LUND UNIVERSITY

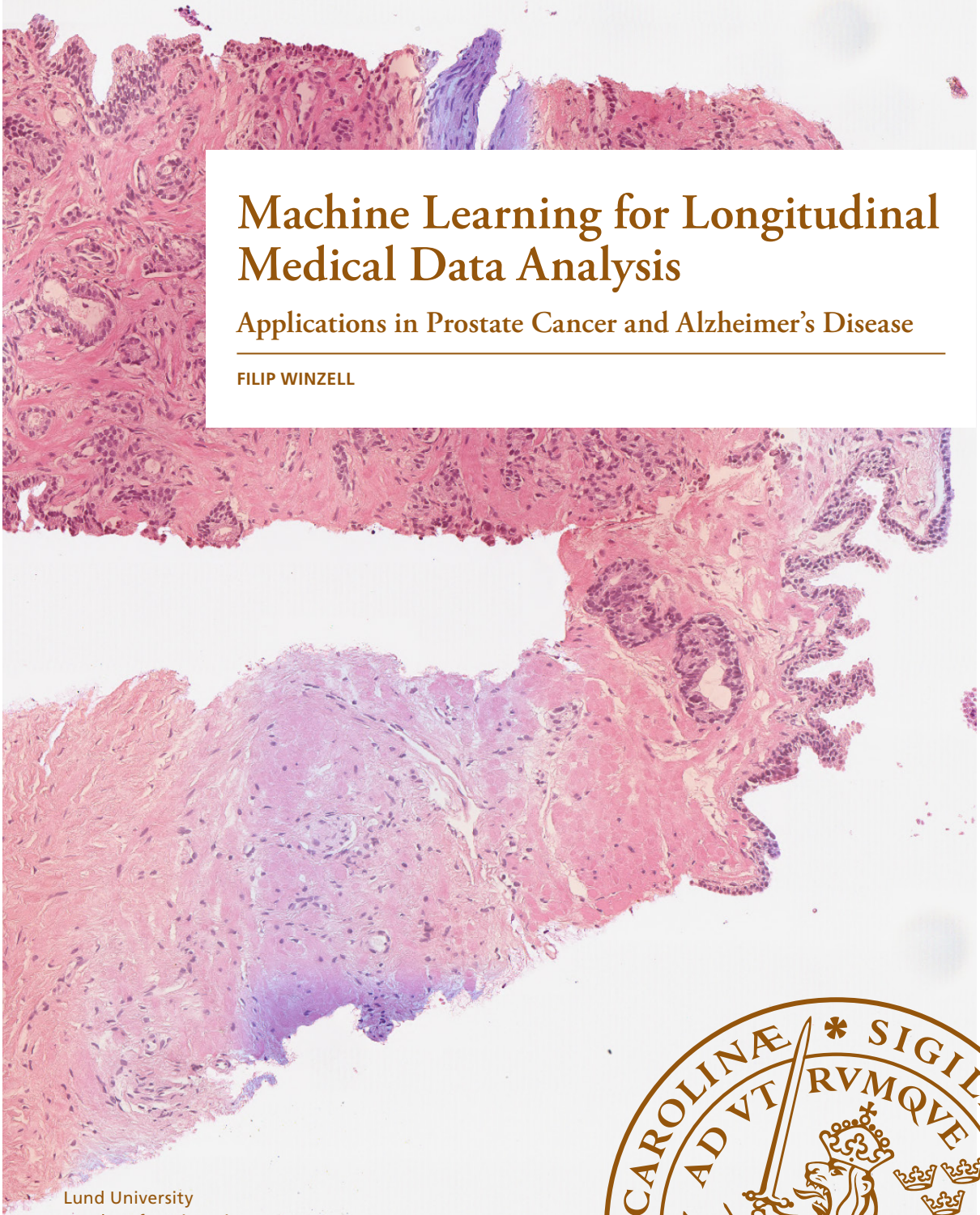
Machine Learning for Longitudinal Medical Data Analysis

Applications in Prostate Cancer and Alzheimer's Disease

Winzell, Filip

2026

– CENTRUM SCIENTIARUM MATHEMATICARUM –



Machine Learning for Longitudinal Medical Data Analysis

Applications in Prostate Cancer and Alzheimer's Disease

FILIP WINZELL

Lund University
Faculty of Engineering
Centre for Mathematical Sciences
Mathematics

Machine Learning for Longitudinal Medical Data Analysis

Applications in Prostate Cancer and Alzheimer's Disease

by Filip Winzell

LUND
UNIVERSITY

ACADEMIC THESIS

which, with due permission of the Faculty of Engineering at Lund University, will be publicly defended on Friday 12th of June, 2026, at 13:00 in lecture hall MH:Hörmander, Centre for Mathematical Sciences, Märkesbacken 4, Lund, for the degree of Doctor of Philosophy in Engineering.

Thesis advisors:
Prof. Anders Heyden, Dr. Ida Arvidsson,
Assoc. Prof. Niels-Christian Overgaard, Prof. Karl Åström

Faculty opponent:
Prof. Rasmus Larsen, Technical University of Denmark (DTU), Denmark

Organization
LUND UNIVERSITY
Centre for Mathematical Sciences
Box 118
SE–221 00 LUND
Sweden

Document name
Doctoral thesis

Date of presentation
2026-06-12

Sponsoring organization
WASP-DDLS, EDAP (EDAP2023-153244)

Author(s)
Filip Winzell

Title and subtitle
Machine Learning for Longitudinal Medical Data Analysis – Applications in Prostate Cancer and Alzheimer's Disease

Abstract

The healthcare systems of today are facing large challenges, with increasing amounts of patients and overworked hospital staff. Prostate cancer and Alzheimer's disease are two of the most prevalent diseases, with incidence numbers expected to rise over the coming decade. Medical imaging plays a central role in current diagnostic procedures for these diseases, enabling the potential use of machine-learning-based image analysis. Overall, computer-aided diagnostics has seen a big increase in research over the last decade, with numerous applications where AI-based methods perform at the same level as experienced physicians. However, with the computational power available today, AI-based methods have the potential to achieve more in terms of prognostication and early detection of diseases. This is something that would be of high value for the treatment of the aforementioned diseases. Thus, the topic of this thesis is to investigate machine learning-based methods for the analysis of longitudinal medical imaging data, with the goal of improving the diagnostic procedures of prostate cancer and Alzheimer's disease.

Currently, there are no general screening programs for prostate cancer, despite the importance of early detection for successful treatment. Screening based on blood measures of prostate-specific antigen (PSA) has been shown to reduce mortality but also significantly increase the levels of over-treatment. To mitigate that, active surveillance has been suggested as an alternative to radical treatment following abnormal PSA values, where patients are monitored with recurring examinations. When a patient is deemed to have a high-risk prostate cancer, as defined by the Gleason grading scale, they receive treatment. However, the Gleason grading system is a subjective grading system with proven inter-observer variability. In this thesis, a method to predict the longitudinal treatment decision of prostate cancer patients on active surveillance is presented. It is based on the popular attention-based multiple instance learning framework, in combination with the state-of-the-art foundation model for pathology called UNI. This model achieved promising results, indicating that it is possible to reliably predict the onset of prostate cancer earlier than trained pathologists.

Alzheimer's disease is a neurodegenerative disease, characterized by an abnormal accumulation of amyloid-β and tau proteins in the brain. The cause of the disease is still unknown and the progression is highly heterogeneous across a population of patients. While it remains incurable, there are recently developed treatments that have been shown to effectively slow down the progression if the disease is detected early. Hence, to better understand why certain patients progress differently than others and how to detect them early is of high interest. In this thesis we developed an algorithm to find patterns of brain atrophy in Alzheimer's disease patients and connect them to other biological abnormalities. We found four distinct subtypes, that could explain parts of this highly complex disease.

Key words
Longitudinal data, Machine Learning, Computational pathology, Prostate cancer, Alzheimer's disease

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language
English

ISSN and key title
1404-0034

ISBN
978-91-90202-04-3 (print)
978-91-90202-05-0 (electronic)

Recipient's notes

Number of pages
xxii+206

Price

Security classification

Machine Learning for Longitudinal Medical Data Analysis

Applications in Prostate Cancer and Alzheimer's Disease

by Filip Winzell

LUND
UNIVERSITY

Cover illustrations: Heatmap of computed attention weights over a prostate biopsy

Centre for Mathematical Sciences
Lund University
Box 118
SE-221 00 Lund
Sweden
www.maths.lu.se

Doctoral Theses in Mathematical Sciences 2026:8
ISSN: 1404-0034
ISBN: 978-91-90202-04-3 (print)
ISBN: 978-91-90202-05-0 (electronic)
LUTFMA-1004-2026

Printed in Sweden by Media-Tryck, Lund University, Lund 2026

Dedicated to my family
Maria, *Peter* and *Emil*

Abstract

The healthcare systems of today are facing large challenges, with increasing amounts of patients and overworked hospital staff. Prostate cancer and Alzheimer's disease are two of the most prevalent diseases, with incidence numbers expected to rise over the coming decade. Medical imaging plays a central role in current diagnostic procedures for these diseases, enabling the potential use of machine-learning-based image analysis. Overall, computer-aided diagnostics has seen a big increase in research over the last decade, with numerous applications where AI-based methods perform at the same level as experienced physicians. However, with the computational power available today, AI-based methods have the potential to achieve more in terms of prognostication and early detection of diseases. This is something that would be of high value for the treatment of the aforementioned diseases. Thus, the topic of this thesis is to investigate machine learning-based methods for the analysis of longitudinal medical imaging data, with the goal of improving the diagnostic procedures of prostate cancer and Alzheimer's disease.

Currently, there are no general screening programs for prostate cancer, despite the importance of early detection for successful treatment. Screening based on blood measures of prostate-specific antigen (PSA) has been shown to reduce mortality but also significantly increase the levels of over-treatment. To mitigate that, active surveillance has been suggested as an alternative to radical treatment following abnormal PSA values, where patients are monitored with recurring examinations. When a patient is deemed to have a high-risk prostate cancer, as defined by the Gleason grading scale, they receive treatment. However, the Gleason grading system is a subjective grading system with proven inter-observer variability. In this thesis, a method to predict the longitudinal treatment decision of prostate cancer patients on active surveillance is presented. It is based on the popular attention-based multiple instance learning framework, in combination with the state-of-the-art foundation model for pathology called UNI. This model achieved promising results, indicating that it is possible to reliably predict the onset of prostate cancer earlier than trained pathologists.

Alzheimer's disease is a neurodegenerative disease, characterized by an abnormal accumulation of amyloid-β and tau proteins in the brain. The cause of the disease is still unknown and the progression is highly heterogeneous across a population of patients. While it remains incurable, there are recently developed treatments that have been shown to effectively slow down the progression if the disease is detected early. Hence, to better understand why certain patients progress differently than others and how to detect them early is of high interest. In this thesis we developed an algorithm to find patterns of brain atrophy in Alzheimer's disease patients and connect them to other biological abnormalities. We found four distinct subtypes, that could explain parts of this highly complex disease.

Popular summary

The artificial intelligence (AI) boom of recent years has led to an exponential development of AI-based methods for analyzing images with numerous applications, such as self-driving cars, face recognition, and image generation. Since we can use these techniques to make a car recognize surrounding objects, what potential do they have for computer-aided diagnosis based on medical images?

Imaging has become a central part of today's healthcare. The development of X-rays, MRI, ultrasound, and other imaging modalities enabled the fundamental opportunity of looking inside the body. Furthermore, modern digital microscopy techniques allow us to study extracted cells at high-resolution. In combination with advancements in computational infrastructure, this has transformed the diagnostic procedures to rely more on the collection, analysis, and storage of medical image data.

Two of the most common diseases of today, where the diagnostic procedure is largely based on imaging, are prostate cancer and Alzheimer's disease. The number of patients with these diseases is increasing every year, and it is expected to rise even further in the future. Additionally, the success of treatment for these diseases is highly dependent on early diagnosis. Several studies have presented promising results of AI-based methods for automated diagnosis at similar levels of performance as trained physicians. By processing vast amounts of data at superhuman speed, AI can enable faster processing of patients and wider screening programs, to detect these diseases at earlier stages than ever before. A key concept of a screening program is the follow-up of patients, where the same type of examination is repeated over several years to track the progression of the disease for each individual. The collection of this type of data is often referred to as *longitudinal data*. A limitation of previous AI-based diagnostic methods is that they can only predict the instantaneous diagnosis. With the use of longitudinal data, we can study the patients long before the diagnosis was set and thereby enable the development of super-human performance in forecasting the progression. In this thesis, that is what we aim to investigate in the context of prostate cancer and Alzheimer's disease.

Prostate cancer is commonly diagnosed with biopsies of the prostate. This means that a sample of cells is extracted from the prostate, which is subsequently studied through a microscope. The tissues are colored with a fluorescent chemical compound, and with the use of a digital microscope, high-resolution images of the tissues can be captured (see Fig. 1 (C)). In these images, regions of cancer can be identified as glands with irregular shapes and uncontrolled growth. The tissues are graded according to the Gleason grading scale, ranging from benign to grade 5 prostate cancer. The highest grade is associated with the highest risk and worst prognosis. This grading system provides the basis for the diagnosis and treatment decision. However, it is also known to be subjective and that the assigned

grades on a case can vary a lot between different pathologists. Hence, training models to predict these grades is not ideal. A better solution could be to predict the outcomes of prostate cancer by utilizing longitudinal data. For example, to answer: ”given this prostate biopsy, what is the probability that this patient will get prostate cancer in 5 years?”. This is a more challenging task for the models to solve, since it requires a deeper understanding of the entire biopsy. To predict the Gleason grade in a location, the model only needs to consider this specific location. To predict the outcome, the whole biopsy needs to be considered. This is illustrated in Fig. 2.

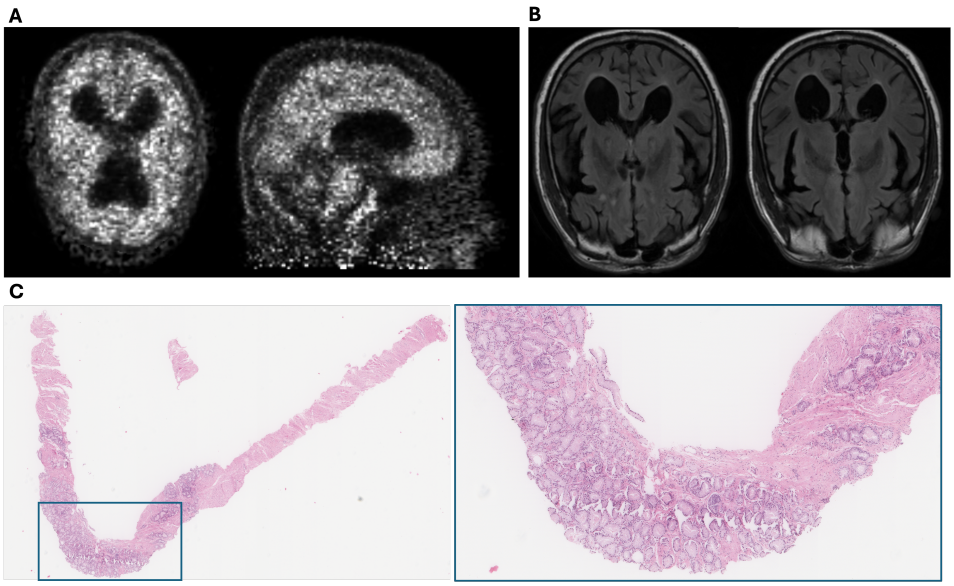


Figure 1: Examples of medical images used in this thesis. **A** Coronal and sagittal views of a positron emission tomography (PET) scan of a brain, showing accumulation of proteins in the brain due to Alzheimer's disease. **B** Two examples of MRIs from patients with Alzheimer's disease. **C** An image of a prostate biopsy, with a zoomed-in region.

Alzheimer's disease is the most common form of dementia, accounting for 60-80% of all cases. The main signs are an abnormal accumulation of certain proteins in the brain and subsequent loss of neurons in the brain, causing symptoms such as loss of memory and other cognitive difficulties. Hence, the diagnosis is based on multiple examinations such as standardized cognitive tests, protein measurements in the cerebrospinal fluid or the blood, and various brain imaging techniques (see Fig. 1 (A-B)). There is strong evidence that the biological alterations occur decades before any clear symptoms can be observed. However, the expression and progression of the disease can differ greatly from person to person, and what causes these differences is still unknown. The cognitive symptoms are naturally connected to loss of neurons in specific regions of the brain. Therefore, discovering different patterns of neurodegeneration and connecting these to other biological alterations could be a key to explaining this complex disease.

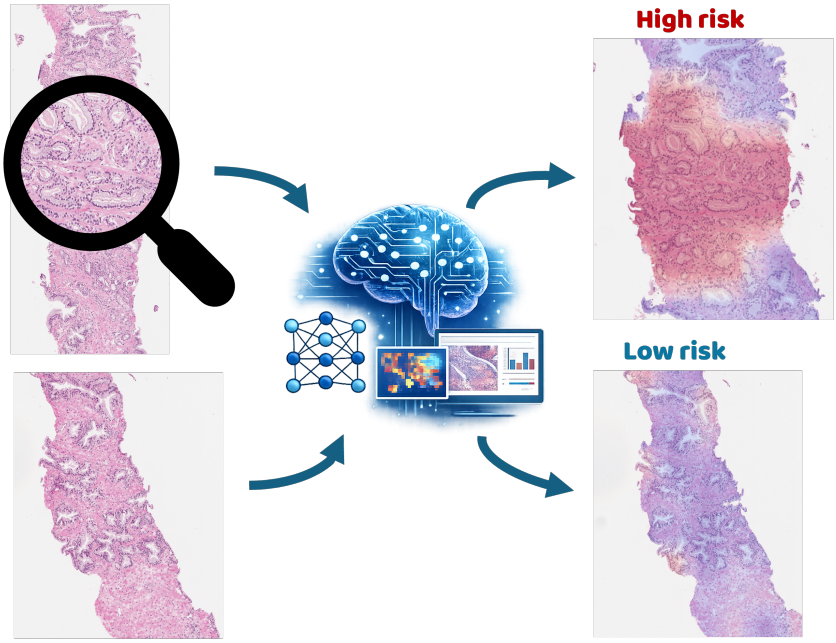


Figure 2: Illustration of an AI-based algorithm for classifying prostate biosies as high or low risk of prostate cancer.

In this thesis, we explore AI-based methods for analysis of longitudinal medical imaging data to enable earlier detection of prostate cancer and Alzheimer's disease. Specifically, we developed an AI-based framework for longitudinal outcome prediction of prostate cancer patients with promising results. We showed that with this method, we can make better long-term estimations than standard Gleason grading, additionally, our algorithm can process images at super-human speed. The benefits of this are two-fold; firstly, it enables finding high-risk prostate cancer patients at an early stage and secondly, it can reduce the risk of treating low-risk patients unnecessarily. However, more work is needed to verify the results on larger, diverse datasets. We also developed an algorithm based on statistical models to find patterns of longitudinal neurodegeneration in Alzheimer's disease. We found distinct types, with varying speed of progression. This shows that it is possible to identify patients with a higher risk of developing the disease before any symptoms. In recent years, several pharmaceuticals have been developed to slow the progression. By identifying the patients who will benefit the most from these drugs, we can avoid unnecessary treatment and take a step towards stopping and curing Alzheimer's disease.

Populärvetenskaplig sammanfattning

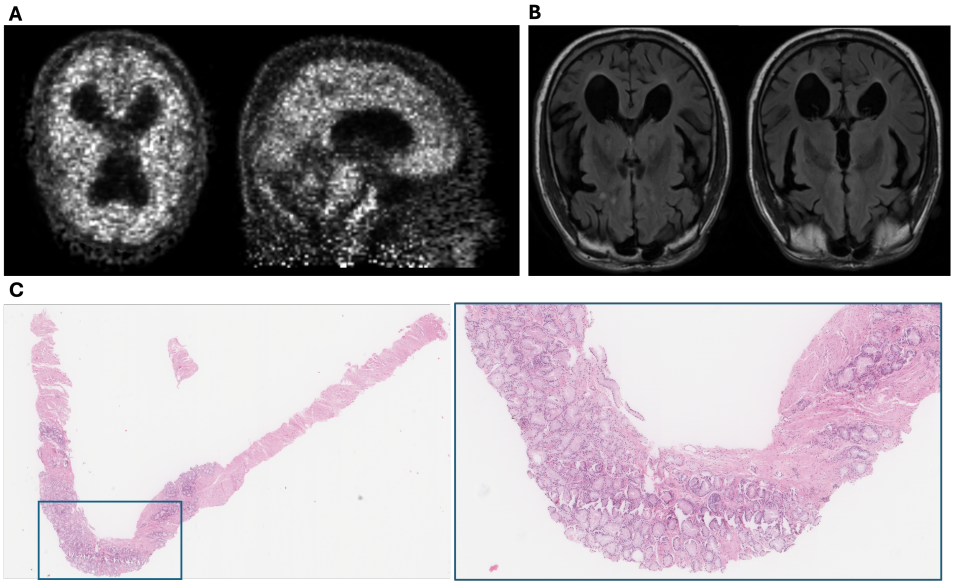
De senaste årens snabba utveckling inom artificiell intelligens (AI) har drivit fram en kraftig ökning av AI-baserade metoder för bildanalys, med tillämpningar inom bland annat själv-körande fordon, ansiktsigenkänning och bildgenerering. Eftersom vi med denna teknik kan få en bil att se omgivande hinder, vilken potential finns det då för datorstödd diagnosticering baserad på medicinska bilder?

Bildbehandling har blivit en central del av dagens hälso- och sjukvård. Utvecklingen av röntgen, MR, ultraljud och andra bildgivande system har gett oss den grundläggande möjligheten att se inuti kroppen. Dessutom, med moderna digitala mikroskop kan vi studera cellprover med hög upplösning. I kombination med bättre datorkraft har detta omvandlat diagnostiska procedurer till att förlita sig mer på insamling, analys och lagring av medicinska bilder.

Två av de vanligaste sjukdomarna idag, där diagnosticeringen till stor del baseras på bildanalys, är prostatacancer och Alzheimers sjukdom. Antalet patienter med dessa sjukdomar ökar varje år och de förväntas fortsätta öka i framtiden. Dessutom är behandlingen av dessa sjukdomar starkt beroende av en tidig diagnos. Flera studier har presenterat lovande resultat av AI-baserade metoder för automatiserad diagnostik med prestanda på liknande nivåer som läkare. Med AI kan stora mängder data analyseras fort, vilket möjliggör ett snabbare patientflöde och därmed bredare screening för att upptäcka dessa sjukdomar tidigare än någonsin förut. Ett screeningprogram innefattar ofta uppföljning av patienter, där samma typ av undersökning upprepas under flera år för att följa sjukdomsförloppet för varje individ. Insamlingen av denna typ av data kallas ofta för *longitudinell data*. En begränsning med tidigare AI-baserade metoder är att de endast kan estimera den omedelbara diagnosen. Med longitudinell data kan vi följa patienter långt före diagnos och därigenom utveckla metoder som gör det möjligt att förutsäga prognoser tidigare än vad en läkare kan göra. I denna avhandling undersöker vi detta för prostatacancer och Alzheimers sjukdom.

Prostatacancer diagnositiceras oftast genom vävnadsprover från prostatan. Dessa prover färgas med ett färgämne som synliggör cellerna och studeras med ett ljusmikroskop. Med digitala mikroskop kan hög-upplösta bilder skapas (se Fig. 1 (C)). I bilderna kan regioner av cancer identifieras som körtlar med oregelbundna former som växer okontrollerat. Vävnaden graderas enligt Gleason skalan från benign vävnad till grad 5, som är associerad med högst risk och sämre prognos. Behandlingsbeslutet baseras till stor del på Gleason graderingen. Dock är det välkänt att graderingen är subjektiv och att det finns en stor variation i bedömmningen som görs av olika patologer. Därför är det inte optimalt att träna AI modeller för att prediktera dessa. En bättre lösning hade varit att prediktera behandlingsbeslutet m.h.a. longitudinell data. Till exempel, för att svara på frågan: ”givet detta vävnadsprovet, vad är sannolikheten att denna patient kommer behöva få behandling för

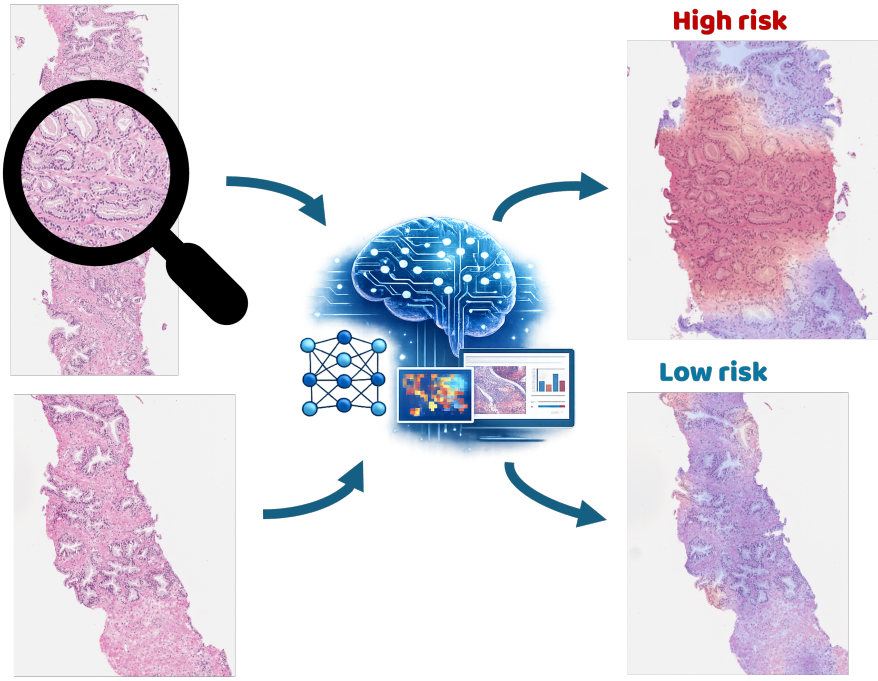
sin prostat cancer inom 5 år?”. Detta är en mer utmananade uppgift för modellerna att lösa eftersom den kräver en djupare kontext. Tillskillnad mot Gleason graderingen, kräver det att hela vävnadsprovet (eller flera vävnadsprov) tas i beaktning av modellen. Detta illustreras i Fig. 2



Figur 1: Exempel på medicinska bilder som används i denna avhandling. **A** Koronala och sagittala vyer av en positronemissionstomografi (PET) av en hjärna, som visar ansamling av proteiner i hjärnan på grund av Alzheimers sjukdom. **B** Två exempel på MR-bilder från patienter med Alzheimers sjukdom. **C** En bild av en prostatabiopsi, med ett inzoomat område.

