Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Robust and trustworthy deep learning-based disease detection and risk assessment on MRI and histopathological images that exceeds predictive performance of human experts

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Informatik
der Universität Hamburg
vorgelegt von Fabian Westhäußer
Hamburg, Mai 2024

Vorsitzender der Prüfungskommission

Prof. Dr. Jan Baumbach

GutachterInnen

Prof. Dr. Stefan Bonn
Prof. Dr. Simone Frintrop
Prof. Dr. Sören Laue

Datum der Disputation

08.10.2024

Abstract

AI-based models for medical image analysis show great potential to assist medical practitioners in their clinical practice, improving diagnostic accuracy, efficiency, and ultimately patient outcome. However, being applicable in a sensitive field such as healthcare comes with several specific challenges and requirements that many existing AI-based models do not meet. These encompass, besides others, a performance that matches those of the human practitioners, robustness to data variance and trustworthiness. Human-level performance builds the foundational incentive to employ an AI-based system for clinical decision support. However, this performance needs to be reliable when confronted with high data variance in clinical practice, caused by factors such as different processing protocols or acquisition devices. Many current models struggle to generalize to data outside of their known training distribution. Finally, practitioners and patients need to be able to put trust into model predictions. While a high predictive accuracy on a validation cohort provides statistical foundation to the models aptitude, the black box character inherent to deep learning models makes interpretation and assessment of individual cases difficult. To this end, this thesis proposes two deep learning models in the field of medical image processing that aim to tackle the previously mentioned requirements of clinical applicability, namely human-level performance, robustness and trustworthiness.

The first model, DeePSC, is a convolutional neural network-based ensemble classifier that detects primary schlerosing cholangitis, an autoimmune liver disease, on magnetic resonance images. It is specifically designed to process MRI images taken from seven angular views around the patient. By comparing against four experienced radiologists, it is shown that DeePSC outperforms the average human rater on two datasets acquired with different magnetic field strengths by 5.5 and 10.3 percentage points in terms of accuracy. Robustness is assessed by showing high predictive accuracy of 92.4% on an external validation cohort acquired at a different scanner device. To further mitigate the black box character of the network and build trust in its predictions, GradCAM activation mappings are employed, which reveal salient regions in the input images in the biologically relevant areas of the biliary tree. Lastly, an extensive technical analysis on multiple aggregation strategies to combine information of the seven angular images of the MRI data is conducted.

The second model, PCAI, is an end-to-end risk prediction network that quantifies the aggressiveness of prostate cancer and the associated risk of patients based on

histopathological microscopy images of prostate tissue. It is trained on one of the largest and most heterogeneous histopathological prostate datasets collected to date, encompassing six cohorts with over 25,591 patients, 83,864 images, and five years of median follow-up from five different centers and three countries. This heterogeneity is utilized by training PCAI in a domain adversarial fashion on digitized tissue microarray spots extracted after operative removal of the prostate. By including further algorithmic extensions such as credibility estimation, color adaptation and cancer indication, PCAI outperforms a separately trained baseline model on eight highly variant internal and external datasets, proving its robustness to distribution shifts encountered in clinical practice. PCAI further systematically outperforms ISUP annotations of multiple highly skilled human experts, which represents the current gold-standard for evaluating cancer aggressiveness, on an unseen spot dataset as well as two pre-operative biopsy datasets by up to 22.3 percentage points in terms of concordance index. Especially the high predictive accuracy on the latter is of great clinical relevance. Finally, the model quantifies its confidence in a prediction by a separate credibility score and highlights cancerous regions on the input images for potential re-evaluation, aiming to build trust and interpetability.

By performing a thorough analysis and evaluation of both proposed models with respect to the initially defined requirements and their aptitude as a clinical decision support system, this thesis aims to contribute to the state of the art of deep learning-based medical image analysis and provide a potential blueprint for decision support systems in clinic practice.

Zusammenfassung

KI-basierte Modelle für die medizinische Bildanalyse weisen ein beträchtliches Potenzial auf, Ärzte in der klinischen Praxis zu unterstützen, die diagnostische Genauigkeit, die Effizienz und letztendlich den Therapieerfolg der Patienten zu verbessern. Die Anwendbarkeit in einem sensiblen Bereich wie dem Gesundheitswesen ist jedoch mit verschiedenen spezfifischen Herausforderungen und Anforderungen verbunden, denen viele KI-basierte Modelle in der Literatur noch nicht gerecht werden. Dazu gehören u. a. eine prädiktive Genaugigkeit, die der des menschlichen Arztes entspricht, Robustheit gegenüber Varianz in den verarbeiteten Daten und Vertrauenswürdigkeit. Die Leistung auf menschlichem Niveau stellt den grundlegenden Anreiz für den Einsatz eines KI-basierten Systems zur klinischen Entscheidungsunterstützungs dar. Die Zuverlässigkeit der prädiktiven Genauigkeit muss jedoch gewährleistet sein, wenn die KI mit Daten konfrontiert wird, welche die in der klinischen Praxis auftretende Varianz widerspiegeln, die durch Faktoren wie unterschiedliche Verarbeitungsprotokolle oder Aufnahmegeräte verursacht wird. Viele KI-basierte Modelle in der Literatur haben Probleme mit der Generalisierbarkeit auf Daten außerhalb ihrer bekannten Trainingsverteilung. Darüber hinaus müssen Ärzte und Patienten in der Lage sein, den Vorhersagen des Modells zu vertrauen. Eine hohe Vorhersagegenauigkeit in einer Validierungskohorte stellt zwar eine statistische Grundlage für die prinzipielle Eignung des Modells dar, jedoch erschwert der Black-Box-Charakter von Deep-Learning-Modellen deren Interpretation und Bewertung von Einzelfällen. Die vorliegende Arbeit präsentiert zwei Deep-Learning-Modelle im Bereich der medizinischen Bildverarbeitung, welche die zuvor genannten Anforderungen an die klinische Anwendbarkeit erfüllen sollen. Dazu zählen die prädiktive Genauigkeit auf menschlichem Niveau, Robustheit und Vertrauenswürdigkeit.

Das erste Modell, DeePSC, ist ein Ensemble-Klassifikator auf der Basis eines Convolutional Neural Networks, der primär sklerondierende Cholangitis, eine Autoimmunerkrankung der Leber, auf Magnetresonanztomographiebildern erkennt und speziell für die Verarbeitung von MRT-Bildern aus sieben Winkelansichten um den Patienten herum konzipiert ist. Ein Vergleich mit vier erfahrenen Radiologen zeigt, dass DePSC bei zwei Datensätzen, die bei unterschiedlichen Magnetfeldstärken aufgenommen wurden, eine um 5.5 und 10.3 Prozentpunkte höhere Genauigkeit aufweist als der durchschnittliche menschliche Experte. Die Robustheit wird durch eine hohe Vorhersagegenauigkeit auf einer externen Validierungskohorte bewiesen, die mit einem

anderen Scannergerät erfasst wurde. Um den Black-Box-Charakter des Netzwerks zu reduzieren und Vertrauen in die Vorhersagen zu schaffen, werden GradCAM-Aktivierungsmappings verwendet, die auffällige Regionen in den Eingabebildern in den biologisch relevanten Bereichen der Gallenwege aufzeigen. Schließlich wird eine umfassende technische Analyse mehrerer Aggregationsstrategien zur Kombination von Informationen aus den sieben Winkelbildern der MRT-Daten durchgeführt.

Das zweite Modell PCAI ist ein End-to-End-Risikovorhersagenetzwerk, welches die Aggressivität von Prostatakrebs und das damit verbundene Risiko von Patienten auf der Grundlage von mikroskopischen Bildern von Prostatagewebe quantifiziert. Es wird auf einem der größten und heterogensten histopathologischen Prostatadatensätze trainiert, die bis heute gesammelt wurden. Dieser umfasst sechs Kohorten mit über 25.591 Patienten, 83.864 Bildern und einem durchschnittlichen Follow-up von fünf Jahren aus fünf verschiedenen Zentren und drei Ländern. Die Heterogenität des Datensatzes wird mithilfe eines Domain-Adversarial Trainingsregime von PCAI mit digitalisierten Spots von Multigewebeblöcken genutzt, die nach der operativen Entfernung der Prostata prozessiert wurden. Mithilfe weiterer algorithmischer Erweiterungen wie Credibility-Schätzung, Farbanpassung und Krebsindikation übertrifft PCAI ein separat trainiertes Referenzmodell auf acht hochgradig variierenden internen und externen Datensätzen und demonstriert damit seine Robustheit gegenüber Varianz in den verarbeiteten Daten, die in der klinischen Praxis auftritt. Darüber hinaus übertrifft PCAI die ISUP-Annotationen mehrerer hochqualifizierter menschlicher Experten, welche den derzeitigen Goldstandard für die Bewertung der Krebsaggressivität darstellen, sowohl auf einem ungesehenen Spot-Datensatz als auch auf zwei präoperativen Biopsiedatensätzen um bis zu 22.3 Prozentpunkte in Bezug auf den Concordance-Index. Insbesondere die hohe Vorhersagegenauigkeit bei letzterem ist von großer klinischer Relevanz. Darüber quantifiziert das Modell seine Konfidenz in eine Vorhersage durch einen separaten Credibility-Score und hebt karzinomatöse Regionen in den Eingabebildern für eine potenzielle Neubewertung hervor, um Vertrauen und Interpretierbarkeit aufzubauen.

Die vorliegende Arbeit zielt darauf ab, durch eine detaillierte Analyse und Bewertung der beiden entwickelten Modelle im Hinblick auf die eingangs definierten Anforderungen sowie deren Eignung als klinisches Entscheidungsunterstützungssystem einen Beitrag zum aktuellen Stand der Technik von Deep Learning basierten Systemen zur medizinischen Bildanalyse zu leisten. Zudem soll sie als Blaupause für künftige Arbeiten dienen, die letztlich ihren Weg in die klinische Praxis finden.

Acknowledgements

Firstly, I want to thank Prof. Dr. Stefan Bonn and Prof. Dr. Simone Frintrop for their supervision, support and guidance and for providing me with the opportunity to pursue a doctoral program.

Further, I want to thank all members of the IMSB for the great working environment. I really enjoyed my time here. I'm especially grateful to Marina Zimmermann, Patrick Fuhlert, Nico Kaiser, Anne Ernst, Esther Dietrich, Robin Khatri, and all other members of the Image Analysis team for their invaluable expertise, inspiration and support, without which this research would not have been possible and that enabled me to learn and grow both professionally and personally.

I sincerely thank all my friends for constantly enriching my life outside of work and for making my time in Hamburg and this PhD journey as delightful as it is.

Finally, I want to express my gratitude to my family, especially my parents, for their love and support and for providing me with everything necessary that lead up to achieving this goal.

Contents

1 Introduction

1.1 Motivation

The application of artificial intelligence (AI) in numerous aspects of society gained exponential momentum in recent years, positively impacting overall productivity and wellbeing [1, 2]. In the specific field of medical image analysis, AI tools have already proved successful in improving efficiency, diagnostic accuracy and ultimately patient outcome when implemented as a so-called clinical decision support (CDS) system [3, 4]. This is mainly made possible by mitigating the influence of various shortcomings in human evaluation of medical images, such as a high subjectivity and inter-rater variability, low reproducibility and increased demands on time and financial resources [5, 6]. Besides helping to overcome those shortcomings, AI-based CDS systems also include the potential to pave the way towards personalized medicine, provide more accurate treatment decisions to patients or even discover previously unknown predictive features of disease progression [3, 7–9]. However, the momentum of implementation of AI-based systems in the medical image processing field lags behind that of AI applications in other sectors, such as finance, media or communication [10, 11]. This is due to the increased requirements machine learning models have to fulfill to be properly applicable in a sensitive field such as healthcare and medicine, where wrongful decisions could have potentially lethal or life changing consequences.

These requirements encompass, besides others, human-level predictive performance, robustness to bias in the data and trustworthiness [11–14]. Human-level performance provides the foundation for a useful CDS – if it performs worse than the clinician, this drastically mitigates its potential usefulness [15]. Current research reveals that clinicians expect AI-based systems to even be superior to the average performing specialist [16]. While many machine learning applications in the literature already reach performance comparable to that of medical experts, their predictive accuracy

degrades significantly when confronted with unseen data expressing a covariate shift that reflects the variance encountered in everyday clinical practice [12, 17]. Rendering AI models robust to this heterogeneity in the data is pivotal when going beyond research and aiming for actual clinical application. Besides that, building trust in a model's prediction is of utmost importance, not only from the perspective of the clinician, but also for the patients [14, 18–20]. Medical doctors can describe their thought process that lead to the final conclusion. Machine learning models often resemble a "black box", that only provides a single prediction value without further information or legitimation [11, 21]. However, how certain can the reader be that this predicted value is valid? One way to increase trustworthiness of those models is therefore to equip them with the necessary means to quantify their confidence in a given prediction, not unlike a human expert that is uncertain about the grading of a particular sample and asks for a second opinion [21]. A model that assesses the confidence in its predictions by identifying problematic input samples that it cannot provide a reliable prediction on can try to fix those samples or defer them for human re-evaluation [18]. Closely related to building trust are the topics of interpretability and explainability. If a model can transparently highlight to the human reader how it came to a conclusion based on the input data, this increases overall trust and even provides the potential of discovering previously unknown predictive features of disease progression, especially if the underlying model is more accurate than currently used clinical protocols [14, 22]. A straightforward way to increase interpretability of those models is by highlighting salient regions in the images that correlate with the final prediction, for example by activation mapping or by highlighting diseased regions [23, 24].

Most research in the field of medical computer vision today still falls short of fulfilling the necessary criteria described above [10, 12]. To this end, this thesis covers two separate projects in the field of medical image processing that aim towards application as a CDS system and are specifically design to tackle the previously mentioned challenges encountered in clinical practice, namely human-level performance, robustness, and trustworthiness.

The first project presents a convolutional neural network-based ensemble classifier for detection of primary sclerosing cholangitis, an autoimmune liver disease, on images acquired by magnetic resonance imaging (MRI). This model, named DeePSC, is specifically designed to process MRI images taken from different angular views around the patient. Predictive accuracy is compared against that of four experienced radiologists.

The second project presents an end-to-end risk prediction network for quantification of the aggressiveness of prostate cancer and the associated risk of patients based on histopathological microscopy images of prostate tissue. While being trained on images acquired from tissue of surgically removed prostates, it allows for inference on significantly larger images of clinically more relevant pre-operative biopsies. The predictive value of the proposed risk score, called the Prostate Cancer Aggressiveness Index (PCAI), is thoroughly evaluated against human assigned grading systems that are currently in clinical use.

1.2 Research Questions

The main research questions addressed in both projects deal with the initially introduced key requirements of clinical applicability. These are:

- **RQ-1:** Is a deep learning model trained on the respective included dataset(s) able to provide a predictive accuracy in its respective task of disease detection or aggressiveness quantification that matches that of human experts?
- **RQ-2:** How does data variance encountered in clinical practice affect such a model and what are suitable algorithmic adaptations to improve generalization capabilities on data outside its training distribution?
- **RQ-3:** What are suitable algorithmic adaptations to increase trustworthiness of the proposed models, including the topics of explainability and interpretability?

Further, a model-specific research question is addressed in each project, namely:

- **RQ-DeePSC:** Does the available information of multiple views in the MRI input data improve performance of the proposed model when assessed in parallel beyond information of a single image? If so, what are suitable methods to aggregate this information?
- **RQ-PCAI:** To what extent can a deep learning model trained on post-operative tissue microarray spot images provide meaningful assessment of cancer aggressiveness on pre-operative biopsy images?

1.3 Thesis Outline

This thesis covers two projects in the field of medical image processing with deep learning, called DeePSC and PCAI, with a focus on clinical applicability, including robustness, trustworthiness and benchmarking against human raters. Both projects are described independently in separate parts of this thesis, which are structured as follows:

First, an introduction into the respective motivation and project aim is given. Then, the necessary medical and technical foundations are described. Next, the used datasets are introduced in detail. The following section then dives deeper into data preprocessing and the experimental setup. The subsequent methods section provides a thorough description of the proposed network architecture, algorithmic extensions as well as the evaluation metrics and procedure. Then, a systematic overview of all conducted experiments with their respective results is given. In the following chapter, the experimental results are thoroughly discussed, with a specific focus on the initially defined requirements for clinical applicability. Lastly, a conclusion, which also addresses the previously stated research questions, is drawn. DeePSC includes an additional chapter describing a technical evaluation study for image aggregation strategies prior to the project's conclusion.

Finally, the overall conclusion summarizes both projects, puts them into context, critically discusses them with respect to the state of the art and their aptitude as a clinical decision support system and provides recommendations for future work to build and extend on the findings of this thesis.

2 Detection of Primary Sclerosing Cholangitis with Deep Learning - DeePSC

2.1 Introduction

Primary sclerosing cholangitis (PSC) is a chronic cholestatic liver disease characterized by progressive multifocal bile duct strictures due to biliary inflammation and fibrosis [25]. This rare liver disorder is often associated with inflammatory bowel disease (IBD) and is considered a premalignant condition, since patients typically are at high risk of developing colorectal and hepatobiliary malignancies [25, 26]. The only curative therapy remains liver transplantation with a disease recurrence rate of up to 25 % [26, 27].

Magnetic resonance imaging (MRI) including cholangiopancreatography (MRCP) has been established as the noninvasive imaging modality of choice for the detection of PSC-compatible bile duct changes and disease related complications [28]. However, reading MRCP scans is subjective, requires expertise and experience, and is often time consuming when it comes to subtle findings. Furthermore, the interpretation of MRCP varies even among PSC experts and experienced radiologists and often shows only poor inter-reader agreement between serial follow-up examinations of patients with PSC [5, 6]. This highlights the need and opportunity for clinical decision support through automated evaluation of MRCP to improve diagnosis of the disease.

Deep learning approaches like convolutional neural networks (CNN) are currently seeing huge advances in medical imaging [4, 29]. Traditionally, these models are trained and validated on single images. Recent research, however, shows that the predictive accuracy of classification tasks can be significantly improved by creating a joint decision from multiple images of the same object of interest [30]. Since MRCP scans usually include several images, each taken from different projections covering the intra- and extrahepatic bile duct system, a multi-view classification network that aggregates information of all views per patient appears to be promising for this task.

A common issue with many existing deep learning applications in the medical field, especially in medical image processing, is the lack of generalizability, where the model performs well on the single dataset it is trained on, but fails to reproduce this performance on previously unseen data. To provide actual decision support in a clinical setting beyond research, this generalizability is fundamental.

The main goal of this work is therefore to verify the aptitude of an AI-based clinical decision support system for the automated classification of PSC-compatible cholangiographic findings on 2D-MRCP by developing a deep-learning model, measuring

its performance across different datasets and putting the results into clinical context. Since, to the best of my knowledge, this is the first study to apply deep-learning methods to 2D-MRCP data, the question of whether a robust automated diagnosis on this type of data is possible with high accuracy is of particular interest. If so, there is opportunity for the model to pave the way to more accurate diagnoses, reducing radiologists' workload, and mitigating inter-reader variability. Specifically, in the course of this work, an end-to-end deep multi-view convolutional neural network ensemble model (DeePSC) is developed and evaluated on 2D-MRCP datasets obtained at magnetic field strengths of 1.5 and 3 Tesla on different MRI scanners by different manufacturers to ensure generalizability. Furthermore, the performance of the classification network is compared to that of four radiologists with varying levels of experience in reading MRCP. Lastly, explainability measures are applied to visualize and verify the model's predictions.

2.2 Foundations

In this chapter, the medical foundations of primary sclerosing cholangitis (PSC) and the corresponding imaging modality used for diagnosis of the disease, magnetic resonance cholangiopancreatography (MRCP), are provided. Furthermore, convolutional neural networks (CNN), the specific SqueezeNet CNN architecture used in this work, as well as gradient-weighted class activation mappings (GradCAM) as a method to increase interpretability of CNNs, are described.

2.2.1 Primary Sclerosing Cholangitis (PSC)

Primary sclerosing cholangitis (PSC) is a chronic cholestatic liver disease characterized by progressive multifocal bile duct strictures due to biliary inflammation and fibrosis [25]. The bile duct scarring narrows the ducts of the biliary tree and impedes the flow of bile to the intestines. Eventually, this leads to bile duct fibrosis, multifocal strictures, cholestasis, and biliary cirrhosis and can result in cirrhosis of the liver and liver failure [25]. This rare liver disorder is often associated with inflammatory bowel disease (IBD) and is considered a premalignant condition, since patients typically are at high risk of developing colorectal and hepatobiliary malignancies [25, 26]. The disease is more prevalent in men (63%) and diagnosed on average at age 54, but can also affect much younger patients [31]. The only curative therapy remains liver transplantation with a disease recurrence rate of up to 25 % [26, 27]. The pathogenesis of PSC is still not fully understood, but it is believed to be a complex interplay between genetic and environmental factors [32]. The diagnosis of PSC is based on clinical, biochemical, radiological, and histological findings, though the gold standard for diagnosis is magnetic resonance cholangiopancreatography (MRCP) [33]. However, the diagnosis of PSC is often challenging due to the heterogeneity of the disease and the lack of specific diagnostic criteria. The disease course is highly variable, ranging from asymptomatic to rapidly progressive disease, with a median survival of 12 years after diagnosis [34]. The only approved medical therapy is ursodeoxycholic acid, which has been shown to improve liver biochemistry, but not patient survival [35]. Therefore, there is an urgent need for new therapeutic approaches.

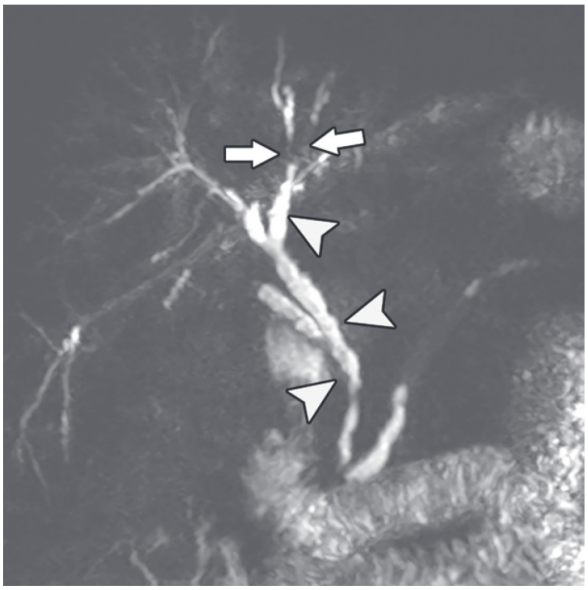

Figure 2.2.1: Beaded appearance of the bile ducts in a 62-year-old man with PSC. Maximum intensity projection from coronal thin-section MRCP shows intra- and extrahepatic bile duct beading (arrowheads). Note the abrupt cutoff (arrows) owing to intrahepatic bile duct strictures. Reprinted with permission from Khoshpouri et al. [36].

2.2.2 Magnetic Resonance Cholangiopancreatography (MRCP)

Magnetic Resonance Cholangiopancreatography (MRCP) is a non-invasive Magnetic Resonance Imaging (MRI) sequence that is specialized at visualizing the biliary and pancreatic ducts [37]. In contrast to other MRI sequences, it is able to generate detailed images of the bile ducts without requiring contrast agents or invasive procedures [37]. Due to its ability to accurately depict the characteristic biliary changes such as multifocal strictures, dilatations, and beading occuring in patients with PSC, it emerged as the non-invasive gold standard for diagnosis of the disease [36]. Figure 2.2.1 depicts an exemplary MRCP image in which the characteristic beaded appearance of the bile ducts in a patient with PSC is visible and highlighted. MRCP can be differentiated in the two main techniques 2D-MRCP and 3D-MRCP [38]. 2D-MRCP acquires images in thin slices in a single plane. By adjusting the orientation of the imaging plane, different aspects of the ductal anatomy can be visualized, resulting in multiple images for a single examination. 3D-MRCP acquires volumetric datasets covering the entire biliary and pancreatic system. It provides higher spatial resolution, but requires longer acquisition times, making it more vulnerable to motion arti-

facts [39]. Besides those techniques, several imaging parameters, including magnetic field strength (e.g., 1.5 Tesla or 3 Tesla), sequence types (e.g., single-shot fast spin-echo, heavily T2-weighted sequences), and contrast enhancement techniques, can be varied depending on the clinical indication and scanner capabilities. A higher magnetic field strength during acquisition is associated with improved signal-to-noise-ratio, spatial resolution, and contrast, whereas the penetration depth in the examined tissue is reduced [40].

2.2.3 Convolutional Neural Networks (CNNs)

A convolutional neural network (CNN) is a special type of neural network that is mostly used for image processing [41]. In contrast to the fully connected layers of classical multi-layer perceptrons, CNNs utilize so called convolutional layers. The learnable weights in these layers are located in filters or kernels of variable size, which are iteratively moved over the incoming image or multi-dimensional feature map to perform the convolution operation at every step. This results in an inherent equivariance to translations in the input, since kernels apply their same weights at every location of the input matrix [42]. The resulting output feature maps for every individual kernel of a convolutional layer are concatenated in the channel dimension. Convolution operations are applied simultaneously to all entries of the incoming channel dimension. The edges of the incoming features maps can be padded to allow for equal number of steps in subsequent layers. The stride parameter defines the step width of the kernel in the spatial dimension. The output size W_{out} of each convolutional layer is derived from the input size W_{in} as

$$W_{out} = \left(\frac{W_{in} - K + 2P}{S} \right) + 1 \tag{2.2.1}$$

where K is the kernel size, P the padding size and S the stride [43]. It is further common practice to apply pooling operations like mean or max pooling after a certain number of convolutional layers in a CNN [41]. These are used for downsampling of the incoming feature map, which increases the so called "field of view" of subsequent convolutional layers at a given kernel size, allowing for extraction of larger and higher level features.

2.2.4 SqueezeNet

SqueezeNet is a CNN architecture specifically aimed for high performance at a very low number of trainable parameters [44]. This makes this architecture especially suitable for applications with limited number of computational power or memory, like e.g. embedded deployment. Additionally, models with a lower number of trainable parameters are better suited for training on datasets with a limited number of samples or level of complexity. The SqueezeNet architecture achieved classification accuracy on the ImageNet dataset similar to the at that time commonly used AlexNet architecture, even though it utilizes a 50-fold reduced number of trainable parameters [45, 46]. This is based on three main design strategies:

- Strategy 1: Replace 3x3 filters with 1x1 filters.
- Strategy 2: Decrease the number of input channels to 3x3 filters.
- Strategy 3: Downsample late in the network so that convolution layers have large activation maps.

The authors proposed the Fire module as their main building block for SqueezeNet, which allowed them to employ above mentioned design strategies. These consist of a *squeeze* convolutional that consists only of 1x1 filters, which aims at reducing the number of input channels to the subsequent *expand* layer (Strategy 2). This includes a mix of 1x1 and 3x3 filters, where the liberal use of 1x1 filters is an application of Strategy 1. The full SqueezeNet architecture includes a convolutional layer at the input, 8 subsequent Fire modules and a final convolutional layer at the output. Max-pooling with stride 2 is performed after the first convolutional layer, the forth and eighth Fire module, as well as the final convolutional layer. The authors claim that those relatively late placements of the pooling layers employ Strategy 3.

2.2.5 Gradient-weighted Class Activation Mapping (GradCAM)

Convolutional neural networks normally resemble a "black box", that only provides a single predicted value without further information or legitimation on how the conclusion was derived [47]. A commonly used approach in the literature to mitigate this intransparency is to increase interpretability by applying Gradient-weighted Class Activation Mapping (GradCAM) [48] to a trained convolutional deep learning model. GradCAM is a method to visualize and localize class-discriminative regions of the in-

put images, i.e. to highlight which regions of the input image contributed most to the overall class prediction. It builds on the Class Activation Mapping (CAM) proposed by Zhou et al. [49] by generalizing their method and making it applicable to a wider range of CNN-based architectures. CAM specifically requires the underlying model to feed global average pooled (GAP) convolutional feature maps directly into a softmax activation, which is not the case in the proposed deep learning models in this work. GradCAM leverages two underlying concepts of deep CNNs. Firstly, deeper representations in a CNN capture higher-level visual constructs, and secondly, convolutional layers retain spatial information which is lost in fully connected layers. With this, the authors expected the last convolutional layers of a CNN to represent the best compromise between high-level semantics and detailed spatial information and the neurons/kernels of those layers to be strongly correlated with semantic class-specific information. Their method then utilizes the gradient information in the final convolutional layer to assign importance values to each neuron with respect to the predicted output class [48]. In detail, to derive the class-discriminative localization map $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ of width u and height v for a given class c, the gradient $\frac{\partial y^c}{\partial A^k}$ of the score y^c for that same class c with respect to the feature map activations A^k of all K neurons of the last convolutional layer is computed. Then, global average pooling is applied over the spatial dimensions to derive the neuron importance weights a^c_k as

$$\alpha^c_k = \overbrace{\frac{1}{u \cdot v} \sum_i \sum_j}^{\text{GAP}} \frac{\partial y^c}{\partial A^k_{ij}}. \tag{2.2.2}$$

Here, i and j are indices for the width and height dimension of the feature map A^k, respectively, and k the index of the neuron in the layer. Since computation of the gradient $\frac{\partial y^c}{\partial A^k}$ requires backpropagation through all subsequent layers of the network, the authors claimed that a^c_k represents a partial linearization of the deep network downstream from the last convolutional layer and therefore captures the 'importance' of the $k - th$ neuron for a target class c. The class-discriminative localization map $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ is then derived as the weighted sum of feature map activations and neuron importance weights as

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right). \tag{2.2.3}$$

Rectified Linear Unit (ReLU) activation is applied to only highlight features that have a positive correlation with the class of interest. The resolution of $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ refers to that of the feature maps in the last convolutional layer. The localization map is finally upscaled to the resolution of the input image and depicted as an overlay to highlight salient regions.

2.3 State of the Art

This chapter describes the current state-of-the-art of deep learning applications on magnetic resonance imaging (MRI) and magnetic resonance cholangiopancreatography (MRCP) data. Then, an overview of machine learning applications in the field of primary sclerosing cholangitis (PSC) is provided. Finally, a review of the literature on machine learning methods for "multi-view learning", referring to the processing of combined information of images from different perspectives of the same object, is delineated.

2.3.1 Deep Learning in Magnetic Resonance Imaging

Deep Learning in magnetic resonance imaging (MRI) is a very active field of research with a broad variety of downstream tasks. One prominent application is the segmentation of brain tumors [50]. Ranjbarzadeh et al. proposed a CNN network that utilizes four combined MRI modalities to achieve state-of-the-art performance on the Brain Tumor Segmentation Challenge 2018 (BraTS18) dataset [51, 52]. They introduced a distance-wise attention mechanism to account for the expected spatial location of the tumor in the brain. In the field of prostate cancer, Pellicer-Valero et al. combined the tasks of detection, segmentation and grading in multiparametric MRI in a single system [53]. They utilized an extended Retina U-Net with ResNet101 as backbone to achieve competitive performance on all three tasks [54, 55]. However, they noted that even though their model was trained on two different data sources, optimum thresholds varied significantly between both domains, highlighting the robustness issue of deep learning models. Another important use case of deep learning is motion artifact correction in MRI. Hossbach et al. proposed the Motion Parameter Estimating Densenet (MoPED) that predicts the best parameters for a subsequent data-consistent correction algorithm [56]. With this method, they avoid the risk of introducing hallucinations, which is otherwise inherent to deep learning based image-to-image translation approaches.

On the specific modality of magnetic resonance cholangiopancreatography MRCP images, research is less abundant. Luo et al. used submodels of YOLOv5 to automatically diagnose common bile duct stones in thick-slab MRCP images [57, 58]. Their method achieved an accuracy of 90.5% on the test data, whereas human raters achieved an accuracy of 92.1%. Muneeswaran et al. proposed a neural network-based

framework for automatic gallbladder segmentation on low contrast MRCP images [59]. They compared to and improved over commonly used network architectures like ResNet50, however, their methodology and results are intransparent. Najjar et al. recently published a review study about AI integration in the clinical practice of radiology [60] . They found that AI already plays a pivotal role, improving diagnostic accuracy, workflow efficiency and personalised patient care. However, challenges about data privacy, security and missing trust due to the 'black box' nature of the algorithms remain to be addressed.

2.3.2 Machine Learning in Primary Sclerosing Cholangitis

While machine learning applications for MRI data in general are a broad and active field of research and ML-generated tools already outperformed conventional tools in predicting patient outcome in other liver diseases [61, 62], very little research has been conducted on ML applications in the specific area of PSC. Eaton et al. developed a PSC Risk Estimate Tool (PREsTo) [63] to predict hepatic decompensation using a gradient boosting model and medical tabular data. Andres et al. introduced a learning algorithm PSSP [64] (patient-specific survival prediction system) to predict patient survival after receiving a liver transplant based on multiple tabular parameters. Similarly, Hu et al. claim to have developed the first AI-based predictive model that performs significantly better than commonly used PSC risk scores [65]. They utilized clinical parameters from a cohort of 1,459 PSC patients with up to 27 years of follow-up. Singh et al. proposed an algebraic topology-based approach to extract vectorized feature representations from MRI images. They trained a decision tree classifier on those features to predict whether a patient will experience hepatic decompensation within one year after image acquisition. Their model achieved an area under the receiver operating characteristic curve (AUROC) of 0.84 on an independent validation cohort of 115 patients. Especially relevant is the work of Ringe et al. on fully automated detection of PSC-compatible bile duct changes in 3D-MRCP [66]. The authors transformed three-dimensional MRCP data into 20 two-dimensional images per patient using maximum intensity projection. An ImageNet-pretrained Inception-Resnet was trained on the single images [67]. Using majority voting over all 20 projections per patient during inference resulted in a very high sensitivity of 95.0% and specificity of 90.9%.

2.3.3 Multi-View Learning

CNNs form the state-of-the-art architectures for computer vision in machine learning. Traditionally, these networks train and evaluate on individual images. While this is successfully applied for various tasks, research shows that taking in and processing the combined information of images from different perspectives of the same object significantly improves predictive accuracy in object classification tasks [30]. The authors refer to this kind of task as "multi-view classification". Since the MRCP data used in this thesis consists of multiple images acquired at different rotational angles around the patient, combination of this multi-view information is expected to benefit overall predictive performance. In the multi-view classification literature, the applied method and position of merging information of multiple views, i.e. performing view-fusion, is a major topic of research. A pioneering work in this field is the multi-view CNN (MVCNN) by Su et al., a CNN that individually processes every view and combines the corresponding latent representations before the final classification layer [68]. Using rendered views of the computer-aided design (CAD) model dataset Modelnet and simple mean and max view pooling as their fusion method, they achieved up to 7% improved accuracy compared to prediction on single views only [69]. Inspired by the success the MVCNN, more sophisticated adaptions of the model using recurrent layers [70], self-attention [71] or feed-forward-fusion methods [72] were developed. Xu et al. [73] also introduced a multi-view-specific loss function. Application of multi-view networks in the field of medical imaging is rare. Here, both Geras et al. and Kaiser et al. proposed similar network architectures for multi-view processing of four mammographic images per patient [74, 75]. Geras et al. used their model for cancer detection, while Kaiser et al. predicted the breast density. In both networks, views are aggregated by concatenation of the latent feature vectors. Finally, although Seeland and Mäder [30] compare and analyse different view-fusion strategies on natural image datasets, to the best of my knowledge, there has been no comparative study performed on medical images to date. To this end, various view-fusion strategies for medical MRCP images will be explored and evaluated in the course of this work.

2.4 Data

This section introduces the patient cohort used in this work, together with the corresponding magnetic resonance cholangiopancreatography (MRCP) data and the protocol used for image acquisition.

2.4.1 Study Design and Data Acquisition

Figure 2.4.1: MRCP data of a patient with PSC taken at three different rotational angles at 3T. For better visibility, contrast-limited adaptive histogram equalization (CLAHE) [76] is applied.

In this work, a total of 897 patients that underwent 2D-MRCP scans at the University Medical Center Hamburg-Eppendorf (UKE) in the time between 2002 and 2022 are retrospectively included. Of these, 596 have a confirmed diagnosis of PSC. The reference standard for diagnosis was established non-invasively according to the guidelines of the European Association for the Study of the Liver (EASL), which are defined as follows: 1. elevated serum markers of cholestasis not otherwise explained, 2. characteristic bile duct changes with multifocal strictures and segmental dilatations visualized by MRCP and/or endoscopic retrograde cholangiopancreatography (ERCP), and 3. exclusion of causes of secondary sclerosing cholangitis and other cholestatic disorders [25]. The corresponding control group consists of 311 patients without any history of immune mediated liver or bile duct disease and without any visible bile duct alterations on MRCP. Table 2.4.1 depicts the indications of MRI for patients of in control group (after filtering, as will be described in Section 2.5.1). Patients and controls were identified via inhouse picture archiving and communication system (PACS) query. Ethical approval was provided by the institutional review board (2021-

100723-B0-ff). The requirement for written informed consent was waived due to the retrospective nature of the study.

Table 2.4.1: Indications for MRI in the included patients of the control group after filtering. NAFLD: Non-alcoholic fatty liver disease. IBD: Inflammatory bowel disease. Reprinted with permission from Ragab and Westhaeusser et al. [77].

Indication	$\mathbf{n = 264}$
Cystic Pancreatic Lesion	62
Pancreatitis	42
Unclear Abdominal Discomfort	42
Hepatic Lesion	39
Cholelithiasis	16
Pancreatic Mass	15
NAFLD	10
Pancreas Divisum	6
Post Distal Pancreatectomy	5
IBD	4
Exogenous Pancreatic Insufficiency	3
Gallbladder Polyp	3
Post Cholecystectomy	2
Other	15

For the 897 patients, a total of 999 MRI examinations including 2D-MRCP scans were collected. These were performed according to routine clinical protocol on different MRI machines from two manufacturers (Philips or Siemens) at either 3 or 1.5 Tesla magnetic field strength. Some patients underwent examination on more than one machine. The distribution of patients to the different MRI machines was based on availability only and has no relation to the medical history or diagnosis. Patients were advised to fast for 4 hours prior to the study to reduce fluid secretion within the gastrointestinal system. In patients without contraindication, 20 mg scopolamine butylbromide (Buscopan, Sanofi-Aventis, Frankfurt, Germany) was additionally intravenously injected to minimize gastrointestinal motility and thus motion artifacts while imaging. Each MRCP examination consists of seven to fourteen radial images from different angular points of view. Figure 2.4.1 depicts exemplary 2D-MRCP images from three different acquisition angles for a sample of 3T scanner. Detailed MRI protocol and imaging parameters are provided in A.1.1.

Table 2.4.2: Metadata and demographic characteristics of the internal (3 & 1.5 Tesla) and external vendor validation (3 Tesla Siemens) datasets used in DeePSC. Denoted as Median (Min - Max).

	3 Tesla		**1.5 Tesla**		**3 Tesla Siemens**	
Patients & MRCPs	361		398		37	
MRI Machines	Philips		Philips, Siemens		Siemens	
PSC/Non-PSC	189/172		283/115		20/17	
Training/Test	322/39		359/39		-/37	
	PSC	**Control**	**PSC**	**Control**	**PSC**	**Control**
Sex [female/male]	74/115	70/102	117/166	58/57	9/11	7/10
Age [years]	47 (16-81)	50 (18-82)	44 (9-79)	53 (15-80)	30 (17-60)	39 (24-71)
Weight [kg]	79 (44-125)	80 (45-165)	76 (46-130)	76 (45-117)	68 (45-104)	79 (58-97)
Image Acquisition Date	2019 (2009-2020)	2019 (2014-2021)	2018 (2002-2021)	2019 (2010-2020)	2021 (2021-2022)	2021 (2021-2022)

2.4.2 Dataset Definition

The total number of 897 patients and 999 MRCPs is separated into an internal dataset of 860 patients and 952 MRCPs and an independent validation cohort, which is referred to in the following as the external vendor validation dataset consisting of 47 patients and MRCPs (see Figure 2.5.1). All MRCPs in the external vendor validation set were collected at a Siemens MRI machine at 3T magnetic field strength, whereas in the internal dataset, all 3T examinations stem from a Philips MRI machine. The 1.5T examinations of the internal dataset were collected both at a Philips and Siemens MRI machine.

Internal and external datasets are further filtered, separated based on magnetic field strength used for image acquisition and split into training and test data according to criteria explained in the subsequent Section 2.5. Baseline demographic characteristics of the final included datasets are presented in Table 2.4.2. There are no significant differences with respect to age or sex when comparing PSC patients and controls. A more detailed visual breakdown of demographic and meta information of the internal datasets can be found in supplemental Figure A.1.1.

2.5 Preprocessing

In this section, the data preprocessing performed in this work, including dataset filtering and curation, separation into training and validation datasets, image normalization, as well as image classification by human experts, is described.

2.5.1 Filtering

Figure 2.5.1 depicts the dataflow from the initially collected cohorts to finally included patients and image datasets. In the internal cohort, 566 patients have an established diagnosis of PSC and 294 are controls without any history of immune mediated liver or bile duct disease and without any visible bile duct alterations on 2D-MRCP. Of these 566 PSC patients, 27 with a diagnosis of small duct PSC and 67 with indefinable PSC are excluded. 115 PSC patients and 2 controls are retrospectively excluded due to incomplete or qualitatively insufficient clinical and MRI data. With this, 649 patients (357 PSC / 292 controls) are subsequently divided into MRI exams obtained with a magnetic field strength of 3 Tesla and 1.5 Tesla. To reduce complexity, these exams are further filtered to only include MRCP scans that follow the clinics standard protocol of exactly seven MRI images taken from different rotational angles with an original image size of 512x512 pixels. Finally, a total of 606 patients (342 PSC / 264 controls) are included in the internal database, resulting in 361 MRCPs taken at 3T (189 PSC / 172 controls) and 398 MRCPs taken at 1.5T (283 PSC / 115 controls). The observed overlap is due to the 113 patients that underwent exams at both magnetic field strengths during clinical follow up. If a patient received multiple examinations at the same magnetic field strength, only the most recent MRCP fulfilling the quality criteria is included in the respective dataset, such that all MRCPs per dataset stem from unique patients.

The 47 patients in the external vendor validation set fulfilled all of the inclusion criteria described above. To strictly separate patients between internal and external data, 10 patients that received multiple MRI examinations and contributed an MRCP to the internal dataset are removed from the external dataset, resulting in a total of 37 included patients and MRCPs (20 PSC / 17 controls).

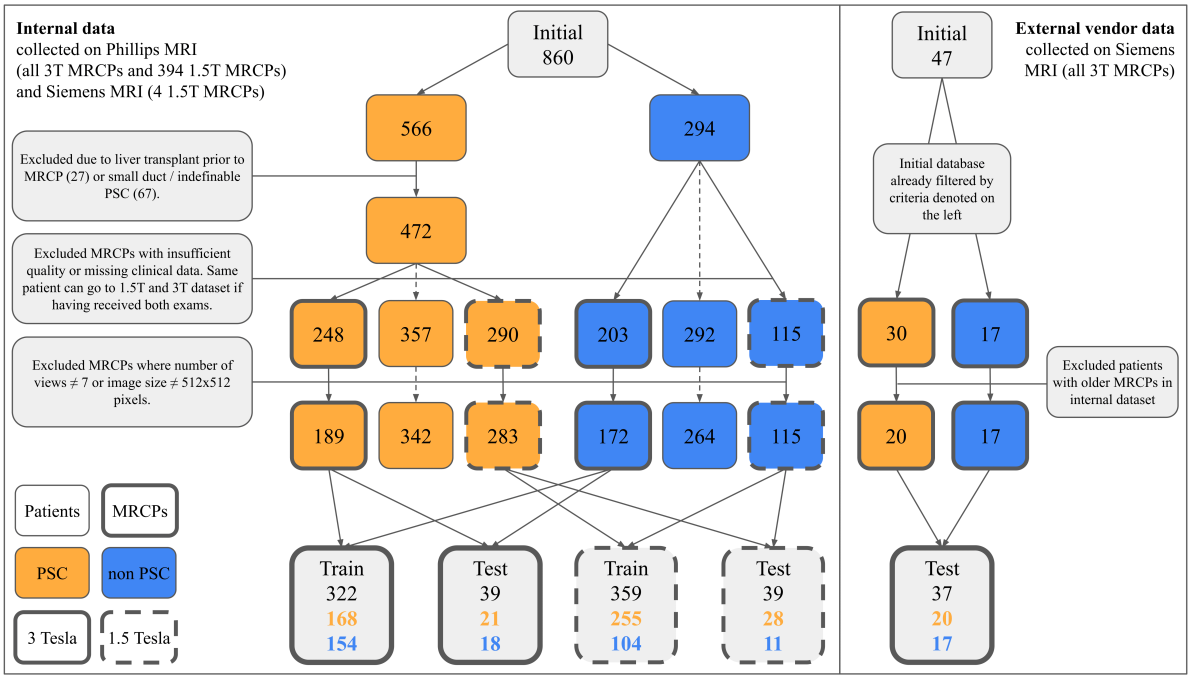

Figure 2.5.1: Overall data-flow. The initial database 897 unique patients is separated into internal and external vendor data. Patients are separated into PSC and non-PSC group. Some patients and their respective MRCP data are excluded due to the denoted reasons. MRCP data of both groups is separated into 3 Tesla and 1.5 Tesla datasets, depending on the magnetic field strength used during acquisition, and finally partitioned into respective training- and test-datasets. Some patients underwent exams at both 3 and 1.5 Tesla during clinical follow up and can therefore contribute MRCPs to both datasets, hence the observed overlap. Reprinted with permission from Ragab and Westhaeusser et al. [77].

2.5.2 Experimental Setup

From the total 361 and 398 samples of the internal 3T and 1.5T dataset, respectively, 39 samples are randomly assigned (with stratification by PSC vs. control, patient age and gender) as test datasets. The remaining 322 and 359 samples of the 3T dataset and 1.5T dataset, respectively, are assigned as training datasets for the deep learning model. From the external vendor validation dataset, all 37 samples are assigned as test data. Figure 2.5.1 and Table 2.4.2 depict the overall data split and distribution into training and test datasets.

2.5.3 Image Normalization

Maximum gray values per MRCP image vary heavily across all samples of all datasets, with a lowest image-wise maximum value of 367 and a highest maximum value of over 20,000 (see Figure A.1.1), but linear rescaling of the maximum and minimum value per image on 1 and 0 respectively did not provide a biologically meaningful homogenization, such that similar structures, e.g., bile ducts, share gray values of a similar range after the preprocessing. One cause for this are artifacts, e.g., the very bright leftover liquid in the stomach present in some MRCPs. To mitigate this issue, after applying Contrast-Limited Adaptive Histogram Equalization (CLAHE) [76] to each MRI image to locally enhance the contrast, the 95th percentile of each image's gray value histogram is mapped to 1 and the 5th percentile is mapped to 0 (see Figure 2.5.2). This smooths out the influence of outlier pixels and areas and provides a robust preprocessing procedure across all used datasets, such that pixel values of the clinically relevant structures in the images are more evenly distributed across individual samples in the dataset.

