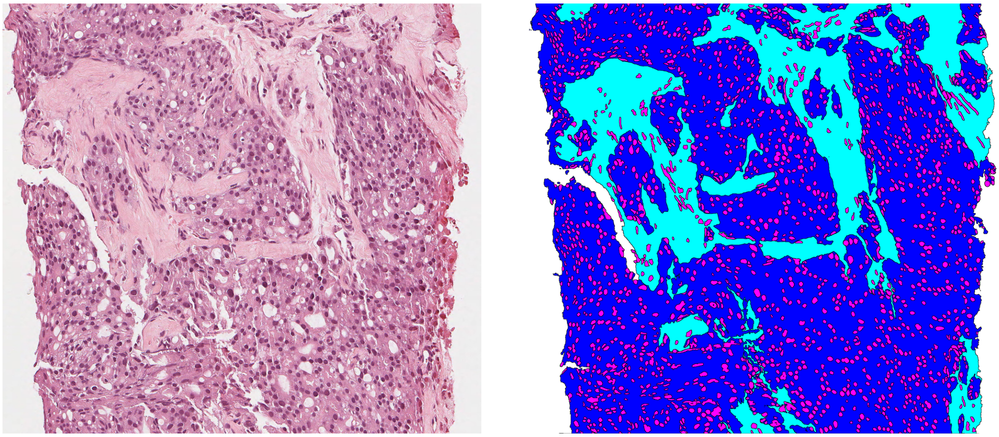


Figure 5.1: The original image (left) and the gold standard segmentation (right) into the classes Background (white), Stroma (light blue), Epithelial Cytoplasm (dark blue) and Nuclei (purple). (©2017 IEEE)

yet different from the original data. Normalisation was done by subtracting each colour channel's mean over the dataset from the same colour channel in each individual patch.

5.1.2 Method

The CNN design used was adapted from the original u-net design, with the main difference that semantic segmentation was performed into four classes instead of only two. Furthermore, zero padding was used in the convolutional layers, which results in that the input and the output have the same size. The up-sampling layers were also replaced with transposed convolutions, see Section 3.4, increasing the resolution a factor two using a 3×3 kernel and 1 zero inserted between the input pixels. Finally, the network was deepened with one more downsampling level. The network design can be seen in Figure 5.2.

The network was implemented using MatConvNet [161]. The final layer in the network calculated the softmax for each pixel and pixel-wise cross-entropy was used as loss function. It was trained from scratch using stochastic gradient descent with momentum, with a learning rate of 0.005 and decreased to 0.0025 after 80 epochs. The network was trained for 106 epochs. The momentum parameter was set to 0.9 and the weight decay parameter was set to 0.0005. The weights were initialised randomly using the standard normal distribution, multiplied by a constant equal to $\sqrt{2/N}$, where N was the total number of weights in the current filter. To evaluate the performance of the network, several metrics were used; the pixel accuracy, the average accuracy and the IoU.

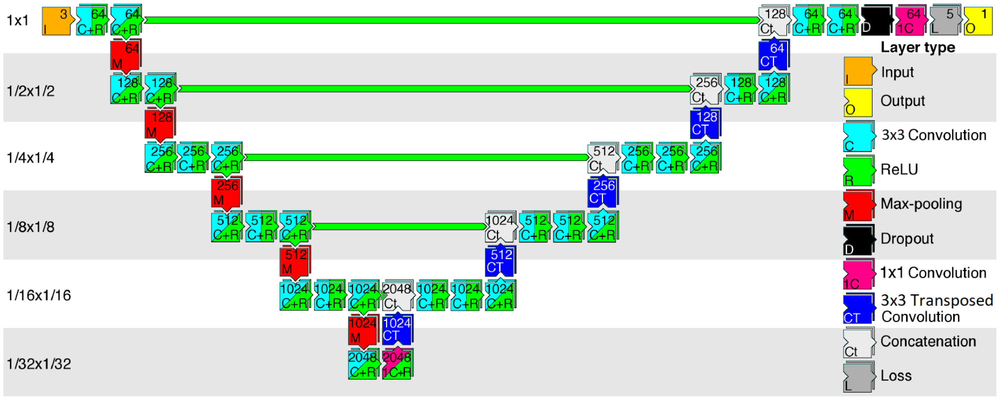


Figure 5.2: The network design used for semantic segmentation of H&E stained prostatic tissue. (©2017 IEEE)

5.1.3 Results

Examples of predicted segmentations and corresponding gold standards are given in Figure 5.3. According to these images the hardest part seems to be differentiation between stroma and epithelial cytoplasm. The obtained pixel accuracies are given in Table 5.2. The

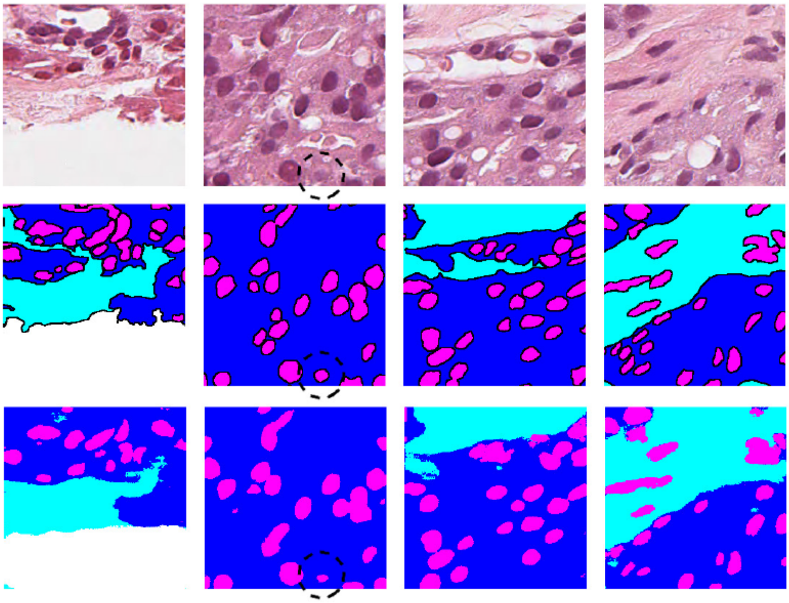


Figure 5.3: Input image (upper), gold standard (middle) and results (bottom). Note the weak, encircled, nucleus which the network is able to detect. (©2017 IEEE)

Table 5.2: Achieved accuracies for semantic segmentation of the prostate tissue. Mean accuracy is the mean of the accuracies for each class, while the pixel accuracy is the accuracy when not considering that there are different number or pixels of the different classes. (©2017 IEEE)

Pixel accuracy	Mean accuracy	Nuclei accuracy
91%	86%	78%

Table 5.3: Achieved results for each class when measuring IoU. (©2017 IEEE)

Background	Stroma	Epithelial Cytoplasm	Nuclei	Mean
84%	75%	88%	72%	80%

intersection over union results for each class are given in Table 5.3. In Figure 5.4 the confusion matrix is given, which specifies how high ratio of each class that has been predicted into the different classes.

5.1.4 Discussion

This was one of the first publications on semantic segmentation into multiple classes for histopathological images. The closest comparisons that could be made were with various methods for nuclei segmentation. For example [67] and [78] review some previous works and [121] used a CNN for this task. They all report pixel accuracies between 80-94%, which can be compared with the result of 78% in this study. However, since the network presented in this study was not only trained for nuclei segmentation the comparison is not fair and better results could have been achieved otherwise. For example, the mean pixel accuracy is 86% which is similar to the previous methods. Furthermore, pixel accuracy has several drawbacks since it only considers true positives. For example, a nuclei accuracy of 100% would have been achieved if all pixels in the image were classified as nuclei, regardless of their gold standard.

The intersection over union results, given in Table 5.3, vary between the different classes. This can to some extent be explained by the different amount of data for the different classes, resulting in that the more dominant classes are better trained, since no weighting of the loss was used and these classes thus affect the loss more. For example, the different amount of

		Predicted class			
		Background	Nuclei	Stroma	Epithelial Cytoplasm
True class	Background	0.84	0.00	0.07	0.09
	Nuclei	0.00	0.78	0.01	0.21
	Stroma	0.00	0.01	0.85	0.14
	Epithelial Cytoplasm	0.00	0.01	0.03	0.96

Figure 5.4: Confusion matrix showing the different misclassifications.

epithelial cytoplasm compared to stroma in the dataset, 50% and 20% respectively, see Table 5.1, promote the classification of epithelial cytoplasm.

Based on the confusion matrix, Figure 5.4, both nuclei and stroma were commonly misclassified as epithelial cytoplasm. Based on the colour and intensity of the pixels from these classes this is not surprising, since the epithelial cytoplasm appears as a mixture of the other two. Another reason for the misclassification could be that bright areas close to epithelial cytoplasm often are stroma, but could also be lumen which in the dataset is marked as epithelial cytoplasm. The results might have been improved if the lumen had been segmented into its own class, but due to the time-consuming gold standard generation this was not done. Some misclassifications, both for background and epithelial cytoplasm, might also be mitigated by increasing the patch size, as the differences sometimes are not apparent without a larger surrounding area to refer to.

Due to the networks design, large patches could have been preferable when training. This to reduce the influence of the zeros from the zero padding. However, to get an adequate amount of patches this size was chosen.

For this study, only data from one slide was used. This is of course a large disadvantage and it is hard to draw fair conclusions on such a limited dataset. One could however suspect that similar results on other slides could be obtained if annotated data from more slides was available to train the network.

5.2 Detection of Small Gleason Grade 5-areas

While G3 and G4 are identified by the pattern of the cells, G5 can sometimes be detected on smaller areas. Since the tumours sometimes split up when the cancer grows, single cells of G5, the highest and most malignant Gleason grade, can occur intermingled with benign tissue. The single cancer cells can be identified by the large nuclei with irregular shape [12]. Since G5 is crucial when setting the diagnosis and determining treatment, it is of great importance to find even very small areas of the highest grade.

In this work, an approach to detect small G5 areas using DL was presented. A CNN was trained to differentiate G5 and non-G5 (i.e. benign, G3 or G4) based on only a small area of tissue but captured in high magnification. The work is based on the master's thesis by Tall [152], who investigated multiple different network designs before choosing the ResNet design described here. The main results were published in Paper IV [153].

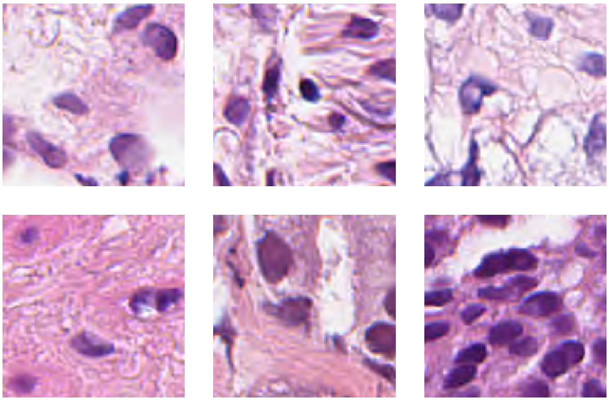


Figure 5.5: Examples of G5 patches (top row) and a benign, G3 and G4 patch (bottom row).

5.2.1 Dataset

The dataset used was created from part of Dataset A. Based on the pathologists' annotations, large image crops containing only tissue with one Gleason grade or benign tissue were manually created. Non-overlapping image patches with size 128 × 128 pixels were automatically created from the large image crops. To expand the dataset, rotation and flipping of the samples were used. Rotation angles were chosen to be integer multiples of ninety degrees in order to avoid the need for using interpolation in the image transformations. Through mirroring and rotation of each patch, the dataset was increased a factor eight. In total the dataset consists of 9840 G5 image patches and 9840 non-G5 image patches, whereof there are equally many benign, G3 and G4 samples. Examples of patches with different grades can be seen in Figure 5.5.

To further expand the dataset and make the algorithm more robust to stain variations, colour augmentation was used. The augmentation process described in Section 3.8 with parameters $h = 1.06, s = 1.3, v = 1.3$ was used. Examples of colour augmented images can be seen in Figure 5.6.

Validation of the network was done using 5-fold cross-validation. All augmented versions of an original image patch were kept in the same fold, to make sure training was not performed on any version of the validation data. The network was trained both with and without colour augmentation, to investigate whether this improved the performance and reduced overfitting.

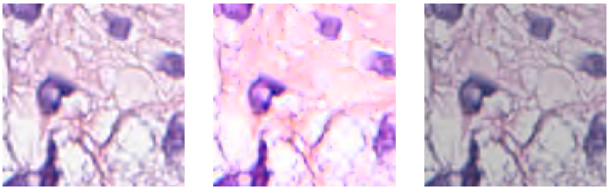


Figure 5.6: Illustration of a G5 image and two colour augmented versions of it.

5.2.2 Method

Multiple CNNs were designed for classification of G5 versus non-G5 images in the master's thesis by Tall [152]. The network which performed the best and was used in this work has a ResNet design with three main sections of ResNet-building block-triplets [68]. It can be seen in Figure 5.7 and the architecture elements are explained in Table 5.4. The sizes of the filter banks for each building block were successively incremented for every section, going from 58, 62, to finally 70. For the two first building blocks of each triplet, batch normalisation was performed after the convolutions, followed by ReLU-activations, as done in [68]. For the final triplet building block, the batch normalisation was instead followed by an addition, after which a ReLU-activation was performed.

The network was implemented and trained from scratch for 250 epochs with a batch size of 72 and the cross-entropy loss in Keras, using the Adam optimiser with the parameters: $\epsilon = 0, \beta_1 = 0.9, \beta_2 = 0.999$, learning rate= 0.000078 [30, 87]. Hyperparameter settings, as well as the number of layers in the architecture, were heuristically determined one by one through iterative tests for the developed architecture.

Table 5.4: Architecture element settings of the constructed ResNet.

Name	Layer	Pooling	Filters	Kernel	Stride	Padding	Act. func.
Add	Addition	N/A	N/A	N/A	N/A	N/A	N/A
Average Pool	2D avg. pool.	7 x 7	N/A	N/A	N/A	N/A	N/A
Batch Norm	Batch norm.	N/A	N/A	N/A	N/A	N/A	N/A
Conv A	2D conv.	N/A	58	3 x 3	1 x 1	Valid	ReLU
Conv B	2D conv.	N/A	58	3 x 3	1 x 1	Same	N/A
Conv C	2D conv.	N/A	62	3 x 3	1 x 1	Valid	ReLU
Conv D	2D conv.	N/A	62	3 x 3	1 x 1	Same	N/A
Conv E	2D conv.	N/A	70	3 x 3	1 x 1	Same	N/A
Dense	Dense	N/A	N/A	N/A	N/A	N/A	Sigmoid
Flatten	Flatten	N/A	N/A	N/A	N/A	N/A	N/A
Max Pool	2D max pool.	2 x 2	N/A	N/A	N/A	N/A	N/A
ReLU	N/A	N/A	N/A	N/A	N/A	N/A	ReLU

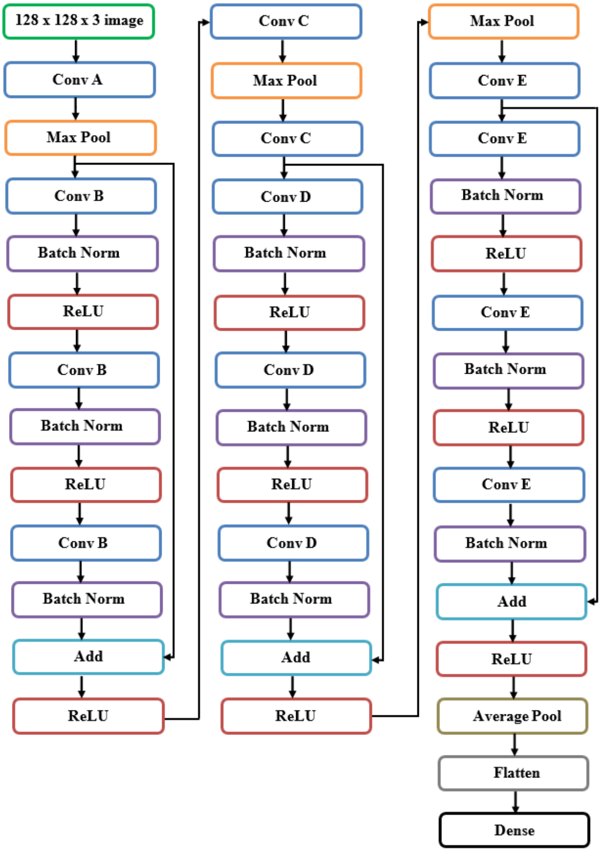


Figure 5.7: Design of the classifier, with a ResNet architecture.

5.2.3 Results

The CNN gave an overall average accuracy of 92% using 5-fold cross-validation without colour augmentation. The corresponding number with colour augmentation was 84%. The confusion matrix in the case without colour augmentation can be seen in Figure 5.8.

		Predicted class	
		Non G5	G5
True class	Benign	87±3%	13±3%
	G3	90±3%	10±3%
	G4	92±2%	8±2%
	G5	7±2%	93±2%

Figure 5.8: Confusion matrices when the network was trained without colour augmentation, given as average correct ratio ± standard deviation for the five folds. The rows represent the correct classes, while the columns give the class predicted by the algorithm, i.e. either G5 or non-G5.

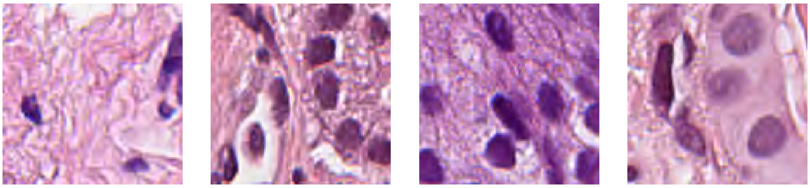


Figure 5.9: Examples of a benign, G3, G4 and G5 image that were misclassified.

In Figure 5.9, some examples of images that were misclassified are shown.

5.2.4 Discussion

The accuracy obtained when using colour augmentation was lower than when no colour augmentation was used. One could expect that colour augmentation would increase the accuracy, since there are more variations in the training data, but this was not the case. However, since the validation data was from the same slides as the training data, although not the same areas, it is not surprising that the accuracy got lower. This is a limitation for this study and indicates that the network probably is overfitted to the dataset. Another reason for high risk of overfitting, is the fact that the hyperparameters and network design were optimised carefully using this dataset. Further evaluation on slides from multiple sites is needed to draw reliable conclusions.

Based on the confusion matrix, it seems to be approximately the same probability that a benign, G3 or G4 image is classified as G5, although it is more common that a benign area is misclassified. One could imagine that G4 was more similar to G5 and thus a more likely misclassification, but this does not seem to be the case.

5.3 Discussion of Related Aspects of Gleason Grading

The two methods described in this chapter could be useful components for an automatic Gleason grading algorithm. They do however both need further evaluation on more data to get reliable results and draw correct conclusions, since they both use training and validation data from the same slides, even if the areas are not overlapping.

The semantic segmentation method could be a useful preprocessing step for a Gleason grading algorithm. The input to the Gleason grading algorithm could be either replaced or combined with semantic segmentation of the tissue, providing additional information and reduce the risk of incorrect conclusions due to inevitable variations in the staining. It efficiently defines where the glands are located, which are the regions of interest for Gleason grading. The obtained results are promising. The most difficult part of the semantic segmentation was differentiation between stroma and epithelial cytoplasm and this part might be improved by adding a fifth class for lumen. However, this method was only evaluated on tissue from one slide. It is reasonable to assume that similar problems with stain variations as described in Chapter 4 would appear for the semantic segmentation algorithm too when applied to slides from different sites.

