Deep Learning-Based Prostate Cancer Risk Prediction on Electronic Health Records and Histopathology Images

Dissertation (monograph) approved by the
Doctoral Degree Committee of
Hamburg University of Technology

in pursuit of the academic degree of

Doktor-Ingenieur(in) (Dr.-Ing.)

written by
Patrick Fuhlert

from
Dortmund

2025

ORCID https://orcid.org/0000-0001-8480-3705
DOI https://doi.org/10.15480/882.14817

Reviewers
Prof. Dr.-Ing. Alexander Schlaefer
Prof. Dr. Stefan Bonn

Date of oral examination
05.02.2025

Abstract

For the selection of optimal patient treatment, survival prediction methods that estimate the expected time to an event of interest can be utilized. Among other applications, it can be used to estimate the influence of covariates on individual survival to develop clinical decision support systems in the context of prostate cancer diagnosis and treatment. Prostate Cancer is among the most common cancers in men with the fifth-highest number of deaths. It is usually slowly growing and often remains undetected at early stages unless screening is performed. Suspicious screening results can lead to diagnostic procedures where a biopsy is taken. The current gold standard in risk assessment for biopsies is done by pathologists that evaluate histopathological images in terms of cancer severity using Gleason grading. Other factors like the blood level of the Prostate-Specific Antigen or the patient's age are also known to be informative regarding patient survival. These factors can be used in survival prediction models to estimate the patient's survival and decide on initial treatment plans. Commonly, low-risk patients are treated with active surveillance that avoids or delays invasive treatment while individuals with a higher risk are treated with more invasive procedures like radical prostatectomy. Adding additional parameters for those patients like the positive resection margin, seminal invasion or others can increase predictive accuracy that can be used for follow-up care of the individual.

In this thesis, alternatives to classical survival estimation using approaches like the Cox Proportional Hazards model are developed and investigated for multiple settings. Specifically, modern methods that use deep learning techniques are analyzed and compared to the classical approaches resulting in the development of a standalone survival prediction network called Discrete Calibrated Survival. This deep learning-based approach enhances the performance regarding relapse prediction by introducing two novel ideas, namely variable temporal output node spacing and an extended loss term that optimizes the use of censored patient data, making the model suitable for datasets with high censoring rates as for example often found in datasets involving prostate cancer. The developed model is further investigated on a high quality dataset of prostate cancer patients that underwent radical prostatectomy. The developed approach can further stratify the patients into risk groups that separate well in terms of patient relapse risk.

Moreover, one of the known most informative factors for prostate cancer relapse, namely the Gleason grading itself, is further investigated. Since it is known that this grading is highly subjective and suffers from high inter-observer variability, recent work led to the realization of artificial intelligence-based automatic grading systems. However, most of these systems focus on emulation of Gleason grading and thus are impacted by the underlying human judgement. This work takes a different approach and instead utilizes the objective survival information for the training process to create a more unbiased risk assessment that does not rely on human annotations. Since it is also known that these models perform poorly on external data, creating a risk estimation that translates beyond images that are included in the training setup is explicitly included in the development process by integrating a total of eleven datasets that either show tissue microarrays or biopsy images. Six of those datasets are used for training and show large variations in the sample acquisition protocol and thus vary in slide thickness, staining protocol or scanner vendor allowing in-depth evaluation and optimization of model robustness to data variation that lead to a risk prediction model called Prostate Cancer Aggressiveness Index. This work shows that the optimization of model robustness has a positive influence overall performance and extends to biopsy images. To find the relevant parts that show cancerous regions of biopsy images, another deep learning-based algorithm called Cancer Indicator is developed. It is shown that the overall pipeline is able to outperform annotations made by human experts for previously unseen datasets. The developed concepts for survival prediction and automated risk assessment can be adapted beyond digital histopathology and prostate cancer applications.

Acknowledgments

I'm extremely grateful to Prof. Dr. Stefan Bonn for granting me the opportunity to pursue my PhD at the Institute of Medical Systems Bioinformatics at the University Medical Center Hamburg-Eppendorf, for his overall supervision, and for serving as a reviewer of my thesis. I also sincerely appreciate the additional supervision and thesis review by Prof. Dr.-Ing. Alexander Schlaefer from the Institute of Medical Technology and Intelligent Systems at the Technical University of Hamburg. Furthermore, I would like to acknowledge Prof. Dr. Moritz Göldner for serving as the chair of my Doctoral Degree Committee, completing the panel overseeing my dissertation.

I extend my gratitude to my supervisors, Dr. Anne Ernst and Prof. Dr. Marina Zimmermann, for their invaluable support and guidance throughout the years. I am also deeply appreciative of my former colleagues, Dr. Fabian Westhaeusser and Dr. Esther Dietrich, for our close collaborations that contributed to this work. Additionally, I would like to acknowledge all members of the Institute of Medical Systems Bioinformatics for their support and the insightful discussions that enriched this journey.

Lastly, I would like to express my deepest gratitude to my wife, Jana, my sister, Alina, and her family, my parents, and my friend Alex for their emotional support and belief in me.

Contents

List of Abbreviations

3DH 3DHistech 50, 114, 115

AE Auto-Encoder 30, 31

AI Artificial Intelligence 1, 2, 6, 31, 34, 90, 92, 109, 111, 139, 143

APE Aperio 50, 114, 115

AS Active Surveillance 1, 14, 138, 167

AUC Area under the Curve 24, 31, 33

AUPRC Area under the Precision-Recall Curve 22, 100, 109

AUROC Area Under the Receiver Operating Characteristic Curve 22–24, 31, 34, 89, 100, 101, 109, 113, 122, 123, 125–127, 135, 136, 174, 177, 178

BASE Baseline PCAI model without robustness extensions 111, 113, 116, 123, 125–127, 135, 136, 177, 178

BCE Binary Cross-Entropy 89, 96, 120

BCR Biochemical Recurrence 14, 38–40, 69–71, 75, 76, 84, 87, 88, 113, 176

BrS Brier Score 23, 25, 58

C-index Concordance Index 19, 23–25, 87, 123, 125–127, 134–136, 139, 177, 178

C-index-td Time-Dependent Concordance Index 23–25, 29, 61–63, 66, 77, 79, 81–83, 87

CA Color Adaptation 119, 121, 122, 126, 135, 136

CDAUC Cumulative-Dynamic AUROC 23–25, 61–64, 66, 74, 75, 77, 79–82, 87, 170

CDSS Clinical Decision Support System 1, 2, 5, 142

CE Credibility Estimation 119–121, 126, 135, 136

CI Cancer Indicator Model 6, 7, 31, 48, 82, 89, 90, 96, 97, 100, 101, 103, 107, 109–111, 115, 117, 122, 126, 127, 135, 136, 138, 139, 141, 142, 173, 174, 177

CNN Convolutional Neural Network 31, 32, 34, 96, 109, 111, 112, 117, 137, 138

CoxPH Cox Proportional Hazards Model 4, 7, 16, 28, 29, 53, 61–63, 66, 67, 69, 72, 75–77, 80–82, 84, 87, 88, 134, 135, 137, 139, 140, 170

CoxTime CoxTime Model 19, 29, 53, 54, 60–63, 66, 168, 169

CV Cross-Validation 75, 76

DA Domain Adversarial 32, 119, 122, 126, 135–137, 142

DCS Discrete Calibrated Survival Model 3, 5–7, 19, 35, 53–55, 58–64, 66, 67, 69–71, 75–77, 80–85, 87, 88, 111, 134–137, 139–142, 168–170, 172

DDC Distributional Divergence for Calibration 19, 23, 25, 61, 63, 64, 66, 74, 75, 82, 87, 170

DeepSurv DeepSurv Model 19, 29, 53, 61–63, 66, 168, 169

DL Deep Learning 1–7, 9, 16, 19, 26, 29–35, 53, 54, 66, 67, 69, 96, 109, 111–113, 123, 135, 139–141, 143

DRE Digital rectal exam 2, 10, 11, 44

DRSA Discrete Recurrent Survival Analysis Model 19, 29, 30, 53, 54, 61–63, 66, 169

EC Event-to-censored Comparisons 53, 58–60, 66, 76, 87, 139

EE Event-to-event Comparisons 53, 58–60, 76, 87, 139

EHR Electronic Health Record 1–7, 9, 16, 18, 26, 30, 31, 35–37, 39, 48, 53, 55, 67, 133, 134, 139, 140, 142, 167

FLC Serum free light chain 36

FLCHAIN Assay of serum free light chain 6, 35, 36, 53, 55, 62–66, 71, 169, 170

FPR False Positive Rate 22

FU Follow-up 38–40, 70, 71, 113, 115, 135, 137, 138, 176

GG Gleason Grade 2, 7, 9, 12–14, 33, 34, 38–41, 45, 46, 69, 71, 76, 79, 80, 82–85, 87–91, 98, 110, 112, 113, 135, 139, 140

GIQ Integrated Quantitative Gleason Score 13, 41, 79, 80, 114, 127, 135, 136, 139, 141, 177

GT Ground Truth 21, 22, 33, 45, 89, 90, 97, 100, 103, 109

H&E Hematoxylin and Eosin staining 2, 6, 11, 32, 33, 39, 40, 89, 109, 167

HAM Hamamatsu 50, 107, 115

HSV Hue, Saturation, Value 32, 48, 116, 121, 137

HT Hormone therapy 1, 14, 39

IBrS Integrated Brier Score 23, 25, 29, 82, 87

IQR Interquartile range 39, 76

ISUP Gleason-based grading system defined by the **I**nternational **S**ociety of **U**rological **P**athology 13, 14, 40, 45, 46, 79, 80, 89, 107, 109, 111–116, 123, 127, 133, 135, 136, 138–141, 176–178

JHU Johns Hopkins University in Baltimore, USA 6, 39, 43, 48, 50, 112, 114, 115, 123, 125, 130, 135, 137, 176, 177

KAM Kamran Model 30, 53, 61–63, 66, 169

KAR Karolinska Institute in Stockholm, Sweden 45, 90, 91, 93, 101, 103, 107, 109

KM Kaplan-Meier 26, 27, 71, 73, 84, 85, 88, 130, 136

LNI Lymph Node Invasion 71, 82, 84, 85, 87, 88, 134, 135, 140

LSTM Long Short-Term Memory 6, 30, 31, 55

MBConv Inverted Linear Bottleneck Layers with Depth-wise Separable Convolution 32

METABRIC Study by the Molecular Taxonomy of Breast Cancer International Consortium 6, 35, 36, 53, 55, 62, 63, 65, 168–170

MICCAI Medical image computing and computer assisted intervention society. 45

MIL Multiple-Instance Learning 33, 34, 118, 137, 138

MK EHR dataset of the Martiniklinik 6, 7, 37, 39, 48, 69–71, 74–77, 80, 81, 85, 111, 139, 140, 142, 170

ML Machine Learning 2, 18

MLP Multilayer Perceptron 6, 29, 31, 54, 55, 118

MMX Prostate Cancer dataset from Malmö, Sweden 6, 39, 44, 47, 48, 50, 107, 111, 112, 114, 115, 126, 127, 130, 135, 136, 176

MSE Mean squared error 25, 58, 60

NLP Natural Language Processing 30, 31

NN Neural Network 6, 7, 30, 31

NYU New York University in New York, USA 6, 39, 43, 48, 50, 112, 114, 115, 123, 125, 130, 135, 176, 177

PANDA **P**rostate c**AN**cer gra**D**e **A**ssessment **C**hallenge 6, 33, 39, 44, 45, 48, 50, 89–91, 97, 100, 109, 139

PCa Prostate cancer 1–7, 9–11, 14, 16, 26, 31, 33–35, 37–39, 44–48, 69–71, 76, 87–89, 109, 111–113, 127, 135, 136, 139–143, 167, 175, 176

PCAI Prostate Cancer Aggressiveness Index 3, 6, 7, 31, 33, 46, 89, 111–113, 116, 117, 119, 121–123, 125–127, 130, 133–139, 141–143, 175, 177, 178

PCBN Prostate Cancer Biorepository Network 43

PDF Probability Density Function 20, 21, 27

PH Proportional Hazards 28, 29, 36, 69, 76, 87, 88, 140, 170

PLCO Prostate, Lung, Colorectal and Ovarian cancer screening trial 138

pp Percentage Points 79–81, 87, 100, 101, 123, 125, 127, 134, 136, 140

PRC Precision-Recall Curve 100

PSA Prostate-specific Antigen 2, 10, 11, 14, 38, 39, 43, 47, 71, 72, 76, 77, 81, 83–85, 87, 88, 134, 135, 140, 171

RAD Radboud University Medical Center in Nijmengen, Netherlands 34, 45, 90, 91, 93, 100, 101, 103, 107, 109

RGB Red, Green, Blue 90

RNN Recurrent Neural Network 30, 33

ROC Receiver Operating Characteristic Curve 22, 100

RP Radical Prostatectomy 1–6, 11–14, 16, 37–40, 44, 48, 69–71, 75, 76, 79, 81, 84, 87, 88, 111–113, 116, 130, 133, 134, 138–141, 167, 171

RPS Rank Probability Score 58

RT Radiation Therapy 1, 14, 39

SA Self-attention 117

SotA State of the Art 26, 31, 53, 66, 141

SPROB20 **S**pear **PRO**state **B**iopsy 20**20** dataset with 2611 biopsy slides 46

SUPPORT Study to Understand Prognoses Preferences Outcomes and Risks of Treatment 6, 35, 36, 53, 55, 61–65, 168–170

SVI Seminal Vesicle Invasion 71, 82, 84, 85, 87, 88

T-stage (path) Pathological T-stage 40, 72, 73, 76, 81, 82, 87

TMA Tissue Microarray 3, 5–7, 11, 33–35, 39–41, 43, 44, 48, 50, 79, 80, 111–116, 119, 122, 123, 125, 127, 130, 134–139, 141, 176–178

TNM Staging system for tumor extension, lymph node involvement and metastasis indication 10, 39, 81

TP True Positive 21

TPR True Positive Rate 21, 22

UKE University Medical Center Hamburg-Eppendorf 3, 5–7, 39, 40, 48, 80, 87, 111–114, 119, 122, 125, 127, 130, 133, 135, 137–139, 141, 176, 177

UKE.first UKE sub-dataset of first TMAs 40, 48, 50, 113, 114, 121–123, 125, 126, 130, 133–136, 175, 177

UKE.long UKE sub-dataset of longer stained (40:00H, 10:00E) TMAs 41, 50, 114, 122, 123, 175, 177

UKE.scanner UKE sub-dataset of scanner TMAs 40, 48, 50, 114, 121–123, 125, 126, 130, 135, 175, 177

UKE.sealed UKE sub-dataset with unknown metadata. 6, 41, 50, 111, 114, 115, 123, 127, 135–137

UKE.second UKE sub-dataset of second TMAs 40, 50, 114, 121, 122, 125, 126, 130, 135, 175, 177

UKE.thick UKE sub-dataset of thicker (10 µm) TMAs 41, 48, 50, 114, 122, 125, 136, 175, 177

UKE.thin UKE sub-dataset of thinner (1 µm) TMAs 40, 48, 50, 114, 122, 123, 136, 175, 177

UKEhv TMA sub-datasets from the UKE with variations in cutting, staining and digitization. 40, 113, 115, 121, 123, 126, 130, 135, 136

UPP Dataset from the SPROB20 image slides enriched with patient-level metadata 6, 39, 44, 46, 48, 50, 111, 112, 114, 115, 126, 127, 130, 135–137, 176

VEN Ventana 50, 107, 115

WSI Whole Slide Image 3, 6, 7, 11, 12, 32–34, 45, 89–91, 100, 103, 107, 109, 110, 167

WW Watchful Waiting 14

List of Figures

List of Tables

List of Algorithms

1 Introduction

1.1 Motivation

Prostate cancer (PCa) is the second most common cancer in men with the fifth-highest number of deaths and an increasing incidence in older age groups and a higher prevalence in developing countries [194]. While Europe registered 470,000 new cases in 2020 [68], Germany recorded almost 66,000 cases with a median age at diagnosis of 71. The survival rate is comparatively high, with a five- and ten-year relative survival rate of 91 % and 89 % respectively in 2020. Age is one of the most important risk factors for PCa. While a 35-year-old man has only a 0.5 % chance to develop PCa in the next ten years, this number increases to 7 % for a 75 year-old man [198]. Most types of PCa are slowly growing. Around 50% of men aged 50 and over 80% of men aged 75 show signs of PCa, but most of those cases remain undiscovered. For less than 20% of patients diagnosed with PCa, it is the primary cause of death [208].

Several therapies including radical prostatectomy (RP), radiation therapy (RT), hormone therapy (HT), active surveillance (AS) and others [22, 82] are considered for an individual patient. For individual treatment decision, not only the cancer needs to be evaluated, but other factors such as complications and the general health of the patient play an important role in the decision-making process.

When physicians prescribe treatments or medications to an individual patient, they must consider a wide variety of data, including, for example, the patient's clinical history, age, or blood type. Due to the increasing amount of data available to the physician, it becomes complex to find the best treatment option for an individual patient. A clinical decision support system (CDSS) that uses machine learning-based algorithms has the potential to guide the physician in finding the best decision for an individual patient. It is able to model disease trajectories, reduce variability, and increase the accuracy of treatment and outcome predictions over those of clinicians [188].

Today, most medical centers use classical statistical models, such as nomograms, that help with an individual patient's therapy decision. A nomogram provides a scalar risk score representing the likelihood of certain disease-specific events such as, for instance, progression-free survival for a defined time period or the probability of a PCa positive biopsy. As the amount of digital data collected in electronic health records (EHR) increases, statistical analysis for individual patient treatments becomes increasingly feasible. However, a current bottleneck that limits the usability of EHRs for (semi-)automated analysis is the lack of data standardization and data sparsity among patients. For efficient analysis by artificial intelligence (AI) methods, it is necessary to build a patient data representation from largely unstandardized formats, considering the large input dimensionality and sparseness of the data. AI models can help to come up with a proper data representation, predicting survival rates as well as giving treatment recommendations based on large amounts of patient data.

Since 20 % to 40 % of patients suffer from cancer relapse after RP [55], a proper adjuvant treatment is of great value for an individual patient. Even though nomograms [125, 220] are an established way to get an estimate of a patient's relapse probability, they lack the necessary flexibility for incorporating complex information that is present in an EHR. Modern deep learning approaches allow the leveraging of optimal features that are not needed to be selected or crafted manually as well as allowing not only linear, but also confounding influence with other features. With deep learning (DL), this high dimensional data can be transformed and used for personalized

risk prediction that can be the basis of decision-making.

Further, PCa biopsies can be extracted after suspicious screening results and used for further diagnosis and treatment recommendations. The gold standard for risk stratification in biopsies is Gleason grading [89] that evaluates the morphology of hematoxylin and eosin (H&E) stained prostate tissue. This process suffers high inter- and intra- observer variability among (expert) pathologists [145]. Computational pathology aims to improve the diagnostic accuracy and reduce this variability for optimal patient care [57]. Automatic grading approaches [36, 170, 222] show how this process is applicable for DL-based approaches. Going beyond the human defined Gleason grade (GG) may allow for additional accuracy in personalized patient care.

The interpretability of underlying DL models is of paramount importance for an actual application in a real world scenario. One or more visualizations of the patient's survival prediction and its most important factors will be developed for assisting physicians and patients in the shared decision-making process. The same holds for image interpretation and automatic guidance to abnormalities in the biopsy images of PCa patients that are the basis for GG.

Even though automated algorithms have gained popularity in recent years, it is often the case that those findings hold for specific clinics or centers, but lack robustness and generalizability when applied to out-of-training data [30]. This holds true for EHR approaches as well as automatic image-based predictions. To create a model that might be used outside the current clinical settings for a specific center, different documentation standards, data formats and protocols need to be accounted for leading to a more robust model. This extends to patient related images that might differ from center to center even though the same biological component is analyzed. For image-based algorithms that try to predict cancer severity, most approaches only emulate the highly variable human grading itself. This leads to the problem that even the best possible model can only achieve human performance. To prevent this problem and surpass human inaccuracies, this work focuses on objective endpoints like the patient's relapse instead.

DL models for survival prediction proved feasible [3] with advantages to classical approaches as they can handle different censoring rates, and do not overfit as easily as other machine learning (ML) approaches. Furthermore, the possibility of DL models to utilize multiple data modalities for example by combining tabular patient information with diagnostic images has additional appeal. No further preprocessing by human grading or feature engineering is required to use for example diagnostic images as additional inputs for risk prediction networks that can then be trained end-to-end without human intervention.

This work

This thesis focuses on PCa patients and how to apply recent results in AI and DL for CDSS. When PCa develops, early detection is possible with a blood test on prostate-specific antigen (PSA) [15]. With regular screenings, a rise in PSA levels can be detected early-on, providing the possibility for an early treatment of PCa patients. Additionally, digital rectal exams (DRE) are a diagnostic tool used in regular PCa screenings. Suspicious PSA or DRE diagnostics are validated by needle biopsy to ascertain the presence and severity of a possible disease, decide on a treatment and estimate the recurrence free survival time after initial treatment [214]. Nonetheless, regular PSA screenings are a controversial topic since they increase the number of diagnosed indolent cancers. These kinds of cancer are slow-growing and do not need to be treated. Nonetheless, the diagnosis leads to a significant amount of stress to affected patients. Studies have shown that regular PSA-screenings do not necessarily lead to a decrease in mortality and may instead expose the patient to additional risks introduced by unnecessary biopsies such as incontinence [6]. The accurate identification of aggressive vs. indolent cancers is an active area of research. Various treatments for aggressive PCa are available. The most common active therapy is RP where the whole prostate is removed via surgery. Widespread

use of minimal invasive robotic assisted RP as well as novel surgical techniques together with advanced pre- and perioperative risk assessment contribute to further popularization of surgical therapy approaches [167, 187]. Notably, around 30 % of patients develop relapse symptoms of cancer after the prostate was completely removed that usually requires additional treatment [55]. Also, side effects such as incontinence or erectile dysfunction can be observed in the procedure [69].

The first part of the thesis concentrates on analyzing risk estimations for PCa patients throughout the patient lifecycle using tabular data. After developing a DL-based survival prediction model, called Discrete Calibrated Survival (DCS), that is based on tabular EHR data in a more general context, the model is applied to PCa patients that received RP at the Martiniklinik to estimate their PCa relapse probability. For this purpose, a unique, large and high quality PCa EHR dataset is needed. Since the Martiniklinik provides such unique access to many high quality EHRs of PCa patients, this data is the basis of this work and rigorously used for model development and testing. This first part of the thesis analyzes the usage of such DL-based models in the context of PCa diagnosis and cancer relapse prediction. It aims to provide smart support for PCa therapy and survival prediction based on tabular patient parameters as well as images to determine cancer aggressiveness. EHRs of PCa patients will be analyzed with DL methods to enable patient-specific predictions. These EHRs cover the patient's status at the time of surgery, meaning no longitudinal EHRs (or dynamic information) is considered. The developed system should be able to aid the clinician in PCa decision-making with better performance than currently used statistical models of low complexity with rigid assumptions. Nonetheless, the recommended decisions should provide means of explanation in order to give a comprehensible result for the clinician and the patient.

The second part of this thesis deals with the most important factor in PCa relapse risk stratification that is based on Gleason grading [145, 207]. As previously mentioned, the grading process suffers high inter- and intra- observer variability among (expert) pathologists [145]. Additionally, this process may take high manual effort since often large images of biopsies need to be analyzed in great detail. This motivates the automation of a grading system that estimates cancer severity only based on these biopsy images. For this purpose, the department of Pathology of the University Medical Center Hamburg-Eppendorf (UKE) provides a large amount of PCa related tissue samples that were used to assess cancer severity. The dataset contains metadata from 17,700 individual patients with 69,251 tissue microarray (TMA) images which are small individual spots of the tissue that was selected after RP for further (mostly research related) analysis. The second part of this thesis now wants to train a DL-based model from those rather small TMA spots and assess them on PCa biopsies. The provided dataset with human annotated samples is analyzed in detail and compared to a DL based alternative that is presented in this thesis as the Prostate Cancer Aggressiveness Index (PCAI). If the learned morphologies of the TMAs translate to external datasets as well as whole slide images (WSIs) that contain whole biopsies, the system may have a clinical application since it then can circumvent the problems of manual Gleason grading.

1.2 Research Questions

This section summarizes the investigated questions that this work tries to answer. The overall question can be summarized to "How can DL models be utilized for risk assessment in the lifecycle of PCa patients?". While RQ1 - RQ3 focus on EHR-based relapse prediction, RQ4 - RQ7 go into information that can be drawn from histopathological images of potentially cancerous prostate tissue without further patient information. The research questions that are analyzed in this work are:

RQ1: How can DL models be utilized in survival prediction to generate better performance in terms of discrimination and calibration compared to classical approaches?

In contrast to classical survival prediction algorithms, DL approaches allow for highly individualized survival predictions that can result in higher discriminative power. These individualized survival estimations often coincide with a loss in model calibration [91, 96] since good calibration can be achieved with a population-based Kaplan-Meier estimation which provides a general survival estimation for a whole population. Is it possible to retain highly calibrated survival predictions as provided by classical models and combine this with better discrimination?

RQ2: Can these DL-based survival prediction models provide additional insights and a better understanding of feature importance?

In classical survival analysis like the Cox Proportional Hazards (CoxPH) model, the impact of covariates for a prediction can be broken down to a scalar factor derived from the hazard ratio. Since the influence of all covariates is linear and constant over time, it is easily interpretable. When building DL-based approaches, this easy to interpret feature importance is not applicable. Can other methods be found to quantify or visualize the impact of independent covariates or the now allowed combinations of those?

RQ3: What patient features and corresponding representations are most relevant in PCa relapse prediction after RP in the given tabular data?

Further, after conceptual research questions are analyzed, application specific problems in the domain of PCa relapse prediction arise. What tabular features of the patients is the most important factor in terms of relapse prediction in the provided PCa dataset from the Martiniklinik? How can those factors be visualized and interpreted even though non-linear dependencies exist?

Beyond survival prediction based on tabular EHR data, the following questions deal with automatic Gleason grading based on TMA spots and the extension to biopsy images of PCa patients.

RQ4: How does the knowledge derived from morphological properties of RP TMA spot images from the UKE translate to other external centers? What adaptations can be applied to improve model generalizability?
Image-based DL models often only work well within the training domain. This thesis will analyze how a model that is trained only on TMA spot images performs on other data sources, namely external TMA spots and biopsy images. Since performance is expected to drop for those datasets, additional strategies are analyzed that aim at performance improvements for those datasets.

RQ5: Do those findings of a digital risk biomarker for TMAs translate to biopsy images?
Since this work has access to a unique high-quality dataset of TMAs with rich follow-up information, training a direct prediction of TMAs to relapse risk is possible. However, clinical applicability would rather utilize such a biomarker at an earlier stage in the PCa patient lifecycle, that is, at the point of biopsy. Can the morphological properties learned on TMAs be translated to biopsies and still provide a meaningful risk prediction?

RQ6: How can the complex findings regarding tabular and image-based risk estimations be presented in an interpretable way as an additional step towards clinical applicability?
Since the developed DL-based approaches allow high interactions between covariates or predict relapse risk from images, they are often hard to interpret. Additional methods regarding model explanation should be provided to improve trustworthiness and reliability in the predictions that are provided to the physicians. How can a physician use the results to generate additional insights for individual patients that is the basis of a clinical decision support system?

RQ7: How can model uncertainty be quantified and used to boost the model's trustworthiness in a clinical setting?
For clinical application, it is crucial that some sort of confidence is provided along with the predictions of the model. If such a measure is provided, it can also increase the trustworthiness of the physician in a CDSS.

1.3 Thesis Outline

Before the thesis introduces and analyzes the aforementioned topics, a chapter-wise thesis outline is presented in the following.

Chapter 2: Background builds upon the aforementioned motivation of this thesis and provides the necessary medical foundations of PCa for this work. Afterwards, EHRs are introduced along with existing clinical approaches in the context of oncology. Further, the most important metrics for evaluation in this thesis are described to provide a basis of comparison for the state-of-the-art section, also highlighting the importance of discriminative and calibrated survival models. Here, recent developments in survival analysis are presented along with the ideas that are used to develop this work's survival prediction model DCS. Further, DL models that focus on histopathology mainly regarding PCa are discussed. Next to model performance approaches, robustness and explainability of DL models are presented before the section is concluded by a section about several related data sources that are not further discussed.

Chapter 3: Data introduces the datasets that are analyzed in this work. The main datasets are the tabular PCa EHRs from the Martiniklinik, called MK (sec. 3.1.4), and the TMA spot dataset provided by the department of Pathology of the UKE. Additional EHR datasets (SUPPORT, METABRIC and FLCHAIN) were included for larger variety in feature quality and censoring rates. Further, to evaluate generalizability and robustness to the image-based models of this work, additional internal and external sources for PCa TMA spots (UKE.sealed, NYU, JHU) and biopsies (MMX, UPP) are also discussed.

Chapter 4: Discrete Calibrated Survival Prediction describes the development of the discrete calibrated survival predictor model for EHRs called DCS. It builds upon the mathematical foundations for continuous and discrete survival prediction presented in the state-of-the-art section before diving into the model's NN architecture and the objective function parts. Afterwards, qualitative and quantitative evaluation results are presented.

Chapter 5: Postoperative Relapse Prediction on Electronic Health Records applies the aforementioned algorithm in rigorous detail to EHR data for RP treated PCa patients provided by the Martiniklinik of the MK dataset. The importance of different features within the EHRs are analyzed as well as different data representations. Further, cancer relapse-based risk stratification for the RP patients is performed and further analyzed.

Chapter 6: Patch-based Cancer Classification on Whole Slide Images shifts the focus away from EHR data to PCa WSI biopsies. Using the PANDA dataset, a patch-based cancer indicator model is used to provide an AI-guided patch importance algorithm. It is trained to determine if a selected patch contains healthy or cancerous tissue. This model can then be used as a tool to highlight the most important areas of WSIs either for humans or as a preselector of the most important region of a biopsy that is used in the next chapter.

Chapter 7: Cancer Risk Estimation from TMA Spots and Biopsies develops a risk prediction based on morphological features from histopathological H&E stained TMA spots that translates well to biopsy WSIs. Additionally, this chapter deals with model generalizability for the model that is exclusively developed on data from the UKE, but evaluated on multiple centers in the USA and Sweden.

Chapter 8: Overall Summary and Conclusion summarizes the results and findings of this thesis focusing on the advancements that lead to the three developed AI models called Discrete Calibrated Survival (DCS), Cancer Indicator (CI) and Prostate Cancer Aggressiveness Index (PCAI). Lastly, this thesis in concluded in terms of the research questions that were raised in this chapter. Additionally, remaining challenges or shortcomings are discussed along with future research ideas.

1.4 Main Contributions

This section provides an overview of the most important contributions of this thesis that analyzes how DL-based survival and risk prediction models can be used in the patient lifecycle of PCa.

Development of a DL-based survival model called DCS

This thesis explains the development of a DL-based survival prediction model called DCS that uses tabular data to predict discrete survival probabilities in time leading to a discrete survival curve. The architecture of DCS consists of a time invariant multilayer perceptron (MLP) based encoder structure, an LSTM-based recurrent module for the predicted time points, and an MLP-based aggregation part that provides the final scalar prediction at a certain point in time. To ensure model generalizability and reproducibility, three open source datasets, namely SUPPORT, METABRIC and FLCHAIN were used in the process of model development.

Analysis of the EHR-based PCa dataset

The large and unique dataset that was provided by the Martiniklinik called MK is analyzed in great detail to generate additional insights for this cohort in terms of PCa relapse prediction. Therefore, the newly developed DCS model is used alongside the classical CoxPH model that is often used in clinical practice. Firstly, the univariate influence of the most common factors regarding PCa relapse prediction are analyzed in terms of feature encoding meaning how should this data be presented to a NN as well as the CoxPH approach. Afterwards, the best representations are combined into a multivariate feature set. This set is further analyzed in terms of feature importance not on an individual, but on a block-wise level. This means that features that are usually obtained together (i.e. GG3-5) are grouped together into one block before the analysis. It is shown that the resulting feature set leads to a higher performance in terms of discrimination and calibration when the DCS model is used over the CoxPH model. Lastly, the patients are stratified into 7 risk groups based on the model's predictions that extends the model's interpretability. The groups are analyzed in terms of the patient features and could be used as an additional factor for further treatment decision.

Cancer localization on biopsy images

As a means for cancer localization on WSIs from PCa biopsies, the CI network was developed. It is a DL-based patch-wise cancer predictor that automatically finds the most cancerous patches of a prostate biopsy.

Image-based cancer aggressiveness estimation

To predict the actual risk of a patient's cancer, the DL network PCAI provides a risk prediction based exclusively on a histopathological input image of prostate tissue. It is shown that even though the model is trained only on UKE data of TMAs, the learned morphological features can be utilized on external TMA datasets as well as much larger biopsy images.

Data collection and integration

For this work, several tabular and image-based datasets, mostly PCa related, were integrated into a database that can be utilized for further research. This database contains tabular- and image-based datasets with a varying level of depth and quality. Most of the data sources remain inaccessible for the public due to patient data privacy.

Public repositories

The DL models developed in this work, namely the survival prediction model DCS[1], the patch-wise cancer prediction model CI and image-based cancer risk prediction PCAI are provided publicly[2] with the intention for further development and usage.

[1] https://github.com/imsb-uke/dcsurv
[2] https://github.com/imsb-uke/pcai

2 Background

This chapter deals with the medical and technical prerequisites of this work. Firstly the medical foundations of PCa with the most common treatment options are described. It follows an introduction of EHRs before survival analysis is introduced. Afterwards, the metrics that were evaluated for the approaches of this thesis are presented. Finally, the state-of-the-art section describes modern approaches for survival analysis, the processing of EHRs, and DL in histopathology in additional detail.

2.1 Medical Background

This section introduces the medical background of the prostate. PCa with initial diagnosis, staging, and analysis of cancer severity, including histopathology, is described before treatment options and patient relapse after initial treatment is discussed.

2.1.1 Prostate Cancer

In the following, the medical details of PCa are introduced. First, the basic anatomy of the prostate is presented before diagnostics staging is further explained. In addition, histopathology imaging for PCa is presented in additional detail along with GG to provide a meaningful measure of cancer aggressiveness based on histopathology images. Lastly, the resulting treatment options are presented.

Anatomy

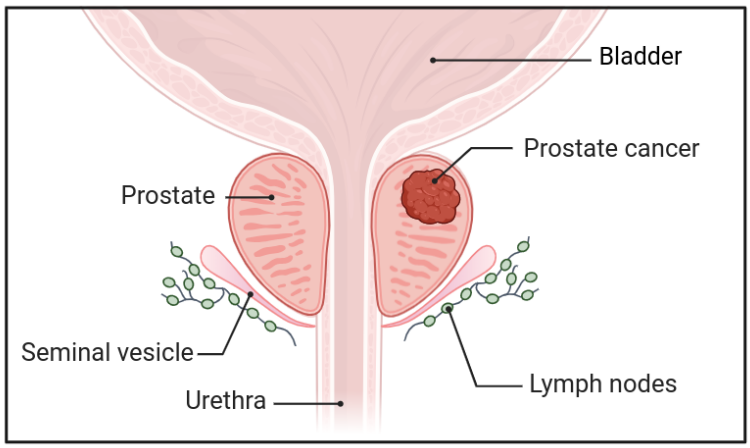


Fig. 2.1: Basic anatomy of the prostate including PCa. Adapted from "Prostate Cancer Risk Assessment", by BioRender.com (2024). Retrieved from https://app.biorender.com/biorender-templates.

The basic anatomy of the walnut-sized prostate gland is shown in fig. 2.1. It is located distal of the bladder and surrounds the proximal part of the urethra. The main purpose of the prostate is the production of prostatic fluid, which is a component of semen. It consists mainly of glandular, connective, and muscle tissue [158].

Diagnosis

Most PCa cases are slow growing and asymptomatic in the early stages [77]. This is why the diagnosis of such a tumor usually comes from a screening process. As an example, the German health care system covers this screening for all men starting at the age of 45 on a yearly basis. A PCa screening process involves a DRE where the prostate and surrounding tissue are evaluated by the palpation of a physician. In addition, the Prostate Specific Antigen (PSA) level is a protein produced by the prostate [101]. It enters primarily semen, but a small amount also goes into the bloodstream, where it can be measured by a blood test where elevated levels of PSA can indicate early forms of PCa. However, suspicious results might have other causes than PCa, including benign hyperplasia, inflammation, or prostatic manipulation. Studies show on the one hand the benefit of PSA-based screening, and on the other that they lead to unnecessary biopsies that should be avoided for healthy patients [215,264]. For diagnosis, there are still different guidelines [209] to determine the exact increase in the PSA level that is considered conclusive. However, a value of 2 ng/mL to 4 ng/mL of PSA in the blood is a typical threshold where values above the threshold are considered suspicious [23,37,168].

Tab. 2.1: Definitions of the T-stage categories following guidelines by the American Joint Committee on Cancer (AJCC) [172].

T-stage	Definition	
TX	Tumor not assessed	
T0	No tumor evident	
	Inapparent tumor	
T1	a	$\leq 5\%$ tumor in resected tissue
	b	$> 5\%$ tumor in resected tissue
	c	Tumor identified in needle biopsy
	Tumor evident inside the prostate	
T2	a	Tumor in one half of one lobe or less
	b	Tumor in more than one half of one lobe, but not both lobes
	c	Tumor present in both lobes
	Tumor extends prostate capsule	
T3	a	No other external structures involved
	b	seminal vesicles are involved
T4	Tumor invades more adjacent structures than seminal vesicles	

Staging

To discriminate the severity of cancer spread for an individual patient, a TNM staging is used. This staging process consists of three categorizations, namely T-stage (tumor extension), N-stage (lymph node spread) and M-stage (metastatic spread) that are combined to stratify the aggressiveness and the risk for an individual patient. The guidelines for the staging process [37, 172] are depicted in tab. 2.1 and described in the following:

The tumor extension is evaluated on a scale from T0 (non-evident tumor) to T4 (tumor invades not only seminal vesicles, but also adjacent structures). It is distinguished between clinical and pathological T-staging that is usually prefixed with c for clinical and p for pathological. While clinical staging is based on physical examination, imaging or biopsies of the suspicious area, pathological staging additionally includes examination of surgical results. An overview of the different classification levels for clinical and pathological T-staging can be found in tab. 2.1. Note that pathological staging uses the same type of definitions but provides a more certain classification since it is not only based on a biopsy, but the whole removed prostate. As a

consequence, T1 does not exist for the pathological T-stage.

The N-stage distinguishes N0 that reflects no lymph node involvement from N1 where metastases are found in the regional lymph node(s). Lastly, the M-stage classifies the presence of more distant metastasis of the cancer. While M0 indicates no presence, M1 describes involvement of distant, non-regional lymph nodes (M1a), bones (M1b) or other regions (M1c). Both previously mentioned stagings include the two additional categories of non-evaluation (NX and MX).

2.1.2 Histopathology

The risk assessment of the patient's cancer needs to be addressed for proper patient treatment. The gold standard is to examine the patient's biopsy and evaluate the cellular structure of the prostate tissue. It is the most important factor regarding treatment planning [241]. The process allows for the classification of the raw tissue in terms of the severity of the cancer by a pathologist and is explained in detail below.

Tissue Acquisition and Preparation

Biopsy samples are taken from patients where suspicions towards PCa are raised, e.g. through a suspicious DRE or a high PSA blood level. Moreover, this work analyzes TMAs that, in contrast to biopsy samples, only show small tissue spots that were sampled from resected prostates after RP. They are usually used in research environments. An advantage of TMAs lies in the homogeneous appearance since multiple spots are first collected into a large block and then processed altogether. This way, the variances between the spots are minimized compared to individual processing [59, 117].

Before the tissue is graded by a pathologist, a workflow is followed to ensure homogeneous analysis of tissue samples. Firstly, the tissue, that either originates from a biopsy or a resected prostate, is formalin-fixed and paraffin-embedded. This way the tissue is prepared for the next step of cutting individual spots from the tissue block. The thickness of those slices varies, but is in the order of 1 µm to 10 µm for the datasets of this work. As an intermediate step for TMAs, individual slices are cut into round spots and then arranged into larger blocks that can contain several hundred TMAs from multiple patients as visualized in fig. 2.2 [131]. An exemplary TMA block used in this work can be found in the data section at sec. 3.2.1, fig. 3.6. The biopsy datasets of this work skip this preprocessing step and are sliced individually.

After slicing, each sample is stained with H&E. While hematoxylin stains cell nuclei in a blue or purple shade, eosin stains the extracellular matrix and cytoplasm in a pink color. This way, the morphologies of the possibly cancerous cells can be analyzed in greater detail. Notably, the slicing and staining process can lead to variances of color, brightness and saturation of the resulting images depending, for example, on the slice thickness and staining time as discussed and analyzed in more detail for the sec. 3.2.1 dataset that varies these attributes in chapter 7 [45].

Scanning After preparing the tissue as presented, the cellular structure can be analyzed by a pathologist either through a microscope or by digitizing the sample. If the sample is digitized, a slide scanner is used that creates a digital copy of the individual sample. This digital whole slide image (WSI) might then contain up to billions of pixels (see, e.g., figs. 3.12 and 3.13). Due to the large image size, it may be advantageous to also store lower-resolution copies of the same image, which allows the analysis of WSIs on a coarser level. This is why a pyramidal file format is a common choice for this kind of data as depicted in fig. 2.2. The scanned slide is saved at different magnification levels that can be accessed dynamically. The lower resolution images are created by downsampling the original image. The specific data format and file extension typically differs by scanner vendor. For example, this thesis deals with WSIs that were scanned

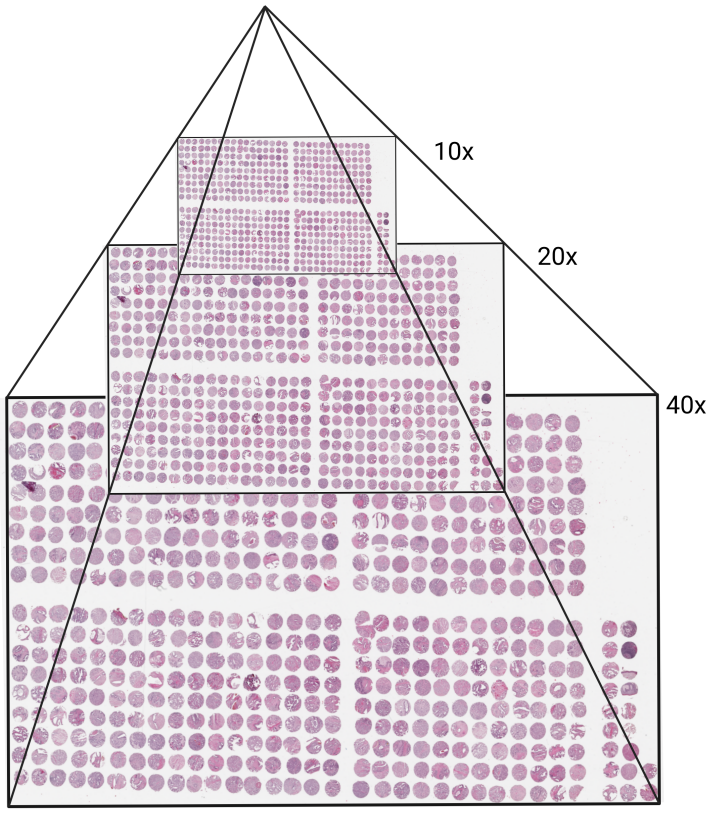


Fig. 2.2: Visualization of a TMA block that was digitized in a pyramidal image format. The digitized image contains three layers with three magnification levels of 40x, 20x RP and 10x with a resolution of approximately 0.25, 0.5 and 1 µm/pixel respectively. Created with BioRender.com

using scanners from Leica Aperio[3], Hamamatsu[4], Ventana[5] and 3DHistech[6]. This work uses the Openslide [7] library to load WSIs at different magnification levels.

Gleason Grading

If a practitioner decides to take a biopsy of the prostate, a sample can be analyzed by a pathologist using the Gleason grading system [89,161] to determine the cancer severity based on the glandular structure of the prostate tissue. The same patterns are also analyzed after RP. To distinguish the two processes, the former is usually called a clinical- the latter a pathological GG. It should be expected that the clinical grading is more inaccurate compared to the pathological grade since only a biopsy sample of the whole prostate is present for evaluation. As shown in fig. 2.3, the GG ranges from a score of 1 to 5 where a higher grade represents rarer glandular structures with poor differentiation. Current consensus classifies the lower GGs 1 and 2 as benign while GGs 3-5 are considered malignant [42].

Observer Variability Even though Gleason grading has proven itself as a reliable source for risk stratification and is considered the most important factor regarding treatment planning [241],

[3]https://www.leicabiosystems.com/de/digitalpathologie/scannen/
[4]https://nanozoomer.hamamatsu.com/
[5]https://diagnostics.roche.com/global/en/article-listing/digital-pathology-slide-scanners.html
[6]https://www.3dhistech.com/research/pannoramic-digital-slide-scanners/pannoramic-1000/
[7]https://openslide.org/

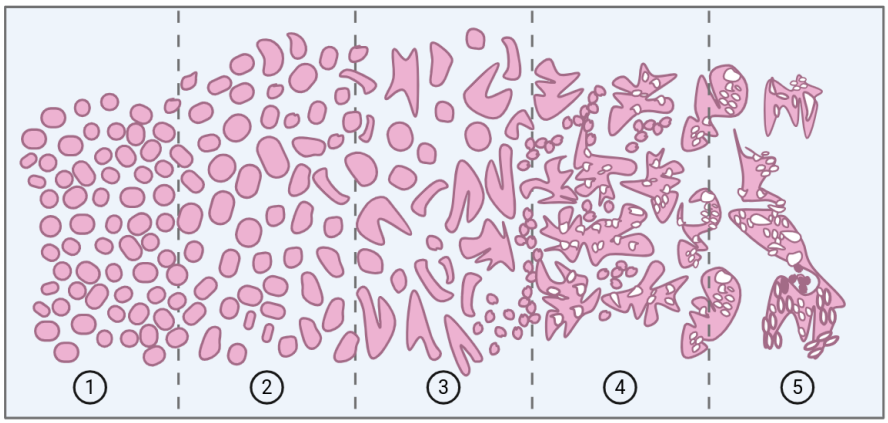


Fig. 2.3: Schematic visualization of Gleason grading based on the glandular cell structure (pink) from GG1 (well differentiated, close structures) with little stroma (white) in between to GG5 (rare regular glandular structure, more spacing, abnormal cells). Created with BioRender.com

a major drawback of the system is the high inter- and intra- observer variability [80]. Typical agreement rates for primary and secondary GG range from 36 % to 81 % between graders or 69 % to 86 % agreement plus or minus one group [70]. Another study shows a concordance rate of 58 % to 64 % [177]. This discrepancy in agreement may lead to a lack of reproducibility of GGs followed by imperfect treatment decisions for the individual patient.

Gleason Score Since a typical tissue sample does not only show a single variant of previously mentioned extends of cancer, the Gleason score is used to classify the tissue in more detail. For biopsies, the most and worst type of GG is used to represent the Gleason score consisting of a primary and secondary grade. If for example, a biopsy contained mainly GG3 and showed some regions of GG4, the primary and secondary Gleason grade of this sample would result in a Gleason score of 3+4. If only a single pattern is found, the Gleason grade is repeated e.g. as Gleason 5+5 [70, 73]. For pathological grading based on the whole prostate specimen (usually from a RP), the most and second most GG is assigned instead of the most and worst. The agreement between clinical and pathological GG remains challenging where the more common case is an undergrading of the biopsy specimen compared to the pathological grading [192, 232]. Moreover, one way to combine the primary and secondary GG is to take the sum of the two grades [49, 220].

ISUP Grading As a successor of the Gleason scoring system, the International Society of Urological Pathology (ISUP) introduced the ISUP grading in 2014 [74]. It translates the primary and secondary GG to the ISUP grade ranging from 1 to 5 for cancerous tissue. The translation between the different aggregation methods is depicted in tab. 2.2. ISUP only considers GGs 3-5. While all Gleason scores up to 3+3 are considered the lowest severity, 3+4 and 4+3 result in the medium ISUP groups 2 and 3. For ISUP4, the Gleason scores 3+5, 4+4 and 5+3 are combined. The same is done for the highest risk group ISUP5 where 4+5, 5+4 and 5+5 are aggregated.

Quantitative Gleason Grading Additional extensions to the grading system exist where a tertiary GG is reported along the first two for higher granularity. One approach is the integrative quantitative Gleason (GIQ) [206, 207] that aggregates the percentages of cancer directly with the formula

$$\mathrm{GIQ} = \mathrm{GG4}_{\%} + \mathrm{GG5}_{\%} + 0.1 \cdot \mathbb{1}_{\mathrm{GG5}_{\%}>0\%} + 0.075 \cdot \mathbb{1}_{\mathrm{GG5}_{\%}>20\%} \tag{2.1}$$

where $\mathrm{GG}i_{\%}$ stands for the ratio of the i-th GG to the total cancerous tissue of a sample and $\mathbb{1}_{(\mathrm{cond})}$ is the indicator function that resolves to 1 if (cond) is true. This way a continuous score

Tab. 2.2: Relation between Gleason score, sum and ISUP groups derived from primary and secondary GG.

Primary GG	Secondary GG	Gleason Score	Gleason Sum	ISUP
3	3	3+3	6	1
3	4	3+4	7	2
4	3	4+3		3
3	5	3+5		
4	4	4+4	8	4
5	3	5+3		
4	5	4+5	9	
5	4	5+4		5
5	5	5+5	10	

from 0 to 1.175 is achieved for a more fine-grained risk differentiation among patients. However, this grading method can be considered more laborious than assigning the Gleason- or ISUP grade and introduces additional challenges regarding inter- and intra- observer variability for the continuous score.

2.1.3 Treatment

If a patient was diagnosed by one or more of the aforementioned diagnostic tools, several therapy options exist. Several diagnostic and treatment options need to be considered in the lifecycle of a PCa patient, and their respective benefits and risks must be carefully weighed. The most suitable treatment option for the individual depends on disease- and patient-related factors among the aggressiveness of the cancer, the patient's age, physical fitness and comorbidities [39]. Fig. 2.4 provides a simplified overview of the most common treatments used for these patients.

The most popular options with curative intent are active surveillance (AS), hormone therapy (HT), radiation therapy (RT) and RP. Another treatment option is watchful waiting (WW). While AS involves monitoring the development of the cancer in certain observation windows before more invasive therapies are considered, WW focuses on monitoring and treatment of arising symptoms with a palliative intent [229].
Lastly, after a patient receives initial therapy, it may be considered whether a patient should receive an adjuvant therapy. The goal will be to stratify patients with respect to the therapy (or combination thereof) that they should receive after cancer diagnosis.

2.1.4 Relapse

The most important measure to determine the treatment quality is the progression-free survival time after surgery. It is the duration from RP until a cancer relapse can be detected for the individual. The most common detection is biochemical recurrence (BCR), indicating a rising PSA level that is found in the patient's blood. This makes the PSA level a valuable biomarker regarding cancer progression [184]. Consequently, regular follow-ups after initial treatment are recommended for patients that received a RP. On rare occasions, a rising PSA level is not detected even though the patient suffers a cancer relapse. The second most common detection of those patients is through occurring symptoms or found metastases in the body or in extreme cases the death of the patient due to PCa.

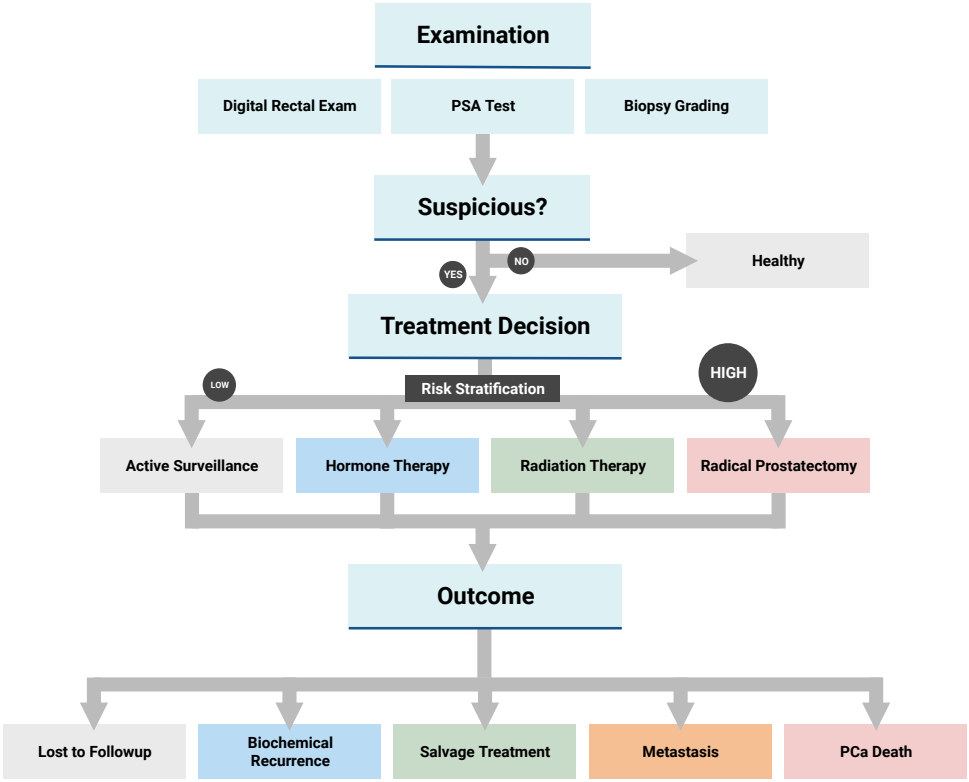


Fig. 2.4: Simplified PCa diagnosis and treatment flowchart. Based on a suspicious examination that typically involves a DRE and a PSA blood test, a biopsy is taken from the patient. A graded biopsy can help on the initial treatment decision. Low risk patients receive AS treatment while higher risk patients with curative treatment intent receive either HT, RT or RP. Subsequently, the most common outcomes are either a loss to follow-up (FU), biochemical recurrence (BCR), additional unplanned (salvage) treatment, diagnosis of metastasis, or PCa related death of the patient. Note that this is only a simplification showing this work's most relevant possible treatments and outcomes. Created with BioRender.com

2.2 Electronic Health Records

In a more general context of patient care in a clinical setting, not only for PCa patients, it is desired to base decisions on high-quality patient data including demographic information, test results, or diagnostic images. This highlights the importance of proper patient documentation and emphasizes the collection of clinical parameters in the necessary level of detail and complexity.

The collection of digital patient data in EHRs throughout a patient's medical timeline saves the known medical events in a patient's history. It combines static information like demographics with chronological information such as diagnoses, treatments, medications, and laboratory results. This enables the application of statistical data analysis to help in treatment decisions for an individual patient based on a cohort of patients with similar conditions. Prominent examples are phenotype prediction [13], patient [146] or disease [67] clustering as well as clinical decision support systems [164].

Statistical machine learning techniques like logistic regression, random forest, gradient boosting or CoxPH models are among the most used. However, the simplicity of these models as well as the requirement to handpick a few individual features that require expert level domain knowledge potentially limits the predictive power of these models. The structure of EHRs typically includes several challenges:

Incomplete data Have additional treatments been performed outside the current hospital setting? Is the patient absent because he is healthy or dead?

Heterogeneity Large differences in individual patient documentation as well as their granularity; tracking in inconsistent intervals.

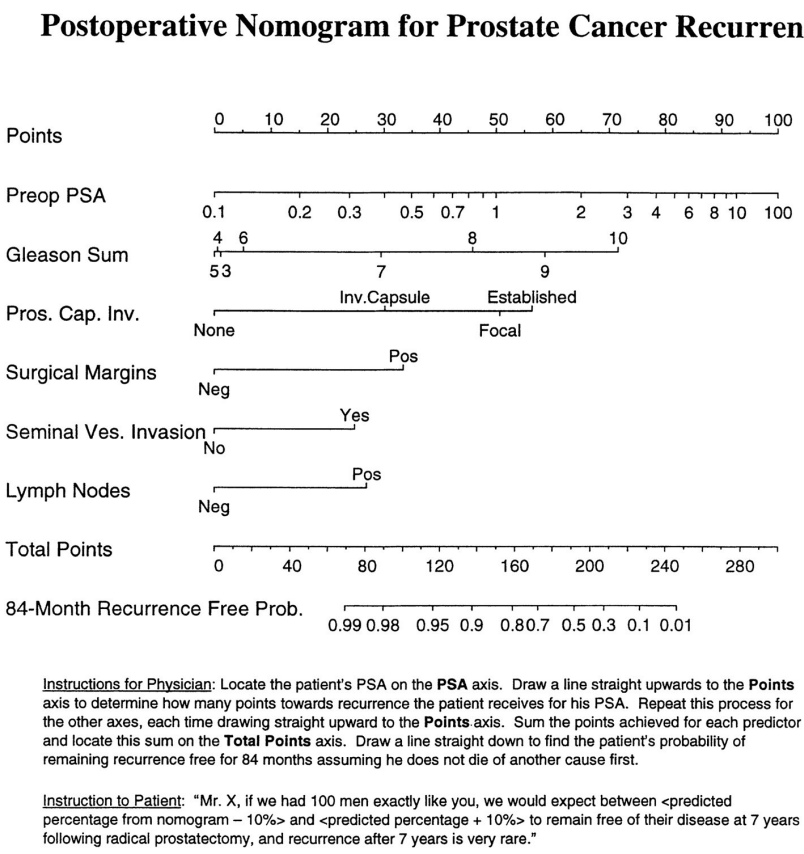
Sparsity The vast number of possible treatments, medications, diagnostic procedures and tests leads to high-dimensional and sparse EHR data, as a patient experiences only a small subset of all possible medical events.

Bias The EHRs for the same patient will most likely differ from doctor to doctor or hospital to hospital. The information stored in the EHR may also be filtered.

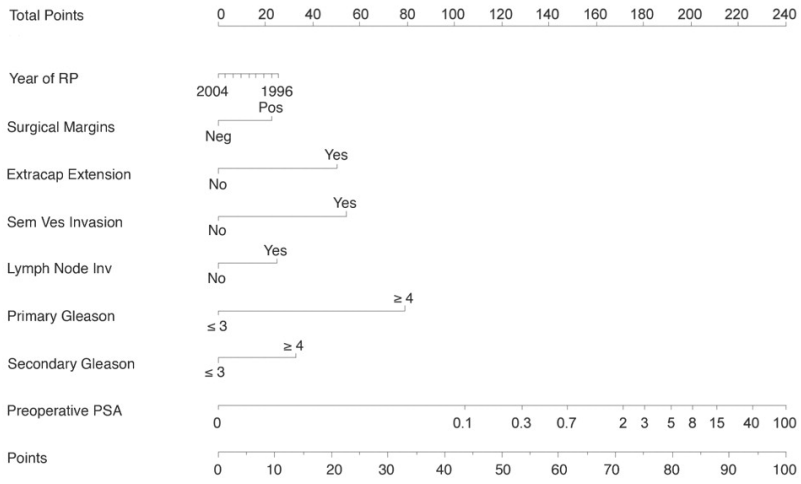
This results in a huge variety of EHR data with different quality. Another limitation of the accessible data comes from the nature of studies. Since most studies have a defined end date or lose patients over time, lots of EHRs do not necessarily contain the specific events this thesis will look for. This data only provides a minimal time for a patient where he did not encounter a certain event. This lack of data is called (right-) censoring. Some of these challenges can be addressed by DL with promising results as introduced in sec. 2.5.1.

2.2.1 Clinical Decision Support Systems

Current clinical decision support in PCa, and in oncology in general, is based on classical multivariate regression methods (known as nomograms), that have proven to be superior to clinical judgement alone [217]. They calculate the probability of certain events in the therapy pathway (such as metastasis, progression-free survival, specific pathologic features that affect surgery procedures) on the basis of patient data collected in EHRs. The nomograms are statistical tools that reflect empirical clinical knowledge about important parameters along disease progression and have grown and evolved over the years with changes in therapeutic interventions and the introduction of novel diagnostic methods. In particular, the nomograms [125, 220] for PCa patients that predict progression-free survival after receiving RP is shown in sec. 2.2.1.



(a) Postoperative relapse free survival nomogram from Kattan et al. Reprinted with permission from [125].



(b) Postoperative relapse free survival nomogram from Stephenson et al. Reprinted with permission from [220].

Fig. 2.5: Two nomograms that use patient characteristics at time of RP to predict seven (a) and ten year (b) relapse free survival probability. A patient can enter his individual features and add the resulting points graphically to obtain an individual survival probability.

Although nomograms are still accepted and used by urologists in clinical practice [8, 247], they suffer from various disadvantages. Nomograms

- were often developed within a (homogeneous) cohort (sometimes on a rather limited number of patients or without external validation), such that the models do not generalize well.
- aim to be accurate enough to convey benefit to patients, but are kept at a simple and explainable level, in order to meet with acceptance of physicians and patients, and to fit into clinical routine. This has the potential to make the model too simple for the complex input data from EHRs.
- use (mostly) a predetermined set of predictors and do not generally look for predictors systematically. Thus, they potentially miss valuable information present in the data like (non-) linear interactions between predictors that are mostly ignored.
- show little flexibility in including multiple data modalities, e.g. images, except in terms of predefined or feature-engineered scores (like the PI-RADS score [2, 94] or Gleason score (as described in sec. 2.1.2)).
- have little flexibility in choosing data models beyond multivariate regression or survival models, ignoring the potential of more complex models that might capture more information from the input data.

2.2.2 Towards Individualized Healthcare

In summary, precision medicine is one of the most revolutionary and promising advances in healthcare today transitioning from one-size-fits-all healthcare to personalized, data-driven treatment that enables improved patient outcomes. Precision or personalized medicine is understood as a medical approach in which patients are stratified based on their disease subtype, risk, prognosis, or treatment response. Medical decisions are based on individual patient characteristics, environment, and lifestyle. Physicians and researchers can use precision medicine to predict more accurately which treatment and prevention strategies will work best for a particular patient. Therefore, precision medicine offers a path to helping people recover from illness faster and stay healthy for a longer time. Precision medicine is deeply connected to and dependent on data science, specifically machine learning, which have been shown in recent years to be promising in predicting disease risk from available multidimensional clinical and biological data. Taking advantage of high-performance computer capabilities, machine learning algorithms can now achieve reasonable success in predicting risk in certain cancers and cardiovascular disease. The convergence of artificial intelligence (AI) and precision medicine promises to revolutionize health care since sophisticated computation and inference techniques are used to generate insights that empower physician decision-making. In this context, EHRs offer great promise in accelerating the predictive analysis needed in precision medicine. In the last decade, predicting the risk of developing certain diseases in patients has become an important research topic in healthcare and accurate identification of similarities among patients based on their historical records is a key step in personalized healthcare. Furthermore, explanations that support the output of an ML model are crucial in precision medicine, where experts require more information from the model than a simple binary prediction to support their diagnosis. Therefore, explainable or interpretable models, which fall within the field of explainable artificial intelligence, allow healthcare experts to make reasonable and data-driven decisions to provide more personalized and precise treatments. [165, 204]

2.3 Survival Analysis

This section introduces fundamental ideas of survival analysis and prediction, as well as the mathematical notation used in this work.

Survival analysis or, in more general terms, time-to-event analysis, is a branch of statistics that analyzes lifetimes of individuals or populations regarding a particular event (like patient death, mechanical failure or customer churn) and infers what determines the underlying distributions [129]. In biomedical statistics, patient data is analyzed to predict, for instance, the impact of specific patient characteristics on survival. This population-level analysis can determine which (groups of) patients have a high event risk and which prognostic features might be responsible for this. Survival prediction deals with predictions of future events under the current conditions of a population or individual. While for some patients the event of interest is recorded, others might not be observed e.g. by dropping out of a study at any point in time. This is called (right-) censoring of the data. This lack of information might also be present before the observation window (left censoring) or even within multiple observation windows (interval censoring). In this work, only right-censoring of the data that is considered that is usually present in medical studies. The observation of participants starts at some point in time, but not all participants experience an event throughout the observation window. Either because their individual event time lies beyond this window of the study or because the individual drops out of the study at any given point in time. Even though some participants do not experience the event of interest, they still contain information about how long they at least stayed event-free. The use of this partial information of censored individuals distinguishes survival analysis from regression problems [162]. The performance of these models is usually evaluated regarding the concordance index (C-index) that measures the discriminative performance for pairs of individuals (i.e. shorter event times should have higher risk predictions). However, especially in clinical settings, the correct prediction of the underlying event time distribution is of high relevance for the patient and medical practitioner, as it can guide their decision-making. The correct temporal prediction of an event can be measured by the calibration of a model, for example, by the Distributional Divergence for Calibration (DDC) [95]. Recent work focuses on the development of deep learning-based algorithms for survival prediction, including models that produce a continuous time output like DeepSurv [126] and CoxTime [136]. Recent DL approaches such as Discrete Recurrent Survival Analysis (DRSA) [195] focus on risk prediction at discrete time points. While these discrete DL models reach state-of-the-art discrimination performance, most disregard model calibration, which makes them of limited value for medical survival prediction. To address this shortcoming, [123] developed an objective function that additionally takes proper calibration into account. This thesis also develops a survival prediction model, called DCS, that is explained in detail in chapter 4 that extends the previously mentioned ideas.

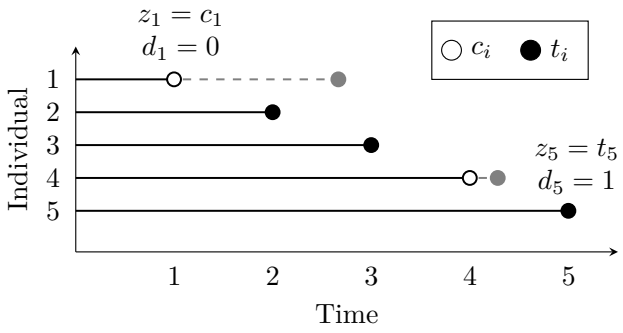


Fig. 2.6: Visualization of partially right-censored observation times for a population of five individuals over time. All censored individuals show an observed time $z_i = c_i$ where the real event time $t_i > c_i$ is hidden (gray). For the uncensored individuals that experience an event, the observed time corresponds to the real time of event ($z_i = t_i$).

The following section describes the mathematical foundation of survival analysis of this work based on various notations and ideas [123, 136, 195].

2.3.1 Input

For each individual i in a population of n individuals, consider a vector of n_{feat} features $\boldsymbol{x}_i \in \mathbb{R}^{n_{\text{feat}}}$, a corresponding event $t_i \in \mathbb{R}_+$ and censoring time $c_i \in \mathbb{R}_+$. The individual's observed time $z_i \in \mathbb{R}_+$ is defined as $z_i = \min(t_i, c_i)$ meaning it is either the censoring or event time. Furthermore, the corresponding censoring indicator $d_i = \mathbb{1}_{y_i = t_i}$ distinguishes censored ($d_i = 0$) and uncensored ($d_i = 1$) observed times z_i. Fig. 2.6 shows an example population with five individuals who experience an event or are censored. With this notation, the population $\mathbb{P}$ that contains n individuals can be represented with the aforementioned definitions as

$$\mathbb{P} = \{(\boldsymbol{x}_i, z_i, d_i) \mid i \in \{0, 1, \ldots, n-1\}\} \tag{2.2}$$

that can now be used as an input for a survival prediction. In this thesis, the most common case of right-censored populations is considered, leaving out other possible scenarios like left- or interval censoring [54, 124, 249].

2.3.2 Survival Prediction

For the observed time z of the event-of-interest, the probability density function $p(z)$, or PDF, represents the probability that the specific event happens at time z. Integrating the PDF yields the cumulative distribution function

$$S(t) = \int_t^\infty p(z)\,\mathrm{d}z \tag{2.3}$$

that describes the probability for an individual to not experience the event until time t as $P(z > t)$ or, in other words, the probability of surviving until time t. In this thesis, this function is called survival function or survival curve. Further, following the notation from [78, 123], the hazard rate at time t can be defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < z \leq t + \Delta t \mid z > t)}{\Delta t} \tag{2.4}$$

that represents the conditional probability to experience the event at time t given that the event was not experienced yet.