2.5.4 Human Evaluation

Four radiologists with varying levels of experience in reading MRCP (2, 3, 4, and 9 years for R1, R2, R3 and R4, respectively) independently and blindly analyzed the 78 samples of both internal test datasets and classified the cases as either PSC or non-PSC, according to previously published PSC-compatible findings based on the 2D-MRCP sequence only [28]. In terms of inter-reader reliability, the evaluations submitted by the pathologists express a Fleiss' kappa of 0.384 [0.223, 0.548 CI] on the 3T and 0.410 [0.254, 0.583 CI] on the 1.5T test-set.

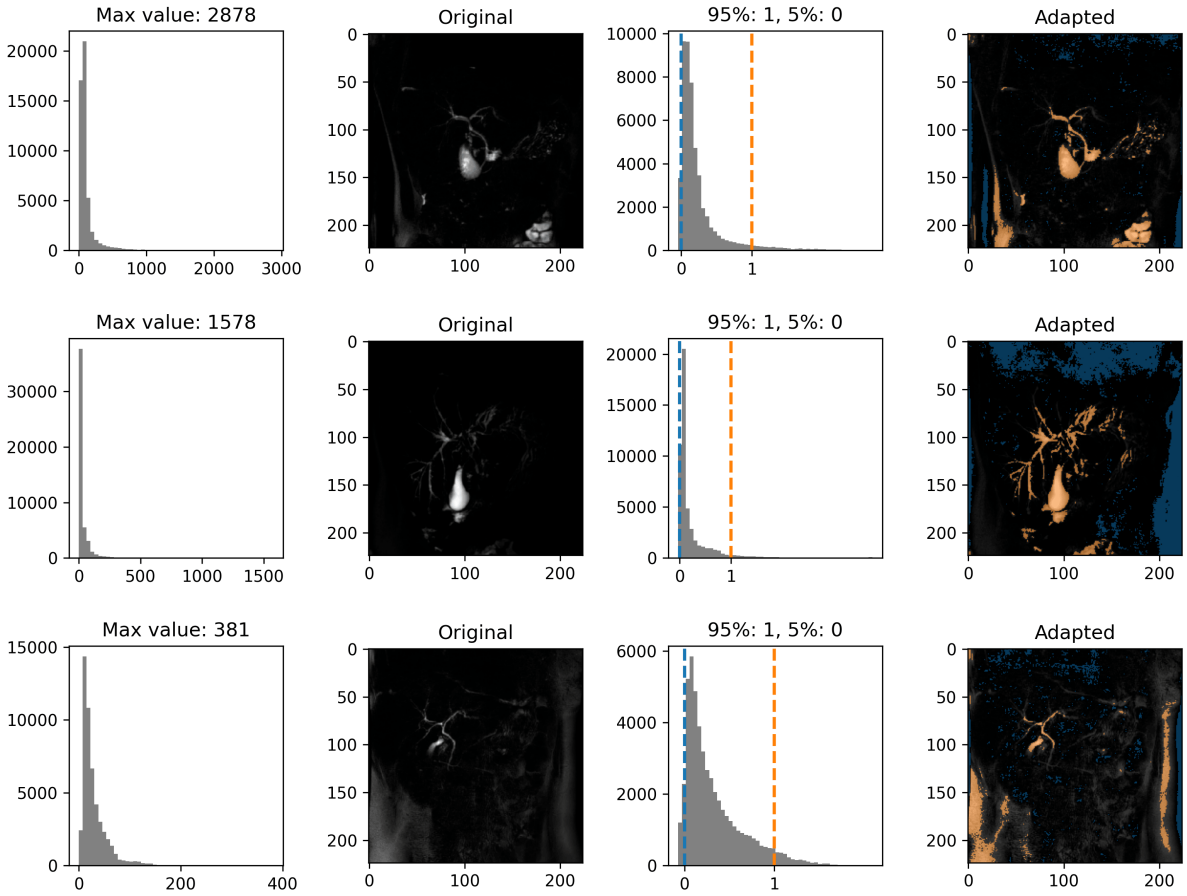

Figure 2.5.2: Original MRCP images of different patients and respective histograms (left) and preprocessed images and histograms after applying contrast-limited histogram equalization (CLAHE) with a contrast-clip limit of 0.015 [76]. Notable are the substantial different maximum gray values of the MRCPs. Before input to the network, all images are normalized to 1 on the 95th percentile of their histogram (dashed orange line) and to 0 on the 5th percentile (dashed blue line). This was found to provide an appropriate dataset homogenization, where similar biological structures share similar gray values across patient samples. This is made visible on the adapted images, where the bile ducts share pixel values > 1 (highlighted in orange) and pixel values < 0 (highlighted in blue) belong to irrelevant structures of the background. Reprinted with permission from Ragab and Westhaeusser et al. [77].

2.6 Methods

This section describes how the proposed deep learning architecture DeePSC in its three levels of complexity, namely the single-view CNN (SVCNN), the multi-view CNN (MVCNN) and the highest confidence ensemble (HCE), is derived. Furthermore, the metrics used for evaluation and the methodology for statistical analysis are delineated.

2.6.1 Deep Learning Architecture (DeePSC)

The main goal of this work is to develop and verify the aptitude of an AI-based clinical decision support system for the automated classification of PSC-compatible cholangiographic findings on 2D-MRCP. To this end, DeePSC is proposed, an end-to-end deep-learning ensemble model for binary classification of PSC 2D-MRCP data. It is specialized in processing the multi-view format of the 2D-MRCP data to combine information across images and improve over classic image-wise classification. DeePSC builds on three increasing levels of complexity, as depicted in Figure 2.6.1: First, a single-view CNN (SVCNN) is trained on all individual images (in the following refered to as MRCP "views") of all patients per dataset until convergence on the task of binary classification of PSC vs. non-PSC. This approach of training and predicting on individual images is the conventional method of AI-based image processing and serves as the baseline. In the second level, the trained SVCNN is extended to form the multi-view CNN (MVCNN) architecture [68]. For this, the feature extraction backbone is duplicated to enable parallel input and processing of all seven MRCP views per patient. An additional attention-based view-fusion layer combines latent information of all images per MRCP into a single representation, which is then used for classification. In the third level, an ensemble of 20 individual instances of the MVCNN is trained on the same data with varying random seeds. To increase predictive robustness of the full model, only the prediction of the MVCNN instance that expresses the highest class probability (and therefore the highest confidence) is considered for the final prediction of the ensemble model. This is further denoted as “highest confidence ensemble” (HCE) and represents the final model DeePSC. DeePSC is implemented using PyTorch in Python 3.6. Conceptual code can be found at https://github.com/imsb-uke/DeePSC. In the following, the different components and sub-models are explained in more detail.

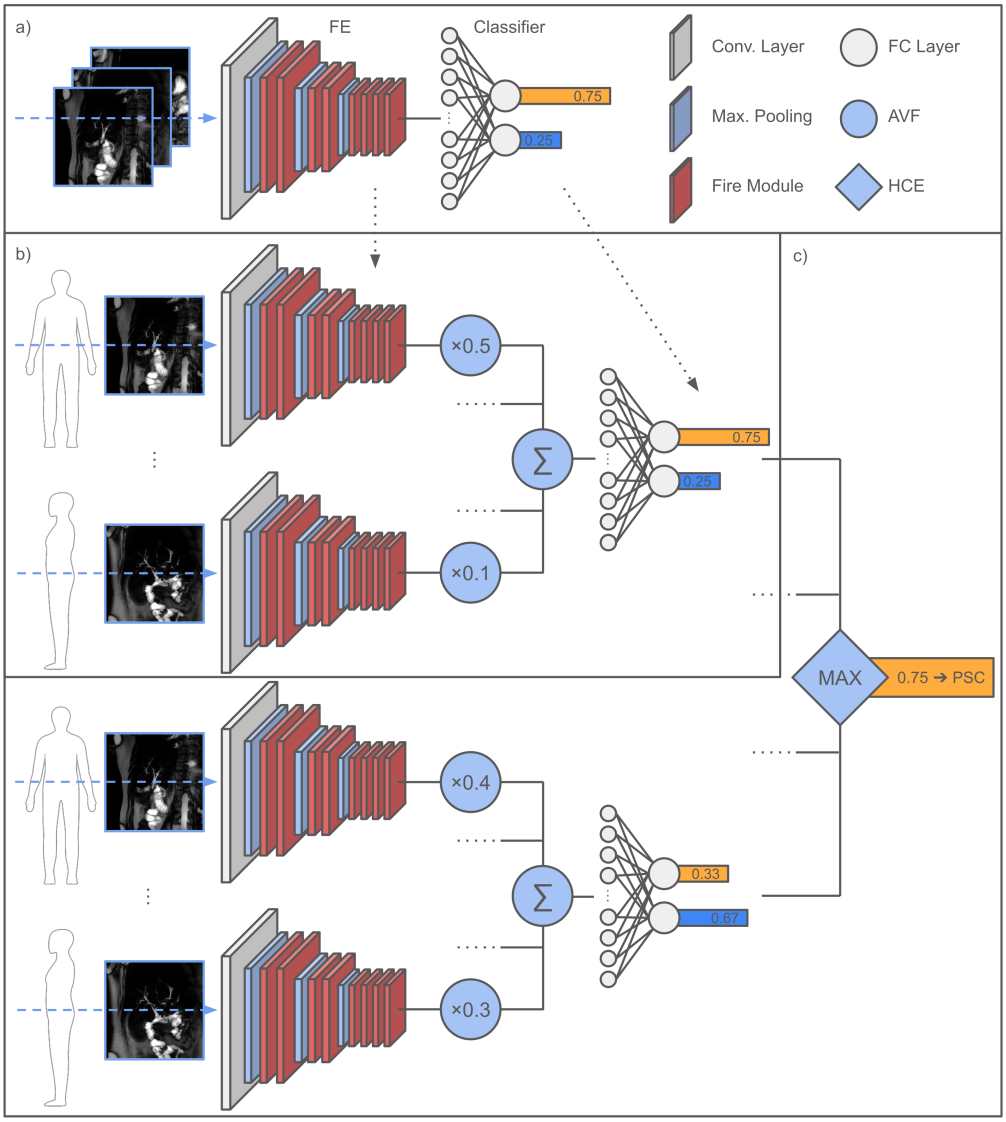

Figure 2.6.1: Overall structure of the DeePSC model. a) The baseline single-view CNN (SVCNN) is trained with a sequential input of all MRCP-views of all patients until convergence. The SVCNN is based on the Squeezenet architecture and consists of the convolutional Feature Extractor (FE) and the fully-connected classification layer [44]. b) The trained SVCNN is extended to the multi-view CNN (MVCNN) by multiplying the FE by the number of MRCP-views per patient and introducing the attention-based view-fusion layer (AVF) between FE and classification layer. Here, all seven views of a single patient are processed in parallel and aggregated in the AVF, while the classification layer derives a single prediction per patient. c) Twenty individually trained MVCNNs are combined in the Highest Confidence Ensemble (HCE) to form the final DeePSC model, where only the prediction with the highest class probability across all twenty MVCNNs is forwarded as the final classification per patient. Reprinted with permission from Ragab and Westhaeusser et al. [77].

Single-view CNN (SVCNN)

In the lowest level of model complexity, all 2D-MRCP images across patients are treated as individual samples and processed separately using the single-view CNN (SVCNN). The SVCNN consists of a convolutional feature extractor (FE) and a subsequent fully connected classification head. All parameters of the SVCNN are trainable.

Feature Extractor (FE): A Squeezenet architecture (see 2.2) pretrained on the ImageNet dataset serves as backbone architecture for the Feature Extractor (FE) [44]. For usage in this work, the final convolutional layer of Squeezenet is replaced by a flatten operation, resulting in a 1-dimensional feature vector of length $L = 512$ in the output of the FE. The FE transforms an MRCP view $V \in \mathbb{R}^{H \times W \times C}$ of height $H = 227$ and width $W = 227$ with three color channels $C = 3$ in the RGB format into a latent feature representation $h^{\text{FE}} \in \mathbb{R}^L$. Before input to the FE, the color channel of the grayscale MRCP images is repeated three times to confer with the RGB input format. Furthermore, the images are downsampled to a resolution of 227x227 pixels from their original resolution of 512x512 pixels, which refers to the expected input shape of the Squeezenet backbone CNN. During the experiments and in contrast to the findings of Geras et al. [74], this reduced resolution had no observable negative impact on training results while greatly improving computational performance. For every image V, the latent representation h^{FE} after the feature extractor is then derived as

$$h^{\text{FE}} = f^{\text{FE}}(V) \text{ with } f^{\text{FE}} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^L. \tag{2.6.1}$$

Classification head: The subsequent classification head consists of a fully connected layer with L input and 2 output nodes, followed by a softmax activation function. In detail, the output logits $z \in \mathbb{R}^2$ of the classification head are derived from the latent feature representations after the FE h^{FE} as

$$z = W^{\text{CH}} h^{\text{FE}} + b^{\text{CH}}. \tag{2.6.2}$$

Here, $W^{\text{CH}} \in \mathbb{R}^{c \times L}$ and $b^{\text{CH}} \in \mathbb{R}^c$ are trainable parameters of the network, where L refers to the length of the input vector and $c = 2$ denotes the number of classes. The probability $\hat{y}_i$ of an image to belong to class i is then derived by the softmax function as

$$\hat{y}_i = \frac{\exp(z_i)}{\sum_{j=0}^{c-1} \exp(z_j)}. \tag{2.6.3}$$

The predicted probabilities refer to the two classes "PSC" and "non-PSC" of the binary classification task.

Multi-view CNN (MVCNN)

The multi-view CNN (MVCNN) is inspired by the work of Su et al. [68]. In contrast to the SVCNN, the MVCNN takes all seven 2D-MRCP images of a single patient as simultaneous input. It extends the SVCNN by introducing the attention-based view-fusion layer (AVF) between the FE and the classification head of the network. The input order of the images per patient is constant and always refers to the order of subsequent rotational acquisition, i.e. images per sample are not shuffled. The images are processed in parallel in the FE. The weights of the FE in the MVCNN are ported from the trained SVCNN and then frozen, such that the same parameters apply to all images. The resulting individual latent representations of all seven views per patient are then merged into a single representation in the AVF. Based on the results of [30], who found view-fusion in the early layers of the network to be significantly less effective, the view-fusion layer is applied late in the network. Finally, the combined latent representation per patient is forwarded to the classification head, which adapts the weights from the trained SVCNN classification, with weights corresponding to those in the trained SVCNN. During training of the MVCNN, only the parameters of the AVF and the classification head are trainable, whereas those of the FE are static.

Attention-based view-fusion (AVF) The Attention-based view-fusion (AVF) layer compresses the input of the seven latent view vectors per patient into a single representation. This method was first introduced by Ilse et al. [78] in the context of multiple instance learning (MIL) on histopathological images. It can be seen as a learnable weighted average function of the input instances, where the respective attention weights are determined inside the network for each sample. Given the set of $N = 7$ latent representations $H^{\text{FE}} = \{h_1^{\text{FE}}, ..., h_N^{\text{FE}}\}$ of shape $h_n^{\text{FE}} \in \mathbb{R}^{Lx1}$ per view, the output $h^{\text{AVF}} \in \mathbb{R}^{Lx1}$ of the AVF layer is derived by

$$h^{\text{AVF}} = \sum_{n=1}^{N} a_n h_n^{\text{FE}}, \tag{2.6.4}$$

where

$$a_n = \frac{\exp(W^{\text{AVF2}} \tanh(W^{\text{AVF1}} h_n^{\text{FE}} + b^{\text{AVF1}}) + b^{\text{AVF2}})}{\sum_{m=1}^{N} \exp(W^{\text{AVF2}} \tanh(W^{\text{AVF1}} h_m^{\text{FE}} + b^{\text{AVF1}}) + b^{\text{AVF2}})}. \tag{2.6.5}$$

Here, $W^{\text{AVF1}} \in \mathbb{R}^{D \times L}$, $b^{\text{AVF1}} \in \mathbb{R}^{D}$, $W^{\text{AVF2}} \in \mathbb{R}^{1 \times D}$ and $b^{\text{AVF2}} \in \mathbb{R}^{1}$ are trainable parameters of the network, where L refers to the length of the input vector and D is a hyperparameter. After optimization, D is set to 64 in this work. The hyperbolic tangent non-linearity aims to prevent the exploding gradient issue, whereas the softmax function ensures that all attention weights sum to 1 [78]. h^{AVF} is then forwarded to the classification head.

Highest Confidence Ensemble (HCE)

The Highest Confidence Ensemble (HCE) refers to an ensemble of 20 individual instances of the MVCNN, trained on the same data and hyperparameters with varying random seeds. This influences the weight initialization, random augmentations as well as sample order during training, which can lead to convergence to a different local minimum of the loss landscape and therefore different results. During inference, all images per patient are processed in all 20 instances of the MVCNNs in the ensemble. Then, only the prediction of the MVCNN instance that expresses the highest class probability (and therefore the highest confidence) is considered as the final prediction of the ensemble model. In detail, let $\hat{y}_m$ represent the probabilities for classes "PSC" and "non-PSC" predicted by the $m-th$ instance of the Multi-view CNN (MVCNN) within the ensemble of 20 individual instances. The output $\hat{y}^{\text{HCE}}$ of the HCE ensemble is then derived as

$$\hat{y}^{\text{HCE}} = \max_i(\hat{y}_m). \tag{2.6.6}$$

The combination of SVCNNs, MVCNNs and HCE refers to the final proposed model DeePSC.

2.6.2 Training and Hyperparameter-Tuning

Separate instances of DeePSC are trained on the 3T and 1.5T datasets. Training is performed with a constant batch-size of 25 patients for a maximum of 500 epochs until

convergence using the AdamW optimizer [79] and the crossentropy loss, defined as

$$L(y, \hat{y}) = -\frac{1}{c}\sum_{i=0}^{c-1}\left(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right), \tag{2.6.7}$$

using automatic mixed precision on a NVIDIA Tesla V100 SXM2 GPU with 16 GB GPU memory. The SVCNNs are trained with a learning rate of 3e-6 and an early stopping patience of 20 epochs on the training loss, while for the MVCNNs a learning rate of 3e-7 and an early stopping patience of 10 epochs is applied. Hyperparameters are tuned via grid searching a predefined parameter-space using stratified 5-fold-cross-validation on the training set of the 3T data over 2 rounds (random seeds) to avoid overfitting. The 3T and 1.5T datasets are always separated during training, but the same hyperparameters are applied to both models. To enhance the model's performance on the minority class (control group) on the highly class imbalanced 1.5T dataset, the minority class in the training dataset is randomly oversampled to achieve a balanced class distribution during training.

A strong emphasis is put on data augmentation. Augmentation methods include random rotation, shifting, shearing and scaling, as well as a random overlay of gaussian noise and random histogram shifts, implemented as per the MONAI python package [80]. The latter is motivated by Billot et al., who found that random histogram shifts in MRI data enables the network to generalize better by focusing more on shapes rather than intensities [81].

2.6.3 Statistical Analysis

Statistical analysis of the demographic data is performed by using GraphPad Prism for MacOS, Version 9.1.1 (2021, GraphPad Software, Inc., USA). Descriptive statistics are used to outline the demographic and clinical characteristics in the PSC and the control group. Continuous variables are compared between groups using unpaired student's t-tests or Mann-Whitney-U-test, respectively. For categorical data, the Chi-square test or the Fisher's exact test is used as appropriate. Statistical analysis of the predictive performances of the deep learning models as well as the radiologists is performed using Welch's t-test as provided by the SciPy package for Python 3.10, Version 1.9.2 (2022, The SciPy Community). Statistical significance is defined as $p < 0.05$.

2.6.4 Metrics

Classification performance of the proposed algorithm as well as the human annotators is evaluated by the following metrics:

- Accuracy, which measures the overall correctness of the classification process. Calculated as $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.
- Sensitivity, which measures the model's ability to correctly identify positive samples. Calculated as $sensitivity = \frac{TP}{TP+FN}$.
- Specificity, which measures the model's ability to correctly identify negative samples. Calculated as $specificity = \frac{TN}{TN+FP}$.
- F1-score, which combines $precision = \frac{TP}{TP+FP}$ and sensitivity in a harmonic mean as $F1 = \frac{2\times(precision\times recall)}{precision+recall}$. This metric provides a more accurate measure of correctness than accuracy on class-unbalanced datasets.

Here, TP denotes true positives, TN true negatives, FP false positives, and FN false negatives.

2.7 Experiments

In the following section, quantitative and qualitative results of all experiments conducted with DeePSC and its submodels MVCNN and SVCNN, as well as the MRCP classifications provided the radiologists, are described.

2.7.1 Evaluation on Internal Data

Firstly, DeePSC, its sub-variants MVCNN and SVCNN as well as the predictions of the four radiologists are evaluated on the task of detecting PSC-compatible findings on 2D-MRCP data on the unseen test-sets of the internal data. Table 2.7.1 depicts the predictive accuracy of all raters in terms of accuracy, F1-score, sensitivity and specificity. Results are derived from five instances of DeePSC per dataset (i.e. magnetic field strength), formed by a hundred instances of MVCNN and SVCNN trained on different random seeds. Performance is reported as mean and standard deviation across instances to enable statistical comparison with the human readers and increase reliability.

The final DeePSC ensemble model achieves an accuracy of 80.51±1.25% on the internal 3T dataset and 82.57±2.99% on the internal 1.5T dataset, while the radiologists achieve an average of 75.00±8.38% and 72.44 ± 8.58%, respectively. DeePSC outperforms the four radiologists on average by 5.51 ($p = 0.338$) percentage points and 10.13 ($p = 0.131$) percentage points, respectively, however, without reaching statistical significance. In comparison to individual radiologists' predictions, the DeePSC model performs slightly better than the best human reader R1 on the 1.5T test-set and beats 3 out of 4 human readers in predictive accuracy on the 3T test-set. Notably, on both datasets, DeePSC outperforms the most experienced radiologist R4 with nine years of experience in reading MRCP. Predictive performance of the proposed algorithm increases among all metrics on both datasets with the increasing level of model complexity. Compared to processing individual MRCP images in the SVCNN, utilizing the combined information of all views per patient in the MVCNN increases average accuracy by 1.42 ($p < 0.0001$) and 3.51 ($p < 0.0001$) percentage points, on the 3T and 1.5T test-set, respectively. This is further enhanced by 2.15 ($p < 0.05$) and 2.39 ($p = 0.189$) percentage points by applying the HCE of the DeePSC model. Taking the individual sample predictions into account, inter-reader reliability among radiologists is low, with a Fleiss' kappa of 0.384 [0.223, 0.548 CI] on the 3T and 0.410 [0.254, 0.583

Table 2.7.1: Predictive performance of the model and the radiologists on the unseen test-set of the 3T and 1.5T datasets, respectively. The SVCNN & MVCNN are trained a hundred times with varying random seeds, where every twenty runs form a DeePSC ensemble (total of five). Underlined numbers highlight the best result in every respective metric. ± refers to the standard deviation. * highlights that these results differ statistically significantly ($p < 0.05$) from those of the row above, n.s indicates that the difference is not statistically significant. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$). Reprinted with permission from Ragab and Westhaeusser et al. [77].

	Accuracy	F1	Sensitivity	Specificity
Train: 3T		Pred: 3T (21 PSC, 18 non PSC)		
SVCNN Mean	76.94 ± 1.76	78.05 ± 1.45	76.11 ± 1.71	77.91 ± 3.80
MVCNN Mean	$78.36 \pm 2.95^{****}$	$79.13 \pm 2.94^{*}$	$76.29 \pm 4.15^{n.s}$	$80.78 \pm 4.87^{****}$
DeePSC Mean	$80.51 \pm 1.25*$	$81.55 \pm 1.18^{**}$	$80.00 \pm 1.90^{*}$	81.11 ± 2.72^{ns}
Radio. Mean	$75.00 \pm 8.38^{n.s}$	$77.39 \pm 6.24^{n.s}$	$77.38 \pm 2.06^{n.s}$	72.22 ± 18.00^{ns}
R1	<u>87.18</u>	<u>86.49</u>	76.19	<u>100.00</u>
R2	64.10	69.57	76.19	50.00
R3	71.79	74.42	76.19	66.67
R4	76.92	79.07	<u>80.95</u>	72.22
Train: 1.5T		Pred: 1.5T (28 PSC, 11 non PSC)		
SVCNN Mean	76.67 ± 2.73	82.63 ± 2.21	77.41 ± 3.29	74.80 ± 5.06
MVCNN Mean	$80.18 \pm 3.92^{****}$	$85.40 \pm 3.00^{****}$	$80.89 \pm 4.41^{****}$	$78.37 \pm 9.69^{**}$
DeePSC Mean	<u>$82.57 \pm 2.99^{n.s}$</u>	<u>$87.33 \pm 2.00^{n.s}$</u>	<u>$83.57 \pm 1.75^{*}$</u>	$80.00 \pm 8.91^{n.s}$
Radio. Mean	72.44 ± 8.58^{ns}	$79.25 \pm 6.30^{n.s}$	$73.22 \pm 5.92^{ns.s}$	$70.46 \pm 17.46^{n.s}$
R1	82.05	86.79	82.14	81.82
R2	61.54	71.70	67.86	45.45
R3	66.67	74.51	67.86	63.64
R4	79.49	84.00	75.00	<u>90.91</u>

CI] on the 1.5T test-set. This shows a strong disagreement among radiologists when asked to base their classification of PSC solely on the MRCP image data. The five DeePSC models trained on different random seeds achieve a fundamentally higher consensus of 0.948 [0.867, 1 CI] and 0.868 [0.736, 0.969 CI], respectively. When comparing the final predictions of the best performing DeePSC model with those of the best performing radiologist R1, a Cohen's kappa of 0.584 [0.328, 0.840 CI] on the 3T and 0.592 [0.304, 0.840 CI] on the 1.5T data is achieved.

2.7.2 Evaluation on External Vendor Validation Data

To quantify the robustness of DeePSC and its potential aptitude in an actual clinical environment, the 3T variant is evaluated on the external vendor validation set. This dataset consists solely from MRCPs that were collected on an MRI machine of a manufacturer that was not used for any sample of the training dataset. Furthermore, all MRCPs stem from new patients unknown to the model, i.e., which are not previously used in the training-set.

Table 2.7.2: Predictive performance of the model trained on the internal 3T training-set and evaluated on the unseen 3T Siemens test-set. The SVCNN & MVCNN are trained a hundred times with varying random seeds, where every twenty runs form a DeePSC ensemble (total of five). ± refers to the standard deviation. * highlights that these results differ statistically significantly ($p < 0.05$) from those of the row above, n.s indicates that the difference is not statistically significant. * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$). Reprinted with permission from Ragab and Westhaeusser et al. [77].

	Accuracy	F1	Sensitivity	Specificity
Train: 3T	Pred: Siemens 3T (20 PSC, 17 non PSC)			
SVCNN Mean	79.62 ± 2.68	82.86 ± 1.85	90.94 ± 4.34	66.30 ± 8.62
MVCNN Mean	$91.51 \pm 3.06^{*****}$	$92.62 \pm 2.52^{****}$	$98.10 \pm 2.98^{****}$	$83.77 \pm 6.23^{****}$
DeePSC Mean	$\mathbf{92.43 \pm 1.08}^{\text{n.s}}$	$\mathbf{93.46 \pm 0.89}^{\text{n.s}}$	$\mathbf{100.00 \pm 0.00}^{****}$	$\mathbf{83.53 \pm 2.36}^{\text{n.s}}$

On the different vendor validation-set, the DeePSC ensemble model trained on the 3T internal dataset achieves 92.43±1.08%, 93.46±0.89%, 100.00±0.00% and 83.53±2.36% for accuracy, F1-score, sensitivity, and specificity, respectively (see Table 2.7.2). The best performing model misclassifies only two of the 17 samples in the control group. Especially noteworthy here is the substantial improvement in average accuracy by 11.89 ($p < 0.0001$) percentage points between SVCNN and MVCNN, proving the positive effect of the proposed attention-based view fusion method when classifying 2D-MRCP, strongly suggesting that the multi-view architecture confers robustness to the network. Applying the HCE of DeePSC leads to a insignificant decrease in specificity by 0.24 ($p = 0.864$) percentage points and a concomitant significant increase in sensitivity by 1.90 ($p > 0.0001$) percentage points.

2.7.3 Explainability

To understand which visual features underlie the decision-making process of DeePSC, gradient-weighted class activation maps (GradCAM) are calculated on the last convolutional layer of the FE for the image receiving the highest attention score in the AVF per MRCP [48].

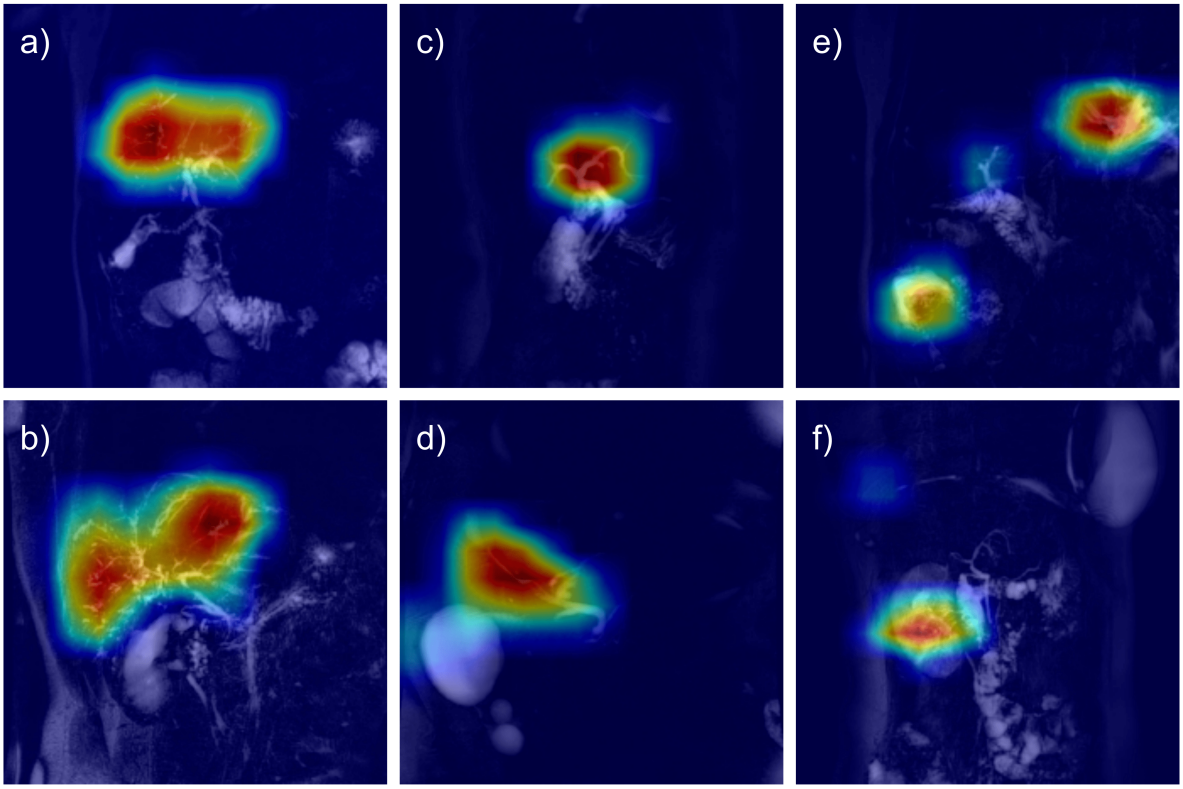

Figure 2.7.1: Class-related activations in the last convolutional layer of the Feature Extractor for six patients in the 3T dataset, calculated with GradCAM [48]. High activation values, on the red end of the spectrum, indicate a high correlation of those features with the model's prediction. Cases a) and b) are correctly classified to the PSC group, cases c) and d) correctly classified to the control group with high activation in the area of the biliary tree. Case e) is correctly classified as belonging to the control group, but the class activation map shows that the model wrongly derives its prediction from the areas of the gastric corpus and the gallbladder. In case f), the model incorrectly classifies the MRCP to the PSC group based on irrelevant features of the colon. Reprinted with permission from Ragab and Westhaeusser et al. [77].

The GradCAM activations are depicted as class-related heatmaps to provide a visual cue of which areas in the image contribute most to the model's class prediction. Figure 2.7.1 presents six exemplary heatmaps of the internal 3T test dataset. For this

dataset, the heatmaps of the best performing DeePSC model reveal high activity values in the anatomic region of the biliary tree in 33 of 39 cases, strongly suggesting that the model learned to base its classification upon biologically relevant features. While in the remaining six cases high activity in the biliary tree can not be observed (see Table 2.7.3), five out of the six cases are still classified correctly. Besides the inherent model uncertainty, data uncertainty such as gastrointestinal fluid and image artifacts can lead to distraction and thus misclassification of the network (see Figure 2.7.1 e) & f)).

Table 2.7.3: Distribution of correct and incorrect classifications (PSC & non-PSC) of DeePSC on the internal 3T testset with respect to the GradCAM activation heatmaps of the last convolutional layer of the CNN feature extractor. Reprinted with permission from Ragab and Westhaeusser et al. [77].

3T GradCAM	Correct Prediction	Incorrect Prediction
High activation in biliary tree	27	6
No high activation in biliary tree	5	1

2.7.4 Ablation

To assess the influence of the proposed preprocessing in DeePSC, an ablation study is performed on the test data of all three datasets, as shown in Figure 2.7.2. Results stem from a single instance of DeePSC, trained on either the 1.5T or 3T data, and 20 underlying MVCNN instances. On the x-axis, the changes to the preprocessing are depicted in an additive fashion. The leftmost setting of [1, 0] refers to input of the data with greyvalue min/max-normalization on 0 and 1, no histogram equalization and no random data augmentation during training. It is visible that for the model trained on 3T data, all proposed additions of contrast-limited adaptive histogram equalization (CLAHE) (see 2.5.3), quantile-based greyvalue normalization (see 2.5.3) and random augmentations (see 2.6.2), consisting of affine transformations, gaussian noise and histogram shifts, lead to a constant increase in classification accuracy on both the internal testset and the different vendor validation set. For the model trained on the 1.5T data, a drop can be observed when applying the quantile-based greyvalue normalization, however, this is mitigated by applying the proposed random data augmentation. In summary, the combination of the proposed preprocessing steps leads to a strongly increased performance on both 3T datasets, while maintaining the baseline

performance on the 1.5T data, proving its beneficial influence to the overall method.

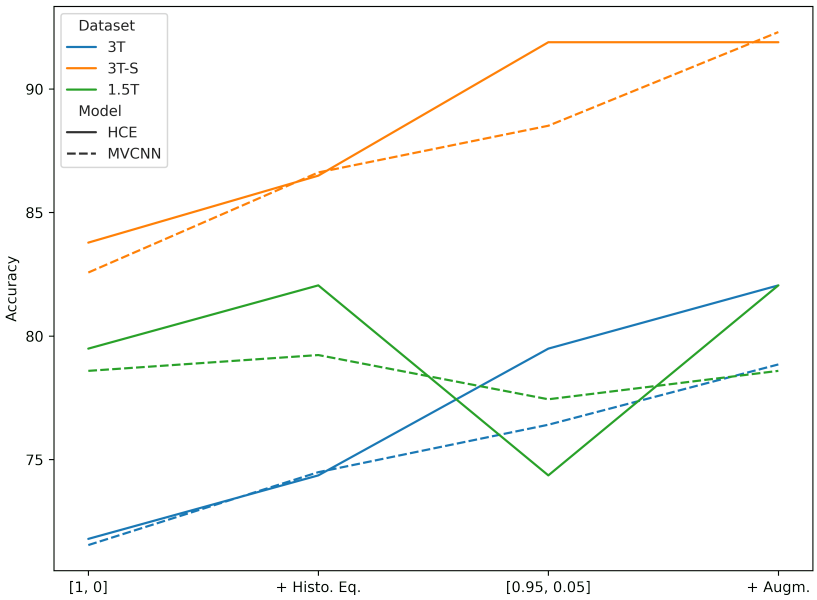

Figure 2.7.2: Ablation study of the individual parts of the proposed MRCP preprocessing for model training, evaluated on the three test datasets. Accuracy of the HCE refers to a single DeePSC entity, while for the MVCNN the mean values of twenty models are depicted. From left to right on the x-axis: Min/Max-normalization on 0/1; Additional CLAHE before normalization; Normalization of the 5th and 95th quantile of each image's histogram on 0/1; Additional image augmentation during training as described in Section 2.6. Even though on the 1.5T dataset accuracy is not monotonously increasing in all steps, the proposed preprocessing procedure is found to provide the best overall performance across all datasets (rightmost point). Reprinted with permission from Ragab and Westhaeusser et al. [77].

2.8 Discussion

In the present study, a multi-view deep convolutional neural network ensemble for the automated classification of PSC-compatible bile duct alterations on 2D-MRCP both taken at 1.5 and 3 Tesla is developed. To the best of my knowledge, this is the first work to apply such a deep learning approach for the assessment of PSC on radial 2D-MRCP datasets. Diagnostic performance of the DL model is compared to that of four diagnostic radiologists at different levels of training and with varying experience in reading MRCP, thus reflecting clinical practice.

2.8.1 Classification Task

The results show that the proposed model identifies patients with PSC based on 2D-MRCP images both at 1.5T and 3T with high reliability. It achieves higher accuracy, sensitivity, and specificity than the radiologists on average. Statistical significance could hereby not be reached, though this is potentially linked to the small compared sample sizes of 5 and 4. When comparing the radiologists' classifications individually, the model performs slightly better than the best radiologist R1 in all metrics except specificity on the 1.5T data and outperforms three out of four radiologists on the 3T data. On the different vendor 3T test-set obtained on a Siemens MRI scanner, the proposed ensemble models reach over 90% accuracy, only misclassifying between two and three out of 37 total samples. While speculative in nature, the high performance on Siemens data may be caused by the higher image quality and homogeneity of the more recently collected MRCP scans. The results show that combining the information of multiple MRCP views per patient with the attention-based view-fusion layer in the MVCNN consistently improves classification performance among all metrics and datasets compared to the classic baseline SVCNN, by up to 17.47 percentage points in specificity on the different vendor validation-set. Similarly, besides a minor decrease in specificity on the different vendor validation-set, performance of the MVCNN is further increased consistently by applying the highest confidence ensemble method of DeePSC on twenty individually trained MVCNNs. Given the aim of providing robust and reliable performance across different datasets, the mentioned drop in specificity of 0.24 percentage points on a single dataset is negligible when taking the gain in other metrics and datasets into account. The predictions provided by the four radiologists are in line with previously reported values. In a meta-analysis,

Dave et al. reported a high diagnostic performance of MRCP for the diagnosis of PSC in cholestatic patients with a sensitivity and specificity of 0.86 and 0.94, respectively [82]. However, objective and reproducible image interpretation remains challenging, especially for less experienced radiologists and low-volume centers. Since PSC is a rare disease with specific and sometimes subtle imaging features, high expertise and thorough knowledge are essential to reliably interpret the characteristic findings. The implementation of deep-learning algorithms is expected to add significant value in the decision-making process, given the high and robust performance of the proposed algorithm and the low inter-reader agreement of the radiologists observed in this study. Of note, all four radiologists are from a large volume referral center where PSC is often seen in the radiology department, underlining the potential unmet need in less experienced services.

According to a recently published meta-analysis on the performance of deep-learning algorithms in medical imaging, many studies on the subject suffer from substantial methodological shortcomings, which limits their translation into real-life clinical setting [83]. Many of the investigated studies, for instance, did not compare the performance of their model with that of a human expert or assessed the performance of their model on a different dataset than the one used for human performance assessment. Also, very few studies included an independent validation-dataset. This might result in incorrectly high values of accuracy due to overfitting and low generalizability of the proposed model [83]. Both shortcomings are addressed in this work.

Foundational work for deep learning-based detection of PSC-compatible bile duct alterations was provided by Ringe et al. [66]. They showed that automated classification of PSC was feasible using majority voting for maximum intensity projections (MIPs) of 3D-MRCP. This thesis represents a significant extension to their work, as it provides realistic benchmarks on the used dataset by comparing the proposed method to the predictions of four radiologists. Furthermore, generalizability of the proposed model to data from a different manufacturer's MRI machine is proven by showing very high performance on an independent validation dataset. While Ringe et al. also improved their performance by combining information of different MIPs per patient by majority voting on the individual predictions of their network, the trainable AVF layer of DeePSC allows for more flexible and sample-related view-fusion and therefore better predictions. Finally, this work shows that by employing an ensemble of multiple individually trained models, performance can be substantially improved compared to a single model instance.

2.8.2 Limitations and Outlook

To the best of my knowledge, this study contains the largest 2D-MRCP dataset of PSC patients and controls published so far. Nevertheless, the patient numbers of this rare disease are still limited to several hundred, which might give reason to believe that network performance could be boosted by including more patient data. While the GradCAM activation heatmaps emphasize that the model learned to correctly base its classification upon biologically relevant features of the biliary tree, the model might still be confused by samples that are outside its learned distribution and mistakenly assess irrelevant feature structures as PSC-related (see Figure 2.7.1). This holds especially true for the differentiation of PSC from other forms of sclerosing cholangitis or biliary malignancy, both not included in the control dataset used in this work. Including more patients and controls with other forms of liver and bile duct diseases therefore is key to providing an even more robust clinical decision support system in the future. Furthermore, this work does not quantify the potential influence of decision support from DeePSC by analyzing diagnostic performance of human readers paired with the proposed model in a clinical setting. However, this might be subject of further studies to come. Lastly, the sample sizes used for comparing the results of DeePSC and the radiologists (five and four respectively) are very small and make valid claims about statistically significant differences ambiguous. Further studies should try to mitigate this issue by increasing the number of human evaluators.

2.9 Excurse: Analysis of Multi-View-Aggregation Strategies

The previous results show that the proposed model is able to robustly classify PSC-compatible findings on multiple datasets and that the utilization of the combined information of all seven MRCP images through the AVF in the MVCNN leads to increased performance. However, the same attention-based view fusion method is applied for all evaluations of DeePSC. The question arises, if and how the observed improvement in predictive performance by utilizing the AVF can be further increased by application of a different method. For this aim, an additional systematic exploration of different view-fusion and view-correlation methods inside and outside of the architecture, which are motivated by the literature, is conducted.

2.9.1 Aggregation Techniques

In the following, the various view-fusion and view-correlation methods analyzed and compared in this study are described in detail.

Single-View CNN Metric Fusion

The multi-view format of the data can already be utilized with the SVCNN by aggregating the individual predicted class probabilities for each separate view, outside the actual DL network. This leaves the SVCNN's architecture unchanged, but evaluates the predictive performance on a per-patient basis instead of a per-image basis. The following view-score fusion methods are analyzed in this work:

Full patient prediction refers to ...

Maximum - ...the maximum value of probabilities per class over all n views [84].

Additive - ...the sum of all probabilities per class over all n views [84].

Multiplicative - ...the product of all probabilities per class over all n views [84].

Majority Vote - ...the majority class when classifying each view separately [66].

Multi-View CNN View Pooling

The View Pooling Layer compresses the input of multiple latent view vectors of shape $L \times n$ into a single representation of shape $L \times 1$ in the MVCNN. It is located between the FE and the classification head. In this work, six pooling methods are examined:

Max Pooling - Basic view pooling method, directly adopted from [68]. Takes the maximum across views of each element along L.

Mean Pooling - Basic view pooling method, directly adopted from [68]. Takes the mean over views of each element along L.

Fully Connected Pooling - A trainable fully connected layer that takes the concatenated latent view vectors as an input size of $n \cdot L$, with an output size of L. Adopted from [74].

LSTM Pooling - An uni-directional long short-term memory (LSTM) layer with two layers that takes the n latent views as sequential input and forwards only the last hidden state [85]. Even though the incoming feature vectors of the different views do not explicitly represent sequential data, the last hidden state combines information of all input feature vectors in a single representation.

Attention Pooling (AVF) - The attention-based view fusion AVF layer as used in DeePSC (see 2.6.1).

Element-wise Attention Pooling - This work further introduces the Element-wise Attention Pooling, which learns as many attention weights as there are features in the latent views. This extends the aforementioned Attention Pooling, where only a single attention weight for each view is computed. In theory, this enables a more fine-graded pooling and allows the network to individually weight and forward specific regions inside each view. The attention-vector $fa_n \in \mathbb{R}^L$ every view is computed as

$$fa_n = \frac{\exp(W^{\text{EAP2}} \tanh(W^{\text{EAP1}} h_n^{\text{FE}} + b^{\text{EAP1}}) + b^{\text{EAP2}})}{\sum_{m=1}^{N} \exp(W^{\text{EAP2}} \tanh(W^{\text{EAP1}} h_m^{\text{FE}} + b^{\text{EAP1}}) + b^{\text{EAP2}})}. \tag{2.9.1}$$

Here, $W^{\text{EAP1}} \in \mathbb{R}^{D \times L}$, $b^{\text{EAP1}} \in \mathbb{R}^D$, $W^{\text{EAP2}} \in \mathbb{R}^{L \times D}$ and $b^{\text{EAP2}} \in \mathbb{R}^L$ are trainable parameters of the network, where L refers to the length of the input vector and D is a hyperparameter. After optimization, D is set to 64 in this work. The hyperbolic tangent non-linearity aims to prevent the exploding gradient issue, whereas the softmax function ensures that all attention weights sum to 1.

After multiplying each feature with its respective attention weight, i.e. the Hadamard-product of the attention vector $fa_n \in \mathbb{R}^L$ and the feature vector $h_n \in \mathbb{R}^L$, the output is summed along the views and pooled to a single representation h^{AVF}, which is then forwarded to the classification head. This is denoted as

$$h^{\text{AVF}} = \sum_{n=1}^{N} fa_n \odot h_n^{\text{FE}}. \tag{2.9.2}$$

Multi-View CNN View Correlation

The view correlation layer is an optional layer in the MVCNN that aims to inject inter-view related information about all views into each latent representations of the extracted views without altering the shape of the tensor. It is located between the FE and the view-fusion layer. Three methods for view-correlation were evaluated in the course of this work:

None - Skip the View Correlation Layer.

LSTM Correlation - A bi-directional LSTM [85] that takes the n latent views as sequential input and forwards all n hidden states. Adopted from [70].

Self-Attention Correlation - For every latent view $h_n \in \mathbb{R}^{Lx1}$, N attention weights a_{nm} are computed, resulting in the attention matrix $A_{SA} \in \mathbb{R}^{N \times N}$. A_{SA} contains information about the relevance of each view h_n in relation to every other view h_m and is multiplied with the incoming feature vector after the FE. This creates context-aware embeddings from every view. This layer is implemented as proposed by Rymarczyk et al. [86]. The output h_n^{SA} for a single latent view h_n^{FE} is derived as

$$h_n^{\text{SA}} = h_n^{\text{FE}} + \sum_{m=1}^{N} a_{nm}(W^{\text{V}} h_m^{\text{FE}} + b^{\text{V}}), \tag{2.9.3}$$

where

$$a_{nm} = \frac{\exp((W^{\text{Q}} h_n^{\text{FE}} + b^{\text{Q}})^T (W^{\text{K}} h_m^{\text{FE}} + b^{\text{K}}))}{\sum_{o=1}^{N} \exp((W^{\text{Q}} h_n^{\text{FE}} + b^{\text{Q}})^T (W^{\text{K}} h_o^{\text{FE}} + b^{\text{K}}))}, \tag{2.9.4}$$

resulting in the set of all latent view features $S^{\text{SA}} = \{h_n^{\text{SA}} \in \mathbb{R}^L\}$ with $n \in \{1, 2, ..., N\}$ after the self-attention correlation layer. Here, $W^{\text{Q}}, W^{\text{K}} \in \mathbb{R}^{D^{\text{SA}} \times L}$, $b^{\text{Q}}, b^{\text{K}} \in \mathbb{R}^{D^{\text{SA}}}$,

$W^{\mathrm{V}} \in \mathbb{R}^{L \times L}$ and $b^{\mathrm{V}} \in \mathbb{R}^{L}$ are trainable parameters of the network, where L refers to the length of the input vector and D^{SA} is a hyperparameter set to 64 in this work. n, m and o are indices of the $N = 7$ views per MRCP.