Alzheimers sjukdom är den vanligaste formen av demens och står för ca. 60–80% av alla fall. De vanligaste tecknen är en onormal ansamling av vissa proteiner i hjärnan och en efterföljande förlust av hjärnvävnad, vilket orsakar symptom som minnesförlust och andra kognitiva svårigheter. Därför baseras diagnosticeringen på flera undersökningar såsom kognitiva tester, mätningar av olika proteinhalter cerebrospinalvätskan eller blodet, och radiologi (se Fig. 1 (A-B)). Det finns starka bevis för att de sjukliga förändringarna i hjärnan uppstår årtionden innan några tydlgia symptom. Samtidigt kan uttrycket och progressionen av sjukdomen variera kraftigt från person till person, och vad som orsakar dessa skillnader är fortfarande okänt. De kognitiva symtomen är kopplade till förlust av hjärnvävnad i specifika regioner. Att upptäcka olika mönster av neurodegeneration och koppla dessa till andra biologiska förändringar kan därför vara en nyckel till att förstå sjukdomsförloppet bättre.

I denna avhandling undersöker vi AI baserade metoder för att analysera longitudinell medicinsk data för att möjliggöra tidigare diagnosticering av prostatacancer och Alzheimers sjukdom. Specifikt, utvecklade vi en metod för att bestämma prognoser för prostatacancer



Figur 2: Illustration av en AI-baserad algoritm för klassificering av prostata biopsier som hög- eller lågrisk.

med lovande resultat. Vi visade att med denna metod får vi bättre långsiktiga prognoser än med endast Gleason gradering. Denna algortimen kan bearbeta bilder i en övermänsklig hastighet. Detta medför två fördelar; den kan hitta högriskpatienter i ett tidigt skede och det kan förhinda onödiga behandlingar av lågriskpatienter. Dock återstår en del arbete i att verifiera resultaten på mer data. Vi utvecklade även en algoritm baserad på statistika modeller för att hitta mönster av longitudinella förändringar i hjärnvävnaden hos Alzheimers patienter. Vi upptäckte olika typer av sjukdomsförlopp. Detta tyder på att det är möjligt att identifiera patienter med högre risk att utveckla sjukdomen innan symptomen syns. De senaste åren har flera läkemedel lanserats för att bromsa utvecklingen av sjukdomen. Genom att identifiera vilka patienter som kan hjälpas av dessa, kan vi undvika onödig behandling och ta ett kliv mot att stoppa och bota Alzheimers sjukdom.

List of Publications

This thesis is based on the following publications, referred to by their Roman numerals. They are reproduced and included in this thesis with the permission of their respective publishers. The author's contributions to each paper is listed below.

I **Systematic Augmentation in HSV Space for Semantic Segmentation of Prostate Biopsies**
F. Winzell, I. Arvidsson, N.C. Overgaard, K. Åström, F.E. Marginean, A. Bjartell, A. Krzyzanowska, A. Simoulis, and A. Heyden
Scandinavian Conference on Image Analysis, Springer Nature Switzerland, Cham, 2023.
Author's contributions: This paper evolved from a previous idea of IA, NCO, AH, KÅ. I did the implementation and analysis and wrote the paper. I computed the annotations, which were verified by FEM. FEM, AB, AK, and AS were involved in the collection and curation of the original dataset.

II **Outcome prediction of prostate cancer patients on active surveillance using weakly supervised deep learning**
F. Winzell, I. Arvidsson, N.C. Overgaard, K. Åström, F.E. Marginean, A. Bjartell, A. Krzyzanowska, A. Simoulis, and A. Heyden
Medical Imaging 2025: Digital and Computational Pathology (SPIE), 2025.

Author's contributions: Myself, IA, NCO, KÅ and AH came up with the idea. I developed the method, implemented it, and wrote the paper. FEM, AB, AK, and AS were involved in the collection and curation of the datasets.

III **Longitudinal outcome prediction of prostate cancer patients on active surveillance using multiple instance learning**
F. Winzell, I. Arvidsson, N.C. Overgaard, K. Åström, F.E. Marginean, A. Bjartell, A. Krzyzanowska, A. Simoulis, and A. Heyden
Journal of Medical Imaging 12(6), 14 October 2025.

Author's contributions: I came up with the idea and the methods, with assistance from IA, NCO, KÅ and AH. I implemented them, analyzed the results and wrote the paper. FEM, AB, AK, and AS were involved in the collection and curation of the PRIAS dataset.

IV **Data-driven clustering of atrophy patterns in Alzheimer's disease**
F. Winzell, L. L. Raket, I. Arvidsson, N. C. Overgaard, K. Åström, A. Heyden, and N. Mattson-Carlgren
Under revision, 2026.

Author's contributions: The idea was developed by NMC. I developed the method with help from LLR and NMC. NCO, KÅ, AH, and IA all provided useful insights during the development of the method and writing of the paper. The current manuscript was mainly written by me.

V **Synthetic Alzheimer's disease dataset generation and evaluation with privacy protection**
F. Winzell, I. Arvidsson, N. C. Overgaard, K. Åström, A. Heyden, L. Karlsson, J. Vogel, O. Hansson and N. Mattson-Carlgren
Under revision, 2026.

Author's contributions: The project idea came from NMC, OH, JV, KÅ and IA. The formulation of this paper was developed by me, LK, NMC, JV, KÅ and IA. NCO and AH assisted with conceptual ideas during the writing process. I did all the method development, experiments and analysis of results. I wrote the paper, with help from LK, JV, and NMC.

VI **Dual energy CT and deep learning for an automated volumetric segmentation of the major intracranial tissues: Feasibility and initial findings**
V. Fransson, F. Winzell, B. Ramgren, S. Christensen, K. Ydström, I. Arvidsson, N. C. Overgaard, K. Åström, A. Heyden, and J. Wassélius
Medical Physics, 53(1), January 2026.

Author's contributions: The idea came from VF, SC, and JW. BR and KY assisted in the collection of the dataset. I developed the methods and computed the results, which VF and I analyzed. IA, NCO, KÅ and AH contributed with conceptual ideas. VF and I wrote the paper.

Acknowledgements

The writing of an academic thesis is a lonely and tedious process. However, the writing of this thesis would not have been possible without the help and support of my colleagues, friends, and family.

First of all, I would like to thank my supervisors: *Niels-Christian Overgaard*, *Kalle Åström*, *Ida Arvidsson*, and *Anders Heyden* for their continuous guidance and support. During my years as a Ph.D student, I have truly learned a lot from working together, which hopefully can be reflected in this thesis. Thank you to the rest of the *Computer Vision and Machine Learning group* (CVML) for the inspiration to pursue this degree. I also want to mention the collaborators I've had during these years. Specifically, *Niklas Mattson-Carlgren* and *Lars Lau Raket* at the Clinical Memory Research Unit. I have learned a lot and gained many new perspectives from working together. Thank you to the people at the *Computational Pathology Group* at Radboud UMC. Even though I only spent a few months in your lab, it was a big inspiration for me, and hopefully, our collaboration can continue in the future.

I also want to thank my fellow Ph.D students and colleagues at the Centre for Mathematical Sciences. Without the many coffee breaks, after-works, and Ph.D parties, working here would have been a lot less fun. A special thank you to *Jonathan Astermark*, my close friend and office-mate for most of my years here. Thank you to *Jennie Karlsson*, my academic twin sister, whom I've known ever since the start of our E-sek days almost 10 years ago. Additionally, I want to thank *Ivar Persson*, *Johanna Engman Granlund*, *Ludvig Dillén*, *Erik Tegler*, *Felix Augustsson*, *Magnus Fries*, *Oskar Åström*, *Anna Gummeson* and *Valentia Schüller*. Together we have been to conferences, winter and summer schools, symposiums of drinking games, operettas, and even weddings. All of you have truly become close friends of mine, and the memorable and fun times we've had together are by far the greatest takeaway from my time here.

I also want to mention my old classmates from BME16, with whom I've shared the first half of my years at LTH. You are the greatest reason why I have remained here for all this time. Particularly, thank you to *Anton*, *Andreas*, *Rasmus*, *Hjalmar*, *Martin*, *Max* and *Fanny*. Thank you for distracting me during tough times, and thank you for your encouragement when I have complained about my work not going the way I hoped. Thank you for being among the closest friends I will ever have.

A big thank you to my old high school friends; *Carl*, *Gabriel*, and *William*. Thank you for sticking by me for all these years since surviving Hvitfeldtska, and thank you for all the laughs and good times we have shared. *Vita vinum est*.

And to my family, *Mamma*, *Pappa*, and *Emil*. Thank you for all the love, support, and encouragement, and for molding me into the person I am today. Thank you *Mamma* for

inspiring me to pursue a Ph.D and for your very helpful advice and uplifting words during the darkest times. Thank you *Pappa* for your constant support and for teaching me how to work hard and to believe in myself. And thank you to *Emil*, my actual twin brother and other-half. Thank you for always being there for me during good and bad times. While this Ph.D has only been a short period of our lives, without you, none of this would have been possible. Shout-out to my family!

And finally, thank you to whoever is reading this for taking an interest in my research. And thank you for not only reading the acknowledgments, but also the rest of this work of art.

Lund, 2026-04-29

Funding

The writing of this thesis has been partly supported by grants from the Wallenberg AI, Autonomous Systems and Software Program (WASP) and the SciLifeLab and Wallenberg National Program for Data-Driven Life Science (DDLS) joint call for research projects (WASP/DDLS22-066). Additionally, from the Early Diagnostics and Prognostics of Alzheimer's disease (EDAP) project (EDAP2023-153244)

Contents

Chapter 1

Introduction

The healthcare systems of today are evolving fast. In particular, the rise of methods based on artificial intelligence (AI) has already transformed the diagnostic procedures in several domains and now enables an increasingly personalized medicine. These systems have superhuman capabilities in processing vast amounts of data in short times and can improve the diagnostic accuracy beyond traditional methods [41]. However, modern healthcare is confronted with significant challenges of rising costs, aging populations, and increasing workload for the hospital staff. There is also a global inequality in healthcare, where lower-income countries face disproportionately severe challenges [41]. The combination of AI with technical advancements for collecting, organizing, and analyzing large medical datasets in recent years holds the potential of mitigating these issues for tomorrow's healthcare [112].

The previous view on health as a state has been replaced by a newer definition, where health is now considered a dynamic process of several correlated states [15]. These states together form a longitudinal trajectory of a person's health, which can reflect a stable condition, a decline in health as a disease progresses, and an increase following recovery. Understanding the transitions between these individual states is crucial for an effective treatment of a specific disease. Therefore, the collection and analysis of longitudinal data is an area of great interest for various medical research fields. Longitudinal medical data consists of repeated observations of a patient's health status, followed over a longer period of time. Examples of such data include repeated analyses of blood samples, medical imaging examinations, such as computed tomography (CT) or magnetic resonance imaging (MRI), and analysis of tissue samples.

While these modalities can be used to describe general health status, the vast amounts of information they generate is challenging. Discerning the important parts of this information when characterizing the progression of a particular disease is a highly complex task, which

usually requires years of medical experience and expertise. With the computational power of today and algorithmic advancements, AI-based methods have capabilities of doing this at comparable levels as trained physicians [41]. Furthermore, these methods also have the ability to discover hidden patterns in the data, enabling predictions of disease progression and identification of unknown risk factors. However, there are several concerns to consider when implementing such methods in the healthcare systems. The large volume of medical data needed to train and test these methods is neither standardized nor openly accessible. Hence, the collection and sharing of high quality data is one of the key challenges for future medical AI development [113]. The monitoring of the performance of these systems in an evolving healthcare system also poses concerns. As newer imaging technologies, software updates, or examination protocols are implemented, we cannot ascertain the same level of safety of an AI-system trained on older data in the previous setting.

Prostate cancer and Alzheimer's disease are examples of two common diseases where the incidence is expected to increase over the following decades, with growing healthcare costs and risks of insufficient therapeutics [55, 103]. The success of treatment is highly dependent on early diagnosis. Therefore, enhancing the efficiency and accuracy of the diagnostic procedures is key to mitigate these issues. Analysis of longitudinal health measurements, such as brain MRI images or microscopy images of prostate tissue samples, with AI-based approaches holds the potential of improving this. Investigating and developing these methods is the overall aim of this thesis. This aim can be divided into the following specific research questions:

- How can the general lack of longitudinal medical data of high quality be addressed?
- Is it possible to predict the necessity of treatment of a prostate cancer patient, given prostate biopsies?
- To what extent can image data from prostate biopsies be used for longitudinal outcome predictions of prostate cancer patients?
- Can biologically different subtypes of Alzheimer's disease be identified from brain MRI images?
- Could these subtypes be used for predicting the disease progression of Alzheimer's disease?

This thesis consists of a collection of publications, aimed at addressing the aforementioned questions in the following manner: Paper I, Paper V and Paper VI provide possible solutions to limitations of data, Paper II and Paper III investigates longitudinal outcome predictions of prostate cancer patients, and Paper IV explores subtypes of Alzheimer's disease based on longitudinal brain MRI imaging (see Fig. 1.1). The relevant clinical background and

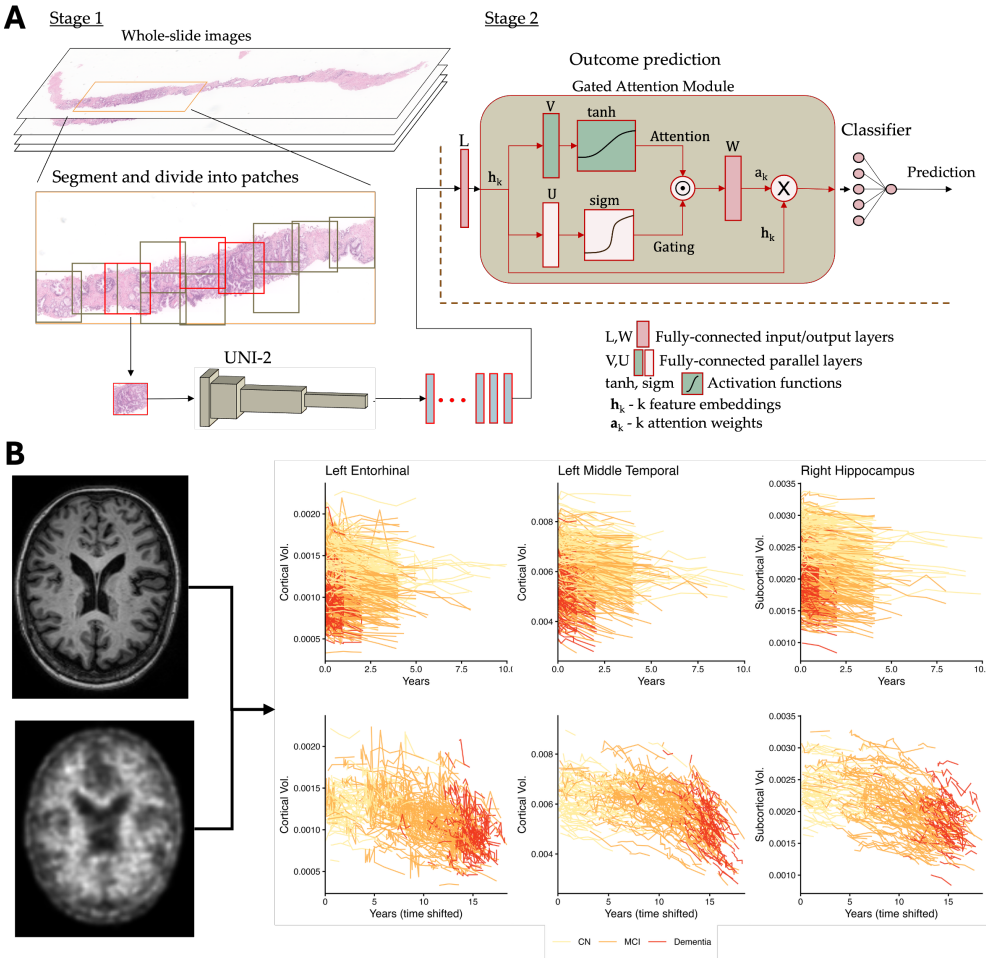


Figure 1.1: **A** Illustration of the framework used in Paper III [139] for predicting longitudinal outcomes of prostate cancer patients. In the first stage, the images are segmented automatically and then divided into patches. The UNI-2 foundation model is used to transform the patches into latent representations. Subsequently, an *attention-based multiple instance learning* module is applied to compute the predictions. **B** Example of the disease progression modeling of Alzheimer's disease from Paper IV. MRI and PET images are used in a statistical framework to find subtypes of disease trajectories.

specific problems are introduced in Chapter 2. The main datasets used in the publications are described in Chapter 3. The theory behind the methodology is presented in Chapters 4-5, divided into machine learning techniques and statistical methods, respectively. The results are summarized and explained in Chapter 6, and finally, a conclusion and future outlook are provided in Chapter 7.

Chapter 2

Clinical background and problem definition

In this chapter, the clinical background and current diagnostic procedures of prostate cancer (Section 2.1) and Alzheimer's disease (Section 2.2) will be discussed. Furthermore, the aim and research questions of this thesis will be motivated.

2.1 Prostate cancer

Prostate cancer is the world's most common male cancer diagnosis, with 1.46 million reported new cases in 2022 [10], and it is expected to rise towards 3 million by 2040 [55, 67]. In Sweden, prostate cancer accounted for 28% of all cancer cases among men in 2024. Earlier detection and improved treatment methods have caused the mortality to drop, and in 2024, 95% of prostate cancer patients in Sweden were expected to survive more than 5 years [18]. However, the anticipated rise in incidence will be challenging for healthcare systems which globally already are under significant strain. Early diagnosis improves outcomes substantially, yet, in many low-to middle-income countries, late or no diagnosis is still the norm [67]. Furthermore, despite the incidence in men of African heritage being significantly higher than for men with European heritage, most studies of the disease are focused on the latter [67, 119]. Screening for prostate cancer has the potential of reducing mortality, however, with the wide range of cancer aggressiveness, screening raises issues with overdiagnosis [58, 9, 13]. The biological processes that affect the progression from indolent cancer to lethal is currently not fully uncovered and treatment of unharmful cancer forms causes unnecessary harm to patients and costs for the healthcare systems [67, 13]. The use of AI systems has the potential of mitigating some of these issues, both in terms

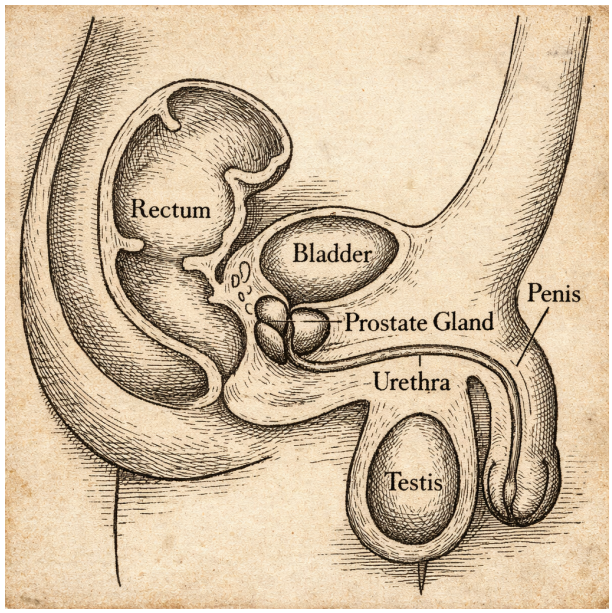


Figure 2.1: Illustration of the anatomy of the prostate. Image adapted from Wikimedia Commons, originally from the National Cancer Institute (NCI). Style edited with *OpenAI* (2026), *DALLE-3* `https://openai.com/dall-e-3/`.

of improving the diagnostic accuracy and speeding up the processing of patient data. This could enable a wider screening program without an unfeasible workload for the healthcare professionals.

2.1.1 Physiology and pathology

The prostate is a small gland, about the size of a walnut, positioned below the bladder, surrounding the urethra (see Fig. 2.1). The prostate is part of the male reproductive system, and its main function is to secrete the fluid of ejaculated semen. This secretion contains numerous nutrients and chemical substances, such as buffers protecting the sperm against the acidic urine, and chemicals increasing the sperm motility [138]. Hence, the prostatic tissue mainly consist of glandular structures, which are tubules surrounded by stroma (see Fig. 2.2) [76]. A gland consists of a layer of epithelial cells that secrete proteins and substances into the lumen of the inside of the tubule, where they are transported and released into the urethra. The epithelial cells and the stroma are separated by a basement membrane with basal cells. The stroma is the connective tissue between the glands, and consists mainly of extracellular matrix, smooth muscle fibers, and fibroblasts [76].

The most common form of prostate cancer is acinar adenocarcinoma, which originates from the epithelial cells [59]. The other non-acinar carcinomas make up around 5-10% of

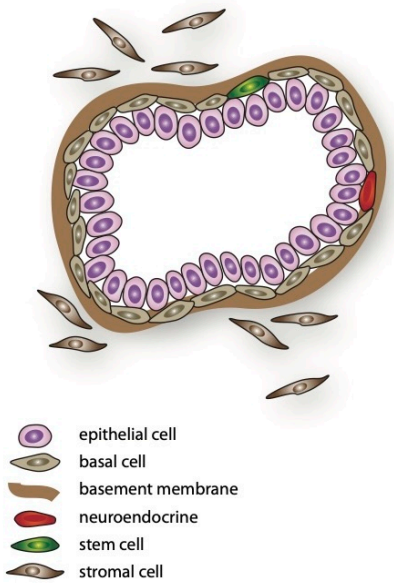


Figure 2.2: Illustration of a normal prostatic glandular structure. One layer of epithelial cells, basal cells and a basement layer surrounds the lumen inside the gland. Outside the basement layer, stromal cells forms the connective tissue. Neuroendocrine and stem cells can occur in the basal cell compartment. Image reproduced with permission from Lippolis [76].

all cases. However, within these groups there are several subtypes, which makes the disease progression highly heterogeneous. The growth of the prostate is partly controlled by the supply of androgens (e.g. testosterone). Prostate-specific antigen (PSA) is a target protein of the androgen receptor (AR) pathway [76]. An increase in activation of the AR and subsequent release of PSA can be associated with abnormal prostate growth, and therefore PSA is used as a blood-based biomarker for prostate cancer.

2.1.2 Diagnosis

It is possible to screen for prostate cancer with blood PSA tests, and this is usually the first test performed when suspecting prostate cancer [2]. The risk of prostate cancer increases with the PSA levels, but there are no general thresholds where prostate cancer either can be ruled out nor confirmed. Usually, a PSA test with elevated levels is followed up with a digital rectal exam and possibly an MRI [2, 28]. However, for a definitive diagnosis, a needle-core biopsy is commonly performed. Today, it is standard practice to do a trans-rectal ultrasonography (TRUS) guided biopsy. The use of multi-parametric MRI (mpMRI) has been suggested by several studies, both to guide the biopsy and as a non-invasive alternative to the biopsy for ruling out prostate cancer [118]. In recent years, traditional methods in pathology have been increasingly digitized in hospitals around the world. This means that

nowadays, instead of being studied through a microscope, tissue samples from the biopsies are scanned into digital images. When a prostate biopsy is scanned, a so-called whole slide image (WSI) is generated, which is usually of gigapixel size. The pathologists then analyze the grade of prostate cancer and determine a diagnosis by studying these large images.

