Aus dem Institut für Pathologie
der Medizinischen Fakultät Mannheim
(Direktor: Prof. Dr. med. Brochhausen-Delius)

Machine Learning zur Gewinnung histomorphologischer Biomarker bei Kolorektalem Karzinom: Ein Vergleich von zentralen und peripheren Tumorarealen

Inauguraldissertation
zur Erlangung des medizinischen Doktorengrades
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Daniel Christian Rusche
aus Nürnberg

2024

Dekan: Prof. Dr. med. Sergij Goerdt

Referent: Prof. Dr. med. Timo Gaiser

Für meine Eltern

Inhaltsverzeichnis

Abkürzungsverzeichnis **1**

1 Einleitung **2**

1.1 Relevanz und Herausforderung bei der histologischen Diagnostik . . 2

1.2 Das Potenzial von Machine Learning bei der histopathologischen Diagnostik von Kolorektale Karzinom (KRK) 3

1.3 Zielsetzung 4

2 Material und Methoden **6**

2.1 Segmentierung 8

2.1.1 Datenakquise und -vorbereitung des Segmentierungsdatensatzes 9

2.1.2 Vorbereitung des Datensatzes 10

2.1.3 Architektur und Training des U-Net 11

2.2 Klassifikation 14

2.2.1 Datenakquise und -vorbereitung 14

2.2.2 Erstellung der Datensätze 17

2.2.3 Architektur und Training des ResNets 20

2.2.4 Class activation maps 21

2.3 Statistische Evaluation 22

2.3.1 Metriken 22

2.3.2 Cohen's Kappa 23

2.3.3 Fläche unter der Receiver Operating Characteristics Kurve . 24

3 Ergebnisse **26**

3.1 Klassifizierung basierend auf Fall-unabhängiger Sortierung 28

3.2 Klassifizierung basierend auf nach Fällen sortierter Gruppierung . . 30

3.3 Zusammenhang zwischen Grading und Nodalstatus 32

4 Diskussion **35**

4.1 Fallstricke des reinen Auswendiglernens von Datensätzen 35

4.2 Der Datensatz und seine Anforderungen an ein Deep Learning Model 37

4.2.1 Noisy Label 37

4.2.2 Kohortengröße 38

4.2.3 Korrelation der prädikativen Parameter 39

4.3 Evaluation der untersuchten Klassifikation im Hinblick auf eine Integration in den klinischen Alltag 40

5 Zusammenfassung **42**

Abbildungsverzeichnis **44**

Tabellenverzeichnis **45**

Literaturverzeichnis **46**

Anhang **57**

Lebenslauf **59**

Danksagung **61**

Abkürzungsverzeichnis

2Bu	Klassizifierung bezüglich Buddingstatus
2No	Klassifizierung bezüglich Nodalstatus
2Pr	Klassifizierung bezüglich Progress
AUROC	Fläche unter der Receiver Operating Characteristics-Kurve
CAM	Class Activation Map
CNN	Convolutional Neural Network
DNN	Deep Neural Network
FFPE	Formalin-fixiert Paraffin-eingebettet
GAN	Generative Adversarial Network
HE	Hämatoxylin-Eosin
HIEN	hochgradige intraepitheliale Neoplasie
KRK	Kolorektale Karzinom
MIL	Multiple Instance Learning
ML	Machine Learning
NIEN	niedriggradige intraepitheliale Neoplasie
PT	periphere Tumorareale
PT2No	Klassifizierung der zufällig sortierten Kacheln peripherer Tumorareale bezüglich deren Nodalstatus
PT2NoGroup	Klassifzierung der nach Fällen gruppierten Kacheln peripherer Tumorareale bezüglich deren Nodalstatus
ResNet	Residual Neural Network
ROC	Receiver Operating Characteristics
UICC	Union for International Cancer Control
WSI	Whole Slide Image
ZT	zentrale Tumorareale
ZT2No	Klassifizierung der zufällig sortierten Kacheln zentraler Tumorareale bezüglich deren Nodalstatus
ZT2NoGroup	Klassifizierung der nach Fällen gruppierten Kacheln zentraler Tumorareale bezüglich deren Nodalstatus

1 Einleitung

Das Kolorektale Karzinom (KRK) stellt in Deutschland den zweithäufigsten malignen Tumor bei Frauen und den dritthäufigsten bei Männern dar. Die Inzidenz soll in den kommenden Jahren zusätzlich steigen [1, 2]. Für die Diagnosestellung wird typischerweise im Rahmen einer Koloskopie Gewebe entnommen und anschließend histopathologisch analysiert. In Zusammenhang mit den zunehmenden Inzidenzen für KRK steigt somit auch die Belastung der Patholog*innen [3]. Hinzu kommt, dass die genaue Beurteilung der einzelnen Präparate im Ergebnis starken Schwankungen unterliegt. Dies zeigt sich sowohl in der Inter- als auch der Intraobserver-Variabilität [4, 5]. Die Kombination aus steigender Arbeitsbelastung der Patholog*innen und der Variabilität der Begutachtung unterstreicht die Notwendigkeit einer robusten Analyse und Befundung von histologischen Präparaten in diesem Bereich.

1.1 Relevanz und Herausforderung bei der histologischen Diagnostik

Im Rahmen des Stagings hat die histopathologische Diagnostik einen großen Stellenwert im Hinblick auf die Therapie als auch die Einschätzung des weiteren Krankheitsverlaufs. Für die grundlegende Einteilung nach den Stadien entsprechend der Union for International Cancer Control (UICC) genügt die gängige Bestimmung von Infiltrationstiefe pT und Nodalstatus pN in Kombination mit der radiologi-

schen Beurteilung von Fernmetastasen M [6]. Im Hinblick auf die Wahl der adjuvanten Therapie sind jedoch zusätzliche Risikofaktoren ausschlaggebend [7]. Neben dem Grad der Differenzierung, lymphatischer oder vaskulärer Infiltration zählt hierzu neuerdings auch ein hochgradiges Budding [8]. Tumorbudding wird dabei definiert als einzelne Zellen oder Kluster bis vier Zellen an der Invasionsfront des Karzinoms. Die absolute Anzahl an Buds soll dabei innerhalb eines Hotspots bestimmt werden und im Rahmen eines drei Stufen Schemas notiert werden [9]. Für UICC Stadium II konnte das Budding mit diesem System schon als unabhängiger prognostischer Faktor bestimmt werden [10, 11]. Darüber hinaus konnte gezeigt werden, dass hochgradiges Tumorbudding bei KRK im Stadium pT1 mit einem erhöhten Risiko für Lymphknotenmetastasen assoziiert ist [12–16]. Durch diese Stellung als Risikofaktor beeinflusst das Tumorbudding sowohl Prognose als auch die individuelle Therapieempfehlung. Jedoch stellt die Bestimmung des Budding immer noch eine Herausforderung dar und ist längst nicht einheitlich. So wurde in den Studien, welche das Budding als Risikofaktor für Lymphknotenmetastasen darstellen, mit uneinheitlichen Methoden zur Bestimmung des Buddingstatus gearbeitet. Zusätzlich gilt es anzumerken, dass auch die aktuelle Konsensus Definition bei der Bestimmung des Hotspots diesen nicht eindeutig definiert [9]. Schlussendlich führend die lückenhaften und wechselnden Bestimmungsrichtlinien zu einer beachtlichen Variabilität bei der Bestimmung des Buddingstatus, welche seiner therapeutischen Tragweite kaum gerecht wird [17, 18].

1.2 Das Potenzial von Machine Learning bei der histopathologischen Diagnostik von KRK

Aus der Sicht von Machine Learning (ML) Modellen entspricht die Bestimmung des Budding- oder Nodalstatus einer Klassifikationsaufgabe. ML kann diese Klassifi-

kationen effektiv durchführen, wobei Deep Neural Networks (DNNs) insbesondere durch ihre Fähigkeit zur eigenständigen Erfassung komplexer Muster und ihrer Anpassungsfähigkeit an nicht-lineare Zusammenhänge einen Vorteil gegenüber anderen klassifizierenden ML Methoden bieten [19]. Dieser Vorteil kommt unter anderem dadurch zustande, dass die Modelle durch Programmierung und teilweise auch in der Hardwarearchitektur dem menschlichen Gehirn nachempfunden sind, indem Prozesse durch tausende Neuronen mit ebenso zahlreichen Verbindungen untereinander ausgeführt werden. Explizit zeigt sich das beispielsweise bei den Convolutional Neural Networks (CNNs) (deutsch: faltende Neuronale Netze) mit ihrer Fähigkeit, Objekte auf Fotos zu erkennen [20]. Erfolgt diese Klassifikation auf Pixelebene, sodass einem bestimmten Bereich eine Bedeutung zugeordnet wird, so spricht man von „semantischer Segmentierung“. Im Sinne einer reinen Klassifikation konnten auf der Basis von Whole Slide Images (WSIs) in Kombination mit klinischen Daten mittels eines CNN Lymphknotenmetastasen bei KRK vorhergesagt werden [21]. DNNs können demnach eine Hilfestellung bei der Diagnostik des KRK darstellen. Die morphologischen Eigenschaften, anhand derer ein DNN seine Entscheidungen trifft, können jedoch nicht definitiv bestimmt werden [19,20].

1.3 Zielsetzung

Das Budding kann eine solches durch DNN-erkennbares Pattern darstellen und ist definitionsgemäß spezifisch an der Invasionsfront lokalisiert [7]. Mittels semantischer Segmentierung sollte eine Unterscheidung der Invasionsfront vom restlichen Gewebe möglich sein. Anschließend kann durch eine selektierte Klassifikation dieser peripheren Tumorareale hinsichtlich prognostischer Marker ein erster Einblick in die Aussagekraft verschiedener Gewebeareale gewonnen werden. Um

im Vergleich sicherzustellen, dass maßgeblich Morphologien der Invasionsfront die Vorhersage beeinflussen, können weitere Klassifikationen auf Basis von zentralen Tumorarealen entwickelt werden. Aufgrund der Abwesenheit von Budding in zentralen Tumorarealen erwarten wir hier eine qualitativ schlechtere Vorhersage.

Zusammenfassend soll im Rahmen dieser Arbeit untersucht werden, ob der Nodalstatus auf Basis von WSI durch ein neuronales Netzwerk zuverlässig vorhergesagt werden kann. Zusätzlich wollen wir die für die Vorhersage ausschlaggebenden Areale eruieren und mit dem Auftreten von Budding korrelieren. Zudem soll untersucht werden, ob bisherige prognostische Marker sowie der Progress selbst durch ein klassifizierendes neuronales Netzwerk allein auf Basis von zentralen oder peripheren Tumorarealen vorhergesagt werden können.

2 Material und Methoden

Im Rahmen dieser Arbeit wurden nacheinander zwei maschinelle Lernmodelle entwickelt, welche Präparate von Kolorektalem Karzinom (KRK) untersuchen. Eine Übersicht über den gesamten Arbeitsablauf ist schematisch in Abbildung 2.1 dargestellt.

Ein erstes Modell wird mit Präparaten aus der Sammlung des Pathologischen Instituts der medizinischen Fakultät Mannheim trainiert, um diese auf Pixelebene in Tumorgewebe, Muskulatur und andere nicht-tumoröse Gewebearten aufzugliedern (siehe Kapitel 2.1). Das etablierte Modell wird dann auf ein Kollektiv der Universität Graz angewendet, um zentrale von peripheren Tumorarealen zu unterscheiden. Ein zweites Modell klassifiziert diese Gruppen anschließend jeweils bezüglich Budding-, Nodalstatus und Progress (siehe Kapitel 2.2).

Um die Entscheidung des Modells bedingt nachvollziehbar zu machen, wurde mittels Class Activation Maps (CAMs) versucht darzustellen, welche Bildausschnitte ausschlaggebend sind für die jeweilige Zuordnung. Abschließend wurden verschiedene Performance-Metriken berechnet (siehe Kapitel 2.3), um eine qualitative Analyse der verschiedenen Klassifikationen und deren Vergleich zu ermöglichen.

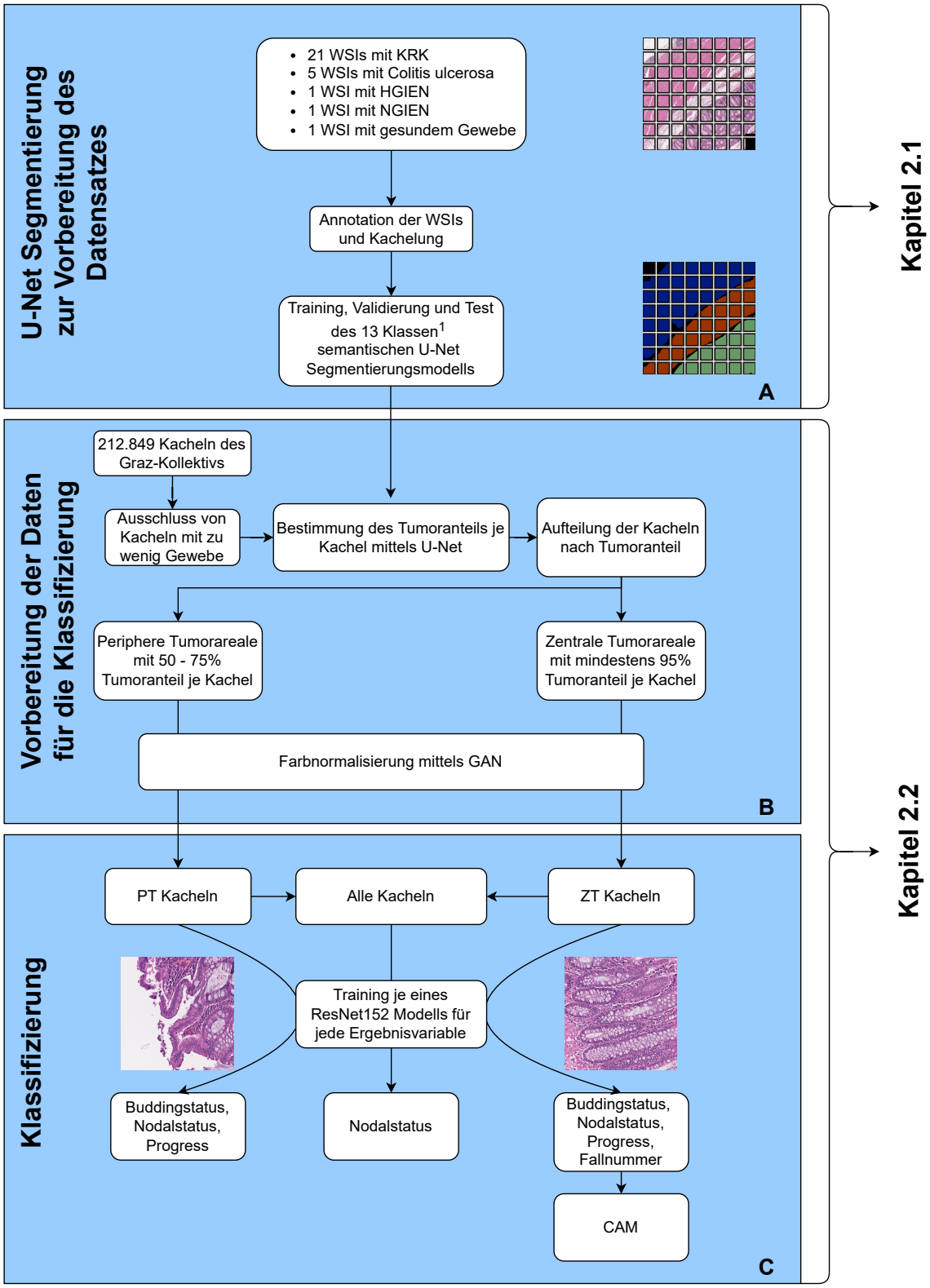
U-Net Segmentierung zur Vorbereitung des Datensatzes
21 WSIs mit KRK
5 WSIs mit Colitis ulcerosa
1 WSI mit HGIEN
1 WSI mit NGIEN
1 WSI mit gesundem Gewebe
Annotation der WSIs und Kachelung
Training, Validierung und Test des 13 Klassen[1] semantischen U-Net Segmentierungsmodells
A
Kapitel 2.1
Vorbereitung der Daten für die Klassifizierung
212.849 Kacheln des Graz-Kollektivs
Ausschluss von Kacheln mit zu wenig Gewebe
Bestimmung des Tumoranteils je Kachel mittels U-Net
Aufteilung der Kacheln nach Tumoranteil
Periphere Tumorareale mit 50 - 75% Tumoranteil je Kachel
Zentrale Tumorareale mit mindestens 95% Tumoranteil je Kachel
Farbnormalisierung mittels GAN
B
Klassifizierung
PT Kacheln
Alle Kacheln
ZT Kacheln
Training je eines ResNet152 Modells für jede Ergebnisvariable
Buddingstatus, Nodalstatus, Progress
Nodalstatus
Buddingstatus, Nodalstatus, Progress, Fallnummer
CAM
C
Kapitel 2.2

