UHH

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Deep learning-based discrete-time survival prediction on prostate cancer histopathology images

Dissertation

submitted to the Universität
Hamburg, Faculty of Mathematics,
Informatics and Natural Sciences,
Department of Informatics, in partial
fulfillment of the requirements for the
degree of Doctor rerum naturalium
(Dr. rer. nat.)

Esther Dietrich

Hamburg, 2022

Date of Submission:
December 5th 2022

Date of Oral Defense:
March 29th 2023

Dissertation Committee:

Prof. Dr. Stefan Bonn (reviewer)
Institute for Medical Systems Biology
University Medical Center Hamburg-Eppendorf

Prof. Dr. Hans-Siegfried Stiehl (reviewer)
Department of Informatics
Universität Hamburg

Prof. Dr. Chris Biemann (reviewer)
Department of Informatics
Universität Hamburg

Prof. Dr. Jan Baumbach (chair)
Department of Informatics
Universität Hamburg

Abstract

In order to enable optimal treatment decisions, physicians can be supported by clinical decision support systems. These aim at providing accurate and objective disease prognoses, e.g., for cancer patients. One of the most prevalent cancers in Germany is prostate cancer. Up to today, prostate cancer severity is, to a large extent, assessed by pathologists from histopathology images by assigning so-called Gleason grades. Since these suffer from high interobserver variability, algorithms for automated tissue analysis have been proposed in the literature. However, these suffer from uncertain and subjective annotations. To eliminate the need for subjective annotations, relapse-free survival times can be predicted as an objective end-point for treatment decision support. Knowing how long a patient lives relapse-free can reduce over- and under-treatment.

Therefore, in this thesis, the topic of automated estimation of relapse times from histopathology images is explored. A literature overview reveals the shortcomings of current approaches to Gleason grade and survival prediction and why a new approach needs to be developed for the given problem statement. In particular, an artificial neural network named eCaReNet (explainable cancer relapse prediction network) is developed, which uses digitized tissue microarray spots extracted after prostatectomy as input and predicts relapse-free survival curves. Multiple datasets are available to train and evaluate the neural network.

In comparison to current state-of-the-art methods, eCaReNet allows for accurate individual survival prediction and outputs biologically reasonable survival curves. It further stratifies patients into up to eight distinct risk groups, in contrast to the usual two to three groups that are stratified in the literature. Also, eCaReNet reaches predictive performance similar to a pathologist, while having access to only a small part of prostate tissue. It is further shown that the pathologist can be outperformed when adding the additional patient parameters prostate-specific antigen (PSA) value, tumor diameter, and volume in addition to the input image.

To integrate a decision support system in clinical workflows, model explainability and robustness to unseen dataset biases are necessary since the variation in the data encountered in a clinic might be greater than in the training dataset. Thus, an ad-hoc explainability is included, which reveals the amount of influence different image regions have on the final prediction. This approach can support pathologists by showing which region to focus on and can build trust in the model's predictions. Furthermore, this work presents an extensive evaluation of the robustness to different data acquisition protocols, which reveals the sensitivity of eCaReNet to dataset biases. An approach for out-of-distribution detection proves proper to assign uncertainty scores to images and decide whether an image is in-distribution or out-of-distribution. It is further proposed to transfer the training dataset color bias to out-of-distribution images with an extension of histogram matching and Macenko adaptation. It is shown that this color adaptation improves results on datasets that mostly include uncertain images.

This thesis provides both a thorough analysis of the state of the art in prostate cancer classification and survival analysis from a multidisciplinary perspective and a proof-of-concept study. It also serves as a starting point for further evaluations on how to obtain robust and accurate survival predictions from histopathology prostate cancer images.

Zusammenfassung

Systeme zur klinischen Entscheidungsunterstützung (vgl. Englisch, “clinical decision support systems”) unterstützen Ärzte und Pathologen im Klinikalltag, indem sie Patientendaten automatisiert auswerten. Sie sind in der Lage objektive Prognosen zu erstellen und Therapieentscheidungen zu unterstützen, beispielsweise bei Krebserkrankungen. Eine der häufigsten Krebsarten in Deutschland ist Prostatakrebs. Um den Schweregrad eines Prostatakrebses zu bestimmen und darauf basierend eine Therapieentscheidung zu treffen, analysiert ein Pathologe einen Teil des Prostatagewebes und klassifiziert es mit dem sogenannten Gleason-Score. Da dieser jedoch sehr subjektiv ist und eine hohe Varianz zwischen Pathologen aufweist, werden derzeit computerbasierte Systeme zur Unterstützung der Gleason Klassifikation erforscht. Allerdings sind diese Systeme dadurch begrenzt, dass sie nur die subjektiven Annotationen von Pathologen lernen können. Um von dieser Subjektivität unabhängig zu werden, wird die rückfallfreie Überlebenszeit als objektiver Endpunkt für eine Entscheidungsunterstützung bevorzugt. Wenn vorhergesagt werden kann wie lange ein Patient überlebt ohne ein Rezidiv zu erleiden, können Über- und Unterbehandlungen von Prostatakrebspatienten verringert werden.

In dieser Arbeit wird daher untersucht, inwieweit eine automatisierte Vorhersage von Rückfallzeiten von Prostatakrebspatienten möglich ist. Mit eCaReNet (explainable cancer relapse prediction network, erklärbares Modell zur Vorhersage von Krebsrezidiven) wird ein künstliches neuronales Netz entwickelt, welches digitalisierte Histopathologie-Bilder verarbeitet, um daraus die Wahrscheinlichkeit eines Prostatarezidivs über die Zeit zu prognostizieren.

Eine ausführliche Untersuchung von eCaReNet zeigt die Vorteile gegenüber derzeitigen Überlebenszeitmodellen auf. Insbesondere ermöglicht eCaReNet eine individuelle, präzise und erklärbare Prognose über einen Zeitraum von sieben Jahren nach der operativen Entfernung der Prostata. Verglichen mit dem derzeitigen Stand der Technik sticht eCaReNet heraus, da eine Risikostratifizierung in bis zu 8 statt der üblichen 2-3 Risikogruppen ermöglicht wird und biologisch realistische Überlebenszeitkurven vorhergesagt werden. Im Vergleich mit der Klassifizierung durch einen erfahrenen Pathologen erreicht eCaReNet eine äquivalente Differenzierung der Patienten. Es wird ebenfalls gezeigt, dass eCaReNets Prognosen durch Zugabe eines zweiten Bildes oder klinischer Daten (PSA-Wert, Tumordurchmesser und -volumen), zusätzlich verbessert werden können. Wenn klinische Daten hinzugenommen werden, übertrifft eCaReNet sogar den Pathologen in seiner Klassifizierung.

Um den Einsatz von Modellen zur Entscheidungsunterstützung in Kliniken zu ermöglichen, sind sowohl Robustheit als auch Erklärbarkeit unerlässlich. Mit einer eingebauten ad-hoc Erklärbarkeit zeigt eCaReNet an, welche Region des Bildes die Entscheidung wie stark beeinflusst. Dies ermöglicht dem Pathologen, die Entscheidungsunterstützung kritisch zu bewerten und zu entscheiden, ob der Prognose vertraut werden kann.

Verglichen mit den zum Anlernen des Modells genutzten Trainingsdatensätzen, können die Daten, die im Klinikalltag verarbeitet werden sollen, vorher nicht beobachtete Variationen aufweisen (zum Beispiel durch Unterschiede in der Gewebeaufbereitung). Da sich die Vorhersagegenauigkeit eines neuronalen Netzes auf unbekannten Daten verringern kann, muss die Robustheit und Generalisierbarkeit genauestens untersucht werden. Eine Untersuchung der Robustheit auf Prostatakrebsdatensätzen mit verschiedenen Protokollen zur Aufbereitung und Digitalisierung von Gewebe zeigt, dass insbesondere unterschiedliche Färbungen durch die Digitalisierung oder dünn geschnittenes Gewebe die Genauigkeit von eCaReNet verschlechtern. Die Unsicherheit der Vorhersage kann mittels eines sogenannten OOD (out-of-distribution, außerhalb der Verteilung) Ansatzes approximiert werden. Es wird vorgeschlagen, Bilder, auf denen keine sichere Vorhersage möglich ist da sie sich außerhalb der Verteilung der Trainingsbilder befinden, durch eine gezielte Änderung der Farbe dem Trainingsdatensatz anzupassen. Eine Kombination aus Anpassung des Histogramms und Macenko Normalisierung trägt dazu bei, dass die Vorhersagegenauigkeit auf Datensätzen mit überwiegend ,,OOD”-Bildern verbessert wird.

Insgesamt präsentiert diese Arbeit eine Machbarkeitsstudie zur Überlebenszeitanalyse von Prostatakrebspatienten auf Grundlage von Histopathologie-Bildern. Sie zeigt eine systematische Herangehensweise auf, um solch ein Vorhersagemodell zu erstellen und zu analysieren. Diese Arbeit soll als Ausgangspunkt zur Erforschung robuster und objektiver Überlebenszeitanalysen dienen.

Contents

List of Figures

List of Tables

List of Abbreviations

AlexNet a CNN architecture for classification

AI artificial intelligence

ANN artificial neural network

AUC (cumulative dynamic) area under the receiver operator curve

AUC^{cd} cumulative dynamic area under the receiver operator curve

AUPRC area under the precision recall curve

AUROC area under the receiver operator curve

BCR biochemical recurrence

C-index concordance index

CAM class activation map

CDOR censoring aware deep ordinal regression (survival prediction model)

ClusterMatch color transformation with histogram and Macenko adaptation including previous training set clustering

CNN convolutional neural network

CT computed tomography

DeepConvSurv survival prediction model that combines a CNN with the Cox model

DeepSurv survival prediction model that combines an ANN with the Cox model

DiagSet public Gleason dataset

DL deep learning

DRE digital rectal exam

eCaReNet explainable cancer relapse prediction network

EfficientNet a CNN architecture for classification

EHR electronic health record

FilterRepr filtering of representative images based on the predicted and annotated Gleason scores as well as the PSA value and relapse time

GAN generative adversarial network

Gleasonaut internal dataset with per-image Gleason annotations

GleasonChallenge public Gleason dataset

GP Gleason pattern

GPU graphics processing unit

GRU gated recurrent unit

H&E hematoxylin and eosin

HR hazard ratio

HSV hue, saturation, and value color space

ID in-distribution

ImageNet public dataset for object classification

Inception a CNN architecture for classification

ipcw inverse probability of censoring weighting

ISUP International Society of Urological Pathology

IQ Gleason integrated quantitative Gleason

KM Kaplan-Meier

LSTM long short-term memory

$\mathbf{M_{Bin}}$ binary survival prediction model trained on Surv1

$\mathbf{M_{ISUP}}$ ISUP classification model trained on the Gleasonaut

MAE mean average error

MobileNet a CNN architecture for classification

MIL multiple instance learning

ML machine learning

MRI magnetic resonance imaging

nlpl negative log partial likelihood

OD optical density

ODIN out-of-distribution detector for neural networks

OOD out-of-distribution

OR odds ratio

PANDA prostate cancer grade assessment (public Gleason dataset)

PCBN prostate cancer biorepository network

PSA prostate-specific antigen

RandHistMatch histogram matching to a random reference image

ResNet a CNN architecture for classification

RGB red, green, and blue color space

RNN recurrent neural network

RPE radical prostatectomy

SICAPv1 public Gleason dataset

SICAPv2 public Gleason dataset

Surv1 internal survival dataset used for training

Surv2 internal survival dataset with tissue from a different prostate area than

Surv1

Surv1AddInfo internal survival dataset with additional patient features

SurvDiff combination of SurvLongStain, SurvThick, and SurvThin

SurvHetero internal survival dataset with multiple images per patient

SurvLongStain internal survival dataset with longer staining time than Surv1

SurvMulti intersection of Surv1 and Surv2

SurvPCBN external dataset

SurvOODCandidates combination of Surv2, SurvDiff, SurvScan, and SurvPCBN

SurvScan internal survival dataset, scanned with a different scanner than Surv1

SurvThick internal survival dataset, cut in thicker slices than Surv1

SurvThin internal survival dataset, cut in thinner slices than Surv1

TCGA the cancer genome atlas (public dataset)

TCGA-PRAD the cancer genome atlas-prostate adenocarcinoma collection (public Gleason dataset)

TMA tissue microarray

TMAZ TMA Zürich (public Gleason dataset)

TNM staging summary of tumor spread, lymph node invasion, and metastases

TRUS transrectal ultrasound

U-Net a CNN architecture for segmentation

UKE University Medical Center Hamburg-Eppendorf

VGG visual geometry group (also a CNN architecture for classification)

WSI whole slide image

XAI explainable artificial intelligence

Chapter 1

Introduction

1.1 Motivation

Computational pathology is a branch of pathology that attempts to automate the disease study of patient specimens (Abels et al., 2019). With the introduction of large tissue scanners and parallel progress in computer vision algorithms, automated analysis of cancer tissue to extract relevant information for decision support has emerged. Using computer-based models to automatically analyze cancer tissue opens up a wide range of opportunities, from increasing objectivity and decreasing time for cancer staging to enabling better treatment decisions and discovering new image features (Abels et al., 2019; Li et al., 2021). One often-studied cancer type is prostate cancer. It affects every 8^{th} male in Germany, and regular screening programs allow detection at a stage when curative treatment is still possible (RKI, 2021; Luiting and Roobol, 2019). In order to evaluate the cancer aggressiveness, pathologists assign Gleason scores to prostate cancer tissue. Since Gleason grading suffers from high interobserver-variability, automated prostate cancer staging models have been developed in the last years to enable objective and accurate Gleason grade predictions (Egevad et al., 2012; Arvaniti et al., 2018; Bulten et al., 2020; Nagpal et al., 2019; Ström et al., 2020). However, an algorithm trained on subjective pathologist annotations can only learn to emulate those. Thus it cannot learn to outperform the pathologist who provided the annotations, which limits such a model's applicability to treatment decision support. Furthermore, the Gleason score considers the size and shape of the glands but does not include all information within the tissue, as it, e.g., neglects the size or shape of nuclei (Egevad et al., 2012). Also, Gleason patterns can stratify patients only into a few discrete groups, leaving no room for individualized predictions. For optimal treatment recommendations, estimation of an individual and objective endpoint, e.g., life expectancy, is preferable (Cheon et al., 2016).

This thesis addresses survival prediction for prostate cancer patients from histopathology images. In particular, an artificial neural network (ANN) to predict patient's relapse-free survival curves after prostatectomy is developed. In the context of this thesis and following the convention in, e.g., Kamran and Wiens (2021), Haider et al. (2020) and Kleinbaum and Klein (2012), a survival curve models the probability that a patient remains relapse-free across time, which is termed survival probability. Digitized prostate cancer histopathology image datasets are available for this task. In clinical practice, such a network should process images obtained from a biopsy to support decisions for or against cancer treatments.

The schematic view of survival prediction with ANNs from patient data in Figure 1.1.1 illustrates possible survival model inputs and the desired output. The left side depicts possible input features, like a histopathology image or the patient's age. These inputs are processed in an ANN to output survival probabilities over time.

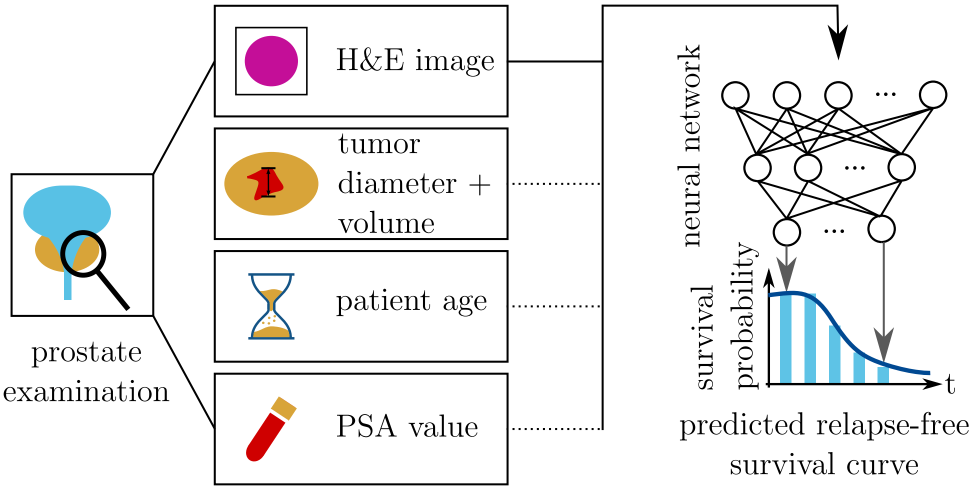


Figure 1.1.1: Schematic workflow for prostate cancer survival prediction with a neural network. As network input, images and clinical patient features can be included. The output is a predicted relapse-free survival curve over time. H&E: hematoxylin and eosin, PSA: prostate-specific antigen.

From a clinical perspective, survival prediction is a meaningful topic since pathologists have to manually analyze and grade prostate cancer tissue today, which is both time-consuming and subjective (Egevad et al., 2012). Furthermore, the staging only indirectly links to patient survival, and if clinicians do directly estimate a patient's survival time, they often overestimate it (Cheon et al., 2016). Healthcare professionals and patients can benefit from better prognoses to avoid over- and under-treatment (Cheon et al., 2016).

From a technical perspective, survival prediction is a relevant task since it has not been explored much on histopathology images. Many researchers simplify survival prediction to a binary task of whether a patient relapses within a given time or not (Huang et al., 2022; Kumar et al., 2017; Yamamoto et al., 2019). However, it is challenging to estimate the relapse probability over time. Difficulties comprise that patients that do not relapse need to be considered. Furthermore, images may not capture all information that affects relapse times. Also, survival model evaluation is complex since the underlying survival probability per individual patient is unknown. With clinical applicability in mind, the model is supposed to be robust toward small dataset biases. Up to today, ensuring robustness is still a challenge in computer vision (Marini et al., 2021a).

This thesis presents a model development path from data preparation over model selection, adaptation, and optimization to generalizability and robustness. An application-oriented perspective is used throughout the thesis since it can only focus on parts of all necessary steps from model development to deployment. In the first step, a model for prostate cancer staging is developed, which predicts Gleason patterns. That model will be extended to a model that predicts accurate survival curves in the second step. In the third step, the model robustness is explored. From data acquisition to model inference, sources of noise, bias, and uncertainty will be revealed. Possible limitations that might influence the performance, such as tissue staining, will be identified and analyzed. The aim is to build a model that is robust to differences in dataset bias. As stated above, the entire search space for a fully-fledged solution toward survival prediction and robustness has to be narrowed to a size feasible for this thesis in accordance with its exploratory nature.

1.2 Research questions

The leading question of this thesis is "How good can a deep learning model predict survival probabilities for a given dataset, and what are the limitations regarding robustness to differences in data acquisition protocols?". It is split into four research questions that will guide this thesis and are outlined in the following:

R1: What is the current state of the art in survival prediction from medical images? A broad overview of current state-of-the-art methods is needed as a starting point for the research in this thesis. It needs to be analyzed how prostate cancer stratification and survival prediction are currently approached in the literature. Further, it is of interest to investigate whether there is a single promising approach that should be used as a model base. This literature research is a starting point to build and improve a survival prediction model.

R2: To what degree can Gleason patterns be predicted accurately in the given dataset of digitized prostate tissue? Currently, a pathologist assigns Gleason scores to prostate tissue for grading. Here, it is hypothesized that a deep learning model can also reach a high accuracy in Gleason grading. To answer this research question, a deep learning model for Gleason classification will be derived from state-of-the-art models, trained, and optimized on the given data. It will be evaluated with different metrics and compared to a pathologist's annotation as ground truth. Therefore, in the best case, the model performs as well as the pathologist.

R3: Can relapse-free survival probability over time after prostatectomy be predicted for individual prostate cancer patients based solely on histopathology images showing part of prostate tissue? For a patient, it is relevant whether a treatment like a prostatectomy is successful and how likely a relapse may occur afterward. The Gleason score, which is purely based on histopathology images, is already highly correlated with prostate cancer relapse but has several flaws. It is hypothesized that an ANN can accurately predict the probability of prostate cancer relapse-free survival over time for individual patients. In order to investigate this research question, a deep learning survival prediction model will be developed, which outputs individual relapse-free survival curves per patient. For evaluation, the predictions will be compared to the ground truth, namely the durations from prostatectomy to biochemical recurrence with respect to calibration and discrimination.

R4: Is it possible to capture the model's limitations in an uncertainty measure and make the model robust toward dataset bias? In order to apply a model in a clinic, it is essential to be aware of cases in which the model cannot provide reliable predictions. Protocols for histopathology image data acquisition are not standardized, wherefore different dataset biases are likely observed in clinical routine. It is hypothesized that it is possible to define a criterion that estimates how certain a survival prediction is, based on the model input. In order to investigate such an uncertainty score, a method to measure the similarity between the training data and new test data during inference will be explored. Furthermore, it will be investigated whether the uncertain predictions are improvable by image adaptation. The evaluation of these methods will be based on the survival prediction metrics.

As a starting point for the research questions, a literature overview of the current state of the art is presented in section 2.2 *State of the art*. It will be explored to which extent these research questions have already been asked and answered, and where current literature is still lacking.

1.3 Thesis outline

The thesis outline is motivated and presented in the following.

Chapter 2: Background In section 2.1 *Medical background*, this thesis is motivated by explaining how prostate cancer is diagnosed presently, stressing the difficulties and drawbacks of current practice. That section emphasizes the clinical relevance of computational pathology models to support pathologists. Afterward, it is described how prostate tissue is extracted, processed, and digitized.

Next, the reader is introduced to the current state-of-the-art machine learning methods for image analysis in section 2.2 *State of the art*. The literature is narrowed down to computational pathology, i.e., automated computational analysis of pathology images. This section explores to which degree the research questions presented above are already discussed in the current literature and where more research is needed, thus addressing R1.

Since this thesis covers methods for prostate cancer grading from histopathology images as well as survival prediction, current approaches for similar tasks are elaborated in section 2.2.3 *Prostate cancer classification for histopathology images* and section 2.2.4 *Deep learning for survival prediction*. Similarities and differences between proposed approaches are stressed. In particular, it is highlighted that no consensus on a common baseline is available, which impedes model development. Furthermore, out-of-distribution detection and bias transfer are introduced as state-of-the-art methods to achieve robust models. Finally, a short overview of explainability methods is given. This chapter aids in putting this thesis's derived models into the current state-of-the-art context.

Chapter 3: Datasets In chapter 3, the datasets used throughout the thesis are introduced. All datasets comprise hematoxylin and eosin (H&E) stained histopathology images. Different internal and external data subsets are introduced in detail, stressing differences and similarities between the sets.

Chapter 4: Gleason grade prediction In order to elaborate on research question R2, different approaches toward automated Gleason grading are introduced in chapter 4. The experiments conducted for this research question and their results are discussed there. The Gleason classification serves as a starting point and preparing task for survival prediction.

Chapter 5: Survival prediction The survival prediction in chapter 5 builds upon the results from the Gleason classification and investigates the research question R3. A theoretical introduction to the problem formulation and the boundary conditions of survival analysis is given. The survival model developed during this work is presented and evaluated in several experiments. It is compared to current state-of-the-art models as well as to a pathologist.

Chapter 6: Robustness For research question R4, the model's robustness to datasets with different biases is addressed. To that end, an out-of-distribution (OOD) detection is proposed to estimate uncertainty. A color transfer method to adjust uncertain model inputs is introduced, evaluated, and compared to a state-of-the-art histogram matching. Further, the advantages of combining color transformation with OOD detection are elaborated.

Chapter 7: Conclusion & discussion The final discussion summarizes this thesis's results and places the presented work in context with the state of the art. The findings for research questions R1 through R4 are elaborated. Open challenges are discussed as a starting point for future work, and an outlook on overcoming these is given.

1.4 Main contributions

The main contributions of this thesis are summarized in the following list.

- Development of a survival prediction model

  A novel ANN named eCaReNet (explainable cancer relapse prediction network) for relapse-free survival prediction is presented in this thesis. The end-to-end trained survival prediction model builds on an InceptionV3 model, which is extended with recurrent layers and performs similarly to a pathologist when using a single TMA spot image per patient as input. The model predictions are explainable by design through the integration of attention-based multiple instance learning. The attention allows insights into how much each image region influences the survival prediction. Explainability is an essential but often neglected aspect of medical image analysis. Compared to state-of-the-art methods, eCaReNet is well-calibrated, shows high discriminative power, and outputs individual and biologically reasonable survival curves for 7 years. The patients can further be stratified into eight distinct risk groups, which allows for more individualized treatment decisions than other approaches, which usually only stratify two to three groups. When additional patient parameters are included, an expert pathologist can be outperformed. The presented eCaReNet is valuable to pathologists and patients as it allows for individual, explainable, and precise survival predictions.

- Robustness

  This thesis extensively evaluates the robustness of survival prediction models. First, the generalizability to datasets with unseen biases is tested. An OOD detection is proposed to estimate the uncertainty of eCaReNet's predictions. It is further proposed to combine OOD detection with a color transfer, which has not been done in the literature before. To this end, a novel and auspicious approach for color transfer is introduced, which extends histogram adaptation and Macenko normalization. The proposed combination of color transformation and OOD detection improves prediction performance on datasets with a significant color bias. To this end, experiments on unique datasets that emulate different data acquisition protocols are conducted.

Chapter 2

Background

In the following section, an introduction to prostate cancer epidemiology, prostate cancer detection, and treatment options is given. The goal is to provide all relevant clinical background information for the scientific problem depicted in this thesis and to motivate the necessity for computer-assisted prostate cancer assessment.

Further, to provide a thorough overview of what research has already been conducted in the field of computational pathology, which questions have already been asked, and which answers have already been found, a brief literature overview on the current state of the art in computational pathology is presented. First, common concepts of deep learning and its application to computer vision are introduced along with state-of-the-art neural network architectures. This is then narrowed down to computational pathology, where existing approaches and concepts are analyzed. Next, the literature on Gleason grade prediction from histopathology images is reviewed, stressing the great variability in models and datasets. Further, algorithms for recurrence-free survival prediction on medical images are presented. Different approaches are contrasted, and the drawbacks of current methods are stressed, which motivates the development of a novel ANN model in this thesis.

Finally, it is evaluated how robustness against domain shifts can be achieved and how these domain differences between a training dataset and new datapoints can be detected. A short overview of explainability completes this chapter. The goal of this chapter is to provide the reader with an overview of the current status of computational pathology and the remaining, unsolved challenges in this field. The research conducted in the work for this thesis will be derived from thie presented state of the art.

2.1 Medical background

2.1.1 Prostate cancer

This section motivates the thesis topic from a medical perspective. An introduction to prostate cancer epidemiology is given, followed by an overview of current diagnosis and treatment-decision standards. Furthermore, the shortcomings of current grading systems are highlighted and the data acquisition process is described.

Epidemiology and screening In 2020, around 68,000 men were newly diagnosed with prostate cancer in Germany, which makes it the second-most frequent cancer after breast cancer and the most frequent cancer in men (Ferlay et al., 2020). According to cancer-related deaths, however, prostate cancer only ranks fifth. The lifetime risk of developing prostate cancer was 12.1% in Germany in 2018, with a mortality risk of 3.3% (RKI, 2021). Since prostate cancer progresses slowly, curative treatment is possible over a long period. Therefore there is a cancer screening program that, in Germany, enables every man starting at the age of 45 years to have one prostate cancer screening each year. The screening includes a digital rectal exam (DRE), whereas the prostate-specific antigen (PSA) value is not measured free of charge in Germany. PSA is a protein produced in the prostate glands and can be measured in the blood (Luiting and Roobol, 2019). In the case of a prostate tumor, the PSA value is increased. However, an increased PSA value does not automatically imply the presence of prostate cancer but can have different, non-cancer-related reasons, like an enlarged prostate. Thus if an increased PSA value is measured, the reason for the increase needs to be further investigated (Sohn, 2015). Therefore, a raised PSA value can hint toward cancer early but at the same time may lead to unnecessary biopsies that may cause side effects or further complications, like incontinence (Sohn, 2015). The significance of the PSA value, and whether it should be included in the screening program, is highly debated in the literature (Andriole et al., 2009; Schröder et al., 2009). Since different studies could not confirm whether regular PSA screenings improve prostate cancer treatment, several guidelines exist until today. A PSA value of $4\ \frac{\mathrm{ng}}{\mathrm{ml}}$ or above is often considered critical. However, this threshold is not defined clearly and consistently across guidelines (Luiting and Roobol, 2019).

Diagnosis and treatment options A patient's possible path from diagnosis to treatment is depicted in Figure 2.1.1. The prostate is shown in orange, below the blue-colored urinary bladder, and surrounding the urethra (Sotelo, 2015). If a male is suspected to have prostate cancer, for example, due to a suspicious DRE or an increased PSA value, a transrectal ultrasound-guided (TRUS) biopsy may be performed to confirm the suspicion (Luiting and Roobol, 2019). Also, magnetic resonance imaging (MRI)-guided biopsies are possible, which allow a

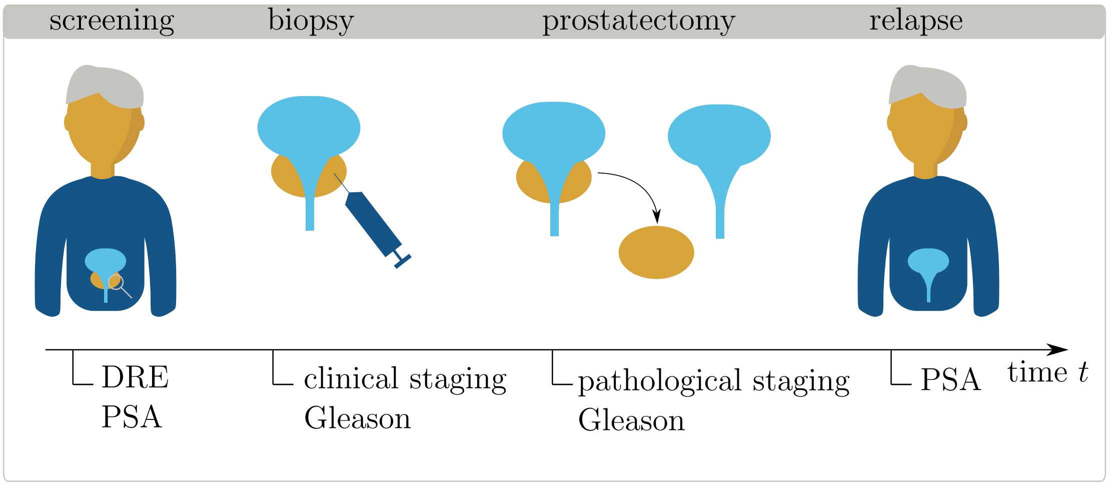


Figure 2.1.1: Possible treatment process of a prostate cancer patient: After a suspicious prostate cancer screening including, e.g., a digital rectal exam (DRE) or prostate-specific antigen (PSA) value measure, a biopsy is performed. Based on the clinical staging, it is decided that a prostatectomy is necessary. After prostate removal, the tissue is examined again, now with the pathological staging to decide on further treatments. The prostate cancer possibly spreads again, which can be detected by an evaluated PSA level (Heidenreich, 2007).

more precise localization of suspicious areas during a biopsy (Luiting and Roobol, 2019). Up to 12 tissue probes are extracted via needle biopsy, and a pathologist examines the tissue sample visually for clinical staging. As a histologic grade, a Gleason score is assigned to the extracted tissue, which is described in detail later. Furthermore, the tumor amount is quantified, for example, by counting the number of malignant cores (Grignon, 2018). Based on the severity estimate, a treatment can be suggested. Options include radical prostatectomy (RPE, i.e., removal of the prostate), radiotherapy, hormonal therapy, or active surveillance, which means no immediate action is taken but the cancer development is monitored closely (Heidenreich, 2007). A treatment like an RPE comes with many possible side effects, such as incontinence and impotence (Rondorf-Klym and Colling, 2003). Hence, overtreatment should be avoided. Loeb et al. (2014) review multiple studies and find that 1.7 % to 67 % of prostate cancer patients are overdiagnosed, which may lead to overtreatment. If the prostate is removed, the extracted tissue is examined to obtain a more thorough impression of the tumor and decide on further treatment (Grignon, 2018). The pathological staging of prostate cancer includes an examination of resection margins, again a histologic grading with Gleason scores, and a TNM staging that includes information on whether the tumor is organ-confined or invades adjacent structures (T, tumor), examination of lymph nodes

(N, nodes), and extraprostatic extension (M, metastasis) (Grignon, 2018). After treatment, the cancer progress is monitored regularly by measuring the PSA level, because the cancer can spread again (Heidenreich, 2007). A biochemical recurrence (BCR) is defined by a rise in PSA level after RPE (Lobel, 2007).

2.1.2 Gleason patterns

In order to classify tumor severity, a pathologist analyzes the appearance of a prostate cancer tissue sample and annotates it with Gleason patterns. Gleason and Mellinger (1974) developed the Gleason grading system that stratifies prostate cancer patients into risk groups based on the glands' architectural pattern. Gleason patterns can be assigned to tissue extracted from a biopsy as well as from an RPE. Different alternatives to stratify risk groups based on the Gleason patterns exist. These are described in the following and summarized in Table 2.1.1.

Gleason grade Gleason patterns stratify prostate tissue into grades ranging from 1 to 5, where grade 1 and grade 2 patterns are classified as no or benign cancer, 3 consists of slightly abnormal glands, 4 is a more malignant cancer, and 5 is the most severe cancer type. Healthy prostate tissue consists of round and regular glands. With higher cancer grades, the glands get more cribriform and ill-defined, and the glandular differentiation decreases (Egevad et al., 2012). This is illustrated in Figure 2.1.2.

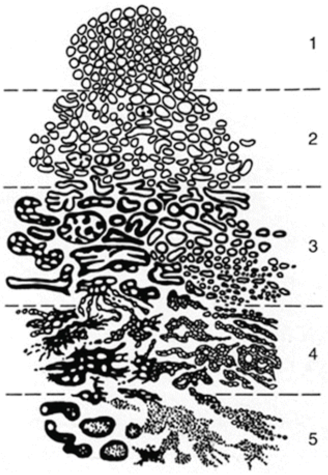


Figure 2.1.2: Gleason grades 1-5 based on the glandular architecture. Reprinted by permission from Springer Berlin Heidelberg, Encyclopedia of Cancer, Gleason grading, Furihata and Takeuchi (2017).

Gleason score In order to grade prostate tissue, a Gleason score is assigned, which generally consists of two Gleason patterns. For tissue extracted after RPE, the most pervasive and second most pervasive Gleason patterns are assigned. Thus, tissue with mostly Gleason pattern 3 and partly Gleason pattern 4 is assigned a Gleason score 3+4. Sometimes, besides the two most common patterns, a small amount of severe pattern 5 is visible; thus, a tertiary pattern is assigned as 3+4 Tert.5 (Egevad et al., 2012).

For tissue obtained from biopsies, the most common and most severe patterns are assigned (Egevad et al., 2012). Thus, if a patient has mostly Gleason pattern 3 but also patterns 4 and 5, the Gleason score is 3+5. Note that no tertiary patterns are assigned to biopsies. Gleason patterns 1 and 2 are typically not assigned anymore to biopsies because they have shown a low correlation with the grading after RPE (Gordetsky and Epstein, 2016).

If the tissue contains only a single pattern, for example, grade 3, this is noted as 3+3 in both cases for biopsy and after RPE. The Gleason score can also be conflated into the sum of both assigned Gleason patterns, leading to a score between 1+1=2 and 5+5=10 (Epstein et al., 2005). Since grades 1 and 2 are not commonly applied, but considered benign, the Gleason scoring for cancerous tissue in practice only ranges from 6 to 10 (Epstein et al., 2016).

ISUP score In most cases, the Gleason score is a valid choice since it shows a great correlation with the time to cancer-related death (Egevad et al., 2002). However, some drawbacks led to the agreement on a new grading system at the 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading (Epstein et al., 2016). Gleason sum 7, e.g., neglects important differences between 3+4 and 4+3, where either pattern 3 or pattern 4 is more pervasive. Chan et al. (2000) showed that the patients with a Gleason sum of 7 are heterogeneous with differences in disease progression and treatment suggestions. Therefore, it is suggested to keep these groups separate. The proposed ISUP scoring system ranges from 1 ($= 3+3$) to 5 ($= 4+5$ / $5+4$ / $5+5$). It is referred to as either ISUP score or Gleason grade group and broadly applied (Epstein et al., 2016; Egevad et al., 2012). The relationship between ISUP score, Gleason sum, and Gleason score is shown in Table 2.1.1.

Quantitative scoring Sauter et al. (2018) developed a more fine-grained system, called the integrated quantitative (IQ) Gleason. The aim is to take into account the Gleason pattern quantities instead of only their presence. Further,

Table 2.1.1: Relationship between Gleason score, Gleason sum, and ISUP score.

Gleason score	3+3	3+4	4+3	4+4	3+5	5+3	4+5	5+4	5+5
Gleason sum	6	7		8			9		10
ISUP score	1	2	3	4			5		

they state that small amounts of Gleason grade 5 already have a large impact on relapse-free survival and should therefore have a higher impact on the grading. To obtain the IQ Gleason, first, the percentages of Gleason patterns (GP) 4 and 5 in the prostate tissue are estimated and added together. If any Gleason pattern 5 is present, 10 points are added to the score, and another 7.5 points are added if there is more than 20 % of Gleason pattern 5. This results in the overall IQ Gleason score ranging from 0 (for 3+3) to 117.5 (for 5+5):

$$\text{IQ} = \%GP4 + \%GP5 + \mathbf{1}_{\%GP5>0} \cdot 10 + \mathbf{1}_{\%GP5>20} \cdot 7.5$$

Here, **1** is the indicator function taking value 1 if the condition is true, 0 otherwise. In their work, Sauter et al. (2018) show that the survival times between IQ Gleason groups stratify well when categorizing the score into 10 distinct risk groups.

Summary All proposed systems rely on Gleason patterns. However, an important drawback of these is that the Gleason pattern is a subjective score and varies considerably among physicians (Epstein, 2018). According to Egevad et al. (2012), interobserver agreement varies between 36 % and 81 %. This is partly because some Gleason patterns are at the margin between two patterns and there are no clear distinctions. Pathologists might not even be able to reproduce their own score, which is reflected in an intraobserver variability of 43 % to 78 % (Egevad et al., 2012). In a recent study, van der Slot et al. (2021) indicate that the reproducibility for the IQ Gleason is lower than for Gleason grade groups. Furthermore, Gleason patterns only take into account the architecture of the glands in the tissue and neglect other features like nuclei size and number, which might also impact the disease status (Egevad et al., 2012). Meyer et al. (2022) conducted an experiment that shows that the accuracy of pathologists' Gleason grading increases when being provided with (correct) artificial intelligence (AI)-based Gleason grade suggestions. This motivates the development of a computer-aided decision support system in this thesis.

2.1.3 Dataset acquisition

Computer-aided decision support systems for prostate cancer stratification and Gleason scoring require digitized histopathology images. The procedure to obtain a digital image from prostate tissue is depicted in Figure 2.1.3. Three possible image sources are included, namely tissue microarray (TMA) spots, prostatectomies, and biopsies. The procedure for TMA spots is described in simplified terms in the following, the workflow for whole prostate tissue and biopsies is similar.

A patient that is believed to have malignant prostate cancer is treated with an RPE, which means that the prostate (shown in orange) is removed. With a hollow needle, a tissue core of the (cancerous area of the) excised prostate is extracted. In a paraffin block, multiple cores are arranged such that the block now shows spot-shaped samples in a grid-like order (Parsons and Grabsch, 2009). The paraffin block is cut into 1-10 µm thin sections (Mescher, 2013). A section is stained with hematoxylin and eosin (H&E staining) sequentially as preparation for light microscopy. H&E is a commonly used coloring method to visualize morphology, resulting in a pink to violet tissue stain. Hematoxylin binds to acidic structures so that the cell nuclei are colored dark blue or purple. Basic components are stained

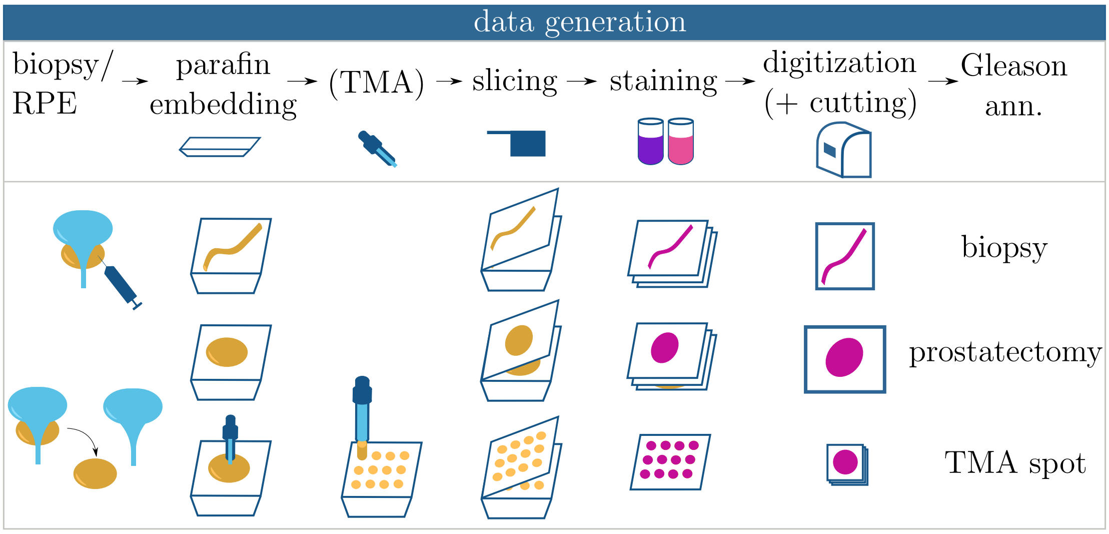


Figure 2.1.3: Overview of data acquisition. *Biopsy and prostatectomy*: The tissue is embedded in a paraffin block and sliced completely. Each slice is stained and digitized afterward. Therefore, the whole prostate or the whole biopsy tissue can be examined. *TMA spot*: From the prostate, only a small tissue core is extracted. This is embedded in a TMA block with cores from different regions of the same prostate and from different patients. Slices are cut from the block, stained, and digitized. Single TMA spots are cut, which are usually obtained for research purposes. RPE: radical prostatectomy, TMA: tissue microarray, ann.: annotation.

pink by eosin (Mescher, 2013). A pathologist usually analyzes the tissue with a microscope to assign Gleason grades. In order to obtain a digitized version of the prostate tissue, the cut and stained section is digitized with a scanner. For a treatment decision, pathologists analyze either slices of the complete prostate or multiple biopsy cores and assign Gleason grades per patient. Single TMA spot annotations are more commonly used for research and teaching purposes (Simon et al., 2004).

Sources of variation In each of the data acquisition steps, artifacts or color variations might be introduced (Wright et al., 2021). Because each TMA is stained separately, the spots of different TMAs may vary in color and intensity. Staining differences occur due to differences in imaging protocols, e.g., staining times, tissue thickness, quality of the reagents, or errors made during the staining process (Chlipala et al., 2021; Janowczyk et al., 2017; Rolls et al., 2008; Wright et al., 2021). Tissue storage also influences the quality since the tissue stain fades with time and light exposure (Azevedo Tosta et al., 2019; Rolls et al., 2008). Chlipala et al. (2021) evaluate differences in staining based on protocol, tissue thickness, day of staining, and reagent quality. They find, e.g., that staining intensity rises with tissue thickness, but the influence is higher in eosin staining compared to hematoxylin. They conclude that manual quality control of the staining is essential to assure high-quality data. A detailed list of sources for artifacts and staining differences can be found in Rolls et al. (2008). Furthermore, differences in scanners may lead to variations in image color intensity or contrast. Rajaganesan et al. (2021) state that not all scanners reproduce the original color of glass slides, but some produce more basophilic or eosinophilic images. Also, they state common artifacts are out-of-focus images.

2.2 State of the art

The following sections give an introduction to the relevant state of the art in order to assess whether current approaches suffice for robust survival prediction and, if necessary, whether already developed models can be extended.

First, an introduction to image analysis with neural networks, in general, is provided. Then, the focus is put on computational pathology, narrowing the view to the current state of the art in prostate cancer stratification and survival prediction. The most commonly used and recent neural network models and approaches will be stressed, while always narrowing the literature on cases similar to this thesis' problem formulation and dataset source. At the end of this section, an introduction to robustness and explainability is given.

2.2.1 Computer vision with neural networks

Computer vision refers to the automated extraction of meaningful information from a digital image using a computer. It can be used, for example, to extract information about where objects are present in an image (detection), to which class an object or image belongs (classification), and which exact pixels an object covers (segmentation).

Deep learning (DL), which is a sub-field of machine learning (ML) and refers to the application of deep neural networks, enabled achieving high accuracy on complex computer vision tasks (O'Mahony et al., 2020). Therefore, the following sections give an introduction to neural networks for computer vision, omitting traditional methods. A comparison of both approaches can be found, for instance, in O'Mahony et al. (2020). It is assumed that the reader is familiar with basic concepts of neural networks, such as neurons, layers, activation functions, and backpropagation. Therefore in the following, the main characteristics of networks designed for computer vision are introduced only shortly. For further background knowledge, the reader is referred to Verdhan (2021).

Introduction to convolutional neural networks Automated image analysis has advanced significantly with the introduction of convolutional neural networks (CNNs). CNNs incorporate convolutional layers, which, in contrast to fully connected layers, are able to spatially exploit the local context of data, which is beneficial for interpreting image data (Alzubaidi et al., 2021). In the convolutional layer, the learned weights are part of filters, which are trained to extract spatial features in the image, like edges or curves (Verdhan, 2021). The filters may have different sizes or scales, which allow for controlling the spatial dimension of extracted features. The network's different consecutive layers detect features of increasing complexity. That means the first layers detect low-level visual structures like edges whereas the later layers detect more complex structures up to objects (Verdhan, 2021). After convolutional layers, pooling layers can be included for dimensionality reduction in image width and height. In these layers, neighboring pixels are combined for example through averaging or using the maximum value (Verdhan, 2021). Eventually, the dimension is reduced to a vector, and fully connected layers perform the final classification task. Similar to fully connected neural networks, CNNs are trained for several iterations (so-called epochs) with a pre-labeled dataset, using the annotations as ground truth and a loss function to update the weights of the connections between neurons in neighboring layers via backpropagation. The filters are convolved with the image (i.e., each filter moves across the image row-wise), while the pooling layers do not have trainable weights. Both these operations reduce the number of trainable parameters (Alzubaidi et al., 2021). With increasing network depth across hidden layers, the image height and width are reduced, while typically increasing the depth (which starts with 3 channels for RGB images).

For image analysis tasks across applications (e.g., healthcare, manufacturing, autonomous driving) and tasks (e.g., image classification, detection, segmentation), CNNs are currently the most prevalent models (Voulodimos et al., 2018). Only in recent years, researchers have started to apply transformer networks to image analysis, which were originally developed for natural language processing (Vaswani et al., 2017; Dosovitskiy et al., 2021; Hu et al., 2021a). While transformers usually require even larger datasets and computational resources for training, the limited availability of annotated high-quality data for training, validating, and testing is a key problem in medical applications.

Common CNN architectures Neural network hyperparameters like the number of convolutional layers, pooling layers, or filter sizes may be combined in various ways, which leads to a great number of existing CNN architectures that have been applied to different tasks. However, some models have shown greater improvements than others through the introduction of novel layers or layer combinations.Since it is impossible to expand on all architectures here, in Table 2.2.1 only some of the most commonly applied CNNs for classification tasks are summarized. Most of these networks were implemented for the classification of objects. They vary in the number of hidden layers and consist mostly of convolutional, pooling, and fully connected layers. The architectures mentioned in that table have been among the most frequently used architectures, at least over a period of time, or up until today (Zhang et al., 2021a).

Krizhevsky et al. (2012) developed AlexNet, a CNN with blocks of convolutional layers and max-pooling, followed by fully connected layers. It was the first deep network to beat a traditional approach on the ImageNet large scale visual recognition challenge in 2012 (Russakovsky et al., 2015; Zhang et al., 2021a). The visual geometry group (VGG) (Simonyan and Zisserman, 2015) introduced deeper architectures, of which the VGG-16 and VGG-19 are nowadays still broadly ap-

Table 2.2.1: Overview of commonly applied CNN architectures. Number (#) of parameters according to the original paper or Chollet et al. (2015) in million (M).

Architecture	**Special feature**	#
AlexNet Krizhevsky et al. (2012)	one of first deep networks	60M
VGG-16 Simonyan and Zisserman (2015)	smaller filter and pooling size	138M
InceptionV3 Szegedy et al. (2016)	inception block: processing with different kernel sizes in parallel	24M
ResNet-50 He et al. (2016)	skip connections between layers	26M
MobileNet Howard et al. (2017)	spatially separable convolutions	4M
EfficientNetB0 Tan and Le (2019)	optimize width, depth, and resolution interdependently	5M

plied. The numbers refer to the number of layers, which are again convolutional, pooling, and fully connected layers. The main difference to AlexNet is, besides the depth, that they use smaller filter sizes.

Novel layers and architectural patterns were implemented over time. Szegedy et al. (2015) introduced an architecture called Inception, which includes “inception” blocks, which process the image with differently sized convolutional filters in parallel. These blocks enable the extraction of features of different sizes. Their updated version InceptionV3 includes batch normalization, which Ioffe and Szegedy (2015) introduced earlier (Szegedy et al., 2016). Instead of only normalizing the input image once, batch normalization normalizes the intermediate output over a batch during training.

He et al. (2016) found that larger architectures do not by definition lead to better results, which is why they introduced residual networks (ResNet) with so-called skip connections that allow weights to skip layers. They also presented different versions with a varying number of layers, e.g., ResNet-50 or ResNet-152.

Since larger networks lead to higher computational complexity in terms of time and memory, and an increased number of parameters to be learned, Howard et al. (2017) introduced MobileNet as a lightweight alternative. The main advantage lies in depthwise differentiable convolutions. For those, the convolutional layer is split into two parts, one for filtering per channel and another for the combination across channels.

When neural networks are expanded to increase complexity, their width (influenced by the filter size), depth (number of hidden layers), or input image size (resolution) might be adapted. However, Tan and Le (2019) point out that only changing a network’s depth, width, or resolution is not beneficial but instead that these parameters are linked. With EfficientNet, they introduce another architecture, which increases width, depth, and resolution interdependently and comes in different sizes (EfficientNetB0 through B7).

For every new task, the best suitable architecture has to be determined and may be adapted further, since each architecture has advantages, which the given problem may benefit from. Besides the given task and dataset, computational resource limits or inference speed might be a critical factor in this decision.

Training and transfer learning The weights of a neural network are not set manually but learned through training after random initialization. This data-driven approach allows discerning complex patterns, which would be difficult to impossible to describe with pre-defined rules. Also, the diverse appearance of objects can be encompassed. A large training dataset is necessary for the neural network to learn the characteristic patterns for a task like classification. A fixed lower limit for the training set size cannot be defined generally for neural network tasks. The training, validation, and test set sizes depend on the complexity of the given problem and dataset. Zhou et al. (2020) train a CNN on 163 samples, while Raghu et al. (2019) already call a dataset with 5,000 training samples "very small".

Available datasets, especially in the medical domain, are often too small to train a network with millions of parameters. To overcome this limitation, a network's weights can be pretrained on a larger dataset, which may be unrelated to the target domain and task. Large datasets are, for example, commonly available for images of everyday objects. Even though the image content in such images and in medical images differs significantly, basic structures like edges need to be recognized in most image classification tasks. Pretraining the network on one dataset, and fine-tuning it on another dataset is also known as transfer learning (Raghu et al., 2019). It is possible to either re-train all model weights or just the last layers since the low-level visual structures might not differ much. However, this needs to be considered on a task-specific basis.

Transfer learning has become a standard in medical image analysis. Raghu et al. (2019) show that it has a positive effect when re-training large models on a small medical dataset, however, the effect decreases when using larger datasets or models with less parameters. Still, pretraining a model on a large dataset is time- and resource-intensive. However, many commonly used architectures are available open-source with weights pre-trained on a large dataset. A prevalently-used dataset for pre-training of weights is ImageNet, which contains images of real-world objects (Deng et al., 2009). It contains millions of images, is designed for classification into 1000 classes, and often serves as a benchmark for the evaluation of new network architectures. These pretrained networks enable a quick adaptation to a new task and image class by fine-tuning on new datasets.

Supervised, unsupervised, and semi-supervised training In the case that the ground-truth classes are known and an annotation exists for each sample in the training dataset, supervised learning can be applied. That means that a direct comparison between network output and annotation/class label in terms of an error metric incorporated in the loss function (Alzubaidi et al., 2021) is possible. In contrast, unsupervised learning does not use any annotations but the goal is to find, e.g., clusters in the training data based on common features (Alzubaidi et al., 2021). A mixture of both is weakly supervised learning. In that case, only coarse/weak annotations are available, e.g., one annotation for a whole image instead of annotations per image region (Otálora et al., 2021). Weakly supervised learning is often applied in medical image analysis when whole slide images (WSI) are available. WSIs show large tissue areas, like a slice of the complete prostate or biopsy core, and typically much background. If these large images (up to more than 100,000 × 100,000 pixels in size) do not fit into memory, they are cut into smaller pieces (patches or tiles) in order to process them in a neural network (Lu et al., 2021; Campanella et al., 2018). If a patient has cancer, usually not the whole tissue is affected. Still, often the whole image is labeled as, for example, cancer, while not every single pixel or patch shows cancer. Therefore, it is not possible to classify the tiles separately.

For this particular problem, multiple instance learning (MIL) was introduced (Dietterich et al., 1997). In MIL, an image is cut into smaller pieces and treated as a bag of patches. The whole bag is processed at once with a neural network, and for the final classification, the results per patch are averaged over the whole bag. In a binary setting, a bag is positive if it contains at least one patch that is labeled positive. Thus a bag is only negative if all its patches are negative. Instead of averaging over all instances, Ilse et al. (2018) add attention to MIL through which the model is trained to automatically weigh the single instances according to their importance. The MIL layer provides an explanation for the prediction by revealing which instance influenced the result by how much. Since the single instances are assumed to be independent of each other (Rymarczyk et al., 2021) introduced self-attention. They leverage the dependencies in-between single instances. More details for both MIL and self-attention are provided in section 5.6 *eCaReNet*.

Recurrent neural networks For image sequences as input or time-dependent outputs, recurrent neural networks (RNNs; Rumelhart et al., 1985) can be used. The dimension of the input or output layer is extended with an additional time dimension. Nodes in the recurrent layer process not only the current input but can also store information from previous time steps in hidden states. Since it is known that backpropagation learning lead to stability problems like exploding or vanishing gradients, advanced architectures have been introduced, including long short-term memory cells (LSTMs; Hochreiter and Schmidhuber, 1997) and gated recurrent units (GRUs; Cho et al., 2014). RNNs can be used in combination with CNNs, e.g., for (per-frame) video classification (sequence of images to class) or image captioning (image to a sequence of words) (Yue-Hei Ng et al., 2015; Mao et al., 2015).

2.2.2 Computational pathology

"Digital pathology" refers to tissue examination, archiving, and reporting in a digitized form (Abels et al., 2019). Azam et al. (2021) show in their study that pathologists' conclusions do not differ significantly when using digitized images compared to direct examination with light microscopy. Since manual image analysis is time-consuming, repetitive, and suffers from high inter-observer variability, computational analysis has been proposed in recent years (Li et al., 2021). The field of "computational pathology" focuses on the computational analysis of patient specimens, like tissue, to study a disease or support decision-making (Abels et al., 2019). Especially with the introduction of whole slide image scanners that allow the digitization of large tissue regions in high resolution, computational pathology applications have emerged. These have further benefited from the advances in computational power (Hanna et al., 2020).

On histopathology images, neural networks can be used for classification (e.g., the decision whether tissue is cancerous; Xu et al., 2017; Campanella et al., 2018), localization (e.g., of cell nuclei; Zhou et al., 2019), quantification (e.g., lymph node quantification; Hu et al., 2021b), or more complex tasks like clinical decision support systems for risk prediction (Wulczyn et al., 2020; Fan et al., 2021). Besides developing models for clinical usage, computational pathology methods can also be used for basic research: Integrating multiple data sources (e.g., genomics and images) can, for example, uncover novel biomarkers (Hanna et al., 2020).

As stated above, deep learning (DL) algorithms have improved performance for many image classification tasks over the last decade. Also in histopathology image analysis, DL models improved the robustness over traditional computer vision methods (Abels et al., 2019). This is especially due to the high complexity of histopathology images, which hinders manual feature extraction (Banerji and Mitra, 2022).

Despite the fast development of research in the field of computational pathology, translation to the clinics has been slow (Rakha et al., 2021). According to Rakha et al. (2021), the main reasons are the gap between development and clinical environment (e.g., more data variation in clinical practice than during model training), missing explainability, and the disruption of current workflows.

Neural networks require large amounts of (annotated) data for training. Photographs from everyday scenes can be searched for and downloaded directly from the internet (Deng et al., 2009). However, access to large histopathology datasets is limited since public histopathology datasets for research usage are rare. Acquiring sufficiently large in-house datasets is also difficult and laborious (Abels et al., 2019). One of the main reasons why digital histopathology images are often not readily available is that during clinical routine tissue is often only examined under the microscope and not digitized (Williams et al., 2018; Nam et al., 2021). According to Williams et al. (2018), in 2018, 58.8 % of institutions in the UK did not produce any digital slides. Furthermore, digital slides are mostly used for teaching, research, or quality assurance, not for diagnosis. Nam et al. (2021) published a study from 2020 showing that only 26 % of Korean pathologists are using digital pathology systems but 78 % see a need for using these.

If digitized histopathology datasets can be collected, data protection, ethics, and privacy concerns need to be considered (Abels et al., 2019). Data needs to be anonymized or pseudonymized, while patients must consent to the data usage (Heesen et al., 2020). Furthermore, a dataset from a single hospital might not include enough (diseased) patients to train a data-driven model. Larger datasets are available when analyzing very common diseases or accessing data from a specialized clinic. Merging images from multiple data sources yields larger but possibly inconsistent datasets since data acquisition protocols are not standardized (Banerji and Mitra, 2022; Howard et al., 2021). Inconsistent datasets might have different biases between data sources and impede neural network performance. For histopathology images, possible sources of variation are detailed in section 2.1.3 *Dataset acquisition*.

One research project that identified the need for a standardized ecosystem is *empaia*. Physicians and AI experts are part of the consortium to encourage an exchange of knowledge and build an ecosystem that pathologists, researchers, and industry can benefit from (empaia.org, 2022).

In addition to high-quality digitized images, for supervised learning, ground truths in the form of image annotations are required. Images need to be annotated by at least one experienced pathologist, e.g., for Gleason classification tasks. This process is time-consuming, expensive, and requires close collaboration with the clinics (Tizhoosh and Pantanowitz, 2018; Montagnon et al., 2020; Kohli et al., 2017). For survival prediction, no manual annotations are required, but follow-up data needs to be recorded, i.e., the patient's disease status over time. Difficulties arise since this data may not be routinely documented in EHRs and requires data to be recorded over a long time.