Discretization

This section deals with the discretization of the aforementioned continuous notation. Given the maximum observation time $t_{\max} = \max(z_i)$, the observation window $(0, t_{\max}]$ can be divided into $L+1$ discrete time points $t_0 = 0 < t_1 < \ldots < t_L \leq t_{\max}$. These discrete time points t_l divide the observation window into L disjoint intervals $V_1, V_2, \ldots, V_L$ where the l-th interval is defined by two adjacent discrete points in time as $V_l = (t_{l-1}, t_l]$. These defined intervals are visualized in fig. 2.7.

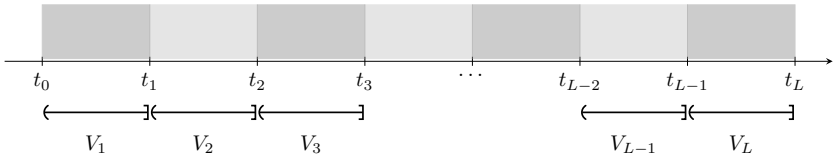


Fig. 2.7: Discretized output grid visualization that divides the observation window $(t_0, t_L]$into L intervals $V_1, V_2, \ldots, V_L$. Round brackets mean exclusive, square brackets inclusive of t_l.

Following [195], this discretization step can also be done for the resulting survival function as

$$S(t_l) = \Pr(z > t_l) = \sum_{j > l} \Pr(z \in V_j). \tag{2.5}$$

Similarly, the discrete PDF p_l can be reformulated to

$$p_l = \Pr(z \in V_l) = S(t_{l-1}) - S(t_l) \tag{2.6}$$

further allowing the definition of discrete hazard rates of the current interval V_l given that the individual did not experience the event until t_{l-1} as

$$h_l = \Pr(z \in V_l \mid z > t_{l-1}) = \frac{\Pr(z \in V_l)}{\Pr(z > t_{l-1})} = \frac{p_l}{S(t_{l-1})}. \tag{2.7}$$

2.4 Metrics and Measurements

The following section introduces the performance metrics and additional measurements that are used throughout this work, among other cases, to evaluate classification and survival prediction estimations.

2.4.1 Classification

For a vector of n predictions $\hat{\boldsymbol{y}} \in [0,1]^n$ and a corresponding classification ground truth (GT) label $y_i \in \{0,1\}$, the following metrics are used in this thesis. Some metrics require a binary prediction for a single class problem. This is usually achieved by thresholding the prediction at $t = 0.5$ unless otherwise stated as $\hat{y}_i^b = (\hat{y}_i > t)$. For accuracy, the number of true positives

$$\mathrm{TP} = \left|\left\{1 : \hat{y}_i^b = 1,\, y_i = 1,\, i \in \{0,1,\ldots,n-1\}\right\}\right| \tag{2.8}$$

is divided by the total number of samples n in the dataset. Since this metric may lead to high scores for bad predictors on imbalanced datasets [205], precision and recall are also considered in this thesis. While

$$\text{precision} = \frac{\mathrm{TP}}{\left|\left\{1 \mid \hat{y}_i^b = 1,\, i \in \{0,1,\ldots,n-1\}\right\}\right|} \tag{2.9}$$

measures the ratio of TPs over all predicted positives where $|\cdot|$ denotes the cardinality of the set,

$$\text{recall} = \frac{\mathrm{TP}}{\left|\left\{1 \mid y_i = 1,\, i \in \{0,1,\ldots,n-1\}\right\}\right|} \tag{2.10}$$

or the true positive rate (TPR) calculates the ratio of TPs over all true positives as illustrated in fig. 2.8.

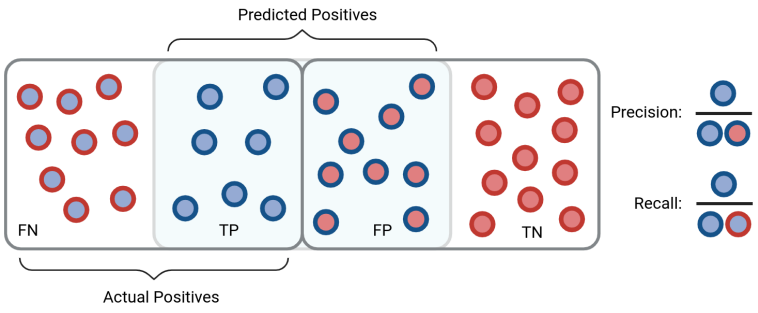


Fig. 2.8: Visualizing precision and recall using true positive (TP), false positive (FP), false negative (FN) and true negative (TN) predictions.

Following the aforementioned definitions, two other metrics are also used in this work that do not require a fixed threshold t for the binary prediction $\hat{y}_i^b$ of a sample, but take into account

all possible thresholds $t \in [0, 1]$. The first metric is the receiver operating characteristic curve (ROC) which relates TPR to the false positive rate that is defined as

$$\mathrm{FPR}(t) = \frac{\left|\left\{1 \mid y_i = 0, \hat{y}_i > t, i \in \{0, 1, \ldots, n-1\}\right\}\right|}{\left|\left\{1 \mid y_i = 0, i \in \{0, 1, \ldots, n-1\}\right\}\right|}. \tag{2.11}$$

The resulting area under the ROC (AUROC) provides a performance indicator for a continuous prediction $\hat{y}_i$ for all samples in the dataset. Another related measure is the precision recall curve that relates recall over the precision similarly. Similarly, it also provides a scalar overall measure with the area under the precision recall curve (AUPRC).

2.4.2 Measure of Agreement

As a measure of agreement between multiple raters that can choose from multiple categories per rating, a common metric is the Cohen's kappa [51] value. For a k-class classification of n predictions $\hat{\boldsymbol{y}}^{(0)} \in \{0, 1, \ldots, k-1\}^n$, the agreement with another prediction of the same shape $\hat{\boldsymbol{y}}^{(1)}$ can be evaluated. Firstly, the observed agreement rate $\hat{p}_o \in \mathbb{R}$ for every individual case i that shows agreement $\hat{y}_i^{(0)} = \hat{y}_i^{(1)}$ between the two predictions is calculated. Afterwards, Cohen's kappa

$$\kappa = 1 - \frac{1 - \hat{p}_o}{1 - \hat{p}_r} \tag{2.12}$$

additionally takes random agreement into account which is defined as

$$\hat{p}_r = \frac{1}{n^2} \sum_{k=0}^{n-1} n_{k_1} n_{k_2} \tag{2.13}$$

where n_{k_i} counts how often class k is predicted in $\hat{y}^{(j)}$ for $j \in \{0, 1\}$. A perfect agreement of the two raters results in $\kappa = 1$ while values around 0 are considered to be not in agreement. Negative values are also possible and would indicate disagreeing annotators.

Moreover, the quadratic weighted kappa extends this approach for ordinal scales where larger discrepancies are punished more than close ones. It accounts for the severity of disagreement where an adjacent predicted category is punished less than a disagreement that spans multiple categories [52]. Another extension to the agreement is called Fleiss kappa [81] that measures the level of agreement among more than two raters.

2.4.3 Segmentation

Since this work does not focus on segmentation metrics, the only metric that is mentioned is the dice score. To calculate the dice score for a binary prediction mask $\hat{\mathbf{Y}} \in \{0, 1\}^{h \times w}$ and a binary GT mask $\mathbf{Y} \in \{0, 1\}^{h \times w}$ with the same width w and height h, the dice score [63, 216] is defined as

$$\mathrm{DICE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{2|\hat{\mathbf{Y}} \cap \mathbf{Y}|}{|\hat{\mathbf{Y}}| + |\mathbf{Y}|} \tag{2.14}$$

which equals 1 for perfect overlap and 0 for no overlap. $|M|$ stands for the total number of non-zero entries in M.

2.4.4 Survival Prediction

In contrast to classification metrics, no GT class label exists when the survival of individuals is predicted. Only the true survival or censoring time of an individual is known. This section deals with the metrics for survival prediction models in this work. This includes metrics for scalar predictions (like a risk or a probability of surviving n years) as well as metrics that evaluate whole survival curves. In contrast to classification problems, a large range of related, but different metrics exist to measure the performance of survival prediction models. Extensive analyses of survival analysis metrics can be found in [95, 183, 196, 233, 234].

Discrimination and Calibration

The most common way to quantify predictive performance is based on the correct (or concordant) ordering of predictions and actual survival or censoring times. This approach is commonly referred to as discriminative performance [221]. Discrimination quantifies if a model predicts the correct order of events, it does not scrutinize if the predicted event occurs at the correct time, since only the relative ordering is taken into account. An example for the usage of discrimination is an organ transplant waiting list, where the patient with the best survival estimate after transplantation may be selected as the most suitable candidate. In this thesis, discrimination is measured with the most commonly used metric, the concordance index [100]. Further, to estimate the discriminative performance of predicted survival curves that potentially cross over time, the time-dependent concordance index (C-index-td) [7] and the cumulative-dynamic AUROC (CDAUC) [234] are used.

In contrast to the discriminative performance, a patient or a physician may be interested if the prediction reflects the true underlying survival distribution. A 2 or 20-year relapse time will make a difference in the selected therapy, regardless of the correctness of the order compared to other patients. Here, the calibration performance is used. Calibration determines whether a scalar prediction, e.g. for a five-year survival rate, actually matches the underlying survival distribution. If this is the case, the physician can use survival probability as a true measure for the individual patient. The most common measure that takes calibration into account is the Brier score (BrS) [33]. It punishes uncertain estimates that would otherwise profit, so a high calibration still needs good discrimination to be practically feasible. Similarly to the C-index, it can also be extended to a time-dependent metric, the Integrated Brier Score (IBrS). Another metric that more measures the calibration performance is the Distributional Divergence for Calibration (DDC) that compares the observed event distribution to the predicted probabilities over time. Additional details that build a greater intuition for calibration can be found in [110,262].

Other metrics such as the absolute difference in predicted event times, or D-calibration [95] exist, but are not further discussed in this thesis. The following section explains the previously mentioned metrics in additional detail.

Concordance Index

The concordance index [99] (C-index), is the most commonly used metric for discriminative performance of individual scalar risks $\hat{r} \in \mathbb{R}$. It is a generalization of the AUROC [7] that evaluates the correct order of the predictions compared to the actual times of events, including censored cases ($d_i = 0$). As an intuition, the C-index measures the number of correctly ordered pairs of individuals over all possible comparisons. A pair of predictions $(\hat{r}_i, \hat{r}_j)$ where $\hat{r_i} > \hat{r_j}$ is considered correctly ordered if the individual with the higher predicted risk $\hat{r}_i$ experiences the event of interest z_i sooner than the other individual's event or censoring time z_j. Consequently, the C-index for a set of n risk predictions $\boldsymbol{r} \in \mathbb{R}^n$ with corresponding observation times $\boldsymbol{z} \in \mathbb{R}_+^n$ and event indicators $\boldsymbol{d} \in \{0, 1\}^n$ is calculated as

$$\text{C-index} = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \mathbb{1}_{z_i < z_j} \mathbb{1}_{d_i = 1} \mathbb{1}_{\hat{r}_i < \hat{r}_j}}{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \mathbb{1}_{z_i < z_j} \mathbb{1}_{d_i = 1}} \tag{2.15}$$

which is the number of concordant pairs over all possible comparisons. As an intuition, this can be interpreted as saying that a comparable pair that is randomly drawn from the sample population is concordant as $\Pr(\hat{r}_i < \hat{r}_j \mid z_i < z_j, d_i = 1)$. A comparison is possible when the i-th individual is not censored ($d_i = 1$) as a censoring time z_i cannot be compared to a longer observation time z_j that might be a censoring ($d_j = 0$) or event ($d_j = 1$) time. Note that this metric does not evaluate the absolute differences between r_i and r_j as long as they are ordered correctly. Furthermore, the C-index is known to be biased towards better scores on populations

with high censoring rates [233] and is in general not a proper score [29]. A proper score ensures that no prediction can achieve a higher discriminative value than the data-generating process itself. This is why additional metrics for discriminative performance are taken into account. A more detailed analysis into the compared pairs is performed in sec. 4.2.3.

Time-dependent Concordance Index If instead of individual risks r_i, a survival curve with $L \in \mathbb{N}^+$ discrete values $\hat{S}(t) \in [0,1]^l$ is predicted and evaluated, the order of those predictions can change over the observation window, i.e. $\hat{S}_0(t_0) < \hat{S}_1(t_0)$, and $\hat{S}_0(t_1) > \hat{S}_1(t_1)$. This is why the C-index is extended to be evaluated at multiple discrete timepoints to incorporate for this behavior of the predicted survival curves. To properly evaluate these time-dependent changes, this work utilized the time-dependent concordance index C-index-td [7], which evaluates the C-index at the event times $\boldsymbol{z}$ of the population:

$$\overline{\text{C-index-td}} = \Pr\left(\hat{S}_i(z_i) < \hat{S}_j\left(z_i\right) \mid z_i < z_j,\, d_i = 1\right) \tag{2.16}$$

$$\tag{2.17}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{n \times l}$ is the matrix of the l predicted discrete timepoints of the individual predicted survival curve $\hat{S}_i$ for all $i \in \{0, 1, \ldots, n-1\}$ individuals of a dataset. Following [29, 257], an extension is made for equal predictions that are counted as half cases to the resulting version of the metric that is used in this thesis:

$$\begin{aligned} \text{C-index-td} =& \Pr\left(\hat{S}_i(z_i) < \hat{S}_j\left(z_i\right) \mid z_i < z_j,\, d_i = 1\right) \\ &+ \frac{1}{2}\Pr\left(\hat{S}_i(z_i) = \hat{S}_j\left(z_i\right) \mid z_i < z_j,\, d_i = 1\right). \end{aligned} \tag{2.18}$$

Time-Specific AUROC

As an alternative measure of how well a risk performs with respect to discriminative performance, a time-specific AUROC at time t within the observation window can be used for a directly predicted scalar risk $r_i \in \mathbb{R}$ or a prediction at that point in time t of a predicted survival curve $\hat{S}_i(t)$ that can be interpreted as a risk by creating the complement to 1 as $r_i = 1 - \hat{S}_i(t)$. With this point in time t, cumulative cases $\text{Ca} = \{(r_i, \bar{y}_i = 1) : z_i \leq t, d_i = 1\}$ that experience the event prior to or at time t can be compared to controls $\text{Co} = \{(r_i, \bar{y}_i = 0) : z_i > t\}$ using AUROC as described in sec. 2.4.1 to achieve a scalar metric for the discriminative performance of the predicted individual risks r_i at time t.

Cumulative-Dynamic Time-Dependent AUROC Building on the previous section, the cumulative-dynamic time-dependent AUROC [29, 234] provides a measure not only for a specific time t, but for the whole observation window for predicted survival curves $\hat{S}_i(t)$.

Therefore, the censoring- and time-dependent area under the curve (AUC) is calculated as

$$\widehat{\text{AUC}}(t) = \frac{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \frac{\mathbb{1}_{z_i \leq t}}{\tilde{C}(t)} \mathbb{1}_{z_j > t} \mathbb{1}_{\hat{S}_i(t) \geq \hat{S}_j(t)}}{\left(\sum_{i=0}^{n-1} \frac{\mathbb{1}_{z_i \leq t}}{\tilde{C}(t)}\right)\left(\sum_{j=0}^{n-1} \mathbb{1}_{z_j > t}\right)} \tag{2.19}$$

$$\tag{2.20}$$

where $\tilde{C}(t)$ is the censoring distribution of the dataset over time that is usually estimated by a Kaplan-Meier estimator as shown in sec. 2.5.1. Further, $\widehat{\text{AUC}}(t)$ can be integrated within a certain time range $[\tau_0, \tau_1]$ to produce a scalar metric as

$$\text{CDAUC} = \frac{1}{\tilde{C}(\tau_0) - \tilde{C}(\tau_1)} \int_{\tau_0}^{\tau_1} \widehat{\text{AUC}}(t)\, d\tilde{C}(t). \tag{2.21}$$

Unlike the C-index, CDAUC is a proper scoring method meaning that no estimating model can be constructed that has a higher discriminative performance than the data-generating process itself [29].

Brier Score

One metric that not only takes discrimination into account, but also calibration, is the Brier score (BrS) [33, 136] that is defined as

$$\mathrm{BrS}(t) = \frac{1}{n}\left(\sum_{i=0}^{n-1} \mathbb{1}_{z_i > t} \frac{\left(1 - \hat{S}_i(t)\right)^2}{\tilde{C}(t)} + \sum_{i=0}^{n-1} \mathbb{1}_{z_i \leq t} \mathbb{1}_{d_i = 1} \frac{\hat{S}_i(t)^2}{\tilde{C}(t)}\right) \tag{2.22}$$

where the mean squared error (MSE) of the predicted survival curve to 1 before, and to 0 after the individual event time z_i is measured and weighted by the Inverse Probability of Censoring Weights [197]. The latter sum is only relevant for uncensored individuals ($d_i = 1$) as the further progress of censored individuals is unknown.

Integrated Brier Score Similarly to CDAUC, a scalar metric over the observed timespan is calculated by integration:

$$\mathrm{IBrS} = \frac{1}{\tilde{C}(\tau_1) - \tilde{C}(\tau_0)} \int_{\tau_0}^{\tau_1} \mathrm{BrS}(t)\, d\tilde{C}(t). \tag{2.23}$$

For this MSE-based metric lower values are better, but a sophisticated survival estimator should achieve a value of the BrS that is below 0.25 [87]. Note that, similarly to C-index, BrS is an improper scoring rule [258].

Distributional Divergence for Calibration

Following [123], this thesis uses an additional measure for calibration, namely the DDC. The estimated survival for each individual at their event time $\hat{S}_i(z_i)$ is mapped into 10 equally sized bins $B = \{[0, 0.1), [0.1, 0.2), \ldots, [0.9, 1)\}$ in the unit interval as shown in [95]. For DDC, the Kullback–Leibler divergence [135] is calculated between the relative frequency P of the observed intervals of the whole distribution compared to a uniform distribution $Q(b) = 0.1$ for $b \in B$ that is expected for perfect calibration as

$$\mathrm{DDC} = \sum_{b \in B} P(b) \log\left(\frac{P(b)}{Q(b)}\right). \tag{2.24}$$

Summary

The metrics that are usually used in survival prediction are focused on discriminative performance. This thesis uses C-index and the time-dependent version, C-index-td along with the CDAUC as discriminative measures. Since the most popular metric among related publications is the C-index, it is also used frequently throughout the thesis even though several drawbacks exist e.g. that C-index overestimates predictive power on datasets with higher censoring rates [233], or that the C-index is not a proper scoring method [29]. This means that prediction models exist that can generate higher discriminative performance than the data generating process itself. This is why CDAUC is introduced as a more unbiased, but complex metric. Further, calibration performance is measured by BrS (with the time-dependent IBrS) as a metric that measures both discriminative and calibration performance. Since BrS is also an improper scoring method [258], the DDC is additionally used as a metric that exclusively aims at measuring model calibration. By evaluating the different metrics, this thesis aims to provide a more complete picture of the

survival predictions analyzed.

In a more general sense, survival metrics show high sensitivity in data sets with, e.g., high censoring rates [237]. Furthermore, the discriminative metrics often depend not only on the individual predictions, but on the dataset itself, since individuals are often compared to each other. This makes the metrics non-linear and only meaningful for model predictions that are evaluated on the same dataset [87].

2.5 State of the Art

This section describes SotA approaches with a general focus on DL methods in survival prediction, the interpretation and usage of EHRs, and how methods can be applied to histopathological images, focusing on PCa-related publications.

2.5.1 Survival Analysis

Following the general ideas of survival analysis that were introduced in sec. 2.3, this section describes the most common survival predictors. The presented estimators can be categorized in the following ways.

Some survival models presented in this thesis estimate the underlying survival distribution $\hat{S}(t)$ where medical research is mostly focused on the discriminative performance of these models [110]. An alternative to the survival curve estimation $\hat{S}_i(t)$ for an individual i is to simplify the prediction of the model to an individual risk estimation $r_i \in \mathbb{R}$ that can also be evaluated in terms of discriminative performance. It is also possible to provide a scalar risk along with a time-dependent survival curve estimation [65,88]. A further distinction between predictors can be made regarding the calibration of the predicted risk. Since discriminative performance only measures the correct risk ordering regardless of scale, such a risk can also be expressed by a survival probability of l years. Only in the latter case it makes sense to address the model's calibration. Further, instead of predicting an individual risk r_i, the expected survival time $\tilde{z}_i$ can also be used as an estimation target that transforms the survival estimation into a classical regression problem [86, 143, 261]. For naive approaches, the latter requires that individuals actually experience the event-of-interest since it cannot be estimated for censored individuals. This makes such a method infeasible for datasets with high censoring rates.

Further, survival prediction models can be distinguished in the following ways: Firstly, does the model predict a risk meaning an arbitrarily scaled scalar where a higher number reflects a higher risk of experiencing an event or does it provide a probability? Next, is this prediction done for a whole timeframe, a specific- or multiple points in time? Lastly, is this prediction based on an individual or a (sub-) population? These questions help to identify what kind of survival prediction is performed [96].

The following section provides information on some of the most commonly used models in survival analysis with special focus regarding DL approaches.

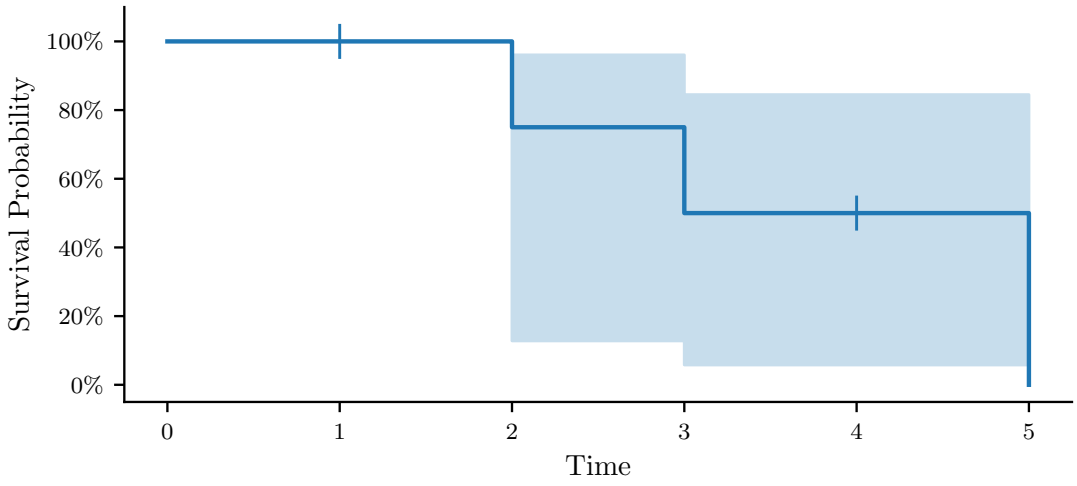
Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator [124] introduces a population-based non-parametric survival estimator by calculating the survival probability for a point in time $t \in \mathbb{R}_+$ as

$$\hat{S}_{\mathrm{KM}}(t) = \prod_{t_i \in P_t,\, t_i \leq t} \frac{n_i - e_i}{n_i} \tag{2.25}$$

where $\mathbb{P}_t$ is the set of all distinct event times z_i in the population $\mathbb{P}$, e_i denotes the number of individuals that experience an event at time t_i and n_i the remaining number of individuals at risk

at t_i. It therefore is based on the relative frequency of individuals that survive a certain point in time over all individuals that are still at risk. Fig. 2.9 depicts an exemplary KM estimation for a population for the right-censored dataset of five individuals that were presented in fig. 2.6.



KM					
At risk 5	4	3	2	1	0
Censored 0	1	1	1	2	2
Events 0	0	1	2	2	3

Fig. 2.9: Exemplary KM curve of censored and uncensored individuals over time. Also shown are censoring events as vertical lines and the 95 % confidence interval and the total number of individuals at risk, censored, and per timestep.

Confidence Interval Estimation

To provide a measure for the confidence intervals of the KM prediction, the exponential Greenwood estimator [92, 121] is commonly used. It assumes a binomial distribution when observing e_i failures among the n_i at-risk cases at a specific time t. For a point in time t, it is defined as

$$\exp\Big(-\exp\Big(\mathrm{CI}^{-}_{\mathrm{KM}}(t)\Big)\Big) < \hat{S}_{\mathrm{CI}}(t) < \exp\Big(-\exp\Big(\mathrm{CI}^{+}_{\mathrm{KM}}(t)\Big)\Big) \tag{2.26}$$

where the time-dependent confidence interval with a corresponding variance can be described as

$$\mathrm{CI}^{\pm}_{\mathrm{KM}}(t) = \log\Big(-\log\Big(\hat{S}_{\mathrm{CI}}\Big)\Big) \mp q_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(t)} \tag{2.27}$$

where α determines the width of the confidence interval that is commonly set to 95 % ($q_{0.025} = 1.96$), and

$$\widehat{\mathrm{Var}}(t) = \frac{1}{\log\Big(\hat{S}_{\mathrm{CI}}(t)\Big)^2} \sum_{t_i \le t} \frac{e_i}{n_i(n_i - e_i)} \tag{2.28}$$

describes the underlying variance of the observed data respectively. The advantage of the exponential formula over the standard Greenwood estimate is that the former ensures that the confidence intervals stay between 0 and 1.

Note that the estimator does not depend on the individual's covariates $\boldsymbol{x}_i$. It is based exclusively on the observed event and censoring times z_i. In contrast to the KM estimator that models the survival distribution using (sub-) populations, the following survival models try to estimate the underlying PDF for each individual i given the corresponding feature vector $\boldsymbol{x}_i$ as $p(z|\boldsymbol{x}_i)$.

Cox Proportional Hazards Model

The classical approach for survival analysis is the CoxPH model [54]. The idea of the CoxPH model is not to perform the regression task on the individual observed event and censoring times, but the hazard rates h based on the individual features $\boldsymbol{x}_i$. Since those rates are not observed directly, they are optimized by maximizing the partial likelihood instead [53]. The CoxPH model finds the parameters that best reproduce the discrimination of individual event times by fitting a logistic regression model that is based on the proportional hazards (PH) assumption and assumes linear and independent covariates [32]. The features of each individual are linearly combined as $r_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$ where the optimal values for $\boldsymbol{\beta}$ are determined using a negative partial log-likelihood approach as

$$\mathcal{L}_{\text{CoxPH}} = \frac{1}{m} \sum_{i:d_i=1} \log \left(\sum_{j:z_j \geq z_i} \exp\left(g(\boldsymbol{x}_j) - g(\boldsymbol{x}_i)\right) \right) \tag{2.29}$$

where m is the number of uncensored individuals. This likelihood approach compares the obtained risk score r_i to all individuals with a greater or equal event time $z_j \geq z_i$ for each uncensored ($d_i = 1$) individual i. Subsequently, the regression model is combined with a time-dependent baseline risk $h_0(t)$ that is the same for the entire population. By this design, predicted survival curves have the same shape for all individuals (effectively discarding individual temporal feature-dependent information) of the population and only differ in a unique parametric scaling factor per patient that depends on the individual's input features.

In detail, the proportional-hazard assumption states that the same hazard function is applicable to all individuals, but a scalar scaling factor discriminates between them. This means that the hazard for each individual can be split into a time-dependent and an individual feature-dependent hazard part as

$$h_i(t) = a_i h_0(t) \tag{2.30}$$

It follows that the comparison between two individuals has no time dependency and can be expressed as

$$\frac{h_i(t)}{h_j(t)} = \frac{a_i}{a_j} \tag{2.31}$$

which allows direct comparison between the hazard rates of two individuals. This also holds for the likelihood function in eq. (2.29) that can as a result be trained independently of the baseline hazard function. To test if the assumption holds for a dataset and its covariates, it can be analyzed using Schoenfeld residuals [180, 219]. If the residuals show a non-random time dependency (like a growing residual error in time), the assumption is most likely violated.

Hazard Ratios One advantage of the CoxPH model is the simple feature importance interpretation based on logarithmic hazard ratios. Since all features contribute to the individual's total hazard independently and linearly, the log coefficients can directly be interpreted in terms of feature importance when compared to the other contributing features in a multivariate survival regression. Note that the hazard ratio usually refers to a unit increase in the encoded covariate. This means that normalization across multiple covariates is necessary if different hazard ratios are compared.

Other Approaches

Extending on the previously mentioned ideas, several other survival estimators have been developed. The accelerated failure time model [252] extends the influence of covariates to accelerated hazard over time instead of only providing a constant factor as in the CoxPH model. Decision-based methods like gradient boosting techniques (such as `XGBoost` [44]), `Random Survival Forest` networks [116] (based on `Random Forests` [31]) or Additive survival least-squares support vector machines [24] became popular approaches. Additionally, hybrid ideas exist that use a neural network model to estimate, the parameters of e.g. Weibull distributions

or accelerated failure time models [26, 179].

However, the performance of these models may drop on datasets with high censoring rates [3]. Furthermore, the dynamic nature DL models that can handle multiple types of input data by, for example, combining tabular patient information with diagnostic images, have additional appeal for multimodal approaches.

DL-based Approaches

Since this thesis develops a survival prediction model based on DL, several related approaches with different levels of complexity are presented.

DeepSurv Since neural networks rather learn the underlying functional dependencies than assuming them, they permit the discovery of non-linear dependencies between the survival model's covariates [126]. DeepSurv retains the partial likelihood function so that all predicted survival curves have the same shape with a different scaling factor, as in the CoxPH model. In other words, the linear encoding of the predictor is replaced by $g(\boldsymbol{x}_i) = \text{MLP}(\boldsymbol{x}_i)$ that produces a similar scalar output. The network is then similarly trained on a partial likelihood approach with L2 regularization. These modifications of the CoxPH model allow for non-linear combinations between the input covariates. The authors show in simulated and real settings that their approach outperforms the aforementioned method with increasingly complex relationships between the covariates.

CoxTime In contrast to DeepSurv, the authors of CoxTime [136] introduce a time-variant degree of freedom in the prediction by adding a scalar point in time t directly into the MLP along the covariates as $g(t, \boldsymbol{x}_i)$. This relaxes the proportionality constraint and allows for potentially crossing survival curves over time at the cost of higher computational complexity in the objective function. In contrast to the likelihood function in DeepSurv, a partial log-likelihood objective function

$$\mathcal{L}_{\text{CoxTime}} = \frac{1}{n_{d_i=1}} \sum_{i:d_i=1} \log \left(\sum_{j:z_j \geq z_i} \exp\big(g(z_i, \boldsymbol{x}_j) - g(z_i, \boldsymbol{x}_i)\big) \right) \tag{2.32}$$

is used where $n_{d_i=1}$ is the number of uncensored individuals in the population. Since $g(z_i, \boldsymbol{x}_j)$ is potentially different for every i, this loss has a complexity of $\mathcal{O}(n^2)$ and is now only computed on a subset rather than the entire population to improve computational performance. This extension to a prediction of potentially crossing survival curves also motivates the use of time-dependent metrics like C-index-td and IBrS that are also used in this thesis. The approach is evaluated on multiple datasets including a churn prediction dataset called KKBox [107]. The authors utilize the aforementioned algorithm to cluster the predicted survival curves and gain additional insights into subgroups of customers. They can, e.g., show that one subgroup of customers is lost at the end of each month, while another shows a large drop after the initial subscription period. These findings motivate the time-dependent modeling of survival data that would not have been found by relying on the PH assumption.

DRSA Another level of complexity is added by the introduction of recurrent decoder networks [88] described in detail for DRSA [195] that introduce a time-dependency in the discrete output predictions of the model. The authors abandon the log-likelihood loss and instead compute the conditional probabilities of the event over time by defining the discrete hazard rates described in eq. (2.7). The authors evaluate the discrete hazard rate predictions before and at the event time for each individual. Furthermore, the index $l \in \{1, \ldots, L\}$ of the current time step is an additional input feature to the encoder following CoxTime to provide a time-dependency into the encoder structure. Since the model predicts hazard rates h_l for each discrete time point t_l,

the resulting discrete survival curve prediction for an individual i is afterwards obtained by

$$\hat{S}(t_l|x_i) = \prod_{j=1}^{l}(1 - h_j). \tag{2.33}$$

The general idea of this work's objective function is to maximize the hazard rate h_{l_i} that includes the event time for the individual while punishing all hazard rates prior to this interval that are greater than 0 while also punishing the predicted survival curve of a censored individual that is smaller than 1 before the censoring time. The objective function is discussed in additional detail in chapter 4.

Kamran The survival prediction model from [123], called KAM in this work, builds upon the previously mentioned ideas by utilizing the architecture of DRSA, but implements a novel loss function that emphasizes the proper calibration of the model. Also, in training it is evaluated directly on the survival curve $\hat{S}(t_l|x_i)$, instead of the hazard rates h_l. Since this thesis extends this work, the loss is explained in detail in sec. 4.2.3.

Other DL models exist but are not further evaluated in this work. Firstly, `DeepHit` [141] considers not only one type, but competing risks for a NN-based survival model. It combines a loss approach based on the maximum likelihood estimator with a regularizer on the order of events. The successor model `Dynamic-DeepHit` [140] extends the approach to include dynamic patient data free of statistical restrictions. As an alternative approach for better calibrated survival models, Goldstein et al. [91] developed `X-CAL`, an explicit calibration approach that utilizes ideas from D-calibration [95] by creating a differentiable version that can be included directly into the model's objective function.

2.5.2 Electronic Health Records

The following section discusses recent advancements in EHRs facilitated by DL models emphasizing their impact on usability despite missing standardization and data sparsity.

Representation Learning

Data representation of sparse and potential high dimensional EHRs is an important problem in itself. A common approach is based on `word2vec` [163] that originates in natural language processing (NLP) to group words that have similar context closer together in an embedding space. A similar approach was used for EHRs in `DeepPatient` [164] that use stacked denoising autoencoders (AEs) that are trained by masking the original input EHR data. An AE is an architecture type of NN that tries to find a robust and dense representation in an unsupervised manner for high dimensional input data in a lower-dimensional latent space that is usually the bottleneck layer of an encoder-decoder network structure [16]. It was shown that a detailed investigation of the latent space generated by an AE combined with task-specific predictors can lead to a more effective and robust model for several prediction tasks [248]. Another example called `Patient2Vec` is developed based on attention blocks that are fed to a gated recurrent unit [46] to encode longitudinal EHR data.

Longitudinal Health Records

Due to the sequential data structure of dynamic parts of EHRs, Recurrent Neural Networks (RNN, especially Long Short-Term Memory networks (LSTMs) [106]) are a popular choice in developments. [150] proposed to take into account not only the sequences of medical events, but also the inter-event time. This work uses LSTMs to embed the patient's medical history that accounts for non-equidistant times between patient records. Another approach was presented by [20]. They account for the chronological component that weighs events not only in terms of order but the time difference by modifying the classical LSTM structure to time aware LSTMs

called T-LSTM. They introduced a time decay function that affects the influence of a medical record on the long and short term memory. Another embedding approach [148], that takes the temporal dimension into account, utilizes positional encoding alongside the clinical event information as input for a transformer architecture similarly to the BERT model [62]. Further, [259] integrates static patient information like demographics or other time-independent patient properties, with dynamic information that changes over time like disease specific therapies. Their work combines an LSTM [106] for the dynamic patient information with an MLP for static features that are concatenated to generate a full representation of a patient's EHR to predict therapy decision in patients with metastatic breast cancer. Similar approaches are used to predict an individual patient's trajectory after kidney transplantation [75] or heart failure prediction [164] with an AUC of 0.845.

Interpretability

Furthermore, an AI guided system should not be a black box for the user especially in the medical context. It must rather be interpretable and explainable. An AI-system that is unable to provide adequate information for a clinician of how the individual result was retrieved will not be accepted in real-world patient treatment [231]. This is why recent development emphasizes the importance of interpretable and explainable DL models that work on this problem [9]. With explainable AI, the physicians and patients have the ability to comprehend the AI's decision and also explore the input factors that led to the individual results. An application of the idea of clinical interpretability is called RETAIN [47]. The model scored an AUC of 0.87 on heart failure diagnosis prediction with the benefit of interpretability of the results. The model was further optimized with improved interpretability by [137] with the RetainVis model and has also been combined with medical ontology mappings to improve AUC on heart failure diagnosis to 0.9. One important DL network architecture regarding this problem is the attention mechanism [238] that enables a specific highlighting of what inputs to a NN provide the most significance for the output prediction. Other recent improvements in NLP like the transformer architecture that introduced BERT [62] can also be applied to EHRs as demonstrated in BEHRT [147]. This model predicts future patient events with state-of-the-art accuracy that scores an AUROC of 0.95 to predict future diseases at the next visit of a patient.

Finally, [138] uses convolutional AEs (in time) to create data representations for unsupervised disease sub-typing e.g. in Parkinson's disease. They propose a DL architecture with a convolutional neural network (CNN) and AE that provides an unsupervised representation of a patient's individual longitudinal EHR that may be used as a basis to answer more specific tasks on EHRs via fine-tuning. This may be a starting point to create a more universal representation of EHRs similar to BERT [62] or GPT-based [34] models for NLP and image-based networks such as ResNet [102].

2.5.3 Computational Histopathology

This section focuses on used or related SotA approaches that are relevant for this work. Especially in the context of cancer detection and survival prediction for PCa histopathology images.

Image Encoding

The first step that is usually done when images are analyzed by DL techniques is a transformation of those images into latent representations that can be further processed regarding individual tasks like classification. Common encoding networks used in histopathology [40, 76, 243] are based on convolutional neural networks (CNNs) like, among others, ResNet18 [102] or EfficientNet [226]. A more recent development uses transformer architectures [103, 212] that originated in natural language processing [62, 238]. Since the models developed in this thesis, namely CI and PCAI (see chapters 6 and 7 respectively), use an EfficientNet encoder network, it is explained in additional detail below. Moreover, it is common to retrain those networks in a transfer-learning approach for

specific tasks that often provide smaller data sets as for example in medical imaging [19, 225, 239]. Common datasets for pre-training are, among others, `ImageNet` [61] or `CIFAR` [133]. As an example, `ImageNet` [61] provides over 14 million training images with one thousand different possible classification labels for natural images. Other pre-training methods like self-supervised learning with contrastive learning [50] exist but are not further considered in this thesis.

EfficientNet A CNN-based architecture following ideas from `MobileNetV2` [203] while taking computational efficiency into account is called `EfficientNet` [226]. While other works like `ResNet` [102] focus more complex models on adding additional layers, the main idea of `EfficientNet` is the introduction of a scaling coefficient ψ controlling the neural network's width (number of channels per convolutional block), depth (total number of layers) and resolution (for the processed input image) jointly. This idea leads to models with different levels of complexity that range from `EfficientNet-b0` (approximately 5 million parameters) to `EfficientNet-b7` with over 66 million parameters. Depending on factors like memory constraints or inference speed, this approach allows an easier selection of the most suitable model complexity. The core building block of this architecture are MBConv layers [203]. This architecture, that is also called an inverted residual block, performs a 1x1 convolution that is followed by a 3x3 depth-wise convolution for each input channel. Afterwards, another 1x1 convolution reduces the number of channels again so that a residual connection can add the input to the resulting output. This approach inverts the common idea of adding a residual connection of two wider blocks and instead connects the bottleneck layers. Moreover, squeeze-and-excitation optimization [108] is additionally used before the last 1x1 convolution as an additional strategy to further improve model efficiency.

Generalizability and Robustness

A common problem in DL-based medical imaging is variance between or even within datasets. Even though lots of variances in the input data are often clinically insignificant, they can lead to a large difference in model estimations [30, 181]. For computational histology, these differences may manifest in different slice thicknesses, variations in staining time or using different scanners for the digitization process that result in different latent representations [218]. This work also introduces a metric to measure the distribution shift from the training data to a dataset from a different domain as the mean distance of all kernel distributions after the last convolutional layer of the CNN-based network. The problem of missing model robustness still offers different solution approaches that were, for example, analyzed in the Mitosis Domain Generalization Challenge [11] that provides WSIs from multiple organs and tumor types that were scanned with various scanners.

One common strategy increases the image variance during training that is achieved, e.g., by augmentation methods [227] where maximum performance is reached when color augmentation in the Hue, Saturation and Value (HSV) color space is combined with color normalization for the input image.

Another approach to overcome this issue is in reducing data variation on unseen datasets by data normalization. For H&E stained histopathological images, stain normalization approaches [65, 153, 236] normalize the color differences in the input images that were introduced due to variations in the staining protocol for each tissue.

Further, the information about different domains can also be included in the training process itself. Given labeled sub-datasets with varying variances as previously discussed, domain-adversarial (DA) networks try to explicitly mix samples from different domains in their latent representation. To achieve this, a second domain classification task is added to the main task of the network that is combined with a gradient reversal layer after the encoder to unlearn this domain information in the latent representations of the shared encoder. [93, 254]

Multiple Instance Learning

Due to large image sizes in computational histopathology, typical DL models do not process those images as a whole due to its impracticality mainly regarding memory usage of multiple GB per image. A common strategy divides the large input image into smaller patches of the same size that are filtered (by removing background patches) and then processed by the DL model. This patching and masking approach is also applied to the image-based DL models of this work (secs. 6.2.2 and 7.2.2). For classification tasks on WSIs, it is not possible to provide individual labels for each of the extracted patches since only a label max exist for the whole slide. As a consequence, the DL architecture needs an approach to combine those individually encoded patches and relate it to the output label of the WSI. A common strategy to achieve this is multiple-instance learning (MIL) where these individual patches are commonly called instances and the collection of all patches of a single WSI is called bag. The approach of [38] describes such an architecture for cancer detection on over 44 thousand WSIs from 15 thousand patients on tissue from PCa, basal cell carcinoma and axillary lymph node metastasis of breast cancer with high accuracy of up to 100 % sensitivity. The approach relies on the identification of the most suspicious patches using an RNN architecture. However, this approach may neglect patches that contain valuable information for the aggregated estimation. An alternative was presented in [114] that combines all individual instances of a bag using a attention-based MIL. This approach creates a weighted average of the individual patch representations based on the individual contents for each patch. The attention weights can additionally be used to create heatmaps that visualize the most important patches for a bag-level representation. This approach is also used in this work for the PCAI model as described in sec. 7.2.3.

Prostate Cancer Applications

Several DL algorithms in the context of PCa exist with several use cases. Compared to TMAs, it can be desirable to identify abnormal or non-healthy tissue regions that are of particular interest to the pathologist. Those regions can guide the practitioners to provide higher quality annotations [127]. It can also help to identify previously missed cancerous regions that need reevaluation [178]. These algorithms focus on PCa detection. Several approaches are presented in the following that try to either create segmentation masks based on the expert (uro-) pathologist annotations of this dataset or try to predict patch-wise labels for small regions of biopsies, as also used in chapter 6. Another possibility is to predict more fine-grained segmentation masks that contain the actual Gleason grades introduced in sec. 2.1.2, namely GG3-5. This approach tries to extract the more fine-grained segmentation masks of the specific GGs. Those two approaches are presented in additional detail below.

Detection Detecting PCa on biopsy images is frequently performed on the PANDA dataset that is explained in additional detail in sec. 3.2.3. Firstly, [79] evaluates an unsupervised meta-learning based approach for patch-level segmentation on this dataset and obtain an AUC of 0.79 with their unsupervised approach. Further, for performance on the whole image, a dice score of 0.432 is achieved. Note that this work does not use any pixel-level ground-truth annotations. Several alternative approaches also exist that use the GT annotation masks of expert (uro-) pathologists in the training DL-based algorithms. [113] produces WSI slide-level segmentation masks using an ensemble `U-Net` model [199] to produce an average dice coefficient of 0.891 for stroma, benign, and cancerous tissue segmentation.

Moreover, the following publications show algorithms for patch-wise cancer detection algorithms. [178] analyzes a similar approach on 3050 internal H&E stained prostate biopsies that were validated on an external dataset with 1627 H&E stained images that were digitized by another scanner. They report AUCs of 0.997 for the internal-, and 0.991 for their external cohort. [120] shows a pipeline that follows the idea of extracting labels in a more general sense of transfer learning techniques called `ChampKit`. As one of many use cases, this work evaluates their algorithm on the PANDA dataset and roughly formulates the same objective in classifying patches as either

benign, GG3 or GG4-5 as their classes of interest that is used to evaluate different patch-based classification models with and without pre-training on different methods. They report patch-level classification AUROCs of up to 0.921 (on GG4-5) for an `ImageNet` pre-trained `ResNet18` [102] model. They find that the pre-trained models perform significantly better on this task compared to those that were trained from scratch, especially regarding false negative rates.

Automated Gleason Grading For the emulation of human assigned GG, [170] develops a DL-based Gleason classification algorithm for biopsies that was trained on 1,226 biopsy WSIs and evaluated on an independent validation dataset of 331 slides. Their algorithm uses a two step approach to first generate a CNN-based GG region classifications for the whole biopsy image that is then postprocessed by a k-nearest neighbor classifier in terms of risk prediction. They claim higher consistency as well as diagnostic accuracy of 0.7 compared to human annotators (with a mean accuracy of 0.61) by comparing the individual predictions with a reference derived from "specialist pathologists". In addition, they show that risk stratification works well using their algorithm compared to the predictions of other pathologists, as well as the reference standard regarding disease progression.

A similar algorithm [36] automatically generates GG annotations for PCa biopsies based on data from RAD and compared to a "reference standard" that was made by three expert (uro-) pathologists on a total of 5759 biopsies from 1243 patients. The algorithm scores a quadratic Cohen's kappa of 0.918 and therefore the 5th highest when compared to 15 expert gradings.

Similarly, [222] developed an algorithm with "clinically acceptable accuracy for PCa detection, localization and Gleason grading". A total of 6,682 PCa related biopsies were digitized from 976 patients. The results claim an AUROC of 0.997 in the identification of malignant compared to benign biopsy WSIs. When comparing the automated Gleason grades to expert pathologists, a mean pairwise kappa of 0.62 was achieved that was comparable to the scores of the pathologists.

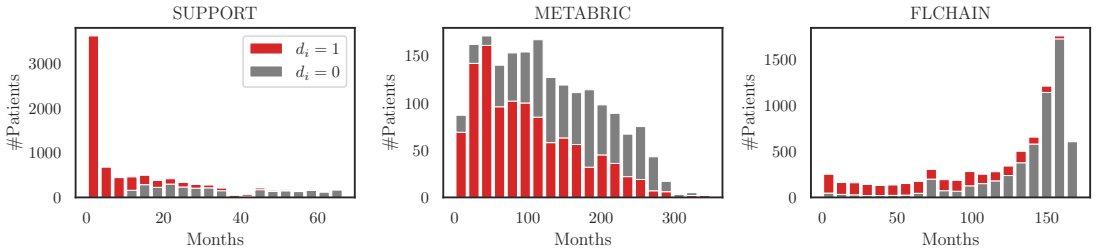
Another approach [10] designed a related algorithm and trains it on 641 patients that was afterwards evaluated on an independent test dataset of 245 patients with annotations from two pathologists. The Cohen's kappa statistic of the algorithm compared to the two pathologists were 0.71 and 0.75 (compared to 0.71 between the pathologists).

In another example [71], a Gleason grading system is compared to 23 international experts on 90 cases. They show the large variability among those experts saying that 41 % fail to reach a consensus in Gleason grading (defined as at least two thirds of expert and AI annotation agreement). Among those cases without agreement, the proposed algorithm ranks sixth regarding Cohen's kappa score.

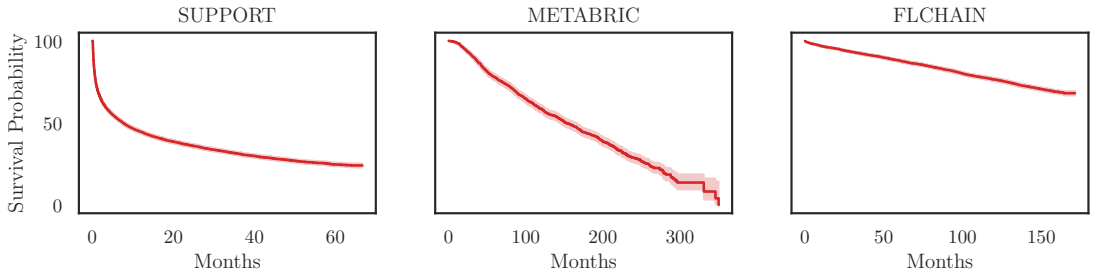
Risk and Survival Prediction in PCa In contrast to the aforementioned approaches, it might be advantageous to ignore the Gleason grading system itself and instead predict directly based on patient outcomes. In [184], the authors create a histological biomarker for PCa TMA spots using a DL trained on the classification of relapse year. They demonstrate that the biomarker can then be used as an additional input for survival prediction and show statistical significance that it provides additional information to the survival model. Furthermore, the authors cluster their latent space predictions to create several concepts that share medical features in the used patches as a means of interpretability. As another approach, [65] developed a DL-based survival prediction model based on TMA spots. With a recurrent network architecture, survival curves of individual patches are modeled and aggregated using an attention-based MIL approach similar to the strategy of this work presented in sec. 7.2.3.

3 Data

The following section describes the analyzed EHR and image-based datasets of this work. Additionally, the different patient populations are compared in terms of PCa related parameters and event-specific survival attributes. While the EHR datasets were used in chapter 4 as well as chapter 5, chapters 6 and 7 focus on the image-related TMA and biopsy datasets.



(a) Histograms of event (red, $d_i = 1$) and censoring (gray, $d_i = 0$) times.



(b) KM curves

Fig. 3.1: Event and censoring distribution for the EHR datasets.

3.1 Tabular EHR Datasets

This section briefly introduces the tabular EHR datasets that were analyzed to develop the DL-based survival model called DCS that is presented in chapter 4. Since real-world datasets may differ significantly in properties like censoring rate, survival distribution and the number of proportional features, this work includes three datasets with varying attributes to obtain a more general overview of typical survival datasets. Tab. 3.1 provides basic attributes of the three EHR survival datasets used in this work. While a significant portion of individuals have an early event time in SUPPORT, the distribution of events over time in METABRIC while FLCHAIN shows most observations towards the end as shown in Fig. 3.1. All analyzed datasets have several thousand patients (8,873 for SUPPORT, 1,904 for METABRIC and 7,874 for FLCHAIN) and a relatively low number of 14, 9 and 8 features for SUPPORT, METABRIC and FLCHAIN respectively. The censoring rate (from 32 % in SUPPORT to 72 % in FLCHAIN) and the ratio of features that fail the Cox Proportionality test (see sec. 2.5.1) varies from 25 % in FLCHAIN to 79 % in SUPPORT along with the median survival time that is only 57 days for the SUPPORT dataset compared to the maximum of 85 months in METABRIC. The three datasets are individually described in more detail in the following.

Tab. 3.1: Basic characteristics for the EHR datasets showing the number of patients, number of features, censoring rate, percent of missing data, median survival time and ratio of covariates that violate the PH assumption (non-prop.).

	#Patients	#Features	Censoring rate	Median surv.	Non-prop.
SUPPORT	8873	14	32 %	57 days	79 %
METABRIC	1904	9	42 %	85 months	56 %
FLCHAIN	7874	8	72 %	71 months	25 %

3.1.1 SUPPORT

The Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments [130] provides information on 8873 seriously ill hospitalized adults with 14 features. The features that are shown for randomly selected patients in tab. A1 contain demographic information, laboratory data like body temperature, heart- and respiration rate, number of white blood cells and different blood levels like bilirubin or albumin. The study analyzed the time of death of a patient.

3.1.2 METABRIC

This dataset from the Molecular Taxonomy of Breast Cancer International Consortium [58] is the smallest analyzed EHR dataset of this work. 1904 breast cancer patients with nine features (demographic, molecular drivers and therapies) were collected to predict long-term clinical outcome of breast cancer. Exemplary patients are shown in tab. A2. The features of this dataset are breast cancer specific markers as well as treatment information (hormone-, radiation- or chemotherapy) along with the patient's age.

3.1.3 FLCHAIN

The assay of serum free light chain (FLCHAIN) from 1995 analyzes the impact of serum free light chain levels (FLC) regarding patient death. A general population study with eight features was conducted on 7,874 individuals mainly from residents of Olmsted County aged 50 or older. The features include patient age, sex, sample group and year, kappa and lambda portion of the serum FLC, serum creatinine level and whether the individual was diagnosed with monoclonal gammopathy. Some sample patients along with the sampled features of this dataset are depicted in tab. A3. The study concluded that elevated FLC levels are associated with higher death rates. More detailed explanations about the study can be found in [66, 228]. This dataset can be considered as the dataset that has the highest ratio of proportional features with only 25 % of the features violating the Cox Proportionality assumption.

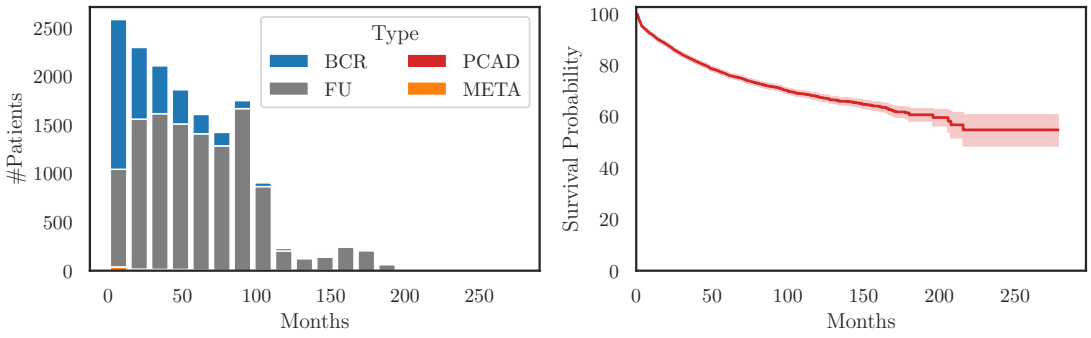


Fig. 3.2: Histogram and KM curve of the MK dataset. BCR: Biochemical recurrence, PCAD: PCa related death, FU: lost to follow-up, META: Found metastases.

3.1.4 Martiniklinik

The main EHR dataset of this work is called MK. It contains information of 16,953 patients that were extracted from the Martiniklinik database with records ranging from 1992 to 2018. All patients were registered within a prospective ethics committee-approved database after informed consent. The patients in this dataset exclusively received RP treatment after being diagnosed with PCa. Fig. 3.2 provides an overview of the patient's survival distribution showing the high censoring rate of the dataset of 78.4 %. A large variety of 86 individual features were collected that range from demographic information such as the patient's age to parameters that were collected during the last PCa biopsy of the patient or parameters that describe the outcome of RP (like positive resection margin) along with tumor-specific characteristics such as seminal vesicle invasion or the number of resected lymph nodes during the operation. Note that the patients at the Martiniklinik are almost exclusively treated with RP. Also, the patients are usually transferred from other urologists which means that the patient documentation before RP is expected to be missing or at least more heterogeneous. Furthermore, tab. 3.2 shows individual exemplary patients of the dataset with some provided covariates.

Tab. 3.2: Sample patients of the MK dataset with exemplary features. The endpoint information (duration z_i from RP until relapse ($d_i = 1$) or censoring ($d_i = 0$)) and some features are shown: Patient age, PSA level at RP, prostate volume, T-stage (path), GG 3-5 volume in mL obtained from the whole removed prostate after RP.

z_i [days]	d_i	age	psa [ng/ml]	pros_vol [ml]	t_stage	GG3 [ml]	GG4 [ml]	GG5 [ml]
1933	1	61	5.2	10	pT3b	2.23	3.53	0.12
1717	0	70	11.0	55	pT2c	7.77	1.10	0.27
1123	0	68	7.1	10	pT3b	0.44	0.69	0.02
3349	0	46	5.2	20	pT2c	5.57	0.29	0.00
				...				
3674	0	74	12.0	40	pT3b	0.50	5.30	3.80
2946	0	63	4.8	20	pT2c	1.80	0.10	0.00

Feature Overview

The following section provides an overview of the most important feature types that were selected in the dataset. The specific analysis of the individual and combined features can be found in chapter 5.

Demographics Information like the age of the patient at time of RP in years, body mass index and others are provided for a general overview of the patient.

Endpoint After RP, the patients are further observed to ensure cancer remission. Minimum information is given in the form of a yearly questionnaire that ensures that no PCa related symptoms (re-)appeared. When any suspicions are found, further actions like a PSA test is performed where a positive result with $> 0.2\,\mathrm{ng/mL}$ is considered as BCR. Additionally, this work defines the event-of-interest as any kind of relapse for the individual. This means that either BCR, found metastases, PCa related death or any additional, unplanned treatment are included in the event definition. If the patient did not show any of those signs on last contact, he is considered to be lost to follow-up (FU) without cancer progression, thus considered censored in this work. Further, all individuals with less than 6 months are removed from the dataset.

PSA Level The PSA level in ng/mL at time of RP that measures the antigen in the blood as introduced in sec. 2.1.1 is also part of this dataset. This work conducts additional experiments on PSA density [210] that normalizes by prostate volume, or including the date where the PSA value was obtained relative to the operation.

Prostate and cancer characteristics Attributes of the extracted prostate with the tumor along with the surrounding tissue of the patient after RP. This includes the prostate and tumor volume. Important factors of the tumor like seminal vesicle invasion, lymph node invasion, capsular extension and surgical resection margin status are also obtained and illustrated in fig. 3.3.

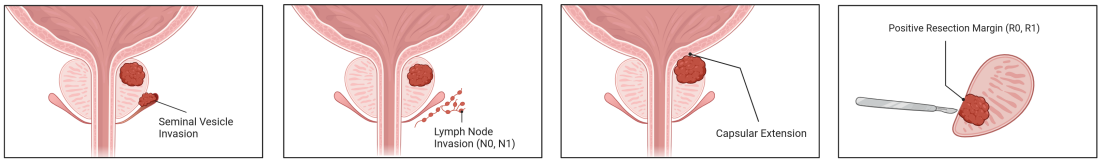


Fig. 3.3: Provided tumor characteristics of the analyzed MK dataset showing seminal vesicle invasion, lymph node invasion, capsular extension, and positive resection margin. Adapted from "Prostate Cancer Risk Assessment", by BioRender.com (2024). Retrieved from https://app.biorender.com/biorender-templates.

Gleason Grades An important characteristic in analyzing PCa severity is Gleason grading as explained in detail in sec. 2.1.2. The given dataset contains clinical (from the last biopsy) and pathological (based on the whole prostate after RP) GG in multiple formats that were graded by a specialized pathologist. Firstly, the primary and secondary grades are provided. Note that by definition, for clinical GG the most and worst grade is reported as primary and secondary while for pathological GG, the most and the second most is considered as the primary and secondary GG respectively. This documentation is extended for several patients by also including the tertiary GG that has shown to include additional information for relapse prediction [98]. Moreover, the pathological GG is also extended for a subset of patients to contain the total volume for GG 3-5 measured in mL for the whole removed prostate. It is to be expected that pathological GG provides a prognostic value since it is based on the whole removed prostate instead of relying on a biopsy that might not be representative of the whole tumor or prostate.

TNM Staging Derived from the most important features discussed above, TNM staging [172] was obtained for the individuals in this dataset that was described in additional detail in sec. 2.1.1. For this dataset, clinical and pathological staging information is provided.

Treatment Information The dataset contains additional information if patients received any form of additional treatment along RP, namely chemotherapy, RT, HT or others. This information can be divided into treatments prior to RP (neoadjuvant) and adjuvant treatments that were applied (but planned) after RP. Also, unplanned treatments after RP (salvage treatments) are documented in this dataset. The thesis considers salvage treatments as disease progression at time of treatment.

3.2 Image Datasets

The second part of this work deals with PCa datasets that contain H&E stained images of biopsy cores or TMA spots that were used for GG (see sec. 2.1.2) in clinical practice or research settings. This thesis deals with TMA datasets from the UKE, NYU and JHU and biopsy data from PANDA, UPP and MMX. Sec. 3.2.4 provides a brief comparison of the different datasets at the end of the section.

3.2.1 UKE Dataset

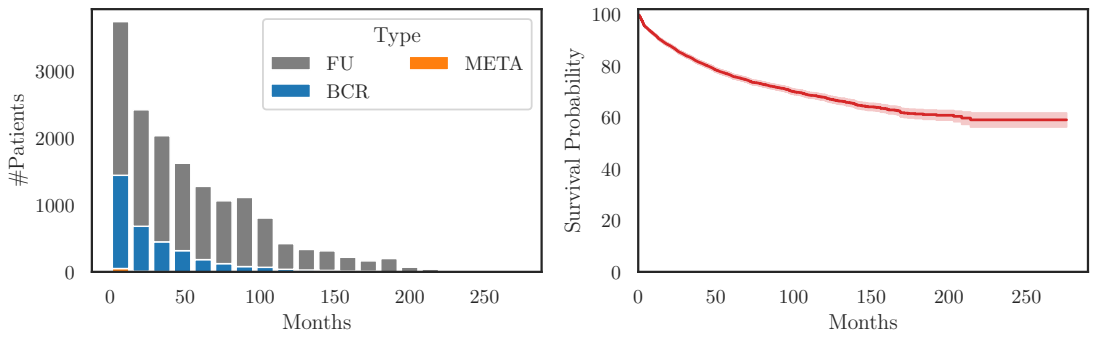


Fig. 3.4: Event distribution of the UKE dataset. While (left) shows a histogram of event types over time, (right) shows the corresponding KM curve. FU: lost to follow-up, BCR: Biochemical recurrence, META: Found metastases.

The main image related dataset of this work is the UKE dataset. The majority of the patient cohort overlaps with the previously presented MK dataset (see sec. 3.1.4), but the focus of this dataset are the TMA spots that are provided along with patient survival information instead of tabular data from the EHRs. Similarly to MK, it contains PCa patients that underwent RP, but no detailed tabular patient information is provided. Since both datasets are pseudonymized, these two datasets cannot be merged. This dataset was provided by the Pathology institute of the UKE and contains 17,700 patients who underwent RP between 1992 and 2014 aged 63.8 ± 6.4 years at the UKE with a maximum FU duration of almost 23 years. The cohort's observed median PSA level at the point of RP is 6.9 ng/mL (Interquartile range (IQR) of 4.8 ng/mL to 10.5 ng/mL).

RP patients receive a regular FU examination for PCa [207] where PSA levels are measured. This work defines recurrence in this dataset as a postoperative PSA level of 0.2 ng/mL and increasing at consecutive measurements. This work defines the individual event label to each patient by combining BCR, additional unplanned therapy, metastasis or PCa-related death as an event-of-interest with a duration from the date of RP to the first of the previously mentioned events. Patients without any record of the previous events are considered censored at the last known

FU date. Regarding the event-of-interest definition, the median survival duration is 19.2 months and the median FU duration is 48.5 months. The event distribution of the dataset is 12,444 cases (78.2 %) lost to FU (censored), 3,392 cases (21.1 %) BCR and 69 cases (0.4 %) metastasis as first recorded event after RP. An overview of the events for this dataset is depicted in fig. 3.4.

Furthermore, ISUP grades are derived from the pathological GG and reported for nearly all patients with a distribution of 510 cases (2.9 %) without cancer and 3,036, 10,228, 2,888, 190, 830 for ISUP grades 1 to 5 respectively. For T-stage (path), 11,509 cases (65.1 %) were staged with pT2, 6,092 cases (34.4 %) with pT3 and 85 cases (0.5 %) with pT4 while 2 cases were assigned to pT0 and pT1 respectively.

UKEhv Sub-Datasets

For the 17,700 patients, a large variety of 69,251 images of TMA spots were obtained from different digitization protocols, which represent the foundation for building a robust prediction model in this work. GGs were assigned by examining the whole prostate after RP for every individual patient. Extending the standard digitization protocol, UKEhv provides additional sub-datasets that offer more variance to the digitized TMA images as depicted in fig. 3.7. The standard slicing and staining protocol was extended to create variations in appearance for the scanned TMAs. An overview for the different number of patients and images for each sub-dataset can be found in fig. 3.5. Note that these datasets consist only of additional images, not spot-specific metadata. For evaluation, all of these datasets use the same patient-level information as the UKE.first dataset that follows the standard protocol. Also, a patient may be included in multiple sub-datasets. The TMA spots are scanned in bulk on 39 different blocks, that received the same staining respectively, thus should be similar regarding staining color and brightness.

	UKE	first	8,123 Images	thin	1,602 Images		sealed
Type	TMA		Primary spot		1 µm Thickness	Patients	826
Patients	8,983	second	7,156 Images	thick	1,574 Images	Images	4104
Images	32,333		Secondary spot		10 µm Thickness	Scanner	A
Scanner	Aperio	scanner	8,114 Images	long	1,667 Images	Thickness	2.5 µm
Thickness	2.5 µm		3DHistech Scanner		Staining 40:00 H, 10:00 E		
Staining	4:00 H, 1:20 E						

Fig. 3.5: Overview of the sub-datasets in UKEhv that extend the standard protocol and what attribute of the sub-dataset varies.

UKE.first The main sub-dataset UKE.first provides 8,123 TMA spots following the standard digitization protocol of the UKE. It contains a selected TMA spot that is characteristic for the disease progression of the individual patient. Tissue samples were sliced at 2.5 µm, stained with H&E for 4 mins and 1.33 mins, respectively, and digitized by an Aperio scanner under a magnification of 40x (0.25 µm/pixel). An exemplary block and individual TMAs are shown in fig. 3.6.

UKE.second A second sub-dataset of 7,156 images called UKE.second contains a different TMA spot that was obtained from another part of the cancerous area of the prostate for the same patient.

UKE.scanner This collection contains 8,114 images of spots that were scanned by a different scanner brand, namely 3DHistech scanner at 80x magnification (0.125 µm/pixel).

UKE.thin A collection of thinner cut spots of 1 µm instead of 2.5 µm from 1,602 patients.

UKE.thick Thicker cut spots at 10 µm instead of 2.5 µm from 1,574 patients.

UKE.long This sub-dataset contains longer stained TMAs with 40 min hematoxylin and 10 min eosin staining from 1,667 patients.

UKE.sealed The last dataset with 827 patients, 4,097 images, and a maximum of 10 images per patient is also included. In order to be able to test the developed algorithms of chapter 7 on a completely unseen internal dataset, only the raw images were available for this work without any additional information. Image-wise predictions were evaluated by an external scientist against GG and GIQ grading from the pathology department. This means that this is the only TMA dataset where this work's predictive performance can be compared to a human annotator, since both utilize the same amount of available images and disregard any additional information.