Histological staining

To examine the tissues of a biopsy, it is fixed in paraffin and sliced into thin sections [4]. In order to visualize the microstructures within these sections, they are stained with chemicals to highlight certain cellular components. The most common stain in histology is hematoxylin-eosin (H&E), which has been used for more than a century [20]. The H&E stain divides tissues into basophillic, eosinophillic and non-staining. The cell nuclei are examples of basophillic cellular components which are colored purple/blue, and the extracellular matrix is eosinophillic and is colored pink (See Fig. 2.3). For difficult cases, immunohistochemical stains can be used to confirm the diagnosis or to definitely rule out prostate cancer [76]. Examples are p63, which is a basal cell marker i.e. benign tissue marker, and AMACR, which is generally expressed in epithelial cancer cells. Although these stains provide more information than H&E, they are costly and do produce false positives and negatives. Hence, H&E remains the gold standard in prostate cancer diagnosis, and in this thesis, only images with this stain will be used.

Gleason grading

The prostate biopsies are graded according to the Gleason grading system. Developed in the 1960s by Dr. Donald Gleason, and revised in 2005 and 2014 by the International Society of Urological Pathology (ISUP), the Gleason grading system now consists of four grades of histological patterns; Benign and Grade 3-5 [116] (see Fig. 2.4). The original grades 1-2 were removed from biopsy reporting in the 2005 revision, but the numbering was kept for grade 3-5 due to historical reasons. For each biopsy core, the two most predominant grades are summed into a Gleason score. In the latest ISUP update, the scores are replaced by Gleason grade groups, ranging from 1-5, where 1 corresponds with the lowest grade of prostate cancer [88]. The ISUP grades and their definitions are briefly summarized in Table 2.1.

Although the introduction of the ISUP grades addressed several issues with the previous Gleason grades, there are concerns regarding intra- and inter-observer variability [35]. Furthermore, studies have shown that ISUP grade 4 tumors with Gleason scores 5+3 can have a worse prognosis than the other grade 4 scores, and therefore should be included in the ISUP grade 5 group [80]. There are inconsistencies in the reporting, where some studies have found that using the most predominant and the highest grade correlates better with

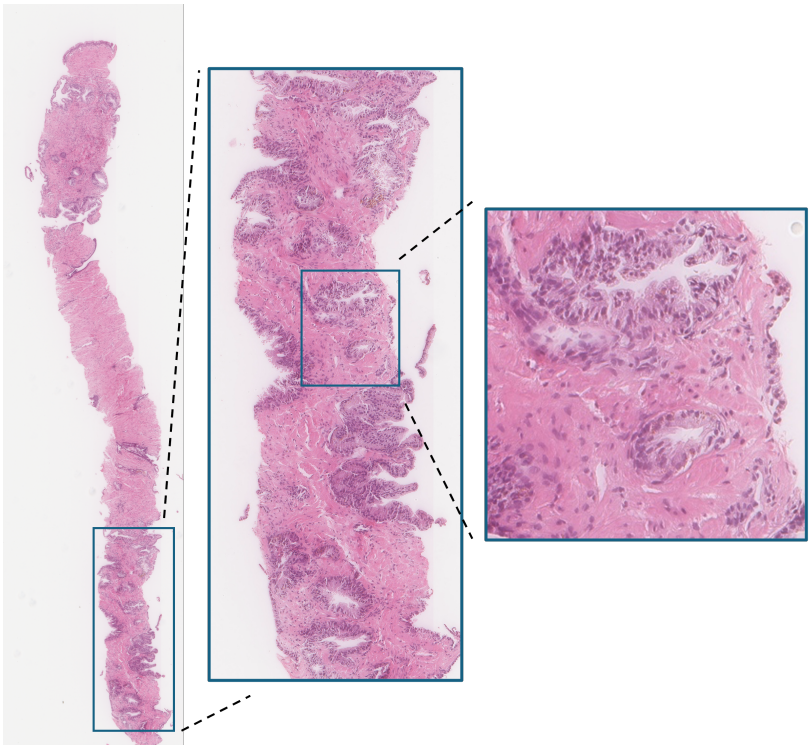


Figure 2.3: An example of a H&E stained WSI with a zoomed-in gland.

the pathological state [76]. There is little consensus on which of these ways of reporting should be used. Additionally, studies have shown that higher tertiary Gleason patterns also affects the prognosis negatively [40, 79]. Yet, it is not included in the ISUP grading system.

2.1.3 Treatment

After a prostate cancer diagnosis, there are several treatment options. To make a decision, in addition to the ISUP grades, the stage of the disease, the general health of the patient, and their life expectancy are considered [76, 3]. If the prostate cancer is localized to the prostate and the patient is deemed fit for curative therapy, the usual treatment options are surgery or radiotherapy. The main type of surgery is radical prostatectomy, where the entire prostate gland and possibly surrounding tissue is removed. This procedure is not trivial, and there are risks of incontinence and erectile dysfunction associated [28]. If the cancer has started to spread outside the prostate, or if the surgery has not completely removed it, radiotherapy can be used. Measuring the blood PSA values is a common way of monitoring patients after prostatectomy. Immediately after surgery, PSA is typically undetectable in the blood, but in approximately one third of patients, the PSA will rise again [26, 46]. This is called

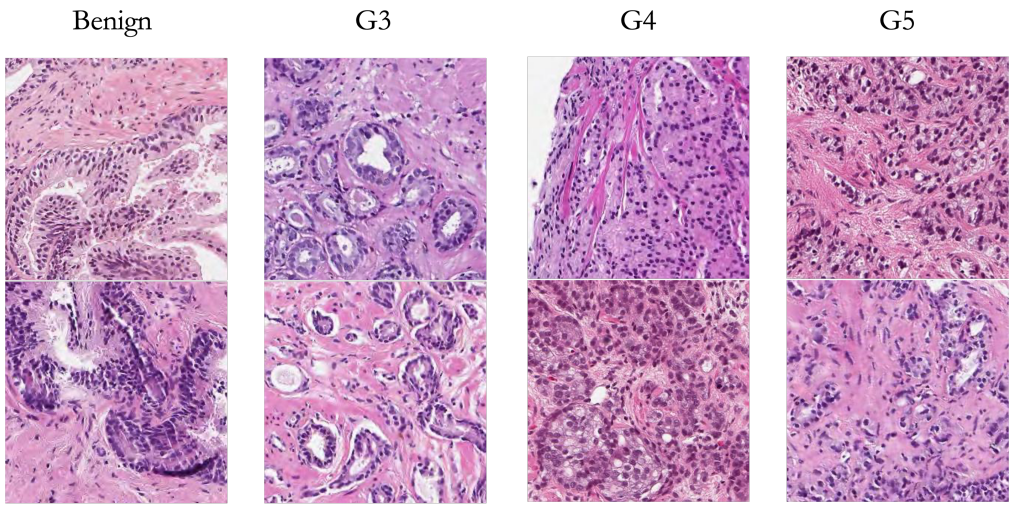


Figure 2.4: Examples of H&E stained prostatic benign tissue and Gleason grades 3-5.

biochemical recurrence, and it is a strong risk factor for prostate cancer death.

For more advanced prostate cancer, hormone therapy or androgen deprivation therapy (ADT) is the most common choice of treatment [76]. The aim of ADT is to slow down the growth of tumors by targeting the AR-androgen pathway, however, this is generally not a curative treatment. There are other palliative treatment options, such as immunotherapy and chemotherapy

2.1.4 Screening and active surveillance

Screening programs for prostate cancer using PSA blood tests have been a heavily debated subject among medical experts. The European Randomized Screening for Prostate Cancer (ERSPC) trial was initialized in 1992 to investigate whether prostate cancer mortality could be reduced by PSA screening [111]. Schröder et al. found that PSA-based screening reduced the rate of prostate cancer death by 20%, although estimated rates of over-treatment were as high as 50% (defined as the amount of men with elevated PSA values who would not show any clinical symptoms in their lifetime) [111]. In a Cochrane review of five randomized studies (conducted between 2004-2018), the main conclusions were that although screening is associated with an increase in prostate cancer detection, no benefit on survival could be observed [60]. On the contrary, the 2019 update on the ERSPC showed a significant reduction in prostate cancer mortality in the screening group [58]. Furthermore, in a more recent 22-year follow-up study, Frånlund et al. found that the cancer mortality could be lowered by 30% with more frequent PSA testing, and that this could likely be increased

Table 2.1: Summary of the ISUP grading system [116]. Note the distinction that the ISUP grade is assigned on biopsy level and Gleason grade on each histopatholgical pattern within each biopsy.

ISUP grade	Gleason scores	Definition
Grade 1	2-6	Well-formed discrete glands of varying sizes
Grade 2	3+4=7	Predominately well-formed glands, with a small proportion of poorly formed/cribriform glands
Grade 3	4+3=7	Predominately poorly formed/cribriform glands
Grade 4	4+4=8	Only poorly formed/cribriform glands
	3+5=8	Predominately well-formed glands with small proportion lacking glands (or with necrosis)
	5+3=8	Predominately lacking glands (or with necrosis) with small proportion of well-formed glands
Grade 5	9-10	Lacking gland formation (or with necrosis) possibly with poorly formed/cribriform glands

even further by including older patients [44]. While more follow-up visits reduced the amount of over-diagnoses, the number of additional patients needed to be diagnosed to prevent one prostate cancer death was found to be 9 cases.

However, active surveillance has been proposed as a viable alternative to radical treatment, following PSA screening, that can mitigate the risk of over-treatment[28, 13, 5]. This means that patients with low-risk prostate cancer are monitored with recurring PSA tests and biopsies to avoid treatment of unharmful tumors. Patients who show indications of disease progression, given pre-defined thresholds, are offered curative treatment. The Prostate Cancer Research International Active Surveillance (PRIAS) study was initiated in 2006 with the aim of worldwide data collection of patients on active surveillance. The study has enabled further investigation of optimal follow-up protocols and selection of high-risk patients [13]. In this thesis, we have utilized a dataset collected within the PRIAS study (see Section 3.1 for more details).

Studies have shown clear benefits of active surveillance in terms of reduction of over-treatment, but the long-term effects of active surveillance are still understudied [122, 125]. The use of enhanced blood test [43, 129] and the use of MRI [37] have been suggested as possible improvements to the diagnostic procedures in the screening programs. In December 2022, the European Union issued a recommendation to researchers to evaluate the effectiveness of organized prostate cancer screening. In a Swedish study, Bratt et al. concluded that these programs are feasible and that with the use of MRI, a biopsy could be avoided for over half of the patients with elevated PSA values [9].

2.1.5 Computational pathology

The introduction of digital pathology has paved the way for the use of computational methods in the diagnostic pipeline [132, 115]. These advances, in combination with the tremendous development of data storage capabilities and AI, have led to an explosive growth of the field of computational pathology [117, 132]. The field has, over the past decade, followed the broader trends of computer vision, i.e. computational analysis of natural images, while applying them to WSIs from digital pathology. In the early 2010s, the development of convolutional neural networks (CNNs) had a particularly big impact, and researchers in computational pathology used these methods for successful applications in various fields [132] (see Section 4.2.1 for an introduction of CNNs). However, digital pathology workflows were not yet widely implemented at this time, and the lack of data was a bottleneck. In 2016, the CAMELYON16 challenge introduced, at the time, one of the largest open-access cohorts in computational pathology, aimed at detection of breast cancer metastases in lymph nodes [36]. Since then, similar challenges have been proposed for numerous applications, such as prostate cancer Gleason grading, kidney transplant assessment, pancreatic tumor segmentation, and more. This led to more efforts at collecting larger, multi-center cohorts in digital pathology.

Following this, several studies presented CNN-based automatic Gleason grading algorithms on-par with experienced pathologists [82, 17, 14, 124]. In 2021, Paige Prostate became the first clinical-grade AI-based solution to receive marketing approval from the Food and Drug Administration [85], and since then, more have followed.

However, a drawback of these methods is that they often require pixel-level annotations of the datasets. Gathering these is a tedious process that is not part of the clinical practice. This led to a shift towards other learning methods that do not require these annotations. Specifically, a deep-learning paradigm called multiple instance learning was employed in several studies in computational pathology with promising results [17, 78] (see Section 4.5 for more details on this method). With these methods, slide-level labels can be used when training the models, e.g. ”does this slide contain cancer or not?”. These labels are always assigned in the standard clinical procedure, and therefore, this data can be leveraged without any additional work for the pathologists.

In recent years, these methods have been further improved by the development of large, general-purpose models, called foundation models. For training these models, various gigantic cohorts of WSIs have been collected. As a comparison, the CAMELYON16 dataset consisted of 299 WSIs [36], and less than 8 years later, for the training of the foundation model Virchow-2, a dataset of 3.1 million WSIs was used [148]. However, these large cohorts are not made public, and still, it is a challenge in computational pathology to obtain data that truly represents clinical practice [132]. More details on the training procedures of foundation models can be found in Section 4.6, and on foundation models in compu-

tational pathology in Section 4.6.2.

2.1.6 Longitudinal outcome prediction

While several studies have shown the strong capabilities of AI models in predicting Gleason grades, they are intrinsically limited by the aforementioned subjectivity and variability of the Gleason grading system. Therefore, adapting these models to predict longitudinal patient outcomes instead would be of great interest. Previous studies have demonstrated strong prognostic capabilities of deep learning-based Gleason grading algorithms, by fitting Cox regression models on the predicted grades [5, 142]. More recent studies have extended these models to predict biochemical recurrence of prostate cancer patients using end-to-end frameworks, with promising results [99, 52]. Pinckaers et al. used an ImageNet pre-trained ResNet-50 for predicting years to recurrence based on prostatectomy tissue microarrays [99]. Grisi et al. extended this by predicting a continuous time-to-recurrence risk using state-of-the-art foundation models, such as UNI and Virchow-2, as feature extractors from WSIs [52]. Furthermore, more sophisticated models have been developed for predicting the occurrence of future prostate cancer [77, 22]. Chelebian et al. developed a method based on multiple instance learning to predict future detection of prostate cancer [22]. They used a cohort of men with elevated PSA values and only slides originally classified as benign, where the positive patients received a prostate cancer diagnosis (any ISUP grade) within 30 months, and the negative remained cancer-free for 8 years. Notably, in 2025, ArteraAI was the first company to receive FDA approval for their AI-based risk stratification tool for patients with non-metastatic prostate cancer [131]. The algorithm combines clinical data with WSIs of prostate biopsies in a multi-modal framework, and its strong capabilities have been demonstrated in an external validation study on phase III clinical trial data [96].

2.2 Alzheimer's disease

Alzheimer's disease (AD) is the most common form of dementia, accounting for 60-80% of all cases [6]. The estimated number of patients globally exceeds 50 million and is expected to rise to 75 million in 2030 [103]. The cost of dementia in the US was estimated at $360 billion with an additional $347 billion for unpaid caregiving in 2023 [6]. AD is a complex neurodegenerative disease, which is characterized by cerebral atrophy (i.e loss of brain neurons, see Fig. 2.5) as well as the accumulation of amyloid-β (Aβ) protein plaques and tau protein tangles in the brain. Common symptoms include memory loss, cognitive decline, personality changes, and language problems. However, the biological changes in the brain are believed to occur as early as 20 years before any symptoms. In recent years, several treatments have been developed to slow the progression of the disease by targeting

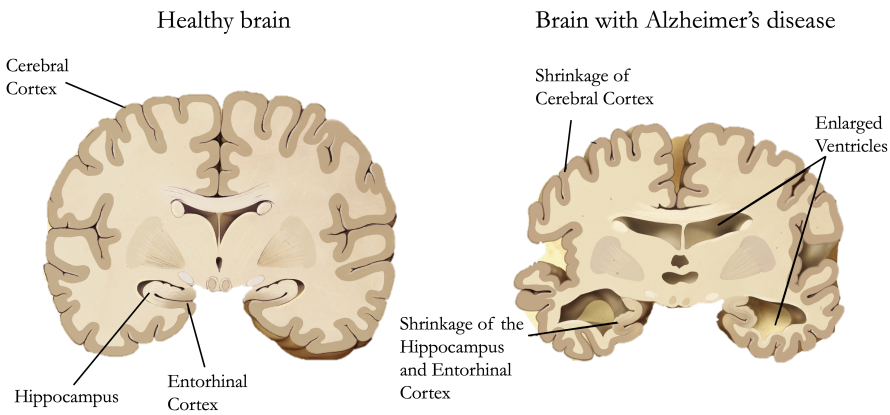


Figure 2.5: A schematic of how different brain regions are affected by Alzheimer's disease. Image adapted from Wikimedia Commons, originally from the Alzheimer's Disease Education and Referral Center.

the brain Aβ plaques [114, 34, 81]. The efficacy of these treatments is highly dependent on early detection of the disease. To be able to do that, a deeper understanding of the early biological changes and the heterogeneous progression of the disease is needed.

2.2.1 Brain anatomy

The human brain consists of three main parts: the cerebrum, the brain stem, and the cerebellum [138]. The cerebrum is divided into left and right hemispheres, with an inner core of white matter and an outer layer of gray matter called the cerebral cortex. Compared with white matter, gray matter contains large amounts of neuronal cell bodies and fewer myelinated axons (the white color of myelin causes the color difference). The cerebral cortex is one of the most complex parts of the central nervous system, which plays a key role in perception, memory, movement etc. Both hemispheres are divided into four lobes: the frontal, parietal, temporal, and occipital lobes. Within each lobe, several regions and sub-regions are defined based on different functions or structures. Deep inside the cerebral hemispheres, there are groups of gray matter, such as the basal nuclei and parts of the limbic system. Additionally, there are four interconnected cavities called ventricles, which produce and circulate the cerebrospinal fluid (CSF). The limbic system consists of both white and gray matter areas associated with behavior, memory, learning, emotion, and a variety of endocrine functions. The hippocampus is an integral part of this system, located in the medial temporal lobe, beneath the lateral ventricle (see Fig. 2.6). In the typical AD pathology, the limbic system and the hippocampus in particular are the primary regions to be affected in the early stages of the disease [33].

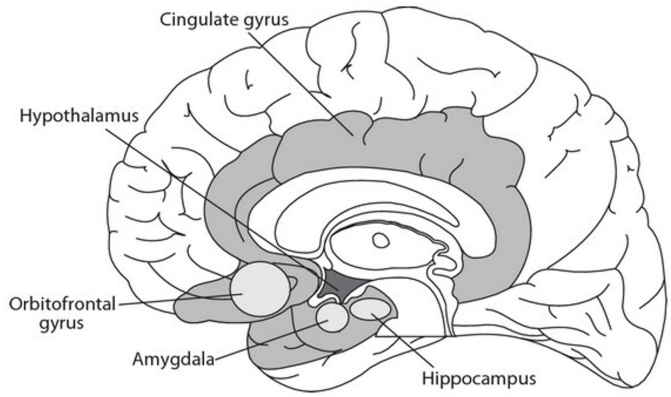


Figure 2.6: Illustration of the limbic system. Image taken from Roxo et al. [108] (CC BY 3.0)

2.2.2 Pathology

AD is usually initiated by Aβ deposition, followed by tau pathology in the preclinical stage of the disease [33, 64, 39]. This stage can last for over 20 years before neurodegeneration occurs to the degree that symptoms start to show. The neurodegeneration and deposition of Aβ can be visualized with MRI and positron emission tomography (PET) (see Fig. 2.7). The Aβ peptide is derived from the larger glycoprotein Amyloid precursor protein (APP) [23]. The role of APP includes neuronal development, signaling, and homeostasis. While Aβ occurs naturally in our bodies, the normal physiological role and the mechanisms that cause Aβ plaques to form in the brain are still unknown. It has been observed that soluble Aβ accumulates gradually, forming oligomers, subsequently larger fibrils, and finally plaques. These earlier forms of Aβ deposits may play pathogenic roles at different stages of the disease [75]. Tau is a microtubule-associated protein that plays an important role in microtubule assembly and stabilization [137] (microtubule forms the cytoskeleton of all cells [138]). Tau can aggregate into paired helical filaments and neurofibrillary tangles, which are associated with a number of neurodegenerative diseases, including AD. Hyper-phosphorylation of tau is believed to enhance the aggregation, as hyper-phosphorylated tau commonly occurs jointly with the aggregated tangles. Still, to date, the mechanism behind the pathogenesis of tau is not fully uncovered [137]. For instance, several studies have shown that the neurofibrillary tangles themselves do not cause brain dysfunction, but rather might be protective responses against soluble toxic tau proteins or reactive oxygen species [137, 110]. The interaction between Aβ and tau is also not fully known, but Aβ aggregation has been shown to induce the tau hyper-phosphorylation [23]. Likely, there are more mechanisms other than Aβ and tau pathology involved and several different pathways in which AD can develop, which explains the heterogeneity of the disease progression.

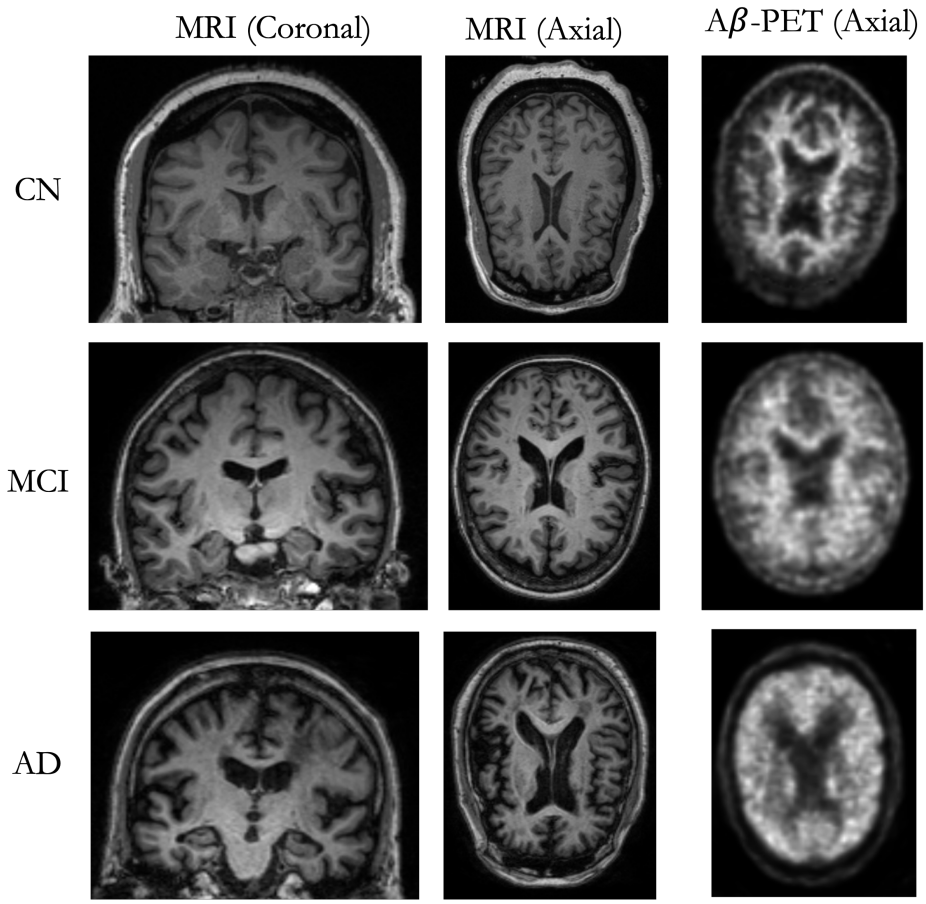


Figure 2.7: Visualization of the progression of AD on MRI and Aβ-PET scans. For the MRI views, a gradual enlargement of the ventricles can be observed, as well as neurodegeneration and loss of gray matter. On the Aβ-PET scans, the characteristic spread of Aβ plaques to the gray matter can be observed. For the AD example, it is difficult to distinguish the white and gray matter based on the PET scan.

Co-pathologies

Most patients with AD also have co-pathologies, which influence the progression of the disease [127]. Cerebrovascular disease is an umbrella term for any vascular disease in the brain, which are common co-pathologies of AD in older patients [6]. Anatomical MRI can be used to find indications of cerebrovascular diseases. Specifically, white matter hyperintensities (WMHs), which appear as brighter regions on T2-weighted MRI scans [29], are associated with cognitive decline and faster progression of AD [72]. Lewy body disease (LBD) is another common comorbidity accompanying AD pathologic changes in 25-50% of cases [104, 127]. Lewy bodies are formations of α-synuclein protein aggregates, which are common in several neurodegenerative diseases, including Parkinson's disease [49]. Re-

cent studies have shown that LBD can be identified in-vivo using a α-synuclein Seed Amplification Assay (SAA) on CSF samples [126]. Another frequent co-pathology of AD is the presence of TAR DNA-binding protein 43 (TDP-43) in the medial temporal lobe and the limbic region [121]. The presence of TDP-43 is associated with atrophy of the hippocampus and a faster progression, compared with typical AD pathology.

2.2.3 Diagnosis and biomarkers

Typically, the first symptoms of AD are subjective cognitive decline (sometimes also referred to as subjective memory concerns), which are self-reported cognitive and memory difficulties. This is followed by mild cognitive impairment (MCI), where the cognitive capabilities are reduced, which can be measured by a reduced performance on cognitive tests [6]. Diagnosing early-stage AD is not easy, with 20-30% being misdiagnosed in specialized care and as high as 40% in primary care [53, 123]. Furthermore, only 8% are correctly diagnosed during the MCI stage [83]. Various biomarkers, such as CSF measurements or PET-scans with Aβ or tau binding tracers, are increasingly being used to improve the diagnostic accuracy. Furthermore, the use of blood-based biomarkers has emerged as a less invasive and less expensive alternative [53, 95].