2.9.2 Evaluation

Table 2.9.1: Classification accuracy of the evaluated architectures on the testsets over 5 rounds. Testset size is 10% of the full data of the 3T and 1.5T datasets. 'gain' refers to the improvement of the fusion-methods (architectures & metrics) over the single-view (SV) accuracy, where ∅ refers to the average improvement, ± refers to the standard deviations of improvements across methods, ↑to the maximum improvement and ↓to the minimum improvement.

3T	**s-att**	**lstm**	**-**	**SV metrics**	
max	82.4±3.2	82.4±2.0	82.4±3.2	**SV**	***78.8±2.9***
mean	81.0±4.4	**82.9±1.7**	82.0±3.3	**maj**	82.4±2.0
fc	79.0±4.4	82.0±2.8	82.4±3.2	**add**	81.5±3.7
lstm	80.5±3.0	82.0±1.3	**83.4±3.2**	**mul**	81.5±3.7
att	82.0±2.8	82.4±1.1	82.4±3.2	**max**	82.0±3.7
e-att	82.4±2.7	**82.9±1.7**	82.4±3.2		
gain	∅3.31	±0.90	↑4.60	↓0.21	

1.5T	**s-att**	**lstm**	**-**	**SV metrics**	
max	80.9±2.5	78.2±3.7	79.6±4.3	**SV**	***78.9±1.9***
mean	80.9±4.0	77.8±4.2	81.3±3.0	**maj**	81.8±3.7
fc	**83.6±2.0**	81.8±2.9	81.8±2.9	**add**	80.0±3.5
lstm	**82.2±4.4**	80.9±3.0	**85.8±1.2**	**mul**	80.0±3.5
att	80.9±3.4	76.4±3.4	80.4±4.0	**max**	79.6±5.3
e-att	79.6±2.9	79.1±4.6	81.3±3.4		
gain	∅1.81	±1.88	↑6.86	↓-2.48	

For the view-aggregation exploration, training of SVCNN and MVCNN is performed as described in 2.6.2. However, for every combination of view-fusion and view-correlation layer, the hyperparameters are tuned independently using 5-fold cross validation on the 3T data. The same hyperparameters are then applied to the 1.5T model. Freezing the weights of the FE in MVCNN ensures that all evaluated aggregation methods receive the same latent representations and that differences in performance are solely based on the respective view-correlation and view-pooling layer.

Table 2.9.2: Classification accuracy of the evaluated architectures using 5-fold cross-validation on the full dataset over 5 rounds (random seeds) per split. 'gain' refers to the improvement of the fusion-methods (architectures & metrics) over the single-view (SV) accuracy, where ∅ refers to the average improvement, ± refers to the standard deviations of improvements across methods, ↑to the maximum improvement and ↓to the minimum improvement.

3T	**s-att**	**lstm**	**-**	**SV metrics**	
max	84.7±2.6	84.6±2.8	84.3±2.5	**SV**	***81.5±1.7***
mean	**85.3±3.1**	85.1±2.9	85.2±2.9	**maj**	84.9±2.1
fc	82.7±2.5	85.1±3.1	84.9±2.5	**add**	**85.6±2.7**
lstm	83.2±2.1	84.8±2.6	84.4±2.9	**mul**	**85.4±2.7**
att	84.8±2.5	85.0±2.6	84.7±2.4	**max**	84.4±2.9
e-att	85.1±2.9	**85.3±2.5**	85.1±2.5		
gain	∅3.19	±0.60	↑3.81	↓1.24	

1.5T	**s-att**	**lstm**	**-**	**SV metrics**	
max	83.1±2.7	**83.4±2.9**	82.8±2.4	**SV**	***81.0±3.0***
mean	82.8±3.3	83.2±3.0	83.3±3.0	**maj**	83.2±3.6
fc	83.2±3.0	**83.4±3.4**	83.0±3.4	**add**	**83.4±2.6**
lstm	82.6±3.0	82.9±3.1	**83.5±3.3**	**mul**	83.3±2.4
att	83.1±2.7	82.7±3.5	82.7±3.0	**max**	82.5±2.6
e-att	83.0±2.8	83.0±3.3	**83.5±2.6**		
gain	∅2.13	±0.25	↑2.59	↓1.70	

Every combination is trained and evaluated five times with varing random seeds. Table 2.9.1 depicts the results of all aggregation methods. Training and test set correspond to those used in DeePSC, though for this evaluation, samples with small-duct PSC are still included, hence the observed shift in absolute performance. On the 3T dataset, applying view fusion methods increase the accuracy in all cases above the SVCNN baseline, with an average improvement across methods of 3.31±0.90 percentage points. On the 1.5T dataset, an average improvement of 1.81±1.88 is achieved. Maximum improvement is at 6.86 percentage points, however, in three out of the 22 methods, a decrease in accuracy can be observed. Since hyperparameters are only optimized on the 3T dataset and ported to the 1.5T data, the decrease might be caused by suboptimal hyperparameter settings. On both datasets, the MVCNN without view-correlation and LSTM-pooling yields the highest accuracy. When ranking the other MV architectures, no common pattern can be observed.

To verify if these findings hold true on a different sample distribution, the evaluation is repeated using 5-fold cross-validation on the full datasets, resulting in five different random splits of 80% training and 20% testset per dataset. Training and evaluation is repeated over five random seeds per split. Table 2.9.2 shows the mean accuracy on the testsets over all splits and rounds per architecture. Here, all fusion methods improve accuracy compared to the SVCNN baseline on both datasets, with an average improvement of 3.19±0.60 percentage points across methods on the 3T data and an average of 2.13±0.25 percentage points on the 1.5T data.

2.9.3 Discussion

The results on the initial datasplit as well as from the cross-validation study indicate that utilizing the combined information of a sample with any view-fusion method improves classification accuracy over the single image baseline in the SVCNN. When interpreting the results on a single dataset with a single datasplit, it might appear that specific view-fusion methods outperform others (Table 2.9.1. While this pattern already does not hold true when moving to another dataset, the differences in performance of view-fusion methods nearly fully vanish when averaging over the results of multiple splits of training- and testset of the same data, as shown in Table 2.9.2. This is highlighted by the increased standard deviations of improvements across methods inside a static datasplit compared to those in the cross-validation case, which indicates a high correlation of the optimum method to the actual sample distribution in training- and testset, rather than a universally best method. When putting the results into relation to the respective model's complexity, there is no visible performance gain whatsoever from utilizing the view-correlation layer with self-attention or a LSTM for more sophisticated architectures. Also the simple max and mean view-pooling as well as the SV score fusion metrics are able to perform on par with the more complex fusion methods, like the proposed highly parameterized element-wise attention-pooling.

In conclusion, it can be stated that utilizing the combined information of multi-view data with any form of view-fusion is able to improve performance compared to single-view classification. With the MRCP data used in this work, the universal performance of a specific view-fusion method appears unpredictable and highly correlated to the underlying sample distribution into training- and testset. When averaging the results per fusion-method over different data splits, the differences between them become

negligible. With this, and according to the theorem of Occam's razor, a recommendation for usage of the simpler methods and architectures, that require no to little effort in finetuning of hyperparameters, can be made. Applying view-fusion metrics to the SVCNN provides the best ratio of simplicity and performance, since this requires no changes to the original SVCNN architecture. If applying a fusion metric is not an option, for example if the network is further extended after the view-fusion to perform survival analysis or other tasks in an end-to-end fashion, MVCNN pooling methods that do not rely on additional hyperparameters, like max or mean pooling, should be preferred. An arguable benefit of the more complex attention pooling is the ability to extract information about the impact and importance of every view for the models prediction, by analysing the respective attention weights, thereby increasing model interpretability. The additional view-correlation layer (i.e. bi-directional LSTM or self-attention) should be omitted, since it introduces unnecessary complexity to the model without improving its predictive performance. However, this might not be true for models trained on significantly larger datasets, as has been shown in the eCareNet proposed by Dietrich et al. [29].

At this point it is important to note an obvious limitation of this study, namely that even though extensive hyperparameter tuning is performed, the search space is not explored to its full extend for every dataset and -split. On the small and sample-sensitive MRCP datasets used in this work, 5-fold cross-validation tuning results vary heavily between splits and do not guarantee a matching test set performance. This is making heavy hyperparameter tuning on such small datasets only useful to a limited extend.

Given the high impact of MRI data preprocessing on overall performance, as depicted in the ablation study in 2.7.4, future work on MRCP multi-view image data should focus their attention more on dataset homogenisation and preprocessing rather than the optimization of the view-fusion architecture. This is also emphasized beyond the specific application of this work by other recent research, like the "data-centric AI movement" [87].

2.10 Conclusion

In conclusion, this work demonstrates that automated classification of PSC-compatible findings based on 2D-MRCP using multi-view deep learning algorithms is achievable with high accuracy for both 1.5T and 3T. The proposed DeePSC model scores higher than the mean classification provided by four radiologists at different levels of training and with varying experience with respect to accuracy, F1-score, specificity and sensitivity (**RQ-1**). It furthermore excels in the classification of previously unseen data from a different manufacturer's MRI machine. The ablation study on both internal datasets and the external vendor validation data reveales how this is enabled by the proposed preprocessing pipeline, consisting of histogram equalization, quantile-based intensity normalization and random data augmentation (**RQ-2**). Examination of the GradCAM activation maps reveal high activation in the relevant region of the biliary tree, positively contributing to the explainability and trustworthiness of the model's prediction (**RQ-3**). Evaluating the aggregation of multiple MRCP images inside and outside the proposed deep learning network leads to the conclusion that utilizing the combined information of all views consistently improves performance over predicting on individual images. However, the influence of exact choice of aggregation method is negligible in comparison to the overall gain. Given the aim of developing a transparent model and the inherent notion of importance per image in the attention-based view fusion method, this choice is the most interpretable of the evaluated options (**RQ-DeePSC**). After further training the network with the inclusion of controls with other liver and bile duct diseases, DeePSC may in the future provide valuable clinical decision support to radiologists, reduce inter-reader variability and therefore contribute to the early and precise diagnosis of PSC based on 2D-MRCP.

3 Prostate Cancer Aggressiveness Index - PCAI

3.1 Introduction

Prostate cancer (PCa) stands as a significant public health concern worldwide, representing a challenge in both clinical management and public health care systems. PCa ranks among the most prevalent cancers in men, with approximately 1.4 million new cases worldwide each year with its incidence steadily increasing over the past decades [88]. Due to the wide variety in growth rates of PCa, histopathology plays a central role in the diagnosis and management of PCa. An essential part of PCa diagnosis is the manual inspection of its severity through biopsies and histopathology. The current gold standard for PCa grading is the ISUP score that is determined by the International Society of Urological Pathology [89] and based on Gleason grades [90] derived from histological examination of prostate biopsies. Gleason grades range from 1-5 with increasing cancer severity to predict disease aggressiveness. For a biopsy, the most frequent and worst Gleason grades are combined to form a Gleason score (e.g. 3+4). ISUP then rearranges the Gleason scores into five ISUP grades, assigning the groups 1 (3+3 and below), 2 (3+4), 3 (4+3), 4 (3+5, 4+4, 5+3) and 5 (4+5, 5+4, 5+5). These categories are used to guide the urologist in treatment decisions. Unfortunately, even between expert pathologists the concordance in Gleason grading suffers from high inter-observer variability leading to possible over- or under-treatment due to the subjective nature of manual assessment [91]. Recent advancements in digital pathology, including the introduction of high throughput digital slide scanners, hold the potential to improve histopathological evaluation of PCa samples. The possibility to use deep learning (DL) based algorithms to automatically analyze pathological samples is not only time and cost effective but offers the potential for a standardized, objective, and accurate evaluation, providing crucial insights into tumor characteristics, aiding in treatment decisions, and assessing disease progression. This offers the opportunity to an efficient and reproducible histopathological assessment, thereby optimizing diagnostic accuracy and streamlining workflows in PCa diagnosis and care. Many automated DL based PCa grading approaches are based on the traditional Gleason or ISUP score of slide biopsies. Several of these algorithms have attained pathologist-level performance in these tasks [92–94]. Model training on ISUP score, however, is limited in its usefulness, as it distinguishes only five subgroups, one of which is characterized by a very good (ISUP 1) and three of which are characterized by a bad prognosis (ISUP 3-5). Moreover, models trained on human ISUP annotations will replicate human error, which is why recent approaches have

shifted towards the more interpretable prediction of objective endpoints, such as time to relapse (e.g. biochemical recurrence or cancer-related death). Some of these studies concentrate on modeling the probability of relapse-free survival over time [29], while others predict the probabilities of relapse up to one or multiple fixed time points [95]. While many promising steps have been conducted in the automated assessment of PCa histopathological data, four prominent challenges still require further attention: Robustness, interpretability, trustworthiness, and above ISUP-level grading performance [12]. The biggest hurdle for automated PCa grading, possibly, is the variation of histopathological protocols, which can result in a lack of robustness of algorithms [12, 17, 96, 97]. Processing tissue for digitization consists of several steps: tissue formalin fixation, paraffin embedding, sectioning, staining and digitization with a slide scanner. Each step involves numerous parameters that can vary between clinics, research institutions and even within the same lab, leading to variations in the appearance of the tissue in the images. Especially AI-based PCa grading seems to be negatively affected by data variation [12], while PCa cancer detection seems to be robust to data variation and is already in clinical use [98, 99]. Given the potential data variation-induced degradation of the predictive performance of AI-based PCa, concepts to improve model trustworthiness have been recently proposed [100]. A trustworthy model that quantifies its confidence or credibility in a prediction allows for the deferral of the PCa grading of problematic samples to a human expert. Of note, a systematic investigation of the impact of data variation on automated PCa grading, as for example conducted for breast cancer [17], is still missing.

This work presents the Prostate Cancer Aggressiveness Index (PCAI), a novel end-to-end PCa risk assessment model that addresses four essential pillars of clinical applicability, robustness, interpretability, trustworthiness, and a PCa grading performance that exceeds the human annotated ISUP score. PCAI utilizes patient relapse information as an objective measure for cancer aggressiveness and is trained on one of the largest and most diverse PCa histopathology datasets collected to date, containing six cohorts with over 25,591 patients, 83,864 images, and 5 years of median follow-up from 5 different centers and 3 countries. An important part of this data is a unique high-variance cohort of 8,157 patients with 28,236 scanned tissue microarrays (TMAs) with variations in sample thickness, staining time, and scanner device. Using this dataset, the robustness of AI-based PCa grading to data variation is systematically evaluated. To this end, first a baseline model BASE is derived, trained on a single TMA data source. Severe performance degradation is observed in the

BASE model on data deviating from its training distribution. PCAI then builds on this BASE model by incorporating several algorithmic adaptations, namely joint domain adversarial training and color adaptation for robustness, credibility estimation for trustworthiness, and cancer localization for interpretability, to exceed its performance on all evaluated datasets. Finally, PCAI outperforms human annotated ISUP grading in both the Concordance Index (C-Index) and area under the receiver operating characteristic curve (AUROC) on unseen and external TMA spot and biopsy images.

PCAI was developed in collaboration with Patrick Fuhlert.

3.2 Foundations

In this chapter, the medical foundations of prostate cancer, its diagnosis using histopathological analysis as well as the clinical workflow to derive at a treatment decision for the patient, are described. Furthermore, the methodologies of survival analysis and the corresponding quantification of cancer aggressiveness based on patient outcome information are outlined. Finally, the Efficientnet CNN architecture as well as the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction technique, which are used in the course of this work, are introduced.

3.2.1 Prostate Cancer

Prostate cancer (PCa) is the second most prevalent cancer in men, with over 1,414,000 new diagnoses worldwide in 2020 and a rising incidence with higher age [88]. The prostate, which is located in the male pelvis below the urinary bladder and immediately anterior to the rectum, consists mainly of glandular tissue that contributes about a third of the total seminal fluid and promotes a healthy alkaline state of the semen [101]. The median age for initial diagnosis of PCa in the United States is 66, and the overall risk to develop a clinically significant cancer over the course of ones life is estimated at 11.6% [102]. PCa is a very slow growing type of cancer, with a 5-year survival rate of over 97% if the cancer did not spread to other parts of the body at time of diagnosis [103]. However, if distant metastasis are present, 5-year survival rate drops close to 30% [104]. Initial or early symptoms are very rare, though in later stages it may cause fatigue due to anemia, bone pain, paralysis from spinal metastases, and renal failure from bilateral ureteral obstruction [102]. To detect PCa in its early stages before it can cause symptoms, most developed countries offer a PCa screening programm. In Germany, these consist mostly of a digital rectal exam (DRE), while in the United States, measuring the prostate-specific antigen (PSA) value is more common. PSA is a protein produced in the prostate glands, that can be measured in the blood and which expresses elevated levels in case of PCa. Healthy levels of PSA are considered around 4 ng/ml. However, PSA screening lacks specificity, since elevated levels can also be caused by an enlarged prostate, which might lead to overdiagnosis. Despite this, it is still the most commonly used and recommended procedure for early detection of PCa [105].

If a DRE or elevated PSA levels hint at the presence of PCa, the gold standard for con-

firmative diagnosis is a a transrectal ultrasound-guided (TRUS) biopsy of the prostate [106]. Newer diagnostic modalities include high precision MRI-guided fusion biopsies or non-invasive MRI-based Prostate Imaging Reporting and Data System (PI-RADS) scoring. Up to 14 tissue cores are extracted during a single biopsy [107]. These cores are then embedded and fixated with paraffin, cut into slices with thicknesses between 1.0 - 10.0 µm and stained with hematoxilin and eosin (H&E) [108]. Hematoxilin binds to acidic components, like cell nuclei, and colors them purple, while eosin binds to basic components and colors them pink. This allows for easier assessment of the tissue through light microscopy by the pathologist. Based on the grade of mutation visible in the cancerous tissue, the examiner assigns a Gleason grade, as explained in the following section. This histopathological grade is then, in combination with other clinical factors, used to derive a treatment decision for the patient. Possible treatments include active surveillance, referring to no immediate action but a close observation of cancer development, hormonal therapy, radiotherapy or radical prostatectomy (RP) [109]. RP refers to the surgical removal of the full prostate. This operation poses major risk of undesired side effects like impotence and incontinence, highlighting the need for accurate patient risk assessment to avoid unnecessary operations [110]. Even after removal of the full prostate, the cancer can spread again. This manifests either in the development of metastases that are visible during radiographic examination or by biochemical recurrence (BCR). The latter points to a recurrence of the disease which, even though it remains radiographically invisible, is indicated by a sudden rise in PSA levels during post-operative monitoring [102]. After RP, histological re-assessment of the removed prostate tissue is performed, where the pathologist analyzes the removed tissue under the microscope and assigns a Gleason grade. This allows for a more thorough assessment than initial needle biopsy and forms the basis for potential additional (adjuvant) treatment decisions after the operation. Besides the described re-assessment in the clinical workflow, the removed prostate tissue can be used to derive tissue microarray (TMA) spots, mostly for teaching and research purposes. To this end, the removed tissue undergoes similar histological processing as the needle biopsy. However, after embedding the full prostate in paraffin, tissue cores are extracted with a hollow needle with a diameter of 0.6 to 2.0 mm [111]. Multiple such extracted cores (of multiple patients) are arranged in a grid-like order in another paraffin block. After this, the blocks are cut into sections of 1.0 - 10.0 µm thickness and stained with H&E, analogue to the processing of the needle biopsy. Figure 3.2.1 depicts the typical workflow from a suspicious PCa screening

over histopathological assessment to treatment decision.

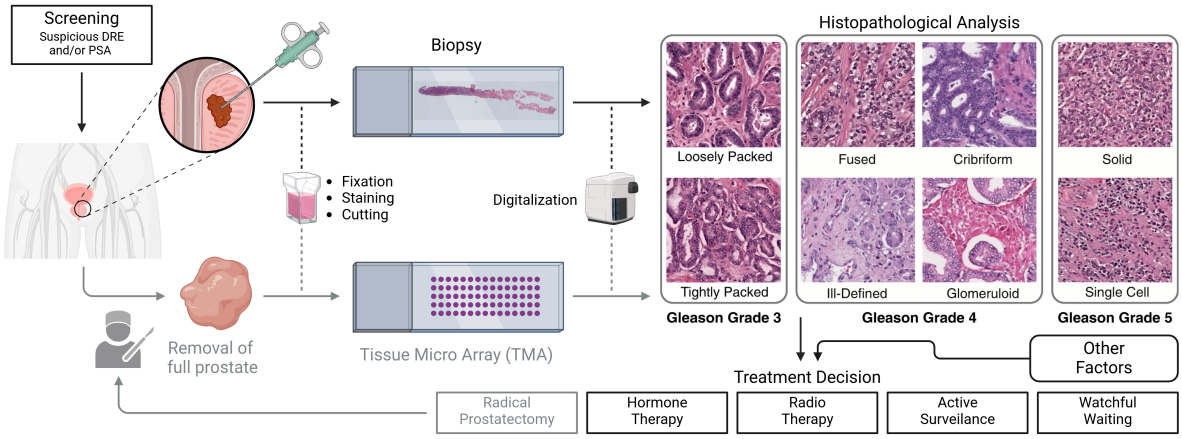

Figure 3.2.1: Workflow from PCa screening to treatment decision. If the screening results in a suspicion of prostate cancer, a biopsy of the prostate tissue is taken. The extracted tissue is further embedded in paraffin, stained with hematoxilin & eosin (H&E) and cut into slices. It is then either assessed directly by the pathologist or digitized using a histopathological scanner for digital evaluation. The pathologist assigns a Gleason grade, ranging from 1-5 based on the grade of visual mutation of the tissue under the microscope. Depending on the Gleason score, as well as other patient related factors, a treatment decision is derived. In case of a radical prostatectomy (RP, depicted in grey), the full prostate is removed. Afterwards, the removed prostate tissue is processed similar to the biopsy tissue for further histopathological analysis, though it is cut and arranged in the form of multiple round spots in a tissue microarray (TMA). Gleason patterns are adapted from Lawson et al. [112] under the Creative Commons Attribution 4.0 International License at `https://creativecommons.org/licenses/by/4.0/`

Prostate Cancer Grading

The Gleason grading system was developed by Dr. Donald Gleason in the 1960's and remains the gold standard for histological PCa risk prediction to date [90, 113]. The Gleason grade is based on the amount of mutation visible by the architectural arrangement or pattern of the glands of the sliced and stained tissue under the microscope. Different grades are assigned by the pathologist to different areas of the tissue, where Gleason grade 1 refers to an almost normal glandular appearance, whereas in Gleason grade 5 no more glandular structure is visible, and only sheets of abnormal cancer cells remain. In Figure 3.2.1, exemplary tissue patches for Gleason grade 3-5 are depicted. In order to derive a Gleason score for the patient based on their

biopsy, the primary and secondary Gleason grade is combined, referring to the most prevalent and the most severe pattern visible in the tissue. So if in a biopsy mostly Gleason 3 patterns are visible, but there are small amounts of Gleason 5, the total score would amount to 3+5. If only a single pattern is visible overall, e.g. only Gleason 4, the score would be written as 4+4. Notably, for prostate tissue analyzed after radical prostatectomy, i.e. TMA spots, the definition for the secondary grade changes from "most severe" to "second most prevalent". In both cases, the Gleason score can also be reported as the sum of both patterns, e.g. Gleason score 3+5=8. However, this summation leads to ambiguity, especially for Gleason score 7, where patients with Gleason score 3+4 express significant differences in disease progression to Gleason score 4+3 [114]. To account for the lack of resolution, the updated ISUP scoring system was proposed at the International Society of Urological Pathology consensus conference in 2014 [89]. The ISUP score ranges between 1 and 5 and transforms all combinations of Gleason grade 3 to 5, with emphasis on better distinguishability in the Gleason grade 3 and 4 groups. Since Gleason grades 1 and 2 are considered benign, these grades are neglected in the ISUP scoring system. Table 3.2.1 depicts the relationship between the Gleason grades, Gleason sum and ISUP.

Table 3.2.1: Translation of Gleason grades to Gleason sum and ISUP score [89].

Risk group	Low	Intermediate		High					
Gleason grades	$3+3$	$3+4$	$4+3$	$4+4$	$3+5$	$5+3$	$4+5$	$5+4$	$5+5$
Gleason sum	6	7		8			9		10
ISUP score	1	2	3	4			5		

Integrative Quantitative Gleason

Sauter et al. proposed a more fine-grained system, called the integrated quantitative Gleason (GIQ), which integrates information about the quantities of the Gleason patterns into account, instead of rating only their presence [115, 116]. In detail, the GIQ is derived from the percentages of Gleason patterns 4 and 5 in the prostate tissue, with additional emphasis on the Gleason 5 pattern. In detail, the GIQ is calculated as

$$\text{GIQ} = \text{GG4}_{\%} + \text{GG5}_{\%} + 10 \cdot \mathbf{1}_{\text{GG5}_{\%}>0\%} + 7.5 \cdot \mathbf{1}_{\text{GG5}_{\%}>20\%}. \tag{3.2.1}$$

Here, $\text{GG4}_{\%}$ and $\text{GG5}_{\%}$ describe the ratio of Gleason pattern 4 and 5, respectively, to the total cancerous tissue of a sample and $\mathbf{1}_X$ is the indicator function. This results in the overall GIQ score ranging from 0 (for 3+3) to 117.5 (for 5+5). The authors

claim that by using this continuous scoring system, increased clinically relevant morphologic information is captured over the standard Gleason grading, especially in borderline cases [115]

Digital Pathology

While histopathological analysis of the formalin fixed, paraffin embedded, cut and stained tissue samples is typically performed by the pathologists through examination under the microscope, in recent years digitalization has found its way into the pathology department. With this, so called digital whole-slide images (WSI) of the prepared biopsy or TMA slides can be created using specialized histopathological scanners [117]. Those WSIs can span several gigapixels in resolution, depending on the microscope objective used for scanning. A typical objective achieves a maximum magnifciation of 40x, which results in a pixel length of roughly 0.25 um. Image files are stored in a multilayered pyramidial format of multiple lower magnification levels, to enable computationally efficient real-time viewing across different resolutions. Notably, due to differences in acquisition hard- and software, different scanner devices introduce intrinsic domain shifts to the images that might exceed apparent differences such as color or resolution [118]. These domain shifts pose a major challenge to deep learning applications in the field of digital pathology.

3.2.2 Survival Analysis

Survival analysis, or more generally, time-to-event-analysis, is a field of statistics that deals with the estimation of the time until a specific event of interest (e.g. death) occurs for an individual or a population. What differentiates survival analysis from regression is the integration of censored data. Right-censored data consists of individuals that did not experience the event of interest by the end of the observation period or that are lost to follow-up prior to the end of the observation period and from there on in an ambiguous and uninformative state. Left-censored data consists of individuals where an event is observed, but the exact time of occurrence is unknown. This work only considers right-censored data and refers to those as "censored" in the following. In the context of prostate cancer disease progression, a patient that dies 5 years after initial diagnosis would be considered an un-censored patient with an event after five years. A patient that survived for four years, but then dropped out

of the study (lost follow-up), would be considered censored after 4 years. If a patient survived until the very end of the observation period, it would be considered censored at that point. A classical, non-parametric approach to model the survival characteristics of a population is the Kaplan-Meier survival curve analysis [119]. It estimates the survival function $S(t)$, i.e. the probability to survive at least until the time t, as

$$\hat{S}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right), \tag{3.2.2}$$

where $\hat{S}(t)$ is the estimated survival probability at time t, d_i is the number of events (deaths) at time t_i, and n_i is the number of individuals that survived (not yet censored nor experienced an event) up to t_i. The survival probability over time of the estimated survival function $\hat{S}(t)$ can then be visualized as a survival curve. Confidence intervals can further be calculated using the exponential Greenwood formula [120]. Since the Kaplan-Meier method estimates a single survival function for the whole population, it is mostly used for comparing the survival probabilities of different groups of patients, e.g. for analysing the impact of a specific treatment on a group of patients (Figure 3.2.2).

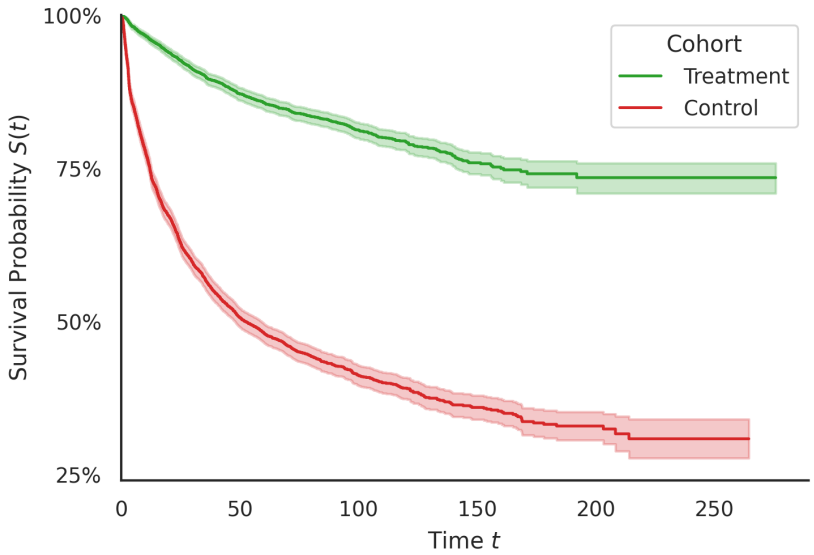

Figure 3.2.2: Kaplan-Meier curve with survival probability $S(t)$ over time t for an exemplary treatment and control cohort. If the treatment is effective, $S(t)$ of that cohort declines slower than for the control group.

In contrast to this population-based survival analysis, individual survival analysis deals with the estimation of survival probability for a single sample or patient based on a set of given input features. This can either be done by directly predicting the individual survival probability $S(t)$ over time, or by prediction of a risk score R that stratifies patients according to their observed follow-up information. In the scope of this work, only risk score based individual survival analysis will be considered.

3.2.3 Quantification of Cancer Aggressiveness

The aim of this project is to build a model that predicts a score that correlates with cancer aggressiveness. However, such a score or notion of aggressiveness is not directly available in patient follow-up data. To transform the clinical follow-up information into a quantifiable measure correlating with cancer aggressiveness, the notion of "relapse-free survival time" is utilized. For this, the indications of PCa-related death, developing a metastasis, and, in the case of patients that underwent radical prostatectomy, biochemical recurrence (BCR) are considered to indicate a worsening of the disease and defined as a relapse. If a patient experienced a relapse shortly after acquisition of the biopsy or radical prostatectomy, the tissue present in the acquired samples is expected to correlate with a highly aggressive cancer, while longer relapse-free survival times are expected to correlate with less aggressive forms of cancer. Following the notion of survival analysis introduced in the previous section, the relapse can also be described as "event". If the patient does not experience an event over the course of the follow-up, he is considered "censored" with a relapse-free survival time until the time of the last known contact in the follow-up.

3.2.4 EfficientNet

Efficientnets are a family of deep learning architectures that are developed by Tan et al. with the aim to achieve state-of-the-art performance while being computationally efficient [121]. The authors claim that previous work mostly focused on scaling up a single dimension of the underlying architecture to increase performance with increasing computational budget, like the depth dimension when stacking more layers in the ResNet family of architectures. They propose a novel principled compound scaling method, that uniformly scales network width, depth and input resolution with a fixed set of scaling coefficients with increasing computational resources. Furthermore, they

develop a novel baseline CNN and use their compound scaling method to derive a family of models at different scales, ranging from Efficientnet-b0 as the smallest instance to Efficientnet-b7 as the largest. Efficientnet-b0 surpassed the widely used ResNet-50 architecture by 1.1% in Top-1 Accuracy on ImageNet, while utilizing only a fifth of the parameters. The main building block of Efficientnets is the mobile inverted bottleneck convolution (MBConv) proposed by Sandler et al., to which the authors additionally add squeeze-and-excitation optimization proposed by Hu et al [122, 123]. The authors claim that these blocks reduce parameters and computational cost without sacrificing performance. In this work, the smallest instance of the Efficientnet family, Efficientnet-b0, is utilized.

3.2.5 Uniform Manifold Approximation and Projection (UMAP)

UMAP (Uniform Manifold Approximation and Projection) is a dimensionality reduction technique designed to capture the underlying structure of high-dimensional data by preserving both local and global relationships [124]. In simple terms, UMAP constructs a high dimensional graph representation of the data then optimizes a low-dimensional graph for maximum structural similarity. Optimization is performed by stochastic gradient descent. To this aim, UMAP assigns the likelihood of being connected to data points in the high dimensional space, based on their actual multi-dimensional distance in the data and a user-definable parameter of nearest neighbors which defines the size of the local neighborhood to consider. This parameter controls how UMAP balances the local versus global structure in the data. A higher setting will push UMAP to capture more of the bigger picture, while smaller values aim to capture more detailed relationships. UMAP then constructs the high dimensional graph of the underlying data and assigns the likelihood of two points being connected as their edge weights. The low-dimensional graph is then iteratively adapted to minimize its discrepancy to the high-dimensional representation. UMAP aims to preserve local and global structure and provide a visually meaningful two-dimensional representation even in complex datasets. This sets it apart from principle component analysis, which works best only if the first two principle components account for the most variability in the data.

In more detail, let $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$ denote a dataset where n is the sample size and d is the dimensionality. UMAP then first constructs the high dimensional graph in the input space using the k-Nearest Neighbors (kNN) algorithm, with the default

value $k = 15$ [125]. The j-th neighbor of $\boldsymbol{x}_i$ is denoted by $\boldsymbol{x}_{i,j}$ and the set of neighbor points $\mathcal{N}_i$ of point $\boldsymbol{x}_i$ as $\mathcal{N}_i = \{\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,k}\}$. Then, the Radial Basis Function kernel is used to measure the similarity of points in the high dimensional space. This results in a probability $p_{j|i}$ for a point $\boldsymbol{x}_i$ to have the point $\boldsymbol{x}_j$ as its neighbor as

$$p_{j|i} = \begin{cases} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 - \rho_i}{\sigma_i}\right) & \text{if } \boldsymbol{x}_j \in \mathcal{N}_i \\ 0 & \text{otherwise} \end{cases}. \tag{3.2.3}$$

Here, $\|.\|_2$ denotes the ℓ_2 norm and ρ_i the distance from $\boldsymbol{x}_i$ to its nearest neighbor as

$$\rho_i = \min\left\{\|\boldsymbol{x}_i - \boldsymbol{x}_{i,j}\|_2 \mid 1 \leq j \leq k\right\}. \tag{3.2.4}$$

The scaling parameter σ_i is calculated such that the total similarity of point $\boldsymbol{x}_i$ to its k nearest neighbors is normalized to $\log_2(k)$, i.e. that σ_i satisfies

$$\sum_{j=1}^{k} \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{i,j}\|_2 - \rho_i}{\sigma_i}\right) = \log_2(k). \tag{3.2.5}$$

Then, equation 3.2.3 is symmetrized to derive the final measure of similarity of points x_i and x_j in the input space as

$$p_{ij} = p_{j|i} + p_{i|j} - p_{j|i}p_{i|j}. \tag{3.2.6}$$

Let then the corresponding set of points in the low dimensional embedding space be $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{p \times n}$ where p is the dimensionality of embedding space. Here, the probability that a point $\boldsymbol{y}_i$ has the point y_j as its neighbor is given by their similarity as

$$q_{ij} = \left(1 + a\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2^{2b}\right)^{-1}, \tag{3.2.7}$$

where $a > 0$ and $b > 0$ are hyperparameters determined by the user with default values of $a \approx 1.929$ and $b \approx 0.7915$ [125].

The graph in the low-dimensional embedding space is then iteratively adapted to maximize similarity to to the graph in the input space by minimizing the cost function

$$L(p,q) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(p_{ij} \ln \left(\frac{p_{ij}}{q_{ij}} \right) + (1 - p_{ij}) \ln \left(\frac{1 - p_{ij}}{1 - q_{ij}} \right) \right). \tag{3.2.8}$$

Throughout this work, UMAP representation with a local neighborhood of $k = 15$ samples will be utilized to provide visual representations of the underlying data.

3.3 State of the Art

The following section describes the current state of the art of deep learning applications in the overall field of digital pathology and the characteristic challenges associated with the processing of histopathological whole slide images (WSIs). It then further provides an overview of AI applications in the specific field of prostate cancer. Finally, a systematic review of the literature focusing on building robustness and generalizability of deep learning models for histopathological image data is conducted.

3.3.1 Deep Learning in Digital Pathology

With digitalization finding its way into clinical pathology in recent years, AI-based applications on histopathological images flourish accordingly, for many different tasks and diseases. However, processing of histopathological images incorporates various characteristics that set it apart from classical deep learning based image processing. The arguably biggest challenge is the sheer size of the WSIs that can span multiple gigapixels. A common method to deal with this in the literature is patch-based Multiple Instance Learning (MIL). Here, the WSI is cut into multiple equally sized patches with resolutions typically used in common encoder architectures like ResNet-50 and processed as a bag of image patches. However, the ground truth, like a cancer severity or the death of a patient in the future, does not correspond with an individual patch, but only with the full patch bag. Therefore, for training of the downstream task, a DL model potentially needs to process all patches at once per iteration. Since this is often not feasible due to computational constraints, different MIL approaches are taken in the literature.

Foundational work of MIL in the field of digital pathology was provided by Campanella et al. [93], who successfully trained a slide-level classifier for prostate, skin and breast cancer detection. They developed a CNN that processes individual patches and performs inference before every training step. Then, only the patches showing the highest predicted class probability are selected and used for training, by assigning the full image label to them. However, this requires a well initialized network and limits the actually utilized information per WSI at every training step to a small fraction of the patches, potentially neglecting valuable information. In contrast, in the attention-based MIL approach proposed by Ilse et al. [78], all patches per bag are processed in parallel in every training step. A learnable attention layer aggregates

the latent patch representations inside the network by assigning an attention weight to every instance and combining them by weighted averaging. The combined bag representation is then processed by the classifier. While this requires training of more parameters than in the method proposed by [93], it enables the network to dynamically decide which and how much information it utilizes for every WSI. Similarly, Lu et al. proposed CLAM, which incorporates multi-class attention-based patch aggregation branches, which, as the authors claimed, improves performance on multi-class classification tasks over the method proposed by Ilse et al. [126]. By visualizing the attention weights per patch, this method also includes an inherent visual interpretability. However, due to the computational demand, end-to-end training was infeasible and the authors relied on latent patch representations from an ImageNet pre-trained ResNet-50 [45, 55]. To improve the quality of patch embeddings, even though end-to-end training with an encoder is not feasible, self-supervised approaches are often utilized in the literature. Ciga et al. utilized the SimCLR method for patch-based contrastive pre-training of multiple ResNet-based encoder architectures on data from 57 unlabelled histopathological datasets [127, 128]. By training a linear classifier on the extracted patch features, they validated their method on various downstream tasks and found that contrastive pre-training on histopathological images significantly improves downstream performance over pre-training on ImageNet. They furthermore found that combining multiple multi-organ datasets with different types of staining and resolution properties improves the quality of the learned features. Only recently, Chen et al. proposed UMI as a general-purpose self-supervised model for computational pathology [129]. It consists of a ViT-Large vision transformer, pretrained under the DINOv2 student-teacher knowledge distillation method using over 100 million tissue patches from over 100,000 WSIs across 20 major tissue types [130, 131]. They evaluated on 33 clinical downstream tasks, using the attention-based patch aggregation method proposed by Ilse et al. for slide level tasks. Alternatively, the same authors proposed the Hierarchical Image Pyramid Transformer (HIPT), which utilizes the inherent pyramidial structure of WSIs to process the full image at multiple magnifications in parallel to create a single slide-level representation [129]. They compared their method to classical MIL aggregation methods and outperformed those among multiple tasks and datasets.

3.3.2 Deep Learning in Prostate Cancer

Risk Assessment without Images

Risk analysis in the field of prostate cancer ranges back to the late 1990s, where Kattan et al. proposed a regression based nomogram to model the 7-year recurrence-free survival probability for patients after radical prostatectomy based on various tabular clinical parameters [132]. These include, beyond others, the PSA value, the pathological Gleason score and information about leftover cancer at the surgical margins. In 2005, Stephensen et al. updated the nomogram to be predictive up to 10 years after radical prostatectomy [133]. In recent years, machine learning based methods for survival analysis on electronic health records emerged [134]. Here, Wang et al. published a meta-analysis of exisiting nomograms and newly developed machine learning models for predicting the risk of lymph node metastasis [135]. They found that predictive accuracy of the machine learning models is superior to existing clinically recommended nomograms, with the former achieving a concordance index on validation data of up to 0.862, while the latter achieve a maximum of 0.745.

Cancer and Gleason Prediction on Images

When utilizing histopathological images of PCa, most recent research in the field of deep learning focused on predicting the subjective Gleason grading or detecting cancerous areas. Foundational work on this task has been provided by Campanella et al. who trained a ResNet-34 model on over 8,000 prostate biopsy WSIs using their proposed MIL approach described in the previous section for binary cancer detection [93]. On roughly 1,800 internal and 12,000 external test images, they achieved an AUROC of 0.991 and 0.932, respectively. Bulten et al. trained an extended U-Net for segmentation of benign tissue and Gleason growth patterns on indvidual patches from 933 prostate biopsy WSIs [94]. By evaluating the normalised volume percentages of each growth pattern, they achieved an AUROC of 0.990 in differentiating benign from malignant biopsies, and an AUROC of 0.974 in differentiating Gleason group 3 from less aggressive forms. Using their model for semi-automatic labelling resulted in the PANDA dataset, which is used in this thesis and will be introduced in later sections. On TMA-spot images, Arvaniti et al. developed a CNN-based patch-wise Gleason classifier. Using a sliding-window approach, they derived pixel level Gleason heatmaps for 245 test images and achieved a human level inter-annotator agreement with two pathologists, measured by Cohen's quadratic kappa, of 0.75 and 0.71 [136]. Pantanowitz et al. developed an algorithm for cancer detection and Glea-

son classification on needle biopsy images [98]. For cancer detection, they achieved an AUROC of 0.991 on external test data. Most importantly, they evaluated their algorithm on 11,429 slides in routine practice, where it lead in 9% of cases to the ordering of additional slides, of which two cases resulted in a third opinion request. In one case, the algorithm detected a case of cancer that would otherwise have been missed by the clinician.

Risk Assessment on Images

Research on objective PCa risk assessment without predicting a subjective intermediate score like Gleason on histopathological images is rare, potentially due to the lack of sufficient patient follow-up information to derive an objective endpoint. To highlight is here the eCareNet developed by Dietrich et al. which models patients risk of relapse after RP by predicting individual 7-year survival curves per patient [29]. To this end, the authors used the time from RP to BCR as objective ground truth and trained their network on TMA-spot images from the same cohort used in the course of this thesis, which will later be referred to as UKE.first. They used a modified InceptionV3 encoder to extract patch features, a recurrent neural network to model the time dependency and the attention-based MIL approach of Ilse et al. to aggregate latent patch features [78, 137]. The predicted individual survival curves can also be transformed into a single risk score. With this approach, they achieved an AUROC of 0.77 and a C-Index of 0.74 on the test data, lacking behind only 3 and 2 percentage points to the risk stratification provided by ISUP scoring of the full prostate tissue. Closely related, Walhagen et al. developed a similar patch-based risk assessment network on the same cohort, however, by predicting the probability of experiencing a biochemical recurrence, metastases, or death from prostate cancer in the first five years after RP in a binary classification task [95]. Using the probability as a continuous risk score, they achieved an AUROC of 0.79 on the test data, lacking 2 percentage points behind ISUP. Both the works of Dietrich et al. and Walhagen et al. provided foundational research to the PCAI model presented in this thesis. To the best of my knowledge, the only other work to perform deep-learning-based risk assessment of PCa patients based on histopathological images was presented by Pinckaers et al. [138]. They transformed the years to biochemical recurrence, metastasis or PCa-related death into a multiclass endpoint, with class 0 referring to relapse within the first year, and class 4 to a relapse after the fourth year. Then, a ResNet50-D based classifier is trained on centered cropouts from the full TMA-spot images, neglecting the need for MIL patch aggregation. Using a nested-case control study, they showed that predictions provided by their

model provides strong correlation with relapse of PCa on test datasets provided by the Johns Hopkins Hospital in Baltimore and the New York Langone Medical Centre. Data from both test cohorts used by the authors will also be utilized in the course of this thesis and referred to as JHU and NYU dataset, respectively.

3.3.3 Building Robustness in Digital Pathology

In addition to the computational difficulties that arise through the sheer size of the images, robustness and generalizability of deep learning models poses a major challenge when utilizing histopathological image data. As Brehler et al. found, differences in the data acquisition procedure, like longer or shorter staining times, varying tissue thickness or different histopathological scanners are detrimental to the performance of deep learning models when exposed to unseen data [12]. Wilm et al. highlighted how especially scanner-induced domain shifts negatively impact model performance on the downstream task of skin cancer segmentation. Their work builds on findings from Stacke et al. who found that the sensitivity of a deep learning model to a covariate shift in the data should be quantified by measuring the discrepancy of domains in the model's latent space [97]. Overcoming this sensitivity to covariate shifts in histopathological image data is a major topic of research and is central to the yearly MItosis DOmain Generalization Challenge (MIDOG) hosted by the Medical Image Computing and Computer Assisted Intervention Society (MICCAI) [139].

Research on methods to increase robustness of deep learning models for histopathological image processing and to mitigate the influence of covariate or domain shifts can roughly be divided into four categories:

Reduce Data Diversity During Inference

The most obvious method to overcome a covariate shift in the data is to directly adapt the images to be more similar before forwarding them to a deep learning network. Popular approaches are the H&E stain color separation and normalization proposed by Macenko et al. and Vahadane et al. [140, 141]. This aims to reduce color variation across images specifically in the range of the purple and pink regions typical for the hematoxilin and eosin staining. Dietrich et al. used the Macenko stain normalization together with a histogram matching approach to successfully increase robustness of eCareNet, as introduced earlier, on multiple datasets expressing a covariate shift

to the training distribution [142]. Besides classic algorithms, Generative Adversarial Networks (GANs) can be utilized to map unseen data from its source domain to the target domain of the downstream model. Thebille et al. developed a CycleGAN to transfer domain specific information of TMA spot data collected at two different hospitals [143]. By adapting images from the unseen domain to the training domain, they improved classification accuracy for the downstream task of Gleason score prediction by 14%. However, adaptation of images, especially using generative deep learning models, poses the risk of introducing undesirable artifacts, such that these methods should be utilized with caution.