Detection of the small G5 areas was successful. Scattered cells of G5 can occur and might be overlooked by algorithms only considering larger areas in lower magnification. Cells from both small and large annotated areas were used, hence the method was not limited to only small cancer areas. A similar approach could potentially be used to make distinction between normal and cancer tissue, since changes in the nuclei can be seen in most cancer cells. This approach has not been used for detection of PCa, since larger areas in general are needed to determine the Gleason grade.

Chapter 6

Evaluation of Automatic PCa Detection and Gleason Scoring

The application of AI to histological images shows promise in cancer detection [22]. Initial studies on automatic PCa detection were aimed at distinguishing cancer from non-cancer regions [13, 41, 99]. In recent years, more studies on automated Gleason grading have emerged, using image patches [83, 96], tissue microarrays [5, 118], prostatectomy specimens [111] as well as biopsies [22, 23, 90, 104, 112, 143]. Most studies have some limitation such as small training datasets, no separate testing sets, training on un-annotated images and lack of testing on images from different hospitals or scanners. In this chapter, two studies evaluating different algorithms for PCa detection are presented. In the first section, a smaller cohort was used for evaluation, where the slides were annotated multiple times and scanned on two different scanners for detailed evaluation of a Gleason grading algorithm. The second section presents preliminary results on cancer detection by a highly sensitive algorithm evaluated on a large cohort with biopsies of mainly normal tissue and low-grade cancer.

6.1 Standardisation of Gleason Scoring

This section is based on the study originally published in Paper VII [107]. Several studies have tested Gleason scoring algorithms on external cohorts scanned on different scanners, but no study that compares the performance on the same slides but scanned on different scanners has been found. The development and evaluation of a Gleason grading algorithm trained on annotated prostate biopsy scans are presented here. The training was done in two consecutive steps, as illustrated in Figure 6.1. The algorithm was evaluated on biopsies

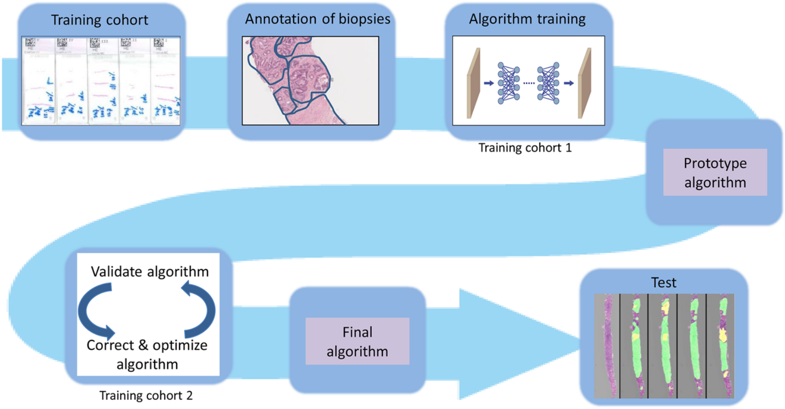


Figure 6.1: The schematic representation of the workflow for the development of the algorithm.

annotated in total three times by two pathologists. The slides were furthermore scanned on two different scanners to assure reproducibility.

Two datasets were used in this study. Extended Dataset A (introduced in Subsection 2.1.3) was used for development of the algorithm and Evaluation Dataset 1 (introduced in Subsection 2.1.4) was used for evaluation.

6.1.1 Development Datasets

The annotations performed by the two pathologists on Extended Dataset A were pooled and used to train the algorithm to recognise four classes: Benign, G3, G4, and G5. The training was performed in two consecutive steps, see Figure 6.1. In the first step, a training cohort consisting of 476 biopsy slides from 119 patients (Train 1; Table 2.2 on page 22) was used to construct the prototype algorithm. From the annotated digital slides, training data was extracted from the annotated regions to create 318 × 318 pixel-sized patches in 10X (0.988 µm/pixel). Three quarters of the annotated regions of each pattern were used to create training patches and the rest for validation of the prototype algorithm (Table 6.1). When extracting patches from the training data, rotation and mirroring of the annotated regions were used to augment the dataset. Rotation was done by every 10° in the range [0°, 350°]. Other types of augmentation, which do not modify the spatial structure of the patches, were also performed, as described under Algorithm Design and Training Procedure (Subsection 6.1.2).

For the second step, an additional subset of 222 slides from a further 55 patients was added for re-training to reach the final algorithm (Figure 6.1 and Train 2; Table 2.2 on page 22). Higher proportion of benign biopsy slides were included in this subset, in order to improve

Table 6.1: The number of patches used to train the two algorithms and for validation.

	Benign	G3	G4	G5
Prototype algorithm	34164	24594	32803	13363
Final algorithm	36554	24912	33543	13430
Validation	9539	2461	3158	947

the specificity of the algorithm. For these slides, the predictions from the first prototype algorithm were used as a starting point when annotating and the pathologists only made corrections to these annotations. Special attention was given to confounders (structures that mimic cancer but are benign): sclerosing adenosis, hyperplasia of glandular and stromal tissue with papillary buds, infoldings, as well as colon and anal mucosa.

When extracting the patches for the second step (final algorithm), only regions where the pathologist corrected the algorithm were included. This was done for two reasons. First, to avoid bias towards what the prototype algorithm predicted and to make sure that the pathologist has checked the region. Secondly, since these patches were incorrectly classified, one could surmise these regions to be harder to classify and thus be of increased value for honing the new algorithm. All of the patches from the second step were used for training. Number of slides used for training and their accompanying clinical diagnoses can be found in Table 2.2 on page 22 (Train 1 and Train 2) and the total number of patches used to train the second algorithm can be found in Table 6.1. As before, rotation and mirroring of the images were used when extracting the patches.

6.1.2 Algorithm Design and Training Procedure

Two different CNNs were trained; the first for the prototype algorithm and the second for the final algorithm. Both received square patches measuring 318 × 318 pixels in 10X magnification as input and predicted the majority class, i.e. which Gleason pattern the majority of pixels in the patch belonged to.

The first prototype algorithm used a CNN design very similar to the one proven successful in Paper V [7] but without the part of the domain classifier. The size of this first network was quite small (only 150196 parameters), which had the inherent advantage of reducing the risk of over-fitting. This first CNN was trained from scratch.

For the final algorithm, a larger CNN, the inception V3 [150], was modified to recognise the four classes. It was pre-trained on the ImageNet dataset [38]. The motivation for switching CNN model was that a larger network can recognise more different structures and thus has higher potential. The relatively large dataset used in the second iteration made it possible to successfully train the network without overtraining.

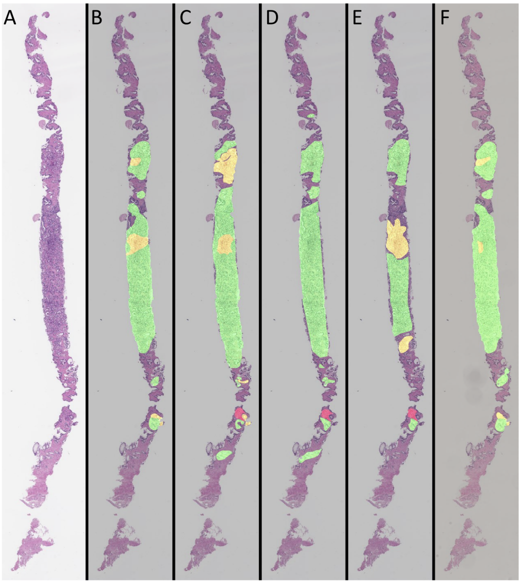
Both networks were implemented in Keras and trained with the Adam optimiser with default parameters [1, 30, 87]. The loss was weighted during training to give each class the same importance although different number of patches. The algorithm was optimised to handle images with noise and different stain appearances by using augmentation. Augmentation was done by blurring, clipping the intensity, adding noise and augmenting the colours of the images, as described in Section 3.8 with parameters $b = 2, t_1 = 0.1, t_2 = 0.9, n = 0.05, h = 1.04, s = 1.25, v = 1.25$.

6.1.3 Evaluation Procedure

Testing was performed on a separate cohort of 37 biopsy sections from 21 patients (Test; Table 2.2 on page 22). The slides were blinded and uploaded for the two pathologists (P1 and P2) to annotate. Pathologist 1 annotated the slides again at a different occasion, 12 months after the first annotation (first occasion: P1-1, second: P1-2). This allowed assessment of intra- and inter-observer differences. Annotations drawn by the pathologists and by the final algorithm were extracted and compared. The GGs for the slides given at time of diagnosis were extracted from clinical records. 36 of the slides were re-scanned on a different scanner 24 months after the original scans were performed. It was not possible to re-retrieve one of the test biopsy slides from the stores. The algorithm was tested on the new scanned images.

The test was done on the whole slide images in 10X by a sliding window approach, i.e. the algorithm was run on each 318×318 pixel patch with stride 50 pixels between the patches. As post-processing, to remove too small regions found as cancer, noise in the initial Gleason grading and to get a smoother and more easily interpretable result from the algorithm, the following steps were used. First, the result was smoothened by convolving with a 50×50 averaging filter. Secondly, the result for each pixel was transformed to a discrete class belonging by assigning the pixel the class with the highest probability. All malignant regions smaller than $(318 \cdot 318 \cdot 0.4 = 40450)$ pixels where removed, corresponding to $199 \times 199\ \mu\text{m}^2$ per patch. Finally, cancer regions of a certain pattern smaller than $(318 \cdot 318 \cdot 0.4 = 40450)$ pixels were changed to the majority class of the surrounding 318×318 pixels. The patch size was determined together with the pathologists based on what they considered the smallest area relevant for diagnosis.

To calculate the percentage of cancer and each Gleason grade pattern individually, the number of pixels containing prostate tissue (as opposed to background pixels) was needed. This was determined by thresholding the intensity of the image, where the pixels had values in the range [0, 1]. Pixels where the sum of the RGB values, i.e. in the range [0, 3], was below 2.5 were considered tissue and remaining pixels were considered background. This threshold was manually determined by visualising the resulting tissue segmentation for a few slides.



	Alg	P1-1	P1-2	P2	Alg Scan2
% cancer	43.4	47.3	43.3	31.7	45.2
% G3	39.0	37.9	42.5	23.0	43.3
% G4	4.4	8.5	0.0	7.9	1.9
% G5	0.0	0.9	0.8	0.8	0.0
Grade group	GG2	GG4	GG4	GG4	GG2

Figure 6.2: (A) Example of a biopsy section with the cancer areas (B) detected by the algorithm and annotated (C) by pathologist 1 at year 1 and (D) year 2 and (E) by pathologist 2. (F) The biopsy slide was rescanned on a different scanner and the algorithm was applied to the new image. Green represents G3, yellow G4, and red G5. The percentage of the biopsy tissue covered by the annotations was compared and the GG was determined. Alg = algorithm; Alg Scan2 = algorithm applied to image from scanner 2; P1-1 = pathologist 1, year 1; P1-2 = pathologist 1, year 2; P2 = pathologist 2.

The quantification of the algorithm output and the pathologists' annotations gave the percentage of G3, G4 and G5 on each biopsy section. The percentage of each Gleason pattern was determined as the ratio of pixels predicted with that pattern and the total number of pixels with tissue. The total percentage of cancer was thus given by the sum of the percentage of each Gleason pattern (Figure 6.2).

To evaluate how well the algorithm agrees with the different pathologists, several different methods were used. The overall percentage of cancer on each biopsy slide was compared between the pathologists (P1-1, P1-2 and P2) and the algorithm. The percentage of each Gleason pattern detected was also compared. The GGs obtained from the two pathologists' annotations and the algorithm were compared to each other as well as to the GG given at time of diagnosis. Specificity and sensitivity were calculated for the algorithm versus each of the pathologists (P1-1, P1-2 and P2) and the result was presented as the average of the three pathologists.

Table 6.2: Sensitivity and specificity of the final algorithm compared to the average of pathologists results. Based on Gleason patterns (G3-G5) annotated by the pathologists (P1-1, P1-2 and P2) and detected by the algorithm.

	Cancer	G3	G4	G5
Sensitivity	100%	89%	91%	80%
Specificity	68%	77%	79%	98%

The agreement between the algorithm and the pathologists was also examined at pixel level, using the agreement index introduced in Section 3.10. To calculate the agreement index for the rescanned slides, they first had to be aligned to the original scans. This was done using the method described in Section 3.9.

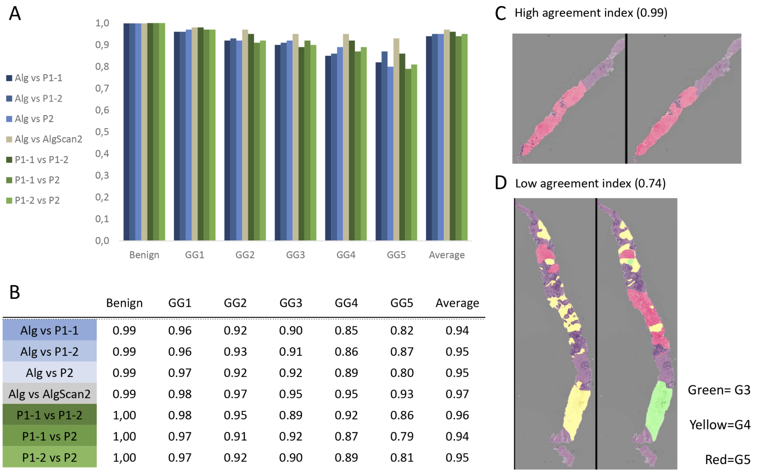
ICC was used to correlate the performance of the algorithm compared with the two pathologists and between the two pathologists in the calculations of per cent cancer and Gleason patterns on each biopsy section. Accuracy (correctly assigned cases/all cases) and Cohen's Kappa method was used to calculate agreement in estimating the GG. Calculations and statistical analysis were performed with SPSS (IBM, Armonk NY, USA) and Python (www.python.org).

6.1.4 Results

The algorithm was highly sensitive in detecting cancer (100%) and identifying the correct Gleason pattern (80–91%, depending on the Gleason pattern; Table 6.2). The specificity was 68–98%, depending on the Gleason pattern. The algorithm identified cancer areas on four out of 13 benign sections, but all false positives detected were very small ($<$ 1.2% of the whole tissue). Upon further investigation, many of these regions were difficult to classify as they were identified by the pathologists as artefacts or patterns that mimicked cancer.

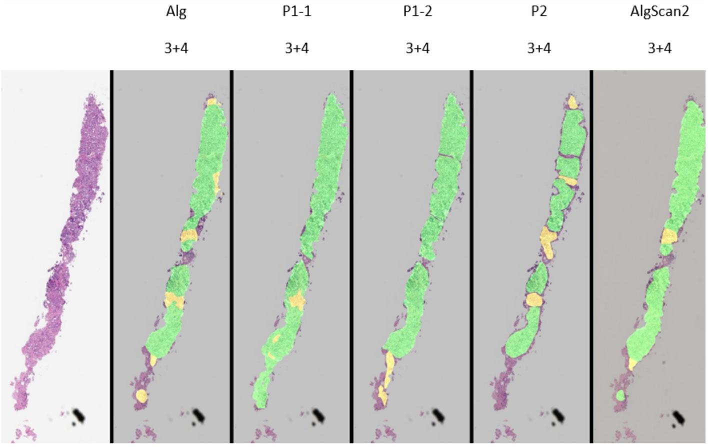
The GG for each biopsy section confirmed the presence of cancer and what pattern it was, but did not take into account whether the cancer was identified in the correct area. This was verified using the agreement index (Figure 6.3). Examples are shown in Figures 6.4, 6.5, 6.6. Complete results are available in the supplementary material to [107].

The agreement index was high in the benign biopsy sections and decreased gradually with increasing GG. The disagreement between the two pathologists and the disagreement between two annotations by the same pathologist were similar to the disagreement with the algorithm. The agreement for the algorithm on the two scanners was at least as high as the agreement for the algorithm and any of the pathologists, showing that the algorithm manages to predict very similar results even for images from different scanners.



	Benign	GG1	GG2	GG3	GG4	GG5	Average
Alg vs P1-1	0.99	0.96	0.92	0.90	0.85	0.82	0.94
Alg vs P1-2	0.99	0.96	0.93	0.91	0.86	0.87	0.95
Alg vs P2	0.99	0.97	0.92	0.92	0.89	0.80	0.95
Alg vs AlgScan2	0.99	0.98	0.97	0.95	0.95	0.93	0.97
P1-1 vs P1-2	1,00	0.98	0.95	0.89	0.92	0.86	0.96
P1-1 vs P2	1,00	0.97	0.91	0.92	0.87	0.79	0.94
P1-2 vs P2	1,00	0.97	0.92	0.90	0.89	0.81	0.95

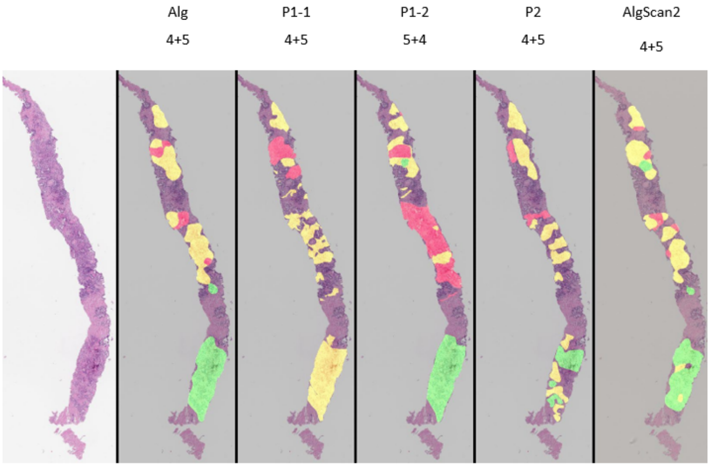
Figure 6.3: Agreement index. (A and B) The agreement index was calculated on a pixel-by-pixel basis. The average result for each GG is presented. Examples of scanned biopsy sections (C) with a high agreement index and (D) in the lower range. Alg = algorithm; Alg Scan2 = algorithm applied to image from scanner 2; P1-1 = pathologist 1, year 1; P1-2 = pathologist 1, year 2; P2 = pathologist 2.



Agreement index

	Alg	P1-1	P1-2	P2	AlgScan2
Alg	1.000	0.877	0.878	0.857	0.914
P1-1	0.877	1.000	0.938	0.881	0.889
P1-2	0.878	0.938	1.000	0.881	0.901
P2	0.857	0.881	0.881	1.000	0.872
AlgScan2	0.914	0.889	0.901	0.872	1.000

Figure 6.4: Biopsy with its corresponding masks generated from the algorithm result (Alg), the annotations from Pathologist 1 at two time points (P1-1 and P1-2), Pathologist 2 (P2) and the algorithm on same section scanned with a different scanner (AlgScan2). The table shows agreement index for the algorithm result and the pathologists' annotations.