In 2024, the Alzheimer's Association Workgroup released an update [66] of the previous clinical framework of AD from 2018 [65]. The 2018 framework built upon the ATN-biomarker classification scheme, where biomarkers are grouped according to their correspondence with A - Aβ pathology, T - tau pathology, or N - neurodegeneration. However, it did not distinguish between CSF and imaging biomarkers within each category, even though evidence now suggests that these biomarkers differ in timing across the AD spectrum. [66]. For example, the T-biomarker phosphorylated tau can become abnormal before a positive tau-PET scan [84]. In the revised framework, the biomarkers are divided into three new groups: core biomarkers of AD neuropathologic change (ADNPC), non-specific biomarkers involved in AD, and biomarkers of non-AD co-pathologies (see Table 2.2) [66]. These biomarkers typically become abnormal at different stages of the disease (see Fig. 2.8). The core biomarkers for AD are those in the previous A and T categories. The A category includes both fluid assays of the Aβ_{42} peptide, and amyloid-PET scans, which usually become abnormal simultaneously. Given evidence that phosphorylated tau (p-tau) fragments may be secreted as a response to the accumulation of Aβ plaques, and that this happens before abnormal tau-PET can be observed, the T category was split into T_1 and T_2. The T_1 group includes these p-tau forms, and together with the A category forms the Core 1 biomarkers. These biomarkers define the initial stage of AD that can be detected in-vivo before any symptoms. The rest of the tau proteinopathy biomarkers constitute the T_2 category and the Core 2 biomarkers. These become abnormal in a later stage, closer to the onset of symptoms. AD can be diagnosed based on abnormal specific biomarkers

in the Core 1 category (for the CSF measurements, usually the ratio $A\beta_{42}/A\beta_{40}$ is used). Core 2 biomarkers should not be used to diagnose AD alone, but they remain very useful in defining the stage of the disease.

The non-specific biomarkers include the N (neurodegeneration) and I (inflammation) categories, which are undoubtedly important steps in the progression of AD. However, they are not AD specific, i.e. they occur in other non-AD neurodegenerative diseases [65, 66]. The influence of co-pathologies on the stages of AD is visualized in Fig. 2.8 (B). Given this, biomarkers of them have gained more attention in recent years. Under this category, WMHs and α-synuclein SAA are listed as biomarkers for vascular brain injury and LBD, respectively. Biomarkers for TDP-43 co-pathology are not mentioned in this category, yet, studies have shown that among patients with biomarker evidence of intermediate to high ADNPC, the rate of TDP-43 co-morbidity is between 35% and 58% [127]. Previously, the presence of TDP-43 could only be detected postmortem, however, MRI-based atrophy has been suggested as a possible biomarker. Specifically, the ratio between the inferior and middle temporal lobe regions divided by the hippocampal volume has been proven as a strong indicator [120].

Table 2.2: The Alzheimer's Association Workgroup's revised categorization of AD-related biomarkers.

Category	CSF or plasma analytes	Imaging
Core biomarkers		
Core 1		
A (Aβ pathology)	$A\beta_{42}$	Amyloid-PET
T_1	p-tau217, p-tau181, p-tau231	
Core 2		
T_2	Other phosphorylated tau forms, Non-phosphorylated tau fragments	Tau-PET
Non-specific AD biomarkers		
N (neurodegeneration)	Neurofilament light chain (NfL)	Anatomic MRI, FDG PET
I (inflammation)	Glial fibrillary acidic protein	
Biomarkers of non-AD co-pathology		
V (vascular brain injury)		MRI/CT white mater hyper-intensities (WMH)
S (α-synuclein)	α-synuclein seed amplification assay (SAA)	

2.2.4 Staging and sub-typing

The biomarkers have been used to define seven clinical stages of AD as the following; (0) Genetic evidence of AD with normal biomarkers and function, (1) Asymptomatic with biomarker evidence, (2) Subjective mild cognitive decline with minimal impact on daily func-

tion, (3) Reduced/abnormal performance on cognitive tests, (4-6) Dementia with mild, moderate and severe functional impairment [66]. However, the progression between these stages is highly heterogeneous, and the mechanisms behind this variance are not understood. To enable treatment of AD at an early disease stage, uncovering the longitudinal processes involved is imperative. Thus, many studies have attempted to investigate the temporal relationships between amyloid deposition, tau accumulation, and regional atrophy patterns, and their contribution to the disease trajectories [66, 39, 27]. This has led to observations of clinical subtypes of AD, such as the typical amnestic phenotype and several non-amnestic phenotypes [51]. The typical phenotype can be characterized by a slower progression, with atrophy starting in the limbic system with neurodegeneration of the hippocampus and entorhinal cortex. The atypical phenotypes generally affect younger patients, with other patterns of brain atrophy [51] as well as divergent regional distributions of tau tangles in the brain [135, 134]. Based on postmortem quantification of tau protein tangles, the AD subtypes typical AD, limbic predominant, and hippocampal-sparing have been defined [90]. The hippocampal-sparing group is characterized by a higher density of tau tangles in cortical areas and a lower density in the hippocampus, compared with typical AD. The limbic-predominant displays the opposite relationship. Naturally, they are connected to different atrophy patterns corresponding with the tau deposition. Several studies have identified them based on measurements of MRI volumes or visual assessments [16, 42]. The typical and limbic-predominant subtypes have been observed to be associated with a worse clinical progression than the hippocampal-sparing and minimal atrophy subgroups.

Since the original three subtypes were identified on autopsies, they only represent terminal states and do not capture the temporal progression. There are likely more subtypes to be found by studying the longitudinal progression based on in-vivo imaging, possibly by applying machine learning or statistical methods. A limitation of some early studies is that they are studying the progression of the disease in terms of time since the baseline visit, while patients will naturally be at different stages of the disease at study baseline. To mitigate this, data-driven models for latent-time disease staging of AD have been developed [71, 145, 105]. Notably, Vogel et al. modeled longitudinal tau-PET measurements and found four distinct subtypes (S1-S4), in addition to one without abnormal tau measurements (S0)[134]. Furthermore, non-linear mixed effects models have been utilized to stage subjects on a continuous time line by estimating a patient-level time-shift corresponding with the time of AD-onset [71, 105] (see Section 5.2). Still, there is little consensus on these subtypes and their contribution to the cognitive decline of AD patients. Likely, with modern imaging techniques, novel methods, and larger datasets, more subtypes of AD can be identified.

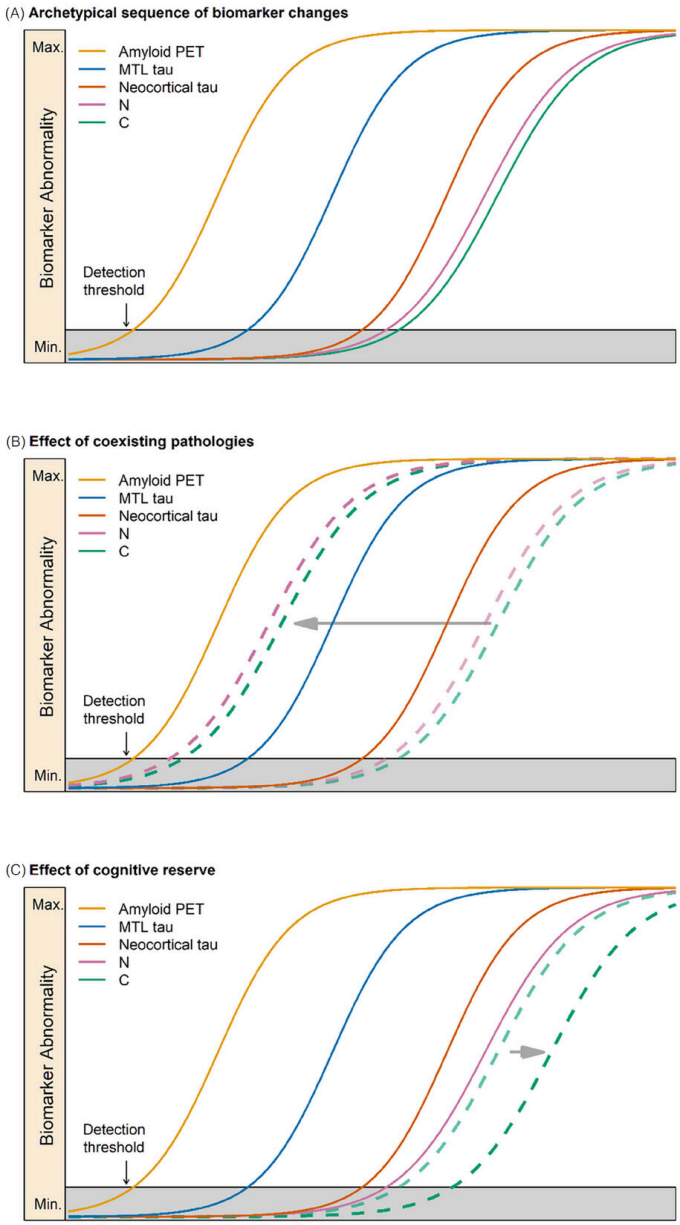


Figure 2.8: Illustration of AD staging with biomarkers. Time is on the x-axis, and the clinical abnormality of a biomarker is on the y-axis. MTL tau and Neocortical tau refers to the uptake of tau-PET for the medial temporal lobe and neocortical regions, respectively. N stands for neurodegeneration and C for clinical symptoms. (A) illustrates the typical AD progression, while (B) and (C) illustrate the effects of co-pathologies and cognitive reserve. Figure reproduced from Jack et al. [66], licensed under CC BY-NC-ND 4.0. No changes were made.

Chapter 3

Data

In this chapter the main datasets used in this thesis will be detailed. The PRIAS dataset was used in Paper I-III and the ADNI dataset in Papers IV-V.

3.1 The PRIAS dataset

To investigate the potential of keeping prostate cancer patients on active surveillance, the Prostate Cancer Research International Active Surveillance (PRIAS) study was initiated in 2006 [13]. The goal of PRIAS was to establish optimal protocols for optimal selection and follow-up on patients following an initial blood PSA test [5]. The inclusion criteria for the study can be summarized as men with histologically proven adenocarcinoma of the prostate who are fit enough for curative therapy. They could at most have two prostate cancer-positive biopsy cores, with at most ISUP grade group 2 (Gleason score 3+3 or 3+4) and at diagnosis, the PSA levels should be $\leq 10\,\mathrm{ng/mm}$ and PSA density < 0.2 [5]. Additional details on the inclusion and exclusion criteria can be found at `prias-project.org` [102]. To date, 138 medical centers across the world have contributed to the study. In this thesis we used a cohort collected under the PRIAS study, which consists of 180 patients monitored at three hospitals (Kristianstad, Lund and Malmö) in Skåne county, Sweden, from 2007 to 2020. Out of these, 35 patients discontinued the study due to some reason unrelated to prostate cancer and were therefore excluded. For the remaining 145 patients, 73 received treatment for prostate cancer and 72 remained on active surveillance through the study (See Fig. 3.1). In total, 4707 WSIs were collected over 307 individual visits. The slides were originally scanned at 40x magnification ($0.247\,\mu\mathrm{m}$/pixel), however, for the experiments in this thesis, the images were downsampled to 10x magnification.

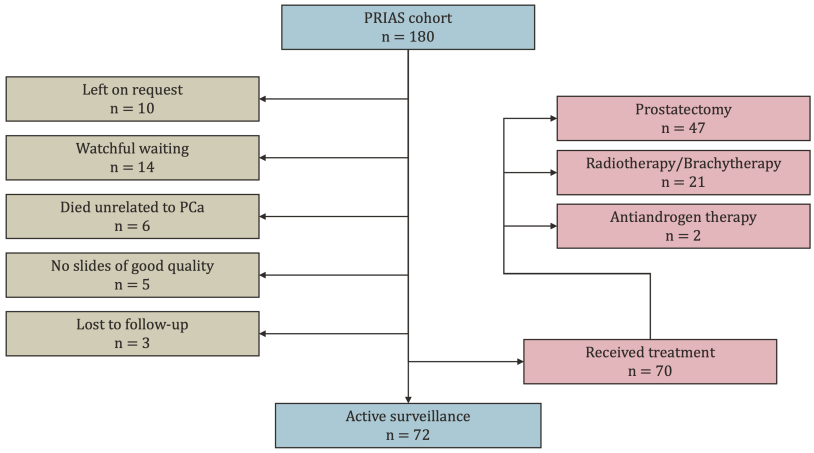


Figure 3.1: Consolidated Standards of Reporting Trials (CONSORT) diagram of the PRIAS study.

3.2 The ADNI dataset

The Alzheimer's disease neuroimaging initiative (ADNI) was launched in 2003 as a public-private partnership with the goal to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment could be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease. As stated by ADNI, current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and to provide data concerning the diagnosis and progression of Alzheimer's disease to the scientific community[1]. The ADNI database has since its release been openly accessible for researchers, hence, it has been used for numerous studies and contributed to several breakthroughs in AD research. The study has five different phases, defined by different funding cycles; ADNI1, ADNIGO, ADNI2, ADNI3 and the on-going ADNI4. All the study participants have done a screening visit consisting of clinical diagnostic evaluation and MRI. If they pass some basic inclusion criteria and accept to be enrolled in the study, they are scheduled for recurring visits approximately every 6 months for longitudinal data collection. The study currently consist of 3788 participants who passed the screening, and got a valid diagnosis (CN: 42%, MCI: 43%, Dementia: 15%). Each visit can consist of one or several examinations e.g cognitive assessment, MRI, PET (Aβ and/or tau), CSF biomarkers, blood-based biomarkers, proteomics.

[1]More up-to date information can be found at `adni.loni.usc.edu`.

Chapter 4

Machine learning techniques

Machine learning (ML) is often defined as a sub-field of the very broad term AI, where the machine teaches itself to perform some task through experience [4, 50]. The idea of machine intelligence dates back to the 19th century [87], however, it would take more than a hundred years until the technical advancements caught up with the theory and realized this idea. The introduction of ML meant moving away from hard-coding knowledge into the machines, and instead allowing them to acquire the knowledge from the data itself [50]. These classical ML algorithms require representations of the raw data. An early example is a logistic regression algorithm that is able to predict the risk of coronary artery disease of a patient [128]. Here, the patient is represented by a number of characteristics such as age, sex, body mass index, and blood values. Naturally, this algorithm's performance will depend heavily on the representations and how well they describe the cardiac health of the patient. To avoid this dependency, a representation learning algorithm can be used instead, where the machine learns both how to extract useful representations from raw data and how to use them to generate the desired output [50].

Deep learning (DL) is a family of representation learning methods that uses deep artificial neural networks (ANN) to extract useful information from the input data. A deep ANN consists of vast amounts of simple processing units (nodes) that form a complex architecture (network). The core idea is that each node is associated with a weight, which can be tuned to weigh this particular node's contribution to the output. During the training of the network, all these weights are tuned to produce the desired output. See [50] for more details on the theoretical background of these concepts. DL methods have been the driving force for the explosion of AI research and development in recent years.

In this thesis, both classical ML and modern DL methods will be used for various applications. The theory behind the concepts of classical ML and DL will be covered briefly in

Sections 4.1 and 4.2, respectively.

4.1 Classical machine learning

ML can be roughly characterized into supervised and unsupervised learning. In supervised learning, each sample of data used for training is associated with a known ground truth label or target.

As a simple example, given a set of input features $\mathbf{x} = \{x_1, x_2, ..., x_n\}$, we have a known target scalar $y \in \mathbb{R}$. The goal is to find a model $f_\theta : \mathbb{R}^n \mapsto \mathbb{R}$ that is able to predict y, given previously unseen features $\mathbf{x}$. Linear regression is a simple ML model, which is a fitted linear function such that:

$$\hat{y} = \mathbf{w}^T\mathbf{x} \ . \tag{4.1}$$

By measuring the error with respect to the target y we get a performance measure of the model. During training, the model learns how to make correct predictions by tuning the parameters $\mathbf{w}$ to minimize this error.

In unsupervised learning, the target is unknown, and the aim is to model the distribution or learn representations of the input data. While this thesis is mainly focused on supervised learning, we will discuss some self-supervised learning techniques in Section 4.6.1.

4.1.1 Support vector machines

Suppose we have a binary classification problem, where the two classes are linearly separable in some high-dimensional feature space. Support vector machines (SVM) are a type of ML model that find the optimal decision boundary [7, 68]. Here, the optimal decision boundary is the hyperplane that maximizes the separation of instances belonging to the two classes (see Fig. 4.1). The algorithm identifies a small set of data points that represent the boundary of each class, called support vectors, and finds the maximal separation between them. Let us define the training data for this model as p pairs of n-dimensional feature vectors $\mathbf{x}_i = \{x_{i1}, ..., x_{in}\}$, $i = \{1, ..., p\}$, and for each, a corresponding binary class $y_i \in \{-1, +1\}$. Similarly to the regression model, the optimal hyperplane can be described by the equation

$$\mathbf{w}\mathbf{x} - b = 0 \tag{4.2}$$

where $\mathbf{w}$ and b are the parameters of the SVM model (i.e. what we will attempt to find). For each sample in the training data we can then write

$$\mathbf{w}\mathbf{x}_i - b \begin{cases} \geq 1, & \text{if } y_i = +1 \\ \leq -1, & \text{if } y_i = -1 \end{cases} \tag{4.3}$$

i.e. we want each sample to be classified correctly by being on the correct side of the decision boundary. The margin is defined as the perpendicular distance between the decision boundary and the closest data points. The data points that are on the margin are the support vectors, and for them we have that $\mathbf{w}\mathbf{x}_i - b = y_i$. Let r be the distance between the boundary and a support vector above it (i.e. $y_i = 1$, see Fig. 4.1). Then the vector between them can be written as $r\frac{\mathbf{w}}{||\mathbf{w}||}$, where $||\cdot||$ denotes the norm of $\mathbf{w}$, and thus $\frac{\mathbf{w}}{||\mathbf{w}||}$ is a unit vector in the direction of $\mathbf{w}$. By adding this to equation 4.2, we get

$$\mathbf{w}(\mathbf{x} + r\frac{\mathbf{w}}{||\mathbf{w}||}) - b = 1 \ . \tag{4.4}$$

We notice that since $\mathbf{w}\mathbf{x} - b = 0$, by solving for r we get $r = \frac{1}{||\mathbf{w}||}$. Since this represents half of the margin, we get that the margin size is $\frac{2}{||\mathbf{w}||}$. The goal of the algorithm is to find the parameters $\mathbf{w}$ that maximize the margin, given the constraints of Eq. 4.3. Thus, this means minimizing $||\mathbf{w}||$ and can be summarized as

$$\min_{\mathbf{w},b} ||\mathbf{w}||, \text{ such that: } y_i(\mathbf{w}\mathbf{x}_i - b) \geq 1 \ \forall i \in \{1, ..., p\} \ . \tag{4.5}$$

This is referred to as the hard-margin SVM since it does not allow data points to be on the wrong side of the boundary. However, if we want to find the optimal decision boundary for a set of data points that are not linearly separable, we can use the soft-margin SVM, which introduces a set of slack variables to allow misclassifications. Details on this can be found in [7, 68]. The optimization process typically uses Lagrange multipliers to combine the objective and constraints of Eq. 4.5 into one function. In modern software such as *sklearn* in Python, this is computed automatically [97]. In Paper V, we used this software to train SVMs for simple classification tasks such as baseline diagnosis or Aβ status of study participants from the ADNI study.

4.2 Deep learning

For high-dimensional, non-linear problems, it is infeasible to fit a linear decision boundary. For these cases, deep learning-based models can be used instead. Similarly to the linear regression example above, the goal can be formulated as estimating a function $f_\theta : \mathbb{R}^n \mapsto \mathbb{R}$ that is able to predict a correct y_i given an input $\mathbf{x} = \{x_1, ..., x_n\}$. However, now the function is parametrized by an ANN. The simplest form of an ANN is the multi-layer perceptron (MLP). This consists of layers of single processing units (perceptrons), which are fully connected to form the network (see Fig. 4.2). Every edge of the network is associated with a weight, and the contributions to every node are summed and passed through a non-linear activation function. Each layer can be seen as its own function, e.g. for three layers we get $y_i = f_\theta(\mathbf{x}_i) = f_{l_3}(f_{l_2}(f_{l_1}(\mathbf{x}_i)))$. Here lies the big advantage of ANNs, by increasing

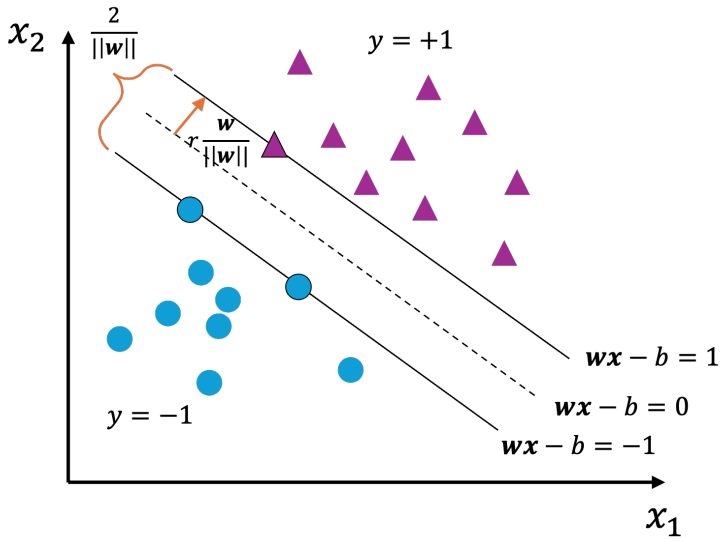


Figure 4.1: Example of a linear decision boundary in 2D, where the margin between the classes is maximized.

the depth (i.e. adding layers) or adding nodes to the layers, we can in theory, approximate any function [50]. The optimization of the parameters (i.e. weights) of the network mainly involves a loss function and an optimizer. The loss function should be a differentiable function that gives some measure of the error between the output of the network and the desired output. There are many different choices for the loss function, which depend on the task and the type of input. The optimizer calculates how the weights should be adjusted to minimize the loss function. There are also multiple choices for optimizer, but Adaptive Moment Estimation (Adam) has become a popular choice in recent years [69]. A deeper introduction to these concepts can be found in [98], and additional details on the mathematical theory can be found in [50].

4.2.1 Convolutional neural networks

Large amount of the medical data of today consists of images from different modalities. To process such data with an ANN, the MLP is not ideal since it cannot utilize spatial information. Furthermore, applying a one-layer fully-connected MLP to an image of 256-by-256 pixels and mapping to a feature vector of size 1000 would result in 65 million edges. Hence, the size of the network would be unfeasible. Convolutional neural networks (CNNs) are designed to handle data in 2D or 3D grids, such as images. As the name implies, CNNs use the mathematical convolution operation to allow parameter sharing and sparse interactions [50]. Each layer of a CNN consists of a kernel of smaller spatial dimensions than the input. The kernel holds the tunable parameters, which are convolved

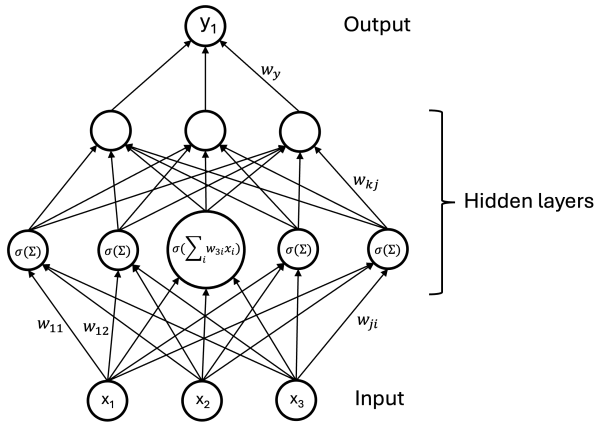


Figure 4.2: Illustration of a simple MLP. Inside each node, the contributions are summed and passed through a non-linear activation function σ. All edges are associated with a tunable parameter (weight) w_{ji}.

with the input and passed through an activation function. The kernel can be viewed as a filter that searches for features across the input array. Each convolutional layer consists of several kernels, and the CNN usually consists of several stacked convolutional layers. The first layer extracts low-level features, such as corner or edge detections, and the subsequent layer merges these into higher-level features representing larger parts of the input image [74]. A key concept of a CNN is the so-called pooling layers, which downsample the spatial dimensions of the input feature map. An example of a common pooling operation is *max pooling*, which selects the maximum output with a smaller grid. By reducing the spatial dimensions, the output becomes more invariant to translation. This means that if the input image is shifted, the output largely remains the same. For example, if we want to find whether an image contains a tumor or not, the exact location should not matter.

Improving the performance of a CNN is not as simple as just stacking more layers. In theory, the performance should not degrade by adding layers, since the network should be able to learn the identity mapping. However, empirical studies showed that for deep architectures this is not the case [56]. To address this issue, He et al. introduced residual connections between blocks of convolutional layers [56]. These are simply identity mappings that skip the convolutional block $\mathcal{F}$ and are subsequently added to the output, such as:

$$y = \mathcal{F}(\mathbf{x}) + \mathbf{x} \ . \tag{4.6}$$

The idea is that if a deeper block is unnecessary, the weights of the block can simply diminish to approximate the identity mapping ($\mathcal{F}(\cdot) \approx 0$). This type of architecture is known as the ResNet, and it has become a popular choice for many computer vision tasks.