Abbildung 2.1 ***(vorherige Seite)*****: Übersicht über den Arbeitsablauf.**
A) Es wurden 29 Whole Slide Images (WSIs) von Präparaten mit kolorektalem Gewebe verwendet. Davon zeigten 21 kolorektales Karzinom (KRK), sieben andere Pathologien und eines gesundes Gewebe. Dieser Datensatz wurde für 13 Klassen[1] annotiert. Nach weiterer Vorbereitung wurden die Kacheln der WSIs verwendet, um ein U-Net für semantische Segmentierung zu trainieren, zu validieren und zu testen. B) Mithilfe des entwickelten U-Nets wurden die Kacheln des Graz-Kollektivs segmentiert, um den Tumoranteil je Kachel zu bestimmen. Anschließend wurde dieser Datensatz in zwei Gruppen für periphere Tumorareale (PT) und zentrale Tumorareale (ZT) aufgeteilt und die Kacheln hinsichtlich ihrer Färbung mittels eines generative adversarial Networks (GAN) normalisiert [22]. C) Ein zusätzlicher Datensatz „Alle Kacheln "wurde generiert, indem beide bereits erwähnten Gruppen PT und ZT zusammengelegt wurden. Klassifikatoren vom Typ eines residual neural Networks (ResNet152) wurden abschließend trainiert, um die Variablen Budding-, Nodalstatus und Progress für jede Gruppe vorherzusagen. Class Activation Maps (CAMs) wurden berechnet, um die Areale innerhalb der Kacheln zu identifizieren, welche den größten Einfluss auf die Entscheidung des Klassifikators hatten. HGIEN: hochgradige intraepitheliale Neoplaise, NGIEN: niedriggradige intraepitheliale Neoplasie
[1] für eine genaue Aufstellung der einzelnen Klassen, siehe Tabelle 2.1.

2.1 Segmentierung

Der erste Schritt dieser Arbeit bestand darin, ein maschinelles Lernmodell zur automatischen Segmentierung von histologischen Präparaten zu erstellen. Dieses Convolutional Neural Network (CNN) mit der Architektur eines U-Nets (siehe Kapitel 2.1.3) wurde in den späteren Schritten verwendet, um einen größeren Datensatz vor zu sortieren. Der Ablauf der Entwicklung des U-Nets kann in Abbildung 2.2 betrachtet werden.

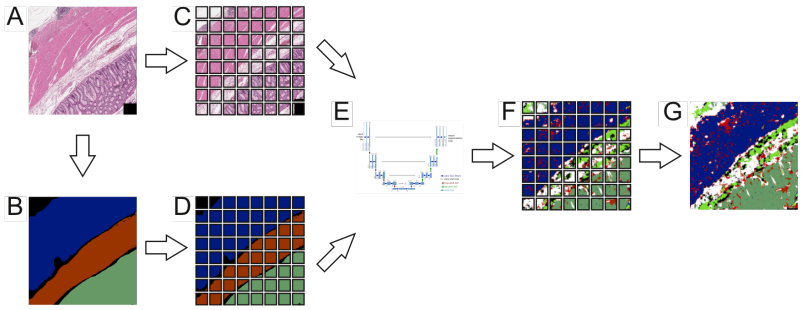

Abbildung 2.2: Ablauf der Segmentierung der HE gefärbten Gewebeschnitte. A) Die ursprünglichen HE Schnitte wurden entsprechend dem zugrundeliegenden Gewebe per Hand annotiert. B) In der entstandenen Annotationsmaske entspricht jede Gewebeklasse wie etwa Tumor oder Mukosa genaue einem Areal mit einer bestimmten Farbe. C,D) Der HE Scan und die Annotationsmaske wurden anschließend je in gleichgroße Kacheln aufgeteilt. E) Als Datensatz dienen diese Bildpaare als Trainingsgrundlage für das U-Net [23]. F) Nach erfolgreichem Training, Validierung und Test des Modells kann das neuronale Netzwerk eigenständig eine Annotationsmaske auf Basis von eingegebenen HE Kacheln produzieren. G) Die vom Netzwerk produzierten Kacheln einer Annotationsmaske können wiederum zusammengesetzt werden, um ein komplettes segmentiertes Bild des ursprünglichen HE Schnitts zu erzeugen.

2.1.1 Datenakquise und -vorbereitung des Segmentierungsdatensatzes

Formalin-fixierte Paraffin-eingebettete (FFPE) und mit Hämatoxylin-Eosin (HE) gefärbte Gewebeschnitte mit kolorektalem Gewebe wurden aus der Sammlung des Pathologischen Instituts der medizinischen Fakultät Mannheim der Universität Heidelberg entnommen und im Folgenden in vollständig anonymisierter Weise verwendet. Patient*inneninformationen wurden in den weiteren Verlauf nicht aufgenommen. Lediglich die histologisch gestellten Diagnosen physiologisches kolorektales Gewebe, KRK, Colitis ulcerosa, hochgradige intraepitheliale Neoplasie (HIEN), niedriggradige intraepitheliale Neoplasie (NIEN) wurden verwendet.

2.1.2 Vorbereitung des Datensatzes

Mit einem M8-Mikroskop und -Scanner (PreciPoint GmbH, Garching b. München, Germany) wurden die Präparate digitalisiert und die erhaltenen Whole Slide Images (WSIs) für die weitere Bearbeitung im .svs-Format abgespeichert. Mit Hilfe der Software QuPath [24] wurden diese digitalen Gewebeschnitte per Hand annotiert und in 13 verschiedene Klassen segmentiert.

Somit ergab sich ein Datensatz aus 29 WSIs, bestehend aus 21 WSIs mit KRK, je ein Schnitt mit HIEN und NIEN, fünf Präparaten mit Colitis ulcerose und einem Schnitt mit gesundem kolorektalem Gewebe. Die Aufteilung der Klassen auf den gesamten Datensatz kann Tabelle 2.1 entnommen werden.

Tabelle 2.1: Farbpalette und Verteilung der segmentierten Klassen.
Insgesamt wurden 13 Klassen annotiert. Innerhalb QuPath wurde jeder Klasse ein spezifischer Farbcode für die weitere Verarbeitung zugewiesen. Die Verteilung der einzelnen Klassen im Vergleich zum gesamten Datensatz wird in der letzten Spalte angegeben.

Farbe	**Klasse**	**Anteil** [%]
	Hintergrund	18.5
	Tumor	33.1
	Mukosa	6.6
	Submukosa	2.8
	Muscularis	12.8
	Fettgewebe	12.6
	Stromale Reaktion	7.1
	Artefakt	0.3
	Entzündung	1.1
	Nekrose	0.5
	Andere	1.4
	Mukus	0.8
	Nicht-tumoröse Pathologie	2.5

Die originalen .svs-Dateien samt dazugehöriger Annotationsmaske wurden mittels eines QuPath-Skripts in Kacheln der Größe 1000×1000 Pixel aufgeteilt. Das Skript wurde von Peter Bankhead veröffentlicht, durch Marlen Runz angepasst und

ist unter [25] abrufbar. Der dadurch erzeugte Datensatz bestand aus Paaren von Kacheln mit je einer Annotationsmaske und dem korrespondierenden Ausschnitt aus dem HE Gewebeschnitt. Anschließend wurden diese Paare von Kacheln zufällig in drei Teile für Training (75%), Validierung (15%) und Test (10%) des neuronalen Netzwerks aufgeteilt.

2.1.3 Architektur und Training des U-Net

Zur Segmentierung wurde die PyTorch-Implementierung eines angepassten U-Nets von Janowczyk verwendet [26, 27]. Die Funktionsweise eines U-Nets basiert auf einem Encoder-Decoder Prinzip und ist schematisch in Abbildung 2.3 dargestellt.

Im ersten Teil, dem Kontraktionspfad, wird eine Abfolge von Convolution- und Pooling-Schichten angewendet. In der Convolution-Schicht wird ein trainierbarer Filter zeilenweise über die Kachel geschoben, um schrittweise die Werte der Feature Map zu berechnen. Beim anschließenden Pooling wird schrittweise das Maximum einer Gruppe von Werten der Feature Map berechnet, wodurch die Pixelzahl der Feature Map vor der nächsten Schicht deutlich reduziert wird. Somit wird die räumliche Auflösung des Datensatzes stark verringert, ein Vorgang der nebenbei Reichenzeit und Speicherplatz spart [20]. Andererseits wird jedoch eine Feature Map mit zunehmenden Kanälen für Eigenschaften, den lernbaren Parametern, generiert. Somit entwickelt das Modell einen detaillierten Einblick in die Eigenschaften des eingegebenen Bildes, kann diese Eigenschaften jedoch kaum räumlich zuordnen. Darum wird die Feature Map vor jedem Pooling, also mit höherer Auflösung, als Kopie an den Decoder weitergegeben [29–31]. Dies entspricht den Querverbindungen in Abbildung 2.3.

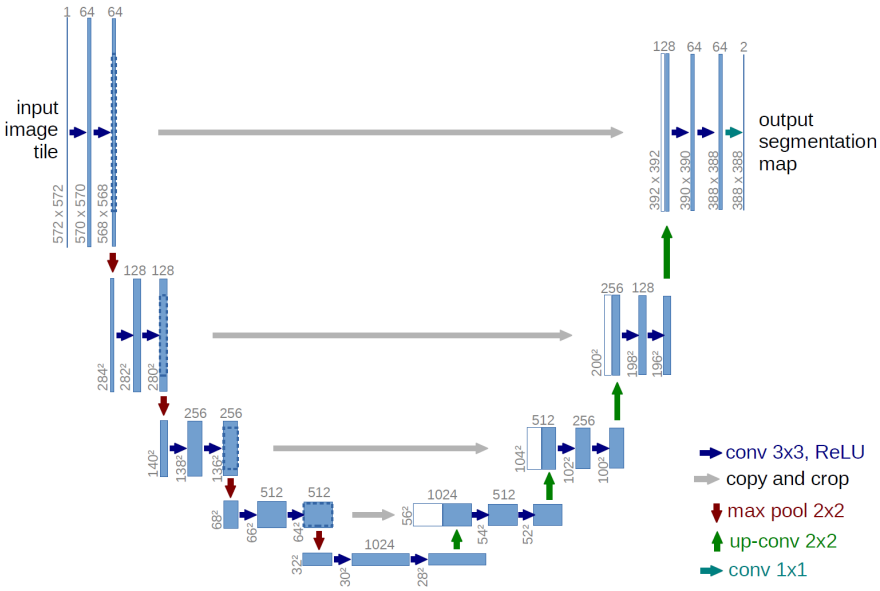

Abbildung 2.3: Schematische Dartellung der Architektur eines U-Nets. Beispiel für eine U-Net Architektur [28] für die Segmentierung einer Eingabe der Größe 572×572 Pixel mit maximal 1024 Eigenschaften. Oberhalb der blauen Flächen, welche Feature Maps mehrerer Kanäle darstellen, ist die Anzahl der Kanäle notiert. Unten links der Flächen ist die Auflösung in x-y-Richtung beschrieben. Weiße Flächen stellen kopierte Feature Maps dar. Operationen: conv: Faltung mittels eines 3×3 Kerns (convolution), ReLu: rectified linear Unit, max pool: Max Pooling, up-conv: transponierte Faltung mittels 2×2 Kern. Abbildung mit freundlicher Genehmigung von [28]

Im zweiten Teil, der Expansion, auch Decoder genannt, wird mittels transponierter Faltung die Auflösung wieder erhöht. Gleichzeitig werden die Feature Maps aus dem entsprechenden Schritt der Kontraktion kopiert und mit einbezogen. Die Gesamtzahl der Feature Map Ebenen wird jedoch entlang des Decoder-Pfads wieder reduziert. Die Feature Map enthält in diesem Fall 13 Kanäle für die 13 Klassen der Segmentierung, die Segmentierungsmaske. Für jede der Operationen entlang des Encoder/Decoder Pfads, wie zum Beispiel (transponierte) Faltung gibt es einen spezifischen Gewichtungsfaktor, der im Verlauf des Trainings optimiert wird [29–31].