Increase Data Diversity During Training

Random data augmentations are a standard technique in the field of machine learning to improve robustness by increasing the variety of data seen by the model during training. Khan et al. evaluated multiple color augmentation and stain normalization strategies indivudally and in combination on the downstream task of classifying benign vs. malignant breast cancer tissue [144]. They found that both stain normalization and random color augmentations are able to improve performance on external data individually and even more so when used in combination. In a larger study conducted by Tellez et al., various combinations of random stain color augmentations and stain color normalization techniques are evaluated on four downstream tasks on histological data from nine different centers [145]. They found that augmenting the color is fundamental for improved performance on images from unseen centers and that stain normalization is neglectable when proper augmentations are in use.

Train Model to Become Domain-Agnostic

Besides the aforementioned data-centric approaches to robustness, model-centric approaches can be taken. One such method often utilized in the literature is domain adversarial training. Here, a second task, like domain classification, is trained in parallel to the main task. By inverting the gradient of the loss of the secondary task, it is aimed to actively "un-learn" to extract discriminative domain-specific information required for that task in the shared part of the model, i.e. make the model domain-agnostic. This approach will be explained in more detail in later sections. Wilm et al. proposed a domain-adversarial trained RetinaNet for mitotic figure detection on H&E stained

breast cancer images as the reference approach to the 2021 MIDOG challenge [17, 146]. Using four different histological scanners as domain ground truth for adversarial training together with applying strong data augmentation, they improved over the baselines of applying only augmentations or no augmentations during training, proving the positive influence of their scanner-adversarial training. Similar findings are reported by Lafarge et al. who analyzed the effect of adversarial training with acquisition center as domain ground truth for mitosis detection and tissue type as domain ground truth for nuclei segmentation [147]. On both tasks, domain adversarial training improved performance on unseen domains. Recently, a thorough analysis of various stain normalization, color augmentation and adversarial training methods has been conducted by Marini et al. [148]. To highlight is here that they additionally analyzed H&E-adversarial training, where instead of discriminating pre-defined domains in the data, regression on the H&E matrix per image, as derived by Macenko's method, is performed. With this, they aimed to make the model agnostic to differences in the staining color and significantly outperformed domain adversarial training with the acquiring center as the domain ground truth.

Transfer Learning

Finally, transfer learning is a common approach when a big and more representative dataset and a smaller target-domain dataset are available. Here, a model that was previously trained on the large representative dataset is fine-tuned on a smaller dataset of the target domain. Aubreville et al. significantly improved mitotic figure detection performance on the target domain when utilizing transfer learning [149]. However, transfer learning is specific to a single domain and poses the risk of degrading model performance on the initial source domain.

3.3.4 Summary

In summary, the state of the art highlights various challenges when utilizing deep learning with histopathological images. To enable processing of the large and heterogeneously shaped WSIs, MIL methods are utilized in the literature. In the course of this work, the attention-based MIL approach proposed by Ilse et al. will be used [78]. This has already proven successful by Dietrich et al. and Walhagen et al., who performed similar experiments on the same cohort and serve as foundational research for

this thesis [29, 95]. In terms of robustness, a multitude of approaches is taken in the literature to overcome the detrimental effects of covariate shifts in the data caused by differences in acquisition protocols, staining times or scanner devices. Most research found that heavy data augmentation paired with domain adversarial training provided the best generalization. Building on this, both approaches will be applied in this work. Stain normalization mostly provided only insignificant or no improvements when applied in combination with aforementioned methods and will be neglected in this thesis. However, a test time image color adaptation procedure will be utilized to further decrease variance before inference and therefore model robustness and predictive accuracy.

3.4 Data

This section introduces the different patient cohorts used in this work, together with their corresponding tissue microarray (TMA) and biopsy data, as well as the respective protocols used for image acquisition. Furthermore, the patient demographics and characteristics as well as image properties of all datasets are described.

3.4.1 Data Acquisition and Composition

The primary aim of this work is to build an algorithm that is robust, explainable, trustworthy and exceeds human PCa grading performance, which necessitates a large, heterogeneous dataset with rich patient follow-up (FU) information of sufficient quality. Gold standard for clinical diagnosis is ISUP grading of preoperative biopsies, typically obtained through transrectal ultrasound-guided biopsy. Multiple tissue samples are collected from different areas of the prostate gland to improve cancer detection rates. After biopsy, specimens are formalin-fixed and paraffin-embedded to preserve structure and stained with Haematoxylin and Eosin (H&E) to enhance cellular visibility for pathologist examination. All biopsies in this work contain ISUP annotations by expert pathologists. In addition to biopsies, this work uses postoperative tissue microarrays (TMAs) from radical prostatectomies (RPs). TMAs consist of many small cylindrical representative samples, termed spots, that are extracted from paraffin-embedded tissue and are widely used in biomarker discovery and validation studies. TMA spots of this work typically have edge lengths of 3,000 to 6,000 pixels with resulting images that contain in the order of 10 million pixels and thus are much smaller than biopsies. Biopsies have long edge lengths in the order of 60,000 pixels with a total of up to 10 billion pixels per image and are therefore carefully selected and preprocessed to represent a patient's cancer status. All ISUP grade annotations for TMA spots were made by expert pathologists from multiple samples of the resected whole prostate to derive a patient-level annotation. Therefore, TMA spots might only partially capture a patient's cancer status, with the notable exception of the UKE.sealed dataset, which contains TMA spot-based ISUP annotations (see UKE.sealed section). To the best of my knowledge, this work collected the biggest and most heterogeneous histopathological PCa dataset to date, with a total of 81,572 TMA spot images and 3,388 biopsy images retrospectively collected from 25,591 patients of five different clinics with follow-up information of up to 23 years and a max-

imum of 8 images for a single patient. This dataset is divided into several subsets acquired with different parameters and used for training or testing the models on images with high variation, as depicted in Table 3.4.1 and Table 3.4.2. Datasets that are used to build and assess robustness of the proposed models are highlighted as "RB" in the tables. All image-level annotated, unseen datasets that are exclusively used for evaluation and benchmarking against human raters are highlighted as "BM". Detailed information on demographics and metadata distributions can be found in Table A.2.1. The largest subset is the UKE-high-variance cohort (UKEhv) provided by the University Medical Center Hamburg Eppendorf which contains 17,700 patients who underwent RP with a FU time up to 23 years. This unique dataset allows to assess differences in acquisition protocol parameters and represents the foundation for building a robust prediction model in this work. As described in Section 3.2.3, a quantifiable measure correlating with cancer aggressiveness that does not rely on subjective human annotations is derived from the patient follow-up information. For this, the objective endpoints of biochemical recurrence (BCR), developing metastasis (META), and PCa-related death (PCAD) are defined as events with corresponding event time relative to the date of RP for TMA spots or to the date of the biopsy procedure for biopsies as time-to-event. If none of those exist, the follow-up time is used as the censoring time. The same metadata and image quality-based filtering steps are applied to all datasets. In brief, patient inclusion is limited to patients that either experienced some kind of relapse or had at least 5 years of follow-up data available. Additionally, images with insufficient quality (e.g. too blurry or no full TMA spot visible on the image) are excluded from the analyses. Exact filtering and preprocessing steps will be described in detail in the subsequent Section 3.5. All datasets used in this study were collected in strict accordance with ethical guidelines and compliance regulations. Data collection was approved by the relevant institutional review boards or ethics committees. Informed consent was obtained from all participants involved in data collection processes or the need for informed consent was waived by the local ethics review board. Additionally, any identifiable information pertaining to participants was anonymized or de-identified prior to analysis. In the following, the datasets used in this study are explained in detail.

Table 3.4.1: Patient metadata for all PCa survival datasets. Age and PSA-level denoted as mean ± standard deviation. Survival and FU denoted as median. RB: Datasets used to assess and build robustness of the proposed deep learning models. BM: Datasets used to benchmark predictive performance of the proposed models against human annotators.

		Patients	Age [years]	PSA-level [ng/µL]	Censoring [%]	Survival [months]	Follow-up [months]	Tissue Acquisition
RB	UKE	8157	63.5 ± 6.1	10.6 ± 38.6	61.4	19.8	95.9	RP (TMA)
	NYU	158	60.9 ± 7.0	7.8 ± 6.8	70.3	46.2	213.5	RP (TMA)
	JHU	879	59.2 ± 6.3	11.8 ± 10.2	0.3	24.0	192.0	RP (TMA)
BM	UKE.sealed	826	unknown	unknown	unknown	unknown	unknown	RP (TMA)
	MMX	269	67.6 ± 8.8	19.8 ± 43.8	88.2	51.8	108.9	Biopsy
	UPP	123	unknown	16.3 ± 13.5	83.7	25.1	83.7	Biopsy

Table 3.4.2: Overview of the image datasets. RB: Datasets used to assess and build robustness of the proposed deep learning models. BM: Datasets used to benchmark predictive performance of the proposed models against human annotators.

		Images (Median per Patient)	Magnification (Resolution [µm/px])	Thickness [µm]	Staining Times [min]	Scanner Types
RB	UKE.first	8123 (1)	40x (0.25)	2.5	4:00H, 1:20E	AP
	UKE.second	7156 (1)	40x (0.25)	2.5	4:00H, 1:20E	AP
	UKE.scanner	8114 (1)	80x (0.125)	2.5	4:00H, 1:20E	3D
	UKE.thin	1602 (1)	40x (0.25)	1.0	4:00H, 1:20E	AP
	UKE.thick	1574 (1)	40x (0.25)	10.0	4:00H, 1:20E	AP
	UKE.long	1667 (1)	40x (0.25)	2.5	40:00H, 10:00E	AP
	NYU	506 (3)	20x (0.5)	5.0	unknown	AP
	JHU	3575 (4)	40x (0.23)	4.0	unknown	VE, HA
BM	UKE.sealed	4095 (6)	40x (0.25)	2.5	4:00H, 1:20E	AP
	MMX	578 (2)	40x (0.23)	unknown	unknown	HA, VE
	UPP	683 (5)	40x (0.25)	unknown	unknown	HA

3.4.2 UKE-high-variance Dataset

The UKEhv cohort provided by the University Medical Center Hamburg Eppendorf contains patients who underwent RP between 1992 and 2014 aged 63.8 ± 6.4 years at the UKE with a FU time up to 23 years. The cohort's observed median PSA level at the point of RP is 6.9 ng/mL (interquartile range of 4.8 to 10.5 ng/mL). In total, 17,700 patients are collected in the dataset providing 69,251 images. Patients received an annual follow-up [115]. PSA values were measured following surgery and biochemical recurrence was defined as a postoperative PSA of 0.2 ng/mL and increasing at subsequent measurements. The individual patient event label are extracted by combin-

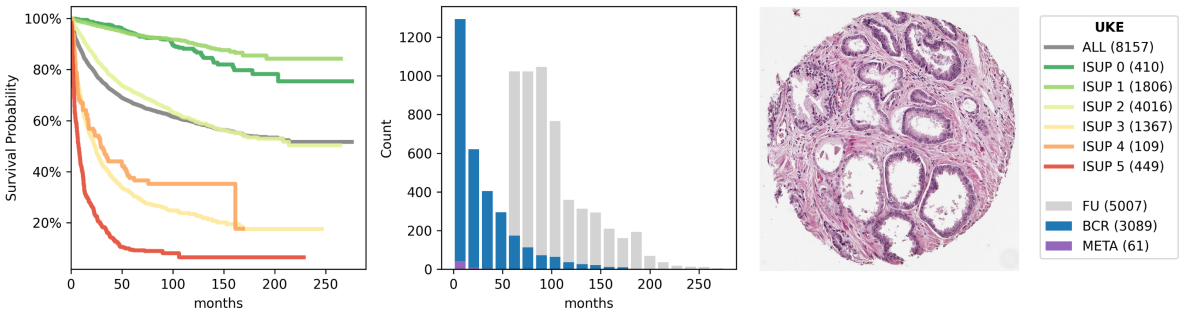

Figure 3.4.1: Kaplan-Meier curves for all patients as well as ISUP sub-cohorts of the UKEhv dataset (left). Distribution of event (BCR, META, PCAD) and censoring (FU) timepoints (middle). Exemplary image of the dataset (right).

ing biochemical recurrence (BCR), metastasis (META) or PCa-related death (PCAD) as an event with a duration from the date of RP to the first of the previously mentioned events. Patients without any of these events are considered censored at the last follow-up date. Further, this dataset includes some patients with healthy tissue who therefore did not obtain an ISUP grading. Building upon this rich information of 17,700 patients, a large variety of 69,251 high-quality images and spots were obtained from different protocols, which represent the foundation for building a robust prediction model in this work. ISUP grades were assigned by examining the whole prostate after RP for every individual patient. After filtering according to the aforementioned criteria (described in more detail in Section 3.5), 8,157 unique patients and 28,236 TMA spot images are included in the analysis. Figure 3.4.1 depicts event distribution and Kaplan-Meier curves for the finally included UKEhv patients. This extracted dataset consists of images with varying attributes, like multiple spots for the same patient, varying scanners, slice thicknesses and staining times and is, to my knowledge, the largest and most variant collection of TMA spot image data paired with rich follow-up data collected to date. The UKEhv data is divided into six sub-datasets, UKE.first, UKE.second, UKE.scanner, UKE.thin, UKE.thick and UKE.long, as depicted in Figure 3.4.2.

In the following, the inherent image characteristics of the UKEhv sub-datasets are described in depth. Note that all sub-datasets stem from the same patient population and a single patient can contribute images to multiple sub-datasets. Detailed patient-level information for the UKEhv sub-datasets can be found in Table A.2.2.

UKE.first: This sub-dataset encompasses 8,123 tissue TMA spots, each selected to represent the most characteristic spot for ISUP grading within each patient. ISUP

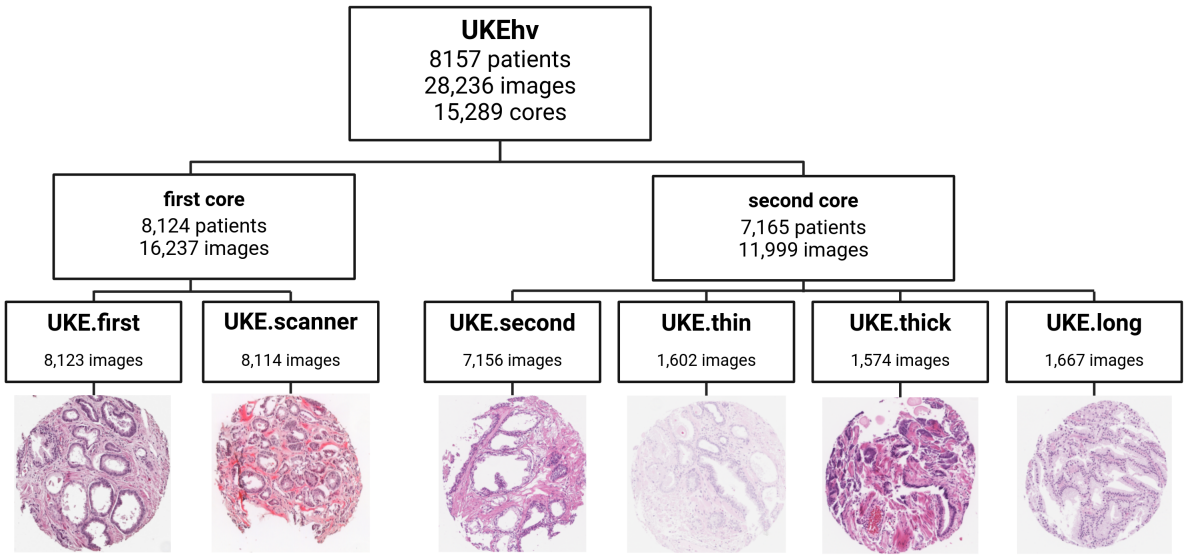

Figure 3.4.2: Composition of the UKEhv dataset that incorporates images with varying acquisition attributes, differentiated into six sub-datasets UKE.first, UKE.second, UKE.scanner, UKE.thin, UKE.thick and UKE.long.

scores were obtained as part of routine diagnostics. The protocol for digitization followed the standard procedure of the University Medical Center Hamburg Eppendorf (UKE), where tissue samples were sliced at a thickness of 2.5µm, stained with Hematoxylin and Eosin for 4 minutes and 1:20 minutes, respectively, and then digitized using an Aperio scanner at a magnification of 40x (0.25µm per pixel). Patient event labels are determined by combining biochemical recurrence (BCR), metastasis META, or prostate cancer-related death PCAD, with patients without any of these events being censored at the last follow-up date.

UKE.scanner: In this sub-dataset TMA images underwent scanning using a 3DHistech scanner. ISUP scores were determined during routine diagnostics. Following the standard digitization protocol of the UKE, the sub-dataset contains 8,114 images scanned at 80x magnification (0.125µm per pixel).

UKE.second: Each of the 7,156 images in the UKE.second sub-dataset represents a secondary batch of TMA spots from the cancerous area of the prostate. These TMAs were processed at a different time and underwent slight variations in the protocol. The ISUP scores were part of routine diagnostics, and the digitization protocol followed the standard procedure of the UKE.

UKE.thin: The UKE.thin sub-dataset comprises 1,602 images, each representing a different TMA spot from the cancerous area of the prostate for every patient. ISUP scores were determined as part of routine diagnostics. Tissue samples were sliced at 1 µm thickness, following the standard digitization protocol of the UKE.

UKE.thick: The UKE.thick sub-dataset, comprising 1,574 images, includes images representing different TMA spots from the cancerous area of the prostate for each patient. ISUP scores were obtained during routine diagnostics, and tissue samples were sliced at a thickness of 10 µm, in line with the standard digitization protocol of the UKE.

UKE.long: In the UKE.long sub-dataset each image represents a different TMA spot from the cancerous area of the prostate for every patient. ISUP scores were determined during routine diagnostics. Tissue samples were stained with Hematoxylin and Eosin for an extended duration of 40 minutes and 10 minutes, respectively, nearly ten times the regular staining time. This experimental sub-dataset contains 1,667 images.

3.4.3 Prostate Cancer Biorepository Network Datasets

Two additional TMA datasets from the Prostate Cancer Biorepository Network (PCBN) in the USA, collected at the New York Langone Medical Centre (NYU) and the Johns Hopkins Hospital in Baltimore (JHU), are included as depicted in Table 3.4.1 and Table 3.4.2, and discussed in detail in the following [150]. Note that every patient received RP treatment for both PCBN datasets, similar to the UKEhv TMA dataset. As for the UKEhv TMA data, expert pathologists ISUP graded the whole prostate.

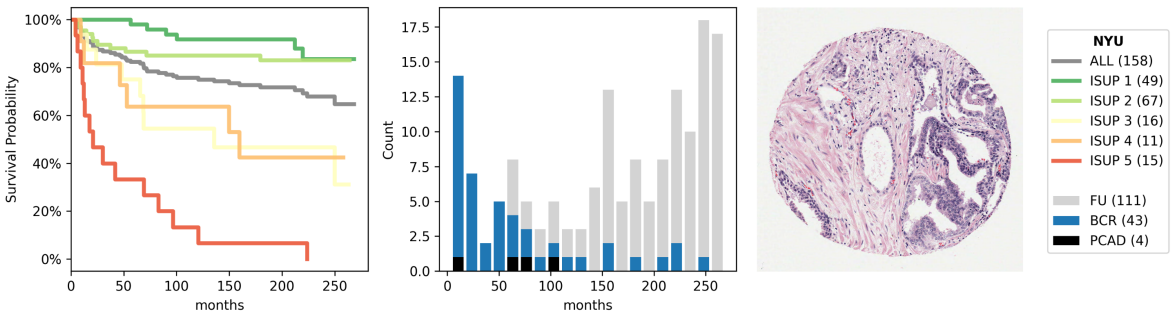

Figure 3.4.3: Kaplan-Meier curves for all patients as well as ISUP sub-cohorts of the NYU dataset (left). Distribution of event (BCR, META, PCAD) and censoring (FU) timepoints (middle). Exemplary image of the dataset (right).

NYU: The TMA cohort from New York University (NYU) contains a total of 204 unique patients arranged in four TMA blocks. ISUP grading is assessed on a patient level and no additional grading details like the number of pathologists are provided. Patients who received any adjuvant therapy are excluded from this dataset. Figure 3.4.3 depicts event distribution and Kaplan-Meier curves for the NYU patients. This dataset includes four TMA blocks that were digitized using an Aperio scanner with a magnification of 20x (0.5 µm per pixel). Notably, these spots were sliced at 5 µm in contrast to 2.5 µm in the internal UKEhv dataset (with the notable exception of UKE.thin and UKE.thick). The TMA blocks are cut into individual images of size 1817x1817 pixels with 0.6mm TMA spots in diameter using QuPath [151]. Spots showing non-neoplastic tissue are excluded. After preprocessing and filtering, this work integrates 515 images of 161 patients with a median of three images per patient.

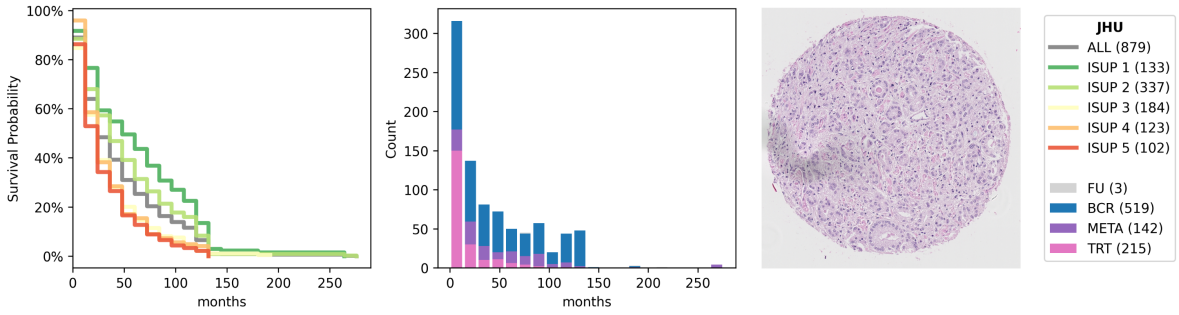

Figure 3.4.4: Kaplan-Meier curves for all patients as well as ISUP sub-cohorts of the JHU dataset (left). Distribution of event (BCR, META, PCAD, TRT) and censoring (FU) timepoints (middle). Exemplary image of the dataset (right).

JHU: The TMA samples from the Johns Hopkins University (JHU) are derived from two datasets named "Case Natural History of Prostate Cancer" (6 TMA blocks) with 235 patients and "Case PSA Progression" (16 TMA blocks) with 726 patients. ISUP scoring is assessed on a patient level and no additional grading details like the number of pathologists are provided. The individual spots were sliced with a thickness of 4 µm and scanned with a Ventana DP2005 and a Hamamatsu NanoZoomer XP6 scanner. In contrast to the other TMA datasets, the endpoint definition of this dataset in terms of event duration is only accessible in a granularity of years instead of days. These two datasets also contain rich metadata information like age, body mass index, race, local recurrence, etc. that are disregarded in this work's analysis. Moreover, for the JHU patients, the aforementioned event indications are extended by salvage (i.e. unplanned) treatment, leading to a censoring rate for this dataset of under 1%. This

means that this cohort can be considered to be biased towards unhealthy individuals, which is further emphasized by the highest relapse rate of JHU patients among all TMA spot datasets in the overall Kaplan-Meier curves depicted in Figure 3.4.4. For integration, the 22 TMA blocks are cut into individual spot images of size 3200x3200 pixels at a magnification of 40x (0.25 µm per pixel) using Qupath [151]. After preprocessing and excluding spots showing control tissue, this work integrates 3,575 TMA spot images that show prostatic adenocarcinoma from 879 patients, with a median of four images per patient.

3.4.4 UKE.sealed Dataset

The UKE.sealed TMA dataset contains 827 patients and 4,097 images with a maximum of 10 images per patient. This dataset is special, in that it contains spot-level quantitative Gleason grading from GS, a renowned PCa pathologist, as opposed to the prostate-level annotations for spots of all other TMA datasets in this study. The information of quantitative Gleason grades was subsequently used to calculate the spot-wise integrated quantitative Gleason (GIQ), the currently best-performing clinical PCa grading system, aggregated as mean or maximum over all images of a single patient [115]. UKE.sealed is therefore the only TMA dataset where an objective comparison of the predictive performance of the proposed algorithm to a human annotator is possible, since both utilize the exact same amount of available images and information. Furthermore, the name UKE.sealed stems from the fact that the access to all patient, metadata, and outcome information was and is restricted exclusively to the department of pathology of the UKE. Also the evaluation of TMA spot predictions of both pathologists and the proposed deep learning models in this work were conducted exclusively by the department of pathology of the UKE.

3.4.5 Malmö Dataset

The MMX biopsy dataset from Malmö, Sweden, contains 716 patients originally collected by Saemundsson et al. [152]. The authors removed all patients with no or less than 2 mm of total cancer in their biopsy, missing follow-up information, inadequate RNA quality, and those that had already developed metastases at the time of diagnosis. Furthermore, for usage in this work, patients that are censored within the first five years as well as images with insufficient quality are removed. In total, 269

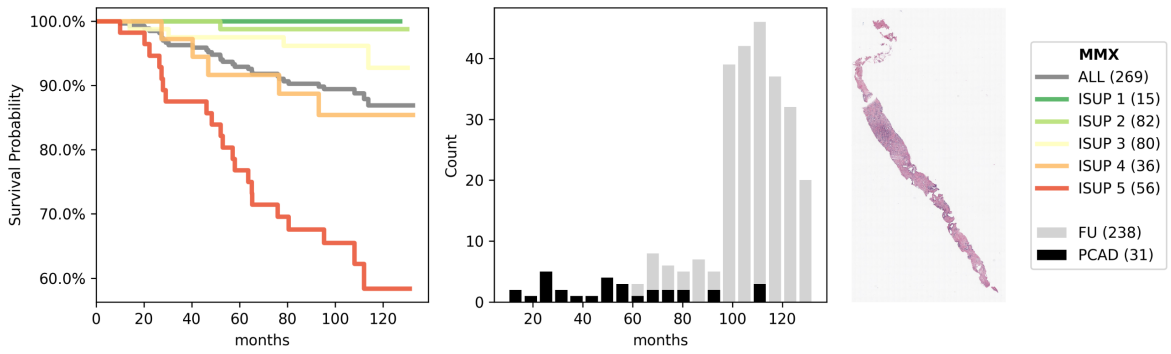

Figure 3.4.5: Kaplan-Meier curves for all patients as well as ISUP sub-cohorts of the MMX dataset (left). Distribution of event (BCR, META, PCAD) and censoring (FU) timepoints (middle). Exemplary image of the dataset (right).

patients with 578 images are included in this work, with up to eight images for a single patient. Figure 3.4.5 depicts event distribution and Kaplan-Meier curves for the MMX patients. The median survival and follow-up time for the remaining patients is 38 and 106 months respectively. Notably, the patients' mean age in this dataset is 4-8 years older than the other datasets. Also, those patients show the highest average PSA values as well as a high variance with 19.9+-44.5 ng/mL. The time-to-event measurement begins with the biopsy date, leading to longer observed time spans in comparison to the TMA datasets, where the reference point is the date of RP. The images of this dataset were digitized using Hamamatsu and Ventana scanners at 40x magnification resulting in individual slide images with a resolution of 0.23 µm per pixel. Image widths and heights vary but consist of up to hundreds of thousands of pixels for the long side of a biopsy. ISUP scores were obtained during routine diagnostics. To further allow for an image-level comparison against the proposed deep learning model PCAI, three individual pathologists annotated all slides independently and blinded from any additional patient information. The ISUP provided by the three pathologists of two centers (Aachen and Uppsala) shows an interrater agreement Fleiss' kappa of 0.199.

3.4.6 Uppsala Dataset

The UPP biopsy dataset from Uppsala, Sweden contains 2,611 unfiltered images of 440 patients from the SPROB20 image dataset that was enriched by patient endpoint information [153]. ISUP scores were obtained from the pathology report of the fu-

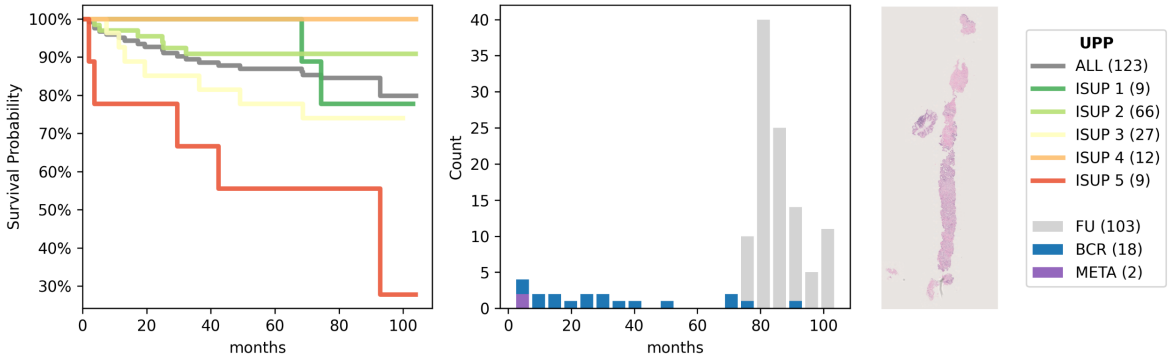

Figure 3.4.6: Kaplan-Meier curves for all patients as well as ISUP sub-cohorts of the UPP dataset (left). Distribution of event (BCR, META, PCAD) and censoring (FU) timepoints (middle). Exemplary image of the dataset (right).

sion biopsies during routine diagnostics. Since some patients in this dataset have had multiple biopsies taken, this work only considers biopsy images from the latest patient visit and excludes all earlier biopsies. Additionally, slides without an assigned ISUP, as well as patients with incomplete or conflicting treatment and follow-up information are excluded from this dataset. In total, 683 images of 123 patients of this dataset are included in the evaluation of PCAI, with up to 10 images per patient at point of biopsy. Figure 3.4.6 depicts event distribution and Kaplan-Meier curves for the UPP patients. The UPP biopsy slides were obtained from a Hamamatsu scanner on a magnification of 40x (0.25 µm per pixel). Since this cohort contains patients from a pilot study for MRI-guided acquisition of prostate biopsies the number of missed biopsies may be different, higher or lower, than it would have been if the conventional procedure had been used.

3.4.7 Prostate Cancer Grade Assessment (PANDA) Dataset

The PANDA dataset contains biopsy slides and corresponding tissue and cancer annotations of expert pathologists. PANDA is one of the largest publicly available whole slide image (WSI) datasets for PCa Gleason grading in the world with 10,616 provided biopsy slides from 2,113 patients. It was published in the Prostate cANcer graDe Assessment challenge and a part of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2020 [154]. The training and validation data for this challenge is provided by two centers, the Karolinska Institute in Stockholm, Sweden (5,456 WSIs) and the Radboud University

Medical Center in Nijmengen, Netherlands (5,160 WSIs). This work uses WSIs with corresponding expert annotations from the PANDA dataset to train the patch-based Cancer Indicator (CI) model that will be described in more detail in Section 3.6.

3.4.8 Color Properties

To the human eye, the most obvious difference between histopathological tissue samples is color. This is caused by, among other things, variance in tissue thickness, H&E staining protocols, or different scanner manufacturers. These differences are visible in the randomly sampled tissue patches of size 256x256 pixels at 40x magnification of five images per dataset in Figure 3.4.7A, as well as the aggregated histograms of the hue, saturation and value channel of all relevant tissue pixels of all images per dataset in Figure 3.4.7B. In the two-dimensional UMAP representation of the HSV histograms of all relevant tissue pixels across datasets, the three larger subdomains of the internal UKEhv data, UKE.first, UKE.second and UKE.scanner form clearly separable clusters from each other, while two of the smaller sub-datasets UKE.thin and UKE.long are intermixed with UKE.second. The third smaller sub-dataset, UKE.thick, forms a separate cluster. UKE.sealed clusters close to UKE.first, though still separable. The external datasets cluster individually. Notably, for the MMX data, which includes images from two different scanners, as well as the JHU data, which includes images from two different scanners as well as two different batches, these differences in acquisition parameters are represented by individual clusters inside the respective datasets.

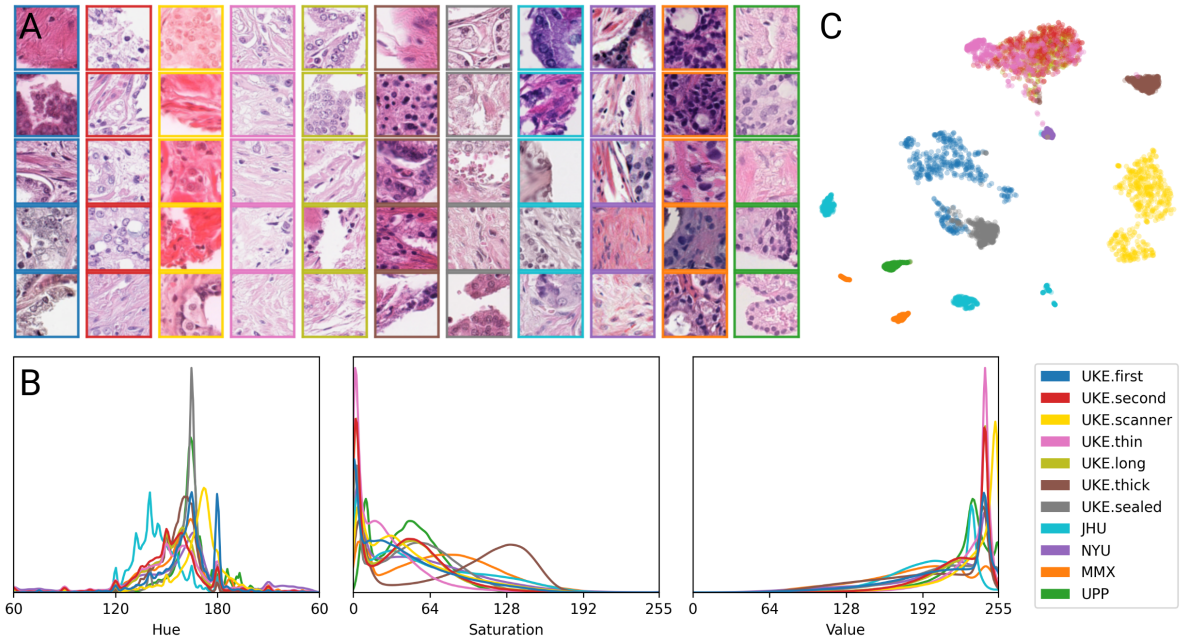

Figure 3.4.7: Visualization of the image properties of all PCa risk datasets used in this work. A: Exemplary patches of five different samples per dataset. B: Hue, Saturation and Value distribution of all valid tissue pixels of all images per dataset. C: UMAP representation of the HSV histograms of all valid tissue pixels per image across datasets.

3.5 Preprocessing

This section describes data preprocessing steps performed in this work. These include the definition of the binary relapse indicator as ground truth for model training, dataset curation and filtering as well as the experimental design with separation into training, validation and test datasets. Furthermore, the preparation of the histopathological images for usage in the deep learning models, with segmentation of relevant tissue areas and division into equally sized patches, is depicted in detail.

3.5.1 Binary Risk Indicator

As described in Section 3.2.3 and Section 3.4, the follow-information per patient is transformed into a quantifiable measure correlating with cancer aggressiveness by defining the objective endpoints of biochemical recurrence (BCR), developing metastasis (META), and PCa-related death (PCAD) as events with corresponding event time relative to the date of RP for TMA spots or to the date of the biopsy procedure for biopsies as time-to-event.

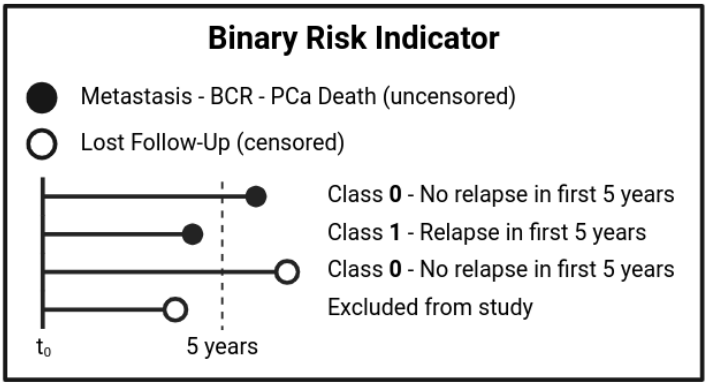

Figure 3.5.1: Definition of the binary cancer relapse indicator for training of the deep learning model. If a patient experiences a relapse prior to five years, defined by developing metastasis, having a biochemical recurrence or dying of PCa, class 1 is assigned. If a patient experienced a relapse later than five years or is censored after that time, class 0 is assigned. Patients that are censored prior to five years are excluded from the study, since no unambiguous indicator can be derived.

To further utilize this information about relapse-free survival time in the proposed deep learning models in this work, it is further transformed into a binary indicator of experiencing a relapse or event in the first five years after sample acquisition. In

detail, if a patient experiences a relapse prior to five years, class 1 is assigned. If a patient experienced an event later than five years or is censored after that time, class 0 is assigned. In case of patients that are censored prior to five years, no unambiguous information about relapse-free survival is available after the time of censoring. Therefore, no binary indicator can be assigned to this subset of patients and they are excluded from analysis in this work. Figure 3.5.1 depicts assignment of the classes based on the patient follow-up information. The five-year relapse indicator can then be utilized to train a binary classification model. In this work, it is assumed that a high predicted probability for class 1 correlates with a low relapse-free survival time and a highly aggressive cancer present at time of sample acquisition. In detail, the binary relapse label y_i is derived as

$$y_i = \begin{cases} 1 & \text{if } T_i < 5 \text{ years and } \delta_i = 1 \\ 0 & \text{if } T_i > 5 \text{ years} \\ discard & \text{otherwise} \end{cases} . \tag{3.5.1}$$

Here, T_i is the follow-up time of the i-th patient and δ_i is its censoring indicator, which is 1 if i is uncensored and 0 if i is censored.

In summary, the task of cancer aggressiveness quantification is transformed into a binary classification problem of predicting the probability or risk of experiencing a relapse within the first five years of sample acquisition. The predicted probability is then interpreted as a continuous risk score, where higher values are expected to correlate with higher cancer aggressiveness.

3.5.2 Filtering

Various preprocessing and filtering steps are applied to the patient metadata and images of all datasets. Figure 3.5.2 depicts the overall workflow. Firstly, only those patients with complete and unambiguous endpoint information are included in the datasets. Ambiguous endpoint information refers to patients with an indicated event without corresponding event time in the metadata, and to patients that are censored within the first 6 months after sample acquisition. In the next step, all patients without an assigned ISUP are removed, since those can't be used for comparison against the deep learning model. Additionally, all patients that received any adjuvant ther-

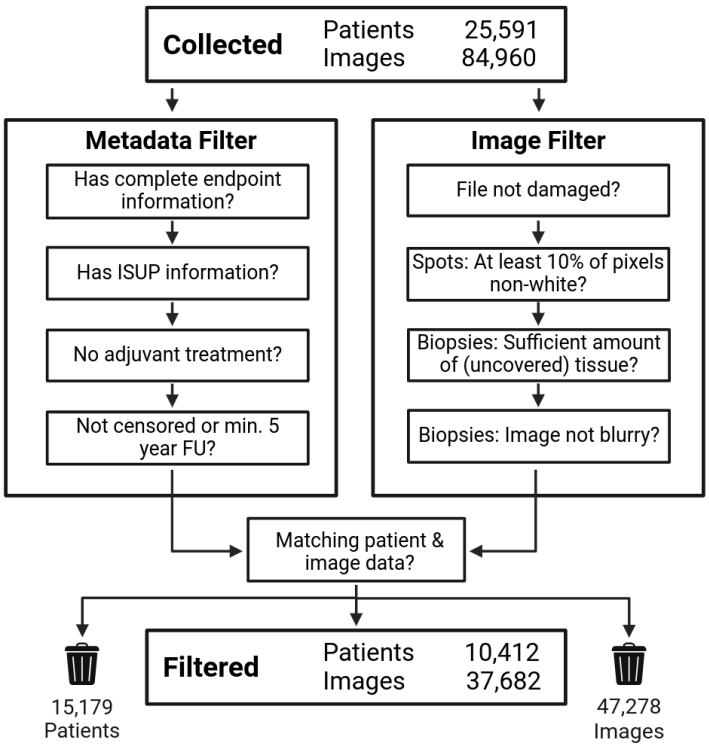

Figure 3.5.2: Dataset curation for patient metadata and images. In total, 84,960 images from 25,591 patients are collected. After applying the depicted filtering steps, 37,682 images from 10,412 patients are finally included in this study.

apy are removed, since this is expected to blur the correlation between cancer aggressiveness at point of sample acquisition and observed event. Finally, endpoint information is transformed into the binary risk indicator as described in Section 3.5.1 and all censored patients with less than 5 years of follow-up are removed, since no clear binary indicator of 5-year relapse-free survival can be derived for those samples.

On the image-level, firstly all damaged files are removed from this study. Next, for the TMA-spots, images that show no or only very little tissue are removed. For this, the fraction of pixels with a lightness value below 235 (for a uint8 value range of 0-255) in the HSL color space is calculated for every TMA-spot image. If less than 10% of the total pixel values are below this threshold, the image is considered empty and excluded from this study. For the biopsy images, where, in contrast to the TMA-spots, image and visualized tissue sizes are very heterogeneous, this automated approach was not applicable. Instead, images that did not contain a sufficient amount of tissue, or only tissue that is to a large portion covered with pen marks, blood or other undesired anomalies, are excluded manually. Additionally, biopsy images that are out of

focus and fully or to a large portion blurred, are also removed manually.

Finally, after filtering metadata and images individually, only those patients with valid data in both groups are included. In total, 15,179 of 25,91 patients and 47,278 of 84,960 images are excluded from analysis in this work, resulting in a final included dataset size of 10,412 patients with 37,682 images.

3.5.3 Masking

Since histopathological images come in arbitrary shapes and sizes and can contain a lot of redundant information, a masking procedure is used to define the relevant tissue areas. For every image, a raw tissue mask, an anomaly mask, a filtered tissue mask and a cancer heatmap is computed. Figure 3.5.3 shows these masks for a sample of a TMA-spot and and a biopsy dataset. In the following, the methodology of deriving each mask is explained in detail.

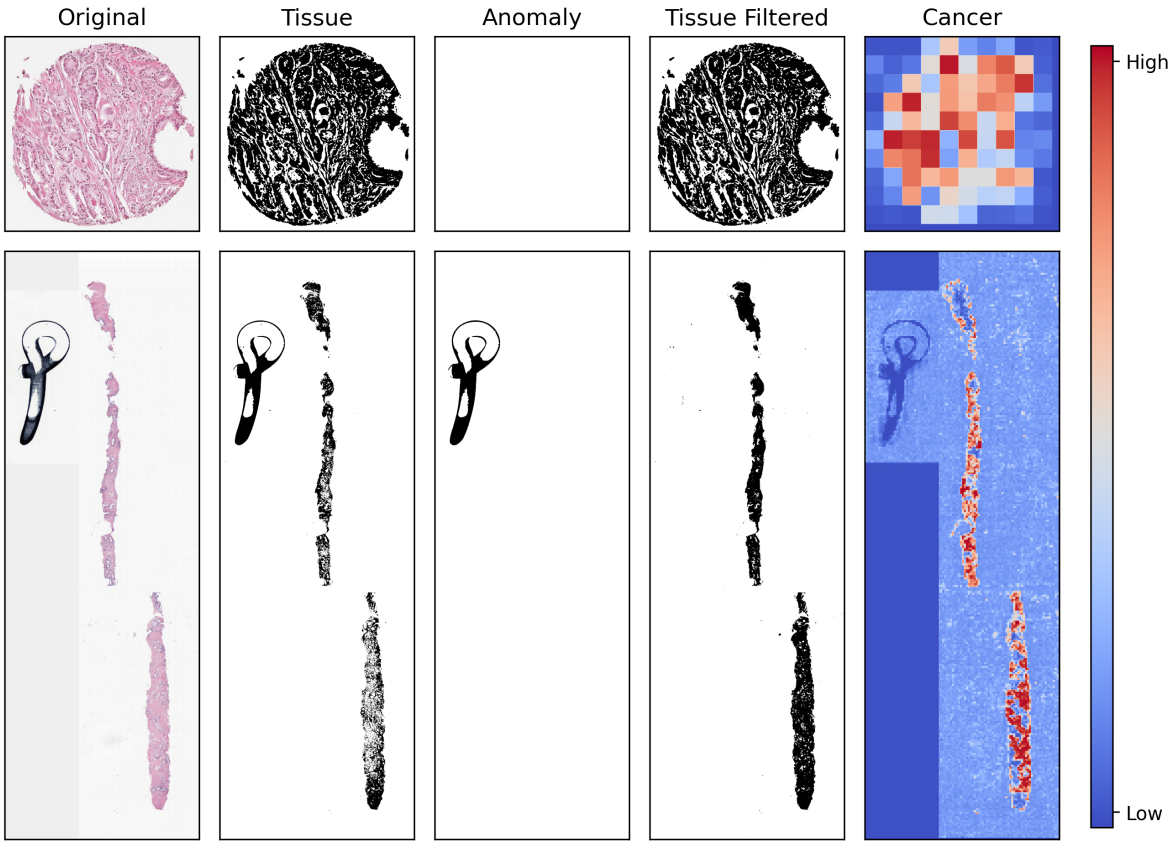

Figure 3.5.3: Exemplary tissue, anomaly, filtered tissue and cancer masks of a sample from the UKE.first TMA spot dataset (top) and the MMX biopsy dataset (bottom).

Tissue Mask: The binary tissue mask is created by separating foreground and background pixels using Otsu's dynamic thresholding on the downsampled histopathological image in the RGB space [155]. This method automatically determines the intensity threshold that maximizes the inter-class variance between foreground and background pixels, and equivalently minimizes the intra-class variance.

Anomaly Mask: Histopathological images can contain undesired artifacts like pen marks, blood or hair. These are especially prevalent on the biopsy slides of the MMX dataset. Since these unwanted artifacts mostly express pixels values that strongly deviate from the background, Otsu's methods assigns those to the foreground class in the tissue mask, as visible in Figure 3.5.3. To detect those unwanted regions, a binary anomaly mask is computed that highlights all foreground pixels with values outside a predefined deviation of the median intensity of the tissue area in the HSV color space. Algorithm 3.5.1 describes the overall procedure.

Filtered Tissue Mask: Artifacts on the images do not only extend the foreground class of the initial tissue mask by undesired regions, it also shifts the optimum threshold defined by Otsu's method such that parts of the actual visible tissue fall into the background class, as it can be observed in the MMX sample depicted in Figure 3.5.3. To mitigate this issue, Otsu's method is applied again on an adapted version of the original image, where the median intensity value of the background class defined by the initial tissue mask is assigned to the anomaly regions defined by the anomaly mask. This results in a less biased threshold for separation of the foreground and background class and a more realistic representation of the actual tissue region in the filtered tissue mask, which could not be achieved by simple subtraction of initial tissue and anomaly mask.