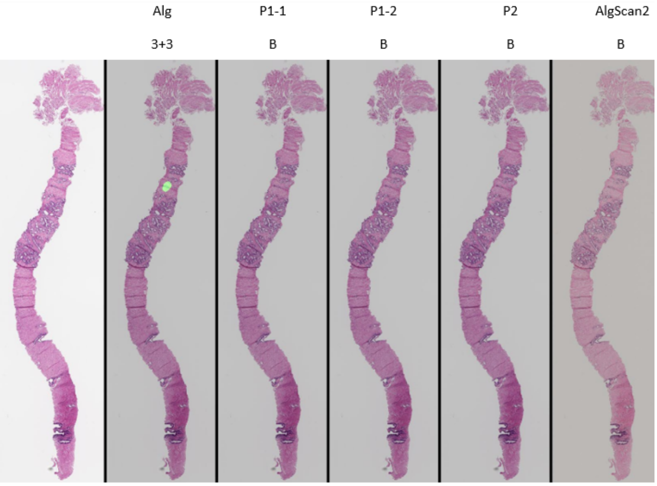
Agreement index

	Alg	P1-1	P1-2	P2	AlgScan2
Alg	1.000	0.818	0.830	0.858	0.916
P1-1	0.818	1.000	0.742	0.853	0.820
P1-2	0.830	0.742	1.000	0.770	0.808
P2	0.858	0.853	0.770	1.000	0.865
AlgScan2	0.916	0.820	0.808	0.865	1.000

Figure 6.5: Biopsy with its corresponding masks generated from the algorithm result (Alg), the annotations from Pathologist 1 at two time points (P1-1 and P1-2), Pathologist 2 (P2) and the algorithm on same section scanned with a different scanner (AlgScan2). The table shows agreement index for the algorithm result and the pathologists' annotations.

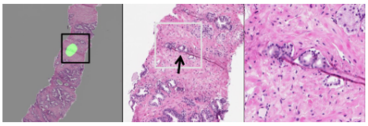
The percentage of cancer detected by the algorithm on test cohort images from the first scanner was strongly correlated with the amount of cancer detected by the pathologists (ICC $\geq$ 0.97; Figure 6.7). The intra- and inter-observer correlation coefficients were $\geq$0.99 and $\geq$0.95, respectively (Table 6.3). Areas identified as Gleason patterns 3 and 4 by the algorithm correlated well with the pathologists' interpretation ($\geq$0.83 for G3 and $\geq$0.88 for G4; Table 6.3). There was, however, varying agreement in the detection of areas identified as G5, whereby the algorithm correlated well with one pathologist's first scoring (P1-1; 0.98) but not with the other pathologist (P2; 0.57). The results of the algorithm on images from the two different scanners were similar (ICC $>$ 0.96).

Cohen's kappa was used to calculate agreement in GGs (Table 6.4). The highest overall agreement (0.69) was between the two time points of pathologist 1 (P1-1 and P1-2), between the two pathologists (P1-1 and P2), as well as between the algorithm and P2. Confusion matrices show that most disagreements differed by one GG (Figure 6.8).



Agreement index

	Alg	P1-1	P1-2	P2	AlgScan2
Alg	1.000	0.998	0.998	0.998	0.998
P1-1	0.998	1.000	1.000	1.000	1.000
P1-2	0.998	1.000	1.000	1.000	1.000
P2	0.998	1.000	1.000	1.000	1.000
AlgScan2	0.998	1.000	1.000	1.000	1.000



Algorithm identifies small suspicious glands as cancer. The fold obscures the architecture.

Figure 6.6: Biopsy with its corresponding masks generated from the algorithm result (Alg), the annotations from Pathologist 1 at two time points (P1-1 and P1-2), Pathologist 2 (P2) and the algorithm on same section scanned with a different scanner (AlgScan2). The table shows agreement index for the algorithm result and the pathologists' annotations. The area which incorrectly was identified as cancer by the algorithm is displayed at the bottom of the figure.

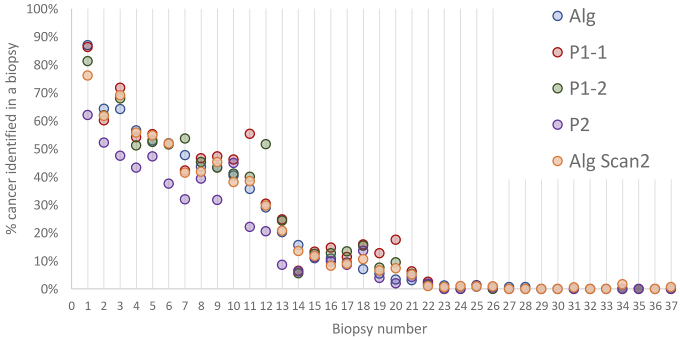


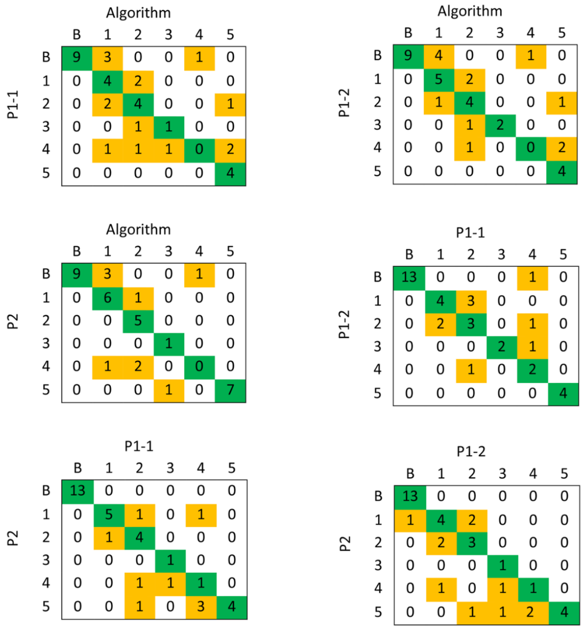
Figure 6.7: The percentage of cancer detected by the pathologists and the algorithm on both the original scan and the new scan of each biopsy section on a different scanner. Biopsy section number 35 was not available for rescan. Alg = algorithm; Alg Scan2 = algorithm applied to image from scanner 2; P1-1 = pathologist 1, year 1; P1-2 = pathologist 1, year 2; P2 = pathologist 2.

Table 6.3: ICC of the per cent of cancer and different Gleason patterns (G3, 4 and 5) detected by the algorithm, the algorithm on re-scanned slides and the two pathologists. All correlations: p≤0.001 apart from §p≤0.05.

	Alg vs P1-1	Alg vs P1-2	Alg vs P2	Alg vs AlgScan2	P1-1 vs P1-2	P1-1 vs P2	P1-2 vs P2	ICC
Cancer	0.99	0.99	0.97	1.00	0.99	0.95	0.95	0.99
GS3	0.83	0.94	0.95	0.99	0.98	0.87	0.86	0.97
GS4	0.89	0.92	0.88	0.99	0.97	0.88	0.84	0.96
GS5	0.98	0.83	0.57	0.96	0.65	0.77	0.51	0.88

Table 6.4: Cohen’s Kappa based on GGs and cancer vs benign.

	Alg vs P1-1	Alg vs P1-2	Alg vs P2	Alg vs AlgScan2	P1-1 vs P1-2	P1-1 vs P2	P1-2 vs P2
GG	0.50	0.56	0.69	0.68	0.69	0.69	0.62
(95% CI)	(0.32-0.68)	(0.37-0.74)	(0.52-0.86)	(0.50-0.86)	(0.52-0.86)	(0.53-0.85)	(0.44-0.79)
CA/B	0.75	0.69	0.75	0.58	0.94	1.00	0.94
(95% CI)	(0.52-0.97)	(0.45-0.93)	(0.52-0.97)	(0.25-0.91)	(0.83-1.05)	(1.00-1.00)	(0.83-1-05)



		Algorithm					
		B	1	2	3	4	5
P1-1	B	9	3	0	0	1	0
	1	0	4	2	0	0	0
	2	0	2	4	0	0	1
	3	0	0	1	1	0	0
	4	0	1	1	1	0	2
	5	0	0	0	0	0	4

		Algorithm					
		B	1	2	3	4	5
P1-2	B	9	4	0	0	1	0
	1	0	5	2	0	0	0
	2	0	1	4	0	0	1
	3	0	0	1	2	0	0
	4	0	0	1	0	0	2
	5	0	0	0	0	0	4

		Algorithm					
		B	1	2	3	4	5
P2	B	9	3	0	0	1	0
	1	0	6	1	0	0	0
	2	0	0	5	0	0	0
	3	0	0	0	1	0	0
	4	0	1	2	0	0	0
	5	0	0	0	1	0	7

		P1-1					
		B	1	2	3	4	5
P1-2	B	13	0	0	0	1	0
	1	0	4	3	0	0	0
	2	0	2	3	0	1	0
	3	0	0	0	2	1	0
	4	0	0	1	0	2	0
	5	0	0	0	0	0	4

		P1-1					
		B	1	2	3	4	5
P2	B	13	0	0	0	0	0
	1	0	5	1	0	1	0
	2	0	1	4	0	0	0
	3	0	0	0	1	0	0
	4	0	0	1	1	1	0
	5	0	0	1	0	3	4

		P1-2					
		B	1	2	3	4	5
P2	B	13	0	0	0	0	0
	1	1	4	2	0	0	0
	2	0	2	3	0	0	0
	3	0	0	0	1	0	0
	4	0	1	0	1	1	0
	5	0	0	1	1	2	4

Figure 6.8: Confusion matrices of GG (1-5) and non-cancer (B) classification. P1-1: Pathologist 1, year 1; P1-2: Pathologist 1, year 2; P2: Pathologist 2.

6.1.5 Discussion

Several studies on the subject of automatic PCa detection have been published in recent years [5, 23, 111, 112, 117, 143], indicating that there is a strong interest and need for the development of such tools. The present study is the first to evaluate an AI algorithm on a cohort scanned on two different slide scanners, comparing their results pixel by pixel. The algorithm performs similarly well on slides scanned on the two scanners, showing robustness despite colour and saturation differences. This is crucial, as scanner variations between different laboratories are one of the problems for adapting image analysis algorithms into practice [56]. Annotations and the resulting diagnoses performed by two pathologists and the same pathologist at two different time points were also compared. The Cohen's kappa values obtained between the two pathologists involved in the present study were in the range of that reported by Egevad and colleagues [47], who found that the kappa values for agreement between 24 expert pathologists evaluating 90 biopsy cases ranged between 0.60 and 0.74. In this study, the same pathologist annotating the same scans at two different dates separated by a year yielded a kappa value of 0.69, showing that there is both inter- and intra-observer variation. Nagpal et al. [111] validated a DL algorithm using diagnosis from 35 pathologists, ten of whom roughly annotated Gleason patterns on radical prostatectomy sections. There was a great variation between the GGs assigned by the ten pathologists; the kappa values were in the range of 0.37–0.62, which is lower than the agreement between the two pathologists involved in this study. These Gleason grading variations are still one of the major controversies among pathologists and standardisation of grading is of high interest.

The results in this study show that the algorithm is highly sensitive for detecting cancer, including the stratification into the different Gleason grade patterns. The cancer versus non-cancer areas detected by the algorithm were highly correlated with both pathologists, confirming the algorithm's good cancer detection performance. The algorithm overestimated cancer presence on several benign biopsy sections (four out of 13 benign sections), reducing the specificity. However, in all cases, the false positives were small (<1.2% of the tissue), and in many of those cases, these small areas consisted of artefacts, folds or mechanical distortions that may be difficult to classify. Increasing the minimum area threshold would help reduce the number of false positives but could risk increasing the chance of missing cancer regions. A recent study by Bulten et al. [23] classified a biopsy section as malignant only if $\geq$10% of the epithelial tissue was predicted as cancer. In pathology practice, even small suspicious areas can have a prognostic value. In this study's testing cohort, ten out of 24 cancerous biopsy sections had $\leq$10% cancer. Changing the threshold to ignore these biopsy sections would significantly change the sensitivity and specificity in this study, but the value of a tool that misses >40% of cancer biopsies is questionable.

The algorithm trained in the present work is similar to many others trained for the same application [5, 111, 118, 143], using a CNN. The robustness of the algorithm used in this study is demonstrated by obtaining very similar results on scans from two different scanners. Nagpal et al. [112] show encouraging results but lack testing on different scanners. Ström et al. [143] tested their algorithm on an external set scanned on a different scanner, but the slides in this set differed from those used in the testing cohort. Ström et al. [143] and Bulten et al. [23] use patches from biopsies with only one type of Gleason pattern (3 + 3, 4 + 4, 5 + 5). Detailed annotations were used for both training and testing in this study. This allowed for training a smaller network requiring less computational power, making this algorithm more practical for routine clinical use. The above studies evaluate their algorithms' performance by obtaining a Gleason grade for the whole slide. In contrast, the algorithm in this study was evaluated at pixel level, which confirmed the cancer grade and that the cancer was found in the correct areas. The agreement index compared the pathologists' and algorithm's pixel-level annotations, considering both the presence of cancer and the Gleason pattern in each pixel. The pixel-by-pixel algorithm output showed as much agreement with the pathologists as the pathologists had between them. The agreement index also allowed for the evaluation of the algorithm on images obtained from two different scanners, as the images could be aligned digitally and the same areas compared.

6.2 Evaluation of PCa Detection Algorithm

This section contains preliminary, unpublished results on automatic cancer detection evaluated on the PRIAS dataset; Evaluation Dataset 2. Instead of focusing on predicting the Gleason grade, the aim is instead to develop and evaluate a tool which is highly sensitive in cancer detection. The evaluation is mainly done on slides with small areas of low-grade cancer as well as benign slides.

6.2.1 Development Datasets

Most of the PCa datasets included in this thesis were used in this study. Extended Dataset A (introduced in Subsection 2.1.3) as well as Dataset B, C and D (introduced in Subsection 2.1.2) were used for training and validation of the algorithm. Three quarters of the slides were used for training and the remaining ones for testing. The division of the slides was done based on the GG of each slide, to get approximately the same ratio of each Gleason grade for both training and validation. Training patches were extracted with random orientations at random locations within the annotated areas from the training slides. In the same way as in described in Subsection 6.1.1, patches from training cohort 2 of Extended Dataset A were only extracted from regions modified by the reviewing pathologists and were only used for training. The validation was also done on patches, created at random locations within the annotated regions of the validation slides, to simplify the validation procedure. The patches had size 299 × 299 in 10X magnification.

6.2.2 Algorithm Design and Training Procedure

The algorithm consists of a CNN predicting 299 × 299 patches in 10X as either benign or malignant and a following postprocessing step to construct a segmentation of the cancer areas of a whole slide. The CNN design used was the DenseNet121. While several of the standard CNN designs probably would work well, this design was chosen mainly because its limited size could reduce the risk of overtraining, while still being large enough to learn the features needed for the limited task. Prediction of only benign versus malignant, and not Gleason grade, was chosen as approach to potentially increase the sensitivity and specificity of the algorithm for this task by getting more patches of each class to train on.

The implementation and training was done in Keras using the Adam optimiser with default parameters [1, 30, 87]. Cross-entropy was used as loss function, weighted to give each of the two classes the same importance. Transfer learning from pre-training on the ImageNet dataset [38] was used, but the final fully connected layers were replaced. The final global average pooling layer was replaced by an average pooling layer with size 9 × 9 and stride

1, to be able to segment larger areas with the same network. The final fully connected layer was replaced with only two output nodes to fit this application. All layers in the network were optimised when training. The network was trained with a batch size of 16 and early stopping based on the loss for the validation data was used to determine the number of epochs. The augmentation methods used were; blurring, clipping the intensity, adding noise and augmenting the colours of the images, as described in Section 3.8 with parameters $b = 2, t_1 = 0.1, t_2 = 0.9, n = 0.05, h = 1.06, s = 1.4, v = 1.4$.

6.2.3 Evaluation Procedure

Evaluation was done on the whole biopsy slides in Evaluation Dataset 2 (introduced in Subsection 2.1.5). The CNN was applied to the whole slide, or large crops of it if the image was too large to fit the GPU memory, to get a rough segmentation of cancer regions. The postprocessing was similar to the procedure described in Section 6.1; first, the result was smoothened by convolving with a 50×50 averaging filter. Secondly, the result for each pixel was transformed to a discrete class-belonging by assigning it the class with the highest probability. All malignant regions smaller than $(299 \cdot 299 \cdot 0.4 = 35760)$ pixels where removed, corresponding to 187×187 µm^2 per patch.

The sensitivity and specificity for cancer detection were measured. Future evaluation would include comparing the length of cancer reported in the medical records with the percentage of cancer predicted by the algorithm. Furthermore, the location of the cancer regions found by the algorithm and the areas marked as cancer by the pathologists (pen marks next to the tissue on most of the cancer slides) could be manually compared.

6.2.4 Results

Evaluation on the 2263 biopsies, whereof 260 with cancer, gives a sensitivity of 96% and a specificity of 69%. Three examples of areas detected by the algorithm and corresponding notes from the pathologists can be seen in Figure 6.9. More detailed examples can be seen in Figure 6.10, 6.11 and 6.12.

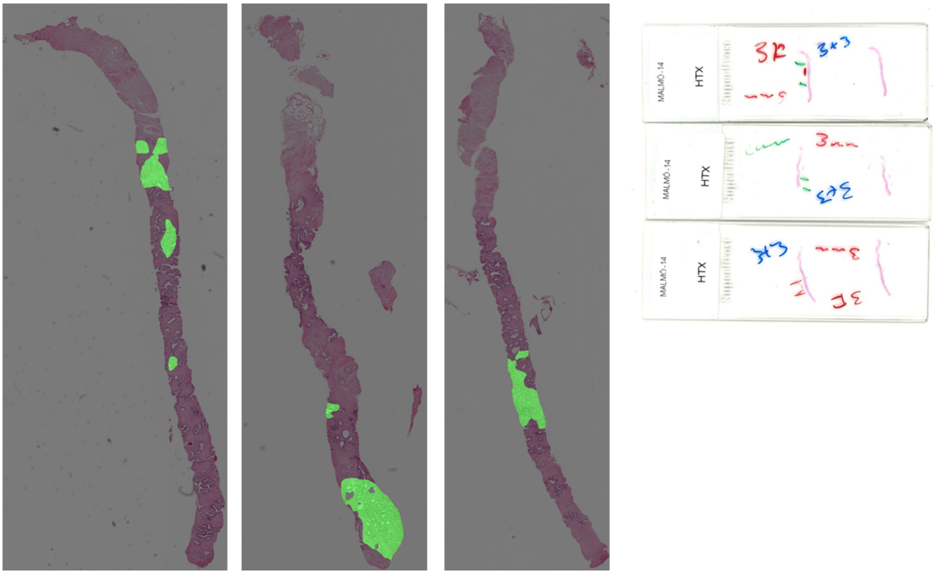


Figure 6.9: Examples of areas detected as cancer by the algorithm on three slides, highlighted in green. The pathologist's notes on the same three slides can be seen to the right, marking approximately the same regions.

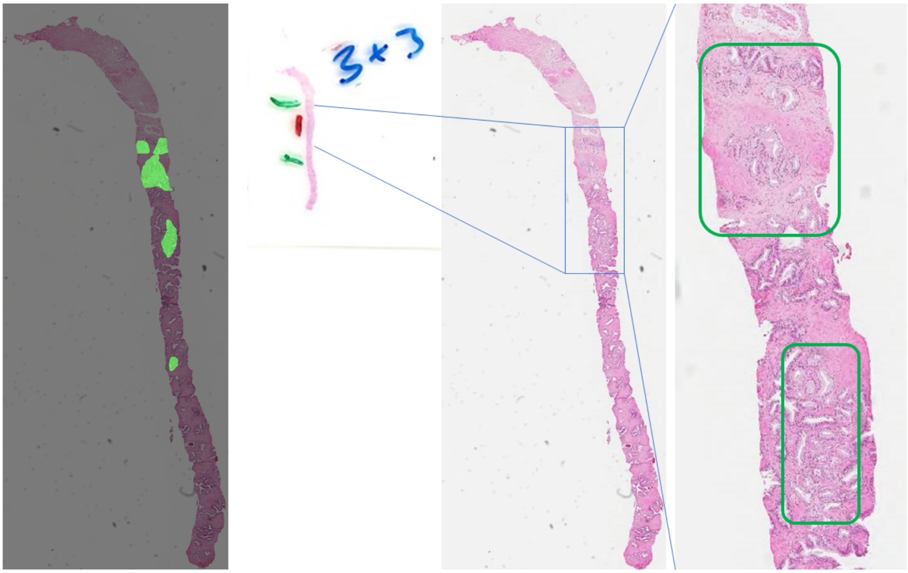


Figure 6.10: Areas detected by the algorithm highlighted in green to the left. The pathologist's note and magnification of the cancer area are also displayed. The areas detected as cancer by the algorithm are approximately given by the green boxes.