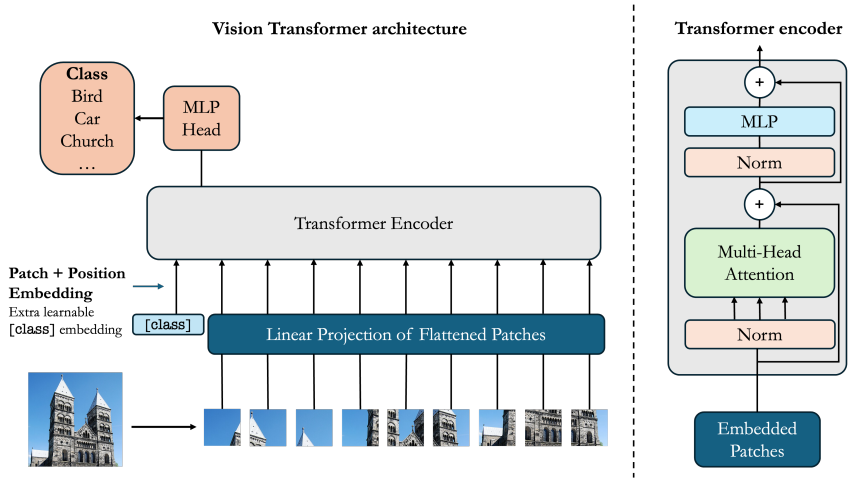


Figure 4.3: Overview of the Vision transformer architecture. The input image is split into smaller patches, which are flattened and linearly embedded. The resulting sequence is fed to the standard transformer encoder, with multi-head attention. Image inspired by Dosovitskiy et al. [32] (`https://arxiv.org/abs/2010.11929`).

4.2.2 Vision transformers

Another DL architecture type with increasing popularity in computer vision is the vision transformer (ViT) [32]. The original transformer model was developed for sequence processing, specifically for natural language processing [133]. It replaced convolutions and recurrent layers with multi-head self-attention to model long-range dependencies within the sequences. The ViT adapts this architecture to images by splitting the images into smaller fixed patches, and processing them as a sequence, using the transformer blocks (see Fig. 4.3). This enables interaction between non-adjacent patches, which is not possible with convolutions. This is particularly useful for many medical applications, where capturing the global pathological context of a larger image could be highly beneficial when making a prediction. In natural language processing, the input sequence of text is often tokenized, meaning that the letters (or combination of letters) are represented as numbers (tokens). For ViT models, the same terminology is often kept, and the input image patches are also referred to as tokens.

4.3 Generalization

The main challenge in ML is to train algorithms that perform well on previously unseen data. This ability, to *generalize* to new data samples that were not part of the training data, can be difficult to estimate, since we typically do not know the variance of the real-world data distribution. A common approach is to set aside a validation set that is used to monitor

the performance of the model during training and then use a hold-out test set to estimate the generalization performance. These datasets have to be large and diverse enough to get a good estimate. Still, we also want to keep as much data as possible for training. This is particularly a challenge in the medical field, where the amount of data is often limited.

Estimating the generalization and the dependence on the actual test data samples is important when comparing methods and evaluating the statistical significance of the results. K-fold cross-validation and bootstrapping are two methods to do this. The former involves splitting the training data into K-folds of even size, and training K separate versions of the model on K-1 folds, leaving one of the folds for validation. Bootstrapping can be used to estimate the variance on the test data by sampling N samples from the test data with replacement. This is repeated several times, and the performance is evaluated on every bootstrap sample.

Data augmentation is commonly used to extend the training data and to force the model to learn how to handle difficult cases [50]. This means that we randomly distort the input data with some transformation $\mathcal{T} : (\mathbf{x}, y) \rightarrow (\hat{\mathbf{x}}, y)$, to generate "new" training pairs. These transformations have to be designed carefully, to not change the meaning of the label y. Exploring different augmentation techniques can be an effective way of improving the generalization performance of an ML model, especially in sparse data settings. In Paper I, we developed a novel method of color augmentation to improve the performance of a segmentation algorithm.

4.4 Segmentation

Semantic segmentation is the task of classifying each pixel of an input image depending on which object depicted in the image they belong to. Semantic segmentation with DL has numerous applications in medical imaging, such as organ segmentation from CT images [8, 63], finding boundaries of tumors [63, 54], rendering 3D models [38], and segmentation of cellular components from microcopy images [62]. Commonly, this requires the usage of a fully-convolutional neural network. The difference from the standard CNN is that the learned representations are up-sampled through a decoder network, which outputs an image representing a segmentation of the input.

4.4.1 U-Net architecture

The U-Net architecture, with its many variations, has become a popular choice for biomedical segmentation tasks [107]. Originally proposed by Ronneberger et al. in 2015, the architecture introduced skip-connections between the encoder and decoder layers, to

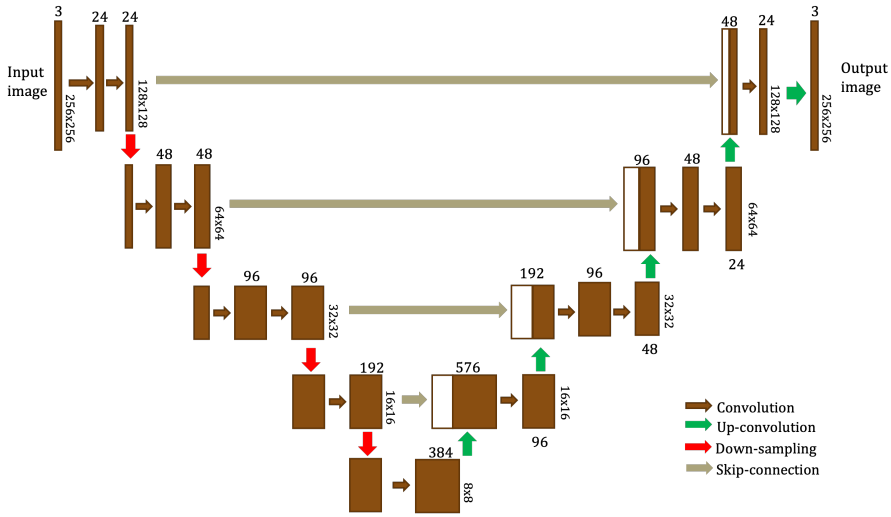


Figure 4.4: The U-Net architecture. The number of layers, the number of filters of each layer, and the step size of pooling operations are common hyper-parameters to tune depending on the application.

guide the decoder in reconstructing objects. They also used an overlapping tiling strategy to process images of arbitrary size without introducing edge effects. Since then, numerous adaptations of the architecture have been developed for different biomedical segmentation tasks [147, 63]. In Papers I and IV, we used variations of this architecture for sematic segmentation of cellular components in H&E-stained images and volumetric brain segmentation from spectral CT images, respectively.

4.5 Multiple instance learning

Application of CNNs on diagnostic tasks such as Gleason grading showed promising results in several studies [82, 14, 91]. However, for regular supervised learning instance or pixel level annotations are required for training these networks. Given that these networks' performances are highly dependent on the quality and the amount of data they are trained on, gathering these annotations quickly turns into a tedious process that requires high medical expertise. On the other hand, any larger healthcare institution has access to large amounts of weakly annotated data. For instance, a cancer patient may have a set of collected images paired with a diagnosis set by the physician. This fits very well with the multiple instance learning (MIL) problem formulation, where a bag of instances is associated with a single label for the whole bag [31]. Recall the supervised learning formulation from Section 4.1. Now, we have a bag of K instances $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_K\}$, and a known label Y associated with the bag. For each instance within the bag, there is a latent label $y_i \in \{0, 1\}$, which

gives the bag label as

$$Y = \begin{cases} 0, \text{ if } \sum_{i=1}^{K} y_i = 0 \\ 1, \text{ otherwise.} \end{cases} \tag{4.7}$$

There are two main approaches for solving this task: the instance-level approach and the embedding-level approach [61]. For the instance-level approach, a classifier returns predictions for all instances, which are pooled to obtain the bag-level prediction. In the embedding-level approach, instances are mapped to low-dimensional embeddings, which are pooled into one representation of the bag. A classifier is then used to predict the bag label based on this representation. In both approaches, the MIL pooling operation needs to be permutation invariant (e.g. the order of images does not affect the diagnosis). If the pooling operation is also differentiable, it can be used as part of an ANN architecture. Two basic operators that fulfill these requirements are max pooling and average pooling. However, both these operations have the major drawbacks that they are non-trainable and cannot be adjusted to fit any particular task. Furthermore, they also lead to a significant loss of information when used in the training of an ANN. For example, with the max pooling operator, the signal used to update the whole network comes from only a single instance, since the rest are discarded. This makes learning slow and data hungry compared with regular supervised learning [17].

To mitigate these issues, Ilse et al. introduced Attention-based MIL pooling [61]. The pooling operation is a weighted average of the instances, where the attention weights are learned by an ANN. For K instance embeddings $\mathbf{h}_1, .., \mathbf{h}_K$, the pooling is defined as

$$\mathbf{z} = \sum_{i=1}^{K} \alpha_i \mathbf{h}_i \tag{4.8}$$

with

$$\alpha_i = \frac{e^{\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_\mathbf{i}^\mathbf{T})}}{\sum_{j=1}^{K} e^{\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_\mathbf{j}^\mathbf{T})}} \tag{4.9}$$

where $\mathbf{w} \in \mathbb{R}^L$ and $\mathbf{V} \in \mathbb{R}^{L \times N}$ are learned parameters of MLPs. The $\tanh(\cdot)$ activation function is used to allow both negative and positive values. To allow learning of more complex relationships, they extend this by adding a gating mechanism to the attention calculation:

$$\alpha_i = \frac{\exp[\mathbf{w}^T(\tanh(\mathbf{V}\mathbf{h}_i) \odot \text{sigm}(\mathbf{U}\mathbf{h}_i))]}{\sum_{j=1}^{K} \exp[\mathbf{w}^T(\tanh(\mathbf{V}\mathbf{h}_j) \odot \text{sigm}(\mathbf{U}\mathbf{h}_j))]}, \tag{4.10}$$

where $\mathbf{U} \in \mathbb{R}^{L \times N}$ is another set of parameters, $\text{sigm}(\cdot)$ is the sigmoid activation function and $\odot$ is an element-wise multiplication. The gated attention mechanism is illustrated in Fig. 4.5.

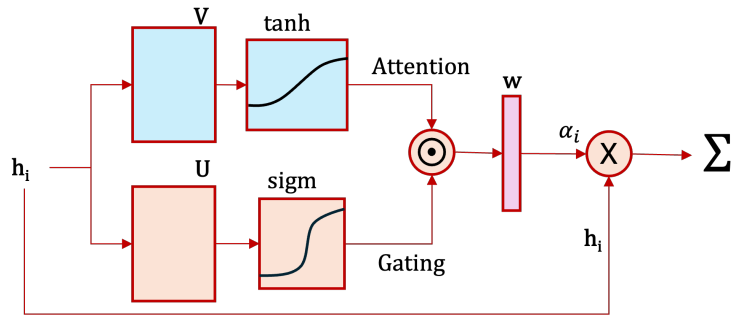


Figure 4.5: Illustration of the gated attention mechanism introduced by [61].

In computational pathology, numerous recent studies have successfully applied MIL approaches to many different applications [17, 89, 78, 47]. Inherently, the MIL formulation fits well with these problems, since a WSI can be seen as a bag of smaller image patches and the cancer grade or diagnosis as the label. Since this type of weak labels are generated in standard clinical practice, these approaches can make use of the large amounts of pathology data collected within the healthcare systems without the need for any instance-level annotations. The attention-based MIL (ABMIL) approach also has the advantage that the attention weights can be used to find key instances with high attention, i.e. the regions of the WSI that influence the cancer grade can be identified [61]. One of the most noteworthy contributions that applied ABMIL in computational pathology is *clustering-constrained-attention multiple instance learning* (CLAM), developed by Lu et al. in 2021 [78]. In their pipeline, they begin by segmenting the tissue of the WSI using Otsu's algorithm, which is a non-learned global thresholding method [94]. The tissue containing parts of the WSI is divided into smaller patches, which are converted into feature vectors by a feature extractor. They used an ResNet model, trained on the ImageNet dataset, which contains over 14 million natural images annotated with thousands of categories [109].

4.6 Foundation models

While MIL frameworks such as CLAM achieved strong results on a variety of tasks, the feature extractor used was not trained on medical data. This begs the question of how this framework would perform if the feature extractor was trained on pathology image data alone. However, there exist no annotated, open-access datasets of comparable size to ImageNet, which would be required for training such a model. On the other hand, many institutions have access to large amounts of unlabeled pathology images, which can be used in a self-supervised approach. These types of general-purpose models, pre-trained on large amounts of data, are often referred to as foundation models.

4.6.1 Self-supervised learning

Modern foundation models are largely built on self-supervised learning, which first demonstrated its tremendous scalability in natural language processing with the development of large language models (LLMs) [11, 30]. This learning approach enabled scraping the enormous amounts of text available on the internet for training these models, and paved the way for the many powerful LLMs of today. In self-supervised learning, the signal used to update the weights of the network is generated from the data itself. This signal comes from defining a pretext task, i.e. a task that has no real-world use, but forces the model to learn useful representations of the data. Masked language modeling is an example from LLMs, where language tokens in the sequences are randomly masked, and the objective for the model is to recover them [30]. This approach was later adapted for image data and pre-training of ViTs in the *Image BERT pre-training with online tokenizer* (iBOT) framework [146]. Similarly, given an image tokenized into a sequence of N tokens (i.e. patches) $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, a random mask $\mathbf{m} \in \{0, 1\}^N$ is applied, yielding a corrupted image $\hat{\mathbf{x}} = \{\hat{\mathbf{x}}_i | (1 - m_i)\mathbf{x}_i + m_i \mathbf{e}_{[mask]}\}_{i=1}^N$, where $\mathbf{e}_{[mask]}$ is an arbitrary mask token. The task is then to recover the masked tokens from this corrupted image.

They also leveraged the concept of self-distillation, previously proposed in the DINO framework [19]. The main idea of knowledge-distillation in DL is that a smaller student network $\mathbf{G}_s$ is trained by matching the output of a teacher network $\mathbf{G}_t$. Given that both networks output probability distributions P_s, P_t, for an input image $\mathbf{x}$, the weights of the student network θ_s are updated by minimizing the cross-entropy between the probabilities as:

$$\min_{\theta_s} -P_t(\mathbf{x}) \log(P_s(\mathbf{x})), \tag{4.11}$$

Furthermore, to make the objective non-trivial, the networks are fed with two separate randomly distorted views $\mathbf{x}'_1, \mathbf{x}'_2$ of $\mathbf{x}$. However, in self-distillation, there is no teacher network, since we have no known ground truth. Instead, the teacher is built upon previous iterations of the student network. In DINO, they found that using exponential moving average updates of the student network gave the best results. The teacher network parameters are updated as $\theta_t \leftarrow \lambda\theta_t + (1 - \lambda)\theta_s$. In iBOT, the same approach is used with the addition of the masked image tokens being fed to the student network, and the teacher making predictions on the non-masked tokens. This pre-training strategy was further refined in the DINOv2 framework, which added several technical additions to improve training stability and efficiency [93]. By also adding a pipeline of data curation, they achieved promising results on several downstream tasks.

4.6.2 Foundation models in histopathology

The first large-scale foundational model for histopathology (CTransPath) introduced a new self-supervised learning strategy and a hybrid model, which combined a CNN and a ViT [136]. Following this, Chen et al. applied the DINOv2 framework when training the computational pathology specific foundation model UNI [24]. They gathered at the time one of the world's largest datasets of histopathology images; an in-house dataset called Mass-100K, with over 100 million image patches from 100 426 diagnostic H&E stained WSIs and over 20 different tissue types. This model was later updated, with a larger ViT model trained on more than 350k WSIs. In papers II and III, we utilized UNI and UNI-2 for feature extraction, respectively.

Chapter 5

Statistical methods

The foundations of ML are rooted in statistical theory, and some of the basic ML algorithms are also frequently applied in statistics. Hence, the key difference between statistical methods and ML is not in the methods themselves, but rather what they are used for. ML models are mainly predictive, and in many cases, the process from input data to prediction is unimportant. For example, DL-based models are often black-box models, where we cannot explain why the model makes a certain prediction. On the contrary, statistical methods are inference-based and are used to explain the world. For example, with statistical methods, we can test hypotheses, quantify uncertainty, investigate effects, etc. An example in the context of this thesis is if we want to investigate whether a specific brain region is associated with the development of Alzheimer's disease. Statistical methods can also be used to validate probabilistic predictions made by an ML model, e.g. with survival analysis, which we discuss in Section 5.3.

5.1 Statistical modeling

Recall the regression model from Section 4.1. This type of model can be applied in a statistical context to investigate the relationship between y and $\mathbf{x}$. With statistical notation, a regression model is typically formulated as

$$y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon_i = \mathbf{x}_i^T \beta + \epsilon_i \,, \tag{5.1}$$

for the i:th observation y_i, where ϵ_i is a random error, assumed to be identically and independently distributed (i.i.d) across observations. The parameters β can be estimated by minimizing the sum of the squared errors, i.e. least squares estimation [106]. In matrix notation we get the estimates; $\hat{\mathbf{y}} = \mathbf{X}\beta$, which gives the least squares estimates of the parameters as $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

5.1.1 Maximum likelihood estimation

If we want to fit a probabilistic model, least squares estimates might not be feasible or ideal. For example, least squares estimation does not restrict predictions to stay within $[0, 1]$ and it assumes constant variance. For a probabilistic model, the variance will be close to zero at the boundary points $p \approx 0$, $p \approx 1$, and large when $p \approx 0.5$. Instead, maximum likelihood estimation (MLE) is commonly used. To explain this, we will use the simple example of a logistic regression model for predicting risk of coronary artery disease from Chapter 4 [128]. The model is formulated as

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi} = \beta^T \mathbf{x}_i \,, \tag{5.2}$$

where p_i is the probability of the i:th subject of getting the disease, and $\mathbf{x}_i$ their predictors (e.g. age, blood values etc.). This gives the formula for the probability $p_i = \frac{e^{\beta^T \mathbf{x}_i}}{1-e^{\beta^T \mathbf{x}_i}}$. The main idea of MLE is to find the parameters β that maximize the likelihood that we get the observations that we have [1]. Since we have a binary outcome, the probability of a subject not getting the disease is simply $q_i = 1 - p_i = \frac{1}{1-e^{\beta^T \mathbf{x}_i}}$. This gives us the likelihood for observing outcome y_i as

$$P(y_i|\mathbf{x}_i, \beta) = \frac{e^{y_i \beta^T \mathbf{x}_i}}{1 - e^{\beta^T \mathbf{x}_i}} \,. \tag{5.3}$$

We can then formulate the likelihood of observing all outcomes in our data $\mathbf{y}$ given our measured data $\mathbf{X}$ as the product

$$\mathcal{L}(\mathbf{y}|\mathbf{X}, \beta) = \prod_i \frac{e^{y_i \beta^T \mathbf{x}_i}}{1 - e^{\beta^T \mathbf{x}_i}} \,, \tag{5.4}$$

which is the likelihood function and what we want to maximize. By taking the logarithm, this translates to the sum

$$\log \mathcal{L}(\mathbf{y}|\mathbf{X}, \beta) = \sum_i y_i \beta^T \mathbf{x}_i - \sum_i \log(1 + e^{\beta^T \mathbf{x}_i}) \,. \tag{5.5}$$

Then the derivative w.r.t β can be simplified into

$$\frac{\partial \log \mathcal{L}}{\partial \beta} = \mathbf{X}^T (\mathbf{y} - P(\mathbf{y} = 1|\mathbf{X}, \beta)) \,, \tag{5.6}$$

which is minimized in the optimization process. This expression does not have a closed form solution, but there are several algorithms that can be used to find a solution, such as the expectation maximum (EM) algorithm [7].

The EM algorithm is an iterative process, where we start with an initial guess or estimate of the model parameters $\hat{\beta}$. This estimate is then used to formulate the expected log-likelihood function as a function, Q, of the next proposed set of parameters β, i.e.

$$Q(\beta, \hat{\beta}) = \mathbb{E}(\log \mathcal{L}|\beta) = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}, \hat{\beta}) \cdot \log \mathcal{L}(\mathbf{y}|\mathbf{x}, \beta)\,. \tag{5.7}$$

The second term of the sum is the log-likelihood given the new suggestion of parameters β, and the first term is the posterior distribution, or in other words, our current best guess. The first term acts as a weight of importance for the data points since a point that is highly unlikely to be observed should not contribute to the updating of the parameters. In the maximization step of the algorithm, we get the new set of parameters $\beta^* = \arg\max_\beta Q(\beta, \hat{\beta})$. If some convergence criterion is not satisfied, the estimated parameters are updated $\hat{\beta} \leftarrow \beta^*$, and the expectation step is repeated (i.e formulation of Eq. 5.7).

5.2 Mixed effects models

For the previous examples, a key assumption is that the errors are i.i.d., which often does not hold in a real-world scenario. As an example, when analyzing the performance of students, their results will depend on which class and which school they belong to [130]. This is called hierarchical data, since we have three levels of grouping, the first being the individual students, the second the classes they belong to, and the third which school. This scenario can be modeled with mixed effects models, which model the outcome with a combination of *fixed* and *random* effects. By fixed effects, we mean effects that are constant across the population, e.g., in the example of students' performance, this can be the amount of time spent studying. The random effects allow variations from the constant fixed effects, e.g. a random intercept depending on which school the students are enrolled in, or a random slope that describes the effect that a particular teacher has on a class's performance. The general form of a mixed linear model is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\nu + \epsilon\,, \tag{5.8}$$

where $\mathbf{y} \in \mathbb{R}^{N\times 1}$ are the responses, $\mathbf{X} \in \mathbb{R}^{N\times p}$ and $\mathbf{Z} \in \mathbb{R}^{N\times q}$ are the design matrices for the fixed and random effects (i.e. observed predictors), and $\beta \in \mathbb{R}^{p\times 1}$ and $\nu \in \mathbb{R}^{q\times 1}$ are the estimated fixed and random parameters respectively. The expected values of the responses are $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$, and the random effects ν, and random residual error $\epsilon \in \mathbb{R}^{N\times 1}$ are assumed to be normally distributed such as

$$\begin{bmatrix} \nu \\ \epsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G} & 0 \\ 0 & \mathbf{R} \end{bmatrix} \right)\,, \tag{5.9}$$

where $\mathbf{G}$ and $\mathbf{R}$ are matrices that depend on a variance parameter θ [106]. During the fitting of the model, the parameters β and ν are estimated along with the variance θ.

5.2.1 Restricted maximum likelihood

Fitting mixed effects models with MLE can provide a biased estimation of the variances. The estimation of the variances in mixed models with MLE is essentially a two-step process where the mean (i.e. $\mathbf{X}\hat{\beta}$) is estimated first, and the variance, θ is subsequently estimated under the assumption that the estimated mean is the true mean. However, this assumption leads to the variances being estimated as too small, especially in a setting with limited data. A less biased option is the *restricted maximum likelihood* (REML) method [106]. We will not go into the mathematical details of this method, but it corrects the variance estimation by accounting for degrees-of-freedom lost in the mean estimation. In other words, for the variance estimation, the data $(\mathbf{X}, \mathbf{y})$ is transformed to remove the influence of the estimated fixed effects.

5.2.2 Disease progression modeling

Due to their ability of modeling temporal relationships while accounting for subject-level variation, mixed effects models are suitable for data-driven disease progression modeling. This type of modeling can be defined as using short-term in-vivo measurements of biomarkers to construct long-term disease trajectories [145] (see Fig. 5.1). It often also involves estimating an individual disease time along this trajectory. If the assumption of linearity does not hold, non-linear mixed effects models are commonly used [71]. These are similar to the linear mixed models discussed, with the addition that the relationship between the observed data and the fixed and random effects no longer has to be linear. These models enable simultaneous modeling of both individual biomarker trajectories and global population-level progression. However, they inherently can be difficult to optimize and need to be carefully designed to fit the task and available data.

In Paper IV, we modeled longitudinal trajectories of MRI measurements of patients with suspected Alzheimer's disease using non-linear mixed effects models. We also employed a disease progression model to predict latent time-shifts for each individual. To facilitate the optimization, we used REML for parameter estimation using the R package 'nlme' [101].