A crucial aspect of survival prediction is choosing the patient cohort. All patients need a common reference time origin of data acquisition (e.g., they enter the study on the day of their treatment), equivalent medical records, and treatments (Gerds and Kattan, 2021). Also, when comparing the tissue of two patients who underwent RPE, some boundary conditions need to match. For example, the time from biopsy to relapse is incomparable if the time between biopsy and RPE is not constant across patients since the cancer can grow during that period.

2.2.3 Prostate cancer classification for histopathology images

Research on ML-based prostate cancer classification has increased in the last years, especially since 2016, as a recent literature review from Denysenko et al. (2022) reveals. A short overview of the most recent works for the classification of prostate cancer histopathology images is given in Table 2.2.3. An extended version is available in Table A.1.1. Due to the extensive literature, this overview makes no claim of being complete, e.g., MRIs and CTs can also be used as an image source for prostate stratification, however, these are not covered here (Ahmed et al., 2017; Bertelli et al., 2022; John et al., 2021; Korevaar et al., 2021). Instead, this overview is restricted to histopathology images since that is the image source used for the experiments in this thesis. The following paragraphs will highlight the main similarities and differences in the approaches to prostate cancer grading from H&E-stained histopathology images.

Model output Automated prostate cancer grading may be approached as a binary classification to differentiate benign from malignant tissue (e.g., Campanella et al., 2018) or low-grade from high-grade cancer (e.g., Jimenez-del Toro et al.,

2017) or as a multi-class classification to differentiate distinct Gleason grade groups (e.g., Bulten et al., 2020). Some research not only aims at classifying a complete image but also at identifying more exact cancer regions, either through segmentation or patch-wise classification (e.g., Arvaniti et al., 2018; Burlutskiy et al., 2019).

Dataset source All prostate cancer datasets in Table 2.2.3 include H&E stained images as input. However, since they stem from different hospitals and countries, and standardized imaging protocols are missing, they differ, for example, in staining, scanning, or image size. Also the tissue origin may differ, as it can be extracted via biopsy or after RPE. Some publications classify small tissue microarray (TMA) spots obtained after RPE (e.g., Arvaniti et al., 2018). The TMA technique was introduced by Kononen et al. (1998) for faster analysis and typically shows round spots of 0.6 mm diameter. Others use larger WSIs (e.g., Li et al., 2021; Campanella et al., 2018), which are either obtained through a needle biopsy (e.g., Marginean et al., 2021; Duran-Lopez et al., 2020) or show prostate slices obtained after RPE (e.g., Nagpal et al., 2019; Jimenez-del Toro et al., 2017). This heterogeneity impedes direct comparisons of models.

Dataset size The used datasets also differ in size, as the number of included patients and images vary significantly. Zhang et al. (2021b) use as little as 54 WSIs, while Campanella et al. (2018) use 12,160 WSIs. Furthermore, most publications use more than one dataset, either for external validation or to train robustness (Bulten et al., 2020; Marini et al., 2021a). The number of images does not necessarily correspond to the number of available patients, since some datasets contain more than one image per patient, while others do not. In contrast, Jimenez-del Toro et al. (2017) state that some of their images contain multiple samples, which further reduces the comparability of the dataset sizes. Unfortunately, the ratio of images to patients is not provided in each paper. Marini et al. (2021a) only state the number of patches used, not how many complete images or patients are included, while Oner et al. (2022) have access to 99 WSIs from 99 patients, Marginean et al. (2021) use 735 images from 195 patients, and Burlutskiy et al. (2019) use 476 images without stating the image/patient ratio.

Dataset splits Besides differing in size, the datasets are all split differently into training, validation, and test sets. Some authors randomly divide the whole dataset into training, validation, and test sets according to fixed percentages (e.g., Campanella et al., 2018; Jimenez-del Toro et al., 2017). Others use datasets from different sources for training and testing (e.g., Bulten et al., 2020; Marini et al., 2021a). Bulten et al. (2020) have multiple test sets: internal, external, and one for comparison to pathologists. Instead of dividing the dataset by source clinic, Arvaniti et al. (2018) split their dataset on the TMA level, using one TMA for testing, one for validating, and the remaining TMAs as training sets. Duran-Lopez et al. (2020) use a three-fold cross-validation, so they only split the data into two training and test sets and use parts of the training set for validation. This heterogeneity impedes the comparison of the models' ability to generalize to unseen test data.

Dataset annotation Image annotations are either provided as single class labels at slide level (e.g., Campanella et al., 2018) or segmentations (e.g., Arvaniti et al., 2018). To overcome high interobserver variability, annotations for the whole dataset, or, due to labeling effort, for a test set, are often collected from multiple pathologists, and a majority vote defines the final ground truth (e.g., Nagpal et al., 2019, 2020; Bulten et al., 2020). Besides obtaining annotations from pathologists directly, some datasets are labeled based on the health records (e.g., Bulten et al., 2020; Li et al., 2021; Campanella et al., 2018). In contrast, Burlutskiy et al. (2019) let pathologists label only one part of their dataset. For the other part, they use a histochemical stain for automated segmentation annotation. They state that the absence of basal cells hints toward cancer and therefore use the color obtained through that staining for the definition of the ground truth.

Public datasets Most authors use non-public datasets that are only evaluated within their own publication. This impedes quality control and hinders the reproducibility of results, but is mostly because patient data requires high data privacy standards. Nevertheless, there exist some publicly available datasets that are used in multiple publications, either for training or as external test sets. However, also here it is not always transparent whether the exact same data partition is used. Some public datasets that are used by multiple authors are summarized in Table 2.2.2 and presented in the following.

The TMAZ dataset (TMA Zürich; Arvaniti et al., 2018) includes 886 TMA spot images with segmentation annotations per Gleason pattern. It is used for Gleason grade prediction by, for instance, Marini et al. (2021a) and Bulten et al. (2020).

Table 2.2.2: Prostate cancer histopathology datasets that are used in multiple publications are listed here. im: number of images, pa: number of patients, TMA: tissue microarray.

Dataset (Source)	**Tissue origin**	**Dataset size**
TMAZ (Arvaniti et al., 2018)	TMAs	886 im, 886 pa
GleasonChallenge (Nir et al., 2018)	TMAs	333 im, 231 pa
TCGA-PRAD (Zuley et al., 2016)	prostatectomies	16,790 im
DiagSet (Koziarski et al., 2021)	biopsies	5,179 im
SICAPv1 (Esteban et al., 2019)	biopsies	79 im, 48 pa
SICAPv2 (Silva-Rodríguez et al., 2020)	biopsies	155 im, 95 pa
PANDA (Bulten et al., 2022)	biopsies	12,625 im

The GleasonChallenge dataset was presented as a challenge at the 22$^{\text{nd}}$ International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2019 (Nir et al., 2018; Karimi et al., 2020). It comprises 333 TMAs from 231 patients with Gleason grade segmentation masks. The dataset is also used by Vuong et al. (2021) and Marini et al. (2021a).

TCGA (The Cancer Genome Atlas) is a composition of multiple cancer types, including prostate cancer (Prostate Adenocarcinoma Collection, TCGA-PRAD; Zuley et al., 2016). It contains whole slide images of prostatectomies and is annotated with Gleason patterns from pathology reports (Jimenez-del Toro et al., 2017).

Further, Koziarski et al. (2021) present DiagSet, a dataset that is available online and consists of biopsy WSIs. 4,675 of those slides are annotated with binary classification (cancer / no cancer), while 668 are with more detail (Gleason grade group, background, benign, artifact). At the moment, Marini et al. (2021a) also use that dataset, especially to estimate the performance on datasets from a different source than the training data.

Li et al. (2021) use SICAPv1, published by Esteban et al. (2019). It consists of 79 prostate biopsies from 48 patients with pixel-level annotations. Silva-Rodríguez et al. (2020) present the extension, SICAPv2, including 155 images of 95 patients, which is publicly available and used by Marini et al. (2021a).

The to-date largest WSI prostate cancer dataset was released for research usage by Bulten et al. (2022) in the prostate cancer grade assessment (PANDA) challenge at MICCAI 2020. It contains whole slide biopsy images from two different institutions and is partly annotated with Gleason pattern segmentation masks, partly with fewer classes (cancer, benign, background).

Model Even though all listed approaches use CNNs for classification, there is no consensus on which architecture leads to the best results. Some researchers report that ResNet performs best (e.g., Campanella et al., 2018; Oner et al., 2022), while others find InceptionV3 outperforming other architectures (e.g., Nagpal et al., 2019; Ström et al., 2020; Marginean et al., 2021). Many approaches also use variations of networks, e.g., Xception-like network (Nagpal et al., 2020). Silva-Rodríguez et al. (2020) develop their own CNN architecture, which they train from scratch. Ikromjanov et al. (2022) apply a vision transformer model on the PANDA dataset. Bulten et al. (2020) use a U-Net for the segmentation of Gleason patterns or benign glands. The final classification is obtained from those segmentations afterward.

The approaches further differ in the way the images are preprocessed. Since the images are relatively large, most authors cut them into patches before processing. Arvaniti et al. (2018) use small patches to predict the Gleason grade per region and later combine the results to obtain pixel-wise segmentation. Nagpal et al. (2019) also classify single patches during training and inference, however, with a different aggregation of the final classification. Ström et al. (2020) use a two-fold approach. First, they classify patches as either benign or malignant, and, with a second model, determine the Gleason grade. Oner et al. (2022) first segment glands in patches and use only these for classification.

Marini et al. (2021a) train a model that is supposed to be invariant to staining (color) differences. For each image, they calculate the H&E stain matrix, which converts the hematoxylin and eosin coloring to RGB colors. They simultaneously train a classifier for Gleason grade group classification and a regressor for the prediction of the H&E matrix components. While the classification is encouraged, correct predictions of H&E matrices are punished in the proposed objective function. Their backbone is a DenseNet (Huang et al., 2017). Ren et al. (2018a) also tackle the problem of different dataset sources. They train a siamese network and use a loss commonly applied in generative adversarial networks (GANs) to encourage learning of domain invariant features.

Metrics Since Gleason grading is a classification task, accuracy is a common metric for evaluation. The accuracy measures the fraction of correctly predicted samples, hence, it ranges from 0 to 1, with 1 being perfect classification (e.g., Ren et al., 2018a; Nagpal et al., 2019; Bhattacharjee et al., 2021). Also, the amount of true and false positives and negatives can be calculated and summarized as precision, recall/sensitivity, specificity, or F1 score. Precision measures how many samples of a predicted class were really of that class. Recall and sensitivity measure how many samples of a class were identified. Specificity measures how many samples of the false class were identified. The F1 score is a mixture of precision and recall. The AUROC (area under the receiver operator curve) summarizes the area under the plot of true positives against false positives, with a maximum value of 1, while the AUPRC (area under the precision recall curve) summarizes the area under the plot of precision and recall. Further, Cohen's kappa (Cohen, 1960) is a metric that is also commonly applied and measures the interobserver agreement of two ratings, taking into account random guessing (e.g., Arvaniti et al., 2018; Bulten et al., 2020; Ström et al., 2020). For random guessing, Cohen's kappa is 0.5, and for perfect agreement, it is 1. Oner et al. (2022) segment glands and therefore also report a Dice score (Dice, 1945) besides accuracy, precision, and recall, which measures the similarity of annotation and prediction. For details on these metrics, consider Vujović (2021).

For the binary Gleason classification task, AUCs as high as 0.999 (Duran-Lopez et al., 2020) are reported, and kappas up to 0.979 for the multi-class classification (Bhattacharjee et al., 2021). Ström et al. (2020) report a kappa of 0.62, which indicates that the results vary widely among different datasets, and thus none of the reported performances can be conceived as a general baseline for clinical assessment.

Performance compared to human pathologists The subjective nature of Gleason grading hinders comparison to a ground truth if the image annotation is provided by a single pathologist. Therefore, some authors compare their model's performance to multiple pathologists. Nagpal et al. (2019) show that their model's accuracy of 0.70 is higher than that of 8 out of 10 pathologists. Ström et al. (2020) show an extensive study, where their ANN outperforms 3 or 7 out of 23 pathologists in terms of kappa, depending on the number of prediction classes (5 or 3 classes, kappa of 0.62 for both). Bulten et al. (2020) show that their model outperforms 10 out of 15 pathologists with a kappa of 0.85.

Summary The approaches toward prostate cancer classification are diverse. Since there is no common dataset for benchmarking, it is difficult to state which approach performs best overall. Each model is optimized to a different dataset and tested on a test set either from the same or an external data source. An objective evaluation of the results is further impeded since not all results are reproducible due to not accessible hyperparameters, imaging protocols, or restricted access to the datasets.

Multiple authors report similar model performance compared to pathologists on the respective datasets and use cases. However, the subjectivity of labels may impede further improvements. For reliable results, large datasets should be accessed, and accuracy as well as kappa should be evaluated since these metrics are most common in the evaluated literature.

In conclusion, a common neural network architecture like ResNet or Inception should be used to build a Gleason classification model, however, for any new prostate cancer dataset it is essential to evaluate first which baseline architecture performs best. Then, the model can be adjusted (i.e., adaptation of preprocessing) and hyperparameters can be optimized to the given dataset.

Table 2.2.3: Literature overview for prostate cancer stratification. AUROC: area under the receiver operator curve, AUPRC: area under the precision recall curve, acc: accuracy, pa: number of patients, im: number of images, n/a: not available.

Paper	**Task**	**Model**	**Metric**	**Dataset**
Campanella et al. (2018)	benign/ malignant	MIL + ResNet, VGG	AUROC 0.98	• non-public • biopsies • 12,160 im (n/a pa)
Jimenez-del Toro et al. (2017)	low/ high	GoogLeNet	acc 0.78	• TCGA • prostatectomies • 235 im (n/a pa)
Burlutskiy et al. (2019)	binary segmentation	U-Net for different resolutions	F1 0.8 AUPRC 0.89	• non-public • biopsies + prostatectomies • 476 im (n/a pa)
Arvaniti et al. (2018)	Gleason grade group	MobileNet	kappa 0.75	• TMAZ • TMAs • 886 im (886 pa)
Nagpal et al. (2019)	Gleason grade group	InceptionV3	acc 0.7 AUROC 0.96	• TCGA + non-public • prostatectomies • 769 im (769 pa)
Ström et al. (2020)	Gleason grade group	InceptionV3	kappa 0.62	• non-public • biopsies • 9,001 im (1,474 pa)
Nagpal et al. (2020)	Gleason grade group	Xception-like	kappa 0.71	• non-public • biopsies • 1,276 im (1,112 pa)
Bulten et al. (2020)	Gleason grade group	U-Net	kappa 0.72-0.85	• non-public + TMAZ • biopsies + TMAs • 6,745 im (2,129 pa)

Marini et al. (2021a)	Gleason pattern	adversarial CNN	kappa 0.47-0.73	• TMAZ + SICAPv2 + GleasonChallenge + DiagSet • TMAs + biopsies • n/a im (n/a pa) 83,091 patches

2.2.4 Deep learning for survival prediction

Predicting the time-to-event, such as the time from surgery to death from a certain disease, is referred to as survival prediction. For clinical decision support, the time to an event is a relevant score to decide for or against treatments or for triage, e.g., for a liver transplant or in intensive care units (Heitz et al., 2021; Andres et al., 2018). In a study, Heinz et al. (2022) asked 75 computational pathology experts which of 12 predefined tasks they find most promising (e.g., cell classification, prediction of gene expression). Survival prediction from H&E images was assigned an importance score of 9 out of 10. Among participants with a medical background, the score reached 10, making it the second-most important task after the prediction of treatment response. These results emphasize the relevance of survival prediction for clinical usage.

There is a large variety of models for survival analysis. A short overview of the literature for survival prediction is presented in Table 2.2.4, with an extended version in Table A.1.2. Since this thesis uses histopathology images as input, the overview is restricted to models using medical images as input. Besides H&E stained images, CT, MRI, or radiography images are considered, since similar models might be applicable. The use case is not restricted to prostate cancer since the approaches can be applied interchangeably to any disease progression and restricting the overview to prostate cancer would be too limiting. Work that does include prostate cancer is highlighted, though. Approaches using only electronic health records (EHRs) or genomics as input are not covered in this thesis (Giunchiglia et al., 2018; Huang et al., 2021b; Katzman et al., 2018; Kvamme et al., 2019) but were researched in a complementary project (Fuhlert et al., 2022).

The chosen papers in Table 2.2.4 allow a subsumption of the approach developed in this thesis into the state of the art, stressing similarities and differences. This list does not include all existing literature but tries to provide a general overview. For an overview exceeding the selection presented in this thesis, consider Wang et al. (2019).

Similar to the overview presented on Gleason classification in section 2.2.3 *Prostate cancer classification for histopathology images*, the presented literature varies in the model architectures and dataset sources as well as evaluation metrics. Furthermore, the problem formulation is not consistent across the presented works, and also the chosen endpoint is either overall survival, time to disease-related death, treatment success, or cancer recurrence (Duanmu et al., 2020; Kumar et al., 2017; Esteva et al., 2022). The broad spectrum of applications, data sources, and the lack of consistent metrics and problem formulation impedes the comparison of model performance. Therefore, the differences and similarities among the approaches are stressed.

In the following, the literature is split into three broader categories based on the problem formulation. Those categories are i) binary prediction (relapse or not), ii) risk score prediction (discrete or continuous), and iii) prediction of relapse probability over time in the form of survival curves. The latter category is split further into approaches that rely on the Cox model (Cox, 1972) and approaches that predict individual survival probabilities for discrete time intervals. Details on the Cox model will be explained in section 5.4.1 *Cox model.* Within these categories, similarities and differences in architectures, losses, and metrics are pointed out. Also, the advantages and limitations of the approaches are stressed. Theoretical aspects, such as the used metrics, are only briefly described in this section for understanding the context, but more details are provided in chapter 5 *Survival prediction.*

Binary classification The simplest strategy to answer the question "Does a patient experience an event (before time t)?" is a binary classification (e.g., Duanmu et al., 2020; Exarchos et al., 2012; Kumar et al., 2017; Yamamoto et al., 2019). Such a classification may support a physician in deciding whether a therapy would be helpful for a patient but neglects the exact time to relapse since only a single cut-off time is considered. Furthermore, the inclusion of censored patients, i.e., patients without an event, is limited to those with a later censoring time than the survival time threshold. (More information on censoring is included in section 5.2 *Censoring.*)

In binary classification, the input images are processed with different CNN architectures and the performance is measured in terms of AUC, accuracy, specificity, or sensitivity. While Exarchos et al. (2012) extract features like "extra nodal spreading" or "bone infiltration" from images to input these together with preprocessed clinical and genomic data to a classifier like random forest, recent work uses neural networks on the input images end-to-end. Huang et al. (2022), e.g., use a CNN and an MIL approach for binary prediction of prostate cancer recurrence in WSIs. Instead of predicting cancer recurrence, Duanmu et al. (2020) predict whether a patient will respond to a therapy. As input to all models, either MRI, CT, or histopathology images are used, while Duanmu et al. (2020) and Exarchos et al. (2012) also include clinical or genomic data.

Kumar et al. (2017), Yamamoto et al. (2019), and Huang et al. (2022) use prostate cancer histopathology images as input. While Kumar et al. (2017) use TMA spots, Yamamoto et al. (2019) and Huang et al. (2022) use WSIs as input. Within the presented literature, the dataset sizes range from 220 to 842 cases. Huang et al. (2022) predict relapse within 3 years, Kumar et al. (2017) within 5 years, and (Yamamoto et al., 2019) evaluate their model on 1-year and 5-year relapse prediction.

Risk score For a more fine-grained survival prediction the problem is formulated as “How high is the risk of experiencing an event?”. To that end, each patient is either assigned a discrete risk score by dividing patients into multiple groups (e.g., long-term, mid-term, short-term survival as in Muhammad et al., 2021) or a continuous risk score between 0 and 1 (Zhou et al., 2020; Wulczyn et al., 2020; Chang et al., 2021; Li et al., 2018; Fan et al., 2021; Huang et al., 2022). Instead of calculating the risk, Pinckaers et al. (2022) directly predict the year of relapse as a score between 0 and 4. Wulczyn et al. (2020) divide the time into intervals of 25 months and aim to predict the risk score per interval. From these, a continuous risk score is created afterward. For these discrete approaches, the authors need to define beforehand how the survival times are split into groups.

For risk group prediction, again accuracy, precision, or F1 score are used as metrics for evaluation. For the continuous risk prediction, no ground truth risk score can be obtained. Most models that predict a continuous score are evaluated with the concordance index, short C-index (Harrell et al., 1982). It ranges from 0 to 1 and measures how well patients are discriminated, hence compares the order of predicted risk scores with the order of survival times. A C-index of 1 indicates perfect discrimination, whereas 0.5 is equivalent to a random guess. Reported C-indices reach up to 0.88 (Muhammad et al., 2021). Pinckaers et al. (2022) report a hazard ratio of 3.02 and an odds ratio of 3.32, which both measure the effect of a feature on the survival time when comparing two groups (George et al., 2020; Kleinbaum and Klein, 2012). For both, a value of 1 states that there is no effect. Based on the predicted continuous risk values, many authors further split the patients into two or three risk groups for evaluation (Chang et al., 2021; Muhammad et al., 2021; Wulczyn et al., 2020). The discrimination power is visually shown with Kaplan-Meier (KM) survival curves per group and sometimes quantified with a log-rank test (Kaplan and Meier, 1958; Chang et al., 2021). With KM curves, populations' probabilities of surviving can be compared. More details on KM curves are given in section 5.3 *Population-based survival prediction*. More details about survival model evaluation metrics can be found in section 5.5 *Metrics*.

As a loss function, the negative log partial likelihood (nlpl) is commonly applied. Since the nlpl loss only compares the ranking of patients it favors the discriminatory power of the models. Muhammad et al. (2021) extend the loss with additional terms to stress distinguishing low and high risk. Patients are stratified into two groups based on the predicted risk, and the difference between both groups' mean risks is optimized. Wulczyn et al. (2020) combine cross-entropy loss, since they use a classification-like output, with nlpl and an additional exponential lower bound on the C-index.

The most commonly used architecture for risk prediction is ResNet (Zhou et al., 2020; Muhammad et al., 2021; Fan et al., 2021), while some papers use EfficientNet (Kiyokawa et al., 2022; Walhagen et al., 2022) or individual approaches (Chang et al., 2021). In contrast, Li et al. (2018) developed a graph network using patch encodings as vertices and thresholded distances as edges. Furthermore, self-supervised pretraining may be included for feature extraction (Chang et al., 2021; Fan et al., 2021). Fan et al. (2021) first train an encoder-decoder architecture on a colorization task. The encoder part is further used for survival prediction. Zhou et al. (2020) and Esteva et al. (2022), for instance, use clinical features in addition to images. Esteva et al. (2022) predict the risk scores either after 5 or 10 years, while Pinckaers et al. (2022) predict in which of the first 4 years a relapse occurs. However, not all authors state the time horizon that is regarded.

Prostate cancer use cases are considered by Walhagen et al. (2022), Pinckaers et al. (2022), and Esteva et al. (2022). Esteva et al. (2022) make use of a non-public dataset with 5,654 digital histopathology image data (16,204 histopathology slides) of pretreatment biopsy samples. For 5-year recurrence prediction, they reach a time-dependent AUC of 0.67, and for 10-year prostate-related death, they reach 0.77. Pinckaers et al. (2022) use two datasets that are available upon request for research purposes and consist of TMA spots from 685 and 204 patients. Walhagen et al. (2022) use a private dataset, which is the same that is used for model development throughout this thesis. They train a network for binary survival prediction but treat the predicted output as a risk score. They achieve AUCs of 0.79 and 0.93, depending on whether the test set is denoised.

From the previous literature on Gleason classification, some papers relate their classification results to the survival of patients afterward, without using survival information during training. Since Gleason grades may be comparable to risk groups those approaches are shortly mentioned here. Ström et al. (2020) show that survival times are equally distributed across their predicted ISUP scores and a pathologist's ISUP annotations. No results in form of a metric are reported. Tolkach et al. (2020) evaluate the KM curves for the predicted Gleason grades and show with a log-rank test that those stratify well. Nagpal et al. (2019) use their classification as risk stratification for BCR and achieve a C-index of 0.65, which is better than the median achieved by pathologists. Arvaniti et al. (2018) group their results into low, intermediate, and high risk groups and show the KM curves for these groups. They also state that their groups are statistically more significant in the log-rank test than those obtained from a pathologist. However, no comparison to models trained on survival times is made.

Survival curve Not only the risk estimation compared to others is relevant for a patient, but also an individual prediction of risk over time is beneficial. A patient would ask, "How long will I live without a relapse? What is the probability of living relapse-free for another three years?". To answer these questions, survival models that predict the probability of an event occurring over time are necessary.

Survival curve with the Cox model

A popular approach to survival prediction on tabular data is the Cox model. It focuses on predicting the correct ordering of patients. Only linear dependencies between patient features are considered and patient survival curves cannot cross. More detailed theory on the Cox model can be found in section 5.4.1 *Cox model*. One of the first approaches to Cox model-based survival analysis on image data is from Zhu et al. (2016). They process images with a CNN and train the model to order the patients correctly. The model output is, like in the Cox model, a single hazard value, which is comparable to a risk score. A survival curve can be derived afterward by modeling a base hazard. As a loss function, the nlpl is applied.

Further extensions of this approach, such as adapting the preprocessing, optimizing the CNN architecture, or including attention are prevalent (Li et al., 2022; Ren et al., 2018b; Wang et al., 2021; Zhu et al., 2016). Ren et al. (2018b) extract features from genomic information and fuse these into an AlexNet, which is used for image processing. The features are combined across patches with a recurrent neural network to obtain a final vector, which is input into a Cox model. Liu and Kurc (2022) include segmentation masks of different structures (tumor region, tumor-infiltrating lymphocytes, and nuclei) in addition to the image so that their input consists of six channels. With these, they first predict a risk score of experiencing an event across five intervals. The result from their model is then further processed with a Cox model to obtain a survival curve. Nam et al. (2022) first predict the survival probability in different intervals with a DenseNet from images and combine this with clinical features. However, they state that adding a Cox regression to those predictions increases the performance. Their prediction intervals cover 600 days each.

For the evaluation of these survival models, the C-index is calculated on the predicted hazard score and reported for all models. Li et al. (2022), Mobadersany et al. (2018), and Liu and Kurc (2022), for instance, draw KM curves for two to three groups to illustrate the discriminative power.

In this category, Ren et al. (2018b) train only on prostate cancer data, reaching a C-index of 0.71 solely on images (0.74 when including genomic information). Sandeman et al. (2022) train a CNN to predict Gleason grading and use these predictions as input for a Cox model. They show that the KM curves stratify well when dividing the patients into two risk groups. Using all 5 Gleason grade groups, the KM curves stratify poorer. However, they do not report a C-index.

Survival curve without the Cox model

Instead of using the Cox model, patients' survival probabilities can be predicted for discrete time intervals, allowing for individual progression estimates (Vale-Silva and Rohr, 2021; Xiao et al., 2020; Nam et al., 2022). Two common approaches are found: Vale-Silva and Rohr (2021) output a hazard score per time interval, while Xiao et al. (2020) and Hermoza et al. (2022) directly model the survival probability per interval. Popescu et al. (2022) estimate a fixed distribution for the individual survival curves with a log-logistic distributed hazard rate. For each patient, they estimate the individual hazard rate with a log-logistic distribution and train to learn its scale μ and shape σ.

If a model is designed to output a prediction per interval, the spacing of the intervals needs to be defined in advance. The interval lengths are inconsistent across the literature presented here since they depend on the use case. Hermoza et al. (2022) divide their bins so that an equal number of patients is included per bin. This results in intervals covering 354, 757, and 1,813 days. Vale-Silva and Rohr (2021) present a study on different interval spacings (1 year, 0.5 years, and quantile spacing based on patients' survival times with different time horizons) and report that results do not differ much when changing the interval spacing.

The most common metrics to evaluate survival curves are again the C-index and the KM curves for visual interpretation. The survival model evaluation metrics are further detailed in section 5.5 *Metrics*. Xiao et al. (2020) use the survival curve to estimate the overall survival time of the patient. They evaluate the mean average error (MAE), which calculates the error between the estimated and the true survival time. If the target is converted into a binary prediction of surviving a specified time, the AUC can be measured. Another metric is the Brier score, which measures the distance between true and observed probability of surviving single intervals. Survival models perform well if they achieve a high discriminatory power and are also calibrated. Good calibration means that the predicted survival probabilities represent the true survival probabilities. However, this is often neglected in the literature (Kamran and Wiens, 2021).

Overall, most models are based on CNNs (Xiao et al., 2020; Hermoza et al., 2022; Lombardo et al., 2021). ResNet is commonly applied (Xiao et al., 2020; Hermoza et al., 2022; Yala et al., 2021). Yala et al. (2021) combine the ResNet with a Transformer, Vale-Silva and Rohr (2021) use a ResNeXt, and Nam et al. (2022) use a DenseNet as base model. Agarwal et al. (2021) use a siamese network structure, to process two patients in parallel, and predict the difference in survival times as the output.

Vale-Silva and Rohr (2021) include 33 different cancer types, among them prostate cancer, in their experiments and reach a C-index of 0.569 when using only WSIs as input. The results are improved to 0.801 when including clinical data. On the prostate cancer dataset, they report a Brier score of 0.079 and a C-index above 0.8 when including additional patient features.

Summary The survival prediction approaches presented in Table 2.2.4 are heterogeneous. Image sources, model architectures, and prediction horizons differ in the current research. It can be concluded that survival prediction from medical images is a research topic that is still under development. The current literature does not provide a clear guidance on which approach is most promising and should be used as a starting point for new developments. While some researchers reduce survival prediction to a binary or risk score prediction, predicting a survival curve provides more detailed information to a pathologist about the patient's progression. The Cox model is often combined with a CNN but restricts the survival curves to be non-crossing. Hence, modeling the survival per interval with a neural network seems more valuable. With such an approach, Xiao et al. (2020) and Hermoza et al. (2022) reach high discrimination performance. However, they model the survival in different intervals independently and thus cannot enforce decreasing survival curves. By modeling the hazard function, Popescu et al. (2022) and Vale-Silva and Rohr (2021) enforce biologically reasonable survival curves. However, both report performances below a C-index of 0.7 when using only images as input. Only by adding clinical or genomic patient data, the prediction performance can be improved. Thus, ways to improve survival prediction models from histopathology images need to be explored.

Table 2.2.4: Literature overview for medical survival prediction using images as input. **Bold** literature uses only prostate cancer data, *italic* literature uses also prostate cancer data, besides others
Data sources: M: MRI image, C: CT image, H: Histopathology image, R: radiograph, c: clinical data, o: omics data (genomics and/or transcriptomics and/or epigenomics and/or radiomics) – Data sizes: pa: patients, im: images – Loss: nlpl: negative log partial likelihood, i.e., Cox-loss, n/a: not available. – Metrics: AUC: area under the receiver operator curve, sp: specificity, se: sensitivity, acc: accuracy, KM: Kaplan-Meier, MAE: mean average error, OR: odds ratio, HR: hazard ratio

Paper	**Task**	**Data**	**Model**	**Loss**	**Metric**
BINARY					
Duanmu et al. (2020)	therapy response	M c o 112 pa	VGG-13 for 3D data	n/a	AUC 0.8 acc 0.89 F1 0.77 sp 0.88 se 0.68
Kumar et al. (2017)	relapse (5 years)	H 220 pa	2 CNNs: detect nuclei + classification	binary cross entropy	AUC 0.81
Yamamoto et al. (2019)	relapse (1 and 5 years)	H 842 pa 9,916 im	Autoencoder + SVM	n/a	AUC 0.76-0.84 pseudo R-squared 0.26
Huang et al. (2022)	relapse (3 years)	H 416 pa 416 im	CNN	cross entropy	AUC 0.78

RISK SCORES

Esteva et al. (2022)	risk score (5 and 10 years)	H c 5,654 pa 16,204 im	self-supervised pretraining + CatBoost fusion	n/a	time dep AUROC 0.67-0.77
Wulczyn et al. (2020)	risk score / risk interval (3 intervals)	H 6,096 pa (15,104 im)	CNN similar to MobileNet	cross entropy (for risk in interval)	c-index 0.61 AUC 5year 0.7 KM (3)
Walhagen et al. (2022)	risk score (event < 3 years)	H 15,238 pa 15,238 im	EfficientNet + MIL	cross-entropy	AUC 0.79-0.93 KM (7)
Pinckaers et al. (2022)	year of relapse (year 0-4)	H 889 pa 2,963 im	ResNet	smooth L1 loss	OR 3.32 HR 3.02 KM curve (2 and 4)

SURVIVAL CURVES WITH COX MODEL

Ren et al. (2018b)	survival	H o 247 im	AlexNet + LSTM	nlpl	HR 5.73 C-index 0.74
Zhu et al. (2016)	survival	H 450 pa	DeepConvSurv (CNN + Cox)	nlpl	C-index 0.63
Liu and Kurc (2022)	5 intervals	H c 978 pa 978 im	6-channel input to MobileNet + Cox	extension of cross-entropy	C-index 0.70 (im only) 0.73 (with c) HR 1.19
Nam et al. (2022)	survival curve (600 days)	R c 5,372 pa	DenseNet + neural net + Cox	negative log likelihood	time dep. 5-year AUC 0.67-0.76 (im only) C-index 0.63-0.72 (im only) KM (2) calibration

SURVIVAL CURVES WITHOUT COX MODEL

Xiao et al. (2020)	survival curve + time	H 769 pa, 1,061 im	CDOR (ResNet for censoring-aware deep ordinal regression)	censor-aware cross entropy	MAE 321.2 C-index 0.74
Hermoza et al. (2022)	survival curve per interval + time	H, Xray 16,013 pa, 49,008 im	ResNet	augmentation of censor aware cross-entropy	MAE 26.28 C-index 0.76
Popescu et al. (2022)	survival curve (to 10 years)	M c 269 pa	multiple network fusion + log-logistic survival model	negative likelihood	C-index 0.63 (im only) 0.74 (with c) Brier 0.19 (im only) 0.14 (with c)
Vale-Silva and Rohr (2021)	survival curve per interval (30 years)	H c o 11,081 pa 8,376 im	ResNext	negative log likelihood	C-index 0.57 (im only) 0.82 (with c) Brier 0.22 (im only) 0.14 (with c) KM (per cancer)
Lombardo et al. (2021)	survival curve	C o c 1,037 pa	2D + 3D-CNN	negative log likelihood	C-index 0.67-0.88 AUC 0.63-0.89 KM (2)

2.2.5 Robustness

Most neural network models are developed under the closed-world presumption: It is assumed that the training, validation, and test data are from the same distribution and valid representations of real-world cases (Hsu et al., 2020; Yang et al., 2021). However, this assumption does not always hold true. When the network is applied in an open-world scenario, the distributions between the training and test datasets can differ (Yang et al., 2021). These differences result from (hidden) biases in the datasets, for instance, due to different staining protocols. A model's performance may decrease when applying it to a dataset with a different bias than the one it was trained on (Tellez et al., 2019). Tschandl et al. (2020) showed that clinicians' decisions improve when being supported by good-quality "AI-based" predictions. However, clinicians' performance degrades when the AI system provides faulty suggestions. This stresses the need for a robust model that performs consistently and equally well on internal and external datasets, independently from the applied staining protocols and scanners used for digitization. By restricting the real-world use cases, e.g., to a certain scanner, some bias may be avoided. However, some sources of domain shift or bias are inevitable.

Three options are commonly applied to tackle the challenge of dataset biases: One option is to adapt the model architecture to make the model robust against bias changes. As a second option, data preprocessing might transfer the dataset bias from one domain to the other. The test set bias can be transferred to all training images prior to training (e.g., Ma, 2021). Also, the training set bias can be transferred to the test images after training (e.g., Varsavsky et al., 2020). A third option is to identify data that differs too much from data that was seen during training to make a reliable prediction. This can be approached with out-of-distribution (OOD) detection. When an input is recognized as OOD, it should not be used for a prediction (e.g., Lee et al., 2018).

The following section first defines and sets the limits of what is covered with the term "robustness" in the scope of this thesis. Then, methods for detecting OOD samples and methods for bias transfer are introduced. In the following, if applicable, example literature using prostate cancer or histopathology images is provided. However, the cited work is not restricted to either prostate cancer as the disease or histopathology as the image source.

Definition The term robustness covers a wide range of applications and meanings. Therefore here, the terminology is restricted to a few use cases. In the scope of this thesis, robustness will be tied to uncertainty. Hence, a model is expected to assign uncertainty scores to predictions in order to differentiate between images on which it can make a prediction and those images on which it cannot make a reliable prediction. Further, a robust model should be able to return correct predictions on images with a great variety of color biases. Thus, the space of uncertain images should be reduced as much as possible. Related research fields like adversarial attacks, detecting non-prostate-tissue images, or other types of robustness are not covered (for an overview on adversarial attacks in medical image analysis, consider Apostolidis and Papakostas, 2021).

Robust models Robustness may be achieved by training on a large dataset that includes various biases. Since such a dataset is not always available, data augmentation is used to expand the variety of training images artificially. The neural network's ability to generalize is increased when trained with a larger variety of images. Image augmentation methods include color adaptations and geometric transforms like flipping or rotation (Shorten and Khoshgoftaar, 2019).

In contrast, color adaptation methods aim at reducing stain variation to assimilate training and test distributions (Tellez et al., 2019). Tellez et al. (2019) perform an extensive study on the effects of stain color augmentation and normalization in histopathology images. As a result, they report that stain color augmentation improves performance, while normalization is negligible. They further train a U-Net to reconstruct an image's original stain after heavy color data augmentation. That network learns to remove the data augmentation effects and thus performs a stain color normalization. The normalized images are used in a classification network for both training and inference.

Different approaches beyond data augmentation have been proposed to make models more robust against domain shifts. Marini et al. (2021a), for example, aim to train a network on prostate cancer histopathology images, which is robust against stain color heterogeneity. Instead of training only a classifier for Gleason grade classification, they in parallel use a regression output to predict the H&E components of the image. They argue that the model learns staining invariant features this way. During training, they use two datasets from different sources.

Ren et al. (2019a) also train an adversarial model for prostate cancer stratification on both the source and target domain, without requiring labels for the target domain. They aim to create a feature space that is discriminative for the task, not for the domain. To achieve this, a siamese CNN network that processes two target domain image patches at once and a CNN to process the source domain patches are implemented. The siamese network is trained to classify two patches from the same image as the same class. A discriminator between both domains is trained with a GAN loss.

Color transfer Instead of training a model to be robust against domain shifts, another approach is to reduce the difference between the training and test domain. To assimilate dataset biases, domain adaptation or color transfer methods can be applied. Some approaches adapt the training dataset such that the images are more similar to the test set domain (Mohseni et al., 2020; Roy et al., 2022; Fort et al., 2021). Ma (2021) show that a model that is trained on MRI images that were adapted to the test set intensities performs better on the test set than a model trained on the original training images. The drawbacks are that this method requires retraining and can only be used when the test domain is known in advance with multiple available test-domain images.

Instead of adapting the training images to match the domain of the test set, the inverse direction is another option. That means the test images are adapted to match the training domain (Ren et al., 2019a; Bulten et al., 2020). An advantage is that neither re-training nor a large test dataset is needed, since single images may be matched to the training domain.

Different approaches for bias transfer have been proposed and can be used on the training or test domain. In the following, domain adaptation is defined as the color transfer from a source domain to a target domain. Biases besides color are neglected.

A simple color transfer approach is histogram matching (in RGB, HSV, or Lab space; Fan et al., 2022). The histogram of a source domain image is matched to the histogram of an arbitrary image from the target domain. However, visual artifacts may be introduced into the images by histogram matching (Ren et al., 2019a). For histopathology images, alternatives have been proposed, for instance, by Macenko et al. (2009) and Vahadane et al. (2016). Macenko et al. (2009) transfer the images into optical density space to obtain two distinct stain matrices. Background pixels that do not exceed a threshold value are omitted. The target image's stain matrices and maximum intensity are adapted to the stain matrices and intensity of the source image to assimilate the source image's appearance. Vahadane et al. (2016) also decompose the image into stain density maps, corresponding to hematoxylin and eosin staining. A structure-preserving color normalization is applied to adapt the source image to the target image.

A reference image needs to be chosen for all these color transfer techniques. Instead of matching to a single random image, Ren et al. (2019a) match the source image to several target images. They combine the predictions per transformation in the feature space to obtain the final classification. With an increasing number of target images, the classification performance improves.

Thebille et al. (2021) and Bulten et al. (2020) developed more complex domain adaptation methods. Both implemented a GAN for bias transfer of external image samples to the training domain. Compared to color transfer via histogram matching, the GAN approach improved test set results. Drawbacks are that GANs need access to a reasonably sized dataset of the new domain for training, they have to be retrained for every new dataset bias, and may also introduce visual artifacts.

For all color transfer methods, it is assumed in advance that the color matching improves the images sufficiently to make a correct prediction afterward. However, the image quality generally is not tested after transformation and no uncertainty measure, with respect to predictive performance, is included.

Out-of-distribution (OOD) detection Making a model robust to domain changes or sufficiently adapting the domain may not always be possible. Furthermore, a transformation does not guarantee reliable and certain model predictions. Therefore, mechanisms are needed to detect samples that do not stem from the same distribution as the training data. When these out-of-distribution (OOD) samples are encountered during inference, the prediction may not be reliable and thus should be discarded or passed on to a human expert.

The term "OOD detection" covers a wide range of applications and has ambiguous definitions in the literature. Yang et al. (2021) summarize out-of-distribution detection, anomaly detection, open set recognition, outlier detection, and novelty detection with the term "generalized OOD detection".

In the case of classification, detecting a semantic shift, that is, a sample coming from an unknown class, is a common task. In contrast, especially in the medical domain, non-semantic shifts like biases in the dataset need to be detected (Yang et al., 2021). Depending on the use case, either the semantic or non-semantic shift is focused on, even though some algorithms aim to detect both equally well (e.g., Hsu et al., 2020). Chen et al. (2020a) aim to detect artificial OOD examples obtained through adversarial attacks. However, robustness to adversarial attacks is considered a separate research field and therefore not covered here. In the scope of this thesis, OOD samples are assumed to have a different color bias on which a model cannot make reliable predictions.

A wide range of approaches toward OOD detection has been developed. The field gained more attention recently as researchers count robustness indispensable for the application of their models (Drenkow et al., 2021). An OOD detection algorithm can either be trained including OOD examples or solely on in-distribution (ID) examples. Exploiting OOD examples eases their detection but is limited to known biases since variations that might occur in real-world data might differ from OOD data present during training. In the following, two often-applied OOD detection approaches are presented, one including density estimation, and one based on classification output (Yang et al., 2021).

In classification tasks, the output of the final layer's softmax activation function is often interpreted as a probability for the input to belong to one of the output classes (e.g., Kumar et al., 2017; Nagpal et al., 2020). The higher the class probability, the higher the confidence. A threshold can be used to decide whether to trust the prediction or not. However, Nguyen et al. (2015) show that using the softmax output as an uncertainty measure is unreliable. It returns overconfident probabilities, even for inputs that are unrecognizable to humans. Therefore, Liang et al. (2018) propose pushing the probabilities of ID and OOD predictions further apart. In their method called ODIN (out-of-distribution detector for neural networks), they add a temperature scaling factor to the softmax function. The second important part of their work is adding a parameter ϵ to all input data to increase the difference between ID and OOD data. For this, they need access to OOD data during training. Hsu et al. (2020) propose a generalized ODIN model that does not need access to OOD samples. They alter the neural network's output to also predict the probability of the image being ID, along with the class prediction. Other classification models explicitly introduce an outlier class in the model so that unknown classes can be classified as outliers.

Lee et al. (2018) propose a different approach for outlier detection. They approximate the training distribution in the latent space (the output of a predefined neural network layer) of each class with a Gaussian distribution. Then they compute the distance from the latent representation of a new data point to each class mean. As a distance metric, the Mahalanobis distance is applied, which incorporates the data correlation. Based on the distance, it is decided whether a sample is considered in- or out-of-distribution. The final decision threshold is chosen to count 95 % of the training data as ID. Furthermore, they apply a data shift to all input data to spread ID and OOD data further apart, similar to Liang et al. (2018). This approach assumes a Gaussian distribution in the latent space, which Sun et al. (2022) propose to circumvent. They propose a k-nearest-neighbors approach and a Euclidean distance. In the latent space, they measure the distance to the k-th nearest neighbor, assuming that OOD samples reveal themselves through a larger distance (less close neighbors) than ID samples. Again, a threshold of 95 % ID training data is set to decide whether a sample is OOD.

However, as can be seen above, most OOD detection models presume an underlying classification task while some even require it, e.g., when adapting the softmax or adding an outlier class node. These approaches can thus not be applied to tasks like survival prediction. Other approaches aim at detecting OOD via the gradient of the model (if the gradient changes a lot, the input must have been different; Huang et al., 2021a) or use GANs to detect anomalies when the GAN cannot reconstruct the original image (Yan et al., 2021).

Summary In the literature, OOD detection, color transfer, and robust models are usually not combined but used separately. There is great potential in combining OOD detection, which only assigns uncertainty scores and dismisses OOD samples, with color transfer, which transfers all images without verifying that the resulting image is reasonable. Thus, evaluating the OOD-ness of a sample after color transfer could increase robustness even further, since uncertain samples can be deferred to an expert and only certain samples are used for prediction.

2.2.6 Explainability

For clinical application, a model needs to provide good performance on varying clinical datasets and studies (Kleppe et al., 2021). Even though this increases trust, neural networks still act as black boxes, without revealing the underlying decision-making processes. The influence of single input features on a prediction is hard to reconstruct due to the non-linear and complex nature of neural networks, which makes them difficult to interpret (Zhang et al., 2018).

To overcome this, explanations of what the neural network focuses on during a prediction can be generated. The decision-making process is better comprehensible if a neural network explains why it predicts a certain model output. For image analysis, e.g., the image regions that were most relevant for classification might be highlighted. Such explanations can be used to explore whether the model focuses on features that a human deems relevant and whether a decision should be trusted. The explanation might then be used during model development or deployment (Barredo Arrieta et al., 2020).

Besides being a beneficial feature, explainable AI (XAI) may be a requirement for clinical usage or acceptance (Heesen et al., 2020). The terms “explainable, trustworthy, understandable, interpretable” suffer from inconsistent connotations in the literature (Barredo Arrieta et al., 2020). For the scope of this thesis, these terms are used interchangeably and the meaning is limited to “showing which image regions are most relevant for a prediction of a neural network”.

There are different options to include explainability in neural networks. One option is to include explainability directly in the model architecture. Using attention MIL, the input image patches are weighted based on their relevance. These weights can reveal which image regions are most important for a prediction (Ilse et al., 2018).

Another option is to apply explainability methods after model development on the black box neural network, without adapting the architecture or training. With the introduction of class activation maps (CAMs), Zhou et al. (2016) established a new method, which has been widely applied and adopted further. By mapping the class activation scores backward through the neural network, they can reveal which image regions are most discriminative for a certain class. Instead of calling their approach explainability, they use the generated heatmaps for object localization. Selvaraju et al. (2017) propose an extension called gradient-weighted class activation mapping (Grad-CAM). They use the gradients of a target concept, such as one class in a classification, and state that their approach is not limited to classification outputs.

Patil et al. (2019) apply interpretability methods on breast cancer histopathology images. They state that attention-based MIL returns better localization results than Grad-CAMs. Li et al. (2021) apply Grad-CAM to the problem of prostate cancer classification to reveal relevant image regions.

A drawback of these methods is that no general statement about feature importance can be made. The importance is only explored and visualized on single images, without further quantification. For further exploration of explainability methods, the reader is referred to Barredo Arrieta et al. (2020), Vilone and Longo (2021), and Singh et al. (2020).

2.2.7 Conclusion

The state of the art illustrates that CNNs are the gold standard in computer vision classification and survival prediction tasks, but there is no single leading architecture. In the field of prostate cancer histopathology, most research focuses on Gleason grade prediction. Gleason prediction, however, requires much annotation effort, and the accuracy is bound by subjective pathologist labels. Instead, survival prediction models enable objective predictions since they can be trained on objective endpoints.

In the literature, survival prediction is often reduced to a binary prediction of whether a relapse occurs. However, this only takes into account a single time horizon, hence, no progression estimation. A more valuable approach is to predict survival curves. Many models that predict survival curves extend the Cox model (e.g., DeepConvSurv; (Zhu et al., 2016)), which is restricted to non-crossing survival curves and thus might not approximate the true underlying survival probability of patients. Therefore, it might be beneficial to pursue an approach that predicts survival probability in discrete intervals. The approach of Xiao et al. (2020), who model a survival curve directly with CDOR based on ResNet, is promising in this field since it only uses histopathology images as input and achieves high predictive performance. The downside of CDOR is that the time-dependency between the intervals is not modeled and the predicted survival per interval is unconstrained, thus it is not restricted to be monotonically decreasing. Vale-Silva and Rohr (2021) present a different promising approach for cancer-related survival prediction. They model the hazard instead of the survival curve and predict the survival for multiple cancer types. However, they also do not model the time dependency between intervals and do not include any explainability. Further, their model performs best when only using clinical patient data like age and genomic features, and the performance significantly decreases when predicting survival from a WSI. Thus, their model is not sufficiently trained to extract valuable image features. They state that the large variability in tissue appearances across cancer types might impede performance.

The robustness of models is not often tackled in histopathology. Marini et al. (2021a) present a classification model that is trained to be robust to staining variations and include a detailed analysis on multiple datasets. However, their model still shows great differences in model performance between internal and external test sets. The model trained by Ren et al. (2019a) achieves significantly improved scores over a baseline model on unseen datasets for prostate cancer stratification. While both models achieve reasonable performance on external datasets both require datasets from two different domains during training. Therefore, a color transfer method that is applied on test set images, independent of model training, is preferable. Current approaches include histogram matching or Macenko adaptation. Since a random reference training image is chosen for those approaches, the transformation is highly dependent on that random choice and not stable. Further, no quality control after color transformation is included. Detection of uncertain images can be tackled with OOD detection. Thus, combining color transfer methods with OOD detection approaches would be preferable to ensure that the model only makes predictions on certain (ID) images while decreasing the number of uncertain (OOD) images by color transfer.

The literature also showed that explainability is still an unsolved research task, which is often approached by visualizing explanations for single inputs. However, it misses quantification.

To date, no model for prostate cancer survival prediction has been developed that provides accurate predictions from only histopathology images, explains its predictions, and includes a robustness analysis on several external or differently biased datasets.

To address these existing challenges, this work will focus on the following research topics:

- Deriving the best-fit model for Gleason classification for a given dataset based on the presented state-of-the-art computer vision architectures (see Table 2.2.1) and Gleason classification models (see Table 2.2.3). This analysis can be found in chapter 4 *Gleason grade prediction*.

- Developing a survival prediction model that, in terms of discrimination and calibration, accurately predicts survival curves using only histopathology images as input. This model is explained in chapter 5 *Survival prediction* and compared to two of the presented state-of-the-art survival prediction models, namely the aforementioned DeepConvSurv and CDOR (see Table 2.2.4).

- Exploring how to obtain predictions that are robust toward changes in dataset bias by combining OOD detection with color transfer. The OOD approach is based on Lee et al. (2018) and Sun et al. (2022), whereas for color transformation, histogram matching and the color adaptation proposed by Macenko et al. (2009) are extended (see section 2.2.5 *Robustness*). This novel approach can be found in chapter 6 *Robustness*.

Chapter 3

Datasets

The datasets used throughout this thesis comprise images of TMA spots obtained after RPE. Since tissue appears similar at biopsy and after RPE, a model trained on RPE tissue is expected to be extendable to process biopsy images, which is the long-term objective for a prostate cancer decision support system. One advantage of these TMA datasets over biopsies is that the patients are comparable since the time of tissue acquisition corresponds to the time of RPE. Also, TMA spot images are smaller in size, therefore, containing less misleading information than WSIs (e.g., background, stroma).

For this thesis, two different dataset sources are available: several non-public internal datasets are provided by the University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany (see section 3.1 *Internal datasets*), while an external dataset is provided by the New York University through the Prostate Cancer Biorepository Network (PCBN; see section 3.2 *External test dataset*).

For all internal datasets, survival annotations and Gleason scores per patient are available from the EHRs, while an experienced pathologist annotated single images with Gleason scores only in one subset. The internal dataset is exceptional for prostate cancer survival prediction, as it is large in size (more than 17,000 patients) and labeled with objective relapse times. That enables training neural network survival models and obtaining meaningful evaluations. Several subsets emulate differences in data acquisition protocols. These are thus suitable for the research questions as they allow a thorough evaluation of robustness and generalizability.

The external dataset is smaller in size (204 patients) and also comprises survival labels. That dataset is used to further evaluate the generalizability to a dataset bias different from the internal datasets.

All datasets and subsets are described in detail in the following.

3.1 Internal datasets

The internal datasets are provided by the UKE Institute of Pathology. The TMA spots have a diameter of 0.6 mm, are 2.5 µm thick, and are stained with hematoxylin for 4 minutes and with eosin for 1:20 minutes unless stated otherwise. A Leica Aperio AT2 scanner[1] is used to digitize the TMA spots with a 20x magnification objective, but the images are digitally magnified to 40x. One dataset is scanned with a 3D Histech scanner[2]. The resulting images are either provided as portable network graphics (.png) files or converted to .png from tagged image file format (.tiff) files. The images are 2490 × 2490 to 3181 × 3181 pixels in size. Patients included in this study underwent an RPE in the UKE between 1992 and 2014. All personal information was removed, so no identification of the patients is possible from these datasets.

For this thesis, there are two kinds of internal datasets, which both comprise TMA spot images: The dataset named **Gleasonaut** contains individual Gleason annotations per TMA spot for Gleason classification, while the **Survival dataset** contains Gleason annotations only for the complete patient, i.e., the whole prostate, along with the patients' times to relapse. The Survival dataset further comprises several subsets. There is an overlap between the patients in the Gleasonaut and the Survival dataset, however, since the TMA spot images differ and no patient IDs are available, it is impossible to match the patients. The composition of all datasets is described in the following.

3.1.1 Gleasonaut

The "Gleasonaut" contains TMA spot images with individual Gleason labels per image. One experienced pathologist annotated each spot with two Gleason grades, sometimes giving a tertiary Gleason grade. The algorithm developed in this thesis is supposed to be applied prior to RPE in the future and thus to biopsy images. Since no tertiary grade is assigned to biopsies, the tertiary grade is integrated into the secondary Gleason grade in this thesis (e.g., 3+4 Tert.5 → 3+5). Even though the pathologist is very experienced, the labels must be considered subjective since some Gleason grades are ambiguous.

This dataset for Gleason grade classification comprises 7 TMAs, each with 195-

[1] `https://www.leicabiosystems.com/de/digitalpathologie/scannen/` (last accessed November 21, 2022)

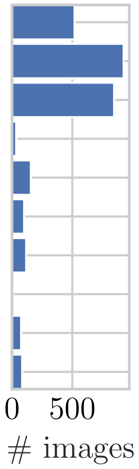
[2] `https://www.3dhistech.com/research/pannoramic-digital-slide-scanners/pannoramic-1000/` (last accessed November 21, 2022)

520 spots, yielding 2,976 images. Images that are corrupted or include (almost) no tissue are dismissed manually, which leaves 2,961 images. For 91 images, no Gleason annotation is available. These are further dismissed, leading to a dataset size of 2,870 images. Since all images are from distinct patients, the number of images equals the number of patients.

6 TMAs are shuffled and split into a training, a validation, and a test set (75 % - 15 % - 15 %). The splitting is stratified and keeps a similar ratio of Gleason grades across the datasets. One TMA (TMA 13.1D) is omitted from this split to evaluate how the algorithm performs on unseen data from a completely separate TMA, with possibly unseen staining variation. The left-out TMA is the one with the fewest images. The data split according to the Gleason scores is shown in Table 3.1.1. It is unbalanced, with most images showing low Gleason grades or benign tissue.

Table 3.1.1: Gleasonaut: Number of patients, i.e., images, per Gleason grade and dataset split. The bar chart on the right illustrates the total numbers. #: number, test13: test13.1D dataset.

ISUP	Gleason grade	training	validation	test	test13	**total**
0	0	343	74	74	28	519
1	3+3	611	131	131	53	926
2	3+4	555	119	120	53	847
4	3+5	21	5	5	4	35
3	4+3	102	22	22	11	157
4	4+4	58	13	13	17	101
5	4+5	72	16	16	14	118
4	5+3	4	1	1	0	6
5	5+4	49	11	11	5	76
5	5+5	55	12	12	6	85
total		1,870	404	405	191	**2,870**



3.1.2 Survival dataset

Each TMA spot image in the Survival dataset is annotated with its patient's recurrence-free survival time. That denotes the time from RPE to BCR, which is available from the EHR. A BCR is defined as a significant rise in PSA value after RPE (Lobel, 2007). Further, the censoring status of a patient is available, stored as censored or uncensored. For censored patients, the occurrence or time of BCR is unknown, and the last known time without relapse is recorded. Censoring is described later in more detail (see section 5.2 *Censoring*). For patients in this Survival dataset, a Gleason label is only available for the whole prostate, not for the individual TMA spots. Since label acquisition is less laborious for this dataset, it comprises more patients than the Gleasonaut described before, where a pathologist had to annotate each image manually.

The Survival dataset includes patients with unknown relapse or censoring times and tissue spots with artifacts. Therefore, all sub-datasets need to be filtered and cleaned before usage. First, patients are removed if their relapse time or censoring status is unknown since they cannot be used for analysis. Furthermore, images with little to no tissue or extreme artifacts, such as overlapping tissue, are removed by manual quality control. Another filtering criterion is applied based on the assumption that not all images are equally informative since a TMA spot only covers a part of the whole prostate and might miss the malignant region. That may result in tissue images not representative of the patient's disease status and outcome. Since Gleason scores per spot (image) are not available to evaluate if the TMA spot grade matches with the aggressiveness seen in the whole prostate, a neural network was used to predict the Gleason score (as ISUP score) for individual TMA spots. The Gleason score prediction per image and the annotation per patient are compared, and an image is omitted if its Gleason prediction is 'no cancer', but the patient's annotated Gleason score is greater than '3+3', his PSA value is greater than $4\frac{\text{ng}}{\text{ml}}$, and he had a relapse within 2 years after the RPE. This criterion filters out the most extreme cases, which are expected to reduce the survival model's ability to learn important image features. It is named FilterRepr (filter representative images) in the following for reference.

The Survival dataset comprises several sub-datasets, which vary in tissue section per patient, scanner, tissue thickness, or staining. The primary dataset used throughout this thesis comprises a single TMA spot per patient and is named **Surv1**. A second dataset, including the same patients but tissue taken from a different region in the patient's prostate, is named **Surv2**. The **SurvHetero** includes multiple images per patient, with partly overlapping patients to Surv1, but distinct images. For SurvHetero, the annotations are only known to the UKE Institute of Pathology. The parameters concerning diameter, staining time, tissue thickness, and scanner for digitization described above in section 3.1 *Internal datasets* apply to Surv1, Surv2, and SurvHetero. For a deeper analysis of the network performance when the tissue thickness or staining time changes, three different datasets, **SurvThin**, **SurvThick**, and **SurvLongStain** are available, summarized as **SurvDiff**. **SurvScan** is scanned with a different scanner. The great variety of images, particularly the emulation of different data acquisition techniques for a single set of patients, is unique and valuable for robustness analyses. The following sections describe the details of the different internal datasets, which are also summarized in Figure 3.1.1.