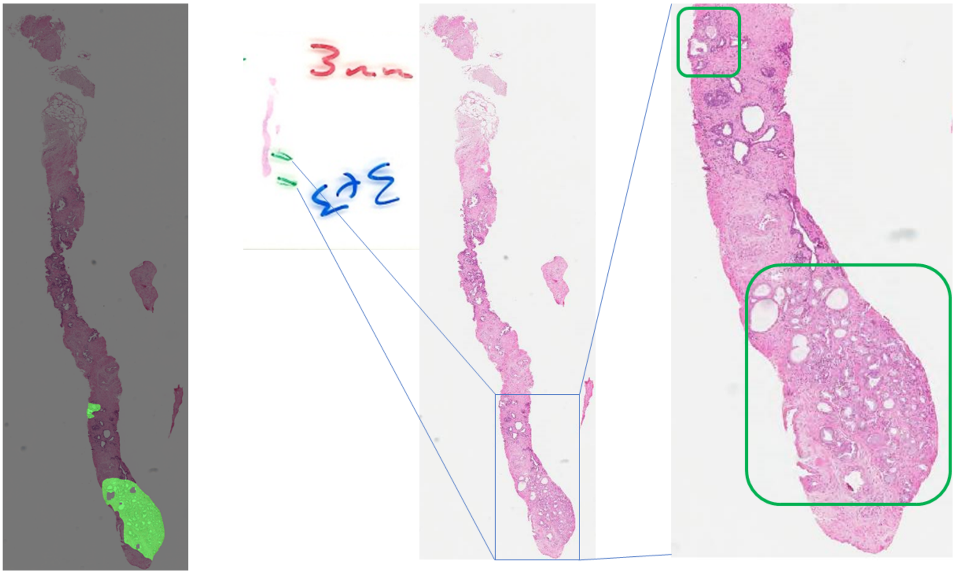


Figure 6.11: Areas detected by the algorithm highlighted in green to the left. The pathologist's note and magnification of the cancer area are also displayed. The areas detected as cancer by the algorithm are approximately given by the green boxes.

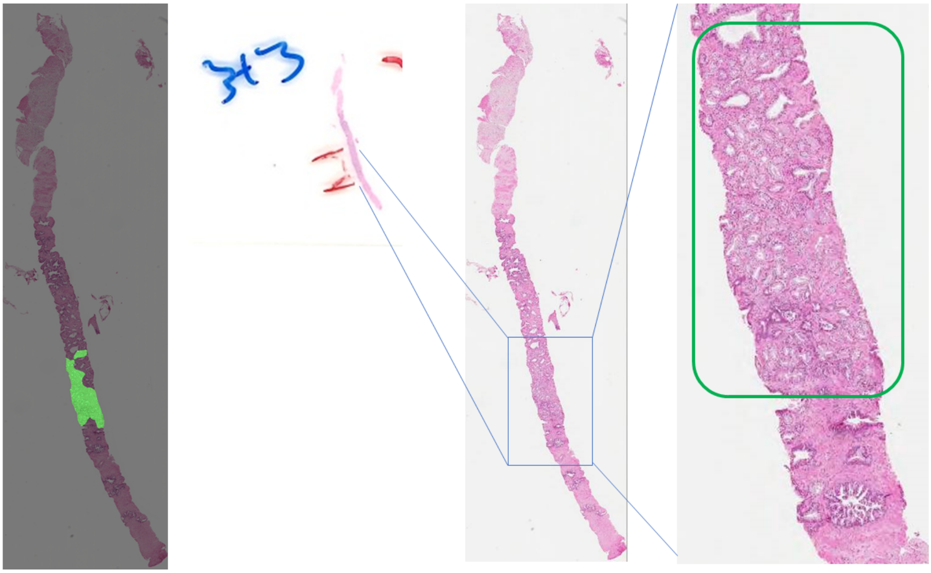


Figure 6.12: Areas detected by the algorithm highlighted in green to the left. The pathologist's note and magnification of the cancer area are also displayed. The areas detected as cancer by the algorithm are approximately given by the green box.

6.2.5 Discussion

The achieved results show promise towards implementation of a highly sensitive automatic PCa detection tool. Most of the slides with cancer in Evaluation Dataset 2 contained only small regions of cancer, the majority less than 2 mm and in average 2.9 mm. Furthermore, there are large stain variations in this dataset making the task even harder. The slides were several years old (from 2011 to 2018, scanned during 2018 – 2021), throughout which years the reagents might have changed and fading occurred. Considering that the algorithm handles these variations successfully, it could possibly be robust in a clinical environment.

The regions detected as cancer by the algorithm coincide well with the notes from the pathologists, see Figure 6.9, 6.10, 6.11 6.12. However, drawings from a pathologist was not always available and not always drawn in the same way. This makes evaluation of the location of the detected cancer regions complicated. Comparing the cancer length reported and the area of the regions detected by algorithm could give some further information of the algorithm's performance. This will be further investigated.

Experiences from the development phase of the algorithm in this study indicates that the CNN design is of minor importance. Instead, maximising the dataset used and expanding it with appropriate augmentation techniques as well as training it without overfitting appears to be more crucial to achieve a well-performing algorithm. The postprocessing step is also necessary to remove very small areas detected by the algorithm, since these typically are caused by e.g. dirt on the slide. Furthermore, using average pooling instead of global average pooling to be able to use the network for segmentation (although trained for classification) makes the evaluation on whole slides computationally much more efficient.

Comparing results from studies using different datasets is always problematic — even more so with this dataset containing mainly low-grade cancer. Bulten et al. [23] classified a biopsy section as malignant only if $\geq$10% of the epithelial tissue was predicted as cancer, which very few of the biopsies in this study had. The slides with small cancer areas are nevertheless of great importance for diagnosis and treatment selection.

6.3 Discussion on PCa Detection and Gleason Scoring

DL is of high interest for improving medical diagnostics. The volume of prostate biopsies collected, the complexity of histopathological assessment, and the lack of pathologists advocate the need for such tools. Two algorithms for automated PCa detection have been presented in this chapter, which could be used as a screening tool, improving the accuracy of PCa diagnosis and reducing pathologists' workload. The first tool detects cancer and assigns a Gleason grade with similar accuracy to that given by experienced pathologists comparing at whole slide and pixel levels. Importantly, the algorithm generated reproducible results on images obtained from a scanner different from that used for training data, suggesting its robustness. The second tool shows high sensitivity on a cohort with low-grade cancers which ought to be hard to detect. Final evaluation on this dataset is desirable. Future usage of the dataset where the training or evaluation is based on the patients' outcome instead of GS would be of great interest. The results are promising towards a tool which improves reproducibility, reduces inter-observer variability and facilitates the diagnostic process in PCa.

The two studies have several limitations. DL depends on a large amount of data and a higher number of annotated samples, particularly those with Gleason pattern 5, would be advantageous for training and testing. As in other studies, there is a problem finding a gold standard due to inter-observer variability among experienced pathologists. In these two studies, the annotations performed by two pathologists were pooled and the algorithms were based on a combination of the annotations. There was a high correlation between the two pathologists for cancer versus non-cancer areas, less so for different GGs. Whilst using annotation data from multiple experts is optimal for training and evaluating machine learning models [118], it is difficult to obtain such data. Using one pathologist for training and three for tuning could be a successful approach to get a general first version, which is subsequently optimised by several experts [111].

The clinical applicability of AI algorithms would be best tested on cohorts with long-term oncological follow-up, such as Evaluation Dataset 2. The chosen course of treatment and clinical outcome of the patient depends on many other factors apart from the biopsy GG. However, very promising results for automatic Gleason grading have been shown in several studies. In particular [23, 112, 143] have shown promising results on large cohorts. The usefulness of an automatic Gleason grading tool has been shown in [24], investigating AI assistance for grading prostate biopsies. This study compared the gradings from a panel of 14 observers grading biopsies with and without AI assistance with an expert reference standard and panel of international experts in prostate pathology, finding that the AI assistance increased the agreement significantly.

Chapter 7

Coronary Artery Disease

MPI is one of the most common cardiological examinations performed today for diagnosis and risk assessment in patients with stable ischaemic heart disease. There are many motivations why an AI algorithm would provide useful input to this task. For example, to reduce the subjectivity and save time for the nuclear medicine physicians working with this time-consuming task. In this chapter, results from three papers are presented: Paper VIII, IX, and X. In these papers, DL algorithms for multi-label classification based on CNNs were developed to detect obstructive CAD in the LAD, LCx and RCA, identify artefacts as well as predict QCA values. The predictions were based on data from MPI studies as well as a number of auxiliary parameters such as angina symptoms and age.

Already in the 1990's Fujita et al. [58] used ANNs to classify CAD using MPI. They used very compressed images, only 16 × 16 pixels, and a small fully connected neural network, with a single hidden layer with 100 nodes and an output layer with 8 nodes. They had a total of 74 images, whereof 58 were used for learning. Despite the very limited amount of resources, they managed to get an accuracy of 77% on their 16 test images, which they state is about as good as a radiologist with two years' experience. However, there is a high risk of overfitting using such a limited dataset.

With increased computational power and availability of large datasets, DL has become the preferred approach to many problems in image analysis. While the algorithm developed by Fujita et al. [58] was impressive for its time, it is also apparent that there were multiple limitations in this field thirty years ago. In the three works presented in this chapter a similar approach is used for similar tasks but using today's state-of-the-art techniques.

7.1 Mutual Methods for CAD Detection

The neural networks presented in this chapter are very similar but had a few differences wherefore these are presented in each corresponding section. The training procedure was however the same in all neural networks included in this chapter and is therefore described here, for all at once.

The neural networks were implemented in Keras [30] and trained with the Adam optimiser [87] with default parameters (Paper VIII and Paper IX) or with learning rate 0.0005 and otherwise default parameters (Paper X). They were trained for 50 epochs and using a batch size of 32. During training, dropout was used before the two first C1 layers. The networks were trained to predict multiple independent binary labels (e.g. CAD in the different arteries or presence of different artefacts), i.e. multiple labels could have a positive result. This was therefore treated as a multi-label problem. Averaged binary cross-entropy was used as loss function. To compensate for the imbalance in label occurrence, i.e. having different number of positive examples of each label, the loss was weighted in the way described in the following paragraph.

Let the number of labels which are predicted, i.e. the number of output nodes, be denoted n ($n \in \{1, 2, 3, 4, 5, 6\}$). Let $y_{t,i}$ denote the true class for label $i \in \{1, ..., n\}$ of an example, i.e., whose value is 0 if the label is normal and 1 otherwise. Let $y_{p,i}$ denote the corresponding prediction, such that $0 \leq y_{p,i} \leq 1$. Furthermore, let T denote the total number of examples and t_i denote the total number of positive examples from label i, that is, the examples with $y_{t,i} = 1$. Define (fixed) weights $w_i = t_i/T$, and assign to each example and each label a weight $\hat{w}_i$ according to the formula

$$\hat{w}_i = \frac{y_{t,i}(1 - w_i) + (1 - y_{t,i})w_i}{2w_i(1 - w_i)} \, . \tag{7.1}$$

Note that the sum of the weights $\hat{w}_i$ over all the t_i positive training examples is $T/2$, which equals the sum of all the $T - t_i$ negative training examples, as wanted. The total sum of all T training examples is thus T for each of the labels, giving them the equal importance. Finally, the weighted loss L_w (per example) is defined by:

$$L_w = -\frac{1}{n} \sum_{i=1}^{n} \hat{w}_i \Big[y_{t,i} \log{(y_{p,i})} + (1 - y_{t,i}) \log{(1 - y_{p,i})} \Big] \, , \tag{7.2}$$

where the weights are given by (7.1) above.

7.2 Prediction of Obstructive Coronary Artery Disease

In this section, an initial DL algorithm for detection of obstructive CAD is presented. The work is published in Paper VIII [9]. A modified CNN with two input layers is presented, which aims to detect obstructive CAD from MPI examinations in each of three regions: LAD, LCx and RCA, see Figure 1.9 on page 13 and Figure 2.6 on page 27. For each examination, two polar images taken under stress are available as input; one in upright and one in supine position. To include other relevant and available information, such as angina symptoms and age, the design of the network was modified to add these scalar values as input.

Previous work in this field include [18, 19]. In [18] a multicentre dataset with 1638 patients from nine sites was used for development of a CNN for automatic prediction of obstructive disease from MPI to be compared with total perfusion deficit. The other paper, [19], uses a similar algorithm for another dataset with 1160 patients. These studies report an AUC of 0.76 on per-vessel and 0.80 on per-patient level in [18] and the corresponding numbers are 0.77 and 0.81 in [19].

In another work [4], a support vector machine was instead used to predict CAD. Since they were using this ML approach, there is a need to manually pick which features to use. The dataset used contained 1181 MPI studies. The authors did not look at the different regions individually, but instead only made a prediction on a per-patient level. For this they report an AUC of 0.92.

Finally, in [140] graph CNNs were used to take advantage of the fact that the intensity values are arranged on a polar grid. They used 946 labelled images from stress and rest and 4-fold cross-validation. Their proposed model achieves an agreement with the human observer with a sensitivity of 83.2% at a specificity of 70.8% when performed on a per-vessel level. They also performed localisation on a 17-segment division of the polar maps with good results.

7.2.1 Dataset

Part of the dataset introduced in Section 2.2 was used. Only data from 588 patients was available, whereof 294 were patients without disease. The data was split into five parts to ensure reliable results while getting the most out of the data, by using 5-fold cross-validation. The partitioning was done such that there was the same number of examples without disease in each fold but otherwise randomly. Possible artefacts were not considered in this work. Each example with disease could have positive results for multiple regions, thus the sum of LAD, RCA and LCx in each fold could be larger than the number of examples with disease. For details, see Table 7.1.

Table 7.1: Details of the dataset used. Each fold had 59 examples with disease and 59 examples without, except the last fold which instead had 58 examples of each kind. (©2021 IEEE)

	LAD	RCA	LCx
Fold 1	23	26	24
Fold 2	26	23	24
Fold 3	28	22	26
Fold 4	31	27	24
Fold 5	25	25	18
Total	133	123	116

The images from stress in up-right and supine position were used. The auxiliary parameters, introduced in Section 2.2, available were: gender, angina symptoms, age, BMI as well as AHA and ESC pre-test probability. For the patients with missing BMI value the average value of the remaining patients was used instead.

7.2.2 Data Preprocessing

For this work all images were available in artificial colouring. A majority of the images were also available in greyscale, where the intensity is directly related to the uptake of nuclide in the myocardium. Since the artificial colouring is only a visual representation of a one channel intensity image, and does not carry any additional information about the patient, the greyscale representation was preferred to use as input to the algorithm. The choice of creating the images in greyscale or colour was done in the QPS software, and at the time of the algorithm development all images were not available in greyscale. A transformation from colour to greyscale was hence needed and it was unfortunately unknown to us how the QPS software generated the colouring. Therefore the transformation, whose details are described in the following paragraphs, had to be learned from the examples where both colour and greyscale images were available.

The transformation was created by constructing a conversion map based on the 884 examples available as both colour and greyscale images (both stress and rest images were used). For each occurring RGB value in these images, the corresponding greyscale intensity was saved as a map, as shown in Figure 7.1a.

For an image available in colour only, this map was used in the following way to achieve a transformation to greyscale: If a pixel's RGB value existed in the map, the pixel was assigned the corresponding greyscale value. If it did not exist, the intensity of the closest RGB point, or the mean of the closest if multiple at the same distance, was used instead. A scatter plot of the RGB values of the 272 patients (543 images) existing in colour only can be seen in Figure 7.1b. The reason for why other RGB values than the ones in the original map (Figure 7.1a) exist in these images is as yet unknown. After transformation, no difference in visual appearance between the transformed colour-only images and the

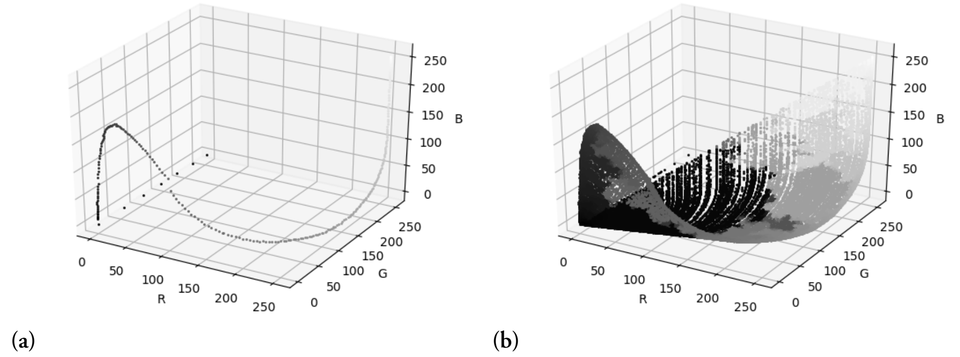


Figure 7.1: The map shown in (a) illustrates the greyscale pixel intensity for different RGB values occurring in the 884 images available in both greyscale and colour. This map was used to convert the remaining colour images into greyscale. The occurrence of RGB values in the images available in colour only, and the greyscale values they were mapped to, are shown as a scatter plot in (b). (©2021 IEEE)

greyscale images from the available colour-greyscale pairs could be seen. All images (with pixel values in the range [0,1]) were finally normalised to the range $[-1, 1]$, by multiplying by 2 and subtracting 1, for better training.

Some of the auxiliary parameters were modified by standardisation, in order to improve training of the neural networks. For the parameters gender, angina symptoms and AHA pre-test probability no modifications were made. Age, BMI and ESC pre-test probability were standardised by subtraction of the mean and division by the standard deviation, see Table 2.4 on page 26.

7.2.3 Augmentation

When training a DL algorithm, augmentation is often used to expand the training dataset in order to reduce the risk of overfitting the model. In [75] different augmentation techniques suitable for medical images are investigated. There are, however, a very limited amount of augmentation approaches that make sense for the MPI images, since the position and orientation of the polar map is fixed in the image. Furthermore, there is no noise or similar artefacts in the images. Therefore, to expand the training dataset the only augmentation techniques used for the images was to rotate by a small angle and use intensity clipping. Rotation of the images were done pairwise, for each pair of images associated with a patient, with a random angle in the range $[-10^\circ, 10^\circ]$ in each epoch. This angle was considered small enough to not alternate which region might be healthy or not, while still augmenting the images to prevent overtraining. Intensity clipping was done as described in Section 3.8, with parameters $t_1 = 0.05$ and $t_2 = 0.95$. An example of an image and one of its most

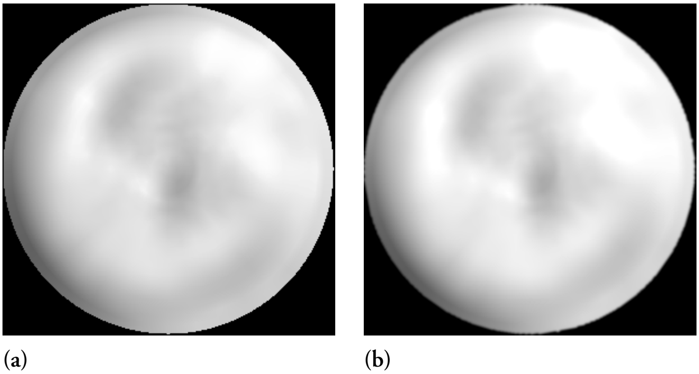


Figure 7.2: An example of a polar map without disease in (a) its original version and (b) rotated 10° counterclockwise as well as intensity clipped from 0.05 to 0.95. (©2021 IEEE)

extreme augmented versions can be seen in Figure 7.2. All augmentation was carried out before the images were normalised to $[-1, 1]$.