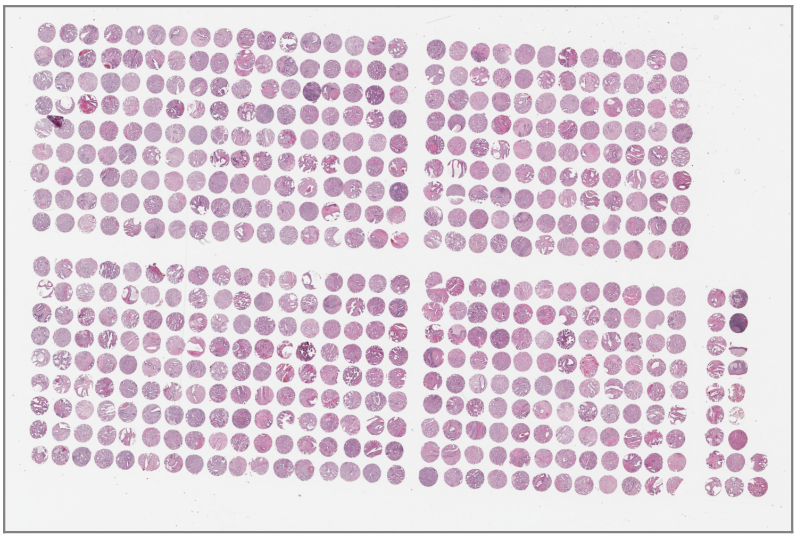
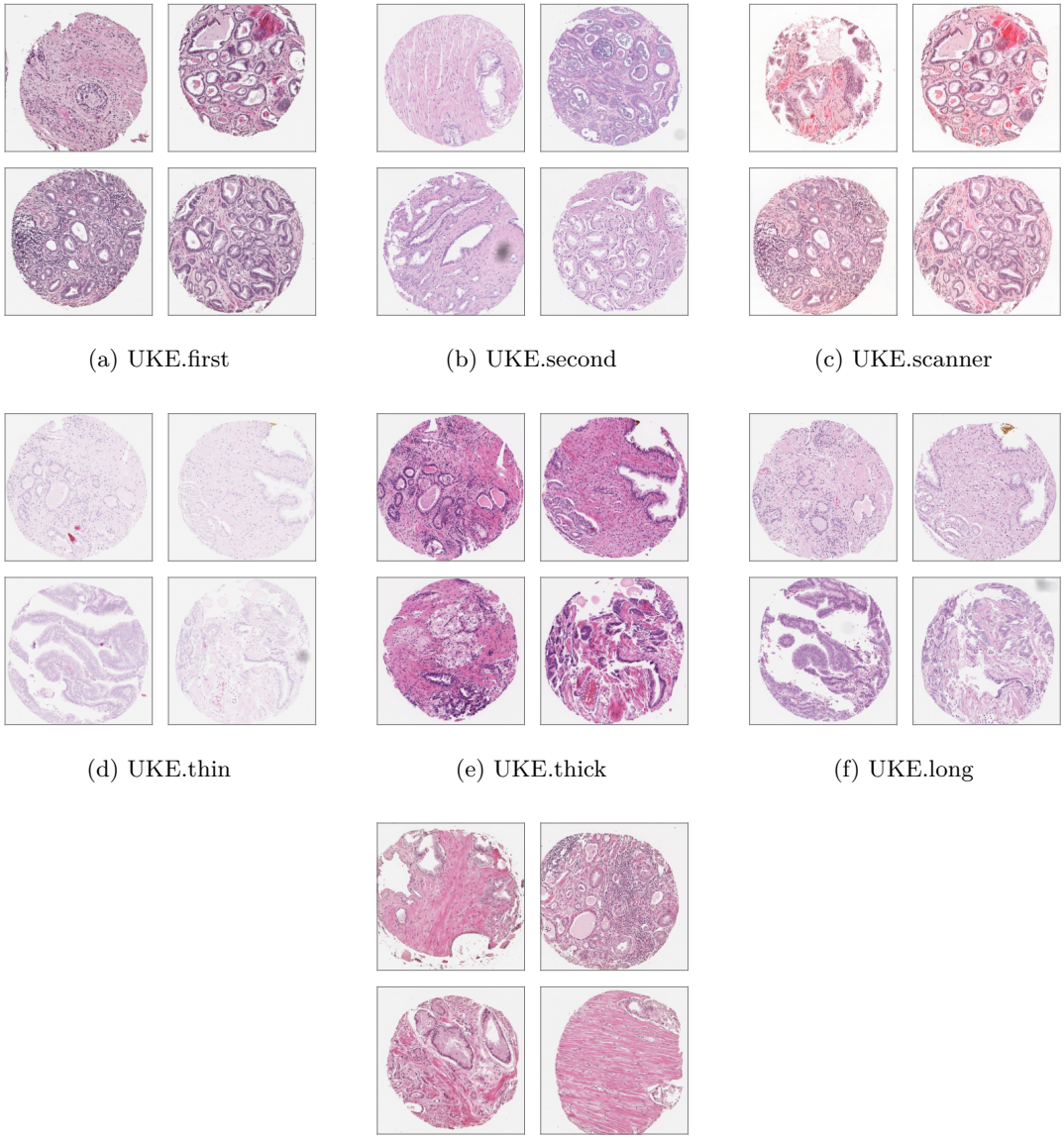


Fig. 3.6: UKE.first exemplary TMA block. The block with 110,000×73,000 pixels is cut into 542 individual images with 1,800×1,800 pixels where every individual image shows only a single TMA spot.



(a) UKE.first (b) UKE.second (c) UKE.scanner

(d) UKE.thin (e) UKE.thick (f) UKE.long

(g) UKE.sealed

Fig. 3.7: Exemplary TMAs of the UKEhv sub-datasets with visible differences in appearance.

3.2.2 Prostate Cancer Biorepository Network

This work presents three datasets from the Prostate Cancer Biorepository Network (PCBN) [160].[8] Specifically, this thesis utilizes three datasets from two centers described in detail below.

New York University Dataset

The cohort from the New York University (NYU) contains a total of 204 unique patients arranged in four TMAs. This work excludes patients that received any adjuvant therapy from this dataset. Additionally, the blocks were digitized using an Aperio scanner with a magnification of 20x (0.5 µm per pixel). The individual spots were sliced at 5 µm. The TMA blocks are cut into individual images using QuPath showing only a single spot with a size of 1817x1817 pixels and a diameter of 0.6 mm. [17] Spots showing non-neoplastic tissue were excluded. After preprocessing and filtering, this work integrated 515 images of 161 patients with a median of 3 images per patient.

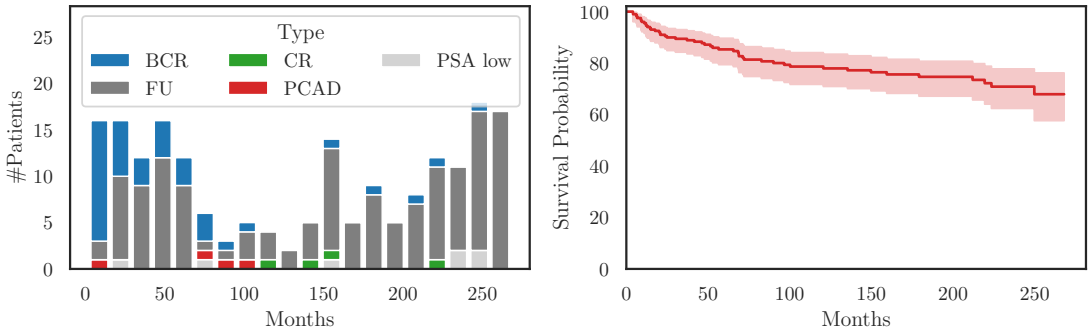


Fig. 3.8: Event distribution of the NYU dataset. BCR: Biochemical recurrence, PCAD: PCa related death, FU: Lost to follow-up, PSA low: Low but detectable PSA level, CR: Clinical Recurrence.

Johns Hopkins University Dataset

The TMA spots provided by the Johns Hopkins University (JHU) were extracted from two datasets called "Case Natural History of Prostate Cancer" (6 TMA blocks, 235 patients) and "Case PSA Progression" (16 TMA blocks, 726 patients). The individual spots were cut at 4 µm and scanned with a Ventana DP200[9] and a Hamamatsu NanoZoomer XP[10] scanner. In contrast to the other TMA datasets, the endpoint definition of this dataset in terms of event duration is only accessible in yearly instead of monthly granularity. These two datasets also contain rich metadata information like age, race, local recurrence, etc. that was disregarded in this work's analysis.

Moreover, the event indication for this dataset was extended to include salvage (unplanned) treatments after initial treatment as additional events, leading to a censoring rate for this dataset of under 1 %. This means that this cohort can be considered to be biased towards unhealthy individuals since it is also showing the highest ratio of M1 (37.2 %) as well as N1 (18.6 %) patients among the TMA spot datasets.

For integration, the 22 TMA blocks were cut into individual spot images of size 3200x3200 pixels at a magnification of 40x (0.25 µm/pixel) using QuPath. After preprocessing and excluding spots showing control tissue, this work integrated 3,575 TMA spot images that show prostatic

[8] https://www.prostatebiorepository.org/
[9] https://diagnostics.roche.com/global/en/products/instruments/ventana-dp-200-ins-6320.html
[10] https://nanozoomer.hamamatsu.com/jp/en.html

adenocarcinoma from 879 patients, with a median of four images per patient.

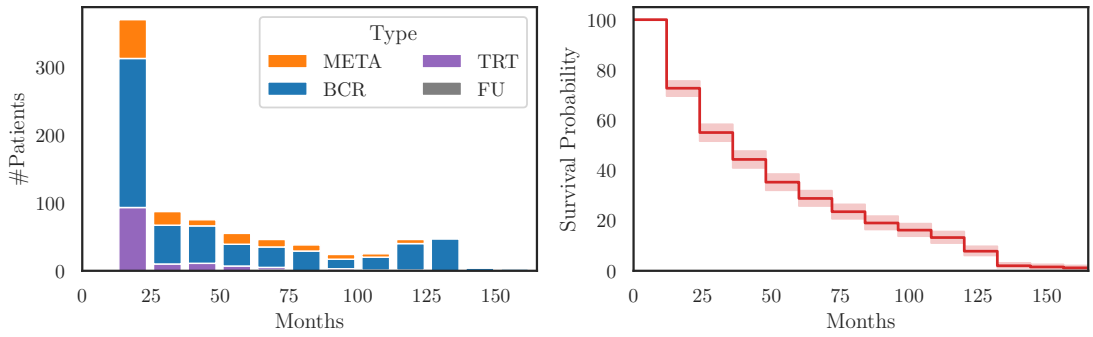


Fig. 3.9: Event distribution of the JHU dataset shown as a histogram over the observed event times (left) and a corresponding KM curve (right). META: found metastases, TRT: any additional treatment, BCR: Biochemical recurrence, FU: lost to follow-up.

3.2.3 Biopsy Datasets

Lastly, this work integrates several datasets that do not show PCa TMA spots, but biopsies. In contrast to TMA spots that are solely obtained after RP, biopsies are collected earlier, usually right after other suspicious findings like a positive DRE. As a result, the patients in these datasets are more diverse since RP is now only one of multiple initial treatment possibilities described in fig. 2.4. Further, the endpoint definition for those patients also changes for the datasets as well. While the PANDA dataset is used for a patch-wise cancer prediction task, UPP and MMX do not use the time from RP to any form of relapse that was used in the TMA spot datasets, but the time from biopsy up to the same events of interest.

Another difference of the biopsy datasets lies in the size of individual images. Image widths and heights vary, but can be hundreds of thousands of pixels for the long side of a biopsy compared to only up to 4,000 pixels in the TMA datasets. This results in up to billions of pixels per slide and makes the preselection of relevant image regions more important for these datasets. A comparison between the image datasets that shows the number of pixels per slide is shown in fig. 3.15.

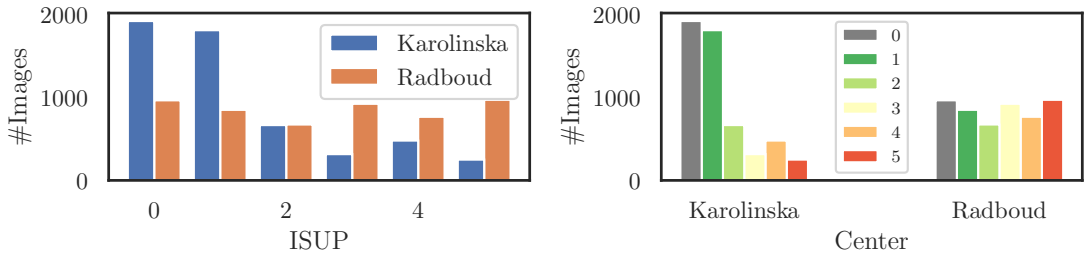
PANDA Dataset

The PANDA dataset contains biopsy slides and GT tissue and cancer related masks of expert (uro-) pathologists. PANDA is one of the largest publicly available WSI datasets for PCa Gleason grading in the world. It was published in the **P**rostate c**AN**cer gra**D**e **A**ssessment challenge and a part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2020. An overview of the challenge results of over 1000 teams was published in [35]. Since the challenge has been completed, the data is publicly available[11] along with a good introduction to the dataset in the official repository[12].

The main objective of the challenge is, given a WSI, to predict the ISUP for this biopsy. The publicly available data consists of 10,616 biopsy WSIs of 2,113 patients and has a total size of 411 GiB with an ISUP annotation illustrated in fig. 3.10. It can be observed that for KAR, most patients were assigned either ISUP0 or ISUP1 while the distribution is approximately equal for ISUP0-5 for RAD.

In addition, segmentation masks from the pathologist are provided for each slide that contain more information about the tissue along with the final ISUP grade for each individual WSI. Note that the test set images were graded by multiple pathologists. The data for this challenge is provided by two centers, namely the Karolinska Institute in Stockholm, Sweden (KAR, 5,456 WSIs) and the Radboud University Medical Center in Nijmengen, Netherlands (RAD, 5,160 WSIs). Note that the granularity of the segmentation masks differ by center. While both KAR and RAD provide segmentation masks for the background, healthy, and cancerous tissue areas, the RAD center additionally includes more fine-grained masks for stroma (connective tissue, non-epithelium tissue), healthy (benign) epithelium and cancerous epithelium separated by GG 3-5. It is worth noting the imperfection of the labels that are used in the dataset. The segmentation masks may contain false positive and false negative regions and the agreement upon GG and ISUP usually suffers from high inter-rater variability.

Data Format The biopsy slides are provided as `tiff` images with a maximum magnification of 20x providing a resolution of 0.486 µm/pixel in pyramidal format (see sec. 2.1.2) where additional downsampled versions of the same image are provided. The biopsy slides are large and have varying sizes. An exemplary slide of the dataset has a size of 8,960 pixel × 28,672 pixel with two additional levels of lower resolution, namely 10x (2,240 pixel × 7,168 pixel) and 5x (560 pixel × 1,792 pixel).



(a) Per ISUP grade

(b) Per center

Fig. 3.10: PANDA image-level ISUP distribution for both medical centers. While (a) compares the number of images per center for each ISUP grade, (b) shows the ISUP distribution for each center separately.

[11] https://www.panda.grand-challenge.org/data/
[12] https://www.github.com/DIAGNijmegen/panda-challenge/

Uppsala Dataset (UPP)

The UPP dataset from Uppsala, Sweden contains 2,611 unfiltered images of 440 patients from the publicly available[13] SPROB20 image dataset [246] that was enriched by patient endpoint information. Since some patients in this dataset have multiple biopsies, this work only considers biopsy images from the latest patient visit and excludes all earlier biopsies. Additionally, slides without an assigned GG or ISUP were considered as healthy and imputed. After this work's filtering steps, up to 10 images per patient from the biopsy are kept. Patients with adjuvant treatments were excluded from this dataset. Since these patients represent a nearly unfiltered cohort of PCa biopsy patients, the censoring rate of this work's event-of-interest in this dataset is, with 84 %, among the highest observed in this work as depicted in fig. 3.11.

The UPP biopsy slides were obtained from an Aperio scanner on a magnification level of 40x (0.25 µm per pixel). In total, 1,013 images of 181 patients from this dataset could be extracted for the evaluation of PCAI. Since this cohort contains patients from a pilot study for magnetic resonance imaging guided acquisition of prostate biopsies, the number of missed biopsies and their overall quality may be different, higher or lower, than it would have been if the conventional procedure had been used. Fig. 3.12 shows randomly selected. It can be observed that the biopsy images are homogeneous in terms of color.

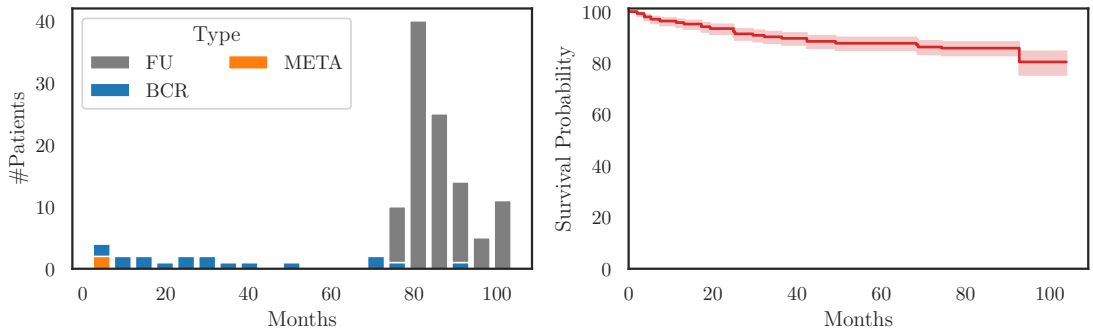


Fig. 3.11: Event distribution of the UPP dataset shown as a histogram over the observed event times (left) and a corresponding KM curve (right). FU: lost to follow-up, META: found metastases, BCR: biochemical recurrence.

Fig. 3.12: Six exemplary slides of the UPP biopsy dataset.

[13] https://datahub.aida.scilifelab.se/10.23698/aida/sprob20

Malmö Dataset (MMX)

The MMX biopsy dataset from Malmö, Sweden contributes 269 patients with 578 biopsy slides to this work with up to eight images for a single patient. The median survival and follow-up time for those patients is 38 and 106 months respectively. The events in this dataset mainly consist of either metastases or PCa related death of the patient. Notably, the patients' mean age of this dataset is 4-8 years older than patients in the other datasets. Also, these patients show the highest average PSA values combined with high variance at $19.9 \pm 44.5\,\mathrm{ng/mL}$. In contrast to the other datasets, MMX does not provide patient-level but slide-level annotations of three pathologists for a fairer comparison of our algorithm to individual humans. The three pathologists of two centers (Aachen and Uppsala) were instructed to annotate the 578 biopsy slides individually without further patient information on patient level. The images of this dataset were digitized using a Hamamatsu and Ventana scanner at 40x magnification resulting in individual slide images with a resolution of $0.23\,\mu\mathrm{m/pixel}$.

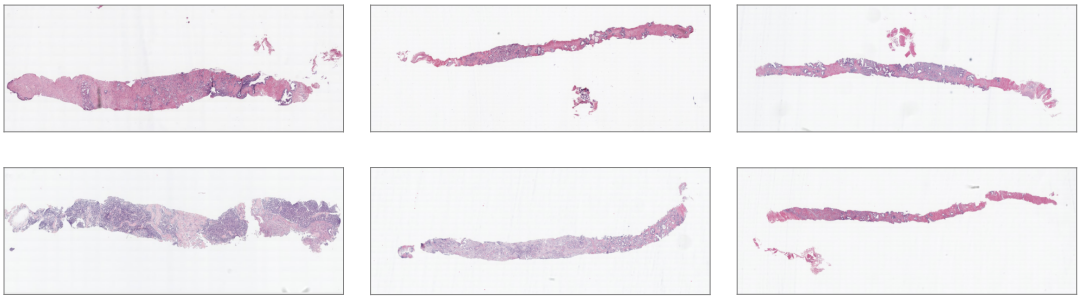


Fig. 3.13: Exemplary slides of the MMX biopsy dataset.

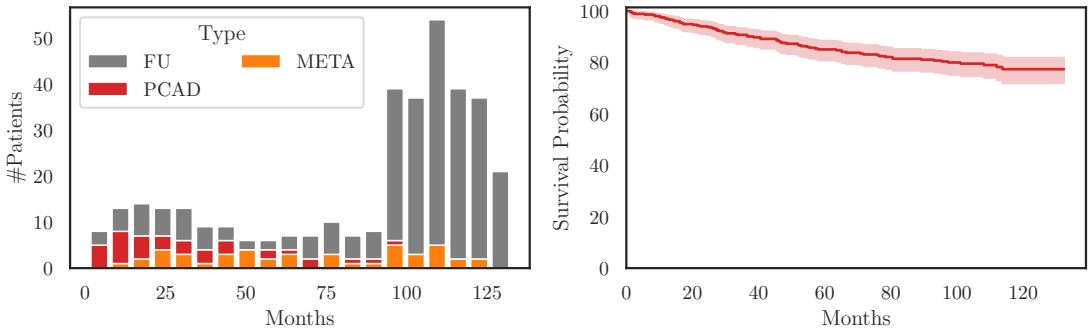


Fig. 3.14: Event distribution of the MMX dataset with a histogram over the observed event times (left) and a corresponding KM curve (right). FU: lost to follow-up, META: found metastases, PCAD: PCa related death.

3.2.4 PCa Database

This section describes the aggregation of all PCa related datasets previously discussed into one database. Also, preprocessing steps like data cleaning and filtering are applied to work with high quality datasets for the individual purposes analyzed in this thesis. This process allows EHR and image data comparison between the different datasets. The following section gives a comparative overview of these datasets in terms of EHR data and analyzes dataset differences in the provided images.

Patient Comparison

Since seven PCa related datasets are used in this work (MK, UKE, NYU, JHU, UPP, MMX, PANDA), a comparison between basic attributes of those patients is performed. Firstly, a distinction can be made regarding the type of data that is provided in those datasets. While MK provides detailed tabular patient information of each individual RP, the remaining datasets are image-related. UKE, NYU and JHU provide one or multiple TMAs per patient, also from the RP. The last three datasets, namely UPP, MMX and PANDA contain images from biopsies that were obtained prior to initial treatment.

Moreover, basic survival characteristics of these datasets can be found in tab. 3.3, In terms of dataset size, UKE and MK contain the most patients with (17,700 and 16,953 respectively) with a large overlap in patients. PANDA provides 2,113 patients while the other datasets contain 959 (JHU), 440 (UPP), 357 (MMX) and 202 (NYU) patients respectively.

Image Comparison

The datasets are divided into TMA spot datasets and biopsy datasets. TMA spot datasets are from UKE, NYU and JHU. The biopsy dataset sources are PANDA, UPP and MMX. All datasets are used for time-to-event prediction except for the PANDA dataset that is used for a patch-wise cancer prediction task for the cancer prediction model (CI) in chapter 6. General image properties among the datasets are illustrated in tab. 3.4. The images are scanned from multiple scanner vendors (Leica Aperio, 3DHistech, Ventana and Hamamatsu) with varying maximum magnification levels from 20x (0.5 µm/pixel) to 80x (0.125 µm/pixel). Additionally, the images may vary in size and resolution as shown in fig. 3.15. Further, the HSV color histograms of the datasets appear different as visualized in fig. 3.16. While JHU is shifted to bluer hues, UKE.scanner and two smaller parts of UKE.first and UPP show a redder hue distribution. Also, the distribution of UKE.thick shows the highest and UKE.thin the lowest saturation distribution respectively. For the value channel, UKE.scanner shows the brightest while JHU and UPP show the darkest value channel distribution respectively.

3.3 Additional Data Sources

Beyond datasets described in this chapter that were used in this work, there are several other related datasets for comparable applications, like in different organs or with longitudinal EHR data. A selection of these datasets that might be valuable for related projects can be found in the appendix at appendix A.

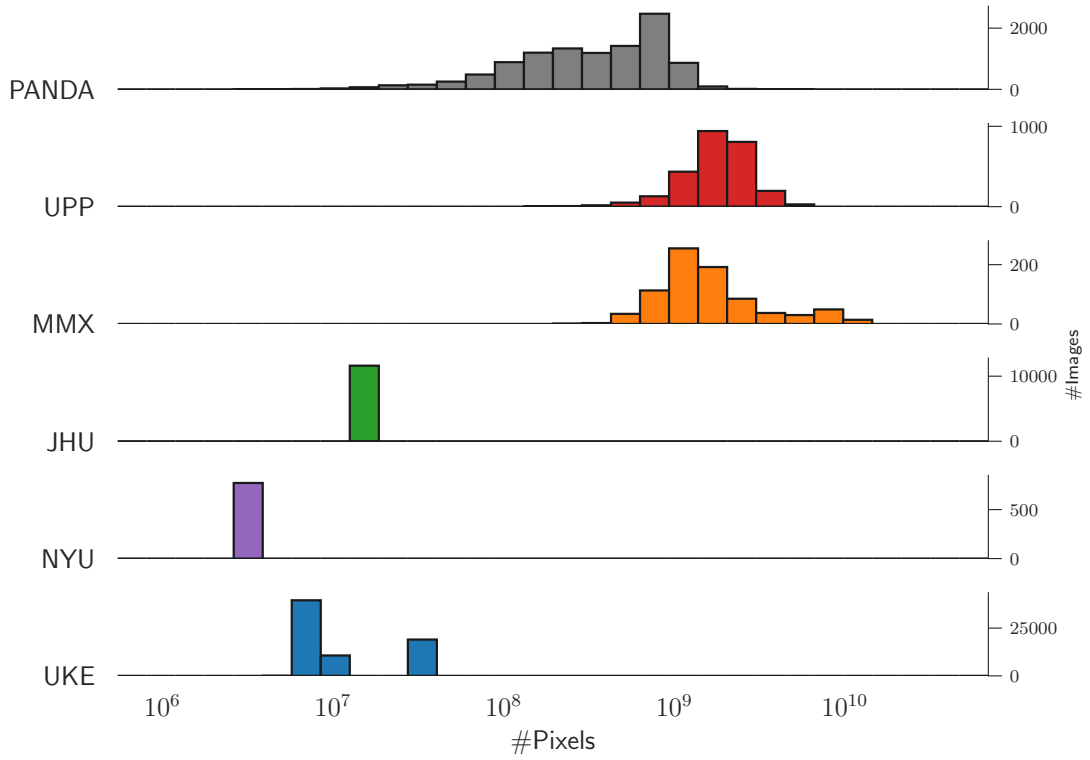


Fig. 3.15: Comparison of the number of pixels per image. A histogram for each image dataset showing the number of pixels per TMA spot or biopsy slide on a logarithmic x-axis. Note that all biopsy datasets (namely, UPP, MMX and PANDA) contain up to three magnitudes more pixels than TMA spot images also with more variance. The TMA spot datasets NYU and JHU were cut into a predefined fixed size.

Tab. 3.3: Survival characteristics for the image datasets showing the (sub-) dataset name, type: image type, number of patients n_{pat}, number of images n_{img}, median number of images per patient $\tilde{\mu}_{\text{img}}$, median survival- $\tilde{\mu}_{\text{surv}}$ and median followup $\tilde{\mu}_{\text{fu}}$ time in months as well as the censoring rate c[%]. The last row aggregates the characteristics over all datasets where the relevant data is present.

(sub-) dataset	type	n_{pat}	n_{img}	$\tilde{\mu}_{\text{img}}$	$\tilde{\mu}_{\text{surv}}$	$\tilde{\mu}_{\text{fu}}$	c[%]
UKE.first	TMA	15905	21093	1	19.2	48.5	78.2
UKE.second	TMA	15905	20510	1	19.2	48.5	78.2
UKE.scanner	TMA	16485	19008	1	20.1	48.7	78.2
UKE.thin	TMA	2603	2880	1	29.7	84.0	74.5
UKE.thick	TMA	2603	2880	1	29.7	84.0	74.5
UKE.long	TMA	2603	2880	1	29.7	84.0	74.5
UKE.sealed	TMA	-	4263	-	-	-	-
NYU	TMA	202	726	3	46.2	175.4	76.7
JHU	TMA	955	6236	6	192.0	24.0	0.3
UPP	Biopsy	440	2611	5	25.2	82.3	92.9
MMX	Biopsy	357	777	2	38.4	106.5	80.7
PANDA	Biopsy	2113	10616	-	-	-	-
total		60171	94480	1	22.3	50.0	76.4

Tab. 3.4: Image-related properties of this work's image datasets showing the dataset's tissue type (TMA=T or biopsy=B), mean ± std of the number of pixels on the long and short edge of each image, used the scanner vendor(s) (APE=Leica Aperio, 3DH=3DHistech, HAM=Hamamatsu, VEN=Ventana), mag.=maximum magnification level and the resulting physical resolution in µm/pixel.

(sub-) dataset	#pixels long edge	#pixels short edge	scanner(s)	mag.	µm/pixel
UKE.first	2900 ± 200	2900 ± 200	APE	40x	0.25
UKE.second	2900 ± 0	2900 ± 0	APE	40x	0.25
UKE.scanner	6100 ± 0	6100 ± 0	3DH	80x	0.125
UKE.thin	2900 ± 0	2900 ± 100	APE	40x	0.25
UKE.thick	2900 ± 0	2900 ± 0	APE	40x	0.25
UKE.long	2900 ± 0	2900 ± 0	APE	40x	0.25
UKE.sealed	3100 ± 200	3100 ± 200	APE	40x	0.25
NYU	1800 ± 0	1800 ± 0	APE	20x	0.5
JHU	3600 ± 0	3600 ± 0	HAM, VEN	40x	0.23
UPP	67100 ± 16300	28800 ± 8100	APE	40x	0.25
MMX	64900 ± 22000	30200 ± 17400	HAM, VEN	40x	0.23
PANDA	26100 ± 8600	15900 ± 8900	APE, 3DH, HAM	20x	0.486

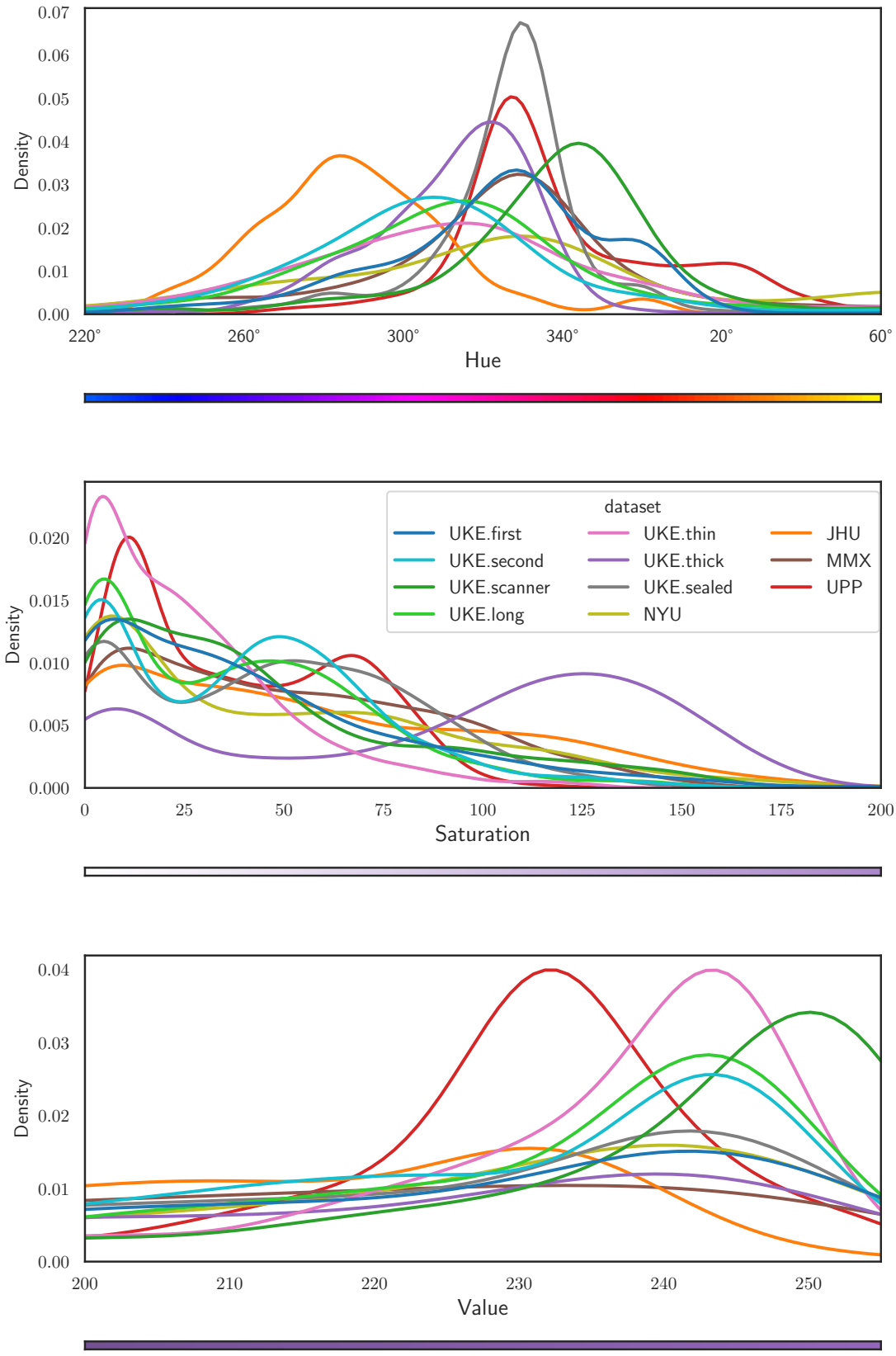


Fig. 3.16: KDE of hue, saturation and value channel for all image datasets after excluding the background.

4 Discrete Calibrated Survival Prediction

4.1 Introduction

This chapter introduces **D**iscrete **C**alibrated **S**urvival (DCS), a DL-based, recurrent survival model that extends the work of DRSA and KAM to achieve SotA discrimination performance while being well calibrated for tabular EHR datasets. The thesis compares DCS to five competing survival models, namely CoxPH, DeepSurv, CoxTime, DRSA and KAM. The evaluation shows that DCS improves discriminative performance on three public medical tabular datasets (SUPPORT, METABRIC, and FLCHAIN) from sec. 3.1, while achieving the best overall calibration among discrete-time models.

The performance gain in discrimination of DCS can be attributed to two novel ideas. First, DCS features a modified DRSA architecture that allows variable temporal output node spacing of discrete predictions over time. The best results were obtained with data-driven node spacing that ensures a uniform distribution of censoring and real events per discrete time step. This equal distribution of events allows for a more balanced training on each of the predicted steps.

Also, DCS features a novel loss extension that optimizes the use of not only comparing uncensored (called event-to-event, or EE) individuals, but also taking into account censored patients (event-to-censored or EC) to improve discriminative performance.

4.2 Methods

4.2.1 Data

The model developed in this chapter uses the three EHR datasets, namely SUPPORT, METABRIC, and FLCHAIN. As described in sec. 3.1 these datasets were selected since they show variations in survival-related properties like median survival duration (from only 57 days to 85 months in METABRIC), censoring rate (from 32 % in SUPPORT to 72 % in FLCHAIN) and number of proportional features (see sec. 2.5.1) that varies from 25 % in FLCHAIN to 79 % in SUPPORT.

4.2.2 Model Architecture

Following the definitions of sec. 2.3, this section explains the development of the discrete time DL-based survival prediction model for a survival curve $\hat{S}(t|\boldsymbol{x})$, given the input vector of features $\boldsymbol{x}^i$ for an individual i at fixed time points $t \in \{t_1, t_2, \ldots, t_L\}$. An overview of the full architecture and objective function for the DCS model is illustrated in fig. 4.1 and described in the following including variations that were analyzed in appendix C.1.

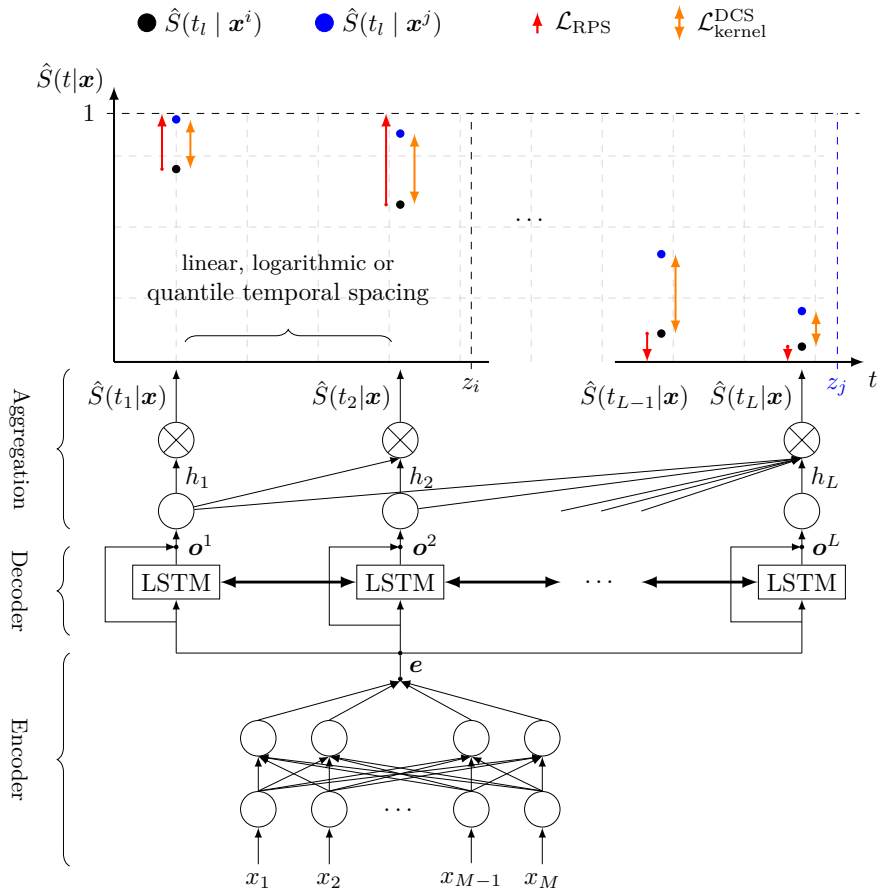


Fig. 4.1: Visualization of the DCS network architecture from the individual feature vector $\boldsymbol{x}$, over an encoder, decoder and aggregation part to the output of a survival curve prediction $\hat{S}(t|\mathbf{x})$ for $t \in \{t_1, t_2, \ldots, t_L\}$. The red arrows indicate the loss $\mathcal{L}_{\text{RPS}}$, which minimizes the distance from the prediction to 1 before and to 0 after the event time z_i. The loss $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ punishes close predictions for individual j compared to i where $(z_j > z_i)$, as shown in orange.

The architecture of DCS can be divided into an encoder, decoder and aggregation part that are explained below.

Encoder

Firstly, a vector $\boldsymbol{x} \in \mathbb{R}^M$ of M concatenated features is fed through a MLP network to produce an encoding $\boldsymbol{e} \in \mathbb{R}^{d_{\text{emb}}}$ depending on the input as $\boldsymbol{e} = \text{enc}(\mathbf{x})$. In contrast to CoxTime and DRSA, no positional information (absolute time t_l or index l) is concatenated. This simplifies the encoded latent representation to be the same for all decoder steps described next.

Decoder

Given an output node spacing vector $\boldsymbol{t} = [t_1, t_2, \ldots, t_L]^T$ that contains the discrete time points for the survival prediction, the decoder part of DCS feeds the encoded feature vector $\boldsymbol{e}$ through a recurrent structure, namely a bidirectional single- or multi-layer LSTM [106] to maximize the use of temporal information. Further, a skip connection is introduced to concatenate the LSTM output $\boldsymbol{r}^l \in \mathbb{R}^{d_{\text{dec}}}$ for each timestep with the feature encoding vector $\boldsymbol{e}$ to produce the output of the decoder part of the model architecture $\boldsymbol{o}^l \in \mathbb{R}^{(d_{\text{enc}}+d_{\text{dec}})}$.

Aggregation

Lastly, the aggregation part of the network reduces the output vector $\boldsymbol{o}^l$ of the recurrent decoder structure to a scalar hazard prediction $h_l \in [0, 1]$ for all timesteps using another MLP with a sigmoid activation in the output. The resulting value for a timestep t_l defined as $h_l = P(t \in V_l \mid t > t_{l-1})$ represents the discrete hazard rate (see eq. (2.7)) which is the probability that the event for the individual happens in V_l given that it did not happen before. Following sec. 2.3, hazard rates $h_1, h_2, \ldots, h_L$ can then be combined to the discrete survival prediction curve over time $\hat{S}(t_l|\mathbf{x}) \in [0, 1]^L$ by

$$\hat{S}(t|\mathbf{x}) = \prod_{j=1}^{l} (1 - h_j). \tag{4.1}$$

Output Node Spacing

The DCS network calculates output predictions $\hat{S}(t_l|\boldsymbol{x})$ for the discrete intervals V_l with $l \in \{1, \ldots, L\}$. Prior works [123, 195] use equidistant output node spacing for the discrete survival prediction task, such that every consecutive inference predicts the survival of a patient for e.g. every month. While equidistant temporal predictions are a natural choice, they might not be optimal for model learning given the distribution of training data. To obtain a model with high-quality predictions for earlier as well as later time-points, this section proposes two alternative approaches taking the event distribution of the training data into account. For some datasets, it is reasonable to assume that the difference between surviving 100 and 200 days should be more important than the difference between 1100 and 1200 days, thus logarithmic spacing of output nodes is introduced. Moreover, quantile spacing ensures that within each output interval V_l, the same number of event and censoring cases are observed leading to a uniform distribution of the training data per predicted time-point. This ensures that the distribution of the individual event or censoring times z_i is approximately equal within each interval V_l. For the EHR datasets, the number of individuals that experience their event or censoring time in the corresponding intervals is illustrated in fig. 4.2. A more homogeneous distribution can be observed for a logarithmic output node distribution compared to the linear spacing for SUPPORT. Nonetheless, for METABRIC and FLCHAIN this distribution does not show a homogeneous event distribution of individuals, thus introducing the data-driven quantile spacing approach. The three strategies are analyzed in additional detail below.

Linear spacing This thesis defines the three aforementioned alternatives for the spacing vector $\boldsymbol{t} = [t_1, t_2, \ldots, t_L]^T$ depending on the number of predicted time points L and the maximum event time $t_{\max}$ in the following. Firstly, linear spacing divides the time interval from 0 to $t_{\max}$ such that

$$t_l = \left\lceil \frac{l}{L} t_{\max} \right\rceil \text{ for } l \in \{0, \ldots, L\} \tag{4.2}$$

defines an equal distance between all output nodes where $\lceil \cdot \rceil$ rounds up to the nearest integer.

Logarithmic spacing Likewise, logarithmic spacing is defined as an equidistant placement of output nodes in the log-transformed time space. Here, t_0 is set to 0 while all others are defined as

$$t_l = \left\lceil \exp_{10}\left(\frac{l}{L}\log_{10} t_{\max}\right)\right\rceil \text{ for } l \in \{1,\ldots,L\} \tag{4.3}$$

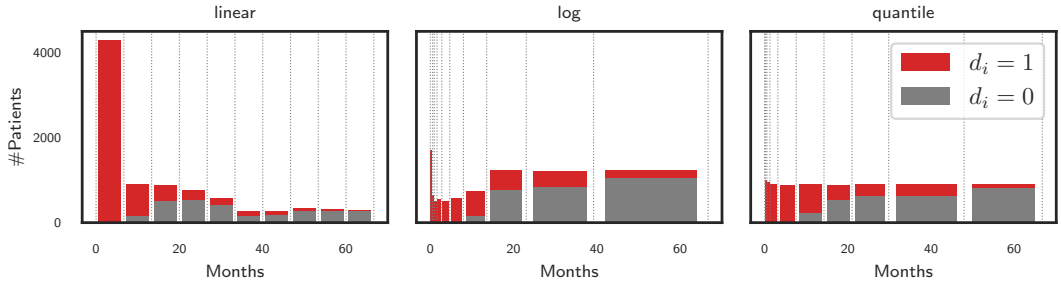
where $\exp_{10}(\cdot)$ is defined as $10^{(\cdot)}$.

Quantile spacing Lastly, quantile spacing is defined in a more data-driven approach by ensuring a uniform event and censoring distribution for all output nodes. This is achieved by dividing the interval $(0, t_{\max}]$ as follows. Recall that $\boldsymbol{z} = [z_0, z_1, \ldots, z_{N-1}]$ contains all event and censoring times for the analyzed dataset. Further, let $\tilde{\boldsymbol{z}} = [z_{\pi_0}, z_{\pi_1}, \ldots, z_{\pi_{N-1}}]^T$ contain those event times z_i for all N individuals in ascending order where z_{π_0} contains the smallest event time, z_{π_2} the second smallest etc. For an equal distribution of event or censoring times, the boundary t_l is then defined as

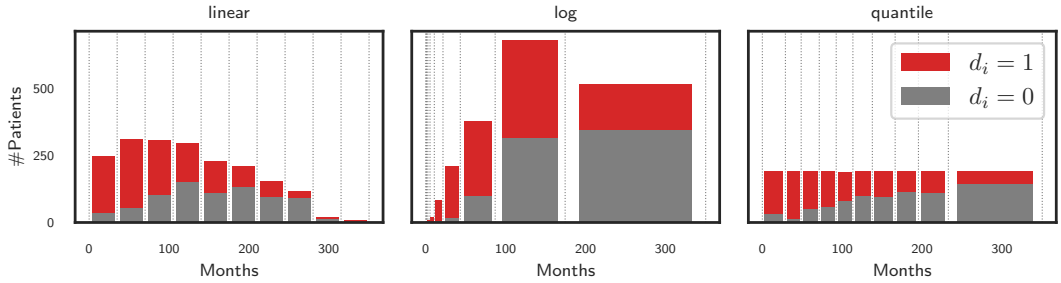
$$t_l = \tilde{\boldsymbol{z}}_{\lfloor \frac{lN}{L} \rfloor} \tag{4.4}$$

where the $\lfloor \frac{lN}{L} \rfloor$-th elements of the sorted event times are chosen as the interval boundaries. This leads to a partition of the observation window into intervals with approximately the same number of observations as

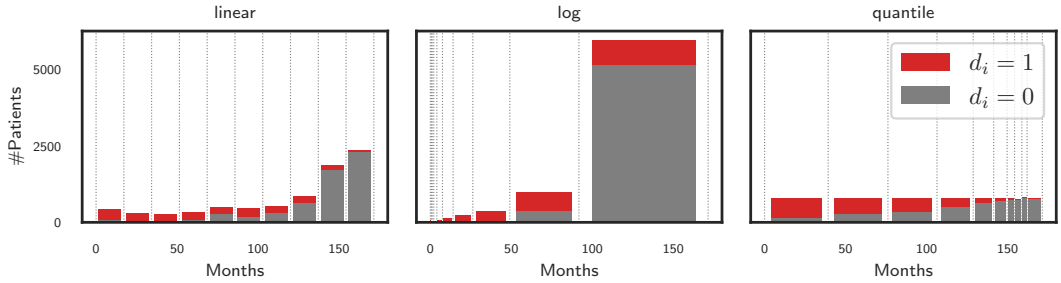
$$V_l = \left(\tilde{\boldsymbol{z}}_{\lfloor \frac{(l-1)N}{L} \rfloor}, \tilde{\boldsymbol{z}}_{\lfloor \frac{lN}{L} \rfloor}\right]. \tag{4.5}$$



(a) SUPPORT



(b) METABRIC



(c) FLCHAIN

Fig. 4.2: Histogram of the number of events (gray for $d_i = 0$, red for $d_i = 1$) per output interval (depicted by dashed lines) on the three EHR datasets and the varying output node spacing (left is linear-, middle is logarithmic- and right is quantile output node spacing).

4.2.3 Objective Function

The objective function of DCS builds upon ideas from [123] where special attention is given to calibration. Likewise, the objective function of DCS is a weighted linear combination of a discriminative and a calibration loss term described below.

Calibration

Firstly, to account for proper calibration, DCS includes the RPS calibration loss introduced in [123]. It punishes individual survival curve predictions $\hat{S}(t|\boldsymbol{x}_i)$ lower than 1 for $t < z_i$, and predictions $\hat{S}(t|\boldsymbol{x}_i)$ greater than 0 for $t \geq z_i$. As previously defined, let $\hat{S}(t_l|\boldsymbol{x}_i)$ denote the estimated survival curve at time t_l with the individual i's feature vector $\boldsymbol{x}_i$. Additionally, let $l_i \in \{0, 1, \ldots, L-1\}$ be the index of the interval that contains the event time so that $z_i \in V_{l_i}$. The first part of the objective function is the rank probability score (RPS) defined as

$$\mathcal{L}_{\text{RPS}} = \frac{1}{nL} \sum_{i=1}^{n} \left[d_i \sum_{l=1}^{L} \left(\hat{S}\left(t_l \mid \boldsymbol{x}_i\right) - \mathbb{1}_{l<l_i} \right)^2 + (1-d_i) \sum_{l=1}^{l_i} \left(\hat{S}\left(t_l \mid \boldsymbol{x}_i\right) - 1 \right)^2 \right]. \tag{4.6}$$

where the loss is normalized to the number of timesteps L and number of individuals n. Uncensored individuals contribute a MSE loss similar to the calculation of the BrS (see eq. (2.22)) that compares the individual's estimated survival function to a prediction that drops from one to zero at the time interval V_{l_i} of the event or censoring time. Moreover, censored individuals ($d_i = 0$) only contribute up to the interval of censoring where a deviation from a constant survival prediction of 1 is punished.

Discrimination

The second loss term from [123] emphasizes proper discrimination of the model predictions by penalizing the wrong order of pairwise comparisons for all uncensored individuals as

$$\mathcal{L}_{\text{kernel}} = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{A}_{i,j} \exp\left(-\frac{1}{\sigma} \left(\hat{S}\left(z_i \mid \boldsymbol{x}_j\right) - \hat{S}\left(z_i \mid \boldsymbol{x}_i\right) \right) \right) \tag{4.7}$$

where every non-zero entry of $\mathbf{A}_{i,j} = \mathbb{1}_{(i \neq j, d_i = d_j = 1, z_i < z_j)}$ denotes the entry in the i-th row and j-th column of $\mathbf{A} \in \{0,1\}^{n \times n}$ with n being the number of individuals in the dataset that corresponds to a single comparison of two uncensored individuals i and j. Since only cases for $z_i < z_j$ are compared, the predicted survival curve for the individuals at z_i, namely $\hat{S}(z_i|x_i)$ is expected to be smaller than the corresponding survival curve of individual j at the same point in time z_i described as $\hat{S}(z_i|x_j)$.

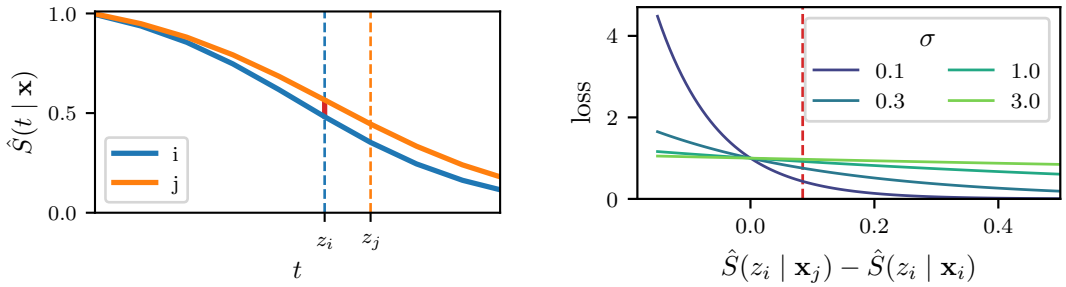
This thesis extends eq. (4.7) to not only include event-to-event (EE), but also event-to-censoring (EC) comparisons of two individuals i and j following the aforementioned definitions. The extension utilizes the fact that for those individuals where $z_i < z_j$, the censoring indicator d_j becomes irrelevant since the individual j lived longer than i which allows the inclusion of this comparison. Formally, the condition of the masking matrix $\mathbf{A}$ in (4.7), that only includes EE comparisons, is relaxed to the new condition $(d_i = 1, z_i < z_j)$ that adds EC comparisons in the novel masking matrix $\mathbf{B} \in \{0,1\}^{n \times n}$, resulting in the discriminative loss part of DCS as

$$\mathcal{L}_{\text{kernel}}^{\text{DCS}} = \frac{1}{|\mathbf{B}|} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{B}_{i,j} \cdot \exp\left[-\frac{1}{\sigma} \left(\hat{S}\left(z_i \mid \mathbf{x}_j\right) - \hat{S}\left(z_i \mid \mathbf{x}_i\right) \right) \right], \tag{4.8}$$

where $\mathbf{B}_{i,j} = \mathbb{1}_{(i \neq j, d_i = 1, z_i < z_j)}$ and an additional normalization by the number of actual comparisons $|\mathbf{B}|$ is introduced. Here $|\mathbf{B}|$ denotes the number of non-zero entries in $\mathbf{B}$ that corresponds to the actual number of comparisons that are performed. A comparison of how the relaxation

of the condition in **B** influences the number of comparisons that are performed is evaluated in additional detail in sec. 4.2.3.

Furthermore, σ controls the size of the loss for wrongly ordered pairwise comparisons depending on the difference between the predictions. In other words, the loss for a comparison between i and j becomes small if $\hat{S}(z_i|x_j)$ is larger than $\hat{S}(z_i|x_i)$. The behavior for different values of σ is analyzed in additional detail in fig. 4.3. For the visualized exemplary predicted survival curves in fig. 4.3a, large values lead to more equal punishment of closer survival predictions while small values for σ diminish the loss influence for a particular difference if the two predictions are sufficiently distanced.



(a) Two exemplary survival curves with the relevant difference highlighted in red.

(b) Resulting loss for varying values of σ. The difference from (a) is highlighted in red.

Fig. 4.3: Visualization of the comparison-based loss $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$.

Overall Loss

Lastly the two loss parts $\mathcal{L}_{\text{RPS}}$ and $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ are combined using a linear combination with a weighting factor λ that leads to the final objective function of DCS:

$$\mathcal{L}_{\text{DCS}} = \mathcal{L}_{\text{RPS}} + \lambda \mathcal{L}^{\text{DCS}}_{\text{kernel}} \tag{4.9}$$

Comparisons

This section analyzes the relationship between the censoring rate and the number of possible comparisons that can be drawn from the population if only EE comparisons are performed or if the additional EC comparisons are considered as introduced in eq. (4.8). The larger the number of comparisons, the more information can be drawn by the individual models for a survival prediction.

Let a dataset $\mathbb{D}$ contain n individuals and have a censoring rate of $c \in [0, 1)$. When drawing two individuals i and j from $\mathbb{D}$ at random, under the assumption that censoring is uniformly distributed throughout the observed event time, the chance of drawing a pair (i, j) that is comparable regarding $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ (4.8) (EE or EC pair) is

$$\begin{aligned} P(\text{comparable}_{\text{DCS}}) = & P(d_i = 1, d_j = 1, z_i < z_j) + \\ & P(d_i = 1, d_j = 0, z_i < z_j) \end{aligned} \tag{4.10}$$

and can be estimated by using the censoring rate c of randomly drawing a censored individual from the dataset. Likewise, an uncensored individual with an actual event is drawn randomly with a probability of $P(E) = 1 - c$. Under the assumption that, when drawing a random pair (i, j), the corresponding event times are equally distributed, the probability that the first event

time z_i is smaller than z_j is assumed as $P(z_i < z_j) = 0.5$. Firstly, the probability of drawing EE pairs which corresponds to the number of comparisons in (4.7) can be estimated as

$$P(\text{comparable}_{\text{EE}}) = P(d_i = 1, d_j = 1, z_i < z_j) = (1 - c)^2. \tag{4.11}$$

Furthermore, the comparisons where the former event time is uncensored and the latter event time is censored can be expressed as

$$P(\text{comparable}_{\text{EC}}) = P(d_i = 1, d_j = 0, z_i < z_j) = c(1 - c). \tag{4.12}$$

Combining (4.11) and (4.12) leads to the final estimation for the new number of comparisons

$$P(\text{comparable}_{\text{DCS}}) = (1 - c)^2 + c(1 - c) = 1 - c\,. \tag{4.13}$$

Since eq. (4.7) only compared all uncensored pairs for i and j that led to $P(\text{comp}_{old}) = (1 - c)^2$, the factor F of more comparisons can be estimated from (4.11) and (4.13) as

$$F_{\text{est}} = \frac{P(\text{comparable}_{\text{DCS}})}{P(\text{comparable}_{\text{EE}})} = \frac{1}{1 - c}. \tag{4.14}$$

This means that, depending on the censoring rate, the number of pairs that can be compared using the added EC comparisons included in the DCS model can be improved by a factor of $1/(1 - c)$ which results in a boost in the number of comparisons of 2 for a censoring rate of $c = 50\,\%$ or a factor of 5 for a censoring rate of $c = 80\,\%$. This number is analyzed in additional detail for the datasets that were utilized in this chapter in the results section at sec. 4.4.3.

Realization

Following CoxTime [136], calculating the comparison matrix for all possible comparisons of a dataset may become infeasible depending on the number of individuals in the dataset since it has a computational complexity of $\mathcal{O}(n^2)$. Thus, comparisons are only made within each individual batch. This reduces the number of comparisons that are actually performed for each epoch during training. Furthermore, the batches are shuffled after each epoch to allow each individual to be compared not only within the same batch but potentially all other samples throughout the training process.

For a performance oriented implementation of the loss eq. (4.9) that was introduced in sec. 4.2.3, a vectorized implementation is used and explained in additional detail in this section. The main idea is to compute losses of each batch for all individuals in parallel using binary masks to extract the relevant information.

The two loss parts $\mathcal{L}_{\text{RPS}}$ and $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ are describe in pseudocode in alg. 1 and alg. 2 respectively. For the individuals with event, the MSE to a curve dropping from 1 to 0 is taken for all predicted timesteps in grid. Likewise, for the censored individuals the MSE to 1 is taken until the time interval of censoring with the index l_i. Likewise, the difference of $\hat{S}\left(z_i \mid \mathbf{x}_j\right)$ and $\hat{S}\left(z_i \mid x_i\right)$ is calculated for all relevant elements of the batch simultaneously. he mathematical operations (not, and, $+$, $-$, $(\cdot)^2$, $\otimes$, exp) are performed element-wise.

Algorithm 1: Implementation of $\mathcal{L}_{\text{RPS}}$.

Input: grid; `// vector of right boundaries of` V_i
Input: $\boldsymbol{d}$; `// vector of all (batch) censoring indicators` d_i
Input: $\boldsymbol{z}$; `// vector of all (batch) observation durations` z_i
Input: $\hat{\mathbf{S}}$; `// matrix of N x L survival curve predictions`

$N \leftarrow \text{length}(\boldsymbol{d})$;
$L \leftarrow \text{length}(\text{grid})$;
$\mathbf{Z}_{is} \leftarrow \text{repeat}(\boldsymbol{z}, \text{n}=L, \text{axis}=1)$;
$\mathbf{D}_{is} \leftarrow \text{repeat}(\boldsymbol{d}, \text{n}=L, \text{axis}=1)$;
$\mathbf{t}_{<z_i} \leftarrow \text{repeat}(\text{grid}, \text{n}=N, \text{axis}=0) < \mathbf{Z}_{is}$;
$\text{losses\_unc} = \mathbf{D}_{is} \text{ and } \left(\hat{\mathbf{S}} - \mathbf{t}_{<z_i}\right)^2 \quad \text{losses\_cen} = [\text{ (not } \mathbf{D}_{is}) \text{ and } \mathbf{t}_{<z_i}] \text{ and } \left(\hat{\mathbf{S}} - 1\right)^2$;
$\text{loss} \leftarrow \sum (\text{losses\_unc} + \text{losses\_cen})$;

Output: RPS loss contribution combined for all individuals of the current batch

4.3 Experiments

4.3.1 Implementation

In the experiments, the CoxPH model from `lifelines`[14], as well as DeepSurv and CoxTime from `pycox`[15] are integrated. DRSA, KAM and DCS models were self-implemented in `Tensorflow 2.5.0` and `Python 3.8.7`. The implementation uses external implementations of the C-index-td (`pycox`) and CDAUC (`scikit-survival`[16]). DDC was self-implemented following sec. 2.4.4. For optimization, `AdamW` [152] was used including optional weight decay and gradient clipping. The overall code with all datasets, baseline models, pipelines, hyperparameter tuning, and metrics can be found in a public repository.[17]

4.3.2 Hyperparameter Tuning

For hyperparameter tuning, `scikit-learn`[18] wrappers for all models were used for the preprocessing workflows and datasets.

Bayesian [255] parameter search from `scikit-optimize`[19] was used to find the optimal hyperparameters for each model and dataset individually. The data was split into 60% training, 20% validation and 20% test data stratified by the event indicator. The hyperparameter tuning objective was set to be validation CDAUC. The test data was only used during the final evaluation of the model performance reported in sec. 4.4. A detailed description of hyperparameter tuning and the best parameters for each model can be found in appendix C.1.

4.3.3 Pre- and Post-processing

The dataset features are standardized and imputed, where required. Numerical features are standardized to a mean of zero and a standard deviation of one. To impute missing values, this work uses median imputation for numerical and most frequent imputation for categorical features. Categorical features are one-hot encoded. For the SUPPORT dataset, zeros in heart rate or

[14] https://github.com/CamDavidsonPilon/lifelines/ [60]
[15] https://github.com/havakv/pycox from [136]
[16] https://github.com/sebp/scikit-survival [185]
[17] https://github.com/imsb-uke/dcsurv
[18] https://github.com/scikit-learn/scikit-learn [182]
[19] https://scikit-optimize.github.io/ [104]

Algorithm 2: Implementation of $\mathcal{L}_{\text{kernel}}$.

Input: grid; `// vector of right boundaries of` V_i
Input: $\boldsymbol{d}$; `// vector of all (batch) censoring indicators` d_i
Input: $\boldsymbol{z}$; `// vector of all (batch) observation durations` z_i
Input: $\hat{\mathbf{S}}$; `// matrix of N x L survival curve predictions`

$N \leftarrow \text{length}(\boldsymbol{d})$;
$\mathbf{M}_{z_i > z_j} \leftarrow \text{repeat}(z, \text{n}=N, \text{axis}=1) < \text{repeat}(z, \text{n}=N, \text{axis}=0)$;
$\mathbf{M}_{d_i=0} \leftarrow \text{repeat}(d, \text{n}=N, \text{axis}=1)$;
$\mathbf{B} \leftarrow \mathbf{M}_{d_i=0}$ and $\mathbf{M}_{z_i > z_j}$
$\mathbf{L}_i \leftarrow \text{get\_interval}(\boldsymbol{z}, \text{grid})$; `// convert durations` z_i `to interval indices`
$\hat{\mathbf{S}}_{\mathbf{L}_i} \leftarrow \text{extract\_at}(\hat{\mathbf{S}}, \mathbf{L}_i)$; `// Only keep` $\hat{\mathbf{S}}(z_i|x_i)$
$\hat{\mathbf{S}}_{\mathbf{L}_i \text{ at } x_i} \leftarrow \text{repeat}(\hat{\mathbf{S}}_{\mathbf{L}_i}, \text{n}=N, \text{axis}=1)$; `// Repeat vector to N x N matrix`
$\hat{\mathbf{S}}_{\mathbf{L}_i \text{ at } x_j} \leftarrow \text{transpose}(\hat{\mathbf{S}}_{\mathbf{L}_i \text{ at } x_i})$;
$\text{losses} \leftarrow \frac{1}{|\mathbf{B}|} \otimes \exp\left[\frac{-1}{\sigma}\left(\hat{\mathbf{S}}_{\mathbf{L}_i \text{ at } x_j} - \hat{\mathbf{S}}_{\mathbf{L}_i \text{ at } x_i}\right)\right]$
$\text{loss} \leftarrow \sum \text{losses}$

Output: Kernel loss contribution combined for all individuals or the current batch

respiration rate were also treated as missing values.

To evaluate the continuous- and discrete time models as fairly as possible, all estimations are evaluated on the same timesteps, namely the corresponding training set event and censoring times.

4.4 Results

This section shows results of the introduced DCS model regarding discriminative and calibration performance on the three datasets SUPPORT, METABRIC and FLCHAIN. Furthermore, an ablation study is performed for the developed overall objective function eq. (4.9).

4.4.1 Quantitative Results

Discrimination and calibration performance of CoxPH, DeepSurv, CoxTime, DRSA, KAM, and three variants of DCS with linear, logarithmic, and quantile output node spacing on the SUPPORT, METABRIC, and FLCHAIN test data are compared and depicted in tab. 4.1. To obtain variance estimations for the model performance, this thesis uses 10-fold bootstrapping and reports the 'mean ± standard deviation'. The quantitative evaluation focuses on three scenarios, namely the overall best model per metric, performance gains achieved by DCS's novel kernel loss and the novel linear, logarithmic and quantile spacing approaches.

Overall, DCS models reach the best discriminative performance for both the C-index-td (tab. 4.1a) and the CDAUC (sec. 4.4.1). More specifically, DCS with quantile output node spacing (DCS-quant) displays top performance in four out of six comparisons. DCS-quant shows the best C-index-td on the SUPPORT (0.628) and METABRIC (0.698) datasets, while DCS-linear reaches first place on the FLCHAIN dataset (0.803) (tab. 4.1a). Regarding the CDAUC, DCS-quant exhibits top performance on the METABRIC (0.773) and FLCHAIN (0.832) datasets, while the DCS-log model is best on the SUPPORT data (0.658) (sec. 4.4.1).

To specifically understand the influence of the modified kernel loss (4.8) on discriminative performance, this thesis compares DCS-linear to KAM. For all three datasets, DCS-linear outperforms

Tab. 4.1: Quantitative results for C-index-td, CDAUC and DDC for the compared models on the three test sets of SUPPORT, METABRIC and FLCHAIN.

		SUPPORT	METABRIC	FLCHAIN
cont.	CoxPH	0.594 ± 0.009	0.638 ± 0.020	0.798 ± 0.007
	DeepSurv	0.604 ± 0.012	0.679 ± 0.014	0.795 ± 0.014
	CoxTime	0.612 ± 0.007	0.675 ± 0.019	0.790 ± 0.010
disc.	DRSA	0.598 ± 0.006	0.661 ± 0.019	0.792 ± 0.018
	KAM	0.610 ± 0.006	0.668 ± 0.023	0.786 ± 0.013
DCS	DCS-linear	0.623 ± 0.009	0.694 ± 0.018	$\mathbf{0.803 \pm 0.011}$
	DCS-log	0.623 ± 0.009	0.674 ± 0.018	0.792 ± 0.008
	DCS-quant	$\mathbf{0.628 \pm 0.009}$	$\mathbf{0.698 \pm 0.019}$	0.794 ± 0.015

(a) C-index-td (↑)

		SUPPORT	METABRIC	FLCHAIN
cont.	CoxPH	0.619 ± 0.013	0.686 ± 0.028	0.797 ± 0.019
	DeepSurv	0.634 ± 0.016	0.700 ± 0.032	0.800 ± 0.016
	CoxTime	0.647 ± 0.010	0.747 ± 0.020	0.799 ± 0.012
disc.	DRSA	0.613 ± 0.008	0.685 ± 0.029	0.814 ± 0.014
	KAM	0.652 ± 0.018	0.727 ± 0.024	0.807 ± 0.018
DCS	DCS-linear	0.641 ± 0.012	0.730 ± 0.021	0.813 ± 0.017
	DCS-log	$\mathbf{0.658 \pm 0.011}$	0.716 ± 0.028	0.824 ± 0.015
	DCS-quant	0.657 ± 0.012	$\mathbf{0.773 \pm 0.023}$	$\mathbf{0.832 \pm 0.017}$

(b) CDAUC (↑)

		SUPPORT	METABRIC	FLCHAIN
cont.	CoxPH	0.009 ± 0.003	0.021 ± 0.007	$\mathbf{0.001 \pm 0.000}$
	DeepSurv	0.010 ± 0.003	0.017 ± 0.006	0.003 ± 0.002
	CoxTime	$\mathbf{0.007 \pm 0.002}$	$\mathbf{0.009 \pm 0.004}$	0.006 ± 0.001
disc.	DRSA	0.193 ± 0.009	0.296 ± 0.024	0.050 ± 0.004
	KAM	0.139 ± 0.006	0.065 ± 0.013	0.006 ± 0.002
DCS	DCS-linear	0.141 ± 0.013	0.138 ± 0.010	0.012 ± 0.002
	DCS-log	0.055 ± 0.005	0.071 ± 0.006	0.035 ± 0.004
	DCS-quant	0.066 ± 0.006	0.027 ± 0.008	0.021 ± 0.004

(c) DDC (↓)

KAM in the C-index-td (SUPPORT 0.628 vs. 0.610, METABRIC 0.694 vs. 0.668, FLCHAIN 0.803 vs. 0.786). Similarly, DCS-linear surpasses KAM in the CDAUC on METABRIC (0.730 vs. 0.727) and FLCHAIN (0.813 vs. 0.807), while it is worse on the SUPPORT (0.641 vs. 0.652) data.

Lastly, it is interesting to observe that the continuous time models CoxPH, DeepSurv, and CoxTime outperform the discrete DRSA and KAM models in the C-index-td. Continuous time models reach the best calibration as measured by DDC, while DCS-quant reaches the best calibration performance of all discrete time models (sec. 4.4.1). For instance, CoxTime features a slightly lower (better) DDC as compared to DCS-quant across all datasets (SUPPORT 0.066 vs. 0.007, METABRIC 0.027 vs. 0.009, FLCHAIN 0.021 vs. 0.006). Among the discrete time models, DCS-quant shows superior performance compared to DRSA on all datasets (SUPPORT 0.066 vs. 0.193, METABRIC 0.027 vs. 0.296, FLCHAIN 0.021 vs. 0.050) and to KAM on two out of three datasets (SUPPORT 0.066 vs. 0.139, METABRIC 0.027 vs. 0.065, FLCHAIN 0.021 vs. 0.006).