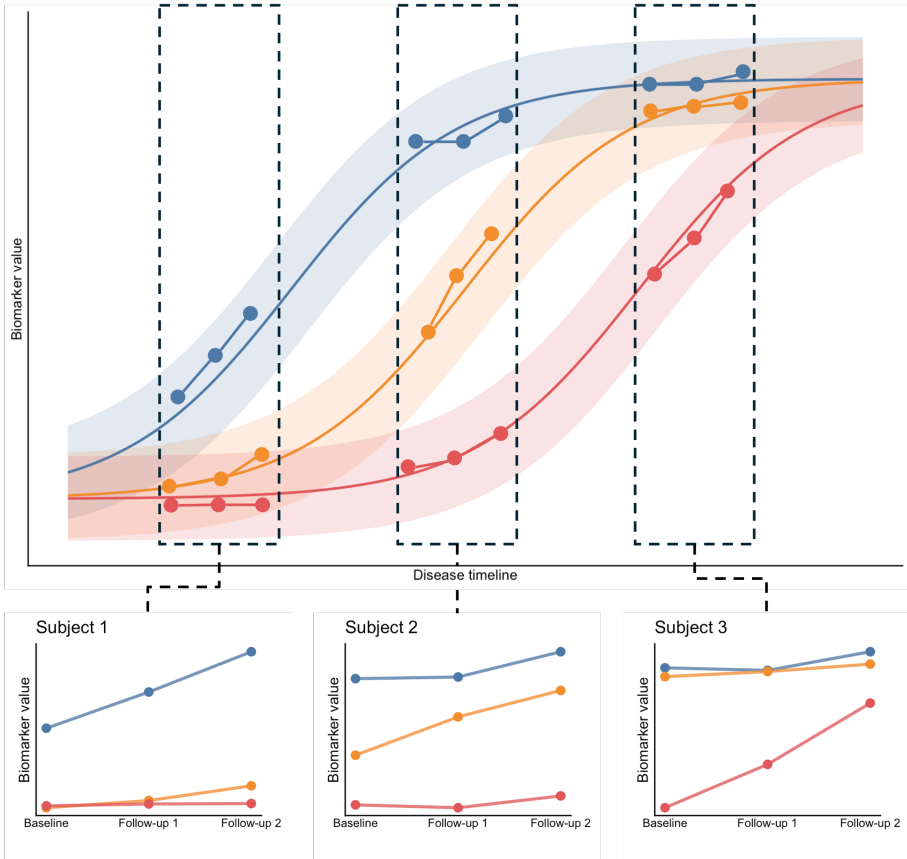


Figure 5.1: Illustration of disease progression modeling of three biomarkers. The bottom shows short-term trajectories of three individuals, indexed by baseline and two follow-up visits. They are staged on the population-level disease trajectory above. Disease progression models simultaneously learn a data-driven time axis (x-axis), a set of biomarker trajectories, and the alignment based on individual disease times.

5.3 Survival analysis

Survival analysis is a collection of statistical methods for data analysis where the outcome variable is time until an event occurs [70]. In medicine, the event can mean disease incidence, disease stage progression, recovery, death, or any other discrete event related to a medical condition. Regardless of which event, the term survival time is often used for the time until the event. These methods can be highly valuable when evaluating treatment responses or finding predictors of certain disease states. Furthermore, they can also be used to bridge the gap between image data and longitudinal outcome measures, as we discuss in this thesis.

Censoring is a key concept in survival analysis, which occurs when we have some informa-

tion about an individual, but we do not know the exact survival time [70]. More concretely, censoring occurs when a person does not experience the event until the end of the study, the person is lost to follow-up during the study, or they withdraw from the study. These are all examples of right-censored data since we only know that they had not experienced the event up until a certain point. While less common, left-censored data can also occur in survival analysis. This means that the survival time can be less than measured. For example, if a patient tests positive for a certain disease, we know that they had the disease at the time of the test, but not exactly when it occurred. If an individual has two tests, one negative at time t_1 and subsequently one positive at t_2, then the data is interval-censored, since we know that the true survival time is within $[t_1, t_2]$.

The goal of survival analysis methods can be summarized as estimating, interpreting, and analyzing the survivor or hazard functions [70]. The survivor function $S(t)$ is defined as the probability that an individual survives longer than the time t, i.e $S(t) = P(T > t)$, where T denotes the random variable that describes an individual's survival time. The hazard function $h(t)$ gives the instantaneous risk that the event will occur at time t for an individual who has survived up until time t. Mathematically, it is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \tag{5.10}$$

where T is the individual survival time variable. Hence, the numerator gives the conditional probability that T will be in the time interval t to $t + \Delta t$, given that $T \geq t$. Since the denominator is the small interval Δt, the hazard function becomes a rate rather than a probability, which depends on the unit of time used. It has the properties that it is always greater than or equal to zero and has no upper bound. The survivor and hazard functions are related as

$$S(t) = e^{-\int_0^t h(u)du} \text{ and } h(t) = -\frac{\frac{d}{dt}(S(t))}{S(t)}, \tag{5.11}$$

i.e. if one is known, the other can be computed directly.

When modeling survival data, it is common to compare the hazard functions between different groups, e.g. to measure how the survival is impacted by a certain treatment. This measure of effect in a survival model is called the hazard ratio, and is defined as the ratio of hazard rates between two groups. Similarly, for the logistic regression model, it is often expressed as an exponential of a parameter, e^{β}. Thus, if there is no difference in hazard between the two groups, the hazard ratio will be close to 1.

5.3.1 Kaplan-Meier curves

A method to estimate and plot the survivor function is the construction of Kaplan-Meier (KM) survival curves [70]. They are non-parametric step-functions that describe the prob-

ability of surviving past time t for possibly censored data. To make the estimation $\hat{S}(t)$, the following columns of a data table is required; ordered survival times t_i of all individuals (including the starting point $t_0 = 0$), number of failures (events) m_{t_i} at a specific survival time, number of censored individuals q_{t_i} up until the next survival time and the number of individuals who survived to at least time t_i, n_{t_i}. If no individuals are censored before t_0 (e.g. they do not have the condition before the start of study), the n_{t_i} column will contain all individuals in the top row and the other columns will be zero, and we get $\hat{S}(t_0) = 1$. The next row will correspond to the time of first failure(s) t_1, and so on. Then the general KM formula is formulated as

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \times P(T > t_i | T \geq t_i) = \hat{S}(t_{i-1}) \frac{n_{t_i} - m_{t_i}}{n_{t_i}}\,, \tag{5.12}$$

where t_{i-1} denotes the previous row. The probability is estimated as the number of individuals who survived past t_i out of the number of individuals at risk. The number of individuals at risk is updated for every row as: $n_{t_i} = n_{t_{i-1}} - (m_{t_{i-1}} + q_{t_{i-1}})$, i.e. the number of censored individuals is also removed.

5.3.2 The Cox proportional hazards model

The Cox proportional hazards (PH) model (sometimes referred to as the Cox regression model) is a popular choice for modeling survival data [70]. It is commonly written as a formula in terms of the hazard function as

$$h(t, \mathbf{x}) = h_0(t) e^{\sum_{i=1}^{p} \beta_i x_i}\,, \tag{5.13}$$

where $\mathbf{x} = \{x_1, ..., x_p\}$ is the set of predictors, $\beta = \{\beta_1, ..., \beta_p\}$ are the parameters and $h_0(t)$ is the baseline hazard function. It is called the proportional hazards model since the hazard ratios remain constant over time, given that the predictors, x_i are time-independent. The baseline hazard function is an undefined function, which makes this a semi-parametric model. This is one of the reasons why the Cox PH model is so popular; we do not have to know the baseline hazard function and still get good estimates of the parameters β. This is possible with an MLE that uses a partial likelihood function, which only considers the ordering of survival times. Similar to eq. 5.4, it is a product of likelihoods, however, each term corresponds with a survival time, i.e. $\mathcal{L} = \mathcal{L}_1 \times \mathcal{L}_2 \times ... \times \mathcal{L}_k$ for a dataset with k unique survival times. For a survival time s, the likelihood term is calculated as

$$\mathcal{L}_s = \frac{h_0(t) e^{\beta \mathbf{x}_s}}{\sum_{i=s}^{k} h_0(t) e^{\beta \mathbf{x}_i}}\,, \tag{5.14}$$

i.e. it is the hazard of the subject who experienced the event at s divided by the hazard of all subjects still at risk. Here, all the baseline hazard terms will cancel out, and therefore,

this can be an arbitrary hazard function. This likelihood is maximized by setting the partial derivatives of the logarithm of $\mathcal{L}$ to zero and solving the resulting system of equations.

The hazard ratio of the i-th predictor is defined as e^{β_i}. It can be interpreted as the change in baseline hazard if the predictor x_i is increased by one unit (i.e. 1). A large hazard ratio indicates that this particular variable increases the risk i.e. the probability of experiencing the event.

5.3.3 The extended Cox model

For many applications, we want to make use of time-dependent variables in survival analysis. For example, in the context of Alzheimer's disease, we know that accumulating Aβ plaques, as measured by PET imaging, is a strong predictor. Thus, to model the survival of subjects at risk for the disease, we need to include a time-dependency in the model, and the PH assumption does not hold anymore. Instead, we can use the extended Cox model, formulated as

$$h(t, \mathbf{x}(t)) = h_0(t) \exp\Big(\sum_{i=1}^{p_1} \beta_i x_i + \sum_{j=1}^{p_2} \delta_j \chi_j(t) \Big), \tag{5.15}$$

where $\mathbf{x}(t) = \{x_1, ..., x_{p_1}, \chi_1(t), ..., \chi_{p_2}(t)\}$ contains both time-independent and time-dependent variables. Similar as with the Cox PH model, the parameters are estimated with an MLE procedure with partial likelihoods. While the computation is more complex for this model, it is still only based on the order of events and does not require a known baseline hazard function. For more details, see Chapter 6 in [70].

For the extended Cox model, the hazard ratio is also time-dependent and has the general formula:

$$\widehat{HR}(t) = \exp\Big(\sum_{i=1}^{p_1} \beta_i [x_i^* - x_i] + \sum_{j=1}^{p_2} \delta_j [x_j^*(t) - x_j(t)] \Big), \tag{5.16}$$

where $\mathbf{x}^*$ and $\mathbf{x}$ denotes the two sets of predictors that we compare. The δ_j term does not depend on the time and thus represents the overall effect of the corresponding time-dependent variable. In Paper III, we evaluated the longitudinal predictive power of our models by fitting both time-independent and time-dependent Cox models.

Chapter 6

Results and discussion

In this chapter, we will summarize the main scientific contributions of this thesis, based on the included publications. We divide this part into three sections according to the thesis aims listed in Chapter 1.

6.1 Addressing data limitations

In the field of medical image analysis, the lack of high-quality and diverse datasets is a recurring limitation when developing machine learning algorithms. In Papers I, V, and VI, we address this issue in different ways.

6.1.1 Paper I: Systematic augmentation in HSV space for semantic segmentation of prostate biopsies

As mentioned in Section 4.3, data augmentation is an effective way of increasing and diversifying the training data when training ML models. Specifically for H&E stained images, effective color augmentation can mitigate the issue of stain variation across different scanners and hospitals. In this paper, we investigated techniques for color augmentation of H&E stained prostate biopsies. We evaluated our methods on the task of semantic segmentation of prostate tissue into three basic components: stroma, epithelial cytoplasm, and nuclei, separated from the background (See Fig. 6.1). For this task, we used a basic U-Net architecture, trained on a small, manually annotated subset of prostate biopsies, taken from the PRIAS dataset. For color augmentation, it is often useful to transform the images to the cylindrical hue-saturation-value (HSV) color space. The main benefit is that the hue, which describes the color, is represented by only one value (the angle of the cylinder) as

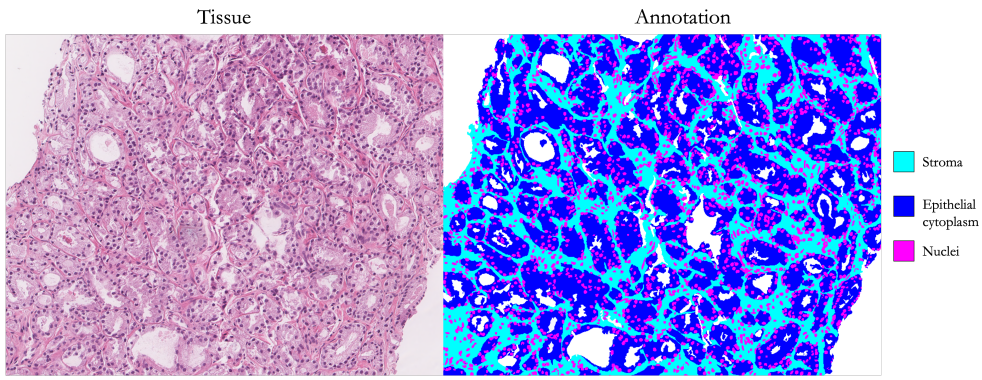


Figure 6.1: Sematic segmentation of prostate tissue. The left-hand image shows the H&E stained tissue and the right-hand image shows the corresponding segmentation classes: stroma, epithelial cytoplasm, and cell nuclei.

opposed to three in RGB. A simple shift of this angle will generate new colors for the image. This will also naturally keep the relative coloring among objects in the image constant, which would not be the case if we shifted the channels in a standard RGB image. This is particularly useful in histopathology, since the colors correlate with relevant histological information.

Our main contribution in this paper was a method designed to ensure an evenly distributed color augmentation based on the stereographic projection of the hue channel. With this method, we achieved the highest average Dice score across the tissue classes. Furthermore, we also found that applying strong color augmentation generally improved results, even if this meant generating images with colors that would never appear in a real H&E stained WSI.

6.1.2 Paper V: Synthetic Alzheimer's disease dataset generation and evaluation with privacy protection

The lack of clinically relevant open-access datasets has become a bottleneck for Alzheimer's disease research. Collecting these datasets is expensive and time-consuming, given the high number of advanced modalities required for characterization of AD (see Section 2.2). Furthermore, these types of studies involve vulnerable research participants, and sharing their sensitive personal data is restricted by privacy agreements. A proposed solution is to generate synthetic datasets, which could be shared openly without privacy concerns. These datasets would need to fulfill key requirements such as protecting the privacy of the real subjects' data used to generate them and preserving the clinically relevant relationships between the variables. In this paper, we leveraged tabular data from the ADNI and A4 studies for synthetic data generation (see Section 3.2). We evaluated four different frameworks for

synthetic tabular data generation: DataSynthesizer [100], CTGAN [143], TabPFN [57], and Synthpop [92]. We evaluated the privacy and utility of the generated synthetic datasets with the open-source framework SynthEval [73]. We found that the Synthpop and DataSynthesizer methods were the only evaluated methods able to generate somewhat realistic synthetic datasets. However, this comes at a cost of lower privacy ratings. Adding a ϵ-differential privacy constraint to the DataSynthesizer model reduced the utility and increased the privacy ratings considerably. However, the utility was reduced to such a degree that this synthetic data would not be useful. Hence, more work is needed to both improve the utility of the synthetic datasets while keeping the privacy of the data at an acceptable level. Furthermore, to generate a clinically relevant synthetic cohort, both longitudinal relationships and higher dimensionality with more complex attributes need to be modeled.

6.1.3 Paper VI: Dual energy CT and deep learning for an automated volumetric segmentation of the major intracranial tissues: Feasibility and initial findings

MRI remains the standard choice of imaging modality for monitoring neurodegenerative brain diseases such as Alzheimer's disease, due to its superior capability in differentiating brain tissues with high resolution. However, CT imaging is often used in clinical practice for other applications because of its fast acquisition times, wide availability, and ability to provide detailed bone imaging. Therefore, a robust method for segmenting brain tissue from CT images could be highly valuable for generating volumetric data [48]. With the use of modern dual-energy CT, virtual mono-energetic images (VMIs) can be generated. These are blends of the two energy levels used, and provide a simulated mono-energetic CT image at a specific energy level [86], which can improve differentiation between gray and white brain matter. In this paper, we investigated the potential of using VMIs for developing a DL-based algorithm for automatic volumetric segmentation of the intracranial brain tissues (white matter (WM), gray matter (GM), cerebrospinal fluid (CSF)). We compared four variations of the U-Net architecture [107], three variations of the U-Net++ architecture [147], and the *CTseg* MATLAB algorithm [12] (for details on each variation, see Paper VI). The best performing model was the U-Net++ model with simply using the VMIs as different augmentations of the same data sample (see Fig. 6.2). This model achieved a Dice score of 0.84 (CI $0.81 - 0.86$), 0.77 (CI $0.76 - 0.79$), and 0.88 (CI $0.84 - 0.92$) for WM, GM, and CSF, respectively. While it did not achieve the smallest Hausdorff distance, it remained low at $2.9\,\mathrm{mm}$ (CI $2.6 - 3.1\,\mathrm{mm}$) for WM, $2.4\,\mathrm{mm}$ (CI $2.2 - 2.7\,\mathrm{mm}$) for GM, and $2.0\,\mathrm{mm}$ (CI $1.9 - 2.2\,\mathrm{mm}$) for CSF. Generally, all models performed similarly with Dice scores ranging from $0.79 - 0.84$ for WM, $0.70 - 0.77$ for GM, and $0.83 - 0.86$ for CSF. We also did not achieve comparable performance to algorithms trained on MRI data. However, with a final cohort of only 26 patients, these results would likely improve with more data.

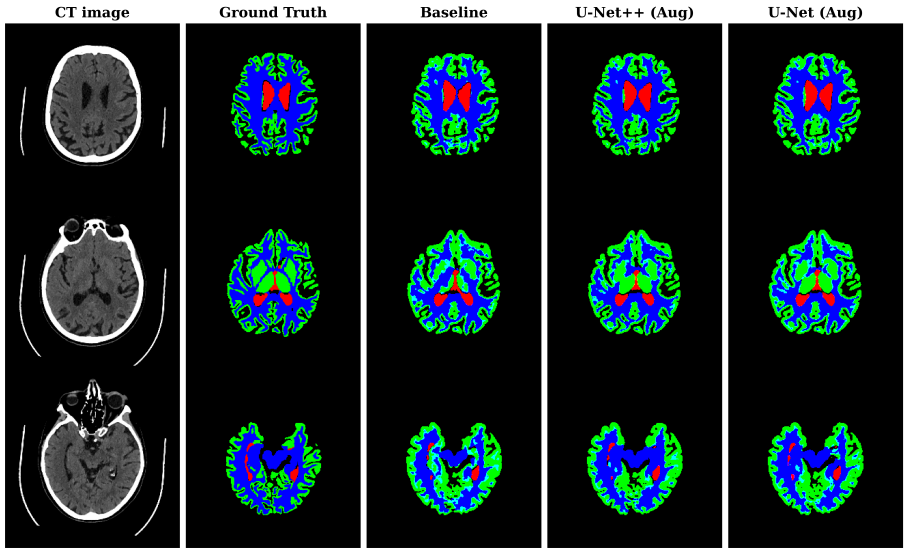


Figure 6.2: Segmentation results from Paper VI. From left to right: CT-image (VMI at 70 keV), MR-based ground truth, baseline U-Net++, best performing U-Net++ (Aug), and U-Net (Aug). Gray matter—green, White matter—blue, Cerebrospinal fluid—red, and Background—black.

6.2 Longitudinal outcome prediction of prostate cancer patients on active surveillance

In Papers II and III, we utilized the PRIAS dataset to develop a framework for longitudinal outcome predictions of prostate cancer patients on active surveillance.

6.2.1 Paper II: Outcome prediction of prostate cancer patients on active surveillance using weakly supervised deep learning

Previous studies have proven the effectiveness of ABMIL-based approaches for slide-level predictions [78, 89, 144]. In Paper II, we attempted to expand this approach to patient-level predictions. This posed a significant challenge in handling the large data compression required, as we had several giga-pixel WSIs as input, and only one binary label as output: should this patient receive treatment for their prostate cancer after this visit. Similar to the CLAM framework [78], our framework consisted of two stages: feature extraction and MIL-based outcome prediction. After dividing the tissue section of each WSI into patches, we used three different feature extractors to convert the patches into representations that could be processed in the subsequent MIL module (see Fig. 6.3). We pre-trained a DenseNet-201 CNN to classify patches from the Gleason cohort into the classes of benign and malignant (we called this the GGNet) [82]. By removing the final classification

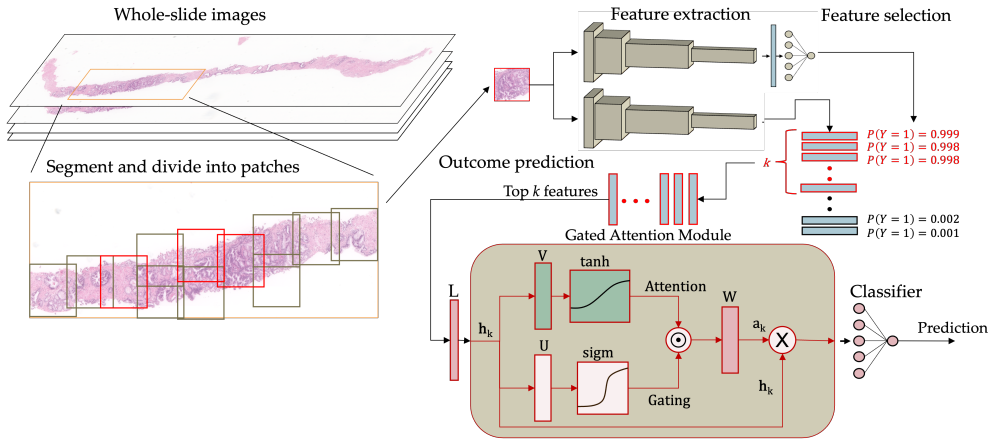


Figure 6.3: Illustration of the framework used in Paper II. In the first stage, the tissue is roughly segmented and divided into patches. The pre-trained feature extractor then transforms the patches into latent representations. The output feature vectors are ranked based on predicted malignancy by our pre-trained GGNet. The top k feature vectors are selected for the ABMIL-based outcome predictor (OP), which makes the final prediction.

layer of this model, we could use its computed representations, which would hypothetically contain relevant information for the treatment decision. We compared this feature extractor with another DenseNet-201 model, pre-trained on the ImageNet database and the UNI foundation model [24]. To refine the set of patch representations, we used the GGNet in parallel to the feature extractors to output probabilities of malignancy for each patch. These probabilities were subsequently used to rank the features, and the top k were selected for the downstream prediction. Lastly, we compared our methods with a baseline where the slide-level Gleason grades were predicted, and the PRIAS protocol was used for making treatment decisions based on these predicted grades.

We found that using UNI as the feature extractor (OP-UNI), clearly outperformed the other methods, with an average AUC of 0.996 (SD: 0.004, 95% CI: $0.996-0.996$) against average AUCs 0.918 (SD: 0.053, 95% CI: $0.915-0.921$) and 0.758 (SD: 0.071, 95% CI: $0.754-0.762$) for OP-GGNet and OP-ImageNet respectively (see Table. 2). Furthermore, OP-UNI consistently achieved the highest accuracy and balanced accuracy (see Fig. 3). The protocol-based method gave a significantly lower average AUC of 0.675 (SD: 0.097, 95% CI: $0.669 - 0.681$) compared with OP-UNI and OP-GGNet. Interestingly, by studying the attention weights computed by the ABMIL framework, we found that for patients who received a treatment decision, the model attended on some occasions, more to patches from benign slides than from slides with higher Gleason grades. This finding is something we studied further in Paper III.

6.2.2 Paper III: Longitudinal outcome prediction of prostate cancer patients on active surveillance using multiple instance learning

While Paper II showed strong results for the immediate treatment prediction, we wanted to investigate whether this method could be extended to provide longitudinal predictions. Thus, we included visits up to 30 months prior to a treatment decision from the PRIAS dataset. We also included all visits for patients who remained on active surveillance, up until the final visit, due to right censoring. We used a similar approach as in Paper II, employing the updated version UNI-2 as our feature extractor. We compared our previous approach of using the GGNet predictions as a feature selector with a basic approach where the ABMIL module is provided with all patches from a visit. Furthermore, we explored two additional extensions to mitigate the tendency of ABMIL to concentrate attention on a few number of instances. One of the approaches introduced a loss term to maximize the entropy of the attention weights (AEM). The other method involved stochastic sampling of k instances (SKS). We evaluated the models, both on the test split of PRIAS and on an external cohort of prostate biopsies with undetected prostate cancer collected at the University Hospital of Umeå [21]. While this did not include prostate cancer patients on active surveillance, it consisted of WSIs originally classified as benign from visits with the binary labels; *Benign*: patient remained prostate cancer free for 8 years, *Undetected cancer*: exhibited any ISUP grade in a re-biopsy within 30 months. Hence, our models could be used to make predictions on this data.

The Basic model outperformed the other models on both the test set (AUROC: 0.958, 95% CI $0.957-0.959$) and the external dataset (AUROC: 0.699, 95% CI: $0.697-0.702$). We evaluated the statistical significance of these results with a pairwise 95% confidence interval of difference test. Here, only the difference between the Basic model and the GG and AEM models on the test set was significant. We selected operating points for our models based on the maximum Youden's J statistic, and subsequently evaluated the accuracy, sensitivity, and specificity. On the test dataset, the Basic model performed the best with an accuracy of 0.871 (95% CI: $0.868-0.873$) and with an even balance between sensitivity (0.835, 95% CI $0.831-0.840$) and specificity (0.905, 95% CI: $0.902-0.908$). However, on the external, the AEM model reached the highest accuracy (0.683, 95% CI: $0.681-0.685$), albeit with a low sensitivity of only 0.263 (95% CI: $0.260-0.267$).