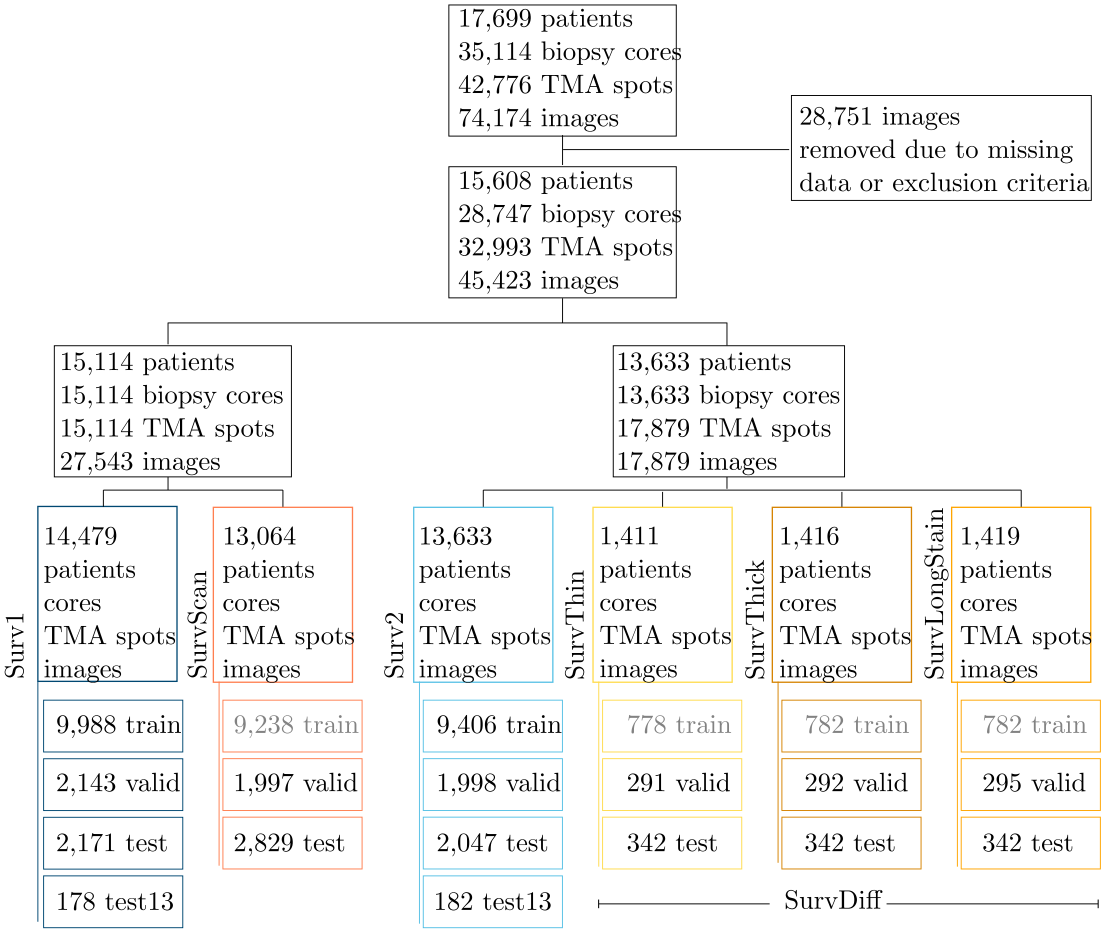


Figure 3.1.1: Composition of the internal survival datasets of which the annotations are not blinded, i.e., excluding SurvHetero. The greyed-out training datasets are not used throughout this thesis but are listed for completeness. The indicated colors per dataset are reused in section 6.4 *Experiments* for clearer visualization. train: training, valid: validation, test13: test13.1D dataset. TMA: tissue microarray.

Surv1 In Surv1, 17,230 images with prostate tissue are available (2,997 images without tissue were neglected). This dataset was further reduced to obtain a clean dataset that can be used for model training as follows: 1,748 of these patients have an unknown relapse time, and 1,741 have an unknown censoring status, which is why these are excluded. Additional 345 images that contain little tissue or are of poor quality (e.g., artifacts in the image) are omitted. The additional Gleason filtering criterion, FilterRepr, is met by 709 patients who are further excluded. The final Surv1 comprises 14,479 patients since some patients fit multiple of the exclusion criteria mentioned above.

The remaining patients are split into 70 % training dataset, 15 % validation dataset, and 15 % test dataset, again leaving out one TMA block (13.1D) as a separate test set. The datasets are stratified by annotated prostate Gleason score to obtain equal distribution of cancer grades across the splits. Table 3.1.2 (a) shows the numbers of censored and uncensored patients per dataset split before and Table 3.1.2 (b) after data cleaning. The distribution of relapse times for each dataset is shown in Figure 3.1.2. It can be seen that the distributions are similar for the training, validation, and test sets, and only the small test set obtained from a single TMA has a different distribution. Most relapses and censored events occur in the first months after RPE, with a median of 26.8 months and a mean of 35.9 months for relapses. In Figure 3.1.3, one random image of each TMA is shown. It is apparent that the staining differs between TMAs.

Table 3.1.2: Number of censored and uncensored patients in Surv1 per dataset split, before and after data cleaning. $c=0$ uncensored with relapse, $c=1$ censored without (known) relapse, and $c=-1$ unknown censoring status.

(a) Surv1, before cleaning.

	training	**validation**	**test**	**test13.1D**	**total**
$c=0$	2,482	558	541	45	3,626
$c=1$	8,217	1,732	1,770	144	11,863
$c=-1$	1,217	255	260	9	1,741
total	11,916	2,545	2,571	198	**17,230**

(b) Surv1, after cleaning.

	training	**validation**	**test**	**test13.1D**	**total**
$c=0$	1,965	445	429	36	2,875
$c=1$	8,023	1,698	1,742	141	11,604
total	9,988	2,143	2,171	177	**14,479**

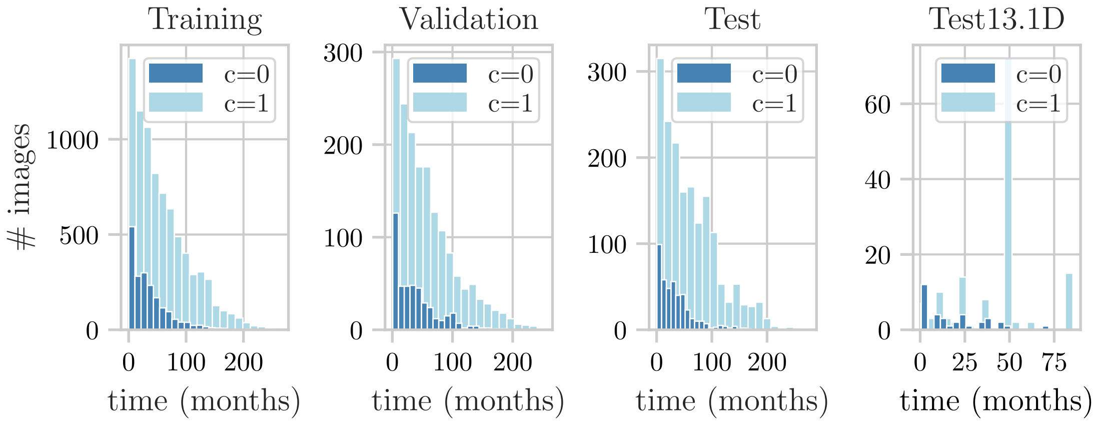


Figure 3.1.2: Histograms of the distribution of time (in months) from RPE to BCR or censoring time for each dataset split of Surv1. Dark blue ($c=0$) are uncensored patients with relapse, and light blue ($c=1$) are censored patients without (known) relapse. #: number, RPE: radical prostatectomy, BCR: biochemical recurrence.

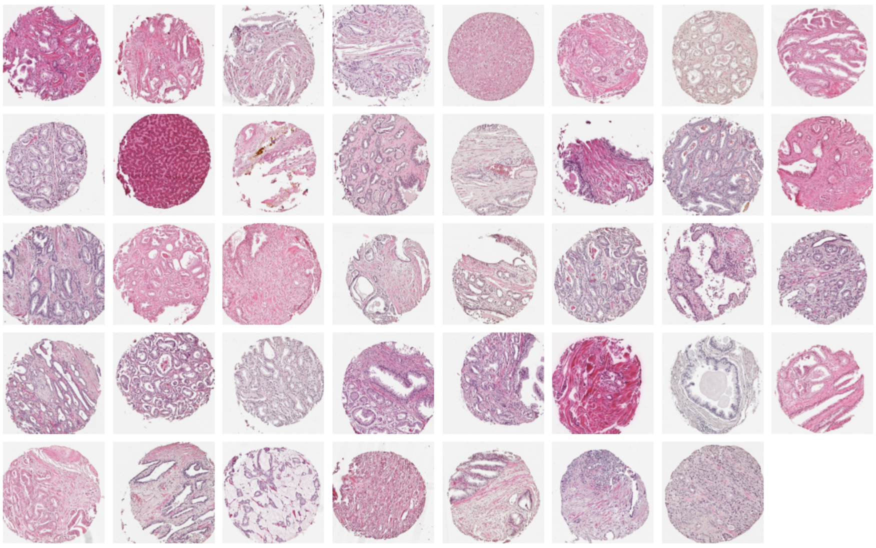


Figure 3.1.3: Sample images per TMA given in Surv1.

Surv2 Surv2 was preprocessed with the same steps as Surv1, the original and final data distribution are shown in Table 3.1.3. Here, it is important to note that the same patients are included in this dataset as in Surv1, hence, the patients are split into the same training, test, and validation sets as before. The tissue samples are, however, extracted from a different region in the prostate. In contrast to Surv1, this dataset is stained and digitized at a single point in time, making the appearance more homogeneous, but different from Surv1, as can be seen in Figure 3.1.4.

Even though both Surv1 and Surv2 initially contain the same patients, the image filtering and selection steps result in some patients for whom only a single image of Surv1 or Surv2 is available, while the majority of patients have two images. For model performance comparisons, it is interesting to create a dataset only including patients that have two images, one from Surv1 and one from Surv2. This dataset is named **SurvMulti** and summarized in Table 3.1.4. Note that the number of images is indicated, which is twice the number of patients.

Table 3.1.3: Number of censored and uncensored patients in Surv2 per dataset split, after data cleaning, $c=0$ uncensored with relapse, $c=1$ censored without (known) relapse.

	training	**validation**	**test**	**test13.1D**	**total**
$c=0$	2,290	469	507	43	3,309
$c=1$	7,116	1,529	1,540	139	10,324
total	9,406	1,998	2,047	182	**13,633**

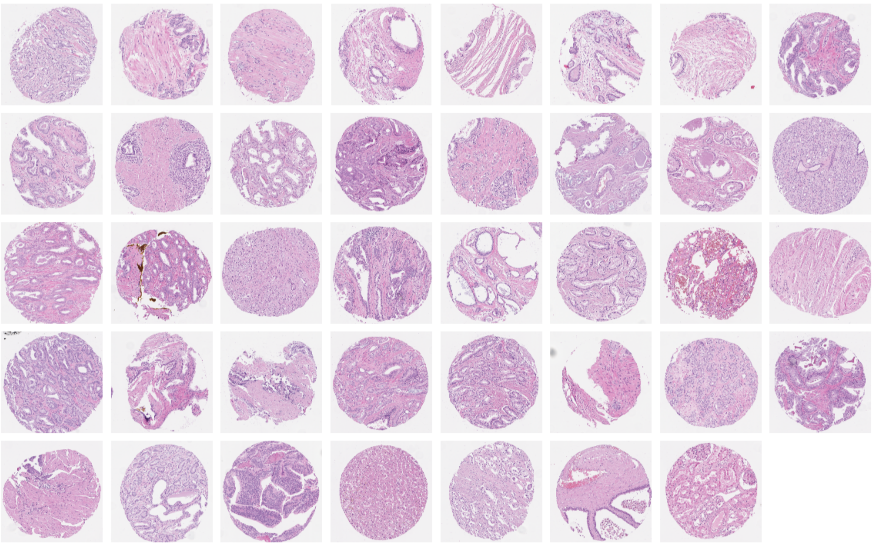


Figure 3.1.4: Sample images per TMA given in Surv2.

Table 3.1.4: Number of censored and uncensored images in SurvMulti (only patients that have an image in both Surv1 and Surv2) per dataset split, after data cleaning, $c=0$ uncensored with relapse, $c=1$ censored without (known) relapse.

	training	**validation**	**test**	**test13.1D**	**total**
$c=0$	3,444	754	770	70	5,038
$c=1$	13,906	3,002	3,020	274	20,202
total	17,350	3,756	3,790	344	**25,240**

SurvDiff Furthermore, tissue cores of five TMA blocks, which were also part of Surv2, were sliced and stained again, emulating different data acquisition processes. While in Surv1 and Surv2, each tissue slice is 2.5 µm thick, for this additional dataset, tissue slices with 1 µm (**SurvThin**) and 10 µm (**SurvThick**) were cut. A third set, **SurvLongStain**, contains images of tissue that was sliced again with 2.5 µm but stained approximately 10 times longer than usual: 40 min with hematoxylin and 10 min with eosin instead of 4 min and 1:20 min, respectively. These tissue samples were obtained by subsequently cutting slices from the same core as was used in Surv2. Therefore, the structures in the images of SurvThin, SurvThick, and SurvLongStain for the same patient look similar (see Figure 3.1.5). The different methods lead to images with lighter (SurvThin), darker (SurvThick), or more saturated appearances (SurvLongStain).

As was shown by Chlipala et al. (2021), different tissue thicknesses and staining protocols influence the optical density of the images, which could reduce the prediction performance of digital image analysis models.

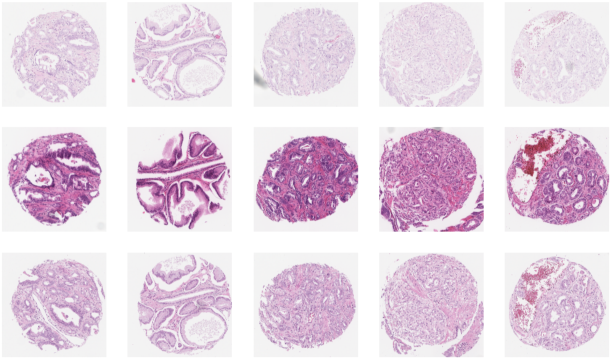


Figure 3.1.5: Sample images per TMA given in SurvDiff. Each column corresponds to the same patient and tissue core, of which slices are cut subsequently. First row: SurvThin, second row: SurvThick, third row: SurvLongStain.

Table 3.1.5: Number of censored and uncensored patients in SurvDiff per dataset split, after data cleaning, train: training set, val: validation set, $c=0$ uncensored with relapse, $c=1$ censored without (known) relapse.

	SurvThin			**SurvThick**			**SurvLongStain**			**total**
	train	**val**	**test**	**train**	**val**	**test**	**train**	**val**	**test**	
$c=0$	226	81	94	229	81	94	229	81	94	1,209
$c=1$	552	210	248	553	211	248	553	214	248	3,037
total	778	291	342	782	292	342	782	295	342	**4,246**

SurvDiff will therefore be used to test the generalizability of models trained on Surv1. It may further approximate the achievable performance on external datasets, for which the data-acquisition conditions are not known in advance.

Since the images included in SurvDiff belong to the same patients as in Surv1 and Surv2, the patients are assigned to the same training, validation, and test sets as before. However, the training split is not used since only the evaluation is of interest. The dataset numbers are shown in Table 3.1.5 for SurvThin, SurvThick, and SurvLongStain. Only a few images are removed due to the filtering criterion so that the three datasets are almost equal in size.

SurvScan In addition to the staining time and tissue thickness, another source of variation for histopathology datasets is the scanner used for digitizing the tissue. The datasets Surv1, Surv2, SurvDiff, and SurvHetero described above are all obtained with a Leica Aperio scanner. For SurvScan, the same tissue as in Surv1 is digitized by a different scanner, the 3D Histech. However, only 33 out of 39 TMAs are available, which leads to a reduced dataset size. The patients' images are split the same way as in Surv1 into training, validation, and test sets. Again, the training set will not be used in this thesis. SurvScan is also cleaned as described above using FilterRep and removing images with little or no tissue, resulting in the dataset distribution shown in Table 3.1.6. An impression of the images is given in Figure 3.1.6. The color varies across TMAs and differs from the previous datasets as it has an orange-pink tint instead of violet.

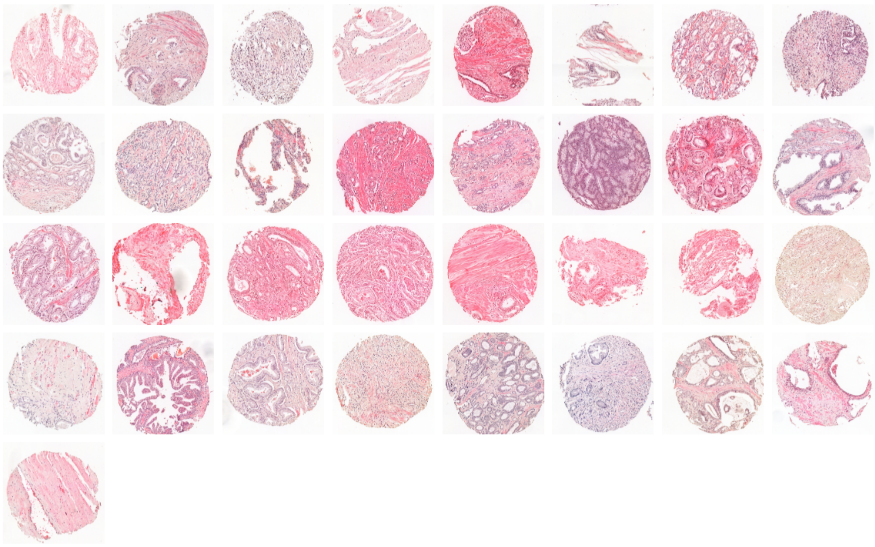


Figure 3.1.6: Sample images per TMA taken with the 3D Histech scanner in SurvScan.

Table 3.1.6: Number of censored and uncensored patients in SurvScan per dataset split, after data cleaning, $c = 0$ uncensored with relapse, $c = 1$ censored without (known) relapse.

	training	**validation**	**test**	**total**
$c=0$	2,309	532	439	3,280
$c=1$	6,929	1,465	1,390	9,784
total	9,238	1,997	1,829	13,064

Surv1AddInfo Besides the prostate tissue TMA spot image, additional information is available for the patients in the internal Survival dataset. These can be used for a clinical decision support system in addition to the H&E images. All clinical features that the dataset comprises are listed in Table 3.1.7 along with their value ranges.

However, some of that additional information is only available after RPE. When adding clinical information to improve a treatment decision model, only those features available before the treatment of a patient have to be selected. Therefore, the feature overview table includes information on whether or not a feature is available at the time of biopsy. For example, whether cancer cells are present in the resection margin can only be known after RPE. In contrast, the patient's age and PSA value are readily available at the time of biopsy. In the Survival dataset, Gleason grades are only available for the whole prostate and are neglected since they comprise more information than is available during the biopsy. Tumor volume and tumor diameter are typically measured after the removal of the prostate. However, research on the approximation of tumor volume and diameter before surgery, for instance, with MRI, exists (Mazaheri et al., 2009; Hsieh et al., 2021). Therefore, these two features may be available at the time of biopsy as model input features.

Tumor volume, tumor diameter, and PSA value are not available for all patients in Surv1. 3,428 patients do not have information about tumor diameter, 4,818 patients lack information about the tumor volume, and 71 patients do not have a recorded PSA value. Therefore, the number of patients in Surv1 reduces to 7,806. An overview of the remaining training, validation, and test set patients in Surv1 is given in Table 3.1.8. This dataset is from now on named Surv1AddInfo.

To estimate whether the data already reveals an association between features and relapse-free survival time, KM curves can be used. The theory of KM curves is described in detail in section 5.3 *Population-based survival prediction.* In short, patients are split into groups according to a single feature, and for each group, the

Table 3.1.7: Additional patient features that are available for parts of the internal Survival dataset, Av. at biopsy: value available at time of biopsy, #: number, PSA: prostate-specific antigen, RPE: radical prostatectomy.

Feature	**Value range**	**Av. at biopsy**
group of PSA level before RPE	<4, 4-10, 10.01-20, >20.01	yes
PSA level before RPE	up to 1,101	yes
patient age at RPE	37.5 - 80.8	yes
size of the tumor	pT1 - pT4	no
positive resection margin	yes/no/unknown	no
vessel invasion	yes/no/unknown	no
lymph vessel invasion	yes/no/unknown	no
lymph node invasion	yes/no/unknown	no
# lymph nodes with metastases	up to 23	no
# lymph nodes total	up to 132	no
diameter of tumor	up to 92	indirect
volume of tumor	up to 3,125	indirect
Gleason pattern	0 - 5+5	yes
% Gleason 3 in prostate	0 - 100	yes
% Gleason 4 in prostate	0 - 100	yes
% Gleason 5 in prostate	0 - 100	yes

Table 3.1.8: Number of censored and uncensored patients in Surv1AddInfo per dataset split, after data cleaning, $c = 0$ uncensored, with relapse, $c = 1$ censored, without (known) relapse.

	training	**validation**	**test**	**total**
$c=0$	890	202	209	1,301
$c=1$	4,569	964	972	6,505
total	5,459	1,166	1,181	7,806

number of patients surviving (relapse-free) is plotted against time. A correlation between the group selection feature and the outcome exists if the KM curves for different groups separate well. When comparing more than two groups, the curves also need to be in the correct order to be meaningful.

For Surv1AddInfo, the patients are grouped according to their features, and the resulting KM curves are plotted in Figure 3.1.7. The last relapse in this dataset occurs at 83.9 months. The patients are split into four quantiles according to the single features, yielding an equal number of patients per group. A higher PSA value correlates with an earlier relapse time since the curves drop earlier the higher the PSA value is in a group. Accordingly, patients with low PSA values ($<5\frac{\text{ng}}{\text{ml}}$) survive longer relapse-free than patients with higher values. The same holds for tumor volume and tumor diameter. The larger the tumor, the lower the survival probability. These correlations are also supported by findings in the literature (Eichelberger et al., 2005; Kattan et al., 1999; Stamey et al., 1999). For the age, however, the correlation is not as clear in Surv1AddInfo since the curves for all patients are very close. Only for patients older than 70 years, a slight difference is visible as the KM curve remains below the others.

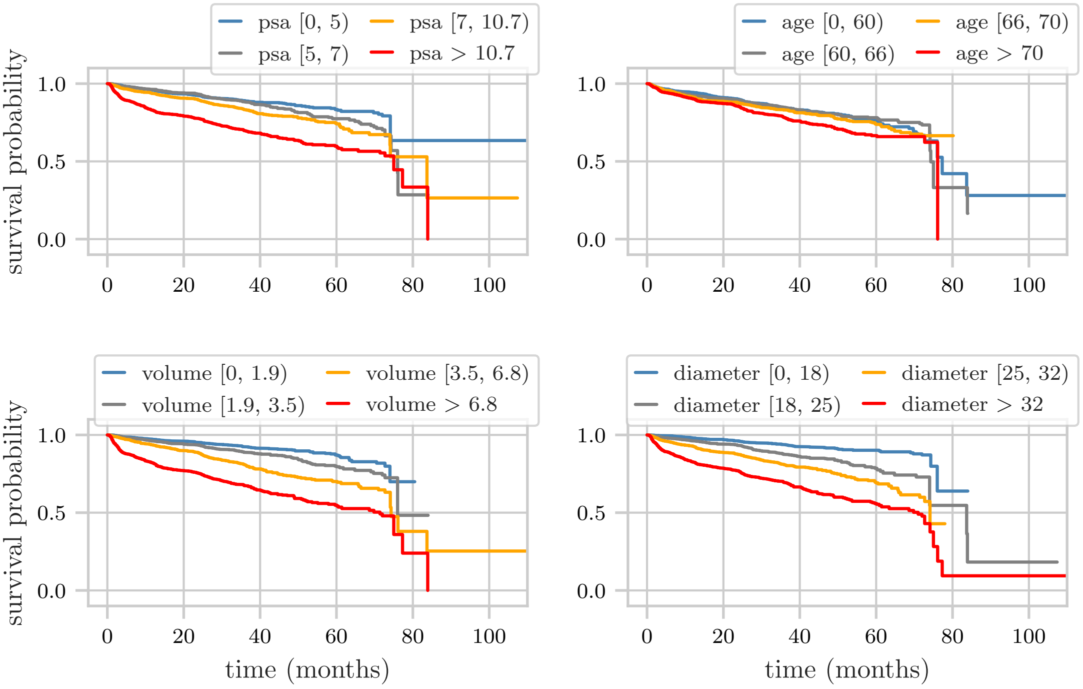
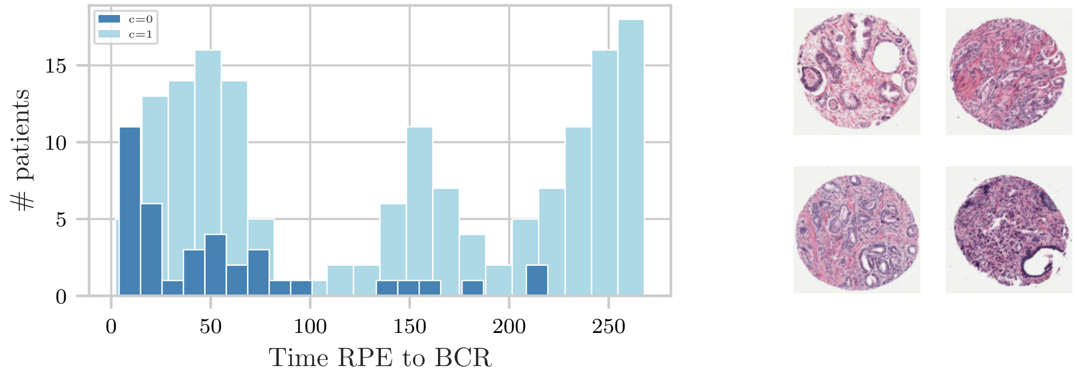


Figure 3.1.7: Kaplan-Meier curves showing how PSA value (in $\frac{\text{ng}}{\text{ml}}$), age (in years), tumor volume (in ml), and tumor diameter (in mm) relate to relapse-free survival time. The plots are cut at 110 months on the y-axis for clarity and without loss of information since only a single patient has a later censoring time. The descriptions of the colors in the line plots are provided in the respective legends.

SurvHetero Another test dataset in the internal Survival dataset is SurvHetero. Like Surv1, the tissue is 2.5 µm thick and scanned with the Leica Aperio. However, it includes multiple TMA spots per patient and therefore resembles the situation of a biopsy more closely, where up to twelve cores are extracted and analyzed. SurvHetero is used to evaluate whether the performance of a model is affected when using multiple images as input during inference. It includes 828 patients that are partly overlapping with the patients in Surv1, but the included images are different. Per patient, 2-6 images are available, in total 4,181. In contrast to the datasets mentioned above, all annotations and additional patient information are blinded. This dataset is also not split but used completely as a test set. In order to evaluate any model on SurvHetero, the predictions are sent to the UKE Institute of Pathology and evaluated there.

3.2 External test dataset

The external dataset also comprises TMA spot images of prostate cancer patients. The dataset is accessed through the Prostate Cancer Biorepository Network (PCBN) and is referred to as SurvPCBN[3]. It is provided by the New York University. This dataset contains 204 patients with up to nine images per patient and includes censoring information as well as the patients' time to BCR. In total, 725 images are available, of which 702 can be used for evaluation after applying FilterRep and filtering out images with few or missing tissue. The patients' survival time distribution is shown in Figure 3.2.1 (a). The images are arranged in four TMA blocks, containing 35, 218, 237, and 235 images, respectively. The tissue is of 0.6 mm diameter and 5 µm thickness. Example images per TMA are shown in Figure 3.2.1 (b). Since SurvPCBN is only used as a test set, no dataset split is performed. The number of censored and uncensored patients and images is summarized in Table 3.2.1.



(a) Distribution of time from radical prostatectomy (RPE) to biochemical recurrence (BCR) or censoring time in SurvPCBN in months. $c=0$ uncensored, $c=1$ censored, #: number.

(b) Sample images per TMA given in SurvPCBN.

Figure 3.2.1: Graphical overview of SurvPCBN.

[3] This work is supported by the Department of Defense Prostate Cancer Research Program, DOD Award No W81XWH-18-2-0013, W81XWH-18-2-0015, W81XWH-18-2-0016, W81XWH-18-2-0017, W81XWH-18-2-0018 and W81XWH-18-2-0019 PCRP Prostate Cancer Biorepository Network (PCBN).

Table 3.2.1: Number of censored ($c = 1$) and uncensored ($c = 0$) patients and images in SurvPCBN.

	patients	**images**
c=0	41	123
c=1	163	579
total	204	702

Comparison of the internal and external datasets The tissue in the external dataset is 5 µm thick, which is thicker than the tissue in Surv1, but thinner than the tissue in SurvThick. The spots in both datasets are 0.6 mm in diameter and stained with H&E. From visual inspection, there seems to be a different color variation in the images, as the images in SurvPCBN appear more red than violet. The images in SurvPCBN are 1817 × 1817 pixels in size and therefore smaller than in Surv1 (images up to 3181 × 3181 pixels). The survival time distributions of Surv1 and SurvPCBN differ since SurvPCBN comprises more censored patients with a long follow-up record (after 200 months). Similar to SurvHetero, SurvPCBN comprises multiple images per patient.

3.3 Data preprocessing

The internal dataset images are provided as single cutout spots from digitized TMA images, whereas the TMA images of the external dataset needed to be cut into single spots using QuPath (Bankhead et al., 2017). Therefore all available images are square but of different sizes (2490 × 2490 to 3181 × 3181 pixels in the internal, 1817 × 1817 pixels in the external dataset). The images show circular tissue areas on white background. Depending on the model used later, different preprocessing steps and data augmentation methods are applied to all images or individually. These methods are described in the following.

Cutting centerpieces Since white background does not include any cancer-related information, a simple method is used to improve the foreground-to-background ratio: Images are reduced to a central square of their circular spot so that most background is removed, but most tissue is kept. It is assumed that the gain by removing background noise that does not include information outweighs the information loss by removing the margin tissue. This will also be analyzed in an experiment in chapter 4 *Gleason grade prediction*. Cutting the centerpieces results in images of size 2048 × 2048 pixels for the internal datasets and 1024 × 1024 pixels for SurvPCBN. The centerpiece is cut by fitting an ellipse to the tissue spot and cutting the square around the ellipse's center since not all spots are perfectly centered. All single steps are shown in Figure 3.3.1. First, the RGB image (a) is converted to grayscale (b) using the OpenCV package for Python (Bradski, 2000). It is then binarized with Otsu-thresholding (c) (Otsu, 1979), and an ellipse is fitted (d) since not all tissue spots are perfectly round. The ellipse's center is used as a center point for the resulting square image (e)+(f). This preprocessing step is applied to all images from all datasets. If it is not indicated otherwise, referring to a dataset image alludes to the centerpiece.

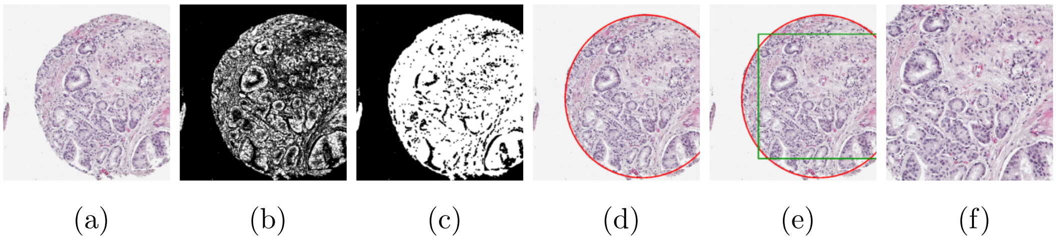


Figure 3.3.1: An example of the preprocessing steps to cut image centerpieces and remove most of the white background. The original image (a) is converted to grayscale (b), and Otsu-thresholding is applied (c). Then, an ellipse is fitted to the tissue spot, here projected in red onto the RGB image (d). In the last step, a center square is cut (e) + (f). Taken from Dietrich et al. (2021).

Patching For some experiments, the image is cut into smaller pieces, also called patches or tiles, instead of using the whole image. For this, an even grid is placed on the image so that non-overlapping patches can be cut (e.g., 8×8 patches of size 128 × 128 pixels each for an image of size 1024 × 1024 pixels).

Data augmentation Data augmentation is applied during model training to increase the variability and artificially expand the given dataset. This includes morphologic transformations, i.e., 90-degree rotation and horizontal or vertical flip.

Normalization All images contain pixel values in the range $[0, 255]$. Since ANNs are trained best when the input values are closer to 0, all images are normalized (Kim, 1999). In the work for this thesis, all images are normalized to a range of $[0, 1]$ simply by dividing the pixel values by the maximum possible pixel value, 255.

3.4 Main dataset challenges

The datasets described above contain several challenges for deep learning algorithms, possibly limiting the performance. A summary of the main challenges due to data size, noise, and uncertainty is given here. Most of the listed dataset biases are unavoidable during data acquisition. However, it is important to be aware of these challenges and evaluate how to treat these. Throughout the thesis, this aspect will be recurring.

Image size After the preprocessing steps, all internal datasets have images with sizes of 2048 × 2048 pixels. The size can be a challenge to computational resources. Reducing image size is a solution, but it comes at the cost of information loss due to decreased resolution. Furthermore, image sizes may vary between dataset sources. For instance, the SurvPCBN contains smaller images of size 1024 × 1024 pixels. Due to the lower resolution, some detail may be lost.

Label noise The Gleasonaut suffers from label noise since the images are annotated with Gleason scores by a single pathologist. As described before, the Gleason score is highly subjective, which likely reduces the accuracy of any model predicting Gleason scores but does not make the labels infeasible as ground truths.

Even though the Survival dataset is annotated with an objective, measurable label, i.e., the reported time to relapse, the time points are noisy to a certain extent. If a patient has regular follow-ups and at one time a rise in PSA value is measured, it is impossible to state at which month exactly between the last and the current follow-up the relapse occurred. The time of the relapse detection serves as an annotation.

Dataset bias Since H&E staining does not follow the same protocol across hospitals, biases might occur due to staining time, tissue thickness, reagents, digitization, or tissue resolution. A model's generalizability to differences in image appearance needs to be evaluated carefully since biases such as unseen colors might occur in clinics. A data-driven model trained on an internal dataset also needs to be evaluated on an external dataset. A model generalizes well if the performance on the external dataset reaches similar values as on the internal dataset. If a generalization is not achievable, minimum requirements for unseen datasets need to be defined to assure high performance. Besides bias from data acquisition, a bias in the patient sample might occur. All patients in the Survival dataset had an

RPE. Therefore, all patients suffered from relatively severe cancer. It needs to be investigated whether data-driven models trained on these images have difficulties generalizing to patients with less severe cancers.

Dataset noise The TMA spots for the Survival dataset are extracted from the tumorous region of the prostate. However, such a spot only covers a small percentage of the complete, inhomogeneous prostate tissue. The malignant tissue of a diseased patient may be missed when extracting a TMA core, resulting in a spot with healthy prostate tissue for a patient with malignant cancer. In that case, the spot itself may not give a good indication of the patient's health status. This kind of dataset noise could be circumvented by a manual effort of a pathologist sorting through all dataset images, which is time-consuming and, therefore, not done in this thesis. A different option to reduce the risk of missing cancerous areas is including multiple tissue samples per patient. Furthermore, some images include artifacts like overlapping tissue, blurry effects, or broken tissue. Careful dataset cleaning removes most of the severely affected images, but probably some cases are still kept inside the dataset, and these are expected to influence the model's performance.

3.4.1 Discussion

The presented datasets allow for comprehensive analyses and experiments, which are explained in the following chapters. The large dataset size enables training artificial neural networks, but the presented dataset challenges and proposed preprocessing steps need to be addressed. The influence of preprocessing steps, like cutting centerpieces and varying the image size, is analyzed in chapter 4 *Gleason grade prediction*, where different possibilities to approach Gleason grade prediction are compared. In chapter 5 *Survival prediction*, it is explored how including additional patient features influences the performance of survival prediction. It is further evaluated how to exploit multiple images from one patient for a prediction. The robustness toward dataset biases is explored using the above-presented datasets with differing data acquisition protocols in chapter 6 *Robustness*. In all experiments, the presented internal and external datasets are used. In particular, the Gleasonaut is used in chapter 4 *Gleason grade prediction*, while the Survival dataset and SurvPCBN are used in both chapter 5 *Survival prediction* and chapter 6 *Robustness*.

Chapter 4

Gleason grade prediction

4.1 Motivation

Pathologists assign Gleason grades to prostate cancer tissue based on the size and shape of prostate glands. Since this manual process is time-consuming and suffers from high interobserver variability, it is expected to benefit from (semi-) automation. In this chapter, the research question R2 is addressed concerning an automated Gleason pattern prediction for the Gleasonaut, which is described in chapter 3 *Datasets*.

Nagpal et al. (2019), Ström et al. (2020), and Bulten et al. (2020) showed that CNNs can outperform pathologists in the task of Gleason grading on several datasets. Therefore, also in this thesis, a CNN will be applied to assign Gleason grades to individual prostate cancer images.

The intention in the context of this thesis is to start with classification as a straightforward computer vision task, for which models already exist off-the-shelf, before moving on to survival prediction afterward. The Gleason classification also serves as a starting point to obtain an estimate of the Gleasonaut quality because that dataset has not been used for computer vision tasks before. If the dataset is of good quality, it is expected that a CNN reaches performances similar to those reported on other datasets in the literature (see section 2.2.3 *Prostate cancer classification for histopathology images*). If the Gleason classification does not yield high performance on the given dataset, going to survival prediction in the next step may not be promising.

Furthermore, pretrained networks exist for neither survival prediction nor histopathology classification. Spanning the gap from ImageNet classification to histopathology survival prediction might be too challenging for transfer learning. First pretraining a network on Gleason classification and using that network as a starting point for the survival prediction is hypothesized to increase performance and robustness since the main characteristics of histopathology images are learned already.

Which network architecture performs best depends on the used dataset, as stated in section 2.2.3 *Prostate cancer classification for histopathology images.* The first task is, therefore, to select the best performing among state-of-the-art architectures on the Gleasonaut. Transfer learning from a model pretrained on ImageNet will be applied to reduce training time. However, several differences between Gleason grading and ImageNet classification may need to be addressed. For example, multiple patterns might be visible in one histopathology image, which in combination affect the final class (e.g., Gleason patterns 3+4 lead to ISUP 2, but 4+3 to ISUP 3).

4.2 Gleason stratification

Gleason grading can be treated as a multi-class classification problem, where one out of k classes is assigned to each image. The label per image is provided in the Gleasonaut, which is why this is a supervised learning task. Since many pretrained neural networks are available open-source, transfer learning from a model pretrained on the ImageNet dataset is easily applied. A single image can contain multiple Gleason patterns (e.g., 3+4). In order to reduce the classification to a single-label prediction, two different approaches are considered in the following:

Single Gleason pattern When a neural network is trained to predict Gleason scores on images that contain mixed Gleason patterns, patterns need to be quantified for the final classification. In order to simplify the problem, all images with mixed Gleason pattern labels are removed, and a model is trained only on images that show single patterns (Bulten et al., 2020; Arvaniti et al., 2018). Leaving out all mixed pattern images reduces the dataset to 1,527 images. It is expected that, for example, each part of an image labeled as 3+3 is of Gleason pattern 3, which means no other patterns are visible anywhere in the image. Thus, when each image region shows only one pattern, single image patches contain all the necessary information for the classification task. This is an advantage since these patches are smaller and can be used individually during training and inference. Thus, the loss of resolution when downsizing the patches is not as great as for the whole image. Disadvantages are that not all information of the image is used, and the applicability to images of mixed patterns during inference is uncertain.

The single Gleason pattern approach reduces the task to a $k = 4$-class classification problem (benign, 3+3, 4+4, 5+5). The annotations of the dataset are converted to one-hot encoded vectors for this approach, from treating no cancer as $[1,0,0,0]$ to Gleason 5+5 as $[0,0,0,1]$. Formally, for a class $c \in [0,\dots,k-1]$ the label l is given by

$$l_c = [\mathbf{1}_{j=c}],\ j = 0...k-1.$$

$\mathbf{1}$ is the indicator function taking value 1 if the condition is true, and 0 otherwise. Class $c = 0$ corresponds to benign tissue, class $c = 1$ to Gleason 3+3, and so on. The final softmax layer of a pretrained classification network will be reduced to 4 output nodes, using maximum voting for the final prediction.

ISUP classification In contrast to single Gleason pattern classification, predicting ISUP classes enables using the complete image dataset. This task is considered to be more complex since the combination and quantity of Gleason patterns influence the ISUP class. The ISUP classes 2 and 3, e.g., only differ by the amount of Gleason patterns 3 and 4. Therefore a more advanced encoding is applied for the ISUP classification. With one-hot encoding, all classes are modeled to be “equally different” and independent. In contrast, ordinal regression includes the notion that ISUP classes 1 and 2 are closer to one another than ISUP classes 1 and 3. The ISUP classification is a task with $k = 6$ classes, encoded with 5 output nodes as

$$l_c = [\mathbf{1}_{j<c}],\ j = 0...k-2$$
$$\text{with } \sum l_c = c$$

for class c. Following this convention, benign is encoded as $[0,0,0,0,0]$, and, for example, ISUP class 2 is encoded as $[1,1,0,0,0]$. In contrast to one-hot encoding, the final prediction is calculated as the sum of all output nodes.

4.3 Metrics and loss function

Metrics For the evaluation of multiclass classification, the number of correctly predicted samples can be counted relative to the total number of samples. That is captured by accuracy, which is defined as the sum of correctly predicted samples over all predictions (Zeng et al., 2010). For N samples, the accuracy

$$\text{Acc} = \frac{1}{N}\sum_{i}^{N} \mathbf{1}_{p_i=y_i} \tag{4.3.1}$$

ranges from 0 to 1, such that a perfectly accurate model reaches 1. Here, p is the predicted class, and y is the annotated class.

Accuracy, however, can be misleading under class imbalance. Another metric that is less prone to class imbalance and which is often used in medical applications is Cohen's kappa (Cohen, 1960)

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}, \tag{4.3.2}$$

with actual agreement $Pr(a)$ and chance agreement $Pr(e)$ (McHugh, 2012). It measures the number of correct classifications compared to a random guess. It was developed to measure the agreement between two physicians for objective measures, but it can also be used to measure the agreement between a model's prediction and the annotation (McHugh, 2012). Note that Cohen's kappa ranges from -1 to 1: If the model's prediction is as good as a random guess, $\kappa = 0$, if it is worse than a random guess, $\kappa < 0$, whereas a model in perfect alignment with the annotation has $\kappa = 1$.

The confusion matrix can be calculated to visualize a model's performance. It is a table showing the amount of correct and incorrect predictions and which classes are confused with each other (Somogyi, 2021).

Loss function As loss function, a categorical cross-entropy loss is applied (Rubinstein, 1999). Since the Gleasonaut is imbalanced, with more samples showing low-grade Gleason patterns than high-grade patterns, the loss is weighted with the inverse class frequency (e.g., Li et al., 2021). Thus, the applied loss is

$$L = \sum_c w_c \, y_c \log(p_c), \tag{4.3.3}$$

using all classes c, predictions p_c and annotations y_c, weighted with w_c, the inverse class frequency ($w_c = N/N_c$ for N total samples and N_c samples per class c).

4.4 Experiments

Several experiments are performed to explore the best-performing neural network model and hyperparameters on the Gleasonaut. All models for the following experiments are trained on an Nvidia Tesla V100 16GB GPU. The algorithms and models are implemented in Python3[1], using keras[2] and TensorFlow[3] as deep learning libraries (van Rossum and Drake, 2009; Chollet et al., 2015; Abadi et al., 2015). The models are trained with Nadam[4] optimizer and a weighted categorical cross-entropy loss[5] with learning rate 0.0003 (Dozat, 2016). The models are trained for 350 epochs with early stopping, so the model that achieves the highest accuracy on the validation dataset is used as the final model. Each model setup is trained three times with different initialization seeds to estimate the reproducibility and stability of the training. All shown results are obtained on the separate test set. The numbers of images per experiment and dataset split are shown in Table 4.4.1.

Table 4.4.1: Number of Gleasonaut images per class for both dataset splits. Gp: Gleason pattern, I: ISUP score, train: training, valid: validation, t13: test13.1D dataset.

Single Gleason patterns						**ISUP scores**					
Gp	train	valid	test	t13	**total**	I	train	valid	test	t13	**total**
0	343	74	74	28	519	0	343	74	74	28	519
3+3	611	131	131	53	926	1	611	131	131	53	926
4+4	58	13	13	17	101	2	555	119	120	53	847
5+5	55	12	12	6	85	3	102	22	22	11	157
						4	83	19	19	21	142
						5	176	39	39	25	279
total	1,067	230	230	104	**1,631**		1,870	404	405	191	**2,870**

4.4.1 Single Gleason grading

For the task of single Gleason grading, only the images with single Gleason patterns (0, 3+3, 4+4, 5+5) are included, reducing the dataset to 1,067 training images and 230 images in both the validation and the test dataset (see Table 4.4.1).

Architecture As shown above in section 2.2.3 *Prostate cancer classification for histopathology images*, different architectures are reported to yield the best results for Gleason classification. Thus, it is difficult to decide a priori which base CNN architecture is suited for the given dataset. Therefore, different architectures are trained here, and their performances are compared to determine which architecture

[1] `https://www.python.org/` (last accessed November 24, 2022)

[2] `https://keras.io/` (last accessed November 24, 2022)

[3] `https://www.tensorflow.org/` (last accessed November 24, 2022)

[4] `https://keras.io/api/optimizers/Nadam/` (last accessed November 24, 2022)

[5] `https://keras.io/api/losses/probabilistic_losses/#categorical_crossentropy-function` (last accessed November 24, 2022)

performs best on the Gleasonaut. Only the most common and recently used architectures ResNet-50[6], InceptionV3[7], and EfficientNetB0[8] are considered here. All architectures are pre-trained on the ImageNet dataset, and all weights are fine-tuned on the Gleasonaut. The centerpiece TMA image is used as input, resized to 224 × 224 pixels, which is the default input size. A batch size of 16 is used during training.

InceptionV3 achieves the highest test set kappa (0.68), before EfficientNet (0.64) and ResNet (0.53). The performance on the validation set reveals low generalizability of EfficientNet (a drop of 0.12 from validation to test set). The kappa of InceptionV3 is 0.01 higher on the test set than on the validation set. Regarding accuracy, ResNet again scores lowest on the validation and the test set (0.74 and 0.75). The test set accuracy is identical for both the InceptionV3 and EfficientNet architectures (0.83). However, again, the performance drop from validation to test data is greater for the EfficientNet, indicating overfitting. Based on these results, it is concluded that InceptionV3 has the best overall performance, considering accuracy, kappa, and generalizability. Therefore, an ImageNet-pretrained InceptionV3 network is used for the following experiments.

Whole image versus centerpiece As explained in chapter 3 *Datasets*, the centerpiece of each image is cut to remove the uninformative background. It is therefore evaluated if centerpiece extraction is improving classification over whole image classification. An ImageNet-pretrained InceptionV3 model is fine-tuned on the complete TMA spots in the Gleasonaut with the white background and on the cut centerpiece without background. The images with the white background are padded with white margins to obtain images of equal sizes of 2525 × 2525 pixels. As input to the model, each image is resized to 224 × 224 pixels, which is the default input size of the InceptionV3 architecture. Using only the centerpiece improves results over using the padded image in both kappa (0.68 versus 0.51 on the test set) and accuracy (0.83 versus 0.74). The gap between validation and test metrics is small (0.01 in kappa, <0.01 in accuracy). It is concluded that the performance gain of removing background is greater than the loss of information in the tissue margin. Therefore, in all other experiments, the cut centerpiece is used.

[6] `https://keras.io/api/applications/resnet/#resnet50-function` (last accessed November 24, 2022)

[7] `https://keras.io/api/applications/inceptionv3/` (last accessed November 24, 2022)

[8] `https://keras.io/api/applications/efficientnet/#efficientnetb0-function` (last accessed November 24, 2022)

Freezing layers In the literature, it has been shown that freezing the first layers of a pretrained CNN and only fine-tuning the last ones leads to good results (Raghu et al., 2019). In the experiments for this thesis, that result could not be confirmed. The performances when fine-tuning all weights and when freezing all convolutional layers to fine-tune only the classification layer are compared. When fine-tuning all weights, the classification performance is better than when freezing the convolutional layers, both in accuracy (0.19 versus 0.83 on the test set) and kappa (0.67 versus 0.68). This can be explained by the large difference between the pretraining dataset (ImageNet) and the fine-tuning dataset (Gleasonaut). The model with frozen weights does not generalize well, which is indicated by a large difference in accuracy on the validation and test sets (0.73 versus 0.19). All model weights will be fine-tuned for the following experiments since this yielded the best results.

Using a single image patch The input to pretrained neural networks is usually much smaller than the original image size of the Gleasonaut images (224×224 instead of 2048×2048 pixels). Downsizing the images results in an information loss due to reduced resolution. Using only a smaller patch of the image as input circumvents this but comes at the cost of loss in image information since a part is cut off. Using a single image patch is a valid approach when using only single-Gleason-pattern images since the whole image consists of a single pattern. Therefore, a small part of the image is also expected to include all information needed for classification. The influence of the patch size on the model's performance is evaluated here. For this experiment, a single squared patch of side lengths 800, 1000, 1400, 1600, or 1800 pixels is cut per input image randomly each epoch, which is then resized to 224×224 pixels as input for the InceptionV3 architecture. For comparison, the complete image is used as input, which has a side length of 2048 pixels.

The accuracy and kappa for the validation and test sets are shown in Figure 4.4.1 for the different patch side lengths. Both accuracy and kappa increase with patch size. This trend indicates that small patches contain too little tissue to recognize the cancer severity or that the glands are only partially visible on the patches. Another reason is that if only a part of the image shows cancerous tissue and another region is benign, a cut patch might only represent the benign tissue. This effect reduces with larger patch sizes. The mean kappa and accuracy are similar for the validation and test splits. Only the standard deviation of kappa values is greater for the test set, indicating lower stability. When using patches of size 800 × 800 pixels, the accuracy on the test set is 0.70, with a kappa of 0.48. When the complete image is used, without cutting patches (2048 × 2048 pixels), an average accuracy of 0.83 and kappa of 0.68 is reached. The rise in accuracy levels off with patches larger than 1600 × 1600 pixels, but the kappa still increases.

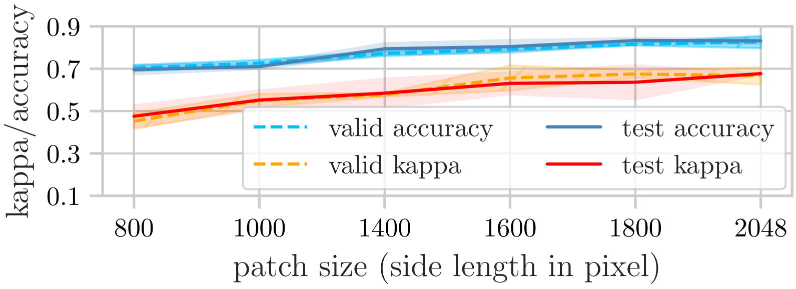


Figure 4.4.1: Comparison of model performance when training with different patch sizes. Mean accuracy is shown in blue, and mean kappa in red. The test set results are indicated by solid lines, and the validation set results with dashed lines. The shaded area indicates the standard deviation. Both results for validation and test set show similar trends. The larger the patches, the more accurate the prediction. Patch size of 2048 corresponds to using the complete image, no patches. The network input shape is kept constant at 224 × 224.

The confusion matrices for the smallest patch size (800 × 800 pixels) and the whole image are shown in Figure 4.4.2. For the smallest patch size in Figure 4.4.2 (a), the model predicts class '3+3' in most cases. Since that class is overrepresented in the dataset, these results indicate that the network cannot learn distinguishing features. Since a TMA spot labeled as malignant might still include some non-cancerous regions, cutting small patches increases the probability that these show only lower cancer grades, and thus some label noise is introduced. The high values on the diagonal of the confusion matrix in Figure 4.4.2 (b) indicate more accurate predictions for the experiment with the whole image. The model most often confuses tissue with Gleason '5+5' as benign. Besides predicting 16 % of benign images as '3+3', the model seldom predicts a higher Gleason grade than the annotation.

Input size When cutting a single patch, image information is lost, and it was shown above that using the whole image is best. However, this has to be downsized to 224 × 224 pixels to fit the default input size of ImageNet, so again, image detail is lost. The less an image is downsized, the more information it retains, which is expected to benefit the neural network. Therefore here, the effect of changing the

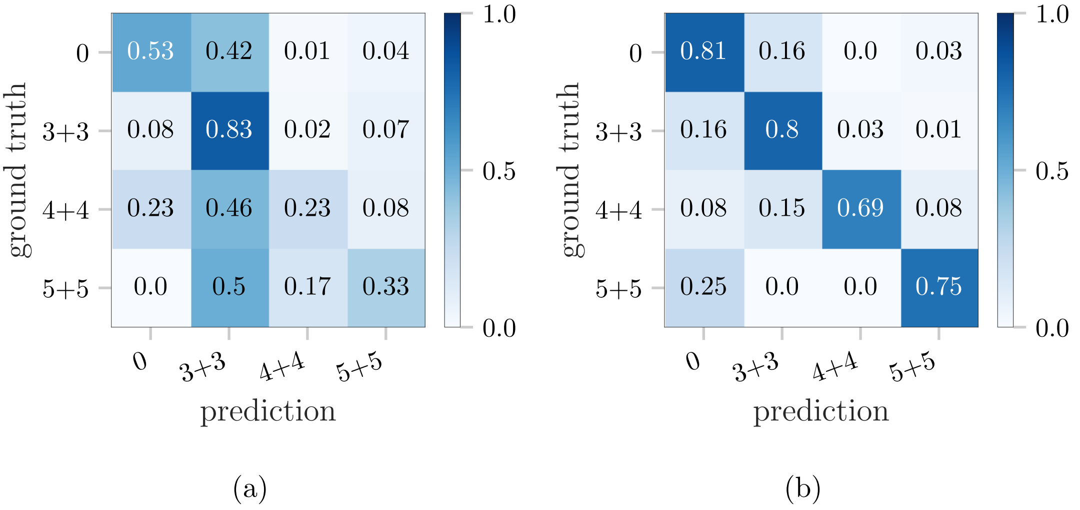


Figure 4.4.2: Comparison of the test set confusion matrices when using different patch sizes. (a) Confusion matrix using patch size 800 × 800 pixels. (b) Confusion matrix using patch size 2048 × 2048 pixels - the whole image.

image input size is evaluated. Again, an ImageNet-pretrained InceptionV3 model is trained, fine-tuning all weights. When downsizing the image to 224 × 224 pixels or 512 × 512 pixels, a batch size of 16 is used, which needs to be reduced to a batch size of 8 for 1024 × 1024 pixel inputs given computational resource constraints.

Downsizing the image to 512 × 512 pixels yields the best accuracy (0.87 on the validation set, 0.86 on the test set) and kappa (0.77 and 0.76). In contrast, the input size of 224 × 224 pixels reduces the test and validation set accuracies to 0.83 and kappa to 0.68 on the validation and 0.67 on the test set. An input size of 1024 × 1024 pixels reduces validation and test set accuracies to 0.82 (kappa validation 0.70, test 0.67). On the one hand, using images of 224 × 224 pixels might reduce the image information so much that essential details are lost. On the other hand, larger input sizes require smaller batch sizes, which might degrade performance. Furthermore, the larger the input size, the higher the number of model parameters, the computational cost, and the training time. Since the final use case does not depend on a fast, real-time analysis, the increased training and inference time with larger input images can be neglected, and the best model solely depends on the accuracy and kappa.

Final best model Concluding from the experiments above, an InceptionV3 performs best when trained on the centerpieces, using the whole available tissue, not single cutout patches. Resizing to 512 × 512 pixels yields the best results. A model trained with different seeds reached an average accuracy of 0.87 on the validation set and 0.86 on the test set. The average kappa is 0.77 on the validation and 0.76 on the test set, see also Table 4.4.2. The results on the left-out test13.1D set consisting of the single TMA drop in both accuracy and kappa to 0.73 and 0.70, respectively.

The single best-performing run achieves an accuracy of 0.88 and a kappa of 0.80 on the test set. The corresponding confusion matrix is shown in Figure 4.4.3. There is little confusion between the classes. The prediction is rarely higher than the ground truth label, however, some '5+5' Gleason tissue is mistaken as benign, and some '4+4' tissue is predicted as '3+3'.

Table 4.4.2: Best results for single Gleason prediction using an InceptionV3 and the cut centerpiece as a whole image, which is resized to 512 × 512 pixels.

	accuracy	kappa
validation set	0.87	0.77
test set	0.86	0.76
test13.1D set	0.73	0.70

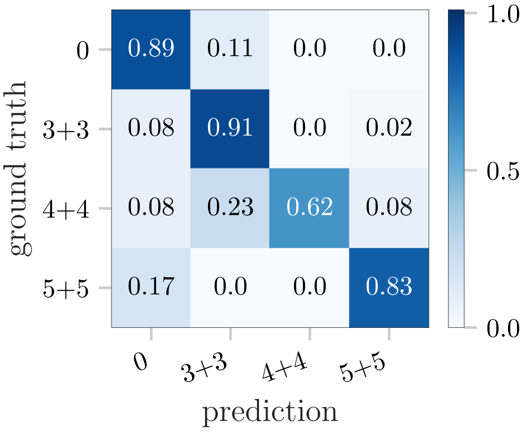


Figure 4.4.3: Confusion matrix for the best single Gleason classification model on the test set.

4.4.2 ISUP grading

Since it has been shown above that using the whole image as input to the network results in high accuracy, it is concluded that this approach should also be applicable to images with mixed Gleason grades. ISUP grade classification is chosen as an endpoint since this is an international standard. Therefore, 6 classes are distinguished (no cancer, ISUP 1-5). These ISUP classes are encoded with ordinal regression, as explained in section 4.2 *Gleason stratification.* For these experiments, all images of the Gleasonaut can be included for training, validation, and testing (see Table 4.4.1).

Again, an InceptionV3 is trained for this task. Again, the loss is a weighted cross-entropy loss (eq. (4.3.3)), the network is optimized with the Nadam optimizer (Dozat, 2016), and early stopping is applied on the validation accuracy.