4.4.2 Ablation Study: Objective Function

To compare the influence of the competing loss parts of the algorithm on the three datasets, an ablative study is performed with 10 individually trained models per case and dataset on otherwise equal hyperparameters. The previously found best DCS quantile models per dataset are ablated in the following way: Four different losses are used to train individual models per dataset that ablate the loss from eq. (4.9). The variants shown include one model that is exclusively trained on $\mathcal{L}_{\text{kernel}}$, one only on $\mathcal{L}_{\text{kernel}}^{\text{DCS}}$, one only on $\mathcal{L}_{\text{RPS}}$, and one is trained on the final combination of both losses in eq. (4.9). Fig. 4.4 shows the results of the ablation study. Detailed numbers for each analyzed loss combination can be foundalet in tab. A5. As expected, it can be observed that only training on $\mathcal{L}_{\text{kernel}}$ leads to the worst results regarding calibration for all datasets, also with large differences in discriminative performance (standard deviation regarding CDAUC of up to 3.8 percentage points) for the FLCHAIN dataset. Further, training exclusively on $\mathcal{L}_{\text{RPS}}$ leads to better calibration performance compared to $\mathcal{L}_{\text{kernel}}$ for all three datasets, but may lead to worse discriminative performance, as can be observed for SUPPORT (CDAUC: $\mathcal{L}_{\text{RPS}}$ 0.639 vs. $\mathcal{L}_{\text{kernel}}$ 0.645). Using the aggregated loss of the two leads to better discriminative performance at the cost of worse calibration scores regarding DDC that is approximately doubled for all three datasets.

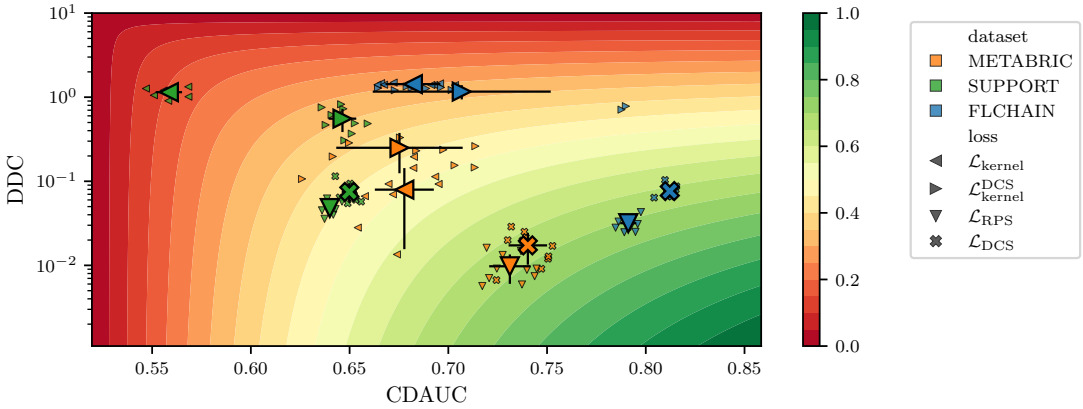


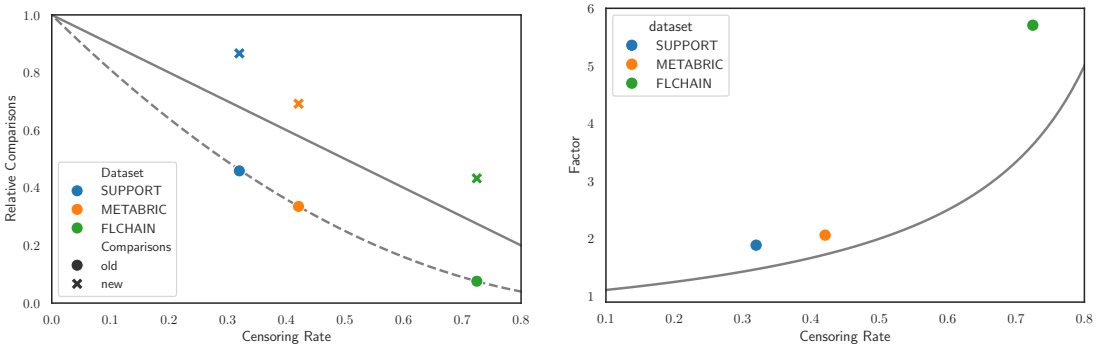
Fig. 4.4: Ablative loss analysis for DCS showing the old kernel loss $\mathcal{L}_{\text{kernel}}$, the extension $\mathcal{L}_{\text{kernel}}^{\text{DCS}}$, $\mathcal{L}_{\text{RPS}}$ and the final combination of $\mathcal{L}_{\text{DCS}}$ per dataset over the log DDC and CDAUC metrics (background colored by their harmonic mean). The standard deviation for the ten models that were trained per variant and dataset are shown as horizontal and vertical error bars.

4.4.3 Number of Comparisons

Tab. 4.2: Number of old ($|\mathbf{A}|$) and new ($|\mathbf{B}|$) possible comparisons for each analyzed dataset. Also included is the number of patients, censoring rate and the factor of old over new comparisons $F = |\mathbf{B}|/|\mathbf{A}|$.

	#Patients	Censoring rate	$\lvert\mathbf{A}\rvert$	$\lvert\mathbf{B}\rvert$	F
SUPPORT	8873	32 %	1.81×10^7	3.41×10^7	1.9
METABRIC	1904	42 %	6.08×10^5	1.25×10^6	2.0
FLCHAIN	7874	72 %	2.35×10^6	1.34×10^7	5.7

One of the novel ideas presented in this chapter is the addition of censored individuals to the discriminative loss part. The estimated number of comparisons can be compared to the actual number of comparisons for the analyzed datasets shown in tab. 4.2. Fig. 4.5 shows that the estimation closely matches the number of old comparisons for all datasets. Fig. 4.5a illustrates the relative amount of comparisons performed compared to all possible pairs that can be drawn from the dataset. It is shown that, for higher censoring rates, the factor F suggests that the approach allows for significantly more comparisons on datasets with high censoring rates (e.g. 70 % censoring leads to a factor of more comparisons of approximately 3 for uniform censoring and over 5 for FLCHAIN) in fig. 4.5b. Note that the observed factor F is even higher than predicted for real-world data sets, indicating that the assumption of a uniform censoring distribution is violated. Furthermore, tab. 4.2 shows that the maximum number of possible comparisons for SUPPORT is raised from 1.81×10^7 to 3.41×10^7, a factor of $F = 1.9$, for METABRIC from 6.08×10^5 to 1.25×10^6 and for FLCHAIN from 2.35×10^6 to 1.34×10^7.



(a) Relative number of comparisons over the censoring rate. The estimated number of EE comparisons and all possible comparisons is depicted with a dashed and a solid gray line respectively.

(b) Factor F of more comparisons over the censoring rate. The gray line depicts the estimated factor for a uniform censoring distribution.

Fig. 4.5: Observed and estimated number of comparisons over the censoring rate for SUPPORT, METABRIC, and FLCHAIN including the estimations for perfectly uniform censoring rates.

4.5 Discussion

The result section shows that all spacing variants of the DCS model show improvements compared to the other models analyzed with respect to discrimination and calibration for the three analyzed datasets. DCS boosts the discriminative performance regarding C-index-td and CDAUC and outperforms the discrete, DL-based baseline models regarding calibration except for KAM on FLCHAIN. These results suggest that both, the novel kernel loss (4.8) and the new output spacing increase discriminative performance robustly outperforming all competing models. Overall, (quantile) output node spacing can be used to boost both, calibration and discrimination performance of DL-based, discrete time survival models.

DCS-quant reaches the best calibration performance for the analyzed discrete time models (DRSA and KAM) for two of three datasets. While DCS-quant shows the best calibration for discrete time models, it is worse than the calibration of the continuous time models (CoxPH, DeepSurv, and CoxTime). Nonetheless, DCS-quant is not able to improve results regarding calibration performance when compared to the continuous time models that were evaluated. A reason for the good calibration of continuous time models, at the cost of inferior discrimination, might be that the time-dependent baseline hazard already defines a reasonably calibrated population wide survival estimate. This could suggest that those models might suffer from bad calibration if subsets with differing survival distributions exist. An explanation of poorer calibration performance for the discrete models lies in the hyperparameter tuning setup. Since the two parts of the overall objective function in eq. (4.9) are combined linearly by λ that is included in hyperparameter tuning, it should not be optimized only on a single metric like CDAUC that focuses on discriminative performance. The optimal value of λ may then neglect adverse effects regarding calibration performance. This problem is addressed in the next chapter by introducing a combined hyperparameter tuning objective metric (see sec. 5.2.5).

The results also indicate that the commonly used C-index-td might not be the best metric to evaluate discriminative performance for survival prediction models with potentially crossing survival curves. While the C-index-td for FLCHAIN yields nearly similar values for all the inspected models, CDAUC shows clearer differences in model performance.

4.6 Conclusion

This chapter presented a novel DL-based survival model, called DCS, with SotA discrimination and good calibration. It utilized ideas of previous approaches like non-linear feature dependency [126], predicting on discrete timesteps [123,195], and allowing more complex input encoding by multiple layers but disregarding positional encoding before the decoder structure. Two novel features that are introduced in DCS seem to be primarily responsible for this performance increase.

Firstly, an extension of the $\mathcal{L}_{\text{kernel}}$ is introduced, which additionally includes event-to-censoring (EC) pairs during training to boost discrimination. It is shown that $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ (from eq. (4.8)) increases the number of performed comparisons for the analyzed datasets by a factor of approximately 2 to 6 depending on the censoring rate of the dataset compared to the baseline from eq. (4.7). It could be shown in sec. 4.2.3 that the number of comparisons can be improved by a factor of $1/(1-c)$ depending on the censoring rate c and assuming a uniform censoring distribution. Comparing this estimate to the real world datasets shows that this number is even higher in practice. These additional comparisons improve discriminative performance with respect to C-index-td and CDAUC in the three analyzed datasets. This can also be observed in the ablation study (see sec. 4.4.2) for the objective functions that allows a direct comparison of the discriminative and calibration performance for the loss combinations. While $\mathcal{L}_{\text{RPS}}$ alone produces reasonable performance regarding CDAUC and DDC, combining it with $\mathcal{L}^{\text{DCS}}_{\text{kernel}}$ boosts the discriminative performance for all three datasets at the cost of lower calibration performance that might be a trade-off worth taking in practice.

Secondly, this thesis introduces three temporal output node spacing options (linear, logarithmic, and quantile spacing) to increase the model discrimination and calibration. Overall, the quantile approach provides the best discriminative performance on the three datasets investigated compared to linear- and logarithmic spacing.

Nonetheless, the difference in calibration performance compared to CoxPH must be further evaluated. The next chapter shows that a proper hyperparameter tuning that incorporates for discriminative and calibration performance can improve the stability of calibration that can even outperform the CoxPH approach.

Overall, DCS shows good discriminative and calibration performance for a DL-based survival prediction model for tabular EHR datasets. The proposed method does not use any specific adaptations for medical datasets indicating that it would also work outside the medical context for other survival prediction problems like customer churning analysis or mechanical failure prediction. The results of this chapter were published in [84] and a public repository is available.[20]

[20] https://github.com/imsb-uke/dcsurv

5 Postoperative Relapse Prediction on Electronic Health Records

5.1 Introduction

The previous chapter described the realization of a DL-based survival prediction model that optimized the utilization of survival information from censored individuals, and additionally allowing non-linear- and time-dependent feature interactions in contrast to the CoxPH model. To date, multiple, continuously evolving tools to predict oncological and functional outcomes and complications in PCa patients treated with RP have been established [125, 220, 242]. However, the majority of widely used prediction tools are still based on classical statistics with restrictions discussed in sec. 2.5.1.

Nonetheless, nomograms, as discussed in sec. 2.3 that estimate the likelihood of specific outcome, which is intricately linked to treatment decisions or recommendations, are still used. While the nomograms were carefully constructed, it is obvious that a larger number of features and varying feature representations will eventually lead to more personalized and better predictions. Survival nomograms are also based on the strong mathematical assumptions of the Cox Proportional Hazard (CoxPH) [54] model presented in sec. 2.5.1. To recap, the model requires that survival estimation for an individual can be attributed to a linear combination of input features, and furthermore that any of these features have a constant effect over time on an individual's risk of having an event. Neither of these assumptions can be taken for granted, in particular with the large and heterogeneous datasets that are available today. A meta-study on phase 3 clinical trials in oncology revealed that 24 % of PH assumptions were violated [191]. Further limitations are homogeneity and selection bias of the development cohort, lack of external validation and as a result failing generalization. Since existing nomograms are limited to a predetermined set of predictors, they potentially miss valuable information present in the data that was lost during feature engineering. In addition, non-linear interactions between predictors are mostly ignored. Finally, nomograms show little flexibility in including different data types or in choosing data models beyond multivariate regression or survival models, which become accessible with modern DL approaches accounting for multimodality.

This chapter evaluates the survival prediction model DCS that reaches state-of-the-art discrimination and calibration on the MK dataset that was introduced in sec. 3.1.4. The dataset with 16,953 PCa patients that received RP at the Martiniklinik contains a rich feature set of over 90 features and followup information of up to more than 20 years. The results regarding relapse prediction are compared to the classical method for PCa relapse prediction, namely the CoxPH model, that is the basis of PCa-related nomograms [125, 220] (see sec. 2.2.1).

To determine the best possible set of features for the BCR relapse prediction task, the large feature set is divided into semantically connected feature blocks. Within each of those blocks, the best representation (e.g. GG represented as primary and secondary, as a sum or as percentage of the total tumor) for the individual features is analyzed by choosing the best univariate survival prediction model. After identifying the best representations based on univariate survival models, the semantic blocks of features are combined into a multivariate model and evaluated in terms of discriminative and calibration performance. Also, feature importance is analyzed using a step-by-step inclusion of the previously defined feature blocks regarding discriminative performance.

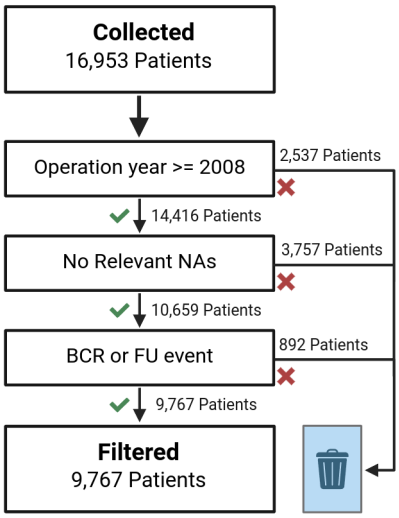
Lastly, with the found DCS-based survival model that yields the best overall discrimination, downstream analysis is performed to generate additional insights in the connections between different features. Therefore, a clustering approach is introduced to obtain risk group estimations based on the predicted survival curves.

5.2 Methods

The following section presents the necessary methods to build and evaluate the DCS survival model for this dataset in additional detail. After the cohort selection is presented, the most important features (including alternative encodings) for the task of relapse free survival after RP are shown. The baseline models that DCS compares to are introduced as well. Further, downstream analysis of the results is presented.

5.2.1 Cohort Selection

The analyzed dataset of this chapter was introduced in sec. 3.2.1. It provides a total of 16,953 PCa patients that were treated with RP in the Martiniklinik in Hamburg, Germany between 1992 and 2018. This chapter further focuses on the sub-cohort of patients who received RP between 2008 and 2018, had no missing values in relevant fields, and participated in the yearly follow-up questionnaires. The detailed selection criteria with the filtering steps are presented in fig. 5.1a, resulting in a final cohort of 9,767 patients.



(a) Filtering steps for proper evaluation of BCR relapse prediction. Created with BioRender.com

MK dataset	
#Patients	9,767
#BCR	2526
Censoring rate	74%
Median time-to-BCR	19 months
Median FU time	61 months

(b) Basic characteristics of the filtered dataset.

Fig. 5.1: MK dataset filtering steps for the final dataset and basic characteristics shown in (b).

Endpoint

Moreover, the follow-up data of these patients was observed for more than 10 years for the following endpoints: biochemical recurrence (measured as PSA level at routine follow-up), lost to follow-up (FU, patients received yearly questionnaires; this endpoint may include lost to follow-up caused by death), metastases, death by PCa, additional therapy.

Fig. 5.1b depicts basic parameters of the extracted MK dataset. It shows a high censoring rate of 74 % which is even higher than the other tabular datasets previously discussed with 72 % for FLCHAIN (see sec. 3.1). This further motivates the usage of a survival model like DCS that extracts as much information of the censored cases as possible. It has a median time-to-event (BCR) of 19 months, a median FU time of 61 months and 2,526 patients experienced BCR that is the event-of-interest of this chapter. A detailed table of patient characteristics in the dataset can be found in tab. A6.

The day of RP was chosen as day 0 for all patients. For further analysis, patients who did not experience any of the other events were characterized as lost to follow-up (FU) and are considered censored individuals. A histogram and KM-curve of the event distribution for the extracted dataset is illustrated in fig. 5.2.

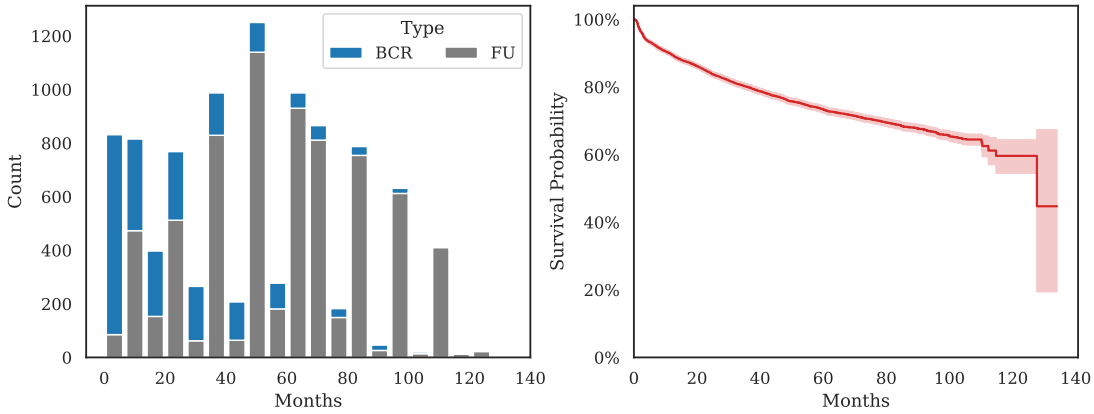


Fig. 5.2: Endpoint distribution of the filtered MK dataset. Histogram of event types (either BCR or FU) over time (left) and KM-curve (right).

5.2.2 Feature Selection

The most important features in terms of postoperative relapse prediction of previous models [123, 220] include preoperative PSA-level, pathological GG in RP specimen, seminal vesicle invasion (SVI), capsular extension, lymph node invasion (LNI) and positive resection margin that are further illustrated in fig. 5.3. This work utilizes these features for the development of the relapse risk prediction models.

5.2.3 Analysis Pipeline

The following section presents detailed information about the data preprocessing pipeline as well as the used survival models for this chapter.

Feature Representation

Since the MK dataset contains numerical and categorical features, they can be preprocessed in different ways before a survival model utilizes their information. This thesis analyzes different strategies to find the best representation of selected features for the aforementioned relapse

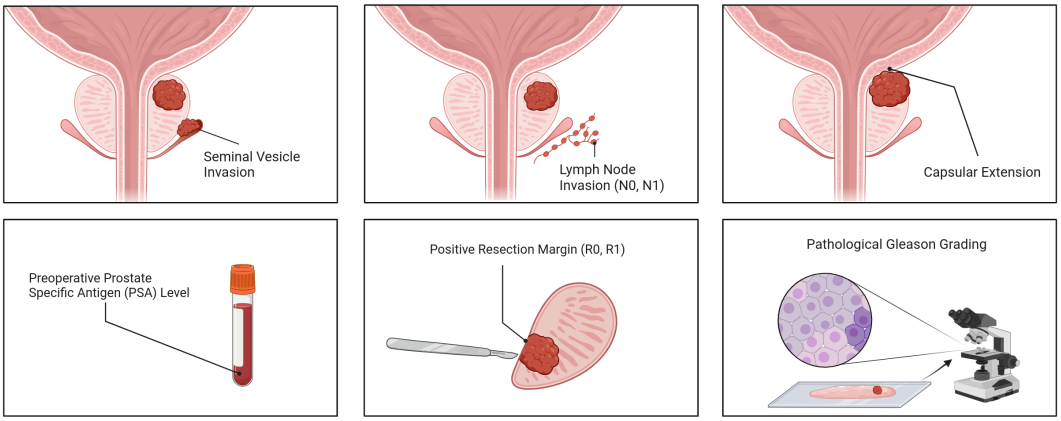


Fig. 5.3: Schematic visualization of relevant predictors in the nomograms: Seminal vesicle invasion (SVI), lymph node invasion (LNI), extracapsular extension, preoperative PSA level, positive resection margin and pathological GG.

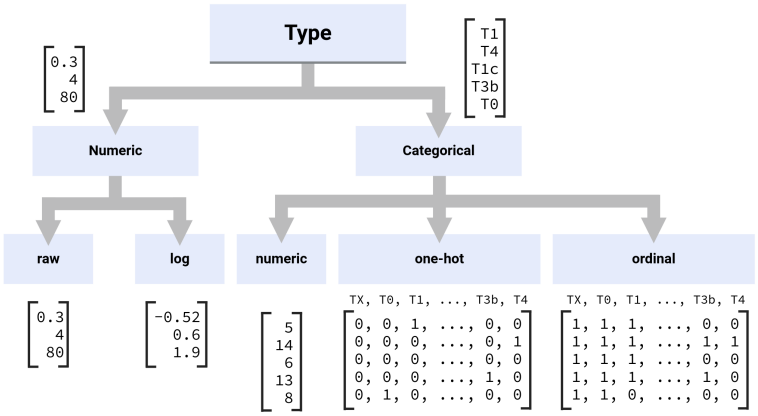


Fig. 5.4: Possible feature encodings shown for a numerical and a categorical feature vector. The numerical data is either used unchanged (raw) or log transformed. For categorical data, either a numeric, one-hot or ordinal representation is used. Created with BioRender.com

prediction task that are shown in fig. 5.4. Numerical features are usually not preprocessed, thus passed along raw into the survival model. An exception is the PSA value of the patient that is usually left skewed and therefore commonly log transformed [25, 240]. Further, categorical features such as T-stage (path) or different variations of Gleason grading need to be preprocessed. Different representations for this data are possible depending on their inherent information.

Firstly, a one-hot encoding where the individual representation of a categorical feature $x_{\text{raw}} \in \{c_0, c_1, \ldots, c_{M-1}\}$ with M distinct categories c_k is converted into M binary feature vectors per individual as

$$\textbf{one-hot}: \quad x_{\text{OH}_k} = \mathbb{1}_{c_k = \boldsymbol{x}_{\text{raw}}} \text{ for } k \in \{0, 1, \ldots, M-1\}. \tag{5.1}$$

This way, all classes c_k are represented independently of each other. A common problem that occurs with this notation is low variance within the least frequent classes that might lead to instable or even non-converging training of the CoxPH model. A suggested countermeasure [60] is to drop those columns from the analysis even though they may contain useful information for the prediction.

Secondly, if an ordering exists such that the classes can be interpreted as $c_0 < c_1 < \ldots < c_{M-1}$, a numerical or ordinal encoding might be preferred to additionally encode the dependency of the feature. For a numerical encoding, the ordered list of classes c_k is assigned the value k in the processed feature vector. This work calls this transformation numeric:

$$\textbf{numeric}: \quad x_{\text{num}_k} = k \text{ for } k \in \{0, 1, \ldots, M-1\}. \tag{5.2}$$

A downside of this approach is that the differences between each categorical value are equidistant. This property might be preferred for example in t-shirt sizes where the distance between S and M might be the same as the distance between L and XL, but not in T-stage (path) where the distance between T1a and T1b is smaller by definition than the distance from T1c to T2a even though both are only one category below the other. This problem can be addressed by using ordinal encoding. For this, the raw categorical vector is transformed into

$$\textbf{ordinal}: x_{\text{ord}_k} = x_{\text{raw}} \leq k \text{ for } k \in \{0, 1, \ldots, M-1\}. \tag{5.3}$$

This way each individual gets assigned M binary indicators x_{ord_k} indicating if at least category k is present in the individual.

These alternatives are investigated in more detail for different features in sec. 5.3. To sum up, the following categorical approaches with the corresponding number of extracted features per individual in parentheses are analyzed:

one-hot (M)
+ Does not assume equal spacing between the feature values since all are encoded individually.
- Sparse vectors might lead to stability problems
- No interaction between the categories

numeric (1) Choose ascending integers for the categories in order.
+ Only one column to present the data
- Assumes equal spacing between categories that is sometimes unreasonable

ordinal (M) For all M classes that are present, create ordinal features that show if the individual's category is at least k.
+ Ordering is kept
+ Distance between categories can vary
- Sparse vectors might lead to stability problems

5.2.4 Risk Stratification

After generating discriminative individual survival curves for all individuals, it is useful to generate clinically interpretable categories. These extracted risk groups can help to compare an individual to a subpopulation with equal risk. To stratify the individual discrete survival curves $\hat{\mathbf{S}}(t|\boldsymbol{x})$ into i risk groups, this work utilizes the individual survival curve predictions for an individual

$$[\hat{S}(t_0|\boldsymbol{x}), \hat{S}(t_1|\boldsymbol{x}), \ldots, \hat{S}(t_L|\boldsymbol{x})]^T \in \mathbb{R}_+^L \tag{5.4}$$

are interpreted as an L-dimensional vector for a clustering algorithm with parameters θ_{Cluster}. The clustering algorithm can then be used to label N individuals into one of K specific risk groups as $\boldsymbol{c} \in \{0, 1, \ldots, K\}^N$ where the i-th entry of $\boldsymbol{c}$ corresponds to the risk group of individual i. The clustering algorithm that is used in this work is K-means clustering [154].

To evaluate if the risk stratification that divides the individuals into disjoint groups have statistically significantly different survival times, a modified log-rank test with Fleming-Harrington weights ($p = 1, q = 0$) is used as encouraged in [144]. The test compares the KM-curves of the

subpopulations that belong to the same risk group for all pairs of risk groups for statistically significant differences. If all pairwise tests succeed, the number of risk groups is incremented by one and the procedure is repeated. This procedure continues with a growing number of disjoint groups in the clustering algorithm until the log-rank test fails for at least one pairwise comparison. The pseudocode of the procedure is shown in alg. 4. To avoid leakage, the clustering algorithm is performed on a training subset while the log-rank evaluation uses a distinct validation set.

Algorithm 3: Log-rank test-based maximum number of risk groups.

Input: Discrete predictions $\hat{\mathbf{S}}(t|\boldsymbol{x}) \in \mathbb{R}^{N \times L}$ for $\boldsymbol{x}$ in $X_{\text{train}} \cup X_{\text{val}}$, $\boldsymbol{z}_{\text{val}}$ and $\boldsymbol{d}_{\text{val}}$, Confidence level α

```
success ← True;
k ← 2;
while success do
    Train θ_Cluster on Ŝ(t|x) to get c_train ∈ {0, 1, ..., k − 1}^{N_train};
    Infer c_val ∈ {0, 1, ..., k − 1}^{N_val} from θ_Cluster;
    pairwise_logrank ← Logrank_{p=1,q=0}(c_val, z_val, d_val);
    if all(pairwise_logrank < α) then
        Infer c_train ∈ {0, 1, ..., k − 1}^{N_train} from θ_Cluster ;
        c_train_max ← c_train;
        c_val_max ← c_val;
        k_max ← k;
        k ← k + 1
    else
        success ← False
if k = 2 then
    raise("No distinction possible!")
```

Output: Vectors $\boldsymbol{c}_{\text{train}_{\max}}$ and $\boldsymbol{c}_{\text{val}_{\max}}$ with $k_{\max}$ different risk groups.

5.2.5 Experimental Setup

This section utilizes the same implementation details as presented in sec. 4.3. Similar to sec. 4.3.3, the MK dataset features are imputed where required. Numerical features are standardized to a mean of zero and a standard deviation of one. Missing values are median imputed for numerical and most frequent imputed for categorical features.

Discrimination and Calibration in Hyperparameter Tuning

For hyperparameter tuning, this chapter aims to find the best model regarding discrimination while also taking calibration performance into account. Since sec. 4.2.3 is a linear combination of a discrimination and a calibration focused loss part, the weighting parameter λ plays an important role in the model's weighting of discriminative and calibration focus. If CDAUC is used as the only hyperparameter tuning objective, the value of λ will be larger to put more emphasis on the discrimination part of the loss described in sec. 4.2.3. To avoid such a behavior, a custom scoring function is introduced from the harmonic mean of CDAUC and DDC as

$$\text{DCS}_{\text{score}} = \frac{2}{\text{CDAUC}^{-1} + \overline{\text{DDC}}^{-1}}. \tag{5.5}$$

This score combines CDAUC and DDC to account for discrimination and calibration at the same time. It includes the log-transformed DDC as

$$\overline{DDC} = \frac{\log_{10}(DDC)}{\log_{10}(1e-6)} \tag{5.6}$$

since resulting scores may be in the order of multiple magnitudes. It normalizes the log-transformed DDC between 0 and 1 for DDC $\in [10^{-6}, 1]$ where a transformed score of 1 represents the best score. The resulting scoring function in relation to CDAUC and DDC is illustrated in fig. 5.5.

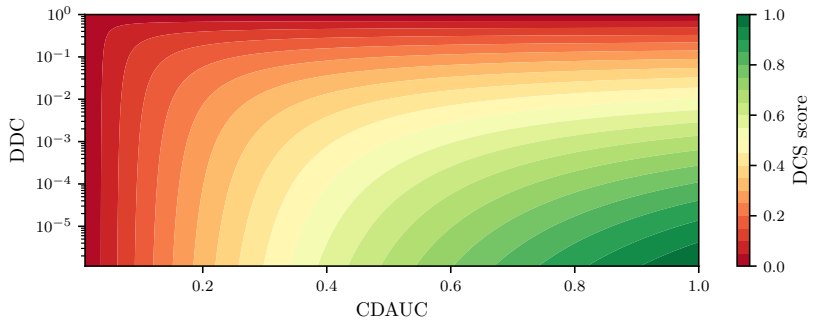


Fig. 5.5: DCS score for hyperparameter tuning based on CDAUC and the log-transformed DDC.

Training Parameters

For DCS, a fixed batch size of 50 is used. All models are trained for 100 epochs. To avoid overfitting during training, early stopping is added with patience of 10 epochs as well as a 20% dropout rate. Also, the CoxPH model is tuned regarding step size and L1 regularization but did not show any model improvement. For hyperparameter tuning and performance evaluation the data is split into 80% training and 20% test data, respectively. The test data was only used during the final evaluation of the models. A Bayesian Hyperparameter Search was performed on the training set using 5-fold cross-validation (CV) with a total of 100 iterations. A composition of CDAUC and DDC as described in sec. 5.2.5 was chosen as the optimization criterion since it yielded reasonable results regarding discrimination while also taking calibration into account. The best model was then retrained on the complete training dataset with the previously found best hyperparameters.

5.3 Results

This section deals with the results that were generated for the MK dataset. Firstly, the patient feature characteristics are discussed before individual feature encodings are analyzed for the most important features using univariate CoxPH and DCS survival models. Further, a multivariate model is trained based on the best found feature representations. Also, the importance of semantically connected features, i.e. feature blocks are presented. The importance is estimated by cumulative inclusion of the feature block that produces the best discriminative performance regarding CDAUC along with previously included feature blocks. Afterwards, downstream analysis is performed on the model predictions that yielded the best discriminative performance by generating risk groups from the predicted survival curves.

Patient Overview

For the filtered dataset that was included in the final analysis, a censoring rate of 78.4 % was observed meaning that 2,526 patients suffered from BCR at some point after RP. Furthermore,

7 features or 21.8 % of the analyzed features violate the PH assumption as shown in appendix D.2.

The median age, preoperative PSA level and prostate volume were 65 years (IQR 59-69), 7.33 ng/mL (IQR 5.2-11) and 25 mL (IQR 20-36) respectively. Furthermore, most patients were diagnosed with a pathological GG of 3+4 (70.1 %), 4+4 (21.9 %) and 4+5 (5.4 %). The observed T-stage (path) of the cohort analyzed for pT2, pT3 and pT4 was 62 %, 37 % and 0.2 % respectively. Furthermore, capsular invasion was observed in 36 %, lymph node invasion in 10 %, lymph vessel invasion in 15 %, positive resection margin in 16 % and seminal vesicle invasion in 14 % of the dataset. A fully detailed table of the PCa-related patient characteristics can be found in tab. A7.

Number of Comparisons

As described in sec. 4.2.3, the number of possible comparisons of individuals within the dataset impacts a survival prediction model's discriminative performance. The DCS model includes more comparisons by including censored individuals which is especially useful for datasets with high censoring rates. For this specific dataset with a censoring rate of 74 %, the ratio of only uncensored (EE) comparisons (n=3.19×10^6) over all possible comparisons (n=4.76×10^7) is 6.7 % while including the EC cases (n=1.84×10^7) boosts this number to 38.6 %. This again emphasizes that the number of comparisons can be increased significantly by including EC cases utilizing 5.8 times more comparisons than the other approaches that only take EE comparisons into account.

5.3.1 Feature Encoding

This section investigates the most important features of the dataset for relapse prediction in additional detail to find the feature representations for a multivariate survival prediction model. Different feature representations are analyzed by univariate CoxPH and DCS models to estimate the predictive value for each representation. Only the best feature representation is used for subsequent analyses.

The MK dataset contains a large variety of PCa related features with multiple representations along with the endpoint definition that are described in more detail in sec. 3.1.4. Since different representations of the same data exist, the best representation for the relapse prediction task of this chapter needs to be found. Therefore, univariate models are trained on the alternative representations. The model and training parameters that were used to find the best representation per feature can be found in tab. A8. Afterwards the feature representation with the highest discriminative score when performing 5-fold CV is selected for further analysis. This section shows and discusses the results of this process for the most interesting features and their representations that are illustrated in fig. 5.6.

PSA level

The preoperative PSA level is one of the most important factors in relapse prediction (see sec. 2.1.4) for RP patients. Since BCR is a recurring rise in this level after RP as described in as described in sec. 2.1.4, the preoperative value contains predictive value [149]. This thesis analyzes three alternative encodings for the preoperative PSA level, namely using the raw feature, using a logarithmic PSA level or calculating the PSA density that relates the PSA level to the prostate volume [211] (with the number of resulting features in parentheses):

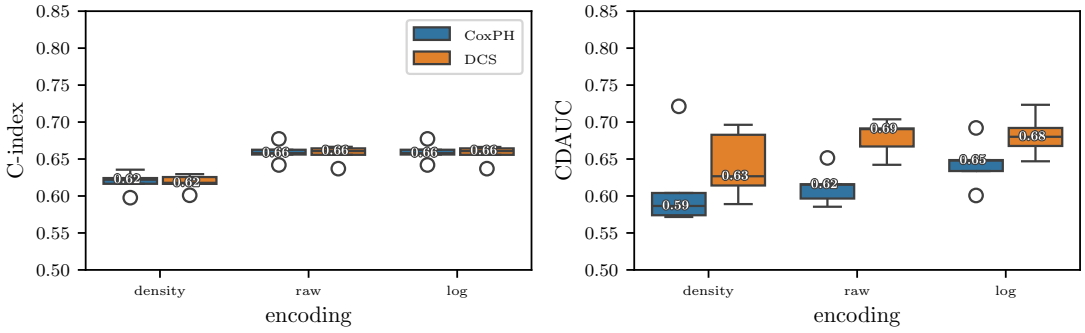
raw (1) The PSA level is provided unchanged in ng/mL

logarithmic (1) As shown in [25, 240], it may be feasible to log-transform the PSA level at least for the CoxPH model

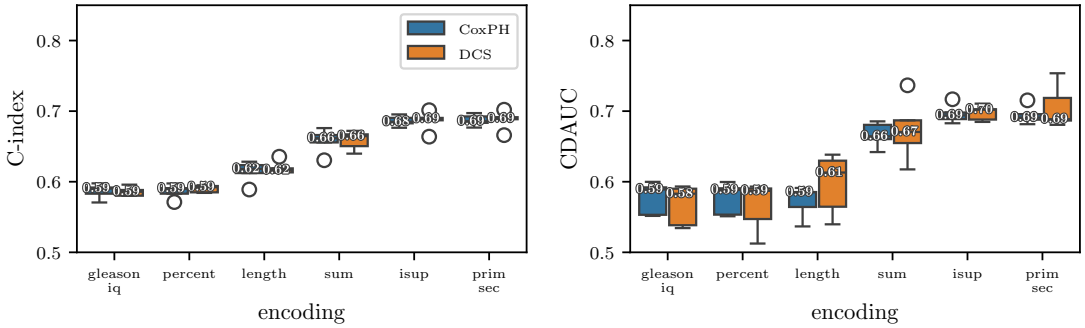
density (1) Normalizing PSA level by the total prostate volume to calculate the PSA density [210, 211]

Fig. 5.6a illustrates the predictive performance of only using one of the above alternative encodings for the PSA level. It can be observed that a raw or logarithmic encoding performs best for the MK dataset. Also note that there is a significant difference in performance between the analyzed models when evaluating the resulting model's CDAUC compared to equal performance regarding C-index-td. It is also evident that log-transforming does not significantly alter the performance of the DCS model, but has a positive effect for the CoxPH approach regarding CDAUC.

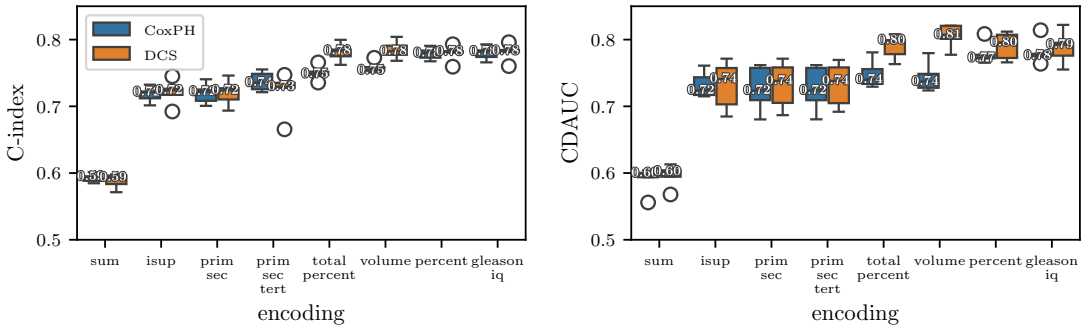
Since the best performance regarding CDAUC is used as the decision criterion, the chosen feature encoding for the preoperative PSA level is using the raw values for the DCS model which achieves a median CDAUC of 0.69.



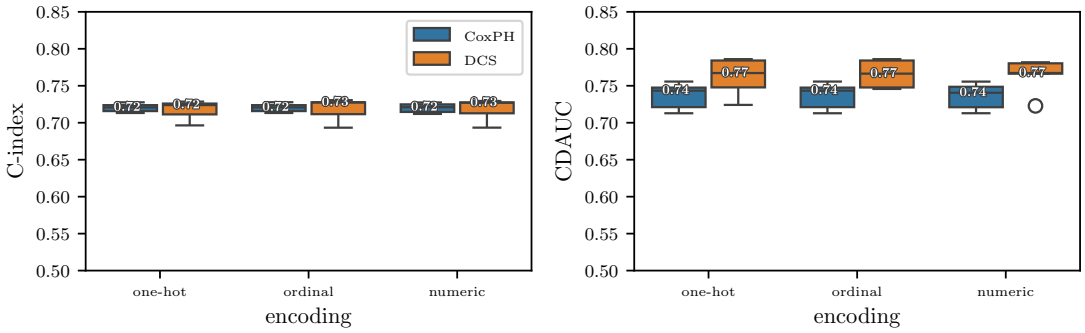
(a) PSA level (at point of RP)



(b) Biopsy-based Gleason score



(c) Pathological Gleason score

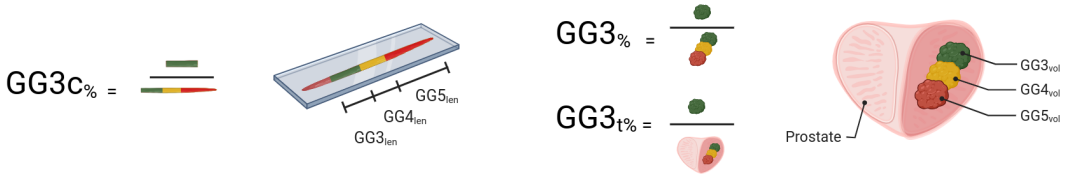


(d) T-stage (path)

Fig. 5.6: Results regarding 5-fold CV C-index (left) and CDAUC (right) for univariate CoxPH (blue) and DCS (orange) models for the different features and feature encodings.

Gleason Grading

The next important factor that is analyzed regarding feature encoding is the Gleason grading described in sec. 2.1.2 for biopsies (clinical GG) and TMAs of the RP patients. It is expected that higher predictive value is found in the pathological GG since it is obtained from the whole prostate instead of only a (potentially non-representative) needle biopsy.



(a) Length-based clinical quantitative GG (b) Volume-based pathological quantitative GG

Fig. 5.7: Visualization of how GG ratios are obtained for the clinical (a) and pathological (b) GG. Created with BioRender.com

Biopsy-based Gleason grading Several encoding strategies with different levels of granularity exist for clinical GG. The obtained grades were documented from the patient's last biopsy prior to RP. In this case, the length (in mm) of the biopsy is used for the quantitative estimates of the different GG encodings (with the number of features in parentheses):

prim+sec (2) Use the primary and secondary GG, namely GGc_p and GGc_s as two independent categorical features. For biopsies, this corresponds to the most and worst GG that was found within the biopsy.

sum (1) The primary and secondary GG are combined into a sum $\text{GGc}_\Sigma = \text{GGc}_p + \text{GGc}_s$ ranging from 6 (3+3) to 10 (5+5).

isup (1) Combine the primary and secondary GGc into the ISUP grading system that was described in sec. 2.1.2 leading to five values ISUP1-5. Further, ISUP0 is additionally defined if no GG3 or above was found.

length (3) The absolute lengths GGc3_{len}, GGc4_{len}, GGc5_{len} in mm.

percent (3) Define $\text{GGc}i_{\%}$ as the length based ratio of one GG to all others as

$$\text{GGc}i_{\%} = \frac{\text{GGc}i_{\text{len}}}{\sum_{j=3}^{5} \text{GGc}j_{\text{len}}} \tag{5.7}$$

iq (1) From $\text{GGc4}_{\%}$ and $\text{GGc5}_{\%}$, calculate GIQ (see sec. 2.1.2) as

$$\text{GIQ} = \text{GGc4}_{\%} + \text{GGc5}_{\%} + 0.1 \cdot \mathbb{1}_{\text{GGc5}_{\%}>0\%} + 0.075 \cdot \mathbb{1}_{\text{GGc5}_{\%}>20\%} \tag{5.8}$$

Fig. 5.6b illustrates the results of the univariate survival models. It can be observed that primary and secondary GG are sufficient to obtain the best discriminative scores regarding relapse prediction for both models along with ISUP encoding. Both approaches yield median scores of 0.68 - 0.70 for both models regarding C-index-td and CDAUC. It is worth noting that the more complex and fine-grained biopsy encodings, namely the length of the different GG inside the biopsy, the transformed percentages, or GIQ showed worse results than both aforementioned approaches with a difference in CDAUC of approximately 8-10 percentage points (pp) with a score of 0.58 - 0.61.

Pathological Gleason grading The MK dataset provides multiple representations of the Gleason grading that originates from the evaluation of a patient's TMA spot by dedicated genitourinary pathologists at the department of pathology of the UKE. Quantitative Gleason grading was performed as initially described in sec. 2.1.2 based on the found volume (in mL) for GG3-5. A visual representation of how $\mathrm{GG}i_{\%}$ and $\mathrm{GG}i_{t\%}$ are obtained can be found in fig. 5.7b. This work compares the following alternative representations with the number of features in parentheses:

prim+sec (2) Use the primary and secondary GG, namely GG_p and GG_s as two independent categorical features. For TMAs, this corresponds to the most and second most GG that was found in the TMA spot.

prim+sec+tert (3) Combine the previously mentioned approach with the tertiary GG.

sum (1) The primary and secondary GG are combined into a sum $GG_\Sigma = GG_p + GG_s$ ranging from 6 (3+3) to 10 (5+5).

isup (1) Combine the primary and secondary GGc into the ISUP grading system that was described in sec. 2.1.2 leading to five values ISUP1-5. Further, ISUP0 is additionally defined if no GG3 or above was found.

volume (3) The absolute volumes $\mathrm{GG3_{vol}}$, $\mathrm{GG4_{vol}}$, $\mathrm{GG5_{vol}}$ in mL.

percent (3) Define $\mathrm{GG}i_{\%}$ as the volume based ratio of one GG to all others as

$$\mathrm{GG}i_{\%} = \frac{\mathrm{GG}i_{\mathrm{vol}}}{\sum_{i=3}^{5} \mathrm{GG}i_{\mathrm{vol}}} \tag{5.9}$$

total percent (3) Define $\mathrm{GG}i_{t\%}$ as the ratio of one GG over the total volume of the removed prostate V_{prostate} (in mL) as

$$\mathrm{GG}i_{t\%} = \frac{\mathrm{GG}i_{\mathrm{vol}}}{V_{\mathrm{prostate}}} \tag{5.10}$$

iq (1) From $\mathrm{GG4_{\%}}$ and $\mathrm{GG5_{\%}}$, calculate the GIQ as described in sec. 2.1.2 as

$$\mathrm{GIQ} = \mathrm{GG4_{\%}} + \mathrm{GG5_{\%}} + 0.1 \cdot \mathbb{1}_{\mathrm{GG5_{\%}>0\%}} + 0.075 \cdot \mathbb{1}_{\mathrm{GG5_{\%}>20\%}} \tag{5.11}$$

As shown in fig. 5.6c, the more complex quantitative representations, namely volume, percent, total percent and GIQ yield the best results for the pathological GG. It can also be observed that the CoxPH model performs worse than the DCS model throughout all representations regarding median CDAUC and especially with the raw tumor volume representation where it lacks 7 pp behind the DCS model. This gap is reduced to 3 pp when the percent encoding is used. This result shows that the non-proportional and dependent features like percentage values can provide more information regarding relapse prediction even though the CoxPH assumptions are not met.

Staging Information

TNM-staging as introduced in sec. 2.1.1 provides a categorical classification of an individual's cancer severity. For the MK dataset, the pathological T-staging contains patients with 6 of a total of 14 different levels, namely pT2c, pT3a, pT3b, pT2a, pT4 and pT2b in descending order of occurrence. This categorical data can be represented in different ways. The following were analyzed with results shown in sec. 5.3.1:

one-hot (13) A patient with T-stage T1b would be assigned the vector $[0, 1, 0, \ldots, 0]$ for the corresponding ordered stages.

ordinal (13) Instead of converting a categorical feature into many unrelated binary indicators, ordinal encoding preserved the ordering. This is achieved by encoding the same number of features by asking "is the current T-stage at least" the corresponding group. This way, feature vectors in the form $[1, 1, \ldots, 0]$ are created for each individual that start with ones on the left as long as the individual's T-stage is at least as bad as the comparing group.

numeric (1) Numeric encoding also preserves the ordering of the groups, but only uses a scalar per individual. This implies that the difference of two adjacent groups is the same throughout the encoded categories.

It can be observed that the encoding of T-stage (path) does not influence the discriminative performance, but that the DCS model outperforms that CoxPH approach regarding CDAUC where a median of 0.77 is achieved for all encodings outperforming the CoxPH approach by 3 pp. Also, it is worth noting that this difference does not show in the C-index-td metric.

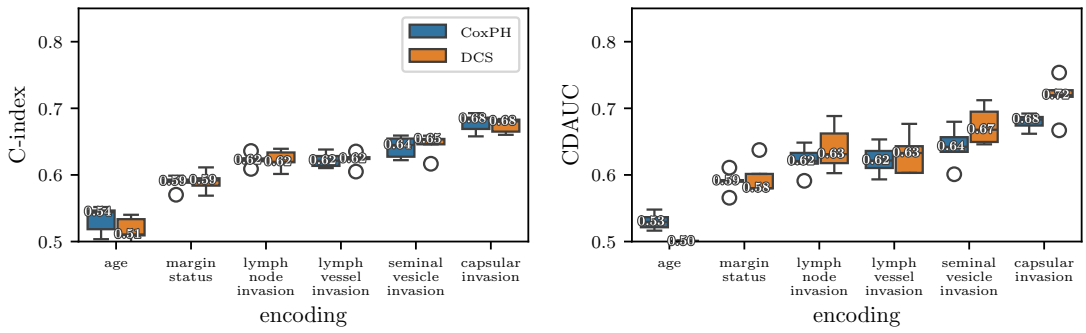


Fig. 5.8: Univariate 5-fold CV results for different predictors shown on C-index and CDAUC performance.

Other Parameters

Other parameters, namely the patient's age, resection status, lymph node invasion, seminal vesicle invasion and capsular invasion are provided to the network as numerical or binary indicators. The individual performances for those features can be found in sec. 5.3.1. It can be observed that capsular invasion shows the highest individual performance also with the largest difference in CDAUC (0.72 for DCS, 0.68 for CoxPH) when comparing the medians the two models.

Summary

The previous section analyzed the individual predictive performance of feature encodings for several parameters that are known to be informative to predict relapse free survival after RP. The chosen features with corresponding encodings are the preoperative PSA level (as a raw feature),

path. GG (with the more complex percentage encoding since it yielded the best median score for both models combined), CI, LNI, positive resection margin and SVI as additional, binary predictors. All other features were discarded from the following evaluation. Specifically the T-stage (path) was excluded since the attributes that define it are already included as individual features (e.g. capsular extension or lymph node invasion) as described in tab. 2.1.

5.3.2 Quantitative Comparison

A quantitative evaluation is performed for the best found features with corresponding encodings. For this, a multivariate DCS and CoxPH model is retrained on the training set and evaluated on the previously unseen test set. To estimate performance variance on the test set, it is bootstrapped 20 times with the same size as the original test dataset. The results are reported as mean ± standard deviation in tab. 5.1.

Tab. 5.1: Performance evaluation of CoxPH vs. DCS regarding discrimination and calibration metrics. (↑) indicates that higher values are better, (↓) that lower values are better.

	CoxPH	DCS
C-index-td (↑)	0.810 ± 0.010	$\mathbf{0.817 \pm 0.009}$
CDAUC (↑)	0.846 ± 0.009	$\mathbf{0.864 \pm 0.011}$
IBrS (↓)	0.127 ± 0.005	$\mathbf{0.120 \pm 0.006}$
DDC (↓)	0.006 ± 0.002	$\mathbf{0.004 \pm 0.001}$

Regarding C-index-td, DCS outperforms CoxPH slightly (0.817 vs. 0.810). When comparing CDAUC, DCS shows higher discriminative performance (0.864) than CoxPH (0.846). Notably, DCS also scores higher regarding calibration performance. Specifically, a higher performance could be observed in IBrS (DCS 0.120 vs. CoxPH 0.127) and DDC (DCS 0.006 vs. CoxPH 0.004). Note that, for IBrS and DDC, lower values are better. In general, the DCS model is able to outperform the CoxPH model.

5.3.3 Block-level Feature importance

Algorithm 4: Feature block importance algorithm.

Input: Set of feature blocks that contain 1 to n features,
metric to evaluate survival model performance

```
remaining_blocks ← all_blocks;
used_blocks ← ∅ ;
all_scores ← ∅ ;
while remaining_blocks ≠ ∅ do
    iteration_scores = [];
    for block in remaining_blocks do
        tmp_features ← used_blocks combined with block;
        tmp_performance ← CV metric trained with tmp_features;
        Append tmp_performance to iteration_scores;
    Set block_best as the block with best performance in iteration_scores;
    Append block_best to used_blocks and remove it from remaining_blocks;
    Append iteration_scores to all_scores;
```

Output: all_scores: CV scores for cumulatively included blocks

As an additional analysis, the selected features are further analyzed in terms of feature importance by feature block. A feature block is a collection of features that semantically belong together.

This means that for example the pathological GGs are not analyzed separately, but together. This analysis also includes additional feature blocks like total prostate and cancer volume, operation parameters (that includes blood loss) or additional demographic information.

In the approach depicted in alg. 3, a survival prediction model for all feature blocks is trained and choose the one with the best discriminative performance. Then subsequently add the feature block that achieves the best discriminative performance together with the previously found features. This process is illustrated for the DCS model in fig. 5.9. It can be observed that the pathological GG already leads to a C-index-td of over 0.8. The next added feature blocks are preoperative PSA level, seminal vesicle invasion and lymph node invasion before the score is saturated.

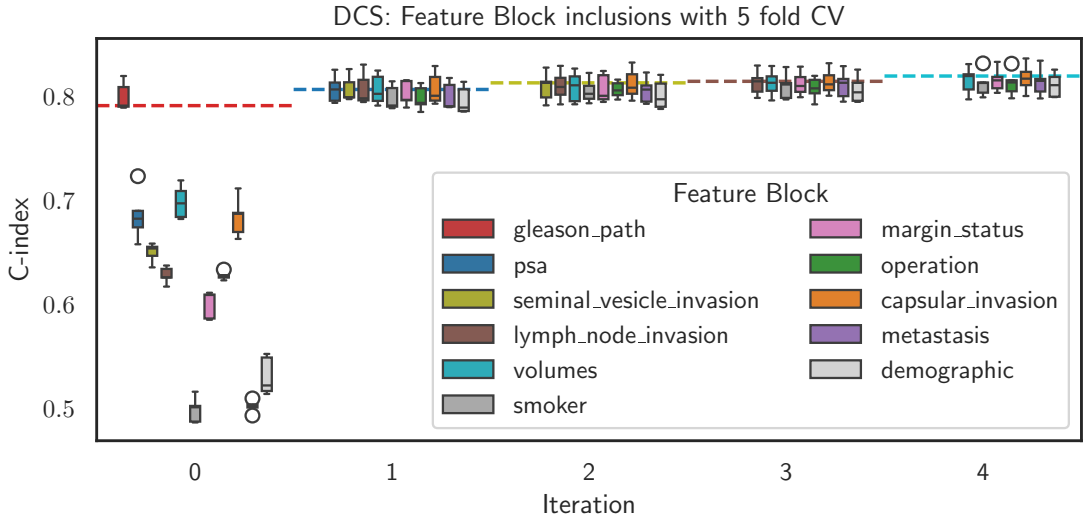


Fig. 5.9: Block-wise feature importance for the MK dataset. For every iteration, take the previous best feature blocks and calculate a new multivariate DCS model regarding relapse prediction evaluated on five folds for all remaining features. Include the additional feature of the best multivariate model in the subsequent iterations. The dashed line is on the height and in the color of the winning feature block for this iteration.

5.3.4 Risk Stratification

The found DCS model and the resulting individual relapse risk predictions can also be used for risk stratification. To obtain distinct risk groups, the patient's predictions are stratified as discussed in sec. 5.2.4 using K-means clustering. Out-of-training patients can then be labeled based on the closest cluster center to their individual survival curve prediction. This method of retrieving risk groups from the predictions enables the usage of a flexible number of risk groups K that can be used as an input for the K-means-clustering algorithm. The first approach is to separate the individuals into a low and a high risk group with $K = 2$ as illustrated in fig. 5.10. When comparing DCS risk groups to CoxPH risk groups that were obtained from the individual hazard rates, one can see that DCS is able to identify approximately double the number of patients in the high risk group while having a more optimistic KM-curve for the low risk sub-cohort. Furthermore, the high risk KM curve of DCS shows a tighter 95 % confidence interval throughout the observed time frame.

DCS identifies 685 of 1905 patients in the high risk group. This includes all 359 patients that were also identified by the CoxPH model thus containing 326 additional patients. Fig. 5.10 illustrates that the overall survival prediction of the DCS high risk group is higher than the corresponding prediction from CoxPH. The same holds true for both low risk groups. However, investigating e.g. the 6 month mark after RP, the identified low risk group experienced an observed 6-month relapse rate of 0.66 % (compared to 1.68 % for CoxPH) whereas 18.7 % (30.64 % for CoxPH) of the high isk sub-cohort suffered BCR in the first six months. This means that the observed relative frequency of BCR in the high risk sub-cohort was approximately 28 times higher (18 times with CoxPH) than in the low risk cohort for the DCS model.

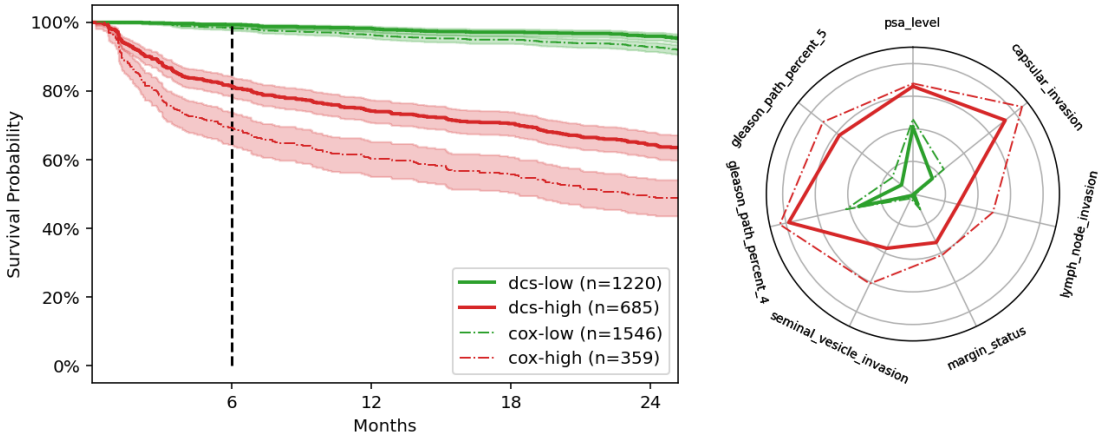


Fig. 5.10: KM curve of the first 2 years after RP for previously unseen test set patients stratified by DCS and CoxPH risk groups and the 6-month mark after RP (left) and the corresponding feature plot (right) that shows the median quantile per feature for the respective model and risk group.

When comparing the features within the risk groups, the following can be observed: DCS high risk patients show higher values for $GG5_{\%}$, LNI and SVI while the percentile in the low risk group for $GG4_{\%}$, PSA, capsular invasion is also different from 0 for the low risk group. This demonstrates that even though a patient has a relatively high $GG4_{\%}$, high PSA level or capsular invasion, he might still not have an elevated risk of relapse regarding the aforementioned analysis.

Maximium Risk Groups

Using the ideas presented in sec. 5.2.4, the number K of distinct risk groups can be increased as long as a difference in survival regarding relapse is still present within the observation window of

the individuals. Following the aforementioned approach of increasing the number of risk groups, KM-curves can be observed for a separate validation set, a maximum of $K = 7$ distinct risk groups can be separated with the survival curves that were predicted by the DCS model. The resulting p-values of $K = 7$ groups for all pairwise log-rank tests are shown in fig. 5.11. It can be observed that p-values of adjacent groups show the highest values.

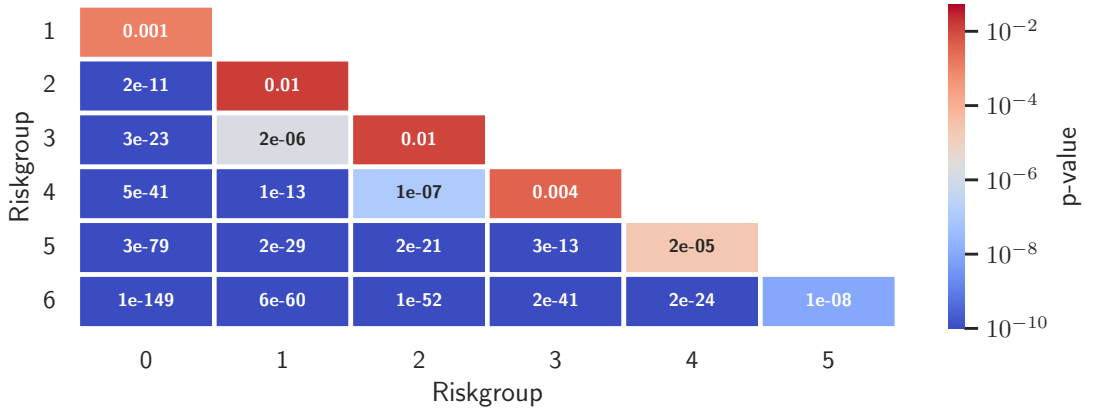


Fig. 5.11: Pairwise log-rank test p-value results for the maximum number of statistically significantly different risk groups obtained from the DCS predictions.

The resulting risk groups of this algorithm on the MK dataset are visualized in fig. 5.12 using the survival information of the unseen (for the risk and clustering algorithm) test set. Further, the risk groups can be analyzed in terms of feature composition. For each of the risk groups, the average values of selected features are visualized in fig. 5.13. It can be observed that certain factors like LNI or pathological GG5 are only present in the three most severe risk groups. Additionally, there is only a small difference in the preoperative PSA level between risk groups 0-3 and slowly increasing mean PSA levels from group 4-6. Furthermore, SVI and LNI are not present in risk groups 0-3 while there are patients, especially in groups 1-3 with positive resection margin status and extracapsular extension. It can be observed that the mean of $GG3_{\%}$ gradually decreases with the risk group while $GG4_{\%}$ and $GG5_{\%}$ increase. Also, $GG5_{\%}$ is only significantly present in risk groups 4 and above.

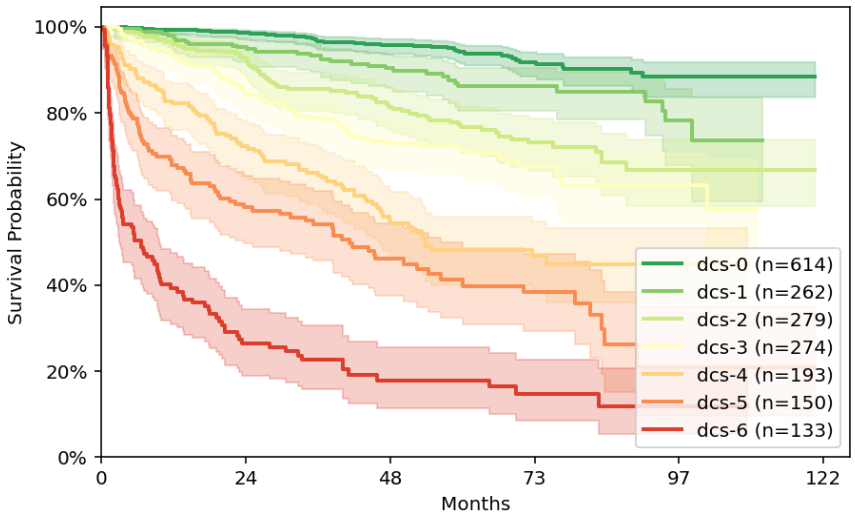


Fig. 5.12: KM-curves of the 7 obtained risk groups for the previously unseen test set patients of the MK dataset.

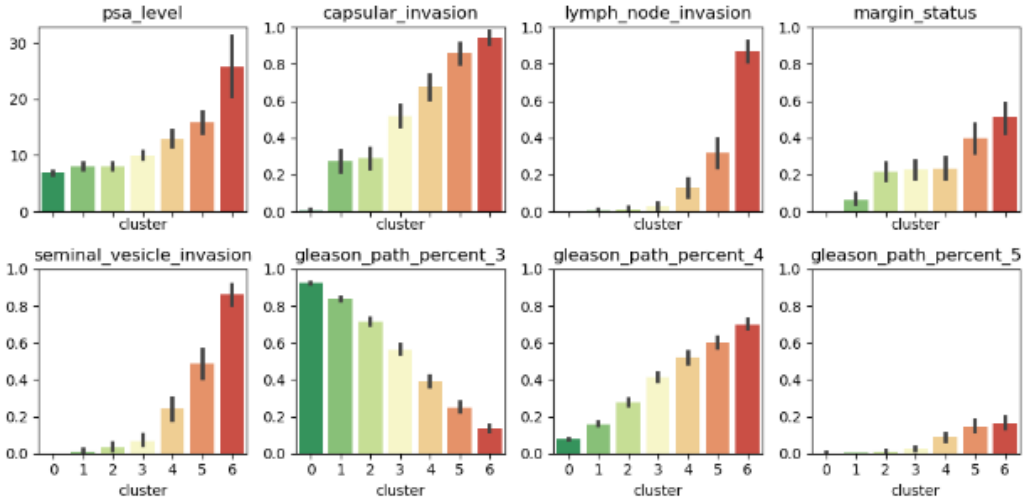


Fig. 5.13: Feature composition after risk grouping the MK dataset showing the mean and standard deviation of the selected features.

5.4 Discussion

This chapter analyzed how the DCS model can help in survival prediction for PCa patients that were treated with RP in terms of BCR prediction. In the dataset, 7 of the 32 features violate the PH assumption and a high censoring rate of 78.4 % is observed.

As shown in the previous chapter, DCS utilizes as much information as possible from the censored cases and does not constrain usage of features that violate the PH assumption allowing time-variant and non-linear interactions. Since censored individuals are fully utilized, the total number of comparisons could be raised by a factor of 5.8 (6.7 % for only EE comparisons to 38.6 % for DCS that includes EC comparisons) for the given dataset.

The results for relapse prediction based on univariate survival models of CoxPH and DCS indicate that some features can better be utilized by the more complex DCS model. For example, this behavior can be observed when only the PSA value is used for relapse prediction. The univariate DCS models showed better median performance with a difference of up to 7 pp regarding CDAUC for the raw PSA value. This difference could not be observed for the biopsy-based Gleason grade encodings where almost identical performance was achieved for the two models on all encodings. However, for the pathological Gleason grade, the more complex encodings lead to better performance in the DCS model with up to 7 pp difference regarding CDAUC when the volume of the different GGs were used. One reason for this behavior could be that the clinical GG are obtained for the individuals outside the Martiniklinik and thus show high variability in terms of quality. This may lead to the loss of additional information in terms of relapse prediction from quantitative Gleason evaluations to noise in the input annotations. Pathological GG annotations are exclusively obtained by the department of pathology of the UKE that provides a high quality standard with less variation.

Moreover, since DCS allows for crossing survival curves, the more complex metrics C-index-td and CDAUC had to be used to measure discriminative performance instead of the C-index. In general, differences for the two models can better be observed for CDAUC then C-index-td throughout the analysis. This can for example be observed for the T-stage (path) where a univariate analysis yields the same discriminative performance regarding C-index-td but a difference of 3 pp when CDAUC is compared.

A better result for DCS over CoxPH can also be observed for the multivariate models that combine the best individual feature representations. In terms of discrimination, DCS slightly outperformed CoxPH on C-index-td (0.817 vs. 0.810) and CDAUC (0.864 vs. 0.846). This also holds true for calibration that was measured with IBrS (0.120 vs. 0.127) and DDC (0.004 vs. 0.006). This result differs from the previous chapter where DCS lacked behind the performance of CoxPH in terms of calibration. It can be explained by the altered hyperparameter tuning that was used. In contrast to the previous chapter, DCS was not only optimized towards CDAUC, but DDC was additionally taken into account with the introduction of the DCS score (see sec. 5.2.5). This way the hyperparameter tuning was able to find a combination of parameters that shows the best performance regarding both metrics. The block-wise feature importance algorithm showed that only using the path. GG already led to good discriminative performance of above 0.8. Additional inclusion of PSA, SVI, and LNI can lead to minor improvements, but quickly leads to a saturation at 0.817 for C-index-td. This further shows the importance of proper pathological Gleason grading.

It would additionally be expected that CoxPH performs worse than DCS on features that violate the PH assumption. This is in general not the case for independent features since for example the univariate survival models for LNI (that fails the PH test) yield almost identical discriminative performance for both models (C-index-td of 0.62 for both, CDAUC is 0.62 for CoxPH, 0.63 for DCS). The same can be observed for resection margin status that also violates the PH assumption where CoxPH (0.59) even slightly outperforms DCS (0.58) regarding CDAUC. On the other hand, capsular invasion shows a difference in CDAUC of 4 pp between CoxPH (0.68) and DCS

(0.72) even though it passes the PH test with a p-value of 0.373 that lies comfortably above the threshold that was used in this work of 0.05.