Furthermore, to evaluate the longitudinal predictive power of our models, we fitted time-dependent Cox proportional hazards models with the predicted probability of treatment as a single co-variate. We defined the unit increase in the probability as a 0.1 increase to make the hazard ratio more interpretable. The Basic model gave the best fit, with a C-index of 0.824 and a hazards ratio of 2.32 (95% CI: $1.32-4.09$) (i.e. an increase in predicted probability of 0.1, by this model increases the risk of treatment by a factor of 2.32). The Basic and SKS models displayed significant effects of the predicted probabilities following

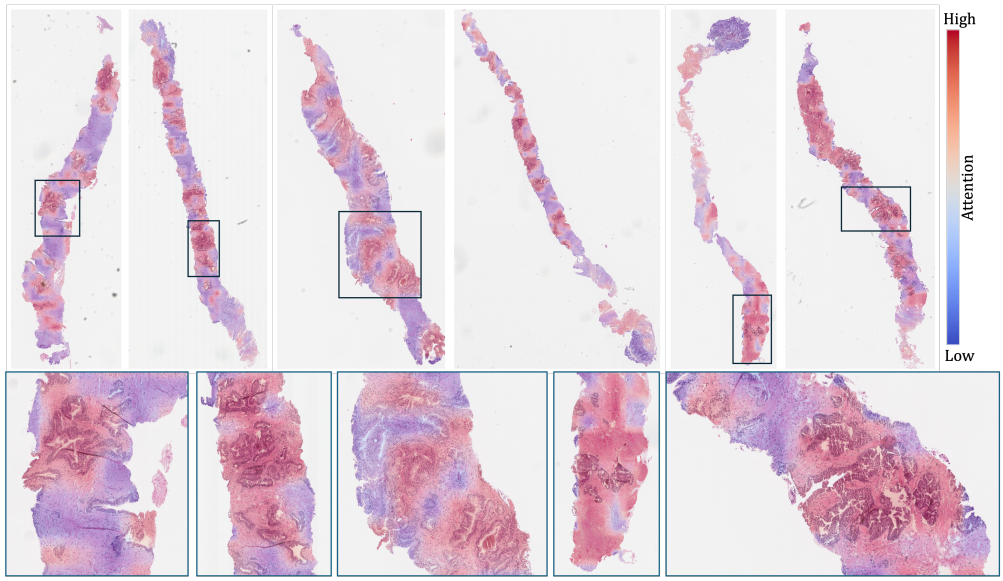


Figure 6.4: Heatmap of attention across one patient visit from the test dataset using the Basic model. Red color means higher attention weights, and blue color means less attention. For regions where patches overlap, the average attention is projected. This patient was correctly classified as Treatment.

the Wald's test at a 95% confidence level. We also fitted Cox regression models to the predicted probabilities on the external data. Since we here only had one prediction (and time point) per patient, we used a time-independent model. Analogously to the AUROC, the Basic model gave the highest C-index of 0.699. However, the GG model produced the highest hazard ratio of 1.45 (95% CI: $1.21 - 1.73$).

Our results suggest that ABMIL on all instances has an advantage over using traditional Gleason grades for longitudinal outcome predictions of prostate cancer patients. Coherent with our findings in Paper II and with other studies [77, 22], this indicates that the benign tissue does contain relevant prognostic information. The predictions made by the basic model are highlighted by the reconstructed attention heatmaps in Fig. 6.4.

6.3 Data-driven modeling of trajectories in Alzheimer's disease

Before presenting the contributions of Paper IV in relation to the aim of discovering AD subtypes from brain MRI images, we will discuss some of the difficulties that this problem entails. An MRI of the brain provides a high-resolution image of the intracranial tissues, which can be segmented volumetrically to get an estimated measure of the size of different regions. By collecting these measurements longitudinally for individuals with suspected or diagnosed AD, the neurodegeneration during the progression of the disease can be ob-

served. By studying these patterns of atrophy, in theory, different subtypes of AD could be identified that would explain the heterogeneous expression of the disease.

However, there are several caveats to consider. First of all, like any imaging modality, noise and artifacts can appear in the MRI scans, which makes the volumetric segmentation less accurate. Some of the most relevant regions make up only small fractions of the brain, for example, the hippocampus is roughly 0.5% of the intra-cranial volume. Hence, these measurements are sensitive to noise, and obtaining high quality data is important. Secondly, in any clinical study, participants will enter the study at different stages of the disease. This needs to be accounted for when modeling temporal relationships. One possible way of mitigating this issue is the use of data-driven disease progression modeling, where a true disease stage is estimated based on the collected clinical data. To date, there is little consensus on which subtypes exist, and given the heterogeneity of AD, there likely is considerable randomness to these trajectories. Furthermore, due to the complexity of the human brain, there are numerous factors and co-pathologies that can affect its function, many of which we have limited knowledge about.

However, in Paper IV, we aimed to uncover some of this heterogeneity in terms of longitudinal progression of atrophy patterns. For this, we utilized longitudinal MRI data from the ADNI database, which had been volumetrically segmented to provide measurements of regional brain volumes. We aimed to model the temporal relationship of neurodegeneration in AD and investigate any clusters of subtypes within this progression. We used an adaptation of the multivariate continuous-time disease progression (MCDP) model developed by Kühnel et al. [71] to stage study participants on a continuous disease timeline. This non-linear mixed effects model was fitted with two cognitive assessments and Aβ-PET measurements as outcomes. It included a random time-shift effect, which we used as an individual estimate of the onset of AD, and we subsequently applied this time shift to the trajectories of MRI data. The effect of this is illustrated in Fig. 6.5.

We developed an algorithm for data-driven clustering of these trajectories in an attempt to find subtypes. We based our approach on non-linear mixed effects models, similar to the MCDP model. However, with 95 different regions as outcomes, fitting one multivariate model was not feasible. Instead, we fitted one exponential mixed effects model independently for each variable, with latent time ($\hat{\mathbf{t}}_i$) as the only predictor. The model was formulated as

$$\mathbf{x}_{ik} = l_k e^{(g_k(1+\beta_{ik})\hat{\mathbf{t}}_{ik})} + v_k + v_{ik} + \epsilon_{ik} \tag{6.1}$$

where l_k, g_k and v_k denote population level fixed effects and v_{ik} and β_{ik} are subject specific random effects, and ϵ_{ik} is the residual error. A benefit of this model formulation is that the only parameter that controls subject-level deviations from the mean curve in terms of atrophy is the time-scaling random effect $(1 + \beta_{ik})$. Thus, we employed a simple K-means clustering approach of these random effects to find clusters of atrophy patterns. To mitigate issues with optimizing the k parameter (i.e. number of clusters) and to enable hierarchical

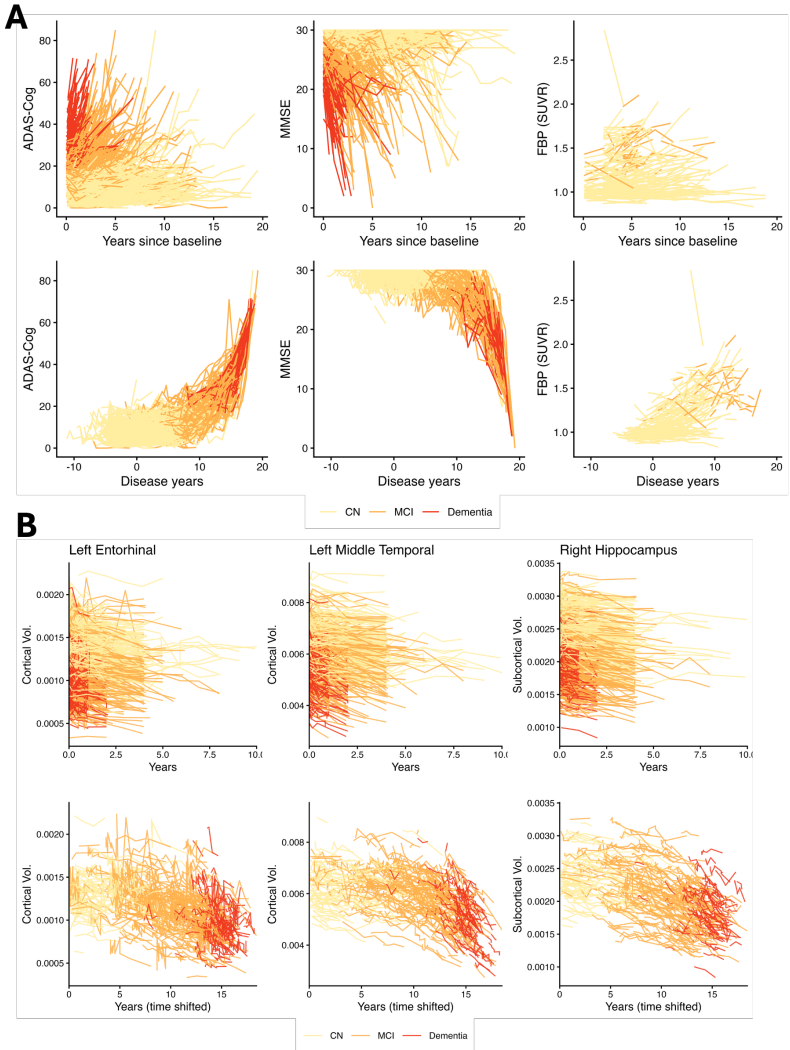


Figure 6.5: Illustration of the effect of the fitted MCDP disease progression model. **A** Longitudinal cognitive assessments of MMSE and ADAS-Cog-13 and Aβ-PET measurements were used to fit the model. The top shows the original time scale, and the bottom illustrates the transformed disease time scale. **B** Examples of MRI-derived volumetric measurements of the left entorhinal, left middle temporal, and right hippocampus. The effect of transforming longitudinal MRI data to the estimated disease timeline can be seen by comparing the top and bottom panels.

clusters, we formulated the algorithm as an iterative process, in which clusters are split into smaller ones. For each split, the mixed effects models were refitted with cluster-specific fixed effects. If the Bayesian information criterion (BIC) was significantly reduced, we considered the new cluster split as useful clusters and kept them. This iteration kept going until the BIC no longer improved. For more details on this algorithm, see Paper IV.

Our algorithm converged to four clusters of longitudinal atrophy patterns, which we labeled as; **A**: Late Atrophy, **B**: Left Temporal, **C**: Stable, and **D**: High Atrophy (see Fig. 6.6). By conducting 5-fold cross-validation, in a leave-one-out fashion, we found that the resulting clusters agreed with an average adjusted Rand index of 0.745 (SD: 0.086). To further validate the clusters, we trained an Extreme Gradient Boosting (XGBoost)[25] classifier on predicting the clusters based on the original MRI data. It was trained with 5-fold cross-validation, and achieved an average AUC of 0.774 (SD: 0.103) across all folds and clusters. Hence, this indicates some overlap between the clusters. The **A** (Late Atrophy) group remained relatively spared until the final stages of the disease, where we saw a rapid neurodegeneration. The **B** (Left Temporal) cluster resembled the **A** cluster, however, with a slight asymmetric atrophy in the temporal regions of the left hemisphere. The **D** (High Atrophy) exhibited higher amounts of atrophy in primarily the typical regions affected in AD, and the **C** (Stable) remained with minimal atrophy.

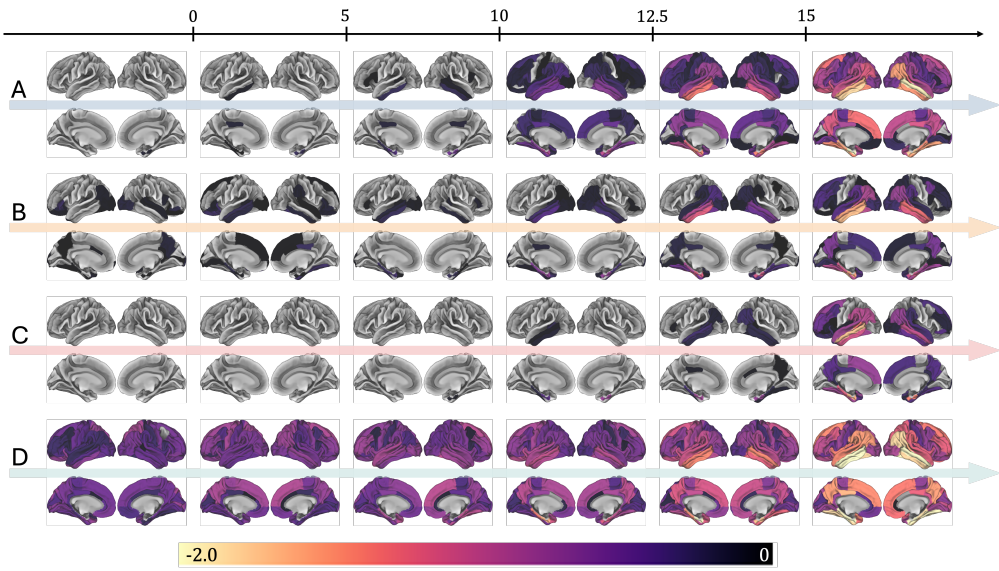


Figure 6.6: Visualization of atrophy patterns. The atrophy of each FreeSurfer region is represented by a z-score relative to the average volume, normalized by ICV, of CN and A$\beta-$ subjects. MRI scans are divided based on the estimated latent-disease time into the intervals: prior to disease onset, 0-5 years after onset, 5-10 years after onset, 10-12.5 years after onset, 12.5-15 years after onset, and after 15 years. For each interval and cluster, the average z-scores are used as the representation of atrophy in that region. If any subject had multiple measurements within one interval, we selected one based on the median latent time (i.e. the middle-most in the interval).

However, we found an apparent discrepancy between these results and the clinical and biological profiles of the found clusters. Cluster **C** (Stable) had larger amounts of baseline AD diagnoses and higher amyloid-β load, which should correspond with more atrophy. Additionally, the **D** (High Atrophy) group had less evidence of AD in terms of the measured biomarkers.

Chapter 7

Conclusions

This thesis presents machine learning-based methods for analyzing longitudinal medical data in prostate cancer and Alzheimer's disease. A major challenge when developing these methods is the lack of high-quality longitudinal medical datasets of sufficient size. Paper I addressed this by investigating methods for data augmentation for the task of semantic segmentation of H&E stained prostate biopsies [141]. This work highlighted how the augmentation methods can improve the performance of deep learning-based segmentation algorithms when the amount of training data is low. However, it is not clear how this would generalize to longitudinal data and other tasks like survival predictions. Here, the augmentation becomes more sensitive since there is a time-dependency to account for. Generating synthetic data that can be openly shared among researchers is another way to mitigate this issue. In paper V, we evaluated several methods of generating tabular data for Alzheimer's disease research. A trade-off between utility and privacy of the synthetic datasets was observed, and we found that none of the methods managed to produce synthetic data with privacy guarantees and sufficiently high utility. In this setting, we also restricted ourselves to non-longitudinal data, which is a limitation of the evaluated methods. To successfully generate meaningful synthetic data for research purposes, this is a requirement that needs to be fulfilled. Paper VI demonstrated the feasibility of dual-energy CT-based segmentation and volumetric estimation of intracranial tissues [45]. By using the VMIs, our best-performing model achieved significantly higher Dice scores and lower volumetric error compared with the baseline model. However, MRI remains the gold standard for brain tissue imaging, and the use of dual-energy CT is limited. Hence, the gains in terms of amounts of longitudinal brain imaging data following these findings are modest. To this end, while we in this thesis have presented many possible solutions, the challenge of gathering longitudinal medical data remains a bottleneck for the development of machine learning-based methods.

To leverage the capabilities of AI for prostate cancer diagnostics, we aimed to develop a model capable of predicting outcomes of patients on active surveillance. In papers II and III, we progressively built this model, which in the end achieved promising results for making longitudinal predictions on the PRIAS dataset [140, 139]. We showed that it is possible to predict the treatment decision based solely on prostate biopsies without using explicit Gleason grades. The use of the UNI foundation model gave a significant boost in performance, which demonstrated the power of these general-purpose feature extractors. We extended our framework to make longitudinal predictions on treatment decisions. An important finding in paper III was that including benign tissue improved performance over selecting regions based on Gleason grades. Previous studies have also found that it is possible to identify prognostic factors from benign prostate tissue [77, 22]. While we did achieve high performance in predicting the treatment decision up to 30 months prior on the PRIAS dataset, we also found a substantial drop in performance when tested on an external dataset. This highlights one of the biggest challenges in current computational pathology: the generalization to real clinical data [132]. In order to implement any AI-based solution clinically, robustness to differences in scanners, tissue staining, and medical protocols has to be demonstrated, since this is what the algorithm would encounter in the real world. Furthermore, this is not easy to evaluate given the limited access to open benchmarking datasets with longitudinal data.

We attempted to identify subtypes of Alzheimer's disease in Paper IV. We developed a data-driven clustering method based on disease progression modeling and non-linear mixed effects models. While the method, in theory should be able to identify subtypes of distinct atrophy patterns, in practice, we did not observe this. When applied to longitudinal data from the ADNI database, we found four clusters of atrophy patterns. However, they were not clearly separated and did not exhibit any distinct biological or clinical profiles. Hence, the question of whether biologically distinct subtypes of AD can be identified from longitudinal brain MRI remains unclear. It is worth noting that the number of participants in our dataset was limited. The strict inclusion criteria required for this study meant only 561 subjects could be used. For a complex and heterogeneous disease like AD, this is a far too small pool to draw any definitive conclusions, regardless of the results. The development of this method involved numerous failed attempts, which highlights the difficulty of this task. Finally, given that we did not manage to identify any distinct subtypes, the question of whether they could be used for predicting the progression of AD also remains unanswered.

7.1 Outlook

In the scientific contributions of this thesis, a common conclusion is that more data is needed to improve upon the results. There are many ethical concerns to consider when

collecting and sharing medical data, such as respecting patient privacy. Especially, with the rise of generative AI over recent years, the protection of sensitive data is increasingly important. Hence, overcoming these issues is not as easy as research institutions releasing their data. Federated learning is an approach for machine learning where only the models are shared between institutions, and the data is kept in-house. While this method has received considerable amounts of attention for many applications, it is often only used as a proof-of-concept. For example, in the field of AD research, there are no larger studies that employed federated learning to develop a method with any significant contribution to the field. Furthermore, there is a clear imbalance in ethnicity and socio-economic status of the patients whose data is used to develop AI, where the current models are not trained on data representative of the entire population.

One of the main challenges for implementing AI in clinical settings is the validation of its performance. Deviations between performance on in-house datasets and external datasets are common and indicate overfitting. In computational pathology, factors such as staining and scanner variations have been shown to have a big impact on the performance of AI-based models. In training, these issues can be mitigated by augmentation techniques such as color transformations. However, without diverse test datasets, this can hardly be validated. Furthermore, with constant technical innovations in both the hardware and software of the imaging systems, robustness to future upgrades will be important. Especially, given the black-box nature of deep learning-based algorithms, detecting a drift in performance can be difficult.

However, there is hope for the use of AI in modern healthcare. In computational pathology, there are numerous studies with promising results across many fields, including prostate cancer. With the analysis of longitudinal data, we can use the power of AI to truly improve our diagnostic procedures. For Alzheimer's disease, the use of machine learning has also contributed to numerous promising results and novel insights. In this thesis, the contributions have been focused on uni-modal settings, where we aimed to extract as much information as possible from one image modality (mainly pathology images or MRI). Nonetheless, for both prostate cancer and Alzheimer's disease, there are many more modalities of data to be leveraged. A clinician working in these fields would not restrict their diagnosis to only one of them. Hence, the future of AI in healthcare lies not only in the use of longitudinal data, but also in the inclusion of multi-modal data.

References

[1] Alan Agresti. *An introduction to categorical data analysis*. Second edition. John Wiley & Sons, 2007.

[2] American Cancer Society. *Tests to Diagnose and Stage Prostate Cancer*. `https://www.cancer.org/cancer/types/prostate-cancer/detection-diagnosis-staging/how-diagnosed`. Last revised March 21, 2025, accessed December 28, 2025. American Cancer Society, 2025.

[3] American Cancer Society. *Treating Prostate Cancer*. `https://www.cancer.org/cancer/types/prostate-cancer/treating`. Last revised November 22, 2023, accessed December 30, 2025. American Cancer Society, 2025.

[4] Ida Arvidsson. "Applications of Deep Learning in Medical Image Analysis: Grading of Prostate Cancer and Detection of Coronary Artery Disease". PhD thesis. Lund University, 2021.

[5] Ida Arvidsson et al. "Artificial intelligence for detection of prostate cancer in biopsies during active surveillance". In: *BJU international* (2024). DOI: `https://doi.org/10.1111/bju.16456`.

[6] Alzheimer's Association. "2024 Alzheimer's disease facts and figures". In: *Alzheimer's & dementia* 20.5 (2024), pp. 3708–3821.

[7] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[8] Sarah de Boer et al. "Robust kidney abnormality segmentation: a validation study of an AI-based framework". In: *arXiv preprint arXiv:2505.07573* (2025).

[9] O. Bratt and et al. "Population-based Organised Prostate Cancer Testing: Results from the First Invitation of 50-year-old Men." In: *European Urology* 85 (2024), pp. 207–214.

[10] Freddie Bray et al. "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: a cancer journal for clinicians* 74.3 (2024), pp. 229–263.

[11] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[12] M. Brudfors. “Generative Models for Preprocessing of Hospital Brain Scans”. PhD thesis. University College London, 2020.

[13] M. Bul et al. “Active surveillance for low-risk prostate cancer worldwide: the PRIAS study.” In: *Eur. Urol.* 63 (2013), pp. 597–603. DOI: {10.1016/j.eururo.2012.11.005}.

[14] Wouter Bulten et al. “Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study”. In: *The Lancet Oncology* 21.2 (2020), pp. 233–241.

[15] Claudine Burton-Jeangros et al. *A life course perspective on health trajectories and transitions*. Vol. 4. Springer, 2015, pp. 1–18. DOI: 10.1007/978-3-319-20484-0_1.

[16] Min Soo Byun et al. “Heterogeneity of Regional Brain Atrophy Patterns Associated with Distinct Progression Rates in Alzheimer’s Disease”. In: *PLOS ONE* 10.11 (Nov. 2015), pp. 1–16. DOI: 10.1371/journal.pone.0142756.

[17] G. Campanella et al. “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”. In: *Nat Med* 25 (2019), pp. 1301–1309. DOI: 10.1038/s41591-019-0508-1.

[18] Cancerfonden. *Statistik prostatacancer – dödlighet & överlevnad*. Accessed: 2025-12-19. Cancerfonden. 2025. URL: https://www.cancerfonden.se/om-cancer/statistik/prostatacancer.

[19] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[20] John KC Chan. “The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology”. In: *International journal of surgical pathology* 22.1 (2014), pp. 12–32.

[21] Eduard Chelebian et al. “A clinical prostate biopsy dataset with undetected cancer”. In: *Scientific Data* 12.1 (2025), p. 423.

[22] Eduard Chelebian et al. “Discovery of tumour indicating morphological changes in benign prostate biopsies through AI”. In: *medRxiv* (2024), pp. 2024–06.

[23] Guo-fang Chen et al. “Amyloid beta: structure, biology and structure-based therapeutic development”. In: *Acta pharmacologica sinica* 38.9 (2017), pp. 1205–1235.

[24] Richard J Chen et al. “Towards a General-Purpose Foundation Model for Computational Pathology”. In: *Nature Medicine* (2024).

[25] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *KDD '16*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[26] Felix K-H Chun et al. "Anatomic radical retropubic prostatectomy—long-term recurrence-free survival rates for localized prostate cancer". In: *World journal of urology* 24.3 (2006), pp. 273–280.

[27] J. Contador et al. "Longitudinal brain atrophy and CSF biomarkers in early-onset Alzheimer's disease." In: *Neuroimage Clin.* 32 (2021). DOI: 10.1016/j.nicl.2021.102804..

[28] Philip Cornford et al. "EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer—2024 update. Part I: screening, diagnosis, and local treatment with curative intent". In: *European urology* 86.2 (2024), pp. 148–163.

[29] C. DeCarli et al. "Anatomical mapping of white matter hyperintensities (WMH): exploring the relationships between periventricular WMH, deep WMH, and total WMH burden." In: *Stroke* (2005). DOI: 10.1161/01.STR.0000150668.58689.f2.

[30] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[31] Thomas G. Dietterich, Richard H. Lathrop and Tomas Lozano-Perez. "Solving the Multiple Instance Problem with Axis-Parallel Rectangles". In: *Artif. Intell.* 89 (1997), pp. 31–71.

[32] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[33] Bruno Dubois et al. "Preclinical Alzheimer's disease: definition, natural history, and diagnostic criteria". In: *Alzheimer's & Dementia* 12.3 (2016), pp. 292–323.

[34] Christopher H. van Dyck et al. "Lecanemab in Early Alzheimer's Disease". In: *New England Journal of Medicine* 388.1 (2023), pp. 9–21. DOI: 10.1056/NEJMoa2212948.

[35] L Egevad et al. "Standardization of Gleason grading among 337 European pathologists". In: *Hisopathology* 62.2 (2013), pp. 247–256. DOI: 10.1111/his.12008.

[36] Babak Ehteshami Bejnordi et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". In: *Jama* 318.22 (2017), pp. 2199–2210.

[37] Martin Eklund et al. "MRI-Targeted or Standard Biopsy in Prostate Cancer Screening". In: *New England Journal of Medicine* 385.10 (2021), pp. 908–920. DOI: 10.1056/NEJMoa2100852.