The auxiliary parameters were augmented, using the approaches described in Section 3.8, for each epoch in the following way. The gender as well as the angina symptoms were altered with a 20% probability. The AHA value was either increased by one if possible with a probability of one third, or similarly decreased by one if possible with the same probability, or unchanged. The standardised values of the age, BMI and pre-test probabilities were added a random value in the range $[-0.2, 0.2]$.

7.2.4 Neural Network

As input to the neural network the clinical intensity images from stress in up-right and supine position were used, without any information of the subdivision of the polar map. The two images were stacked to form a two-channel input image. As additional input, the six auxiliary parameters were given to further increase the algorithm's ability. The aim for the algorithm was to predict the probability for myocardial ischaemia for each of the three coronary arterial territories (LAD, RCA and LCx). Thus, this was treated as a multi-label problem, where multiple or none of the classes could be the expected output. Compare this with [58], where eight output nodes instead were used for the same task.

The design of the neural network can be seen in Figure 7.3. The main part consists of a CNN with filters of size 3×3 and 4×4 as well as 2×2 max pooling layers. The fully connected part of the network was modified, compared to traditional architectures, by adding a concatenation layer to include the auxiliary parameters. The CNN was tested both with and without this additional input, to investigate if it improved the performance.

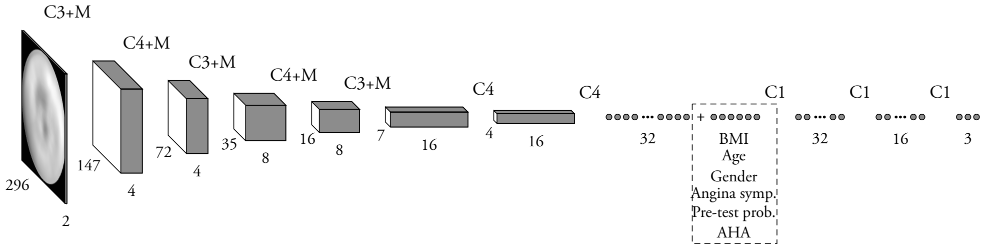


Figure 7.3: The design of the CNN. The spatial size and the number of channels for each layer can be seen in the figure. The auxiliary parameters, which could be excluded, are concatenated with the rest of the data when the spatial resolution is one, see the dashed box. Here *CX* denotes a convolution with an $X \times X$ filter, followed by a ReLU activation function. The symbol *M* denotes the occurrence of a 2 × 2 max pooling layer with stride 2. After the final *C*1 layers (i.e. fully connected), a sigmoid activation function is used. (©2021 IEEE)

The weighting of the loss was done as described in Section 7.1. The design is similar but deepened compared to [19].

7.2.5 Results

ROC curves were produced by varying the classification threshold of the output from the neural network. Based on these the AUCs were calculated. The accuracy was determined using the threshold 0.5. Both these measures were done both on per-vessel level, but also on per-patient level. The per-patient level results were achieved by merging the results of all regions as one result, either no disease or disease in at least one of the three regions. The algorithm was tested both with and without the auxiliary parameters, and both with and without augmentation. An overview of the AUC results and accuracy can be seen in Table 7.2. The results were similar for many of the configurations. Looking at the AUC results, the best performing configuration was achieved when using the auxiliary parameters and augmentation of the images, but with almost identical results when excluding the augmentation. When looking at the accuracy instead, the best result was achieved when including the auxiliary parameters and no augmentation was used. Overall, it seems that this configuration gave the best result. It was therefore stated that the best performing algorithm in this study was the one using the auxiliary parameters and trained without any augmentation. The corresponding ROC curve can be seen in Figure 7.4.

To investigate which of the auxiliary parameters gave the most information and affected the result most, each one of them was excluded one at a time when using the best performing configuration, i.e. no augmentation. The results can be seen in Table 7.3.

To compare with previous work the average accuracy was calculated, using a threshold of 0.5 for the classifier. Furthermore, the sensitivity when the specificity was 56% as well as 71% was determined. These results can be seen in Table 7.4.

Table 7.2: Results for different configurations of the algorithm; trained with and without augmentation and including or excluding the auxiliary parameters. Inclusion is denoted with an x and exclusion with a dash. The standard deviation is given in parenthesis. (©2021 IEEE)

Algorithm			AUC					Accuracy				
Incl.	Augm.											
Aux.	Im.	Aux.	LAD	RCA	LCx	Average	Patient	LAD	RCA	LCx	Average	Patient
-	-	-	.84 (.03)	.85 (.04)	.86 (.03)	.85 (.02)	.89 (.02)	.77 (.03)	.77 (.06)	.80 (.04)	.78 (.03)	.83 (.03)
-	x	-	.83 (.03)	.85 (.04)	.83 (.03)	.83 (.01)	.88 (.02)	.74 (.05)	.77 (.05)	.75 (.04)	.75 (.02)	.80 (.02)
x	-	-	.88 (.03)	**.89** (.01)	**.90** (.04)	**.89** (.02)	**.95** (.02)	.78 (.03)	**.80** (.02)	.83 (.04)	**.81** (.02)	**.89** (.03)
x	x	-	**.89** (.03)	**.89** (.03)	**.90** (.04)	**.89** (.02)	**.95** (.02)	**.79** (.04)	.79 (.03)	.82 (.04)	.80 (.01)	.88 (.03)
x	-	x	.88 (.03)	.88 (.03)	**.90** (.03)	**.89** (.01)	.94 (.01)	.78 (.02)	.79 (.03)	**.84** (.04)	.80 (.01)	**.89** (.01)
x	x	x	.86 (.03)	**.89** (.02)	.89 (.04)	.88 (.02)	.93 (.01)	.77 (.04)	**.80** (.04)	**.84** (.04)	.80 (.01)	.87 (.02)

Table 7.3: Results when including all the auxiliary parameters, and then excluding them one by one. No augmentation was used when training these algorithms. The standard deviation is given in parenthesis. (©2021 IEEE)

	AUC					Accuracy				
Aux. param.	LAD	RCA	LCx	Average	Patient	LAD	RCA	LCx	Average	Patient
Excluding all	.84 (.03)	.85 (.04)	.86 (.03)	.85 (.02)	.89 (.02)	.77 (.03)	.77 (.06)	.80 (.04)	.78 (.03)	.83 (.03)
Including all	.88 (.03)	.89 (.01)	.90 (.04)	.89 (.02)	.95 (.02)	.78 (.03)	.80 (.02)	.83 (.04)	.81 (.02)	.89 (.03)
Excl. gender	.87 (.04)	.89 (.01)	.89 (.05)	.88 (.02)	.94 (.02)	.77 (.05)	.80 (.04)	.81 (.06)	.80 (.04)	.87 (.02)
Excl. angina symp.	.87 (.03)	.87 (.01)	.88 (.05)	.88 (.02)	.93 (.02)	.78 (.02)	.78 (.03)	.81 (.05)	.79 (.02)	.87 (.02)
Excl. AHA	.86 (.03)	.89 (.03)	.88 (.05)	.88 (.02)	.94 (.02)	.78 (.03)	.79 (.04)	.79 (.05)	.79 (.02)	.87 (.02)
Excl. age	.86 (.05)	.90 (.02)	.89 (.04)	.88 (.02)	.94 (.02)	.78 (.02)	.82 (.01)	.81 (.03)	.80 (.02)	.88 (.02)
Excl. BMI	.88 (.04)	.89 (.02)	.89 (.03)	.88 (.02)	.95 (.02)	.77 (.03)	.80 (.03)	.81 (.03)	.79 (.01)	.87 (.02)
Excl. Pre-test prob.	.89 (.02)	.87 (.02)	.90 (.05)	.88 (.01)	.95 (.02)	.79 (.02)	.79 (.03)	.83 (.04)	.80 (.02)	.88 (.03)

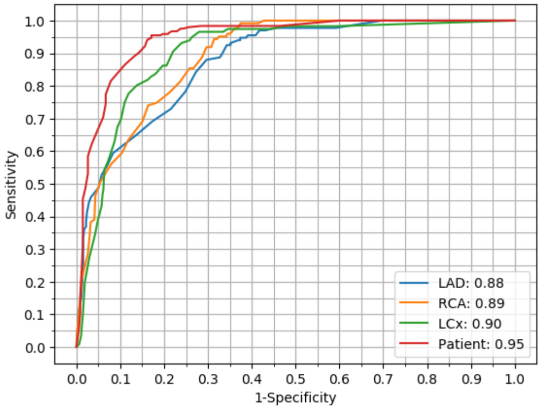


Figure 7.4: ROC curves for the three regions and per-patient level, using the best performing configuration, i.e. training without augmentation but including the auxiliary parameters. The curve was obtained by varying the classification threshold of the neural network. (©2021 IEEE)

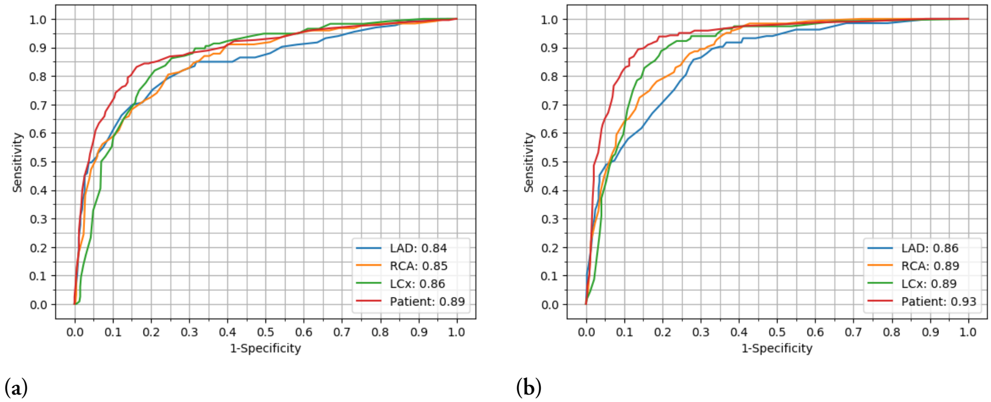


Figure 7.5: ROC curves from the algorithm (a) without auxiliary parameters and augmentation, as well as (b) with both auxiliary parameters and augmentation. (©2021 IEEE)

Table 7.4: Sensitivities for the different regions when the specificity is 56% and 71% respectively. The first specificity is stated to be the same as for a normal clinical read in [19]. The latter is the value reported by [140]. The algorithm trained without augmentation but including the auxiliary parameters is used. (©2021 IEEE)

Specificity	Sensitivity				
	LAD	RCA	LCx	Average	Patient
0.56	0.97	1.00	0.97	0.98	0.98
0.71	0.86	0.90	0.97	0.91	0.98

7.2.6 Discussion

The best performing configuration gives a maximum AUC of 0.95 on per-patient level and 0.89 on average per-vessel level. The values can be compared with [19], stating an AUC of 0.81 on per-patient level and 0.77 as average per-vessel level, and [4], stating an AUC of 0.92 on per-patient level. It should however be noted that, since different datasets have been used in the different studies, this might not be an entirely fair comparison. The corresponding ROC curve can be seen in Figure 7.4, showing that the algorithm is approximately equally good at predicting disease in each of the three regions although there is a performance drop for LAD. The same performance drop could be seen in all configurations of the algorithm, cf. Figure 7.5, indicating that either this region is harder to classify or that the available data is sub-optimal for this region.

Using the best performing configuration, an average accuracy of 81% for the three regions and 89% on patient level was achieved, using a threshold of 0.5 for the classifier. The individual accuracies for the three regions can be seen in Table 7.2. The results can be compared with the mean accuracy of 77% reported by [58], which however was determined based on only 16 test images.

The study in [140] report a sensitivity of 83% at a specificity of 71% on per-vessel level. For the same specificity, an average sensitivity of 91% on per-vessel level was achieved in this study when using the optimal configuration. In [19], the sensitivity for when the specificity is 56%, which is stated to be the specificity for a normal clinical read, is reported to be 85% for their DL algorithm. This should be compared with 83% for a clinical read. For the same specificity, the algorithm in this study with optimal configuration achieves a sensitivity of 98% on per-patient level and 98% on average per-vessel level. See Table 7.4 for complete results.

As can be seen by the results in Table 7.2, the auxiliary parameters certainly improved the performance. While this data gives a good indication on whether there is any disease, it does, however, not give any information about which of the regions that are affected or not. When excluding the auxiliary parameters one at a time (see Table 7.3), the results are almost identical independent of which parameter is excluded, with only a small performance drop compared to when all are included. It is therefore concluded that there is not one single parameter that is of most importance, but rather the combination of them.

The use of augmentation of the training data did not affect the performance that much. The motivation for using augmentation was to prevent the network from overfitting to the training data. To not distort the images in an unrealistic way, only small amounts of augmentations were used. This was probably the reason that the results were not affected more and unfortunately did not improve them.

7.3 Detection of Left Bundle Branch Block and Artefacts

There is a need for a tool which can facilitate the work for physicians detecting CAD. To make such a tool reliable, it is important that it can handle various artefacts. In this section Paper IX [10] is presented. It is a continuation of Paper VIII [9] presented in Section 7.2, but extended with a larger dataset and furthermore the detection of LBBB is added to the detection of obstructive CAD in each of the three major coronary arteries. No previous study investigating automatic detection of LBBB from MPI has been found. It is investigated whether individual prediction is beneficial, or if simultaneous prediction of both myocardial ischaemia, LBBB and additional artefacts is more efficient.

In this work a design for such an automatic tool based on DL, using images from MPI as well as auxiliary parameters is presented. More specifically, the aim is to estimate the probability of obstructive CAD in each of the main coronary artery regions (LAD, LCx and RCA), see Figure 1.9 on page 13. In addition, LBBB is also detected from the same polar maps, as it is a common artefact in the daily practice. LBBB is usually detected from an ECG examination. However, LBBB has a typical pattern in the MPI that is important to know since it can be the source of false positive results in CAD detection [145]. There are also two types of artefacts, from breast and diaphragm respectively, which can cause incorrect conclusions. See Figure 2.6 on page 27 for examples of CAD, LBBB and artefacts.

7.3.1 Dataset

The dataset described in Section 2.2 was used. Of the 759 patients, 387 were patients without disease and 66 were patients with LBBB. The data was split into five parts and stratified 5-fold cross-validation was used, see Table 7.5 for details. Each fold was constructed such that there was the same number of examples without disease and the same number of examples with LBBB in each fold, but otherwise randomly. For all patients, there could be artefacts from any combination of LBBB, breast and diaphragm. For a patient with disease, there could be positive result for multiple regions.

As in the previous section, images from stress in up-right and supine position were used together with the six auxiliary parameters (Table 2.4 on page 26). For patients where the BMI value was unknown, the average value of the remaining patients (BMI = 28) was used instead. Normalisation of the images and auxiliary parameters was done in the same way as described in Subsection 7.2.2.

Table 7.5: The partitioning obtained for the different diseases and artefacts when the dataset was split into five folds.

Fold	Total	Normal	LAD	RCA	LCx	LBBB	Breast	Diaphragm
1	152	78	30	26	20	13	11	12
2	151	77	36	31	23	13	11	18
3	153	78	26	33	28	14	9	18
4	151	77	32	26	23	13	20	13
5	152	78	30	25	29	13	12	13
Total	759	387	154	141	123	66	55	80

7.3.2 Method

All images were available in greyscale. The auxiliary parameters were standardised in the same way as described in the previous section in order to improve training of the neural networks. Since augmentation did not improve the results presented in the previous section, no augmentation was used in this work.

The network design and the training procedure is similar to what described in Paper VIII [9], but modified to include the additional output labels for LBBB and artefacts from breast or diaphragm. The neural network design for detection of myocardial ischaemia, LBBB and artefacts can be seen in Figure 7.6. The network has two input layers, one for the polar image pairs and another for the six auxiliary parameters and the output is prediction of myocardial ischaemia for each of the three coronary arterial territories (LAD, RCA and LCx), LBBB or artefacts from breast and diaphragm. Since multiple, or none, of these six labels could have positive results, this was treated as a multi-label problem. Both to predict all labels with the same network, but also to train networks to only predict one or a few of them, was tested. Therefore, the number of output nodes can vary between one and six. The loss weights described in Section 7.1 were used.

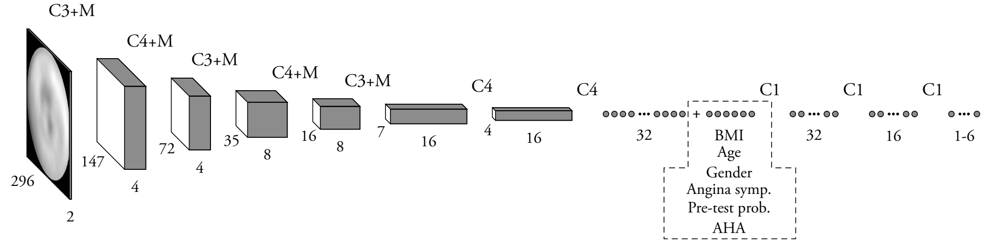


Figure 7.6: Illustration of the CNN with two input layers, using both images and the auxiliary parameters to detect CAD and LBBB. The following notations are used: *CX* denotes a convolution with an $X \times X$ filter, *M* denotes a 2 × 2 max pooling layer with stride 2. Each convolutional layer is followed by a ReLU activation, except the last one where sigmoid activation is used. The spatial size and the number of channels for each layer can also be seen.

7.3.3 Results

The algorithm was tested both with and without the auxiliary parameters, and the best results were achieved when including them, see Table 7.6. An overview of the AUC results when predicting different combinations of CAD and artefacts can be seen in Table 7.7. Figure 7.7 shows ROC curves for the network predicting both CAD in the three regions, LBBB and artefacts.

Table 7.6: AUC results when not detecting any of the artefacts, with and without the auxiliary parameters. The standard deviation is given in parenthesis.

Aux.	LAD	RCA	LCx	Average	Patient
Excl.	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.89 (.02)
Incl.	.89 (.02)	.88 (.03)	.90 (.04)	.89 (.02)	.94 (.01)

Table 7.7: AUC results for different configurations of the algorithm; including different labels (excluded labels are denoted with a dash). The auxiliary parameters were included in all of the configurations. The standard deviation is given in parenthesis.

LAD	RCA	LCx	Average	Patient	LBBB	Breast	Diaphragm
.89 (.02)	.88 (.03)	.90 (.04)	.89 (.02)	.94 (.01)	-	-	-
.88 (.02)	.88 (.03)	.90 (.04)	.89 (.02)	.94 (.01)	.82 (.09)	-	-
.87 (.03)	.90 (.03)	.91 (.03)	.89 (.03)	.94 (.01)	-	.68 (.09)	.64 (.09)
.87 (.02)	.88 (.03)	.90 (.03)	.89 (.02)	.93 (.01)	.80 (.07)	.62 (.07)	.63 (.07)
-	-	-	-	-	.82 (.06)	-	-
-	-	-	-	-	-	.64 (.04)	.60 (.07)
-	-	-	-	-	.74 (.07)	.63 (.08)	.57 (.06)

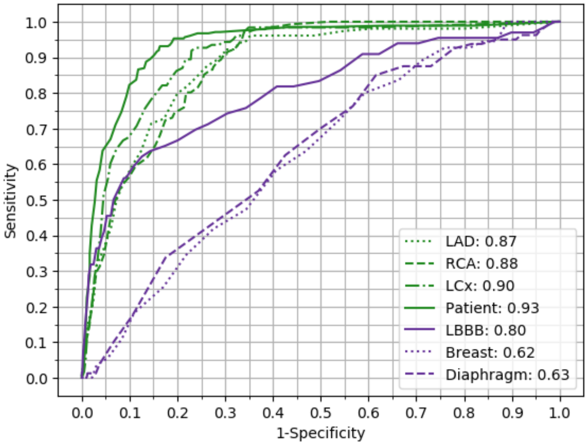


Figure 7.7: ROC curves for detection of CAD in the three regions and CAD at per-patient level as well as for detection of LBBB and the two artefacts.