The downstream analysis, dividing the individuals into a low and a high risk group yielded additional inclusion of 326 patients into the high risk groups when DCS was used over CoxPH. This way the survival of the low risk group could further be improved where only 0.66 % of individuals experience a relapse after six months. This further shows the discriminative abilities of DCS that finds additional patients at higher risk inside the low risk cohort that was identified by the CoxPH model. When this approach was used to find the maximum number of statistically significantly distinguishable sub-cohorts, a total of 7 groups could be obtained. This further generated insights on the distribution of feature characteristics throughout the groups. Features like $GG5_{\%}$, LNI or SVI are almost exclusively found in groups 4-6 indicating that their presence yields to a significantly lower chance of relapse-free survival.

5.5 Conclusion

This chapter analyzed BCR relapse prediction after RP based on the Martiniklinik dataset for the containing 9,767 PCa patients that received RP between 2008 and 2018 using the DCS model in terms of relapse prediction. DCS can boost discriminative and calibration performance compared to CoxPH that is commonly used in clinical practice on the given dataset. Optimal feature representations for the most commonly used factors that contain information towards relapse prediction were obtained and combined for multivariate survival models. For the univariate analysis, Gleason grading showed the highest results regarding discriminative power. While the best performance for the biopsy-based GG was achieved with primary and secondary GG, it was shown that a more complex pathological GG that additionally includes quantitative information increased the discriminative ability. When combining the best found feature representations into a multivariate model, DCS outperformed CoxPH. Further, it could be shown that the quantitative encoding for the path. GG as an individual feature already reaches nearly the same discriminative performance as a model that includes PSA value, tumor characteristics and other characteristics. The discriminative performance of DCS was further verified by constructing 7 statistically significantly distinguishable risk groups that show good separation regarding patient survival using the pairwise log-rank test.

For future work, the presented approach can be evaluated in a prospective study to further ensure the validity of the approach. The tool can further be integrated in the clinical workflow to improve the communication between the practitioner and patient. Especially the predicted risk groups with corresponding survival curves can be used to inform the patient about his status by assigning him one of the predicted risk groups based on his current parameters. Since the model is well calibrated and would predict on the same patient cohort in the same center, it is unexpected to observe significant patient bias for future predictions. This way the well calibrated survival probabilities can also be used in patient communication the same way that nomograms offer. A sentence like "If we had 100 patients exactly like you, we would expect X of them to remain relapse free after 10 years" could be formulated to further inform the patient about his current status. The value for X can be found using the KM estimation of the corresponding risk group of similar patients. The additional confidence interval that is provided by this estimation can further be used to communicate lower and upper boundaries for the relapse free survival probability to improve trust of the patient in the decisions that were made with the practitioner.

Moreover, the risk groups may also be useful for the practitioner in terms of follow-up examination planning. It could be used to shorten or prolong intervals between follow-up examinations. Since the identified low risk sub-cohort shows nearly no relapse, the interval can be prolonged for those patients while high risk patients would benefit from smaller intervals and resulting earlier relapse detection.

6 Patch-based Cancer Classification on Whole Slide Images

6.1 Introduction

The next two chapters shift the focus towards Gleason grading itself without taking any other modalities into account. GG estimates the cancer severity from H&E stained histopathological images of the prostate. In contrast to Gleason grading where a pathologist inspects the tissue, this and the following chapter automate this process.

Firstly, this chapter presents the development of the CI model that is able to classify sections, or patches, of biopsy images as cancerous or healthy tissue. The main purpose of the CI model is to function as a patch selector for PCAI described in chapter 7 to identify the most relevant (or cancerous) regions of biopsy samples for risk assessment from up to tens of thousands of patches per biopsy WSI. For model development, PCa biopsy slides and segmentation masks from the PANDA dataset as described in sec. 3.2.3 are used. For this work, the GG-specific GT segmentation masks of this dataset are used to generate a single mask representing cancerous area of the tissue for all patches as a basis for the binary cancer classification label.

Since most other related work focuses on predicting GG or ISUP on WSIs as presented in sec. 2.5.3, publications that predict patch-wise cancer classification independent of Gleason grading are rare. However, a similar algorithmic approach can be found in [200] that developed a patch-wise classification model called `HistoCAE` for WSIs of liver tissue. Detailed segmentation masks of pathologists are then used to extract patch-wise labels based on the cancerous area inside each given patch. Additionally, this chapter also builds upon ideas published by colleagues in [245] that utilize a patch preselection model for PCa risk estimation.

6.1.1 This work

This work builds the CI to predict a patch-wise PCa-cancer label on WSIs that can afterwards be used as a coarse segmentation mask. It utilizes an `Efficientnet-b0` backbone to encode 256x256 pixel sized patches combined with a binary classification head that utilizes a weighted binary cross-entropy (BCE) loss. In total, the CI model is trained and evaluated on over 5.5 million extracted and labeled patches from 10,616 WSIs split on slide level into 8,492 (80 %) training-, 1,062 (10 %) validation- and 1,062 (10 %) test WSIs. As a result, the model achieves an overall patch-based AUROC of 0.94 on the unseen test set biopsy WSIs.

6.2 Methods

The following section describes the methods that were used to create the CI model starting with the data derived from the PANDA dataset (see sec. 3.2.3) including preprocessing to extract patch-wise labels, explaining the model architecture and objective function as well as the final experimental setup.

6.2.1 Dataset

The PANDA dataset (as introduced in detail in sec. 3.2.3) contains 10,616 WSI biopsy slides from 2,113 patients along with corresponding segmentation masks. The masks differ by the providing medical centers. On the one hand, the Karolinska Institute (KAR) in Stockholm, Sweden provides human-annotated, individual masks of the background, tissue and cancerous areas. On the other hand, Radboud (RAD) University Medical Center in Nijmengen, Netherlands contributes additional, more fine-grained segmentation masks that were generated by an AI model [36] with individual masks for GG3-5. Regarding metadata of this dataset, no further information about the relapse of those patients is given.

6.2.2 Preprocessing

This work extracts individual foreground patches from the provided slides based on two ground-truth masks that are used to firstly select each individual patch based on the amount of tissue that is present on the patch and secondly assign a label derived from the relative amount of cancerous area that is present on each individual patch. This way, the patches of a WSI that do not contain enough tissue are discarded and not used during training and evaluation. The selected patches further receive a label indicating if the patch shows healthy or cancerous tissue. This approach is described in more detail in the following.

Patch Selection

Since the biopsy WSIs of the PANDA dataset vary in size from approximately 10 million to 3 bilion pixels, they are not processed as a whole. Patches are extracted from each slide and afterwards individually processed. To generate equally sized patches, the images are cut into corresponding patches with a side length of p_s pixels starting in the top left corner of the slide. Further, the bottom and right part of the image are truncated so that the image width w_i and height h_i are a multiple of the patch size p_s. The developed CI model of this thesis uses a fixed patch size of $p_s = 256$ pixels.

Formally, for all n slides of a dataset $\{\mathbf{S}_i \mid i \in \{0, 1, \ldots, n-1\}\}$, the i-th slide $\mathbf{S}_i \in \mathbb{R}_+^{h_i \times w_i \times 3}$ with individual side width $w_i \in \mathbb{N}_+$, height $h_i \in \mathbb{N}_+$ and 3 red, green and blue (RGB) color channels is cut into patches $\mathbf{P}_{(i,m,n)} \in \mathbb{R}_+^{p_s \times p_s \times 3}$ with constant square patch width and height $p_s \in \mathbb{N}_+$. Here, $\mathbf{P}_{(i,m,n)}$ denotes a patch of slide $\mathbf{S}_i$ in the m-th row and the n-th column of a grid with a horizontal and vertical distance p_s. For each slide $\mathbf{S}_i$, a total of $\lfloor w_i/p_s \rfloor$ times $\lfloor h_i/p_s \rfloor$ patches can be extracted.

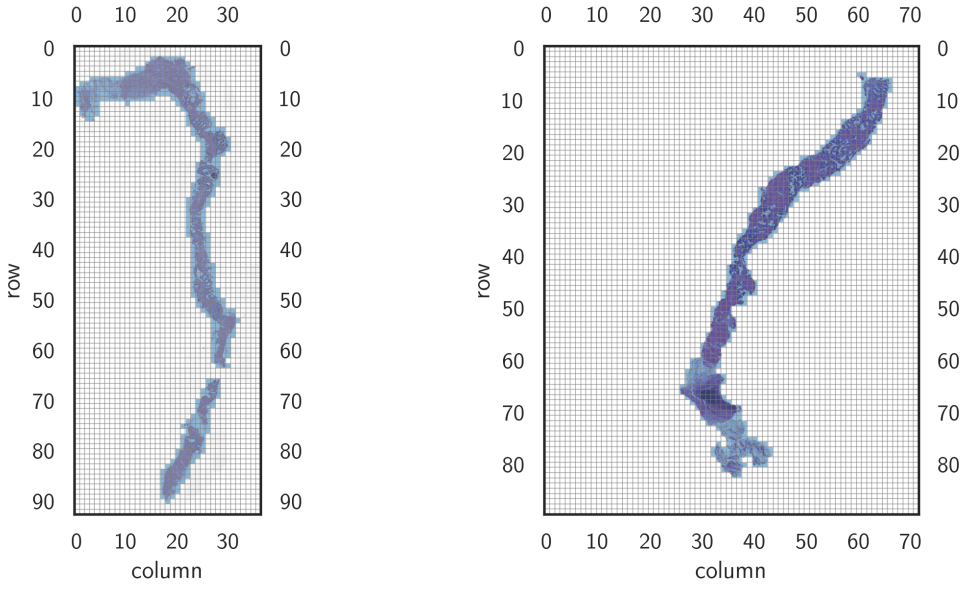
In the next step, patches that do not contain a minimum amount of tissue are discarded based on the corresponding tissue mask for each slide. Note that the original GT masks are downsampled by a factor of 16 compared to the WSI. This is why the tissue and cancer masks are upsampled by nearest neighbor interpolation to generate masks of the same size as the provided actual biopsy slides. Formally, let a patch $\mathbf{P}_{(i,m,n)}$ of the i-th slide in the m-th row and the n-th column be selected according to the corresponding tissue segmentation mask $\mathbf{T}_i \in \{0, 1\}^{h_i \times w_i}$ that contains a 1 for a pixel inside the tissue mask and 0 otherwise. Further, let $\mathbf{T}_{(i,m,n)} \in \{0, 1\}^{p_s \times p_s}$ be the corresponding tissue mask patch that corresponds to $\mathbf{P}_{(i,m,n)}$. A patch is selected if the covered

tissue area of the whole patch is greater than a threshold as

$$\frac{|\mathbf{T}_{(i,m,n)}|}{p_s^2} \geq T_{\text{th}} \tag{6.1}$$

where $T_{\text{th}} \in [0, 1]$ and $|\mathbf{M}|$ denotes the number of non-zero entries in $\mathbf{M}$. In this thesis, $T_{\text{th}} = 0.1$ is used. This means that at least 10 % of pixels per patch must contain tissue to be selected as a foreground patch.

The selected patches contain enough information to be processed further, and are called foreground patches as shown for two examples in fig. 6.1. It can be observed that the presented slides have different sizes and a large amount of patches only contain background that are therefore discarded. Moreover, a distribution of how much foreground patches remain per biopsy slide in the PANDA dataset can be found in fig. 6.2. For the whole dataset, the maximum number of extracted foreground patches is 2844 and the mean is at approximately 466 patches per slide.



(a) RAD (b) KAR

Fig. 6.1: Exemplary WSIs of the PANDA dataset from RAD (left) and KAR (right) with a patch grid of 256 pixels side length and the corresponding patch row and column coordinates. The WSIs vary in overall size leading to different amounts of patches that can be extracted per image. Foreground patches according to the GT segmentation masks are highlighted with a blue overlay.

Cancer Mask Preprocessing

Along with the tissue mask that is provided for each slide, the cancer indication mask is utilized to provide patch-level label definitions. However, the cancer masks of the two centers KAR and RAD differ in granularity and therefore require a preprocessing step. While WSIs from RAD contain detailed, pixel-level segmentation masks of the different GGs, KAR offers coarser outlines of cancerous or healthy regions mostly for whole pieces of tissue. To unify those two centers in terms of mask granularity, the different GG of the RAD slides are combined to represent a single cancer segmentation mask. However, the cancer mask granularity still differs significantly as illustrated in fig. 6.3. Following preliminary work from [245], the RAD masks are processed such that they look more like those from KAR as shown in fig. 6.3c using morphological opening and closing.

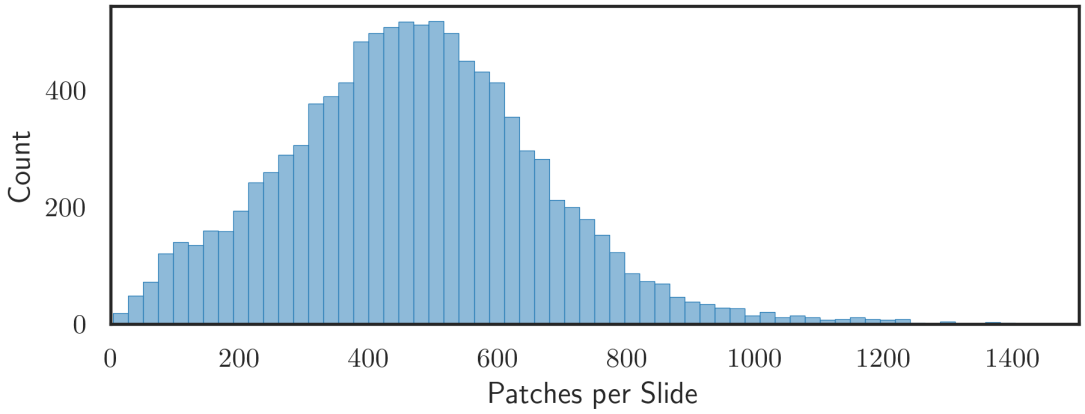
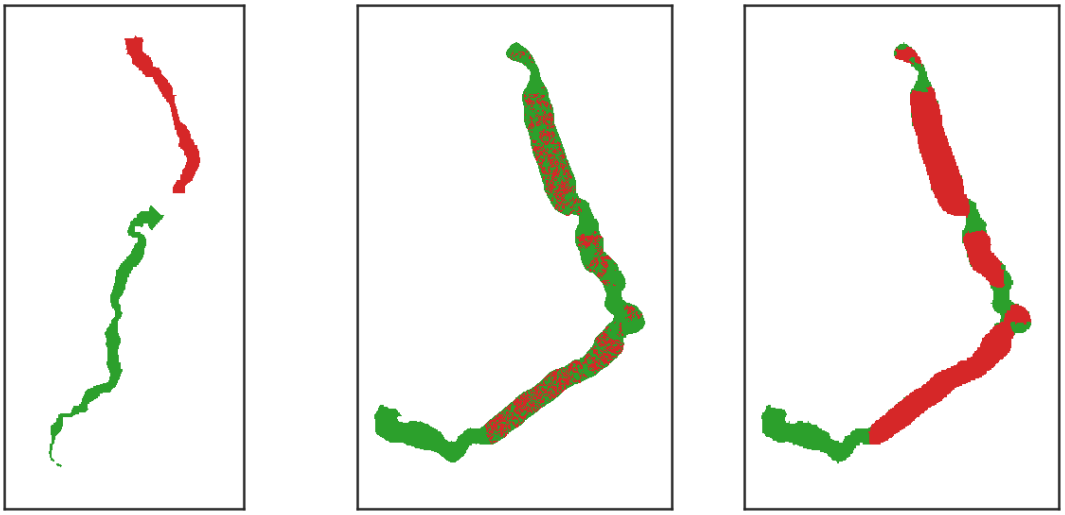


Fig. 6.2: Distribution of the extracted number foreground patches with a side length of 256 pixels per slide in the PANDA dataset.



(a) GT human-annotated mask provided from KAR where the whole tissue parts are annotated.

(b) AI-generated mask from RAD showing pixel-level cancer annotation.

(c) Preprocessed mask based on (b) with adapted granularity.

Fig. 6.3: Exemplary GT masks for the two centers of the PANDA dataset. Tissue that was annotated as healthy is shown in green, cancerous tissue in red, and background in white.

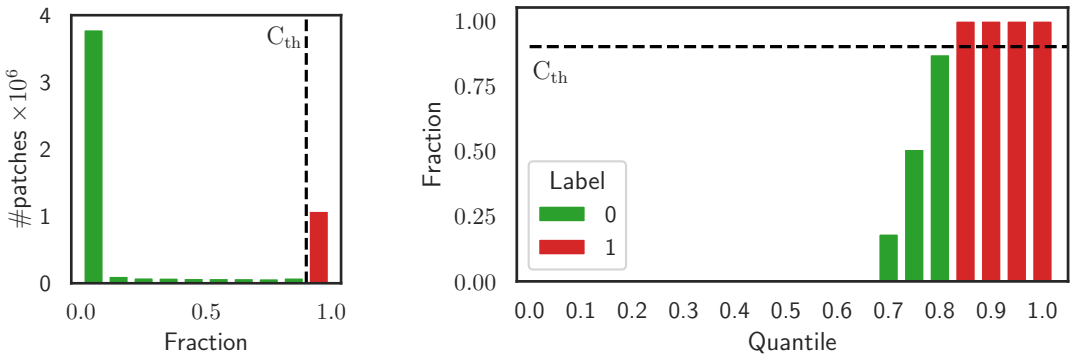
Label Generation

After the preprocessing step for the cancer masks of RAD, the resulting cancer masks can be utilized to define a patch-based label for each selected patch of the dataset.

Let the cancer segmentation mask $\mathbf{C}_i \in \{0,1\}^{h_i \times w_i}$ contain a 1 for the regions with cancerous, and a 0 for healthy tissue area. Moreover, let $\mathbf{C}_{(i,m,n)} \in \{0,1\}^{p_s \times p_s}$ denote the corresponding cancer mask for patch $\mathbf{P}_{(i,m,n)}$. This work assigns the binary cancer indication label $y_{(i,m,n)} \in \{0,1\}$ based on the thresholded area of the cancerous region in $\mathbf{C}_{(i,m,n)}$ normalized by the tissue region $\mathbf{T}_{(i,m,n)}$ of a specific patch as

$$y_{(i,m,n)} = \begin{cases} 1 & \text{if } \frac{|\mathbf{C}_{(i,m,n)}|}{|\mathbf{T}_{(i,m,n)}|} > C_{\text{th}} \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

where $C_{\text{th}} \in [0,1]$ defines the threshold for the patch to obtain an assigned label of 1. By normalizing on the amount of tissue on a patch $|\mathbf{T}_{(i,m,n)}|$ instead of the total patch area, it is ensured that the cancer label is relative to the amount of tissue that is actually shown in a patch. This proposed definition ensures that only regions that actually show tissue are relevant for the label generation process. For illustration, fig. 6.5 and fig. 6.6 show two exemplary slides of the two providing centers. It is noticeable that the mask provided by KAR misses parts of the tissue area leading to wrongly rejected patches while the preprocessed masks from RAD shows extension of the tissue mask into the background.



(a) Histogram of the cancerous fraction for all extracted patches

(b) Quantile plot of the observed cancerous fractions per patch. Patches up to a quantile of 0.6 contain no cancerous area while mixed fraction are found up to a quantile of 0.85.

Fig. 6.4: Distribution of the observed cancerous fraction per patch. The chosen label threshold of $C_{\text{th}} = 0.9$ is indicated with the dashed black line and defines the coloring of the bars according to the assigned healthy (green, $y = 0$) or cancerous (red, $y = 1$) label.

The distribution of the fraction $|\mathbf{C}_{(i,m,n)}|/|\mathbf{T}_{(i,m,n)}|$ is illustrated in fig. 6.4. In total, approximately 4 million patches contain a pure label of $y = 0$ and 1 million the label $y = 1$. This leaves approximately 0.9 million patches (20 %) with mixed areas of healthy and cancerous tissue.

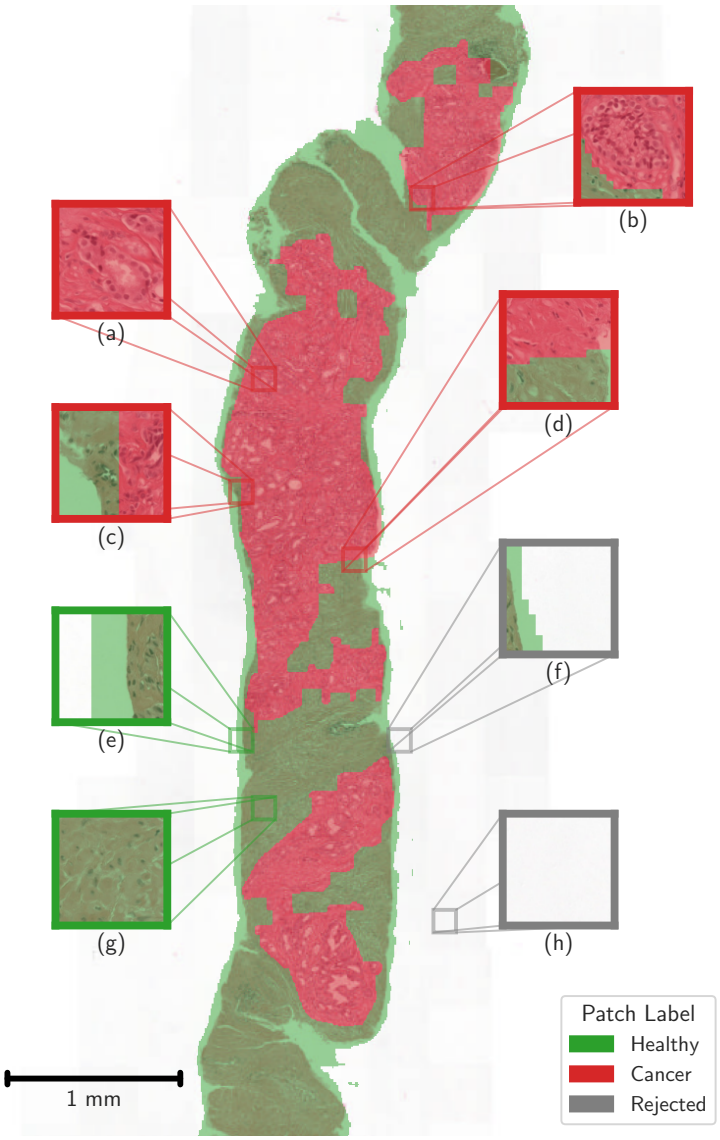


Fig. 6.5: WSI with overlapping cancer and tissue mask from RAD. Patches (a) and (g) only contain cancerous and healthy tissue respectively. The segmentation masks show non-zero fractions in (b), (c), and (d) where all three contain enough cancerous area for an overall label of $y = 1$. Patches (e) and (f) contain both foreground and background, with patch (f) lacking enough foreground to be selected. Additionally, the foreground segmentation masks for these patches extend into the white background. Patch (h) contains no tissue at all and is therefore rejected.

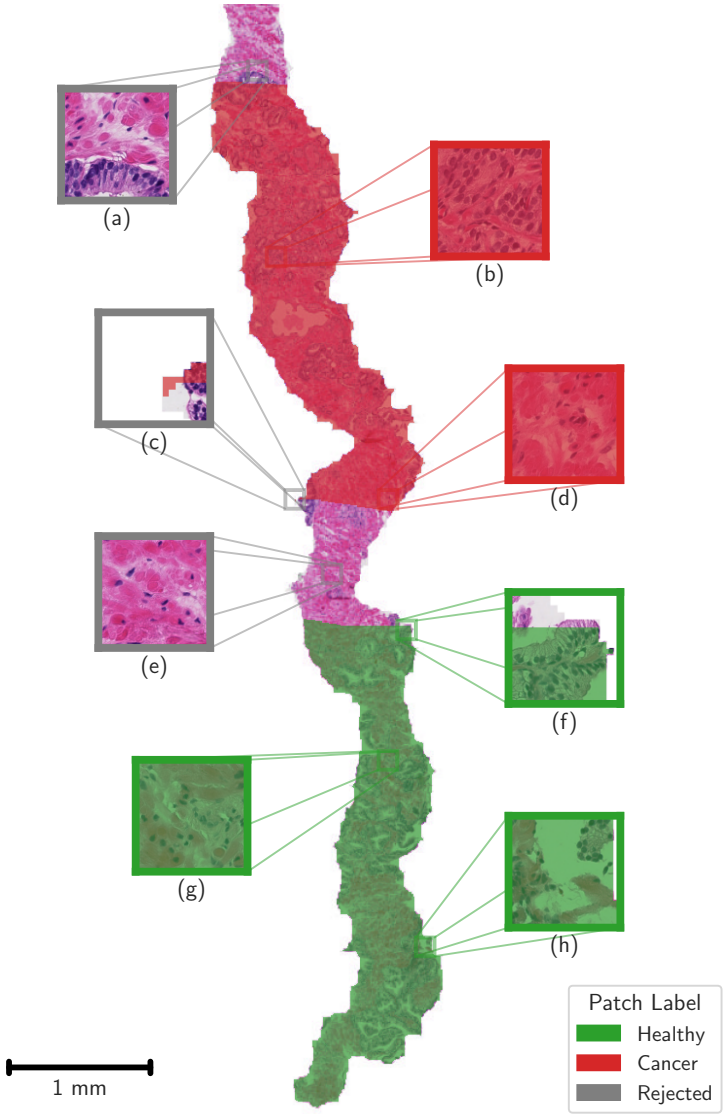


Fig. 6.6: WSI from KAR where the annotated cancer and healthy tissue areas do not overlap. Note that the tissue mask misses a significant part of the visible tissue, leading to the rejection of patches (a) and (e).

6.2.3 Model

This section describes the building blocks of the CI model in additional detail. The DL model combines a CNN-based encoder network with a classification head that can then be trained for patch-wise classification of the labels that were presented in the previous section.

Architecture

The architecture of the CI model consists of a CNN-based image encoding network and a prediction part. The patches of an image are processed individually meaning all information of neighboring patches is discarded for the patch-level label prediction of cancerous vs. healthy tissue. This is why this section ignores the slide origin and coordinate information of all n_P patches $\mathbf{P}_i$ where $i \in \{0, 1, \ldots, n_P - 1\}$.

For the encoder, this thesis chooses the CNN-based `EfficientNet-b0` architecture (see sec. 2.5.3) and combines it with a dense layer to generate two output nodes to predict the probability of the patch showing cancerous or healthy tissue. The encoder part of the network achieved good performance on histopathological images [83, 122, 169] and showed promising initial results in this work's setting compared to `ResNet` [102] and `InceptionV3` [224]. The selected encoder architecture comprises a total of approximately 6.5 million trainable parameters.

An extracted patch $\mathbf{P} \in R_+^{p_s \times p_s \times 3}$ is passed to the encoder that transforms the input patch with

$$\boldsymbol{e} = \text{emb}(\mathbf{P}) \tag{6.3}$$

where $\text{emb} : R_+^{p_s \times p_s \times 3} \to \mathbb{R}^{d_{\text{emb}}}$ defines the function that transforms an input patch $\mathbf{P}$ to a latent representation vector $\boldsymbol{e} \in \mathbb{R}^{d_{\text{emb}}}$. Afterwards, the classification head of the architecture is used to transform the latent representation $\boldsymbol{e}$ to a two-dimensional output vector

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\mathbf{W}\boldsymbol{e} + \boldsymbol{b}\right) \tag{6.4}$$

where $\mathbf{W} \in \mathbb{R}^{2 \times d_{\text{emb}}}$, $\boldsymbol{b} \in \mathbb{R}^2$ are learnable parameters and softmax is defined element-wise as

$$\hat{y}_j = \text{softmax}(o_j) = \frac{\exp\left(o_j\right)}{\exp\left(o_0\right) + \exp\left(o_1\right)} \tag{6.5}$$

where $\boldsymbol{o} = \mathbf{W}\boldsymbol{e} + \boldsymbol{b}$ is the output of the dense layer with $\boldsymbol{o} \in \mathbb{R}^2$ and j is either 0 or 1 for healthy or cancerous tissue respectively. $\hat{\boldsymbol{y}} \in \mathbb{R}^2$ generates the final output of the network architecture with the respective predicted probabilities that the input patch contains healthy ($\hat{y}_0$) or cancerous ($\hat{y}_1$) tissue.

Objective Function

To train the CI model, the objective function is defined as a weighted BCE loss as

$$\begin{aligned} \mathcal{L}_{\text{WBCE}} = \frac{-1}{w_0 + w_1} \big[& w_0 \left(y \log(\hat{y}_1) + (1 - y) \log(1 - \hat{y}_1) \right) + \\ & w_1 \left((1 - y) \log(\hat{y}_0) + y \log(1 - \hat{y}_0) \right) \big] \end{aligned} \tag{6.6}$$

where y corresponds to the true generated label of each patch that was defined in sec. 6.2.2 and $w_0 \in \mathbb{N}_+$, $w_1 \in \mathbb{N}_+$ are the number of negatively and positively labeled samples in the training dataset, respectively.

6.3 Experimental Setup

For training, the PANDA dataset was split into training, validation, and test sets with 80 % (8,492 slides), 10 % (1,062 slides) and 10 % (1,062 slides) of the total of 10,616 slides respectively. By splitting on slide-level instead of patch-level, leakage between the different data splits is ensured.

The highest available magnification level of 20x (with a resolution of 0.486 µm/pixel) with a patch size $p_s = 256$ pixels was used. These parameters resulted in an extraction of 3,949,222 training, 510,452 validation, and 504,027 test patches respectively. This thesis uses a tissue threshold $T_{\text{th}} = 0.1$ and a label threshold of $C_{\text{th}} = 0.9$ meaning that the final CI model was trained on patches where the tissue covers at least 10 % of the overall patch with at least 90 % of cancerous tissue. As previously discussed, approximately 0.9 million patches contain mixed cancerous and healthy tissue areas that were excluded to ensure model training on patches with pure labels.

For the encoder part of the network, `ImageNet` pre-trained weights were used for weight initialization that generates a latent embedding size $d_{\text{emb}} = 1280$.

Extensive Hyperparameter tuning with over 4 million training patches was performed. The parameters of the best found model can be found in appendix E.1. It uses an Adam optimizer with added weight decay with a factor of 6.8×10^{-4}, a learning rate of 5.17×10^{-4}, 50 % dropout in the classification layer, and a batch size of 256.

6.4 Results

6.4.1 Patch-level

To evaluate the model's patch-wise predictions, classification metrics indicating the correct output labeling regarding cancerous or healthy tissue of each patch are analyzed on the previously unseen test dataset (see sec. 6.3). Note that classification labels that are used for the evaluation are based on the GT segmentation masks, but depend on other factors like the patch size p_s, the minimum amount of tissue T_{th}, and the amount of cancerous area C_{th} within a patch meaning that the results may change for variations in those settings.

Furthermore, the resulting patch classifications can also be used to produce a cancer heatmap for whole biopsy slides. For a comparison to the original GT segmentation mask and the cancer heatmap, segmentation metrics are additionally presented.

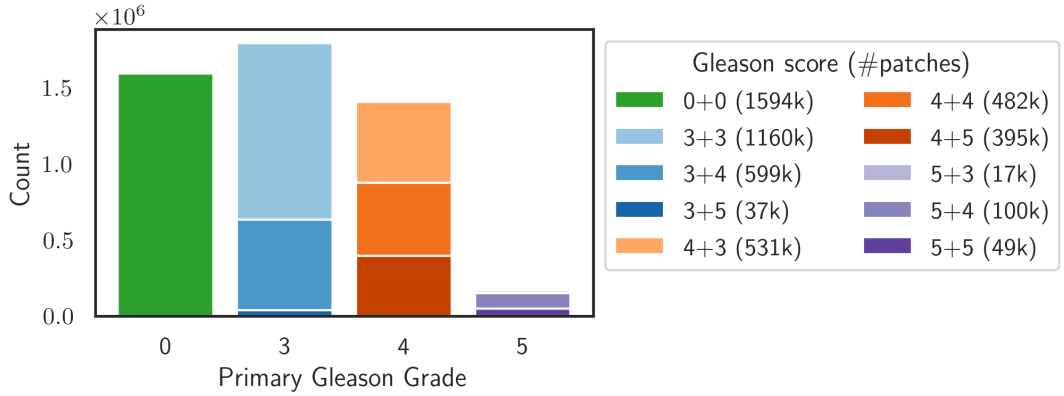


Fig. 6.7: Number of extracted patches per slide-level primary Gleason grade.

Label Distribution

The distribution of the assigned cancerous and healthy labeled patches is analyzed regarding Gleason scores on slide level. Fig. 6.7 depicts the number of patches over the primary GG for all extracted patches. It shows that most patches originate from a slide-level Gleason score of 0+0 (1.594 million), 3+3 (1.16 million), and 3+4 (599 thousand) while the least samples are provided for Gleason scores 5+5 (49 thousand), 3+5 (37 thousand), 5+3 (17 thousand).

Moreover, it can be observed that the ratio of patches for positive and negative labels is not constant over the different Gleason scores. Fig. 6.8 shows that a Gleason score of 0+0 contains no cancerous patches while the remaining ratios vary between 0.26 for Gleason score 3+3 and 0.49 for Gleason score 4+5.

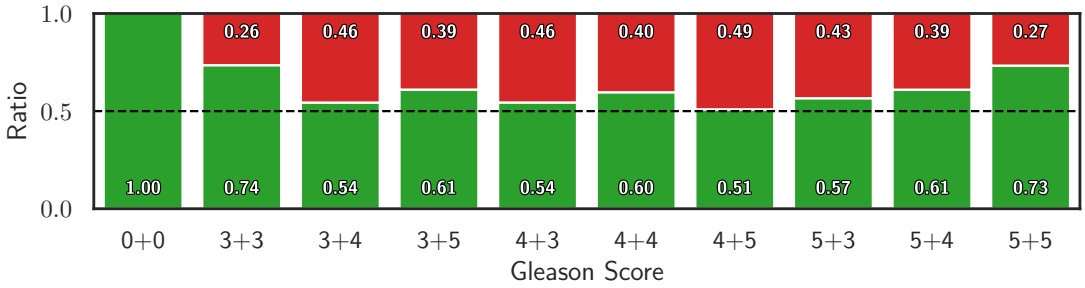


Fig. 6.8: Relative label ratio per Gleason Score. The red part of each bar represents the ratio of patches with cancerous areas, green stands for healthy tissue patches per Gleason Score. The balanced ratio of 0.5 is depicted by the dashed line.

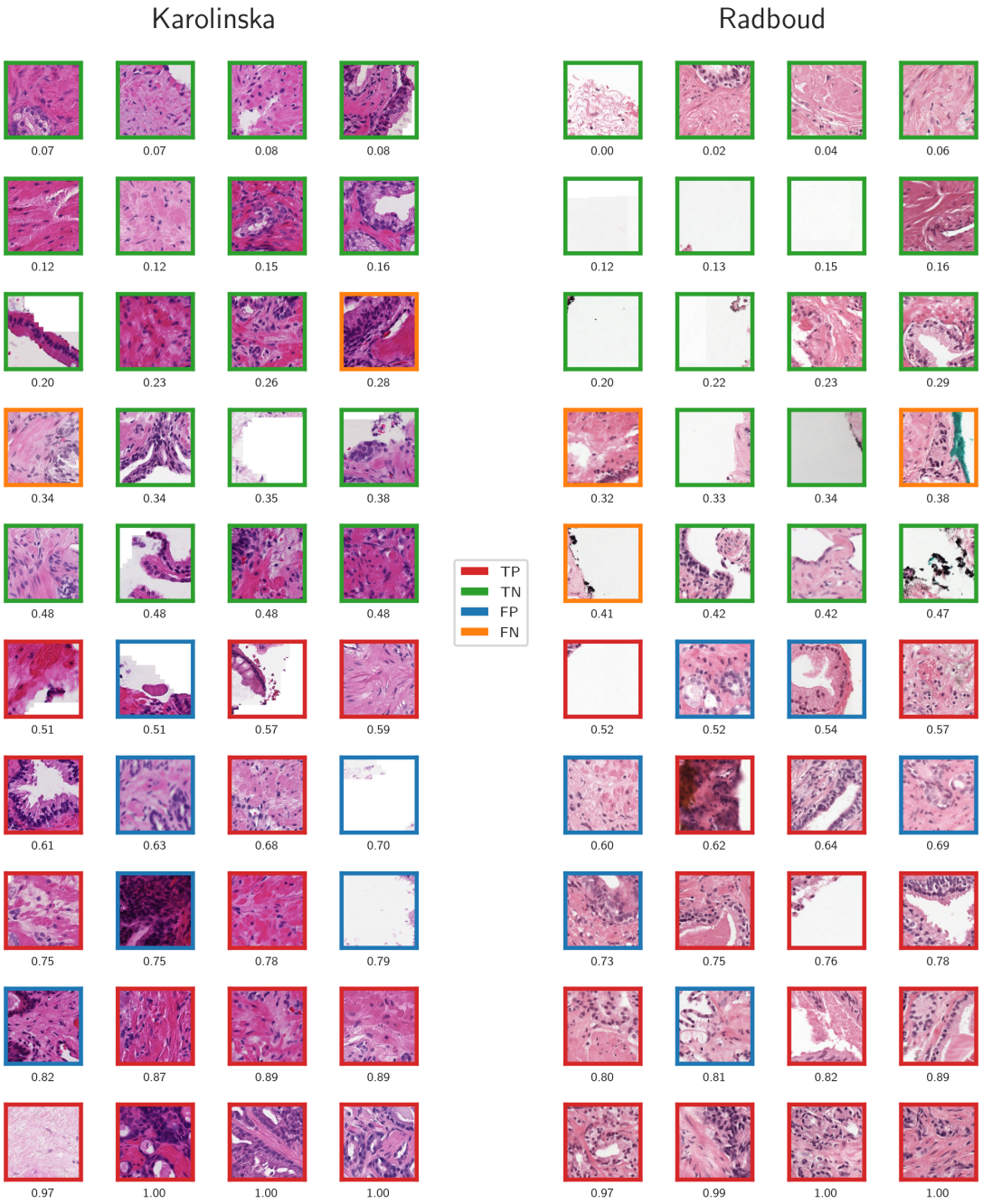
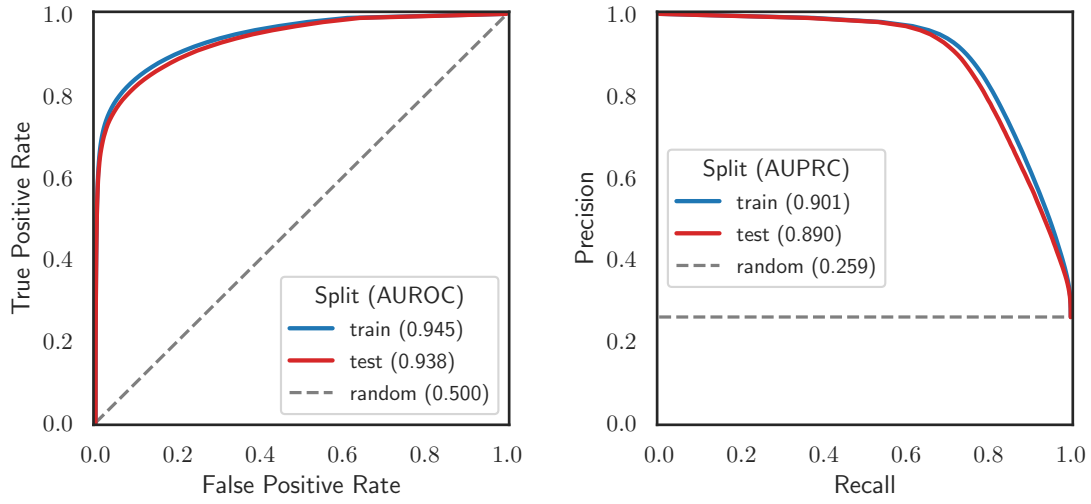


Fig. 6.9: Patch-level prediction examples of the CI model for KAR (left) and RAD (right) with increasing predictions from top to bottom. The outline color of each patch indicates the prediction type (TP in red, TN in green, FP in blue, and FN in orange for thresholded predictions at 0.5).

Classification Performance

Since the CI model is trained on patch-wise cancer classification, it is evaluated on the prediction of the extracted individual patch labels as shown in fig. 6.9 for visual inspection. The CI model achieves an AUROC of 0.938 and a AUPRC of 0.890 on the 504,027 patches of 1,062 test set WSIs. The corresponding ROC and Precision-Recall Curve (PRC) for the training and test splits are shown in fig. 6.10. Note that random performance for PRC takes the class imbalance into account (number of positives over all samples). Further, it can be observed that training performance is slightly higher (0.7 pp for AUROC and 1.1 pp for AUPRC).



(a) Receiver Operating Characteristic Curve

(b) Precision-Recall Curve

Fig. 6.10: Patch-wise ROC and PRC evaluation for the final CI model for the train and test split.

Performance per Gleason Score To generate additional insights on how the CI model performs, the previous ROC and PRC curves were additionally calculated for each Gleason score separately for the test set. Note that Gleason score 0+0 had to be excluded from this analysis since no positive labels are present. The resulting curves are presented in fig. 6.11. A performance difference can be observed ranging from an AUROC of 0.869 (3+3) to 0.977 (5+4) compared to the AUROC over all samples of 0.938.

Segmentation Mask Quality Furthermore, there are limitations in the provided segmentation masks that this work uses to identify foreground patches and assign the patch-wise labels as described in sec. 6.2.2.

Firstly, the tissue masks of this work are inaccurate, as stated by the authors of the PANDA challenge [35]. This can be observed, for example, in fig. 6.5 where parts of the background are included in the GT tissue mask. To estimate the quality of the provided tissue masks, the GT mask is compared to a brightness-based filtering approach for foreground patches. The brightness based filter removes all patches with an average brightness above 0.92 or 235/255 which is only expected if no tissue is present in the individual patch. This way, 31,722 patches or 6.2 % of the test set were removed indicating that a significant fraction of the tissue mask is inaccurate. It is worth noting that the majority of the identified patches originate from RAD (92 %) where a median of 13.8 % of patches were removed from each individual slide as depicted in fig. 6.12b. Reevaluating on the remaining 478,730 patches improves the performance of the model only marginally by 0.2 percentage points to 0.940 as shown in fig. 6.12a. Nonetheless, it can be observed that the AUROC performance drops significantly when it is evaluated on the

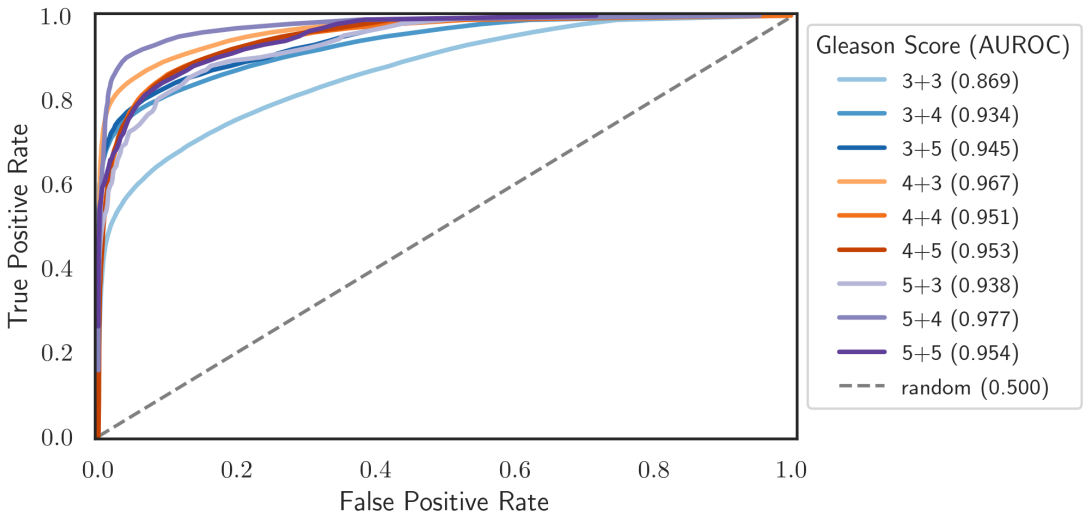
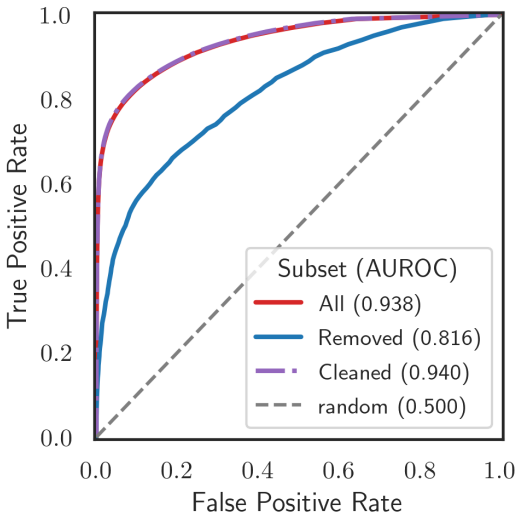


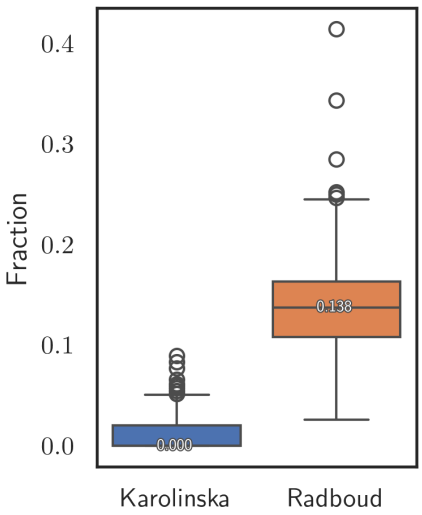
Fig. 6.11: Patch-wise test set ROC by Gleason score.

filtered patches to 0.816 indicating that the information within those patches contains less useful information for the cancer indication task.

Performance per Center Moreover, another difference can be observed when comparing the CI model performance per center as depicted in fig. 6.13. It can be observed that the model's performance is higher for biopsies from RAD compared to KAR. While the combined AUROC for all biopsies from the RAD dataset reaches 0.978, biopsies from KAR lack behind by 13.3 pp and only show an AUROC of 0.845. An additional analysis of the individual performance for the two centers by Gleason score is illustrated in fig. A1 where for RAD, AUROCs of up to 0.994 are reached.



(a) ROC for brightness-based filtering subsets.



(b) Fraction of patches per slide that were removed by the brightness filter.

Fig. 6.12: Dataset cleaning results when removing wrongly labeled foreground patches by a brightness filter. After cleaning of the 504,027 patches in the test set (All), 478,730 patches remain (Cleaned) and 31,722 (Removed) were removed.

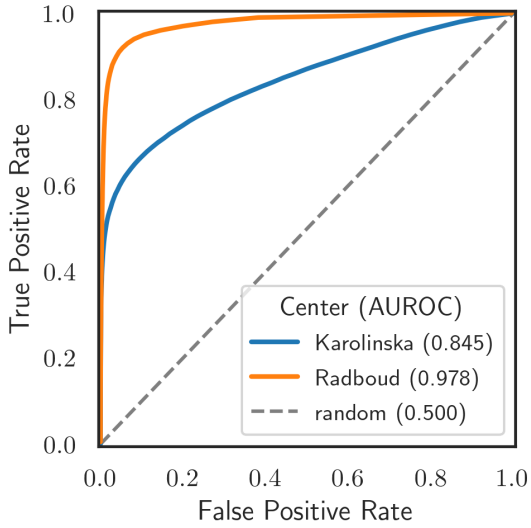


Fig. 6.13: AUROC per data center on the test dataset.

6.4.2 Slide-level Segmentation

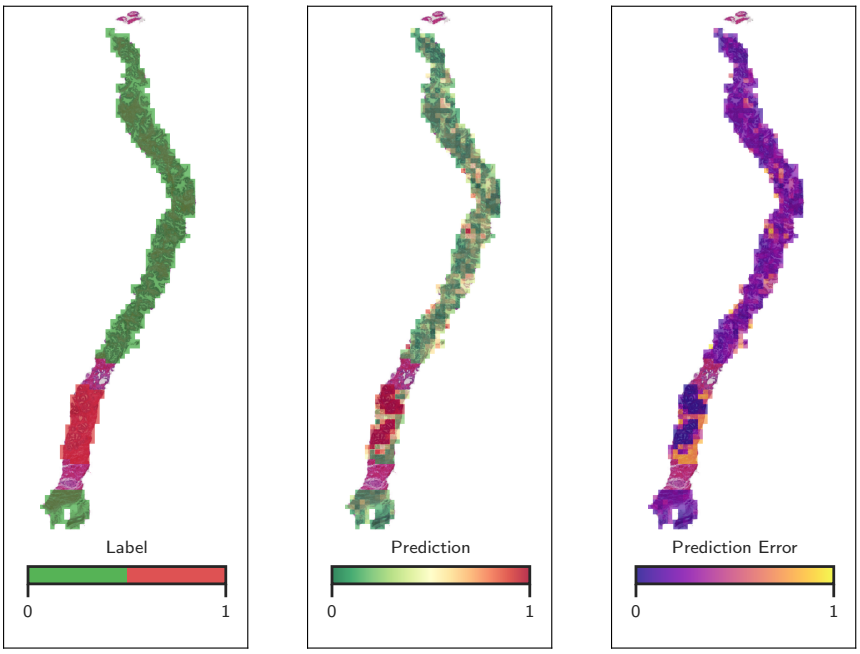
To further visualize the predictive performance of the CI model, this section shows WSI predictions that were obtained by stitching the individual patch predictions of the CI model to generate cancer heatmaps. The individual patch predictions were not further processed to take the predictions of neighboring patches into account (e.g. by averaging across multiple adjacent patches). This enables the comparison of the patch-wise predictions to the generated labels on full WSIs.

Firstly, fig. 6.14 shows two heatmaps with high prediction errors. The first slide contains some high errors in the cancerous area with low predictions on some patches as well as some errors in the area at the top that is annotated as healthy. For the second slide, almost the entire region is predicted to have a cancerous area, while the GT labels only show two smaller and distinct regions.

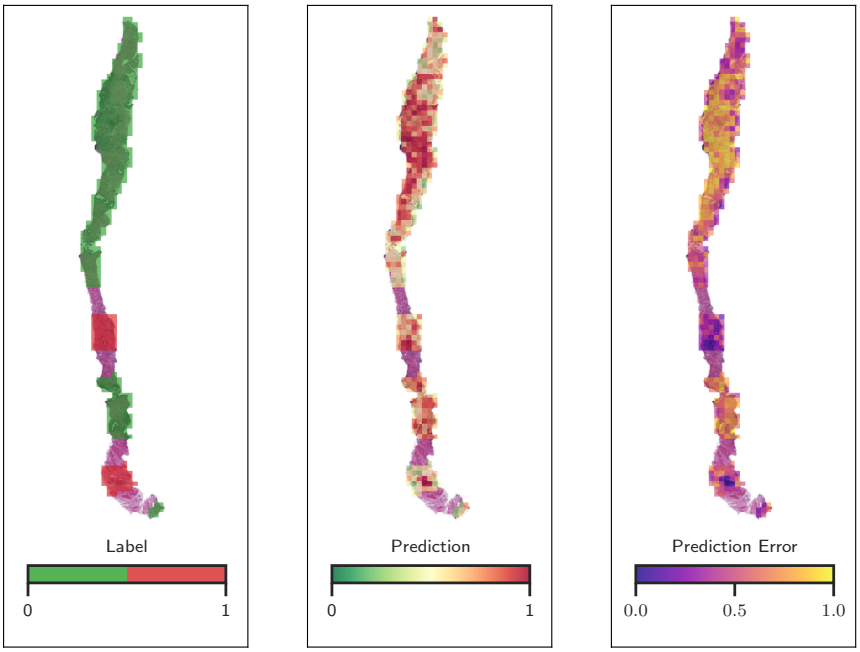
Moreover, fig. 6.15 shows two examples where the predicted heatmaps show almost no error in the predictions. The CI predictions in Fig. 6.15a closely match the GT area. In fig. 6.15b, only a single patch does not match the GT annotation that did not see any cancerous tissue on the biopsy slide.

Comparing to Segmentation Masks

The generated cancer heatmap does not provide the same level of resolution as the original GT masks, especially for the fine-grained masks from RAD. This makes an evaluation of segmentation metrics infeasible. However, fig. 6.16 shows the patch-wise cancer heatmap compared to an original fine-grained segmentation mask. It can be observed that the correct regions are identified. Mainly false positive regions exist since the GT cancer labels are small. Further, the CI can be evaluated using the segmentation masks on the unseen test data for all slides that contain cancer (i.e., excluding all 0+0 cases on slide-level). A mean dice score of 0.54 with a standard deviation of 0.24 could be achieved. Note that the performance differs for the two centers since the GT masks are provided in different granularity. On the coarser segmentation masks provided by KAR, a higher mean dice score of 0.65 ± 0.23 was achieved, while the finer segmentation masks of RAD lead to a mean dice score of 0.43 ± 0.19.

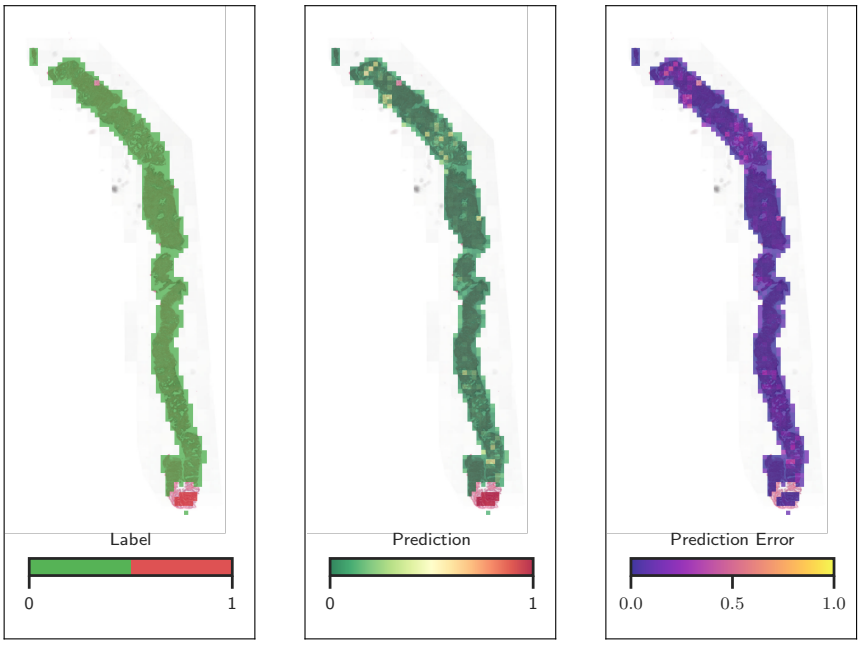


(a) High prediction error in parts of the annotated cancerous area.

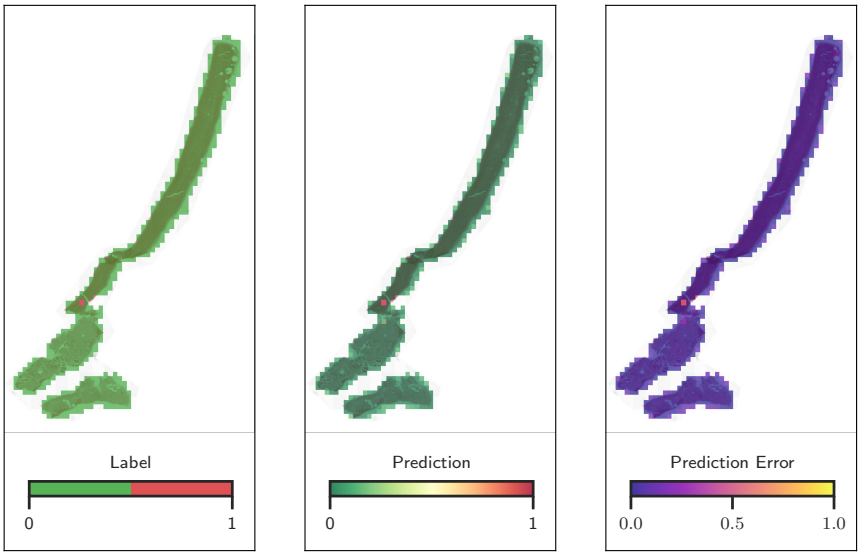


(b) High prediction error in the non-cancerous area at the top.

Fig. 6.14: Example WSIs with high mean prediction error when compared to the extracted label.



(a) Only some patches show a prediction error in the lower and upper region of the biopsy.



(b) Almost no prediction error except for a single patch in the middle of the tissue.

Fig. 6.15: Example WSIs with low mean prediction error when compared to the extracted label.

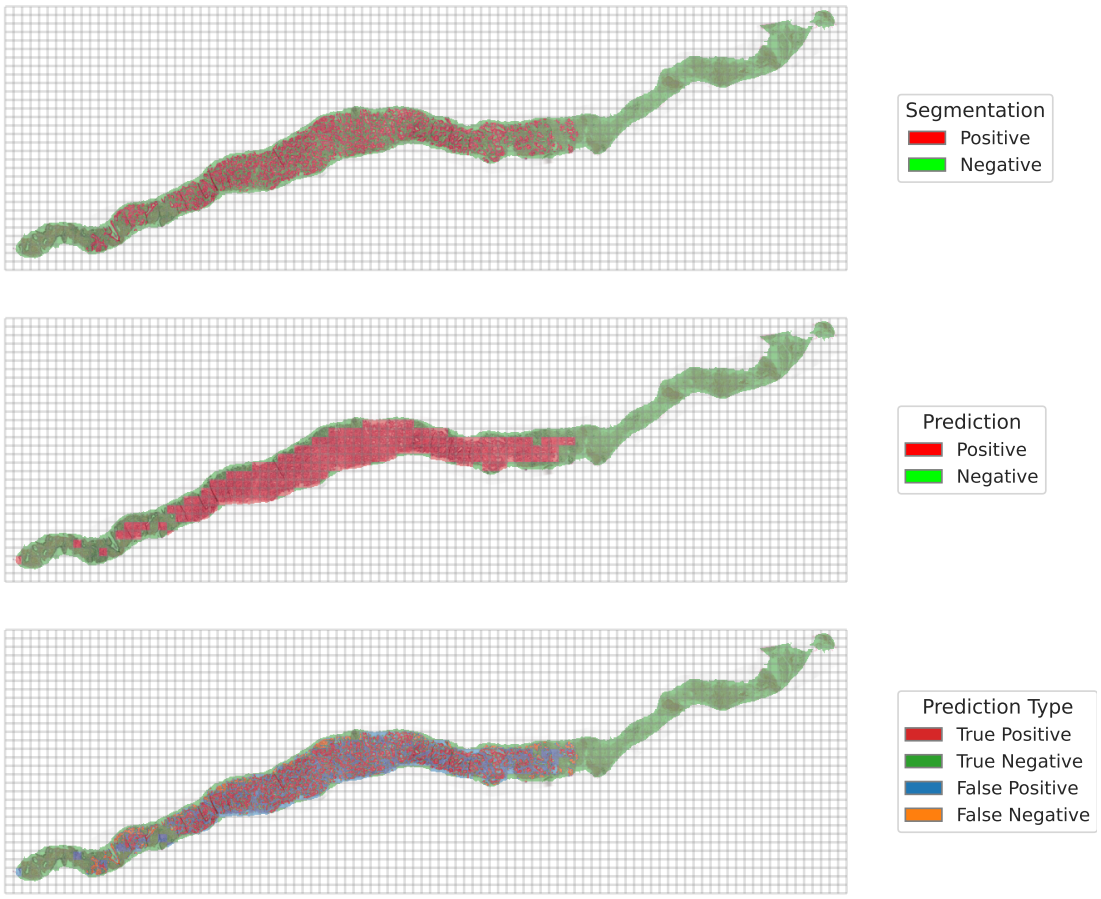


Fig. 6.16: Exemplary GT segmentation mask (top), stitched patch-wise predictions of the CI model (middle) and the pixel-wise prediction type overlay (bottom).

Prediction and Cancer Severity

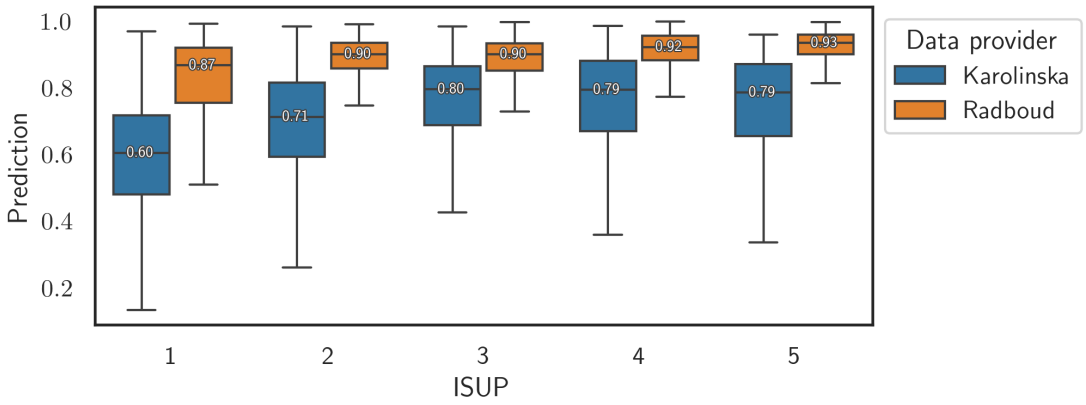


Fig. 6.17: Mean patch-level CI prediction of cancerous regions per slide split by data center and slide-level ISUP grade.

To analyze if the cancer severity increases along with the CI model predictions is shown in fig. 6.17. For this analysis, ISUP is used as an indicator of cancer severity. To compare the CI predictions to slide-level ISUP, the mean prediction for all cancerous patches per slide is calculated. These are compared to the corresponding ISUP grade and separated by the two centers. For KAR, the median of the extracted mean prediction per biopsy slide rises from ISUP1 (0.60) to ISUP3 (0.80) before being approximately constant for ISUP groups 4 (0.79) and 5 (0.79). For RAD, the median prediction slightly increases with ISUP grade from ISUP1 (0.87) to ISUP5 (0.93).

Furthermore, it can be observed that the mean prediction per biopsy for the KAR center is constantly higher than the mean slide prediction of the RAD center. This shift from one center to the other is caused by the coarseness of the segmentation mask. Since the mean predictions of cancerous areas are used, the human-annotated masks from KAR additionally include more healthy tissue that decreases the mean prediction for those biopsy slides.

Nonetheless, it can be observed that the mean predicted CI value per cancerous area of the individual slides increases from lower ISUP to higher ISUP grades hinting that a higher prediction of the CI model corresponds to higher cancer aggressiveness.

Background Prediction

Lastly, since the CI model is trained only on patches that contain at least 10 % tissue, it shows undefined behavior on patches that do not contain any tissue. Fig. 6.18 shows qualitative examples for the external MMX dataset where two other scanner vendors, namely HAM and VEN, were used. It can be observed that the background regions of the WSIs that were scanned with VEN show significantly higher values for background patches compared to the slides that were scanned with the HAM scanner. This shows that the CI that was trained for this chapter should only be used on patches that contain at least 10 % of tissue.

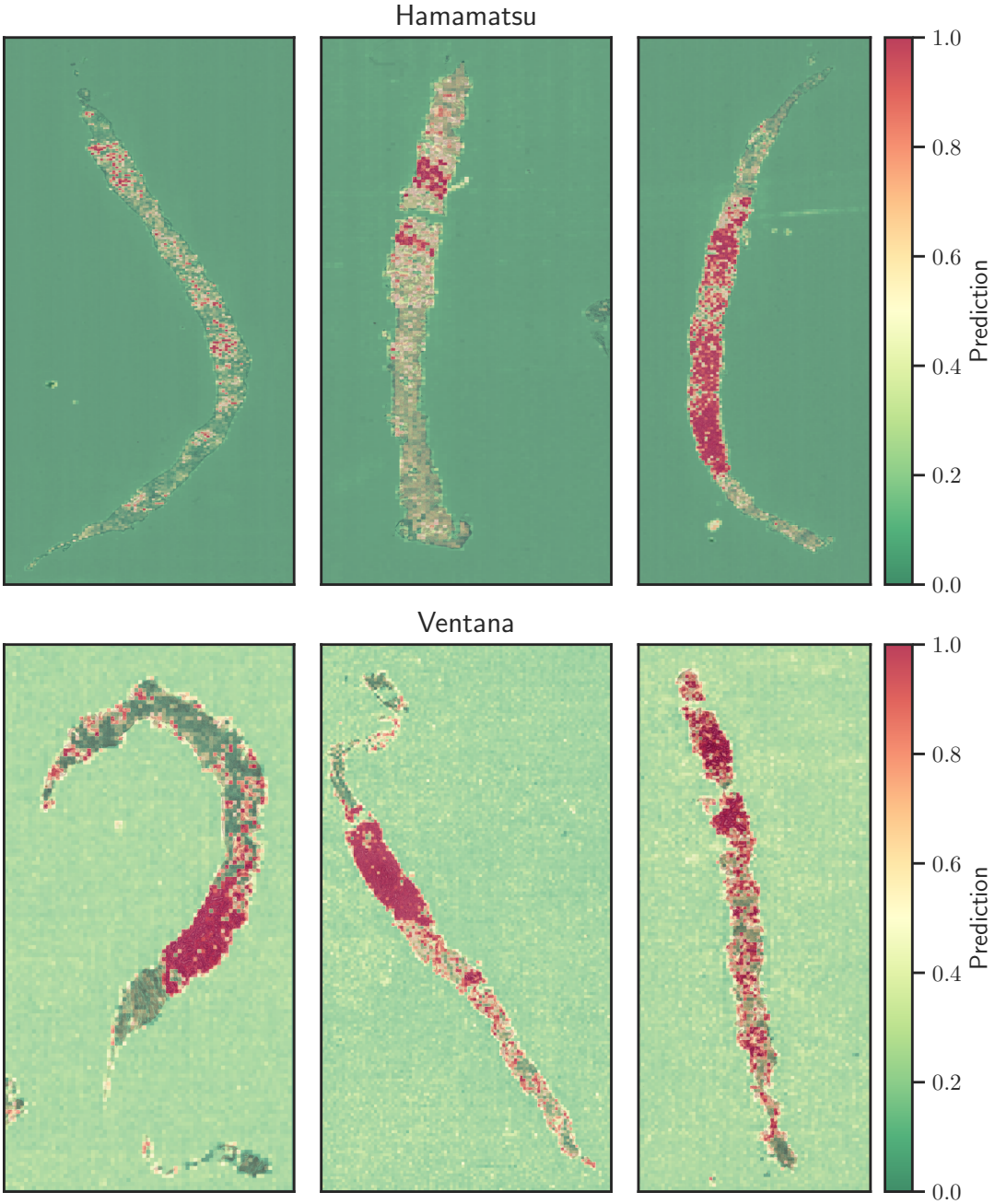


Fig. 6.18: Qualitative examples of WSIs from the MMX dataset with CI prediction overlay showing high predictions in background areas for the VEN scanner (bottom) compared to HAM (top).

6.5 Discussion

This chapter showed that the CI model is able to detect PCa on WSIs using individual patches with a test set AUROC of 0.938 proving the general applicability. A drawback of the segmentation mask-based label generation approach of this work lies in the tissue masks themselves since the authors state that the mask definitions may be imperfect [35]. To analyze the mask quality, an analysis that excluded all patches based on the average brightness revealed that over 6 % of patches that only show background were incorrectly included in the tissue mask although a minimum of 10 % of each patch should have been covered by tissue according to the tissue annotation mask. Nonetheless, the model's overall classification performance is high enough to provide a basis for selecting patches from cancerous regions of a biopsy WSI. Differences in predictive performance could be observed for different Gleason scores where the lowest performance of the CI model was achieved for the 3+3 group. This shows that the CI model struggles the most when distinguishing healthy from cancerous patches because Gleason score 3+3 shows the least difference from healthy tissue among the Gleason scores analyzed in the dataset.