Input size Since multiple Gleason patterns within a single image form the ISUP grading, the image size reduction when downsizing might have a different influence on the model performance compared to the single Gleason pattern experiments. Therefore again, the influence of the model input size on the performance is compared. Two image sizes are considered, 512 × 512 pixels and 1024 × 1024 pixels. In contrast to the above results, for validation and test data, the performance increases with the larger input size 1024 × 1024 pixels in both accuracy (0.62 versus 0.64) and kappa (0.75 versus 0.79) on the test set and the validation set (accuracy 0.63 versus 0.66, kappa 0.80 versus 0.83). This indicates that more detailed information is exploited for ISUP classification.

Best model The best model for ISUP classification is an InceptionV3 network with an input size of 1024 × 1024 pixels. It is called M_{ISUP} from now on. Training an M_{ISUP} on the Gleasonaut results in an accuracy of 0.66 on the validation set and 0.64 on the test set. The kappa is 0.83 on the validation and 0.79 on the test set, see Table 4.4.3. On the separate test13.1D set, the model's average accuracy reduces to 0.57 and the kappa to 0.76.

The best-performing single model achieves an accuracy of 0.68 with a kappa of 0.85 on the test set. That model's confusion matrix is shown in Figure 4.4.4. The high values on the diagonal indicate accurate model performance. It can be seen that the confusion between nearby classes is higher than between classes that are farther apart. That corresponds to human understanding of the ordinal ranking of the classes. Pathologists are expected to confuse close ISUP grades following

Table 4.4.3: Best results for ISUP classification with M_{ISUP} (InceptionV3, whole image, resized to 1024 × 1024 pixels).

	accuracy	kappa
validation set	0.66	0.83
test set	0.64	0.79
test13.1D set	0.57	0.76

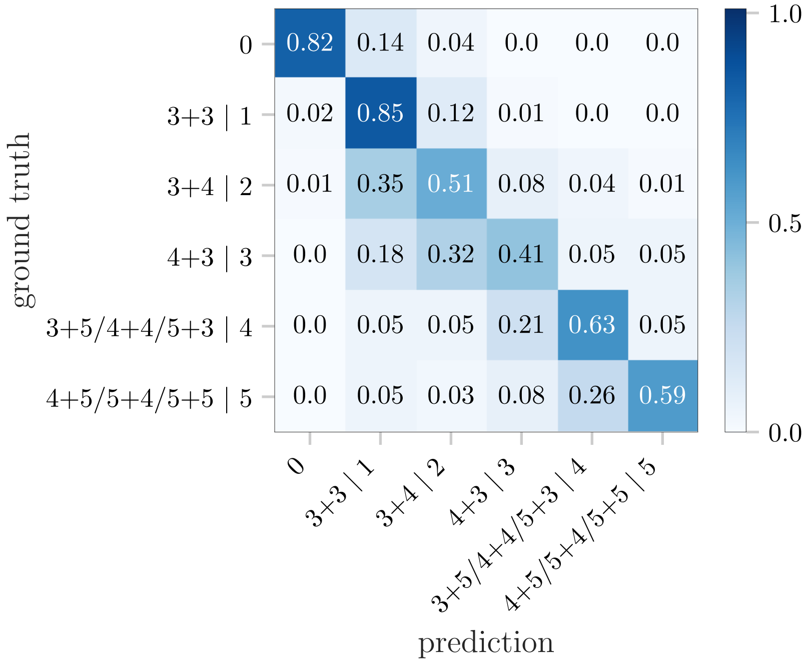


Figure 4.4.4: Confusion matrix for the best ISUP classification model, M_{ISUP}, on the test set.

a similar pattern. The model most often confuses ISUP 2 tissue as ISUP 1, and ISUP 3 tissue as ISUP 2. If the model's prediction does not match the annotation, the predicted ISUP is lower than the annotation in most cases.

4.5 Discussion

The possibility of predicting Gleason grades in the Gleasonaut was evaluated in this chapter. The research question R2 "To what degree can Gleason patterns be predicted accurately in the given dataset of digitized prostate tissue?" has been answered. When considering single Gleason patterns, an InceptionV3 reaches an accuracy of 0.87 and a kappa of 0.76 on the test set. Since these scores are only 0.01 below the validation set results, it is concluded that the model is good at generalization to unseen data. On the left-out test set 13.1D, the accuracy and kappa slightly decrease, which might result from the fact that relatively more Gleason grades 4+4 are in that test set. It could be seen in the confusion matrix in Figure 4.4.3 that ground truth class 4+4 is the greatest source of error, as it is often confused with 3+3. Also, it would be interesting to investigate whether that separate TMA has a different color stain to which the model is not robust.

When more classes are differentiated in the form of ISUP scores, the InceptionV3 reaches an accuracy of 0.64 and a kappa of 0.79. The performance drop from validation to test set is only slightly higher than in the single Gleason case. Again, the performance drops for the test set 13.1D, which is again attributed to the higher ratio of ISUP 4 and 5. It also hints at sensitivity to a dataset bias since that TMA was not included in the training set.

The best parameters for the neural network were found in a detailed analysis. Experiments revealed that the InceptionV3 has higher generalizability than an EfficientNetB0 and better overall performance than a ResNet50. The results show clear evidence that it is best to reduce the image background by cutting the centerpiece and training on that. Further, it is best to include all available centerpiece information, hence the whole image, instead of cutting a smaller patch. Since only random patches are cut in the experiments, it should be further tested whether an informative selection could improve the results. The information loss due to downsizing is lower than the information loss due to cutting out tissue parts. The results on ISUP grading indicate that less downsizing leads to better results since more details are kept in the images. In contrast to findings in the literature (Raghu et al., 2019), freezing layers pretrained on ImageNet did not improve the results. It needs to be further investigated whether this can be attributed to the large differences between images from the ImageNet database and the Gleasonaut. Further, it is worth experimenting whether freezing the first convolutional layers and training the deeper convolutional and fully-connected layers of InceptionV3 would improve results. The presented results also motivate using the ISUP prediction model as a pretrained network for the survival prediction in the next step instead of using an ImageNet-pretrained model.

It is concluded that Gleason grading is possible on the given dataset with an InceptionV3 network. The results reach comparable scores to the ones reported in the literature in Table 2.2.3. It further leads to the conclusion that the dataset is well suited for Gleason prediction as it includes high-quality images that are rich in information. Reaching a perfect classification accuracy and kappa closer to 1 is impeded by the variations in image bias as well as the label noise of Gleason scoring since this is a subjective label.

Since the results for Gleason classification are promising, moving on to survival prediction is a reasonable succeeding step to be discussed in the next chapter.

Chapter 5

Survival prediction

Survival prediction is an important research area in computational pathology and crucial for optimal treatment recommendations (Cheon et al., 2016). It is also called time-to-event prediction since the endpoint is not always patient death but could be, for example, time to relapse. The following sections motivate survival prediction and introduce the main concepts. Then, two different categories of survival prediction are introduced: Survival prediction on a population and an individual level. While population-based methods are advantageous for comparing groups that are, e.g., treated with two different medications, individualized predictions take into account multiple patient features and allow for precise treatment decisions per patient (Kumar et al., 2022). Further, the metrics used in this thesis for evaluating individual survival prediction models are presented.

A model for survival prediction of prostate cancer patients will be derived from the state of the art and introduced in section 5.6 *eCaReNet*. The following experiments section explores ablation studies, comparisons to a pathologist, explainability, and extensions to improve model performance to investigate research question R3.

5.1 Motivation

In oncology, estimating the patient's life expectancy is crucial since it can improve treatment decisions, give patients a better estimate of their current situation, and avoid over- and under-treatment (Cheon et al., 2016). Currently, for prostate cancer patients, life expectancy is not estimated directly but is captured indirectly with the Gleason score. With an image-based survival prediction model, it might be possible to extract more disease-related details from a prostate tissue image than what is captured by the Gleason score alone. In contrast to Gleason prediction, the annotations available for survival prediction are objective. In the scope of this thesis, the term survival refers to relapse-free survival after RPE.

Challenges in survival prediction arise since the probability of having a relapse depends on many factors, of which the tissue image captures only a part. Since many different factors influence the survival time, including dietary factors, genetics, and family history, there are always unknowns when predicting an individual patient's survival time (Rawla, 2019). Further, the actual probability of having a relapse over time is not measurable, but only quantifiable is whether a relapse occurs at discrete time points. In contrast to overall survival prediction, where the event is death, relapse prediction has to deal with the problem that not every patient experiences an event sooner or later since some remain relapse-free (Kleinbaum and Klein, 2012). These challenges are addressed in the following.

5.2 Censoring

In survival analysis, not every patient experiences an event since endpoints are disease-related. Furthermore, not all events are observed, for example, if patients drop out of a study and do not have further follow-ups recorded (Kleinbaum and Klein, 2012). Instead of removing those patients without (known) events from the dataset, they are marked as censored ($c = 1$) and remain in the analysis. That is important, on the one hand, because any model should also apply to healthy or cured patients that will never experience an event. On the other hand, these censored patients still have regular follow-ups, which is valuable information since it is known until when a relapse did not occur (Kleinbaum and Klein, 2012). In order to illustrate censoring, Figure 5.2.1 shows an exemplary study time with three patients. In this example, patient A has an RPE at the beginning of the study and regular follow-ups afterward. He remains relapse-free until he drops out of the study for unknown reasons. The dashed line indicates that the state of relapse is unknown afterward. Patient B has a later RPE, lives relapse-free until the end of the study, and is censored when the recording stops. He has a relapse later, which is, however, not recorded. In contrast, the relapse of patient C is observed during a follow-up. He is, therefore, uncensored. Only the relative time from RPE to BCR or censoring is considered for relapse-free survival time. Thus, the time origin is the time of RPE.

In this thesis, only right censoring (as illustrated) is considered, meaning the time of RPE is known for every patient, while the time of BCR may be unknown. Censoring can further be distinguished into independent versus non-independent, random versus non-random, and non-informative versus informative censoring. Random censoring means patients that are censored at time t should be representative of all patients surviving time t. Thus, censoring is not depending on a

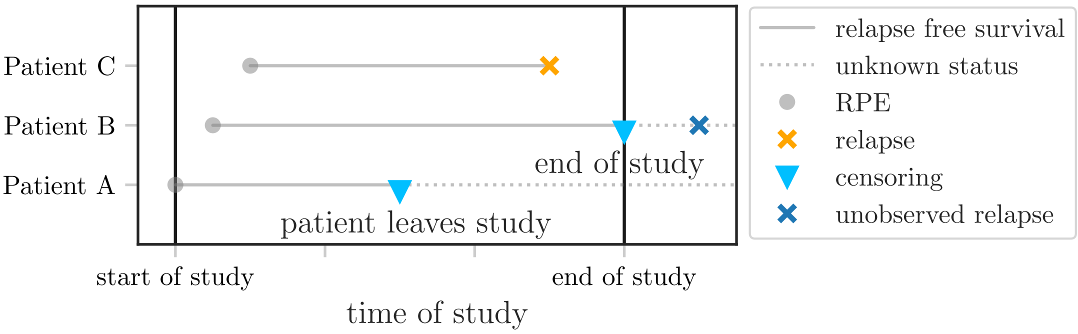


Figure 5.2.1: Example to illustrate right-censoring: A and B drop out of the study without an event and are thus censored, C experiences an event. Solid lines indicate the relapse-free survival time, dashed lines indicate an unknown status of relapse. The orange and blue “x” indicate relapses that are or are not observed during the study, respectively. A triangle indicates censoring. The gray dot is at the time of RPE, i.e., the time a patient enters the study. RPE: radical prostatectomy.

patient’s features. Independent means censoring is random within any subgroup of patients. For example, when splitting the patients according to their age, the percentage of censored patients should be equal in all groups. If the censoring is non-informative, the distribution of the event times does not give any information about the distribution of censoring times. The same holds in reverse. Often, it is assumed that independent and random censoring assure non-informativeness (Kleinbaum and Klein, 2012). For this thesis, censoring is considered random, independent, and non-informative.

5.3 Population-based survival prediction

For a population, the number of individuals that survive a time t can be estimated with the Kaplan-Meier (KM) method, which also accounts for censored patients (Kaplan and Meier, 1958). KM curves enable a graphical interpretation of the survival probabilities of a population over time and are used to compare the survival of different subpopulations. In order to calculate the KM curves, the following values need to be known for each individual: time-to-event, censoring status, and the subgroup the individual belongs to (e.g., whether the individual received a treatment). Per subgroup, one KM curve is estimated and visualized, enabling the analysis of differences in survival probability (Rich et al., 2010). For

KM curves $\hat{S}$, a population's failure times t_f are sorted. The curve is defined by

$$\begin{aligned}\hat{S}(t_f) &= \hat{S}\left(t_{f-1}\right) P(t^* > t_f | t^* \geq t_f) \\ &= \prod_{i}^{f} P(t^* > t_i | t^* \geq t_i),\end{aligned} \tag{5.3.1}$$

with the product of all conditional probabilities to survive a current failure time t_f, given the event did not occur before (Kleinbaum and Klein, 2012). Here, t^* is the individual patient's survival time, which is a random variable. Since the KM curve is evaluated at every discrete event time in the dataset, it is a step-function that drops at each measured event time. In general, the KM curves are calculated as

$$\hat{S}(t_f) = \prod_{t_i < t_f} \frac{n_i - d_i}{n_i}, \tag{5.3.2}$$

with d_i uncensored patients with an event $t^* = t_i$ and n_i patients at risk, i.e., without any event until time t_i $(t^* \geq t_i)$(May, 2017). If KM curves of two population groups stratify well, the underlying stratification correlates with the survival outcome.

Figure 5.3.1 shows an example of KM curves in which one group that received treatment (blue) is compared to a control group (gray). The survival probability curve of the control group decreases earlier, indicating that the patients in this group have, on average, an earlier event time. Therefore it is concluded that patients receiving treatment will outlive patients without treatment. Since differences between patients in other characteristics are neglected, the control and treatment groups should have equal distributions in their remaining patient features (e.g., age). That assures that no other factors influence the event time. In this figure, confidence intervals are plotted as shaded areas and calculated following the Greenwood formula for the variance (Yuan and Rai, 2011; Greenwood, 1926):

$$\mathrm{Var}[\hat{S}(t)] = \hat{S}(t)^2 \left(\prod_{t_i < t} \frac{d_i}{n_i(n_i - d_i)} \right). \tag{5.3.3}$$

Log-rank test Besides visually inspecting KM curves, a log-rank test can be used to assess whether the difference between two curves is statistically significant.

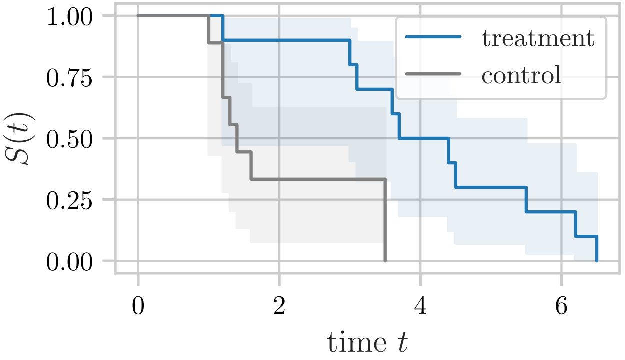


Figure 5.3.1: Example of Kaplan-Meier curves for two different subgroups. The survival probability of the control group (gray) is lower than the survival probability of patients in the treatment group (blue).

The tested hypothesis is that the probability of experiencing an event is equal for two populations at all times t:

$$H_0 : S_1(t) = S_2(t) \quad \forall t \tag{5.3.4}$$

(Li et al., 2015; Bland and Altman, 2004). For large samples, the log-rank test is a chi-square test, with

$$\text{log-rank} = \sum_i \frac{(O_i - E_i)^2}{\text{var}(O_i - E_i)} \tag{5.3.5}$$

$$\approx \sum_i \frac{(O_i - E_i)^2}{O_i} = \chi^2 \tag{5.3.6}$$

for i population groups, O_i observed and E_i expected events (Kleinbaum and Klein, 2012). A p-value indicates whether the hypothesis is rejected. If it is, the two populations do not have similar survival probabilities, their survival curves stratify well, and the feature used for splitting the population into groups is related to survival.

A drawback of the log-rank test is that it performs best under the proportional hazards assumption, which means that survival curves do not cross. Since KM survival curves may cross, the log-rank test should be modified. Li et al. (2015) compare several log-rank test adaptations under different scenarios, like varying amounts of censoring, number of samples per subgroup, and the time point of crossing for the survival curves (early, middle, late). They show that Fleming-Harrington weights (Fleming and Harrington, 1991) are appropriate to adapt the log-rank test in situations with more censored than uncensored patients, as is the case in the given Survival dataset in this thesis. In that case, the survival times are not all weighted equally but according to the failure time. The test statistic changes to

$$\text{log-rank}_{FH} = \frac{(\sum w(O_i - E_i))^2}{\text{var}(\sum w(O_i - E_i))}, \tag{5.3.7}$$

$$\text{with} \quad w = \hat{S}(t)^p \times [1 - \hat{S}(t)]^q, \tag{5.3.8}$$

(Kleinbaum and Klein, 2012). Here, $p = 1$ and $q = 0$ holds (Li et al., 2015).

5.4 Individual survival prediction

For personalized healthcare, it is beneficial to predict the survival outcome for individuals instead of comparing groups of patients. As described in section 2.2.4 *Deep learning for survival prediction*, survival prediction can be interpreted in different ways. Some research is concerned with the binary prediction of whether a patient survives a certain time, while other aims at stratifying patients into different risk groups. In the context of this thesis, however, the aim is to develop a model that predicts individual survival curves for patients based on their input features, i.e., a diagnostic image.

If a patient's survival probability is predicted over time, this is a continuous-time problem. The Cox model is a typically-used continuous-time survival prediction model and is thus also introduced here as a baseline model for comparison. Extending the Cox model with neural networks is also possible. An alternative is treating survival prediction as a discrete-time problem, for which neural networks can also be trained end-to-end. The derivations of the most important formulas and concepts, starting from continuous and then moving on to discrete-time survival prediction are given in the following.

5.4.1 Continuous-time survival prediction

Let t^* be the event time of a patient. That event time is defined as either the time of censoring t^c or the time of relapse t^r as $t^* = \min(t^c, t^r)$. Therefore,

$$t^* = \begin{cases} t^c & \text{if } c = 1, \\ t^r & \text{if } c = 0, \end{cases}$$

with the censoring indicator c. The patient's survival function

$$S(t) = P(t^* > t). \tag{5.4.1}$$

defines the probability to survive event-free up to time t. It is a monotonically decreasing function, defined for $t \geq 0$ (Emmert-Streib and Dehmer, 2019). Since the event time t^* is a random variable, it has a probability density function $f(t)$ and the cumulative density function $F(t)$, which describes the probability that the event occurs before time t. Thus, in reverse, the survival probability can be written as

$$S(t) = 1 - F(t) = \int_t^\infty f(s)\mathrm{ds} \tag{5.4.2}$$

(Rodriguez, 2007). The survival time of a patient with features $\mathbf{x}$ can be calculated as the expected value of the survival function:

$$\mathbf{E}[\mathbf{x}] = \int_0^\infty t f(t)\mathrm{dt} = \int_0^\infty S(t)\mathrm{dt} \tag{5.4.3}$$

(Emmert-Streib and Dehmer, 2019). Instead of modeling the survival probability directly, an option is to describe the risk of the event occurring at time t, given that it did not occur until time t. This is described by the conditional hazard rate

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq t^* < t + \Delta t | t^* \geq t)}{\Delta t} = \frac{f(t)}{S(t)}, \tag{5.4.4}$$

which is a non-negative function (Emmert-Streib and Dehmer, 2019). The survival function can be expressed in terms of the hazard function as

$$S(t) = \exp\left(- \int_0^t h(t) \right). \tag{5.4.5}$$

Many survival models estimate $h(t)$ first and the above relation eq. (5.4.5) is applied afterward to obtain a survival curve. Note that the expectation value cannot be calculated for censored patients since t is undefined. The survival and hazard functions are still defined though (Rodriguez, 2007).

Cox model A popular approach for continuous-time survival prediction is the Cox proportional hazards (CoxPH) model, in short, Cox model (Cox, 1972). It separates the time dependency of the survival curves from the influence of patient features. Let $\mathbf{x}$ be the vector of p input features for a patient. In the Cox model, the hazard function for this patient is calculated as

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t)\exp\left(g(\mathbf{x})\right) \\ &= h_0(t)\exp\left(\sum_{i}^{p}\beta_p x_p\right) \end{aligned} \tag{5.4.6}$$

with $h(t) \geq 0\,\forall$ t (Kleinbaum and Klein, 2012; Emmert-Streib and Dehmer, 2019). Here, $h_0(t)$ is called the baseline hazard since h reduces to $h_0(t)$ if all input features $\mathbf{x} = 0$ (Kleinbaum and Klein, 2012). The baseline hazard is equal for all patients and only dependent on t, not on patient features $\mathbf{x}$. In contrast, the exponential part $g(\mathbf{x})$ is independent of time t. It can be modeled with a logistic regression, fitting the parameters β_p with a maximum likelihood estimator. It limits the model to linear dependencies between the patient features and survival time. Since the baseline hazard $h_0(t)$ is not defined in particular, the Cox model is a semiparametric model. In practice, the baseline hazard is often neglected since it is unnecessary for risk stratification. Several assumptions are made when using a Cox model. The proportional hazards assumption states that the hazard rate of two patients remains constant over time (Kleinbaum and Klein, 2012). As stated above, it is assumed that the baseline hazard is equal for all patients and only scaled individually through $g(\mathbf{x})$. Therefore the survival curves of two patients cannot cross. During the optimization of the model, only the exponential part is used to compare the ranking between patients.

The Cox model can be extended to nonlinear functions when modeling $g(\mathbf{x})$ with a more complex function. In the DeepSurv model, Katzman et al. (2018) propose to train a neural network with a single output neuron to model $g(\mathbf{x})$. As an extension to processing images instead of tabular data, Zhu et al. (2016) replace that neural network with a CNN, calling their model DeepConvSurv accordingly.

When training such a Cox (-based) model, the ranking of the patients is compared since the progress of all patients' curves over time is equal and therefore negligible. As an objective function, the Cox partial likelihood

$$L = \prod_{i:c_i=0} \frac{\exp\left(g(x_i)\right)}{\sum_{j:t_j^* \geq t_i^*} \exp\left(g(x_j)\right)} \tag{5.4.7}$$

is applied. For each uncensored patient ($c_i = 0$) with features x_i and event time t_i^*, his event probability given the patients that are still in the risk set at time t_i^* is considered (Kvamme et al., 2019). Using the negative logarithm of the loss simplifies the product to a sum in the negative log partial likelihood (nlpl)

$$L = -\frac{1}{n} \sum_{i:c_i=0} \left(g(x_i) - log \sum_{j:t_j^* \geq t_i^*} \exp\left(g(x_j)\right) \right) \tag{5.4.8}$$

(Katzman et al., 2018). Note that when training a neural network, only patients within the same batch are compared in the loss, setting n to the patients with events within a batch, not in the population. These adaptations of the Cox model using neural networks still do not allow crossing survival curves.

5.4.2 Discrete-time survival prediction

Instead of adapting the Cox model with a neural network, in this thesis survival curves are modeled end-to-end. That means, not only a risk factor is predicted with a neural network, but the survival curve per discrete time step. Sloma et al. (2021) showed that a discrete model does not result in a loss in performance compared to a continuous model. An advantage is that the proportional hazards assumption can be ignored and survival curves are allowed to cross.

In contrast to continuous-time survival prediction, in discrete-time survival prediction the time of interest t has discrete values $0 = t_0 < t_1 < \ldots < t_k$ (Kvamme and Borgan, 2021). These times define the boundaries of k intervals $I_j = (t_{j-1}, t_j]$, $j = 1, \ldots,\ k$ (Suresh et al., 2022). The survival function over time for a patient with features $\mathbf{x}$ and a true survival time t^*,

$$S(t_j|\mathbf{x}) = P(t^* > t_j|\mathbf{x}), \tag{5.4.9}$$

now describes the probability of surviving the time step t_j and thus the interval I_j. Equivalent to the continuous-time case, the probability mass function

$$f(t_j|\mathbf{x}) = P(t^* \in I_j|\mathbf{x}) \tag{5.4.10}$$

can be used to define the survival function

$$S(t_j|\mathbf{x}) = \sum_{k>j} f(t_k|\mathbf{x}) \tag{5.4.11}$$

(Kvamme and Borgan, 2021; Suresh et al., 2022). Now t^* corresponds to the interval in which the event occurs. The hazard function defines the probability of an event to occur in the current interval, given it has not yet occurred:

$$h(t_j|\mathbf{x}) = P(t^* \in I_j | t^* > t_{j-1}) \tag{5.4.12}$$

(Kvamme and Borgan, 2021; Suresh et al., 2022). The survival probability can be written in terms of hazard as

$$S(t_j) = \prod_{k=1}^{j} (1 - h(t_k)). \tag{5.4.13}$$

(Kvamme and Borgan, 2021). When training a neural network on survival prediction in an end-to-end approach, either $S(t_j)$ or $h(t_j)$ can be the prediction endpoint. The advantage of modeling the hazard rate over time is that the constraint of having a monotonically decreasing survival function does not need to be accounted for, as this results from eq. (5.4.13) directly.

As an output for a neural network, one neuron per time interval is chosen. In the ground truth, each interval that a patient survives is annotated as $S = 1$, and each interval that he does not survive is annotated as $S = 0$. After censoring, S is not defined. The hazard is annotated with $h = 1$ in the interval with the event and $h = 0$ before. In the intervals after the event, the hazard rate is not defined (n.d.) anymore. This can be written as

$$S_{annotation}(t_j) = \begin{cases} 1 & \text{if } t^* > t_j, \\ 0 & \text{if } t^* \leq t_j \text{ and } c_i = 0, \\ \text{n.d.} & \text{if } t^* \leq t_j \text{ and } c_i = 1, \end{cases} \tag{5.4.14}$$

$$h_{annotation}(t_j) = \begin{cases} 1 & \text{if } t^* \in I_j \text{ and } c_i = 0, \\ 0 & \text{if } t^* > t_j, \\ \text{n.d.} & \text{if } t^* \in I_j \text{ and } c_i = 1, \\ \text{n.d.} & \text{if } t^* < t_{j-1}. \end{cases} \tag{5.4.15}$$

Risk estimation Estimating the life expectancy with S can be a problem for censored patients, as it would grow to infinity. In the literature, that measure is nevertheless sometimes applied. Xiao et al. (2020), e.g., calculate the overall survival time as the sum of the predicted survival probabilities per time interval. By limiting the number of intervals, an infinite survival probability is prevented.

Instead of using the area under the survival curve as a measure of life expectancy, in this thesis, it is proposed to use it as a measure of patients' risk.

The higher the area under the curve, the lower the risk. Therefore, a risk score between 0 and 1

$$r = 1 - \frac{1}{t_k} \sum_{i=1}^{k} S(t_i) \cdot |t_i - t_{i-1}| \tag{5.4.16}$$

can be estimated when summing over all k intervals. For a censored patient, whose survival probability ground truth is $S(t) = 1 \ \forall t$, the risk is 0. A patient with an event in the first interval, therefore $S(t) = 0 \ \forall t$, has risk 1.

Loss function In this setting of discrete-time survival prediction, the aim is not only to correctly order the patients but also to correctly estimate the survival probability over time. Thus, a loss depending only on the order of patients, like the nlpl, is not sufficient. Instead, the predicted hazard or survival probability per interval needs to be included in the loss function. Here, a likelihood function is formulated per patient as

$$L = f(t_i^*)^{(1-c_i)} S(t_i^*)^{c_i} \tag{5.4.17}$$

(Kvamme and Borgan, 2021). The mean negative log-likelihood

$$L = -\frac{1}{n} \sum_{i=1}^{n} ((1 - c_i) \log[f(t_i^* | \mathbf{x}_i)] + c_i \log[S(t_i^* | \mathbf{x}_i)]), \tag{5.4.18}$$

over n patients, each with features $\mathbf{x}$, is minimized by a neural network (Kvamme and Borgan, 2021). This loss can further be expressed in terms of the hazard function and split into a censored and an uncensored part:

$$L_{c=0} = \sum_{c=0} [log(h(t^*)) + \sum_{t_i : t_i < t^*} log(1 - h(t_i))], \tag{5.4.19}$$

$$L_{c=1} = \sum_{c=1} [log(S(t^*))] \tag{5.4.20}$$

$$= \sum_{c=1} [\sum_{t_i : t_i \leq t^*} log(1 - h(t_i))]. \tag{5.4.21}$$

Thus, the predicted survival probabilities and hazard rates until the event interval are considered. Ren et al. (2019b) propose to weight both losses separately as

$$L = \alpha L_{c=0} + (1 - \alpha) L_{c=1}. \tag{5.4.22}$$

This loss will be used in section 5.7 *Experiments* with α=0.5, which empirically led to the best results.

Definition of time intervals There are several options to define the discrete intervals I_j used for survival analysis. The most common option is using equally spaced intervals, each of the same length. Another option is using intervals divided into quantiles, with an equal number of patients' event and censoring times per interval (Kvamme and Borgan, 2021).

Since the division into quantiles is data-driven, this may lead to problems when applying the model to new datasets. Further, in a dataset like the one illustrated in Figure 3.1.2, this spacing would lead to very small intervals at the beginning and large intervals at the end. Distinguishing between very small differences in survival times is not reasonable since these might result from the time of follow-up rather than the disease. Large intervals in the end, reflect an uncertainty for far-away predictions but are also less meaningful. In this thesis, equal spacing of intervals is considered.

5.5 Metrics

For the evaluation of survival prediction models, multiple metrics are considered in this thesis. A survival model should be able to discriminate between patients while being well-calibrated at the same time. While discrimination measures how well the order of the predicted survival curves (or risk scores) matches with the actual order of patients' event times, calibration measures how well the prediction matches with the ground truth (Gerds and Kattan, 2021). Metrics for both cases are presented in the following.

C-index As a discriminating metric for survival analysis, the concordance index, short C-index, is commonly applied (Harrell et al., 1982). When randomly drawing two patients A and B, with survival times $t^*_A < t^*_B$, the C-index estimates the probability that the risk for patient A is higher than for patient B (Gerds and Kattan, 2021). If the predicted order of the two patients matches the true survival times' order, the pair is concordant. Only comparable pairs of data points are used, which means that in this example, patient A may not be censored. The C-index is applied to measure how many patient pairs are concordant, relative to all comparable pairs. In case the prediction is not a single risk score but a measure that changes over time, like a survival curve, the evaluation is performed at the event times of the patients. This means that at the event time t_A, it is expected that the survival probability of patient B is higher than that of patient A ($S_B > S_A$ at time t^*_A), see also Figure 5.5.1. That is especially necessary in the case of non-proportional hazards, with the possibility of crossing survival curves. The time-dependent (td) C-index follows as

$$\text{C-index}^{td} = P\left(S_i(t^*_i) < S_j(t^*_i)|t^*_i < t^*_j, c_i = 0\right) = \frac{\#\text{ concordant pairs}}{\#\text{ comparable pairs}} \tag{5.5.1}$$

for patients i and j, where $S_i(t) = S(t|\mathbf{x}_i)$ (Antolini et al., 2005). Counting tied predictions only half is proposed in the literature:

$$\begin{aligned}\text{C-index}^{td} = {} & P(S_i(t^*_i) < S_j(t^*_i)|t^*_i < t^*_j, c_i = 0) \\ & + 0.5P(S_i(t^*_i) = S_j(t^*_i)|t^*_i < t^*_j, c_i = 0)\end{aligned} \tag{5.5.2}$$

(Yan and Greene, 2008; Longato et al., 2020). A C-index of 0.5 is equivalent to a random guess, and a measure of 1 is the highest possible concordance (Harrell et al., 1982). A problem with the C-index occurs if the risk of experiencing an event within a time horizon t_h is of interest. The event horizon does not affect the calculation of the C-index and therefore can result in misleading conclusions (Gerds and Kattan, 2021). It has been shown by Blanche et al. (2019) that the C-index is not a proper score. That means the true underlying data distribution does not necessarily lead to the best C-index. Therefore, this measure needs to be considered with care. Furthermore, as can be seen in Figure 5.5.1, the C-index does not account for crossing survival curves, since two patients are only compared at a single time. The predictions are considered concordant, even though after a time, the survival curve of patient B drops below the curve of patient A.

AUC^{cd} Another measure of discrimination is the cumulative-dynamic time-dependent area under the receiver operator curve (AUC^{cd}), see Figure 5.5.2. In contrast to the C-index, the AUC^{cd} takes into account a time horizon t_h for the evaluation

$$\begin{aligned}AUC^{cd}(t) = {} & P\left(S_i(t) < S_j(t) \mid t^*_i \leq t, t^*_j > t\right) \\ & + 0.5P\left(S_i(t) = S_j(t) \mid t^*_i \leq t, t^*_j > t\right),\end{aligned} \tag{5.5.3}$$

when $t < t_h$ (Blanche et al., 2019; Kamarudin et al., 2017). For a certain time t, the AUC^{cd} discriminates between cases (patients with an event before t) and

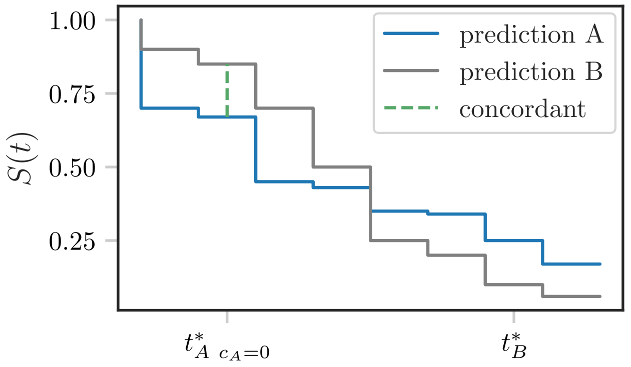


Figure 5.5.1: The two survival curves of patients A and B are compared at t^*_A for the C-index since $t^*_A < t^*_B$.

controls (patients with an event later than t). Equivalent to the C-index, tied predictions are only counted as half.

As the name suggests, cumulative sensitivity and discriminative specificity are included in the definition. The cumulative sensitivity describes the probability that a patient's survival prediction is lower than a threshold z, given an event $t^*_i < t$. The discriminative specificity is defined as the probability of a patient having a survival probability greater than z, among all patients who are event-free at time t. Thus,

$$Se(z,t) = P(S_i(t) \leq z | t^*_i \leq t), \tag{5.5.4}$$

$$Sp(z,t) = P(S_i(t) > z | t^*_i > t), \tag{5.5.5}$$

(Kamarudin et al., 2017). Using these definitions, the AUC^{cd} can be formulated as

$$\begin{aligned} AUC^{cd}(t) &= P(S_i(t) < S_j(t) | t^*_i = t, t^*_j > t), \quad i \neq j \\ &= \int_{-\infty}^{\infty} Se(z,t)\; d[1 - Sp(z,t)]. \end{aligned} \tag{5.5.6}$$

An adaptation is needed for the sensitivity in the case of censoring, which is why an inverse probability of censoring weighting (ipcw) can be performed (Vock et al., 2016). A KM estimate of the censoring distribution $\hat{S}_C$ is calculated with eq. (5.3.2), giving the probability that a patient is censored instead of the probability that a patient has an event. This is also applied when using Uno's estimator of sensitivity and specificity (Uno et al., 2007). According to Blanche et al. (2013)

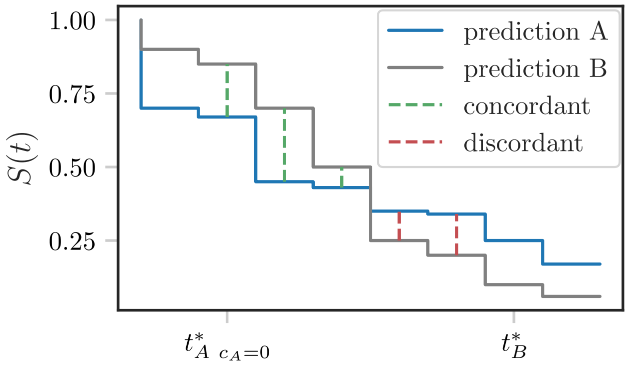


Figure 5.5.2: The two survival curves of patients A and B are compared at multiple times $t^*_A \leq t < t^*_B$ for the AUC^{cd}. Concordant and discordant pairs are indicated in green and red, respectively.

and Kamarudin et al. (2017), these are defined as

$$Se(c,t) = \frac{\sum_{i=1}^{n} \mathbf{I}(S_i(t) \leq z, t^*_i \leq t)\frac{(1-c_i)}{n\hat{S}_C(t^*_i)}}{\sum_{i=1}^{n} \mathbf{I}(t^*_i \leq t)\frac{(1-c_i)}{n\hat{S}_C(t^*_i)}}, \tag{5.5.7}$$

$$Sp(c,t) = \frac{\sum_{i=1}^{n} \mathbf{I}(S_i(t) > z, t^*_i > t)}{\sum_{i=1}^{n} \mathbf{I}(t^*_i > t)}, \tag{5.5.8}$$

for n samples and with the indicator function $\mathbf{I}$ taking either 0 or 1 as value.

To obtain a single score over time span (t_1, t_2), the AUC^{cd}(t) is further integrated

$$\overline{AUC^{cd}}(t_1, t_2) = \frac{1}{S(t_1) - S(t_2)} \int_{t_1}^{t_2} AUC^{cd}(t)\, dS(t) \tag{5.5.9}$$

(Lambert and Chevret, 2016). Throughout the thesis, the $\overline{AUC^{cd}}$ will be referred to as AUC to ease readability.

Brier score A metric that combines discrimination and calibration is the Brier score (Brier, 1950). For uncensored cases, the Brier score is equivalent to the mean squared error and calculates the distance between the true survival curve and the prediction. Since the true underlying survival probability cannot be observed, the difference between the annotation (a survival curve that drops from 1 to 0 in the event interval) and the predicted survival probability per time step is calculated, as illustrated in Figure 5.5.3. A Brier score below 0.25 is considered meaningful (Gerds and Kattan, 2021). Following Kvamme et al. (2019), the Brier score can

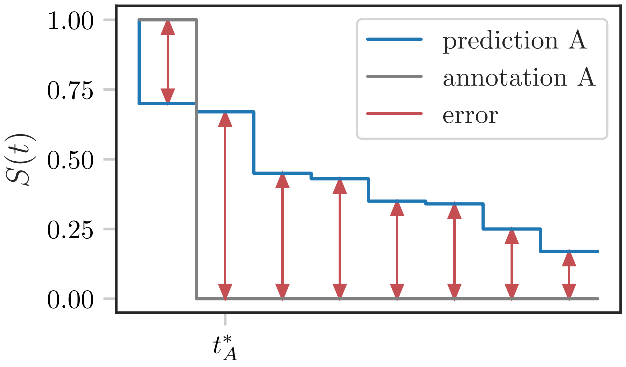


Figure 5.5.3: The survival curve of patient A is compared to the ground truth curve, which drops from one to zero at t_A^* for the Brier score. The absolute error of the prediction is indicated with red arrows.

also be weighted with ipcw to account for censored cases

$$\mathrm{Brier}(t) = \frac{1}{n}\sum_{i=1}^{n}[\frac{(1 - S_i(t))^2 \mathbf{I}(t_i^* > t)}{\hat{S}_C(t)} + \frac{S_i(t)^2 \mathbf{I}(t_i^* \leq t, c_i = 0)}{\hat{S}_C(t^*)}]. \tag{5.5.10}$$

Also in this case, a single Brier score over time is obtained through integration,

$$\overline{\mathrm{Brier}} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \mathrm{Brier}(s)ds, \tag{5.5.11}$$

for a time span from t_1 to t_2. Throughout this thesis, the Brier score refers to this integrated Brier score. The perfect survival curve with a Brier score of 0 would match the ground truth annotation. Haider et al. (2020) state that the Brier score is not sufficient to estimate the goodness-of-fit of predicted survival curves. They argue that the perfect survival curve would model an unrealistic probability for the stochastic event time t. Further, it cannot provide an estimate of the probability of surviving past a specific time. Therefore, they introduce the d-calibration.

D-calibration For an individual patient, it is not only relevant if a model predicts the order of patients' events correctly, but it is of interest if the predicted survival curve matches the underlying ground truth. That can be assessed using a calibration measure. A model is d-calibrated if the survival functions per patient reflect the probability of relapse over time (Haider et al., 2020). If according to the prediction, a patient's survival probability after two years is 90 %, that should reflect the true probability. That is valid if 90 % of patients survive past the time at which their survival probability is 90 %. It is expected that 10 % of patients relapse while their survival probability is predicted to be 90-100 %. The same holds true for each probability interval: The survival probability is divided into 10 bins $[p_{k-1}, p_k), k = 1, \ldots, 10$ each covering 10 %, i.e., $p_0 = 0, p_1 = 0.1$, etc. Since the d-calibration is easiest understood graphically, Figure 5.5.4 shows an exemplary plot. In order to calculate calibration, the predicted survival probability at each patient's event time, $S_i(t^*)$, is recorded. Each patient is assigned to one of the bins based on his survival probability. It is expected that 10 % of the patients are assigned per bin. For the calculation of d-calibration, Haider et al. (2020) treat censored patients differently from uncensored ones since their t^* is not equal to the time of the event.

They calculate the number of patients per 10 %-bin, b_k, for a dataset with N patients as

$$\begin{aligned} b_k = \frac{1}{|N|} \sum_i^N \Bigg[& \qquad\qquad\quad \mathbf{I}(S_i(t_i^*) \in [p_k, p_{k+1}) \; c_i = 0) \\ & + \frac{S_i(t_i^*) - p_k}{S_i(t_i^*)} \cdot \mathbf{I}(S_i(t_i^*) \in [p_k, p_{k+1}) \; c_i = 1) \\ & + \frac{p_{k+1} - p_k}{S_i(t_i^*)} \cdot \; \mathbf{I}(S_i(t_i^*) \geq p_{k+1} \qquad c_i = 1) \Bigg]. \end{aligned} \tag{5.5.12}$$

Whether the resulting distribution over the bins is uniform is tested with a chi-square test. If the test passes ($p > 0.05$), the model is well calibrated (Haider et al., 2020).

Remarks Besides the Brier score, all metrics consider the whole population and cannot evaluate the model's performance for single patients. The scores are best used to compare models trained on the same dataset (Gerds and Kattan, 2021). Since calibration and discrimination are two contradictory scores, usually a trade-off between both is needed. To evaluate and compare survival models, all the above metrics, AUC, C-index, Brier score, and d-calibration, need to be considered.

5.6 eCaReNet

For survival prediction on prostate cancer patients, eCaReNet (explainable Cancer Relapse prediction Network) was developed for this thesis and published (Dietrich et al., 2021). It is a neural network that takes as input images of prostate tissue and predicts individual hazard rates. The hazard rates are used to calculate

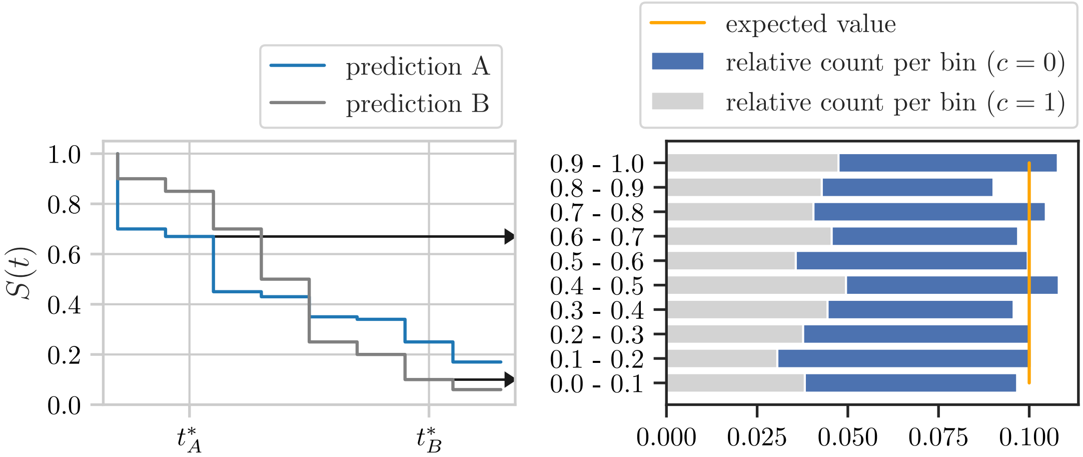


Figure 5.5.4: The survival prediction of patient A is 0.67 at t_A^* and therefore contributes to the 0.6-0.7 bucket. $S_B(t_B^*) = 0.09$ thus contributes to the 0-0.1 bucket in the d-calibration plot on the right. A model is calibrated if the distribution on the right is uniform. Inspired by Haider et al. (2020).

both a survival curve and a risk score per patient, using eq. (5.4.13) and eq. (5.4.16). It builds upon the pretrained model M_{ISUP} from section 4.4.2 *ISUP grading*. The version predicting ISUP scores from the whole TMA spot is used since it is supposed to capture more detail than the single-Gleason-pattern model.

5.6.1 Overview

Figure 5.6.1 shows a complete overview of eCaReNet. In part A, M_{ISUP} (from section 4.4.2 *ISUP grading*) is used for supervised pretraining on ISUP grades. Part B shows a binary relapse prediction network, M_{Bin}, which approximates the probability of a patient having a relapse within two years. Its output can be used as an additional input for eCaReNet in part C. Part C is the main model, eCaReNet, which outputs individual survival curve predictions. The overview figure also shows which datasets and annotations are used for each model part. Details for parts B and C are provided in the following since part A was already detailed in section 4.4.2 *ISUP grading*.

5.6.2 Binary survival prediction

As a preparatory task, a binary survival prediction model M_{Bin} is trained. It predicts whether or not a patient will have a relapse before time t_x. The output of this network will be used as additional input to eCaReNet. As a model, M_{ISUP} from

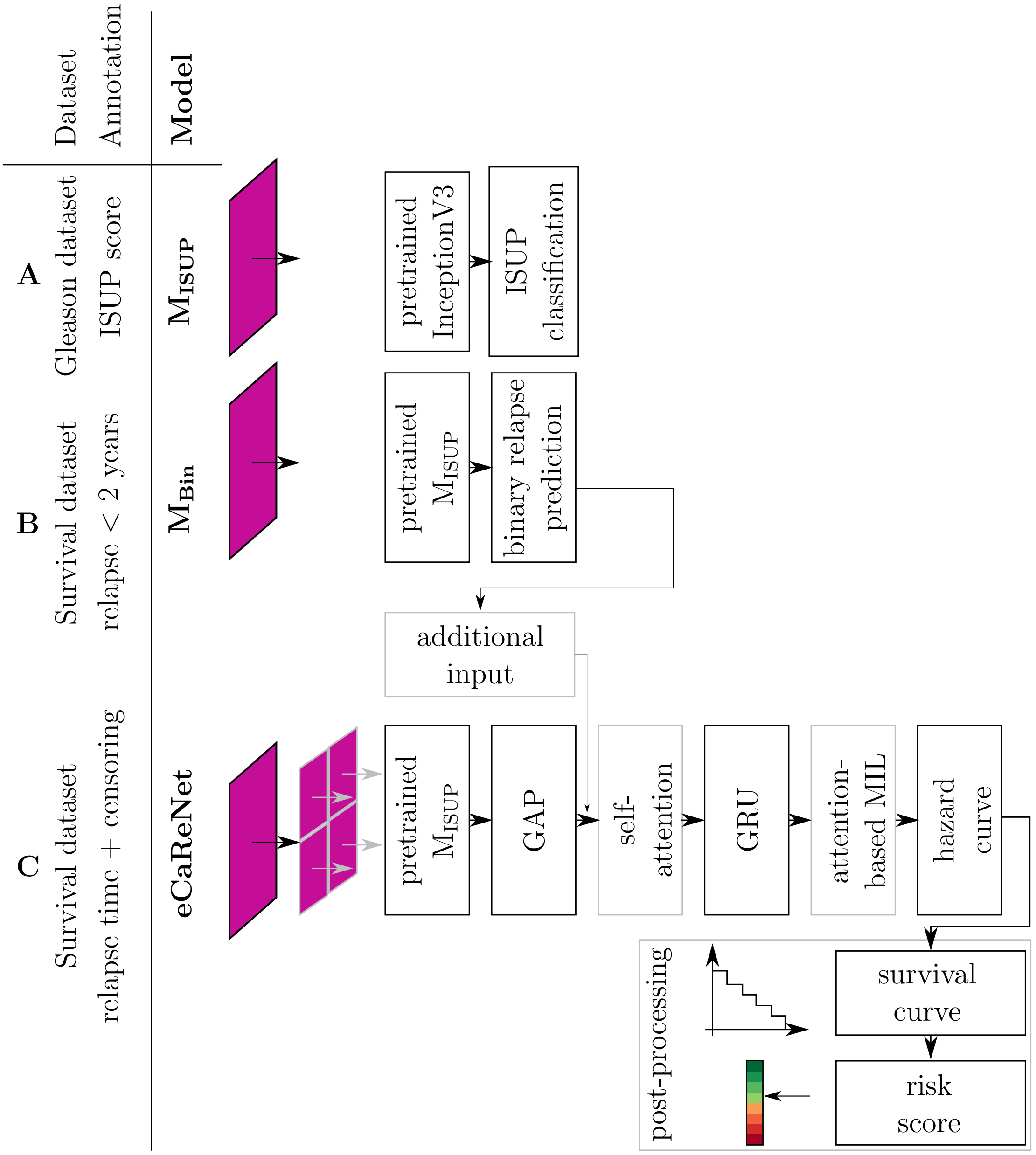


Figure 5.6.1: Overview of the complete model, including the three steps M_{ISUP}, M_{Bin} and eCaReNet. On the left, the dataset and the annotation for training are indicated. The input image is shown in pink, optional model parts are drawn in gray, and necessary parts are in black. For clarity, only four image patches are drawn, while the final eCaReNet uses 64 patches. ISUP: International Society of Urological Pathology, Bin: binary, GAP: global average pooling, GRU: gated recurrent unit, MIL: multiple instance learning.

section 4.4.2 *ISUP grading* is used but the last layer is reduced to 2 output neurons and it is retrained. The input size is 1024 × 1024 pixels. During model training with cross-entropy loss, all network weights are fine-tuned using the complete Surv1 training data for which the survival status at t_x is known. That means that patients with a censoring time before t_x are excluded. The binary survival prediction model is shown in part B in Figure 5.6.1.

5.6.3 Survival curve prediction

The same pretrained $\mathrm{M}_{\mathrm{ISUP}}$ is used again as a base model for the survival curve prediction in eCaReNet. Images with three color channels are used as input, and hazard rates per time interval are returned. However, some more adaptations are made to transform the classification network into a survival curve prediction network.

$\mathrm{M}_{\mathrm{ISUP}}$ is cut after four inception blocks to decrease the number of parameters, required computational resources, and training time. In experiments, this came with no loss in prediction performance compared to using the complete architecture.

After the last convolutional layer, a global average pooling layer is attached. That reduces the 3-dimensional output to a vector. To this vector, the output of the binary prediction model $\mathrm{M}_{\mathrm{Bin}}$ is concatenated.

Time is discretized into intervals so the model can output hazard rates per interval. Following the approach of Ren et al. (2019b), a recurrent layer is used to account for the time dependencies between time interval hazard rates. As input to the recurrent layer, the current vector needs to be repeated so it matches the number of time intervals and a time step is concatenated to each of these repetitions, representing the interval limits. The recurrent layer then outputs one hazard prediction per time interval. In this thesis, GRUs (Cho et al., 2014) are chosen for the recurrent layer since these yielded the best results compared to LSTMs in experiments. GRUs comprise an update gate and a reset gate. During a forward pass, the reset gate is updated as

$$r_j = \sigma\left([\mathbf{W}_r\mathbf{x}]_j + [\mathbf{U}_r\mathbf{h}^{t-1}]_j\right), \tag{5.6.1}$$

with the logistic sigmoid function σ, the weight matrices $\mathbf{W}$ and $\mathbf{U}$, the input $\mathbf{x}$ and the previous hidden state $\mathbf{h}^{t-1}$. Thus, superscript t denotes the current time and each $[.]_j$ indicates a vector's j-th element (Cho et al., 2014). The elements of the update gate $\mathbf{z}$ are computed similarly as

$$z_j = \sigma\left([\mathbf{W}_z\mathbf{x}]_j + [\mathbf{U}_z\mathbf{h}^{t-1}]_j\right). \tag{5.6.2}$$

The final activation at time t, $\mathbf{h}^t$, is then calculated element-wise as

$$\begin{aligned} h_j^t &= z_j h_j^{t-1} + (1 - z_j)\tilde{h}_j^t, & (5.6.3)\\ \text{with} \quad \tilde{h}_j^t &= \tanh\left([\mathbf{W}_h\mathbf{x}]_j + [\mathbf{U}_h(\mathbf{r} \odot \mathbf{h}_{t-1})]_j\right), & (5.6.4) \end{aligned}$$

with the element-wise product $\odot$. The reset gate allows the GRU to ignore previously seen information that is stored in the hidden state.

In order to include ad-hoc explainability, an attention-based multiple instance learning (MIL; Ilse et al., 2018) approach is included in the survival model as follows. Instead of inputting the whole image at once into the network, each image is cut into non-overlapping square tiles, which form a bag of instances. These tiles are processed individually until an MIL layer after the GRU weights the hazard rates per image patch with attention a_k and adds those together with

$$\begin{aligned} a_k &= \frac{\exp(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_k^T))}{\sum_{j=1}^{K} \exp(\mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_j^T))}, & (5.6.5)\\ \mathbf{o} &= \sum_{k=1}^{K} a_k \mathbf{h}_k. & (5.6.6) \end{aligned}$$

Here, $\mathbf{V}$ is a matrix of network weights, $\mathbf{w}$ contains the learned attention weights as a column vector, $\mathbf{o}$ is the layer output, and $\mathbf{h}$ is the output vector per image patch out of K patches (row vectors forming a bag $\mathcal{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_K\}$). The hyperbolic tangent tanh is applied element-wise (Ilse et al., 2018). The weights per image patch reveal how much each image region influences the final prediction. The network outputs one prediction per patient.

For eCaReNet, a further adaption with self-attention is included (Rymarczyk et al., 2021). The self-attention layer accounts for dependencies in-between patches and is included right before the GRU. That way, the information about other patches of the same image can be considered in the hazard prediction. For self-attention, per patch a key vector $\mathbf{k}_i = \mathbf{W}_k\mathbf{h}_i$, a query vector $\mathbf{q}_j = \mathbf{W}_q\mathbf{h}_j$, and a value vector $\mathbf{v}_i = \mathbf{W}_v\mathbf{h}_i$ are defined. The dot product of the key and query is

$$s_{ij} = \langle \mathbf{k}(\mathbf{h}_i), \mathbf{q}(\mathbf{h}_j) \rangle. \tag{5.6.7}$$

Scores β per patch are calculated with

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N} \exp(s_{ij})} \tag{5.6.8}$$

for each j, i combination of N patches. Finally, the output of the self-attention layer is calculated patch-wise as

$$\hat{\mathbf{h}}_j = \gamma \sum_{i=1}^{N} \beta_{j,i} \mathbf{v}_i + \mathbf{h}_j, \tag{5.6.9}$$

where γ is trainable (Rymarczyk et al., 2021; Ramachandran et al., 2019).

The whole architecture of eCaReNet is depicted in Figure 5.6.1, part C. Some parts of eCaReNet are necessary for survival curve prediction, while others are not essential but are expected to increase the performance (shown in gray). The influence of the optional parts, e.g., MIL, is evaluated in an ablation study in section 5.7 *Experiments*.

The model predicts a hazard score per time interval to avoid modeling the monotonicity of survival curves. These hazard scores are converted to survival probabilities per time interval with eq. (5.4.13). A risk score per patient is obtained with eq. (5.4.16). The aforementioned maximum likelihood loss from eq. (5.4.22) is applied during training. The hazard and survival annotations are obtained following eq. (5.4.14) and eq. (5.4.15). For both censored and uncensored patients, the hazard remains undefined after t^*. That is also reflected in the loss function, which does not take into account any intervals after the event time.

5.6.4 Risk score

A risk score per patient can be estimated with eq. (5.4.16). It follows that risk $r = 0$ if $\forall\ t_j : S(t_j) = 1$ and $r = 1$ if $\forall\ t_j : S(t_j) = 0$, or - in words - the later the survival curve drops, the lower the risk. Still, a single scalar risk value is difficult to interpret if it is not put in relation to other patients. That is especially true since the risk values may not be distributed uniformly among the population.

In order to obtain a comparable risk value, patients can be grouped into discrete risk groups. The number of risk groups needs to be chosen deliberately since there is a tradeoff between good stratification and clinical utility. For example, defining only two well-stratifying groups, low-risk versus high-risk, is insufficient to decide on an individual treatment. Also, assigning each patient their own risk group might yield individual predictions, but the stratification suffers. In the preferred case, the maximum number of groups is chosen subject to sufficient patient stratification. This means risk groups need to be defined based on the dataset. Each risk group is limited by an interval with a lower and an upper risk-score limit.

In this thesis, an exploratory approach will be applied to find the optimal intervals: First, multiple possible interval limits between the lowest and highest training set risk scores (minimum 0 and maximum 1) are defined in steps of, e.g., 0.03 (0.03, 0.06, 0.09, 0.12, etc.), and d is chosen as the desired number of intervals. For each possible combination of d intervals, the patients are assigned to their appropriate risk groups. For example, if $d = 4$, the explored interval limit combinations are ([0.03,0.06,0.09], [0.03,0.06,0.12], [0.03,0.06,0.15], etc.), where the first combination yields the 4 intervals [0,0.03), [0.03,0.06), [0.06,0.09), and [0.09,1]. A patient with a risk of 0.02 would be assigned to the first group. The patients' survival times are summarized per group with KM curves and all curves are tested for discrimination power with a log-rank test. Then it is counted how many of the proposed risk groups stratify well. If no interval combination leads to perfect stratification (i.e., at least one log-rank test fails), the number of intervals is decreased to $d - 1$. The procedure is repeated with all possible combinations of $d - 1$ intervals, and the number of intervals is decreased until all log-rank tests pass.

Multiple interval combinations for d groups may lead to a perfect stratification between groups on the training set. In that case, these combinations are further evaluated on the validation set. As before, the patients are assigned to the d risk groups, and the stratification of the KM curves is tested with log-rank tests. The interval combination that yields the best result on the validation set (the least failed log-rank tests) is chosen for the final risk group assignment. The evaluation of the risk group stratification is performed on the test set. As mentioned above, survival prediction with neural networks allows for non-proportional hazards and therefore crossing survival curves. Therefore the log-rank test is modified with Fleming-Harrington weights (see section 5.3 *Population-based survival prediction*).