Die Segmentierungsmaske wird durch das Netzwerk anschließend mit der Annotationsmaske verglichen. Mittels Dice Loss [32] wird der Verlust des Netzwerks zwischen manueller Annotation und Segmentierung durch das U-Net pixelweise bestimmt. Der Dice Loss zeigte sich in verschiedenen Studien als besonders geeignet für das Training einer semantischen Segmentierung [33]. Um das Ungleichgewicht bei der Verteilung der unterschiedlichen Gewebeklassen auszugleichen, wird der Dice Loss mit einer Gewichtung berechnet. In der Trainigsphase korrigiert ein Adam Optimizer [34] nach jedem Durchlauf auf Grundlage des Loss die Netzwerkparameter. Die Lernrate skaliert diese Korrektur. Ein eigenes Steuerprogramm verringert im Verlauf des Trainings die Lernrate abhängig von der Änderung des Loss. Das Ausmaß der Korrekturen durch den Optimizer nimmt somit im Trainingsverlauf ebenfalls ab [29, 35].

Da die Präparate und die daraus erstellten Kacheln keine einheitliche Sättigung durch die HE Färbung aufwiesen und das Modell sich nicht spezifische Kacheln auswendig merken sollte, wurden im Rahmen der Datenaugmentation zufällig Transformationen der Kacheln bezüglich Farbe, Größe und Orientierung durchgeführt [29, 36].
Der endgültige Code und das im weiteren Verlauf verwendete Netzwerk können unter [36] eingesehen werden. Das U-Net erreichte auf Testdaten eine pixelweise Genauigkeit von 72%. Für die weitere Verwendung des Modells war dieser Wert ausreichend, da der Fokus weniger auf einer pixelgenauen Analyse lag. Es sollten lediglich Areale verschiedenen Gewebes unterschieden werden. Der Code kann unter [37] eingesehen werden.

2.2 Klassifikation

Im zweiten Teil dieser Arbeit wurde ein Modell zur Klassifizierung eines Datensatzes (siehe Kapitel 2.2.1) hinsichtlich verschiedener klinischer Verlaufsparameter entwickelt. Die WSIs wurden dabei entsprechend einem Schema in Abbildung 2.4 vorbereitet. Die Aufteilung der Daten für die jeweilige Klassifikation wird in Abbildung 2.6 und Kapitel 2.2.2 dargestellt.

2.2.1 Datenakquise und -vorbereitung

Die HE Gewebeschnitte des FFPE Tumorgewebes wurden digital zusammen mit einer anonymisierten Liste der Patient*innendaten durch das Pathologische Institut der medizinischen Fakultät Graz der Universität Graz zur Verfügung gestellt. Von den Patient*innendaten wurden im weiteren Verlauf Budding-, Nodalstatus, Grading und Progress verwendet.

Definition klinischer Parameter

Nodalstatus: Der in dieser Arbeit beschriebene Nodalstatus N gleicht dem durch eine Pathologin oder einen Pathologen erhobenen Status pN. Im Einklang mit der TNM Klassifikation ist der Nodalstatus definiert als N0 für Fälle ohne jegliche Lymphknotenmetastasen, N1 für Fälle mit ein bis drei regionalen Lymphknotenmetastasen und N2 bei mehr als drei betroffenen regionalen Lymphknoten [38].

Buddingstatus: Der Buddingstatus wurde basierend auf Satoh's fünf-Stufen-Modell eingeteilt [39]. Hierbei werden Fälle ohne Budding der Klasse B0 zugeordnet, Fälle mit ein bis drei Buds der Klasse B1, mit fünf bis neun Buds der Klasse B2, mit zehn bis 19 Buds der Klasse B3 und ab 20 Buds der Klasse B4. Diese Klassifizierung ist im Vergleich zu den aktuellen Guidelines [7, 9] veraltet,

konnte in verschiedenen Studien aber trotzdem als Prädikator für Lymphknotenmetastasen bestätigt werden [39–41].

Grading: Die Definition des Grading entspricht dem in der Pathologie gängigen vier-Stufen-Schema. G1 beschreibt dabei Tumoren mit hoher Differenzierung und G4 Tumoren mit keiner Differenzierung, beziehungsweise Anaplasie [42].

Progress: Die Eigenschaft klinischer Fortschritt wurde entweder auf Eins für Fälle mit Fortschritt gesetzt oder auf Null für Fälle bei denen kein weiteres Wachstum des Tumors beobachtet werden konnte.

Bildvorverarbeitung

Die WSIs aus dem Kollektiv wurden in Kacheln der Größe 1000×1000 Pixel aufgeteilt und anschließend jene Kacheln aussortiert, welche aufgrund der randständigen Lage innerhalb des Präparats eine ungleiche Kantenlänge oder nur wenige Pixel mit Gewebe aufwiesen (siehe Abbildung 2.4). Für diese Schritte wurde der Code aus den Projekten wsi-tools und Openslide verwendet [43, 44].

Das U-Net aus Kapitel 2.1 wurde herangezogen, um die verbliebenen Kacheln automatisch zu annotieren. Der Anteil der als Tumor klassifizierten Pixel je Kachel wurde ermittelt und die Kacheln dementsprechend in zwei Gruppen aufgeteilt:

- Gruppe PT: Kacheln mit $50-75\%$ Tumoranteil, entsprechend den peripheren Regionen des Tumorareals
- Gruppe ZT: Kacheln mit $\geq 95\%$ Tumoranteil, entsprechend den zentralen Regionen des Tumorgebiets.

Bildeigenschaften wie Helligkeit, Farbsättigung variieren zwischen den einzelnen WSIs beispielsweise aufgrund unterschiedlicher Akquirierungszeitpunkte oder auch Färbeprotokolle. Für eine möglichst stabile Klassifizierung, welche sich nicht auf

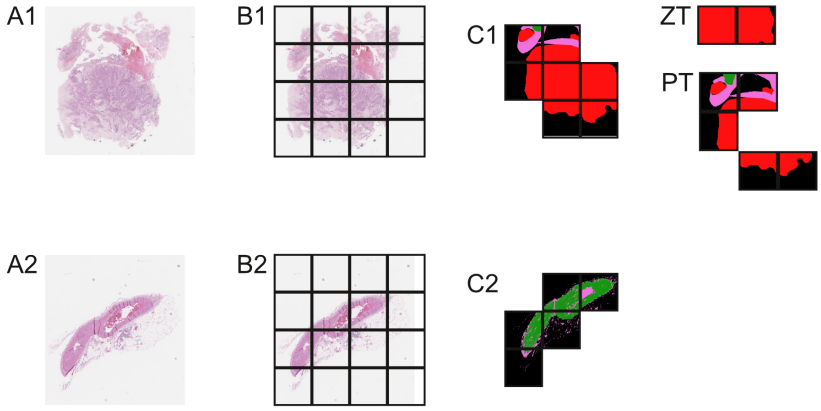

Abbildung 2.4: Ablauf der Klassifikation der HE Schnitte.
Die Whole Slide Images (WSIs) (A1 mit Tumor, A2 ohne Tumor) des Graz-Kollektivs wurden gekachelt (B1, B2) und Kacheln mit der falschen Kantenlänge oder zu wenig Gewebe aussortiert. Anschließend wurde das vorhandene Gewebe mittels des oben erwähnten U-Net segmentiert (C1, C2). Der Anteil des Tumorareals je Kachel wurde bestimmt und Kacheln mit $\geq 95\%$ Tumoranteil wurden der Gruppe zentraler Tumorareale (ZT) zugeordnet. Kacheln mit $50 - 75\%$ Tumoranteil wurden der Gruppe peripherer Tumorareale (PT) zugeordnet. Kacheln ohne Tumor wurden für die weiteren Untersuchungen nicht benötigt (C2). Farblegende: rot - Tumor, schwarz - Hintergrund, grün - Submukosa, pink - Stromale Reaktion

spezifische Farbschemata der einzelnen Fälle im Patient*innenkollektiv fokussiert, wurden die Kacheln deshalb hinsichtlich ihrer Färbung normalisiert. Hierfür wurde ein Ansatz basierend auf einem Generative Adversarial Network (GAN) verwendet, welches durch Marlen Runz zur Verfügung gestellt wurde [22].

2.2.2 Erstellung der Datensätze

Aus den bisher ausgewählten Kacheln wurden nun für jede Klassifikation spezifische Datensätze erstellt. Dabei wurden zwei Verfahren exploriert, deren maßgeblicher Unterschied in Abbildung 2.5 dargestellt ist.

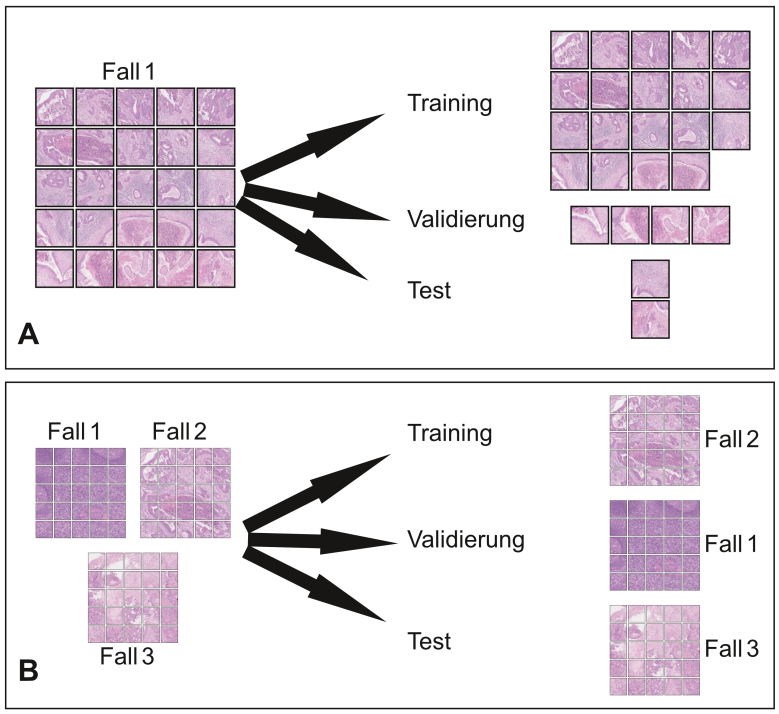

Abbildung 2.5: Vergleich der Verfahren zur Sortierung der Kacheln.
A) Für die meisten Klassifikationen wurden die Daten entsprechend dieses Schemas sortiert. Hierbei konnten ein Fall so aufgeteilt werden, dass Trainings-, Validierungs- und Testset jeweils Kacheln des Falls enthalten. B) Bei diesem Verfahren wurden die Fälle als ganzes auf die verschiedenen Sets verteilt, so dass keine Aufteilung eines Falles wie in A) stattfinden konnte.

Datensätze mit gemischten Fällen (vergleiche *A)* in Abbildung 2.5)

In einem ersten Durchlauf wurden die Kacheln ohne Berücksichtigung des zugehörigen Falls auf Training, Validierung und Test verteilt. Für die Klassifizierung der Kacheln aus der Gruppe zentraler Tumorareale (ZT) hinsichtlich deren Nodalstatus wurden beispielsweise alle Kacheln aus dem Zentrum des Tumors entsprechend ihres in den Patient*innendaten vermerkten Nodalstatus auf drei Klassen (N0−N2) aufgeteilt. Danach wurden die Kacheln jeweils innerhalb einer Klasse jeweils zufällig auf Sets für Training (75%), Validierung (15%) und Test (10%) aufgeteilt, sodass am Ende jedes der drei Sets Kacheln aus allen Klassen enthielt (siehe Abbildung 2.6). Somit lagen zu einem einzelnen Fall beziehungsweise WSI verschiedene Kacheln im Datensatz, welcher zum Training des Klassifikators verwendet wurde und weitere Kacheln desselben WSI in jenem Datensatz, mit welchem der trainierte Klassifikator getestet wurde.

Datensätze getrennt nach Fällen (vergleiche *B)* in Abbildung 2.5)

In einem zweiten Verfahren wurden die Präparate so sortiert, dass alle Kacheln zu einem Fall beziehungsweise WSI entweder nur im Datensatz für das Training des Klassifikators auftauchen oder nur im Datensatz für die Validierung beziehungsweise den Test des Klassifikators. Die Zuordnung der jeweiligen Fälle zu den entsprechenden Phasen der Entwicklung erfolgt dabei wieder zufällig. Die auf diesen Datensätzen trainierten Klassifikatoren tragen in der Kennzeichnung das Suffix „Group“.

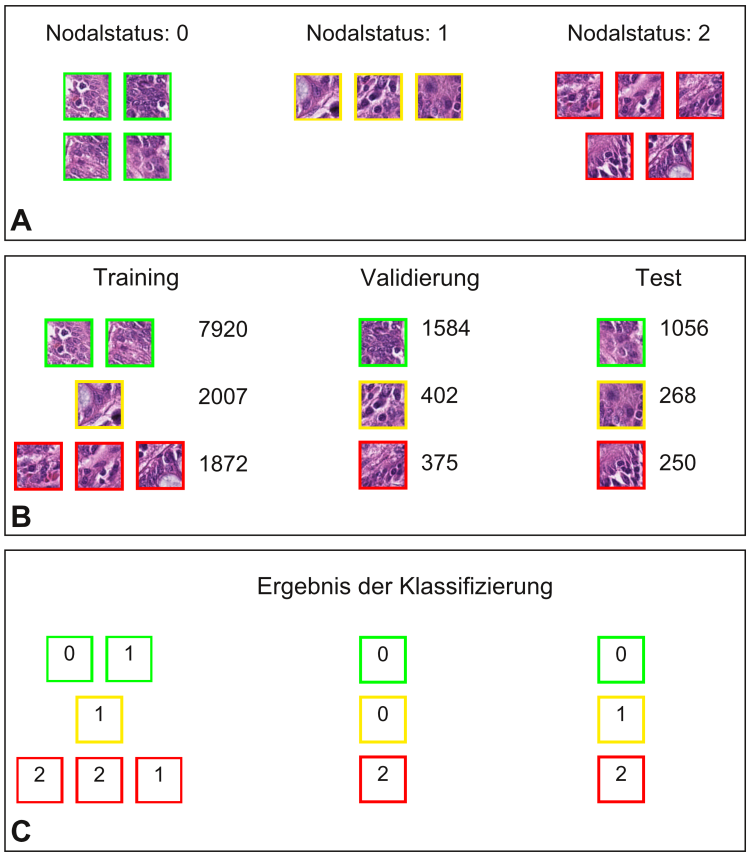

Abbildung 2.6: Sortierung der Daten für die Klassifikation.
Hier wird beispielsweise die Sortierung der Kacheln für die Klassifizierung ZT zu Nodalstatus dargestellt. A) Im ersten Schritt werden die Kacheln der Gruppe ZT in ihre jeweiligen Klassen sortiert, in diesem Fall der entsprechende Nodalstatus. B) Anschließend werden die Kacheln je Nodalstatus zufällig auf die Phasen Training, Validierung und Test mit dem festen Verhältnis von 75%, 15% und 10% aufgeteilt. Für die dargestellte Klassifikation wurden somit 7920 Kacheln mit N0 dem Training zugeordnet, 402 Kacheln mit N1 der Validierung und 250 Kacheln mit N2 dem Test. C) Für jede Kachel trifft der Klassifikator eine Vorhersage bezüglich des Nodalstatus.