This information can be combined with the finding that biopsy slides with higher slide-level ISUP grades also show a higher slide-level CI prediction by averaging the over the cancerous area per slide. It is an indication that a higher prediction of the CI model is also associated with cancer severity.

It is also worth noting that mean patch-wise prediction per biopsy slide differs for the two centers indicating that some kind of bias exists. The overall AUROC for the two centers shows differences with an overall test set AUROC of 0.978 for RAD and 0.845 for KAR. This difference can most likely be explained by the origin of the segmentation mask that was used to generate the patch-level GT labels of this work. While expert (uro-) pathologists provided rather coarse segmentation masks for KAR, more fine-grained AI-generated masks were used for the RAD dataset. This model's prediction combines the two annotation approaches to produce predictions where the granularity is between the two approaches provided, namely on patch-level.

6.6 Conclusion

This chapter presented the cancer indicator or CI model. A DL-based approach to detect cancer from a single patch of H&E stained prostate biopsy tissue. The model architecture was kept simple and contained a CNN-based image encoder, namely `EfficientNet-b0`, and a classification head with two output nodes for the binary classification problem of healthy vs. cancerous tissue that was trained using a binary cross-entropy loss for the two classes. Model training utilized data from the PANDA challenge described in sec. 3.2.3. 10,616 biopsies were provided and split into a training, validation and a test set with 8,492 (80 %), 1,062 (10 %), 1,062 (10 %) slides respectively. Afterwards, all WSIs were cut into individual patches where only those that show at least 10 % of foreground tissue were selected. Afterwards, based on the provided GT masks of expert (uro-) pathologists, the amount of cancerous area within the tissue was the basis to define a binary label for each patch if it contains at least 90 % cancerous tissue ($y = 1$) or not ($y = 0$). This definition led to a total extraction of 4,963,701 individual patches with a corresponding label indicating if it shows either healthy or cancerous tissue. For training, around 20 % that contained healthy and cancerous areas were excluded to ensure label purity. This filtering was not applied on the test dataset. Regarding quantitative test set performance, the final CI model achieves a patch-level classification AUROC of 0.938 and a AUPRC of 0.890. Moreover, the patch-wise predictions for a whole biopsy image were combined to produce heatmaps indicating cancer location for whole images. Qualitative analysis of the heatmaps shows, that a large overlap with the corresponding GT masks exists up to the granularity of individual patches.

It was shown that the tissue selection and cancer annotation masks did contain a significant

amount of errors for both centers. These factors motivate the usage of a more homogeneous source of segmentation masks to enable the creation of a more accurate model. Nonetheless, the large number of 10,616 WSIs could then be included for pre-training of the network. Since another drawback of the presented approach is missing external validation, a new version should consider adding additional data sources. This can also increase the model's generalizability and robustness towards biopsy slides from unknown sources or also the undefined behavior for predictions on background patches. This way the clinical applicability of such a network would be greatly enhanced since the performance would not depend on the exact protocols of a particular pathology institute, but redirect the network's attention towards the biologically influenced morphologies that are visible on the tissue.

Further, this chapter did not analyze how different magnification levels of the images and patches might be utilized. Since different biological structures can be observed in different magnification levels, a more complex approach would consider multiple magnification levels and aggregate this information before the prediction step as presented in [200].

Future work could also extend the model's capabilities to not only distinguish between healthy and cancerous tissue, but take additional classes like cell types into account. The chapter showed in sec. 6.4.2 that CI predictions may get unreliable when it is used to infer on out of training data that include areas of background. It was emphasized that the current model should only be used on patches that contain at least 10 % of tissue. To avoid this problem, the CI model could be extended to not only predict on a binary cancer vs. healthy tissue endpoint, but on multiple classes. Those multiple classes could include the background itself as an additional target class which would get rid of the undefined behavior when these parts of the biopsy image are actively learned. Furthermore, tissue types could also be used as additional classes for prediction such as epithelium and stroma to further classify a patch on a more fine-granular level. The generated information could then also be used and be compared to the GG definition presented in sec. 2.1.2 that might lead to additional biological insights.

The creation of cancer heatmaps for the whole biopsy slide from independent patch predictions is suboptimal. Adding information of neighboring patches in the aggregation step should be done to improve slide-level heatmaps. A straightforward approach could for example use a moving window and calculate averages for each patch and adjacent neighbors, effectively blurring the resulting heatmap. Another idea would be to overlap the predicted patches to avoid the independence between neighboring patches. These simple ideas could be extended to any degree to produce more reliable slide-level heatmaps that reflect reality better than independent patch predictions.

Nonetheless, the model that was developed is able to reliably predict cancerous tissue on patch selections from biopsies. A public repository of the code that was used for this chapter was published in a public repository[21]. The final CI model is a valuable source to estimate patch importance for biopsy slides which is demonstrated in the following chapter.

[21] https://github.com/imsb-uke/pcai

7 Cancer Risk Estimation from TMA Spots and Biopsies

7.1 Introduction

This thesis showed in chapter 5 that pathological Gleason grading with more detailed information, namely quantitative Gleason grading yielded the best discriminative performance for an individual feature analysis regarding relapse prediction of the MK cohort of PCa patients after RP. It is worth taking a deeper look into Gleason grading itself since it suffers from high inter- and intra-observer variability that may lead to wrongful treatment decisions for the individual patient. Advancements in the field of digital pathology [156] lead to a growing number of digitized slides that open up the possibility to replace the human-guided, subjective annotation by a more objective approach like a DL-based model that requires sufficiently large datasets for reliable predictions.

This chapter presents the Prostate Cancer Aggressiveness Index (PCAI), a CNN-based approach to automatically derive relapse risk from histopathology images of prostate tissue. More specifically, the proposed model is trained end-to-end on TMA spots from the UKE dataset provided by the Institute of Pathology at the UKE to extract morphological features of the tissue structure related to relapse risk of the patient at the time of RP. To improve model robustness, it is trained on TMA spots with differences in acquisition protocols that include variations in staining time, slicing or the used scanner. Further, it is shown that this approach also can be applied beyond the domain of TMA spots into estimating the cancer aggressiveness of histopathological images of whole biopsy slides. These have greater importance for the urologist and open the potential for clinical application e.g. in terms of initial therapy decision. However, the biopsy images are much larger than individual TMAs. Biopsies are obtained at an earlier stage in the lifecycle of a PCa patient as discussed in sec. 2.1.1. This motivates the use of the previously presented CI model (see chapter 6) to select relevant parts of biopsy images that are afterwards used in the risk estimation process of PCAI.

The chapter is divided into the following parts. After presenting how this thesis contributes to the model development, the methods section describes how the PCa related datasets were utilized from the sources that were introduced in chapter 3. Further, the DL-based model architecture and strategies regarding model robustness and generalizability are presented. To show the validity of the robustness extensions, the results section compares the developed model to a baseline (BASE) that does not take these extensions into account. Further, the PCAI is compared to ISUP grading of human annotators on TMA spot and biopsy images in terms of cancer aggressiveness. Note that some datasets only provide a patient-level annotation that may be aggregated from multiple images. Nonetheless, three of the presented datasets, namely UKE.sealed, UPP, and MMX, provide individual slide-level annotations that are used for a fairer comparison. Qualitative results show ideas on how the DL model can aid the urologist in the decision-making progress that mainly involves the initial treatment planning after diagnosis from the biopsy. It is presented how the derived model can further be used to visualize relevant cancer locations on biopsies to guide manual inspection, derive risk groups from the PCAI prediction, and provide a credibility score for a presented sample based on the model's training data distribution to enhance the trustworthiness of the AI model towards the urologist.

In contrast to the DCS model development in chapter 4 that predicts individual survival curves

based on multiple input features, this chapter focuses on an individual scalar risk prediction that is generated exclusively from individual images. This allows for a better comparison to pathological GG or ISUP grades for a fairer comparison. Also, the resulting predictions of the model can afterwards be combined with other patient features and be interpreted as a digital biomarker in a joint survival model (as shown in chapter 5) of tabular- and image-based features of each individual patient.

7.1.1 Contribution

This chapter is based on preliminary work of the colleagues Esther Dietrich [64, 65] and Peter Walhagen [245] that analyzed DL-based risk prediction approaches for PCa from TMA spots and biopsies. This chapter was developed in collaboration with Fabian Westhaeusser.

7.2 Methods

The aim of this chapter is to develop a DL-based model that can predict cancer severity of histopathology images, namely TMAs and biopsies by taking robustness into account. The developed PCAI model utilizes TMA spot data of different qualities and staining protocols to train a robust CNN encoder-based model that estimates five-year cancer relapse after RP on TMAs. After training, the model is extended to the related, and clinically more relevant use case of biopsy risk prediction.

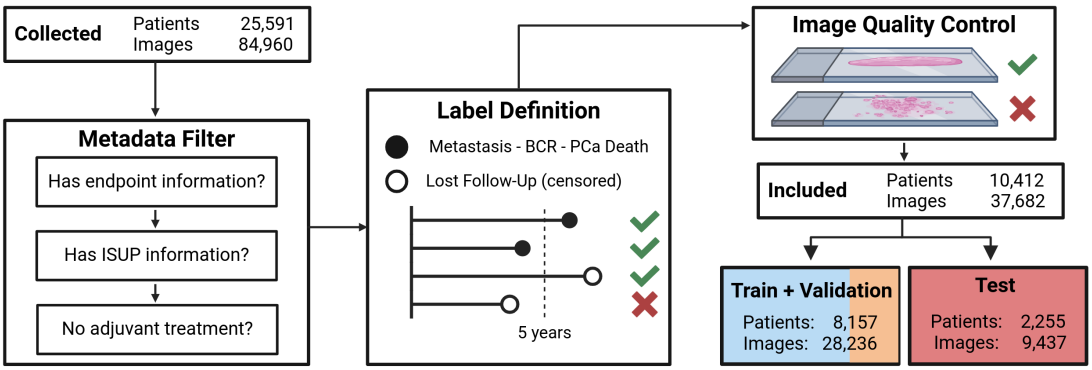


Fig. 7.1: Overview of the metadata and image-based quality control filtering steps to obtain the filtered datasets that are used for development and evaluation of the PCAI model. Created with BioRender.com

7.2.1 Dataset Preparation and Selection

This chapter utilizes the PCa datasets that contain TMAs or biopsies introduced in sec. 3.2. The images originate from five different clinics and contain up to eight different images for a single patient. The clinics from three different countries that are involved in this work are the UKE (Hamburg, Germany), NYU (New York City, USA), JHU (Baltimore, USA), UPP (Uppsala, Sweden) and MMX (Malmö, Sweden). A total of 10,412 patients and 37,675 TMA and biopsy images were analyzed. The following section describes how the datasets that were used in this work to train and evaluate the PCAI model are derived based on their metadata and image quality as shown in fig. 7.1.

Firstly, patients with insufficient endpoint information are filtered. To compare to patient-level ISUP grading, only patients that provide this feature are kept for all datasets. Further, individuals that received adjuvant treatment are removed from further analysis. For all datasets, it is

documented if the patients developed BCR, metastasis, had an additional, unplanned therapy, or are lost to FU. Patients that are lost to FU are considered censored at that time while all other events are considered as an event leading to an event indicator of $d_i = 1$ with the corresponding duration z_i from either the RP (for TMA datasets) or the date of biopsy taken (for biopsy datasets) until the event-of-interest. In addition, patients who have a censoring time less than 6 months are excluded in this chapter. Pathological GG were obtained from the whole prostate after RP.

Moreover, this work utilizes a binary five-year survival indicator (similarly to the calculation of an AUROC for a single point in time, see sec. 2.4.4). For training, the indicator is introduced based on the event time z_i and indicator d_i for each individual as

$$y_i = \begin{cases} 0 & \text{if } z_i \geq 5 \text{ years} \\ 1 & \text{if } z_i < 5 \text{ years and } d_i = 1 \\ \text{undefined} & \text{otherwise} \end{cases} . \tag{7.1}$$

This step cannot provide a valid label for censored individuals with less than 5 years of follow-up information that can therefore not be used during training, thus are filtered out.

Further, all images with insufficient quality are removed. This includes the exclusion of all images that do not show tissue in at least 10 % of the presented area. Since no tissue masks exist for the datasets, a brightness filter is used where images that show a lightness value of more than 0.92 (or 235/255) in over 90 % of the total area are removed. This automated approach was not feasible for biopsies that may show even higher fractions of non-tissue area, but still contain valuable information. Instead, images were manually analyzed and removed if they did not show enough tissue, were blurred or had pen marks on relevant parts of the slide.

The results part of this chapter will compare model predictions of DL-based models to human annotated ISUP grades that are present for each image. Some datasets only provide an ISUP grade that was assigned on patient-level by combining information that may have been obtained from multiple images. This makes a comparison to the predictions that were made on individual images hard to compare. To overcome this problem, three datasets are included that also provide image-level ISUP grading for a fairer comparison.

The slide-level human annotations and the developed DL-based algorithm predictions are then aggregated to patient level by using the worst prediction per patient among the images. This way the resulting predictions can be evaluated and compared to the patient's actual survival information.

Training Data

For model training, only TMAs from the UKE are used that show single TMAs of RP patients. Since this dataset is the only source of training data for the model of this chapter, it is explained in additional detail. In total, 39 blocks with 69,251 of 17,700 patients were collected and observed for over 20 years after RP offering a high quality dataset regarding PCa relapse prediction after RP.

For research, the TMAs obtained from those patients were digitized using variations in the acquisition protocol by the department of pathology of the UKE (see sec. 2.1.2). In total, 8,157 patients with 28,236 images that are divided into sub-datasets categorized by variations in the staining protocol as described in sec. 3.2.1 are used for model development and evaluation. The sub-datasets that were analyzed in this work are visualized in fig. 7.2 and briefly explained below. In the following, the set of these sub-datasets is called UKEhv. This figure also includes the data that was used to train the BASE and final PCAI model that are described in sec. 7.2.3.

The first sub-dataset is called UKE.first and contains 8,123 TMAs that were most representative for the individuals. They were cut at 2.5 µm and stained with hematoxylin for four minutes

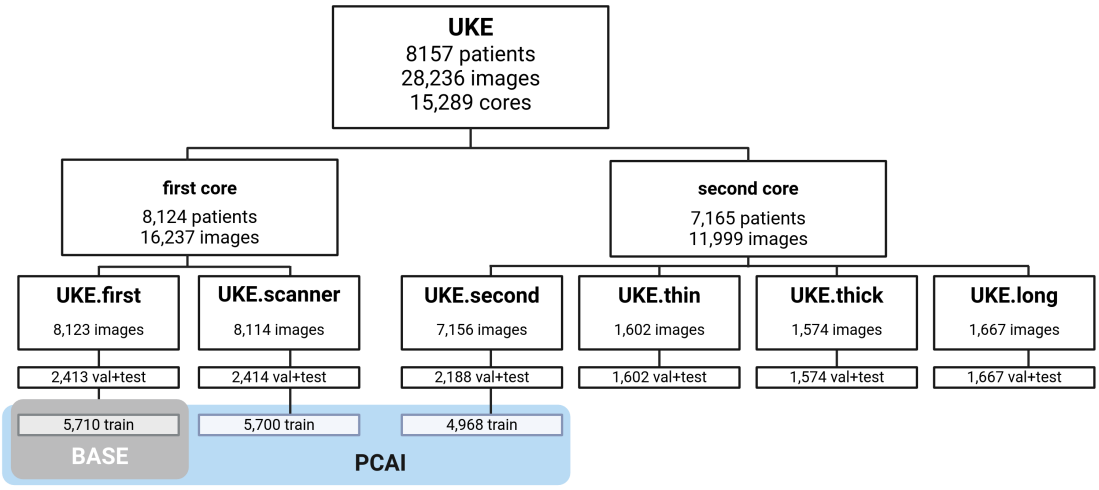


Fig. 7.2: Image distribution of UKEhv sub-datasets that originate from two cores. Additionally, the training data for BASE (gray) and PCAI (blue) are shown.

and with eosin for 1 minute and 20 seconds respectively. Afterwards, the TMA blocks were scanned with an Aperio (APE) scanner. Further, UKE.second uses the same staining protocol, but with a secondary TMA spot that was obtained from a different core of the cancerous region leading to 7,156 extracted TMA images. Moreover, UKE.scanner contains the same TMAs as UKE.first that were scanned with a different scanner vendor (3DH) at a higher magnification of 80x (0.125 µm/pixel) instead leading to 8,114 extracted TMAs.

Additionally, three smaller datasets with variations in TMA thickness and staining time are included. UKE.thin contributes 1,602 TMAs that were cut at a thinner thickness of 1 µm whereas UKE.thick provides 1,574 images of TMAs that were cut at 10 µm. Lastly, UKE.long contains 1,667 TMAs that were stained for an almost ten times longer staining time of 40 minutes for hematoxylin and 10 minutes eosin.

Lastly, additional TMAs were provided without any corresponding metadata information which that is only used for a final evaluation of the algorithm by the department of pathology of the UKE. This thesis calls this sub-dataset UKE.sealed that provides 4,097 TMAs with hidden patient information. This is the only sub-dataset of the UKE that contains multiple images per patient and provides an image-level annotation of ISUP as well as GIQ grades that are used in the evaluation.

External Evaluation Data

Additional datasets are included for external model evaluation. This includes data as described in sec. 3.2 from NYU, JHU, UPP and MMX. For those datasets, the same filtering steps as explained above are applied to ensure the same dataset quality. While NYU and JHU are used to evaluate the robustness extensions of the developed model, UKE.sealed, UPP and MMX provide image-level annotations for a final model evaluation compared to image-based human ISUP grading.

The main characteristics of the resulting datasets are visualized in tab. 7.1 (additional information can be found in appendix F.2). Note that only 18 % of patients from the UPP dataset could be included mostly due to lack of sufficient followup information. Also, all patients that received multiple therapies are excluded. In total, the filtered datasets that are used in this work contain 36,414 TMAs and 1,261 biopsies from 10,412 patients. Fig. 7.3 shows the fraction of patients

per dataset that were included in this study meaning that more than 5 years of FU information was present for censored individuals, ISUP grading was provided, no adjuvant treatment was performed, and the image quality was sufficient.

Tab. 7.1: Basic characteristics of the datasets that are used for the development and evaluation of PCAI. Type differentiates between TMAs and biopsies, annotation level between patient- and image-level ISUP annotation.

	UKEhv	NYU	JHU	UKE.sealed	UPP	MMX
Type	TMA	TMA	TMA	TMA	Biopsy	Biopsy
Annotation level	Patient	Patient	Patient	Image	Image	Image
#Patients	8,157	158	879	826	123	269
#Images	28,236	506	3,575	4,097	683	578
Scanner	APE, 3DH	APE	HAM, VEN	APE	HAM, VEN	HAM
Thickness [µm]	1, 2.5, 10	5	4	2.5	unknown	4-5

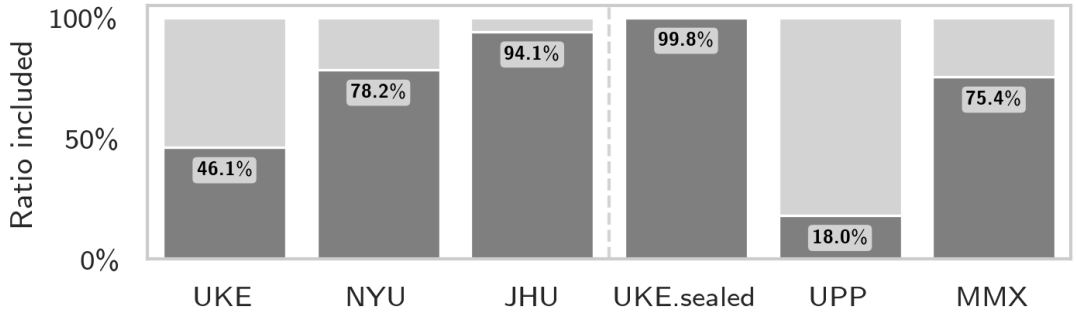


Fig. 7.3: Fraction of included patients for the analyzed datasets.

7.2.2 Image preprocessing

Magnification Level For the integration of the datasets, several preprocessing steps are necessary. Firstly, the highest shared magnification level among all datasets of 20x (0.5 µm/pixel) is used prior to patch extraction to ensure the same level of detail among all datasets. For TMAs, this leads to images with a side length of approximately 1,800 pixels.

Masking The resulting images in the final datasets may contain arbitrary shapes and sizes especially for the provided biopsy slides. To ensure that consistently sized images can be provided to the model, the same patching strategy that was also used for the CI model is used as presented in chapter 6. This means that the images are cut into equally sized patches that were processed individually. This chapter uses a similar approach, but keeps the individual patches of a specific slide inside a bag since only a slide-level label exist.

To select the relevant patches of the images, several masks are generated:

Tissue Mask The tissue mask distinguishes between foreground and background of each image. It is generated by using Otsu's thresholding [176]. The main idea of this approach is to find a dynamic threshold that minimizes the sum of weighted within-class variance in the histogram for each individual image.

Anomaly Mask Some selected images might contain pen markers, blood stains, or other impurities that do not show actual tissue. To detect and remove those, the HSV representation of an image calculates the median intensities $\bar{H}$, $\bar{S}$ and $\bar{V}$ of the foreground pixels for each channel (using the aforementioned tissue mask). Afterwards, all pixels that deviate more than empirically determined values of $(-60, +60)$ from $\bar{H}$, $(-30, +70)$ from $\bar{S}$, and $(-60, +60)$ from $\bar{V}$, are considered anomalies. Note that the total ranges of the HSV space in this work are defined as $H \in [0, 360]$, $S \in [0, 255]$, $V \in [0, 255]$. Final morphological opening and closing remove small regions from the image to obtain the final anomaly mask.

Filtered Tissue Mask Since the anomaly mask may remove foreground and background regions, the first tissue mask is regenerated without the anomalies to provide the final tissue mask $\mathbf{T}_i$ for each individual image.

Patch Selection Based on the definitions from sec. 6.2.2 and the aforementioned mask definitions, individual patches $\mathbf{P}_{(i,m,n)} \in \mathbb{R}_+^{p_s \times p_s \times 3}$ of the i-th image in the m-th row and n-th column with a patch size of p_s that contain at least 10 % of filtered tissue are obtained from each individual image. All patches from a single image are contained in the corresponding bag

$$\mathbb{B}_i = \left\{ \mathbf{P}_{(i,m,n)} \;\middle|\; \frac{|\mathbf{T}_{(i,m,n)}|}{p_s^2} > 0.1 \right\} . \tag{7.2}$$

where $|\mathbf{T}_{(i,m,n)}|$ defines the number of non-zero entries of the corresponding filtered tissue mask of the i-th image in the m-th row and the n-th column.

Finally, combining the binary five-year relapse indicator y_i from eq. (7.1), the final dataset of n_{img} individual images that is used for model training can now be defined as tuples of the individual bag and the five-year relapse indicator as $\{(\mathbb{B}_i, y_i) \mid i \in \{0, 1, \ldots, n_{\text{img}}\}$.

7.2.3 Model Architecture

The following section explains the different parts of the PCAI model in more detail. It provides the foundations of BASE before the extensions of PCAI are introduced. The main idea of this chapter is to develop an end-to-end risk assessment model for histopathological images of cancerous prostate tissue. The goal is to compare the predictive performance of the model to ISUP grading based on human annotations that is known to face challenges like a high inter observer variability as discussed in sec. 2.1.2. Unlike other models, this work does not try to emulate the Gleason grading itself, but provide an alternative risk estimation for the individual images. To achieve this, these human annotations are not used during model training. The model rather utilizes objective information, namely if the patient survived at least five years after RP without any form of cancer relapse after RP. During inference, the predicted value of the model can be interpreted as the probability of cancer relapse after RP within the first five years for a given TMA. This defines the idea of the BASE model as depicted in fig. 7.4 that is described in additional detail below. Note that layer biases were omitted to simplify some equations.

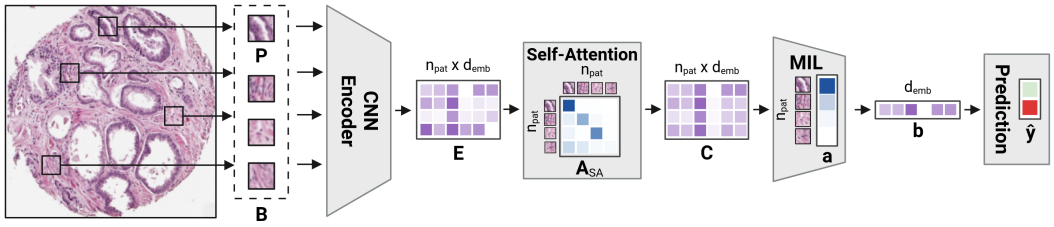


Fig. 7.4: Overview of the BASE model architecture. After n patches of the i-th image are selected and preprocessed, the encoder transforms the bag of patches $\mathbb{B}$ into the embedding matrix $\mathbf{E}$ by concatenating of the individually encoded patch embedding vectors. The SA layer creates the context-aware embedding matrix $\mathbf{C}$ that is aggregated via attention-based MIL to the bag embedding vector $\boldsymbol{b}$ before the classification layers produce the final prediction vector $\hat{\boldsymbol{y}}$. Created with BioRender.com

Patch-Level Encoder

The first building block of PCAI is the image encoding network that transforms all n_{patches} patches of an input image into a patch-level latent representation.

Using the embedding function

$$\text{emb}\colon \mathbb{R}^{p_s \times p_s \times 3} \to \mathbb{R}^{d_{\text{emb}}}, \tag{7.3}$$

every patch $\mathbf{P}_n$ of the bag $\mathbb{B}$ is passed to the encoder independently which transforms $\mathbf{P}_n$ into a latent vector $\boldsymbol{e}_n \in \mathbb{R}^{d_{\text{emb}}}$ with the embedding size d_{emb}. This creates the set of all patch embedding vectors for one image

$$\mathbb{E} = \{\text{emb}(\mathbf{P}_n) \mid \mathbf{P}_n \in \mathbb{B}\} \tag{7.4}$$

that are further concatenated as row vectors in the embedding matrix

$$\mathbf{E} = \left[\boldsymbol{e}_0, \boldsymbol{e}_1, \ldots, \boldsymbol{e}^{n_{\text{patches}}-1}\right]^T . \tag{7.5}$$

Following the same approach that was chosen for the CI model (see sec. 6.2.3), this chapter also utilizes the CNN-based `EfficientNet-b0` [226] as introduced in additional detail in sec. 2.5.3 for patch-wise image embedding.

Patch-Level Self-Attention

Next, a self-attention (SA) layer following [193, 202] is used that introduces interdependencies between a bag's individual patch embeddings. For the i-th individual patch embedding vector $\boldsymbol{e}_i \in \mathbb{E}$, a key vector $\boldsymbol{k}_i = \mathbf{W}^k \boldsymbol{e}_i$, query vector $\boldsymbol{q}_i = \mathbf{W}^q \boldsymbol{e}_i$, and value vector $\boldsymbol{v}_i = \mathbf{W}^v \boldsymbol{e}_i$ are defined to obtain the self-attention weights

$$a_{nm} = \frac{\exp\left(\boldsymbol{q}_n{}^T \boldsymbol{k}_m\right)}{\sum_{o=0}^{n_{\text{patches}}} \exp\left(\boldsymbol{q}_n{}^T \boldsymbol{k}_o\right)} \tag{7.6}$$

in the n-th row and m-th column of the self-attention matrix $\mathbf{A}_{\text{SA}} \in \mathbb{R}^{n_{\text{patches}} \times n_{\text{patches}}}$. The output of context-aware embeddings $\boldsymbol{c}_n$ of the SA layer for all patch-level embeddings $\boldsymbol{e}_n$ can then be obtained by

$$\boldsymbol{c}_n = \sum_{m=0}^{n_{\text{patches}}-1} a_{nm} \boldsymbol{v}_m + \boldsymbol{e}_n \tag{7.7}$$

where the weight matrices $\mathbf{W}^k$ and $\mathbf{W}^q \in \mathbb{R}^{d_{\text{emb}} \times d_{\text{SA}}}$ along with $\mathbf{W}^v \in \mathbb{R}^{d_{\text{SA}} \times d_{\text{SA}}}$ are learnable parameters of the model and d_{SA} refers to the size of the self-attention embedding.

Bag-Level Attention-Based Multiple Instance Learning

It follows an attention-based MIL layer [114] that aggregates all context-aware patch embeddings of an input image into a one-dimensional feature vector. The context-aware patch embeddings are transformed by the attention-weighted sum as

$$\boldsymbol{b} = \sum_{k=0}^{n_{\text{patches}}-1} a_k \boldsymbol{c}_k \tag{7.8}$$

with individual attention weights

$$a_k = \frac{g(\boldsymbol{c}_k)}{\sum_{m=0}^{n_{\text{patches}}-1} g(\boldsymbol{c}_m)} \tag{7.9}$$

and $g(\boldsymbol{c})$ defined as

$$g(\boldsymbol{c}) = \exp\Big(\boldsymbol{w}^T \tanh\big(\mathbf{V}\boldsymbol{c}^T\big)\Big) \,. \tag{7.10}$$

Moreover, $\boldsymbol{w} \in \mathbb{R}^{d_{\text{MIL}}}$ and $\mathbf{V} \in \mathbb{R}^{d_{\text{MIL}} \times d_{\text{emb}}}$ are learnable weights of the model architecture. To sum up, this process generates an attention-weighted sum of the individual context aware embeddings $\boldsymbol{c}$ for the vector $\boldsymbol{b} \in \mathbb{R}^{d_{\text{emb}}}$ that now contains a bag-level representation of all provided patches for one image.

Target Prediction Head

As the last step in the model pipeline, an MLP is used to further process the bag-level latent representation to two output neurons. The bag-level representation $\boldsymbol{b}$ is transformed to the final prediction of the model with the two fully connected layers as

$$\hat{\boldsymbol{y}} = \text{softmax}\left(\mathbf{Y}^{(2)}\text{ReLU}(\mathbf{Y}^{(1)}\boldsymbol{b})\right) \tag{7.11}$$

where ReLU is defined as the rectified linear unit $\text{ReLU}(x) = \max(0, x)$ and the learnable weight matrices are defined as $\mathbf{Y}^{(1)} \in \mathbb{R}^{d_{\text{hid}} \times d_{\text{emb}}}$ and $\mathbf{Y}^{(2)} \in \mathbb{R}^{2 \times d_{\text{hid}}}$. The MLP maps the bag-level latent representation $\boldsymbol{b}$ to the binary prediction vector $\hat{\boldsymbol{y}} \in [0,1]^2$ that estimates the probability of having a relapse before five years or not.

7.2.4 Objective Function

For the output of the network, the objective function is based on $\hat{\boldsymbol{y}} \in [0,1]^2$ with two probabilities that estimate the probability of a certain input image if a relapse will occur within the first five years $\hat{y}_1$ or not $\hat{y}_0$. Note that the two events are the only possibilities for each patient leading to $\hat{y}_0 + \hat{y}_1 = 1$. For training, a weighted two-class cross-entropy loss

$$\mathcal{L}_{\text{y}} = -\sum_{i=0}^{1} w_i \big[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\big] \tag{7.12}$$

$$\tag{7.13}$$

is used where the one-hot encoded vector $\boldsymbol{y}$ indicates an individual's five-year relapse ($\boldsymbol{y} = [0,1]^T$) or not ($\boldsymbol{y} = [1,0]^T$) and $\hat{\boldsymbol{y}} \in [0,1]^2$ is the two-dimensional vector of the two corresponding predicted probabilities. To ensure that the two possibly imbalanced classes are equally weighted

during training, the number of positive samples n_1 and the number of negative samples n_0 is used to calculate the respective loss weights

$$w_i = \frac{n_{\text{total}}}{n_{\text{i}}} \tag{7.14}$$

based on the inversely proportional absolute frequencies n_{i} for the two classes and $n_{\text{total}} = n_0 + n_1$ as the total number of training samples.

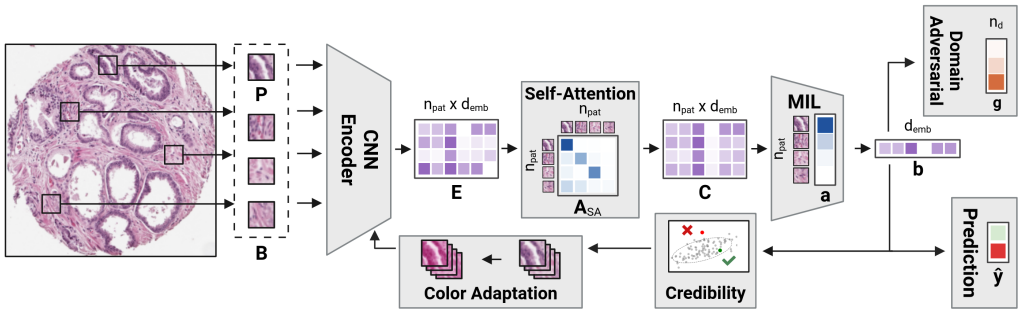


Fig. 7.5: The final PCAI architecture with the added robustness extensions of an additional domain adversarial head and credibility estimation-based color adaptation. Created with BioRender.com

7.2.5 Improving Model Robustness

The aforementioned model trained on TMAs from the UKE dataset can already produce good discriminative performance on the internal dataset as long as the model is used to predict on images from the same sub-dataset. When samples with higher variations are presented, the performance drops significantly impeding the model's usability for a clinical setting. The next section describes the strategies that were taken to overcome the aforementioned robustness problems. Therefore, the PCAI model utilizes additional training data from two more sub-datasets, that extends the diversity of images for the model during training, also introducing a second domain-adversarial (DA) head. In addition, credibility estimation (CE) is introduced that identifies images that are too far away from the training distribution based on the bag-level latent representation $\boldsymbol{b}$. Furthermore, histogram matching-based color adaptation (CA) aims to push those samples closer to the training data distribution. Fig. 7.5 shows these extensions schematically that are explained in additional detail below.

Domain Adversarial Training

As one extension towards better model robustness, a domain adversarial head is added to the model network during training. This adds another classification task to the network that distinguishes between multiple input domains of the training data.

Following [85, 254], a gradient reversal layer is introduced before the domain classification head that, during backpropagation, reverses the gradient of the learned domain distinction. The goal of this method is to merge the latent representations of those different domains that are used in training in the shared encoder part of the network.

Similarly to the target prediction head, the additional domain classification head consists of two dense layers that estimate to which domain an individual bag-level representation $\boldsymbol{b}$ belongs as

$$\hat{\boldsymbol{d}} = \text{softmax}\left(\mathbf{D}^{(2)}\text{ReLU}\left(\mathbf{D}^{(1)}\boldsymbol{b}\right)\right) \tag{7.15}$$

where $\mathbf{D}^{(1)} \in \mathbb{R}^{d_{\text{dom}} \times d_{\text{emb}}}$ and $\mathbf{D}^{(2)} \in \mathbb{R}^{n_{\text{domains}} \times d_{\text{dom}}}$ are learnable parameters of the network architecture with the number of dimensions d_{dom} and the number of target domains n_{domains}. Also, this domain head is trained using a BCE-based loss as

$$\mathcal{L}_{\text{d}} = \sum_{i=0}^{n_{\text{d}}-1} w_i \Big[d_i \log(\hat{\boldsymbol{d}}_i) + (1 - d_i) \log(1 - \hat{\boldsymbol{d}}_i) \Big] \tag{7.16}$$

where $w_i = n_{\text{domains}}/n_{\text{i}}$ are the inversely proportional weights of the different training domains based on the absolute frequencies n_{i} for each one-hot encoded target domain vector $\boldsymbol{d}$ where $i \in \{0, \ldots, n_{\text{domains}} - 1\}$. This loss is then added to eq. (7.13) to produce the new overall objective function of the model

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{y}} + \lambda \mathcal{L}_{\text{d}} \tag{7.17}$$

that is used to train both heads in parallel to produce less domain-specific latent representations of the bag embeddings $\boldsymbol{b}$.

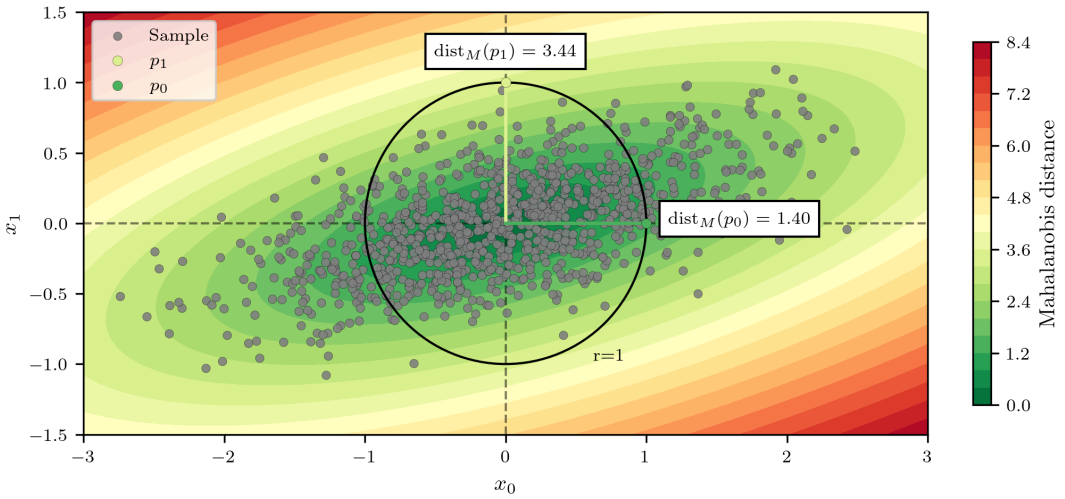


Fig. 7.6: Example of Mahalanobis distances for a closer and a further sample regarding a reference distribution with the same euclidean distance of 1 to the distribution's center in two dimensions

Credibility Estimation

With credibility estimation (CE), this work introduces a latent-space-based measurement to determine how close a new input image's latent representation is to the training distribution following [141, 173, 244]. This work uses the measure on the latent distance of an image's bag representation $\boldsymbol{b}$ from the bag-level latent representations of all training samples. As a distance measure, the Mahalanobis distance is used that measures the distance in standard deviations from the mean of the training distribution [186]. An example of the Mahalanobis distance is shown in fig. 7.6 for a 2-dimensional distribution that compares a euclidean distance of 1 for the center of a distribution to the Mahalanobis distance. The image shows why this metric is better suited to estimate the distance from a distribution compared to the Euclidean distance indicated by the circle since the two points with equal euclidean distance from the distribution's center should be treated differently regarding in-domain classification.
The Mahalanobis distance is defined as

$$\|\boldsymbol{b}\|_M = \sqrt{(\boldsymbol{b} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{b} - \boldsymbol{\mu})} \tag{7.18}$$

where $\boldsymbol{\mu} = [\mu_0, \mu_1, \ldots, \mu_{d_{\text{emb}}}]^T$ is the vector of mean values per individual dimension and $\boldsymbol{\Sigma} \in \mathbb{R}^{d_{\text{emb}} \times d_{\text{emb}}}$ is the covariance matrix of a reference distribution. For this chapter, this distribution is the training dataset's bag-level representations.

To derive the credibility of a sample $\boldsymbol{b}_j$, a calibration set of bag-level representations is defined following [230] as the label-dependent bag-level representation subsets

$$\mathbb{B}_k = \left\{ \boldsymbol{b}_i \mid y_i = k,\, i \in \{0, 1, \ldots, n_{\text{patches}}\} \right\} \tag{7.19}$$

for $k \in \{0, 1\}$. The Mahalanobis distance is applied to each predicted sample in the label specific subsets to obtain

$$\mathbb{M}_k = \left\{ \left\| \boldsymbol{b} \right\|_M \mid \boldsymbol{b} \in \mathbb{B}_k \right\}. \tag{7.20}$$

This further allows the derivation of the credibility as

$$\text{cred}(\boldsymbol{b}_j) = \max_k \left(\frac{\left| \{ m \in \mathbb{M}_k,\, m \geq \left\| \boldsymbol{b}_j \right\|_M \} \right|}{|\mathbb{M}_k|} \right) \tag{7.21}$$

where $|\cdot|$ denotes the cardinality of a set. This calculates the fraction of samples with greater Mahalanobis distance regarding the label-separated calibration subsets $\mathbb{B}^k$. For the credibility measure, the maximum fraction of the calibration samples $\boldsymbol{b}$ that are farther away from the training center compared to the current sample's distance $\left\| \boldsymbol{b}_j \right\|_M$ of the two target class labels is provided as a sample's credibility in this thesis. This way, the output prediction of the model gets an additional measure of trustworthiness that is useful in a clinical context for the interaction with the clinicians. A credibility close to 1 now indicates a sample that is close to the training distribution while a low credibility indicates that no comparable sample was seen during model training according to the latent representation.

Color Adaptation

Since samples that are far from the training distribution can now be identified using CE from above, it is desirable to alter those samples in a way that brings their bag-level latent representation closer to the training distribution to ensure a more credible representation. One strategy that proved successful is to introduce color adaptation on those samples with histogram matching to the closest training histogram cluster in the HSV space as explained below.

For all training images, the set of all HSV histograms is divided into k clusters by performing a k-means clustering based on the channel-wise Wasserstein distance [201]. The Wasserstein distance measures the cost of transforming one distribution (here a histogram) into another. For the found clusters, a mean histogram is calculated that represents the k-th cluster of images of the training data. For a new source image, the HSV histogram of the k calculated mean histograms with minimal Wasserstein distance to the source histogram is then used as a reference. Histogram matching is then performed on the source's three independent channel histograms to transform them to the reference histogram. The matching uses the cumulative distribution function of the channel-wise histograms to match the source to the reference distribution.

For the final PCAI model, this method of CA is only used for samples that are considered non-credible by the first pass through the network as shown in fig. 7.5. If a non-credible sample is identified, one additional pass through the network using the CA altered image is performed to determine the output prediction.

7.2.6 Experimental Setup

Data Splitting

The three largest datasets of UKEhv, namely UKE.first, UKE.second and UKE.scanner are split stratified by the event indicator into 70 % training, 15 % validation, and 15 % test data on

patient-level to avoid leakage of patient information into multiple splits. This also defines the three domains that are used for the DA task of the network. For the actual training, samples from UKE.first are weighted twice since they are supposed to be most characteristic for the individual patient.

Moreover, the three smaller datasets, namely UKE.thin, UKE.thick, and UKE.long are not used for training, but are considered during validation. These datasets are split (again stratified by the event indicator) into 50 % validation and 50 % test data. All other datasets are used for model evaluation only thus belonging exclusively to the test dataset.

Patch selection

Since this model should also be used for predictions on biopsies, the preprocessing for those images adds an additional step of patch-selection due to the size of the biopsies which might contain up to 10 thousand patches instead of 100 for the TMA datasets. In the training step on TMAs, 100 over- or under-sampled patches of a single image are selected. To keep this consistent for inference, the same number of patches is preselected. Here, the 100 patches with the highest prediction by the CI model are used as an additional preselection step. This way the PCAI model is extended to not only work on TMAs, but also on the much larger scale of biopsy images.

Model Training

For the training and validation process of the models, all TMAs of the UKE sub-datasets are preprocessed as follows. With a fixed patch size of $p_s = 128$ pixels, a fixed number of $n_{\text{patches}} = 100$ patches is selected from all available patches per TMA. While some TMAs are oversampled if less than 100 patches are present, TMAs with more than 100 potential patches are undersampled.

Furthermore, the number of internal dimensions to generate the attention-based bag-level representation for $\boldsymbol{b}$ is set to $d_{\text{MIL}} = 128$. The patches are augmented using `AugMix` augmentation [105] to obtain the final input images for the training procedure. Further, the `EfficientNet-b0` model is initialized with `ImageNet` pre-trained weights. Also, the patches are normalized such that they show the same mean and standard deviation that was found in `ImageNet` pre-training.

During training, an Adam optimizer is used with a batch size of 16, a learning rate of $2.75e-06$ and a maximum of 200 epochs, but including early stopping based on the mean validation five-year relapse prediction AUROC. Further, dropout is used as an additional regularization strategy at a value of 0.34.

Robustness Extensions

For the PCAI model, the training splits of the three biggest sub-datasets, namely UKE.first, UKE.second and UKE.scanner, were included in the DA training providing $n_d = 3$ domains. During training, a custom training dataset is created that contains UKE.first twice and the other two only once to emphasize the most important sub-dataset that follows the default protocol for digitization of TMAs without any alterations.

For CA, this thesis defines a sample as an outlier if its credibility lies below 0.75. For color adaptation (CA), $k = 8$ clusters are used. These parameters were optimized by the mean five-year AUROC of the six validation sub-datasets.

7.3 Results

The PCAI model that was presented in the previous section is evaluated regarding several quantitative and qualitative measures. Results of BASE without the robustness approaches are shown for a comparison on all extracted datasets. A detailed table of the patient characteristics for the extracted datasets is depicted in tab. A10.

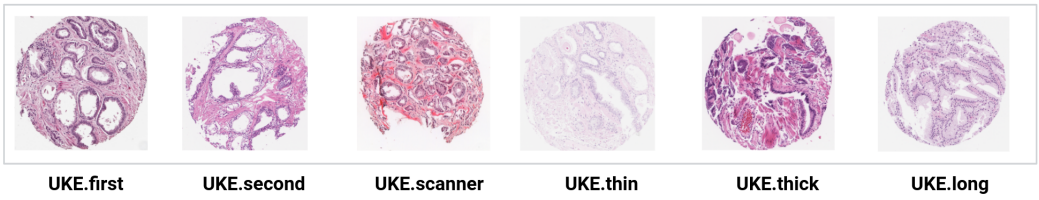
Firstly, this chapter deals with the influence of input data variation for the BASE and PCAI model to analyze where the robustness extensions of the model lead to better performance. This is mainly performed on the different sub-datasets of UKEhv as well as the NYU and JHU datasets. Afterwards, this analysis is extended to the datasets where image-level human ISUP annotations are given. The discriminative model performance of BASE and PCAI is compared to one or multiple human annotators based on individual images.

7.3.1 Effect of Input Data Variation

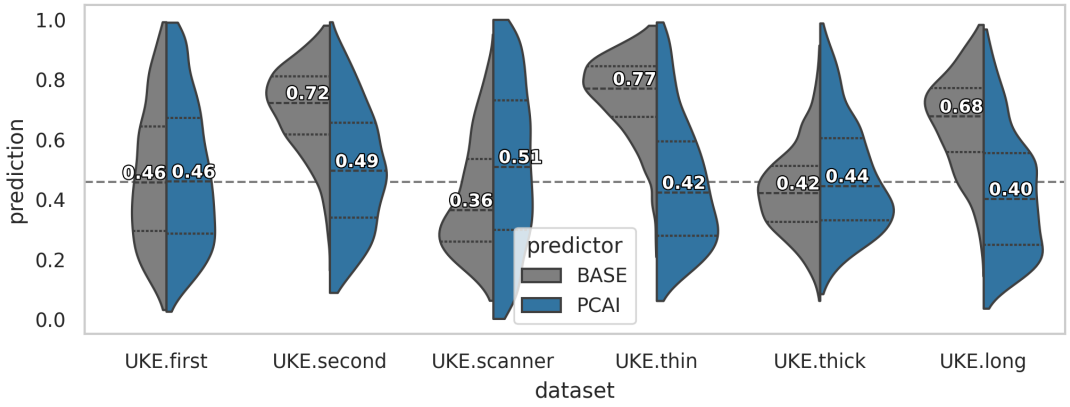
Variations in the input images can significantly reduce the performance of DL-based algorithms compared to the influence on humans [11, 30]. The sub-datasets in UKEhv show variations in the histopathological images that originate from differences in formalin fixation, paraffin embedding, sectioning staining, and the digitization process. The first comparison is done on a subset of the UKEhv dataset where only patients are included that provide TMA images in all six sub-datasets leading to a total of 1,537 patients to analyze the robustness extensions that were described in sec. 7.2.5. For a comparison, exemplary TMAs of the UKEhv sub-datasets excluding UKE.sealed are illustrated in fig. 7.7a with visible differences in image brightness and saturation.

Fig. 7.7b shows the prediction distributions for BASE and PCAI on each sub-dataset in UKEhv. It can be observed that the BASE and PCAI prediction distribution for UKE.first look similar. BASE shows a larger difference compared to the UKE.first distribution than the PCAI model. The largest median prediction difference with 31 pp of the BASE model is found in the UKE.thin sub-dataset with a median prediction of 0.77 compared to 0.46 in UKE.first. In contrast, the PCAI prediction distributions differ less from the UKE.first predictions with a maximum median difference of 6 pp in the UKE.long sub-dataset with 0.40 compared to a median of 0.46 in UKE.first. These findings show that the robustness extensions help to keep the prediction distribution of the model more consistent with the varying input image attributes for overlapping patients.

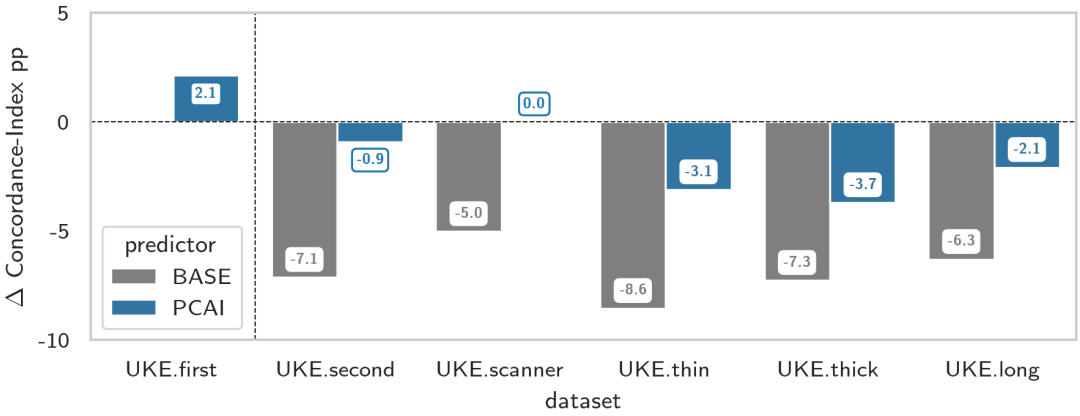
In terms of discriminative performance on the internal UKEhv sub-datasets as shown in fig. 7.7c, it can be observed that the performance of BASE drops significantly for the five out-of-training datasets, even though this model is already trained on 8,123 images and patients. The performance regarding the C-index drop significantly from 0.645 for UKE.first by 5 pp (UKE.scanner) and 8.6 pp (UKE.thin). Similar results can be observed for the 3-, 5-, 7-, and 10 year AUROC. This finding demonstrates the difficulty of the BASE model to deal with previously unseen dataset variations in the TMA images.



(a) Exemplary TMA samples.



(b) Prediction distributions for the overlapping patients.



(c) Relative C-index for BASE (gray) and PCAI (blue) compared to BASE C-index on UKE.first.

Fig. 7.7: Comparison of BASE and PCAI on UKEhv sub-datasets. Variations in spot appearance (a), the resulting prediction distribution (b) and the C-index difference compared BASE on UKE.first for all UKEhv sub-datasets excluding UKE.sealed with 1,537 patients that provide images in each sub-dataset to visualize influence of data variation on BASE and PCAI model predictions.

This drop in performance motivates the robustness extensions from sec. 7.2.5 for the PCAI model. Note that PCAI was not only trained on UKE.first, but also included training samples from UKE.second and UKE.scanner. Fig. 7.7c and fig. 7.8 show that the discriminative performance of the BASE model can be improved by the PCAI model throughout the UKE sub-datasets. A detailed table with the obtained scores can be found in tab. A11. This also holds true for the external TMA datasets NYU and JHU (see fig. 7.9). Regarding C-index, the performance could even be raised by 2.2 pp on UKE.first which is the only dataset that is used in training for the BASE model. Further, C-index was enhanced across all sub-datasets from 3.5 pp on UKE.thick (0.573 to 0.609) to 6.2 pp on UKE.second (0.602 to 0.661). The same behavior can be observed for the yearly AUROCs except for years 8-10 in UKE.scanner depicted in fig. 7.8. Additionally, an improvement regarding all metrics shown can also be observed for NYU and JHU e.g. regarding C-index with 5.3 pp (0.641 vs. 0.694) and 1.0 pp (0.577 vs. 0.587) respectively.

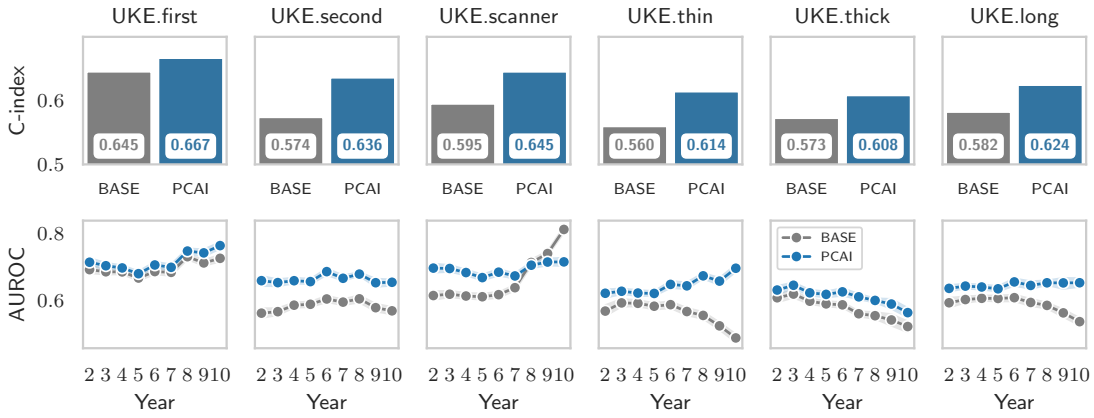


Fig. 7.8: Discriminative performance of BASE (gray) and PCAI (blue) regarding C-index and AUROC over time for the sub-datasets in UKEhv excluding UKE.sealed.

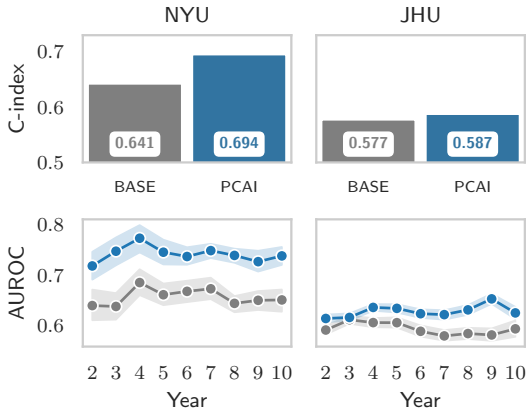


Fig. 7.9: Discriminative performance of BASE (gray) and PCAI (blue) regarding C-index and AUROC over time for the external TMA datasets NYU and JHU.

Ablation: Robustness Extensions

The robustness extensions of PCAI that consists of including two additional sub-datasets into the training, domain-adversarial training and (CE-guided) color adaptation are evaluated in an ablation study. Since the extensions aim to make the model more robust on external data, the individual influence towards performance on the external datasets is shown. Note that the CI model was used to preselect the 100 patches with the highest cancer prediction for the biopsy datasets UPP and MMX. The models presented are the 9 combinations (with the number of choices in parentheses) of:

Training domains (2) The model is either trained on images from UKE.first (5,710 images) exclusively or additionally includes UKE.second and UKE.scanner (5,700 and 4,968 images respectively), but weights images from UKE.first twice.

DA (2) Whether domain-adversarial training was used as described in sec. 7.2.5 or all samples were used without further domain knowledge. Note that DA can only be used in combination with the larger training dataset (sec. 7.2.5).

CA (3) If color adaptation was used on either none, all or CE-guided samples (sec. 7.2.5).

Tab. 7.2 shows the mean and standard deviation of the C-index and AUROC for the external datasets on 9 combinations of the robustness extensions that were independently trained and hyperparameter tuned on the UKEhv sub-datasets. It can be observed that the more robustness measures are taken, the better the mean discriminative model performance becomes on the external datasets regarding C-index and five-year AUROC. The biggest performance gain compared to the first row that represents BASE is observed by adding the two largest additional sub-datasets UKE.second and UKE.scanner (approximately tripling the available training data) joined by minor performance boosts when DA and selective CA based on CE are added. However, adding the robustness extensions separately might lead to a drop in performance when using DA training and no CA.

To sum up, adding the additional training data already gives a significant performance on the external datasets without introducing more complexity in the model. However, this removes the credibility score (as part of the CE module) that makes the model more interpretable by providing an additional score during inference showing how close the new image is to the training data distribution.

Tab. 7.2: PCAI ablation study results are presented with the best performance highlighted in bold. The mean C-index and 5-year AUROC on all external datasets (NYU, JHU, UPP, MMX) is used to evaluate the robustness of the extensions (if included, indicated by •. CE means selective CA based on CE results) of the PCAI model.

training (sub-)domains	DA	CA	C-index	AUROC5
UKE.first			0.642 ± 0.095	0.667 ± 0.114
		•	0.614 ± 0.101	0.637 ± 0.102
		CE	0.614 ± 0.101	0.637 ± 0.102
UKE.first UKE.second UKE.scanner			0.665 ± 0.076	0.686 ± 0.091
		•	0.674 ± 0.124	0.686 ± 0.130
		CE	0.667 ± 0.131	0.678 ± 0.145
	•		0.655 ± 0.071	0.682 ± 0.083
		•	0.674 ± 0.134	0.694 ± 0.136
		CE	**0.688 ± 0.126**	**0.703 ± 0.133**

7.3.2 PCAI vs. Image-Level Annotations

UKE.sealed (TMA dataset)

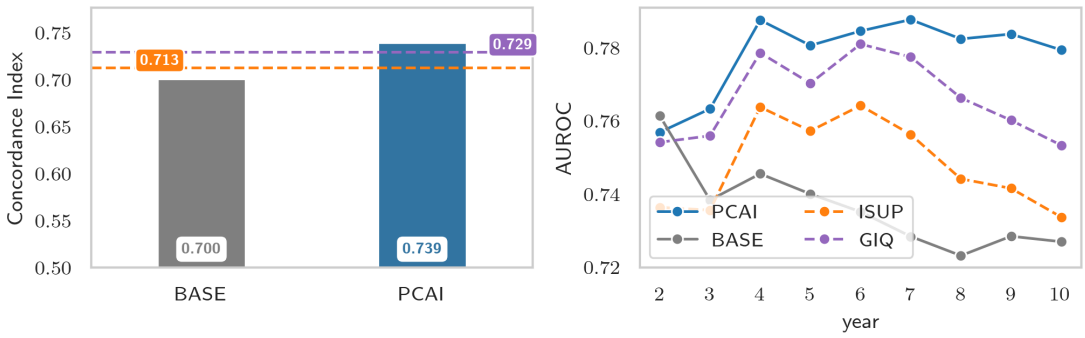
This section evaluates BASE and PCAI predictions to individual image-level human annotated TMAs from the UKE.sealed dataset. This dataset does not only contain image-level ISUP annotations from the department of pathology of the UKE, but also GIQ annotations as presented in sec. 2.1.2 which is considered the best performing grading system in PCa histopathology [207]. Note that the evaluation of this dataset was performed externally after model development to avoid leakage into the training and development process of the PCAI model.

To evaluate image-level annotations and compare them to patient-level survival information, the existing annotations are aggregated as follows. For ISUP grading, the maximum image-level annotation among all images for a single patient is used. For GIQ, BASE and PCAI, the mean and max of the image-wise predictions for each individual patient are considered. The results in fig. 7.10 (with detailed numbers provided in tab. A12) show that PCAI outperforms BASE on all presented metrics significantly, for example the C-index by 3.2 pp and 3.9 pp for mean and max aggregation respectively. Similarly, expert ISUP annotations are outperformed by the max aggregated PCAI across all presented metrics, with 2.6 pp regarding C-index (0.713 vs. 0.739). When comparing PCAI to GIQ, mean aggregation shows similar discriminative performance regarding C-index (0.744 vs. 0.743). For AUROC, the performance of GIQ is mostly better before dropping below PCAI at 8 years. Most notably, for max aggregation, PCAI performs slightly better than GIQ in all presented metrics by 1 pp (C-index), 0.7 pp (3 year AUROC), 1.1 pp (5- and 7- year AUROC) and 2.6 pp for 10-year AUROC. In general, GIQ and ISUP show a drop in performance over time while PCAI remains more stable. Overall, slightly higher scores are achieved across the presented metrics when mean aggregation is used for GIQ and PCAI.

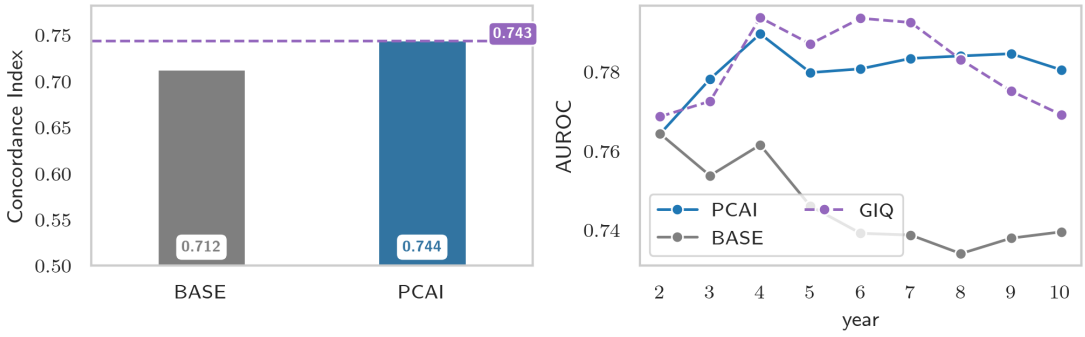
UPP and MMX (Biopsy datasets)

The clinically relevant evaluation of the algorithm is performed on the biopsy datasets that may help the urologist regarding risk prediction and treatment planning. This thesis evaluates the performance of PCAI on UPP and MMX that were described in detail in sec. 3.2.3. As discussed in sec. 7.1, biopsies are significantly larger and may contain over 10000 valid patches. This thesis only uses the 100 most relevant patches according to the patch-based CI prediction from the previous chapter were used for this analysis. Similar to sec. 7.3.2, these datasets provide image-level ISUP annotations that are aggregated (by using the worst annotation) to patient-level to compare it to the patient's actual survival information that is shown in fig. 7.11 (an additional table for the obtained results is provided in tab. A13).

For UPP, PCAI scores a C-index of 0.604 which is slightly above the ISUP annotation (0.597). Further, BASE only reaches a C-index of 0.581. Regarding yearly AUROC, PCAI achieves better performance than ISUP after year 2 with a five-year AUROC of 0.672 vs. 0.659. Moreover, the MMX dataset shows the same trend. PCAI achieves the best C-index of 0.864 while the BASE model only scores 0.779. Further, PCAI scores higher than all human annotators for C-index (0.864 for PCAI, 0.838, 0.834 and 0.641 for A1, A2, and A3 respectively). The same results can be observed throughout the yearly AUROC for example at year 5 where PCAI shows the highest AUROC of 0.868 compared to the human annotation with 0.819, 0.817, and 0.657 for A1, A2, and A3 respectively. The maximum AUROC of PCAI is achieved for 10 years with a value of 0.890.

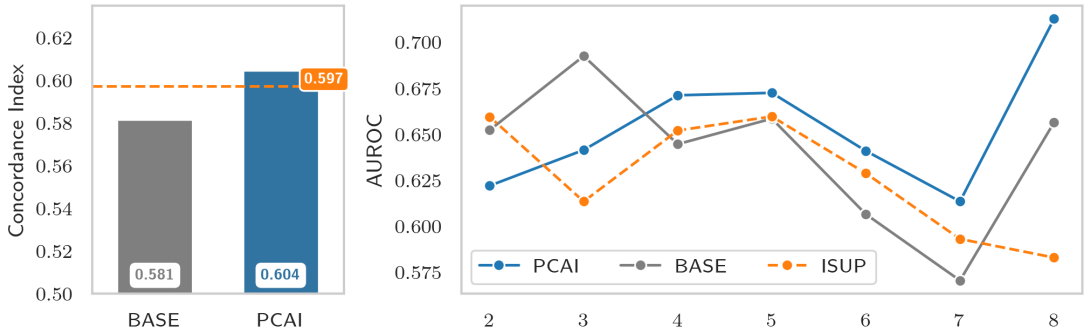


(a) Max aggregation

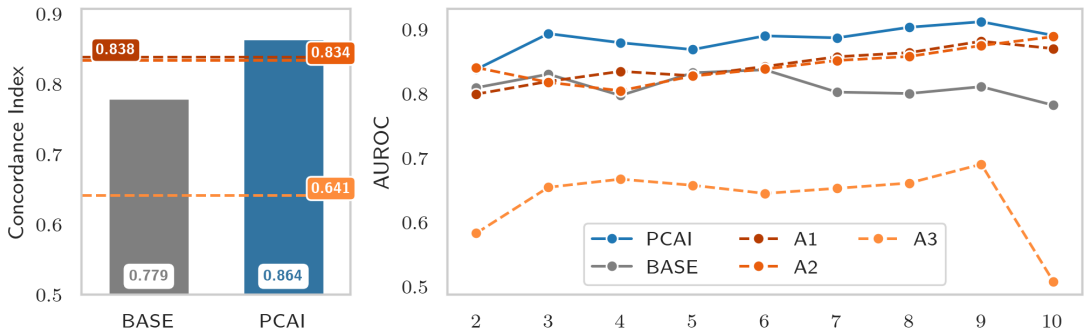


(b) Mean aggregation

Fig. 7.10: Discriminative performance of ISUP (orange), GIQ (purple), BASE (gray) and PCAI (blue) with max (a) and mean (b) aggregation regarding C-index (left) and 2-10-year AUROC (right) for UKE.sealed.



(a) UPP dataset



(b) MMX dataset

Fig. 7.11: Discriminative performance of human ISUP annotations (orange shades), BASE (gray) and PCAI (blue) regarding C-index and 2-10-year AUROC for UPP (a) and MMX (b).

7.3.3 Risk Stratification

Another step towards clinical applicability of PCAI's predictions lies in risk stratification based on the continuous prediction similar to the risk group assessment described in sec. 5.2.4 for discrete output survival curves. Firstly, a median-based comparison of low vs. high risk is conducted, followed by an analysis of the maximum number of risk groups that can be generated for the UKEhv dataset. It is shown that this method creates up to seven statistically significantly different risk groups.

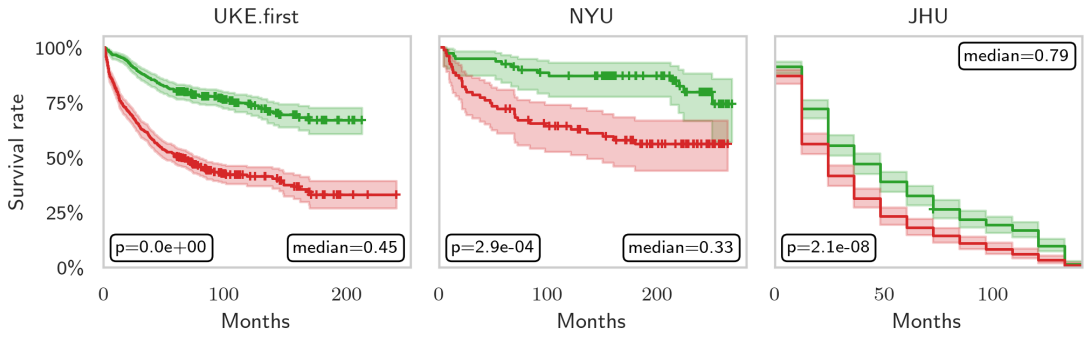
Median-based Stratification

A simple approach that allows stratification into low and high risk groups is comparing the patient-level PCAI predictions to the median of each individual dataset. This creates a risk stratification into a low and a high risk group as visualized in fig. 7.12. It can be observed that this method allows a simple stratification of the datasets into a low and a high risk group that is statistically significant (with all p-values being below 5 %).