7.3.4 Discussion

The best performing configuration for CAD detection gives a maximum AUC of 0.94 on per-patient level and 0.89 on average per-vessel level, about the same as in previous section which used a smaller dataset [9].

For LBBB the best results (AUC 0.82) are achieved when training the network for either prediction of only LBBB or for prediction of myocardial ischaemia as well. While these three additional labels seem to help the network to learn how to recognise LBBB, the two artefact labels breast and diaphragm did instead worsen the results (AUC 0.80 and 0.74, to be compared with 0.82).

Prediction of artefacts from breast and diaphragm did not improve the performance for prediction of the other labels. Overall, these predictions were not very successful; the best AUC results were 0.68 for breast and 0.64 for diaphragm. The reason for this could for example be that the artefacts where not clearly visible or that there was inconsistency in the gold standard.

7.4 Prediction of Quantitative Coronary Angiography Values

ANNs and MPI technology have evolved showing promising results for the diagnosis of CAD [4, 18, 19, 73, 140]. While efforts have been done for the diagnosis of CAD using ANNs and MPI, no previous work in the current scientific literature has been found attempting to predict QCA from MPI, an information which could give great aggregated value to the MPI study. Using ANNs the study's primary aim was to predict the degree of coronary artery stenosis from the main coronary artery regions — LAD, LCx and RCA — using stress MPIs in the upright and supine position, with QCA from ICA as reference. The study is contained in Paper X.

7.4.1 Dataset

Part of the dataset described in Section 2.2 was used. Subjects with LBBB, congenital heart disease and cardiac transplantation were excluded. Due to the risk of bronchospasm from regadenoson [127] all patients with known asthma or chronic obstructive lung disease were asked to perform a lung spirometry test if the vasodilator was going to be used for the MPI study. Consequently, all patients with a forced expiratory volume in 1 second <1L were excluded from the study. All patients who underwent ICA within six months after the MPI study were identified (N=354). From those 275 were evaluated with QCA and included in the study. For the control group 275 patients were included providing that

Table 7.8: Patients' characteristics. Values are n (%) or mean ± standard deviation.

	Development Cohort		Test Cohort	
	No Obstructive CAD	Obstructive CAD	No Obstructive CAD	Obstructive CAD
Number	250	250	25	25
0-vessel disease	250 (100%)	50 (20%)	25 (100%)	5 (20%)
1-vessel disease	-	96 (38%)	-	11 (44%)
2-vessel disease	-	62 (25%)	-	9 (36%)
3-vessel disease	-	42 (17%)	-	0 (0%)
Age, yrs	63.5±11.6	68.2±9.6	64.0±12.9	65.0±10.6
Male	144 (58%)	187 (75%)	19 (76%)	18 (72%)
Female	106 (42%)	63 (25%)	6 (24%)	7 (28%)
BMI	27.3±4.6	28.4±4.4	27.8±4.9	29.4±3.0
Exercise MPI	149 (60%)	116 (46%)	10 (40%)	11 (44%)
Pharmacological MPI	93 (37%)	106 (42%)	14 (56%)	13 (52%)
Combination of exercise and pharmacologic MPI	8 (3%)	28 (11%)	1 (4%)	1(4%)

they fulfilled the inclusion criteria, had no symptoms of angina pectoris according to the ESC pre-test probability estimate and had a clinical follow-up period of at least six months after the MPI study was performed, showing no cardiovascular events during this time.

All patients had MPI images taken from stress in upright and supine position. The following auxiliary parameters were used: BMI, age and gender, summarised in Table 7.8. In each of the eight cases where BMI values were missing, the average BMI value from the remaining patients was used instead. The data was split into a development and an evaluation cohort, consisting of 500 and 50 patients respectively with equal amounts of examples with disease and without in each part. The development cohort was divided into five parts and 5-fold cross-validation was used. The partitioning was done such that there were 50 pathological examples and 50 examples without disease in each fold, but otherwise randomly. The number of examples with stenosis in different intervals can be seen in Table 7.9.

ICA was routinely performed according to standard techniques. Per cent lumen area reductions due to intracoronary atheromatous plaques were first determined visually on end-diastolic frames and with the help of a QCA software (General Electric Advantage Work-

Table 7.9: Distribution of obstructed coronary arteries with QCA-values in different intervals.

	Development Cohort			Test Cohort		
QCA interval	LAD	RCA	LCx	LAD	RCA	LCx
50-59%	9	0	1	1	0	0
60-69%	17	9	7	4	0	3
70-79%	15	6	12	2	2	0
80-89%	33	24	28	1	2	2
90-100%	54	82	53	3	8	2

station, Cardiac X-Ray Applications, Stenosis Analysis v1.6) for stenosis visually assessed to be around the 50% threshold by an experienced angiographer physician. Where applicable, two separate measurements in orthogonal views of the same stenotic segment were obtained and values were averaged to represent an approximate measurement of the per cent (%) vessel area stenosis. Any stenosis $\geq$ 50% was considered significant and regarded as a positive QCA test. Total coronary vessel occlusions were marked as 100% lumen area stenosis. When no visible stenotic lumen was seen on angiography with a marginally patent vessel (with other than normal flow) the stenosis was also regarded as a total occlusion as well.

7.4.2 Method

The aim of the algorithm was to estimate the degree of coronary artery stenosis in each of the three main coronary arteries compared to QCA and evaluate the improvement of the algorithm when adding information regarding the auxiliary parameters. As input to the neural network the intensity images from stress in upright and supine position were used, stacked as a two-channel image. As additional input, the three auxiliary parameters BMI, age and gender could be included to evaluate if these parameters would improve the result. To handle the multiple independent classes, this was treated as a multi-label problem, where multiple or none of the classes could be the expected output.

The design of the neural network consists mainly of 3 × 3 and 4 × 4 convolutional layers as well as 2 × 2 max pooling layers and ReLU activations, see Figure 7.8. The convolutional part of the neural network, used to extract features from the images, is concatenated with a second input layer where the auxiliary parameters are introduced. The training procedure described in Section 7.1 was used. For the main task, to predict whether the stenosis was above the 50% threshold, one such network was trained. To further predict the per cent vessel area stenosis, identical networks were also trained to predict whether the stenosis was above the thresholds 60%, 70%, 80% and 90%.

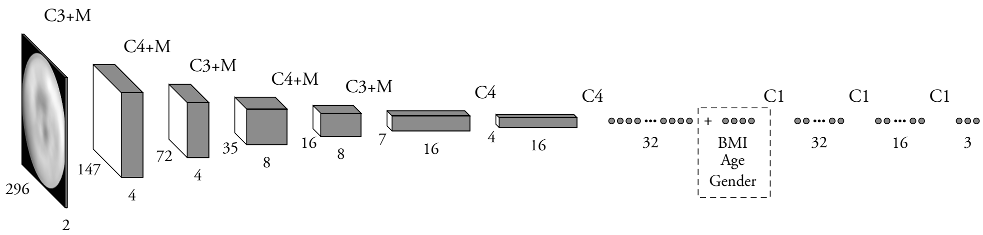


Figure 7.8: The CNN used for prediction of QCA value. Three auxiliary parameters are included by a concatenation layer. Here *CX* denotes a convolution with an $X \times X$ filter, followed by a ReLU activation function. The symbol M denotes the occurrence of 2 × 2 max pooling layers with stride 2. After the final *C*1 layers (i.e. fully connected), a sigmoid activation function is used.

The auxiliary parameters BMI and age were standardised by subtraction of the average and division by the standard deviation. Similar augmentation as in Paper VIII was used; the images were augmented randomly in each epoch, by rotating by an angle in the interval [-10°, 10°]. Intensity clipping and blurring was done as described in Section 3.8, with parameters $t_1 = 0.05, t_2 = 0.95, b = 1$. Furthermore, the auxiliary parameters were augmented in each epoch by altering the gender with a 20% probability and the standardised values (i.e. subtracted the mean and divided by the standard deviation) of the age and BMI were added a random value in the range [-0.2, 0.2].

7.4.3 Results

An overview of the AUC for the prediction of QCA, with a 50% narrowing of the artery, from the cross-validation both with and without the auxiliary parameters can be seen in Table 7.10. The results for other thresholds of the stenosis can be seen in Table 7.11. The corresponding results for the test cohort can be seen in Table 7.12 and Table 7.13. The ROC curves for the different regions and the thresholds 50%, 70% and 90% for the test cohort can be seen in Figure 7.9.

Table 7.10: AUC results from five-fold cross-validation when predicting quantitative coronary angiography (QCA>50%) excluding or including the auxiliary parameters for the different regions (LAD, RCA, LCx), the average for the three regions as well as combining all of them on per-patient level. The standard deviation is given in parentheses.

Settings	LAD	RCA	LCx	Average	Patient
No auxiliary param.	0.77 (0.02)	0.84 (0.04)	0.79 (0.05)	0.80 (0.02)	0.85 (0.03)
Incl. age, BMI, gender	0.75 (0.03)	0.84 (0.03)	0.81 (0.05)	0.80 (0.02)	0.86 (0.02)

Table 7.11: AUC results from five-fold cross-validation when predicting QCA-value for the different coronary artery regions (LAD, RCA, LCx), the average for the three regions as well as combining all of them on per-patient level. The standard deviation is given in parentheses. The auxiliary parameters were included.

Settings	LAD	RCA	LCx	Average	Patient
50%	0.75 (0.03)	0.84 (0.03)	0.81 (0.05)	0.80 (0.02)	0.86 (0.02)
60%	0.75 (0.04)	0.83 (0.06)	0.80 (0.06)	0.79 (0.04)	0.84 (0.03)
70%	0.77 (0.03)	0.82 (0.03)	0.80 (0.04)	0.79 (0.02)	0.85 (0.01)
80%	0.78 (0.03)	0.83 (0.05)	0.81 (0.06)	0.81 (0.03)	0.84 (0.04)
90%	0.80 (0.05)	0.77 (0.06)	0.83 (0.07)	0.80 (0.05)	0.82 (0.04)

Table 7.12: AUC results for the test cohort when predicting quantitative coronary angiography (QCA>50%) excluding or including the auxiliary parameters for the different regions (LAD, RCA, LCx), the average for the three regions as well as combining all of them on per-patient level. The standard deviation is given in parentheses.

Settings	LAD	RCA	LCx	Average	Patient
No auxiliary param.	0.60 (0.04)	0.81 (0.04)	0.87 (0.01)	0.76 (0.03)	0.83 (0.04)
Incl. age, BMI, gender	0.55 (0.03)	0.75 (0.03)	0.83 (0.02)	0.71 (0.02)	0.78 (0.02)

Table 7.13: AUC results for the test cohort when predicting QCA-value for the different coronary artery regions (LAD, RCA, LCx), the average for the three regions as well as combining all of them on per-patient level. The standard deviation is given in parentheses. The auxiliary parameters were included.

Settings	LAD	RCA	LCx	Average	Patient
50%	0.55 (0.03)	0.75 (0.03)	0.83 (0.02)	0.71 (0.02)	0.78 (0.02)
60%	0.61 (0.01)	0.78 (0.02)	0.84 (0.04)	0.74 (0.02)	0.84 (0.01)
70%	0.68 (0.04)	0.72 (0.01)	0.86 (0.04)	0.75 (0.01)	0.79 (0.02)
80%	0.76 (0.05)	0.86 (0.03)	0.83 (0.03)	0.82 (0.01)	0.85 (0.01)
90%	0.76 (0.09)	0.80 (0.03)	0.97 (0.04)	0.85 (0.03)	0.82 (0.03)

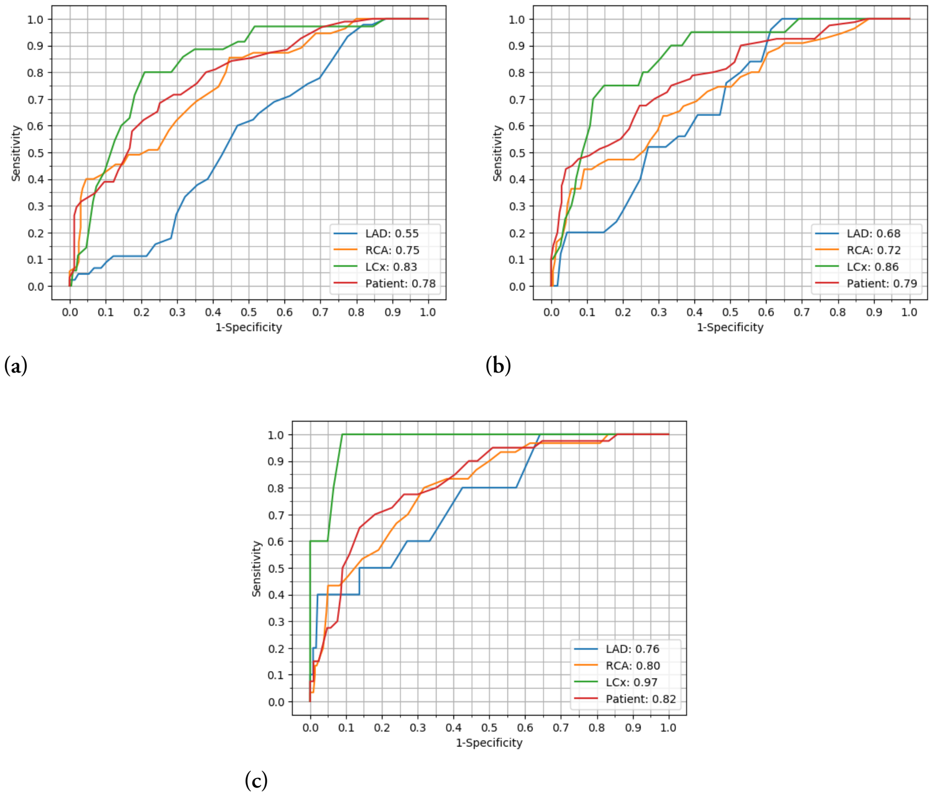


Figure 7.9: ROC curves when predicting whether the stenosis was (a) ≥50%, (b) ≥70% and (c) ≥90%, using the auxiliary parameters as well as the two MPI images as input to the CNN.

7.4.4 Discussion

In this study, ANNs have been created for automatic prediction of QCA from stress MPI polar maps and compared with QCA from ICA, which is the gold standard for diagnosis. The results demonstrate that DL achieves high diagnostic efficacy in predicting the percentage of coronary artery stenosis compared to QCA. While other studies have developed

ML for the prediction of obstructive CAD using MPI data [4, 18, 19, 58, 73, 140], no previous study with DL that works with the prediction of QCA from stress MPI has been found.

The implications of this finding for the healthcare system are sustainable since the information provided from the QCA, by means of the MPI using a non-invasive approach, could result in advantages like avoiding the ICA intervention. This could in turn be directly translated in less radiation exposure for the patients and fewer number of hospitalisations.

While the created DL algorithm performs very well when predicting the degree of artery stenosis compared to QCA from ICA, the algorithm does not perform as well in locating the affected coronary artery territory and the majority of the errors occur for patients with at least one narrowed vessel, where the wrong vessel or too many vessels are predicted positive by the algorithm. The performance in these cases could be improved by including more such examples when training the CNN, which unfortunately was not possible in this study.

A limitation of the study is the number of cases included (N=550). The data was collected from only one site. However, no larger cohort has been found reported in the literature to date performed in a D-SPECT cardiac camera compared with ICA [15, 16, 114, 122] and evaluated with QCA.

Regarding diagnostic purposes, ICA is only necessary in patients with suspected CAD in cases of inconclusive non-invasive testing. ICA may also be indicated if non-invasive assessment suggests high event risk for determination of options for revascularisation. Invasive functional assessment should complement ICA, especially in patients with coronary stenosis of 50-90% or multivessel disease, given the frequent mismatch between the angiographic and haemodynamic severities of coronary stenosis [88]. While the most common invasive measurement used in clinical praxis is Fractional Flow Reserve and some authors [33, 36, 69] have worked with ML and Fractional Flow Reserve by means of computed tomography, QCA has its advantages in terms of reproducibility [35, 115], being this work the first one that attempts to obtain this kind of information using MPI.

The developed algorithms can predict coronary artery stenosis of 50, 70 and 90% with promising accuracy (see Table 7.13). It could play an important role in the clinical scenario in the future if fully developed and clinically applied. Future prospective studies comparing the results from the QCA obtained from the MPI and Fractional Flow Reserve from ICA are warranted in order to establish the haemodynamic relevance of the information provided by the QCA values predicted from MPI.

Three different CNNs were used for prediction using different QCA thresholds. Instead of having three different models, a regression network could be used for the same task. It is however common with regression via classification, an approach that has been shown to possibly improve the accuracy compared to standard regression, using different binning variants of the target values [17]. Further development of the algorithm predicting QCA could include using such an approach.

7.5 Discussion of Automatic CAD Detection

In the three studies for automatic detection of CAD, algorithms have been developed to estimate the probability of obstructive CAD from MPI examinations. The input to the algorithms were two images from stress, one in upright and one in supine position, as well as other relevant auxiliary parameters; gender, angina symptoms, age, BMI and pre-test probability according to AHA and ESC. State of the art results were achieved with an AUC of 0.95 on per-patient level and average AUC for the regions of 0.89. Additionally, LBBB was detected from the same data with an AUC result of 0.82. The performance for LBBB detection was the same when the network was trained to detect myocardial ischaemia too, but was worsened when artefacts from breast and diaphragm also should be predicted. Furthermore, the achieved results showed that the auxiliary parameters provide important information and improve the performance. To investigate how much information could be extracted from the six auxiliary parameters, an additional network was trained using only those as input. The network was trained using the same dataset as used in Section 7.3, but the network was reduced such that only the part using the auxiliary parameters was used. The achieved results can be seen in Table 7.14, showing that the auxiliary parameters hold a lot of information about obstructive CAD and that they for this dataset were correlated with which artery had disease.

An algorithm which can estimate the percentage of coronary artery stenosis with promising results from MPI images was presented, compared to QCA from ICA. The auxiliary parameters were excluded in Paper X, since the aim was that the prediction should be possible without prior information. The result from cross-validation was better than the result on the test dataset, indicating that some overtraining occurred. More research is needed in order to establish the haemodynamic significance of the QCA values determined from MPI, develop and clinically validate this promising novel technology.

Due to the limited datasets available, data augmentation could have been an important approach to train a more reliable algorithm. However, it is hard to find suitable augmentation

Table 7.14: AUC results from neural networks using only auxiliary parameters (Aux.), only images (Im.) as well as both (Im. & Aux.) for prediction of CAD. The standard deviation is given in parenthesis.

Input	LAD	RCA	LCx	Average	Patient
Aux.	.79 (.03)	.81 (.03)	.85 (.02)	.82 (.02)	.88 (.02)
Im.	.86 (.02)	.86 (.02)	.86 (.02)	.86 (.02)	.89 (.02)
Im. & Aux.	.89 (.02)	.88 (.03)	.90 (.04)	.89 (.02)	.94 (.01)

techniques for this data and the ones tested in these studies did not improve the performance significantly. Another limitation due to the small dataset is that small networks are preferable to avoid overtraining. It is however likely that there is more information in the data that a larger network could have extracted.