Cancer Heatmap: The cancer heatmap is created by inference of all risk datasets with the separate cancer indicator (CI) model trained on the PANDA dataset, which will be explained in more detail in 3.6.3. The CI predicts the probability of containing cancerous prostate tissue for individual patches of size 256x256 pixels at 20x magnification. This results in a resolution of a single value per patch for the cancer heatmap. Additionally, the cancer heatmap contains values on a continuous scale, in contrast to the binary values of the other masks.

Input: Image img, Tissue Mask $tissue\_mask$
Output: Anomaly Mask $anom\_mask$
Function `DetectAnomalies` ($hsv\_dev\_lower = (30, 30, 60)$, $hsv\_dev\_upper = (30, 70, 60)$, $blur\_ks = 3$, $morph\_ks = 3$, $size\_thresh = 40$):
  Convert img to HSV color space;
  Reduce noise by blurring img with kernel size $blur\_ks$;
  Define foreground pixels p_f of img based on $tissue\_mask$;
  Calculate HSV median m_f of p_f;
  Derive anomaly mask $anom\_mask$ from foreground pixels where $p_f < m_f$ - hsv_dev_lower or $p_f > m_f$ + hsv_dev_upper;
  Perform morphological closing and dilation on $anom\_mask$ with kernel size $morph\_ks$;
  Remove small connected components on $anom\_mask$ with size < $size\_thresh$;
  return $anom\_mask$;

Algorithm 3.5.1: Pseudocode for computation of the anomaly mask $anom\_mask$ based on the original histopathologcial image img and its corresponding tissue mask $tissue\_mask$.

3.5.4 Patching

For usage in the deep learning model, the images are finally cut into equally sized patches based on the masks described in the previous section. For this, a grid with patch size 128x128 pixels at 20x magnification is drawn over the masked images, starting at the top left corner. If at least 10% of the pixels inside a single field of the grid are assigned to the foreground class in the underlying mask, this patch is selected as a valid foreground patch. Non-square patches at the right and bottom edge are discarded by default. For the MMX dataset, the filtered tissue mask serves as underlying mask for patch selection, while for all other datasets, the initial tissue mask is used. Figure 3.5.4 depicts the patch selection procedure for a TMA-spot and biopsy image.

The number of valid patches varies across samples of a single dataset, as well as across datasets, though the biggest difference can be observed, as expected, between TMA-spots and biopsy data. The number of valid foreground patches ranges between 12 and 140, with a median of 92.0, for the TMA-spot data, and between 305 and 7483, with a median of 2474.5, for the biopsy data. Figure 3.5.5 illustrates the sample-wise distribution across datasets.

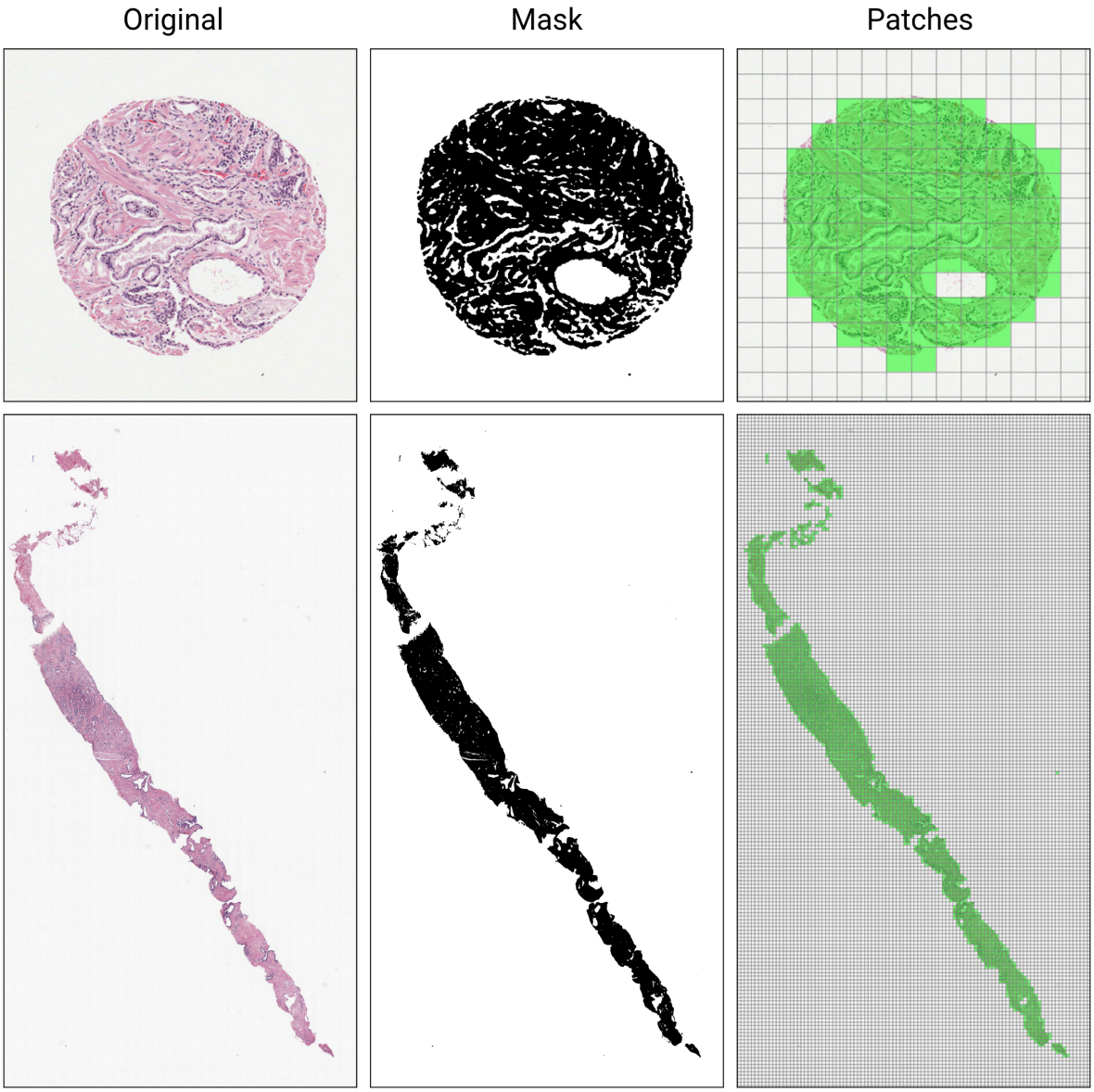

Figure 3.5.4: Original image, tissue mask and finally selected patches (green) for an exemplary sample from the UKE.first TMA spot dataset (top) and the MMX biopsy dataset (bottom).

3.5.5 Experimental Setup

In the course of this work, two deep learning models for PCa risk assessment are trained and evaluated, a baseline model BASE, trained on a single data domain, and the robust model PCAI, trained on three data domains. To this end, the three larger sub-datasets of the UKEhv data, UKE.first, UKE.second and UKE.scanner, are split into 70% training, 15% validation and 15% test set, and the three smaller sub-dataset

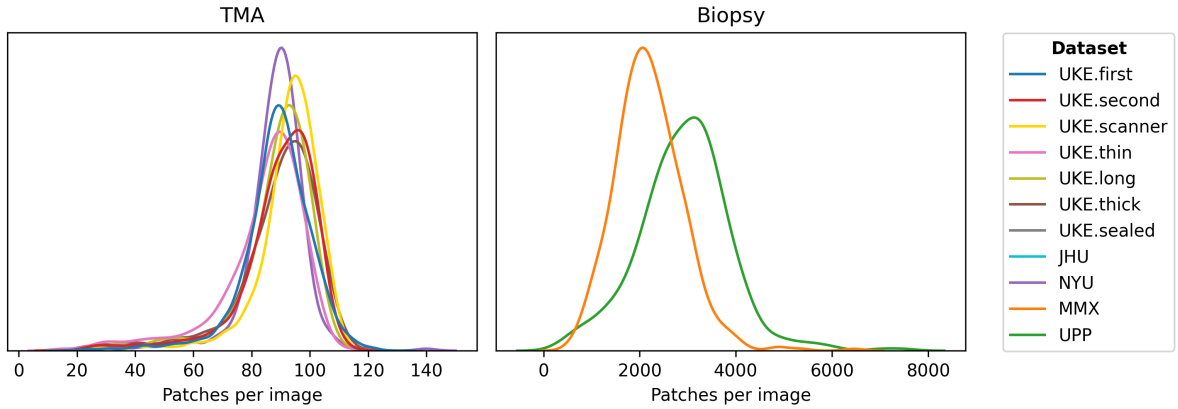

Figure 3.5.5: Number of valid tissue patches per image across datasets selected by the proposed masking and patching procedure.

UKE.thin, UKE.thick and UKE.long are split into 50% validation and 50% test set. The data is split stratified by the binary 5-year relapse indicator. Patients that contribute images to multiple sub-datasets are strictly separated across data splits to avoid leakage. This means that TMAs of the same patient are present in either the training or test data but never in both. Final numbers per split slightly deviate from the initial percentages since some images are excluded retrospectively after assigning the split due to the image filter criteria. The remaining datasets UKE.sealed, JHU, NYU, UPP and MMX are exclusively assigned as test data. In total, 16378 images are included in the training set, 5959 in the validation set and 15336 in the test set. Figure 3.5.6A illustrates the distribution of images per dataset into training, validation and test set.

As illustrated in Figure 3.5.6B, the training set of the UKE.first data is used to train the BASE model and the training sets of the UKE.first, UKE.second and UKE.scanner data are used to train the PCAI model. For online validation and hyperparameter optimization, the validation set of UKE.first is utilized for the BASE model. In case of the PCAI model, the validation set of all six UKEhv sub-domains is used for online validation and hyperparameter optimization. Here, the three smaller UKEhv sub-domains are additionally included to the validation data to optimize for performance on data domains not seen during training.

Finally, as described in Section 3.4, image-wise ISUP or GIQ annotations are only available for the UKE.sealed, UPP and MMX dataset. Therefore, these three test datasets are used to benchmark PCAI against the human assigned ground truth, since only on these images an objective comparison is possible. The remaining test data

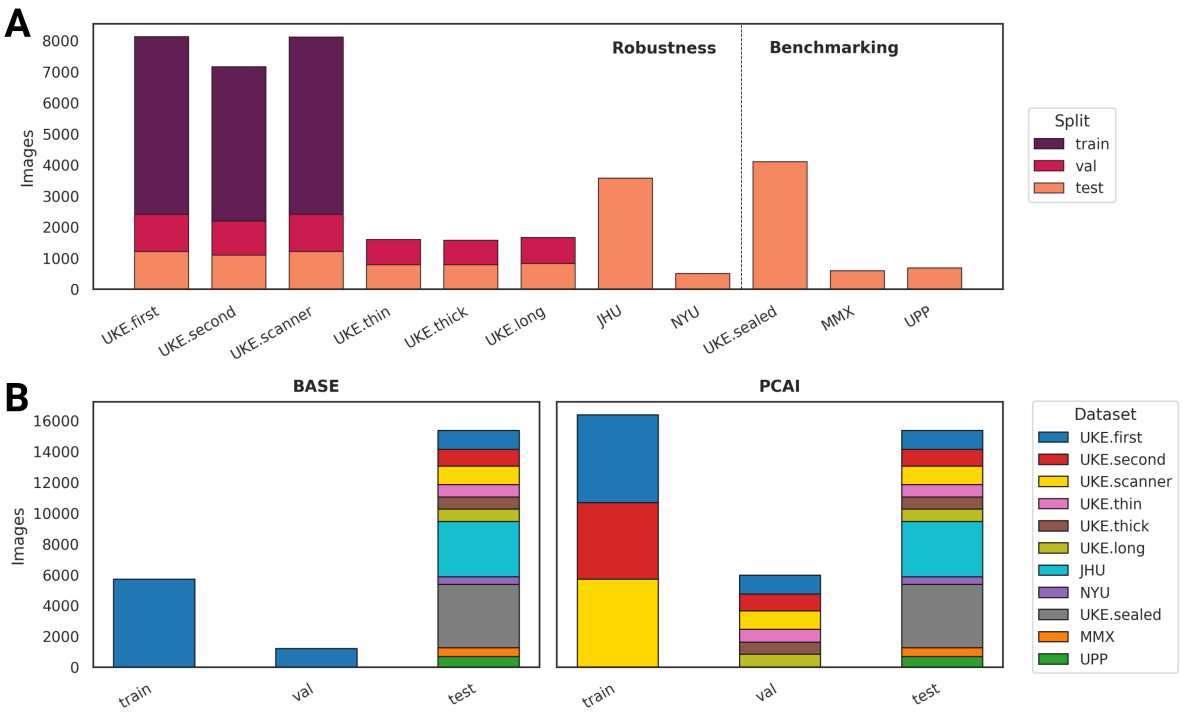

Figure 3.5.6: Experimental design. A: Separation into training-, validation- and test-set across all included datasets. In total, 16378 images are included in the training set, 5959 in the validation set and 15336 in the test set. The UKEhv sub-datasets and the JHU and NYU datasets are used to derive and assess robustness of the proposed models. The UKE.sealed, MMX and UPP datasets are used to benchmark against human assigned annotations. B: The BASE model is trained and validated only on images from the UKE.first dataset. The PCAI model is trained on images from the UKE.first, UKE.second and UKE.scanner datasets, and validated on all UKEhv sub-domains.

from UKEhv, NYU and JHU, where no image-wise annotations are available, are used to build and assess robustness of PCAI in comparison to the BASE model.

3.6 Methods

This section presents the methodology and network architectures of proposed deep learning models of this work, BASE and PCAI. For PCAI, the algorithmic adaptations aiming for clinical applicability, namely domain adversarial training, credibility estimation, color adaptation and cancer indication, are described in detail. Finally, the metrics used for evaluation of the predictive performance of the deep learning models and the cancer grading of the pathologists are introduced.

3.6.1 Prostate Cancer Aggressiveness Index

PCAI is an end-to-end risk assessment model for histopathological prostate cancer data, aimed at actual implementation in a real-world environment (Figure 3.6.1). PCAI is built upon four pillars of clinical applicability, which form the basis of all design decisions. First, the model's risk prediction performance should exceed that of the current subjective scoring system. It is hypothesized that this is only possible if the model learns how to predict objective patient outcomes over time instead of replicating a subjective Gleason or ISUP grade. To this end, all datasets used in this work contain at least 5 years of follow up information for all patients. PCAI then grades the cancer by predicting a potential disease relapse in the future. A leading cause for model prediction errors are variations in the processing of histopathological samples. To render PCAI robust to these changes, it is further hypothesized that domain adversarial training on the large and heterogeneous UKEhv dataset will result in stable predictions across unseen datasets and domains that reflect the variance encountered in everyday clinical practice. Even though PCAI is trained and optimized for stable predictions across different sample processing protocols, it might still encounter histopathological slides of e.g. bad quality, for which it cannot provide a reliable grading. A relevant feature for the proposed model is therefore the notion of confidence or trustworthiness via credibility estimation, not unlike a human expert that is uncertain about the grading of a particular sample and asks for a second opinion. With the aim to stabilize PCAI on images where it shows a low confidence, a credibility-guided color adaptation procedure is further introduced that maps the color scheme of low-credible samples to that of the model's training distribution. Finally, a very relevant feature of PCAI is its interpretability, which lets human experts understand and trust model predictions. PCAI achieves this via its cancer indicator

(CI) module, which highlights and selects cancerous regions of the input images, as well as distinct risk groups derived from the predicted score.

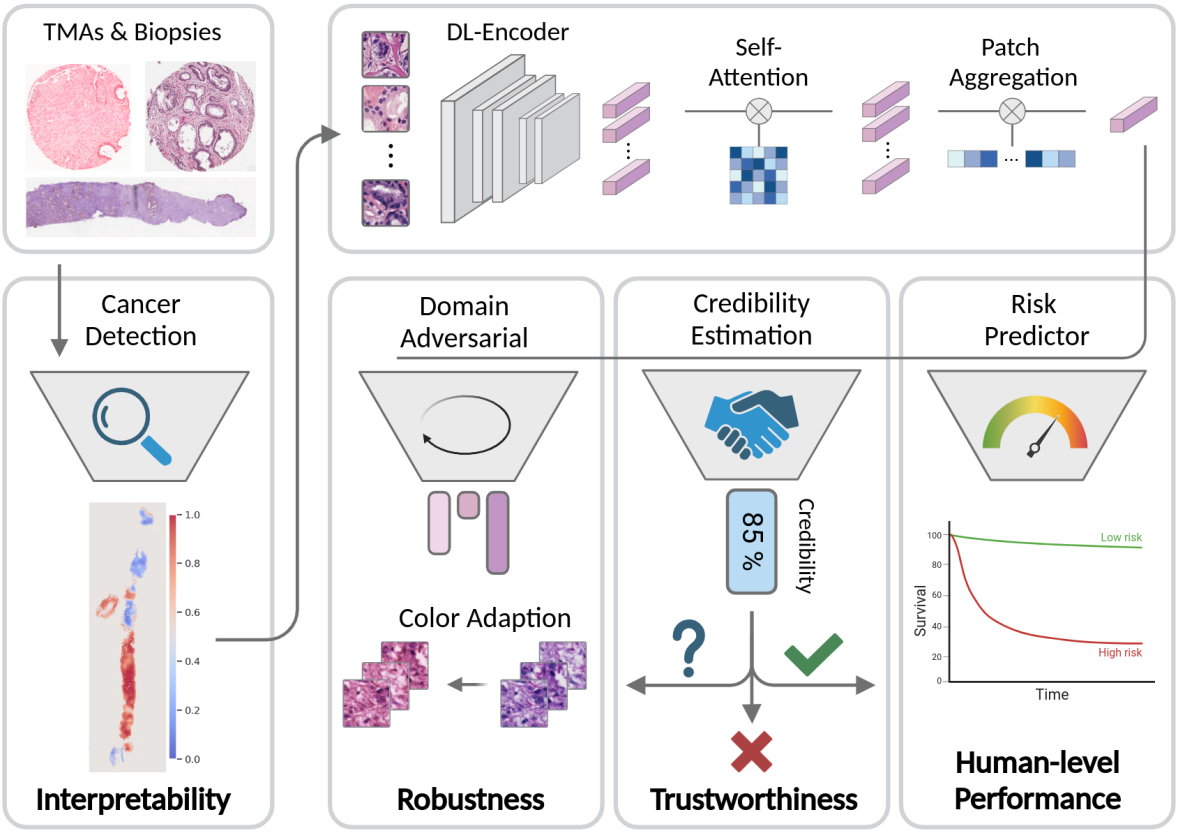

Figure 3.6.1: Overview of the proposed PCAI prostate cancer risk prediction model. PCAI is built and designed on four pillars of clinical applicability, namely interpretability, robustness, trustworthiness and human-level performance.

With these design features in mind, first a reference model BASE is derived, trained only on the single internal data domain UKE.first, containing the most predictive TMA spot per patient, according to the collecting pathologist. In the next step, the aforementioned adaptations of domain adversarial (DA) training on the UKE.first, UKE.second, and UKE.scanner datasets, credibility estimation (CE), color adaptation (CA), and cancer indication (CI) are applied to the BASE model to derive the final proposed risk prediction model PCAI. In the following, the BASE model and the PCAI model with its individual algorithmic extensions are described in detail.

Further, the notions of "patient-level", "image-level" and "patch-level" are used to describe whether the mentioned context correlates to the full patient (e.g. relapse information), a single WSI or a single patch of a WSI.

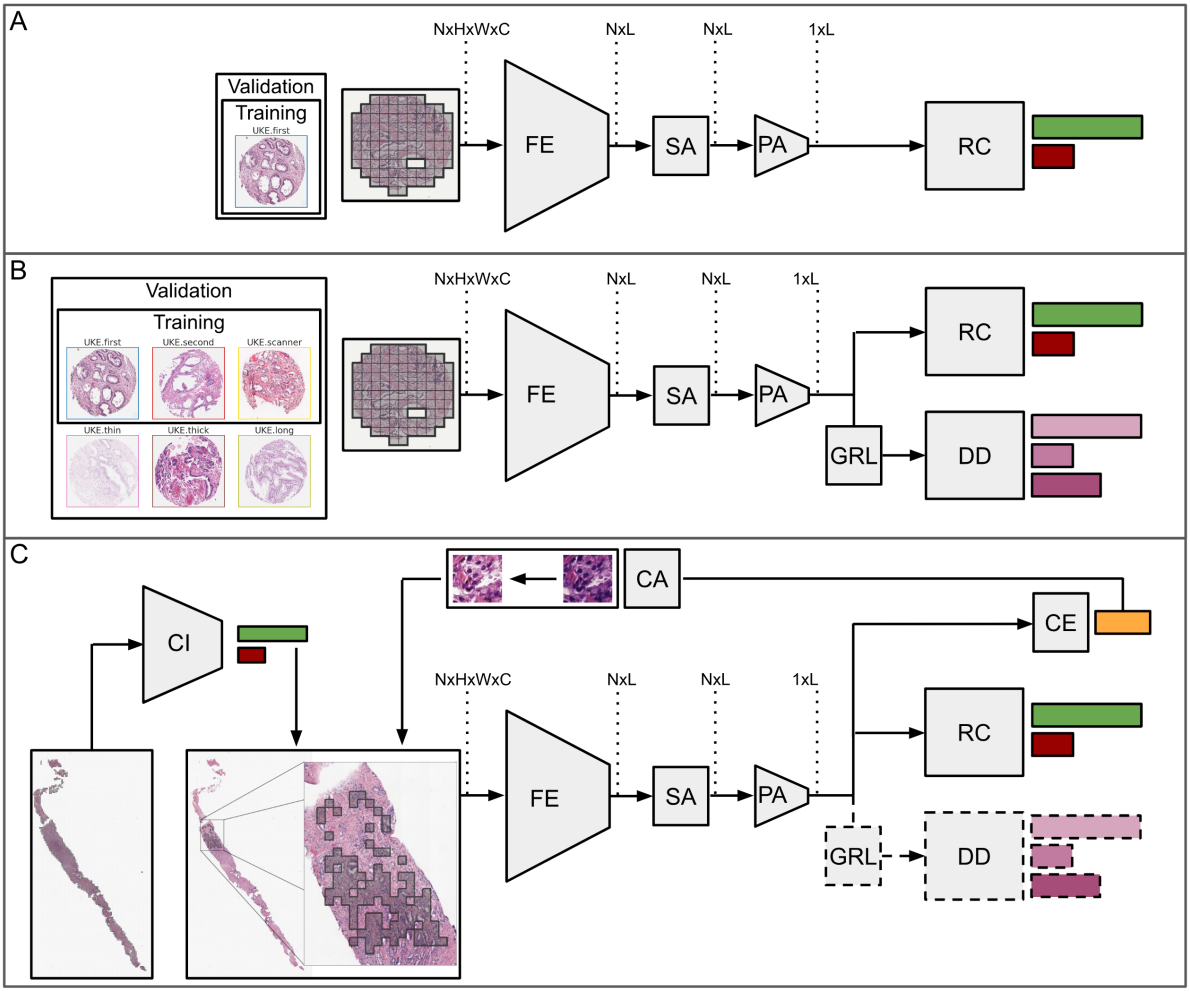

Figure 3.6.2: Schematic of the proposed deep learning models in this work. A: The BASE model, trained and validated on the single data domain UKE.first. Images are fed as bag of patches through the CNN-based feature extractor (FE), cross-correlated in the self-attention layer (SA) and aggregated in the patch aggregation layer (PA), before the risk classification head (RC) assigns the final risk score. B: Extension for domain adversarial (DA) training on the three data domains UKE.first, UKE.second and UKE.scanner and validation on all UKEhv sub-datasets. The domain discriminator (DD) and the gradient reversal layer (GRL) are attached for dual task training. C: Schematic of the full PCAI pipline, which combines the trained DA model with cancer indicator (CI) based patch selection, credibility estimation (CE) and credibility-guided color adaptation (CA) during inference.

3.6.2 Baseline Model (BASE)

The baseline risk prediction network BASE is a binary classifier that assigns the probability of having a relapse in the first 5 years after examination to a histopathological

WSI showing PCa. As described in Section 3.3, a key challenge in the field of digital pathology arises from the size and resolution of the WSIs, especially biopsies, which makes processing of those images in their entirety practically infeasable due to hardware limitations. As further explained, this is mostly approached by multiple instance learning (MIL), where the input image is cut in equally sized patch images and forwarded to the network as a bag of patches, where it is aggregated to derive a single prediction. In this work, the patch-based attention MIL method proposed by Ilse et al. is used in the BASE and PCAI risk prediction model presented in this thesis [78]. Segmentation of the relevant regions in the images and subsequent patching is performed as described in 3.5. The deep learning architecture of the BASE model consists of a CNN-based feature extractor (FE) that transforms all patches per image in parallel and independently into a latent representation. Next, a self-attention layer (SA), as proposed in Rymarczyk et al., accounts for cross-dependencies between all patches of a bag by creating context-aware embeddings from every patch [86]. This bag of patch embeddings is further aggregated into a single latent representation in the attention-based patch aggregation (PA), as proposed by Ilse et al. [78]. Finally, the fully connected risk classification (RC) head predicts the probability for both classes of the binary risk ground truth. The predicted probability for the class 1 (see 3.5.1), corresponding to having a relapse prior to five years, represents the final risk score. In the following, the components of the BASE model are described in more detail.

Feature Extractor

An Efficientnet-b0 architecture (see 3.2) pretrained on the ImageNet dataset serves as backbone architecture for the feature extractor (FE) part of the risk prediction model [121]. For usage in this work, the final fully connected classification layer of the Efficientnet-b0 architecture is removed. The FE transforms bags of N patches $P_n \in \mathbb{R}^{H \times W \times C}$ of height $H = 128$ and width $W = 128$ with three color channels $C = 3$ in the RGB format into bags of N latent feature vectors $h_n^{\text{FE}} \in \mathbb{R}^L$ of length $L = 1280$. Formally, let $S^{\text{img}} = \{P_n \in \mathbb{R}^{H \times W \times C}\}$ with $n \in \{1, 2, ..., N\}$ denote the set of all N patches per image. For every patch P_n, the latent representation h_n^{FE} after the feature extractor is derived as

$$h_n^{\text{FE}} = f^{\text{FE}}(P_n) \text{ with } f^{\text{FE}} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^L, \tag{3.6.1}$$

resulting in the set of all latent patch features $S^{\text{FE}} = \{h_n^{\text{FE}} \in \mathbb{R}^L\}$ with $n \in \{1, 2, ..., N\}$.

Self-Attention Layer

The self-attention (SA) layer aims to correlate patch-embeddings and inject inter-patch related information to all embeddings without altering their shape. In detail, for every patch embedding h_n^{FE} in $S^{\text{FE}} = \{h_n^{\text{FE}} \in \mathbb{R}^L\}$ in the output of the feature extractor, N attention weights a_{nm} are computed, resulting in the attention matrix $A^{\text{SA}} \in \mathbb{R}^{N \times N}$. A^{SA} contains information about the relevance of each patch embedding h_n^{FE} in relation to every other patch embedding h_m^{FE} and is multiplied with the incoming feature vector after the encoder. This creates context-aware embeddings from every patch. This layer is implemented as proposed by Rymarczyk et al. [86], though in this work their proposed trainable γ that weights the influence of the self-attention is set to a static 1. The output h_n^{SA} for a single patch embedding h_n^{FE} is derived as

$$h_n^{\text{SA}} = h_n^{\text{FE}} + \sum_{m=1}^{N} a_{nm}(W^{\text{V}} h_m^{\text{FE}} + b^{\text{V}}), \tag{3.6.2}$$

where

$$a_{nm} = \frac{\exp((W^{\text{Q}} h_n^{\text{FE}} + b^{\text{Q}})^T (W^{\text{K}} h_m^{\text{FE}} + b^{\text{K}}))}{\sum_{o=1}^{N} \exp((W^{\text{Q}} h_n^{\text{FE}} + b^{\text{Q}})^T (W^{\text{K}} h_o^{\text{FE}} + b^{\text{K}}))}, \tag{3.6.3}$$

resulting in the set of all latent patch features $S^{\text{SA}} = \{h_n^{\text{SA}} \in \mathbb{R}^L\}$ with $n \in \{1, 2, ..., N\}$ after the self-attention layer. Here, $W^{\text{Q}}, W^{\text{K}} \in \mathbb{R}^{D^{\text{SA}} \times L}$, $b^{\text{Q}}, b^{\text{K}} \in \mathbb{R}^{D^{\text{SA}}}$, $W^{\text{V}} \in \mathbb{R}^{L \times L}$ and $b^{\text{V}} \in \mathbb{R}^L$ are trainable parameters of the network, where L refers to the length of the input vector and D^{SA} is a hyperparameter set to 160 in this work. n, m and o are indices of the N patch embedding instances.

Patch Aggregation Layer

The attention-based patch aggregation (PA) layer compresses the input of N latent patch embeddings of length L into a single representation of the same length. This method was first introduced by Ilse et al. [78]. It can be seen as a learnable weighted average function of the input instances, where the respective attention weights are determined inside the network for each sample. Given the set of latent patch embeddings $S^{\text{SA}} = \{h_n^{\text{SA}} \in \mathbb{R}^L\}$ after the SA layer, the output $h^{\text{PA}} \in \mathbb{R}^L$ of the PA layer is derived by

$$h^{\text{PA}} = \sum_{n=1}^{N} a_n h_n^{\text{SA}}, \tag{3.6.4}$$

where

$$a_n = \frac{\exp(W^{\text{PA2}} \tanh(W^{\text{PA1}} h_n^{\text{SA}} + b^{\text{PA1}}) + b^{\text{PA2}})}{\sum_{m=1}^{N} \exp(W^{\text{PA2}} \tanh(W^{\text{PA1}} h_m^{\text{SA}} + b^{\text{PA1}}) + b^{\text{PA2}})}. \tag{3.6.5}$$

Here, $W^{\text{PA1}} \in \mathbb{R}^{D^{\text{PA}} \times L}$, $b^{\text{PA1}} \in \mathbb{R}^{D^{\text{PA}}}$, $W^{\text{PA2}} \in \mathbb{R}^{1 \times D^{\text{PA}}}$ and $b^{\text{PA2}} \in \mathbb{R}^1$ are trainable parameters of the network, where L refers to the length of the input vector and D is a hyperparameter. After optimization, D^{PA} is set to 128 in this work. The hyperbolic tangent non-linearity aims to prevent the exploding gradient issue, whereas the softmax function ensures that all attention weights sum to 1 [78].

Risk Prediction Head

Finally, the risk classification head (RC) predicts the probability for both binary risk classes. It consists of a fully connected hidden layer with 100 neurons followed by a ReLU activation function and a fully connected layer with 2 neurons followed by a softmax activation function. In detail, the output logits $z^{\text{R}} \in \mathbb{R}^2$ of the final fully connected layer of the risk prediction head are derived from the aggregated latent patch representation in the output of the PA layer $h^{\text{PA}} \in \mathbb{R}^L$ as

$$z^{\text{R}} = W^{\text{R2}} \cdot \text{ReLU}(W^{\text{R1}} h^{\text{PA}} + b^{\text{R1}}) + b^{\text{R2}}. \tag{3.6.6}$$

Here, $W^{\text{R1}} \in \mathbb{R}^{100 \times L}$, $b^{\text{R1}} \in \mathbb{R}^{100}$, $W^{\text{R2}} \in \mathbb{R}^{c^{\text{R}} \times 100}$ and $b^{\text{R2}} \in \mathbb{R}^{c^{\text{R}}}$ are trainable parameters of the network, where L refers to the length of the input vector and $c^{\text{R}} = 2$ denotes the number of classes in the risk prediction head. The probability $\hat{y}_i^{\text{R}}$ of an image to belong to risk class i is then derived by the softmax function as

$$\hat{y}_i^{\text{R}} = \frac{\exp(z_i^{\text{R}})}{\sum_{j=0}^{c_{\text{R}}-1} \exp(z_j^{\text{R}})}. \tag{3.6.7}$$

The probability for class 1, which corresponds to having a relapse prior to five years, represents the final predicted risk score R^{BASE} of the baseline model, such that $R^{\text{BASE}} = \hat{y}_1^{\text{R}}$.

Risk Prediction Loss

For training the baseline network, cross-entropy of the predicted class probabilities $\hat{y}_i^{\mathrm{R}}$ of the risk prediction head is used as the loss function L^{R}, such that

$$L^{\mathrm{R}}(y^{\mathrm{R}}, \hat{y}^{\mathrm{R}}) = -\frac{1}{ns^{\text{train}}} \sum_{i=0}^{c^{\mathrm{R}}-1} \frac{ns^{\text{train}}}{ns_i} \left(y_i^{\mathrm{R}} \log(\hat{y}_i^{\mathrm{R}}) + (1 - y_i^{\mathrm{R}}) \log(1 - \hat{y}_i^{\mathrm{R}}) \right). \tag{3.6.8}$$

Here, $ns_i \in \mathbb{N}$ denotes the number of samples of the $i-th$ class in the training data, and $ns^{\text{train}} \in \mathbb{N}$ the total number of training samples.

Baseline Training Regime

The BASE model is derived exclusively from the UKE.first TMA-spot dataset. During training, 100 valid tissue patches are randomly sampled from every image based on the area defined by the tissue mask. If less than 100 valid tissue patches are available, patches are randomly oversampled. The static number of 100 patches allowed for training with batch sizes > 1 and is chosen to be close to the median number of valid tissue patches across samples in the dataset (see Figure 3.5.5). Patches are further randomly transformed with AugMix augmentation before input to the network [156]. Training is performed using Adam optimizer with a batch size of 16 and a learning rate of 2.75e-06 for a maximum of 200 epochs, with early stopping on the 5-year relapse AUROC of the UKE.first validation split. Dropout probability is set to 0.34. During inference, all valid patches per image and no AugMix augmentations are used. Hyperparameters are optimized for maximum 5-year AUROC on the UKE.first validation data using a Bayesian search paradigm.

3.6.3 Building Clinical Applicability (PCAI)

In pursuit of building a model based on the four initially defined pillars of clinical application of interpretability, robustness, trustworthiness and a predictive performance that exceeds human annotated ISUP, the BASE model is extended by four algorithmic adaptations that aim to provide these features. Firstly, domain adversarial (DA) training, which utilizes the remaining sub-datasets and the inherent heterogeniety of the internal UKEhv data and aims to increase robustness as well as overall performance. Secondly, credibility estimation (CE) aims to equip the model with the ability to quantify uncertainty in its predictions and therefore increase trustworthiness and

reliability. Thirdly, color adaptation (CA) of samples outside the training distribution aims to further boost robustness and predictive accuracy. Lastly, cancer probability heatmaps provided by the cancer indicator (CI) module are used to focus the risk prediction model on relevant tissue regions, provide visual interpretability and further boost robustness by reduction of noise and redundancy in the input data. The combination of these techniques with the BASE model represents the advanced risk prediction model PCAI. In the following, the individual algorithmic adaptations are explained in more detail.

Domain Adversarial Training (DA)

The highly heterogeneous sub-datasets of the internal UKEhv data are utilized to apply a domain adversarial (DA) training regime to the risk prediction network. This method was first introduced by Ganin et al. [157]. They key idea of domain adversarial training is to apply a multi-task training regime, that aims at creating domain-agnostic feature representations in the model that are still predictive for the main task. For this, the initial BASE architecture is extended by an adversarial part, consisting of a gradient reversal layer (GRL) and a subsequent domain discrimination (DD) head, which aims to classify the images based on previously assigned domain labels. The classifier of the main task and the adversarial part access the same feature space. The part of the network between input layer and shared feature space of main task head and adversarial head is referred to as the shared part of the network. Both main task and domain discrimination task are trained in parallel to minimize the main task loss and the domain classification loss. However, during backpropagation, the GRL inverts the sign of the gradient flowing from the domain discrimination (DD) head into the shared part. This means that the weights of the DD are adapted to improve classification performance of separating domains in the training data, while the weights of the shared part are adapted in the exact opposite direction, leading to a shared feature space that is increasingly less predictive for the DD, i.e. contains less domain-specific information. Through parallel training of both heads, weights of the shared part are adapted to produce domain-agnostic features that are still predictive for the main task. In the field if digital pathology, domain adversarial training was successfully applied for mitotic figure detection on histopathological images by Wilm et al., where the scanner used for image acquisition served as the domain ground truth [17].

For usage this work, the BASE architecture is extended by the adversarial part consist-

ing of the gradient reversal layer (GRL) and the domain discriminator (DD) after the PA layer, such that the adversarial part and the risk classifier access the same latent space of aggregated patch-embeddings. Therefore, the shared part of the network consists of the FE, the SA layer and the PA layer, the main task part of the risk classifier (RC) and the adversarial part of the GRL and the DD. Let from here on θ_S denote the parameters of the shared part, θ_R the parameters of the RC and θ_D the parameters of the DD. Figure 3.6.2B depicts the updated architecture for the DA training. In the following, the GRL and the DD are described in detail.

Domain Discrimination Head

The domain discrimination (DD) head predicts the probability of a sample belonging to each of the predefined domain classes present the training data. Analogous to the risk classification (RC) head, it consists of a fully connected hidden layer with 100 neurons followed by a ReLU activation function and a fully connected layer with 3 neurons followed by a softmax activation function. In detail, the output logits $z^{\mathrm{D}} \in \mathbb{D}^3$ of the final fully connected layer of the DD head are derived from the aggregated latent patch representation in the output of the PA layer $h^{\mathrm{PA}} \in \mathbb{R}^L$ as

$$z^{\mathrm{D}} = W^{\mathrm{D2}} \cdot \mathrm{ReLU}(W^{\mathrm{D1}} h^{\mathrm{PA}} + b^{\mathrm{D1}}) + b^{\mathrm{D2}}. \tag{3.6.9}$$

Here, $W^{\mathrm{D1}} \in \mathbb{R}^{100 \times L}$, $b^{\mathrm{D1}} \in \mathbb{R}^{100}$, $W^{\mathrm{D2}} \in \mathbb{R}^{c^{\mathrm{D}} \times 100}$ and $b^{\mathrm{D2}} \in \mathbb{R}^{c^{\mathrm{D}}}$ are trainable parameters of the network, where L refers to the length of the input vector and $c^{\mathrm{D}} = 3$ denotes the number of classes in the DD head. The probability $\hat{y}_k^{\mathrm{D}}$ of an image to belong to domain k is then derived by the softmax function as

$$\hat{y}_k^{\mathrm{D}} = \frac{\exp(z_k^{\mathrm{D}})}{\sum_{l=0}^{c^{\mathrm{D}}-1} \exp(z_l^{\mathrm{D}})}. \tag{3.6.10}$$

Gradient Reversal Layer

The Gradient Reversal Layer (GRL) is a non-parameterized layer that acts as an identity transformation in the forward pass, but inverts the sign of the gradient of the subsequent layer during backpropagation before passing it to the preceding layer [157]. The GRL is inserted between the shared part, ending with the PA layer, and the DD, resulting in the architecture depicted in Figure 3.6.2B. When backpropagating the domain classification loss L^{D}, the partial derivatives of the loss that are downstream the GRL w.r.t. the layer parameters θ^{S} of the shared part upstream the GRL get mul-

tiplied by -1 , i.e., $\frac{\partial L^{\mathrm{D}}}{\partial \theta^{\mathrm{S}}}$ is effectively replaced with $-\frac{\partial L^{\mathrm{D}}}{\partial \theta^{\mathrm{S}}}$. Mathematically, the GRL can be formulated as a "pseudo-function" $\mathcal{F}^{\mathrm{GRL}}(\mathbf{x})$ defined by the equation $\mathcal{F}^{\mathrm{GRL}}(\mathbf{x}) = \mathbf{x}$, describing its behaviour during the forward pass, and $\frac{d\mathcal{F}^{\mathrm{GRL}}}{d\mathbf{x}} = -\mathbf{I}$, describing its behaviour during backpropagation, where $\mathbf{I}$ is an identity matrix.

Domain Adversarial Loss

Cross-entropy of the domain class probabilities $\hat{y}_k^{\mathrm{D}}$ predicted by the DD head is used as the loss function L^{D}, weighted by inverse occurrence frequency of the domain labels in the training data, such that

$$L^{\mathrm{D}}(y^{\mathrm{D}}, \hat{y}^{\mathrm{D}}) = -\frac{1}{ns^{\mathrm{train}}} \sum_{k=0}^{c^{\mathrm{D}}-1} \frac{ns^{\mathrm{train}}}{ns_k} \left(y_k^{\mathrm{D}} \log(\hat{y}_k^{\mathrm{D}}) + (1 - y_k^{\mathrm{D}}) \log(1 - \hat{y}_k^{\mathrm{D}}) \right). \tag{3.6.11}$$

Here, $ns_k \in \mathbb{N}$ denotes the number of samples of the $k-th$ domain in the training data, and $ns^{\mathrm{train}} \in \mathbb{N}$ the total number of training samples.

During training of the domain adversarial model, both loss functions L^{R} and L^{D} of the RC and the DD head are minimized in parallel, such that the total loss function L^{PCAI} amounts to

$$L^{\mathrm{PCAI}} = L^{\mathrm{R}} + \lambda \cdot L^{\mathrm{D}}. \tag{3.6.12}$$

Here, the additional hyperparameter $\lambda \in \mathbb{R}$ regularizes the influence of the domain loss L^{D} to the overall training. Unlike the original implementation by Ganin et al., who differentiated between source and target domains and only used the source domain to train the main task, in this work all training samples are used to train both the risk classification and the domain discrimination task.

A total objective pseudo function $\tilde{E}\left(\theta^{\mathrm{S}}, \theta^{\mathrm{R}}, \theta^{\mathrm{D}}\right)$ that is minimized during training can then be defined as

$$\begin{aligned} \tilde{E}\left(\theta^{\mathrm{S}}, \theta^{\mathrm{R}}, \theta^{\mathrm{D}}\right) = \frac{1}{ns^{\mathrm{train}}} \sum_{s=1}^{ns^{\mathrm{train}}} & L^{\mathrm{R}}\left(G^{\mathrm{R}}\left(G^{\mathrm{S}}\left(\mathbf{x}_s; \theta^{\mathrm{S}}\right); \theta^{\mathrm{R}}\right), y_s^{\mathrm{R}}\right) \\ & -\lambda L^{\mathrm{D}}\left(G^{\mathrm{D}}\left(\mathcal{F}^{\mathrm{GRL}}\left(G^{\mathrm{S}}\left(\mathbf{x}_s; \theta^{\mathrm{S}}\right)\right); \theta^{\mathrm{D}}\right), y_s^{\mathrm{D}}\right), \end{aligned} \tag{3.6.13}$$

where G^{S}, G^{R} and G^{D} describe the shared part, RC head and DD head, respectively.

Domain Adversarial Training Regime

As described in the experimental design in Section 3.5.5, the training dataset of the DA model is extended by the other two large UKEhv sub-domains, such that the total training dataset consists of data from UKE.first, UKE.second and UKE.scanner. Data from the UKE.first domain is fed twice per epoch to put a stronger emphasis on the data containing the most representative spot per patient. The validation dataset is extended accordingly by data from UKE.first, UKE.second and UKE.scanner. Furthermore, to evaluate performance on domains not seen during training, the validation splits of the three smaller UKEhv sub-datasets UKE.thin, UKE.thick and UKE.long are also included in the overall validation data. The overall training procedure is analogue to the BASE model, though here a learning rate of 9.87e-07, dropout rate of 0.5 and stochastic depth of 0.5 is used. Early stopping as well as hyperparameter optimization is performed on the combined 5-year AUROC of the validation splits of all six UKEhv sub-domains.

Credibility Estimation (CE)

To be applicable in an actual clinical setting, the predicted risk score should be accompanied with some notion of trustworthiness that quantifies how certain the model is when predicting on a given image. For this, a credibility estimation (CE) setup is introduced, which computes a credibility score for every unseen sample based on the distance to the learned distribution of the model. The underlying assumption is that samples that differ strongly from the data seen during training should receive a lower credibility score than those close to the learned distribution, independent of the actual predicted risk score.

Distance Measure

The Mahalanobis distance d^{M} in the space of latent representations in the output of the PA layer is employed as distance measure. This is motivated by the work of Lee et al., who successfully utilized the latent Mahalanobis distance for detection of out-of-distribution data [158]. In detail, let $h_u^{\mathrm{PA}} \in \mathbb{R}^L$ denote the aggregated latent patch representation of an unseen sample u. The Mahalanobis distance d^{M} of h_u^{PA} to the center of the training distribution $\bar{H}_{\text{train}}^{\mathrm{PA}} \in \mathbb{R}^L$ is then derived as

$$d^{\mathrm{M}} = \sqrt{\left(h_u^{\mathrm{PA}} - \bar{H}_{\text{train}}^{\mathrm{PA}}\right) V_{\text{train}}^{-1} \left(h_u^{\mathrm{PA}} - \bar{H}_{\text{train}}^{\mathrm{PA}}\right)^T}. \tag{3.6.14}$$

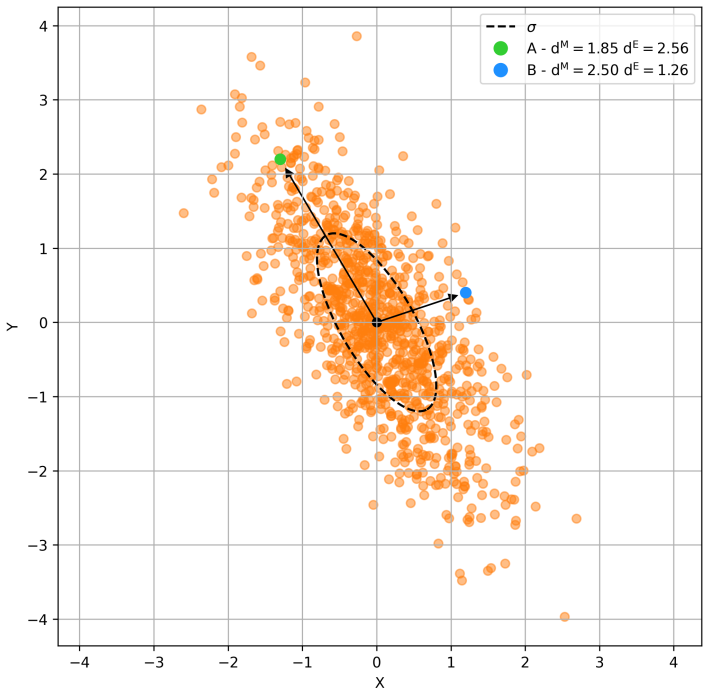

Figure 3.6.3: Example of the Mahalanobis distance d^{M} in the two-dimensional space. Point A expresses a higher Euclidian distance d^{E} to center of the distribution (orange) than Point B, though the Mahalanobis d^{M} distance is smaller. Standard deviation σ of the underlying distribution is depicted by the dotted line.

Here, V_{train}^{-1} refers to the inverse of the covariance matrix of the aggregated patch representations of all ns^{train} samples in the training set, while $\bar{H}_{\text{train}}^{\text{PA}}$ refers to their average in each of the L dimensions. Since the training set differs for the DA model and the BASE model, V_{train}^{-1} and $\bar{H}_{\text{train}}^{\text{PA}}$ are derived from the UKE.first, UKE.second and UKE.scanner domains for the former and from UKE.first only for the latter.