5.6.5 Multiple images

In a study by Vollmer (2009), 75 % of patients had a tumor with a volume less than 25 % of the prostate. Thus, a core extracted from a prostate affected by cancer may not contain tumor tissue. Even though the TMA spots in Surv1 are taken from the tumorous area of the prostate after RPE, it is possible that not all tumor information is captured within a single image. Analyzing more tissue may give a better overview and lead to more precise predictions. Also in practice, up to twelve cores of tissue are extracted during a biopsy, and it has been shown that using more cores yields higher detection rates (Hu et al., 2019). The prediction with a second core per patient can be simulated with Surv2. It consists of a second image per patient for most patients in Surv1. The images are obtained from a different region in the patients' prostates. Furthermore, an evaluation dataset, SurvHetero, is available, which comprises up to 6 images per patient. These patients are partly distinct from the other datasets, and the annotations are only known to the UKE Institute of Pathology. Multiple images of the same patient can be combined for the survival prediction inference in different ways. One option is to use all image patches at once, thus only inputting more patches into eCaReNet than before (image concatenation). Another option is to process multiple images sequentially and then combine the predictions (mean, pessimistic, or optimistic vote). All options are introduced here:

Image concatenation Since eCaReNet uses patches, only the size per patch but not the number of patches is predefined. Therefore, it is possible to input fewer or more image patches during inference than were used during training. For the approach of image concatenation, all image patches of both images are input at once into the network, so that the trained attention layer decides which patches to weight more or less for the final prediction.

Mean Instead of using both images at once, another option is to input each image separately into eCaReNet. The survival predictions can then be averaged per interval to obtain the final survival curve.

Pessimistic Imagine a patient for whom two images of prostate cancer are given, one with, the other without cancer. The image with cancer may be more important for the final decision than the benign image. When averaging the results, this effect is lost. Therefore, another option is to use the pessimistic vote, that is, the survival curve with the lowest survival probability at each time interval.

Optimistic The opposite of the pessimistic vote is an optimistic vote, which uses the highest predicted survival probability per time interval across the input images. That method might outperform the mean or pessimistic voting if most images show cancer, and a benign image reduces the predicted risk.

5.6.6 Multimodal data input

Complex processes in the human body influence the occurrence of prostate cancer relapse. Hence the tumor tissue may not include all the relevant information to predict cancer progression. As described in chapter 3 *Datasets*, some datasets also contain additional patient information and the KM curves in Figure 3.1.7 indicated a correlation between some patient features (e.g., tumor volume) and survival time. Thus, it is interesting to examine whether a multimodal approach that includes additional patient features influences the survival prediction. Furthermore, it was shown in a complementary project (Fuhlert et al., 2022) and in the work of other researchers that relapse prediction from EHRs alone is successful (e.g., Vale-Silva and Rohr, 2021).

Since the model developed for this thesis should be applicable at the time of a biopsy, only features available before an RPE should be taken into account, as indicated in Table 3.1.7. A Gleason score can be obtained from the biopsy, but for all patients in the Survival dataset, only the Gleason score for the whole prostate is available, not for the individual images. Since that score includes more information than is available during a biopsy, it is neglected. Furthermore, a per-image Gleason score would be redundant since it is derived from the tissue, and the input image contains the necessary information already. In this thesis, PSA value, age, tumor diameter, and tumor volume are considered since these features are available or can be estimated at the time of biopsy.

For the multimodal analysis, these additional patient features can be attached to the output of the global average pooling layer, equivalent to the binary survival prediction from M_{Bin}. The model's predictive performance will be evaluated with respect to adding these patient features.

5.7 Experiments

In the following, the survival prediction experiments that are conducted during the work for this thesis are presented. Extensive evaluation of metrics and comparisons to pathologists are performed. Further, experiments using multiple images and multimodal input data are shown. Parts of the shown results have been published in Dietrich et al. (2021).

5.7.1 Setup: eCaReNet

The models are all implemented in Python3[1] with TensorFlow[2] and keras[3] (van Rossum and Drake, 2009; Abadi et al., 2015; Chollet et al., 2015). They are partly trained on an NVIDIA Tesla V100 with 16GB, and partly on an Nvidia Quadro RTX 8000 with 48 GB. The code for eCaReNet is publicly available at `www.github.com/imsb-uke/ecarenet`. All survival prediction models are trained with a Nadam optimizer (Dozat, 2016), the loss introduced in eq. (5.4.22), a learning rate of 0.00005, and early stopping such that the epoch with the lowest validation loss is used for the best model. For all conducted experiments no model layer was frozen since it showed lower performance in the previous Gleason classification task. Since a random initialization might influence the training, the models are trained five times with different initialization seeds to enable a comparison of the variation between runs of the same model and between models with different setups. If not stated differently, a single TMA spot image, cut into patches, is used as input for a patient. As data augmentation, the images are randomly flipped and rotated by 90 degrees. Modifying the images' hue, saturation, brightness, or blurring the images did not increase performance and is therefore neglected in the experiments.

The survival time is encoded into a hazard annotation per interval with eq. (5.4.15). The model predicts a hazard rate per time interval, which is always postprocessed to a survival curve showing the relapse-free survival probability over a time of 7 years. That period covers the 90 % of relapses that occur before 7 years after RPE. The intervals should not be chosen too small (e.g., one week per interval) since usually patients are not monitored that closely. Large intervals, however, e.g., 1 year, impede meaningful differentiation between patients, which is needed for treatment recommendations. Therefore, for this thesis, equal intervals of three months are chosen. Thus, to cover a span of 7 years (84 months), 28 intervals are used.

5.7.2 Binary survival prediction

First, the binary prediction model $\mathrm{M}_{\mathrm{Bin}}$ of relapse-free survival beyond two years is trained on Surv1. The predictions will be used as extra input in addition to the images for eCaReNet. A time horizon of 2 years is chosen since this is close

[1] `https://www.python.org/` (last accessed November 24, 2022)

[2] `https://www.tensorflow.org/` (last accessed November 24, 2022)

[3] `https://keras.io/` (last accessed November 24, 2022)

to the median (26.8 months) of the relapse times (44 % of relapses occur before 2 years) in Surv1. Since all patients with a censoring time before 2 years need to be removed from the dataset, it reduces to 8,200 images in the training set, 1,765 in the validation set, and 1,790 in the test set. It is unbalanced, with 89 % of patients surviving beyond 2 years since many patients are censored afterward.

The confusion matrices for the best M_{Bin} on the validation and test sets are shown in Figure 5.7.1. The accuracy reaches 0.71 on the validation and 0.67 on the test set. The confusion matrix reveals that the wrong predictions are evenly spread among both classes in the validation set, but more patients with a relapse within two years were predicted to not have a relapse on the test set. This is attributed to the unbalanced dataset. Nevertheless, M_{Bin} does not completely overfit on the most prevalent class and its predictions are thus included in eCaReNet as described in the following section.

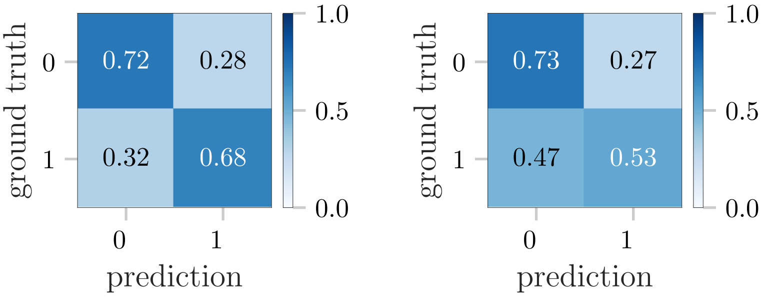


Figure 5.7.1: Confusion matrices for the best run of M_{Bin} for binary survival prediction. Left: validation set, right: test set.

5.7.3 eCaReNet ablation study

As described above, not all parts of eCaReNet are mandatory to enable survival prediction but some parts are expected to contribute to increased performance or usability. While the InceptionV3 backbone and the GRU layer are mandatory, attention MIL, self-attention, and binary survival prediction are optional.

Therefore, in the following, eCaReNet is adapted to evaluate which architectural parts contribute most to model discrimination power and calibration. Furthermore, the influence of pretraining the model on the histopathology images from the Gleasonaut instead of ImageNet is evaluated.

Table 5.7.1 summarizes all ablation study results, where a higher AUC and C-index but a lower Brier score indicate better model performance. For d-calibration, it is only indicated whether the chi-square test passes or fails. As model input, the images are downsized to 1024×1024 pixels, which is the resolution used in $\mathrm{M}_{\mathrm{ISUP}}$ (see section 4.4.2 *ISUP grading*).

Results As base model $\mathrm{M}_{\mathrm{base}}$, a GRU layer is added to the InceptionV3 network, and the H&E image is input as a whole. That model is pretrained only on the ImageNet dataset and does not include the previous binary prediction. The base model $\mathrm{M}_{\mathrm{base}}$ reaches an AUC of 0.74, a C-index of 0.72, and a Brier score of 0.116. The d-calibration chi-square test passes.

As a first adaptation ($\mathrm{M}_{\mathrm{pretr}}$), the pretrained weights of $\mathrm{M}_{\mathrm{ISUP}}$ from section 4.4.2 *ISUP grading* are used. The results of $\mathrm{M}_{\mathrm{pretr}}$ show that pretraining on histopathology images has a positive effect on all metrics. The AUC increases to 0.76, the C-index increases to 0.73, and the Brier score decreases to 0.109. This is expected since the visual content of ImageNet images and histopathology images differs, and a model pretrained on histopathology images produces more valuable latent space representations than a model pretrained on ImageNet. Since all model weights are fine-tuned, without freezing any layers, it is concluded that the transfer learning from both a different dataset source (ImageNet images) and task (classification) to histopathology survival prediction is too challenging to obtain optimal results.

Next, an attention-based MIL layer is added after the GRU layer for $\mathrm{M}_{\mathrm{MIL}}$. The MIL layer enables insights into the model's predictions by revealing the contribution of each image patch to the final prediction. While this approach has the advantage of adding transparency to the survival model, it also adds more complexity. The input image is cut into non-overlapping patches for this approach. Using 16 patches with size 256×256 pixels showed the best results. With the patch-based MIL approach $\mathrm{M}_{\mathrm{MIL}}$, the C-index improves to 0.74 on both the validation and test set, while the AUC and the Brier score remain unchanged. That alone would suggest no clear advantage of using image patches. However, the attention in the MIL includes explanations about which image regions contribute most to the final prediction. Therefore, it is concluded that adding explainability does not reduce prediction performance and should hence be kept in the model.

Next, the prediction output from $\mathrm{M}_{\mathrm{Bin}}$ (section 5.7.2 *Binary survival prediction*) is added to the global average pooling layer in $\mathrm{M}_{\text{MIL-Bin}}$. That binary task can already hint toward an early or late relapse time and therefore support the survival prediction. The AUC improves to 0.77 on the validation and test sets, the Brier score improves to 0.107 on the validation set, and 0.109 on the test set. The C-index remains as high as before (0.74). Therefore, also the binary prediction

Table 5.7.1: Comparison of model adaptations. Values are the mean of five training runs with the standard deviation in parentheses for the validation (Valid) and test sets. When models M_{ISUP} (M_I) or M_{Bin} (M_B), or M_{MIL} (M_M) or self-attention (s) layers are included, it is indicated with a dot (•). The best results are marked in bold. MIL=multiple instance learning, Bin=including binary relapse prediction from M_{Bin}. All d-calibration (D) chi-squared tests pass (p).

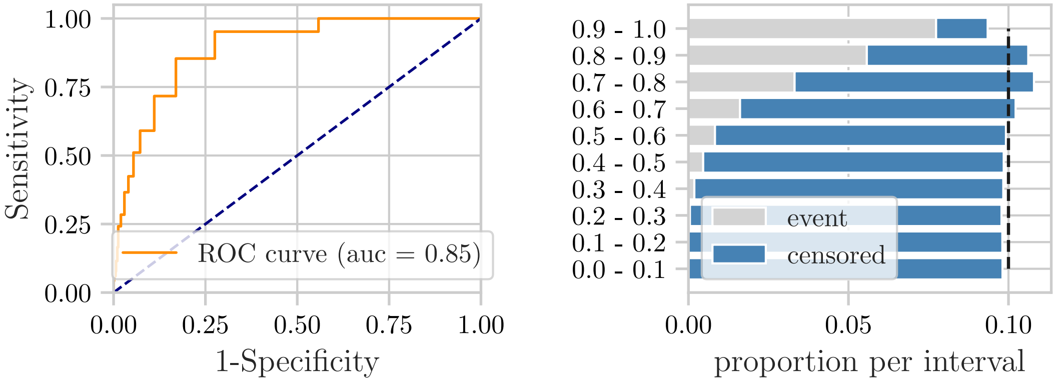
Valid set	$\mathbf{M_I}$	$\mathbf{M_M}$	$\mathbf{M_B}$	**s**	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D**
M_{base}					0.74 (0.0042)	0.72 (0.0008)	0.116 (0.0038)	**p**
M_{pretr}	•				0.76 (0.0018)	0.73 (0.0023)	0.109 (0.0005)	**p**
M_{MIL}	•	•			0.76 (0.0004)	0.74 (0.0000)	0.109 (0.0000)	**p**
$M_{MIL\text{-}Bin}$	•	•	•		0.77 (0.0012)	0.74 (0.0026)	**0.107** (0.0003)	**p**
eCaReNet	•	•	•	•	**0.78** (0.0041)	**0.75** (0.0016)	**0.107** (0.0004)	**p**
Test set								
M_{base}					0.74 (0.0054)	0.71 (0.0031)	0.115 (0.0007)	**p**
M_{pretr}	•				0.76 (0.0031)	0.73 (0.0018)	0.110 (0.0004)	**p**
M_{MIL}	•	•			0.76 (0.0002)	**0.74** (0.0003)	0.110 (0.0000)	**p**
$M_{MIL\text{-}Bin}$	•	•	•		**0.77** (0.0011)	**0.74** (0.0022)	**0.109** (0.0003)	**p**
eCaReNet	•	•	•	•	**0.77** (0.0048)	**0.74** (0.0037)	**0.109** (0.0006)	**p**

should be kept to ensure high discrimination performance. It has to be kept in mind that the binary prediction only discriminates between surviving two years or not, while the survival model $M_{MIL\text{-}Bin}$ is evaluated over a time of 7 years.

Finally, for eCaReNet, a self-attention layer is included to account for inter-patch influences. These influences are relevant since the final survival curve is not predictable from single patches, but, e.g., the ratio of benign and malignant image regions is relevant, too. Here, using 64 patches (with 128×128 pixels each) showed the best results. The AUC and C-index on the validation set improve to 0.78 and 0.75, respectively, however, this effect cannot be seen on the test set. In addition, the variance increases. This shows that network initialization has a greater influence on the final prediction performance with self-attention. It can be concluded that the inter-patch dependencies add little additional information.

For all model variations, the d-calibration chi-square test passes, assuring calibration. Furthermore, the models generalize well, as there is only a slight performance decrease observable from the validation set to the test set. Evaluation of the results on the separate test set 13.1D, only containing a single TMA, also results in AUC scores of 0.74-0.76 for all adaptations, confirming equally good performance and generalizability.

Conclusion: Best variant In conclusion, the best model, eCaReNet, includes self-attention and attention-based MIL layers, the prediction from M_{Bin}, and is pretrained on ISUP scores (built on M_{ISUP}). As input, the H&E images are cut into 64 patches of size 128×128 pixels. The model reaches an AUC of 0.78 on the validation and 0.77 on the test set. That is an AUC integrated over time, but the AUC can also be calculated at single time points. The single best eCaReNet (initialization seed that led to the best AUC) reaches the highest AUC on the test set at 18 months, where it reaches 0.85, see Figure 5.7.2 (a). The time-dependent C-index reaches 0.75 and 0.74 on the validation and the test set, respectively. The Brier scores are 0.107 on the validation set and 0.109 on the test set, thus, below 0.25 and indicate a good model performance. The d-calibration plot for the best model is shown in Figure 5.7.2 (b). The chi-square test passes, indicating a uniform distribution and thus a well-calibrated model.



(a) Test-set AUC at $t = 18$ months.

(b) Test-set d-calibration.

Figure 5.7.2: Results of eCaReNet on the Surv1 test set for a single trained model.

5.7.4 Comparison to baseline and pathologist

In order to evaluate eCaReNet further, its performance is compared to models and loss functions proposed in the literature. It is also compared to a pathologist's annotations, which are not survival estimates but Gleason scores for the patients' whole prostates. All results are summarized in Table 5.7.2 using the same metrics as before. Since the pathologist only assigned a single score per patient and does not predict a survival curve, only discrimination can be evaluated, not calibration.

Two models from the literature, DeepConvSurv (Zhu et al., 2016) and CDOR (Xiao et al., 2020), are used for comparison. These are chosen among the models presented in section 2.2.4 *Deep learning for survival prediction* since DeepConvSurv is the standard Cox-based survival model and CDOR shows the best performance on only histopathology images for survival curve prediction. Both those models have different underlying CNNs, optimized for the applied datasets. In order to enable a fair performance comparison in this thesis, both base models are exchanged with an InceptionV3 as used for eCaReNet. In particular, the InceptionV3 model pretrained on ISUP scores, $\mathrm{M}_{\mathrm{ISUP}}$ (see section 4.4.2 *ISUP grading*), is used since the ablation study revealed that the pretraining already improves the performance. Comparing ImageNet-pretrained models from the literature with eCaReNet pretrained on histopathology images would yield misleading conclusions.

Comparison to baseline models For DeepConvSurv, a single output node is added after the global average pooling layer of the pretrained InceptionV3. That output node returns the patient's risk and models the exponential part of the Cox model (see eq. (5.4.6)). DeepConvSurv is trained with the negative log partial likelihood loss from eq. (5.4.8). That model reaches an AUC of 0.69 and a C-index of 0.65 on the validation set (see Table 5.7.2). The AUC on the test set is 0.71 with a C-index of 0.64. The test for d-calibration fails, and the Brier score of 0.305 on the validation set (0.296 on the test set) indicates a non-calibrated model, too.

For CDOR, fully connected layers are attached to the pretrained InceptionV3 after the global average pooling. Since it has the same intervals as eCaReNet, 28 output nodes are needed. CDOR predicts the survival probability per interval directly without modeling the hazard first. However, a constraint that enforces monotonically decreasing survival curves is missing. That model reaches an AUC of 0.77 on the validation and 0.78 on the test set. The C-index is 0.73 for both sets. Like DeepConvSurv, it fails in terms of calibration. Furthermore, the resulting survival curves are not monotonically decreasing, therefore biologically unreasonable (see Figure 5.7.3).

Compared to both previously described models from the literature, eCaReNet shows the best performance for all measures on the validation set (AUC 0.78, C-index 0.75, Brier score 0.107). On the test set, it also obtains the best C-index (0.74) and Brier score (0.109) and passes the chi-square test for d-calibration. CDOR performs best on the test set's AUC. In contrast to CDOR, eCaReNet predicts monotonically decreasing survival functions. Figure 5.7.3 compares the predicted survival curves for three uncensored patients. CDOR and eCaReNet both predict the order of the patients correctly, but the individual predictions per interval differ. Furthermore, eCaReNet passes the d-calibration chi-square test and is thus better calibrated than CDOR and DeepConvSurv. Another advantage is that eCaReNet offers an intuitive explanation for its decisions through patch-wise attention weights, which will be evaluated later in more detail.

Table 5.7.2: Comparison of eCaReNet to state-of-the-art models and a pathologist's ISUP annotations. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D), only failure (f) or pass (p) is indicated. The best results are marked in bold.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D**
ISUP	**0.78**	**0.75**	-	-
DeepConvSurv[Zhu et al. (2016)]	0.69 (0.0207)	0.65 (0.0173)	0.305 (0.0146)	f
CDOR[Xiao et al. (2020)]	0.77 (0.0089)	0.73 (0.0046)	0.111 (0.0014)	f
eCaReNet	**0.78** (0.0041)	**0.75** (0.0016)	**0.107** (0.0004)	**p**
Test set				
ISUP	**0.80**	**0.76**	-	-
DeepConvSurv	0.71 (0.0232)	0.64 (0.0132)	0.296 (0.0227)	f
CDOR	**0.78** (0.0005)	0.73 (0.0003)	0.110 (0.0001)	f
eCaReNet	0.77 (0.0048)	**0.74** (0.0037)	**0.109** (0.0006)	**p**

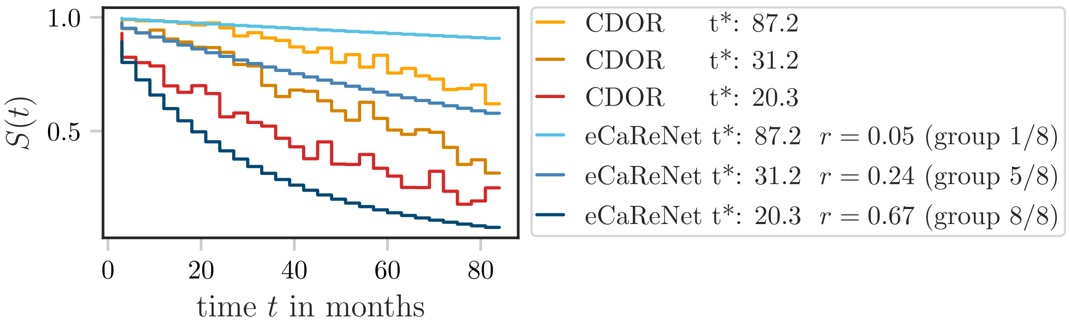


Figure 5.7.3: Predicted survival curves of eCaReNet compared to CDOR. While the order of the three patients is correctly predicted with both models, only the survival curves of eCaReNet are monotonically decreasing. t^*: ground truth survival time, r: predicted risk score, group a/b: predicted risk group a out of b groups.

Comparison to pathologist Pathologists assign ISUP scores to patients instead of directly estimating the time to relapse, whereby higher ISUP scores correspond to higher risk and, thus, lower survival probability. Therefore, the discrimination power of the assigned ISUP scores and eCaReNet's prediction can be compared. The pathologist's annotations are assumed to be constant over time for AUC and C-index calculation. In terms of validation set discrimination, eCaReNet reaches on-par performance with the pathologist's annotations (AUC 0.78, C-index 0.75). On the test set, the pathologist reaches a higher AUC and C-index than eCaReNet (AUC 0.80, C-index 0.76). In this context, it is important to note that in contrast to eCaReNet, which analyzes single TMA spots, the pathologist had access to the patient's whole prostate tissue.

Even though the pathologist's assigned risk scores are constant over time, the AUC can be evaluated at different time points, see eq. (5.5.8). The performance of eCaReNet and the pathologist is therefore compared in terms of AUC for a time of 84 months (7 years) in Figure 5.7.4. The curves on the validation and test sets are similar and, therefore, not addressed separately in this analysis. In the time range from 6 to 26 months after RPE, eCaReNet's AUC is higher than 0.8, with the best test set AUC at 18 months, as stated above. Hence, eCaReNet's predictions are best in a time range from half a year to two years after RPE. Overall, eCaReNet's AUC is similar to the expert pathologist. In the first months, where eCaReNet's survival curves are still close to 1 for all patients, it is outperformed by the pathologist, whose predictions are constant over time. The pathologist and eCaReNet perform similarly in the time range from 15 to 42 months. Afterward, the pathologist again outperforms eCaReNet in terms of AUC. Since eCaReNet and the pathologist achieve similar discrimination, it is concluded that eCaReNet has the potential to support pathologists' decisions.

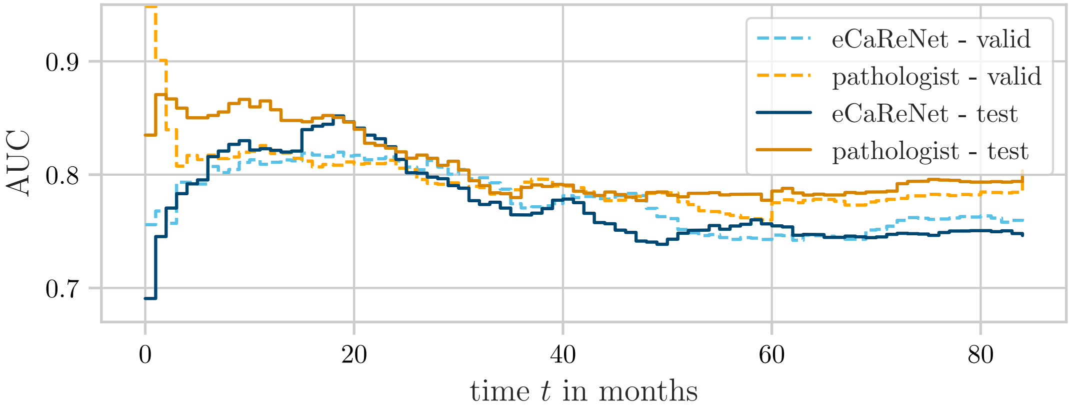


Figure 5.7.4: AUC over time of eCaReNet compared to a pathologist. Test set results are shown with solid lines, validation set results with dashed lines. The pathologist's performance is shown in orange, and eCaReNet's performance is shown in blue.

Conclusion It was shown that eCaReNet achieves similar discrimination on Surv1 as a pathologist and outperforms state-of-the-art approaches by being calibrated and predicting biologically reasonable survival curves. However, the performance does not increase beyond the predictive performance of the ISUP annotations from a pathologist. It should be noted again, however, that eCaReNet only uses a single image per patient for its prediction, while the pathologist had access to the whole prostate. These results are encouraging for future work. It needs to be investigated further whether pathologists' patient stratification and decision-making can improve when being supported by eCaReNet's predictions.

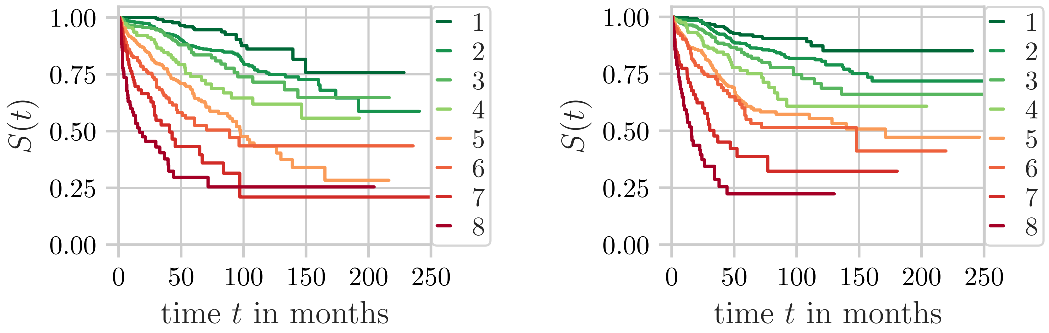
5.7.5 Risk group evaluation

Besides describing a patient's relapse-free survival probability with a survival curve, his overall risk can be condensed to a single risk score, and he can be assigned to a risk group. An exploratory approach for defining the number of differentiable risk groups is applied for eCaReNet, as described in section 5.6.4 *Risk score*. As initial possible interval limits, steps of 0.03 between the lowest training set risk 0.009 and the highest risk 0.85 are defined as 0.03, 0.06, ..., 0.84. As a starting number of differentiable groups, $d = 10$ is chosen. d is decreased as long as the log-rank tests do not all pass between the KM curves per risk group on the training set.

That approach leads to a maximum of eight risk groups that are still well stratified in the training set, according to the log-rank test. The limit for the p-value is 0.05. The eight intervals that lead to the best training set stratification are, from low to high risk, [0, 0.06) [0.06, 0.12),[0.12, 0.15), [0.15, 0.18),[0.18,0.3),[0.3, 0.42), [0.42,0.51), and [0.51, 1]. The lowest risk group is named group 1, and the highest is named group 8. The uneven spacing of the intervals confirms that a data-driven approach for defining the risk groups is beneficial since pure uniform partitions would not yield such well-stratified groups. In contrast to state-of-the-art models, where mostly two to three groups are distinguished, differentiating eight groups enables a more individualized prognosis. That again stresses the discriminatory power of eCaReNet. Figure 5.7.3 shows an example of survival curves and their associated risk scores and groups.

The log-rank tests pass for all risk groups in the training dataset. The KM curves on the validation and test set are shown in Figure 5.7.5. For the validation set, only one log-rank test fails (p-value 0.219), indicating that groups 2 and 3 do not stratify sufficiently. The log-rank test for the test set fails for groups 3/4 (p-value 0.07) and 5/6 (p-value 0.306). Therefore, 6 groups are still well-stratified in the test set.

Most test set patients are assigned to risk group 2 (835 patients - 38 % of the test set data). The group with the fewest patients is risk group 8, which includes 66 patients (3 % of the test data). That is reasonable since most patients' cancer progresses slowly, and in the given dataset most patients did not report a relapse throughout the study.



(a) Validation set.

(b) Test set.

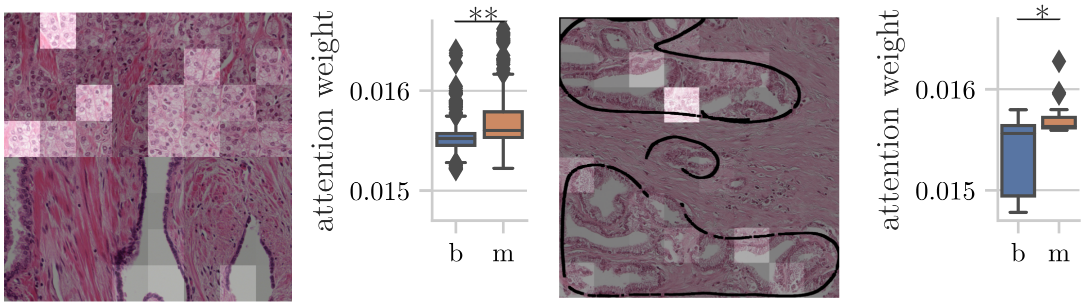
Figure 5.7.5: KM survival curves per risk group on the validation and test sets of Surv1.

Conclusion With eCaReNet, it is possible to stratify up to eight risk groups in the training set and six risk groups in the test set, which allows for more individualized decisions in comparison to current state-of-the-art methods, which often only discriminate two or three risk groups.

5.7.6 Attention evaluation

So far it has been demonstrated that eCaReNet has high discriminatory power, is well-calibrated, and generalizes to an unseen test set. In order to apply such a model in a clinic, trustworthy and explainable predictions are crucial besides high prediction performance. Since deep neural networks act as black boxes, their decisions are incomprehensible. As described in chapter 2 *Background*, the term "explainability" has no clear connotation in the literature. Here, explainability is included in eCaReNet by inspecting the attention weights assigned to single image patches in the MIL layer. It is expected that image regions showing cancer receive more attention than benign tissue.

Synthetic example with stitched patches For the first experiment, synthetic images are stitched from cancerous and non-cancerous patches. For this, an image that is labeled as benign (Gleason 0) and an image that is labeled as highly cancerous (Gleason pattern 5), are stitched together in such a way that half the image shows benign and half the image shows cancerous tissue. Since the images in Surv1 are not labeled image-wise, the Gleasonaut test set is used for this experiment, of which 12 example images are stitched. Each stitched image is cut into patches, input to eCaReNet, and the attention weights that are assigned per patch in the MIL layer are calculated. In the example image in Figure 5.7.6 (a), patches with higher attention are highlighted in a lighter color. It is shown that the malignant, upper half receives more attention than the lower, benign part. The attention that malignant and benign patches are assigned is plotted in the boxplot in Figure 5.7.6 (b). It confirms that malignant image regions receive significantly more attention than benign patches (p-value < 0.01).



(a) (b) (c) (d)

Figure 5.7.6: Evaluation of attention scores in benign and malignant patches. (a) Example stitched image with Gleason pattern 0 in the lower and 5 in the upper half, where patches are lighter the more attention they are assigned. (b) Comparison of attention assigned to benign and malignant patches. The box indicates the upper and lower quartiles, and the black line is the median value. b: benign, m: malignant, ** indicates that the difference is significant with a p-value < 0.01. (c) Example image from Surv1. Patches with higher attention are shown in a lighter color. The cancerous areas are circled in black. (d) Comparison of attention assigned to benign and malignant patches (*: $p < 0.05$). Adapted from Dietrich et al. (2021).

Real examples In order to further explore whether the attention weights match with cancerous regions in the Surv1, two images of each of 38 TMAs of Surv1 were randomly selected (one from the validation and the other from the test set) and shown to a pathologist (Prof. Dr. Guido Sauter, UKE Institute of Pathology). In each image, the pathologist marked the cancerous regions with a pen. Each image is then cut into 64 patches with 128 × 128 pixels (input size for eCaReNet), and a patch is considered malignant if at least 2/3 of its tissue are within the marked area. Images showing no cancerous structures are left out since the attention cannot focus on any malignant parts, hence, the results might be misleading.

The original image patches (without marks) are processed with eCaReNet and the attention weight per patch is extracted. The average attention the patches receive is calculated separately for all patches annotated as malignant and for all patches annotated as benign. The average attention of malignant patches compared to benign patches is shown in a boxplot in Figure 5.7.6 (d) for the test set. It can be seen that malignant patches are assigned significantly higher attention weights (p-value < 0.05) than benign patches. Figure 5.7.6 (c) shows an example test set image with superimposed attention. Patches with high attention are illustrated bright whereas patches with low attention are darker. The regions circled with black lines are annotated as cancerous by a pathologist. As can be seen, not all image regions that are marked as malignant obtain high attention weights. However, all patches that receive the highest attention are within the malignant regions. Still, the attention weights are all within a small range around 0.015625, which would be an equal attention across 64 patches.

Conclusion The attention assigned per image patch in the MIL layer correlates to malignancy. On average, malignant patches are assigned higher attention values. Thus, eCaReNet incorporates an explainability that makes it more transparent and could help pathologists to decide whether to trust the provided predictions. It can also assist pathologists' decisions by determining the most relevant image regions the pathologist should focus on.

5.7.7 Multiple images per patient

Since a single TMA spot might not be representative of a patient's disease status, in the following experiments, multiple images per patient are used during evaluation. The different options to combine the information of multiple images per patient during inference explained in section 5.6.5 *Multiple images* are compared on SurvMulti and SurvHetero.

SurvHetero First, eCaReNet is trained on Surv1 and evaluated on SurvHetero. Since only the institute of pathology has access to the annotations for that dataset, SurvHetero can be used only for evaluation, not for training. All patient survival curves are converted into scalar risk scores, and the results are sent to the institute of pathology for evaluation. Instead of calculating metrics over time, the AUC at 5 years after RPE is evaluated, which is why no other metrics are available here.

SurvHetero contains 828 patients and up to 6 images per patient. A risk score and a risk group are calculated per image, which is why, for one patient, multiple risk scores and groups are available that need to be combined. For most patients, the minimum and maximum predicted risk groups in all their images are adjacent. However, for some patients, one of their TMA spots is predicted as risk group 1 (no cancer) while another is predicted as the highest risk group 8. This shows the heterogeneity of the images in SurvHetero. For the following evaluations, all available images per patient (4,181 images in total) are used to predict relapse-free survival. Calculating the mean or the pessimistic vote of the survival curves is compared to inputting all image patches at once to eCaReNet. It has to be noted that eCaReNet is again trained on a single patient image from Surv1 and evaluated on multiple images. All results are summarized in Figure 5.7.7.

When using the mean of all predictions and the most pessimistic prediction per patient, the AUCs are equal at 0.78. Concatenating all images leads to a similar AUC of 0.77. It is concluded that the method of combining multiple predictions does not influence the results much since the differences between the results are within the model variance observed in the previous experiments (compare Table 5.7.2). Since the AUC calculates the discriminative power, it is reasonable that using the pessimistic vote for all patients or the mean for all patients does not make a great difference in the predicted patient order. The discrimination is similar to the results on Surv1, which confirms eCaReNet's generalizability to unseen data.

The pathologist annotated every single image in SurvHetero, and the annotations were combined in different ways. Thus, the pathologist and eCaReNet use the same information, as opposed to Surv1, for which the pathologist had access to the whole prostate tissue. When assigning the maximum Gleason score of all images per patient, the pathologist reaches an AUC of 0.76 and thus is outperformed by eCaReNet. If the more differentiated IQ Gleason score is used and summed up over all images, the pathologist and eCaReNet (using the pessimistic or mean voting) are on par with an AUC of 0.78. The pathologist only scores higher than eCaReNet when he assigns the maximum IQ Gleason score over all images to a patient, reaching an AUC of 0.79.

Different aspects could contribute to keeping the AUC below 0.8 for both

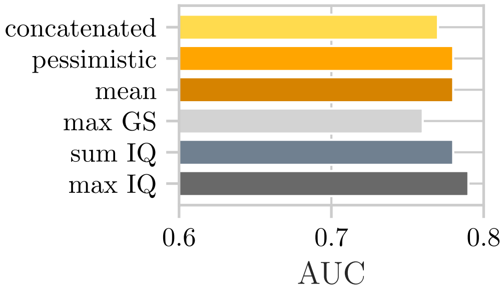
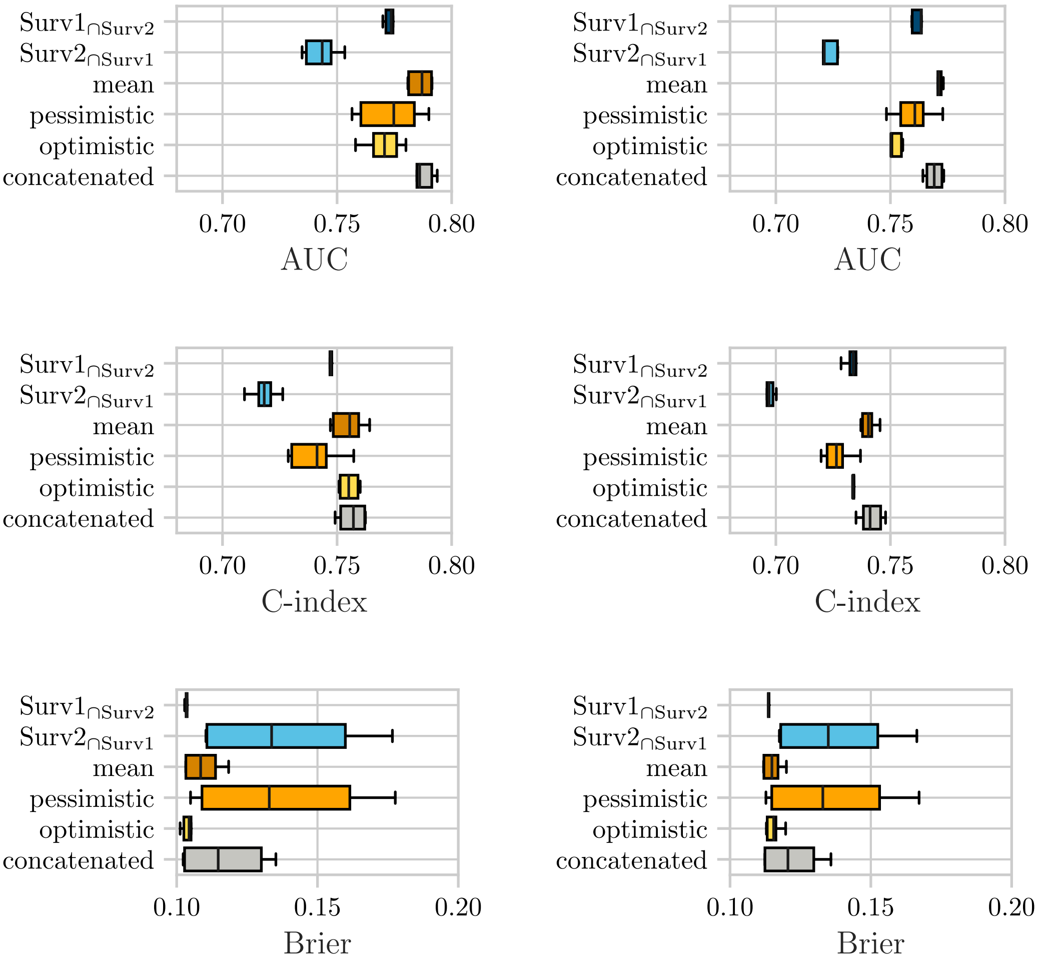


Figure 5.7.7: AUC for 5-year relapse-free survival of eCaReNet (concatenated, pessimistic, and mean voting) on SurvHetero, compared to the pathologist's annotation (max GS, sum IQ, and max IQ). GS: Gleason score, IQ: integrated quantitative Gleason score.

eCaReNet and the pathologist. One limitation is that the tissue alone does not include all information necessary to predict relapse times since also the patient's lifestyle or other clinical features are relevant.

SurvMulti For most patients in the internal datasets, two images are available with the "regular" data acquisition protocol (tissue 2.5 µm thick, normal staining time, Leica Aperio scanner), one image in Surv1 and one in Surv2, showing tissue from different regions in the prostate. These are used to evaluate if a second image per patient can improve predictive performance. eCaReNet is trained as before on the complete Surv1, so the images from Surv2 are only used during inference.

For evaluation, both datasets are reduced to only comprise patients with images in each of both datasets. The resulting datasets are therefore called $\text{Surv1}_{\cap\text{Surv2}}$ and $\text{Surv2}_{\cap\text{Surv1}}$. While $\text{Surv1}_{\cap\text{Surv2}}$ contains images from Surv1 but is reduced to those patients who also have an image in Surv2, $\text{Surv2}_{\cap\text{Surv1}}$ contains images from Surv2 but is reduced to those patients who also have an image in Surv1. Both datasets comprise 1,878 images each in the validation set and 1,895 images each in the test set. To allow for a fair comparison of prediction performance, as a baseline eCaReNet needs to be evaluated on $\text{Surv1}_{\cap\text{Surv2}}$ since that is only a subset of Surv1. Thus, eCaReNet is first evaluated on both these subsets $\text{Surv1}_{\cap\text{Surv2}}$ and $\text{Surv2}_{\cap\text{Surv1}}$ separately and then on a combination, using two images (one from each subset) for per-patient predictions. Again, using all patches of both images at once (concatenating the images) is compared to mean, pessimistic and optimistic voting when predicting both images separately. All results are shown in Figure 5.7.8 as a boxplot over five runs, and with the mean as a black line. For quantitative results consider Table A.2.1.



(a) Results on the validation set patients.

(b) Results on the test set patients.

Figure 5.7.8: Model performances when using two images per patient for evaluation. The boxplots show the results over five runs with the mean value and different initialization seeds on (a) the validation and (b) the test set patients. $\text{Surv1}_{\cap\text{Surv2}}$ contains images from Surv1 but is reduced to patients who also have an image in Surv2.

The performances on the $\text{Surv1}_{\cap\text{Surv2}}$ test set are similar to the results on Surv1. The AUC on the test set is 0.77, the C-index 0.73, and the Brier score 0.114. The validation set results are also similar to before, with an AUC of 0.77, a C-index of 0.75, and a Brier score of 0.104. On both test and validation sets, the model shows good calibration.

The performance on $\text{Surv2}_{\cap\text{Surv1}}$ decreases compared to $\text{Surv1}_{\cap\text{Surv2}}$. The AUC and C-index on the test set reduce to 0.72 and 0.70, respectively, and the Brier score increases to 0.135. The same holds for the validation set (AUC 0.74, C-index 0.72, Brier 0.134). It needs to be investigated whether the performance degradation is due to differences in staining or to the images stemming from a different location in the prostate, thus including less meaningful information.

When combining the images of both datasets for evaluation, the results improve slightly. On the validation set, the performance in all metrics is best when predicting both images separately and using the mean across both survival curves per patient. The AUC reaches 0.79, the C-index reaches 0.76, and the Brier score reaches 0.109. When concatenating both images, the discrimination is equal to using the mean, but the Brier score increases to 0.115. The chi-square test fails for both variants. It only passes in the case of optimistic voting. In that case, the AUC only reaches 0.77, which is equal to the results on $\text{Surv1}_{\cap\text{Surv2}}$. The pessimistic vote reaches the lowest performance on the validation set (AUC 0.77, C-index 0.74, Brier 0.133).

The performance on the test set is also slightly higher than on $\text{Surv1}_{\cap\text{Surv2}}$ when combining both images. Again, using the mean across both predictions yields the best results while failing the d-calibration test (AUC 0.77, C-index 0.74, Brier 0.115). When concatenating both images for a prediction, the discrimination performance on the test set is again equal to the performance when using the mean of both predicted survival curves. However, the Brier score increases to 0.121, and the d-calibration test passes in this case. Using a pessimistic vote again scores lower than the best approaches (AUC 0.76, C-index 0.73, Brier 0.133).

The best results when combining two images are summarized in Table 5.7.3 and compared to a pathologist. The pathologist did not annotate two different images but each patient's whole prostate. Therefore, the pathologist's scores do not differ between $\text{Surv1}_{\cap\text{Surv2}}$ and $\text{Surv2}_{\cap\text{Surv1}}$. Regarding AUC, the discrimination performance of the pathologist's ISUP annotation is above eCaReNet. The ISUP score achieves an AUC of 0.80 on the validation and 0.82 on the test set patients. On the validation set, the pathologist achieves a C-index of 0.75. Hence, eCaReNet outperforms the pathologist. The C-index achieved with the ISUP score on the

Table 5.7.3: Model performances when using two images per patient, one from Surv1, one from Surv2. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D-cal.), only failure (f) or pass (p) is indicated. The best results are marked in bold. An extended version including pessimistic and optimistic voting is included in Table A.2.1.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
ISUP	**0.80**	0.75	-	
Surv1$_{\cap\text{Surv2}}$	0.77 (0.0039)	0.75 (0.0013)	0.104 (0.0005)	**p**
Surv2$_{\cap\text{Surv1}}$	0.74 (0.0069)	0.72 (0.0055)	0.134 (0.0287)	f
mean	**0.79** (0.0049)	**0.76** (0.0067)	**0.109** (0.0064)	f
concatenated	**0.79** (0.0066)	**0.76** (0.0056)	0.115 (0.0147)	f
Test set				
ISUP	0.82	0.76	-	
Surv1$_{\cap\text{Surv2}}$	0.76 (0.0050)	0.73 (0.0038)	0.114 (0.0006)	**p**
Surv2$_{\cap\text{Surv1}}$	0.72 (0.0077)	0.70 (0.0029)	0.135 (0.0205)	**p**
mean	**0.77** (0.0019)	**0.74** (0.0030)	**0.115** (0.0032)	f
concatenated	**0.77** (0.0036)	**0.74** (0.0049)	0.121 (0.0102)	**p**

test set is 0.75, thus, again above eCaReNet's performance.

The presented results show that eCaReNet's performance on Surv2$_{\cap\text{Surv1}}$ decreases, which can have multiple reasons. It is worthwhile exploring whether Surv2 includes less informative images than Surv1 or has a dataset bias to which eCaReNet cannot generalize. Since combining both images increases the performance over using only a single image from either dataset, it is concluded that Surv2 does include some useful additional information. In order to further investigate whether Surv2 contains less informative images or if a dataset bias hinders the generalizability, it is evaluated next how a model performs when being trained on Surv2.

Effect of using a different tissue core

It is evaluated whether the performance decreases on Surv2 because its images are less informative of the disease status or if the performance decreases because eCaReNet fails to generalize. Therefore, eCaReNet is trained on the training partition of Surv2 and evaluated on the respective validation and test sets. If the performance reaches a similar performance to Surv1, it hints at the model being sensitive to a dataset bias. If the performance remains below Surv1, the tissue in Surv2 might not be as informative as the tissue in Surv1.

The results when training and evaluating eCaReNet on Surv2 are summarized in Table 5.7.4. The AUC on the validation set reaches 0.73. That is below the performance when training and evaluating on Surv1. The C-index (0.71) and the Brier score (0.125) are also below the previously reported performances. On the test set, the model shows low generalizability since the AUC decreases to 0.69, the C-index to 0.66, and the Brier score increases to 0.142. These results indicate that the images in Surv2 include information to discriminate patients well, but the images contain less meaningful information for survival prediction than the images in Surv1.

Table 5.7.4: Performances of eCaReNet trained on Surv2. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D-cal.), the chi-square tests pass (p).

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
Surv2	0.73 (0.0132)	0.71 (0.0086)	0.125 (0.0018)	**p**
Test set				
Surv2	0.69 (0.0147)	0.66 (0.0080)	0.142 (0.0024)	**p**

Conclusion The experiments show that using multiple images per patient increases predictive performance. Thus, a single TMA spot cannot include all relevant information for survival prediction. In particular, the information included in a single TMA spot per patient varies with the location where the tissue is extracted. This was indicated by the performance difference between $\text{Surv1}_{\cap\text{Surv2}}$ and $\text{Surv2}_{\cap\text{Surv1}}$. The best way to combine multiple images is to make one prediction per image and average the predicted survival curves.

On SurvHetero, it could be seen that if more than two images per patient are available and the pathologist and eCaReNet use the same information, both perform similarly. However, there still is a gap to perfect discrimination. It is concluded that the prostate tissue alone does not include all relevant information that influences relapse times. Thus, the influence of including additional patient information along with a TMA spot image in a multimodal approach is evaluated next.

5.7.8 Multimodal analysis

All the above experiments are conducted using only TMA spot images as input. However, a TMA spot only covers a part of the disease picture. Thus, it is evaluated whether adding clinical patient features increases the model performance by providing a broader overview of the patient's disease status.

Setup As additional patient-specific features, age, PSA value, tumor volume, and diameter are used. Their influence on the prediction performance is analyzed individually and in combination. The age is available for every patient, but the PSA value, tumor volume, and diameter are unknown for some patients in Surv1. Therefore, the dataset reduces to patients for whom the values for all these four features are known, thus to Surv1AddInfo (see Table 3.1.8). Due to the removal of patients, only two patients with a censoring time after 7 years and no patients with a relapse later than 7 years remain in the training set of Surv1AddInfo. Therefore the previously chosen time horizon of 7 years is no longer reasonable and, for these experiments, the time horizon is adjusted to 5 years (60 months). That makes 20 intervals of 3 months in length, thus, 20 output nodes.

As input to eCaReNet, the additional values need to be normalized to the range $[0, 1]$. Therefore, the age is divided by 100, PSA by 1,000, diameter by 90, and volume by 110. These denominators are the maximum values in the dataset, rounded up. Again, the model is trained with an early stopping criterion on the validation loss. Furthermore, for a fair comparison, the model trained only on images needs to be re-trained on the images included in this sub-dataset Surv1AddInfo so that all differences in performance can be attributed to the additional features, not to the included patients and dataset size.

Results The results in Figure 5.7.9 show the values per metric over five runs with different seeds on the validation and test sets. All results are also summarized in Table A.2.2, with mean values and standard deviations.

The model trained only on the Surv1AddInfo images achieves a mean validation set AUC and a C-index of 0.77. On the test set, the AUC is 0.75, and the C-index is 0.74. The Brier score is 0.105 on the validation and 0.112 on the test set. These results are slightly below the previously reported results on the complete Surv1.

Adding age as a feature slightly lowers the model's predictive performance in all metrics on the Surv1AddInfo validation set (AUC 0.76, C-index 0.76, Brier 0.107). On the test set, the discrimination remains equal to using the image only when adding the age, and the Brier score increases slightly (AUC 0.75, C-index 0.74, Brier 0.113). In other words, age does not seem to impact model performance significantly. That corresponds to the KM estimate in Figure 3.1.7, which already indicated a low correlation between age and survival time.

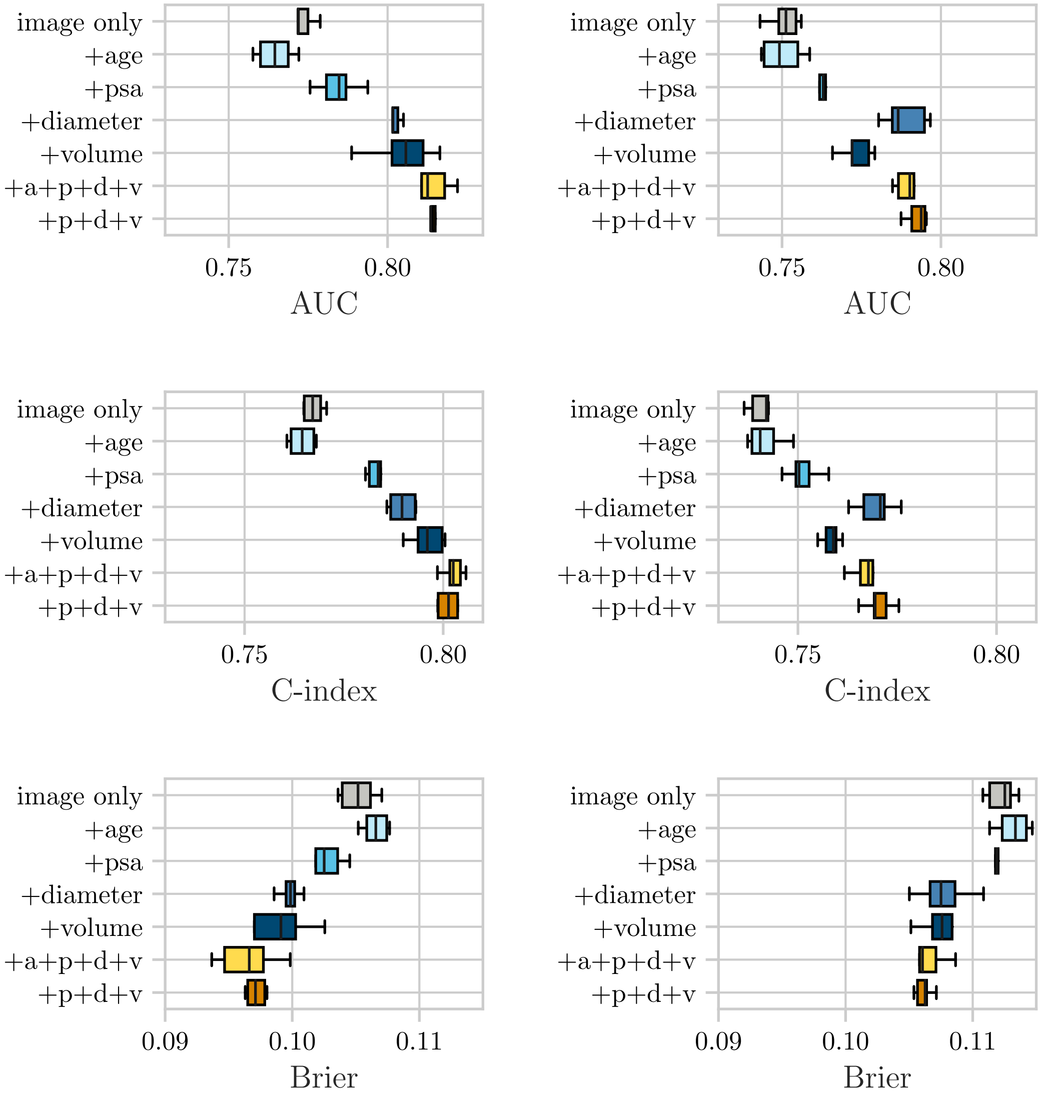
Adding only the PSA value increases the model performance in all validation (AUC 0.78, C-index 0.78, Brier 0.103) and test set metrics (AUC 0.76, C-index 0.75, Brier 0.111). That confirms the expectations based on the KM curves shown in Figure 3.1.7 and reinforces the intuitive idea that the prostate tissue alone cannot cover all relevant information for relapse prediction. It also stresses the need for a multimodal analysis of patients. Adding only the tumor volume or diameter as an additional feature also increases performance, as can be seen in Figure 5.7.9. That is reasonable since the TMA spot image only shows whether there is cancer in the tissue, but the ratio of benign and malignant tissue in that spot may not be representative of the whole prostate. Thus, including information on how much cancer is in the prostate improves predictive performance.

When only inputting a single patient feature in addition to the image to eCaReNet, adding the tumor volume results in the best performance on the validation set (AUC 0.81, C-index 0.80, Brier 0.099), adding only the tumor diameter leads to the best results on the test set (AUC 0.79, C-index 0.77, Brier 0.108).

When inputting multiple features along with the histopathology image, the best combination is using the PSA value, the tumor volume, and the tumor diameter. The discrimination equals adding only volume in the validation set and only diameter in the test set. However, the Brier score is slightly lower, and the standard deviation decreases.

The results for using the image only and the best feature combination are contrasted quantitatively in Table 5.7.5. The best model reaches a mean AUC of 0.81 on the validation and 0.79 on the test set. The best C-indices are 0.80 and 0.77, and the Brier scores are 0.097 and 0.106 for the validation and test sets, respectively. Adding the age in addition to the PSA value, tumor volume, and diameter does not affect the average model performance but increases the standard deviation. As can be seen, the model that includes the PSA value, tumor volume, and tumor diameter as additional features outperforms the pathologist's ISUP annotation per patient. The pathologist achieves AUCs of 0.79 and 0.77 and C-indices of 0.77 and 0.76 on the validation and test sets, respectively.

Conclusion Adding additional patient information increases eCaReNet's predictive performance significantly. That was shown already when only using a single additional patient feature. When jointly adding the PSA value, tumor volume, and tumor diameter, eCaReNet outperforms the pathologist on both the validation and the test set. That emphasizes that a TMA spot can only give a first impression of the disease status. In particular, information about how much of the prostate tissue is cancerous supports an accurate prediction. It is expected that predictive performance could be increased further if more patient features were available and included.



(a) Results on the validation sets.

(b) Results on the test sets.

Figure 5.7.9: Model performance when inputting patient features in addition to the TMA spot image. The boxplots show the results over five runs with different initialization seeds. A black line indicates the mean value. The more patient features are added, the better the results. Only age is decreasing performance. a: age, p: PSA value, d: tumor diameter, v: tumor volume.

Table 5.7.5: Model performances when using only an image or adding PSA value, tumor diameter (dm), and tumor volume (vol) to eCaReNet. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D-cal.), the chi-square tests pass (p). The best results are marked in bold.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
ISUP	0.79	0.77	-	
image only	0.77 (0.0069)	0.77 (0.0026)	0.105 (0.0015)	**p**
image+psa+dm+vol	**0.81** (0.0040)	**0.80** (0.0025)	**0.097** (0.0008)	**p**
Test set				
ISUP	0.77	0.76	-	
image only	0.75 (0.0051)	0.74 (0.0028)	0.112 (0.0012)	**p**
image+psa+dm+vol	**0.79** (0.0033)	**0.77** (0.0038)	**0.106** (0.0007)	**p**

5.7.9 Evaluation on different datasets

Next, eCaReNet is evaluated on different datasets to evaluate the generalizability further. For these experiments, a single image is used as input, no additional patient information is included, and eCaReNet is trained on the complete Surv1. The following evaluation results are again averaged over five runs with different initialization seeds on eCaReNet, and a summary of all metrics for all Survival datasets is provided in Table 5.7.6. The following paragraphs give detailed analyses of the results.

Surv2 Since the results on Surv2$_{\cap \text{Surv1}}$ were already described in the previous section 5.7.7 *Multiple images per patient*, they are not elaborated here again but listed in Table 5.7.6 for completeness.

SurvDiff The influence of thin, thick, and differently stained tissue is evaluated next. The datasets SurvThin, SurvThick, and SurvLongStain comprise a subset of the same tissue cores as Surv2. The TMA spots differ slightly since the spots are sliced subsequently from the tissue cores. Since it was shown in section 5.7.8 *Multimodal analysis* that eCaReNet's performance drops on the second core per patient, the following results are compared to the performance on the regularly sliced, stained, and scanned dataset Surv2. Thus, changes in performance can more clearly be related to the slicing or staining instead of the tissue core selection.