2.2.3 Architektur und Training des ResNets

Zur Klassifizierung wurde ein auf dem ImageNet Datenset vortrainiertes Residual Neural Network (ResNet) mit 152 Ebenen verwendet [45, 46]. Das Kernelement dieses Netzwerks sind die sogenannten Residual Blocks, welche parallel zu zwei Faltungen jeweils die Identitätsfunktion anwenden und somit am Ende des Blocks die erhaltene Feature Map zum eigentlichen Eingangswert addieren. Somit wird das Problem umgangen, dass bei der Aneinanderreihung von sehr vielen Operationen deren Gewichtungsfaktoren im Laufe des Trainings gegen null tendieren und verschwinden. Dieses Verhalten wird auch gemeinhin als „vanishing Gradient“bezeichnet [47]. Im Gegensatz dazu können die Gradienten direkt entlang der Identitätsfunktion retrograd propagiert werden. Das Ergebnis der Faltungen und Identitätsfunktion dient abschließend als Eingangswert für eine Aktivierungsfunktion, meist eine Rectified Linear Unit (siehe Abbildung 2.7) [20, 47].

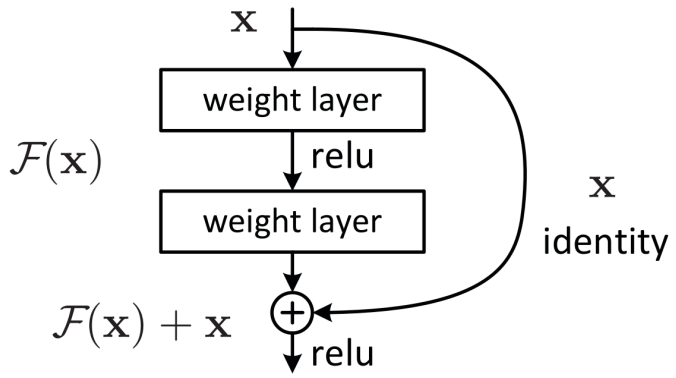

Abbildung 2.7: Residual Block.
Der Residual Block stellt das Grundelement des ResNet dar. Eine Eingangsvariable x durchläuft zwei Faltungen („weight Layer“). Parallel dazu wird die Identität von x weitergegeben und anschließend zum Ergebnis der Faltung addiert als $\mathcal{F}(x) + x$. Abschließend wird eine Rectified Linear Unit („relu“) angewendet, bevor der nächste residual Block angewendet werden kann [47].

Von diesen Residual Blocks können theoretisch beliebig viele hintereinander ausgeführt werden, in Abbildung 5.1 im Anhang ist beispielsweise ein ResNet mit 34 Schichten dargestellt. Am Ende der Residual Blocks erfolgt noch ein Average Pooling und ein Dense Layer rechnet die erstellte Feature Map in eine Liste aus Vorhersagewahrscheinlichkeiten für alle Klassen um. Das Maximum dieser Wahrscheinlichkeiten wird als Ergebnis der Klassifizierung ausgegeben. Die verschiedenen Operationen innerhalb des ResNet haben ebenfalls eine Gewichtung, welche im Training nach jedem Durchlauf entsprechend dem Fehler durch einen Optimizer angepasst wird. [20, 47]

Das Training wurde über jeweils 50 Epochen durchgeführt. Es wurde jeweils ein Modell für jede der Variablen Budding-, Nodalstatus, Progress für jede der Datengruppen ZT und periphere Tumorareale (PT) trainiert. Zusätzlich wurde ein Modell basierend auf sowohl zentralen als auch peripheren Tumorarealen für die Variable Nodalstatus und ein weiteres basierend auf zentralen Tumorarealen für die individuelle Fallkennung trainiert.

2.2.4 Class activation maps

Um die Areale innerhalb einer Kachel zu bestimmen, welche ausschlaggebend für das Ergebnis der Klassifikation sind, wurden CAMs nach einer Vorlage von Pointer [48] berechnet. Es wurde für jedes Pixel eine Wahrscheinlichkeit für die getroffene Vorhersage kalkuliert und die Ergebnisse mittels einer Heatmap dargestellt. Areale mit hohen Vorhersagewerten treten somit farblich hervor und der Fokus verschiedener Klassifikationen kann verglichen oder die Kacheln auf bestimmte morphologische Charakteristiken untersucht werden.

2.3 Statistische Evaluation

Das Hauptaugenmerk der statistischen Evaluation lag darauf, die Vorhersage-Güte der verschiedenen Klassifikationen zu vergleichen. In einem zweiten Schritt wurde die Fähigkeit einzelner Klassifikationen untersucht, genau zwischen den jeweiligen Klassen zu unterscheiden.

2.3.1 Metriken

Zwei gängige Maße zur Beurteilung einer Klassifikation sind die Genauigkeit und das F-Maß. Die Genauigkeit A berechnet sich dabei aus [49]:

$$A = \frac{TP + TN}{N} \tag{2.1}$$

mit TP der Anzahl an richtig positiv klassifizierten, TN der Anzahl an richtig falsch klassifizierten Ergebnissen und N der Größe des Datensatzes.

Das F1-Maß basiert auf dem positiven Vorhersagewert PPV und der richtig-positiv Rate TPR, welche wie folgt definiert sind:

$$PPV = \frac{TP}{TP + FP} \tag{2.2}$$

$$TPR = \frac{TP}{TP + FN} \tag{2.3}$$

mit TP der Anzahl an richtig positiven Zuordnungen, FN der Menge an falsch negativen Ergebnissen, FP den falsch positiven Ergebnissen.

Damit ergibt sich nach [50] das F1-Maß durch:

$$F_1 = 2 \times \frac{PPV \times TPR}{PPV + TPR} \tag{2.4}$$

Bei der Genauigkeit A liegt der Schwerpunkt auf den richtig klassifizierten Ergebnissen. Beim F1-Maß werden hingegen auch die falsch klassifizierten Ergebnisse mit in die Gewichtung einbezogen und die ungleichmäßige Verteilung der Klassen durch das harmonische Mittel aus PPV und TPR berücksichtigt [51].

2.3.2 Cohen's Kappa

Die Qualität der Klassifikationen wird maßgeblich durch die Übereinstimmung zwischen Vorhersage und der wahren Eingabe bestimmt. Ein Maß für diese Übereinstimmung ist Cohen's Kappa. Die Berechnung basiert auf der k mal k großen Konfusionsmatrix, wobei k der Anzahl der jeweiligen Klassen innerhalb der Klassifikation entspricht. Innerhalb der Matrix definiert ein Element f_{ij} die Anzahl an Kacheln, welche laut den Patient*innendaten der Klasse i zugeordnet werden und durch das ResNet der Klasse j zugeordnet wurden. Somit stellt f_{jj} die Anzahl an Übereinstimmungen zwischen Patient*innendaten und Vorhersage durch das ResNet für die Kategorie j dar. Mit [52] folgt:

$$P_o = \frac{1}{N} \sum_{j=1}^{k} f_{jj}, \tag{2.5}$$

$$r_i = \sum_{j=1}^{k} f_{ij}, \forall i, \text{ and } c_j = \sum_{i=1}^{k} f_{ij}, \forall j, \tag{2.6}$$

$$P_e = \frac{1}{N^2} \sum_{i=1}^{k} r_i c_i, \tag{2.7}$$

mit P_o der beobachteten proportionalen Übereinstimmung, r_i und c_j die Summe der jeweiligen Zeilen und Spalten für Klasse i und j und P_e die zu erwartende zufällige Übereinstimmung.

Der Wert für Kappa berechnet sich dann mit (2.8).

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{2.8}$$

Die Standardabweichung von κ ist gegeben durch [52]:

$$std(\kappa) = \sqrt{\frac{P_o(1 - P_o)}{N(1 - P_e)^2}} \tag{2.9}$$

Landis und Koch bezeichnen eine Übereinstimmung mit Werten von $\kappa = [0,6; 0,8]$ als substanzielle Übereinstimmung und $\kappa > 0,8$ als nahezu perfekte Übereinstimmung zwischen zwei verschiedenen Beurteilungen [53].

2.3.3 Fläche unter der Receiver Operating Characteristics Kurve

Um die Fähigkeit der Klassifikation, zwischen den einzelnen Klassen zu unterscheiden beziehungsweise zu beurteilen, wurde die Fläche unter der Receiver Operating Characteristics-Kurve (AUROC) berechnet. Sie basiert auf der Kurve der Receiver Operating Characteristics (ROC), welche bei einem Grenzwert, hier die Vorhersagewahrscheinlichkeit, die $TPR = TP/(TP + FN)$ gegen die falsch positiv Rate (FPR) aufträgt [51].

Die FPR ist hierbei definiert durch:

$$FPR = \frac{FP}{FP + TN} = 1 - Spezifität \tag{2.10}$$

Da die Definition der ROC für binäre Klassifikationen ausgelegt wurde, kann entweder eine Klasse mit einer anderen Klasse im Sinne eines „one vs one“ (OvO) Settings verglichen werden oder man vergleicht eine Klasse mit den restlichen Klassen (OvR) [54].

Zusammen mit der jeweiligen Vorhersage gibt die Klassifikation einen Wert für die Vorhersagewahrscheinlichkeit aus. Nimmt man diese Wahrscheinlichkeit als Grenzwert für die Zuordnung in eine Klasse, kann man die oben genannten Variablen für den gesamten Datensatz berechnen. Iteriert man diesen Schritt für alle Kacheln und Vorhersagewahrscheinlichkeiten des Datensatzes, erhält man eine Reihe an Paaren aus TPR und FPR, welche man als ROC-Kurve darstellen kann.

In Zusammenhang mit der ROC-Kurve beschreibt AUROC die Fläche unter der ROC-Kurve. Eine AUROC von 1 steht dabei für komplett separierbare Klassen, wohingegen eine AUROC von $0,5$ einer Klassifikation entspricht, welche genauso gut ist wie eine zufällige Aufteilung. Zur Berechnung wurde eine Anpassung von Trevisan's Code verwendet [55].

3 Ergebnisse

Im Rahmen dieser Arbeit wurde die Klassifizierung von WSIs hinsichtlich prognostischer Faktoren in zwei verschiedenen Settings untersucht:
a) Im ersten Setting wurden Kacheln aller Fälle gemischt und unabhängig von der Fallzugehörigkeit in Datensätze für Modell-Training, -Validierung und -Test aufgeteilt (siehe Kapitel 3.1).
b) In einem zweiten Ansatz wurden die Kacheln nach Fällen gruppiert auf die entsprechenden Sets aufgeteilt (vergleiche Kapitel 3.2).

Für die Zuordnung der prognostischen Faktoren werden zwei Klassifikationen betrachtet:
a) Ein erster Klassifikator betrachtet nur Kacheln histologischer Präparate, welche laut der Segmentierung (siehe Kapitel 2.1) mindestens 95% Tumoranteil aufweisen und als zentrale Tumorareale (ZT) definiert wurden.
b) In einer zweiten Klassifikation wurden ausschließlich jene Kacheln der histologischen Präparate betrachtet, welche laut der Segmentierung einen Tumoranteil von $50 - 75\%$ aufwiesen. Diese Kacheln werden als periphere Tumorareale (PT) bezeichnet.

Auf die Ergebnisse und Qualität des entwickelten U-Nets sowie des Generative Adversarial Network (GAN) und den entsprechenden Vorarbeiten [22] wird nicht genauer eingegangen, da sie nur als Werkzeug dienten, um die Datensätze vorzubereiten. Nach der Aufbereitung der Daten (siehe Kapitel 2) ergaben sich für

die verschiedenen Klassifizierungen Datensätze, deren genaue Zusammensetzung in Tabelle 3.1 veranschaulicht wird. Da die klinischen Parameter Nodal- und Buddingstatus sowie klinischer Progress nicht für alle untersuchten Fälle jeweils vollständig vorhanden waren, gibt es Unterschiede in der Anzahl der Fälle zwischen den einzelnen Klassifizierungen. Die im späteren Verlauf genannten Klassifizierungen bezüglich des Nodalstatus mit nach Fällen gruppierten Kacheln basierend auf zentralen (ZT2NoGroup) und peripheren (PT2NoGroup) Tumorarealen besitzen dieselbe Anzahl an Fällen beziehungsweise Kacheln wie die entsprechenden Klassifizierungen mit einer zufälligen Verteilung der Kacheln (ZT2No, PT2No).

Tabelle 3.1: Verteilung der Kacheln bezüglich der jeweiligen Klassifikation. Nachdem die Kacheln mittels U-Net Segmentierung in eine Gruppe mit zentralen Tumoranteilen (ZT) und eine andere mit peripheren Tumoranteilen (PT) aufgeteilt worden waren, wurden sie für die jeweiligen Klassifikationen entsprechend der klinischen Daten sortiert. Die Anzahl der Fälle ist dabei gleichzusetzen mit der Anzahl der Whole Slide Images (WSIs). Die Benennung der entstandenen Datensätze folgt einem zweiteiligen Schema: Zuerst wird das Tumorareal bezeichnet (ZT oder PT) und anschließend die zu klassifizierende Charakteristik Budding- (Bu), Nodalstatus (No), Progress (Pr), Fallnummer (Fn).

Klassifikation	Anzahl an Fällen	- Kacheln
PT zu Buddingstatus (PT2Bu)	176	39419
ZT zu Buddingstatus (ZT2Bu)	167	15718
PT zu Nodalstatus (PT2No)	178	39600
ZT zu Nodalstatus (ZT2No)	169	15734
ZT und PT zu Nodalstatus (Al2No)	178	55334
PT zu Progress (PT2Pr)	162	36371
ZT zu Progress (ZT2Pr)	154	14448
Fälle aus ZT2No zu Fallnummer (No2Fn)	169	15734

3.1 Klassifizierung basierend auf Fall-unabhängiger Sortierung

Die Klassifikation basierend auf zentralen Tumoranteilen ordnete den Buddingstatus mit einer Accuracy von 91% zu (siehe Abbildung 3.1). Die Klassifizierung basierend auf PT Kacheln und dem gleichen Ziel erreichte eine Accuracy von 86%.