In contrast to the Euclidian distance d^{E}, which is derived as

$$d^{\mathrm{E}} = \sqrt{\left(h_u^{\text{PA}} - \bar{H}_{\text{train}}^{\text{PA}}\right)\left(h_u^{\text{PA}} - \bar{H}_{\text{train}}^{\text{PA}}\right)^T}, \tag{3.6.15}$$

the Mahalanobis distance takes the covariance of the underlying distribution into account and is expected to provide a more meaningful representation of how strongly an unseen sample deviates from the learned distribution of the model. Figure 3.6.3 illustrates the difference between both distance measures for an examplary two-dimensional distribution. The Mahalanobis distance refers to a multivariate generalization of the square of the standard score $z = (x - \mu)/\sigma$, measuring distances in multiples of stan-

dard deviation σ from the mean μ of the underlying distribution.

To mitigate the influence of outliers, the fifth percentile of training samples expressing the highest Mahalanobis distance to the training center are removed. Then, the training center is re-calculated on the remaining training samples for usage in the CE setup.

Calculation of Credibility

To further transform the Mahalanobis distance to the training center into a normalized representation of model uncertainty, ideas from the concept of Conformal Prediction (CP) are employed [159]. CP is a post-hoc method to measure uncertainty in pre-trained prediction models by providing sets of valid class predictions that exceed a given significance level. Here, first a non-conformity measure with score α that assesses the strangeness of an unseen sample is derived from the underlying model. Next, a separate calibration set $S^{\text{calib}} = \{(x_1, y_1), ..., (x_m, y_m)\}$ with samples that stem from the same distribution as the training data but are unseen to the model is defined. The non-conformity score α_j is computed for every sample in the calibration set. To evaluate how different an unseen sample x_u is from the training distribution, its non-conformity score α_u is then compared to the non-conformity scores α_j of the calibration set for every class label y_c, such that

$$p_c(\alpha_u) = \frac{|\{j = 1, \ldots, m : y_j = y_c \text{ and } \alpha_j \geq \alpha_u\}| + 1}{|\{j = 1, \ldots, m : y_j = y_c\}| + 1}, \tag{3.6.16}$$

where $p_c(\alpha_u)$ refers to the p-value (distinct from the statistical p-value) for a given class y_c. High p-values indicate high conformity with the training distribution, since most calibration examples expressed higher non-conformity scores than x_u [160]. For a given significance level ϵ, the conformal prediction set T^ϵ then contains all classes with $p_c(\alpha_u) > \epsilon$.

In the literature, CP mostly utilizes non-conformity scores α based on the softmaxed class probabilites in the output of the model (e.g. $\alpha = 1 - \hat{y}_c$). However, any heuristic notion of uncertainty in the underlying model is applicable to it [161]. For usage in this work, the Mahalanobis distance d_M to the center of the latent train distribution in the output of the PA layer, as described above, is chosen as the non-conformity score α. Since this measure does not directly correlate with the class prediction, the prediction sets T^ϵ are neglected in this work. Instead, the maximum p-value among both 5-year relapse classes is defined as the credibility $Cred_u$ of an unseen sample,

such that

$$Cred_u = \max_c \left(p_c \left(\alpha_u \right) \right) \text{ with } \alpha_u = d_M. \tag{3.6.17}$$

This credibility score quantifies how close a given sample is to the model's learned distribution, based on the unseen calibration dataset, and is expected to correlate with the validity of the final risk prediction.

The validation split of the UKEhv sub-datasets that are also present in the training data of the respective model serves as the calibration data when applying the CE setup to PCAI and the BASE model, such that it consists of data from the UKE.first, UKE.second and UKE.scanner domains for the former and of data from the UKE.first domain for the latter.

Color Adaptation (CA)

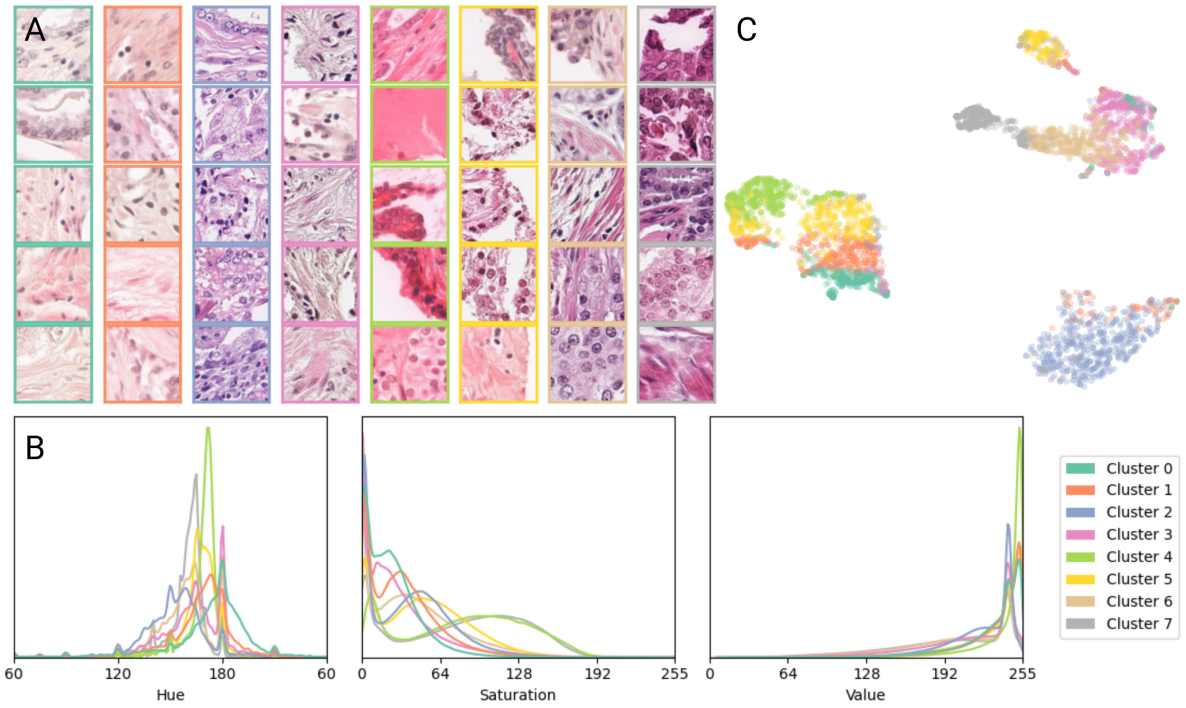

Figure 3.6.4: Visualization of the eight histogram clusters in the training data defined by the color adaptation setup. A: Exemplary patches of five different samples per cluster. B: Hue, Saturation and Value distribution of all valid tissue pixels of all images per cluster. C: UMAP representation of the HSV histograms of all valid tissue pixels per image across clusters.

As shown in Section 3.4.8, color information of the images, represented by their histograms in the HSV space, is a strong separator of the individual datasets used in this work and indicates the covariate shift between those. The proposed domain adversarial training regime aims to mitigate the influence of this shift as a model-centric approach, by making the encoder less sensitive to domain specific biases. To further boost the robustness of the proposed model with an additional data-centric approach, a color adaptation (CA) setup is used in this work.

More specifically, cluster-guided histogram matching of unseen images with the training distribution is performed. For this, 8 k-means cluster are derived from the histograms of all valid tissue patches per image of the training data in the HSV space, using the Wasserstein distance as the distance measure. The Wasserstein distance, which is also referred to as the "earth movers distance", quantifies the cost of transforming one distribution, e.g. a histogram, into another. For every cluster, histograms of all training samples belonging to that cluster are aggregated into a single cluster histogram. Figure 3.6.4 depicts exemplary patches from different samples per cluster, the aggregated histograms of every cluster and a UMAP representation of HSV histograms of the training images of PCAI. The number of 8 clusters is determined by an elbow plot and is used for both PCAI and the BASE model, however, the samples of the training set differ accordingly.

Histogram Matching

During inference of the model, the Wasserstein distance of the histogram of an unseen sample to all 8 training clusters is calculated. Then, the histogram of the input image is matched with the aggregated histogram of the closest cluster. Matching histograms with aggregated histograms of clusters has proven superior to matching on random histograms of the training data in the literature [142]. This smooths the effect of outliers while preserving inherent type differences inside the training dataset.

In detail, let $H_u(k)$ denote the histogram of an unseen sample x_u in the HSV space and $H_r(k)$ the aggregated histogram of the nearest histogram cluster in the training dataset. Here, k represents the intensity level, ranging from 0 to 255 for the saturation and value channel and from 0 to 179 for the hue channel. Next, the cumulative distribution functions $CDF_u(k)$ and $CDF_r(k)$ are calculated for both histograms as

$$CDF(k) = \sum_{j=0}^{k} \frac{H(j)}{N}, \tag{3.6.18}$$

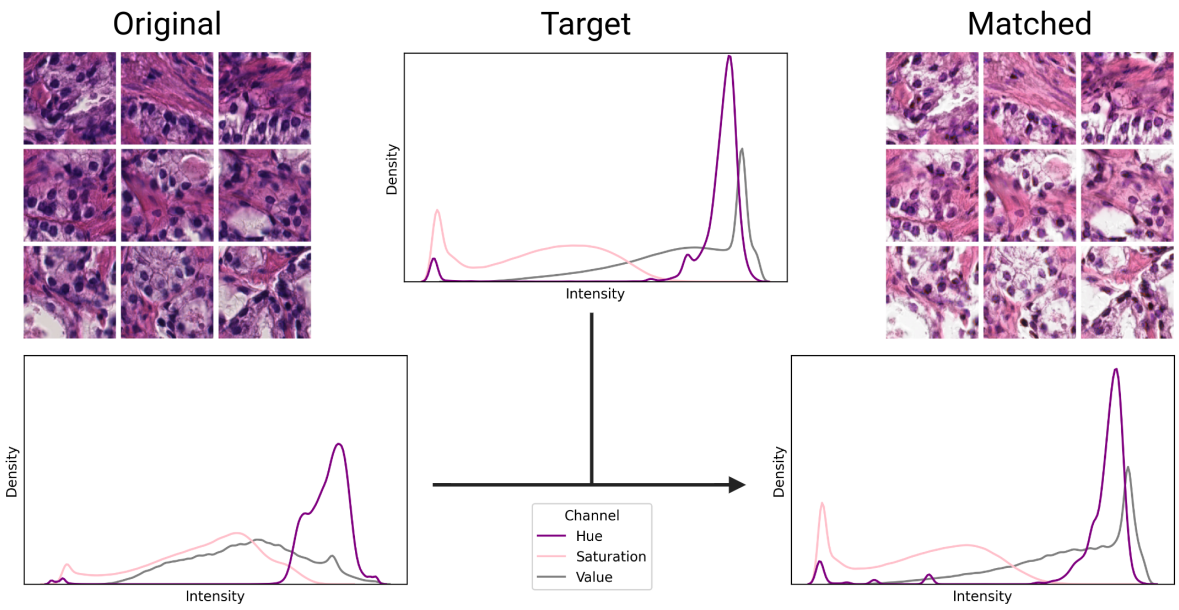

Figure 3.6.5: Schematic of the proposed color adaptation procedure for a sample of the MMX dataset. The combined histogram of all patches is matched with the histogram of the closest training set cluster. Adapted patches are then forwarded for inference with the risk prediction model.

where N denotes the total number of pixels per histogram.

For each intensity level k in the unseen image, a corresponding intensity level g is determined that minimizes the absolute difference between $CDF_u(k)$ and $CDF_r(k)$. This dependency is captured by the transformation function $G(k)$ and can be expressed as

$$G(k) = \arg\min_g |CDF_u(k) - CDF_r(g)|. \tag{3.6.19}$$

Finally, the transformation function $G(k)$ is applied to every pixel of the input image x_u to map it to its corresponding transformed intensity level g. Figure 3.6.5 depicts patches and histograms of a sample of the MMX dataset before and after CA.

Credibility-guided Histogram Matching (CE-CA)

Since histogram matching poses the potential risk of introducing unwanted artifacts to the images, a selective approach is further proposed. Here, only those images are adapted where the model constitutes the necessity, meaning it cannot provide a reliable prediction on the unaltered sample. To quantify this necessity, the Credibil-

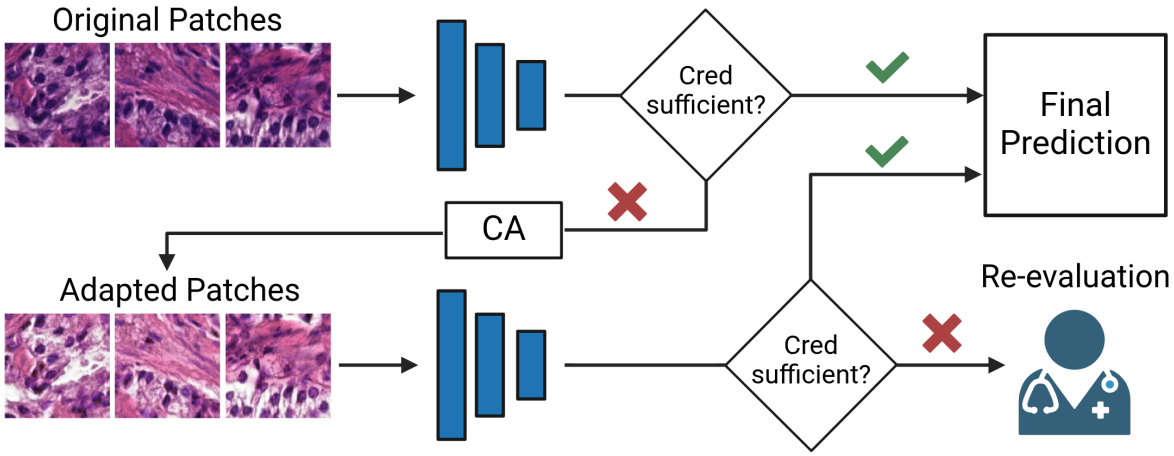

Figure 3.6.6: Schematic of the proposed credibility-guided color adaptation workflow with the human in the loop. After inference on the unaltered patches, the credibility score is assessed. If sufficient credibility is reached, the corresponding risk score is defined as the final prediction. Otherwise, the patches are adapted with the proposed color adaptation setup. Then, inference and assessment of the credibility score of the adapted image is performed again. If the score is still not sufficient, the sample is regarded as ungradable and should be examined by the pathologist.

ity score $Cred$ is utilized, such to only adapt images that deviate strongly from the learned distribution. In detail, using the CE setup described above, a threshold on the credibility scores is defined such that 75% of the calibration data of the underlying model express higher credibility scores. This threshold on the credibility score amounts to 0.279 for the PCAI model (and to 0.255 for the BASE model, however, the color adaptation procedure is not applied to the BASE model).

During inference of the PCAI model, a feedback loop is applied. Images that express a credibility score below the defined threshold during initial prediction are adapted based on the CA procedure described above and fed through the network a second time. The final risk score and credibility score then refer to the model's output for the adapted image.

In clinical practice, this aims to allow the model to identify problematic images and then try to recover those for a credible prediction by adapting its color. If sufficient credibility is not reached by then, the image should be deferred and evaluated by a human rater. Figure 3.6.6 depicts the credibility-guided color adaptation CE-CA procedure with the human in the loop.

Cancer Indication (CI)

A separate cancer indicator model is derived in the course of this work and combined with the risk prediction model in the cancer indicator module (CI). The expected benefit of applying the CI module to PCAI is two-fold: First, it is used to reduce noise and redundancy in the data and to focus the risk prediction model on the most relevant regions in the image. Second, it provides an additional means of visual interpretability to the overall setup by highlighting cancerous regions on the images.

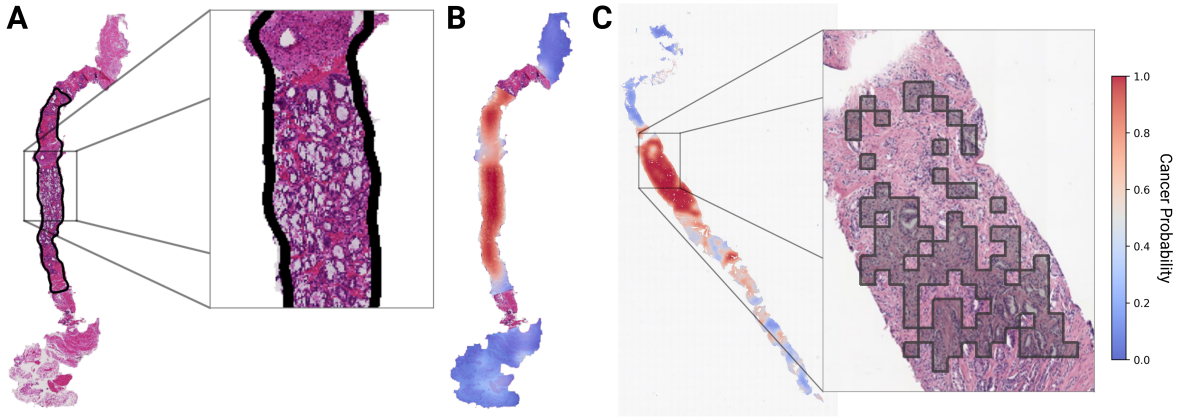

Figure 3.6.7: Components of the cancer indicator (CI) module. A: Ground truth annotation of the cancerous area on a sample of the PANDA dataset used to train the cancer indicator model. B: Cancer probability heatmap on the same PANDA sample predicted by the CI. C: Cancer probability heatmap on a sample of the MMX dataset and resulting patch selection of the 100 most cancerous patches used for subsequent inference in PCAI.

Cancer Indicator Model

The CI model is derived exclusively from images of the PANDA dataset (see 3.4). First, images are cut into equally sized patches $P \in \mathbb{R}^{H \times W \times C}$ of height $H = 256$ and width $W = 256$ with three color channels $C = 3$ at 40x magnification in the RGB format. Then, the available pixel-wise segmentation masks of the PANDA dataset are used to derive a binary cancer label $y^{\mathrm{C}} \in \{0, 1\}$ as ground truth for every patch P. In detail, a patch is assigned to class 1, i.e. cancerous, if more than 90% of its pixels are labelled as cancerous in the corresponding segmentation mask. Analogue, a patch is assigned to class 0, i.e. healthy (or benign), if none of its pixels are labelled as cancerous and it contains less than 90% background pixels. Patches that do not fall in either of the mentioned categories are discarded. Figure 3.6.7A depicts the the ground truth annotation of the cancerous area on an exemplary slide from the

PANDA dataset.

The network architecture of the CI model is a CNN-based binary classifier that takes individual patches as input and assigns the probability of being cancerous to every patch. It consists of a feature extractor with an EfficientNet-b0 backbone and a subsequent fully connected layer with two output nodes and softmax activation. The cancer indicator can be described by the function $f^{\mathrm{CI}}(P)$ and the cancer probability $\hat{y}^{\mathrm{C}}$ per patch P is derived as

$$\hat{y}^{\mathrm{C}} = f^{\mathrm{CI}}(P) \text{ with } f^{\mathrm{CI}} : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}. \tag{3.6.20}$$

The CI model is trained on 3.94 million randomly sampled patches from 8492 images on the task of patch-wise binary cancer detection. It achieved an AUROC of 0.938 and an area under the precision-recall curve (AUPRC) of 0.890 on 0.51 million test patches of 1062 unseen images. However, detailed examination of the cancer indicator model is out of the scope of this work.

Cancer Heatmaps

Inference of the trained CI model is performed on patches from all remaining datasets in this work. The predicted probabilities $\hat{y}^{\mathrm{C}}$ for every patch P are used to create a cancer mask for every image, as shown in 3.5.3. By defining relevant regions based on the tissue masks, cancer probability heatmaps can be visualized as an overlay on the original slide image, as depicted in Figure 3.6.7B and C. Gaussian filtering is applied on the probability maps to smooth the patch-based resolution. These cancer heatmaps are expected to assist clinical practice by guiding the clinician to salient regions in the image, e.g. if human re-evaluation of an image is necessary due to a low credibility score provided by the risk prediction model.

Cancer-based Patch Selection

Besides providing visual support to clinical practice through cancer heatmaps, the CI is further used in PCAI to reduce noise and redundancy in the input data. In detail, when performing inference on the large biopsy images of the MMX and UPP dataset, the cancer masks are used to select the 100 patches expressing the highest cancer probability in the relevant tissue region (see Figure 3.6.7C). Only these patches are then forwarded to the risk prediction network. Since cancerous tissue supposedly incorporates the most relevant information for risk prediction and biopsy images often contain large amounts of non-cancerous tissue, this is expected to focus PCAI on the

salient regions of the input image and improve overall predictive accuracy as well as robustness. Assuming that the TMA spot datasets used in this work contain mostly cancerous tissue, the CI-based patch selection is not applied to those datasets. Furthermore, since patch selection and therefore risk prediction of biopsy images is based on the CI, the heatmaps can also be regarded as a means of visual interpretability that highlight salient regions in the overall PCAI model.

3.6.4 Aggregation of Multiple Images per Patient

If multiple images are available for a single patient, in both PCAI and the BASE model, predictions are aggregated by taking only the highest risk score across images as the final patient score.

In detail, if N images are available for a patient, the final patient level risk score R is then derived as

$$R = \max_{n} \left(R_n \right). \tag{3.6.21}$$

Accordingly, the final credibility score $Cred$ refers to the credibility of the image expressing the highest risk, such that

$$Cred = Cred_n \text{ with } n = \arg\max_{n} \left(R_n \right). \tag{3.6.22}$$

3.6.5 Metrics

Two main metrics are used in the course of this work to evaluate the predictive performance of the deep learning models as well as the ISUP rating assigned by the pathologists.

Area under the Receiver-Operator Characteristic Curve (AUROC)

The area under the receiver operator characteristic curve (AUROC) is a metric that is widely used in binary classification tasks. It summarizes the model's discriminative capability across all possible classification thresholds in a single scalar. It is calculated by integrating the receiver operator characteristic (ROC) curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at each classification threshold setting of the predicted class probabilities. The AUROC is therefore a

threshold-independent measure and robust against class imbalance in the data. It can be calculated by using the trapezoidal rule as

$$\text{AUROC} = \sum_{p=1}^{N_p-1} \frac{(FPR_{p+1} - FPR_p) \times (TPR_p + TPR_{p+1})}{2}. \tag{3.6.23}$$

Here, FPR_p is the False Positive Rate at threshold p, defined as the ratio of false positives to the total number of actual negatives. TPR_p is the True Positive Rate at threshold p, also known as sensitivity, defined as the ratio of true positives to the total number of actual positives. The sum is taken over all thresholds p from 1 to $N_p - 1$, where N_p is the number of unique predicted probabilities. An AUROC of 1 corresponds to the best model prediction, and a value of 0.5 represents a random prediction.

In the course of this work, positive instances are defined as patients who experienced a relapse prior to five years, as described in 3.5.1, and negative instances as those who did not. This is referred to as the 5-year AUROC throughout this thesis and the main metric used for model training and optimization.

Concordance Index (C-Index)

Additionally, the concordance index (C-Index) is evaluated in this work, which is a generalization of the AUROC that can take into account censored data [162]. The C-Index measures the proportion of comparable pairs of individuals for which the predicted risk scores are consistent with the observed outcomes over the total number of comparable pairs. It is independent of the binary relapse label required for the AUROC and can evaluate the full available test data. In detail, it can be calculated as

$$\text{C-Index} = \frac{\sum_{i,j}^{n} 1_{T_j<T_i} \cdot 1_{R_j>R_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j<T_i} \cdot \delta_j}, \tag{3.6.24}$$

where

$$1_{T_j<T_i} = \begin{cases} 1 & \text{if } T_j < T_i \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad 1_{R_j>R_i} = \begin{cases} 1 & \text{if } R_j > R_i \\ 0 & \text{otherwise} \end{cases}. \tag{3.6.25}$$

Here, R_i and R_j are the predicted risk scores for the i-th and j-th instances, T_i and T_j are their respective follow-up times and δ_j is the censoring indicator of the j-th

instance, which is 1 if j is uncensored and 0 if j is censored. n depicts the total number of samples. Similarly to the AUROC, a C-Index of 1 corresponds to the best model prediction, and a value of 0.5 represents a random prediction.

Evaluation of Cancer Grading Provided by the Pathologists

For evaluation of the ISUP annotations of the pathologists, the ISUP scale ranging from 1 to 5 is rescaled into a risk score between 0 and 1, where ISUP 1 corresponds to a risk score of 0 and ISUP 5 correpsonds to a risk score of 1. AUROC and C-Index are then calculated as described above.

3.7 Experiments

This section describes all experiments conducted with the proposed deep learning models BASE and PCAI, as well as the cancer grading provided the pathologists, with their quantitative and qualitative results.

3.7.1 Building a Clinically Applicable Model

In the following, the proposed deep learning models BASE and PCAI are evaluated with respect to the four initially defined pillars of clinical applicability, namely robustness, trustworthiness, human-level performance and interpretability.

Robustness

In the first experiment, the BASE model, trained on the single data domain UKE.first, and the PCAI model, trained on the three data domains UKE.first, UKE.second and UKE.scanner and including the algorithmic extensions of domain adversarial training (DA), credibility estimation (CE), credibility-guided color adaptation (CE-CA and cancer indication (CI), are evaluated on the unseen test sets of all internal UKEhv subdatasets, as depicted in Table 3.7.1. It can be seen that PCAI scores higher than the BASE model on every dataset in terms of concordance index and 5-year relapse AUROC, even on the UKE.first domain, which was included in the training data of both models, indicating a general stronger predictive capability of PCAI over BASE. Notably though, performance of both models is lowered on images from datasets other than UKE.first, especially on the three experimental sub-datasts UKE.thin, UKE.thick and UKE.long. This suggests how data variation can significantly reduce the performance of AI-based decision systems.

Table 3.7.1: Discriminative performance of the PCAI and BASE model in terms of concordance index (C-Ind.) and five year relapse AUROC (AUC) on the unseen TMA spot test images of the internal UKEhv sub-datasets.

	UKE.first		UKE.second		UKE.scanner		UKE.thin		UKE.thick		UKE.long	
	C-Ind.	AUC	C-Ind.	AUC	C-Ind.	AUC	C-Ind.	AUC	C-Ind.	AUC	C-Ind.	AUC
PCAI	**0.695**	**0.723**	**0.661**	**0.698**	**0.674**	**0.702**	**0.617**	**0.630**	**0.609**	**0.619**	**0.634**	**0.654**
BASE	0.685	0.719	0.602	0.623	0.637	0.663	0.564	0.581	0.573	0.584	0.586	0.601

To systematically investigate this effect of data variation on AI-based PCa grading performance, evaluation of the BASE and PCAI models is repeated on the subset of 1,537 patients that contribute an image to each of the six UKEhv sub-datasets. This sub-cohort is referred to as UKEhv_6 in the following and constitutes an ideal basis for the evaluation of data variation on model performance. Assuming images that are equally gradable and only differ by their domain specific bias, a perfectly robust model should predict similar score distribution for all sub-datasets and achieve similar discriminative performance, since the same ground truth accounts to the images of every protocol variation. When evaluating the BASE model on the UKEhv_6 patients of the UKE.first dataset, the same domain it was trained on, it achieves a concordance-index of 0.645 and predicts a median risk score across samples of 0.46. Figure 3.7.1 depicts the difference in concordance-index to that initial score when evaluating the BASE and PCAI model on the same UKEhv_6 patient cohort on all remaining UKEhv sub-datasets. Additionally, the distribution of predicted risk scores is displayed. Notably, performance of the BASE model drops significantly on images that are acquired with a different clinical protocol, by a minimum of 5 percentage points on the UKE.scanner images to 8.6 percentage points the on UKE.thin. When analyzing the distribution of predicted risk scores on the UKEhv_6 patients of the remaining subdatasets, strong deviations from the reference distribution on UKE.first are visible, with a median risk score of up to 0.77 on the UKE.thin dataset. This highlights the sensitivity to data variation in the BASE model and indicates that AI-based models, even when trained on large datasets, have significant difficulties with data variations they were not trained on. When evaluating the proposed PCAI model on the same images, the drop in performance on data from different acquisition protocols is reduced to a maximum of 3.7 percentage points on the UKE.thick data, while it performs on par on the UKE.scanner data. On the main domain UKE.first, it performs even 2.1 percentage points better than the BASE model. When analyzing the distribution of risk scores, PCAI predicts a similar distribution and median of 0.46 on the UKE.first data. However, in comparison to the BASE model, this distribution only shifts marginally across the remaining sub-datasets, down to a minimum median score of 0.40 on the UKE.long data. These results strongly suggest that the algorithmic modifications for robustness and credibility in PCAI increase its generalization capabilities over the BASE model and contribute to robustness to a broader spectrum of data variations.

These findings on the internal UKEhv images further translate to the performance

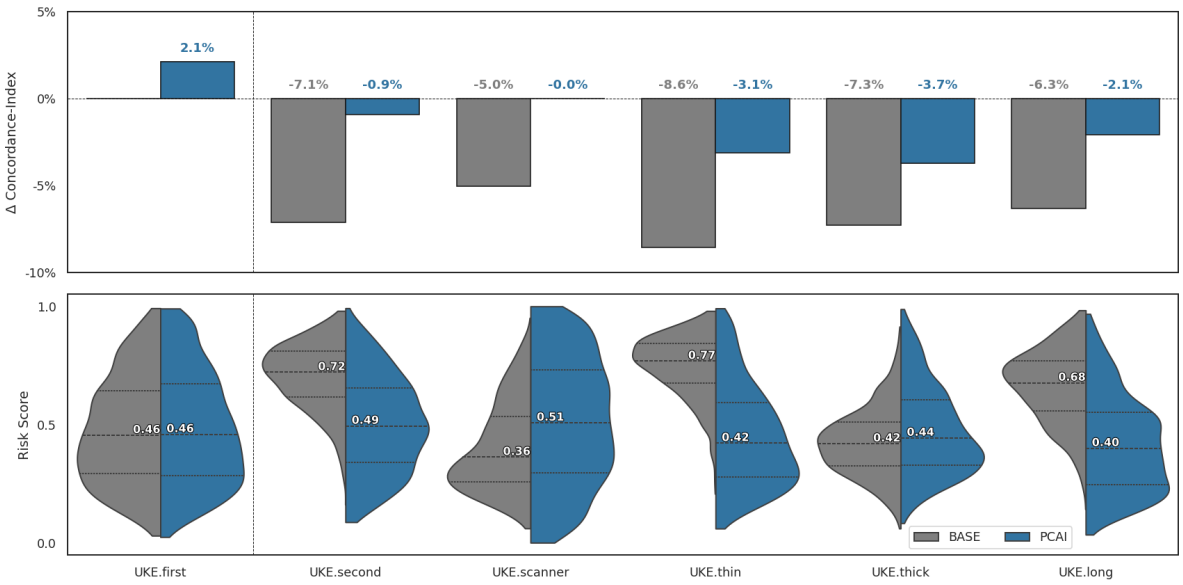

Figure 3.7.1: The effect of data variation on AI-based PCa grading performance: The UKEhv dataset contains all protocol variations for a subset of the same 1,537 $UKEhv_6$ patients, taken from the same RP sample. Assuming images that are equally gradable and only differ by their domain specific bias, a perfectly robust model should predict similar score distribution for all sub-datasets and achieve similar discriminative performance. Difference in performance of the PCAI and BASE model on all six protocol variants is depicted with respect to the concordance index of 0.645 that BASE achieved on the UKE.first dataset (top). Distribution of the predicted risk scores of PCAI and BASE across images for every protocol variation (bottom).

of BASE and PCAI on the TMA test datasets NYU and JHU, which are employed to assess the generalizability of the proposed models on external data, as depicted in Table 3.7.2. Here, PCAI outperformes BASE by 5.3 percentage points and 8.0 percentage points on the NYU data in terms of concordance index and 5-year relapse AUROC, respectively, and by 1.0 percentage points and 3.2 percentage points on the JHU data.

Table 3.7.2: Discriminative performance of the PCAI and BASE model in terms of concordance index (C-Ind.) and five year relapse AUROC (AUC) on the unseen TMA spot images of the external NYU and JHU datasets.

	NYU		JHU	
	C-Ind.	AUC	C-Ind.	AUC
PCAI	**0.694**	**0.744**	**0.587**	**0.638**
BASE	0.641	0.664	0.577	0.606

Trustworthiness

The risk stratification results on the TMA datasets show that PCAI is able to provide stable predictions across on a broad spectrum of sample processing protocols. However, it might still encounter histopathological slides of e.g. bad quality, for which it cannot provide a reliable grading. A relevant feature for PCAI is therefore the notion of confidence or trustworthiness via credibility estimation (CE), not unlike a human expert that is uncertain about the grading of a particular sample and asks for a second opinion. To this end, PCAI assigns a credibility score to every image with the aim to detect problematic samples. It then alters those images by matching its color with the training distribution in the attempt to enable a credible prediction and improve performance. To verify whether the proposed credibility-guided color adaptation (CE-CA) setup enables PCAI to correctly detect and correct problematic samples, the improvement in AUROC when applying credibility-guided color adaptation (CE-CA) as well as color adaptation of all samples (all-CA) over unaltered predictions of PCAI is depicted in Figure 3.7.2 for all datasets used to assess its robustness. Percentages of adapted samples per dataset for CE-CA are noted on top of every bar. Here, the proposed CE-CA improves overall performance across datasets by 7 percentage points, while un-guided adaptation of all images (all-CA) improves in sum by only 2 percentage points. Notably, on the NYU data, CE-CA turns a decrease by 1.7 percentage points into an increase of 4.5 percentage points, highlighting how the credibility score correctly guides PCAI towards problematic samples. While on the UKE.thin, UKE.long and JHU data adaptation of all samples achieves a stronger improvement than CE-CA, it leads to a decrease in performance in five out of eight datasets. Guiding the color adaptation by the credibility score only leads to a neglectable decrease by 0.3 percentage points on the UKE.thick data, while otherwise increasing performance. On the UKE.thin dataset, half of the potential performance increase is fulfilled by adapting only 15% of overall images based on the assigned credibility. In summary, it can be observed that CE-CA improves over un-guided adaptation, indicating that the proposed Mahalanobis-distance-based credibility score is beneficial in identifying problematic samples. Furthermore, the proposed histogram-matching-based color adaptation (CA) procedure enables PCAI to successfully alter images in a way that allow for a increasingly correct risk assessment and more credible predictions.

The CE setup is further applied to the BASE model to allow for a quantitative com-

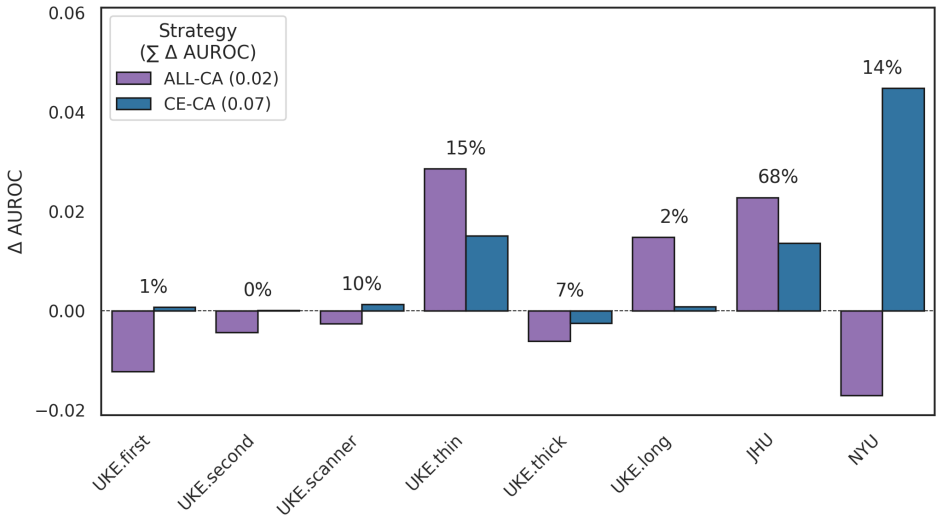

Figure 3.7.2: Difference in AUROC in PCAI when adapting the color of all samples (all-CA) of every dataset by default (purple) or when using credibility-guided color adaptation (CE-CA) (blue). Percentage of adapted samples for CE-CA per dataset is depicted on top of each bar.

parison of credibility scores to the PCAI model. Figure 3.7.3 depicts the average credibility score across all test samples for the datasets used to assess the models robustness. It can be seen that credibility in the BASE model drops drastically close to zero when predicting on data outside its training domain UKE.first. In PCAI, credibility scores are significantly higher across all domains and surpass the average credibility of BASE on its training domain in five out of eight cases. Notably, the lowest credibility scores in PCAI were assigned to samples of the UKE.scanner domain, even though images from this domain were used to train the domain adversarial network. This highlights the significant domain shift that is introduced by different scanner types into histopathological images.

Human-level Performance

After assessing the capability of PCAI over the BASE model to provide robust and credible predictions across a wide variety of data, the litmus test for the proposed deep learning system lies in whether it can provide a competitive stratification of cancer aggressiveness to the currently used medical gold standards of Gleason, ISUP and GIQ assigned by the pathologists. While the datasets used to build and assess robustness of the PCAI model include an ISUP score per patient, this score corresponds

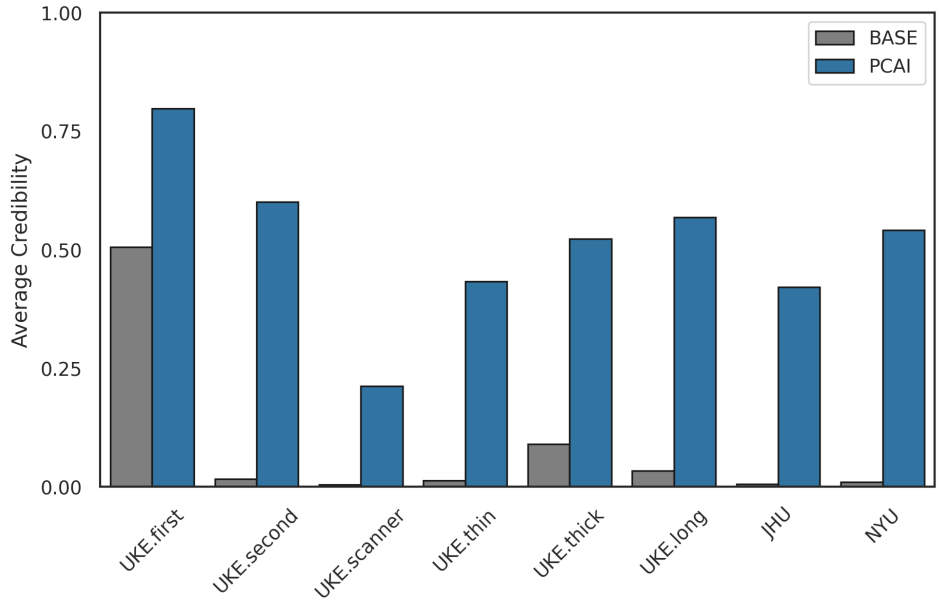

Figure 3.7.3: Average credibility scores of PCAI and BASE on the datasets used to build and assess the model's robustness.

to the full patient rather than an individual image and was derived from analyzing significantly more tissue than the single TMA spot image available to PCAI. Comparing predictions of PCAI on a single image to this patient-level ISUP would be disadvantageous to the deep learning model and result in an inconclusive evaluation. To ensure a fair comparison under equal conditions, the three datasets with image-level human annotations are utilized for benchmarking PCAI, namely UKE.sealed, UPP and MMX.

PCAI Surpasses Expert Cancer Grading on TMAs

The UKE.sealed TMA spot dataset contains a spot-level ISUP grading from UKE pathologists. In addition, the UKE.sealed TMA spots were graded using the Integrated Quantative Gleason (GIQ) grading system by an internationally renowned pathologist. The GIQ is currently one of the best performing grading systems for PCa histopathology [116]. In this dataset, up to eight annotated images are available for a single patient. To derive the final patient-level score, the maximum ISUP across images is taken as the final prediction. For the GIQ as well as PCAI, both mean and maximum aggregation strategies are evaluated. Table 3.7.3A depicts the risk stratification performance of PCAI and human annotations in terms of concordance index and AUROC. Comparing similar image-level prediction aggregation strategies, using

maximum aggregation, PCAI outperforms ISUP annotations by 2.6 and 2.4 percentage points in terms of C-Index and AUROC, respectively, and the more advanced GIQ score by 1.0 and 1.1 percentage points. Using mean aggregation, PCAI and GIQ perform on par in terms of C-Index, while GIQ scores with 0.7 percentage points slightly higher than PCAI. These results indicate that PCAI is able to keep up with existing human-annotated scoring systems on TMA images, even in terms of the advanced GIQ score provided by an highly experienced pathologist. It even surpasses the widely employed ISUP score by more than two percentage points in both metrics.

PCAI Surpasses Expert Cancer Grading on Biopsies

While the results on the UKE.sealed dataset indicate the capabilities of PCAI to extract highly predictive features from post-operative TMA spot images, a clinically relevant application of PCAI would most likely involve pre-operative biopsies. It is therefore of particular interest if the feature representations learned by training on TMA spot data translate their predictive value when performing inference on biopsy images. To verify this, PCAI is evaluated on the two biopsy datasets UPP and MMX. On both datasets, image-wise ISUP annotations, obtained during routine diagnostics, are available. Additionally, on the MMX dataset, three individual pathologists annotated all slides independently and blinded from any additional patient information, to obtain a notion of human inter-rater variability and assess significance when comparing to the proposed model. The ISUP provided by the three pathologists of two centers (Aachen and Uppsala) show an interrater agreement Fleiss kappa of 0.199. Maximum aggregation of scores in the case of multiple images for a single patients is utilized both for PCAI and ISUP.

As shown in Table 3.7.3C, PCAI achieves a C-Index of 0.604, 0.7 percentage points higher than ISUP, on the UPP dataset. In terms of AUROC, PCAI outperforms ISUP by 1.3 percentage points.

Table 3.7.3B depicts performance of PCAI and the human annotators on the MMX dataset. Here, PCAI achieves an C-Index of 0.864, 4.7 percentage points higher than ISUP, and an AUROC of 0.868, 5.5 percentage points higher than ISUP. When comparing to the image-wise ISUP grading of three highly skilled pathologists from Germany and Sweden (A1, A2, A3), PCAI significantly exceeds the performance of the expert ISUP grading (A1: 0.838, A2: 0.834, A3: 0.641, mean: 0.771) by 9.3 (mean) percentage points. This holds also true for in terms of AUROC, where PCAI (0.868) significantly surpasses expert ISUP grading (A1: 0.827, A2: 0.827, A3: 0.657, mean:

0.770) by 9.8 (mean) percentage points.

Taken together, these results on biopsy-derived whole slide images indicate not only that feature representations learned by training PCAI on TMA-spot data remain predictive on biopsy images, they also enable PCAI to surpass predictive performance of expert ISUP grading in all cases on both datasets.

Table 3.7.3: Discriminative performance of the PCAI model and image-wise ISUP and GIQ annotations assigned by various pathologists in terms of concordance index (C-Ind.) and five year relapse AUROC (AUC). A: Performance on the unseen UKE.sealed TMA spot images. Aggregation of multiple images is performed by either taking the maximum score (-max) or the average score (-mean) for both PCAI and GIQ. B: Performance on the unseen MMX biopsy images. C: Performance on the unseen UPP biopsy images.

UKE.sealed	C-Ind.	AUC
PCAI-max	0.739	0.781
PCAI-mean	**0.744**	0.780
GIQ-max	0.729	0.770
GIQ-mean	0.743	**0.787**
ISUP	0.713	0.757

(a) UKE.sealed

MMX	C-Ind.	AUC
PCAI	**0.864**	**0.868**
A1	0.838	0.827
A2	0.834	0.827
A3	0.641	0.657
ISUP	0.817	0.813

(b) MMX

UPP	C-Ind.	AUC
PCAI	**0.604**	**0.672**
ISUP	0.597	0.659

(c) UPP

Interpretability

The previous results show that PCAI is able to predict a robust and credible risk score that surpasses human-annotated ISUP on TMA and biopsy datasets. However, for actual clinical applicability it is pivotal that these predictions are accompanied with notions of interpretability, which allows the expert to trust or ignore a model's prediction. PCAI approaches this two-fold: Firstly, by deriving distinct risk groups to transform the continuous cancer grade score into clinically interpretable categories. Secondly, by using the cancer indicator to provide visual cues on images as to the location of cancerous areas.

Risk Groups

On the output side, correlation of risk with therapeutic options might be challenging in a clinical setting using a continuous score. To enhance interpretability of the predicted risk score, distinct risk groups can be derived by separating patients into a "low" and "high" risk group based on the median predicted risk score per dataset. Figure 3.7.4 depicts the Kaplan-Meier curves for the low and high risk patients on the UKE.first as well as all external datasets. It can be seen that both risk groups separate patients well over time, indicating their potential for correlation with respective treatment options.

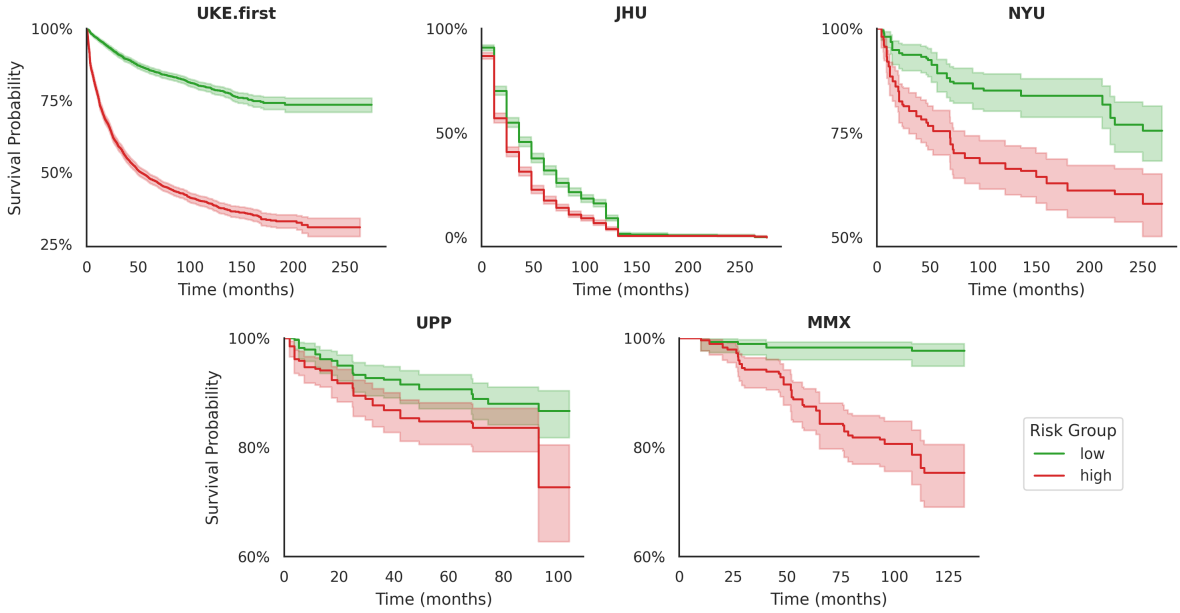

Figure 3.7.4: Kaplan-Meier curves of the low and high PCAI risk groups on UKE.first as well as all external datasets. Risk groups are separated on the medium predicted risk score per dataset.