Since SurvThin, SurvThick, and SurvLongStain only encompass a subset of all patients, only those patients with images in Surv2 and in all three different-acquisition datasets (SurvDiff: SurvThin, SurvThick, and SurvLongStain) are used for the comparisons. Those datasets are therefore named Surv2$_{\cap \text{SurvDiff}}$,

Table 5.7.6: Results of eCaReNet on different datasets. eCaReNet is trained on Surv1. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D), only failure (f) or pass (p) is indicated. T indicates the trend of whether the performance on the dataset remains equal or decreases.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D**	**T**
Surv1	0.78 (0.0041)	0.75 (0.0016)	0.107 (0.0004)	p	
Surv1$_{\cap\text{Surv2}}$	0.77 (0.0039)	0.75 (0.0013)	0.104 (0.0005)	p	
Surv2$_{\cap\text{Surv1}}$	0.74 (0.0069)	0.72 (0.0055)	0.134 (0.0287)	f	↘
Surv2$_{\cap\text{SurvDiff}}$	0.77 (0.0139)	0.74 (0.0085)	0.133 (0.0271)	p	
SurvLongStain$_{\cap\text{SurvDiff}}$	0.76 (0.0108)	0.71 (0.0053)	0.134 (0.0245)	p	→
SurvThick$_{\cap\text{SurvDiff}}$	0.70 (0.0063)	0.66 (0.0045)	0.130 (0.0095)	p	↘
SurvThin$_{\cap\text{SurvDiff}}$	0.50 (0.0221)	0.53 (0.0185)	0.157 (0.0244)	p	↓
Surv1$_{\cap\text{SurvScan}}$	0.78 (0.0034)	0.75 (0.0017)	0.108 (0.0003)	p	
SurvScan$_{\cap\text{Surv1}}$	0.53 (0.0137)	0.54 (0.0103)	0.146 (0.0042)	f	↓
Test set					
Surv1	0.77 (0.0048)	0.73 (0.0048)	0.113 (0.0008)	p	
Surv1$_{\cap\text{Surv2}}$	0.76 (0.0050)	0.73 (0.0038)	0.114 (0.0006)	p	
Surv2$_{\cap\text{Surv1}}$	0.72 (0.0077)	0.70 (0.0029)	0.135 (0.0205)	p	↘
Surv2$_{\cap\text{SurvDiff}}$	0.71 (0.0116)	0.68 (0.0105)	0.140 (0.0163)	p	
SurvLongStain$_{\cap\text{SurvDiff}}$	0.65 (0.0120)	0.65 (0.0109)	0.146 (0.0176)	p	↘
SurvThick$_{\cap\text{SurvDiff}}$	0.67 (0.0096)	0.66 (0.0097)	0.134 (0.0020)	p	↘
SurvThin$_{\cap\text{SurvDiff}}$	0.61 (0.0315)	0.61 (0.0210)	0.144 (0.0076)	p	↓
Surv1$_{\cap\text{SurvScan}}$	0.78 (0.0048)	0.74 (0.0042)	0.104 (0.0006)	p	
SurvScan$_{\cap\text{Surv1}}$	0.53 (0.0121)	0.53 (0.0087)	0.139 (0.0038)	f	↓
SurvPCBN	0.53 (0.0054)	0.56 (0.0138)	0.104 (0.0007)	p	↓

SurvThin$_{\cap\text{SurvDiff}}$, SurvThick$_{\cap\text{SurvDiff}}$, and SurvLongStain$_{\cap\text{SurvDiff}}$, and comprise 302 patients per test set.

On the Surv2$_{\cap\text{SurvDiff}}$ test set, eCaReNet reaches an AUC of 0.71, a C-index of 0.68, and a Brier score of 0.140. These performances are below the previously reported results on Surv2$_{\cap\text{Surv1}}$. On the validation set, however, the AUC increases to 0.77, the C-index to 0.74, and the Brier score remains similar (0.133). That stresses how sensitive all scores are to the dataset composition and size.

On the SurvLongStain$_{\cap\text{SurvDiff}}$ validation set, eCaReNet achieves an AUC of 0.76, therefore similar to Surv2$_{\cap\text{SurvDiff}}$. However, the AUC decreases to 0.65 on the test set. The C-index also decreases, to 0.71 on the validation and 0.65 on the test set. The Brier score increases slightly to 0.134 on the validation and 0.146 on the test set. The d-calibration test passes for the validation and test sets. These results indicate that the longer staining time degrades model performance slightly.

The performance on $\mathrm{SurvThick}_{\cap\mathrm{SurvDiff}}$ decreases in comparison to $\mathrm{Surv2}_{\cap\mathrm{SurvDiff}}$. The AUC on the validation set is 0.70, the C-index decreases to 0.66, and the Brier score decreases to 0.130. On the test set, discrimination is lower than on $\mathrm{Surv2}_{\cap\mathrm{SurvDiff}}$, while the Brier score remains similar (AUC 0.67, C-index 0.66, Brier 0.134). On validation and test sets, the results are well-calibrated. It is concluded that the tissue thickness has a slightly negative influence on eCaReNet's performance.

The results on $\mathrm{SurvThin}_{\cap\mathrm{SurvDiff}}$ are more evident since a more significant decrease in AUC in both the validation and the test set is visible. The AUC reaches 0.61 on the test set and 0.50 on the validation set, which is equally good as randomly ordering the patients and below the test set performance. The C-index decreases to 0.61 on the test and 0.53 on the validation set. The Brier score increases to 0.157 on the validation and 0.144 on the test set. These results indicate that the image coloring or contrast influences the results much since the images of tissue with 1 µm thickness are pale compared to Surv1. Moreover, the standard deviation of the discrimination is higher on this dataset, indicating unstable performance.

SurvScan To compare the performances on SurvScan and Surv1, subsets with overlapping patients are created ($\mathrm{Surv1}_{\cap\mathrm{SurvScan}}$ and $\mathrm{SurvScan}_{\cap\mathrm{Surv1}}$). These subsets each comprise 1,737 patients in the test set (1,895 in the validation set) and consist of the same TMA spots in both sets, which are only scanned with different scanners. Every change in performance can therefore be related to a bias created by the scanner.

The results on the subset $\mathrm{Surv1}_{\cap\mathrm{SurvScan}}$ are still well calibrated according to the d-calibration chi-square test. The AUC is 0.78 on the validation and the test set. The C-index is 0.75 on the validation and 0.74 on the test set, hence also similar to the above-reported performance on the complete Surv1 validation and test sets. The Brier score is 0.108 on the validation and 0.104 on the test set.

The performance on $\mathrm{SurvScan}_{\cap\mathrm{Surv1}}$ is much lower than on $\mathrm{Surv1}_{\cap\mathrm{SurvScan}}$ for the validation and test sets. The AUC drops to 0.53 on the validation and the test set, and the C-index drops to 0.54 on the validation and 0.53 on the test set. The Brier score increases to 0.146 on the validation and 0.139 on the test set. On this dataset, the d-calibration test fails, indicating a non-calibrated model. Since $\mathrm{SurvScan}_{\cap\mathrm{Surv1}}$ comprises the exact same TMA spots as $\mathrm{Surv1}_{\cap\mathrm{SurvScan}}$, those results reveal the model's sensitivity to the dataset source and hint at a significant influence of color bias on the performance.

SurvPCBN The performance on the external dataset SurvPCBN is evaluated next. Multiple images per patient are available, and the predictions per image are averaged to obtain a survival curve per patient since that yielded the best results before. The Brier score indicates a good performance at 0.104. Furthermore, the d-calibration chi-square test shows good calibration. However, the model only reaches an AUC of 0.53 on SurvPCBN and a C-index of 0.56. The drop indicates an insufficient discrimination power, which might result from the differences in staining and could also be due to differences in patient survival time distributions.

Conclusion It was shown that the model is sensitive to data acquisition. When tissue is cut thinner or scanned differently, eCaReNet's performance decreases significantly, while longer staining time and thicker tissue decrease model performance only slightly. The model also fails to achieve good performance on the external SurvPCBN. It can be concluded that eCaReNet is not robust to dataset biases. In the presented cases, differences are particularly visible as color biases.

5.8 Discussion

This chapter presented eCaReNet as a survival prediction model that reaches discriminative performance on par with a pathologist, using a single TMA spot image as input. Compared to state-of-the-art models, eCaReNet provides more clinically reasonable and well-calibrated survival curves. Further, the results allow individualized predictions since patient survival curves cover 7 years, and up to eight distinct risk groups are distinguishable. This detailed stratification is an advantage over many proposed methods in the literature, which often only distinguish two to three groups (Nam et al., 2022; Lombardo et al., 2021; Pinckaers et al., 2022).

By including an MIL layer, explainability is included in eCaReNet. A quantitative analysis showed that malignant image patches receive higher attention weights than benign patches. Thus, eCaReNet is expected to support pathologists by indicating the most relevant, probably malignant, image regions that should be focused on when analyzing the tissue.

It was further evaluated how model performance can be improved to reach a higher discrimination power. Including a second patient image from Surv2 in addition to the image from Surv1 during evaluation increases the model performance. That shows that a single TMA spot may not represent the patient's whole prostate. Thus, it is suggested to include more than a single TMA spot for clinical decision support. Also, it should be investigated further whether training eCaReNet on multiple images per patient improves model performance.

Providing the model with additional patient information increases the performance, reaching higher discrimination than with the ISUP annotations. That confirms that a single image spot cannot encompass all information about a patient's disease status but that more information is needed for a thorough picture of the patient. In particular, how much of the prostate tissue is cancerous, included in tumor volume and diameter, is relevant to increase performance. If more patient information was available, for example, about the family cancer history, a further increase in predictive performance is expected.

For the Survival dataset, the image quality was controlled manually to remove images with artifacts. It needs to be investigated if automated detection of artifacts like blurry spots, tissue folds, or missing tissue can improve overall performance further. Shakhawat et al. (2020), for example, propose a quality estimation that first detects artifacts and then classifies these into artifacts that could be removed by rescanning (e.g., out-of-focus) and artifacts that are due to slide preparation (e.g., tissue folds). However, some differences in the datasets are not due to artifacts but result from differences in data acquisition protocols. Extending the evaluation to these datasets that differ in data acquisition reveals the model's limitations. eCaReNet is sensitive to the scanner used for digitization, and the prediction performance also decreases for thin tissue. The datasets on which eCaReNet achieves the lowest performances are SurvPCBN and SurvScan. Since, compared to the training dataset, these show a great difference in color, it is concluded that color bias is the most significant source of error. The low performance on SurvThin appears to indicate a sensitivity to decreased color intensity and contrast. Since the color is a systematic bias introduced by the data acquisition and staining, it is proposed to adjust the color of images that appear different to the training set in chapter 6 *Robustness*. Since a dataset with a consistent dataset acquisition protocol was used for training (Surv1), it would be interesting to evaluate the performance when training on multiple datasets with differing biases.

Furthermore, the distribution of relapse times in SurvPCBN is different from the training dataset of Surv1 since the external dataset contains more patients with late relapse times. Possibly, eCaReNet cannot predict well on patients who remain relapse-free over a long time. That stresses the need for further evaluations of different biases to ensure robustness when applying a survival prediction model in a clinic.

Chapter 6

Robustness

It was shown in section 5.7.9 *Evaluation on different datasets* that eCaReNet's predictive performance is sensitive to dataset biases. Thus, research question R4 is addressed here ("Is it possible to capture the model's limitations in an uncertainty measure and make the model robust toward dataset bias?") by introducing an uncertainty measure and a color transfer method to increase model robustness. The chapter starts with a theoretical introduction before presenting the experiment results.

6.1 Motivation

In order to apply a trained neural network to new data, which may come from a different source than the training data, it is necessary to estimate whether the prediction is reliable. Especially in pathology, the input images might vary, for example, in color, intensity, resolution, and sharpness due to differences in data acquisition, as described in section 2.1.3 *Dataset acquisition.* Since images obtained during clinical routine are likely to have a greater variance in appearance than a dataset used for training a model, building a model that is robust to most biases and identifies uncertain predictions is of uttermost importance. In this thesis, the focus is on color differences since this is the most prevalent difference in the Survival dataset's images. Recognizing corrupt or adversarial images is also important, but not covered in this thesis (for adversarial attacks, see Apostolidis and Papakostas, 2021).

An option to distinguish samples coming from the same distribution as the training data from samples coming from a different distribution than the training data (due to domain bias), is OOD detection. Usually, OOD samples are detected in a latent space representation and neglected during inference. Besides only deciding that an image is OOD and denying a prediction, in this thesis, it is investigated whether the OOD samples can be shifted closer to the training distribution through a color adaptation.

This thesis proposes a novel workflow, shown in Figure 6.1.1, to decide how to proceed with newly presented images, which may differ from the training data in coloring. In the first step during inference, it is decided whether a new image is OOD or ID based on its latent space representation (for details, see section 6.2 *Out-of-distribution detection*). If an image is classified as ID, a survival curve is predicted and made available to the pathologist. If the sample is OOD, the image color is adjusted to match that of the training set with a color transfer method (for details, see section 6.3 *Color transfer*). Afterward, the OOD score of the transformed image is reassessed. If the image is still OOD, it will be deferred to a pathologist without providing a survival prediction. In contrast, survival predictions on ID images can be passed on to the pathologist for decision support. In the following, the methods for OOD detection and color adaptation that are used and developed in this thesis are motivated and introduced. Differences to current state-of-the-art methods are stressed in this context.

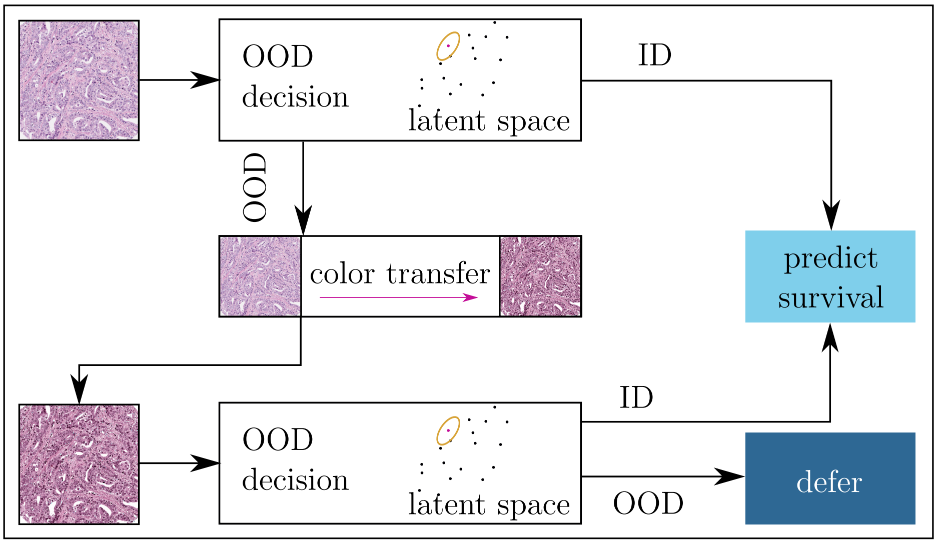


Figure 6.1.1: Proposed workflow to combine OOD detection and color transfer during inference: First, an image is transferred into the latent space to decide whether it is OOD or ID. If it is OOD, the color bias of the training set is transferred to the new image. The transformed image is again tested for OOD-ness. ID images are used for survival prediction, while OOD images are deferred to a pathologist. OOD: out-of-distribution, ID: in-distribution.

6.2 Out-of-distribution detection

Lee et al. (2018) introduce a method for OOD detection using the Mahalanobis distance, which is now commonly applied. First, each sample from the training dataset is transformed into the latent space, i.e., the output after a predefined layer in the neural network. Per annotated class, these latent space representations are used to fit a multivariate class-conditional Gaussian distribution. Therefore, each class is represented by a mean in the latent space, with the covariance tied among all classes. Instead of using a single latent space, the authors propose using the average across multiple neural network layers (the number of layers depends on the network architecture). In order to estimate whether an unseen test image is close to the training distribution, it is also transformed into its latent space representation first. Then, the minimum Mahalanobis distance M from the sample to the distribution of each class (c) is calculated to serve as an OOD score as

$$M(\mathbf{x}) = \min_{c} \left(f(\mathbf{x}) - \hat{\mu}_c\right)^{\mathrm{T}} \hat{\mathbf{\Sigma}}^{-1} (f(\mathbf{x}) - \hat{\mu}_c), \tag{6.2.1}$$

with the latent representation f of test sample input features $\mathbf{x}$. The class-mean $\hat{\mu}_c$, and covariance matrix $\hat{\mathbf{\Sigma}}$ are defined as

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\mathbf{x}_i), \tag{6.2.2}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{c} \sum_{i:y_i=c} (f(\mathbf{x}_i) - \hat{\mu}_c)(f(\mathbf{x}_i) - \hat{\mu}_c)^{\mathrm{T}}, \tag{6.2.3}$$

for N total training samples and N_c samples of class c, each sample with class annotations y_i and features $\mathbf{x}_i$ (Lee et al., 2018). This Mahalanobis distance M is chosen as the final score to estimate the prediction confidence or "OOD-ness": the smaller the distance, the higher the probability that the sample is ID. A threshold to binarize the samples into ID and OOD according to their distances is defined such that 95 % of the training samples are counted as ID (true positive rate of 95 %). An AUROC measures the performance of this method for ID/OOD classification if clearly-defined ID and OOD test sets are available.

Ren et al. (2021) also use a Mahalanobis distance but adapt it to being a relative Mahalanobis distance. They fit a distribution per class for the per-class distance as Lee et al. (2018) and additionally fit a single distribution for the whole training set for the overall distance with

$$\mu = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i), \tag{6.2.4}$$

$$\mathbf{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (f(\mathbf{x}_i) - \mu)(f(\mathbf{x}_i) - \mu)^{\mathrm{T}} \tag{6.2.5}$$

(Ren et al., 2021). The relative Mahalanobis distance is the smallest per-class distance relative to the distance to the overall distribution.

To stress the difference between ID and OOD samples, Lee et al. (2018) add a small perturbation ϵ to each input during training, with the goal of driving ID and OOD samples further apart in the latent space and improving discriminative performance. Therefore, Lee et al. (2018) need ID and OOD samples during model training.

Sun et al. (2022) propose an approach with a similar underlying idea as Lee et al. (2018). They also suggest estimating the uncertainty in the latent space representation with a distance measure. However, they state that the Mahalanobis distance might not always be applicable since it expects the latent space to be Gaussian-distributed, which may not hold true for all datasets. Instead of fitting distributions to the latent space representations of the training dataset, they propose to use the distance from a new sample to its k-th nearest neighbor in the training dataset. Instead of calculating the distances to all training samples, which would be computationally expensive, they only use a subset. Further, they use a Euclidean distance instead of the Mahalanobis distance and normalize the feature space. Again, a threshold is defined to decide which sample is ID or OOD.

Adaptations for this thesis Since no distinct classes are available for survival analysis, the Mahalanobis approach needs to be adapted. In this thesis, the latent space representation of the training data is approximated with only a single multivariate Gaussian, and possible classes are neglected. Hence eq. (6.2.4) and eq. (6.2.5) are used. In their setup, Lee et al. (2018) define images as OOD when these come from classes that are not part of the training set. Instead, for survival prediction, detecting OOD samples that show dataset biases for which the model cannot output confident predictions is necessary. It is difficult to obtain such a dataset that can be clearly labeled as OOD since a dataset bias may or may not lead to model performance degradation. Therefore, the input perturbation part in Lee et al. (2018), which needs OOD samples during training, is disregarded in this thesis.

Further, the k-th nearest neighbor approach of Sun et al. (2022) is combined with the Mahalanobis distance from Lee et al. (2018). That means a new sample is first transformed into its latent space representation. The distances from this representation to all latent space representations of a training data subset are calculated with the Mahalanobis distance. The distance to a sample's k-th nearest neighbor serves as its OOD score. If a sample's distance exceeds a threshold, meaning that the k-th neighbor is too far away, that sample is classified as OOD and a prediction on the image is expected to be unreliable. For this thesis, $k = 0$, thus, the distance to the closest neighbor is chosen as a reference since this led to the best results. As a threshold, a 95 % true positive rate on the training data is chosen.

Evaluation of OOD detection The quality of OOD detection will be estimated with Surv2, SurvDiff, SurvScan, as well as the external SurvPCBN (all introduced in chapter 3 *Datasets*, and denoted from now on as SurvOODCandidates). These datasets do not represent semantic shifts (same tissue type, same disease) but non-semantic shifts due to data acquisition differences. Defining these images as OOD is not straightforward since the model could be robust toward the changes. Therefore, calculating the AUC of the "classification" into ID and OOD is misleading. Instead, the model's predictive performance on all SurvOODCandidates' samples will be evaluated and compared to the performance on only ID samples. It is expected that images with higher OOD scores are more often incorrectly predicted than images with lower OOD scores.

6.3 Color transfer

Color transfer methods transfer test images into the color space of the training images. The goal is to improve a model's prediction by imitating the training dataset bias during inference.

One of the simplest color transfer approaches is histogram matching. For that method, at least one training and one test set image need to be available. In histogram matching, the histogram of a source image is adjusted to the histogram of a target image. An image's histogram counts the number of pixels per color value, thus neglecting the spatial distribution. For color transfer during model inference, a test image's color space is manipulated such that after the color transfer, its histogram matches that of a randomly chosen reference training image. That method is named RandHistMatch in the remainder of this thesis.

For histogram matching, the cumulative distribution functions of the reference and test image are matched. In theory, continuous pixel intensities r and t, for the reference and test image, can be seen as random variables, with the histogram as the probability density function. For matching, the cumulative histograms G for both images are computed. A transformation function between the test image's histogram and the reference histogram needs to be calculated. To this end, both histogram cumulative distribution functions are equated, using the random variable s:

$$\begin{aligned}
&\text{reference :} && G(r) &&= (L-1)\int_0^r p_r(w)dw = s, && (6.3.1)\\
&\text{target :} && G(t_{transf}) &&= (L-1)\int_0^t p_t(w)dw = s, && (6.3.2)\\
&\text{match target :} && G(t_{transf}) &&\overset{!}{=} G(r), && (6.3.3)\\
&\text{inverse mapping :} && t_{transf} &&= G^{-1}[G(r)], && (6.3.4)
\end{aligned}$$

(Gonzalez and Woods, 2002). Here, p_r is the probability density function for r, and L is the number of different intensity values, thus, 256 for an 8-bit image. In the case of $k = L - 1$ discrete pixel values, the equations change to

$$\begin{aligned}
G(t_k) &= (L-1)\sum_{j=0}^{k} p_t(t_j) = s_k && (6.3.5)\\
&= \frac{L-1}{MN}\sum_{j=0}^{k} n_j,\\
t_{transf} &= G^{-1}(s_k) && (6.3.6)
\end{aligned}$$

for an image with $M \times N$ pixels, and n_j pixels with intensity t_j. The computation of G^{-1} is not necessary in the discrete case, but all values for G can be calculated directly and stored in a lookup table. The closest match for s_k in the lookup table is used (Gonzalez and Woods, 2002).

Histogram matching adapts the intensity of the target image. However, histogram matching is not restricted to gray values but can be performed in any color space. For multichannel images, like RGB, the color channels are treated independently (van der Walt et al., 2014).

For H&E-stained images, it is relevant that equivalent structures are colored similarly. Therefore, approaches exist that decompose the RGB channels of a histopathology image into the hematoxylin and eosin channels and adapt those instead of the RGB channels directly. Macenko et al. (2009) introduced such a method. They transform an RGB image into optical density space (OD) and adapt the H and E channels to a randomly chosen reference training image. The optical density space is defined as

$$OD = -\log_{10}(I), \qquad (6.3.7)$$

where I is the image as an RGB color vector, normalized to $[0, 1]$ (Macenko et al., 2009). The matrices of stain vectors V and saturations of each stain S are defined as

$$OD = VS, \tag{6.3.8}$$
$$S = V^{-1}OD. \tag{6.3.9}$$

The stain matrix converts between RGB and H&E colors. In their experiments, they show that removing background which is mostly white improves results. The stain matrix of the reference image is used as the new stain matrix for the test image. Furthermore, the 99^{th} percentiles of both intensity values in H and E of the test image are matched to those of the reference image. Then, the new image is converted back from optical density into RGB color space.

Both approaches RandHistMatch and Macenko adaptation do not require a large known test set, which is an advantage over other methods, e.g., those using a GAN. Further, they are applied on the test set, so no model retraining is required.

Adaptations for this thesis In this thesis, both Macenko normalization and histogram matching are used in consecutive steps to adjust the staining color of a test image to the training dataset. The proposed color transformation method is named ClusterMatch from now on and illustrated in Figure 6.3.1.

Since all training images vary in color and the amount of background on the image, adapting the color of a test image to a random training sample might introduce artifacts and cannot yield reproducible results. Instead, an improved method for the selection of the reference histogram and H&E stain matrices is proposed. For both transfer methods, adapting the test image's color to the most similar sample in the training dataset leads to as little color transformation as possible, therefore also limiting the chances of introducing undesired artifacts. Since comparing the histogram and stain matrix to every training sample is costly, it is proposed to cluster similar training samples first and represent the training dataset with the cluster centers.

The histograms for histogram matching and stain matrices for Macenko adaptation of the training set are k-means clustered independently (Figure 6.3.1 (a)). For Macenko adaptation, each training sample is reduced to its stain matrix, and the 99^{th} percentile of the intensity per H and E channel is calculated. The stain matrix and 99^{th} percentiles are concatenated to one vector and clustered with k-means, using a Euclidean distance as the metric. For histogram matching, the

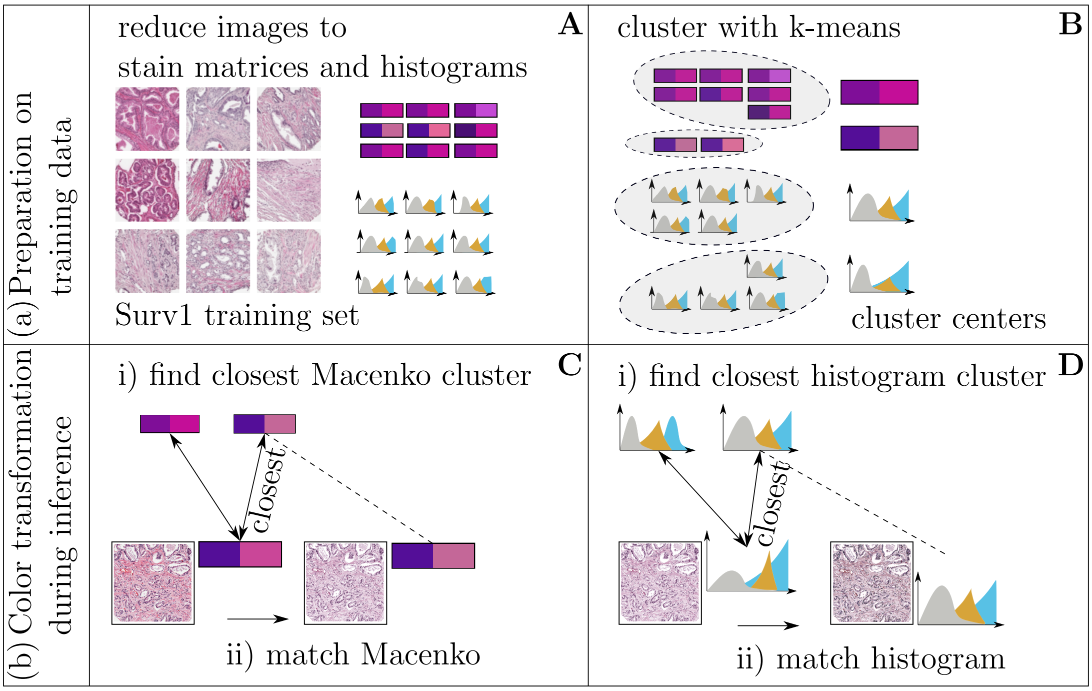

(a) Preparation on training data
reduce images to stain matrices and histograms
A
Surv1 training set
cluster with k-means
B
cluster centers
(b) Color transformation during inference
i) find closest Macenko cluster
C
closest
ii) match Macenko
i) find closest histogram cluster
D
closest
ii) match histogram

Figure 6.3.1: Workflow for ClusterMatch. (a) For each training image, the stain matrix and HSV histogram are calculated (A). These are k-means clustered so the cluster centers represent the training dataset (B). (b) During inference, the image is first adapted to the most similar stain matrix among the cluster centers with the Macenko transformation (C). Then, the test image histogram is adapted to the closest histogram cluster center of the training set (D).

RGB training images are first transformed into HSV color space, as the separation of hue, saturation, and value allows for independent color tuning. Then, the Wasserstein distance per color channel is used to find the distance between two histograms. It measures the amount of work that is needed to transform one distribution into another. For two distributions H_1 and H_2 with cumulative density functions F_1 and F_2, the Wasserstein distance W is defined as

$$W(H_1, H_2) = \int_0^1 |F_1(y) - F_2(y)| dy \tag{6.3.10}$$

(Chan et al., 2007). The distances per color channel are weighted in this thesis to emphasize saturation, which reflects the white background. Therefore, the overall distance d between two images with color channels HSV is calculated as

$$d = W(H_1, H_2) + 2W(S_1, S_2) + W(V_1, V_2). \tag{6.3.11}$$

Weighting the saturation twice as much as the other color channels led to the best results in experiments, compared to equal weighting or emphasizing hue or value.

The optimal number k of clusters needs to be identified for both methods so that the training set is represented well. The elbow method is used to plot the total distance from each training sample to its cluster center against the number of clusters. The optimal number of clusters, k, is defined by the elbow of that plot. During inference, the closest clusters for Macenko adaptation and histogram matching are computed and each image's color is adapted consecutively (see Figure 6.3.1 (b)). Note that the closest histogram cluster is computed after Macenko color adaptation. The color adaptation is expected to shift the test image closer to the training distribution. However, it is possible that an image is still OOD after applying ClusterMatch, which is why the OOD detection is applied again after color transformation.

Evaluation of color transfer Using ClusterMatch in combination with the OOD detection is evaluated again on the SurvOODCandidates with the AUC, C-index, and Brier score. The predictive performance of eCaReNet is expected to increase on those datasets after color transfer and OOD removal.

6.4 Experiments

The experiments for OOD detection and color transfer are evaluated in the following. They are performed on a single best-performing model (see section 5.7.3 - *Conclusion*), not with models including different seeds since single OOD scores need to be assigned to images. Using multiple trained models with different seeds would not lead to a clear ID/OOD discrimination. Using a single model is reasonable since the variation caused by differing initialization seeds has shown to be smaller than the variation in-between datasets in section 5.7 *Experiments*. For completeness, the prediction performances on all SurvOODCandidates and for Surv1 for that model are illustrated in Figure 6.4.1 since, in section 5.7.9 *Evaluation on different datasets*, only the mean over five runs was evaluated. The model is trained on Surv1 and evaluated again on the subsets with overlapping patients to enable performance comparison. Recall that Surv2$_{\cap \text{Surv1}}$ contains images from Surv2, restricted to those patients that also have an image in Surv1.

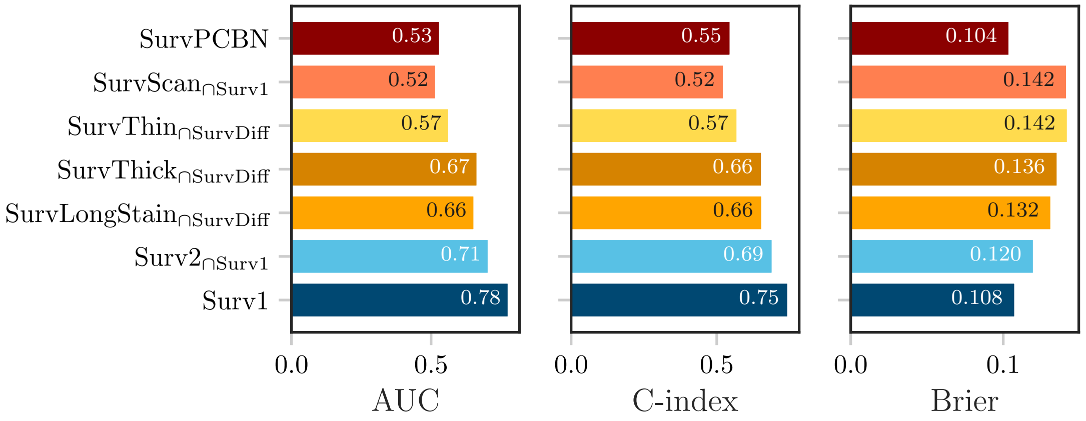


Figure 6.4.1: Results of a single trained eCaReNet model on all test datasets.

6.4.1 OOD scores

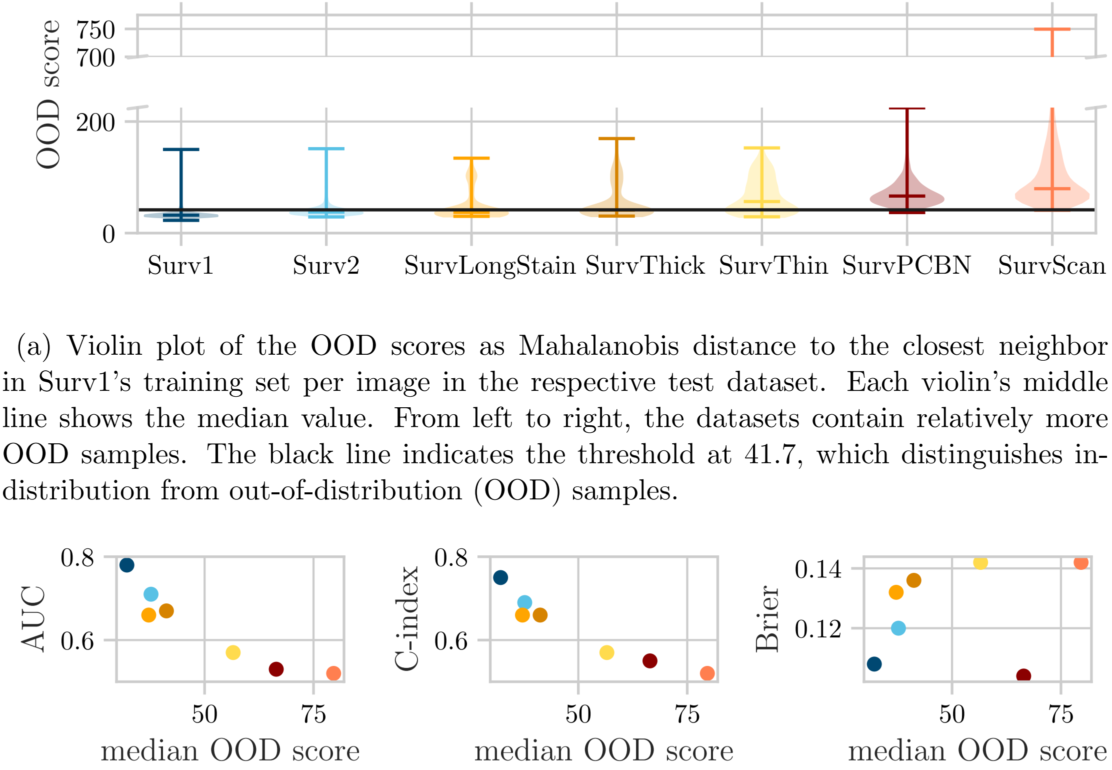
Setup There is no ground truth for the OOD scores of the SurvOODCandidates' images. Therefore, to evaluate whether the calculated OOD scores (Mahalanobis distances) can be used as uncertainty measures, it is evaluated whether those correlate to eCaReNet's predictive performance. It is expected that the datasets on which eCaReNet shows the lowest performance include more OOD samples or have higher OOD scores than those on which eCaReNet shows performances similar to Surv1.

The OOD-ness of each image in SurvOODCandidates and Surv1 is calculated with the nearest-neighbor approach described in section 6.2 *Out-of-distribution detection*. First, eCaReNet transforms all image patches into their latent space representations, using the global average pooling layer as output, and taking the average over all patches for an image. As OOD measure, the Mahalanobis distance to the closest neighbor in the Surv1 training set is calculated in the latent space for each image from the SurvOODCandidates. Instead of calculating the distance to every training sample, which would be computationally expensive, a subset of training samples is selected randomly, as proposed by Sun et al. (2022). Here, 1,000 training samples yield stable results. The Mahalanobis distance to the closest neighbor is then used as a score for an image's OOD-ness. The threshold to count a sample as ID or OOD is a distance of 41.7, which is the threshold at which 95 % of the training samples are classified as ID, as proposed by Lee et al. (2018).

OOD score as uncertainty measure To compare the OOD-ness of all SurvOODCandidates' test sets, their images' OOD scores are depicted as violin plots in

Figure 6.4.2 (a). By definition of the OOD threshold, 95 % of Surv1's images are ID. Surv2 does not have many OOD samples, which fits the above hypothesis that the performance decrease is mostly due to the tissue selection instead of the image staining. The proportions of OOD samples in SurvThin, SurvThick, and SurvLongStain coincide with the measured discrimination and Brier scores: SurvThin contains most OOD samples while having the lowest performance. SurvThick and SurvLongStain have less OOD samples, while having performances closer to Surv2. Nearly all images of SurvScan and SurvPCBN are OOD. That also matches the low AUCs on those datasets. The lowest test set AUC and C-index, as well as the highest Brier score, are achieved on SurvScan, which has the highest OOD scores in the violin plot and fewest ID samples (2 images).

The relation between model performance and OOD scores is further depicted in Figure 6.4.2 (b). A negative correlation between discrimination and OOD scores is shown. The Pearson correlation is $r = -0.93$ and $r = -0.95$ for the AUC and C-index, respectively, with p-values smaller than 0.05. The Brier score



(a) Violin plot of the OOD scores as Mahalanobis distance to the closest neighbor in Surv1's training set per image in the respective test dataset. Each violin's middle line shows the median value. From left to right, the datasets contain relatively more OOD samples. The black line indicates the threshold at 41.7, which distinguishes in-distribution from out-of-distribution (OOD) samples.

(b) Model performance against median dataset OOD scores. The same dataset colors as in (a) are chosen.

Figure 6.4.2: Evaluation of the out-of-distribution (OOD) scores per test dataset.

seems to show a nonlinear correlation, wherefore the Kendall τ is measured. It is $\tau = 0.39$, indicating only a slight correlation, which is probably due to the outlier SurvPCBN. Due to the small sample size, the significance of these results needs to be considered with caution. Based on the ratio of ID and OOD images, Surv2, SurvLongStain, and SurvThick can be called near-OOD datasets, whereas SurvThin, SurvPCBN, and SurvScan are far-OOD datasets.

OOD score threshold If the OOD score correlates with the uncertainty, the performance on ID samples should be higher than the performance on OOD samples. Therefore, eCaReNet's performance should increase when removing OOD samples gradually. That hypothesis is tested by varying the threshold of counting a sample as either OOD or ID and evaluating the model performance only on the samples with a lower OOD score than the threshold. Since the dataset size and included patients change when removing OOD samples, only evaluating the discrimination alone could be misleading. That is because the AUC and C-index compare the ranking of patients and are thus strongly dependent on the chosen cohort. The Brier score evaluates single survival curves and therefore depends less on the patient distribution.

Figure 6.4.3 depicts the Brier scores and AUCs that eCaReNet reaches on the SurvThin test dataset for varying thresholds. The metrics are evaluated only on the ID images. From left to right, the threshold is reduced such that the number of images counted as ID decreases. The Brier score and AUC improve when more and more OOD samples are removed. The number of ID samples reduces to 104, while the Brier improves to 0.113 and the AUC increases to 0.65. Equivalent plots for the remaining test datasets are in the supplementary material in Figure A.3.1. The results are not as clear for every dataset, as sometimes the Brier score or the AUC decreases slightly.

Conclusion It can be concluded that the proposed OOD measure correlates with eCaReNet's predictive performance, particularly, it is negatively correlated with discrimination. The more OOD samples are in a dataset, the lower the predictive performance. Thus, the OOD score can be seen as an uncertainty measure. Only the Brier score of SurvPCBN does not follow this pattern, as the Brier score is low for that dataset, which contains almost only OOD samples.

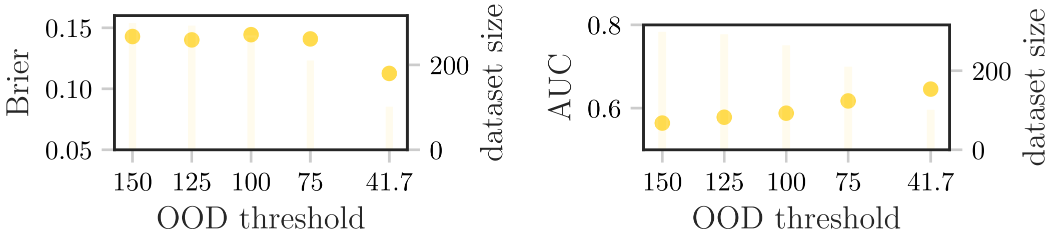


Figure 6.4.3: The model performance increases on SurvThin, when removing more and more OOD samples. The dots show the Brier score in the left, and AUC in the right plot. The bars indicate the test set size, which decreases with removing OOD samples until it reaches 104. The Brier improves to 0.113, the AUC increases to 0.65.

6.4.2 Color transformation

Setup It is now evaluated whether a color transformation improves performance on the SurvOODCandidates. As a baseline, all images are transformed with RandHistMatch, which means an image's histogram is matched to a randomly chosen training image. The results are compared to using ClusterMatch only for the images identified as OOD before. When applying ClusterMatch, the OOD score is re-calculated after color transfer since a transformed image might still be OOD. Only those images that are ID afterward are included for a final prediction and performance evaluation. Images still calculated as OOD need to be deferred to a pathologist in clinical practice.

In order to evaluate the color transfer, the model's performances on the original and the transformed images are compared. One difficulty in the evaluation is that discrimination and calibration are both relevant, but one does not require the other. That means that, e.g., the AUC might decrease with color adaptation while the Brier score improves, or vice versa.

For the histogram matching, the algorithm provided by scikit-image[1] is used (van der Walt et al., 2014). The Macenko color adaptation is performed with an online repository[2]. Scikit-learn is used to find the k-means clusters[3] of stain matrices for the Macenko algorithm (Pedregosa et al., 2011). Since the clustering of the histograms requires the Wasserstein distance as a custom distance function,

[1] `https://scikit-image.org/docs/stable/api/skimage.exposure.html#skimage.exposure.match_histograms` (last accessed November 24, 2022)

[2] `http://github.com/wanghao14/Stain_Normalization` (last accessed November 21, 2022)

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html` (last accessed November 24, 2022)

pyclustering[4] is used (Novikov, 2019). The Wasserstein distance is calculated with scipy[5] (Virtanen et al., 2020).

Baseline: RandHistMatch As a baseline color transformation method, RandHistMatch is applied to all test images in SurvOODCandidates. For each image, a random training image is chosen as a reference for histogram matching in the HSV space. In Figure 6.4.4, eCaReNet's predictive performances are compared before and after applying RandHistMatch to each dataset. The same subsets as in Figure 6.4.1 are used to enable performance comparison.

For the near-OOD datasets Surv2, SurvLongStain, and SurvThick, eCaReNet's performance decreases after color transformation in AUC, C-index, and Brier score. The AUC on Surv2$_{\cap \text{Surv1}}$ and SurvThick$_{\cap \text{SurvDiff}}$ both decrease by 0.07 to 0.64 and 0.60, respectively. Their C-indices decrease to 0.63 and 0.58, and the Brier scores increase to 0.126 and 0.143, respectively. For SurvLongStain$_{\cap \text{SurvDiff}}$, the AUC decreases to 0.62 and the C-index to 0.61. The Brier score increases to 0.135. These results confirm that the near-OOD datasets are already very close to the original training dataset and cannot benefit from color adaptation.

For the far-OOD datasets SurvThin, SurvScan, and SurvPCBN, the performance improves in AUC, Brier score, and C-index. After applying RandHistMatch, the discrimination metrics are close to those on the near-OOD datasets. On SurvThin$_{\cap \text{SurvDiff}}$, the C-index increases by 0.05 to 0.62, and the AUC by 0.04 to 0.61. The C-index on SurvScan$_{\cap \text{Surv1}}$ reaches 0.60 and the AUC 0.61. On SurvPCBN, the C-index is 0.61 after applying RandHistMatch with an AUC of 0.58. Also the Brier scores improve to 0.135 on SurvThin$_{\cap \text{SurvDiff}}$, to 0.127 on SurvScan and 0.087 on SurvPCBN. Thus, increased performance is shown on all datasets that include more OOD than ID images.

ClusterMatch For ClusterMatch, both the Surv1 training set's stain matrices and histograms are clustered separately with k-means.

Each training sample is converted to HSV color space, and the histogram per color channel is calculated. Those histograms are clustered with a k-means clustering algorithm using the Wasserstein distance. For best results, the saturation is weighted twice as much as the hue and value. The optimal number of clusters in the training set is determined as 20 with an elbow plot. When assigning each

[4] `https://pyclustering.github.io/docs/0.8.2/html/da/d22/classpyclustering_1_1cluster_1_1kmeans_1_1kmeans.html` (last accessed November 24, 2022)

[5] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wasserstein_distance.html` (last accessed November 24, 2022)

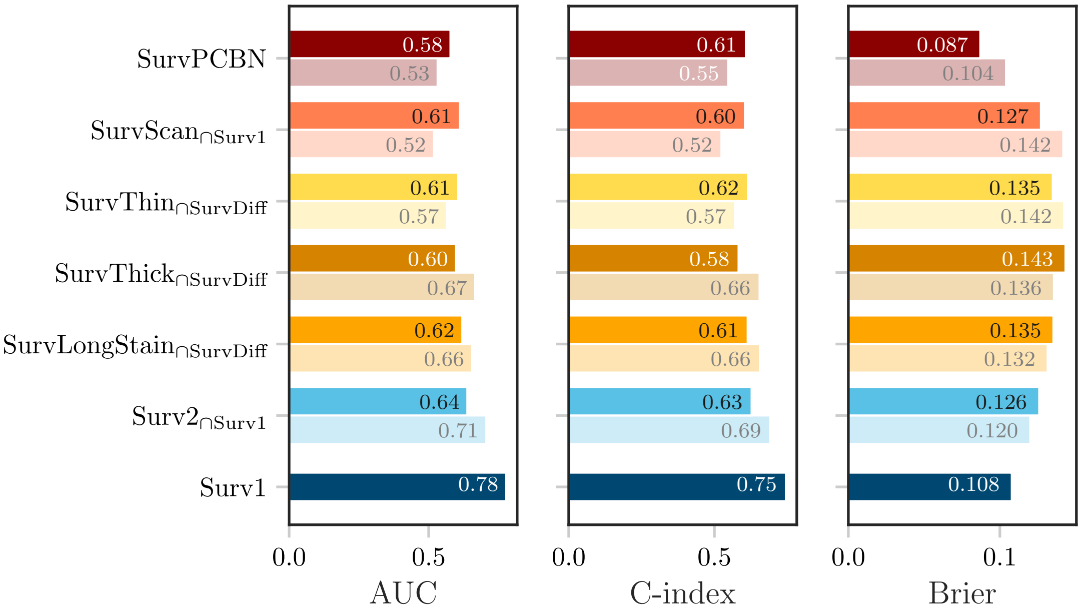


Figure 6.4.4: Performance of eCaReNet when applying RandHistMatch to all test datasets is shown in dark color and the results before applying RandHistMatch in light color.

training sample to a cluster, the smallest cluster contains 143 samples, and the largest one contains 880. Thus, each cluster contains 1.4 to 8.8 % of the training data, and each cluster's histogram represents a reasonable proportion of the training set.

For the Macenko color matching, clusters of H&E stain matrices combined with the intensity values are found. Each training sample is converted to a vector with 8 elements: the first three elements correspond to the RGB encoding of hematoxylin, the following three elements correspond to the RGB encoding of eosin, and the last two values are the image's lowest and highest intensity values. For k-means clustering, the Euclidean distance is used. Again, the optimal number of clusters is determined as 20 with an elbow plot. When assigning a training sample per cluster, the clusters include 24 to 819 samples. Since the smallest cluster only includes 0.24 % of the training set, that cluster is not representative of the training set. Therefore, only the remaining 19 clusters are used in the following experiments.

Examples of ClusterMatch image transformations are provided in Figure 6.4.5. There, (a) shows the original image, (b) shows the image after Macenko adaptation, and (c) is the final image after additional histogram adaptation. The upper

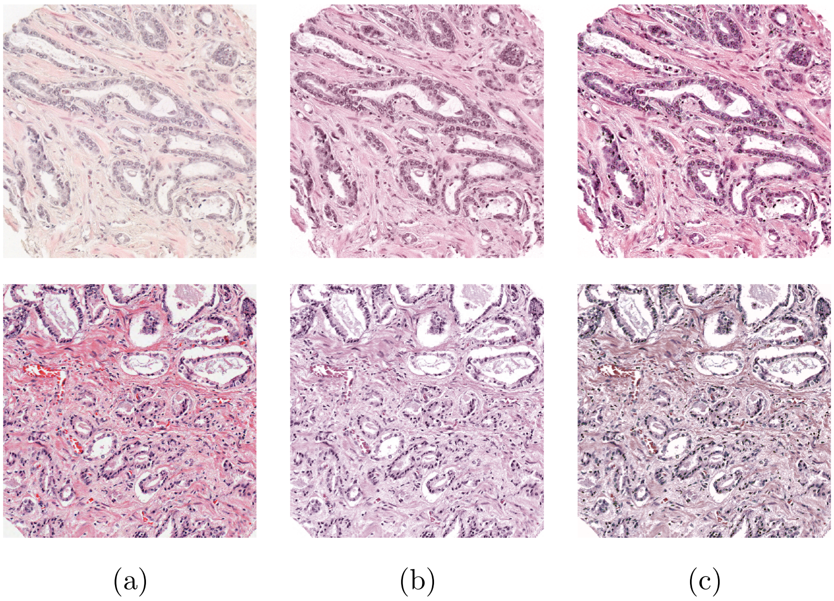


Figure 6.4.5: Examples of ClusterMatch. (a) original image, (b) after Macenko adaptation, (c) after histogram matching. Upper row: SurvScan, lower row: SurvPCBN.

row shows the color transformation of an example image from SurvScan. It is transformed to be more intense and pink. The lower row shows an example image from SurvPCBN. The red blood cells in the image are colored less red after color transformation.

The results of the experiments are illustrated in Figure 6.4.6 and also provided in the supplements in Table A.3.1 and Table A.3.2. They are detailed in the following.

Surv2 26 % of Surv2$_{\cap \text{Surv1}}$ are counted as OOD (492 images), which is why their color is transformed. The results on Surv2$_{\cap \text{Surv1}}$ are not improving with ClusterMatch, the AUC remains at 0.71 while the C-index is slightly increasing from 0.69 to 0.70 (Surv2$_{CM}$ in Figure 6.4.6 (a)). The Brier score also improves slightly from 0.120 to 0.119, which is within the model variance. Compared to RandHistMatch (Surv2$_{RHM}$), the decrease in performance is low. It is concluded that Surv2 is neither benefiting from nor impaired by applying ClusterMatch. That supports the hypothesis that the main reason for a performance decrease compared to Surv1$_{\cap \text{Surv2}}$ is not the color but the tissue content. Removing images that are still OOD after color transformation (393 images) does not improve the metrics further (see Surv2$_{ID}$). This appears to indicate a dataset bias in the remaining images, which is not captured by the OOD detection.

SurvLongStain SurvLongStain$_{\cap \text{SurvDiff}}$ with longer H&E staining time has only 72 OOD samples (24 %). The performance was already close to Surv2$_{\cap \text{SurvDiff}}$ in terms of AUC, C-index, and Brier score without color transformation. As shown in Figure 6.4.6 (b), the AUC slightly reduces from 0.66 to 0.64 with ClusterMatch and the C-index to 0.64, while the Brier score remains constant at 0.132 (SurvLongStain$_{CM}$). Still, the performance is better than when applying RandHistMatch (SurvLongStain$_{RHM}$). Removal of the remaining 65 OOD samples improves the Brier score to 0.126. The AUC increases to 0.67, which is still slightly below the performance before applying ClusterMatch. The same holds for the C-index, which only increases to 0.66 after removing OOD samples (SurvLongStain$_{ID}$).

SurvThick SurvThick$_{\cap \text{SurvDiff}}$ has 125 OOD samples (41 %), which are adjusted for color. The performance without color transformation is similar to the performance on Surv2$_{\cap \text{SurvDiff}}$. ClusterMatch improves the Brier score from 0.136 to 0.133 (SurvThick$_{CM}$), while even better results can be obtained by removing OOD samples (0.129, SurvThick$_{ID}$) (Figure 6.4.6 (c)). Applying ClusterMatch does not influence the AUC, which is still 0.67. It increases slightly to 0.68 when removing all remaining 91 OOD samples. The C-index only improves to 0.68 by removing OOD samples. It is concluded that the images in SurvThick neither benefit nor suffer from ClusterMatch application, while RandHistMatch clearly reduces performance in all three metrics (SurvThick$_{RHM}$).

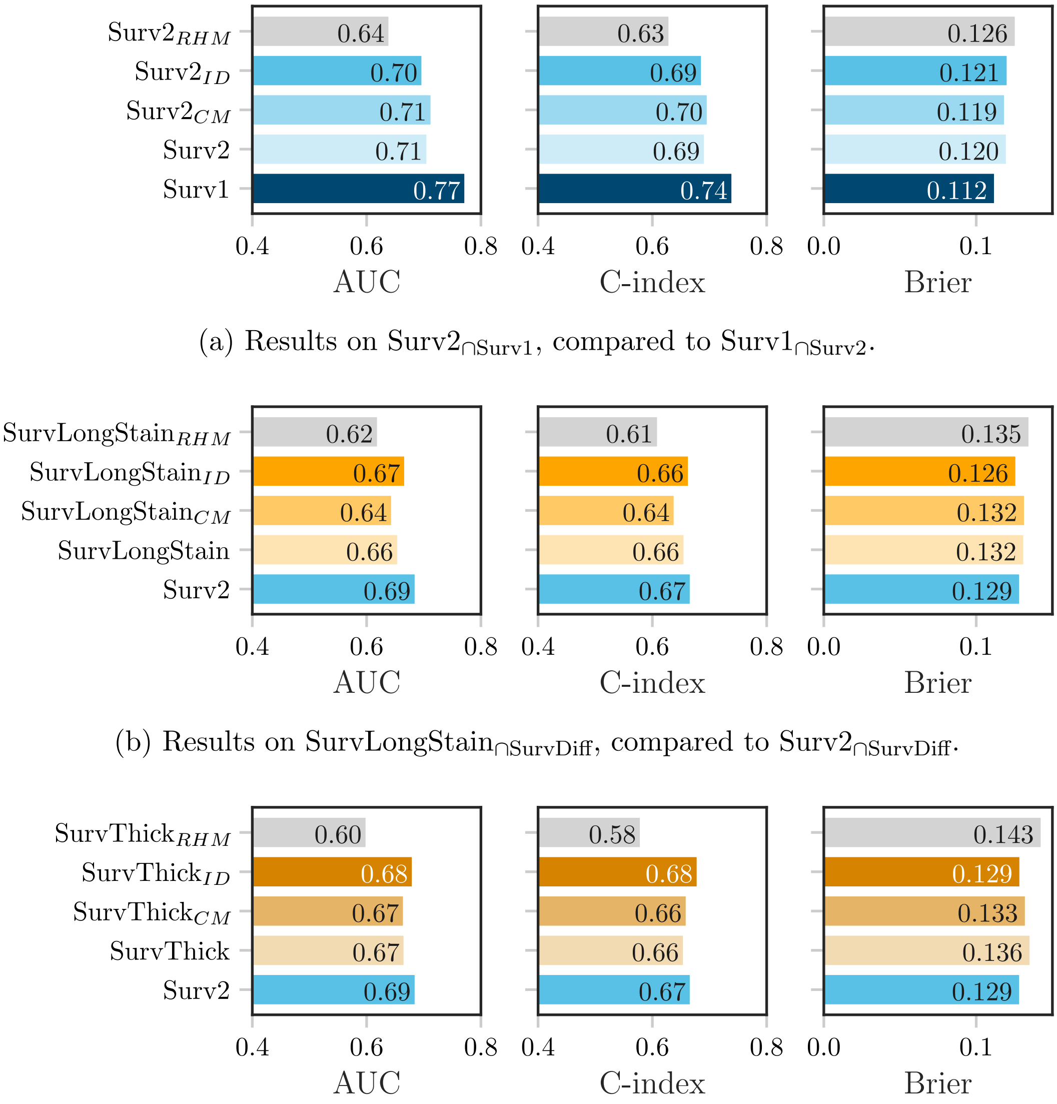
SurvThin SurvThin$_{\cap \text{SurvDiff}}$ contains 66 % OOD images (198 images). As shown in Figure 6.4.6 (d), ClusterMatch improves the AUC from 0.57 to 0.61 and the C-index from 0.57 to 0.62 (SurvThin$_{CM}$). The Brier score reduces from 0.142 to 0.135, providing evidence that both calibration and discrimination profit from ClusterMatch. Interestingly, the performance after applying ClusterMatch is equal to the performance when applying RandHistMatch (SurvThin$_{RHM}$). After color transformation, most images remain OOD (190). Removing these OOD samples and evaluating the metrics only on the ID images increases the AUC further to 0.68 and the C-index to 0.67 (SurvThin$_{ID}$). The final Brier score is 0.113, which is below the performance on Surv2$_{\cap \text{SurvDiff}}$. Since all metrics now show similar values compared to the performance on Surv2$_{\cap \text{SurvDiff}}$, it is concluded that the SurvThin dataset contains a bias besides color, which is not compensated by ClusterMatch but revealed by the OOD detection and which decreases eCaReNet's performance.

SurvScan The positive effect of color transformation with ClusterMatch also shows in SurvScan$_{\cap\text{Surv1}}$, which comprises 99.9 % OOD images. Instead of reaching an AUC of 0.52, the performance increases to 0.71 with ClusterMatch, and the C-index improves to 0.69 (SurvScan$_{CM}$, Figure 6.4.6 (e)). The Brier score improves from 0.141 to 0.114. These results are better than after applying RandHistMatch (SurvScan$_{RHM}$). After color adaptation, 74 % of the images are still OOD. When only keeping the ID samples after matching, the AUC slightly reduces to 0.70, but the Brier score further improves to 0.105, which is close to the results on Surv1$_{\cap\text{SurvScan}}$. Also, the C-index improves to 0.71 on the ID images (SurvScan$_{ID}$).

SurvPCBN Similar results are achieved on the external dataset SurvPCBN. As shown in Figure 6.4.6 (f), ClusterMatch increases performances from 0.53 to 0.61 in AUC, from 0.54 to 0.64 in C-index. This is a larger improvement than with RandHistMatch. The Brier score improves from 0.104 to 0.097. Removing OOD samples increases the overall performance to an AUC of 0.69 and a C-index of 0.71. The Brier score improves to 0.090, which is better than on Surv1. It is important to note that SurvPCBN contains multiple images per patient, hence removing an image does not equal removing a patient. Removing OOD samples excludes 51 % of the images but keeps 79 % of the patients for evaluation. A reason for the AUC being lower than on Surv1 while the Brier score is better might be the difference in relapse time distributions, as shown in Figure 3.2.1 (a). Also, it needs to be analyzed whether a dataset bias besides the color is present in SurvPCBN.