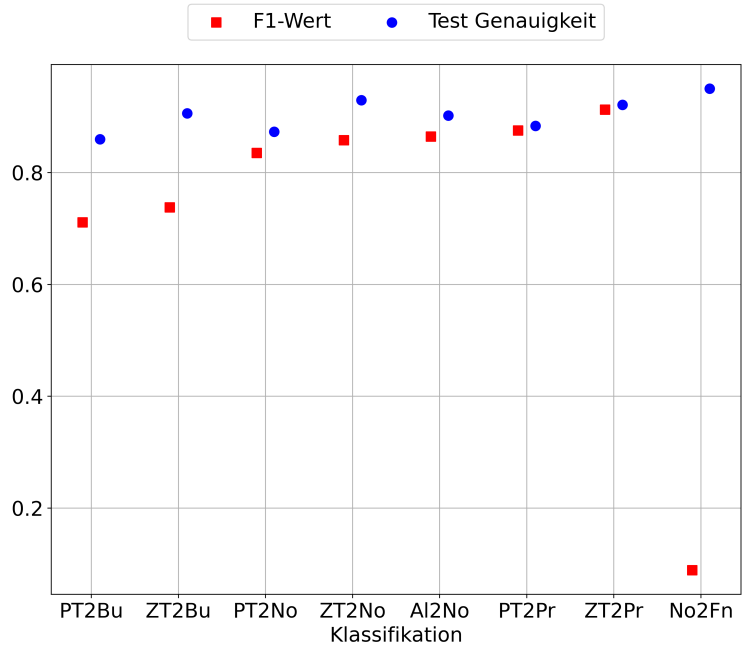

Abbildung 3.1: Accuracy und F1-Wert für die Klassifikation von HE Kacheln. Nach einer Aufteilung des Datensatzes entsprechend Kapitel 2 wurden acht verschiedene Klassifikatoren trainiert. Die Werte für Accuracy und F1 wurden für die Zuordnung von separaten Kacheln, welche nicht für das Training verwendet wurden, durch das beste Modell aus dem Training berechnet. Ausgangswerte: PT: periphere Tumorareale, ZT: zentrale Tumorareale, Al: zentrale und periphere Tumorareale; Abhängige Variablen: Bu: Buddingstatus, No: Nodalstatus, Pr: Progress, No2FN: Klassifikation von Fällen aus dem Set ZT2No zur Fallnummer (siehe auch Tabelle 3.1).

Mithilfe von Cohens Kappa (siehe Kapitel 2.3.2) wurde die Zuverlässigkeit der Zuordnungen bestimmt und ein qualitativer Vergleich ermöglicht. Die Ergebnisse zeigen hier eine substanzielle Übereinstimmung zwischen den klinischen Daten und der Zuordnung basierend auf PT (siehe Abbildung 3.2). Für die Klassifikationen, welche ihre Zuordnung auf ZT basierten, ergab sich eine nahezu perfekte Übereinstimmung mit den klinischen Daten. Den besten Wert von $0,8716$ und einem 95% Konfidenzintervall von $[0,8519;\ 0,8914]$ erreichte die Zuordnung des Buddingstatus zu Kacheln aus dem Bereich ZT.

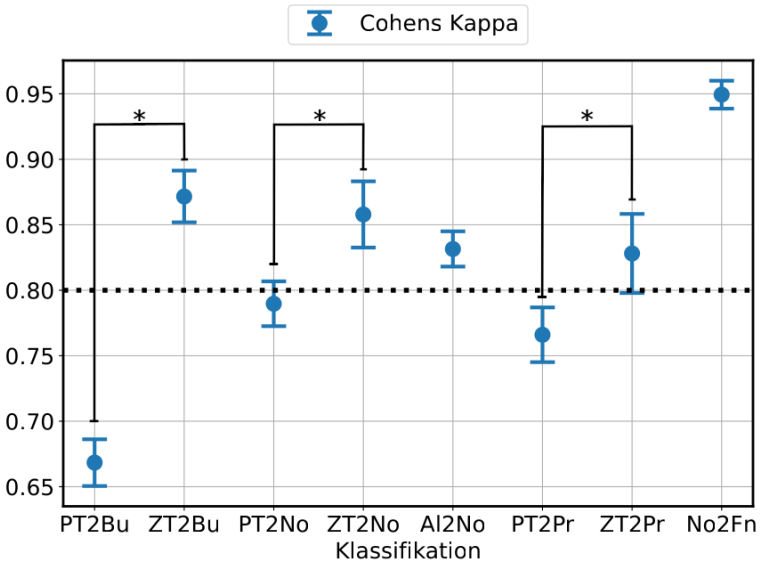

Abbildung 3.2: Cohens Kappa und 95% Konfidenzintervall für die Klassifizierung von HE Präparaten.
Cohens Kappa wurde für jeden Klassifikator berechnet, um die verschiedenen Zuordnungen zu evaluieren und anschließend miteinander vergleichen zu können. Alle Modelle erfüllten die Anforderung mit substanzieller ($\kappa = 0,6 - 0,8$) bis nahezu perfekter ($\kappa = 0,8 - 1$) Übereinstimmung [53]. PT: periphere Tumorareale, ZT: zentrale Tumorareale, Al: zentrale und periphere Tumorareale, Bu: Buddingstatus, No: Nodalstatus, Pr: Progress, No2FN: Klassifikation von Fällen aus dem Set ZT2No zur Fallnummer (Für weitere Informationen zum Datensatz, inklusive Anzahl der jeweiligen Kacheln, siehe auch Tabelle 3.1).
Signifikante Unterschiede zwischen den verschiedenen Klassifikatoren mit $p < 0,001$ sind gekennzeichnet mit $*$.

Wie man in Abbildung 3.2 erkennen kann, erfolgt die Zuordnung basierend auf ZT mit $p < 0,001$ jeweils signifikant besser für die Parameter Budding- (ZT2Bu), Nodalstatus (ZT2No) und Progress (ZT2Pr) im Vergleich zu den Klassifikatoren, deren Zuordnung auf PT fußt (PT2Bu, PT2No, PT2Pr). Die Zuordnung des Nodalstatus zu dem gemeinsamen Datensatz von ZT und PT (Al2No) wurde mit einem Kappa von $0,8315$ samt 95% Konfidenzintervall von $[0,8181;\ 0,845]$ erzielt. In diesem Fall überschneidet sich das Konfidenzintervall mit der Klassifikation des Nodalstatus basierend auf ZT mit einem Kappa von $0,8579$ $[0,8326;\ 0,8832]$.

3.2 Klassifizierung basierend auf nach Fällen sortierter Gruppierung

Die bisherigen Ergebnisse wurden auf Basis von Datensätzen berechnet, bei welchen die Kacheln eines Falles über Trainings-, Validierungs- und Testset verteilt sein konnten (siehe Kapitel 2.2.2). Im Vergleich dazu wurden für die Klassifikation bezüglich des Nodalstatus zusätzlich zwei Datensätze erstellt, indem die Kacheln nach Fällen gruppiert auf Training, Validierung und Test aufgeteilt wurden. Bei den erhaltenen Datensätzen ZT2NoGroup und PT2NoGroup sind Kacheln eines Falles somit nur noch entweder im Trainingsset vorhanden oder im Validierungsset oder im Testset. Die qualitative Auswertung der Klassifikationen ist in Tabelle 3.2 dargestellt. Mit einer Accuracy von $32,8\%$ für ZT2NoGroup und $46,8\%$ für PT2NoGroup wurden weniger als die Hälfte der Vorhersagen korrekt getroffen. Mit Kappa-Werten kleiner 0,5 erreichen beide Modelle eine Übereinstimmung mit den klinischen Daten, welche kleiner ist als eine zufällige Vorhersage.

Tabelle 3.2: Ergebnisse der Klassifikationen auf Basis der nach Fällen gruppierten Datensätze.
ZT2NoGrouped: Klassifikation der nach Fällen gruppierten Kacheln aus zentralen Tumorarealen bezüglich deren Nodalstatus, PT2NoGroup: Klassifikation der nach Fällen gruppierten Kacheln aus peripheren Tumorarealen bezüglich deren Nodalstatus

Klassifikation	Accuracy	F1-Maß	Kappa [95% Konfidenzintervall]
ZT2NoGroup	$0,328$	$0,255$	$0,044[0,016; 0,072]$
PT2NoGroup	$0,468$	$0,379$	$0,126[0,098; 0,153]$

Der Qualitätsunterschied zu den bisher betrachteten Modellen (vergleiche Kapitel 3.1) zeigt sich auch im Setting einer binären Klassifikation, in diesem Fall in der Fähigkeit, zwischen Kacheln mit negativem oder positivem Nodalstatus zu entscheiden. Hierzu wurden in Abbildung 3.3 jeweils die Receiver Operating Characteristics (ROC)-Kurve (siehe Kapitel 2.3.3) für das Setting Nodalstatus negativ gegen Nodalstatus positiv berechnet. Für die Klassifikatoren, welche mit nach Fällen gruppierten Datensätzen arbeiten, ergeben sich Fläche unter der Receiver Operating Characteristics-Kurve (AUROC)-Werte, die kaum besser sind als eine zufällige Zuordnung ($0,642$ für ZT2NoGroup; $0,626$ für PT2NoGroup). Die Klassifikation der nicht gruppierten Kacheln zeigt hingegen eine nahezu perfekte Unterscheidung zwischen nodal negativen und positiven Fällen ($0,997$ für ZT2No; $0,991$ für PT2No).

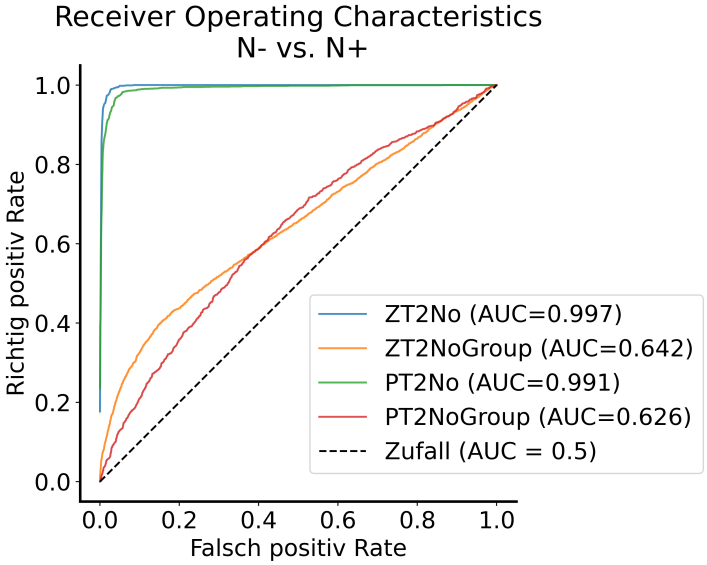

Abbildung 3.3: Receiver Operating Characteristics-Kurven der Klassifikationen für Nodalstatus.
Die Klassifikationen für den Nodalstatus unterscheiden sich hauptsächlich in Abhängigkeit des Datensatzes. Die Klassifikationen ZT2No und PT2No, deren Datensätze nicht nach Fällen gruppiert wurden, erreichen eine deutlich höhere AUROC. Die Klassifikationen ZT2NoGroup und PT2NoGroup, welche auf nach Fällen gruppierten Datensätzen basieren, erreichen eine Leistung, die nur wenig besser ist als eine zufällige Klassifizierung. AUC: Area Under the Curve, ZT: zentrale Tumorareale, PT: periphere Tumorareale, No: Nodalstatus, Group: Datensatz mit nach Fällen gruppierten Kacheln

3.3 Zusammenhang zwischen Grading und Nodalstatus

Um die Funktionsweise des Convolutional Neural Network (CNN) hinsichtlich der Erkennung von Morphologien innerhalb der Whole Slide Image (WSI) Kacheln zu untersuchen, wurden Class Activation Maps (CAMs) angefertigt. Mithilfe der CAMs konnte dargestellt werden, welche Regionen innerhalb einer Kachel ausschlaggebend für die Entscheidung des Klassifikators sind. Die farbliche Darstellung der Vorhersagewahrscheinlichkeiten wurde mit dem ursprünglichen Hämatoxylin-Eosin (HE) Bild überlagert. Im Ergebnis stellen orange bis rote Be-

reiche der Kacheln eben die Regionen dar, welche eine hohe Wahrscheinlichkeit für die letztendliche Zuordnung aufweisen (siehe Abbildung 3.4).