However, it is to note that the predicted risk scores on the external datasets express a distribution shift in comparison to the UKE.first dataset, as depicted in Table 3.7.4. Here, PCAI predicts higher median risk scores on the JHU (0.654), UPP (0.875) and MMX (0.714) dataset and a lower median risk score for the NYU (0.265) dataset. Notably, except for the JHU data, not the full available range of of risk scores between 0 and 1 is utilized. On both biopsy datasets, the minimum predicted risk score is significantly higher than on the TMA spot data. Conversely, on the NYU dataset, the maximum predicted risk score is 19.7 percentage points lower than on UKE.first.

Table 3.7.4: Minimum, median and maximum predicted risk score of PCAI on all test samples of the UKE.first, JHU and NYU TMA spot datasets and the UPP and MMX biopsy datasets.

	Min	Median	Max
UKE.first	0.020	0.454	1.000
JHU	0.093	0.654	0.997
NYU	0.056	0.265	0.803
UPP	0.272	0.875	0.997
MMX	0.193	0.714	0.999

Cancer Indication

To provide additional interpretability on the image level, the CI module is utilized to create cancer probability heatmaps on all images. These aim to provide visual cues to the human reader, especially in cases where a low credibility score in PCAI suggests human re-evaluation of an image. Figure 3.7.5 depicts exemplary heatmaps for TMA spot and biopsy data.

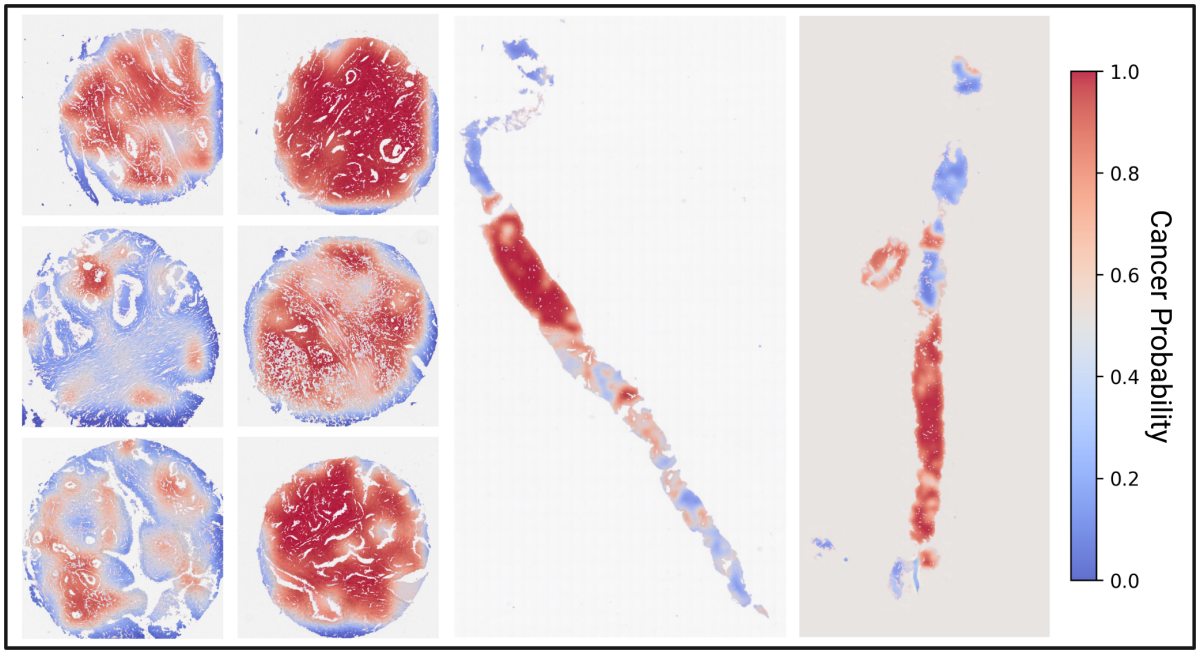

Figure 3.7.5: Cancer probability heatmaps provided by the CI module of PCAI on exemplary TMA spot and biopsy images.

Since the CI module is also utilized to guide PCAI on highly cancerous regions for the biopsy datasets, the influence of this CI-guided patch selection on predictive accuracy

is depicted in Table 3.7.5. Here, concordance index, average credibility score and Mahalanobis distance d_M on the UPP and MMX dataset are given for the default PCAI model, where only the 100 most cancerous patches according to the CI are forwarded during inference, and the PCAI model without CI sampling (PCAI \CI), where all valid patches per biopsy image are forwarded. It can be seen that in terms of concordance index, CI-based patch selection improves performance on both datasets, by 3.1 percentage points on the UPP data and 7.2 percentage points on the MMX data. However, CI-based patch selection significantly reduces average credibility scores on both datasets. This is further emphasized by correlating increased Mahalanobis distances d_M to latent center of the training distribution.

Table 3.7.5: Concordance index as well as average credibility score and Mahalanobis distance to the training center across samples of the UPP and MMX biopsy datasets when forwarding the 100 most cancerous patches according to the CI module (PCAI) and when forwarding all patches per image (PCAI \CI).

	UPP			MMX		
	C-Ind. ↑	Cred ↑	d_M ↓	C-Ind. ↑	Cred ↑	d_M ↓
PCAI	0.604	0.084	66.96	0.864	0.178	57.55
PCAI \CI	0.573	0.393	37.74	0.792	0.669	28.84

3.7.2 Extended Analysis

While the previous experiments highlight the rationale behind building PCAI with the aim for clinical applicability, the following evaluations aim to analyse the final PCAI model and its individual components more in depth from a technical perspective.

Latent Space Analysis

Additional analysis of the algorithmic adaptations for robustness in PCAI is performed in accordance with findings from Stacke et al. who claimed that the sensitivity of a deep learning model to covariate shifts in the input data should be quantified by measuring the discrepancy of domains in the model's latent space [97]. Here, less clustered and entangled latent space representations of individual data domains are expected to generally indicate a less domain sensitive and more robust model. To this

end, UMAP representations of the latent features after the patch aggregation layer of all test images used in this work are derived for the incremental levels of complexity from the BASE model to PCAI and depicited in Figure 3.7.6. Additionally, the sum of Mahalanobis distances of the latent centers per dataset (corresponding to the mean feature vector across samples per dataset) to the latent training center is shown for every step.

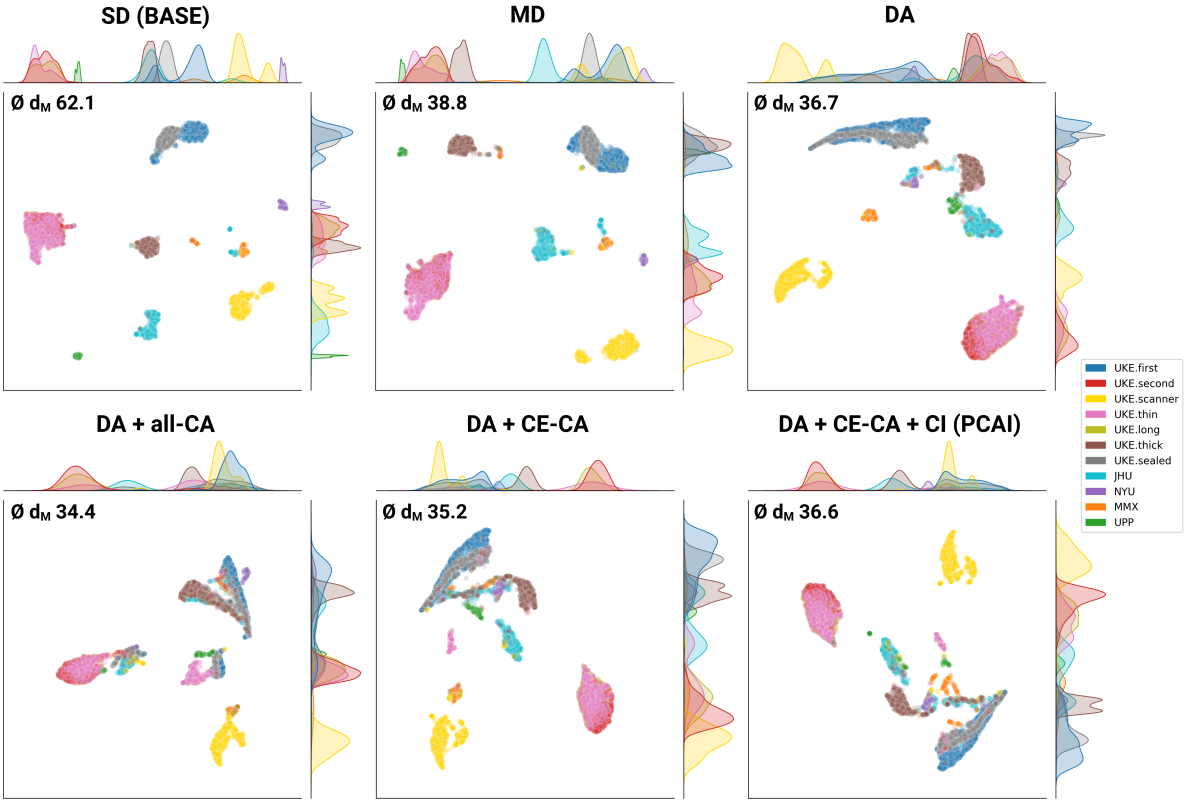

Figure 3.7.6: UMAP representations of the image-wise latent features after the patch aggregation layer for incremental increases in model complexity from BASE to PCAI. SD: single domain, MD: multi domain, DA: domain adversarial, all-CA: color adaptation of all images, CE-CA: credibility-guided color adaptation, CI: cancer indicator guided patch selection.

SD (BASE): For the single domain (SD) training of the BASE model, a strong clustering of latent features based on the dataset of origin can be observed. Samples from UKE.first and UKE.sealed seem to express high similarity, since they cluster together. Similarly, samples from UKE.second, UKE.thin and UKE.long form a single cluster. The other datasets form separate clusters. Notably, those dataset which contain images from multiple scanners, namely MMX and JHU, form multiple clusters, highlighting the covariate shift caused by a different image acquisition device. The

overall Mahalanobis distance of dataset centers to the training center in the SD case is 62.1.

MD: The subsequent UMAP depicts the latent space that was trained jointly under a multi domain (MD) training regime, with data from UKE.first, UKE.second and UKE.scanner, but without utilizing the domain adversarial (DA) training method. Qualitatively, no significant difference in clustering can be observed in comparison to the SD case, indicating that the model still expresses high sensitivity to the input domain. In contrast, the average dataset Mahalanobis distance to the training center is drastically reduced to 38.8. However, this is mainly linked to the absolute shift in position of the training center, since the training data now consists of three domains instead of only one in the SD case.

DA: When applying the domain adversarial (DA) method to the multi domain training regime, qualitative analysis of the UMAP reveals a noticeable difference to the MD case. While the UKE.scanner cluster, UKE.first & UKE.sealed cluster and the UKE.second & UKE.thin & UKE.long cluster are still present, data from UPP, JHU, NYU and UKE.thick moves significantly closer together and becomes less distinguishable. Only one of the two MMX clusters, representing one of the used scanners, still separates clearly from the remaining external data. This indicates the beneficial influence of the domain adversarial training method to the model to become less sensitive to the origin of the input data. The reduced average Mahalanobis distance of 36.7 additionally quantitatively emphasizes this finding.

DA + all-CA: Applying color adaptation to all images (DA + all-CA) further reduces the average Mahalanobis distance to 34.4. This highlights how matching the color of samples with histograms of the training distribution moves them indeed closer to the learned distribution. Notably, MMX forms no separate cluster in that case.

DA + CE-CA: When utilizing credibility-guided color adaptation (DA + CE-CA), the average Mahalanobis distance increases again to 35.2. However, this is expected behaviour due to the reduced number of adapated samples compared with the all-CA case. When qualitatively comparing the UMAP with the DA case without color adaptation, the CE-CA case appears more entangled, proving its robustness-conferring influence.

DA + CE-CA + CI (PCAI): Finally, additionally utilizing the cancer indicator for patch selection represents the overall proposed risk prediction model PCAI. Average Mahalanobis distance increased further to 36.6, which is in line with the findings in Table

3.7.5. Qualitatively, samples from the MMX data are slightly more separable than in the previous case.

In summary, it can be observed that applying the proposed algorithmic adaptations in PCAI leads qualitatively to a more entangled latent space and quantitatively moves latent representations closer together, when comparing with the SD or the MD model. This highlights the decreased sensitivity of PCAI to the dataset origin of the input images and corresponding covariate shifts caused by variations in clinical acquisition protocols. It is however to note that the findings in Table 3.7.5 show that a higher Mahalanobis distance to the training center does not necessarily correlate with a lower predictive accuracy, and can, in case of applying the CI for patch selection, actually improve risk stratification results. Finally, while the relative improvement between BASE and PCAI is clearly visible, especially the UKE.scanner and UKE.second & UKE.thin & UKE.long data still forms highly distinguishable clusters, indicating remaining domain sensitivity of the model even after applying the proposed adaptations for robustness.

Color Dependency of the Risk Prediction

While the previous analysis of the latent space evaluated the sensitivity of the model to inter-domain differences, building robustness also aims to decrease sensitivity to intra-domain variation, since differences in staining, lighting or acquisition protocol can also be present across images of a single dataset. It is hypothesized that these differences mostly manifest in terms of a color shift of the images. To understand whether the adaptations for robustness in PCAI led to a reduced sensitivity of its predictions to intra-domain-specific variations in color, the squared Pearson correlation coefficient r^2 between the predicted risk scores per image and the Wasserstein distance of their histogram to the center of the training distribution in the HSV color space is calculated for BASE and PCAI. It is assumed that a high Wasserstein distance of an image's histogram to the average histogram of the training images indicates a stronger color-based covariate shift. Figure 3.7.7 depicts the Pearson r^2 of BASE and PCAI for all datasets used in this work. While PCAI shows an increased sensitivity to color on the UKE.thin and UKE.second images, on the UKE.scanner the Pearson r^2 is reduced almost 8-fold, from 0.151 to 0.019. Similarly, on the MMX data, the Pearson r^2 of PCAI is reduced to a fraction of the BASE model, from 0.144 to 0.002. The sensitivity to color is also lower in PCAI for all remaining external datasets. These findings

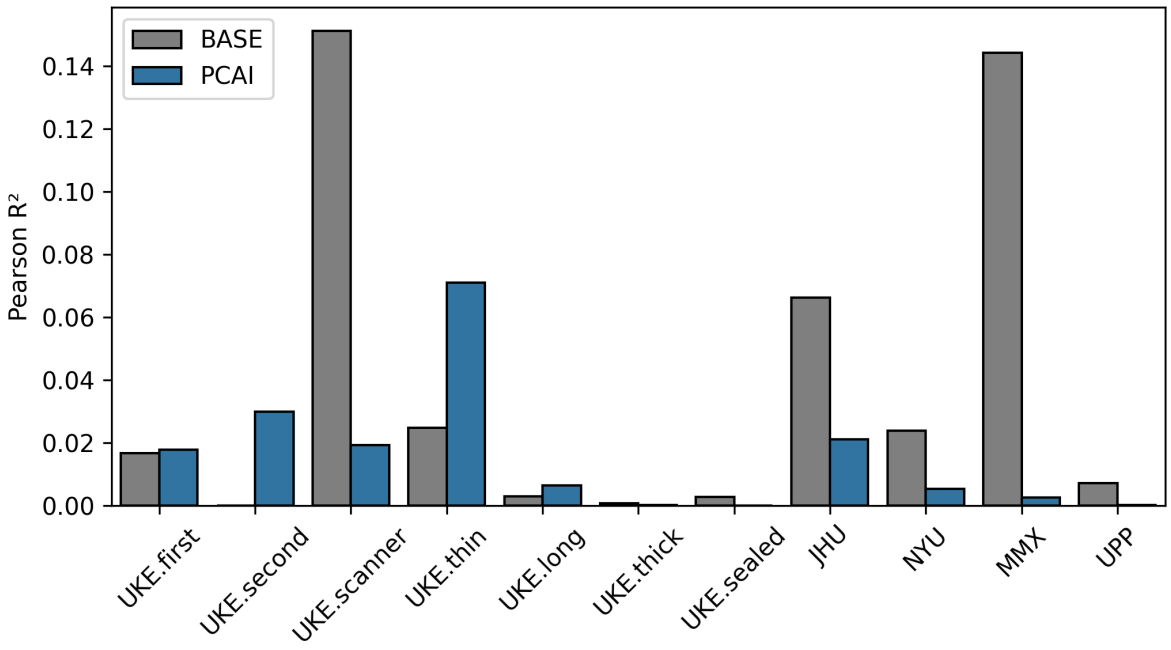

Figure 3.7.7: Color dependency of the risk score: Squared Pearson correlation coefficient r^2 between the predicted risk scores of PCAI and BASE per image and the Wasserstein distance of the corresponding HSV color histogram to the center of the training distribution.

indicate that the algorithmic adaptations in PCAI successfully decrease sensitivity to intra-domain-specific variations in color in the data and analogously increase robustness of the proposed model.

Discarding Images Based on Credibility

The proposed credibility score based on the latent Mahalanobis distance is inspired mainly by the work of Dietrich [142], who utilized a similar method for detection of in-distribution (ID) and out-of-distribution (OOD) samples. Their proposed eCareNet survival prediction model is closely related to the BASE model of this thesis and trained on the same patient cohort, though not completely identical. Furthermore the authors combined the works of Lee et al. [158] and Sun et al. [163] to define OOD samples based on the Mahalanobis distance to the nearest neighbor of the training distribution, whereas in this thesis the Mahalanobis distance to the training center is utilized. Dietrich et al. hypothesized that images with higher OOD scores are more often incorrectly predicted than images with lower OOD scores. They found that re-

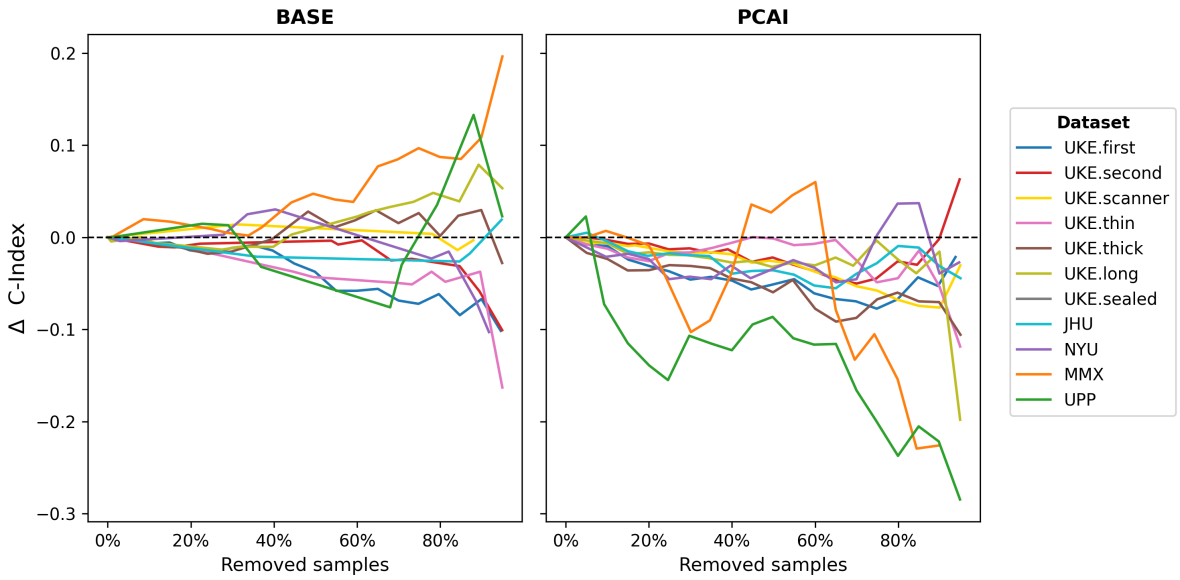

Figure 3.7.8: Difference in concordance index when removing increasing percentiles of images with lowest credibility scores from all datasets in BASE and PCAI. From left to right, remove no images up to removing 95%. Evaluation is performed in increments of 5 percentage points.

moving samples based on their OOD score slightly increased performance on some of the evaluated datasets, however, with conflicting results on others. Analogue evaluation of the credibility score as OOD measure on all datasets is performed for the BASE and PCAI models developed in the course of this work. Figure 3.7.8 depicts the difference in C-Index when predicting on datasets where increasing percentiles of samples expressing the lowest credibility scores are removed (see Figure A.2.1 for AUROC). It can be seen that with decreasing dataset size, absolute difference in performance increases for the BASE model. In line with the findings in the eCareNet, a positive difference in performance is visible for some datasets and a negative difference for others. For the PCAI model, removing samples based on their credibility score leads to more consistently negative results across datasets over varying thresholds. A potential explanation for this is that PCAI is specifically trained towards robustness and a more entangled latent space, as shown in Section 3.7.2. This might lead to a state where latent Mahalanobis distances are less predictive in terms of out-of-distribution detection as compared to the highly clustered latent space of the BASE model. It is however to note that C-Index and AUROC are discriminative metrics which compare the ranking of patients and are thus strongly dependent on the underlying cohort. Since dataset sizes change during the evaluation, comparing predictive performance

in terms of absolute numbers on different cohort sizes might be misleading [142]. When adapting the color of samples expressing a low credibility score instead of discarding them, as shown in Section 3.7.1, an overall increase is predictive accuracy is visible, indicating that the credibility score detects samples far from the learned distribution. These results are more representative, since the discriminative evaluation metrics were computed on the same underlying patient cohorts.

In summary, discarding samples based on the credibility score does not lead to a consistent improvement in the BASE model, which is in line with the literature. In the PCAI model, a decrease can be observed. However, previous evaluations in this work showed that the credibility score is an appropriate measure to detect samples that benefit from color adaptation. Future work should explore more suited methods to detect OOD samples, if an increase in performance is aspired. Different nonconformity functions than the Mahalanobis latent distance used in this work can be easily explored through the plug-and-play nature in proposed conformal prediction (CP) based credibility estimation (CE) setup of PCAI.

3.8 Discussion

In the present study, PCAI is proposed as a fully automated end-to-end PCa risk assessment pipeline that aims at clinical applicability, defined by four key criteria of robustness, trustworthiness, human-level performance and interpretability. It builds on one of the largest and most heterogeneous PCa image datasets paired with patient follow-up information collected to date, consisting of 83,864 images from 25,591 patients and five distinct centers. Several algorithm adaptations over the additionally derived BASE model are applied in PCAI, which utilize the rich and diverse underlying dataset collected in this work. This is made possible by the extensive mask-based image preprocessing and metadata integration performed in the course of this work, which allows to utilize image data from various centers and acquisition protocols, as well as TMA spot images and up to two orders of magnitude larger biopsy images in a single deep learning model. In the following, it will be discussed to which extend the four initially defined key requirements of PCAI are achieved.

Robustness

The effect of data variance encountered in clinical practice on the predictive accuracy of the DL models developed in this work is shown on the UKEhv$_6$ subset, which contains all six acquisition protocol variations for a subset of the same 1,537 patients, taken from the same RP sample. Here, discriminative performance in terms of C-Index of the BASE model is reduced on all datasets that lie outside its training distribution, up to 8.6 percentage points on UKE.thin. In PCAI, this decrease is significantly reduced across datasets, to a maximum of 3.7 percentage points. This robustness to variance in the input images further extends to the external JHU and NYU TMA data, where PCAI outperforms the BASE model in both C-Index and AUROC. Additionally, while the BASE model expresses noticeable distribution shifts in risk score prediction when confronted with data acquired with a protocol variation, PCAI predicts more homogeneous distributions across the UKEhv$_6$ images, which is desired behaviour for data of the same underlying patient population. These results prove the robustness-conferring influence of the algorithmic adaptations in PCAI over BASE. When moving further from evaluating absolute performance, evaluation of the UMAPs and average dataset center Mahalanobis distances to the training center reveal a more dense and entangled latent space in PCAI, suggesting that the encoder in PCAI extracts less domain-specific information from the images. Furthermore, analysis of the histogram Wasserstein distances shows that sensitivity of

the predicted risk score to shifts in the input color space is reduced. However, even though PCAI scores high in discriminative performance on all datasets, it is to note that on the external datasets a shift in the predicted ranges of the risk score is visible. While some variation in distribution is expected due to differences in the underlying cohorts, especially the high median risk scores on the two biopsy datasets UPP and MMX indicate a remaining sensitivity to the input domain of the images in PCAI. Future work might address this issue with a re-calibration when applying PCAI on a novel patient cohort. In summary, the results and evaluations show that the goal of robustness in PCAI is achieved.

Human-level Performance

The predicted risk score of PCAI is compared to image-wise ISUP and GIQ annotations of multiple renowned pathologists on one unseen TMA-spot dataset and two unseen biopsy datasets. On all three datasets, the risk score predicted by PCAI provides a more exact patient risk stratification than expert assigned ISUP grading. This shows that the feature representations learned in PCAI by training on post-operative TMA spot images remain predictive on pre-operative biopsy images with high accuracy. On the TMA-spot data of UKE.sealed, the PCAI risk score even surpasses PCa aggressiveness grading of the GIQ score assigned by an internationally renowned pathologist, when taking the maximum score across images per patient, and performes on par when taking the mean of scores across images per patient. These results on three unseen datasets strongly suggest that the risk score predicted by PCAI matches the performance of the currently most predictive PCa grading system GIQ, and even surpasses the predictive capabilities of ISUP. It can therefore be stated that the goal of human-level performance is achieved.

Trustworthiness

Building trust in a model's prediction is of utmost importance in clinical practice and a model should be equipped with the necessary means to quantify its confidence. In PCAI, this is attempted by the conformal prediction (CP) inspired credibility estimation (CE) setup, which computes a credibility score based on the latent Mahalanobis distance of an image to the learned distribution. PCAI scores higher credibility scores on all datasets than BASE, indicating that the algorithmic adaptations taken in PCAI enable it to confidently predict on an increased range and variety of data. These algorithmic adaptations include the feedback loop of color adaptation of low credible images. Using this credibility-guided CA procedure leads to a higher improvement in

AUROC across datasets than adapting all images by default, which concludes that the proposed credibility score does in fact detect samples that fall far from PCAI's learned distribution and underpins its aptitude as a measure for the models confidence in its prediction. If PCAI assigns a low credibility score to a potentially problematic image even after trying to fix it by adapting its color, the proposed workflow (depicted in Figure 3.6.6) suggests to defer the image to evaluation by a clinician. However, discarding images that express low credibility scores does not lead to an improvement in discriminative performance of PCAI on the remaining samples. Here it is to note that discriminative performance on cohorts of different sizes is comparable to a limited extend only [142]. In summary, PCAI is able to quantify its confidence in a prediction and to self-sufficiently correct problematic images it encounters, increasing discriminative performance and overall trust that can be placed in its predictions. However, on the task of OOD detection to discard samples based on the credibility score, further research should be conducted to find a more suited measure than the proposed Mahalanobis distance to the training center. The goal of trustworthiness is therefore, in my opinion, partly achieved.

Interpretability

Correlation of a continuous risk with therapeutic options might be challenging in a clinical setting, even if that score is highly discriminative. In this work, distinct "low" and "high" risk groups are derived based on the median predicted risk per dataset. It can be seen that both risk groups separate patients well over time, indicating their potential for correlation with respective treatment options. The exact number of risk groups to derive from the score and the treatment options to correlate with each group is subject of future work and requires further exchange with the clinical practitioners. However, as already mentioned when discussing PCAI's robustness, the predicted risk score expresses a distribution shift on the external datasets, especially on the biopsy data. Therefore, a definition of score boundaries for the risk group on the training data might not translate well to external data. Here, re-calibration of the risk scores on a new patient cohort can potentially address this issue. Finally, the cancer probability heatmaps provided by the cancer indicator provide visual cues and guide the clinician to the relevant regions of the biopsy images. Since only the most cancerous regions are forwarded to PCAI during inference, these heatmaps further highlight which region of the images contributed to the overall prediction, providing a means of explainability to the network. Furthermore, the attention weights assigned inside the PA layer of the model provide additional insights into the impor-

tance of every patch, achieving an even more fine-graded transparency. The goal of interpretability is therefore, in my opinion, partly achieved.

Limitations & Outlook

It is to note that this work encompasses some limitations. First and foremost, the very long training of up to over a week for the domain adversarial regime make hyperparameter optimization challenging. While the results indicate a well optimized model, a more extensive hyperparameter search in the future might potentially yield further improvements in predictive accuracy and robustness. Moreover, only a single instance with a single random seed of PCAI and BASE is evaluated. To get a better estimation of variance in the found parameter setting, evaluation of multiple trained instances with varying random seeds is desirable. Additionally, the literature showed the highly beneficial influence of strong color augmentations to the robustness of histopathological deep learning models [17]. While the AugMix augmentations used in this work are specifically designed for robustness of deep learning models in natural image processing, it is to note that they do not adapt the hue of the images. It is expected that additional augmentation of the image's hue channel is an easy to apply adaptation for increased robustness in future versions of PCAI. Moreover, the domain adversarial training regime is performed using the dataset of origin as label for domain classification. While this leads to the desired robustness and utilizes the inherent heterogeneity of the UKEhv sub-datasets, recent literature showed that performing adversarial regression on the H&E-stain matrix computed by Macenko's method further improved generalization capabilities [164]. Future work should utilize this over the adversarial domain classification, since this can be utilized to render the model even more robust to intra-dataset specific variation. Furthermore, the proposed credibility estimation setup is based on the concept on conformal prediction [159, 165]. However, in contrast to the original method, this work does not utilize the predicted class probabilities to derive the prediction sets of class point predictions. This stems from the fact that the predicted probability for class 1 (i.e. experiencing a relapse prior to five years) is interpreted as a continuous score and the class prediction just serves as a proxy during training. Future work should explore how to utilize the full potential of conformal prediction in PCAI, similar as has been shown by the work of Olsson et al. [100]. Finally, while it is shown that PCAI is a highly robust and accurate predictor that even outperforms ISUP as the current human assigned gold standard for PCa grading, it is to note that not all algorithmic adaptations always lead to a consistent increase in performance across datasets when applied individu-

ally to the underlying model. Instead, they function in unison to mitigate potential weaknesses and form a strong combined setup.

3.9 Conclusion

This work presents PCAI, a comprehensive deep learning based pipeline for prostate cancer risk assessment on histopathological images, designed specifically to meet the initially defined criteria of clinical applicability. PCAI builds on one of the largest histopathological prostate cancer dataset collected to date, encompassing 83,864 images from 25,591 patients and five different centers. With the goal of building a clinically applicable model, design decisions are based on meeting four key requirements, namely robustness, trustworthiness, interpretability and human-level performance. To this aim, several algorithmic adaptations that specifically utilize the heterogeneity and quantity of the underlying dataset are included into PCAI. These adaptations set it apart from a separately derived BASE model, which is trained on data from a single source and emulates the shortcomings of commonly used approaches in the literature. By applying the proposed adaptations of joint domain adversarial training, conformal prediction based credibility estimation, credibility guided color adaptation and cancer indication, PCAI improves over the BASE model for all defined requirements (**RQ-2**). In detail, by utilizing data from the UKEhv$_6$ subset of 1,537 patients, which includes TMA spot images of six different clinical protocol variations taken from the same RP sample, it is shown that performance of the BASE model degrades significantly on images acquired by a protocol outside its training domain, by up to 8.6 percentage points. In PCAI, domain adversarial training and credibility guided color adaptation reduce this drop to a maximum of 3.7 percentage points. Robustness further extends to two unseen external TMA spot datasets (**RQ-2**). Furthermore, predictive capabilities of the PCAI, which is trained on TMA-spot images from removed prostate specimen after RP, extends to biopsy images of two unseen external datasets (**RQ-PCAI**). This is assessed by comparing the predictive accuracy of the model's risk score against image-wise ISUP annotations of multiple human experts. Here, the predicted PCAI risk score provides a better risk stratification than ISUP in all cases, strongly suggesting its clinical value (**RQ-1**). This is further emphasized when comparing against the more sophisticated GIQ score on unseen TMA spot images, where PCAI performes on par with annotations of one of the worlds most renowned pathologists (**RQ-1**). Finally, with the added notion of trustworthiness in the models predictions through the credibility score and the interpretability for human re-evaluation provided by the cancer heatmaps (**RQ-3**), this work aims to provide a potential blueprint and pave the way for future real-world implemen-

tation of deep-learning based prostate cancer risk prediction models into the clinical workflow.

4 Overall Conclusion and Outlook

This thesis presents two projects in the field of medical image processing with deep learning, DeePSC and PCAI, with a specific focus on clinical applicability. In both projects, emphasis is put on fulfilling the initially defined key requirements of reaching human-level predictive accuracy, robustness, and trustworthiness.

DeePSC is a convolutional neural network based ensemble classifier that detects PSC on MRCP images and is specifically designed to process multiple images taken from different angular views around the patient in parallel. PCAI is an end-to-end risk prediction network that quantifies the cancer aggressiveness of prostate cancer patients based on their microscopic TMA-spot and biopsy images.

Both models surpass predictive accuracy of various human experts in their specific tasks of disease detection and patient risk stratification. Additionally, both models generalize on unseen data acquired with clinical protocols differing from that of the training distribution, thus reflecting robustness to data variance encountered in clinical practice. For PCAI, re-calibration to a new cohort might be advisable to allow for absolute interpretation of the predicted risk score. High predictive accuracy and robustness are arguably the most relevant requirements for an AI-based CDS: Scheetz et al. found in a survey that most clinicians had universally high expectations on the model's predictive accuracy, which forms the basis for acceptance of such a system in clinical practice [16]. To this end, the robust, superhuman performance of both proposed models in this thesis, DeePSC and PCAI, is assessed and proven on multiple datasets.

Besides designing the models to be highly accurate, they are further build such that their predictions are likely to be trusted in a clinical setting. Creating trust in the models is a nonlinear problem and the literature often only vaguely describes what clinicians actually expect from a 'trustworthy' model [166].

One commonly mentioned approach to this is to build the model 'explainable' or 'interpretable', e.g. to equip it with the necassary means to highlight how it arrived

at a particular decision [24]. In the proposed deep learning models of this thesis, this is attempted by providing visual cues of salient regions in the input images that contribute to the model's prediction. This is in line with the works of Evans et al., who found that pathologists preferred visual explanations in a CDS since those relate to their own way of thinking [22]. In DeePSC, GradCAM class activation heatmaps are derived on the last convolutional layer of the network. Here, high activation in the regions of the biliary tree reveal a focus of the model on the biologically relevant areas of the images. In PCAI, cancer probability heatmaps are derived on the full biopsy images, which are further used for selecting the relevant regions for the subsequent risk prediction. Besides revealing which areas contribute to the overall risk score, these aim to guide and support the physician if re-evaluation of an image is necessary. Additionally, in both projects, attention-based aggregation is utilized, for combining the latent representations of the seven MRCP views per patient in DeePSC and for combining the latent representations of patches per WSI in PCAI. Since the attention aggregation amounts to a weighted average function, the attention weights further allow for assessment of how much an individual MRCP view or image patch contributed to the overall final score. However, it is to note that in the literature, explainability as a means to build trustworthy models is also perceived critically: Markus et al. found that the evidence of usefulness of transparent or explainable models in terms of trustworthiness is still lacking, and it should rather be assessed by performing extensive external validation, data quality control and regulation [167]. Similarly, Ghassemi et al. argued 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. [168]. Interestingly, too much transparency might even lead to averse effects: Dietvorst et al. showed that people tend to distrust an algorithm more quickly than a human forecaster after seeing them do the same mistake [169]. It is to note that the proposed risk groups in PCAI are also categorized as a means of interpretability. However, they aim at making the model's predicted continuous risk score more interpretable in a clinical setting, by allowing for an easier correlation with treatment options, rather than making the deep learning model itself interpretable or explainable.

Besides building trust in the models by making them explainable or interpretable, quantifying the uncertainty of a model's predictions is an additional approach suggested by the literature [18, 21]. While transparency-boosting methods like GradCAM saliency maps require the human reader to verify the legitimacy of a given

output, confidence estimation aims to equip the model with the necassary means to self-sufficiently quantify the validity of its predictions. To this end, the credibility estimation setup is introduced into PCAI, which is inspired by the concept of conformal prediction [159]. It is shown that the credibility score successfully guides PCAI towards problematic samples that benefit from color adaptation. However, discarding samples that PCAI assigned a low credibility score to does not improve overall performance. It is to mention that certainty quantification and out-of-distribution detection is a multi-layered and complex problem, and the proposed solution is just a first attempt that aims to detect samples that deviate from the learned distribution in the latent space [170]. Future work should extend on this and find a more suited non-conformity metric than the proposed latent Mahalanobis distance, potentially by also including predicted class probabilities as done by Olsson et al. [100]. Ideally, future versions of DeePSC should also encompass a method to quantify its predictive confidence.

Finally, the models need be transferred from research to actual practical application and evaluated in a hands-on setting where the physician is paired with the AI to see how the CDS affects decision making. As Meyer et al. found, this can reveal surprising behaviour of clinicians and expectations towards model features that differ from what was initially assessed by evaluating a questionnaire or when building on beliefs from the literature [171].

For such an implementation study in clinical practice, the CDS might need to meet further requirements not accounted for in this thesis, like software maturity, accessability in terms of human-computer interaction or getting official certification [172, 173]. Bozyel et al. emphasize in a recent study the importance of data quality, representativeness and up-to-dateness of clinical CDS systems [174]. A potential way for future work to incorporate this into a running CDS system based on models as proposed in this thesis is by applying a continuous federated learning approach [175]. This aims to continuously keep the models knowledge up to date, while in parallel boosting robustness by utilizing highly heterogeneous data from multiple medical centers and protocols.

In summary, this thesis proposes two models for disease diagnosis and risk estimation, that both robustly outperform multiple human experts in their respective tasks. With the proposed adaptations to increase trustworthiness in the AI and given the outlook on remaining challenges and next steps, it is hoped that this thesis provides

a potential blueprint for future work that ultimately finds its way into clinical practice to aid clinicians as well as patients into more precise diagnosis, better treatment decisions and improved overall outcomes.

List of Publications

Publications included in this thesis:

- **DeePSC**: Ragab*, H., **Westhaeusser***, F., Ernst, A., Yamamura, J., Fuhlert, P., Zimmermann, M., Sauerbeck, J., Shenas, F., Özden, C., Weidmann, A., Adam, G., Bonn, S. & Schramm, C. DeePSC: A Deep Learning Model for Automated Diagnosis of Primary Sclerosing Cholangitis at Two-dimensional MR Cholangiopancreatography. *Radiology: Artificial Intelligence* **5** (2023)
- **PCAI**: **Westhaeusser***, F., Fuhlert*, P., Dietrich, E., Lennartz, M., Khatri, R., Kaiser, N., Röbeck, P., Bülow, R., von Stillfried, S., Ladjevardi, S., Drotte, A., Severgardh, P., G. Puelles, V., Häggman, M., Brehler, M., Boor, P., Walhagen, P., Dragomir, A., Busch, C., Graefen, M., Bengtsson, E., Sauter, G., Zimmermann, M. & Bonn, S. Robust, credible, and interpretable AI-based Prostate cancer digital pathology with above expert detection and grading performance. *tba* (tba)

Publications generated in the period of the PhD but not included in this thesis:

- Monedero*, S. M., **Westhaeusser***, F., Yaghoubi, E., Frintrop, S. & Zimmermann, M. *RADR: A Robust Domain-Adversarial-based Framework for Automated Diabetic Retinopathy Severity Classification* in *Medical Imaging with Deep Learning* (2024). `https://openreview.net/forum?id=2Q3mTp6a6T`
- Fuhlert, P., Ernst, A., Dietrich, E., **Westhaeusser**, F., Kloiber, K. & Bonn, S. *Deep learning-based discrete calibrated survival prediction* in *IEEE International Conference on Digital Health (ICDH)* (2022), 169–174

* Equal contribution

Bibliography

1. Peng, S., Kalliamvakou, E., Cihon, P. & Demirer, M. The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590* (2023).
2. Gao, X. & Feng, H. AI-Driven Productivity Gains: Artificial Intelligence and Firm Productivity. *Sustainability* **15** (2023).
3. Kulkarni, S., Seneviratne, N., Baig, M. S. & Khan, A. H. A. Artificial intelligence in medicine: where are we now? *Academic radiology* **27,** 62–70 (2020).
4. Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., *et al.* A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and Meta-Analysis. *The Lancet Digital Health* **1,** e271–e297 (2019).
5. Zenouzi, R., Liwinski, T., Yamamura, J., Weiler-Normann, C., Sebode, M., *et al.* Follow-up Magnetic Resonance Imaging/3D-magnetic Resonance Cholangiopancreatography in Patients with Primary Sclerosing Cholangitis: Challenging for Experts to Interpret. *Alimentary Pharmacology & Therapeutics* **48,** 169–178 (2018).
6. Grigoriadis, A., Morsbach, F., Voulgarakis, N., Said, K., Bergquist, A., *et al.* Inter-Reader Agreement of Interpretation of Radiological Course of Bile Duct Changes between Serial Follow-up Magnetic Resonance Imaging/3D Magnetic Resonance Cholangiopancreatography of Patients with Primary Sclerosing Cholangitis. *Scandinavian Journal of Gastroenterology* **55,** 228–235 (2020).
7. Johnson, K. B., Wei, W.-Q., Weeraratne, D., Frisse, M. E., Misulis, K., *et al.* Precision medicine, AI, and the future of personalized health care. *Clinical and translational science* **14,** 86–93 (2021).

8. Chakravarty, K., Antontsev, V., Bundey, Y. & Varshney, J. Driving success in personalized medicine through AI-enabled computational modeling. *Drug Discovery Today* **26,** 1459–1465 (2021).
9. Khosravi, M., Zare, Z., Mojtabaeian, S. M. & Izadi, R. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews. *Health Services Research and Managerial Epidemiology* **11** (2024).
10. Akinrinmade, A. O., Adebile, T. M., Ezuma-Ebong, C., Bolaji, K., Ajufo, A., *et al.* Artificial Intelligence in Healthcare: Perception and Reality. *Cureus* **15** (2023).
11. Zhang, P. & Kamel Boulos, M. N. Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet* **15** (2023).
12. Brehler, M., Walhagen, P., Busch, C., Bonn, S. & Bengtsson, E. Difficulties and Recommendations for AI-Based Prediction of Prostate Cancer Aggressiveness in Digital Pathology. *Medical Research Archives* **11** (2023).
13. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., *et al.* Deep learning-enabled medical computer vision. *NPJ digital medicine* **4** (2021).
14. Kundu, S. AI in medicine must be explainable. *Nature medicine* **27,** 1328–1328 (2021).
15. Au Yeung, J., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., *et al.* AI chatbots not yet ready for clinical use. *Frontiers in digital health* **5** (2023).
16. Scheetz, J., Rothschild, P., McGuinness, M., Hadoux, X., Soyer, H. P., *et al.* A survey of clinicians on the use of artificial intelligence in ophthalmology, dermatology, radiology and radiation oncology. *Scientific reports* **11** (2021).
17. Wilm, F., Marzahl, C., Breininger, K. & Aubreville, M. *Domain adversarial RetinaNet as a reference algorithm for the MItosis DOmain generalization challenge* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), 5–13.
18. Ghoshal, B., Tucker, A., Sanghera, B. & Lup Wong, W. Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection. *Computational Intelligence* **37,** 701–734 (2021).
19. Somashekhar, S., Sepúlveda, M.-J., Puglielli, S., Norden, A., Shortliffe, E. H., *et al.* Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology* **29,** 418–423 (2018).

20. Hoff, K. A. & Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* **57,** 407–434 (2015).

21. Dolezal, J. M., Srisuwananukorn, A., Karpeyev, D., Ramesh, S., Kochanny, S., *et al.* Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology. *Nature communications* **13** (2022).

22. Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., *et al.* The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems* **133,** 281–296 (2022).

23. Suara, S., Jha, A., Sinha, P. & Sekh, A. A. *Is grad-CAM explainable in medical images?* in *International Conference on Computer Vision and Image Processing* (2023), 124–135.

24. Pierce, R. L., Van Biesen, W., Van Cauwenberge, D., Decruyenaere, J. & Sterckx, S. Explainability in medicine in an era of AI-based clinical decision support systems. *Frontiers in genetics* **13** (2022).

25. Karlsen, T. H., Folseraas, T., Thorburn, D. & Vesterhus, M. Primary Sclerosing Cholangitis – a Comprehensive Review. *Journal of Hepatology* **67,** 1298–1323 (2017).

26. Lu, L. Guidelines for the Management of Cholestatic Liver Diseases (2021). *Journal of Clinical and Translational Hepatology* **10,** 757–769 (2022).

27. Hildebrand, T., Pannicke, N., Dechene, A., Gotthardt, D. N., Kirchner, G., *et al.* Biliary Strictures and Recurrence after Liver Transplantation for Primary Sclerosing Cholangitis: A Retrospective Multicenter Analysis. *Liver Transplantation* **22,** 42–52 (2016).

28. Schramm, C., Eaton, J., Ringe, K. I., Venkatesh, S. & Yamamura, J. Recommendations on the Use of Magnetic Resonance Imaging in PSC-A Position Statement from the International PSC Study Group. *Hepatology* **66,** 1675–1688 (2017).

29. Dietrich, E., Fuhlert, P., Ernst, A., Sauter, G., Lennartz, M., *et al. 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 *Machine Learning for Health* (2021), 38–53.

30. Seeland, M. & Mäder, P. Multi-View Classification with Convolutional Neural Networks. *PLOS ONE* **16** (Dec. 2021).

31. Liang, H., Manne, S., Shick, J., Lissoos, T. & Dolin, P. Incidence, prevalence, and natural history of primary sclerosing cholangitis in the United Kingdom. *Medicine* **96** (2017).

32. Ferri, P. M., Simões e Silva, A. C., Campos Silva, S. L., Aquino, D. J. Q. d., Fagundes, E. D. T., *et al.* The role of genetic and immune factors for the pathogenesis of primary sclerosing cholangitis in childhood. *Gastroenterology Research and Practice* **2016** (2016).

33. Singh, S. & Talwalkar, J. A. Primary sclerosing cholangitis: diagnosis, prognosis, and management. *Clinical Gastroenterology and Hepatology* **11,** 898–907 (2013).

34. Carbone, M., Kodra, Y., Rocchetti, A., Manno, V., Minelli, G., *et al.* Primary sclerosing cholangitis: burden of disease and mortality using data from the national rare diseases registry in Italy. *International Journal of Environmental Research and Public Health* **17** (2020).

35. Triantos, C., Koukias, N., Nikolopoulou, V. & Burroughs, A. Meta-analysis: ursodeoxycholic acid for primary sclerosing cholangitis. *Alimentary pharmacology & therapeutics* **34,** 901–910 (2011).

36. Khoshpouri, P., Habibabadi, R. R., Hazhirkarzar, B., Ameli, S., Ghadimi, M., *et al.* Imaging features of primary sclerosing cholangitis: from diagnosis to liver transplant follow-up. *Radiographics* **39,** 1938–1964 (2019).

37. Mudgal, P., Niknejad, M. & Chieng, R. *Magnetic resonance cholangiopancreatography (MRCP). Reference article, Radiopaedia.org* `https://doi.org/10.53347/rID-32557` (2024).