[38] Tobias Ekman, Arthur Barakat and Einar Heiberg. “Generalizable deep learning framework for 3D medical image segmentation using limited training data”. In: *3D printing in medicine* 11.1 (2025), p. 9.

[39] U. Ekman, D. Ferreira and E. Westman. “The A/T/N biomarker scheme and patterns of brain atrophy assessed in mild cognitive impairment.” In: *Sci Rep* 8 (2018). DOI: 10.1038/s41598-018-26151-8.

[40] Jonathan I Epstein. “Prostate cancer grading: a decade after the 2005 modified system”. In: *Modern Pathology* 31 (2018), pp. 47–63.

[41] Yosri A Fahim et al. “Artificial intelligence in healthcare and medicine: clinical applications, therapeutic advances, and future perspectives”. In: *European Journal of Medical Research* 30.1 (2025), p. 848.

[42] D. Ferreira et al. “Distinct subtypes of Alzheimer’s disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications.” In: *Sci Rep* 7 (2017). DOI: https://doi.org/10.1038/srep46263.

[43] N.M. Fonseca et al. “Prediction of plasma ctDNA fraction and prognostic implications of liquid biopsy in advanced prostate cancer”. In: *Nat Commun* 15 (2024). DOI: 10.1038/s41467-024-45475-w.

[44] M. Frånlund and et al. “Results from 22 years of Followup in the Göteborg Randomized Population-Based Prostate Cancer Screening Trial.” In: *J Urol.* 208 (2022), pp. 292–300.

[45] Veronica Fransson et al. “Dual energy CT and deep learning for an automated volumetric segmentation of the major intracranial tissues: Feasibility and initial findings”. In: *Medical Physics* 53.1 (2026), e70217.

[46] Stephen J Freedland et al. “Risk of prostate cancer–specific mortality following biochemical recurrence after radical prostatectomy”. In: *Jama* 294.4 (2005), pp. 433–439.

[47] Michael Gadermayr and Maximilian Tschuchnig. “Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential”. In: *Computerized Medical Imaging and Graphics* 112 (2024), p. 102337.

[48] A. Giorgio and N. De Stefano. “Clinical use of brain volumetry”. In: *J Magn Reson Imaging* 37.1 (2013), pp. 1–14.

[49] Stephen N Gomperts. “Lewy body dementias: dementia with Lewy bodies and Parkinson disease dementia”. In: *Continuum: Lifelong Learning in Neurology* 22.2 (2016), pp. 435–463.

[50] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[51] Jonathan Graff-Radford et al. “New insights into atypical Alzheimer’s disease in the era of biomarkers”. In: *The Lancet Neurology* 20.3 (2021), pp. 222–234.

[52] Clément Grisi et al. “Deep learning from routine histology improves risk stratification for biochemical recurrence in prostate cancer”. In: *arXiv preprint arXiv:2603.14187* (2026).

[53] O. Hansson et al. “Blood biomarkers for Alzheimer’s disease in clinical practice and trials.” In: *Nat Aging* 3 (5 2023), pp. 506–519. DOI: 10.1038/s43587-023-00403-3.

[54] Intisar Rizwan I Haque and Jeremiah Neubert. “Deep learning approaches to biomedical image segmentation”. In: *Informatics in Medicine Unlocked* 18 (2020), p. 100297. DOI: https://doi.org/10.1016/j.imu.2020.100297.

[55] Emily Harris. “Prostate Cancer Cases Might Rise to 3 Million Globally by 2040”. In: *JAMA* 331.20 (May 2024), pp. 1698–1698. ISSN: 0098-7484. DOI: 10.1001/jama.2024.6729.

[56] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 770–778.

[57] Noah Hollmann et al. “Accurate predictions on small data with a tabular foundation model”. en. In: *Nature* 637.8045 (2025), pp. 319–326. DOI: 10.1038/s41586-024-08328-6.

[58] J. Hugosson and et al. “A 16-yr Follow-up of the European Randomized study of Screening for Prostate Cancer.” In: *European Urology* 76 (2019), pp. 43–51.

[59] Peter A Humphrey. “Histological variants of prostatic carcinoma and their significance”. In: *Histopathology* 60.1 (2012), pp. 59–74.

[60] Dragan Ilic et al. “Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis”. In: *bmj* 362 (2018).

[61] Maximilian Ilse, Jakub M. Tomczak and Max Welling. “Attention-based Deep Multiple Instance Learning”. In: *CoRR* (2018). URL: http://arxiv.org/abs/1802.04712.

[62] Johan Isaksson et al. “Semantic segmentation of microscopic images of H&E stained prostatic tissue using CNN”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. 2017, pp. 1252–1256. DOI: 10.1109/IJCNN.2017.7965996.

[63] Fabian Isensee et al. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature methods* 18.2 (2021), pp. 203–211.

[64] C. R. Jack et al. “Tracking pathophysiological processes in Alzheimer’s disease: an updated hypothetical model of dynamic biomarkers.” In: *Lancet Neurol.* 12 (2013). DOI: 10.1016/S1474-4422(12)70291-0.

[65] Clifford R. Jack et al. “NIA-AA Research Framework: Toward a biological definition of Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 14.4 (2018), pp. 535–562. ISSN: 1552-5260. DOI: 10.1016/j.jalz.2018.02.018.

[66] Clifford R Jack Jr et al. “Revised criteria for diagnosis and staging of Alzheimer’s disease: Alzheimer’s Association Workgroup”. In: *Alzheimer’s & Dementia* 20.8 (2024), pp. 5143–5169.

[67] Nicholas D James et al. “The Lancet Commission on prostate cancer: planning for the surge in cases”. In: *The Lancet* 403.10437 (2024), pp. 1683–1722.

[68] Ameet V. Joshi. “Support Vector Machines”. In: *Machine Learning and Artificial Intelligence*. Cham: Springer International Publishing, 2023, pp. 89–99. ISBN: 978-3-031-12282-8. DOI: 10.1007/978-3-031-12282-8_8.

[69] Diederik P Kingma. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[70] David G. Kleinbaum and Mitchel Klein. *Survival Analysis: A Self-Learning Text*. Third edition. Springer: Statistics for Biology and Health, 2012. ISBN: 978-1-4419-6645-2. DOI: 10.1007/978-1-4419-6646-2.

[71] Line Kühnel et al. “Simultaneous modeling of Alzheimer’s disease progression via multiple cognitive scales”. In: *Statistics in Medicine* 40 (2021), pp. 3251–3266. DOI: https://doi.org/10.1002/sim.8932.

[72] Sharon Lam et al. “White matter hyperintensities and cognition across different Alzheimer’s biomarker profiles”. In: *Journal of the American Geriatrics Society* 69.7 (2021), pp. 1906–1915. ISSN: 1532-5415. DOI: 10.1111/jgs.17173.

[73] A. D. Lautrup et al. “SynthEval: A Framework for Detailed Utility and Privacy Evaluation of Tabular Synthetic Data”. In: *arXiv* (2024). DOI: 10.48550/arXiv.2404.15821.

[74] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.

[75] Sylvain E Lesne et al. “Brain amyloid-β oligomers in ageing and Alzheimer’s disease”. In: *Brain* 136.5 (2013), pp. 1383–1398.

[76] Giuseppe Lippolis. “Image analysis of prostate cancer tissue biomarkers”. PhD thesis. Lund University, 2015.

[77] B. Liu et al. “Using deep learning to detect patients at risk for prostate cancer despite benign biopsies.” In: *iScience* 25(7) (2022). DOI: 10.1016/j.isci.2022.104663.

[78] M.Y. Lu et al. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nat Biomed Eng* 5 (2021), pp. 555–570.

[79] Ilaria Lucca et al. “Validation of tertiary Gleason pattern 5 in Gleason score 7 prostate cancer as an independent predictor of biochemical recurrence and development of a prognostic model”. In: *Urologic Oncology: Seminars and Original Investigations*. Vol. 33. 2. Elsevier. 2015, 71–e21.

[80] Brandon A Mahal et al. “Gleason score 5+ 3= 8 prostate cancer: much more like Gleason score 9?” In: *BJU international* 118.1 (2016), pp. 95–101.

[81] C Mallinckrodt et al. “Investigating partially discordant results in phase 3 studies of aducanumab”. In: *The Journal of Prevention of Alzheimer’s Disease* 10.2 (2023), pp. 171–177.

[82] Felicia Marginean et al. “An Artificial Intelligence–based Support Tool for Automation and Standardisation of Gleason Grading in Prostate Biopsies”. In: *European Urology Focus* 7.5 (2021), pp. 995–1001. ISSN: 2405-4569. DOI: 10.1016/j.euf.2020.11.001.

[83] Soeren Mattke et al. “Expected and diagnosed rates of mild cognitive impairment and dementia in the US Medicare population: observational analysis”. In: *Alzheimer’s Research & Therapy* 15.1 (2023), p. 128.

[84] Niklas Mattsson-Carlgren et al. “Aβ deposition is associated with increases in soluble and phosphorylated tau that precede a positive Tau PET in Alzheimer’s disease”. In: *Science advances* 6.16 (2020), eaaz2387.

[85] Lauren-Jei McCarthy. *FDA Authorizes Software that Can Help Identify Prostate Cancer*. FDA News Release, Online, 21 September 2021 https://www.fda.gov/news-events/press-announcements/fda-authorizes-software-can-help-identify-prostate-cancer. (Accessed: 3 July 2024).

[86] C. H. McCollough et al. “Dual- and Multi-Energy CT: Principles, Technical Approaches, and Clinical Applications”. In: *Radiology* 276.3 (2015), pp. 637–653.

[87] Luigi Federico Menabrea and Ada Lovelace. “Sketch of the analytical engine invented by Charles Babbage”. In: *Sci Mem* 3 (1843), pp. 666–731.

[88] James L. Mohler and et. al. “Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology”. In: *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw* 17.5 (2019), pp. 479–505. DOI: 10.6004/jnccn.2019.0023.

[89] Y. Mun et al. “Yet Another Automated Gleason Grading System (YAAGGS) by weakly supervised deep learning”. In: *NPJ Digit. Med.* 4 (2021).

[90] Melissa E Murray et al. “Neuropathologically defined subtypes of Alzheimer’s disease with distinct clinical characteristics: a retrospective study”. In: *The Lancet Neurology* 10.9 (2011), pp. 785–796. ISSN: 1474-4422. DOI: https://doi.org/10.1016/S1474-4422(11)70156-9.

[91] Kunal Nagpal et al. “Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens”. In: *JAMA oncology* 6.9 (2020), pp. 1372–1380.

[92] Beata Nowok, Gillian M Raab and Chris Dibben. “synthpop: Bespoke creation of synthetic data in R”. In: *Journal of statistical software* 74 (2016), pp. 1–26.

[93] Maxime Oquab et al. “Dinov2: Learning robust visual features without supervision”. In: *arXiv preprint arXiv:2304.07193* (2023).

[94] Nobuyuki Otsu. “A Threshold Selection Method from Gray-Level Histograms”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.

[95] Sebastian Palmqvist et al. “Blood biomarkers to detect Alzheimer disease in primary care and secondary care”. In: *Jama* 332.15 (2024), pp. 1245–1257.

[96] Charles TA Parker et al. “External validation of a digital pathology-based multimodal artificial intelligence-derived prognostic model in patients with advanced prostate cancer starting long-term androgen deprivation therapy: a post-hoc ancillary biomarker study of four phase 3 randomised controlled trials of the STAMPEDE platform protocol”. In: *The Lancet Digital Health* 7.7 (2025).

[97] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[98] Ivar Persson. “Efficient Monocular 3D Localisation Using Machine Learning: with Additional Studies on Pose Estiamtion and Shape Reconstruction”. PhD thesis. Lund University, 2026.

[99] Hans Pinckaers et al. “Predicting biochemical recurrence of prostate cancer with artificial intelligence”. In: *Communications Medicine* 2.1 (2022), p. 64.

[100] Haoyue Ping, Julia Stoyanovich and Bill Howe. “DataSynthesizer: Privacy-Preserving Synthetic Datasets”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. SSDBM ’17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 1–5. ISBN: 978-1-4503-5282-6. DOI: 10.1145/3085504.3091117. (Visited on 25/10/2024).

[101] José Pinheiro, Douglas Bates and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-168. 2025. DOI: 10.32614/CRAN.package.nlme. URL: https://CRAN.R-project.org/package=nlme.

[102] PRIAS website. https://prias-project.org/. 2024.

[103] Martin Prince et al. “World Alzheimer report 2015. The global impact of dementia: an analysis of prevalence, incidence, cost and trends.” In: *Alzheimer’s Disease International* (2015).

[104] Gil D Rabinovici et al. “Multiple comorbid neuropathologies in the setting of Alzheimer’s disease neuropathology and implications for drug development”. In: *Alzheimer’s & dementia: translational research & clinical interventions* 3.1 (2017), pp. 83–91.

[105] Lars Lau Raket et al. “Estimating the time course of biomarker changes in Alzheimer’s disease”. In: *Brain* (2025).

[106] John O Rawlings, Sastry G Pantula and David A Dickey. *Applied regression analysis: a research tool*. Springer, 1998.

[107] Olaf Ronneberger, Philipp Fischer and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[108] Marcelo R Roxo et al. “The limbic system conception and its historical evolution”. In: *The scientific world journal* 11.1 (2011), pp. 2427–2440.

[109] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[110] K Santacruz et al. “Tau suppression in a neurodegenerative mouse model improves memory function”. In: *Science* 309.5733 (2005), pp. 476–481.

[111] F.H. Schröder et al. “Screening and Prostate-Cancer Mortality in a Randomized European Study”. In: *New England Journal of Medicine* 360.13 (2009), pp. 1320–1328. DOI: 10.1056/NEJMoa0810084.

[112] Smadar Shilo, Hagai Rossman and Eran Segal. “Axes of a revolution: challenges and promises of big data in healthcare”. In: *Nature medicine* 26.1 (2020), pp. 29–38.

[113] Christina Silcox et al. “The potential for artificial intelligence to transform healthcare: perspectives from international health leaders.” In: *npj Digit. Med.* 7.88 (2024).

[114] John R. Sims et al. “Donanemab in Early Symptomatic Alzheimer Disease: The TRAILBLAZER-ALZ 2 Randomized Clinical Trial”. In: *JAMA* 330.6 (2023), pp. 512–527. ISSN: 0098-7484. DOI: 10.1001/jama.2023.13239.

[115] Andrew H Song et al. “Artificial intelligence for digital and computational pathology”. In: *Nature Reviews Bioengineering* 1.12 (2023), pp. 930–949.

[116] John R. Srigley et al. “One is the new six: The International Society of Urological Pathology (ISUP) patient-focused approach to Gleason grading”. In: *Canadian Urological Association Journal* 10.9-10 (Oct. 2016), pp. 339–41. URL: 10.5489/cuaj.4146.

[117] Chetan L Srinidhi, Ozan Ciga and Anne L Martel. “Deep neural network models for computational histopathology: A survey”. In: *Medical image analysis* 67 (2021), p. 101813.

[118] Armando Stabile et al. “Multiparametric MRI for prostate cancer diagnosis: current status and future directions”. In: *Nature reviews urology* 17.1 (2020), pp. 41–61.

[119] Harold Evelyn Taitt. “Global trends and prostate cancer: a review of incidence, detection, and mortality as influenced by race, ethnicity, and geographic location”. In: *American journal of men’s health* 12.6 (2018), pp. 1807–1823.

[120] Stefan Teipel, Michel J Grothe, Alzheimer’s Disease Neuroimaging Initiative et al. “MRI-based basal forebrain atrophy and volumetric signatures associated with limbic TDP-43 compared to Alzheimer’s disease pathology”. In: *Neurobiology of disease* 180 (2023), p. 106070.

[121] Stefan J Teipel, Michel J Grothe and Alzheimer’s Disease Neuroimaging Initiative. “Antemortem basal forebrain atrophy in pure limbic TAR DNA-binding protein 43 pathology compared with pure Alzheimer pathology”. In: *European Journal of Neurology* 29.5 (2022), pp. 1394–1401.

[122] Frederik B Thomsen et al. “Active surveillance for clinically localized prostate cancer—-A systematic review”. In: *Journal of surgical oncology* 109.8 (2014), pp. 830–835.

[123] Pontus Tideman et al. “Primary care detection of Alzheimer’s disease using a self-administered digital cognitive test and blood biomarkers”. In: *Nature Medicine* (2025), pp. 1–9.

[124] Yuri Tolkach et al. “High-accuracy prostate cancer pathology using deep learning”. In: *Nature Machine Intelligence* 2.7 (2020), pp. 411–418.

[125] Jeffrey J Tosoian et al. “Intermediate and longer-term outcomes from a prospective active-surveillance program for favorable-risk prostate cancer”. In: *Journal of Clinical Oncology* 33.30 (2015), pp. 3379–3385.

[126] Duygu Tosun et al. “A cross-sectional study of α-synuclein seed amplification assay in Alzheimer’s disease neuroimaging initiative: Prevalence and associations with Alzheimer’s disease biomarkers and cognitive function”. In: *Alzheimer’s & Dementia* 20.8 (2024), pp. 5114–5131.

[127] Duygu Tosun et al. “Identifying individuals with non-Alzheimer’s disease co-pathologies: a precision medicine approach to clinical trials in sporadic Alzheimer’s disease”. In: *Alzheimer’s & Dementia* 20.1 (2024), pp. 421–436.

[128] Jeanne Truett, Jerome Cornfield and William Kannel. “A multivariate analysis of the risk of coronary heart disease in Framingham”. In: *Journal of Chronic Diseases* 20.7 (1967), pp. 511–524. ISSN: 0021-9681. DOI: https://doi.org/10.1016/0021-9681(67)90082-3.

[129] B. Trujillo et al. “Blood-based liquid biopsies for prostate cancer: clinical opportunities and challenges.” In: *Br J Cancer* 127 (2022), pp. 1394–1402. DOI: 10.1038/s41416-022-01881-9.

[130] Jos WR Twisk. *Applied mixed model analysis: a practical guide*. Cambridge University Press, 2019.

[131] Urology Times. *FDA grants de novo authorization to ArteraAI prostate test*. Accessed: 2026-05-05. 2025. URL: https://www.urologytimes.com/view/fda-grants-de-novo-authorization-to-arteraai-prostate.

[132] Jeroen Van der Laak, Geert Litjens and Francesco Ciompi. "Deep learning in histopathology: the path to the clinic". In: *Nature medicine* 27.5 (2021), pp. 775–784.

[133] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[134] J.W. Vogel et al. "Four distinct trajectories of tau deposition identified in Alzheimer's disease." In: *Nat Med* 27 (2021), pp. 871–881. DOI: 10.1038/s41591-021-01309-6.

[135] Jacob W Vogel and Oskar Hansson. "Subtypes of Alzheimer's disease: questions, controversy, and meaning". In: *Trends in neurosciences* 45.5 (2022), pp. 342–345.

[136] Xiyue Wang et al. "Transformer-based unsupervised contrastive learning for histopathological image classification". In: *Medical image analysis* 81 (2022), p. 102559.

[137] Yipeng Wang and Eckhard Mandelkow. "Tau in physiology and pathology". In: *Nature reviews neuroscience* 17.1 (2016), pp. 22–35.

[138] Eric Widmaier, Hershel Raff and Kevin Strang. *Vander's Human Physiology: The Mechanics of Body Function*. Fourteenth edition. McGraw-Hill Education, 2016.

[139] Filip Winzell et al. "Longitudinal outcome prediction of prostate cancer patients on active surveillance using multiple instance learning". In: *Journal of Medical Imaging* 12.6 (2025), pp. 061408–061408.

[140] Filip Winzell et al. "Outcome prediction of prostate cancer patients on active surveillance using weakly supervised deep learning". In: *Medical Imaging 2025: Digital and Computational Pathology*. Vol. 13413. SPIE. 2025, pp. 116–126.

[141] Filip Winzell et al. "Systematic Augmentation in HSV Space for Semantic Segmentation of Prostate Biopsies". In: *Scandinavian Conference on Image Analysis*. Springer Nature Switzerland, 2023, pp. 293–308.

[142] E. Wulczyn et al. "Predicting prostate cancer specific-mortality with artificial intelligence-based Gleason grading". In: *Commun Med 1* 10 (2021). DOI: https://doi.org/10.1038/s43856-021-00005-3.

[143] Lei Xu et al. "Modeling tabular data using conditional gan". In: *Advances in neural information processing systems* 32 (2019).

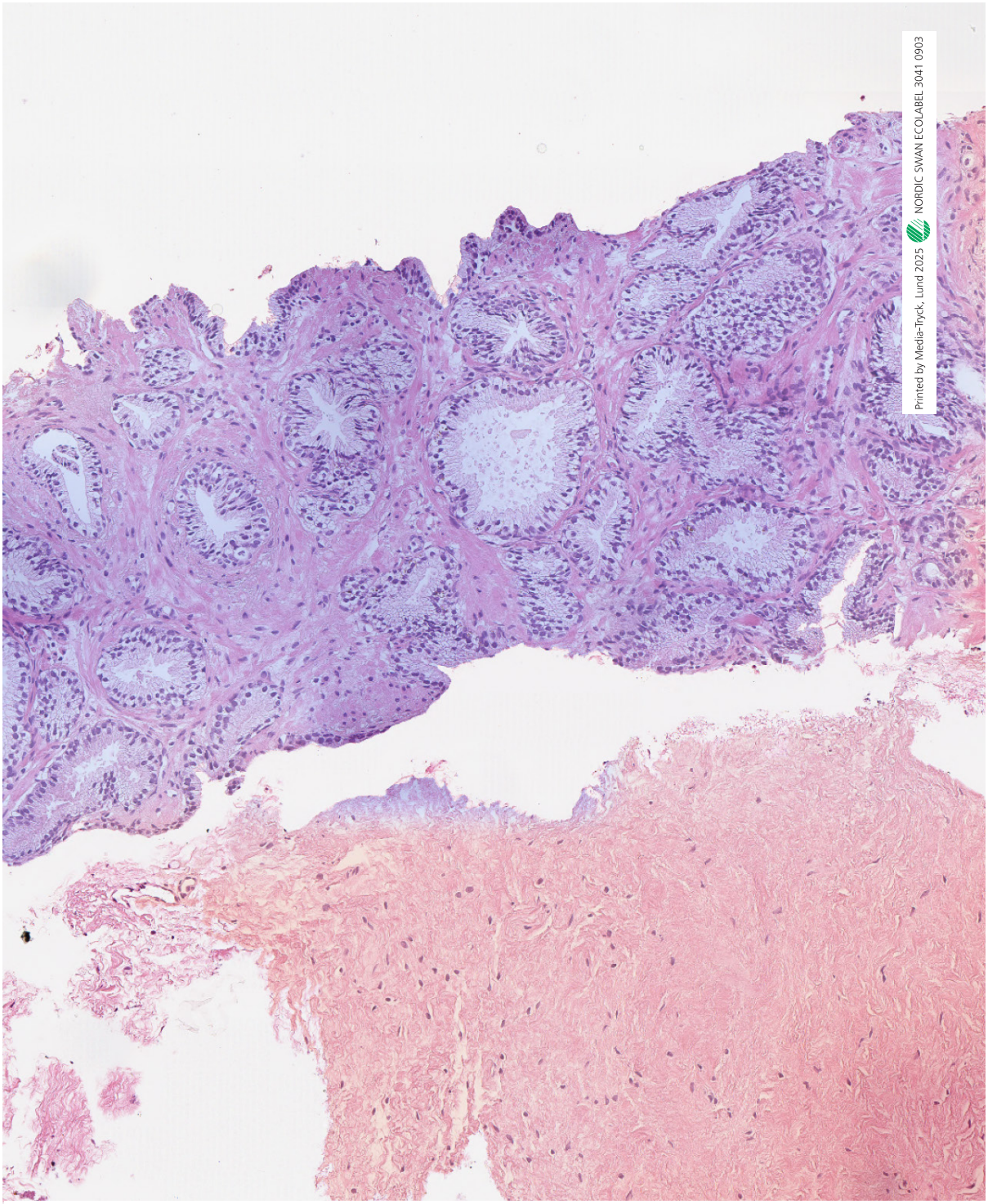
[144] Zhongyi Yang et al. “The devil is in the details: a small-lesion sensitive weakly supervised learning framework for prostate cancer detection and grading”. In: *Virchows Archiv* 482.3 (2023), pp. 525–538.

[145] Alexandra L Young et al. “Data-driven modelling of neurodegenerative disease progression: thinking outside the black box”. In: *Nature Reviews Neuroscience* 25.2 (2024), pp. 111–130.

[146] Jinghao Zhou et al. “ibot: Image bert pre-training with online tokenizer”. In: *arXiv preprint arXiv:2111.07832* (2021).

[147] Z. Zhou et al. “UNet++: A Nested U-Net Architecture for Medical Image Segmentation”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer, 2018.

[148] Eric Zimmermann et al. “Virchow2: Scaling self-supervised mixed magnification models in pathology”. In: *arXiv preprint arXiv:2408.00738* (2024).



Printed by Media-Tryck, Lund 2025 NORDIC SWAN ECOLABEL 3041 0903

LUND UNIVERSITY

Doctoral Theses in Mathematical Sciences 2026:8
ISBN 978-91-90202-04-3
ISSN 1404-0034