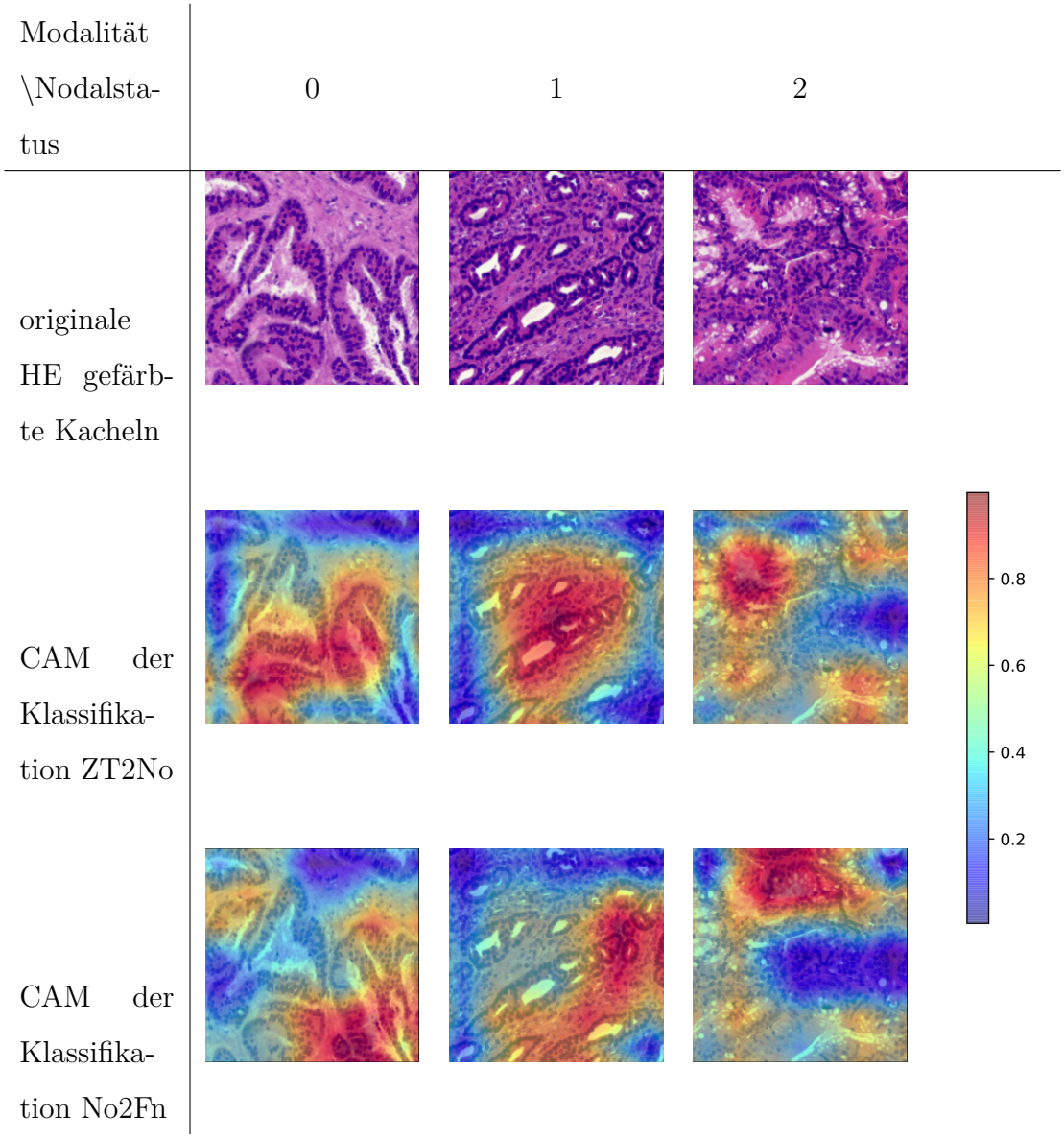

Abbildung 3.4: Class activation map (CAM) für verschiedene Klassifikatoren und Nodalstatus.
Die erste Zeile stellt für jeden möglichen Nodalstatus $0-2$ eine originale Kachel des HE Präparats dar. Die mittlere Zeile zeigt CAMs für die Zuordnung des Nodalstatus zu zentralen Tumoranteilen (ZT2No). In der letzten Reihe sind CAMs für die Klassifikation der Fallnummer basierend auf dem gleichen Datensatz wie (ZT2No) dargestellt. In der gewählten Farbcodierung entsprechen orange bis rote Areale einer hohen Vorhersagewahrscheinlichkeit und blaue Areale einer niedrigen.

Die Betrachtung der CAMs zeigt, dass die Klassifikatoren sich an morphologischen Eigenschaften des Bildes orientieren und nicht einfach nur eine Zuordnung der Kacheln zum jeweiligen Fall lernen. Hierzu wurden für einen Datensatz zwei Klassifikatoren trainiert. In einem ersten Durchlauf wurden ZT Kacheln ihrem Nodalstatus zugeordnet. In einem zweiten Schritt wurde denselben Kacheln ihre Fallnummer zugeordnet. Für beide Klassifikatoren wurden CAMs erstellt und exemplarisch in Abbildung 3.4 verglichen. Es ist deutlich zu erkennen, dass die Klassifikatoren auf jeweils unterschiedliche Areale der Kacheln fokussieren. Für eine genauere Beschreibung der morphologischen Charakteristiken wurde eine Auswahl an Kacheln mit hohem Vorhersagewert durch einen erfahrenen Pathologen (C. W.) untersucht. Diese Analyse ergab, dass sich das Grading in gewisser Weise proportional zum Nodalstatus verhält. So wiesen Kacheln von Fällen ohne Lymphknotenmetastasierung eine höhere Differenzierung auf als Kacheln von Fällen mit Lymphknotenmetastasierung. Bei der Betrachtung des gesamten Patient*innenkollektivs hinsichtlich dieser Merkmale zeigte sich über den gesamten Datensatz hinweg eine ähnliche Verteilung von Grading zu Nodalstatus (siehe Abbildung 3.5).

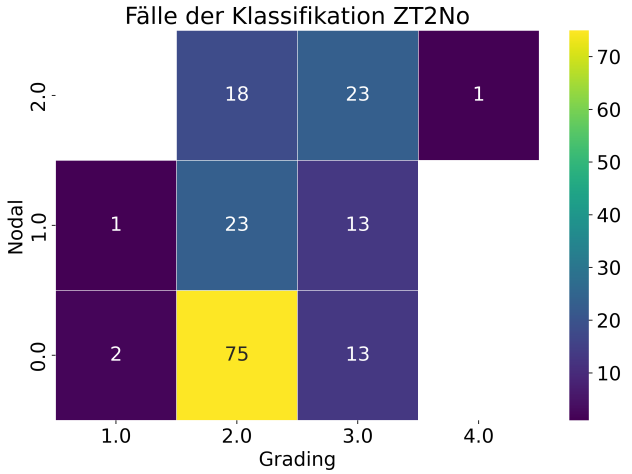

Abbildung 3.5: Nodalstatus und Grading der Fälle der Klassifikation ZT2No. Die Anzahl für jede Kombination aus Grading und Nodalstatus wird für den Datensatz der Klassifikation zentraler Tumorareale (ZT) bezüglich deren Nodalstatus farblich dargestellt. Weiße Felder bedeuten, dass zu dieser Kombination aus Nodalstatus und Grading kein Fall existierte.

4 Diskussion

In den Ergebnissen zeigt sich eine deutliche Differenz in Abhängigkeit der Vorsortierung der Daten. Bei einer Verteilung der Daten unabhängig der entsprechenden Fälle konnten die prognostischen Marker fast perfekt bestimmt werden (vergleiche Kapitel 3.1). Wurden die Kacheln jedoch nach Fällen gruppiert und in Trainings- und Testsets verteilt, so erwies sich die Vorhersage des Nodalstatus als deutlich schwieriger, mit einer Genauigkeit wenig besser als eine zufällige Bestimmung (vergleiche Kapitel 3.2).

4.1 Fallstricke des reinen Auswendiglernens von Datensätzen

Grundlegend ist es schwer möglich, mit Gewissheit fehlerhafte Stellen innerhalb eines neuronalen Netzwerks zu benennen [20]. Eine abschließende Interpretation der verschiedenen Layer des Convolutional Neural Network (CNN), vor allem der aus dem Bildmaterial extrahierten abstrakten Eigenschaften, gestaltet sich noch schwierig. Der hier verwendete Ansatz, mittels Class Activation Maps (CAMs) ausschlaggebende Morphologien hervorzuheben, bleibt auf einzelne punktuelle Hervorsagen begrenzt. Zwar zeigt sich in Kapitel 3.3, dass die Klassifizierung bezüglich des Nodalstatus sich auf andere Bildbereiche fokussiert als eine Klassifizierung, welche den Kacheln deren jeweilige Fallnummer zuordnet. Dies lässt hoffen, dass das Modell morphologische Charakteristiken erkannt hat, welche für den Nodal-

status prädikativ sind. Zeitgleich geht damit die Vermutung einher, dass sich diese Morphologien von fallspezifischen Präparateigenschaften unterscheiden. Betrachtet man jedoch ein Setting wie in Kapitel 3.2, bei dem fallspezifische Eigenschaften definitiv keinen Einfluss auf die Vorhersage haben können, so ergibt sich ein ernüchterndes Ergebnis. Die statistische Auswertung ergibt mit einer Fläche unter der Receiver Operating Characteristics-Kurve (AUROC) von 0,642 für zentrale Tumorareale und 0,626 bei peripheren Tumorarealen eine Bestimmung des Nodalstatus, welche wenig besser ist, als eine zufällige Vorhersage. Auch die Betrachtung der Receiver Operating Characteristics (ROC)-Analyse in Abbildung 3.3 lässt erahnen, dass die scheinbar perfekten Ergebnisse der erstgenannten Modelle eher durch Überanpassung zustande kommen als durch ein ausgeklügeltes Training von morphologischen Eigenschaften.

In der Zusammenschau legen die Ergebnisse nahe, dass die Modelle aus Kapitel 3.1 die einzelnen Fälle auswendig gelernt haben, anstatt relevante histologische Morphologien zu erkennen. Dieses Problem ist nicht neu und vor allem bei den in der Medizin verbreitet auftretenden Datensätzen geringer Größe verbreitet. Durch den im Training auftretenden Prozess des „Overfitting“ passt sich das Modell progressiv spezifisch an den Trainingsdatensatz an. Treten nun im Testdatensatz die gleichen Merkmale auf, so kann das Modell diese mit den gleichen Fällen aus dem Training in Verbindung bringen und nahezu perfekte Vorhersagen treffen. Dies geht jedoch zulasten der Generalisierbarkeit, was sich durch das schlechte Vorhersageverhalten auf einem bisher unbekannten Testset zeigt. Können die im Training auswendig gelernten Eigenschaften des Datensatzes im Testdatensatz nicht mehr wiedergefunden werden, ist das Modell nicht imstande, zuverlässige Prognosen zu treffen [56]. Wie Abschnitt 3.2 darlegt, kann das Residual Neural Network (ResNet) die komplett neuen Fälle des Testsets nicht den Fällen des Trainingssets zuordnen.

Die erlernten Parameter des CNN zielen auf Eigenschaften ab, welche in den Fällen des Testsets nicht vorhanden sind. Klassifikationen sind somit nur mit geringen Vorhersagewahrscheinlichkeiten möglich. Es gilt somit, eine Deep Learning Architektur zu finden, welche spezifisch fallunabhängige morphologische Eigenschaften mit hohem prädikativem Wert erlernen kann.

4.2 Der Datensatz und seine Anforderungen an ein Deep Learning Model

4.2.1 Noisy Label

Die in Kapitel 4.1 geforderten morphologischen Eigenschaften sind jedoch nicht in allen Kacheln ausgeprägt. Für die Klassifikation wurden die Bilddatensätze in zentrale und periphere Tumorareale getrennt, mit der Absicht, dass Kacheln mit peritumoralem Budding definitiv nur dem Datensatz peripherer Tumorareale (PT) zugeordnet werden. Somit sollte anschließend eine Unterscheidung möglich sein, ob die Klassifikation von Kacheln, welche Budding enthalten, bessere Ergebnisse liefert als Klassifikationen von Kacheln aus dem Datensatz zentraler Tumorareale (ZT), welcher definitiv kein Budding aufweist. Bei genauerer Betrachtung muss man jedoch feststellen, dass die Kacheln aus dem PT Datensatz selbst nur in geringen Mengen Budding darstellen. Die Separationskriterien verlangten nicht spezifisch nach Budding, sondern lediglich nach einem geringen Tumoranteil, welcher meistens mit dem Tumorrand übereinstimmt. Dieser Tumorrand kann auch zum Lumen hin liegen, sodass Kacheln aus dem PT Datensatz abgesehen vom Tumor gar kein Gewebe erfassen. Bei der Untersuchung des Datensatzes in einer weiteren Studie konnte gezeigt werden, dass nur circa 10% aller Kacheln in dieser Hinsicht signifikante Informationen tragen [57]. Folglich sind 90% der Kacheln mit

einem Label versehen, wofür das Bildmaterial selbst keine Informationen liefert. Die Kennzeichnung dieser Kacheln muss aus Sicht des Models als fehlerhaft betrachtet werden. Da die Kacheln mit relevanten Informationen jedoch a priori nicht von jenen ohne signifikante Morphologien zu unterscheiden sind beziehungsweise eine solche Vorauswahl das Ergebnis beeinflussen könnte, ist ein Deep Learning Model zu wählen, welches eine fehlerhafte Kennzeichnung zu kompensieren vermag. Die Literatur liefert verschiedene Verfahren, um mit sogenannten „Noisy Label“ umzugehen [58–60]. Zu diesen Verfahren zählen eine ausgedehnte Augmentation der Daten, wie etwa die hier verwendete Farbnormalisierung (siehe Kapitel 2.2.1). Jedoch wurde auch unter diesen Bedingungen der Datensatz auswendig gelernt. Andere Methoden, wie ein früher Stopp des Trainings, führten zu einem ähnlichen Ergebnis [57]. Es sei noch angemerkt, dass auf den Kacheln des ZT Datensatzes jedoch sehr sicher kein Budding zu beobachten war. Die Klassifikationen erwiesen sich hier mit einer Genauigkeit von bis zu $0,93$ und Kappa von $0,858$ mit einer Standardabweichung von $[0,833; 0,883]$ dennoch sehr treffsicher. Die Abwesenheit von Tumorbudding scheint hier keine negativen Auswirkungen auf die Qualität des ResNet zu haben.

4.2.2 Kohortengröße

Wendet man sich ab von der Beurteilung einzelner Kacheln hin zu der Beurteilung des Falls als Ganzes, können wiederum andere Deep Learning Architekturen verwendet werden. Nebenbei entspricht diese Herangehensweise auch eher dem klinischen Setting, in dem stets das Präparat in seiner Gesamtheit beurteilt wird, anstatt nur einzelner Ausschnitte. Rusche et al. konnten für diesen Datensatz mit dem Multiple Instance Learning (MIL) Ansatz in Kombination mit einer ausgeklügelten Datenaugmentation, welche den Umfang der Dissertation übersteigen

würde, eine AUROC von $0,794$ erreichen [57]. Jedoch zeigt sich auch dieser Ansatz limitiert durch die insgesamt geringe Größe des Datensatzes. Für eine langfristige Validierung der Methodik ist somit eine Erweiterung der Kohorte nötig. Ein multizentrischer Datensatz kann hier zusätzlich dazu beitragen, die Variabilität der Grunddaten zu erhöhen. Dies hat in Experimenten mit histologischen Präparaten von Kolorektalem Karzinom (KRK) zu einer verbesserten Accuracy und einer gesteigerten Generalisierbarkeit des Modells geführt [61].