Conclusion The proposed ClusterMatch improves the performance of eCaReNet, particularly for far-OOD datasets. If the datasets include many ID samples and only a few near-OOD samples, there is little to no performance gain by color transformation. On SurvLongStain, the discrimination decreases slightly when adapting the color. It was further shown that ClusterMatch outperforms RandHistMatch in most cases.

Removing OOD samples after applying ClusterMatch further improves results on far OOD dataset, so eCaReNet's performance is shifted closer to that on the Surv1 test set. Since many samples are still OOD after color transformation, it needs to be investigated whether there is a different dataset bias that needs to be addressed besides the color. The images in SurvPCBN are of size 1024×1024 pixels after cutting the centerpiece, whereas the training images in Surv1 are downsized from 2048 × 2048 pixels to 1024 × 1024 pixels. Therefore, possible differences in the level of detail need to be evaluated in further experiments.

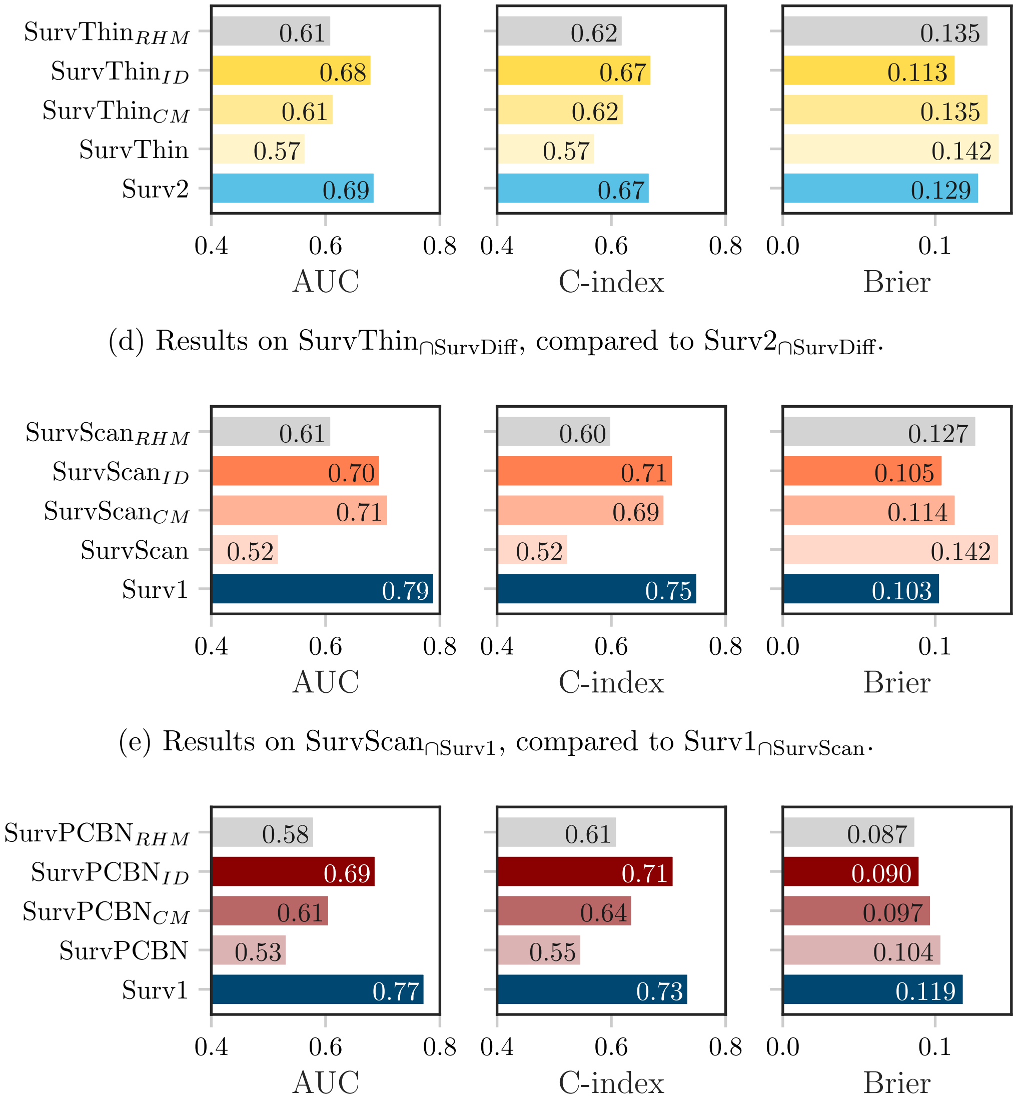


(a) Results on Surv2$_{\cap \text{Surv1}}$, compared to Surv1$_{\cap \text{Surv2}}$.

(b) Results on SurvLongStain$_{\cap \text{SurvDiff}}$, compared to Surv2$_{\cap \text{SurvDiff}}$.

(c) Results on SurvThick$_{\cap \text{SurvDiff}}$, compared to Surv2$_{\cap \text{SurvDiff}}$.

Figure 6.4.6: For figure caption, see next page.



(d) Results on SurvThin$_{\cap\mathrm{SurvDiff}}$, compared to Surv2$_{\cap\mathrm{SurvDiff}}$.

(e) Results on SurvScan$_{\cap\mathrm{Surv1}}$, compared to Surv1$_{\cap\mathrm{SurvScan}}$.

(f) Results on SurvPCBN, compared to Surv1.

Figure 6.4.6: Test set performances of the single best performing eCaReNet model when inputting the original images, applying ClusterMatch, and additionally removing OOD samples afterward. For comparison, the performances when using RandHistMatch are shown. For better readability, subscripts like $_{\cap\mathrm{Surv1}}$ are omitted in the y-axis labels. CM: applying ClusterMatch, ID: keeping only ID samples after color adaptation, RHM: applying RandHistMatch.

6.5 Discussion

The experiments show that the proposed OOD measure can be used as an uncertainty score since eCaReNet performs best on datasets that contain more ID than OOD images. The performance decreases with the number of OOD images in terms of discrimination. The OOD detection might be further improved when instead of using the distance to the nearest neighbor, an average distance to k nearest neighbors is used. Also, modeling the training distribution with a multimodal Gaussian could be considered.

Differences between the images from the training dataset Surv1 and the SurvOODCandidates are visible, particularly as color biases. Therefore, a color transformation approach, ClusterMatch, is proposed. It is shown that ClusterMatch improves performances on all datasets that contain many OOD samples. In contrast to the state-of-the-art RandHistMatch, ClusterMatch rarely decreases the predictive performance. Only on SurvLongStain, on which the performance is already close to Surv2, the performance decreases slightly. It needs to be further investigated whether other color transformation methods, such as applying a GAN, lead to similar or improved results. The combination of OOD detection and ClusterMatch increases the predictive performance on all datasets. Since eCaReNet does not always reach similar performances on the SurvOODCandidates after removing OOD samples as on Surv1, it is concluded that the proposed OOD detection method cannot detect all images that eCaReNet makes imprecise predictions on.

In summary, the proposed framework is beneficial for digital histopathology in clinical applications to ensure that a model only makes predictions when it has sufficient confidence and certainty. The method is independent of the model architecture and thus further applicable to other endpoints besides survival prediction. It is expected to be transferable to other image types, such as radiology. Single images of different color biases can be processed with the proposed approach since no model fitting to an OOD dataset is needed. However, since the performance did not always reach the reference performance of Surv1, it is concluded that color transfer alone cannot cover all variations in the SurvOODCandidates' images. In practice, additional variations, like blurred images or artifacts on images, are expected to occur and need to be addressed.

Chapter 7

Conclusion and discussion

A quick review of the whole thesis is presented here, along with the answers to the research questions posed in the beginning. Afterward, a critical analysis of the thesis' limitations is presented and future research directions are proposed.

7.1 Summary

This thesis contributes to the state of the art by proposing a comprehensive systematic method to approach deep learning-based relapse-free survival prediction for prostate cancer patients. First, literature research revealed that the approaches for Gleason grade prediction and survival prediction from medical images are heterogeneous. However, none of the existing models is directly applicable to the given problem in this thesis due to differences in the dataset or problem formulation.

In contrast to models for classification of everyday objects, pretrained models on histopathology prostate cancer images or for survival prediction are not available open-source. Hence, an ImageNet-pretrained model was first optimized on the Gleasonaut dataset for Gleason grade classification before moving on to survival prediction. It was shown that this model, $\mathrm{M}_{\mathrm{ISUP}}$, achieves performances similar to the state of the art. In the next step, eCaReNet was introduced as a survival prediction model, building on $\mathrm{M}_{\mathrm{ISUP}}$. An extensive evaluation revealed that attention-based MIL, binary survival prediction, and self-attention are beneficial for survival prediction. eCaReNet outperforms state-of-the-art models and is on par with pathologists.

Further, it was shown that the performance increases when using more data during inference, such as additional images or clinical patient information. Including clinical patient information, eCaReNet could even outperform a pathologist. In the end, a robustness analysis revealed the model's sensitivity to unseen biases, for example, due to color differences. An approach to increase prediction performance was presented as a combination of OOD detection and color transfer with

ClusterMatch. The proposed method incorporates an uncertainty measure and significantly increases eCaReNet's predictive performance on far-OOD datasets.

Since eCaReNet is not limited to prostate cancer histopathology images it is worthwhile analyzing how it can be applied to different survival prediction endpoints using medical image input. Also, the proposed method to increase robustness is expected to be applicable other histopathology datasets, regardless of the endpoint.

7.2 Answers to research questions

The research questions posed in section 1.2 are addressed here along with so far achieved answers.

R1: What is the current state of the art in survival prediction from medical images? The literature overview in section 2.2 *State of the art* showed great heterogeneity in cancer stratification and survival prediction approaches. Survival prediction can be interpreted as a binary prediction, a risk score prediction, or a prediction of survival probability over time. Many models build upon the Cox model, however, this has drawbacks since it does not allow survival curves to cross. Approaches that predict the survival probability in discrete intervals are emerging, but these are still rare. Since no common ground can be built upon, there is a need for a transparent and thorough exploration of survival prediction for prostate cancer histopathology images. Furthermore, robustness and explainability are often not addressed in computational pathology.

R2: To what degree can Gleason patterns be predicted accurately in the given dataset of digitized prostate tissue? ISUP prediction was performed with $\mathrm{M}_{\mathrm{ISUP}}$, an InceptionV3 network pretrained on ImageNet and finetuned on the Gleasonaut dataset (chapter 4 *Gleason grade prediction*). Kappas of 0.83 on the validation set and 0.79 on the test set were reached. That is in line with the results presented in the literature. It is concluded that $\mathrm{M}_{\mathrm{ISUP}}$ can successfully support pathologists in their Gleason grading decisions. Another conclusion is that the Gleasonaut has high quality, which leads to the assumption that also the Survival dataset provided by the UKE is suited for training neural networks.

R3: Can relapse-free survival probability over time after prostatectomy be predicted for individual prostate cancer patients based solely on histopathology images showing part of prostate tissue? For survival prediction, eCaReNet was developed, which predicts relapse-free survival curves over 7 years after prostatectomy (chapter 5 *Survival prediction*). Besides predicting individual survival curves, eCaReNet can stratify patients into up to eight distinct risk groups. Given only single TMA spot images per patient, the model reaches a performance close to a pathologist, with AUCs of 0.78 and 0.77 on the validation and the test set, respectively, while being well calibrated. The predictive performance could be increased by including a second image per patient during inference. The pathologist was even outperformed by eCaReNet when adding additional patient information about the tumor volume, tumor diameter, and PSA value. It remains unclear whether the performance could be increased beyond a pathologist's performance on the given dataset when using only images or if an upper limit has already been reached, for example, due to dataset noise. It is further interesting to explore whether the pathologist's patient stratification can be improved when being supported by eCaReNet.

R4: Is it possible to capture the model's limitations in an uncertainty measure and make the model robust toward dataset bias? Color bias was identified as a crucial factor influencing eCaReNet's performance. That effect was studied on several datasets, showing that the performance decreases for images digitized with a different scanner, containing thinner tissue, or stemming from other clinics. A measure was proposed to decide for single images whether they are OOD. It was shown that this OOD measure correlates to model performance and can thus serve as an uncertainty score for single test images during inference (chapter 6 *Robustness*). Such an uncertainty measure is crucial in clinical practice to avoid misleading predictions. Further, a color transformation was proposed, ClusterMatch, which could successfully increase the performance on datasets including many OOD samples. The combination of color transfer and OOD detection was shown to lead to prediction performances close to those on reference test datasets. Biases besides color, like image sharpness, were not explored. Thus, it needs to be further investigated which other data variations influence model performance and how to make eCaReNet robust against those.

7.3 Future research directions

This thesis is a proof-of-concept that survival prediction from histopathology images alone provides predictions that reach performance close to a pathologist. However, to create added value, it is desired to build a model that outperforms pathologists. Only then can decision support systems be integrated into a clinical workflow and improve human decisions. A remaining question is whether eCaReNet reached an upper limit of performance achievable with Surv1, for example, due to dataset noise, or if eCaReNet can be improved further to outperform a pathologist in terms of discrimination. Also, it needs to be explored whether eCaReNet's robustness to unseen dataset biases can be improved.

Therefore, the following discussion explores possibilities to improve the implemented model and its prediction performance. Since the vast space of options for solving computational pathology problems could not be explored entirely in this thesis, this section elaborates on approaches to continue with the research or steer it in a different direction.

Adaptation of model and training process A significant flaw of computational pathology is the currently limited amount of data. The experiments conducted in the work for this thesis show that pretraining on Gleason scores improves results over pretraining on ImageNet. Current approaches in medical image analysis explore a different kind of pretraining, which is called self-supervised learning.

In self-supervised learning, a neural network is trained on a large dataset without having annotations of the final (classification) task. The idea is to pretrain a network to learn common patterns characteristic of the tissue or classes. One approach for self-supervised learning is contrastive learning (Ciga et al., 2022; Chen et al., 2020b). In contrastive learning, a network is trained with pairs of images, where a pair could be an image and its augmented counterpart or an image patch and a neighboring patch. With a contrastive loss, the model is trained such that two instances of a pair are considered similar, and two instances from different pairs are learned to be dissimilar. A comparable approach is contrastive predictive coding, where pairs or sequences of images are used for training. However, the goal is to input one image and predict the paired image (encoded as a latent space representation) or to input a sequence of images and predict the sequence of following images (van den Oord et al., 2018; Lu et al., 2019). The pretrained encoder that computes latent space representations can be used afterward for training on a supervised task.

For both approaches, the intention is to increase the focus on structures that are characteristic of the images instead of learning local biases. For self-supervised learning in computational pathology, any large database with histopathology images can be used without requiring annotations. Images from different organs can also be mixed to increase the dataset size (Ciga et al., 2022). Self-supervised pretraining is a promising approach to fill the gap between missing large annotated datasets and histopathology-pretrained, openly available models. Self-supervised learning as a pretraining could thus increase the performance of eCaReNet and outperform the current model that is pretrained on the Gleasonaut only.

It is also possible to improve the model architecture itself. eCaReNet builds upon a state-of-the-art CNN, but transformers are evolving in the field of computer vision. Transformers emerged in natural language processing and outperform classic approaches like RNNs (Vaswani et al., 2017). Instead of recurrent layers, transformers include (self-)attention layers, reducing the number of operations and increasing inference speed. A drawback is that transformers need large datasets for training and have more trainable parameters than common CNN architectures. Dosovitskiy et al. (2021), for instance, state that pretraining on a dataset of 14 - 300 million images gives good results. Nevertheless, recent approaches using vision transformers exist also in computational pathology, showing promising results (e.g., Shao et al., 2021; Chen et al., 2021; Ikromjanov et al., 2022; Lv et al., 2022). Thus, it is interesting to analyze whether using a Transformer as a base model improves performance for the given survival prediction task.

It remains an open question how accurately a model trained solely on histopathology images from Surv1 can predict the time to relapse. This thesis showed that including the PSA value, prostate volume, and prostate diameter as model input increases performance metrics. Thus, it is expected that including more features, for instance, about the patient's lifestyle (e.g., whether he smokes) or family history (i.e., whether a relative is affected by prostate cancer), further increases performance. Also, the fusion of EHR and image data needs to be evaluated in more detail. In this thesis, the EHR data is concatenated after the global average pooling layer. Duanmu et al. (2020), for example, show that fusing the information at multiple layers inside the model improves predictive performance. A general limitation is in the evaluation of survival models since it was shown that a single metric is not sufficient for performance evaluation and multiple metrics may be contradictory. Reinke et al. (2021) elaborate that choosing the correct metric is a commonly underappreciated topic in image analysis.

Dataset improvements The extracted TMA spots of Surv1 are not all representative of the patients' disease statuses. Since eliminating the white background at the margins of the TMA spots already increases performance, it is assumed that a model's predictive performance can profit from further data cleaning. A more sophisticated elimination of less informative regions, like a classification of background, stroma, and artifacts, should be integrated (Arvaniti et al., 2018; Li et al., 2021). Such a dataset cleaning would ensure that eCaReNet is only trained on relevant regions of interest and thus can reach improved performances. Manual quality control of the images by a pathologist who annotates whether the TMA spot of the patient seems representative would be an ideal setting. However, this is very time-consuming and could again lead to problems due to a subjective selection. It is also possible to train eCaReNet with multiple images per patient to increase the probability of including representative images.

Further, all evaluations in this thesis are performed on TMA spot images, but a prediction on biopsy images is more relevant in clinical routine for treatment decision support. Thus, digitized biopsy images of patients and their relapse times need to be obtained to evaluate whether the results are transferable to biopsy tissue as expected. Difficulties may arise again due to differences in color bias. Also, biopsies are larger in size and might include more non-cancerous tissue, which could decrease eCaReNet's predictive performance. It needs to be evaluated whether the predictive performance on biopsy images is similar to that on TMA spots. If not, the reasons for the performance decrease need to be explored, and eCaReNet has to be retrained with biopsy images.

Improvement of robustness The proposed approach for color transformation, ClusterMatch, increases performance on datasets with color biases, but the model performance on most datasets remains below that on Surv1. These results suggest that non-color biases in the data still reduce prediction performance, while color biases are efficiently dealt with by ClusterMatch. Therefore, an improvement of model robustness to dataset biases is needed.

An option to improve generalizability is to use all SurvOODCandidates during training. By extending the training dataset, eCaReNet could adjust its weights to these datasets' biases. Besides only adding the images to the training dataset, the survival prediction model can be adapted to focus mainly on the tissue content instead of the tissue color. Using images from different domains to deliberately train a model that does not concentrate on the staining but only on the structural appearance of histopathology images is possible, as shown by Ren et al. (2019a) and Marini et al. (2021a). When using paired data from Surv1 and SurvScan, hence, one TMA spot scanned with different scanners, a loss could be included that pushes the model to produce equal latent space representations for both images. Using unpaired data is possible when the samples' domains are known. The model can then be trained with a loss to decrease domain discrimination performance like Ren et al. (2019a) propose.

When explicitly providing different datasets during model training, the space of biases counted as ID increases. However, it needs to be explored whether such a model is able to generalize to new, unseen biases. As stated above, datasets have many different sources of variation, and the presented datasets can only cover parts thereof. This thesis builds upon the hypothesis that most biases stem from color variations. However, other biases, like blurriness, are also imaginable for a neural network to count as OOD. Collecting a dataset that covers all possible variations to extend the training dataset is far too time- and resource-intensive. Instead, explicitly exploring the space of OOD samples is a promising approach. In the latent space, one can explore which OOD regions are covered by the current datasets and which region in the latent space is not covered yet. GANs could artificially create OOD samples not covered by current datasets. Instead of creating samples in the latent space, which is not interpretable by humans, the GAN should be trained to generate input images. The GAN needs to be conditioned to generate OOD images based on the Mahalanobis kth-nearest-neighbor approach. Using these generated images during training extends the ID space; thus, it is possible to control manually to which biases the network should be insensitive. A difficulty in synthetically creating OOD images is that these need to be of very high quality and medically realistic. Thus, a close collaboration with pathologists is necessary. Suppose a pathologist states that a histopathology image cannot be analyzed, for instance, since it is too blurry or medically unreasonable. Such images should also be excluded from a training dataset.

Toward clinical application Since this thesis shows that an automated survival prediction from prostate cancer tissue images is possible, the next step for such a model is one toward clinical application. However, aspects like evaluating eCaReNet on other external datasets, ensuring robustness, and obeying regulations need to be considered (Homeyer et al., 2021).

The developed survival prediction model should support pathologists' decision-making in clinical practice. Thus, it needs to be explored whether pathologists trust the predictions when provided with a single risk score, risk group, or a survival curve per patient. Meyer et al. (2022) state that pathologists' decisions do not change when provided with model accuracy and information like stating the model focuses on cells. Evans et al. (2022) present a study where a questionnaire was sent to pathologists, and they find that visual explanations are preferred since those relate to the way pathologists think. Thus, it needs to be investigated if visual explanations for single images can build trust in eCaReNet.

Besides highlighting the patches that get the most attention in the MIL layer, GradCAM can be used to reveal relevant regions for the prediction in more detail (Selvaraju et al., 2017). The GradCAM algorithm tracks the gradient of the target class and thus highlights the regions that are most important for predicting that class. GradCAM might need to be adapted for survival prediction since no single classes are predicted. Another explainability method that can be applied, particularly when adding patient features to the model, is LIME (local interpretable model-agnostic explanations, Ribeiro et al., 2016). For LIME, the input is perturbed in various ways, and it is observed how the outcome changes. Thus, each input feature's importance can be estimated. Whether these local image-based explanations improve decision-making needs to be evaluated in a clinical study.

Ghassemi et al. (2021) argue that local explanations (i.e., explanations for single samples) are ambiguous and that proving robust model performance in a thorough validation study is sufficient for clinical usage. However, explanations might be necessary for certification (Heesen et al., 2020). In the study by Evans et al. (2022), the pathologists also state that trust would increase when showing reliable predictions in extensive tests and comparisons to physicians. Another opportunity that comes with explanations is to generate clinical insights. Pathologists can evaluate regions relevant to the model to find if the model discovered patterns that are yet unknown to contribute to cancer relapse. Visual explanations can also reveal which images the model mistakes, thus giving insights to improve training.

Challenges remain in the non-uniform, as yet non-standardized, data acquisition processes. Thus, a clinical study in one clinic might not suffice to prove the added value of a clinical decision support system. The research community has recognized the need for consistent standards, but the cost of reagents and scanners, pathologists' coloring preferences, time constraints for staining, and other aspects hinder a standardized tissue acquisition (Lang, 2006; Kanwal et al., 2022; Wright et al., 2021).

Appendix A

Supplementary material

A.1 State of the art

Table A.1.1: Extended overview of papers for prostate cancer stratification. AUROC: area under the receiver operator curve, AUPRC: area under the precision recall curve, acc: accuracy, pr: precision, sp: specificity, se: sensitivity, npv: negative predictive value, ppv: positive predictive value, pa: number of patients, im: number of images, n/a: not available.

Paper	**Task**	**Model**	**Metric**	**Dataset**
Campanella et al. (2018)	benign/ malignant	MIL + ResNet, VGG	AUROC 0.98	• non-public • biopsies • 12,160 im (n/a pat.)
Jimenez-del Toro et al. (2017)	low/ high	GoogLeNet	acc 0.78	• TCGA • prostatectomies • 235 im (n/a pa)
Burlutskiy et al. (2019)	binary segmentation	U-Net for different resolutions	F1 0.8 AUPRC 0.89	• non-public • biopsies + prostatectomies • 476 im (n/a pa)
Oner et al. (2022)	benign/ malignant glands	Mask R-CNN + multi resolution ResNet	AUROC 0.996 pr 0.997	• non-public • biopsies + prostatectomies • 99 im (99 pa)
Duran-Lopez et al. (2020)	benign/ malignant	feature extraction, wide and deep network	acc 0.999 sp 1 se 0.999 pr 1 F1 0.999	• non-public • biopsies • 97 im (n/a pat.)
Chen et al. (2019)	benign/ malignant	InceptionV4	AUROC 0.99	• TCGA + non-public • prostatectomies • 451 im (n/a pat.)

Bhattacharjee et al. (2022)	benign / malignant	custom CNN	AUROC 0.98 acc 0.93 pr 0.96 re 0.94 F1 0.93	• PANDA + non-public • biopsies • 1,900 im (n/a pa)
Karimi et al. (2020)	benign/ malignant +low/high	combination of 3 CNNs	acc 0.86-0.92	• GleasonChallenge • TMAs • 333 im (231 pa)
Ren et al. (2018a)	low/ high	AlexNet + Siamese network	acc 0.83	• TCGA + non-public • prostatectomies • 990 im (less pa)

Arvaniti et al. (2018)	Gleason grade group	MobileNet	kappa 0.75	• TMAZ • TMAs • 886 im (886 pa)
Nagpal et al. (2019)	Gleason grade group	InceptionV3	acc 0.7 AUROC 0.96	• TCGA + non-public • prostatectomies • 769 im (769 pa)
Ström et al. (2020)	Gleason grade group	InceptionV3	kappa 0.62	• non-public • biopsies • 9,001 im (1,474 pa)
Nagpal et al. (2020)	Gleason grade group	Xception-like	kappa 0.71	• non-public • biopsies • 1,276 im (1,112 pa)
Bulten et al. (2020)	Gleason grade group	U-Net	kappa 0.72-0.85	• TMAZ + non-public • TMAs + biopsies • 6,745 im (2,129 pa)
Marini et al. (2021b)	Gleason group	student/ teacher (ResNext)	kappa 0.45-0.76	• TMAZ + TCGA • TMAs + prostatectomies • 1,187 im (1,185 pa)
Marginean et al. (2021)	Gleason group + percent b/3/4/5	InceptionV3	se 0.8-1 sp 0.77-0.98 kappa 0.5-0.69	• non-public • biopsies • 735 im (195 pa)
Mun et al. (2021)	Gleason grade group	DenseNet (2stage: benign/ malginant, then groups)	kappa 0.88-0.90 acc 0.67-0.78	• non-public + GleasonChallenge • biopsies + TMAs • 7,844 im (1,032 pa)

Otálora et al. (2021)	Gleason grade group (strong + weak)	MobileNet	kappa 0.69	• TMAZ + TCGA • TMAs + prostatectomies • 1,187 im (n/a pa)
Tolkach et al. (2020)	Gleason pattern and ISUP	NASNetLarge + special evaluation of unsure patches	kappa 0.59-0.67	• TCGA + non-public • prostatectomies • 1,233 pa (n/a im)
Bhattacharjee et al. (2021)	Gleason 0,3,4,5	dual-channel CNN	acc 0.98 kappa 0.98 pr 0.98 re 0.98 F1 0.98	• PANDA + non-public • 6,000 patches (n/a pa)
Li et al. (2021)	Gleason group	VGG + two stage MIL, first b/m, then classification	acc 0.93 kappa 0.82	• SICAPv1 + non-public • biopsies • 20,308 im (909 pa)
Vuong et al. (2021)	Gleason grade	EfficientNet with categorical and ordinal classification	acc 0.70-0.80 F1 0.62-0.66 kappa 0.62-0.71	• TMAZ + GleasonChallenge • TMAs • 1,130 im (n/a pa)
Nir et al. (2018)	Gleason pattern	UNet adaptation, gland + nuclei segmentation + logistic regression	kappa 0.51	• GleasonChallenge + non-public • TMAs + prostatectomies • 563 im (287 pa)
Ikromjanov et al. (2022)	Gleason group	Vision Transformer	pr 0.8 re 0.8 F1 0.8	PANDA biopsies > 5,000 im

Singhal et al. (2022)	Gleason group	UNet, active learning, multi task	kappa 0.92-0.96 acc 0.83-0.89	• PANDA + non-public • biopsies • 6,670 im (n/a pa)
Leon and Martinez (2021)	Gleason score	InceptionV3, Xception + triplet loss	acc 0.62	• TMAZ • TMAs • 886 im (886 pa)
Zhang et al. (2021b)	benign, G3, G4/5	attention net + InceptionV3	acc 0.91 AUROC 0.98 pr 0.96	• TCGA • 54 im (n/a pa)
Salman et al. (2022)	ISUP	Yolo detection	pr 0.84-0.97 re 0.85-0.97 F1 0.94-0.97 acc 0.89-0.97	• non-public • biopsies • 500 im
Marini et al. (2021a)	Gleason pattern	adversarial CNN	kappa 0.47-0.73	• TMAZ + SICAPv2 + GleasonChallenge + DiagSet • TMAs+biopsies • n/a im (n/a pa) 83,091 patches
Koziarski et al. (2021)	binary + Gleason grade group	VGG19	acc 0.95	• DiagSet • biopsies • 5,179 im (n/a pa)
Silva-Rodríguez et al. (2020)	Gleason grade, sum, (cribriform)	custom CNN architecture	kappa 0.77-0.81 acc 0.67 F1 0.65	• SICAPv2 + TMAZ non-public • biopsies + TMAs • 648 im + 625 patches (608 pa)

Sandeman et al. (2022)	Gleason grade, area	custom CNN architecture	acc 0.67 kappa 0.77 se 0.58 sp 0.94 npv 0.94 ppv 0.57	• non-public • biopsies • 4,221 im (750 pa)

Table A.1.2: Extended literature overview for medical survival prediction using images as input. **Bold** literature uses only prostate cancer data, *italic* literature uses also prostate cancer data, besides others.
Data sources: M: MRI image, C: CT image, H: Histopathology image, R: radiograph, c: clinical data, o: omics data (genomics and/or transcriptomics and/or epigenomics and/or radiomics) – Data sizes: pa: patients, im: images – Metrics: AUC: area under the receiver operator curve, sp: specificity, se: sensitivity, acc: accuracy, KM: Kaplan-Meier, MAE: mean average error, OR: odds ratio, HR: hazard ratio – Loss: nlpl: negative log partial likelihood, i.e., Cox-loss, n/a: not available.

Paper	Task	Data	Model	Loss	Metric
BINARY					
Duanmu et al. (2020)	therapy response	M c o 112 pa	VGG-13 for 3D data	n/a	AUC 0.8 acc 0.89 F1 0.77 sp 0.88 se 0.68
Kumar et al. (2017)	relapse (5years)	H 220 pa	2 CNNs: detect nuclei + classification	binary cross entropy	AUC 0.81
Yamamoto et al. (2019)	relapse (1 and 5 years)	H 842 pa 9,916 im	Autoencoder + SVM	n/a	AUC 0.76-0.84 pseudo R-squared 0.26
Huang et al. (2022)	relapse (3 years)	H 416 pa 416 im	CNN	cross entropy	AUC 0.78
Exarchos et al. (2012)	relapse yes/no	C M 41 pa	feature extraction + Bayesian network	n/a	acc 0.83 se 0.79 sp 0.86

Yang et al. (2020)	response	C c o 99 pa 793 CTs	simple Temporal Attention	MSE	AUC 0.47 (im only) 0.8 (with c+o)

RISK SCORES

Esteva et al. (2022)	risk score (5 and 10 years)	H c 5,654 pa 16,204 im	self supervised pretraining + CatBoost fusion	n/a	time dep AUROC 0.67-0.77
Wulczyn et al. (2020)	risk score / risk interval (3 intervals)	H 6,096 pa 15,104 im	CNN similar to MobileNet	cross entropy (for risk in interval)	c-index 0.61 AUC 5year 0.7 KM (3)
Walhagen et al. (2022)	risk score (event < 3 years)	H 15,238 pa 15,238 im	EfficientNet + MIL	cross-entropy	AUC 0.79-0.93 KM (7)
Pinckaers et al. (2022)	year of relapse (year 0-4)	H 889 pa 2,963 im	ResNet	smooth L1 loss	OR 3.32 HR 3.02 KM curve (2 and 4)
Han et al. (2022)	3 groups: 0-3-5, >5 years	C c 198 pa	multibranch spatiotemporal ResNet	n/a	acc 0.87 F1 0.86
Zhou et al. (2020)	3 groups: 0-10, 10-15, >15 months	M c 163 pa	ResNet (per image, combine)	custom	acc 0.66 F1 0.59 pr 0.58 re 0.61
Abbet et al. (2020)	risk score	H 374 pa 660 im	divide and rule	custom: mean squared error, cross-entropy, reconstruction and rule loss	Brier 0.27 C-index 0.69

Chang et al. (2021)	risk score	H 53,454 pa	self supervision + custom hybrid aggregation network	nlpl	c-index 0.67-0.73 KM (2)
Di et al. (2020)	risk score	H 1,573 pa 4,790 im	hypergraph	ranking loss	c-index 0.66-0.68 AUROC 0.66-0.71 KM (2)
Fan et al. (2021)	risk score	H 1,368 pa 4,405 im	weakly pretraining, ResNet	rank loss + consistency	C-index 0.67-0.70
Kiyokawa et al. (2022)	risk score	H 68 pa 550 im	EfficientNet	n/a	acc 0.97 pr 0.96 re 0.97 F1 0.96
Laleh et al. (2021)	risk score	H 775 pa	ResNet per patch	Cox prop hazard loss	C-index 0.67-0.75 KM (2)
Li et al. (2018)	risk score	H 1,090 pa 1,451 im	Graph Network	Cox negative likelihood loss	c-index 0.66-0.71
Liu et al. (2022)	risk score	M c 649 pa	Risk Attention Network + Segmentation	neg log part likelihood	C-index 0.70 (im only) 0.73 (+c)
Muhammad et al. (2021)	risk score	H 265 pa	ResNet for patches + fusion	nlpl + cluster (own loss)	c-index 0.88 KM (2)
Yao et al. (2020)	risk score at time points	H 1,533 pa 2,791 im	VGG + clustering + siamese network + MIL	nlpl	c-index 0.70 AUC 0.71 KM (2)
Agarwal et al. (2021)	survival time as value/ difference	H c 389 pa 860 im	siamese network	pairwise ranking loss	C-index 0.62

SURVIVAL CURVES WITH COX MODEL

Ren et al. (2018b)	survival	H o 247 im	AlexNet + LSTM	nlpl	HR 5.73 C-index 0.74
Zhu et al. (2016)	survival	H 450 pa	DeepConvSurv (CNN + Cox)	nlpl	C-index 0.63
Liu and Kurc (2022)	5 intervals	H c 978 pa 978 im	6-channel input to MobileNet + Cox	extension of cross-entropy	C-index 0.70 (im only) 0.73 (with c) HR 1.19
Nam et al. (2022)	survival curve (600 days)	R c 5,372 pa	DenseNet + neural net + Cox	negative log likelihood	time dep. 5year AUC 0.67-0.76 (im only) 0.72-0.83 (+c) C-index 0.63-0.72 (im only) 0.68-0.79 (+c) KM (2) calibration
Jiang et al. (2021)	survival	H 2,375 pa 6,162 im	ResNet + Multihead attention	nlpl	C-index 0.64 HR 2.27
Li et al. (2019)	survival	C 84 pa	CNN with one risk output / Cox	nlpl	C-index 0.64
Li et al. (2022)	survival	H+g 1015 pa	custom architecture	nlpl	C-index 0.77 AUC 0.81 KM (2)
Lv et al. (2022)	survival	H o c 520 pa	ResNet + Linear model to Transformer fusion + Cox	nlpl	C-index 0.822 KM (2)
Mobadersany et al. (2018)	survival	H o 769 pa 1061 im	Cox on detected ROIs	NLL	C-index 0.74 (im only) 0.78 (+o) HR 7.15 KM (3)

Tabibu et al. (2019)	survival	H 2,093 im	extracted features + Cox	nlpl	KM (2)
Wang et al. (2018)	survival	C 129 pa	Residual Conv Autoencoder (weak pretraining) + LASSO Cox	nlpl	C-index 0.70 KM (2)
Wang et al. (2021)	survival	H o 627 pa 1993 im	Multi-modal Transformer-like	nlpl	C-index 0.69-0.70 (im only) 0.75-0.76 (+o) KM (2)
Zhu et al. (2017)	survival	H 651 pa 1,844 im	DeepConvSurv for patch clusters	nlpl	C-index 0.60-0.70

SURVIVAL CURVES WITHOUT COX MODEL

Xiao et al. (2020)	survival curve + time	H 769 pa, 1,061 im	CDOR (ResNet for censoring-aware deep ordinal regression)	censor-aware cross-entropy	MAE 321.2 C-index 0.74
Hermoza et al. (2022)	survival curve per interval + time (1-6423 days)	H Xray 16,013 pa 49,008 im	ResNet	adapt. of censor-aware cross-entropy	MAE 26.28 C-index 0.76
Popescu et al. (2022)	survival curve (to 10 years)	M c 269 pa	multiple network fusion + log-logistic survival model	negative likelihood	C-index 0.63 (im only) 0.74 (+c) Brier 0.19 (im only) 0.14 (+c)

Vale-Silva and Rohr (2021)	survival curve per interval (30 years)	H c o 11,081 pa 8,376 im	ResNext	negative log likelihood	C-index 0.57 (im only) 0.82 (+c) Brier 0.22 (im only) 0.14 (+c) KM (between cancers)
Lombardo et al. (2021)	survival per interval	C o c 1,037 pa	2D + 3D-CNN	negative log likelihood	C-index 0.67-0.88 AUC 0.63-0.89 KM (2)
Yala et al. (2021)	risk classification per year	Xray 91,520 pa 295,002 im	ResNet + Transformer aggregation + hazard pred	log likelihood	C-index 0.76-0.81 1-5-year AUROCs (0.76-0.90)

A.2 Survival prediction

This section includes additional result tables and plots for the experiments conducted in section 5.7 *Experiments*.

A.2.1 Multiple images

Table A.2.1: Results when combining the prediction of multiple images per patient. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D-cal.), only failure (f) or pass (p) is indicated. The best results are marked in bold.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
ISUP	**0.80**	0.75	-	
Surv1$_{\cap\text{Surv2}}$	0.77 (0.0039)	0.75 (0.0013)	0.104 (0.0005)	**p**
Surv2$_{\cap\text{Surv1}}$	0.74 (0.0069)	0.72 (0.0055)	0.134 (0.0287)	f
mean	**0.79** (0.0049)	**0.76** (0.0067)	**0.109** (0.0064)	f
pessimistic	0.77 (0.0136)	0.74 (0.0106)	0.133 (0.0305)	f
optimistic	0.77 (0.0078)	**0.76** (0.0039)	**0.104** (0.0027)	**p**
concatenated	**0.79** (0.0066)	**0.76** (0.0056)	0.115 (0.0147)	f
Test set				
ISUP	0.82	0.76	-	
Surv1$_{\cap\text{Surv2}}$	0.76 (0.0050)	0.73 (0.0038)	0.114 (0.0006)	**p**
Surv2$_{\cap\text{Surv1}}$	0.72 (0.0077)	0.70 (0.0029)	0.135 (0.0205)	**p**
mean	**0.77** (0.0019)	**0.74** (0.0030)	**0.115** (0.0032)	f
pessimistic	0.76 (0.0085)	0.73 (0.0061)	0.133 (0.0227)	f
optimistic	0.75 (0.0060)	0.73 (0.0015)	**0.115** (0.0025)	**p**
concatenated	**0.77** (0.0036)	**0.74** (0.0049)	0.121 (0.0102)	**p**

A.2.2 Multimodal data input

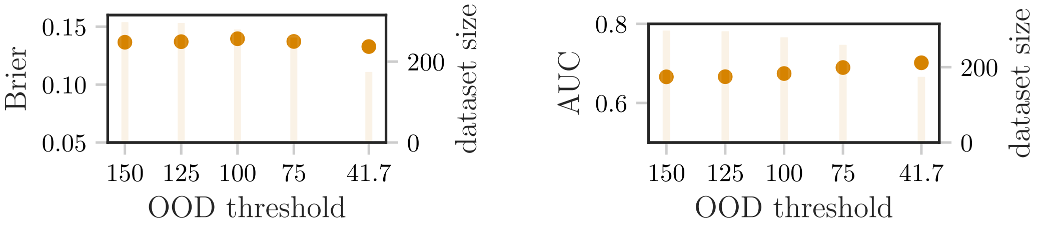
Table A.2.2: Results when using only the image or adding PSA value, tumor diameter, and volume to eCaReNet. Values are the mean of five training runs with the standard deviation in parentheses. For d-calibration (D-cal.), all chi-square tests pass (p). The best results are marked in bold. dm: tumor diameter, vol: tumor volume.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
ISUP	0.79	0.77	-	
image only	0.77 (0.0069)	0.77 (0.0026)	0.105 (0.0015)	**p**
+age	0.76 (0.0065)	0.76 (0.0037)	0.107 (0.0011)	**p**
+psa	0.78 (0.0069)	0.78 (0.0035)	0.103 (0.0021)	**p**
+dm	0.80 (0.0038)	0.79 (0.0034)	0.100 (0.0009)	**p**
+vol	**0.81** (0.0110)	**0.80** (0.0044)	0.099 (0.0024)	**p**
+age+psa+dm+vol	**0.81** (0.0085)	**0.80** (0.0027)	**0.097** (0.0025)	**p**
+psa+dm+vol	**0.81** (0.0040)	**0.80** (0.0025)	**0.097** (0.0008)	**p**
Test set				
ISUP	0.77	0.76	-	
image only	0.75 (0.0051)	0.74 (0.0028)	0.112 (0.0012)	**p**
+age	0.75 (0.0073)	0.74 (0.0051)	0.113 (0.0015)	**p**
+psa	0.76 (0.0058)	0.75 (0.0044)	0.111 (0.0016)	**p**
+dm	**0.79** (0.0069)	**0.77** (0.0051)	0.108 (0.0021)	**p**
+vol	0.77 (0.0052)	0.76 (0.0024)	0.108 (0.0022)	**p**
+age+psa+dm+vol	**0.79** (0.0063)	**0.77** (0.0068)	**0.106** (0.0020)	**p**
+psa+dm+vol	**0.79** (0.0033)	**0.77** (0.0038)	**0.106** (0.0007)	**p**

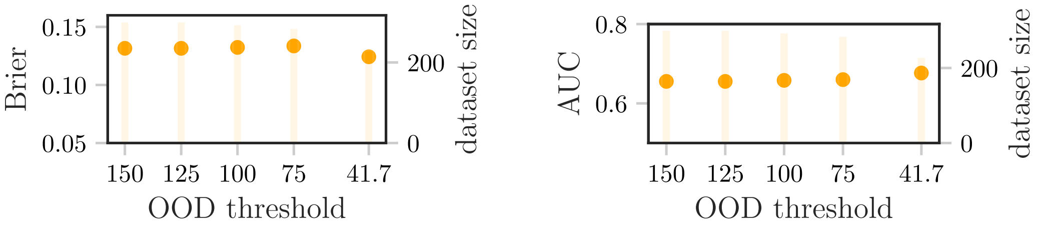
A.3 Robustness

This section includes additional result tables and plots for the experiments conducted in section 6.4 *Experiments*.

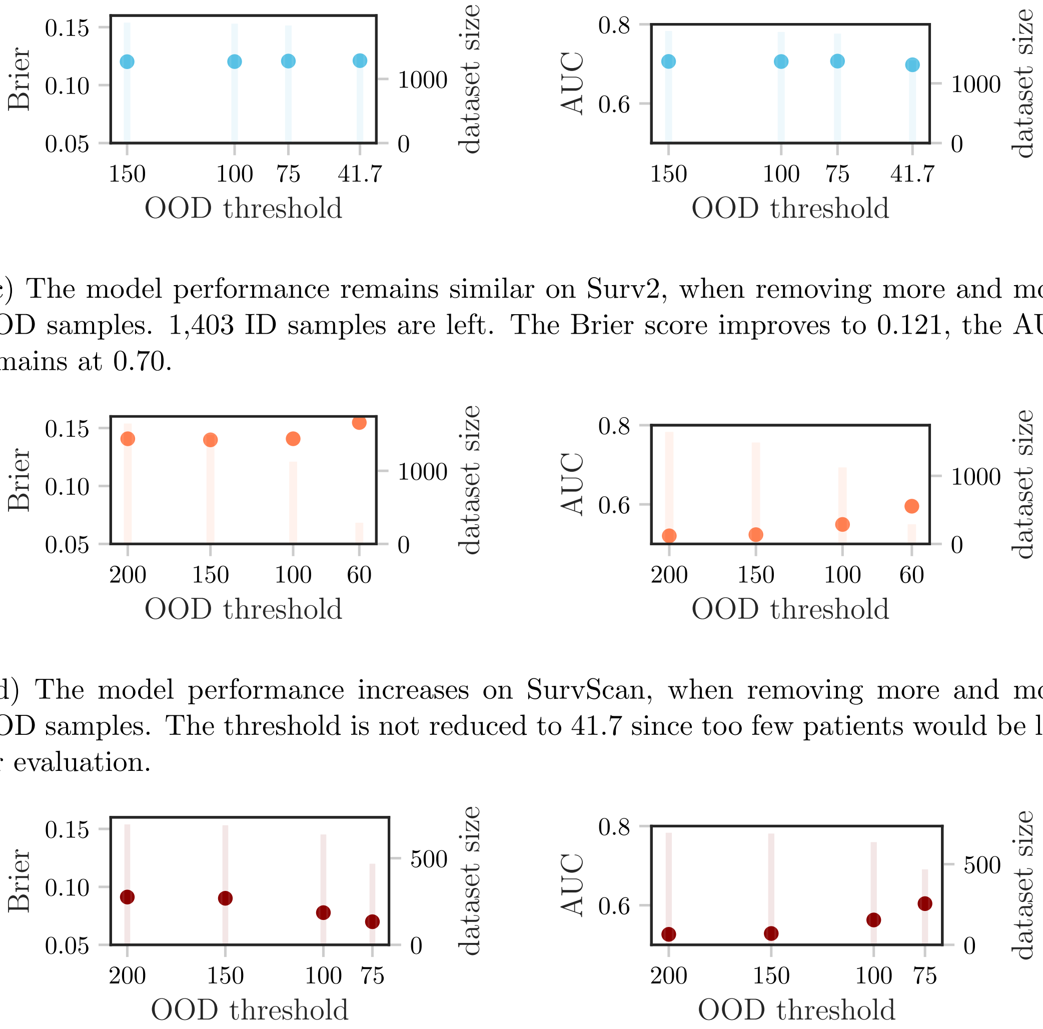
A.3.1 OOD threshold variation



(a) The model performance increases on SurvThick, when removing more and more OOD samples. 177 samples are left in the ID set. The Brier score improves to 0.133, the AUC increases to 0.70.



(b) The model performance increases on SurvLongStain when removing more and more OOD samples. 230 patients are left in the ID set. The Brier score improves to 0.124, the AUC increases to 0.68.



(c) The model performance remains similar on Surv2, when removing more and more OOD samples. 1,403 ID samples are left. The Brier score improves to 0.121, the AUC remains at 0.70.

(d) The model performance increases on SurvScan, when removing more and more OOD samples. The threshold is not reduced to 41.7 since too few patients would be left for evaluation.

(e) The model performance increases on SurvPCBN, when removing more and more OOD samples. The threshold is not reduced to 41.7 since too few patients would be left for evaluation.

Figure A.3.1: For different datasets, the influence of the OOD threshold on Brier score and AUC are visualized. The dots show the Brier scores in the left, AUCs in the right plots. The bars indicate the test set sizes, which decrease with removal of OOD samples.

A.3.2 Results using OOD detection and ClusterMatch

Table A.3.1: Results of eCaReNet on different test datasets after applying ClusterMatch (CM) and keeping only in-distribution (ID) samples afterward. Best results are marked in bold. For d-calibration (D-cal.), only pass (p) or failure (f) is indicated. For better readability, subscripts like $_{\cap\text{Surv1}}$ are omitted when CM or ID is indicated.

Test set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
Surv1$_{\cap\text{Surv2}}$	0.77	0.74	0.112	**p**
Surv2$_{\cap\text{Surv1}}$	**0.71**	0.69	0.120	**p**
Surv2$_{CM}$	**0.71**	**0.70**	**0.119**	**p**
Surv2$_{ID}$	0.70	0.69	0.121	**p**
Surv2$_{\cap\text{SurvDiff}}$	0.69	0.67	0.129	**p**
SurvLongStain$_{\cap\text{SurvDiff}}$	0.66	**0.66**	0.132	**p**
SurvLongStain$_{CM}$	0.64	0.64	0.132	**p**
SurvLongStain$_{ID}$	**0.67**	**0.66**	**0.126**	**p**
SurvThick$_{\cap\text{SurvDiff}}$	0.67	0.66	0.136	**p**
SurvThick$_{CM}$	0.67	0.66	0.133	**p**
SurvThick$_{ID}$	**0.68**	**0.68**	**0.129**	**p**
SurvThin$_{\cap\text{SurvDiff}}$	0.57	0.57	0.142	**p**
SurvThin$_{CM}$	0.61	0.62	0.135	**p**
SurvThin$_{ID}$	**0.68**	**0.67**	**0.113**	**p**
Surv1$_{\cap\text{SurvScan}}$	0.79	0.75	0.103	**p**
SurvScan$_{\cap\text{Surv1}}$	0.52	0.52	0.142	f
SurvScan$_{CM}$	**0.71**	0.69	0.114	f
SurvScan$_{ID}$	0.70	**0.71**	**0.105**	**p**
SurvPCBN	0.53	0.55	0.104	f
SurvPCBN$_{CM}$	0.61	0.64	0.097	**p**
SurvPCBN$_{ID}$	**0.69**	**0.71**	**0.090**	**p**

Table A.3.2: Results of eCaReNet on different validation datasets after applying ClusterMatch (CM) and keeping only in-distribution (ID) samples afterward. Best results are marked in bold. For d-calibration (D-cal.), only pass (p) or failure (f) is indicated. For better readability, subscripts like $_{\cap\text{Surv1}}$ are omitted when CM or ID is indicated.

Validation set	**AUC ↑**	**C-index ↑**	**Brier ↓**	**D-cal.**
Surv1$_{\cap\text{Surv2}}$	0.77	0.75	0.103	**p**
Surv2$_{\cap\text{Surv1}}$	**0.73**	**0.72**	0.111	f
Surv2$_{CM}$	**0.73**	0.70	0.111	**p**
Surv2$_{ID}$	**0.73**	0.71	**0.109**	**p**
Surv2$_{\cap\text{SurvDiff}}$	0.75	0.73	0.112	**p**
SurvLongStain$_{\cap\text{SurvDiff}}$	**0.74**	**0.71**	**0.115**	**p**
SurvLongStain$_{CM}$	0.73	0.70	0.118	**p**
SurvLongStain$_{ID}$	0.73	**0.71**	0.121	**p**
SurvThick$_{\cap\text{SurvDiff}}$	0.69	0.66	**0.124**	**p**
SurvThick$_{CM}$	**0.72**	0.67	0.125	**p**
SurvThick$_{ID}$	**0.72**	**0.69**	0.131	**p**
SurvThin$_{\cap\text{SurvDiff}}$	0.48	0.51	0.138	**p**
SurvThin$_{CM}$	**0.68**	**0.65**	0.128	**p**
SurvThin$_{ID}$	0.64	0.62	**0.109**	**p**
Surv1$_{\cap\text{SurvScan}}$	0.79	0.75	0.108	**p**
SurvScan$_{\cap\text{Surv1}}$	0.52	0.53	0.150	f
SurvScan$_{CM}$	0.71	0.68	0.120	f
SurvScan$_{ID}$	**0.72**	**0.69**	**0.126**	**p**

Appendix B

Publications originating from this thesis

Dietrich, E., Fuhlert, P., Ernst, A., Sauter, G., Lennartz, M., Stiehl, H. S., Zimmermann, M., & Bonn, S. (2021). Towards Explainable End-to-End Prostate Cancer Relapse Prediction from H&E Images Combining Self-Attention Multiple Instance Learning with a Recurrent Neural Network. In *Proceedings of Machine Learning for Health* (Vol. 2020, pp. 38–53). https://ml4health.github.io/2021/poster_A1.html

Fuhlert, P., Ernst, A., **Dietrich, E.**, Westhaeusser, F., Kloiber, K., & Bonn, S. (2022). Deep Learning-Based Discrete Calibrated Survival Prediction. *IEEE International Conference on Digital Health (ICDH), 2022*, 1–6. https://doi.org/10.1109/ICDH55609.2022.00034

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from `https://www.tensorflow.org`.

Abbet, C., Zlobec, I., Bozorgtabar, B., and Thiran, J.-P. (2020). Divide-and-rule: Self-supervised learning for survival analysis in colorectal cancer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 480–489. `http://arxiv.org/abs/2007.03292`.

Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck, A. H., and Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *The Journal of Pathology*, 249(3):286–294. `https://doi.org/10.1002/path.5331`.

Agarwal, S., Eltigani Osman Abaker, M., and Daescu, O. (2021). Survival prediction based on histopathology imaging and clinical data: A novel, whole slide CNN approach. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, volume 12905, pages 762–771. Springer International Publishing. `https://doi.org/10.1007/978-3-030-87240-3_73`.

Ahmed, H. U., El-Shater Bosaily, A., Brown, L. C., Gabe, R., Kaplan, R., Parmar, M. K., Collaco-Moraes, Y., Ward, K., Hindley, R. G., Freeman, A., Kirkham, A. P., Oldroyd, R., Parker, C., and Emberton, M. (2017). Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389:815–822. `https://doi.org/10.1016/S0140-6736(16)32401-1`.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53). `https://doi.org/10.1186/s40537-021-00444-8`.

Andres, A., Montano-Loza, A., Greiner, R., Uhlich, M., Jin, P., Hoehn, B., Bigam, D., Shapiro, J. A. M., and Kneteman, N. M. (2018). A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PLoS ONE*, 13(3):1–14. `https://doi.org/10.1371/journal.pone.0193523`.

Andriole, G. L., Crawford, E. D., Grubb, R. L., Buys, S. S., Chia, D., Church, T. R., Fouad, M. N., Gelmann, E. P., Kvale, P. A., Reding, D. J., Weissfeld, J. L., Yokochi, L. A., O'Brien, B., Clapp, J. D., Rathmell, J. M., Riley, T. L., Hayes, R. B., Kramer, B. S., Izmirlian, G., Miller, A. B., Pinsky, P. F., Prorok, P. C., Gohagan, J. K., and Berg, C. D. (2009). Mortality results from a randomized prostate-cancer screening trial. *New England Journal of Medicine*, 360(13):1310–1319. `https://doi.org/10.1056/nejmoa0810696`.

Antolini, L., Boracchi, P., and Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24:3927–3944. `https:doi.org/10.1002/sim.2427`.

Apostolidis, K. D. and Papakostas, G. A. (2021). A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, 10(2132). `https:doi.org/10.3390/electronics10172132`.

Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., Fankhauser, C., Wey, N., Wild, P. J., Rüschoff, J. H., and Claassen, M. (2018). Automated gleason grading of prostate cancer tissue microarrays via deep learning. *Nature Scientific Reports*, 8(1):1–11. `https:doi.org/10.1038/s41598-018-30535-1`.

Azam, A. S., Miligy, I. M., Kimani, P. K.-U., Maqbool, H., Hewitt, K., Rajpoot, N. M., and Snead, D. R. (2021). Diagnostic concordance and discordance in digital pathology: A systematic review and meta-analysis. *Journal of Clinical Pathology*, 74(7):448–455. `https:doi.org/10.1136/jclinpath-2020-206764`.

Azevedo Tosta, T. A., de Faria, P. R., Alves Neves, L., and Zanchetta do Nascimento, M. (2019). Color normalization of faded H&E-stained histological images using spectral matching. *Computers in Biology and Medicine*, 111:103344. https:doi.org/10.1016/j.compbiomed.2019.103344.

Banerji, S. and Mitra, S. (2022). Deep learning in histopathology: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(1):1–13. https:doi.org/10.1002/widm.1439.

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., and Hamilton, P. W. (2017). QuPath: Open source software for digital pathology image analysis. *Nature Scientific Reports*, 7(16878). https:doi.org/10.1038/s41598-017-17204-5.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115. https:doi.org/10.1016/j.inffus.2019.12.012.

Bertelli, E., Mercatelli, L., Marzi, C., Pachetti, E., Baccini, M., Barucci, A., Colantonio, S., Gherardini, L., Lattavo, L., Pascali, M. A., Agostini, S., and Miele, V. (2022). Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric MRI. *Frontiers in Oncology*, 11(January):1–14. https:doi.org/10.3389/fonc.2021.802964.

Bhattacharjee, S., Hwang, Y.-B., Ikromjanov, K., Sumon, R. I., Kim, H.-C., and Choi, H.-K. (2022). An explainable computer vision in histopathology: Techniques for interpreting black box model. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 392–398. IEEE. https:doi.org/10.1109/ICAIIC54071.2022.9722656.

Bhattacharjee, S., Ikromjanov, K., Hwang, Y.-B., Sumon, R. I., Kim, H.-C., and Choi, H.-K. (2021). Detection and classification of prostate cancer using dual-channel parallel convolution. In *Proceedings of the Future Technologies Conference (FTC) 2021, Volume 2*, volume 2, pages 66–83. Springer International Publishing. http://doi.org/10.1007/978-3-030-89880-9_6.

Blanche, P., Dartigues, J.-F., and Jacqmin-Gadda, H. (2013). Review and comparison of ROC curve estimators for a time-dependent outcome with marker-dependent censoring. *Biometrical Journal*, 55(5):687–704. https:doi.org/10.1002/bimj.201200045.

Blanche, P., Kattan, M. W., and Gerds, T. A. (2019). The c-index is not proper for the evaluation of t -year predicted risks. *Biostatistics*, 20(2):347–357. https:doi.org/10.1093/biostatistics/kxy006.

Bland, J. M. and Altman, D. G. (2004). The logrank test. *BMJ (Clinical research ed.)*, 328(7447):1073. https:doi.org/10.1136/bmj.328.7447.1073.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*. https://opencv.org/.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3. https:doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., van Boven, H., Vink, R., Hulsbergen-van de Kaa, C., 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. O., 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. T., Li, W., Li, J., Speier, W., Arnold, C., Kim, K., Bae, B., Kim, Y. W., Lee, H.-S., and Park, J. (2022). Artificial intelligence for diagnosis and gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*. https://doi.org/10.1038/s41591-021-01620-2.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., and Litjens, G. (2020). Automated deep-learning system for gleason grading of prostate cancer using biopsies : a diagnostic study. *The Lancet Oncology*, 21(2):233–241. https://doi.org/10.1016/S1470-2045(19)30739-9.

Burlutskiy, N., Pinchaud, N., Gu, F., Hägg, D., Andersson, M., Björk, L., Eurén, K., Svensson, C., Wilén, L. K., and Hedlund, M. (2019). Segmenting potentially cancerous areas in prostate biopsies using semi-automatically annotated data. In *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*. https://proceedings.mlr.press/v102/burlutskiy19a.html.

Campanella, G., Silva, V. W. K., and Fuchs, T. J. (2018). Terabyte-scale deep multiple instance learning for classification and localization in pathology. *arxiv preprint*. http://arxiv.org/abs/1805.06983.