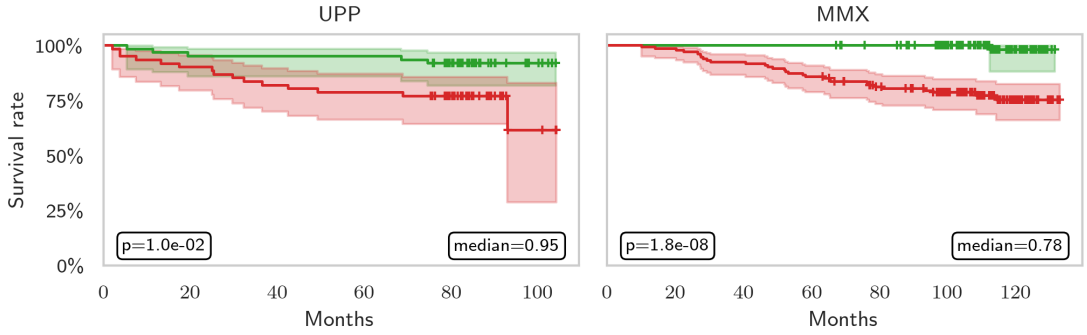
However, it can be observed that the value of the median differs significantly for the TMA datasets in fig. 7.12a and the biopsy datasets in fig. 7.12b. This indicates that the risk predictions of individual images is not independent of the dataset and has to be considered for an interpretation of the derived risk groups. However, when comparing the median prediction for the TMA datasets in fig. 7.12a, the order of medians (0.79 for JHU, 0.45 for UKE.first, and 0.33 for NYU) indicates that the patient's with the worst expected survival can be found in the JHU dataset which can also be observed in the corresponding survival curves that show only a 10 % survival rate at 100 months for the high risk group compared to corresponding values of 45 % for UKE and 70 % for NYU. However, this behavior does not extend to the biopsy datasets in fig. 7.12b that show even higher median values (0.95 for UPP, 0.78 for MMX), but also relatively high survival predictions for the high risk group at 100 months of 65 % and over 75 % for UPP and MMX respectively.

Maximum Risk Stratification

To achieve the maximum number of risk groups for the UKEhv dataset, alg. 3 is reused based on the scalar output predictions of the PCAI model. A maximum number of statistically significantly different risk groups can be obtained in the same manner. The algorithm is used on the estimated risk scores for the UKE training and validation dataset of PCAI. The images of UKE.first, UKE.second and UKE.scanner are used for the approach that is then evaluated by the log-rank test on the validation sets of the six sub-datasets. The results of how the predictions translate into the different risk groups, the KM curves of the observed patient survival, and the pairwise comparisons of the log-rank test for a maximum of 7 obtained risk groups are shown in fig. 7.13. As expected, more adjacent groups show less statistically significant stratification quantified in fig. 7.13b. Additionally, the distribution of all PCAI predictions for the validation sets is shown in fig. 7.13a. In summary, this approach enables a direct risk group prediction from 1 to 7 based on the PCAI risk prediction for an individual image which can be useful for clinical application of the approach. This is shown in fig. 7.14 for the test set of UKE.first that was not part of the training or validation of the aforementioned algorithm. It can be observed that the risk groups stratify well for the observed survival visualized in the KM-curves regarding relapse after RP. However, as shown in sec. 7.3.3, those risk groups cannot be extended to other datasets or domains without proper recalibration. This is why this section only shows the risk group estimation for the training data domain.

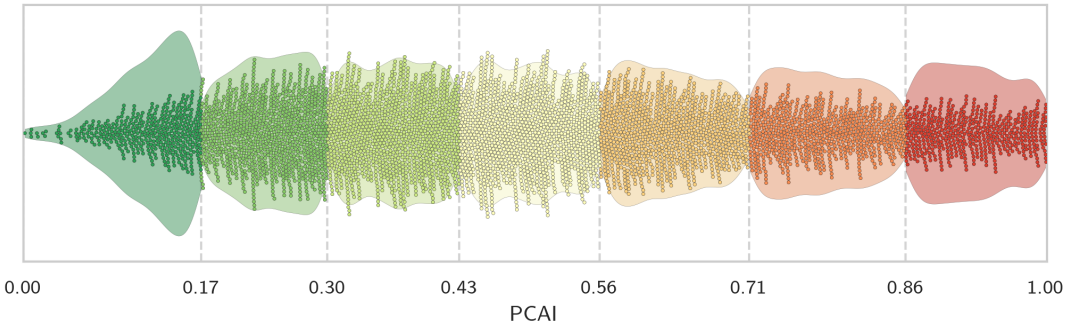


(a) TMA datasets

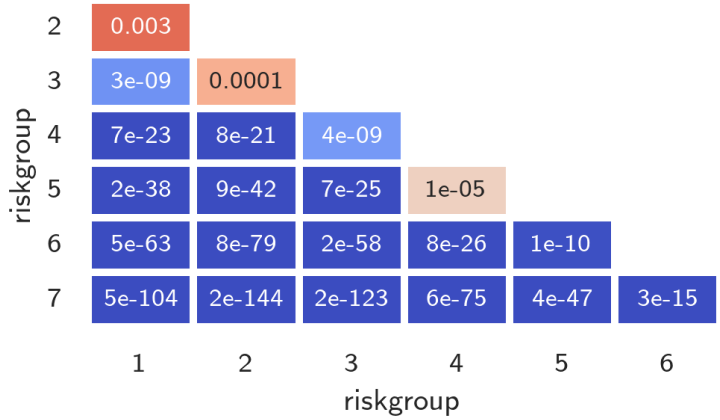


(b) Biopsy datasets

Fig. 7.12: KM-curves of TMA and biopsy datasets split at patient-level median prediction for the low and high risk group. Also shown is the p-value of the log-rank test the indicates statistically significant differences in the two respective groups.



(a) PCAI risk distribution and assigned risk groups from 1 (left, green) to 7 (right, red).



(b) Pairwise risk group log-rank test p-values.

Fig. 7.13: Distributions with corresponding risk group selection and log-rank test results for predictions obtained from UKEhv for the 7 obtained risk groups defined by PCAI for the validation data that was used to determine the maximum number of risk groups.

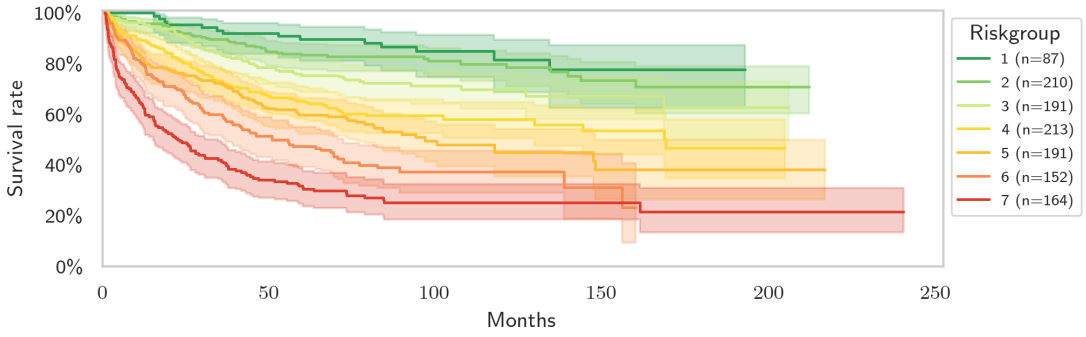


Fig. 7.14: KM-curves of the observed survival for the unseen UKE.first test dataset patients showing good separation among the obtained risk groups.

Comparing PCAI groups to ISUP

To better understand the relationship between the two grading systems in additional detail, this section compares the derived PCAI risk groups of the UKE.first images to the corresponding patient-level ISUP groups of the dataset.

Fig. 7.15 shows how the newly derived PCAI risk groups are composed of the different ISUP grades. On the left and right, the patients of UKE.first are analyzed in terms of five year relapse ratio per ISUP and PCAI risk group. The red part of each bar represents the ratio of patients that experienced a relapse within 5 years after RP while the green fraction of each bar represents relapse-free patients. A good risk score would show an increasing ratio for higher risk groups. It can be observed that this is violated for ISUP with a slightly worse percentage for ISUP1 compared to ISUP0 and a higher relapse percentage for ISUP3 compared to ISUP4. This may be caused by variance due to only 24 patients within the ISUP4 group in the whole UKE dataset compared to 602 patients in the ISUP2 group. For PCAI, the relapse ratio corresponds well to the obtained risk groups showing higher relapse ratios for higher risk groups.

The middle part of the image shows the flow from ISUP grade to PCAI risk group. It can be observed that the highest risk group of PCAI, namely PCAI7, is mainly composed of ISUP grade 3, 4 and 5 while the lower ISUP groups of ISUP0, ISUP1 and ISUP2 are the main contributors to the lowest risk group PCAI1. However, a few patients show conflicting results since there exists for example a flow from ISUP5 to the lowest risk group PCAI1 or from ISUP0 to PCAI7. This may be explained by the difference in annotation level of the two grading systems. While the shown PCAI risk groups originate exclusively from the individual images, ISUP is taken from the EHR on patient level. This means that the image that was used for PCAI may not contain the information that led to the specific ISUP grading.

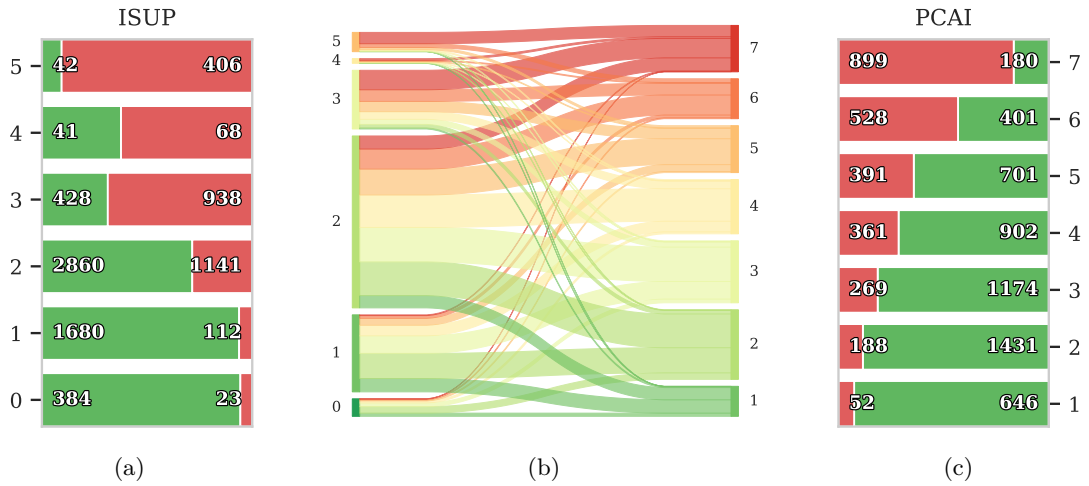


Fig. 7.15: Comparison of ISUP grade to PCAI risk group for UKE.first. (a) and (c) show the relapse ratios as horizontal bars per ISUP and PCAI risk group respectively. The red part of each bar represents the ratio of patients that experience a relapse within the first five years. Additionally, the absolute number of cases with and without relapse are shown. Also, (b) illustrates the flow from the patient's ISUP grade to each PCAI risk group, where the size of each flow represents the corresponding number of patients.

7.3.4 PCAI Biomarker on UKE.first

As a proof of concept, the prediction of the PCAI algorithm can also be used as an independent biomarker together with other covariates of the individual patient's EHR. This way the same kind of analyses that were presented in chapter 5 can be performed using the survival estimation models CoxPH and DCS. One dataset where this can be shown is the UKE.first dataset since it provides some additional patient information along with the individual TMAs. In detail, this dataset provides the characteristics age, positive resection status, LNI, and PSA level at RP.

For this analysis, the patient-level PCAI risk prediction is added as an additional input and interpreted as a digital biomarker for training the two survival models regarding cancer relapse after RP. This leads to the scores presented in fig. 7.16. The univariate models using PCAI both score a C-index of 0.695. This discriminative performance can be improved by approximately 2 pp when PCAI is combined with one of the additional features PSA level, positive resection status (margin status), or LNI. This boost can be observed for the three additional patient-level features in the DCS model. For CoxPH, only a combination with positive resection status increases the score to approximately the same value. Combining more than two features leads to additional boosts in performance up to 0.745 for DCS when all four features are combined. Also, DCS outperforms CoxPH in every case and performs on par when only PCAI is used.

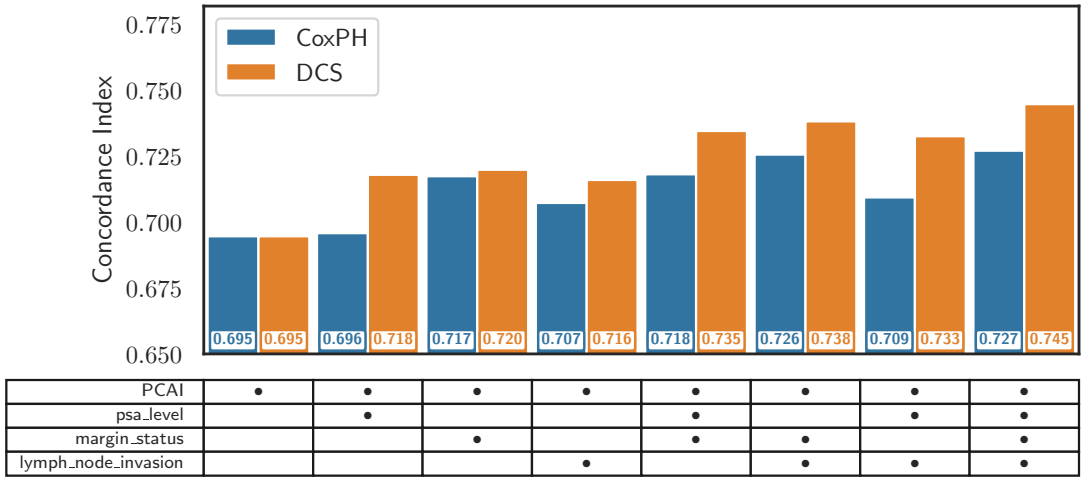


Fig. 7.16: Combining PCAI with additional patient information from UKE.first for multivariate DCS and CoxPH models.

7.4 Discussion

This chapter described the development of a DL-based end-to-end cancer risk prediction model, called PCAI, for PCa images from TMAs or biopsies that is trained using a variety of TMAs from the department of pathology of the UKE that provided one of the largest TMA datasets with detailed FU information in the world. Model performance, robustness, trustworthiness, and interpretability are improved by adding additional input data variety and adapting the model architecture. The approach is trained using objective survival information to avoid human-based Gleason annotations.

In general, PCAI generalizes better than the BASE implementation on unseen TMA datasets from the same center in UKEhv and on external data from NYU and JHU. The robustness extensions, namely including additional TMAs with variations in the acquisition protocol, DA training and CE-based CA improve overall results regarding discriminative performance.

When comparing PCAI to human annotators, individual, image-based annotations are used for a fair comparison. Firstly, PCAI shows better performance in terms of C-index and yearly AUROC than individual, image-based ISUP grades on UKE.sealed. Further, PCAI performs on par with the more complex GIQ that takes the individual percentages of GG3-5 into account.

Additionally, it is analyzed if the model also produces reasonable results for biopsy images. As an additional preprocessing step, the CI model from chapter 6 is utilized to select the most cancerous patches of each biopsy. With this adaptation, PCAI outperforms for the two external biopsy datasets with one ISUP annotation for UPP and three ISUP annotations for MMX. Overall, PCAI yields the best discriminative performance regarding C-index and shows the best results regarding yearly AUROC except for years 2 and 3 in the UPP dataset. The maximum AUROC for PCAI is achieved for a ten-year prediction with a value of 0.89.

The derived risk groups based on UKE.first, UKE.second, and UKE.scanner show high discrimination on the corresponding validation set such that a total of 7 distinguishable risk groups could be derived. This risk group can further be used in clinical reports and provides easier interpretability than the continuous score. However, it was also shown that the absolute predictions show bias when other datasets are analyzed. The median prediction for each dataset varies from 0.33 (NYU) to 0.95 (UPP). This shows that further adaptations or calibration need to be made if the provided risk groups with fixed intervals should be utilized further. However, the finding that a model that is trained exclusively on TMAs also performs well in terms of discriminative performance when it is applied to biopsy datasets proves the validity of the presented approach.

Furthermore, the CE itself can also be reported and thus provide a measure of credibility for the model's prediction. A high credibility indicates that similar samples were observed during model training while samples with low credibility might contain unseen cases that could be passed to a human pathologist for another opinion. This further can further improve the interpretability and trustworthiness. However, such a measure is not provided for ISUP grades that is currently used in clinical practice. A report for the practitioner could additionally contain cancer heatmaps for the individual biopsy slides to further improve interpretability and trustworthiness not only in PCAI predictions but the overall pipeline.

As a proof of concept, PCAI can also be utilized as an independent biomarker together with other clinical factors. This approach can be useful when additional information that is not present in the images is provided like the PSA level, knowledge about metastatic spread, LNI, or a positive resection margin. For UKE.first, this approach can be tested since some of those features are present. The combination of these factors yields better results when the DL-based survival prediction model DCS from chapter 4 is utilized compared to CoxPH. This combination of input features further improves the results in discriminative performance.

7.5 Conclusion

This chapter showed the development of the image-based PCa relapse risk prediction model PCAI. The scalar prediction endpoint based on five year survival that was used during training proved feasible and extended to external TMA datasets and to biopsy images.

The robustness extensions of the model, namely the inclusion of datasets with variations in sample preparation, DA training and CE guided CA improved overall discriminative performance compared to the BASE model. PCAI proved more robust than the baseline implementation throughout multiple internal datasets with variations in sample preparation, e.g. staining protocol, slice thickness or the scanner vendor that was used for digitization. Further, it could be observed that the discriminative performance regarding C-index for the internal UKEhv sub-datasets was significantly reduced for the BASE dataset up to 8.6 pp (UKE.thin) when different sample acquisition protocols were used. For PCAI, this drop could be reduced to a maximum of 3.7 pp (UKE.thick). The results were further confirmed by an ablation study that showed that adding additional training data from other sample acquisition protocols improves model performance the most, domain adversarial training has a minor influence and CE-guided color adaptation brings another small boost in mean C-index for the four external datasets.

For a comparison to the gold standard of Gleason grading, three datasets with image-level annotations were integrated with one (UKE.sealed, UPP) or even three (MMX) annotations for each individual image. To enable the model to predict on much larger biopsy images, patch selection was performed using the CI model that was presented in chapter 6. The CI model identified the most cancerous areas of each biopsy that were afterwards used in the PCAI risk prediction. It was further shown that PCAI is able to outperform the human annotators regarding ISUP annotation (2.6 pp vs. the best annotation) on MMX, be slightly better on UPP (0.7 pp) and score at least on par with the more complex GIQ grading on UKE.sealed (0.1 pp for mean, 1 pp for maximum aggregation respectively). Similar results can be observed when yearly AUROC performance is analyzed.

Moreover, the results showed that PCAI risk predictions can be used to stratify the cohorts that were analyzed in this dataset. A stratification in low and high risk groups can be achieved by median splitting of the predicted patient-level risks for the individual datasets. The resulting stratified KM curves show good separation which is additionally quantified by the pairwise log-rank test that yields values below 5 % for all five analyzed datasets. However, the different median values for the datasets hint that further recalibration of the model output is necessary to obtain a more universal risk stratification for individual domains. Nonetheless, this thesis optimized the risk stratification for the internal UKEhv dataset. Similar to the analysis in sec. 5.3.4, a maximum number of seven risk groups could be found that still yield statistically significant differences in the observed survival curves of those patients.

It could be shown that a risk interpretation of the PCAI prediction can be used together with additional patient variables to improve overall relapse risk prediction on the UKE.first dataset. Discriminative performance could further be improved when it is combined with the DCS survival prediction model that was introduced in chapter 4.

Lastly, the PCAI model is available in a public repository.[22]

7.5.1 Future Work

However,several drawbacks of the presented approach can be addressed. Firstly, the model architecture might be reviewed. One idea would be to change the endpoint of the network to directly

[22] https://github.com/imsb-uke/pcai

predict a risk rather than using the five-year endpoint as demonstrated in [88]. This approach would be closely related to the actual training of the CoxPH model and could be achieved by emphasizing a comparing loss like the kernel loss that was used in the development of this work's DCS survival model. Further, the integration of cell type (e.g. glandular cells or stroma) and nuclei density information might be beneficial for model development and interpretability. As an example, an approach for pancreatic cancer showed the influence of desmoplastic stromal tissue areas on patient outcome [157]. While this approach might need higher magnification levels as the 20x magnification that was used in this work, it can lead to the integration of previously hidden information that can be useful for risk prediction. A similar approach can be found in [5] that derives biologically meaningful classifications of tissue regions from hand-crafted, biologically relevant features based on cell- and nuclei- segmentation masks. Due to the human intervention in feature derivation, the interpretability of the resulting model can be improved. This information can further be used in a graph neural network as presented in [223] to capture the underlying spatial structure of the tissue. Also, more complex transformer architectures [112, 251] might be able to boost the patch-based encodings themselves compared to the CNN-based approach that was used in this work. These architectures might provide better latent representations of the images, given that enough training data is accessible. Other architectures like hierarchical vision transformers [256, 263] might also be a reasonable alternative to the attention-based MIL that was used in this work.

Since the robustness extensions that were implemented in this work proved successful, it is worth investigating them further. The presented approach boosts model performance when compared to the baseline implementation. This indicates that the strategies to improve robustness might be worth exploring in additional detail. For PCAI, one measure is the credibility-guided color adaptation where the original image, the corresponding latent representation and a reference histogram based on the closest HSV cluster of the training is used to alter the histogram of the original image. Other approaches might be able to perform these adaptations in a less complex manner. One approach, among others, would be to allow a manipulation of the latent representation itself which would require a continuous latent space. This would simplify the steps that need to be taken to push one sample closer to the training distribution. Regarding DA training, this work used the differences in the sample acquisition protocol from the three largest sub-datasets as targets on the bag level representations for the TMAs. This approach might be worth revisiting since differences in color might be more consistent among the 39 blocks that could also be used in a more complex DA approach. As an alternative, unsupervised learning approaches can be considered to ensure better image representations by contrastive predictive coding [175]. Also, augmentation techniques were not the focus of this work even though especially color augmentation proved feasible to reduce a model's generalization error in histopathology [227]. Stronger augmentation might improve the robustness and generalizability of the presented approach with or without the DA strategy.

To provide patient-level predictions, this chapter usually used the maximum prediction that was found on the image-level to represent the risk for an individual. For the UKE.sealed dataset it was shown that mean aggregation led to improved discriminative performance which is a non-intuitive result. However, this finding might hint that more complex aggregation methods are feasible to determine a patient-level aggregation of the image-level predicted risks. One approach would be to feed all images at once to the PCAI model to avoid the post-processing step of aggregating image-level predictions.

In general, this work integrated a large variety of TMAs and biopsy images that were used for model training as well as internal and external validation. The TMAs that were provided from the department of pathology of the UKE provide a high quality basis for model development. However, the external datasets that were analyzed vary in quality. While the JHU dataset contains high quality FU data and up to eight TMAs for a single patient, the biopsy datasets do not provide this level of detail. The data integration part showed that around 82 % of the patients in the UPP dataset had to be removed due to insufficient FU information. Optimally, a

biopsy dataset with high quality images and long FU data would improve at least the model evaluation. Such a dataset could also be used directly in training as described above.

Moreover, the patch selection for biopsies could be revisited. For now, only the patches that produce the highest cancer predictions based on the CI model are used. Various other approaches exist, like educated cluster-based sampling [190] or the removal of confident true negative instances [189,190]. The latent representation of individual patches could also be used for increased model interpretability as demonstrated in [184] that derive tissue concepts based on a clustering approach based on this latent representations that is further correlated in terms of cancer severity. Also, alternatives to the used attention-based MIL approach might be worth revisiting like CNN-based feature maps [1] or hierarchical transformer architectures [43] might produce better bag-level representations for an individual slide.

It could be shown that the model that is exclusively trained on TMAs can also be used for whole biopsy slides. This opens the path to possible clinical applications since the cancer severity estimate for biopsies might be used in the decision process for a urologist regarding initial treatment decision. However, if such a prediction were used in clinical practice, further development of the current model would become necessary. Since biopsy datasets with long follow-up times are rare and often rather small compared to the used TMA dataset from the UKE, a possible approach could be a retraining of the present PCAI model directly on biopsy images. A dataset that was not used in this work but might be beneficial in this regard is the PLCO dataset as mentioned in appendix A. This would lead to better calibrated predictions for the target domain. These calibrated predictions could also be used for a biopsy-based risk stratification similar to the approach presented in this thesis. With such a well calibrated model, the next steps towards clinical applicability can be taken. For the integration of the PCAI system into clinical workflow, the predictions could be used as an additional measure for initial treatment decisions for urologists. It should be analyzed if such a system can improve initial treatment decisions when a biopsy shows an ISUP grading of 2 or 3, which is between a passive approach like AS and a radical treatment like RP.

Lastly, dataset cleaning could be improved by the CI model for TMAs. The main use case would be the removal of TMAs with high patient-level ISUP annotations that do not show any cancerous regions according to the CI model. This combination is likely to show at least non-representative TMAs for the documented ISUP grading. This can be observed in fig. A3, where the different TMA blocks that were used for model training show different behavior regarding the maximum CI prediction that is present for individual TMAs. Patients that already show cancerous spread as metastases, capsular extension, or seminal vesicle invasion could either be removed from the image-based learning process or this additional information could be embedded into the model itself since those factors highly influence patient survival but cannot be learned from individual images. If the aim of such a network is the development of a purely image-based prediction of cancer aggressiveness, the former should be preferred.

8 Overall Summary and Conclusion

8.1 Summary

This thesis contributes to state-of-the-art survival analysis and risk prediction in PCa risk assessment. First, the discrete calibrated survival model DCS [84] was presented that is able to boost discriminative and calibration performance compared to similar DL-based approaches. This model was then used on MK, a tabular EHR dataset containing information of 16,953 patients that received RP at the Martiniklinik. It was shown that the developed model can improve discriminative and calibration performance regarding survival prediction compared to the CoxPH model that is usually used for this task. The most important factors for PCa relapse prediction were further analyzed in terms of feature encoding by developing univariate models to compare their predictive performance. It was shown that the more granular quantitative Gleason grading is beneficial for pathologic grading after RP, but does not improve predictive performance compared to Gleason or ISUP grading of biopsies. The pathological Gleason grading is known to be one of the most predictive factors for PCa relapse [125, 207, 220] which was also shown in this work. The feature importance analysis showed that Gleason Grading alone already nearly meets the predictive performance of a model where additional information of the patient is added. The second half of this thesis therefore took a deeper look into the Grading process from histopathological images themselves. Firstly, the CI network was developed that serves as a patch-level predictor for healthy vs. cancerous tissue that utilizes AI generated and human annotations of the prostate regarding cancerous and healthy tissue that was provided in the PANDA dataset [35]. It is used in the development of PCAI to preselect the most important patches of biopsy images. PCAI is DL-based risk prediction model for histopathological images. It was shown that the discriminative performance regarding PCa relapse prediction outperforms image-level ISUP grading for TMAs and biopsies. Further it was shown that the model performs equally well as quantitative GG, namely GIQ scores. It is worth noting that the PCAI model was developed on TMAs of a single center, namely the UKE, while predictions on external TMA datasets and biopsies were evaluated. The performance of PCAI can be credited at least partly to the robustness extensions that were introduced to ensure the model's performance on previously unseen datasets. Mainly domain adversarial training and sample credibility-guided color adaptation of the input images lead to a discriminative performance boost compared to a baseline model.

8.2 Research Questions

RQ1: How can DL models be utilized in survival prediction to generate better performance in terms of discrimination and calibration compared to classical approaches?

The developed DCS model from chapter 4 shows that DL-techniques can be utilized in the context of survival prediction for multiple tabular datasets. On the one hand, calibration performance could be improved among the DL-based algorithms, but did not reach the performance of the CoxPH model. On the other hand, discriminative performance that exceeded comparable DL models as well as the approach that is most frequently used, namely the CoxPH model, could be generated. This was shown with the aid of a modified loss part that boosts the number of comparisons based on the censoring rate of the dataset (see sec. 4.2.3). This idea of using EC comparisons along the EE comparisons was recently also analyzed regarding the C-index itself [3]. The ablation study in sec. 4.4.2 that analyzes the different loss parts of the model revealed, that even though the discrimination and calibration focus of different losses might lead

to conflicting optimization results, a combination of both can lead to better overall performance. This proofed especially helpful for datasets of this work with high censoring rates. Chapter 5 analyzed the tabular MK EHR dataset that showed that DCS outperforms CoxPH regarding all discriminative and calibration metrics. The additional performance boost in calibration compared to the aforementioned datasets can be contributed to proper weighting of discriminative and calibration loss during hyperparameter tuning (see sec. 5.2.5).

RQ2: Can these DL-based survival prediction models provide additional insights and a better understanding of feature importance?

To analyze this question, chapter 5 presented a detailed analysis of the MK dataset. Since DCS only allows non-linear dependencies that might involve multiple covariates, the contribution of individual factors is a more challenging task compared to the CoxPH model where only the linear, independent contribution of each covariate is measured by the hazard ratios. Consistent with the literature, this thesis showed that the PH assumption is violated for the given tabular datasets for 25 % to 79 % of features. As an example, the individual contribution of quantitative Gleason grades 3-5 is not meaningful since those features are codependent by design. This is why this work combined those features into semantically connected feature blocks. As an alternative way to measure feature importance, the cumulative contribution to discriminative performance of those feature blocks was analyzed in sec. 5.3.3, where iteratively added feature blocks with the highest discriminative score were added to produce a ranking.

Another step towards explainability was achieved by generating risk groups based on the predicted discrete survival curves as explained in sec. 5.3.4. Since the presented clustering approach is based on the predicted potentially crossing survival curves, more fine-grained sub-cohorts might be identified that show different behaviors regarding early and late relapse. For the MK data, a total of 7 statistically significantly distinct risk groups could be identified. Together with the feature distributions of those sub-cohorts, additional insights were generated that enable the analysis of the trajectory from a representative low risk patient to a high risk individual.

However, those analyses only provide different angles for a task where the interaction between covariates might not be as easily explainable as it is often done using CoxPH models. It remains a challenging task to adequately represent the complex feature interactions such that they can be better understood by physicians and patients. Other methods exist that analyze the local linear influence around individual predictions to estimate feature importance [132]. Another approach that might be worth exploring further lies in allowing more complex feature interactions as demonstrated in [4] that can also be extended to include time-dependency [134].

RQ3: What patient features and corresponding representations are most relevant in PCa relapse prediction after RP in the given tabular data?

The univariate feature encoding analysis in sec. 5.3.1 revealed that for CoxPH and DCS, the encoding of input features played an important role regarding discriminative performance of the model. Firstly, the PSA level performed better when used without taking the prostate volume into account for PSA density. Further, Gleason scores that were obtained from biopsies did not gain predictive power when more fine-grained, quantitative information was added for the individuals. Primary and secondary GG as well as ISUP provided the best possible feature representation. For pathological GG that were obtained after RP, it could be shown that the more complex quantitative information significantly improves the predictive performance regarding relapse prediction by approximately 6 pp for the CoxPH and the DCS model. Individual analysis of binary variables showed that capsular invasion yields the best results followed by seminal vesicle invasion. The trajectory of the most important features from the obtained lowest to highest identified risk group further reveals that LNI is almost exclusively present in the worst three identified risk groups while other factors like capsular invasion can already be found in the second-lowest risk group.

RQ4: How does the knowledge derived from morphological properties of RP TMA spot images from the UKE translate to other external centers? What adaptations can be applied to improve model generalizability?

Findings of the unique and high quality dataset with 69,251 TMAs from 17,700 patients can be translated to a digital risk biomarker that is also applicable for external datasets if model generalizability and robustness as introduced in sec. 7.2.5 are taken into account. The extensions made the model more robust for out-of-training data and boosted discriminative performance for external TMA datasets and also proved beneficial for the internal datasets (sec. 7.3.1). The main contributions to boost model robustness are provided by a domain adversarial training method with the aim to merge samples of different acquisition protocols in the bag-level latent representation, and a credibility-guided color adaptation approach for samples that are far from the training distribution. These extensions were necessary since the model was only trained on TMAs of a single center, namely the UKE. Without taking model robustness into account, it was shown that the discriminative performance does not translate even to other internal TMA datasets with differences in the acquisition protocol. The resulting model is even able to perform at least as well as the SotA method of GIQ when image-level annotations are compared as shown in sec. 7.3.2.

RQ5: Do those findings of a digital risk biomarker for TMAs translate to biopsy images?

With the given measures for model robustness, the learned morphological features that were picked up by the model also translate to biopsy images. With the guidance to cancerous regions of biopsies that are preselected by the CI model that was presented in chapter 6, the resulting pipeline outperforms expert ISUP grading for image-based annotations on two biopsy datasets.

RQ6: How can the complex findings regarding tabular and image-based risk estimations be presented in an interpretable way as an additional step towards clinical applicability?

After proving the discriminative performance of the tabular risk prediction model DCS as well as the cancer severity biomarker PCAI, both predictions were used to generate additional insights regarding risk stratification by grouping the adjacent risk predictions. For the tabular model, it could then be explored which features contribute to what extent for the different risk groups.

RQ7: How can model uncertainty be quantified and used to boost the model's trustworthiness in a clinical setting?

The complex feature interdependency for the DCS model turns feature importance into a challenging task. This work tried to assign feature importance based on discriminative performance or additional information regarding relapse prediction with the feature block importance algorithm. To enhance the trustworthiness in the application, further steps need to be taken to improve interpretability of the risk prediction network. The same can be said for the image-based relapse risk prediction network PCAI. While the CI generates heatmaps that are used for the predictions of PCAI, further analysis needs to be performed. One further approach would be to find representative types of patches in the latent space of the model that could be ordered by cancer severity. These patches could further be analyzed in collaboration with a pathologist to derive biological attributes of the identified subcategories of patches or tissue.

8.3 Future Work

This thesis provides a first step towards automated risk assessment of PCa patients regarding tabular parameters as well as risk prediction directly on images. A DL-based survival prediction model was developed that demonstrates the predictive power of the approach in the field of survival prediction in general and specifically for PCa datasets. Further, a proof-of-concept for biopsy cancer risk prediction model was presented that could be used in a clinical setting

as a part of a clinical decision support system with the main aim to help the urologist in the initial treatment decision for an individual patient. While individual discussions and outlooks of the developed approaches were presented in sec. 4.5, sec. 5.4, sec. 6.5 and sec. 7.5, this section discusses the next research directions from a broader point of view.

Towards Clinical Applicability

This thesis focused on developed survival and risk prediction models that proved successful in the context of PCa patients. However, to integrate such a CDSS into clinical practice, additional steps need to be taken. A first application could be made for the initial treatment decision for PCa patient based on automatic PCAI-based biopsy grading that may be combined with a DCS-based risk prediction network for cancer severity from information that is available at point of biopsy. Such a system would require external evaluation which is crucial for the success of such a system. While PCAI already proved abilities towards generalizability, the same approach needs to be taken for DCS that was trained only on the internal MK dataset. Furthermore, model explanations like cancer heatmaps that were provided by the CI model play an important role in the acceptance of such a system [159]. Additional steps need to be taken that provide better explanations of the complex risk prediction that is performed by the models.

Survival Prediction Model Equivalently to the PCAI model, it would be beneficial if a survival prediction model provides estimation certainty along with the prediction. While a latent-based credibility approach as presented in sec. 7.2.5 would work to identify input samples that were unseen during model training, another approach could be a Bayesian survival model which includes uncertainty prediction in the model itself [18, 111, 179]. Similar to the credibility estimation, such an additional output of the network would lead to more trustworthiness in the model's prediction while samples that could not be predicted with a minimum certainty could be discarded or analyzed in collaboration with a urologist to identify conflicting documentations within a EHR.

Furthermore, the developed model could be improved in terms of model interpretability. Several tools exist that can be applied to any survival prediction model. Some examples are Shapley value-based [213] approaches like `SurvShap` [4], a time-dependent extension called `SurvSHAP(t)` [134], or `SurvLIME-Inf` [235] might provide additional details to the model predictions or on the feature importance by analyzing the local neighborhood of an individual prediction.

Individual Treatment Estimation A possible step for further development would be the integration of competing risks as demonstrated in [140, 141]. This would allow the model to integrate multiple possible event types in the predicted survival curves. A related topic would be the exploration of treatment options within the survival prediction model as additional parameters to explore the effectiveness for the individual. This treatment recommendation would enable the exploration of treatment alternatives and relate them with success for each individual patient as shown in [126, 142, 171], leading to better clinical decision support. Developing such a system generates an additional level of complexity that only one treatment per individual can be performed and documented and thus utilized for subsequent analysis. While some approaches analyze treatment effectiveness in terms of side effects [229], others try to emulate human decision itself [97].

None of the aforementioned methods would allow a reliable comparison of treatment effectiveness for an individual. To achieve this, different treatment effects are emulated for the same individual while keeping the selection bias of the treatment options into account. Developing such a system may involve statistical methods like propensity score matching [12, 109, 115, 166] or counterfactual inference approaches [41]. As an alternative, the effect of treatment information can also be effectively ignored or unlearned in the data representation as shown in [118, 260]. One of these approaches [28] utilizes treatment information adversarially to generate treatment invariant latent representations for each individual similarly to the DA approach that was used to build better

robustness in the PCAI model that was presented in this thesis. Further, exemplary publications in the context of PCa exist for example by comparing radical prostatectomy to external beam radiation therapy approaches in terms of patient survival [72] or a comparison of treatment options for benign prostatic hypertrophy in terms of surgical outcome [14].

Multimodal Risk Prediction Moreover, the aforementioned systems could also take the image-based risk predictions as presented in chapter 7 into account. This objective evaluation of images might include information that is relevant for diagnosis or treatment decision especially for biopsy images. Even though intermediate results like risk predictions on those individual images do not yield treatment recommendations, they would help to build trust in the complex overall model structure by providing interpretable risks. This step away from the black box nature of end-to-end AI models can boost the trustworthiness of such a system.

Lastly, additional modalities could also be included in such a risk prediction model. A candidate for biopsy diagnosis and treatment suggestion would be genetic profiling. Genetic markers show predictive capabilities in the context of PCa [48, 56, 139]. These genetic biomarkers might contain useful information that extend visible information in images and the tabular data that is present for an individual at risk of PCa that can be integrated using DL-based approaches as shown for example in [250].

Bibliography

[1] Agarwal, S., Abaker, M.E.O., Daescu, O.: Survival prediction based on histopathology imaging and clinical data: A novel, whole slide cnn approach. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12905 LNCS, 762–771 (2021), https://link.springer.com/chapter/10.1007/978-3-030-87240-3_73

[2] Ahmed, H.M., Ebeed, A.E., Hamdy, A., El-Ghar, M.A., Razek, A.A.K.A.: Interobserver agreement of prostate imaging–reporting and data system (pi-rads–v2). Egyptian Journal of Radiology and Nuclear Medicine 52, 1–8 (2021)

[3] Alabdallah, A., Ohlsson, M., Pashami, S., Rögnvaldsson, T.: The concordance index decomposition: A measure for a deeper understanding of survival prediction models. Artificial Intelligence in Medicine 148, 102781 (2 2024), https://linkinghub.elsevier.com/retrieve/pii/S093336572400023X

[4] Alabdallah, A., Pashami, S., Rognvaldsson, T., Ohlsson, M.: Survshap: A proxy-based algorithm for explaining survival models with shap. Proceedings - 2022 IEEE 9th International Conference on Data Science and Advanced Analytics, DSAA 2022 (2022)

[5] Amgad, M., Hodge, J.M., Elsebaie, M.A., Bodelon, C., Puvanesarajah, S., Gutman, D.A., Siziopikou, K.P., Goldstein, J.A., Gaudet, M.M., Teras, L.R., Cooper, L.A.: A population-level digital histologic biomarker for enhanced prognosis of invasive breast cancer. Nature Medicine 2023 pp. 1–13 (11 2023), https://www.nature.com/articles/s41591-023-02643-7

[6] Andriole, G.L., Crawford, E.D., Grubb, R.L., Buys, S.S., Chia, D., Church, T.R., Fouad, M.N., Isaacs, C., Kvale, P.A., Reding, D.J., Weissfeld, J.L., Yokochi, L.A., O'Brien, B., Ragard, L.R., Clapp, J.D., Rathmell, J.M., Riley, T.L., Hsing, A.W., Izmirlian, G., Pinsky, P.F., Kramer, B.S., Miller, A.B., Gohagan, J.K., Prorok, P.C.: Prostate cancer screening in the randomized prostate, lung, colorectal, and ovarian cancer screening trial: Mortality results after 13 years of follow-up. JNCI: Journal of the National Cancer Institute 104, 125–132 (1 2012), https://dx.doi.org/10.1093/jnci/djr500

[7] Antolini, L., Boracchi, P., Biganzoli, E.: A time-dependent discrimination index for survival data. Statistics in Medicine 24, 3927–3944 (12 2005), http://doi.wiley.com/10.1002/sim.2427

[8] Arafa, M.A., Farhat, K.H., Khan, F.K., Rabah, D.M., Elmorshedy, H., Mokhtar, A., Al-Taweel, W.: Development and internal validation of a nomogram predicting significant prostate cancer: Is it clinically applicable in low prevalent prostate cancer countries? a multicenter study. The Prostate 84(1), 56–63 (2024)

[9] Arrieta, A.B., Díaz-Rodríguez, N., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion 58, 82–115 (6 2020)

[10] Arvaniti, E., Fricker, K.S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P.J., Rueschoff, J.H., Claassen, M.: Automated gleason grading of prostate cancer tissue microarrays via deep learning. Scientific reports 8(1), 12054 (2018)

[11] Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., Breen, J., Ravikumar, N., Chung, Y., Park, J., Nateghi, R., Pourakpour, F., Fick, R.H., Hadj, S.B., Jahanifar, M., Shephard, A., Dexl, J., Wittenberg, T., Kondo, S., Lafarge, M.W., Koelzer, V.H., Liang, J., Wang, Y., Long, X., Liu, J., Razavi, S., Khademi, A., Yang, S., Wang, X., Erber, R., Klang, A., Lipnik, K., Bolfa, P., Dark, M.J., Wasinger, G., Veta, M., Breininger, K.: Mitosis domain generalization in histopathology images — the midog challenge. Medical Image Analysis 84, 102699 (2 2023)

[12] Austin, P.C.: An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research 46(3), 399–424 (2011)

[13] Austin, P.C., Tu, J.V., Ho, J.E., Levy, D., Lee, D.S.: Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. Journal of Clinical Epidemiology 66, 398–407 (4 2013), /pmc/articles/PMC4322906/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322906/

[14] Ayoub, C.H., Haber, R., Amine, R., Mikati, D., Mahfoud, Z.R., Hajj, A.E.: Comparison of postoperative outcomes of trans-urethral resection of the prostate, laser vaporization, and laser enucleation: A double propensity score matched analysis. Urology 177, 148–155 (7 2023)

[15] Balk, S.P., Ko, Y.J., Bubley, G.J.: Biology of prostate-specific antigen. Journal of Clinical Oncology 21, 383–391 (1 2003), http://ascopubs.org/doi/10.1200/JCO.2003.02.083

[16] Ballard, D.H.: Modular learning in neural networks. In: Proceedings of the sixth National conference on Artificial intelligence-Volume 1. pp. 279–284 (1987)

[17] Bankhead, P., Loughrey, M.B., Fernández, J.A., Dombrowski, Y., McArt, D.G., Dunne, P.D., McQuaid, S., Gray, R.T., Murray, L.J., Coleman, H.G., James, J.A., Salto-Tellez, M., Hamilton, P.W.: Qupath: Open source software for digital pathology image analysis. Scientific Reports 2017 7:1 7, 1–7 (12 2017), https://www.nature.com/articles/s41598-017-17204-5

[18] Bartoš, F., Aust, F., Haaf, J.M.: Informed bayesian survival analysis. BMC Medical Research Methodology 2022 22:1 22, 1–22 (9 2022), https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01676-9

[19] Bayramoglu, N., Heikkilä, J.: Transfer learning for cell nuclei classification in histopathology images. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14. pp. 532–539. Springer (2016)

[20] Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware lstm networks. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 65–74 (2017)

[21] Beckmann, K., Garmo, H., Lissbrant, I.F., Stattin, P.: The value of real-world data in understanding prostate cancer risk and improving clinical care: Examples from swedish registries. Cancers 2021, Vol. 13, Page 875 13, 875 (2 2021), https://www.mdpi.com/2072-6694/13/4/875/htmhttps://www.mdpi.com/2072-6694/13/4/875

[22] Beerlage, H.P.: Alternative therapies for localized prostate cancer. Current Urology Reports 4(3), 216–220 (2003)

[23] Bell, K.J., Mar, C.D., Wright, G., Dickinson, J., Glasziou, P.: Prevalence of incidental prostate cancer: A systematic review of autopsy studies. International Journal of Cancer 137, 1749–1757 (10 2015), https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.29538https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.29538https://onlinelibrary.wiley.com/doi/10.1002/ijc.29538

[24] Belle, V.V., Pelckmans, K., Suykens, J.A., Huffel, S.V.: Additive survival least-squares support vector machines. Statistics in Medicine 29, 296–308 (1 2010), https://onlinelibrary.wiley.com/doi/full/10.1002/sim.3743https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3743https://onlinelibrary.wiley.com/doi/10.1002/sim.3743

[25] Benecchi, L., Pieri, A.M., Pastizzaro, C.D., Potenzoni, M.: Optimal measure of psa kinetics to identify prostate cancer. Urology 71, 390–394 (3 2008)

[26] Bennis, A., Mouysset, S., Serrurier, M.: Dpwte: A deep learning approach to survival analysis using a parsimonious mixture of weibull distributions. In: Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30. pp. 185–196. Springer (2021)

[27] van den Bergh, R.C., Roemeling, S., Roobol, M.J., Roobol, W., Schröder, F.H., Bangma, C.H.: Prospective validation of active surveillance in prostate cancer: The prias study. European Urology 52, 1560–1563 (12 2007)

[28] Bica, I., Alaa, A.M., Jordon, J., van der Schaar, M.: Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In: International Conference on Learning Representations (2019)

[29] Blanche, P., Kattan, M.W., Gerds, T.A.: The c-index is not proper for the evaluation of t-year predicted risks. Biostatistics 20, 347–357 (2019), https://academic.oup.com/biostatistics/article/20/2/347/4864363

[30] Brehler, M., Walhagen, P., Busch, C., Bonn, S., Bengtsson, E.: Difficulties and recommendations for ai-based prediction of prostate cancer aggressiveness in digital pathology. Medical Research Archives 11 (12 2023), https://esmed.org/MRA/mra/article/view/4586

[31] Breiman, L.: Random forests. Machine Learning 45, 5–32 (10 2001), https://link.springer.com/article/10.1023/A:1010933404324

[32] Breslow, N.E.: Analysis of survival data under the proportional hazards model. International Statistical Review / Revue Internationale de Statistique 43, 45 (4 1975)

[33] Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthly weather review 78(1), 1–3 (1950)

[34] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. Advances in Neural Information Processing Systems 2020-December (5 2020), https://arxiv.org/abs/2005.14165v4

[35] Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., van Boven, H., Vink, R., van de Kaa, C.H., van der Laak, J., Amin, M.B., Evans, A.J., van der Kwast, T., Allan, R., Humphrey, P.A., Grönberg, H., Samaratunga, H., Delahunt, B., Tsuzuki, T., Häkkinen, T., Egevad, L., Demkin, M., Dane, S., Tan, F., Valkonen, M., Corrado, G.S., Peng, L., Mermel, C.H., Ruusuvuori, P., Litjens, G., Eklund, M., Brilhante, A., Çakır, A., Farré, X., Geronatsiou, K., Molinié, V., Pereira, G., Roy, P., Saile, G., Salles, P.G., Schaafsma, E., Tschui, J., Billoch-Lima, J., Pereira, E.M., Zhou, M., He, S., Song, S., Sun, Q., Yoshihara, H., Yamaguchi, T., Ono, K., Shen, T., Ji, J., Roussel, A., Zhou, K., Chai, T., Weng, N., Grechka, D., Shugaev, M.V., Kiminya, R., Kovalev, V., Voynov, D., Malyshev, V., Lapo, E., Campos, M., Ota, N., Yamaoka, S., Fujimoto, Y., Yoshioka, K., Juvonen, J., Tukiainen, M., Karlsson, A., Guo, R., Hsieh,

C.L., Zubarev, I., Bukhar, H.S., Li, W., Li, J., Speier, W., Arnold, C., Kim, K., Bae, B., Kim, Y.W., Lee, H.S., Park, J.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature Medicine 2022 28:1 28, 154–163 (1 2022), https://www.nature.com/articles/s41591-021-01620-2

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

[37] Buyyounouski, M.K., Choyke, P.L., McKenney, J.K., Sartor, O., Sandler, H.M., Amin, M.B., Kattan, M.W., Lin, D.W.: Prostate cancer – major changes in the american joint committee on cancer eighth edition cancer staging manual. CA: A Cancer Journal for Clinicians 67, 245–253 (5 2017), https://onlinelibrary.wiley.com/doi/full/10.3322/caac.21391https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21391https://acsjournals.onlinelibrary.wiley.com/doi/10.3322/caac.21391

[38] Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Silva, V.W.K., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine 2019 25:8 25, 1301–1309 (7 2019), https://www.nature.com/articles/s41591-019-0508-1

[39] Cetin, B., Ozet, A.: The potential for chemotherapy-free strategies in advanced prostate cancer. Current Urology 13(2), 57–63 (2019)

[40] Chai, J., Zeng, H., Li, A., Ngai, E.W.: Deep learning in computer vision: A critical review of emerging techniques and application scenarios. Machine Learning with Applications 6, 100134 (2021)

[41] Chapfuwa, P., Assaad, S., Zeng, S., Pencina, M.J., Carin, L., Henao, R.: Enabling counterfactual survival analysis with balanced representations. In: Proceedings of the Conference on Health, Inference, and Learning. pp. 133–145 (2021)

[42] Chen, N., Zhou, Q.: The evolving gleason grading system. Chinese Journal of Cancer Research 28, 58 (2 2016), /pmc/articles/PMC4779758//pmc/articles/PMC4779758/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779758/

[43] Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)

[44] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August-2016, 785–794 (8 2016), https://dl.acm.org/doi/10.1145/2939672.2939785

[45] Chlipala, E.A., Butters, M., Brous, M., Fortin, J.S., Archuletta, R., Copeland, K., Bolon, B.: Impact of preanalytical factors during histology processing on section suitability for digital image analysis. Toxicologic Pathology 49, 755–772 (6 2021), https://journals.sagepub.com/doi/full/10.1177/0192623320970534

[46] Cho, K., Merriënboer, B.V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference pp. 1724–1734 (6 2014), https://arxiv.org/abs/1406.1078v3

[47] Choi, E., Bahadori, M.T., Kulas, J.A., Schuetz, A., Stewart, W.F., Sun, J., Health, S.: Retain: An interpretable predictive model for healthcare using reverse time attention mechanism (2016)

[48] Choudhury, A.D., Eeles, R., Freedland, S.J., Isaacs, W.B., Pomerantz, M.M., Schalken, J.A., Tammela, T.L., Visakorpi, T.: The role of genetic markers in the management of prostate cancer. European Urology 62, 577–587 (10 2012)

[49] Chun, F.K.H., Steuber, T., Erbersdobler, A., Currlin, E., Walz, J., Schlomm, T., Haese, A., Heinzer, H., McCormack, M., Huland, H., et al.: Development and internal validation of a nomogram predicting the probability of prostate cancer gleason sum upgrading between biopsy and radical prostatectomy pathology. European urology 49(5), 820–826 (2006)

[50] Ciga, O., Xu, T., Martel, A.L.: Self supervised contrastive learning for digital histopathology. Machine Learning with Applications 7, 100198 (2022), https://www.sciencedirect.com/science/article/pii/S2666827021000992

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

[52] Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin 70, 213–220 (10 1968)

[53] Cooke, R.: Cox proportional hazard model. Encyclopedia of Statistics in Quality and Reliability (1 2008), https://www.academia.edu/10604864/Cox_Proportional_Hazard_Model

[54] Cox, D.R.: Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological) 34(2), 187–202 (1972)

[55] Crawford, E.D.: Epidemiology of prostate cancer. Urology 62, 3–12 (12 2003)

[56] Cucchiara, V., Cooperberg, M.R., Dall'Era, M., Lin, D.W., Montorsi, F., Schalken, J.A., Evans, C.P.: Genomic markers in prostate cancer decision making. European Urology 73, 572–582 (4 2018)

[57] Cui, M., Zhang, D.Y.: Artificial intelligence and computational pathology. Laboratory Investigation 2021 101:4 101, 412–422 (1 2021), https://www.nature.com/articles/s41374-020-00514-0

[58] Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Aparicio, S., Brenton, J.D., Ellis, I., Huntsman, D., Pinder, S., Murphy, L., Bardwell, H., Ding, Z., Jones, L., Liu, B., Papatheodorou, I., Sammut, S.J., Wishart, G., Chia, S., Gelmon, K., Speers, C., Watson, P., Blamey, R., Green, A., MacMillan, D., Rakha, E., Gillett, C., Grigoriadis, A., De Rinaldis, E., Tutt, A., Parisien, M., Troup, S., Chan, D., Fielding, C., Maia, A.T., McGuire, S., Osborne, M., Sayalero, S.M., Spiteri, I., Hadfield, J., Bell, L., Chow, K., Gale, N., Kovalik, M., Ng, Y., Prentice, L., Tavaré, S., Markowetz, F., Langerød, A., Provenzano, E., Purushotham, A., Børresen-Dale, A.L., Caldas, C.: The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012 486:7403 486(7403), 346–352 (apr 2012), https://www.nature.com/articles/nature10983

[59] Datta, M.W., Kajdacsy-Balla, A.A.: Tissue microarrays in prostate cancer research. Reviews in Urology 6, 44 (11 2004), /pmc/articles/PMC1472679/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1472679/

[60] Davidson-Pilon, C.: lifelines: survival analysis in python. Journal of Open Source Software 4(40), 1317 (2019), https://doi.org/10.21105/joss.01317

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

[62] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1, 4171–4186 (10 2018), http://arxiv.org/abs/1810.04805

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

[64] Dietrich, E.: Deep learning-based discrete-time survival prediction on prostate cancer histopathology images. Ph.D. thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky (2022)

[65] Dietrich, E., Fuhlert, P., Ernst, A., Sauter, G., Lennartz, M., Stiehl, H.S., Zimmermann, M., Bonn, S.: Towards Explainable End-to-End Prostate Cancer Relapse Prediction from H&E Images Combining Self-Attention Multiple Instance Learning with a Recurrent Neural Network. Proceedings of Machine Learning Research pp. 1–16 (nov 2021), https://ml4health.github.io/2021/poster_A1.html

[66] Dispenzieri, A., Katzmann, J.A., Kyle, R.A., Larson, D.R., Therneau, T.M., Colby, C.L., Clark, R.J., Mead, G.P., Kumar, S., Melton, L.J., Rajkumar, S.V.: Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. Mayo Clinic Proceedings 87, 517–523 (6 2012)

[67] Doshi-Velez, F., Ge, Y., Kohane, I.: Comorbidity clusters in autism spectrum disorders: An electronic health record time-series analysis. Pediatrics 133 (2014), https://pubmed.ncbi.nlm.nih.gov/24323995/

[68] Dyba, T., Randi, G., Bray, F., Martos, C., Giusti, F., Nicholson, N., Gavin, A., Flego, M., Neamtiu, L., Dimitrova, N., Carvalho, R.N., Ferlay, J., Bettio, M.: The european cancer burden in 2020: Incidence and mortality estimates for 40 countries and 25 major cancers. European Journal of Cancer 157, 308 (11 2021), /pmc/articles/PMC8568058//pmc/articles/PMC8568058/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8568058/

[69] Eastham, J.A., Scardino, P.T., Kattan, M.W.: Predicting an optimal outcome after radical prostatectomy: The "trifecta" nomogram. J Urol 179, 2207–2211 (2008)

[70] Egevad, L., Mazzucchelli, R., Montironi, R.: Implications of the international society of urological pathology modified gleason grading system. Archives of Pathology & Laboratory Medicine 136, 426–434 (4 2012), https://dx.doi.org/10.5858/arpa.2011-0495-RA

[71] Egevad, L., Swanberg, D., Delahunt, B., Ström, P., Kartasalo, K., Olsson, H., Berney, D.M., Bostwick, D.G., Evans, A.J., Humphrey, P.A., et al.: Identification of areas of grading difficulties in prostate cancer and comparison with artificial intelligence assisted grading. Virchows Archiv 477, 777–786 (2020)

[72] Ennis, R.D., Hu, L., Ryemon, S.N., Lin, J., Mazumdar, M.: Brachytherapy-based radiotherapy and radical prostatectomy are associated with similar survival in high-risk localized prostate cancer. Journal of Clinical Oncology 36(12), 1192–1198 (2018)

[73] Epstein, J.I.: An update of the gleason grading system. The Journal of Urology 183, 433–440 (2 2010)

[74] Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., Srigley, J.R., Humphrey, P.A., Al-Hussain, T., Algaba, F., Aron, M., Berman, D., Berney, D., Brimo, F., Cao, D., Cheville, J., Clouston, D., Colecchia, M., Comperat, E., Cunha, I.W.D., Marzo, A.D., Ertoy, D., Fine, S., Foster, C., Grignon, D., Gupta, N., Gupta, R., Kench, J., Kristiansen, G., Kunju, L., Leite, K.R.M., Loda, M., Lopez-Beltran, A., Lotan, T., Lucia, M.S., Magi-Galluzzi, C.,

Montironi, R., McKenney, J., Merrimen, J., Netto, G., Orozco, R., Paner, G., Parwani, A., Pizov, G., Reuter, V., Ro, J., Samaratunga, H., Schultz, L., Shanks, J., Sesterhenn, I., Shen, S., Simko, J., Suzigan, S., Suryavanshi, M., Tan, P.H., Takahashi, H., Tomlins, S., Trpkov, K., Troncoso, P., True, L., Tsuzuki, T., Kwast, T.V.D., Varma, M., Warren, A., Wheeler, T., Yang, X., Zhou, M., Kantoff, P., Eisenberger, M., Stadler, W., Andriole, G., Klein, E., Benson, M., Montorsi, F., Crawford, D., Loeb, S., Catto, J., Schaeffer, E., Nacey, J.N., DeWeese, T., Sandler, H., Zietman, A., Pollack, A., Rodrigues, G.: The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma definition of grading patterns and proposal for a new grading system. American Journal of Surgical Pathology 40, 244–252 (2016), https://journals.lww.com/ajsp/fulltext/2016/02000/the_2014_international_society_of_urological.10.aspx

[75] Esteban, C., Staeck, O., Yang, Y., Tresp, V.: Predicting clinical events by combining static and dynamic information using recurrent neural networks. Proceedings - 2016 IEEE International Conference on Healthcare Informatics, ICHI 2016 pp. 93–101 (2 2016), http://arxiv.org/abs/1602.02685

[76] Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. NPJ digital medicine 4(1), 5 (2021)

[77] Etzioni, R., Cha, R., Feuer, E.J., Davidov, O.: Asymptomatic incidence and duration of prostate cancer. American Journal of Epidemiology 148, 775–785 (10 1998), https://dx.doi.org/10.1093/oxfordjournals.aje.a009698

[78] Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in Medicine 14(1), 73–82 (jan 1995), http://doi.wiley.com/10.1002/sim.4780140108

[79] Fetisov, N., Hall, L., Goldgof, D., Schabath, M.: Unsupervised prostate cancer histopathology image segmentation via meta-learning. Proceedings - IEEE Symposium on Computer-Based Medical Systems 2023-June, 838–844 (2023)

[80] Flach, R.N., Willemse, P.P.M., Suelmann, B.B., Deckers, I.A., Jonges, T.N., van Dooijeweert, C., van Diest, P.J., Meijer, R.P.: Significant inter-and intralaboratory variation in gleason grading of prostate cancer: A nationwide study of 35,258 patients in the netherlands. Cancers 13, 5378 (11 2021), https://www.mdpi.com/2072-6694/13/21/5378/htmhttps://www.mdpi.com/2072-6694/13/21/5378

[81] Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619 (10 1973), https://journals.sagepub.com/doi/10.1177/001316447303300309

[82] Fleming, C., Wasson, J.H., Albertsen, P.C., Barry, M.J., Wennberg, J.E., Bubolz, T., Lindsay, C.C., Littenberg, B., Flood, A.B., Lu-Yao, G.L., et al.: A decision analysis of alternative treatment strategies for clinically localized prostate cancer. Jama 269(20), 2650–2658 (1993)

[83] Folorunso, S.O., Awotunde, J.B., Rangaiah, Y.P., Ogundokun, R.O.: Efficientnets transfer learning strategies for histopathological breast cancer image analysis. https://doi.org/10.1142/S1793962324410095 (5 2023), https://www.worldscientific.com/worldscinet/ijmssc

[84] Fuhlert, P., Ernst, A., Dietrich, E., Westhaeusser, F., Kloiber, K., Bonn, S.: Deep Learning-Based Discrete Calibrated Survival Prediction. In: ICDH 2022. pp. 1–6 (2022)

[85] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of machine learning research 17(59), 1–35 (2016)

[86] George, B., Seals, S., Aban, I.: Survival analysis and regression models. Journal of nuclear cardiology 21(4), 686–694 (2014)

[87] Gerds, T.A., Kattan, M.W.: Medical Risk Prediction : With Ties to Machine Learning. Medical Risk Prediction (jan 2021), https://www.taylorfrancis.com/books/mono/10.1201/9781138384484/medical-risk-prediction-thomas-gerds-michael-kattan

[88] Giunchiglia, E., Nemchenko, A., van der Schaar, M.: Rnn-surv: A deep recurrent model for survival analysis. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27. pp. 23–32. Springer (2018)

[89] Gleason, D.F., Mellinger, G.T.: Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. The Journal of urology 111(1), 58–64 (1974)

[90] Gohagan, J.K., Prorok, P.C., Hayes, R.B., Kramer, B.S.: The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. Controlled clinical trials 21 (2000), https://pubmed.ncbi.nlm.nih.gov/11189683/

[91] Goldstein, M., Han, X., Puli, A., Perotte, A.J., Ranganath, R.: X-cal: Explicit calibration for survival analysis (2021), https://github.com/rajesh-lab/X-CAL

[92] Greenwood, M.: A Report on the Natural Duration of Cancer. London: H.M.S.O. (1927)

[93] Guan, H., Liu, M.: Domainatm: Domain adaptation toolbox for medical data analysis. NeuroImage 268, 119863 (3 2023)

[94] Gupta, R.T., Mehta, K.A., Turkbey, B., Verma, S.: Pi-rads: Past, present, and future. Journal of Magnetic Resonance Imaging 52(1), 33–53 (2020)

[95] Haider, H., Hoehn, B., Davis, S., Greiner, R.: Effective ways to build and evaluate individual survival distributions (2018)

[96] Haider, H., Hoehn, B., Davis, S., Greiner, R.: Effective ways to build and evaluate individual survival distributions. Journal of Machine Learning Research 21(85), 1–63 (2020)

[97] Han, J.H., Lee, S., Lee, B., kee Baek, O., Washington, S.L., Herlemann, A., Lonergan, P.E., Carroll, P.R., Jeong, C.W., Cooperberg, M.R.: Explainable ml models for a deeper insight on treatment decision for localized prostate cancer. Scientific Reports 2023 13:1 13, 1–8 (7 2023), https://www.nature.com/articles/s41598-023-38162-1

[98] Harnden, P., Shelley, M.D., Coles, B., Staffurth, J., Mason, M.D.: Should the gleason grading system for prostate cancer be modified to account for high-grade tertiary components? a systematic review and meta-analysis. Lancet Oncology 8, 411–419 (5 2007), http://www.thelancet.com/article/S1470204507701365/fulltexthttp://www.thelancet.com/article/S1470204507701365/abstracthttps://www.thelancet.com/journals/lanonc/article/PIIS1470-2045(07)70136-5/abstract

[99] Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L., Rosati, R.A.: Evaluating the yield of medical tests (5 1982), https://jamanetwork.com/journals/jama/fullarticle/372568https://jamanetwork.com/

[100] Hartman, N., Kim, S., He, K., Kalbfleisch, J.D.: Pitfalls of the concordance index for survival outcomes. Statistics in Medicine 42, 2179–2190 (6 2023), https://onlinelibrary.wiley.com/doi/full/10.1002/sim.9717https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.9717https://onlinelibrary.wiley.com/doi/10.1002/sim.9717

[101] Hashimoto, T., Ohori, M., Shimodaira, K., Kaburaki, N., Hirasawa, Y., Satake, N., Gondo, T., Nakagami, Y., Namiki, K., Ohno, Y.: Prostate-specific antigen screening impacts on biochemical recurrence in patients with clinically localized prostate cancer. International Journal of Urology 25, 561–567 (6 2018)

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

[103] He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., Shen, D.: Transformers in medical image analysis. Intelligent Medicine 3(1), 59–78 (2023)

[104] Head, T., Kumar, M., Nahrstaedt, H., Louppe, G., Shcherbatyi, I.: scikit-optimize/scikit-optimize (Oct 2021), https://doi.org/10.5281/zenodo.5565057

[105] Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. 8th International Conference on Learning Representations, ICLR 2020 (12 2019), https://arxiv.org/abs/1912.02781v2

[106] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9, 1735–1780 (11 1997), http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735

[107] Howard, A., Chiu, A., McDonald, M., MSLA, Kan, W., Yianchen: Wsdm - kkbox's churn prediction challenge (2017), https://kaggle.com/competitions/kkbox-churn-prediction-challenge

[108] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks (2018), http://image-net.org/challenges/LSVRC/2017/results

[109] Hu, L., Ji, J., Ennis, R.D., Hogan, J.W.: A flexible approach for causal inference with multiple treatments and clustered survival outcomes. Statistics in medicine 41(25), 4982–4999 (2022)

[110] Huang, Y., Li, W., Macheret, F., Gabriel, R.A., Ohno-Machado, L.: A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association 27, 621–633 (4 2020), https://dx.doi.org/10.1093/jamia/ocz228

[111] Ibrahim, J.G., Chen, M.H., Sinha, D., Ibrahim, J., Chen, M.: Bayesian survival analysis, vol. 2. Springer (2001)

[112] Ikromjanov, K., Bhattacharjee, S., Hwang, Y.B., Sumon, R.I., Kim, H.C., Choi, H.K.: Whole slide image analysis and detection of prostate cancer using vision transformers. In: 2022 international conference on artificial intelligence in information and communication (ICAIIC). pp. 399–402. IEEE (2022)

[113] Ikromjanov, K., Bhattacharjee, S., Sumon, R.I., Hwang, Y.B., Rahman, H., Lee, M.J., Kim, H.C., Park, E., Cho, N.H., Choi, H.K.: Region segmentation of whole-slide images for analyzing histological differentiation of prostate adenocarcinoma using ensemble efficientnetb2 u-net with transfer learning mechanism. Cancers 2023, Vol. 15, Page 762 15, 762 (1 2023), https://www.mdpi.com/2072-6694/15/3/762/html

[114] Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. 35th International Conference on Machine Learning, ICML 2018 5, 3376–3391 (2 2018), https://arxiv.org/abs/1802.04712v4

[115] Imai, K., Ratkovic, M.: Covariate balancing propensity score. Journal of the Royal Statistical Society Series B: Statistical Methodology 76(1), 243–263 (2014)

[116] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The Annals of Applied Statistics 2, 841–860 (9 2008), https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-3/Random-survival-forests/10.1214/08-AOAS169.full

[117] Jawhar, N.M.: Tissue microarray: A rapidly evolving diagnostic and research tool. Annals of Saudi Medicine 29, 123 (2009), /pmc/articles/PMC2813639//pmc/articles/PMC2813639/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813639/

[118] Johansson, F., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. In: International conference on machine learning. pp. 3020–3029. PMLR (2016)

[119] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific Data 2016 3:1 3, 1–9 (5 2016), https://www.nature.com/articles/sdata201635

[120] Kaczmarzyk, J.R., Gupta, R., Kurc, T.M., Abousamra, S., Saltz, J.H., Koo, P.K.: Champkit: A framework for rapid evaluation of deep neural networks for patch-based histopathology classification. Computer Methods and Programs in Biomedicine 239, 107631 (9 2023)

[121] Kalbfleisch, J.D., Prentice, R.: Survival analysis (1980)

[122] Kallipolitis, A., Revelos, K., Maglogiannis, I.: Ensembling efficientnets for the classification and interpretation of histopathology images. Algorithms 2021, Vol. 14, Page 278 14, 278 (9 2021), https://www.mdpi.com/1999-4893/14/10/278/htmhttps://www.mdpi.com/1999-4893/14/10/278

[123] Kamran, F., Wiens, J.: Estimating Calibrated Individualized Survival Curves with Deep Learning. TheThirty-Fifth AAAI Conference on Artificial Intelligence 35(1), 240–248 (2021)

[124] Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association 53(282), 457–481 (1958)

[125] Kattan, M.W., Wheeler, T.M., Scardino, P.T., et al.: Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. Journal of clinical oncology 17(5), 1499–1507 (1999)

[126] Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. BMC Medical Research Methodology 18, 1–12 (2 2018), https://link.springer.com/articles/10.1186/s12874-018-0482-1

[127] Kim, I., Kang, K., Song, Y., Kim, T.J.: Application of artificial intelligence in pathology: Trends and challenges. Diagnostics 12(11) (2022), https://www.mdpi.com/2075-4418/12/11/2794

[128] Kim, Y.J., Jang, H., Lee, K., Park, S., Min, S.G., Hong, C., Park, J.H., Lee, K., Kim, J., Hong, W., Jung, H., Liu, Y., Rajkumar, H., Khened, M., Krishnamurthi, G., Yang, S., Wang, X., Han, C.H., Kwak, J.T., Ma, J., Tang, Z., Marami, B., Zeineh, J., Zhao, Z., Heng, P.A., Schmitz, R., Madesta, F., Rösch, T., Werner, R., Tian, J., Puybareau, E., Bovio, M., Zhang, X., Zhu, Y., Chun, S.Y., Jeong, W.K., Park, P., Choi, J.: Paip 2019: Liver cancer segmentation challenge. Medical Image Analysis 67, 101854 (1 2021)

[129] Klein, J.P., Moeschberger, M.L., et al.: Survival analysis: techniques for censored and truncated data, vol. 1230. Springer (2003)

[130] Knaus, W.A., Harrell, F.E., Lynn, J., Goldman, L., Phillips, R.S., Connors, A.F., Dawson, N.V., Fulkerson, W.J., Califf, R.M., Desbiens, N., Layde, P., Oye, R.K., Bellamy, P.E., Hakim, R.B., Wagner, D.P.: The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Annals of Internal Medicine 122(3), 191–203 (1995)

[131] Kononen, J., Bubendorf, L., Kallioniemi, A., Bärlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M.J., Sauter, G., Kallioniemi, O.P.: Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nature medicine 4, 844–847 (7 1998), https://pubmed.ncbi.nlm.nih.gov/9662379/

[132] Kovalev, M.S., Utkin, L.V., Kasimov, E.M.: Survlime: A method for explaining machine learning survival models. Knowledge-Based Systems 203, 106164 (2020)

[133] Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009), https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[134] Krzyziński, M., Spytek, M., Baniecki, H., Biecek, P.: Survshap(t): Time-dependent explanations of machine learning survival models. Knowledge-Based Systems 262, 110234 (2 2023)

[135] Kullback, S., Leibler, R.A.: On information and sufficiency. https://doi.org/10.1214/aoms/1177729694 22, 79–86 (3 1951), https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Sufficiency/10.1214/aoms/1177729694.fullhttps://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-1/On-Information-and-Su

[136] Kvamme, H., Ørnulf Borgan, Scheel, I.: Time-to-event prediction with neural networks and cox regression. Journal of Machine Learning Research 20, 1–30 (2019), http://jmlr.org/papers/v20/18-424.html.