4.2.3 Korrelation der prädikativen Parameter

Wie in Abschnitt 3.3 dargestellt wird, besteht bei dem betrachteten Kollektiv ein Zusammenhang zwischen Buddingstatus und Grading. Andere Untersuchungen der Kohorte ergaben zudem eine vergleichbare Korrelation der beiden Parameter mit dem Nodalstatus [57]. Zwar kann man anmerken, dass der Buddingstatus in dieser Kohorte nicht entsprechend der aktuellen Empfehlungen bestimmt wurde. Diese geben laut „Consensus Conference on Tumor Budding in CRC“ ein drei-Stufen-System zur Einteilung des Buddingstatus vor [9, 62]. Die Hotspot-Methode soll dabei zur Berechnung des Buddingstatus herangezogen werden. Sie berücksichtigt nur jene Tumorbuds, welche innerhalb des Feldes mit der größten Dichte an Tumorbuds liegen [12, 63]. Zusätzlich zu den vorgeschlagenen drei Stufen weist dieser Datensatz eine weitere Stufe B0 für Fälle ohne beobachtetes Budding und eine Aufteilung der höhergradigen Buddingstatus in B3 mit $10 - 19$ Buds und B4 mit 20 oder mehr Buds auf. Seine Wertigkeit als Prädikator für Lymphknotenmetastasierung büßt der Buddingstatus durch die andere Berechnung jedoch nicht ein [39–41]. Dies wurde für die betrachteten Daten schon von Harbaum et al. beschrieben [64]. Schafft man es in einer erweiterten Kohorte, Zusammenhänge zwischen Prädikatoren zu entkoppeln, kann man einen Klassifikator dazu drän-

gen, sich auf diverse morphologische Eigenschaften zu fokussieren. Die Relevanz des Buddingstatus als Prädikator kann somit gefestigt werden, aber auch neue histologische Marker könnten somit identifiziert werden.

4.3 Evaluation der untersuchten Klassifikation im Hinblick auf eine Integration in den klinischen Alltag

Ein klarer Vorteil der auf Machine Learning (ML) basierten Klassifizierung von Hämatoxylin-Eosin (HE) Präparaten mit KRK scheint seine Zeiteffizienz und Interrater-Reliabilität zu sein. So konnten Modelle auf Basis des Datensatzes bereits nach wenigen Stunden Training den Nodalstatus zuordnen [57]. Der Nodalstatus konnte dabei direkt bestimmt werden, ohne den zeitintensiven Zwischenschritt der Bestimmung des Buddings. Für diese Vorhersage waren darüber hinaus keine Entscheidungsregeln notwendig. Ganz anders präsentiert sich hier das aktuelle klinische Vorgehen. Die Bestimmung des Buddingstatus gilt weiterhin als Goldstandard, wird jedoch fortwährend neu definiert [9, 65]. Zusammen mit der menschlichen Komponente führt dies zu einer vergleichbar hohen inter-observer Variabilität [17, 18]. Die Fähigkeit der ML Modelle, außerhalb von Entscheidungsalgorithmen ähnlich einer Blackbox zu agieren, ermöglicht es ihnen ohne jegliche durch Fachleute definierte Bildeigenschaften auszukommen. Dies spart zwar die wertvolle Zeit von erfahrenen Patholog*innen, andererseits ist eine Bestätigung der ML Ergebnisse nicht möglich. Um die Ergebnisse der Klassifikation verständlich zu machen, wurden in dieser Arbeit retrospektiv mittels CAMs relevante Bildareale bestimmt. Ein anderer Ansatz in der Literatur basiert auf der Klassifikation mit-

tels MIL [57]. Diese Methode ermöglicht schon allein durch ihre Architektur die Bestimmung aussagekräftiger Kacheln und Bildareale [66].

Für eine verlässliche Integration der Klassifikation in den klinischen Alltag muss deren Generalisierbarkeit noch deutlich weiterentwickelt werden. Neben den in Abschnitt 4.1 aufgeführten Herausforderungen, gilt es auch grundlegende Unterschiede zwischen Institutionen zu kompensieren. Hierzu zählen vor allem unterschiedliche Protokolle bei der Herstellung der HE Präparate, speziell in Bezug auf die Protokolle zur Färbung. Um einen Einsatz der Klassifikation in verschiedenen Institutionen mit verschiedenen Protokollen sicher zu ermöglichen, ist für das Training ein ebenso diverser Datensatz mit Präparaten aus mehreren Instituten nötig. In dieser Arbeit wurde versucht, die Auswirkungen der verschiedenen Färbungen durch eine Farbnormalisierung zu kompensieren [22]. Dieser Schritt zeigte sich nicht ausreichend, um im betrachteten Setting eine belastbare Klassifikation zu ermöglichen. Auch gilt es, weitere prädikative Eigenschaften zu berücksichtigen. Beispielsweise könnte mittels eines Graphen-basierten Modells die räumliche Verteilung der Buds berücksichtigt werden [67, 68]. Andere Studien konnten zeigen, dass Lymphozyteninfiltrate wiederholt in diagnostisch wegweisenden Bildarealen auftreten [21, 69–72]. Insgesamt gilt es, alle Möglichkeiten auszuschöpfen, um ein Maximum an diagnostischer Accuracy zu erreichen.

5 Zusammenfassung

Budding bei Kolorektalem Karzinom konnte sich als Prädikator bezüglich Nodalstatus und als genereller Risikofaktor etablieren. Aufgrund unterschiedlicher Bestimmungsrichtlininen sowie Interratervariabilität ist die klinische Bestimmung noch nicht präzise möglich.
Machine Learning Modelle versprechen nutzerunabhängige Klassifikationen von histologischen Präparaten. Darum soll untersucht werden, ob Budding-, Nodalstatus und Progress auf Basis von zentralen oder peripheren Tumorarealen in Whole Slide Images durch ein Residual Neuronal Network mit 152 Ebenen zuverlässig vorhergesagt werden können. Zusätzlich wollen wir die für die Vorhersage ausschlaggebenden Areale eruieren und mit dem Auftreten von Budding korrelieren.

Das Training der Klassifikatoren wurde mit bis zu 178 Fällen beziehungsweise Whole Slide Images durchgeführt. Mit Datensätzen, bei denen Kacheln nicht nach Fällen gruppiert wurden, ergaben sich substanzielle bis nahezu perfekte Übereinstimmungen zwischen den klinischen Daten und den Zuordnungen durch das Residual Neuronal Network.

Den besten Kappa-Wert von $0,8716$ und einem 95 % Konfidenzintervall von $[0,8519;\ 0,8914]$ erreichte die Zuordnung des Buddingstatus zu Kacheln aus dem Bereich zentraler Tumorareale mit einer Accuracy von 91 %. Die Ergebnisse unterschieden sich dabei jeweils signifikant hinsichtlich des Ursprungs der Kacheln zwischen peripheren und zentralen Tumorarealen mit besseren Ergebnissen aus

den zentralen Bereichen. Die Area Under the Curve lag bei 0,997, beziehungsweise 0,991. Mit Klassifikatoren, welche auf Datensets trainiert wurden, die nach Fällen gruppiert erstellt wurden, wurde bei der Klassifikation des Nodalstatus nur eine Accuracy von 32,8 % basierend auf zentralen und 46,8 % für periphere Tumorareale erreicht. Bei beiden Modellen lag der Kappa Wert unter 0,5 mit einer Area Under the Curve von 0,642, beziehungsweise 0,626. Bei der Analyse ausschlaggebender Bildareale zeigte sich eine gesteigerte Gewichtung von Ausschnitten mit schlechterer Differenzierung. Für die Kohorte konnte bei der Auswertung der klinischen Daten ein proportionales Verhalten von Grading zu Nodalstatus beobachtet werden.

Die Klassifizierung von Hämatoxylin-Eosin Präparaten von Kolorektalem Karzinom bezüglich deren Nodal- oder Buddingstatus basierend auf zentralen oder peripheren Tumorarealen muss in Zusammenschau der Ergebnisse vor allem kritisch eingeordnet werden. Die Klassifikation kann zwar die prognostischen Variablen treffsicher bestimmen, erscheint dabei jedoch abhängig von der Sortierung des Bildmaterials. Für den klinischen Kontext scheint die Bestimmung des Nodalstatus mittels Multiple Instance Learning einem reinen Residual Neuronal Network mit 152 Ebenen überlegen. Ersteres neigt weniger zu Overfitting und kompensiert die Noisy Label Eigenschaften des Datensatzes. Im Hinblick auf eine klinische Implementierung sollte die Diversität und der Umfang der Kohorte gesteigert werden, um eine verbesserte Generalisierbarkeit zu erreichen.

Abbildungsverzeichnis

2.1 Übersicht über den Arbeitsablauf. 8
2.2 Ablauf der Segmentierung der HE gefärbten Gewebeschnitte. 9
2.3 Schematische Dartellung der Architektur eines U-Nets. 12
2.4 Ablauf der Klassifikation der HE Schnitte. 16
2.5 Verfahren zur Sortierung der Kacheln. 17
2.6 Sortierung der Daten für die Klassifikation. 19
2.7 Residual Block. 20

3.1 Accuracy und F1-Wert für die Klassifikation von HE Kacheln. . . . 28
3.2 Cohens Kappa und 95% Konfidenzintervall für die Klassifizierung von HE Präparaten. 29
3.3 ROC-Kurven für N^- vs. N^+ 32
3.4 Class activation map (CAM) für verschiedene Klassifikatoren und Nodalstatus. 33
3.5 Nodalstatus und Grading der Fälle der Klassifikation ZT2No. . . . 34

5.1 Architektur eines ResNet34. 58

Tabellenverzeichnis

2.1 Farbpalette und Verteilung der in QuPath segmentierten Klassen. . 10

3.1 Verteilung der Kacheln bezüglich der jeweiligen Klassifikation. . . . 27

3.2 Ergebnisse der Klassifikationen auf Basis der nach Fällen gruppierten Datensätzen 31

Literaturverzeichnis

[1] J. Douaiher, A. Ravipati, B. Grams, S. Chowdhury, O. Alatise, and C. Are, "Colorectal cancer—global burden, trends, and geographical variations," *Journal of surgical oncology*, vol. 115, no. 5, pp. 619–630, 2017.

[2] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[3] D. A. Joseph, R. G. Meester, A. G. Zauber, D. L. Manninen, L. Winges, F. B. Dong, B. Peaker, and M. van Ballegooijen, "Colorectal cancer screening: estimated future colonoscopy need and current volume and capacity," *Cancer*, vol. 122, no. 16, pp. 2479–2486, 2016.

[4] P. G. Van Putten, L. Hol, H. Van Dekken, J. Han van Krieken, M. Van Ballegooijen, E. J. Kuipers, and M. E. Van Leerdam, "Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening," *Histopathology*, vol. 58, no. 6, pp. 974–981, 2011.

[5] L. J. Smits, E. Vink-Börger, G. van Lijnschoten, I. Focke-Snieders, R. S. van der Post, J. B. Tuynman, N. C. van Grieken, and I. D. Nagtegaal, "Diagnostic variability in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population," *Histopathology*, vol. 80, no. 5, pp. 790–798, 2022.

[6] J. D. Brierley, M. K. Gospodarowicz, and C. Wittekind, *TNM classification of malignant tumours*. John Wiley & Sons, 2017.

[7] A. Benson, Venook, Chen, *et al.*, "NCCN clinical practice guidelines in oncology (NCCN guidelines): colon cancer," 2024.

[8] A. Costas-Chavarri, S. Temin, and M. A. Shah, "Treatment of patients with early-stage colorectal cancer: ASCO resource-stratified guideline summary," *Journal of Oncology Practice*, vol. 15, no. 5, pp. 290–292, 2019.

[9] A. Lugli, R. Kirsch, Y. Ajioka, F. Bosman, G. Cathomas, H. Dawson, H. El Zimaity, J.-F. Fléjou, T. P. Hansen, A. Hartmann, *et al.*, "Recommendations for reporting tumor budding in colorectal cancer based on the International Tumor Budding Consensus Conference (ITBCC) 2016," *Modern pathology*, vol. 30, no. 9, pp. 1299–1311, 2017.

[10] V. W. K. Lee and K. F. Chan, "Tumor budding and poorly-differentiated cluster in prognostication in Stage II colon cancer," *Pathology-Research and Practice*, vol. 214, no. 3, pp. 402–407, 2018.

[11] A. Romiti, M. Roberto, P. Marchetti, A. Di Cerbo, R. Falcone, G. Campisi, M. Ferri, G. Balducci, G. Ramacciato, L. Ruco, *et al.*, "Study of histopathologic parameters to define the prognosis of stage ii colon cancer," *International Journal of Colorectal Disease*, vol. 34, pp. 905–913, 2019.

[12] V. H. Koelzer, I. Zlobec, and A. Lugli, "Tumor budding in colorectal cancer — ready for diagnostic practice?," *Human pathology*, vol. 47, no. 1, pp. 4–19, 2016.

[13] S. L. Bosch, S. Teerenstra, J. H. de Wilt, C. Cunningham, and I. D. Nagtegaal, "Predicting lymph node metastasis in pt1 colorectal cancer: a systematic

review of risk factors providing rationale for therapy decisions," *Endoscopy*, vol. 45, no. 10, pp. 827–841, 2013.

[14] R. K. Pai, Y. Chen, M. A. Jakubowski, B. L. Shadrach, T. P. Plesec, and R. K. Pai, "Colorectal carcinomas with submucosal invasion (pT1): analysis of histopathological and molecular factors predicting lymph node metastasis," *Modern Pathology*, vol. 30, no. 1, pp. 113–122, 2017.

[15] I. S. Brown, M. L. Bettington, A. Bettington, G. Miller, and C. Rosty, "Adverse histological features in malignant colorectal polyps: a contemporary series of 239 cases," *Journal of clinical pathology*, vol. 69, no. 4, pp. 292–299, 2016.

[16] Y. Backes, S. G. Elias, J. N. Groen, M. P. Schwartz, F. H. Wolfhagen, J. M. Geesing, F. Ter Borg, J. van Bergeijk, B. W. Spanier, W. H. d. V. tot Nederveen, *et al.*, "Histologic factors associated with need for surgery in patients with pedunculated T1 colorectal carcinomas," *Gastroenterology*, vol. 154, no. 6, pp. 1647–1659, 2018.

[17] J. Bokhorst, A. Blank, A. Lugli, I. Zlobec, H. Dawson, M. Vieth, L. Rijstenberg, S. Brockmoeller, M. Urbanowicz, J. Flejou, *et al.*, "Assessment of individual tumor buds using keratin immunohistochemistry: moderate interobserver agreement suggests a role for machine learning," *Modern pathology*, vol. 33, no. 5, pp. 825–833, 2020.