Chan, T., Esedoglu, S., and Ni, K. (2007). Histogram based segmentation using wasserstein distances. In *Scale Space and Variational Methods in Computer Vision (SSVM 2007)*, pages 697–708. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-72823-8_60.

Chan, T. Y., Partin, A. W., Walsh, P. C., and Epstein, J. I. (2000). Prognostic significance of gleason score 3+4 versus gleason score 4+3 tumor at radical prostatectomy. *Urology*, 56(5):823–827. https://doi.org/10.1016/S0090-4295(00)00753-6.

Chang, J.-R., Lee, C.-Y., Chen, C.-C., Reischl, J., Qaiser, T., and Yeh, C.-Y. (2021). Hybrid aggregation network for survival analysis from whole slide histopathological images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, volume 12905, pages 731–740. Springer International Publishing. https://doi.org/10.1007/978-3-030-87240-3_70.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. (2020a). Robust out-of-distribution detection for neural networks. *AAAI 2022, Workshop on Adversarial Machine Learning and Beyond*. http://arxiv.org/abs/2003.09711.

Chen, P.-H. C., Gadepalli, K., MacDonald, R., Liu, Y., Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G. S., Hipp, J. D., Mermel, C. H., and Stumpe, M. C. (2019). An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine*, 25:1453–1457. https://doi.org/10.1038/s41591-019-0539-7.

Chen, R. J., Lu, M. Y., Weng, W.-h., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. (2021). Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3995–4005. https://doi.org/10.1109/ICCV48922.2021.00398.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR. https://proceedings.mlr.press/v119/chen20j.html.

Cheon, S., Agarwal, A., Popovic, M., Milakovic, M., Lam, M., Fu, W., DiGiovanni, J., Lam, H., Lechner, B., Pulenzas, N., Chow, R., and Chow, E. (2016). The accuracy of clinicians' predictions of survival in advanced cancer: A review. *Annals of Palliative Medicine*, 5(1):22–29. https://doi.org/10.3978/j.issn.2224-5820.2015.08.04.

Chlipala, E. A., Butters, M., Brous, M., Fortin, J. S., Archuletta, R., Copeland, K., and Bolon, B. (2021). Impact of preanalytical factors during histology processing on section suitability for digital image analysis. *Toxicologic Pathology*, 49(4):755–772. https://doi.org/10.1177/0192623320970534.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1724–1734. https://doi.org/10.3115/v1/d14-1179.

Chollet, F. et al. (2015). Keras. https://keras.io.

Ciga, O., Xu, T., and Martel, A. L. (2022). Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7. https://doi.org/10.1016/j.mlwa.2021.100198.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, XX(1):37–46. https://doi.org/10.1177/001316446002000104.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220. http://www.jstor.org/stable/2985181.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

Denysenko, A., Savchenko, T., Dovbysh, A., Romaniuk, A., and Moskalenko, R. (2022). Artificial intelligence approach in prostate cancer diagnosis : Bibliometric analysis. *Galician medical journal*, 29(2). `https://doi.org/10.21802/gmj.2022.2.5`.

Di, D., Li, S., Zhang, J., and Gao, Y. (2020). Ranking-based survival prediction on histopathological whole-slide images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 428–438. `https://doi.org/10.1007/978-3-030-59722-1_41`.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302. `https://doi.org/10.2307/1932409`.

Dietrich, E., Fuhlert, P., Ernst, A., Sauter, G., Lennartz, M., Stiehl, H. S., Zimmermann, M., and Bonn, S. (2021). Towards explainable end-to-end prostate cancer relapse prediction from H&E images combining self-attention multiple instance learning with a recurrent neural network. In *Proceedings of Machine Learning for Health, PMLR*, volume 2020, pages 38–53. PMLR. `https://ml4health.github.io/2021/poster_A1.html`.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71. `https://doi.org/10.1016/s0004-3702(96)00034-3`.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021*. `https://iclr.cc/virtual/2021/poster/3013`.

Dozat, T. (2016). Incorporating nesterov momentum into Adam. *4th International Conference on Learning Representations, ICLR 2016, Workshop Track*, 1. `https://openreview.net/forum?id=OM0jvwB8jIp57ZJjtNEZ`.

Drenkow, N., Sani, N., Shpitser, I., and Unberath, M. (2021). Robustness in deep learning for computer vision: Mind the gap? *arxiv preprint*, pages 1–22. `http://arxiv.org/abs/2112.00639`.

Duanmu, H., Huang, P. B., Brahmavar, S., Lin, S., Ren, T., Kong, J., Wang, F., and Duong, T. Q. (2020). Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging , molecular and demographic data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 242–252. https://doi.org/10.1007/978-3-030-59713-9_24.

Duran-Lopez, L., Dominguez-Morales, J. P., Conde-Martin, A. F., Vicente-Diaz, S., and Linares-Barranco, A. (2020). PROMETEO: A cnn-based computer-aided diagnosis system for WSI prostate cancer detection. *IEEE Access*, 8:128613–128628. https://doi.org/10.1109/ACCESS.2020.3008868.

Egevad, L., Granfors, T., Karlberg, L., Bergh, A., and Stattin, P. (2002). Prognostic value of the gleason score in prostate cancer. *BJU International*, 89:538–542. https://doi.org/10.1046/j.1464-410X.2002.02669.x.

Egevad, L., Mazzucchelli, R., and Montironi, R. (2012). Implications of the international society of urological pathology modified gleason grading system. *Archives of Pathology & Laboratory Medicine*, 136(April). https://doi.org/10.5858/arpa.2011-0495-RA.

Eichelberger, L. E., Koch, M. O., Eble, J. N., Ulbright, T. M., Juliar, B. E., and Cheng, L. (2005). Maximum tumor diameter is an independent predictor of prostate-specific antigen recurrence in prostate cancer. *Modern Pathology*, 18:886–890. https://doi.org/10.1038/modpathol.3800405.

Emmert-Streib, F. and Dehmer, M. (2019). Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1:1013–1038. https://doi.org/10.3390/make1030058.

empaia.org (2022). empaia project. https://www.empaia.org/ (last accessed October 18, 2022.

Epstein, J. I. (2018). Prostate cancer grading: a decade after the 2005 modified system. *Modern Pathology*, 31:47–63. https://doi.org/10.1038/modpathol.2017.133.

Epstein, J. I., Allsbrook, W. C., Amin, M. B., Egevad, L. L., Bastacky, S., López Beltrán, A., Berner, A., Billis, A., Boccon-Gibod, L., Cheng, L., Civantos, F., Cohen, C., Cohen, M. B., Datta, M., Davis, C., Delahunt, B., Delprado, W., Eble, J. N., Foster, C. S., Furusato, M., Gaudin, P. B., Grignon, D. J., Humphrey, P. A., Iczkowski, K. A., Jones, E. C., Lucia, S., McCue, P. A., Nazeer, T., Oliva, E., Pan, C. C., Pizov, G., Reuter, V., Samaratunga, H., Sebo, T., Sesterhenn, I., Shevchuk, M., Srigley, J. R., Suzigan, S., Takahashi, H., Tamboli, P., Tan, P. H., Tètu, B., Tickoo, S., Tomaszewski, J. E., Troncoso, P., Tsuzuki, T., True, L. D., Van Der Kwast, T., Wheeler, T. M., Wojno, K. J., and Young, R. H. (2005). The 2005 international society of urological pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. *American Journal of Surgical Pathology*, 29(9):1228–1242. `https://doi.org/10.1097/01.pas.0000173646.99337.b1`.

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., Humphrey, P. A., and the Grading Committee (2016). 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(2):244–252. `https://doi.org/10.1097/PAS.0000000000000530`.

Esteban, Á. E., López-Pérez, M., Colomer, A., Sales, M. A., Molina, R., and Naranjo, V. (2019). A new optical density granulometry-based descriptor for the classification of prostate histological images using shallow and deep gaussian processes. *Computer Methods and Programs in Biomedicine*, 178:303–317. `https://doi.org/10.1016/j.cmpb.2019.07.003`.

Esteva, A., Feng, J., van der Wal, D., Huang, S.-C., Simko, J. P., DeVries, S., Chen, E., Schaeffer, E. M., Morgan, T. M., Sun, Y., Ghorbani, A., Naik, N., Nathawani, D., Socher, R., Michalski, J. M., Roach, M., Pisansky, T. M., Monson, J. M., Naz, F., Wallace, J., Ferguson, M. J., Bahary, J.-P., Zou, J., Lungren, M., Yeung, S., Ross, A. E., Kucharczyk, M., Souhami, L., Ballas, L., Peters, C. A., Liu, S., Balogh, A. G., Randolph-Jackson, P. D., Schwartz, D. L., Girvigian, M. R., Saito, N. G., Raben, A., Rabinovitch, R. A., Katato, K., Sandler, H. M., Tran, P. T., Spratt, D. E., Pugh, S., Feng, F. Y., and Mohamad, O. (2022). Prostate cancer therapy personalization via multi-modal deep learning on randomized phase iii clinical trials. *npj Digital Medicine*, 5(1). `https://doi.org/10.1038/s41746-022-00613-w`.

Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., and Holzinger, A. (2022). The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems*, 133:281–296. `https://doi.org/10.1016/j.future.2022.03.009`.

Exarchos, K. P., Goletsis, Y., and Fotiadis, D. I. (2012). Multiparametric decision support system for the prediction of oral cancer reoccurrence. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1127–1134. `https://doi.org/10.1109/TITB.2011.2165076`.

Fan, L., Sowmya, A., Meijering, E., and Song, Y. (2021). Learning visual features by colorization for slide-consistent survival prediction from whole slide images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021*, volume 12908, pages 592–601. Springer International Publishing. `https://doi.org/10.1007/978-3-030-87237-3_57`.

Fan, X., Sun, Z., and Tian, E. (2022). Histological image color normalization using a skewed normal distribution mixed model. *Journal of the Optical Society of America A*, 39(3):441 – 451. `https://doi.org/10.1364/josaa.446221`.

Ferlay, J., Ervik, M., Lam, F., Colombet, M., Mery, L., Piñeros, M., Znaor, A., Soerjomataram, I., and Bray, F. (2020). Global cancer observatory: Cancer today. lyon, france: International agency for research on cancer. `https://gco.iarc.fr/today`(last accessed May 28, 2022).

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons. `https://doi.org/10.1002/9781118150672`.

Fort, S., Ren, J., and Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Neural Information Processing Systems - NeurIPS 2021*, volume 100. `https://proceedings.neurips.cc/paper/2021/hash/3941c4358616274ac2436eacf67fae05-Abstract.html`.

Fuhlert, P., Ernst, A., Dietrich, E., Westhaeusser, F., Kloiber, K., and Bonn, S. (2022). Deep learning-based discrete calibrated survival prediction. *IEEE International Conference on Digital Health (ICDH), 2022*, pages 1–6. `https://doi.org/10.1109/ICDH55609.2022.00034`.

Furihata, M. and Takeuchi, T. (2017). Gleason grading. In *Encyclopedia of Cancer*, pages 1904–1907. Springer Berlin Heidelberg, Berlin, Heidelberg. `https://doi.org/10.1007/978-3-662-46875-3_2415`.

George, A., Stead, T. S., and Ganti, L. (2020). What's the risk: Differentiating risk ratios, odds ratios, and hazard ratios? *Cureus*, 12(8). `https://doi.org/10.7759/cureus.10047`.

Gerds, T. and Kattan, M. (2021). *Medical Risk Prediction: With Ties to Machine Learning*. CRC Press. https://doi.org/10.1201/9781138384484.

Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3:e745–e750. https://doi.org/10.1016/s2589-7500(21)00208-9.

Giunchiglia, E., Nemchenko, A., and van der Schaar, M. (2018). RNN-SURV: A deep recurrent model for survival analysis. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 23–32. Springer International Publishing. https://doi.org/10.1007/978-3-030-01424-7_3.

Gleason, D. F. and Mellinger, G. T. (1974). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *Journal of Urology*, 111(1):58–64. https://doi.org/10.1016/s0022-5347(17)59889-4.

Gonzalez, R. C. and Woods, R. E. (2002). *Digital Image processing*, volume 3. Pearson Education. https://dl.acm.org/doi/10.5555/1076432.

Gordetsky, J. and Epstein, J. (2016). Grading of prostatic adenocarcinoma: Current state and prognostic implications. *Diagnostic Pathology*, 11(1):1–9. http://doi.org/10.1186/s13000-016-0478-2.

Greenwood, M. (1926). A report on the natural duration of cancer. In *Reports on Public Health and Medical subjects*. http://doi.org/10.1001/jama.1927.02680330059037.

Grignon, D. J. (2018). Prostate cancer reporting and staging: needle biopsy and radical prostatectomy specimens. *Modern Pathology*, 31:96–109. http://doi.org/10.1038/modpathol.2017.167.

Haider, H., Hoehn, B., Davis, S., and Greiner, R. (2020). Effective ways to build and evaluate individual survival distributions. *Journal of Machine Learning Research*, 21:1–63. http://jmlr.org/papers/v21/18-772.html.

Han, J., Xiao, N., Yang, W., Luo, S., Zhao, J., Qiang, Y., Chaudhary, S., and Zhao, J. (2022). MS-ResNet: disease-specific survival prediction using longitudinal ct images and clinical data. *International Journal of Computer Assisted Radiology and Surgery*. http://doi.org/10.1007/s11548-022-02625-z.

Hanna, M. G., Parwani, A., and Sirintrapun, S. J. (2020). Whole slide imaging: Technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259. http://doi.org/10.1097/PAP.0000000000000273.

Harrell, Frank E., J., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. (1982). Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546. http://doi.org/10.1001/jama.1982.03320430047030.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. http://doi.org/10.1109/CVPR.2016.90.

Heesen, J., Müller-Quade, J., Wrobel, S., Beyerer, J., Brink, G., Faisst, W., Hoffmann, M., Huchler, N., Kirchner, E., Matzner, T., Peissner, M., Peylo, C., Schauf, T., Straube, S., Suchy, O., Wolfgram, S., Schmid, U., Rostalski, F., Poretschkin, M., and Birnstill, P. (2020). Zertifizierung von KI-systemen. *Whitepaper aus der Plattform Lernende Systeme*. https://publica.fraunhofer.de/entities/publication/aa1d90e6-3da0-4336-b2ac-503849368d97/details.

Heidenreich, A. (2007). Guidelines and counselling for treatment options in the management of prostate cancer. In *Recent Results in Cancer Research*, volume 175, pages 131–162. Springer-Verlag Berlin Heidelberg. http://doi.org/10.1007/978-3-540-40901-4_9.

Heinz, C. N., Echle, A., Foersch, S., Bychkov, A., and Kather, J. N. (2022). The future of artificial intelligence in digital pathology – results of a survey across stakeholder groups. *Histopathology*, 80(7):1121–1127. http://doi.org/10.1111/his.14659.

Heitz, J., Ficek, J., Faltys, M., Merz, T. M., Rätsch, G., and Hüser, M. (2021). WRSE – a non-parametric weighted-resolution ensemble for predicting individual survival distributions in the icu. *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*. https://proceedings.mlr.press/v146/heitz21a.html.

Hermoza, R., Maicas, G., Nascimento, J. C., and Carneiro, G. (2022). Censor-aware semi-supervised learning for survival time prediction from medical images. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. https://doi.org/10.1007/978-3-031-16449-1_21.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780. http://doi.org/10.1162/neco.1997.9.8.1735.

Homeyer, A., Lotz, J., Weiss, N., Romberg, D., Höfener, H., Zerbe, N., and Hufnagl, P. (2021). Artificial intelligence in pathology: From prototype to product. *Journal of Pathology Informatics*, 12(13). http://doi.org/10.4103/jpi.jpi_84_20.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arxiv preprint*. http://arxiv.org/abs/1704.04861.

Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., Cipriani, N., Grossman, R. L., and Pearson, A. T. (2021). The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(4423):1–13. https://doi.org/10.1038/s41467-021-24698-1.

Hsieh, P. F., Li, T. R., Lin, W.-C., Chang, H., Huang, C.-P., Chang, C.-H., Yang, C. R., Yeh, C.-C., Huang, W.-C., and Wu, H.-C. (2021). Combining prostate health index and multiparametric magnetic resonance imaging in estimating the histological diameter of prostate cancer. *BMC Urology*, 21(1):1–8. http://doi.org/10.1186/s12894-021-00928-y.

Hsu, Y.-C., Shen, Y., Jin, H., and Kira, Z. (2020). Generalized ODIN: Detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. http://doi.org/10.1109/CVPR42600.2020.01096.

Hu, S., Fridgeirsson, E. A., van Wingen, G., and Welling, M. (2021a). Transformer-based deep survival analysis. *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, 2021:1–16. https://proceedings.mlr.press/v146/hu21a.html.

Hu, Y., Su, F., Dong, K., Wang, X., Zhao, X., Jiang, Y., Li, J., Ji, J., and Sun, Y. (2021b). Deep learning system for lymph node quantification and metastatic cancer identification from whole-slide pathology images. *Gastric Cancer*, 24(4):868–877. http://doi.org/10.1007/s10120-021-01158-9.

Hu, Z., Wang, J., Sun, D., Cui, L., and Ran, W. (2019). How Many Cores Does Systematic Prostate Biopsy Need?: A Large-Sample Retrospective Analysis. *Journal of Ultrasound in Medicine*, 38:1491–1499. http://doi.org/10.1002/jum.14834.

Huang, G., Liu, Z., van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 2261–2269. http://doi.org/10.1109/CVPR.2017.243.

Huang, R., Geng, A., and Li, Y. (2021a). On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, pages 677–689. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/063e26c670d07bb7c4d30e6fc69fe056-Abstract.html.

Huang, S., Chaudhary, K., and Garmire, L. X. (2021b). More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 11. http://doi.org/10.3389/fgene.2017.00084.

Huang, W., Randhawa, R., Jain, P., Hubbard, S., Eickhoff, J., Kummar, S., Wilding, G., Basu, H., and Roy, R. (2022). A novel artificial intelligence–powered method for prediction of early recurrence of prostate cancer after prostatectomy and cancer drivers. *JCO Clinical Cancer Informatics*. http://doi.org/10.1200/CCI.21.00131.

Ikromjanov, K., Bhattacharjee, S., Hwang, Y.-B., Sumon, R. I., Kim, H.-C., and Choi, H.-K. (2022). Whole slide image analysis and detection of prostate cancer using vision transformers. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 399–402. IEEE. http://doi.org/10.1109/ICAIIC54071.2022.9722635.

Ilse, M., Tomczak, J. M., and Welling, M. (2018). Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*. https://proceedings.mlr.press/v80/ilse18a.html.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on Machine Learning – ICML 2015*, pages 448–456. https://dl.acm.org/doi/10.5555/3045118.3045167.

Janowczyk, A., Basavanhally, A., and Madabhushi, A. (2017). Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. *Computerized Medical Imaging and Graphics*, 57:51–60. http://doi.org/10.1016/j.compmedimag.2016.05.003.

Jiang, S., Suriawinata, A. A., and Hassanpour, S. (2021). MHAttnSurv : Multihead attention for survival prediction using whole-slide pathology images. *arxiv preprint*. https://arxiv.org/abs/2110.11558.

Jimenez-del Toro, O., Atzori, M., Otálora, S., Andersson, M., Eurén, K., Hedlund, M., Rönnquist, P., and Müller, H. (2017). Convolutional neural networks for an automatic classification of prostate tissue slides with high-grade gleason score. *Medical Imaging 2017: Digital Pathology*, 10140(March):101400O. http://doi.org/10.1117/12.2255710.

John, J., Ravikumar, A., and Abraham, B. (2021). Prostate cancer prediction from multiple pretrained computer vision model. *Health and Technology*. http://doi.org/10.1007/s12553-021-00586-y.

Kamarudin, A. N., Cox, T., and Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: Current methods and applications. *BMC Medical Research Methodology*, 17(53). http://doi.org/10.1186/s12874-017-0332-6.

Kamran, F. and Wiens, J. (2021). Estimating calibrated individualized survival curves with deep learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, volume 35, pages 240–248. https://doi.org/10.1609/aaai.v35i1.16098.

Kanwal, N., Pérez-Bueno, F., Schmidt, A., Engan, K., and Molina, R. (2022). The devil is in the details: Whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: A review. *IEEE Access*, 10:58821–58844. https://doi.org/10.1109/ACCESS.2022.3176091.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481. http://www.jstor.org/stable/2281868.

Karimi, D., Nir, G., Fazli, L., Black, P. C., Goldenberg, L., and Salcudean, S. E. (2020). Deep learning-based gleason grading of prostate cancer from histopathology images - role of multiscale decision aggregation and data augmentation. *IEEE Journal of Biomedical and Health Informatics*, 24(5):1413–1426. https://doi.org/10.1109/jbhi.2019.2944643.

Kattan, M. W., Wheeler, T. M., and Scardino, P. T. (1999). Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of Clinical Oncology*, 17(5):1499–1507. https://doi.org/10.1200/jco.1999.17.5.1499.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(24):1–12. https://doi.org/10.1186/s12874-018-0482-1.

Kim, D. (1999). Normalization methods for input and output vectors in backpropagation neural networks. *International Journal of Computer Mathematics*, 71(2):161–171. https://doi.org/10.1080/00207169908804800.

Kiyokawa, H., Abe, M., Matsui, T., Kurashige, M., Ohshima, K., Tahara, S., Nojima, S., Ogino, T., Sekido, Y., Mizushima, T., and Morii, E. (2022). Deep learning analysis of histologic images from intestinal specimen reveals adipocyte shrinkage and mast cell infiltration to predict postoperative Crohn disease. *The American Journal of Pathology*. https://doi.org/10.1016/j.ajpath.2022.03.006.

Kleinbaum, D. G. and Klein, M. (2012). *Survival Analysis - A Self-Learning Text*. Springer New York, 3 edition. https://doi.org/10.1007/978-1-4419-6646-9.

Kleppe, A., Skrede, O.-J., De Raedt, S., Liestøl, K., Kerr, D. J., and Danielsen, H. E. (2021). Designing deep learning studies in cancer diagnostics. *Nature Reviews Cancer*, 21:199–211. http://doi.org/10.1038/s41568-020-00327-9.

Kohli, M. D., Summers, R. M., and Geis, J. R. (2017). Medical image data and datasets in the era of machine learning—whitepaper from the 2016 C-MIMI meeting dataset session. *Journal of Digital Imaging*, 30:392–399. http://doi.org/10.1007/s10278-017-9976-3.

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O.-P. (1998).

Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4(7):884–847. `http://doi.org/10.1038/nm0798-844`.

Korevaar, S., Tennakoon, R., Page, M., Brotchie, P., Thangarajah, J., Florescu, C., Sutherland, T., Kam, N. M., Bab-Hadiashar, A., and Bab-Hadiashar, A. (2021). Incidental detection of prostate cancer with computed tomography scans. *Nature Scientific Reports*, 11(7956). `http://doi.org/10.1038/s41598-021-86972-y`.

Koziarski, M., Cyganek, B., Olborski, B., Antosz, Z., Żydak, M., Kwolek, B., Wąsowicz, P., Bukała, A., Swadźba, J., and Sitkowski, P. (2021). DiagSet: a dataset for prostate cancer histopathological image classification. *arxiv preprint*. `http://arxiv.org/abs/2105.04014`.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. `https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html`.

Kumar, N., Qi, S.-a., Kuan, L.-H., Sun, W., Zhang, J., and Greiner, R. (2022). Learning accurate personalized survival models for predicting hospital discharge and mortality of COVID-19 patients. *Nature Scientific Reports*, 12(4472). `https://doi.org/10.1038/s41598-022-08601-6`.

Kumar, N., Verma, R., Arora, A., Kumar, A., Gupta, S., Sethi, A., and Gann, P. H. (2017). Convolutional neural networks for prostate cancer recurrence prediction. In *Medical Imaging 2017: Digital Pathology*, volume 10140, pages 106 – 117. `http://doi.org/10.1117/12.2255774`.

Kvamme, H. and Borgan, Ø. (2021). Continuous and discrete-time survival prediction with neural networks. *Lifetime Data Analysis*, 27(4):710–736. `https://doi.org/10.1007/s10985-021-09532-6`.

Kvamme, H., Borgan, Ø., and Scheel, I. (2019). Time-to-event prediction with neural networks and cox regression. *Journal of Machine Learning Research*, 20:1–30. `https://jmlr.org/papers/v20/18-424.html`.

Laleh, N. G., Echle, A., Muti, H. S., Hewitt, K. J., Schulz, V., and Kather, J. N. (2021). Deep learning for interpretable end-to-end survival prediction in gastrointestinal cancer histopathology. In *COMPAY 2021: The third MICCAI* *workshop on Computational Pathology*, pages 1–15. https://proceedings.mlr.press/v156/ghaffari-laleh21a.html.

Lambert, J. and Chevret, S. (2016). Summary measure of discrimination in survival models based on cumulative/dynamic time-dependent roc curves. *Statistical Methods in Medical Research*, 25(5):2088–2102. https://doi.org/10.1177/0962280213515571.

Lang, G. (2006). *Histotechnik: Praxislehrbuch für die Biomedizinische Analytik*. Springer Vienna, Vienna. https:doi.org/10.1007/3-211-33142-5.

Lee, K., Lee, K., Lee, H., and Shin, J. (2018). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177. https://papers.nips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html.

Leon, F. and Martinez, F. (2021). Learning a triplet embedding distance to represent gleason patterns. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 3229–3232. https://doi.org/10.1109/EMBC46164.2021.9630755.

Li, H., Boimel, P., Janopaul-Naylor, J., Zhong, H., Xiao, Y., Ben-Josef, E., and Fan, Y. (2019). Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 846–849. IEEE. https://doi.org/10.1109/ISBI.2019.8759301.

Li, H., Han, D., Hou, Y., Chen, H., and Chen, Z. (2015). Statistical inference methods for two crossing survival curves: A comparison of methods. *PLoS ONE*, 10(1). https://doi.org/10.1371/journal.pone.0116774.

Li, J., Li, W., Sisk, A., Ye, H., Wallace, W. D., Speier, W., and Arnold, C. W. (2021). A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine*, 131. https://doi.org/10.1016/j.compbiomed.2021.104253.

Li, R., Wu, X., Li, A., and Wang, M. (2022). HFBSurv : Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics online preprint*, pages 1–9. https://doi.org/10.1093/bioinformatics/btac113.

Li, R., Yao, J., Zhu, X., Li, Y., and Huang, J. (2018). Graph CNN for survival analysis on whole slide pathological images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 174 – 182. Springer International Publishing. https://doi.org/10.1007/978-3-030-00934-2_20.

Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=H1VGkIxRZ.

Liu, H. and Kurc, T. (2022). Deep learning for survival analysis in breast cancer with whole slide image data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btac381.

Liu, J., Chen, Y., Yan, J., Zhang, Z., Zhang, H., and Li, Z.-C. (2022). Risk attention network : Weakly-supervised learning for joint tumor segmentation. In *Digital TV and Wireless Multimedia Communications*, pages 96–107. Springer Singapore. https://doi.org/10.1007/978-981-19-2266-4_8.

Lobel, B. (2007). Does localized prostate cancer exist? In *Prostate Cancer*. Springer Berlin Heidelberg New York. https://doi.org/10.1007/978-3-540-40901-4_7.

Loeb, S., Bjurlin, M. A., Nicholson, J., Tammela, T. L., Penson, D. F., Carter, H. B., Carroll, P., and Etzioni, R. (2014). Overdiagnosis and overtreatment of prostate cancer. *European Urology*, 65(6):1046–1055. https://doi.org/10.1016/j.eururo.2013.12.062.

Lombardo, E., Kurz, C., Marschner, S., Avanzo, M., Gagliardi, V., Fanetti, G., Franchin, G., Stancanello, J., Corradini, S., Niyazi, M., Belka, C., Parodi, K., Riboldi, M., and Landry, G. (2021). Distant metastasis time to event analysis with CNNs in independent head and neck cancer cohorts. *Nature Scientific Reports*, 11(6418):1–12. https://doi.org/10.1038/s41598-021-85671-y.

Longato, E., Vettoretti, M., and Di Camillo, B. (2020). A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models. *Journal of Biomedical Informatics*, 108:103496. https://doi.org/10.1016/j.jbi.2020.103496.

Lu, M. Y., Chen, R. J., Wang, J., Dillon, D., and Mahmood, F. (2019). Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding. *arxiv preprint*. http://arxiv.org/abs/1910.10825.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570. https://doi.org/10.1038/s41551-020-00682-w.

Luiting, H. B. and Roobol, M. J. (2019). Prostatakrebs-Früherkennung: Stand und Evidenz der Methoden. In *Versorgungs-Report Früherkennung*, pages 147–164. Medizinisch Wissenschaftliche Verlagsgesellschaft. https://doi.org/10.32745/9783954664023-10.

Lv, Z., Lin, Y., Yan, R., Wang, Y., and Zhang, F. (2022). TransSurv: Transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–10. https://doi.org/10.1109/TCBB.2022.3199244.

Ma, J. (2021). Histogram matching augmentation for domain adaptation with application to multi-centre, multi-vendor and multi-disease cardiac image segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC*, pages 177–186. https://doi.org/10.1007/978-3-030-68107-4_18.

Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., Schmitt, C., and Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. *IEEE International Symposium on Biomedical Imaging*, pages 1107–1110. https://doi.org/10.1109/ISBI.2009.5193250.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., and Yuille, A. (2015). Deep captioning with multimodal recurrent neural networks (m-RNN). In *3rd International Conference on Learning Representations, ICLR 2015*. https://arxiv.org/abs/1412.6632.

Marginean, F., Arvidsson, I., Simoulis, A., Overgaard, N. C., Åström, K., Heyden, A., Bjartell, A., and Krzyzanowska, A. (2021). An artificial intelligence–based support tool for automation and standardisation of gleason grading in prostate biopsies. *European Urology Focus*, 7(5):995–1001. https://doi.org/10.1016/j.euf.2020.11.001.

Marini, N., Atzori, M., Otálora, S., Marchand-Maillet, S., and Müller, H. (2021a). H&E-adversarial network: A convolutional neural network to learn stain-invariant features through hematoxylin eosin regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 601–610. https://doi.org/10.1109/ICCVW54120.2021.00073.

Marini, N., Otálora, S., Müller, H., and Atzori, M. (2021b). Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. *Medical Image Analysis*, 73. https://doi.org/10.1016/j.media.2021.102165.

May, W. L. (2017). Kaplan-meier survival analysis. In *Encyclopedia of Cancer*, pages 2383–2386. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-46875-3_3196.

Mazaheri, Y., Hricak, H., Fine, S. W., Akin, O., Shukla-Dave, A., Ishill, N. M., Moskowitz, C. S., Grater, J. E., Reuter, V. E., Zakian, K. L., Touijer, K. A., and Koutcher, J. A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252(2):449–457. https://doi.org/10.1148/radiol.2523081423.

McHugh, M. L. (2012). Interrater reliability : the kappa statistic. *Biochemica Medica*, 22(3):276–282. https://doi.org/10.11613/BM.2012.031.

Mescher, A. L. (2013). *Junqueira's Basic Histology: Text and Atlas*, volume 1. McGraw-Hill Education. https://www.worldcat.org/title/junqueiras-basic-histology-text-and-atlas/oclc/820107038.

Meyer, J., Khademi, A., Têtu, B., Han, W., Nippak, P., and Remisch, D. (2022). Impact of artificial intelligence on pathologists' decisions: an experiment. *Journal of the American Medical Informatics Association*, pages 1–8. https://doi.org/10.1093/jamia/ocac103.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J., and Cooper, L. A. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(13):E2970–E2979. https://doi.org/10.1073/pnas.1717139115.

Mohseni, S., Pitale, M., Yadawa, J., and Wang, Z. (2020). Self-supervised learning for generalizable out-of-distribution detection. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 5216–5223. https://doi.org/10.1609/aaai.v34i04.5966.

Montagnon, E., Cerny, M., Cadrin-Chênevert, A., Hamilton, V., Derennes, T., Ilinca, A., Vandenbroucke-Menu, F., Turcotte, S., Kadoury, S., and Tang, A. (2020). Deep learning workflow in radiology: a primer. *Insights into Imaging*, 11(22). https://doi.org/10.1186/s13244-019-0832-5.

Muhammad, H., Xie, C., Sigel, C. S., Doukas, M., Alpert, L., and Fuchs, T. J. (2021). EPIC-survival: end-to-end part inferred clustering for survival analysis, featuring prognostic stratification boosting. In *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, pages 520–531. https://proceedings.mlr.press/v143/muhammad21a.html.

Mun, Y., Paik, I., Shin, S.-J., Kwak, T.-Y., and Chang, H. (2021). Yet another automated gleason grading system (YAAGGS) by weakly supervised deep learning. *npj Digital Medicine*, 4(99). https://doi.org/10.1038/s41746-021-00469-6.

Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., Wren, J. H., Corrado, G. S., MacDonald, R., Peng, L. H., Amin, M. B., Evans, A. J., Sangoi, A. R., Mermel, C. H., Hipp, J. D., and Stumpe, M. C. (2019). Development and validation of a deep learning algorithm for improving gleason scoring of prostate cancer. *npj Digital Medicine*, 2(1):1–10. https://doi.org/10.1038/s41746-019-0112-2.

Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P. H. C., Steiner, D. F., Manoj, N., Olson, N., Smith, J. L., Mohtashamian, A., Peterson, B., Amin, M. B., Evans, A. J., Sweet, J. W., Cheung, C., van der Kwast, T., Sangoi, A. R., Zhou, M., Allan, R., Humphrey, P. A., Hipp, J. D., Gadepalli, K., Corrado, G. S., Peng, L. H., Stumpe, M. C., and Mermel, C. H. (2020). Development and validation of a deep learning algorithm for gleason grading of prostate cancer from biopsy specimens. *JAMA Oncology*, 6(9):1372–1380. https://doi.org/10.1001/jamaoncol.2020.2485.

Nam, J. G., Kang, H.-R., Lee, S. M., Kim, H., Rhee, C., Goo, J. M., Oh, Y.-M., Lee, C.-H., and Park, C. M. (2022). Deep learning prediction of survival in patients with chronic obstructive pulmonary disease using chest radiographs. *Radiology*, 000. https://doi.org/10.1148/radiol.212071.

Nam, S. J., Chong, Y., Jung, C. K., Kwak, T.-Y., Lee, J. Y., Park, J., Rho, M. J., and Go, H. (2021). Preference and demand for digital pathology and computer-aided diagnosis among korean pathologists: A survey study focused on prostate needle biopsy. *Applied Sciences*, 11(16). https://doi.org/10.3390/app11167380.

Nguyen, A., Yosinski, J., and Clune, J. (2015). Deep neural networks are easily fooled: High confidence prediction for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436. https://doi.org/10.1109/CVPR.2015.7298640.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B. F., Tavassoli, P., Turbin, D., Villamil, C. F., Wang, G., Wilson, R. S., Iczkowski, K. A., Lucia, M. S., Black, P. C., Abolmaesumi, P., Goldenberg, S. L., and Salcudean, S. E. (2018). Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. *Medical Image Analysis*, 50:167–180. https://doi.org/10.1016/j.media.2018.09.005.

Novikov, A. (2019). PyClustering: Data mining library. *Journal of Open Source Software*, 4(36):1230. https://doi.org/10.21105/joss.01230.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., Riordan, D., and Walsh, J. (2020). Deep learning vs. traditional computer vision. *Advances in Intelligent Systems and Computing*, 943(April):128–144. https://doi.org/10.1007/978-3-030-17795-9_10.

Oner, M. U., Ng, M. Y., Giron, D. M., Chen Xi, C. E., Yuan Xiang, L. A., Singh, M., Yu, W., Sung, W.-K., Wong, C. F., and Lee, H. K. (2022). An AI-assisted tool for efficient prostate cancer diagnosis. *bioRxiv preprint*, pages 1–14. https://www.biorxiv.org/content/10.1101/2022.02.06.479283v1.

Otálora, S., Marini, N., Müller, H., and Atzori, M. (2021). Combining weakly and strongly supervised learning improves strong supervision in gleason pattern classification. *BMC Medical Imaging*, 21(77):1–14. https://doi.org/10.1186/s12880-021-00609-0.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076.

Parsons, M. and Grabsch, H. (2009). How to make tissue microarrays. *Diagnostic Histopathology*, 15(3):142–150. https://doi.org/10.1016/j.mpdhp.2009.01.010.

Patil, A., Tamboli, D., Meena, S., Anand, D., and Sethi, A. (2019). Breast cancer histopathology image classification and localization using multiple instance learning. In *IEEE International WIE Conference on Electrical and Computer Engineering, WIECON-ECE*, pages 1–4. IEEE. https://doi.org/10.1109/WIECON-ECE48653.2019.9019916.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. https://scikit-learn.org.

Pinckaers, H., van Ipenburg, J., Melamed, J., De Marzo, A., Platz, E. A., van Ginneken, B., van der Laak, J., and Litjens, G. (2022). Predicting biochemical recurrence of prostate cancer with artificial intelligence. *Nature Communications Medicine*, 2(64). https://doi.org/10.1038/s43856-022-00126-3.

Popescu, D. M., Shade, J. K., Lai, C., Aronis, K. N., Ouyang, D., Moorthy, M. V., Cook, N. R., Lee, D. C., Kadish, A., Albert, C. M., Wu, K. C., Maggioni, M., and Trayanova, N. A. (2022). Arrhythmic sudden death survival prediction using deep learning analysis of scarring in the heart. *Nature Cardiovascular Research*, 1:334 – 343. https://doi.org/10.1038/s44161-022-00041-9.

Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*. https://papers.nips.cc/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.

Rajaganesan, S., Kumar, R., Rao, V., Pai, T., Mittal, N., Sahay, A., Menon, S., and Desai, S. (2021). Comparative assessment of digital pathology systems for primary diagnosis. *Journal of Pathology Informatics*, 12(25). https://doi.org/10.4103/jpi.jpi_94_20.

Rakha, E. A., Toss, M., Shiino, S., Gamble, P., Jaroensri, R., Mermel, C. H., and Chen, P.-H. C. (2021). Current and future applications of artificial intelligence in pathology: A clinical perspective. *Journal of Clinical Pathology*, 74:409–414. https://doi.org/10.1136/jclinpath-2020-206908.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. (2019). Stand-alone self-attention in vision models. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, volume 32. https://papers.nips.cc/paper/2019/hash/3416a75f4cea9109507cacd8e2f2aefc-Abstract.html.

Rawla, P. (2019). Epidemiology of Prostate Cancer. *World Journal of Oncology*, 10(2):63–89. https://doi.org/10.1159/000423644.

Reinke, A., Tizabi, M. D., Sudre, C. H., Eisenmann, M., Rädsch, T., Baumgartner, M., Acion, L., Antonelli, M., Arbel, T., Bakas, S., Bankhead, P., Benis, A., Cardoso, M. J., Cheplygina, V., Christodoulou, E., Cimini, B., Collins, G. S., Farahani, K., van Ginneken, B., Glocker, B., Godau, P., Hamprecht, F., Hashimoto, D. A., Heckmann-Nötzel, D., Hoffman, M. M., Huisman, M., Isensee, F., Jannin, P., Kahn, C. E., Karargyris, A., Karthikesalingam, A., Kainz, B., Kavur, E., Kenngott, H., Kleesiek, J., Kooi, T., Kozubek, M., Kreshuk, A., Kurc, T., Landman, B. A., Litjens, G., Madani, A., Maier-Hein, K., Martel, A. L., Mattson, P., Meijering, E., Menze, B., Moher, D., Moons, K. G. M., Müller, H., Nichyporuk, B., Nickel, F., Noyan, M. A., Petersen, J., Polat, G., Rajpoot, N., Reyes, M., Rieke, N., Riegler, M., Rivaz, H., Saez-Rodriguez, J., Gutierrez, C. S., Schroeter, J., Saha, A., Shetty, S., van Smeden, M., Stieltjes, B., Summers, R. M., Taha, A. A., Tsaftaris, S. A., Van Calster, B., Varoquaux, G., Wiesenfarth, M., Yaniv, Z. R., Kopp-Schneider, A., Jäger, P., and Maier-Hein, L. (2021). Common limitations of image processing metrics: A picture story. *arxiv preprint*, 1. http://arxiv.org/abs/2104.05642.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. (2021). A simple fix to mahalanobis distance for improving near-OOD detection. In *ICML 2021 Uncertainty and Robustness in Deep Learning, workshop*. https://icml.cc/Conferences/2021/ScheduleMultitrack?event=11964.

Ren, J., Hacihaliloglu, I., Singer, E. A., Foran, D. J., and Qi, X. (2018a). Adversarial domain adaptation for classification of prostate histopathology whole-slide images. In *Medical Image Computing and Computer Assisted*

Intervention – MICCAI 2018, pages 201–209. https://doi.org/10.1007/978-3-030-00934-2_23.

Ren, J., Hacihaliloglu, I., Singer, E. A., Foran, D. J., and Qi, X. (2019a). Unsupervised domain adaptation for classification of histopathology whole-slide images. *Frontiers in Bioengineering and Biotechnology*, 7(102). https://doi.org/10.3389/fbioe.2019.00102.

Ren, J., Karagoz, K., Gatza, M. L., Singer, E. A., Sadimin, E., Foran, D. J., and Qi, X. (2018b). Recurrence analysis on prostate cancer patients with Gleason score 7 using integrated histopathology whole-slide images and genomic data through deep neural networks. *Journal of Medical Imaging*, 5(4). https://doi.org/10.1117/1.JMI.5.4.047501.

Ren, K., Qin, J., Zheng, L., Yang, Z., Zhang, W., Qiu, L., and Yu, Y. (2019b). Deep recurrent survival analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:4798–4805. https://doi.org/10.1609/aaai.v33i01.33014798.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “Why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, page 1135–1144, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., and Wang, E. W. (2010). A practical guide to understanding kaplan-meier curves. *Otolaryngology - Head and Neck Surgery*, 143(3):331–336. https://doi.org/10.1016/j.otohns.2010.05.007.

RKI (2021). *Krebs in Deutschland für 2017/2018*, volume 13. Zentrum für Krebsregisterdaten, Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V. . https://www.krebsdaten.de/Krebs/DE/Content/Publikationen/Krebs_in_Deutschland/krebs_in_deutschland_inhalt.html.

Rodriguez, G. (2007). Lecture notes on generalized linear models. https://data.princeton.edu/wws509/notes/ (last accessed October 28, 2022).

Rolls, G. O., Farmer, N. J., and Hall, J. B. (2008). *Artifacts in Histological and Cytological Preparations.* Scientia - Leica Microsystems’ Education Series. http://dp000393.ferozo.com/Bibliografia/Leica%20.Artefactos%20en%20histo%20y%20cito%20preparaciones.pdf.

Rondorf-Klym, L. M. and Colling, J. (2003). Quality of life after radical prostatectomy. *Oncology nursing forum*, 30(2):24–32. https://doi.org/10.1188/03.ONF.E24-E32.

Roy, A. G., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., Vo, N., Bui, P., Winter, S., MacWilliams, P., Corrado, G. S., Telang, U., Liu, Y., Cemgil, T., Karthikesalingam, A., Lakshminarayanan, B., and Winkens, J. (2022). Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75. https://doi.org/10.1016/j.media.2021.102274.

Rubinstein, R. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology And Computing In Applied Probability*, 1(2):127–190. https://doi.org/10.1023/A:1010091220143.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, volume Vol. 1, pages 318–362. MIT Press. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. In *ImageNet Large Scale Visual Recognition Challenge, IJCV, 2015*, pages 211–252. https://doi.org/10.1007/s11263-015-0816-y.

Rymarczyk, D., Borowa, A., Tabor, J., and Zieliński, B. (2021). Kernel self-attention in deep multiple instance learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1721–1730. https://doi.org/10.1109/WACV48630.2021.00176.

Salman, M. E., Çakirsoy Çakar, G., Azimjonov, J., Kösem, M., and Cedimoğlu, I. H. (2022). Automated prostate cancer grading and diagnosis system using deep learning-based yolo object detection algorithm. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2022.117148.

Sandeman, K., Blom, S., Koponen, V., Manninen, A., Juhila, J., Rannikko, A., Ropponen, T., and Mirtti, T. (2022). AI model for prostate biopsies predicts cancer survival. *Diagnostics*, 12(5). `https://doi.org/10.3390/diagnostics12051031`.

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., and Schlomm, T. (2018). Integrating tertiary gleason 5 patterns into quantitative Gleason grading in prostate biopsies and prostatectomy specimens. *European Urology*, 73(5):674–683. `https://doi.org/10.1016/j.eururo.2017.01.015`.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., Denis, L. J., Recker, F., Berenguer, A., Määttänen, L., Bangma, C. H., Aus, G., Villers, A., Rebillard, X., van der Kwast, T., Blijenberg, B. G., Moss, S. M., de Koning, H. J., and Auvinen, A. (2009). Screening and prostate-cancer mortality in a randomized european study. *New England Journal of Medicine*, 360(13):1320–1328. `https://doi.org/10.1056/NEJMoa0810084`.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM : Visual explanations from deep networks. *IEEE International Conference on Computer Vision - ICCV 2017*, pages 618–626. `https://doi.org/10.1109/ICCV.2017.74`.

Shakhawat, H. M., Nakamura, T., Kimura, F., Yagi, Y., and Yamaguchi, M. (2020). Automatic quality evaluation of whole slide images for the practical use of whole slide imaging scanner. *ITE Transactions on Media Technology and Applications*, 8(4):252–268. `https://doi.org/10.3169/MTA.8.252`.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., and Zhang, Y. (2021). TransMIL: Transformer based correlated multiple instance learning for whole slide image classication. *Advances in Neural Information Processing Systems*, pages 1–14. `https://papers.nips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html`.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(60). https://doi.org/10.1186/s40537-019-0197-0.

Silva-Rodríguez, J., Colomer, A., Sales, M. A., Molina, R., and Naranjo, V. (2020). Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer Methods and Programs in Biomedicine*, 195. https://doi.org/10.1016/j.cmpb.2020.105637.

Simon, R., Mirlacher, M., and Sauter, G. (2004). Tissue microarrays. *BioTechniques*, 36(jan):98–105. https://doi.org/10.2144/04361rv01.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations - ICLR 2015*, pages 1–14. https://arxiv.org/abs/1409.1556.

Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(52):1–19. https://doi.org/10.3390/JIMAGING6060052.

Singhal, N., Soni, S., Bonthu, S., Chattopadhyay, N., Samanta, P., Joshi, U., Jojera, A., Chharchhodawala, T., Agarwal, A., Desai, M., and Ganpule, A. (2022). A deep learning system for prostate cancer diagnosis and grading in whole slide images of core needle biopsies. *Nature Scientific Reports*, 12(3383). https://doi.org/10.1038/s41598-022-07217-0.

Sloma, M., Syed, F. J., Nemati, M., and Xu, K. S. (2021). Empirical comparison of continuous and discrete-time representations for survival prediction. In *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*, volume 2021. https://proceedings.mlr.press/v146/sloma21a.html.

Sohn, E. (2015). Screening: Diagnostic dilemma. *Nature*, 528:2014. https://doi.org/10.1038/528S120a.

Somogyi, Z. (2021). *The Application of Artificial Intelligence*. Springer Cham. https://doi.org/10.1007/978-3-030-60032-7.

Sotelo, R. (2015). Introduction. In *Prostate Cancer - A Patient's Guide*, pages 1–5. Springer Cham Heidelberg New York Dordrecht London. https://doi.org/10.1007/978-3-319-05600-5.

Stamey, T. A., McNeal, J. E., Yemoto, C. M., Sigal, B. M., and Johnstone, I. M. (1999). Biological determinants of cancer progression in men with prostate cancer. *Jama*, 281(15):1395–1400. https://doi.org/10.1001/jama.281.15.1395.

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., Iczkowski, K. A., Kench, J. G., Kristiansen, G., van der Kwast, T. H., Leite, K. R. M., McKenney, J. K., Oxley, J., Pan, C.-C., Samaratunga, H., Srigley, J. R., Takahashi, H., Tsuzuki, T., Varma, M., Zhou, M., Lindberg, J., Lindskog, C., Ruusuvuori, P., Wählby, C., Grönberg, H., Rantalainen, M., Egevad, L., and Eklund, M. (2020). Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology*, 21(2):222–232. https://doi.org/10.1016/S1470-2045(19)30738-7.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. *Proceedings of the 39th International Conference on Machine Learning, PMLR*, (162). http://arxiv.org/abs/2204.06507.

Suresh, K., Severn, C., and Ghosh, D. (2022). Survival prediction models: an introduction to discrete-time modeling. *BMC Medical Research Methodology*, 22(207). https://doi.org/10.1186/s12874-022-01679-6.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA. https://doi.org/10.1109/CVPR.2016.308.

Tabibu, S., Vinod, P. K., and Jawahar, C. V. (2019). Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Nature Scientific Reports*, 9(10509). https://doi.org/10.1038/s41598-019-46718-3.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learn-*

ing - ICML 2019, pages 10691–10700. https://proceedings.mlr.press/v97/tan19a.html.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J. M., Ciompi, F., and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58. https://doi.org/10.1016/j.media.2019.101544.

Thebille, A. K., Dietrich, E., Klaus, M., Gernhold, L., Lennartz, M., Kuppe, C., Kramann, R., Huber, T. B., Sauter, G., Puelles, V. G., Zimmermann, M., and Bonn, S. (2021). Deep learning-based bias transfer for overcoming laboratory differences of microscopic images. In *Medical Image Understanding and Analysis*, volume 12722, pages 322–336. Springer International Publishing. https://doi.org/10.1007/978-3-030-80432-9_25.

Tizhoosh, H. R. and Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics*, 9(38). https://doi.org/10.4103/jpi.jpi_53_18.

Tolkach, Y., Dohmgörgen, T., Toma, M., and Kristiansen, G. (2020). High-accuracy prostate cancer pathology using deep learning. *Nature Machine Intelligence*, 2(7):411–418. https://doi.org/10.1038/s42256-020-0200-7.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., Paoli, J., Puig, S., Rosendahl, C., Soyer, H. P., Zalaudek, I., and Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234. https://doi.org/10.1038/s41591-020-0942-0.

Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102(478):527–537. https://doi.org/10.1198/016214507000000149.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N. (2016). Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Transactions on Medical Imaging*, 35(8):1962–1971. https://doi.org/10.1109/TMI.2016.2529665.

Vale-Silva, L. A. and Rohr, K. (2021). Long-term cancer survival prediction using multimodal deep learning. *Nature Scientific Reports*, 11(1):1–12. https://doi.org/10.1038/s41598-021-92799-4.

van den Oord, A., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arxiv preprint*. http://arxiv.org/abs/1807.03748.

van der Slot, M. A., Hollemans, E., den Bakker, M. A., Hoedemaeker, R., Kliffen, M., Budel, L. M., Goemaere, N. N., and van Leenders, G. J. (2021). Interobserver variability of cribriform architecture and percent Gleason pattern 4 in prostate cancer: relation to clinical outcome. *Virchows Archiv*, 478:249–256. https://doi.org/10.1007/s00428-020-02902-9.

van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., Yu, T., and the scikit-image contributors (2014). scikit-image: image processing in Python. *PeerJ*, 2:e453. https://doi.org/10.7717/peerj.453, https://scikit-image.org/.

van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA. https://www.python.org/.

Varsavsky, T., Orbes-Arteaga, M., Sudre, C. H., Graham, M. S., Nachev, P., and Cardoso, M. J. (2020). Test-time unsupervised domain adaptation. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 428–436. https://doi.org/10.1007/978-3-030-59710-8_42.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems - NIPS 2017*, pages 5999–6009. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Verdhan, V. (2021). *Computer Vision Using Deep Learning.* Apress, Berkeley, CA, 1 edition. https://doi.org/10.1007/978-1-4842-6616-8.

Vilone, G. and Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3:615–661. https://doi.org/10.3390/make3030032.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272. `https://doi.org/10.1038/s41592-019-0686-2` `https://scipy.org/`.

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., and O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, 61:119–131. `https://doi.org/10.1016/j.jbi.2016.03.009`.

Vollmer, R. T. (2009). Percentage of tumor in prostatectomy specimens a study of american veterans. *American Journal of Clinical Pathology*, 131:86–91. `https://doi.org/10.1309/AJCPX5MAMNMFE6FQ`.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018. `https://doi.org/10.1155/2018/7068349`.

Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606.

Vuong, T. T. L., Kim, K., Song, B., and Kwak, J. T. (2021). Joint categorical and ordinal learning for cancer grading in pathology images. *Medical Image Analysis*, 73. `https://doi.org/10.1016/j.media.2021.102206`.

Walhagen, P., Bengtsson, E., Lennartz, M., Sauter, G., and Busch, C. (2022). AI-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples. *Journal of Pathology Informatics*. `https://doi.org/10.1016/j.jpi.2022.100137`.

Wang, P., Li, Y., and Reddy, C. K. (2019). Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6). `https://doi.org/10.1145/3214306`.

Wang, R., Huang, Z., Wang, H., and Wu, H. (2021). AMMASurv: Asymmetrical multi-modal attention for accurate survival analysis with whole slide images and gene expression data. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021*. `https://doi.org/10.1109/BIBM52615.2021.9669382`.

Wang, S., Liu, Z., Chen, X., Zhu, Y., Zhou, H., Tang, Z., Wei, W., Dong, D., Wang, M., and Tian, J. (2018). Unsupervised deep learning features for lung cancer overall survival analysis. In *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2583–2586. `https://doi.org/10.1109/EMBC.2018.8512833`.

Williams, B. J., Lee, J., Oien, K. A., and Treanor, D. (2018). Digital pathology access and usage in the UK: Results from a national survey on behalf of the national cancer research institute's CM-path initiative. *Journal of Clinical Pathology*, 71(5):463–466. `https://doi.org/10.1136/jclinpath-2017-204808`.

Wright, A. I., Dunn, C. M., Hale, M., Hutchins, G. G. A., and Treanor, D. E. (2021). The effect of quality control on accuracy of digital pathology image analysis. *IEEE Journal of Biomedical and Health Informatics*, 25(2):307 – 314. `https://doi.org/10.1109/JBHI.2020.3046094`.

Wulczyn, E., Steiner, D. F., Xu, Z., Sadhwani, A., Wang, H., Flament-Auvigne, I., Mermel, C. H., Chen, P.-H. C. C., Liu, Y., and Stumpe, M. C. (2020). Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS ONE*, 15(6):1–18. `https://doi.org/10.1371/journal.pone.0233678`.

Xiao, L., Yu, J.-G., Liu, Z., Ou, J., Deng, S., Yang, Z., and Li, Y. (2020). Censoring-aware deep ordinal regression for survival prediction from pathological images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 449–458. `https://doi.org/10.1007/978-3-030-59722-1_43`.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., and Chang, E. I.-C. (2017). Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics*, 18(281). `https://doi.org/10.1186/s12859-017-1685-x`.

Yala, A., Mikhael, P. G., Strand, F., Lin, G., Smith, K., Wan, Y. L., Lamb, L., Hughes, K., Lehman, C., and Barzilay, R. (2021). Toward robust mammography-based models for breast cancer risk. *Science Translational Medicine*, 13. https://doi.org/10.1126/scitranslmed.aba4373.

Yamamoto, Y., Tsuzuki, T., Akatsuka, J., Ueki, M., Morikawa, H., Numata, Y., Takahara, T., Tsuyuki, T., Tsutsumi, K., Nakazawa, R., Shimizu, A., Maeda, I., Tsuchiya, S., Kanno, H., Kondo, Y., Fukumoto, M., Tamiya, G., Ueda, N., and Kimura, G. (2019). Automated acquisition of explainable knowledge from unannotated histopathology images. *Nature Communications*, 10(1). https://doi.org/10.1038/s41467-019-13647-8.

Yan, G. and Greene, T. (2008). Investigating the effects of ties on measures of concordance. *Statistics in Medicine*, 27:4190 – 4206. https://doi.org/10.1002/sim.3257.

Yan, X., Zhang, H., Xu, X., Hu, X., and Heng, P.-A. (2021). Learning semantic context from normal samples for unsupervised anomaly detection. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v35i4.16420.

Yang, J., Chen, J., Kuang, K., Lin, T., He, J., and Ni, B. (2020). MIA -prognosis : A deep learning framework to predict therapy response. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 4, pages 211–220. Springer International Publishing. https://doi.org/10.1007/978-3-030-59713-9_21.

Yang, J., Zhou, K., Li, Y., and Liu, Z. (2021). Generalized out-of-distribution detection: A survey. *arxiv preprint*, pages 1–20. http://arxiv.org/abs/2110.11334.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., and Huang, J. (2020). Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65. https://doi.org/10.1016/j.media.2020.101789.

Yuan, X. and Rai, S. N. (2011). Confidence intervals for survival probabilities: A comparison study. *Communications in Statistics: Simulation and Computation*, 40(7):978–991. https://doi.org/10.1080/03610918.2011.560732.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR.2015.7299101.

Zeng, X.-D., Chao, S., and Wong, F. (2010). Optimization of bagging classifiers based on SBCB algorithm. In *Proceedings of the Ninth International Conference on Machine Learning and Cybernetics*, number Jul, pages 262–267. https://doi.org/10.1109/ICMLC.2010.5581054.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. (2021a). *Dive into Deep Learning*. https://d2l.ai/ (last accessed October 19, 2022).

Zhang, J., Ma, K., Van Arnam, J., Gupta, R., Saltz, J., Vakalopoulou, M., and Samaras, D. (2021b). A joint spatial and magnification based attention framework for large scale histopathology classification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 3771–3779. https://doi.org/10.1109/CVPRW53098.2021.00418.

Zhang, Z., Beck, M. W., Winkler, D. A., Huang, B., Sibanda, W., and Goyal, H. (2018). Opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Annals of Translational Medicine*, 6(11):216–216. https://doi.org/10.21037/atm.2018.05.32.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929. IEEE. https://doi.org/10.1109/CVPR.2016.319.

Zhou, T., Fu, H., Zhang, Y., Zhang, C., Lu, X., Shen, J., and Shao, L. (2020). M^{2}Net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 221–231. https://doi.org/10.1007/978-3-030-59713-9_22.

Zhou, Y., Onder, O. F., Dou, Q., Tsougenis, E., Chen, H., and Heng, P.-A. (2019). CIA-Net: Robust nuclei instance segmentation with contour-aware information aggregation. In *Information Processing in Medical Imaging*, pages 682–693. Springer International Publishing. https://doi.org/10.1007/978-3-030-20351-1_53.

Zhu, X., Yao, J., and Huang, J. (2016). Deep convolutional neural network for survival analysis with pathological images. In *2016 IEEE International*

Conference on Bioinformatics and Biomedicine (BIBM), pages 544–547. IEEE. https://doi.org/10.1109/BIBM.2016.7822579.

Zhu, X., Yao, J., Zhu, F., and Huang, J. (2017). WSISA: Making survival prediction from whole slide histopathological images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7234–7242. https://doi.org/10.1109/CVPR.2017.725.

Zuley, M. L., Jarosz, R., Drake, B. F., Rancilio, D., Klim, A., Rieger-Christ, K., and Lemmerman, J. (2016). The cancer genome atlas prostate adenocarcinoma collection (tcga-prad) (version 4) [dataset]. the cancer imaging archive. https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=6884022.

Eidesstattliche Versicherung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, den ____________ Unterschrift ______________________________