One limitation of the studies is the patients for which the BMI value was missing. For those patients, the average BMI value was used instead. This meant that both training and evaluation of the networks was done with possibly incorrect data. To avoid this, one possibility would be to exclude those patients. But due to the limited dataset, this was not an appealing option. Another option would be to vary the BMI value for those patients while training, but this would not solve the problem when evaluating on a patient with unknown BMI. As the BMI was only missing from a small part of the patients, the possible errors this could cause were estimated to be minor and therefore the average BMI value was used instead.

Another possible limitation is that only low-probability patients were used as healthy cases. This to be sure that the ground truth was correct. However, this might make the classification too easy and the results better than they actually are. But since the classification was not only done as healthy patients versus patients with CAD the algorithm could not only rely on this. Instead, it had to predict which regions that were affected, were each patient with CAD could have one or two normal arteries.

The achieved results are very promising towards an automatic system, which would be of great use for the healthcare system. Future work would include to increase the dataset, train a larger network for possibly increased performance capability and additional evaluation.

Chapter 8

Concluding Remarks

Each study presented in this thesis has been followed by a discussion and a few conclusions. In this chapter, some additional aspects of DL and in particular its utilisation to medical applications will be discussed. Some final conclusions of the most important findings are also presented.

8.1 General Discussion

The field of DL is progressing impressively fast. Excellent DL software such as MATLABs Deep Learning Toolbox or the Python-library Keras make implementation of ANNs easy and fast. Furthermore, the increased computational power makes training faster. This increases the possibility to optimise for example hyperparameters, instead of relying on good guesses.

The main constraint when applying DL to a new field in medicine is nowadays typically not technology, but instead limited datasets. The field of medical image analysis using DL has more and more evolved into dealing with the datasets, collecting and handling those in the most efficient way. The fact that what works well under some circumstances might not work well in another environment, such as another hospital, is an important aspect of the research. There are still few publicly available datasets, which is a limitation for e.g. research reproducibility. One aspect which probably is limiting the available datasets is the sensitive personal data it might include and associated regulations. A related question is the ownership of the data. While one could argue that either the patient or the care provider is the owner, [93] suggests that when the data has been used for its primary purpose, to provide care, it should be used to benefit future patients. The datasets used in this thesis are unfortunately not publicly available but might be in the future.

When applying AI to healthcare, there are of course ethical perspectives to consider. The patients' medical treatment should not be affected negatively. However, developed countries struggle with an ageing population and increased demand for healthcare. If the possible improvements and efficiencies in healthcare are restrained, this might result in less healthcare for the population in the future. This raises the question whether it is ethically defensible not to develop and utilise AI tools for medical applications.

Another important aspect of AI in healthcare is when it should be considered good enough to be used. Should it be better than the average medical doctor or outperform even the best? The first alternative is probably achievable, while the latter would not only require reproducing the medical doctors work but also surpass the same. It might be possible to accomplish this, using application dependent approaches. If the ground truth can be achieved via some other examination, the DL system can be trained to reproduce it. This is the case for CAD predictions from MPI, where a ground truth is available from the ICA and QCA values. Another option is to use the outcome for the patients. This is seldom possible, since when a disease is detected, it is typically treated as soon as possible. With the PRIAS dataset, including patients with low-grade PCa and follow-up without treatment, this might be possible. Some of the cancers will progress while others will not. Both these approaches have the potential to capture important information that currently is unknown. However, this requires large datasets, which are hard to collect and often contain biases. In the case of PCa, it is in general a slow-growing cancer form, especially so in patients like those in the PRIAS cohort which have a mild version of the cancer. In order to get good follow-up data for PCa, one should scan slides from more than 10 years ago. But, as observed, this material is not of good quality due to faded staining. This means that datasets collected now, might be more valuable in many years' time from now. To use the outcome, progression (which in PCa is relatively slow) from many patients with disease is needed.

It is often discussed whether DL is a black box and therefore has unpredictable behaviour. However, a trained network contains no randomness even though the behaviour sometimes can be hard to explain. This fact is discussed in a blog post by Lundström [105], pointing out that there are several other technologies used in medicine which are similar in that aspect. For example, a computed tomography scanner is complex and few people can explain the full pipeline, although it does have a detailed description for each part. It is considered reliable, thanks to quality assurance where everything is thoroughly tested. In the same way, there will be a need to test AI systems carefully and continuously. However, it is also of uttermost importance that the medical doctors and the society as a whole trust the AI systems.

For correct evaluations of AI algorithms, many aspects should be investigated. For the case of Gleason grading, it is important that the evaluation data is obtained from different patients than the training data [118]. Furthermore, variations in appearance can be expected in images from different labs and different scanners. For fair conclusions the existing variability should be considered and future utilisation might require site specific and continuous evaluation, to handle these inevitable variations.

There are several possible ways the DL algorithms presented in this thesis could be used. To be used as sole autonomous systems is less probable in the near future. Instead, they can guide or assist medical doctors in many ways, such as highlighting severe cases or interesting areas for faster diagnosis or excluding the obviously healthy cases. In [142], breast cancer was detected by both unassisted pathologists and pathologists assisted by a DL algorithm. The comparison was done both in terms of time spent per image and sensitivity, with the conclusion that the DL assisted approach was superior in both aspects. The possibility to continuously optimise the algorithm based on feedback from the medical doctors, similarly to what was described in Section 6.1 but in an automatic way, is an appealing approach. This will however increase the requirements on continuous evaluation since the algorithm will not be static anymore. Thus, while DL has great potential for many medical applications and the possible methods of use are many, it needs to be handled with care.

8.2 Conclusion

Several different studies for generalisation of Gleason grading have been presented in Chapter 4. The three methods with most promising results were to use the encoding part of an autoencoder for efficient downsampling, training the classifier as a DANN or using extensive augmentation. A study comparing all the methods on the same dataset would be informative to draw final conclusions. The simplest but still well-performing approach with augmentation was chosen in the two evaluation studies presented in Chapter 6. The related aspects described in Chapter 5 have not been further investigated, since the semantic segmentation approach will most likely experience the same problem with stain differences as the Gleason grading algorithm and from the results presented in Chapter 6 the separate G5 detector does not seem necessary.

The comparison of the Gleason scoring algorithm and the two pathologists in Section 6.1 shows promising results and future usefulness for a Gleason scoring algorithm. There are both inter- and intra-observer variability and an automatic algorithm could be used to reduce these uncertainties or as a second opinion. The evaluation cohort was however quite limited. The second evaluation study, Section 6.2, was significantly larger but lacks detailed and repeated annotations. The results are nevertheless very promising for this dataset consisting of mainly patients with low-grade cancers, thus being hard to detect. The evaluation was in this case limited to only distinguishing between normal and cancer cases. For the future, it would be interesting to instead use this dataset to develop an algorithm which not only is trained on pathologists' annotations, but also on information on whether the patients' cancer progressed or not.

The three studies for detection of CAD presented in Chapter 7 are similar but with slightly different goals. The results for automatic detection of CAD from MPI are impressive. The fact that further information such as QCA values can be predicted with good performance based on MPI is thrilling and indicates future utility. Reducing ICA would save suffering for patients and free resources within the healthcare system. Further evaluation is however needed to ensure the findings.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] W. C. Allsbrook Jr, K. A. Mangold, M. H. Johnson, R. B. Lane, C. G. Lane, and J. I. Epstein. Interobserver reproducibility of Gleason grading of prostatic carcinoma: general pathologist. *Human pathology*, 32(1):81–88, 2001.

[3] S. Antti and K. Juhani. ESC 2019 guidelines for the diagnosis and management of chronic coronary syndromes. *Herz*, 45(5):409–420, 2020.

[4] R. Arsanjani, Y. Xu, D. Dey, V. Vahistha, A. Shalev, R. Nakanishi, S. Hayes, M. Fish, D. Berman, G. Germano, et al. Improved accuracy of myocardial perfusion SPECT for detection of coronary artery disease by machine learning in a large population. *Journal of Nuclear Cardiology*, 20(4):553–562, 2013.

[5] E. Arvaniti, K. S. Fricker, M. Moret, N. Rupp, T. Hermanns, C. Fankhauser, N. Wey, P. J. Wild, J. H. Rueschoff, and M. Claassen. Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports*, 8(1):1–11, 2018.

[6] I. Arvidsson, N. C. Overgaard, F.-E. Marginean, A. Krzyzanowska, A. Bjartell, K. Åström, and A. Heyden. Generalization of prostate cancer classification for multiple sites using deep learning. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 191–194. IEEE, 2018.

[7] I. Arvidsson, N. C. Overgaard, K. Åström, and A. Heyden. Comparison of different augmentation techniques for improved generalization performance for Gleason grading. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 923–927. IEEE, 2019.

[8] I. Arvidsson, N. C. Overgaard, A. Krzyzanowska, F.-E. Marginean, A. Simoulis, A. Bjartell, K. Åström, and A. Heyden. Domain-adversarial neural network for improved generalization performance of Gleason grade classification. In *Medical Imaging 2020: Digital Pathology*, volume 11320, page 1132016. International Society for Optics and Photonics, 2020.

[9] I. Arvidsson, N. C. Overgaard, A. Davidsson, J. Frias Rose, M. Ochoa Figueroa, K. Åström, and A. Heyden. Prediction of obstructive coronary artery disease from myocardial perfusion scintigraphy using deep neural networks. *IEEE 2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4442–4449, 2021.

[10] I. Arvidsson, N. C. Overgaard, A. Davidsson, J. F. Rose, K. Åström, M. O. Figueroa, and A. Heyden. Detection of left bundle branch block and obstructive coronary artery disease from myocardial perfusion scintigraphy using deep neural networks. In *Medical Imaging 2021: Computer-Aided Diagnosis*, volume 11597, page 115970N. International Society for Optics and Photonics, 2021.

[11] C. Avenel, A. Tolf, A. Dragomir, and I. B. Carlbom. Glandular segmentation of prostate cancer: an illustration of how the choice of histopathological stain is one key to success for computational pathology. *Frontiers in bioengineering and biotechnology*, 7:125, 2019.

[12] A. I. Baba and C. Câtoi. Tumor cell morphology. In *Comparative Oncology*. The Publishing House of the Romanian Academy, 2007.

[13] P. H. Bartels, H. G. Bartels, R. Montironi, P. W. Hamilton, and D. Thompson. Machine vision in the detection of prostate lesions in histologic sections. *Analytical and quantitative cytology and histology*, 20(5):358–364, 1998.

[14] B. E. Bejnordi, G. Litjens, N. Timofeeva, I. Otte-Höller, A. Homeyer, N. Karssemeijer, and J. A. van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2015.

[15] S. Ben-Haim, K. Kacperski, S. Hain, D. Van Gramberg, B. F. Hutton, K. Erlandsson, T. Sharir, N. Roth, W. A. Waddington, D. S. Berman, et al. Simultaneous dual-radionuclide myocardial perfusion imaging with a solid-state dedicated cardiac camera. *European journal of nuclear medicine and molecular imaging*, 37(9): 1710–1721, 2010.

[16] S. Ben-Haim, O. Almukhailed, J. Neill, P. Slomka, R. Allie, D. Shiti, D. S. Berman, and J. Bomanji. Clinical value of supine and upright myocardial perfusion imaging in obese patients using the D-SPECT camera. *Journal of Nuclear Cardiology*, 21(3): 478–485, 2014.

[17] A. Berg, M. Oskarsson, and M. O'Connor. Deep ordinal regression with label diversity. *arXiv preprint arXiv:2006.15864*, 2020.

[18] J. Betancur, F. Commandeur, M. Motlagh, T. Sharir, A. J. Einstein, S. Bokhari, M. B. Fish, T. D. Ruddy, P. Kaufmann, A. J. Sinusas, et al. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. *JACC: Cardiovascular Imaging*, 11(11):1654–1663, 2018.

[19] J. Betancur, L.-H. Hu, F. Commandeur, T. Sharir, A. J. Einstein, M. B. Fish, T. D. Ruddy, P. A. Kaufmann, A. J. Sinusas, E. J. Miller, et al. Deep learning analysis of upright-supine high-efficiency SPECT myocardial perfusion imaging for prediction of obstructive coronary artery disease: A multicenter study. *Journal of Nuclear Medicine*, 60(5):664–670, 2019.

[20] L. P. Bokhorst, R. Valdagni, A. Rannikko, Y. Kakehi, T. Pickles, C. H. Bangma, and M. J. Roobol. A decade of active surveillance in the PRIAS study: an update and evaluation of the criteria used to recommend a switch to active treatment. *European urology*, 70(6):954–960, 2016.

[21] C. Boran, E. Kandirali, F. Yilmaz, E. Serin, and M. Akyol. Reliability of the 34βE12, keratin 5/6, p63, bcl-2, and AMACR in the diagnosis of prostate carcinoma. *Urologic Oncology: Seminars and Original Investigations*, 29(6):614–623, 2011.

[22] W. Bulten, P. Bándi, J. Hoven, R. van de Loo, J. Lotz, N. Weiss, J. van der Laak, B. van Ginneken, C. Hulsbergen-van de Kaa, and G. Litjens. Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard. *Scientific reports*, 9(1):1–10, 2019.

[23] W. Bulten, H. Pinckaers, H. van Boven, R. Vink, T. de Bel, B. van Ginneken, J. van der Laak, C. Hulsbergen-van de Kaa, and G. Litjens. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020.

[24] W. Bulten, M. Balkenhol, J.-J. A. Belinga, A. Brilhante, A. Çakır, L. Egevad, M. Eklund, X. Farré, K. Geronatsiou, V. Molinié, et al. Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 34(3):660–671, 2021.

[25] R. Cao, A. M. Bajgiran, S. A. Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung. Joint prostate cancer detection and gleason score prediction in mp-MRI via FocalNet. *IEEE transactions on medical imaging*, 38(11):2496–2506, 2019.

[26] N. M. Carleton, G. Lee, A. Madabhushi, and R. W. Veltri. Advances in the computational and molecular understanding of the prostate cancer cell nucleus. *Journal of cellular biochemistry*, 119(9):7127–7142, 2018.

[27] B. R. Chaitman, M. G. Bourassa, K. Davis, W. J. Rogers, D. H. Tyras, R. Berger, J. Kennedy, L. Fisher, M. Judkins, M. Mock, et al. Angiographic prevalence of high-risk coronary artery disease in patient subsets (CASS). *Circulation*, 64(2):360–367, 1981.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[29] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[30] F. Chollet et al. Keras. `https://github.com/fchollet/keras`, 2015.

[31] F. Chowdhury, S. Vaidyanathan, M. Bould, J. Marsh, C. Trickett, K. Dodds, T. Clark, R. Sapsford, C. Dickinson, C. Patel, et al. Rapid-acquisition myocardial perfusion scintigraphy (MPS) on a novel gamma camera using multipinhole collimation and miniaturized cadmium–zinc–telluride (CZT) detectors: prognostic value and diagnostic accuracy in a 'real-world'nuclear cardiology service. *European Heart Journal–Cardiovascular Imaging*, 15(3):275–283, 2014.

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[33] A. Coenen, Y.-H. Kim, M. Kruk, C. Tesche, J. De Geer, A. Kurata, M. L. Lubbers, J. Daemen, L. Itu, S. Rapaka, et al. Diagnostic accuracy of a machine-learning approach to coronary computed tomographic angiography–based fractional flow reserve: result from the MACHINE consortium. *Circulation: Cardiovascular Imaging*, 11(6):e007217, 2018.

[34] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[35] C. Collet, M. J. Grundeken, T. Asano, Y. Onuma, W. Wijns, and P. W. Serruys. State of the art: coronary angiography. *EuroIntervention: journal of EuroPCR in collaboration with the Working Group on Interventional Cardiology of the European Society of Cardiology*, 13(6):634–643, 2017.

[36] J. De Geer, M. Sandstedt, A. Björkholm, J. Alfredsson, M. Janzon, J. Engvall, and A. Persson. Software-based on-site estimation of fractional flow reserve using standard coronary CT angiography data. *Acta Radiologica*, 57(10):1186–1192, 2016.

[37] O. J. del Toro, M. Atzori, S. Otálora, M. Andersson, K. Eurén, M. Hedlund, P. Rönnquist, and H. Müller. Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade Gleason score. In *Medical Imaging 2017: Digital Pathology*, volume 10140, page 101400O. International Society for Optics and Photonics, 2017.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[39] G. A. Diamond and J. S. Forrester. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. *New England Journal of Medicine*, 300(24): 1350–1358, 1979.

[40] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[41] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE transactions on biomedical engineering*, 59(5):1205–1218, 2010.

[42] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan. Medical image segmentation based on u-net: A review. *Journal of Imaging Science and Technology*, 64(2):20508–1, 2020.

[43] V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

[44] W. L. Duvall, L. B. Croft, E. S. Ginsberg, A. J. Einstein, K. A. Guma, T. George, and M. J. Henzlova. Reduced isotope dose and imaging time with a high-efficiency CZT SPECT camera. *Journal of Nuclear Cardiology*, 18(5):847–857, 2011.

[45] W. L. Duvall, J. M. Sweeny, L. B. Croft, M. H. Barghash, N. K. Kulkarni, K. A. Guma, and M. J. Henzlova. Comparison of high efficiency CZT SPECT MPI to coronary angiography. *Journal of Nuclear Cardiology*, 18(4):595–604, 2011.

[46] L. Egevad, B. Delahunt, J. R. Srigley, and H. Samaratunga. International society of urological pathology (ISUP) grading of prostate cancer — an ISUP consensus on contemporary grading, 2016.

[47] L. Egevad, B. Delahunt, D. M. Berney, D. G. Bostwick, J. Cheville, E. Comperat, A. J. Evans, S. W. Fine, D. J. Grignon, P. A. Humphrey, et al. Utility of pathology

imagebase for standardisation of prostate cancer grading. *Histopathology*, 73(1):8–18, 2018.

[48] J. I. Epstein. An update of the Gleason grading system. *The Journal of urology*, 183 (2):433–440, 2010.

[49] J. I. Epstein. International society of urological pathology (ISUP) grading of prostate cancer: author's reply. *The American journal of surgical pathology*, 40(6):862–864, 2016.

[50] J. I. Epstein, M. J. Zelefsky, D. D. Sjoberg, J. B. Nelson, L. Egevad, C. Magi-Galluzzi, A. J. Vickers, A. V. Parwani, V. E. Reuter, S. W. Fine, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. *European urology*, 69(3):428–435, 2016.

[51] F. P. Esteves, J. R. Galt, R. D. Folks, L. Verdes, and E. V. Garcia. Diagnostic performance of low-dose rest/stress Tc-99m tetrofosmin myocardial perfusion SPECT using the 530c CZT camera: quantitative vs visual analysis. *Journal of Nuclear Cardiology*, 21(1):158–165, 2014.

[52] M. Fiechter, J. R. Ghadri, S. M. Kuest, A. P. Pazhenkottil, M. Wolfrum, R. N. Nkoulou, R. Goetti, O. Gaemperli, and P. A. Kaufmann. Nuclear myocardial perfusion imaging with a novel cadmium-zinc-telluride detector SPECT/CT device: first validation versus invasive coronary angiography. *European journal of nuclear medicine and molecular imaging*, 38(11):2025, 2011.

[53] M. Fiechter, C. Gebhard, T. A. Fuchs, J. R. Ghadri, J. Stehli, E. Kazakauskaite, B. A. Herzog, A. P. Pazhenkottil, O. Gaemperli, and P. A. Kaufmann. Cadmium-zinc-telluride myocardial perfusion imaging in obese patients. *Journal of Nuclear Medicine*, 53(9):1401–1406, 2012.