[18] J.-M. Bokhorst, I. D. Nagtegaal, I. Zlobec, H. Dawson, K. Sheahan, F. Simmer, R. Kirsch, M. Vieth, A. Lugli, J. van der Laak, *et al.*, "Semi-supervised learning to automate tumor bud detection in cytokeratin-stained whole-slide images of colorectal cancer," *Cancers*, vol. 15, no. 7, p. 2079, 2023.

[19] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[20] W. Ertel and N. T. Black, *Grundkurs Künstliche Intelligenz*, vol. 4. Springer, 2016.

[21] L. Kiehl, S. Kuntz, J. Höhn, T. Jutzi, E. Krieghoff-Henning, J. N. Kather, T. Holland-Letz, A. Kopp-Schneider, J. Chang-Claude, A. Brobeil, *et al.*, "Deep learning can predict lymph node status directly from histology in colorectal cancer," *European Journal of Cancer*, vol. 157, pp. 464–473, 2021.

[22] M. Runz, D. Rusche, S. Schmidt, M. R. Weihrauch, J. Hesser, and C.-A. Weis, "Normalization of HE-stained histological images using cycle consistent generative adversarial networks," *Diagnostic Pathology*, vol. 16, no. 1, pp. 1–10, 2021.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, eds.), vol. 9351, pp. 234–241, Cham: Springer International Publishing, 2015.

[24] P. Bankhead, M. B. Loughrey, J. A. Fernández, Y. Dombrowski, D. G. McArt, P. D. Dunne, S. McQuaid, R. T. Gray, L. J. Murray, H. G. Coleman, J. A. James, M. Salto-Tellez, and P. W. Hamilton, "QuPath: Open source software for digital pathology image analysis," *Scientific Reports*, vol. 7, p. 16878, Dec. 2017.

[25] M. Runz, "qupath scripting with Groovy for histological image processing." `https://github.com/m4ln/qupath-scripts`. (abgerufen am 11.04.2023).

[26] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and

S. Chintala, "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[27] A. Janowczyk, "Digital Pathology Segmentation using Pytorch + U-Net." `https://github.com/choosehappy/PytorchDigitalPathology/tree/master/segmentation_epistroma_unet`, 2021. (abgerufen am 11.04.2023).

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241, Springer, 2015.

[29] A. Janowczyk, "Digital Pathology Segmentation using Pytorch + Unet." `http://www.andrewjanowczyk.com/pytorch-unet-for-digital-pathology-segmentation/`, 2018. (abgerufen am 08.03.2023).

[30] H. Lamba, "Understanding semantic segmentation with UNET," *Towards Data Science*, pp. 1–28, 2019.

[31] R. Olaf, "U-NET: Convolutional Networks for Biomedical Image Segmentation." `https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/`. (abgerufen am 11.04.2023).

[32] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[33] S. Kato and K. Hotta, "Adaptive t-vMF dice loss: An effective expansion of dice loss for medical image segmentation," *Computers in Biology and Medicine*, vol. 168, p. 107695, 2024.

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

[35] "ReduceLROnPlateau — PyTorch 1.13 documentation." `https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html`. (abgerufen am 08.03.2023).

[36] D. Rusche, "Segmentation and classification of HE-stained colorectal carcinoma tissue." `https://github.com/cpheidelberg/proj_buddingCRC-MIL-pytorch`. (abgerufen am 11.03.2023).

[37] D. Rusche, "U-Net based multiclass segmentation." `https://github.com/cpheidelberg/proj_buddingCRC-MIL-pytorch/tree/main/Segmentation`. (abgerufen am 18.03.2024).

[38] C. Wittekind, *TNM: Klassifikation maligner Tumoren*. John Wiley & Sons, 2016.

[39] K. Satoh, S. Nimura, M. Aoki, M. Hamasaki, K. Koga, H. Iwasaki, Y. Yamashita, H. Kataoka, and K. Nabeshima, "Tumor budding in colorectal carcinoma assessed by cytokeratin immunostaining and budding areas: possible involvement of c-met," *Cancer Science*, vol. 105, no. 11, pp. 1487–1495, 2014.

[40] A.-H. Benson, Venook, Azad, *et al.*, "NCCN clinical practice guidelines in oncology (NCCN guidelines): colon cancer," 2022.

[41] S. J. Lee, A. Kim, Y. K. Kim, W. Y. Park, H. S. Kim, H.-J. Jo, N. Oh, G. Am Song, *et al.*, "The significance of tumor budding in T1 colorectal

carcinoma: the most reliable predictor of lymph node metastasis especially in endoscopically resected T1 colorectal carcinoma," *Human Pathology*, vol. 78, pp. 8–17, 2018.

[42] U. Kellner, S. O. Frahm, C. Mawrin, M. Krams, and S. Schüller, *Kurzlehrbuch Pathologie*. Georg Thieme Verlag, 2019.

[43] J. Jiang and S. N. Hart, "WSITools." `https://github.com/smujiang/WSITools`. (abgerufen am 13.10.2022).

[44] A. Goode, B. Gilbert, and J. Harkes, "OpenSlide." `https://openslide.org/`. (abgerufen am 13.10.2022).

[45] C.-A. Weis, "Assessment of glomerular morphological patterns by deep learning." `https://github.com/catweis/Assessment-of-glomerular-morphological-patterns-by-deep-learning`, 2021. (abgerufen am 11.04.2023).

[46] Torch-Contributors, "torchvision.models." `https://pytorch.org/vision/0.8/models.html`, 2017. (abgerufen am 16.04.2023).

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

[48] I. Pointer, "Class Activation Mapping In PyTorch." `http://snappishproductions.com/blog/2018/01/03/class-activation-mapping-in-pytorch.html`. (abgerufen am 13.10.2022).

[49] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, pp. 1–13, 2020.

[50] N. Chinchor, "Proceedings of the 4th Conference on Message Understanding," 1992.

[51] M. Gusarova, *Introduction to Data Science with python: A complete guide to learn data science and machine learning step by step based on a Fintech end-to-end project*. Independently published, 2023.

[52] D. G. Altman, *Practical Statistics for Medical Research*. London: Chapman & Hall / CRC, 1991.

[53] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[54] V. Trevisan, "Multiclass classification evaluation with ROC Curves and ROC AUC." `https://towardsdatascience.com/multiclass-classification-evaluation-with-roc-curves-and-roc-auc-294fd4617e3a`, 2022. (abgerufen am 13.10.2022).

[55] V. Trevisan and F. Bottoni, "ROC Curve and ROC AUC." `https://github.com/vinyluis/Articles/tree/main/ROCCurveandROCAUC`. (abgerufen am 13.10.2022).

[56] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168, p. 022022, IOP Publishing, 2019.

[57] D. Rusche, N. Englert, M. Runz, S. Hetjens, C. Langner, T. Gaiser, and C.-A. Weis, "Unraveling a Histopathological Needle-in-Haystack Problem: Exploring the Challenges of Detecting Tumor Budding in Colorectal Carcinoma Histology," *Applied Sciences*, vol. 14, no. 2, p. 949, 2024.

[58] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Medical image analysis*, vol. 65, p. 101759, 2020.

[59] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[60] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *International conference on machine learning*, pp. 312–321, PMLR, 2019.

[61] R. Therrien and S. Doyle, "Role of training data variability on classifier performance and generalizability," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, pp. 58–70, SPIE, 2018.

[62] H. Kawachi, Y. Eishi, H. Ueno, T. Nemoto, T. Fujimori, A. Iwashita, Y. Ajioka, A. Ochiai, S. Ishiguro, T. Shimoda, *et al.*, "A three-tier classification system based on the depth of submucosal invasion and budding/sprouting can improve the treatment strategy for T1 colorectal cancer: a retrospective multicenter study," *Modern Pathology*, vol. 28, no. 6, pp. 872–879, 2015.

[63] B. Mitrovic, D. F. Schaeffer, R. H. Riddell, and R. Kirsch, "Tumor budding in colorectal carcinoma: time to take notice," *Modern Pathology*, vol. 25, no. 10, pp. 1315–1325, 2012.

[64] L. Harbaum, M. J. Pollheimer, P. Kornprat, R. A. Lindtner, C. Bokemeyer, and C. Langner, "Peritumoral eosinophils predict recurrence in colorectal cancer," *Modern Pathology*, vol. 28, no. 3, pp. 403–413, 2015.

[65] A. Lugli, I. Zlobec, M. D. Berger, R. Kirsch, and I. D. Nagtegaal, "Tumour budding in solid cancers," *Nature Reviews Clinical Oncology*, vol. 18, no. 2, pp. 101–115, 2021.

[66] Z. Li, W. Zhao, F. Shi, L. Qi, X. Xie, Y. Wei, Z. Ding, Y. Gao, S. Wu, J. Liu, *et al.*, "A novel multiple instance learning framework for COVID-19 severity assessment via data augmentation and self-supervised learning," *Medical Image Analysis*, vol. 69, p. 101978, 2021.

[67] G. Jaume, P. Pati, V. Anklin, A. Foncubierta, and M. Gabrani, "Histocartography: A toolkit for graph analytics in digital pathology," in *MICCAI Workshop on Computational Pathology*, pp. 117–128, PMLR, 2021.

[68] C.-A. Weis, J. N. Kather, S. Melchers, H. Al-Ahmdi, M. J. Pollheimer, C. Langner, and T. Gaiser, "Automatic evaluation of tumor budding in immunohistochemically stained colorectal carcinomas and correlation to clinical outcome," *Diagnostic pathology*, vol. 13, pp. 1–12, 2018.

[69] S. Brockmoeller, A. Echle, N. Ghaffari Laleh, S. Eiholm, M. L. Malmstrøm, T. Plato Kuhlmann, K. Levic, H. I. Grabsch, N. P. West, O. L. Saldanha, *et al.*, "Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer," *The Journal of pathology*, vol. 256, no. 3, pp. 269–281, 2022.

[70] M. Zhao, S. Yao, Z. Li, L. Wu, Z. Xu, X. Pan, H. Lin, Y. Xu, S. Yang, S. Zhang, *et al.*, "The crohn's-like lymphoid reaction density: a new artificial intelligence quantified prognostic immune index in colon cancer," *Cancer Immunology, Immunotherapy*, pp. 1–11, 2022.

[71] C. Bian, Y. Wang, Z. Lu, Y. An, H. Wang, L. Kong, Y. Du, and J. Tian, "ImmunoAIzer: a deep learning-based computational framework to characterize

cell distribution and gene mutation in tumor microenvironment," *Cancers*, vol. 13, no. 7, p. 1659, 2021.

[72] M. S. Kwak, H. H. Lee, J. M. Yang, J. M. Cha, J. W. Jeon, J. Y. Yoon, and H. I. Kim, "Deep convolutional neural network-based lymph node metastasis prediction for colon cancer using histopathological images," *Frontiers in Oncology*, vol. 10, p. 619803, 2021.

Anhang

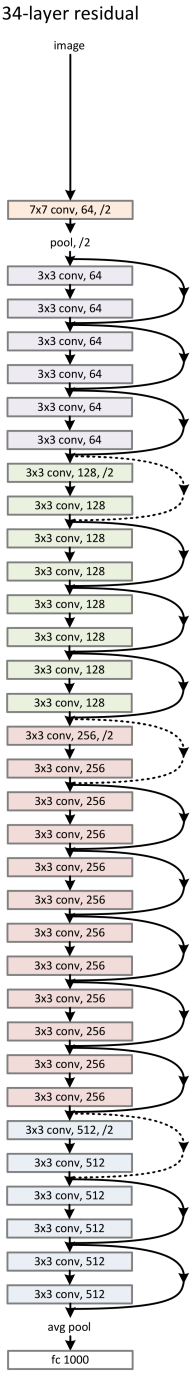
34-layer residual
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000

Abbildung 5.1 ***(vorherige Seite)*****: Schematische Darstellung eines ResNet mit 34 Ebenen.**
Das Eingangsbild durchläuft zuerst eine Faltung mit einem Filter der Größe 7×7. Anschließend folgt eine Reihe an Residual Blocks mit Faltungslayern (Rechtecke) und Identitätsfunktion (gebogene Pfeile). Nach einer bestimmten Anzahl an Residual Blocks kommt es zu einer Änderung der Dimensionen des Ergebnisses, so dass die Identitätsfunktion neu skaliert werden muss (gestrichelte Pfeile). Ist die gewünschte Tiefe des Netzwerks mittels ausreichend Residual Blocks erreicht, erfolgt ein Average Pooling. Hierbei wird für eine Gruppe Neuronen deren Mittelwert an die Feature Map weitergeleitet. Die erhalten Feature Map wird mit einem Fully Connected Layer mit 1000 Neuronen (fc. 1000) zusammengefasst [47].

Lebenslauf

Personalien

Name und Vorname:	Rusche, Daniel
Geburtsdatum:	04. Oktober 1994
Geburtsort:	Nürnberg

Schulischer Werdegang

2001 - 2006:	Siedlerschule Nürnberg
2006 - 2013:	Willstätter-Gymnasium Nürnberg
28.06.2013:	Abitur

Universitärer Werdegang

2013 - 2016:	Bachelorstudium Physik an der Friedrich-Alexander-Universität Erlangen-Nürnberg
28.09.2016:	Bachelor: Experimentelle Untersuchung und Simulation des Signals von Knochenfissuren im Röntgen Phasenkontrast-Hochenergieaufbau
2016 - 2021:	Medizinstudium an der Universität Heidelberg, Medizinische Fakultät Mannheim
05.09.2018:	Erster Abschnitt der Ärztlichen Prüfung
07.10.2021:	Zweiter Abschnitt der Ärztlichen Prüfung
2021 - 2023:	Medizinstudium an der Universität Regensburg
06.12.2022:	Dritter Abschnitt der Ärztlichen Prüfung
seit 2023:	Promotionsstudium an der Universität Heidelberg, Medizinische Fakultät Mannheim

Danksagung

Zuallererst möchte ich mich bei meinem Doktorvater Prof. Dr. med Gaiser und meinem Betreuer PD Dr. med Weis für die Überlassung des Themas bedanken. Ich bin vor allem Herrn PD Dr. med. Weis sehr dankbar für seine Geduld und seine unermüdliche Unterstützung während des gesamten Prozesses.

Auch meinen Eltern möchte ich herzlich danken. Ohne ihre ausdauernde Unterstützung in jeder Hinsicht wäre diese Arbeit nicht möglich gewesen.

Darüber hinaus möchte ich mich bei allen bedanken, die meine Arbeit gelesen und kommentiert haben. Ihre wertvollen Rückmeldungen haben dazu beigetragen, dass ich meine Arbeit verbessern konnte und ich bin ihnen dafür sehr dankbar.