[137] Kwon, B.C., Choi, M.J., Kim, J.T., Choi, E., Kim, Y.B., Kwon, S., Sun, J., Choo, J.: Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. IEEE Transactions on Visualization and Computer Graphics 25, 299–309 (1 2019), https://ieeexplore.ieee.org/document/8440842/

[138] Landi, I., Glicksberg, B.S., Lee, H.C., Cherng, S., Landi, G., Danieletto, M., Dudley, J.T., Furlanello, C., Miotto, R.: Deep representation learning of electronic health records to unlock patient stratification at scale. npj Digital Medicine 3, 96 (12 2020), http://arxiv.org/abs/2003.06516https://doi.org/10.1038/s41746-020-0301-zhttp://www.nature.com/articles/s41746-020-0301-z

[139] Leapman, M.S., Carroll, P.R.: New genetic markers for prostate cancer. Urologic Clinics of North America 43, 7–15 (2 2016), http://www.urologic.theclinics.com/article/S0094014315000841/fulltexthttp://www.urologic.theclinics.com/article/S0094014315000841/abstracthttps://www.urologic.theclinics.com/article/S0094-0143(15)00084-1/abstract

[140] Lee, B., Chun, S.H., Hong, J.H., Woo, I.S., Kim, S., Jeong, J.W., Kim, J.J., Lee, H.W., Na, S.J., Beck, K.S., Gil, B., Park, S., An, H.J., Ko, Y.H., Hoon Chun, S., Hyung Hong, J., Sook Woo, I., Kim, S., Won Jeong, J., Jun Kim, J., Woo Lee, H., Jung Na, S., Sarah Beck, K., Gil, B., Park, S., Jung An, H., Ho Ko, Y.: DeepBTS: Prediction of Recurrence-free Survival of Non-small Cell Lung Cancer Using a Time-binned Deep Neural Network. Scientific Reports 10(1) (dec 2020), /pmc/articles/PMC7005286/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7005286/

[141] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems 31 (2018)

[142] Lee, K.H., Choi, G.H., Yun, J., Choi, J., Goh, M.J., Sinn, D.H., Jin, Y.J., Kim, M.A., Yu, S.J., Jang, S., Lee, S.K., Jang, J.W., Lee, J.S., Kim, D.Y., Cho, Y.Y., Kim, H.J., Kim, S., Kim, J.H., Kim, N., Kim, K.M.: Machine learning-based clinical decision support system for treatment recommendation and overall survival prediction of hepatocellular carcinoma: a multi-center study. npj Digital Medicine 2024 7:1 7, 1–8 (1 2024), https://www.nature.com/articles/s41746-023-00976-8

[143] Leung, K.M., Elashoff, R.M., Afifi, A.A.: Censoring issues in survival analysis. Annual review of public health 18(1), 83–104 (1997)

[144] Li, H., Han, D., Hou, Y., Chen, H., Chen, Z.: Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One 10(1), e0116774 (2015)

[145] Li, J., Ettel, M., Amin, A., Bhalla, R., Das, K., Deng, F.M., Lee, P., Matoso, A., Melamed, J., Mendrinos, S., Tian, W., Yaskiv, O., Shah, R.B., Zhou, M.: Interobserver reproducibility of quantifying gleason pattern 4 cancer in prostate biopsy: Implications for clinical practice. http://www.xiahepublishing.com/ 3, 4–9 (3 2023), http://www.xiahepublishing.com/2771-165X/JCTP-2022-00026

[146] Li, L., Cheng, W.Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P., Dudley, J.T.: Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Science Translational Medicine 7 (10 2015), https://pubmed.ncbi.nlm.nih.gov/26511511/

[147] Li, Y., Rao, S., Solares, J.R.A., Hassaine, A., Ramakrishnan, R., Canoy, D., Zhu, Y., Rahimi, K., Salimi-Khorshidi, G.: Behrt: Transformer for electronic health records. Scientific Reports 10, 1–12 (12 2020), https://www.nature.com/articles/s41598-020-62922-y

[148] Li, Y., Nair, P., Lu, X.H., Wen, Z., Wang, Y., Dehaghi, A.A.K., Miao, Y., Liu, W., Ordog, T., Biernacka, J.M., Ryu, E., Olson, J.E., Frye, M.A., Liu, A., Guo, L., Marelli, A., Ahuja, Y., Davila-Velderrain, J., Kellis, M.: Inferring multimodal latent topics from electronic health records. Nature Communications 11, 1–17 (12 2020), https://doi.org/10.1038/s41467-020-16378-3

[149] Lilja, H., Ulmert, D., Vickers, A.J.: Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. Nature Reviews Cancer 2008 8:4 8, 268–278 (4 2008), https://www.nature.com/articles/nrc2351

[150] Lipton, Z.C., Kale, D.C., Elkan, C.P., Wetzel, R.C.: Learning to diagnose with lstm recurrent neural networks. CoRR abs/1511.03677 (2015), https://api.semanticscholar.org/CorpusID:13880

[151] Litjens, G., Bandi, P., Bejnordi, B.E., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J.: 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. GigaScience 7, 1–8 (6 2018), https://dx.doi.org/10.1093/gigascience/giy065

[152] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. 7th International Conference on Learning Representations, ICLR 2019 (11 2017), https://arxiv.org/abs/1711.05101v3

[153] Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009 pp. 1107–1110 (2009)

[154] MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Cam, L.M.L., Neyman, J. (eds.) Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability. vol. 1, pp. 281–297. University of California Press (1967)

[155] Mariotto, A.B.: Cancer Survival from a Policy and Clinical Perspective: US Surveillance, Epidemiology, and End Results (SEER) Program, 1975-2010. Oxford University Press (2014)

[156] Marletta, S., Eccher, A., Martelli, F.M., Santonicco, N., Girolami, I., Scarpa, A., Pagni, F., L'Imperio, V., Pantanowitz, L., Gobbo, S., Seminati, D., Tos, A.P.D., Parwani, A.: Artificial intelligence–based algorithms for the diagnosis of prostate cancer: A systematic review. American Journal of Clinical Pathology (2 2024), https://dx.doi.org/10.1093/ajcp/aqad182

[157] Mascharak, S., Guo, J.L., Foster, D.S., Khan, A., Davitt, M.F., Nguyen, A.T., Burcham, A.R., Chinta, M.S., Guardino, N.J., Griffin, M., et al.: Desmoplastic stromal signatures predict patient outcomes in pancreatic ductal adenocarcinoma. Cell Reports Medicine 4(11) (2023)

[158] McNeal, J.E.: The zonal anatomy of the prostate. The Prostate 2, 35–49 (1 1981), https://onlinelibrary.wiley.com/doi/full/10.1002/pros.2990020105https://onlinelibrary.wiley.com/doi/abs/10.1002/pros.2990020105https://onlinelibrary.wiley.com/doi/10.1002/pros.2990020105

[159] Meacham, S., Isaac, G., Nauck, D., Virginas, B.: Towards explainable ai: Design and development for explanation of machine learning predictions for a patient readmittance medical application. Advances in Intelligent Systems and Computing 997, 939–955 (2019), https://link.springer.com/chapter/10.1007/978-3-030-22871-2_67

[160] Melamed, J., of Medicine, N.Y.U.S.: Prostate cancer biorepository network (pcbn) (2019)

[161] Mellinger, G.T., Gleason, D., Bailar, J.: The histology and prognosis of prostatic cancer. The Journal of urology 97(2), 331–337 (1967)

[162] Meng, Z., Xu, J., Li, Z., Wang, Y., Chen, F., Wang, Z.: A multi-task kernel learning algorithm for survival analysis. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 298–311. Springer (2021)

[163] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality (2013)

[164] Miotto, R., Li, L., Kidd, B.A., Dudley, J.T.: Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports 6, 1–10 (5 2016), www.nature.com/scientificreports/

[165] Moncada-Torres, A., van Maaren, M.C., Hendriks, M.P., Siesling, S., Geleijnse, G.: Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. Scientific reports 11(1), 6968 (2021)

[166] Moon, H.H., Kim, H.S., Park, J.E., Lee, J.S.: Do findings on immediate post-radiation advanced mri change treatment decision and clinical outcome in glioblastoma? a propensity score-matched study (1 2023), https://www.researchsquare.comhttps://www.researchsquare.com/article/rs-2505778/v1

[167] Moran, P.S., O'Neill, M., Teljeur, C., Flattery, M., Murphy, L.A., Smyth, G., Ryan, M.: Robot-assisted radical prostatectomy compared with open and laparoscopic approaches: a systematic review and meta-analysis. International Journal of Urology 20(3), 312–321 (2013)

[168] Mottet, N., Bellmunt, J., Briers, E., Van den Bergh, R., Bolla, M., Van Casteren, N., Cornford, P., Culine, S., Joniau, S., Lam, T., et al.: Guidelines on prostate cancer. European Association of Urology 56, e137 (2015)

[169] Munien, C., Viriri, S.: Classification of hematoxylin and eosin-stained breast cancer histology microscopy images using transfer learning with efficientnets. Computational Intelligence and Neuroscience 2021 (2021)

[170] Nagpal, K., Foote, D., Liu, Y., Chen, P.H.C., Wulczyn, E., Tan, F., Olson, N., Smith, J.L., Mohtashamian, A., Wren, J.H., et al.: Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. NPJ digital medicine 2(1), 48 (2019)

[171] Nezhad, M.Z., Sadati, N., Yang, K., Zhu, D.: A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. Expert Systems with Applications 115, 16–26 (1 2019)

[172] Olawaiye, A.B., Baker, T.P., Washington, M.K., Mutch, D.G.: The new (version 9) american joint committee on cancer tumor, node, metastasis staging for cervical cancer. CA: a cancer journal for clinicians 71, 287–298 (7 2021), https://pubmed.ncbi.nlm.nih.gov/33784415/

[173] Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E.A., Blilie, A., Egevad, L., Spjuth, O., Eklund, M.: Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature Communications 2022 13:1 13, 1–10 (12 2022), https://www.nature.com/articles/s41467-022-34945-8

[174] Omar, M.I., Roobol, M.J., Ribal, M.J., Abbott, T., Agapow, P.M., Araujo, S., Asiimwe, A., Auffray, C., Balaur, I., Beyer, K., Bernini, C., Bjartell, A., Briganti, A., Butler-Ransohoff, J.E., Campi, R., Cavelaars, M., Meulder, B.D., Devecseri, Z., Voss, M.D., Dimitropoulos, K., Evans-Axelsson, S., Franks, B., Fullwood, L., Horgan, D., Smith, E.J., Kiran, A., Kivinummi, K., Lambrecht, M., Lancet, D., Lindgren, P., MacLennan, S., MacLennan, S., Nogueira, M.M., Moen, F., Moinat, M., Papineni, K., Reich, C., Reiche, K., Rogiers, S., Sartini, C., van Bochove, K., van Diggelen, F., Hemelrijck, M.V., Poppel, H.V., Zong, J., N'Dow, J., Andersson, E., Arala, H., Auvinen, A., Bangma, C., Burke, D., Cardone, A., Casariego, J., Cuperus, G., Dabestani, S., Esperto, F., Fossati, N., Fridhammar, A., Gandaglia, G., Tandefelt, D.G., Horn, F., Huber, J., Hugosson, J., Huisman, H., Josefsson, A., Kilkku, O., Kreuz, M., Lardas, M., Lawson, J., Lefresne, F., Lejeune, S., Longden-Chapman, E., McVie, G., Moris, L., Mottet, N., Murtola, T., Nicholls, C., Pang, K.H., Pascoe, K., Picozzi, M., Plass, K., Pohjanjousi, P., Reaney, M., Remmers, S., Robinson, P., Schalken, J., Schravendeel, M., Seisen, T., Servan, A., Shiranov, K., Snijder, R., Steinbeisser, C., Taibi, N., Talala, K., Tilki, D., den Broeck, T.V., Vassilev, Z., Voima, O., Vradi, E., Waldeck, R., Weistra, W., Willemse, P.P., Wirth, M., Wolfinger, R., Kermani, N.Z.: Introducing pioneer: a project to harness big data in prostate cancer research. Nature Reviews Urology 2020 17:6 17, 351–362 (5 2020), https://www.nature.com/articles/s41585-020-0324-x

[175] van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR abs/1807.03748 (2018), http://arxiv.org/abs/1807.03748

[176] Otsu, N., et al.: A threshold selection method from gray-level histograms. Automatica 11(285-296), 23–27 (1975)

[177] Ozkan, T.A., Eruyar, A.T., Cebeci, O.O., Memik, O., Ozcan, L., Kuskonmaz, I.: Interobserver variability in gleason histological grading of prostate cancer. Scandinavian Journal of Urology 50, 420–424 (11 2016), https://www.tandfonline.com/doi/abs/10.1080/21681805.2016.1206619

[178] Pantanowitz, L., Quiroga-Garza, G.M., Bien, L., Heled, R., Laifenfeld, D., Linhart, C., Sandbank, J., Shach, A.A., Shalev, V., Vecsler, M., Michelow, P., Hazelhurst, S., Dhir, R.: An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. The Lancet Digital Health 2, e407–e416 (8 2020)

[179] Paolucci, I., Lin, Y.M., Albuquerque, J., Silva, M., Brock, K.K., Odisio, B.C.: Bayesian parametric models for survival prediction in medical applications. BMC Medical Research Methodology 2023 23:1 23, 1–14 (10 2023), https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-02059-4http://creativecom-mons.org/publicdomain/zero/1.0/

[180] Park, S., Hendry, D.J.: Reassessing schoenfeld residual tests of proportional hazards in political science event history analyses. American Journal of Political Science 59, 1072–1087 (10 2015), https://onlinelibrary.wiley.com/doi/full/10.1111/ajps.12176https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12176https://onlinelibrary.wiley.com/doi/10.1111/ajps.12176

[181] Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 493–501. Springer International Publishing, Cham (2018)

[182] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)

[183] Pencina, M.J., D'Agostino, R.B.: Evaluating discrimination of risk prediction models: the c statistic. Jama 314(10), 1063–1064 (2015)

[184] Pinckaers, H., van Ipenburg, J., Melamed, J., Marzo, A.D., Platz, E.A., van Ginneken, B., van der Laak, J., Litjens, G.: Predicting biochemical recurrence of prostate cancer with artificial intelligence. Communications Medicine 2022 2:1 2, 1–9 (6 2022), https://www.nature.com/articles/s43856-022-00126-3

[185] Pölsterl, S.: scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research 21(212), 1–6 (2020), http://jmlr.org/papers/v21/20-729.html

[186] Prasanta Chandra, M., et al.: On the generalised distance in statistics. In: Proceedings of the National Institute of Sciences of India. vol. 2, pp. 49–55 (1936)

[187] Qin, Y., Han, H., Xue, Y., Wu, C., Wei, X., Liu, Y., Cao, Y., Ruan, Y., He, J.: Comparison and trend of perioperative outcomes between robot-assisted radical prostatectomy and open radical prostatectomy: nationwide inpatient sample 2009-2014. International braz j urol 46, 754–771 (2020)

[188] Qiu, S., Joshi, P.S., Miller, M.I., Xue, C., Zhou, X., Karjadi, C., Chang, G.H., Joshi, A.S., Dwyer, B., Zhu, S., Kaku, M., Zhou, Y., Alderazi, Y.J., Swaminathan, A., Kedar, S., Saint-Hilaire, M.H., Auerbach, S.H., Yuan, J., Sartor, E.A., Au, R., Kolachalama, V.B.: Development and validation of an interpretable deep learning framework for alzheimer's disease classification. Brain : a journal of neurology 143, 1920–1933 (2020), https://www.ncbi.nlm.nih.gov/pubmed/32357201https://academic.oup.com/brain/article/143/6/1920/5827821

[189] Qu, L., Ma, Y., Luo, X., Wang, M., Song, Z.: Rethinking multiple instance learning for whole slide image classification: A good instance classifier is all you need. arXiv preprint arXiv:2307.02249 (2023)

[190] Qu, L., Yang, Z., Duan, M., Ma, Y., Wang, S., Wang, M., Song, Z.: Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21463–21473 (2023)

[191] Rahman, R., Fell, G., Ventz, S., Arfé, A., Vanderbeek, A.M., Trippa, L., Alexander, B.M.: Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: Prevalence, associated factors, and implications. Clinical Cancer Research 25, 6339–6345 (11 2019), /clincancerres/article/25/21/6339/82099/Deviation-from-the-Proportional-Hazards-Assumptionhttps://dx.doi.org/10.1158/1078-0432.CCR-18-3999

[192] Rajinikanth, A., Manoharan, M., Soloway, C.T., Civantos, F.J., Soloway, M.S.: Trends in gleason score: Concordance between biopsy and prostatectomy over 15 years. Urology 72, 177–182 (7 2008)

[193] Ramachandran, P., Bello, I., Parmar, N., Levskaya, A., Vaswani, A., Shlens, J.: Stand-alone self-attention in vision models. Advances in Neural Information Processing Systems 32 (6 2019), https://arxiv.org/abs/1906.05909v1

[194] Rawla, P.: Epidemiology of prostate cancer. World journal of oncology 10, 63–89 (2019), https://pubmed.ncbi.nlm.nih.gov/31068988/

[195] Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., Yu, Y.: Deep recurrent survival analysis. 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019 pp. 4798–4805 (9 2019), http://arxiv.org/abs/1809.02403

[196] Rindt, D., Hu, R., Steinsaltz, D., Sejdinovic, D.: Survival regression with proper scoring rules and monotonic neural networks. Proceedings of Machine Learning Research 151, 1190–1205 (3 2021), https://arxiv.org/abs/2103.14755v2

[197] Robins, J.M., Finkelstein, D.M.: Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. Biometrics 56(3), 779–788 (2000)

[198] Ronckers, C., Spix, C., Trübenbach, C., Katalinic, A., Christ, M., Cicero, A., Folkerts, J., Hansmann, J., Kranzhöfer, K., Kunz, B., et al.: Krebs in deutschland für 2019/2020 (2023)

[199] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351, 234–241 (2015), https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28

[200] Roy, M., Kong, J., Kashyap, S., Pastore, V.P., Wang, F., Wong, K.C., Mukherjee, V.: Convolutional autoencoder based model HistoCAE for segmentation of viable tumor regions in liver whole-slide images. Scientific Reports 2021 11:1 11(1), 1–10 (jan 2021), https://www.nature.com/articles/s41598-020-80610-9

[201] Rubner, Y., Tomasi, C., Guibas, L.J.: A metric for distributions with applications to image databases. In: Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). pp. 59–66. IEEE (1998)

[202] Rymarczyk, D., Borowa, A., Tabor, J., Zielinski, B.: Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1721–1730 (2021)

[203] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks (2018)

[204] Saraswat, D., Bhattacharya, P., Verma, A., Prasad, V.K., Tanwar, S., Sharma, G., Bokoro, P.N., Sharma, R.: Explainable ai for healthcare 5.0: opportunities and challenges. IEEE Access 10, 84486–84517 (2022)

[205] Sasaki, Y., et al.: The truth of the f-measure. Teach tutor mater 1(5), 1–5 (2007)

[206] Sauter, G., Clauditz, T., Steurer, S., Wittmer, C., Büscheck, F., Krech, T., Lutz, F., Lennartz, M., Harms, L., Lawrenz, L., Möller-Koop, C., Simon, R., Jacobsen, F., Wilczak, W., Minner, S., Tsourlakis, M.C., Chirico, V., Weidemann, S., Haese, A., Steuber, T., Salomon, G., Matiu, M., Vettorazzi, E., Michl, U., Budäus, L., Tilki, D., Thederan, I., Pehrke, D., Beyer, B., Fraune, C., Göbel, C., Heinrich, M., Juhnke, M., Möller, K., Bawahab, A.A.A., Uhlig, R., Huland, H., Heinzer, H., Graefen, M., Schlomm, T.: Integrating tertiary gleason 5 patterns into quantitative gleason grading in prostate biopsies and prostatectomy specimens. European Urology 73, 674–683 (5 2018)

[207] Sauter, G., Steurer, S., Clauditz, T.S., Krech, T., Wittmer, C., Lutz, F., Lennartz, M., Janssen, T., Hakimi, N., Simon, R., et al.: Clinical utility of quantitative gleason grading in prostate biopsies and prostatectomy specimens. European urology 69(4), 592–598 (2016)

[208] Schlomm, T., Sauter, G.: Beurteilung des prostatakarzinoms: Gleason-score – status 2016. Deutsches Ärzteblatt Online (8 2016)

[209] Sciarra, A., Santarelli, V., Salciccia, S., Moriconi, M., Basile, G., Santodirocco, L., Carino, D., Frisenda, M., Pierro, G.D., Giudice, F.D., Gentilucci, A., Bevilacqua, G.: How the management of biochemical recurrence in prostate cancer will be modified by the concept of anticipation and incrementation of therapy. Cancers 2024, Vol. 16, Page 764 16, 764 (2 2024), https://www.mdpi.com/2072-6694/16/4/764/htmhttps://www.mdpi.com/2072-6694/16/4/764

[210] Seaman, E., Whang, M., Olsson, C.A., Katz, A., Cooner, W.H., Benson, M.C.: Psa density (psad). role in patient evaluation and management. The Urologic Clinics of North America 20, 653–663 (11 1993), https://europepmc.org/article/med/7505973

[211] Sfoungaristos, S., Perimenis, P.: Evaluating psa density as a predictor of biochemical failure after radical prostatectomy: Results of a prospective study after a median follow-up of 36 months. ISRN Urology 2013, 1–5 (5 2013)

[212] Shamshad, F., Khan, S., Zamir, S.W., Khan, M.H., Hayat, M., Khan, F.S., Fu, H.: Transformers in medical imaging: A survey. Medical Image Analysis p. 102802 (2023)

[213] Shapley, L., Corporation, R.: Notes on the N-person Game. No. 2 in Notes on the N-person Game, Rand Corporation (1951)

[214] Smith, M.R., Saad, F., Egerdie, B., Sieber, P.R., Tammela, T.L., Ke, C., Leder, B.Z., Goessl, C., Taaffe, D.R.D.R., Newton, R.U.R.U., Spry, N., Joseph, D.J., Galvão, D.A.D., Owen, P.J., Daly, R.M., Livingston, P.M., Fraser, S.F., Veni, T., Leroy, V., Pradère, B., Rébillard, A., Mathieu, R., Tzortzis, V., Samarinas, M., Zachos, I., Oeconomou, A., Pisters, L.L., Bargiota, A., Keating, N.L., O'Malley, A.J., McNaughton-Collins, M., Oh, W.K., Smith, M.R., Rawla, P., Sprod, L.K., Mohile, S.G., Demark-Wahnefried, W., Janelsins, M.C., Peppone, L.J., Morrow, G.R., Lord, R., Gross, H., Mustian, K.M., Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E., Forman, D., Cormie, P., Newton, R.U.R.U., Spry,

N., Joseph, D.J., Taaffe, D.R.D.R., Galvão, D.A.D., Schwartz, L.H., Litière, S., Vries, E.D., Ford, R., Mandrekar, S., Shankar, L., Bogaerts, J., Chen, A., Hayes, W., Hodi, F.S., Hoekstra, O.S., Huang, E.P., Buffart, L.M., Kalter, J., Sweegers, M.G., Courneya, K.S., Newton, R.U.R.U., Aaronson, N.K., Jacobsen, P.B., May, A.M., Galvão, D.A.D., Chinapaw, M.J., Steindorf, K., Irwin, M.L., Stuiver, M.M., Hayes, S.C.S., Griffith, K.A., Lucia, A., Mesters, I., van Weert, E., Knoop, H., Goedendorp, M.M., Mutrie, N., Daley, A.J., McConnachie, A., Bohus, M., Thorsen, L., Schulz, K.H., Short, C.E., James, E.L., Plotnikoff, R.C., Arbane, G., Schmidt, M.E., Potthoff, K., van Beurden, M., Oldenburg, H.S., Sonke, G.S., van Harten, W.H., Garrod, R., Schmitz, K.H., Winters-Stone, K.M., Velthuis, M.J., Taaffe, D.R.D.R., van Mechelen, W., Kersten, M.J., Nollet, F., Wenzel, J., Wiskemann, J., de Leeuw, I.M.V., Brug, J., Newton, R.U.R.U., Spence, R.R., Galvão, D.A.D., Fairman, C., Kendall, K., Newton, R.U.R.U., Hart, N., Taaffe, D.R.D.R., Chee, R., Tang, C., Galvão, D.A.D.: Sarcopenia during adt for pc. Prostate Cancer and Prostatic Diseases 16, 1077–1083 (2019), https://doi.org/10.14740/wjon1191

[215] Sohn, E.: Screening: Diagnostic dilemma. Nature 2015 528:7582 528, S120–S122 (12 2015), https://www.nature.com/articles/528S120a

[216] Sorensen, T.: A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. Biologiske skrifter 5, 1–34 (1948)

[217] Specht, M.C., Kattan, M.W., Gonen, M., Fey, J., Zee, K.J.V.: Predicting nonsentinel node status after positive sentinel lymph biopsy for breast cancer: Clinicians versus nomogram. Annals of Surgical Oncology 12, 654–659 (8 2005), https://link.springer.com/article/10.1245/ASO.2005.06.037

[218] Stacke, K., Eilertsen, G., Unger, J., Lundström, C.: A closer look at domain shift for deep learning in histopathology. In: MICCAI 2019 Computational Pathology Workshop COMPAY (2019), https://openreview.net/forum?id=rkeVzGaobS

[219] Stensrud, M.J., Hernán, M.A.: Why test for proportional hazards? Jama 323(14), 1401–1402 (2020), https://doi.org/10.1001/jama.2020.1267

[220] Stephenson, A.J., Scardino, P.T., Eastham, J.A., Bianco, F.J., Dotan, Z.A., DiBlasio, C.J., Reuther, A., Klein, E.A., Kattan, M.W.: Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. Journal of Clinical Oncology 23(28), 7005–7012 (2005), /pmc/articles/PMC2231088/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2231088/

[221] Steyerberg, E.W., Vickers, A.J., Cook, N.R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M.J., Kattan, M.W.: Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.) 21, 128 (1 2010), /pmc/articles/PMC3575184//pmc/articles/PMC3575184/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3575184/

[222] Ström, P., Kartasalo, K., Olsson, H., Solorzano, L., Delahunt, B., Berney, D.M., Bostwick, D.G., Evans, A.J., Grignon, D.J., Humphrey, P.A., et al.: Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. The Lancet Oncology 21(2), 222–232 (2020)

[223] Su, Y., Bai, Y., Zhang, B., Zhang, Z., Wang, W.: Hat-net: A hierarchical transformer graph neural network for grading of colorectal cancer histology images. In: BMVC. p. 412 (2021)

[224] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)

[225] Talo, M.: Automated classification of histopathology images using transfer learning. Artificial intelligence in medicine 101, 101743 (2019)

[226] Tan, M., Le, Q.V.: Efficientnet: Rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, ICML 2019 2019-June, 10691–10700 (5 2019), https://arxiv.org/abs/1905.11946v5

[227] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis 58, 101544 (12 2019)

[228] Therneau, T.M., Lumley, T.: Package 'survival'. R Top Doc 128(10), 28–33 (2015)

[229] Tiruye, T., O'callaghan, M., Ettridge, K., Moretti, .K., Jay, A., Higgs, .B., Santoro, K., Kichenadasse, .G., Beckmann, K.: Clinical and functional outcomes for risk-appropriate treatments for prostate cancer. BJUI Compass (9 2023), https://onlinelibrary.wiley.com/doi/full/10.1002/bco2.288https://onlinelibrary.wiley.com/doi/abs/10.1002/bco2.288https://bjui-journals.onlinelibrary.wiley.com/doi/10.1002/bco2.288

[230] Toccaceli, P., Gammerman, A.: Combination of inductive mondrian conformal predictors. Machine Learning 108, 489–510 (2019)

[231] Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: Contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning Research (5 2019), http://arxiv.org/abs/1905.05134

[232] Truesdale, M.D., Cheetham, P.J., Turk, A.T., Sartori, S., Hruby, G.W., Dinneen, E.P., Benson, M.C., Badani, K.K.: Gleason score concordance on biopsy-confirmed prostate cancer: is pathological re-evaluation necessary prior to radical prostatectomy? BJU International 107, 749–754 (3 2011), https://onlinelibrary.wiley.com/doi/full/10.1111/j.1464-410X.2010.09570.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.1464-410X.2010.09570.xhttps://bjui-journals.onlinelibrary.wiley.com/doi/10.1111/j.1464-410X.2010.09570.x

[233] Uno, H., Cai, T., Pencina, M.J., D'Agostino, R.B., Wei, L.J.: On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine 30, 1105–1117 (5 2011), https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4154https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4154https://onlinelibrary.wiley.com/doi/10.1002/sim.4154

[234] Uno, H., Cai, T., Tian, L., Wei, L.J.: Evaluating prediction rules for t-year survivors with censored regression models. Journal of the American Statistical Association 102, 527–537 (6 2007), https://www.tandfonline.com/doi/abs/10.1198/016214507000000149

[235] Utkin, L.V., Kovalev, M.S., Kasimov, E.M.: Survlime-inf: A simplified modification of survlime for explanation of machine learning survival models. arXiv preprint arXiv:2005.02387 (2020)

[236] Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Transactions on Medical Imaging 35, 1962–1971 (8 2016)

[237] Vasilev, I., Petrovskiy, M., Mashechkin, I.: Sensitivity of survival analysis metrics. Mathematics 2023, Vol. 11, Page 4246 11, 4246 (10 2023), https://www.mdpi.com/2227-7390/11/20/4246/htmhttps://www.mdpi.com/2227-7390/11/20/4246

[238] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[239] Vesal, S., Ravikumar, N., Davari, A., Ellmann, S., Maier, A.: Classification of breast cancer histology images using transfer learning. In: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27–29, 2018, Proceedings 15. pp. 812–819. Springer (2018)

[240] Vickers, A.J., Brewster, S.F.: Psa velocity and doubling time in diagnosis and prognosis of prostate cancer. British journal of medical & surgical urology 5, 162 (7 2012), /pmc/articles/PMC3375697//pmc/articles/PMC3375697/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3375697/

[241] Vickers, A.J., Fine, S.W.: Three things about gleason grading that just about everyone believes but that are almost certainly wrong. Urology 143, 16–19 (9 2020), http://www.goldjournal.net/article/S0090429520303678/fulltexthttp://www.goldjournal.net/article/S0090429520303678/abstracthttps://www.goldjournal.net/article/S0090-4295(20)30367-8/abstract

[242] Vickers, A.J., Kent, M., Scardino, P.T.: Implementation of dynamically updated prediction models at the point of care at a major cancer center: Making nomograms more like netflix. Urology 102, 1–3 (4 2017), http://www.goldjournal.net/article/S0090429516309049/fulltexthttp://www.goldjournal.net/article/S0090429516309049/abstracthttps://www.goldjournal.net/article/S0090-4295(16)30904-9/abstract

[243] Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E.: Deep learning for computer vision: A brief review. Computational intelligence and neuroscience 2018 (2018)

[244] Vovk, V., Gammerman, A., Shafer, G.: Conformal prediction: General case and regression. Algorithmic Learning in a Random World pp. 19–69 (2022), https://link.springer.com/chapter/10.1007/978-3-031-06649-8_2

[245] Walhagen, P., Bengtsson, E., Lennartz, M., Sauter, G., Busch, C.: Ai-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples. Journal of Pathology Informatics 13, 100137 (2022), https://www.sciencedirect.com/science/article/pii/S2153353922007313

[246] Walhagen, P., Pontus, R., Ewert, B., Christer, B., Michael, H.: Spear prostate biopsy 2020 (sprob20) (2020), https://datahub.aida.scilifelab.se/10.23698/aida/sprob20

[247] Wang, H., Xia, Z., Xu, Y., Sun, J., Wu, J.: The predictive value of machine learning and nomograms for lymph node metastasis of prostate cancer: a systematic review and meta-analysis. Prostate Cancer and Prostatic Diseases 2023 26:3 26, 602–613 (7 2023), https://www.nature.com/articles/s41391-023-00704-z

[248] Wang, L., Tong, L., Davis, D., Arnold, T., Esposito, T.: The application of unsupervised deep learning in predictive models using electronic health records. BMC Medical Research Methodology 20, 37 (2 2020), https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-020-00923-1

[249] Wang, P., Li, Y., Reddy, C.K.: Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR) 51(6), 1–36 (2019)

[250] Wang, T.H., Lee, C.Y., Lee, T.Y., Huang, H.D., Hsu, J.B.K., Chang, T.H.: Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion. Cancers 13, 2528 (6 2021), https://www.mdpi.com/2072-6694/13/11/2528/htmhttps://www.mdpi.com/2072-6694/13/11/2528

[251] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Huang, J., Yang, W., Han, X.: Transpath: Transformer-based self-supervised learning for histopathological image classification. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 186–195. Springer (2021)

[252] Wei, L.J.: The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine 11, 1871–1879 (1 1992), https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780111409https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780111409https://onlinelibrary.wiley.com/doi/10.1002/sim.4780111409

[253] Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics 45(10), 1113–1120 (2013)

[254] Wilm, F., Marzahl, C., Breininger, K., Aubreville, M.: Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization challenge. In: Aubreville, M., Zimmerer, D., Heinrich, M. (eds.) Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. pp. 5–13. Springer International Publishing, Cham (2022)

[255] Wu, J., Chen, X.Y., Zhang, H., Xiong, L.D., Lei, H., Deng, S.H.: Hyperparameter optimization for machine learning models based on bayesian optimization. Journal of Electronic Science and Technology 17(1), 26–40 (2019)

[256] Xu, H., Xu, Q., Cong, F., Kang, J., Han, C., Liu, Z., Madabhushi, A., Lu, C.: Vision transformers for computational histopathology. IEEE Reviews in Biomedical Engineering (2023)

[257] Yan, G., Tom, G.: Investigating the effects of ties on measures of concordance. Statistics in Medicine 27, 4190–4206 (9 2008), https://onlinelibrary.wiley.com/doi/full/10.1002/sim.3257https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3257https://onlinelibrary.wiley.com/doi/10.1002/sim.3257

[258] Yanagisawa, H.: Proper scoring rules for survival analysis (7 2023), https://proceedings.mlr.press/v202/yanagisawa23a.html

[259] Yang, Y., Fasching, P.A., Tresp, V.: Predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. In: 2017 IEEE International Conference on Healthcare Informatics (ICHI). pp. 46–55 (2017)

[260] Yao, L., Li, S., Li, Y., Huai, M., Gao, J., Zhang, A.: Representation learning for treatment effect estimation from observational data. Advances in neural information processing systems 31 (2018)

[261] Ying, Z., Jung, S.H., Wei, L.J.: Survival analysis with median regression models. Journal of the American Statistical Association 90, 178–184 (1995)

[262] Yu, C.N., Greiner, R., Lin, H.C., Baracos, V.: Learning patient-specific cancer survival distributions as a sequence of dependent regressors (2011)

[263] Zelenčík, M.: Vision transformers as a support for prostate cancer detection. Master's thesis, NTNU (2023)

[264] Zhang, J., Kowsari, K., Harrison, J.H., Lobo, J.M., Barnes, L.E.: Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. IEEE Access 6, 65333–65346 (2018)

Appendix

A Additional Datasets

This section provides a list of related EHR and PCa related datasets that were considered in this thesis, but did not meet the requirements for this thesis like insufficient follow-up durations. Other datasets could be a good source for similar analyses on different organs.

PLCO The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial [90] is a collection of approximately 148,000 participants aged 55-74 at 10 centers in the United States that determines the effect of screening on cancer-related mortality in a randomized study. Regarding PCa, the dataset contains approximately 15,600 diagnostic procedures for PCa, 8,800 patients with diagnosed PCa and treatment information. Additionally, data from 3,200 RPs including complications are documented.

CAMELYON The CAMELYON dataset [151] contains 1,399 H&E-stained WSI images of lymph node sections related to breast cancer patients. It can be used for the detection and classification of breast cancer metastases.

PAIP2019 The PAIP challenge [128] dealt with liver cancer segmentation and risk estimation on 50 samples showing tumor lesions in H&E stained WSIs along with additional metadata.

PIONEER The European Network of Excellence for Big Data in Prostate Cancer [174] provides a large collection for PCa related data in Europe.

PRIAS The PRIAS (Prostate Cancer Research International: Active Surveillance) dataset [27] provides information of approximately 10,000 early-stage PCa patients under AS.

SEER The Surveillance, Epidemiology, and End Results (SEER) Program [155] contains detailed, cancer-related tabular data of US citizens.

PcBaSe The dataset [21] provides information of the National Prostate Cancer Register of Sweden and contains 110,000 cases with detailed tumor and treatment information.

MIMIC-III The Medical Information Mart for Intensive Care dataset [119] contains information on 112,000 clinical records diagnoses and corresponding procedures performed along with additional patient information like demographics, vital signs or survival information.

TCGA The Cancer Genome Atlas (TCGA) [253] is a public dataset that is frequently used in survival analysis mainly regarding genetic markers of over 20,000 primary cancer sampled for 33 different cancer types with additional survival information and additional patient information. For example, the dataset provides 592 patients with 12 features regarding glioblastoma, 170 patients with 18 features regarding rectal adenocarcinoma and 1,095 patients with 61 features in the context of breast invasive carcinoma.

B Tabular EHR Datasets

Tab. A1: Sample patients and features of the SUPPORT dataset containing age, sex and race, resp: respiratory rate, temp: body temperature as well as sodium- and creatinine level.

z_i [days]	d_i	age	sex	race	resp	temp	...	sodium	creatinine
30	1	83	1	2	16	38.2		19.0	1.1
1527	0	80	1	0	16	38.0		10.0	0.9
96	1	23	1	2	45	37.3		5.2	1.2
892	0	53	1	4	18	36.0		8.7	0.8
					...				
10	1	75	0	2	30	36.2		15.4	0.9
879	0	40	1	3	24	38.0		20.4	1.3

Tab. A2: Sample patients of the METABRIC dataset. MKI67: proliferation marker, EGFR: epidermal growth factor receptor, PGR: progesterone receptor protein, ERBB2: gene marker, ER+: indicator for a cancer with estrogen receptor, HT: hormone therapy, RT: Radiation Therapy, CT: Chemotherapy

z_i [days]	d_i	MKI67	EGFR	PGR	ERBB2	ER+	HT	RT	CT	age
2980	0	5.60	7.81	10.80	5.97	1	1	1	0	57
2872	1	5.28	9.58	10.20	5.66	1	1	0	0	86
4207	0	5.92	6.78	12.43	5.87	1	0	1	0	48
7179	0	6.65	5.34	8.65	5.66	0	0	0	0	67
					...					
5953	0	6.82	5.37	11.65	6.08	1	1	0	0	59
4223	0	5.73	5.45	9.68	6.60	0	1	1	0	61

C DCS

This chapter provides additional information to the DCS model introduced in chapter 4 .

C.1 Hyperparameter Tuning

Tab. A4 shows the best found hyperparameters per model for each dataset. For all discrete-time models, the optimal network architectures tend to be rather shallow. In some cases, some parts of the architecture, namely encoder or decoder parts are dismissed. Notice that DeepSurv and CoxTime almost always have the best performance with shallow neural networks – usually with one hidden layer, even though the search space included up to five. For the aggregation part of the network, one fully connected layer is necessary to map the decoder's output to a single hazard rate h_l. In our approach, we allow additional fully connected aggregation layers for a deeper model architecture. We also analyzed the influence of the total number of output nodes L in hyperparameter tuning. Here we can see incoherent results that might depend on the dataset as well as the temporal spacing of output nodes. In some cases, increasing the number of output nodes yields better results (DCS-linear, DCS-log on SUPPORT and METABRIC), in others (DCS-quant on SUPPORT and METABRIC) less total output nodes were selected.

Tab. A3: Sample patients of the FLCHAIN dataset. The features contain the demographic information (age, sex, sampling year and group information), measurements include the serum free light chain kappa and lambda portion, serum creatinine level and whether the patient was diagnosed with monoclonal gammopathy (MGUS).

z_i [days]	d_i	age	sex	group	year	kappa	lambda	creatinine	mgus
85	1	97	1	10	1997	6	5	1.70	0
1281	1	92	1	1	2000	1	1	0.90	0
69	1	94	1	10	1997	4	4	1.40	0
115	1	92	1	9	1996	2	2	1.00	0
					...				
4982	0	53	1	9	1995	2	3	-	0
3995	0	50	1	4	1998	1	1	0.70	0

Tab. A4: Best hyperparameters for the analyzed datasets and models in chapter 4.

SUPPORT	DeepSurv	CoxTime	DRSA	KAM	DCS-linear	DCS-log	DCS-quant
encoder_num_layers	1	1	0	0	1	2	2
encoder_nodes_per_layer	64	32	-	-	32	128	64
decoder_num_layers			1	1	1	0	1
decoder_nodes_per_layer			128	32	32	-	64
decoder_bidirectional					0	-	1
decoder_use_lstm_skip					0	-	0
aggregation_num_layers				1	1	1	2
additional_aggregation_nodes_per_layer				-	-	-	32
output_grid_num_nodes			280	67	140	140	35
α			0.36				
λ				2.00	0.85	0.25	2.00
σ				1.08	1.55	2.00	1.00

METABRIC	DeepSurv	CoxTime	DRSA	KAM	DCS-linear	DCS-log	DCS-quant
encoder_num_layers	1	1	0	0	2	1	0
encoder_nodes_per_layer	8	8	-	-	128	64	-
decoder_num_layers			1	1	2	1	2
decoder_nodes_per_layer			128	64	64	64	32
decoder_bidirectional					0	1	0
decoder_use_lstm_skip					1	1	0
aggregation_num_layers				1	2	1	1
additional_aggregation_nodes_per_layer				-	32	-	-
output_grid_num_nodes			350	350	700	350	60
α			0.03				
λ				1.94	1.19	1.08	0.25
σ				1.63	0.25	1.45	2.00

FLCHAIN	DeepSurv	CoxTime	DRSA	KAM	DCS-linear	DCS-log	DCS-quant
encoder_num_layers	5	1	1	0	1	0	1
encoder_nodes_per_layer	64	8	64	-	64	-	64
decoder_num_layers			1	1	1	1	0
decoder_nodes_per_layer			16	32	64	128	-
decoder_bidirectional					1	1	-
decoder_use_lstm_skip					1	0	-
aggregation_num_layers				1	1	1	1
additional_aggregation_nodes_per_layer				-	-	-	-
output_grid_num_nodes			42	171	125	42	85
α			0.39				
λ				1.34	2.00	0.67	1.13
σ				0.34	1.00	0.46	0.73

C.2 Loss Ablation

Tab. A5: Performance results for the ablation loss combinations per dataset showing mean and standard deviation of CDAUC and DDC.

dataset	loss	CDAUC	DDC
	$\mathcal{L}_{\text{all}}$	0.650 ± 0.004	0.075 ± 0.019
SUPPORT	$\mathcal{L}_{\text{kernel}}$	0.646 ± 0.007	0.555 ± 0.170
	$\mathcal{L}^{EE}_{\text{kernel}}$	0.558 ± 0.007	1.146 ± 0.127
	$\mathcal{L}_{\text{RPS}}$	0.640 ± 0.003	0.048 ± 0.008
	$\mathcal{L}_{\text{all}}$	0.740 ± 0.010	0.017 ± 0.007
METABRIC	$\mathcal{L}_{\text{kernel}}$	0.675 ± 0.032	0.250 ± 0.126
	$\mathcal{L}^{EE}_{\text{kernel}}$	0.678 ± 0.015	0.080 ± 0.064
	$\mathcal{L}_{\text{RPS}}$	0.731 ± 0.010	0.010 ± 0.004
	$\mathcal{L}_{\text{all}}$	0.812 ± 0.003	0.077 ± 0.012
FLCHAIN	$\mathcal{L}_{\text{kernel}}$	0.707 ± 0.045	1.165 ± 0.225
	$\mathcal{L}^{EE}_{\text{kernel}}$	0.682 ± 0.014	1.430 ± 0.047
	$\mathcal{L}_{\text{RPS}}$	0.791 ± 0.004	0.032 ± 0.006

D MK Dataset

D.1 Patient Characteristics

A detailed overview of the patient characteristics in MK is depicted in tab. A6.

D.2 Proportional Hazards Assumption

Since this chapter uses CoxPH estimations, it needs to be tested if the PH assumption holds true for the individual features as described in sec. 2.5.1. For a feature to pass the assumption, no time-varying effect should be observable. The test results can be found in tab. A7. Note that 7 features, namely `gleason\_path\_percent\_3`, `gleason\_path\_score\_sec`, `lymph\_node\_invasion`, `margin\_status`, `op\_year`, `path\_m\_stage`, `prostate\_volume`, and `seminal\_vesicle\_invasion`, fail the proportionality test. Further, patient age and SVI are also close to a p-value of 0.05.

D.3 Feature Encoding Model

The training and model architecture parameters for the DCS model that was used for the feature encoding analysis is depicted in tab. A8.

Tab. A6: Patient characteristics of the analyzed MK dataset for all patients and for the patients that experienced BCR. For age and prostate volume, median and IQR are shown.

Baseline variable	**bin**	**Overall (n=9,767)**	**With event (n=2,526)**
Age, years		65 (59 - 69)	65 (60 - 70)
prostate volume, mL		25 (20 - 36)	28 (20 - 40)
	<10	70.41% (n=6877)	54.39% (n=1374)
PSA intervals in ng/mL, % (n)	10 - 20	22.05% (n=2154)	30.48% (n=770)
	≥ 20	7.54% (n=736)	15.12% (n=382)
	3+3	0.29% (n=28)	0.04% (n=1)
	3+4	70.15% (n=6852)	39.83% (n=1006)
	3+5	0.40% (n=39)	0.40% (n=10)
	4+3	21.86% (n=2135)	39.94% (n=1009)
RP Gleason Score, % (n)	4+4	0.04% (n=4)	0.16% (n=4)
	4+5	5.39% (n=526)	14.85% (n=375)
	5+3	0.22% (n=21)	0.32% (n=8)
	5+4	1.65% (n=161)	4.47% (n=113)
	5+5	0.01% (n=1)	0.00% (n=0)
	<25	14.53% (n=1419)	36.70% (n=927)
RP Gleason 3 percent, % (n)	25 - 50	15.38% (n=1502)	24.35% (n=615)
	50 - <75	17.35% (n=1695)	17.93% (n=453)
	≥ 75	52.74% (n=5151)	21.02% (n=531)
	<25	55.74% (n=5444)	25.14% (n=635)
RP Gleason 4 percent, % (n)	25 - 50	18.86% (n=1842)	23.36% (n=590)
	50 - 75	18.73% (n=1829)	35.35% (n=893)
	≥ 75	6.68% (n=652)	16.15% (n=408)
	<25	95.60% (n=9337)	88.99% (n=2248)
RP Gleason 5 percent, % (n)	25 - 50	2.89% (n=282)	7.05% (n=178)
	50 - 75	1.13% (n=110)	2.89% (n=73)
	≥ 75	0.39% (n=38)	1.07% (n=27)
	≤ pT2	62.30% (n=6085)	34.05% (n=860)
Pathological tumor stage, % (n)	pT3a	24.05% (n=2349)	31.91% (n=806)
	pT3b	13.44% (n=1313)	33.49% (n=846)
	pT4	0.20% (n=20)	0.55% (n=14)
RP extracapsular extension, % (n)	0	64.40% (n=6290)	38.20% (n=965)
	1	35.60% (n=3477)	61.80% (n=1561)
RP residual tumor status, % (n)	R0	84.04% (n=8208)	72.33% (n=1827)
	R1	15.96% (n=1559)	27.67% (n=699)
RP seminal vesicle invasion, % (n)	0	86.38% (n=8437)	66.07% (n=1669)
	1	13.62% (n=1330)	33.93% (n=857)
	N0	79.20% (n=7735)	67.81% (n=1713)
RP nodal spread, % (n)	N1	10.29% (n=1005)	27.75% (n=701)
	NX	10.51% (n=1027)	4.43% (n=112)
RP lymphatic invasion, % (n)	L0	84.92% (n=8294)	67.06% (n=1694)
	L1	15.08% (n=1473)	32.94% (n=832)
RP metastatic spread, % (n)	M0	99.88% (n=9755)	99.64% (n=2517)
	M1	0.12% (n=12)	0.36% (n=9)

Tab. A7: Individual PH-test results for all analyzed features in the filtered MK dataset. The 7 features that failed the test (p-value below 0.05) are highlighted in bold. Note that patient age and SVI are also close to a p-value of 0.05.

feature	p_value	feature	p_value
age_op	0.087	gleason_path_score_tert	0.514
capsular_invasion	0.373	gleason_path_sum	0.632
gleason_biop_length_3	0.259	gleason_path_volume_3	0.321
gleason_biop_length_4	0.762	gleason_path_volume_4	0.751
gleason_biop_length_5	0.456	gleason_path_volume_5	0.522
gleason_biop_percent_3	0.796	lymph_node_invasion	**0.007**
gleason_biop_percent_4	0.680	lymph_vessel_invasion	0.414
gleason_biop_percent_5	0.909	margin_status	**0.000**
gleason_biop_score_prim	0.566	op_year	**0.007**
gleason_biop_score_sec	0.393	path_m_stage	**0.038**
gleason_biop_sum	0.857	path_n_stage	0.482
gleason_path_percent_3	**0.016**	path_t_stage	0.171
gleason_path_percent_4	0.256	prostate_volume	**0.022**
gleason_path_percent_5	0.111	psa_level	0.745
gleason_path_score_prim	0.125	seminal_vesicle_invasion	0.054
gleason_path_score_sec	**0.018**	tumor_volume	0.336

Tab. A8: DCS model and training parameters that were used for the feature encoding analysis.

	DCS
encoder_num_layers	1
encoder_nodes_per_layer	64
decoder_spacing	quantile
decoder_num_layers	2
output_grid_num_nodes	120
decoder_nodes_per_layer	64
decoder_bidirectional	1
decoder_use_lstm_skip	1
aggregation_num_layers	2
additional_aggregation_nodes_per_layer	64
include_EE_comparisons	1
λ	1
σ	0.7
learning_rate	4.4×10^{-4}
batch_size	50
weight_decay	3×10^{-5}
dropout_rate	0.2

E CI

E.1 Hyperparameter Tuning

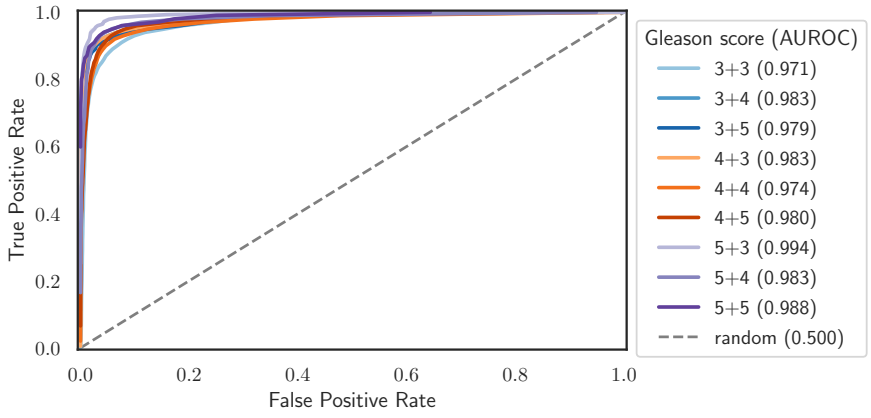
The final parameters for the CI model were obtained using extensive hyperparameter and are shown in tab. A9.

Tab. A9: Hyperparameters for the final CI model.

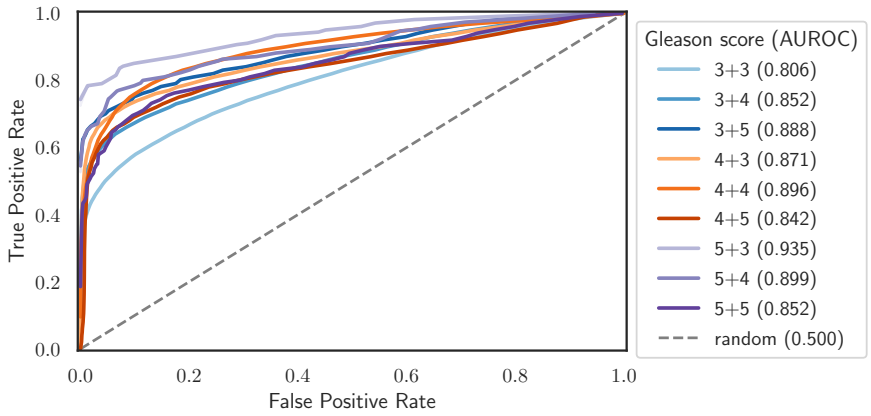
	CI	comment
p_s	256	Patch size
T_{th}	0.1	Min tissue area
C_{th}	0.9	Min cancerous area in tissue region
max_epochs	100	Epochs to train
learning_rate	5.17×10^{-4}	Learning rate
batch_size	256	Batch size
weight_decay	6.8×10^{-4}	Weight decay
dropout_rate	0.5	Dropout rate
early_stopping_on	val/auroc	Metric to use early stopping on
early_stopping_patience	5	Min. number of epochs before metric increase before early stopping

E.2 Performance per Center and Gleason Score

The results for the CI model regarding AUROC per center and split by Gleason score is depicted in fig. A1.



(a) RAD



(b) KAR

Fig. A1: AUROC performance per Gleason score compared individually for both centers.

F PCAI

F.1 Experiment Design and Filtering

The experiment design of PCAI depicted in fig. A2 utilizes training data from UKE.first, UKE.second, and UKE.scanner. For validation, the smaller sub-datasets of UKE.thin, UKE.thick, and UKE.long are added. All external datasets are only evaluated in the test split. Not that splitting was performed on patient level to avoid leakage between the splits.

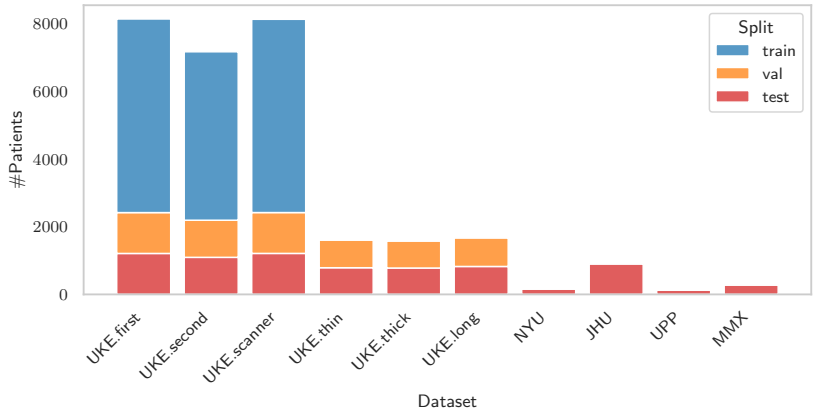


Fig. A2: Data splits for the datasets used for the development and evaluation of PCAI. UKE sub-datasets are used for model training and validation while the external datasets, namely NYU (TMA), JHU (TMA), UPP (biopsies) and MMX (biopsies) are used exclusively for evaluation.

F.2 Patient Characteristics

The detailed PCa patient characteristics of the extracted datasets are depicted in tab. A10.

Tab. A10: Patient characteristics of the extracted datasets showing the number of unique patients, the image type, censoring rate in percent, age (with mean ± standard deviation), median survival and follow-up time in years, the event type classification (BCR=biochemical recurrence, META=metastasis, TRT=any additional treatment, FU=lost to follow-up, PCAD=PCa related death), primary and secondary Gleason grade, ISUP, pathological (TMAs), and clinical (biopsy) T-, N- and M-stage

		UKE	NYU	JHU	UPP	MMX
patients		8157	158	879	123	269
image type		TMA	TMA	TMA	Biopsy	Biopsy
censoring		61.4	70.3	0.3	83.7	88.5
age		63.5 ± 6.1	60.9 ± 7	59.2 ± 6.3	-	67.6 ± 8.9
median survival		1.6	3.9	2	2.1	4.3
median FU		8	17.8	16	7	9.1
ISUP	0	410 (5.03%)	-	-	-	-
	1	1806 (22.14%)	49 (31.01%)	133 (15.13%)	9 (7.32%)	15 (5.58%)
	2	4016 (49.23%)	67 (42.41%)	337 (38.34%)	66 (53.66%)	82 (30.48%)
	3	1367 (16.76%)	16 (10.13%)	184 (20.93%)	27 (21.95%)	80 (29.74%)
	4	109 (1.34%)	11 (6.96%)	123 (13.99%)	12 (9.76%)	36 (13.38%)
	5	449 (5.50%)	15 (9.49%)	102 (11.60%)	9 (7.32%)	56 (20.82%)
Gleason Score	≤ 3+3	2216 (27.17%)	49 (31.01%)	134 (15.28%)	9 (7.32%)	15 (5.58%)
	3+4	4016 (49.23%)	67 (42.41%)	334 (38.08%)	66 (53.66%)	82 (30.48%)
	3+5	37 (0.45%)	7 (4.43%)	28 (3.19%)	1 (0.81%)	10 (3.72%)
	4+3	1367 (16.76%)	16 (10.13%)	185 (21.09%)	27 (21.95%)	80 (29.74%)
	4+4	55 (0.67%)	3 (1.90%)	84 (9.58%)	11 (8.94%)	26 (9.67%)
	4+5	366 (4.49%)	10 (6.33%)	85 (9.69%)	8 (6.50%)	46 (17.10%)
	5+3	17 (0.21%)	1 (0.63%)	10 (1.14%)	-	-
	5+4	82 (1.01%)	5 (3.16%)	15 (1.71%)	1 (0.81%)	8 (2.97%)
	5+5	1 (0.01%)	-	2 (0.23%)	-	2 (0.74%)
Event Type	BCR	3089 (37.87%)	43 (27.22%)	521 (59.27%)	18 (14.63%)	
	FU	5007 (61.38%)	111 (70.25%)	3 (0.34%)	103 (83.74%)	226 (84.01%)
	META	61 (0.75%)	-	142 (16.15%)	2 (1.63%)	42 (15.61%)
	PCAD	-	4 (2.53%)	-	-	1 (0.37%)
	TRT	-	-	213 (24.23%)	-	-
T-stage	≤ T1	2 (0.02%)	-	-	95 (78.51%)	122 (45.35%)
	T2	4966 (60.88%)	104 (65.82%)	134 (15.42%)	26 (21.49%)	90 (33.46%)
	T3	3128 (38.35%)	52 (32.91%)	735 (84.58%)	-	54 (20.07%)
	T4	61 (0.75%)	2 (1.27%)	-	-	3 (1.12%)
N-stage	N0	4306 (86.41%)	56 (35.44%)	700 (80.18%)	-	-
	N1	677 (13.59%)	1 (0.63%)	163 (18.67%)	-	-
	N2	-	-	2 (0.23%)	-	-
	NX	-	101 (63.92%)	8 (0.92%)	-	-
M-stage	M0	6335 (78.47%)	-	509 (60.89%)	7 (5.69%)	79 (29.48%)
	M1	1738 (21.53%)	-	327 (39.11%)	7 (5.69%)	-
	MX	-	-	-	109 (88.62%)	189 (70.52%)

F.3 Results

Comparing BASE to PCAI

The discriminative performance of BASE and PCAI is compared on all TMA datasets in tab. A11 and tab. A12. A similar evaluation is performed for the biopsy datasets in tab. A13.

Tab. A11: Discriminative performance of BASE and PCAI regarding C-index, 3-, 5-, and 7-year AUROC for UKEhv, NYU and JHU.

	model	C-index	AUROC3	AUROC5	AUROC7
UKE.first	BASE	0.645	0.679	0.664	0.687
	PCAI	**0.667**	**0.699**	**0.678**	**0.707**
UKE.second	BASE	0.574	0.561	0.584	0.598
	PCAI	**0.636**	**0.654**	**0.661**	**0.665**
UKE.scanner	BASE	0.595	0.610	0.604	0.638
	PCAI	**0.645**	**0.685**	**0.661**	**0.677**
UKE.thin	BASE	0.560	0.587	0.576	0.568
	PCAI	**0.614**	**0.629**	**0.625**	**0.646**
UKE.thick	BASE	0.573	0.619	0.583	0.564
	PCAI	**0.608**	**0.649**	**0.620**	**0.615**
UKE.long	BASE	0.582	0.594	0.597	0.599
	PCAI	**0.624**	**0.649**	**0.642**	**0.644**
NYU	BASE	0.641	0.631	0.664	0.663
	PCAI	**0.694**	**0.745**	**0.744**	**0.742**
JHU	BASE	0.577	0.611	0.606	0.575
	PCAI	**0.587**	**0.617**	**0.638**	**0.624**

Tab. A12: Discriminative performance of ISUP, GIQ, BASE and PCAI with mean and max aggregation regarding C-index, 3-, 5-, 7-, and 10-year AUROC for the UKE.sealed dataset. The best score per aggregation method for each metric is highlighted in bold.

aggregation	annotation	C-index	AUROC3	AUROC5	AUROC7	AUROC10
	ISUP	0.713	0.735	0.757	0.756	0.734
max	GIQ	0.729	0.756	0.770	0.777	0.753
	BASE	0.700	0.738	0.740	0.728	0.727
	PCAI	**0.739**	**0.763**	**0.781**	**0.788**	**0.779**
	GIQ	0.743	0.772	**0.787**	**0.792**	0.769
mean	BASE	0.712	0.754	0.746	0.739	0.739
	PCAI	**0.744**	**0.778**	0.780	0.783	**0.780**

F.4 Utilizing CI on TMAs from the UKE

The TMAs that were used for training PCAI are not scanned individually, but together on blocks of up to several hundred TMAs. Using the predictions of the CI model it can be shown, that some scanned blocks from the UKE datasets contain TMAs where no or less cancerous TMAs were indicating wrongly labeled data. A next step would be to exclude the TMA spots that received a cancer classification (ISUP1-5), but have a close to zero prediction of containing cancer according to the CI model from the overall training process. The same can be done the other

Tab. A13: Discriminative performance of human ISUP annotations (including A1-A3), BASE and PCAI regarding C-index, 3-, 5-, 7-, and 10-year (only for MMX) AUROC for the UPP and MMX datasets. The best score per dataset for each metric is highlighted in bold.

dataset	annotation	C-index	AUROC3	AUROC5	AUROC7	AUROC10
	ISUP	0.597	0.613	0.659	0.593	-
UPP	BASE	0.581	**0.692**	0.658	0.570	-
	PCAI	**0.604**	0.641	**0.672**	**0.613**	-
	A1	0.838	0.819	0.827	0.857	0.870
	A2	0.834	0.817	0.827	0.851	0.888
MMX	A3	0.641	0.654	0.657	0.652	0.507
	BASE	0.779	0.830	0.832	0.802	0.782
	PCAI	**0.864**	**0.893**	**0.868**	**0.886**	**0.890**

way around by removing all TMA spots with a high cancer predictions, but an assigned label of ISUP0.

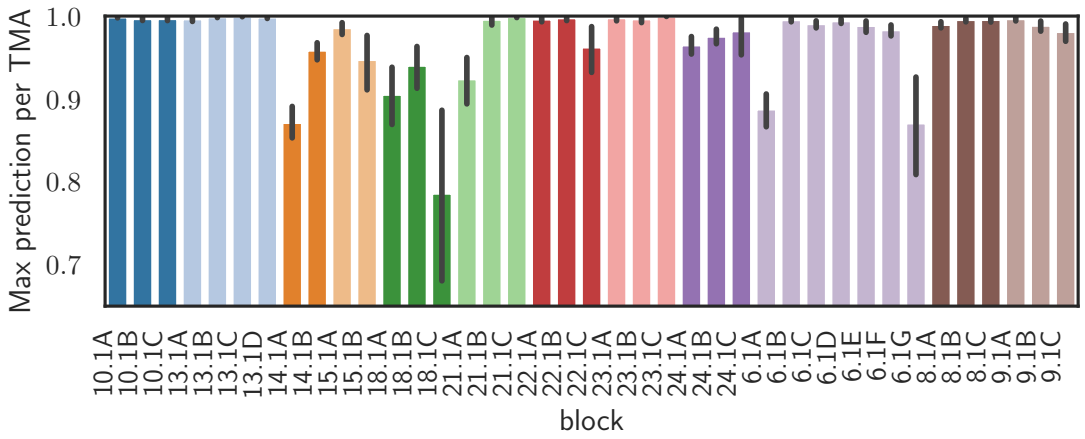


Fig. A3: Maximum cancer prediction per TMA spot for each block in the UKE dataset that was used for training the PCAI model. Where blocks that start with the same number (e.g. 10.1) are highlighted in the same color.