[54] S. D. Fihn, J. M. Gardin, J. Abrams, K. Berra, J. C. Blankenship, A. P. Dallas, P. S. Douglas, J. M. Foody, T. C. Gerber, A. L. Hinderliter, et al. 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease: a report of the american college of cardiology foundation/american heart association task force on practice guidelines, and the american college of physicians, american association for thoracic surgery, preventive cardiovascular nurses association, society for cardiovascular angiography and interventions, and society of thoracic surgeons. *Journal of the American College of Cardiology*, 60(24):e44–e164, 2012.

[55] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[56] J. Folmsbee, S. Johnson, X. Liu, M. Brandwein-Weber, and S. Doyle. Fragile neural networks: the importance of image standardization for deep learning in digital pathology. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 1095613. International Society for Optics and Photonics, 2019.

[57] H. Fox. Is H&E morphology coming to an end? *Journal of clinical pathology*, 53 (1):38–40, 2000.

[58] H. Fujita, T. Katafuchi, T. Uehara, and T. Nishimura. Application of artificial neural network to computer-aided diagnosis of coronary artery disease in myocardial SPECT bull's-eye images. *Journal of Nuclear Medicine*, 33(2):272–276, 1992.

[59] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[60] P. Garrone, G. Biondi-Zoccai, I. Salvetti, N. Sina, I. Sheiban, P. R. Stella, and P. Agostoni. Quantitative coronary angiography in the current era: principles and applications. *Journal of interventional cardiology*, 22(6):527–536, 2009.

[61] R. J. Gibbons, G. J. Balady, J. W. Beasley, J. T. Bricker, W. F. Duvernoy, V. F. Froelicher, D. B. Mark, T. H. Marwick, B. D. McCallister, P. D. Thompson, et al. ACC/AHA guidelines for exercise testing: a report of the american college of cardiology/american heart association task force on practice guidelines (committee on exercise testing). *Journal of the American College of Cardiology*, 30(1):260–311, 1997.

[62] P. Gjertsson, M. Lomsky, J. Richter, M. Ohlsson, D. Tout, A. Van Aswegen, R. Underwood, and L. Edenbrandt. The added value of ECG-gating for the diagnosis of myocardial infarction using myocardial perfusion scintigraphy and artificial neural networks. *Clinical physiology and functional imaging*, 26(5):301–304, 2006.

[63] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[64] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[65] A. Gummeson. Prostate cancer classification using convolutional neural networks. *Master's Theses in Mathematical Sciences*, 2016.

[66] A. Gummeson, I. Arvidsson, M. Ohlsson, N. C. Overgaard, A. Krzyzanowska, A. Heyden, A. Bjartell, and K. Åström. Automatic Gleason grading of H and E stained microscopic prostate images using deep convolutional neural networks. In

Medical Imaging 2017: Digital Pathology, volume 10140, page 101400S. International Society for Optics and Photonics, 2017.

[67] M. N. Gurcan, L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009.

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

[69] H. S. Hecht, J. Narula, and W. F. Fearon. Fractional flow reserve and coronary computed tomographic angiography: a review and critical analysis. *Circulation research*, 119(2):300–316, 2016.

[70] C. Hindorf, J. Oddstig, F. Hedeer, M. J. Hansson, J. Jögi, and H. Engblom. Importance of correct patient positioning in myocardial perfusion SPECT when using a CZT camera. *Journal of Nuclear Cardiology*, 21(4):695–702, 2014.

[71] M. Horvath, 2006. URL `https://commons.wikimedia.org/wiki/File:RGBCube_b.svg`.

[72] M. Horvath, 2010. URL `https://en.wikipedia.org/wiki/HSL_and_HSV#/media/File:HSV_color_solid_cone_chroma_gray.png`.

[73] L.-H. Hu, J. Betancur, T. Sharir, A. J. Einstein, S. Bokhari, M. B. Fish, T. D. Ruddy, P. A. Kaufmann, A. J. Sinusas, E. J. Miller, et al. Machine learning predicts per-vessel early coronary revascularization after fast myocardial perfusion SPECT: results from multicentre REFINE SPECT registry. *European Heart Journal-Cardiovascular Imaging*, 21(5):549–559, 2020.

[74] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[75] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin. Differential data augmentation techniques for medical imaging classification tasks. In *AMIA Annual Symposium Proceedings*, volume 2017, page 979. American Medical Informatics Association, 2017.

[76] F. Hyafil, A. Gimelli, R. H. Slart, P. Georgoulias, C. Rischpler, M. Lubberink, R. Sciagra, J. Bucerius, D. Agostini, and H. J. Verberne. EANM procedural guidelines for myocardial perfusion scintigraphy using cardiac-centered gamma cameras. *European Journal of Hybrid Imaging*, 3(1):1–27, 2019.

[77] A. Ibrahim, P. Gamble, R. Jaroensri, M. M. Abdelsamea, C. H. Mermel, P.-H. C. Chen, and E. A. Rakha. Artificial intelligence in digital breast pathology: Techniques and applications. *The Breast*, 49:267–273, 2020.

[78] H. Irshad, A. Veillard, L. Roux, and D. Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a review — current status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014.

[79] J. Isaksson. Recognizing microscopic structures: Dense semantic segmentation of multiple histopathological classes using fully convolutional neural networks. *Master's Theses in Mathematical Sciences*, 2016.

[80] J. Isaksson, I. Arvidsson, K. Åström, and A. Heyden. Semantic segmentation of microscopic images of H&E stained prostatic tissue using CNN. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1252–1256. IEEE, 2017.

[81] P. Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11 (2):37–50, 1912.

[82] H. Källén. *Applications of Machine Vision—Quality Control, Cancer Detection and Traffic Surveillance*. PhD thesis, Lund University, 2016.

[83] H. Källén, J. Molin, A. Heyden, C. Lundström, and K. Åström. Towards grading Gleason score using generically trained deep convolutional neural networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1163–1167. IEEE, 2016.

[84] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International conference on information processing in medical imaging*, pages 597–609. Springer, 2017.

[85] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

[86] A. M. Khan, N. Rajpoot, D. Treanor, and D. Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014.

[87] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[88] J. Knuuti, W. Wijns, A. Saraste, D. Capodanno, E. Barbato, C. Funck-Brentano, E. Prescott, R. F. Storey, C. Deaton, T. Cuisset, et al. 2019 ESC guidelines for the diagnosis and management of chronic coronary syndromes. *European heart journal*, 41(3):407, 2020.

[89] T. K. Koo and M. Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[90] O. Kott, D. Linsley, A. Amin, A. Karagounis, C. Jeffers, D. Golijanin, T. Serre, and B. Gershman. Development of a deep learning algorithm for the histopathologic diagnosis and Gleason grading of prostate cancer biopsies: a pilot study. *European urology focus*, 2019.

[91] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25: 1097–1105, 2012.

[92] M. W. Lafarge, J. P. Pluim, K. A. Eppenhof, P. Moeskops, and M. Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer, 2017.

[93] D. B. Larson, D. C. Magnus, M. P. Lungren, N. H. Shah, and C. P. Langlotz. Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology*, 295(3):675–682, 2020.

[94] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[95] R. M. Levenson, E. A. Krupinski, V. M. Navarro, and E. A. Wasserman. Pigeons (columba livia) as trainable observers of pathology and radiology breast cancer images. *PLOS ONE*, 10(11):e0141357, 2015.

[96] J. Li, W. Speier, K. C. Ho, K. V. Sarma, A. Gertych, B. S. Knudsen, and C. W. Arnold. An EM-based semi-supervised deep learning approach for semantic segmentation of histopathological images from radical prostatectomies. *Computerized Medical Imaging and Graphics*, 69:125–133, 2018.

[97] G. Lippolis. *Image analysis of prostate cancer tissue biomarkers*. PhD thesis, Lund University, 2015.

[98] G. Lippolis, A. Edsjö, L. Helczynski, A. Bjartell, and N. C. Overgaard. Automatic registration of multi-modal microscopy images for integrative analysis of prostate tissue sections. *BMC cancer*, 13(1):1–11, 2013.

[99] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-van de Kaa, P. Bult, B. van Ginneken, and J. van der Laak. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6, 2016.

[100] G. Litjens, F. Ciompi, J. M. Wolterink, B. D. de Vos, T. Leiner, J. Teuwen, and I. Išgum. State-of-the-art deep learning in cardiovascular image analysis. *JACC: Cardiovascular Imaging*, 12(8 Part 1):1549–1565, 2019.

[101] Y. Liu, K. Gadepalli, M. Norouzi, G. E. Dahl, T. Kohlberger, A. Boyko, S. Venugopalan, A. Timofeev, P. Q. Nelson, G. S. Corrado, et al. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442*, 2017.

[102] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[103] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[104] M. Lucas, I. Jansen, C. D. Savci-Heijink, S. L. Meijer, O. J. de Boer, T. G. van Leeuwen, D. M. de Bruin, and H. A. Marquering. Deep learning for automatic Gleason pattern classification for grade group determination of prostate biopsies. *Virchows Archiv*, 475(1):77–83, 2019.

[105] C. Lundström. AIDA blog: AI's black box: Trick or treat?, 2019. URL `https://medtech4health.se/aida-blog-ais-black-box-trick-or-treat/`.

[106] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[107] F. Marginean, I. Arvidsson, A. Simoulis, N. C. Overgaard, K. Åström, A. Heyden, A. Bjartell, and A. Krzyzanowska. An artificial intelligence–based support tool for automation and standardisation of Gleason grading in prostate biopsies. *European Urology Focus*, 2020.

[108] K. O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological methods*, 1(1):30, 1996.

[109] G. Montalescot, U. Sechtem, S. Achenbach, F. Andreotti, C. Arden, A. Budaj, R. Bugiardini, F. Crea, T. Cuisset, C. D. Mario, et al. 2013 ESC guidelines on the management of stable coronary artery disease: the task force on the management of

stable coronary artery disease of the european society of cardiology. *European heart journal*, 34(38):2949–3003, 2013.

[110] R. Montironi, R. Mazzuccheli, M. Scarpelli, A. Lopez-Beltran, G. Fellegara, and F. Algaba. Gleason grading of prostate cancer in needle biopsies or radical prostatectomy specimens: contemporary approach, current clinical significance and sources of pathology discrepancies. *BJU international*, 95(8):1146–1152, 2005.

[111] K. Nagpal, D. Foote, Y. Liu, P.-H. C. Chen, E. Wulczyn, F. Tan, N. Olson, J. L. Smith, A. Mohtashamian, J. H. Wren, et al. Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ digital medicine*, 2(1):1–10, 2019.

[112] K. Nagpal, D. Foote, F. Tan, Y. Liu, P.-H. C. Chen, D. F. Steiner, N. Manoj, N. Olson, J. L. Smith, A. Mohtashamian, et al. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA oncology*, 6(9):1372–1380, 2020.

[113] R. Nakazato, B. K. Tamarappoo, X. Kang, A. Wolak, F. Kite, S. W. Hayes, L. E. Thomson, J. D. Friedman, D. S. Berman, and P. J. Slomka. Quantitative upright–supine high-speed SPECT myocardial perfusion imaging for detection of coronary artery disease: correlation with invasive coronary angiography. *Journal of Nuclear Medicine*, 51(11):1724–1731, 2010.

[114] R. Nakazato, P. J. Slomka, M. Fish, R. G. Schwartz, S. W. Hayes, L. E. Thomson, J. D. Friedman, M. Lemley, M. L. Mackin, B. Peterson, et al. Quantitative high-efficiency cadmium-zinc-telluride SPECT with dedicated parallel-hole collimation system in obese patients: results of a multi-center study. *Journal of Nuclear Cardiology*, 22(2):266–275, 2015.

[115] V. G. Ng and A. J. Lansky. Novel QCA methodologies and angiographic scores. *The international journal of cardiovascular imaging*, 27(2):157–165, 2011.

[116] M. K. K. Niazi, K. Yao, D. L. Zynger, S. K. Clinton, J. Chen, M. Koyutürk, T. LaFramboise, and M. Gurcan. Visually meaningful histopathological features for automatic grading of prostate cancer. *IEEE journal of biomedical and health informatics*, 21(4):1027–1038, 2017.

[117] G. Nir, S. Hor, D. Karimi, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, R. S. Wilson, et al. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical image analysis*, 50:167–180, 2018.

[118] G. Nir, D. Karimi, S. L. Goldenberg, L. Fazli, B. F. Skinnider, P. Tavassoli, D. Turbin, C. F. Villamil, G. Wang, D. J. Thompson, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA network open*, 2(3):e190442–e190442, 2019.

[119] F. Nudi, A. E. Iskandrian, O. Schillaci, M. Peruzzi, G. Frati, and G. Biondi-Zoccai. Diagnostic accuracy of myocardial perfusion imaging with CZT technology: systemic review and meta-analysis of comparison with invasive coronary angiography. *JACC: Cardiovascular Imaging*, 10(7):787–794, 2017.

[120] C. Olsson, O. Enqvist, and F. Kahl. A polynomial-time bound for matching and registration with outliers. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[121] B. Pang, Y. Zhang, Q. Chen, Z. Gao, Q. Peng, and X. You. Cell nucleus segmentation in color histopathological imagery using convolutional networks. In *2010 Chinese Conference on Pattern Recognition (CCPR)*, pages 1–5. IEEE, 2010.

[122] M. Perrin, W. Djaballah, F. Moulin, M. Claudin, N. Veran, L. Imbert, S. Poussier, O. Morel, C. Besseau, A. Verger, et al. Stress-first protocol for myocardial perfusion SPECT imaging with semiconductor cameras: high diagnostic performances with significant reduction in patient radiation doses. *European journal of nuclear medicine and molecular imaging*, 42(7):1004–1011, 2015.

[123] J. Persson, U. Wilderäng, T. Jiborn, P. N. Wiklund, J.-E. Damber, J. Hugosson, G. Steineck, E. Haglind, and A. Bjartell. Interobserver variability in the pathological assessment of radical prostatectomy specimens: Findings of the laparoscopic prostatectomy robot open (LAPPRO) study. *Scandinavian journal of urology*, 2014.

[124] R. G. Pontius Jr and M. Millones. Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429, 2011.

[125] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

[126] S. G. Rathod, D. G. Jaiswal, and R. S. Bindu. Diagnostic utility of triple antibody (AMACR, HMWCK and p63) stain in prostate neoplasm. *Journal of family medicine and primary care*, 8(8):2651, 2019.

[127] E. Reyes and S. R. Underwood. Regadenoson myocardial perfusion scintigraphy for the evaluation of coronary artery disease in patients with lung disease: A series of five cases. *Journal of Nuclear Cardiology*, 27(1):315–321, 2020.

[128] E. Richardson and Y. Weiss. The surprising effectiveness of linear unsupervised image-to-image translation. *arXiv preprint arXiv:2007.12568*, 2020.

[129] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, J. Teuwen, M. Broeders, G. Gennaro, P. Clauser, T. H. Helbich, M. Chevalier, T. Mertelmeier, et al. Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. *European radiology*, 29(9): 4825–4832, 2019.

[130] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[131] A. C. Ruifrok, D. A. Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[132] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[133] C. Senaras, B. Sahiner, G. Tozbikian, G. Lozanski, and M. N. Gurcan. Creating synthetic digital slides using conditional generative adversarial networks: application to Ki67 staining. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058103. International Society for Optics and Photonics, 2018.

[134] A. Sethi, L. Sha, A. R. Vahadane, R. J. Deaton, N. Kumar, V. Macias, and P. H. Gann. Empirical comparison of color normalization methods for epithelial-stromal classification in H and E images. *Journal of pathology informatics*, 7, 2016.

[135] R. B. Shah and M. Zhou. *Prostate biopsy interpretation: An illustrated guide*. Springer, 2019.

[136] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[137] A. Shrivastava, W. Adorno, L. Ehsan, S. A. Ali, S. R. Moore, B. C. Amadi, P. Kelly, S. Syed, and D. E. Brown. Self-attentive adversarial stain normalization. *arXiv preprint arXiv:1909.01963*, 2019.

[138] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[139] P. J. Slomka, H. Nishina, D. S. Berman, C. Akincioglu, A. Abidov, J. D. Friedman, S. W. Hayes, and G. Germano. Automated quantification of myocardial perfusion SPECT using simplified normal limits. *Journal of nuclear cardiology*, 12(1):66–77, 2005.

[140] N. Spier, S. Nekolla, C. Rupprecht, M. Mustafa, N. Navab, and M. Baust. Classification of polar maps from cardiac perfusion imaging with graph-convolutional neural networks. *Scientific reports*, 9(1):1–8, 2019.

[141] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021.

[142] D. F. Steiner, R. MacDonald, Y. Liu, P. Truszkowski, J. D. Hipp, C. Gammage, F. Thng, L. Peng, and M. C. Stumpe. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology*, 42(12):1636, 2018.

[143] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, D. G. Bostwick, A. J. Evans, D. J. Grignon, P. A. Humphrey, et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2):222–232, 2020.

[144] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 2021.

[145] E. Surkova, L. P. Badano, R. Bellu, P. Aruta, F. Sambugaro, G. Romeo, F. Migliore, and D. Muraru. Left bundle branch block: from cardiac mechanics to clinical and diagnostic challenges. *Ep Europace*, 19(8):1251–1271, 2017.

[146] N. Suzuki, T. Asano, G. Nakazawa, J. Aoki, K. Tanabe, K. Hibi, Y. Ikari, and K. Kozuma. Clinical expert consensus document on quantitative coronary angiography from the japanese association of cardiovascular intervention and therapeutics. *Cardiovascular intervention and therapeutics*, 35(2):105–116, 2020.

[147] E. Svanemur. A prospective study of the outcome of prostate cancer patients under active surveillance. *Scientific Project, Faculty of Medicine and Health Sciences, Linköping University*, 2020.

[148] J. Swartling, J. Axelsson, G. Ahlgren, K. M. Kälkner, S. Nilsson, S. Svanberg, K. Svanberg, and S. Andersson-Engels. System for interstitial photodynamic therapy with online dosimetry: first clinical experiences of prostate cancer. *Journal of biomedical optics*, 15(5):058003, 2010.

[149] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[150] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[151] K. Tägil, M. Bondouy, J. Chaborel, W. Djaballah, P. Franken, S. Grandpierre, B. Hesse, M. Lomsky, P. Marie, T. Poisson, et al. A decision support system improves the interpretation of myocardial perfusion imaging. *European journal of nuclear medicine and molecular imaging*, 35(9):1602–1607, 2008.

[152] K. Tall. Automatic Gleason classification of prostate cancer-classification of small regions. *Master's Theses in Mathematical Sciences*, 2018.

[153] K. Tall, I. Arvidsson, N. C. Overgaard, K. Åström, and A. Heyden. Automatic detection of small areas of Gleason grade 5 in prostate tissue using CNN. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560E. International Society for Optics and Photonics, 2019.

[154] D. Tellez, M. Balkenhol, N. Karssemeijer, G. Litjens, J. van der Laak, and F. Ciompi. H and E stain augmentation improves generalization of convolutional networks for histopathological mitosis detection. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 105810Z. International Society for Optics and Photonics, 2018.

[155] D. Tellez, M. Balkenhol, I. Otte-Höller, R. van de Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[156] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, and J. van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis*, 58:101544, 2019.

[157] R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[158] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[159] J. S. Uebersax. Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological bulletin*, 101(1):140, 1987.

[160] A. Vahadane, T. Peng, A. Sethi, S. Albarqouni, L. Wang, M. Baust, K. Steiger, A. M. Schlitter, I. Esposito, and N. Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

[161] A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[162] World Health Organization. Fact sheet: Cardiovascular diseases (CVDs). `https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)`, 2017.

[163] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[164] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.