38. Wallnoefer, A., Herrmann, K., Beuers, U., Zech, C., Gourtsoyianni, S., *et al.* Comparison of 2D and 3D sequences for MRCP: Clinical value of the different techniques. *Der Radiologe* **45,** 993–1003 (2005).

39. Chien, C.-P., Chiu, F.-M., Shen, Y.-C., Chen, Y.-H. & Chung, H.-W. Magnetic resonance cholangiopancreatography at 3T in a single breath-hold: comparative effectiveness between three-dimensional (3D) gradient-and spin-echo and two-dimensional (2D) thick-slab fast spin-echo acquisitions. *Quantitative Imaging in Medicine and Surgery* **10** (2020).

40. Gaillard, F., Tho, D. & Luong, D. *1.5 T vs 3.0 T. Reference article, Radiopaedia.org* `https://doi.org/10.53347/rID-801` (2024).

41. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).

42. Lenc, K. & Vedaldi, A. *Understanding image representations by measuring their equivariance and equivalence* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), 991–999.

43. Dumoulin, V. & Visin, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).

44. Iandola, F. N. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).

45. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., *et al. ImageNet: A Large-Scale Hierarchical Image Database* in *IEEE Conference on Computer Vision and Pattern Recognition* (2009), 248–255.

46. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60,** 84–90 (2017).

47. Liang, Y., Li, S., Yan, C., Li, M. & Jiang, C. Explaining the black-box model: A survey of local interpretation methods for deep neural networks. *Neurocomputing* **419,** 168–182 (2021).

48. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* **128,** 336–359 (2020).

49. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. *Learning deep features for discriminative localization* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 2921–2929.

50. Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L. & Erickson, B. J. Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of digital imaging* **30,** 449–459 (2017).

51. Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., *et al.* Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Scientific Reports* **11,** 1–17 (2021).

52. Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., *et al.* The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging* **34,** 1993–2024 (2014).

53. Pellicer-Valero, O. J., Marenco Jimenez, J. L., Gonzalez-Perez, V., Casanova Ramon-Borja, J. L., Martin Garcia, I., *et al.* Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. *Scientific reports* **12** (2022).

54. Jaeger, P. F., Kohl, S. A., Bickelhaupt, S., Isensee, F., Kuder, T. A., *et al. Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection* in *Machine Learning for Health Workshop* (2020), 171–183.

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

56. Hossbach, J., Splitthoff, D. N., Cauley, S., Clifford, B., Polak, D., *et al.* Deep learning-based motion quantification from k-space for fast model-based magnetic resonance imaging motion correction. *Medical physics* **50,** 2148–2161 (2023).

57. Luo, B., Li, Z., Zhang, K., Wu, S., Chen, W., *et al.* Using deep learning models in magnetic resonance cholangiopancreatography images to diagnose common bile duct stones. *Scandinavian Journal of Gastroenterology* **59,** 118–124 (2024).

58. Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., Kwon, Y., *et al.* ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. *Zenodo* (2022).

59. Muneeswaran, V., Nagaraj, P. & Ijaz, M. F. *An Articulated Learning Method Based on Optimization Approach for Gallbladder Segmentation from MRCP Images and an Effective IoT Based Recommendation Framework* in *Connected e-Health: Integrated IoT and Cloud Computing* (Springer, 2022), 165–179.

60. Najjar, R. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics* **13** (2023).

61. Trépo, E., Goossens, N., Fujiwara, N., Song, W.-M., Colaprico, A., *et al.* Combination of gene expression signature and model for end-stage liver disease score predicts survival of patients with severe alcoholic hepatitis. *Gastroenterology* **154,** 965–975 (2018).

62. Konerman, M. A., Zhang, Y., Zhu, J., Higgins, P. D., Lok, A. S., *et al.* Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology* **61,** 1832–1841 (2015).

63. Eaton, J. E., Vesterhus, M., McCauley, B. M., Atkinson, E. J., Schlicht, E. M., *et al.* Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. *Hepatology* **71,** 214–224 (2020).

64. Andres, A., Montano-Loza, A., Greiner, R., Uhlich, M., Jin, P., *et al.* A novel learning algorithm to predict individual survival after liver transplantation for primary sclerosing cholangitis. *PloS one* **13** (2018).

65. Hu, C., Iyer, R. K., Juran, B. D., McCauley, B. M., Atkinson, E. J., *et al.* Predicting cholangiocarcinoma in primary sclerosing cholangitis: using artificial intelligence, clinical and laboratory data. *BMC gastroenterology* **23** (2023).

66. Ringe, K. I., Vo Chieu, V. D., Wacker, F., Lenzen, H., Manns, M. P., *et al.* Fully Automated Detection of Primary Sclerosing Cholangitis (PSC)-Compatible Bile Duct Changes Based on 3D Magnetic Resonance Cholangiopancreatography Using Machine Learning. *European Radiology* **31,** 2482–2489 (2021).

67. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. *Inception-v4, inception-resnet and the impact of residual connections on learning* in *Proceedings of the AAAI conference on artificial intelligence* **31** (2017).

68. Su, H., Maji, S., Kalogerakis, E. & Learned-Miller, E. *Multi-view convolutional neural networks for 3d shape recognition* in *Proceedings of the IEEE international conference on computer vision* (2015), 945–953.

69. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., *et al. 3d shapenets: A deep representation for volumetric shapes* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), 1912–1920.

70. Ma, C., Guo, Y., Yang, J. & An, W. Learning multi-view representation with LSTM for 3-D shape recognition and retrieval. *IEEE Transactions on Multimedia* **21,** 1169–1182 (2018).

71. Liang, Q., Wang, Y., Nie, W. & Li, Q. MVCLN: multi-view convolutional LSTM network for cross-media 3D shape recognition. *IEEE Access* **8** (2020).

72. Feng, Y., Zhang, Z., Zhao, X., Ji, R. & Gao, Y. *Gvcnn: Group-view convolutional neural networks for 3d shape recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 264–272.

73. Xu, J., Zhang, X., Li, W., Liu, X. & Han, J. *Joint multi-view 2D convolutional neural networks for 3D object classification* in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (2021), 3202–3208.

74. Geras, K. J., Wolfson, S., Shen, Y., Wu, N., Kim, S., *et al.* High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047* (2017).

75. Kaiser, N., Fieselmann, A., Vesal, S., Ravikumar, N., Ritschl, L., *et al. Mammographic breast density classification using a deep neural network: assessment based on inter-observer variability* in *Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment* **10952** (2019), 156–161.

76. Zuiderveld, K. *Contrast limited adaptive histogram equalization* in *Graphics gems IV* (1994), 474–485.

77. Ragab*, H., **Westhaeusser***, F., Ernst, A., Yamamura, J., Fuhlert, P., *et al.* DeePSC: A Deep Learning Model for Automated Diagnosis of Primary Sclerosing Cholangitis at Two-dimensional MR Cholangiopancreatography. *Radiology: Artificial Intelligence* **5** (2023).

78. Ilse, M., Tomczak, J. & Welling, M. *Attention-based deep multiple instance learning* in *International conference on machine learning* (2018), 2127–2136.

79. Loshchilov, I. & Hutter, F. *Decoupled Weight Decay Regularization* in *International Conference on Learning Representations* (2019). `https://openreview.net/forum?id=Bkg6RiCqY7`.

80. The MONAI Consortium. Project MONAI. *Zenodo* (2020).

81. Billot, B., Greve, D. N., Van Leemput, K., Fischl, B., Iglesias, J. E., *et al. A Learning Strategy for Contrast-agnostic MRI Segmentation* in *Medical Imaging with Deep Learning* (2020), 75–93.

82. Dave, M., Elmunzer, B. J., Dwamena, B. A. & Higgins, P. D. R. Primary Sclerosing Cholangitis: Meta-Analysis of Diagnostic Performance of MR Cholangiopancreatography. *Radiology* **256,** 387–396 (2010).

83. Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y. & Park, S. H. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. *Korean journal of radiology* **20,** 405–410 (2019).

84. Do, T.-B., Nguyen, H.-H., Vu, H., Le, T.-L., *et al. Plant identification using score-based fusion of multi-organ images* in *2017 9th International conference on knowledge and systems engineering (KSE)* (2017), 191–196.

85. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9,** 1735–1780 (1997).

86. Rymarczyk, D., Borowa, A., Tabor, J. & Zieliński, B. Kernel self-attention in deep multiple instance learning. *arXiv preprint arXiv:2005.12991* (2020).

87. Ng, A., Laird, D. & He, L. Data-centric ai competition. *DeepLearning AI* (2021).

88. Gandaglia, G., Leni, R., Bray, F., Fleshner, N., Freedland, S. J., *et al.* Epidemiology and prevention of prostate cancer. *European urology oncology* **4,** 877–892 (2021).

89. Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., *et al.* 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. *The American journal of surgical pathology* **40,** 244–252 (2016).

90. Gleason, D. F. & Mellinger, G. T. Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. *The Journal of urology* **111,** 58–64 (1974).

91. Flach, R. N., Willemse, P.-P. M., Suelmann, B. B., Deckers, I. A., Jonges, T. N., *et al.* Significant inter-and intralaboratory variation in gleason grading of prostate cancer: a nationwide study of 35,258 patients in the Netherlands. *Cancers* **13** (2021).

92. Nagpal, K., Foote, D., Tan, F., Liu, Y., Chen, P.-H. C., *et al.* Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA oncology* **6,** 1372–1380 (2020).

93. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., *et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine* **25,** 1301–1309 (2019).

94. Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21,** 233–241 (2020).

95. Walhagen, P., Bengtsson, E., Lennartz, M., Sauter, G. & Busch, C. AI-based prostate analysis system trained without human supervision to predict patient outcome from tissue samples. *Journal of Pathology Informatics* **13** (2022).

96. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering* **6,** 1330–1345 (2022).

97. Stacke, K., Eilertsen, G., Unger, J. & Lundström, C. Measuring domain shift for deep learning in histopathology. *IEEE journal of biomedical and health informatics* **25,** 325–336 (2020).

98. Pantanowitz, L., Quiroga-Garza, G. M., Bien, L., Heled, R., Laifenfeld, D., *et al.* An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. *The Lancet Digital Health* **2,** e407–e416 (2020).

99. Sandeman, K., Blom, S., Koponen, V., Manninen, A., Juhila, J., *et al.* AI model for prostate biopsies predicts cancer survival. *Diagnostics* **12** (2022).

100. Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., *et al.* Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nature communications* **13** (2022).

101. Toivanen, R. & Shen, M. M. Prostate organogenesis: tissue induction, hormonal regulation and cell type specification. *Development* **144,** 1382–1398 (2017).

102. Leslie, S. W., Soon-Sutton, T. L., R I, A., Sajjad, H. & Skelton, W. P. *Prostate Cancer* in *StatPearls [Internet]* (StatPearls Publishing, 2024). `https://www.ncbi.nlm.nih.gov/books/NBK470550/` (2024).

103. Siegel, R. L., Miller, K. D., Wagle, N. S. & Jemal, A. Cancer statistics, 2023. *CA: a cancer journal for clinicians* **73,** 17–48 (2023).

104. Siegel, D. A. Prostate cancer incidence and survival, by stage and race/ethnicity — United States, 2001–2017. *MMWR. Morbidity and mortality weekly report* **69** (2020).

105. Wei, J. T., Barocas, D., Carlsson, S., Coakley, F., Eggener, S., *et al.* Early detection of prostate cancer: AUA/SUO guideline part I: prostate cancer screening. *The Journal of Urology* **210,** 46–53 (2023).

106. Luiting, H. & Roobol, M. in *Versorgungs-Report: Früherkennung* (eds Günster, C. *et al.*) (Medizinisch Wissenschaftliche Verlagsgesellschaft, 2019).

107. Das, C., Razik, A., Sharma, S. & Verma, S. Prostate biopsy: when and how to perform. *Clinical radiology* **74,** 853–864 (2019).

108. Mescher, A. L. in *Junqueira's Basic Histology: Text and Atlas, 17th Edition* (McGraw Hill, 2024).

109. Heidenreich, A. Guidelines and counselling for treatment options in the management of prostate cancer. *Prostate Cancer,* 131–162 (2007).

110. Rondorf-Klym, L. M. & Colling, J. Quality of life after radical prostatectomy. *Oncology Nursing Forum* **30,** E24–E32 (2003).

111. Eskaros, A. R., Egloff, S. A. A., Boyd, K. L., Richardson, J. E., Hyndman, M. E., *et al.* Larger core size has superior technical and analytical accuracy in bladder tissue microarray. *Laboratory Investigation* **97,** 335–342 (2017).

112. Lawson, P., Sholl, A. B., Brown, J. Q., Fasy, B. T. & Wenk, C. Persistent homology for the quantitative evaluation of architectural features in prostate cancer histology. *Scientific reports* **9** (2019).

113. Delahunt, B., Miller, R. J., Srigley, J. R., Evans, A. J. & Samaratunga, H. Gleason grading: past, present and future. *Histopathology* **60,** 75–86 (2012).

114. Chan, T. Y., Partin, A. W., Walsh, P. C. & Epstein, J. I. Prognostic significance of Gleason score 3+ 4 versus Gleason score 4+ 3 tumor at radical prostatectomy. *Urology* **56,** 823–827 (2000).

115. Sauter, G., Steurer, S., Clauditz, T. S., Krech, T., Wittmer, C., *et al.* Clinical utility of quantitative Gleason grading in prostate biopsies and prostatectomy specimens. *European urology* **69,** 592–598 (2016).

116. Sauter, G., Clauditz, T., Steurer, S., Wittmer, C., Büscheck, F., *et al.* Integrating tertiary Gleason 5 patterns into quantitative Gleason grading in prostate biopsies and prostatectomy specimens. *European urology* **73,** 674–683 (2018).

117. Farahani, N., Parwani, A. V. & Pantanowitz, L. Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. *Pathology and Laboratory Medicine International,* 23–33 (2015).

118. Wilm, F., Fragoso, M., Bertram, C. A., Stathonikos, N., Öttl, M., *et al. Mind the Gap: Scanner-induced domain shifts pose challenges for representation learning in histopathology* in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023), 1–5.

119. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association* **53,** 457–481 (1958).

120. Greenwood, M. The natural duration of cancer. *Reports on public health and medical subjects* **33,** 1–26 (1926).

121. Tan, M. & Le, Q. *Efficientnet: Rethinking model scaling for convolutional neural networks* in *International conference on machine learning* (2019), 6105–6114.

122. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. *Mobilenetv2: Inverted residuals and linear bottlenecks* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 4510–4520.

123. Hu, J., Shen, L. & Sun, G. *Squeeze-and-excitation networks* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), 7132–7141.

124. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3,** 861 (2018).

125. Ghojogh, B., Ghodsi, A., Karray, F. & Crowley, M. Uniform manifold approximation and projection (UMAP) and its variants: tutorial and survey. *arXiv preprint arXiv:2109.02508* (2021).

126. Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering* **5,** 555–570 (2021).

127. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications* **7** (2022).

128. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. *A simple framework for contrastive learning of visual representations* in *International conference on machine learning* (2020), 1597–1607.

129. Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F., Jaume, G., *et al.* A general-purpose self-supervised model for computational pathology. *arXiv preprint arXiv:2308.15474* (2023).

130. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

131. Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., *et al.* DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research.* `https://openreview.net/forum?id=a68SUt6zFt` (2024).

132. Kattan, M. W., Wheeler, T. M., Scardino, P. T., *et al.* Postoperative nomogram for disease recurrence after radical prostatectomy for prostate cancer. *Journal of clinical oncology* **17,** 1499–1507 (1999).

133. Stephenson, A. J., Scardino, P. T., Eastham, J. A., Bianco Jr, F. J., Dotan, Z. A., *et al.* Postoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **23** (2005).

134. Fuhlert, P., Ernst, A., Dietrich, E., **Westhaeusser**, F., Kloiber, K., *et al. Deep learning-based discrete calibrated survival prediction* in *IEEE International Conference on Digital Health (ICDH)* (2022), 169–174.

135. Wang, H., Xia, Z., Xu, Y., Sun, J. & Wu, J. The predictive value of machine learning and nomograms for lymph node metastasis of prostate cancer: a systematic review and meta-analysis. *Prostate Cancer and Prostatic Diseases* **26,** 602–613 (2023).

136. Arvaniti, E., Fricker, K. S., Moret, M., Rupp, N., Hermanns, T., *et al.* Automated Gleason grading of prostate cancer tissue microarrays via deep learning. *Scientific reports* **8** (2018).

137. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. *Rethinking the inception architecture for computer vision* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 2818–2826.

138. Pinckaers, H., van Ipenburg, J., Melamed, J., De Marzo, A., Platz, E. A., *et al.* Predicting biochemical recurrence of prostate cancer with artificial intelligence. *Communications Medicine* **2** (2022).

139. Aubreville, M., Stathonikos, N., Bertram, C. A., Klopfleisch, R., Ter Hoeve, N., *et al.* Mitosis domain generalization in histopathology images—the MIDOG challenge. *Medical Image Analysis* **84** (2023).

140. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., *et al. A method for normalizing histology slides for quantitative analysis* in *2009 IEEE international symposium on biomedical imaging: from nano to macro* (2009), 1107–1110.

141. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., *et al.* Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging* **35,** 1962–1971 (2016).

142. Dietrich, E. *Deep learning-based discrete-time survival prediction on prostate cancer histopathology images* PhD thesis (Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 2022).

143. Thebille, A.-K., Dietrich, E., Klaus, M., Gernhold, L., Lennartz, M., *et al. Deep learning-based bias transfer for overcoming laboratory differences of microscopic images* in *Medical Image Understanding and Analysis: 25th Annual Conference* (2021), 322–336.

144. Khan, A., Atzori, M., Otálora, S., Andrearczyk, V. & Müller, H. *Generalizing convolution neural networks on stain color heterogeneous data for computational pathology* in *Medical Imaging 2020: Digital Pathology* (2020), 173–186.

145. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., *et al.* Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical image analysis* **58** (2019).

146. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal loss for dense object detection* in *Proceedings of the IEEE international conference on computer vision* (2017), 2980–2988.

147. Lafarge, M. W., Pluim, J. P., Eppenhof, K. A. & Veta, M. Learning domain-invariant representations of histological images. *Frontiers in medicine* **6** (2019).

148. Marini, N., Otalora, S., Wodzinski, M., Tomassini, S., Dragoni, A. F., *et al.* Data-driven color augmentation for H&E stained images in computational pathology. *Journal of Pathology Informatics* **14** (2023).

149. Aubreville, M., Bertram, C. A., Donovan, T. A., Marzahl, C., Maier, A., *et al.* A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. *Scientific data* **7** (2020).

150. Melamed, J. *Prostate cancer biorepository network (PCBN)* 2019.

151. Bankhead, P., Loughrey, M. B., Fernández, J. A., Dombrowski, Y., McArt, D. G., *et al.* QuPath: Open source software for digital pathology image analysis. *Scientific Reports* **7** (1 2017).

152. Saemundsson, A., Xu, L.-D., Meisgen, F., Cao, R. & Ahlgren, G. Validation of the prognostic value of a three-gene signature and clinical parameters-based risk score in prostate cancer patients. *The Prostate* **83,** 1133–1140 (2023).

153. Walhagen, P., Pontus, R., Ewert, B., Christer, B. & Michael, H. *Spear Prostate Biopsy 2020 (SPROB20)* 2020. `https://datahub.aida.scilifelab.se/10.23698/aida/sprob20`.

154. Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature medicine* **28,** 154–163 (2022).

155. Otsu, N. *et al.* A threshold selection method from gray-level histograms. *Automatica* **11,** 23–27 (1975).

156. Hendrycks*, D., Mu*, N., Cubuk, E. D., Zoph, B., Gilmer, J., *et al. AugMix: A Simple Method to Improve Robustness and Uncertainty under Data Shift* in *International Conference on Learning Representations* (2020). `https://openreview.net/forum?id=S1gmrxHFvB`.

157. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., *et al.* Domain-adversarial training of neural networks. *Journal of machine learning research* **17,** 1–35 (2016).

158. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems* **31** (2018).

159. Shafer, G. & Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research* **9** (2008).

160. Pereira, T., Cardoso, S., Silva, D., de Mendonça, A., Guerreiro, M., *et al. Towards trustworthy predictions of conversion from mild cognitive impairment to dementia: a conformal prediction approach* in *11th International Conference on Practical Applications of Computational Biology & Bioinformatics* (2017), 155–163.

161. Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* (2021).

162. Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B. & Wei, L.-J. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine* **30,** 1105–1117 (2011).

163. Sun, Y., Ming, Y., Zhu, X. & Li, Y. *Out-of-distribution detection with deep nearest neighbors* in *International Conference on Machine Learning* (2022), 20827–20840.

164. Marini, N., Atzori, M., Otálora, S., Marchand-Maillet, S. & Müller, H. *H&E-adversarial network: a convolutional neural network to learn stain-invariant features through Hematoxylin & Eosin regression* in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), 601–610.

165. Vovk, V., Gammerman, A. & Shafer, G. *Algorithmic learning in a random world* (Springer, 2005).

166. Jones, C., Thornton, J. & Wyatt, J. C. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Medical law review* **31,** 501–520 (2023).

167. Markus, A. F., Kors, J. A. & Rijnbeek, P. R. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of biomedical informatics* **113** (2021).

168. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3,** e745–e750 (2021).

169. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144** (2015).

170. Yang, J., Zhou, K., Li, Y. & Liu, Z. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision,* 1–28 (2024).

171. Meyer, J., Khademi, A., Têtu, B., Han, W., Nippak, P., *et al.* Impact of artificial intelligence on pathologists' decisions: an experiment. *Journal of the American Medical Informatics Association* **29,** 1688–1695 (2022).

172. Heesen, J., Müller-Quade, J., Wrobel, S., *et al. Zertifizierung von KI-Systemen – Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme* Whitepaper from Plattform Lernende Systeme. München, 2020.

173. Homeyer, A., Lotz, J., Schwen, L. O., Weiss, N., Romberg, D., *et al.* Artificial intelligence in pathology: from prototype to product. *Journal of pathology informatics* **12** (2021).

174. Bozyel, S., Şimşek, E., Koçyiğit, D., Güler, A., Korkmaz, Y., *et al.* Artificial Intelligence-Based Clinical Decision Support Systems in Cardiovascular Diseases. *Anatolian Journal of Cardiology* **28** (2024).

175. Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., *et al.* Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal* (2023).

176. **Westhaeusser***, F., Fuhlert*, P., Dietrich, E., Lennartz, M., Khatri, R., *et al.* Robust, credible, and interpretable AI-based Prostate cancer digital pathology with above expert detection and grading performance. *tba* (tba).

177. Monedero*, S. M., **Westhaeusser***, F., Yaghoubi, E., Frintrop, S. & Zimmermann, M. *RADR: A Robust Domain-Adversarial-based Framework for Automated Diabetic Retinopathy Severity Classification* in *Medical Imaging with Deep Learning* (2024). https://openreview.net/forum?id=2Q3mTp6a6T.

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence. 3, 8, 17, 26, 48, 53, 65, 116–118, 141, 143

all-CA Unguided Color Adaptation of All Samples. 119, 120

AUROC Area Under the Receiver Operating Characteristic Curve. 17, 54, 67, 68, 100, 104, 112–116, 118–123, 131, 133, 135

AVF Attention-based View Fusion. 27, 29, 36, 40, 42, 43

BASE Baseline Risk Prediction Network. 53, 54, 91–97, 100, 101, 104, 105, 107, 108, 110, 113, 116–121, 127, 129–134, 136, 138

BCR Biochemical Recurrence. 56, 61, 68, 74, 76–78, 81, 82, 85

C-Index Concordance Index. 54, 68, 114, 115, 122, 131, 133

CA Color Adaptation. 95, 96, 101, 108–110, 119, 134

CDS Clinical Decision Support. 3, 4, 141–143

CE Credibility Estimation. 95, 96, 100, 104, 106, 107, 110, 116, 119, 132, 134

CE-CA Credibility-guided Color Adaptation. 110, 116, 119, 120

CI Cancer Indicator. 83, 89, 95, 96, 101, 111–113, 116, 125, 126, 129

CLAHE Contrast-limited Adaptive Histogram Equalization. 19, 24, 25, 37, 38

CNN Convolutional Neural Network. 8, 10, 12–14, 16, 18, 26, 28, 37, 55, 62, 65, 67, 96, 97, 112

CP Conformal Prediction. 106, 132, 134

DA Domain Adversarial. 95, 96, 100–102, 104, 105, 116, 128

DD Domain Discriminator. 96, 101–103

DeePSC Deep learning classification network for primary sclerosing cholangitis. 4, 6, 9, 26, 27, 30, 33–43, 46, 49, 141–143

DL Deep Learning. 39, 42, 52, 65, 133

DRE Digital Rectal Exam. 55

ERCP Endoscopic Retrograde Cholangiopancreatography. 19

FE Feature Extractor. 27–29, 36, 43–45, 96, 97, 102

FU Follow-Up. 73–76, 78, 81, 82

GIQ Integrated Quantative Gleason. 58, 80, 92, 120–123, 134, 138

GPU Graphics Processing Unit. 31

GradCAM Gradient-weighted Class Activation Mapping. 10, 13, 14, 36, 37, 41, 49, 142

GRL Gradient Reversal Layer. 96, 101–103

H&E Hematoxilin and Eosin. 56, 57, 69–71, 73, 83, 136

HCE Highest Confidence Ensemble. 26, 27, 30, 33, 35, 38

HSL Hue-Saturation-Lightness Colorspace. 87

HSV Hue-Saturation-Value Colorspace. 83, 84, 89, 107, 108, 130

IBD Inflammatory Bowel Disease. 8, 10

ID In Distribution. 130

ISUP International Society of Urological Pathology. 52–54, 58, 68, 73, 76–79, 81, 82, 86, 92, 94, 100, 113, 115, 120–123, 134, 136, 138

JHU Johns-Hopkins University TMA-spot Dataset. 69, 78–80, 83, 92, 93, 118, 119, 124, 125, 127, 128, 133

LSTM Long Short-Term Memory. 43, 44, 46–48

META Metastasis. 74, 76–78, 81, 82, 85

MIL Multiple Instance Learning. 29, 65–68, 71, 97

ML Machine Learning. 17

MMX Malmö Biopsy Dataset. 80, 81, 83, 88–93, 109, 111, 112, 121–129, 134

MRCP Magnetic Resonance Cholangiopancreatography. 8–11, 16–29, 33–36, 38–42, 45, 47–49, 141, 142

MRI Magnetic Resonance Imaging. 4, 8, 9, 11, 16, 17, 19–22, 31, 35, 39, 48, 49, 56, 82

MVCNN Multi-view Convolutional Neural Network. 18, 26, 27, 29–31, 33–35, 37–39, 42–46, 48

NYU New York University TMA-spot Dataset. 69, 78, 79, 92, 93, 118, 119, 124, 125, 128, 133

OOD Out of Distribution. 130–132, 135

PA Patch Aggregation. 96–99, 102, 104, 106, 135

PANDA Prostate cANcer graDe Assessment Biopsy Dataset. 67, 82, 83, 89, 111, 112

PCa Prostate Cancer. 52, 53, 55–57, 61, 67–69, 73–76, 80, 82, 84, 85, 91, 97, 117, 118, 133, 134, 136

PCAD PCa-related Death. 74, 76–78, 81, 82, 85

PCAI Prostate Cancer Aggressiveness Index (Risk Prediction Network). 5, 6, 53, 54, 68, 81, 82, 91–97, 101, 107, 108, 110–113, 116–136, 138, 141–143

PCBN Prostate Cancer Biorepository Network. 78

PIRADS Prostate Imaging Reporting and Data System. 56

PSA Prostate-specific Antigen. 55, 56, 67, 75, 79, 81

PSC Primary Sclerosing Cholangitis. 8, 10, 11, 16, 17, 19, 21–24, 26, 29–31, 33, 34, 36, 37, 39–42, 46, 49, 141

RC Risk Classifier. 96, 97, 99, 102, 103

ReLU Rectified Linear Unit. 15, 99, 102

RGB Red-Blue-Green Colorspace. 28, 89, 97, 111

RP Radical Prostatectomy. 56, 57, 68, 73–76, 78, 81, 85, 118, 133, 138

SA Self Attention. 96–98, 102

SVCNN Single-view Convolutional Neural Network. 26–31, 33–35, 39, 42, 45–48

TMA Tissue Microarray. 53, 54, 56–59, 67, 68, 70, 73, 74, 76–81, 85, 87, 88, 90–92, 100, 113, 116, 118, 119, 121–125, 133, 134, 138, 141

TRUS Transrectal Ultrasound-guided Biopsy. 56

UKE University Hospital Hamburg-Eppendorf. 19, 74, 75, 77, 78, 80, 121

UKE.first First Batch Subset of the UKE-high-variance TMA-spot Dataset. 68, 76, 77, 83, 88, 91–93, 95, 96, 100, 104, 105, 107, 116–118, 120, 124, 125, 127, 128

UKE.long Long Stained Subset of the UKE-high-variance TMA-spot Dataset. 76–78, 83, 92, 104, 116, 117, 119, 127–129

UKE.scanner Different Scanner Subset of the UKE-high-variance TMA-spot Dataset. 76, 77, 83, 91–93, 95, 96, 104, 105, 107, 116, 117, 120, 128, 129

UKE.sealed TMA Spot Dataset with Follow-Up Ground Truth only available to the Pathology Department of the UKE. 73, 80, 83, 92, 93, 121, 122, 127, 128, 134

UKE.second Second Batch Subset of the UKE-high-variance TMA-spot Dataset. 76, 77, 83, 91–93, 95, 96, 104, 105, 107, 116, 127–129

UKE.thick Thick Cut Subset of the UKE-high-variance TMA-spot Dataset. 76–79, 83, 92, 104, 116, 117, 119, 128

UKE.thin Thin Cut Subset of the UKE-high-variance TMA-spot Dataset. 76–79, 83, 92, 104, 116, 117, 119, 127–129, 133

UKEhv UKE-high-variance TMA-spot Dataset. 74–79, 83, 91–94, 96, 100, 101, 104, 107, 116–118, 136

UKEhv$_6$ Subset of patients present in all six sub-datasets of the UKEhv data. 117, 118, 133, 138

UMAP Uniform Manifold Approximation and Projection. 55, 62, 64, 83, 84, 107, 108, 127, 128, 133

UPP Uppsala Biopsy Dataset. 81, 82, 92, 93, 112, 121–126, 128, 134

WSI Whole-Slide Image. 59, 65–67, 71, 82, 83, 95, 97, 142

A Appendix

A.1 DeePSC

A.1.1 Image Acquisition Protocol

MRCP imaging was obtained in routine fashion with rapid acquisition with relaxation enhancement sequences (RARE) using 2D single-shot (SS) thick-slab and 2D thin-slice (multi-slice) techniques. First, axial T1-weighted dual gradient-echo in-phase and opposed-phase images and breath-hold T2-weighted sequences in axial and coronal plane covering the liver were acquired. For thick-slab technique, a fat-suppressed single-shot (SS) turbo spin echo (TSE) sequence was obtained in high slice thickness (40-80 mm). Multi-slice MRCP was performed with an axial breath-hold T2-weighted Half-Fourier acquisition single-shot turbo spin echo (HASTE) sequence and postprocessed with isometric maximum intensity projections (MIP). Both sequences were obtained in coronal planes and angulated along the hepatobiliary ducts to allow for optimal visualization of all relevant anatomical structures. Finally, a total of 7-14 radial MRCP rotations from different angular points of view were reconstructed and reviewed for each exam. Imaging parameter were TR 6000 - 8000 ms, TE 740 - 920 ms, FA 90°, and FOV 300 x 300 mm for the internal dataset and TR 4000 ms, TE 696 ms, FA 178°, and FOV 300 x 300 mm for the different vendor validation dataset, respectively. There were no fundamental technical differences between the internal and external dataset beyond the known manufacturer specifications.

TR: Time of repetition, TE: time of echo, FA: flip angle, FOV: field of view.

A.1.2 Demographic Characteristics and Metadata

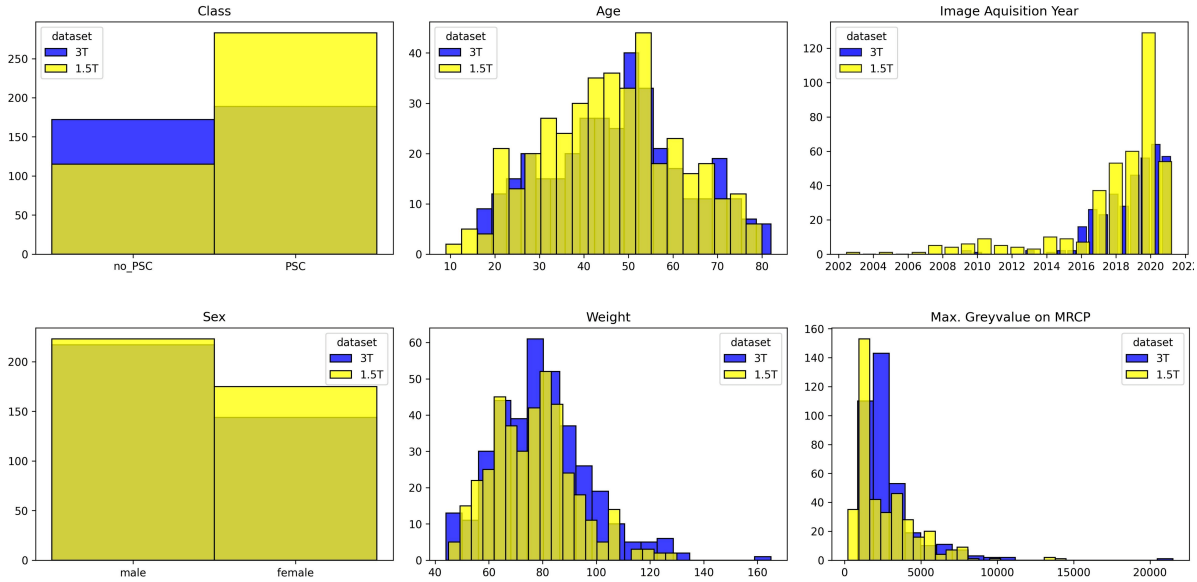

Figure A.1.1: Histograms of demographic and meta-information in the internal 3T (blue) and 1.5T (yellow) dataset. Reprinted with permission from Ragab and Westhaeusser et al. [77].

A.1.3 Experiments

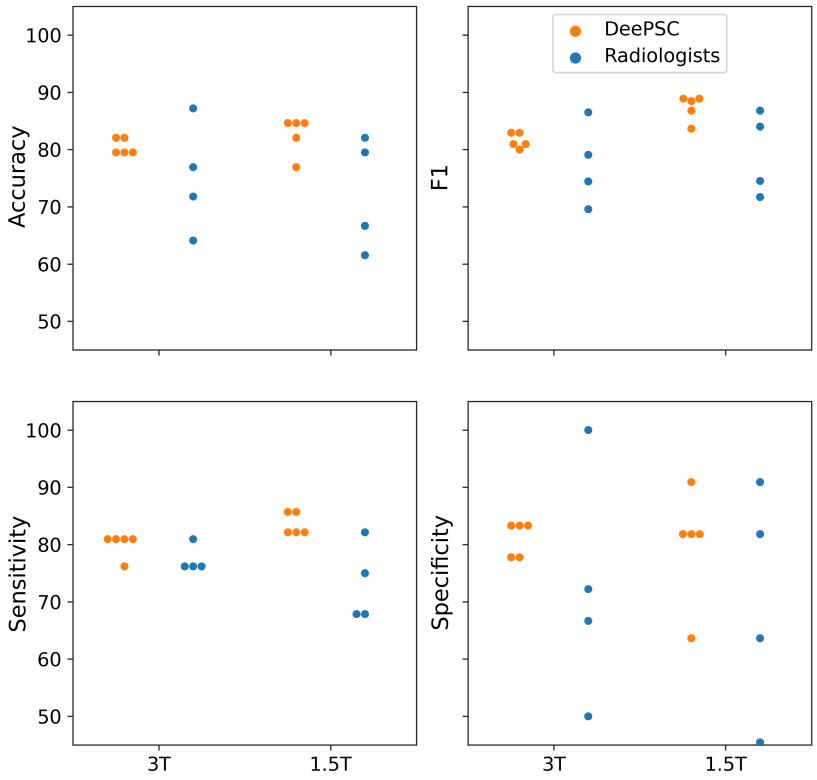

Figure A.1.2: Results of DeePSC (five ensemble models, orange) and the radiologists (four readers, blue) on the internal 3T and 1.5T test-sets. Reprinted with permission from Ragab and Westhaeusser et al. [77].

A.1.4 Extended Analysis

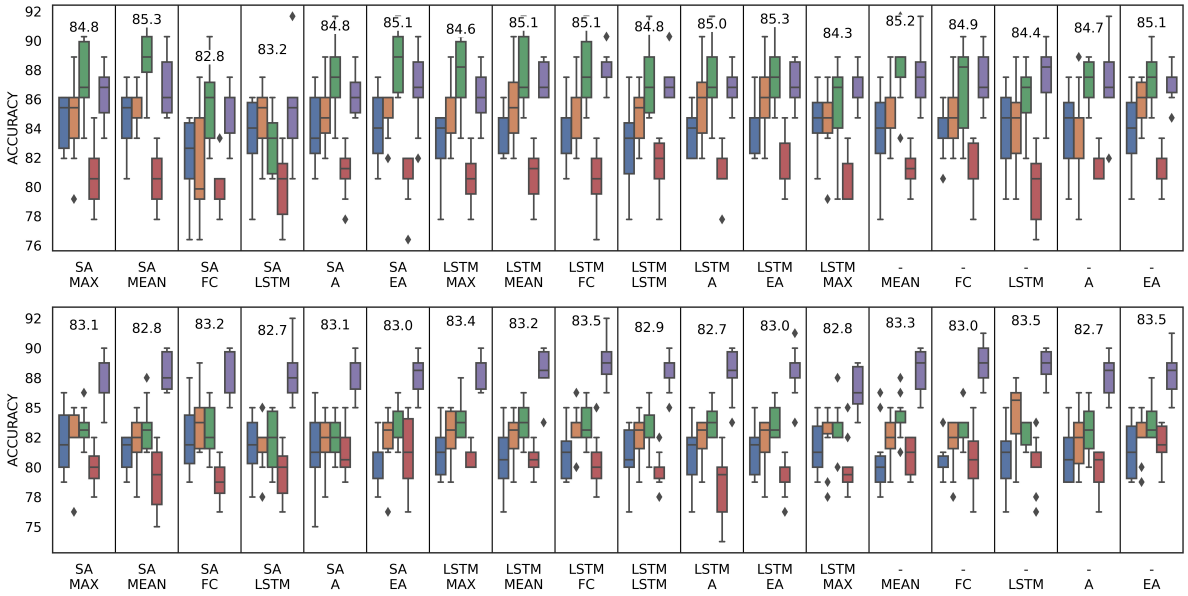

Figure A.1.3: Results of five training/test splits per architecture on the 3T (top) and 1.5T (bottom) dataset. Same colors per dataset refer to the same datasplit. Boxes per architecture include values of five training rounds with different random seeds. The mean over splits per architecture is depicted above the respective boxes. Notably, the main influence on performance is from the underlying datasplit, whereas choice of architecture does not show a consistent pattern.

A.2 PCAI

A.2.1 Metadata

Table A.2.1: Patient metadata for all PCa survival datasets.

		UKE	NYU	JHU	UPP	MMX
patients		8157	158	879	123	269
image type		TMA	TMA	TMA	Biopsy	Biopsy
age [years], mean ± SD		63.5 ± 6.1	60.9 ± 7	59.2 ± 6.3		67.6 ± 8.9
censoring [%]		61.4	70.3	0.3	83.7	88.5
median survival [years]		1.6	3.9	2	2.1	4.3
median followup [years]		8	17.8	16	7	9.1
ISUP	0	410 (5.03%)				
	1	1806 (22.14%)	49 (31.01%)	133 (15.13%)	9 (7.32%)	15 (5.58%)
	2	4016 (49.23%)	67 (42.41%)	337 (38.34%)	66 (53.66%)	82 (30.48%)
	3	1367 (16.76%)	16 (10.13%)	184 (20.93%)	27 (21.95%)	80 (29.74%)
	4	109 (1.34%)	11 (6.96%)	123 (13.99%)	12 (7.32%)	36 (13.38%)
	5	449 (5.50%)	15 (9.49%)	102 (11.60%)	9 (7.89%)	56 (20.82%)
event type	BCR	3089 (37.87%)	43 (27.22%)	521 (59.27%)	18 (14.63%)	
	FU	5007 (61.38%)	111 (70.25%)	3 (0.34%)	103 (83.74%)	226 (84.01%)
	META	61 (0.75%)		142 (16.15%)	2 (1.63%)	42 (15.61%)
	PCAD		4 (2.53%)			1 (0.37%)
	TRT			213 (24.23%)		
T-stage	≤ T1	2 (0.02%)			97 (78.86%)	122 (45.35%)
	T2	4966 (60.88%)	104 (65.82%)	134 (15.42%)	26 (21.14%)	90 (33.46%)
	T3	3128 (38.35%)	52 (32.91%)	735 (84.58%)		54 (20.07%)
	T4	61 (0.75%)	2 (1.27%)			3 (1.12%)
N-stage	N0	4306 (86.41%)	56 (35.44%)	700 (80.18%)		
	N1	677 (13.59%)	1 (0.63%)	163 (18.67%)		
	N2			2 (0.23%)		
	NX		101 (63.92%)	8 (0.92%)		
M-stage	M0	6335 (78.47%)		509 (60.89%)	7 (5.69%)	79 (29.48%)
	M1	1738 (21.53%)		327 (39.11%)	7 (5.69%)	
	MX				109 (88.62%)	189 (70.52%)

Table A.2.2: Patient metadata for the UKEhv survival sub-datasets.

		UKE.first	UKE.second	UKE.scanner	UKE.thin	UKE.thick	UKE.long
patients (images)		8123	7156	8114	1602	1574	1667
age [years], mean ± SD		63.5 ± 6.1	63.6 ± 6.1	63.5 ± 6.1	63.2 ± 6	63.2 ± 5.9	63.2 ± 6
censoring [%]		61.3	60.9	61.3	67.4	67.6	67.4
median survival [years]		1.6	1.6	1.6	2.4	2.4	2.4
median followup [years]		8	8	8	7.2	7.2	7.7
ISUP	0	407 (5.01%)	370 (5.17%)	405 (4.99%)	98 (6.12%)	96 (6.10%)	103 (6.18%)
	1	1792 (22.06%)	1557 (21.76%)	1789 (22.05%)	305 (19.04%)	304 (19.31%)	322 (19.32%)
	2	4001 (49.26%)	3512 (49.08%)	3997 (49.26%)	879 (54.87%)	864 (54.89%)	911 (54.65%)
	3	1366 (16.82%)	1223 (17.09%)	1366 (16.84%)	253 (15.79%)	243 (15.44%)	262 (15.72%)
	4	109 (1.34%)	94 (1.31%)	109 (1.34%)	19 (1.19%)	19 (1.21%)	21 (1.26%)
	5	448 (5.52%)	400 (5.59%)	448 (5.52%)	48 (3.00%)	48 (3.05%)	48 (2.88%)
event type	BCR	3084 (37.97%)	2745 (38.36%)	3081 (37.97%)	518 (32.33%)	506 (32.15%)	539 (32.33%)
	FU	4978 (61.28%)	4355 (60.86%)	4972 (61.28%)	1080 (67.42%)	1064 (67.60%)	1123 (67.37%)
	META	61 (0.75%)	56 (0.78%)	61 (0.75%)	4 (0.25%)	4 (0.25%)	5 (0.30%)
T-stage	≤ T1	2 (0.02%)	2 (0.03%)	2 (0.02%)			
	T2	4940 (60.81%)	4301 (60.10%)	4932 (60.78%)	976 (60.92%)	958 (60.86%)	1021 (61.25%)
	T3	3120 (38.41%)	2796 (39.07%)	3119 (38.44%)	610 (38.08%)	601 (38.18%)	628 (37.67%)
	T4	61 (0.75%)	57 (0.80%)	61 (0.75%)	16 (1.00%)	15 (0.95%)	18 (1.08%)
N-stage	N0	4290 (86.39%)	3719 (85.61%)	4284 (86.37%)	923 (90.22%)	907 (90.25%)	971 (90.49%)
	N1	676 (13.61%)	625 (14.39%)	676 (13.63%)	100 (9.78%)	98 (9.75%)	102 (9.51%)
M-stage	M0	6306 (78.44%)	5499 (77.67%)	6298 (78.43%)	1260 (78.95%)	1237 (78.89%)	1313 (79.05%)
	M1	1733 (21.56%)	1581 (22.33%)	1732 (21.57%)	336 (21.05%)	331 (21.11%)	348 (20.95%)

A.2.2 Extended Analysis

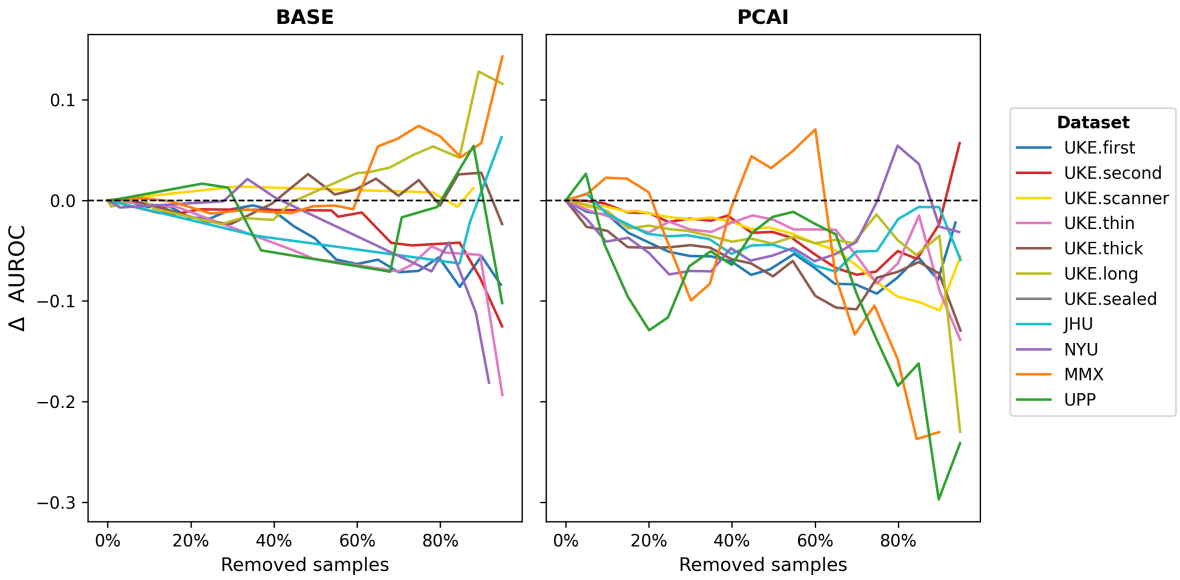

Figure A.2.1: Difference in five year relapse AUROC when removing increasing percentiles of images with lowest credibility scores from all datasets in BASE and PCAI. From left to right, remove no images up to removing 95%. Evaluation is performed in increments of 5 percentage points.