Aus dem Institut für Pathologie
der Universitätsmedizin der Johannes Gutenberg-Universität Mainz

Maschinelles Lernen zur Diagnostik von Siegelringzellkarzinomen des oberen Gastrointestinaltraktes

Inauguraldissertation
zur Erlangung des Doktorgrades der
Medizin
der Universitätsmedizin
der Johannes Gutenberg-Universität Mainz

vorgelegt von

Franziska Hartmann
aus Hermeskeil

Mainz, 2023

Wissenschaftlicher Vorstand:

1. Gutachter:

2. Gutachter:

Tag der Promotion: 01.12.2023

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

Abb. ..... Abbildung
AEG ..... Adenokarzinome des ösophagogastralen Überganges
CDH1 ..... Cadherin 1
CNN ..... Convolutional Neural Networks
ConvNet ..... Convolutional Neural Networks
DALY ..... disability-adjusted life years
EU ..... Europäische Union
GERD ..... gastroösophageale Refluxkrankheit
H.E. ..... Hämatoxylin und Eosin
H.p. ..... Helicobacter pylori
HNPCC ..... hereditäres nicht-polypöses Kolonkarzinom
ICD .. International Statistical Classification of Diseases and Related Health Problems
ÖGD ..... Ösophagogastroduodenoskopie
PAH ..... polycyclische aromatische Kohlenwasserstoffe
PDL1 ..... Programmed cell death 1 ligand 1
R ..... Tumorrest nach Resektion
SRCC ..... Siegelringzellkarzinome
TCGA ..... The Cancer Genome Atlas
TNM ..... Tumor, Nodus, Metastasen
UICC ..... Union for International Cancer Control
USA ..... United States of America
WHO ..... world health organization

Tabellenverzeichnis

1 Einleitung

Magenkarzinome sind häufig vorkommende Malignome mit hoher Mortalität, deren Therapie und Prognose von einer prompten und korrekten Diagnose abhängt. Insbesondere diffus infiltrierende Karzinome wie das Siegelringzellkarzinom stellen eine große diagnostische Herausforderung dar und zeichnen sich durch ein aggressives tumorbiologisches Verhalten aus. Schlüsseldisziplin für Diagnostik und Therapieentscheidungen ist die Pathologie. Während die Arbeitsbelastung bei steigenden Fallzahlen in Zeiten zeitaufwändiger, individualisierter Diagnosen und Therapien zunimmt, sinkt zugleich die Zahl der Nachwuchspathologen. Abhilfe könnte eine Digitalisierung der noch weitgehend analog arbeitenden Pathologie schaffen. Damit würde mittels maschinellem Lernen, einem Teilgebiet der künstlichen Intelligenz, auch eine automatisierte Analyse von Gewebeschnitten ermöglicht.
Ziel dieser Arbeit ist die Evaluation des möglichen Einsatzes von maschinellem Lernen für die Gewebeklassifizierung bei Magen- beziehungsweise Siegelringzellkarzinomen mit Unterscheidung der Gewebetypen Epithel, Fett, Immunzellen, Muskel und Karzinom. Die Ergebnisse des Modells sollen in einer klinisch nutzbaren Form dargestellt werden. In einem zweiten Teil sollen mögliche Verwendungsmöglichkeiten eines solchen Klassifikationsmodelles recherchiert, der potenzielle Einfluss auf den klinischen Alltag evaluiert und mögliche - einer Implementation noch im Wege stehende - Hürden diskutiert werden.

Zur besseren Lesbarkeit wird in dieser Arbeit das generische Maskulinum verwendet. Die in dieser Arbeit verwendeten Personenbezeichnungen und personenbezogenen Hauptwörter beziehen sich – sofern nicht anders kenntlich gemacht – auf alle Geschlechter.

2 Literaturdiskussion

2.1 Magenkarzinome

2.1.1 Inzidenz, Prävalenz, Mortalität und Entwicklungen

Jährlich wird 990.000 Menschen weltweit die Diagnose Magenkarzinom gestellt (1), insgesamt handelt es sich bei acht Prozent aller bösartigen Tumore um Magenkarzinome (2). Im Vergleich werden die höchsten Inzidenzen in Ostasien, Osteuropa und Südamerika festgestellt, die niedrigsten in Nordamerika und Afrika (3). Gründe hierfür könnten in den unterschiedlichen Ernährungsgewohnheiten der Nationen liegen, vor allem bezüglich des Risikofaktors des ausgeprägten Konsums salziger und geräucherter Nahrung in Asien und Lateinamerika (4) (vgl. Abschnitt 2.1.2) und in der geographisch unterschiedlichen Verbreitung des Risikofaktors Helicobacter pylori (H.p.) (vgl. Abschnitt 2.1.2) mit erhöhter Prävalenz in Entwicklungsländern (4). Bezüglich der Inzidenz zeigt Deutschland für beide Geschlechter leicht höhere Zahlen als die Mittelwerte für die Europäische Union (EU): 2013 wurden etwa 15.600 Fälle, 9.300 Männer und 6.300 Frauen neu diagnostiziert (4). Dies entspricht einer Halbierung der Inzidenz im Vergleich zu 40 Jahre zuvor, was dem allgemein rückläufigen Verlauf der Inzidenz von Magenkarzinomen weltweit entspricht (4). Eine Ausnahme des allgemeinen Trends stellen die Magenkarzinome der Kardia, des Mageneingangs, dar, deren Inzidenzen in westlichen Ländern als stabil oder vor allem beim männlichen Geschlecht sogar als steigend beschrieben werden (5). Die Ursachen dafür werden in der unterschiedlichen Relevanz der Risikofaktoren vermutet. H.p., dessen Prävalenz seit der Verbesserung der hygienischen Standards und der Einführung der Eradikationstherapie Ende der 1990er gesunken ist, stellt eher für die abnehmende Zahl der Karzinome außerhalb der Kardia einen Risikofaktor dar. Dagegen sind Übergewicht und gastroösophagealer Reflux, die beide in westlichen Ländern zunehmen, Risikofaktoren für Kardiakarzinome (5). Möglicherweise sind auch die durch die Verbreitung von Kühlschränken veränderten Ernährungsgewohnheiten von Bedeutung, womit Einfluss auf gleich zwei bekannte Risikofaktoren genommen werden konnte: Herkömmliche Konservierungsmethoden wie Räuchern und Pökeln wurden abgelöst, ganzjähriger Zugang zu frischem Obst und Gemüse geschaffen (4) (vgl. Abschnitt 2.1.2).

Die 5-Jahresprävalenz in Deutschland wurde 2013 mit 33.200 Fällen beschrieben, das mittlere Erkrankungsalter lag dabei bei ungefähr 72 Jahren (4).

738.000 Menschen versterben weltweit jährlich am Magenkarzinom, was diese Entität zum zweithäufigsten Grund der krebsbedingten Mortalität macht (1). In Deutschland

starben im Jahr 2013 rund 9.600 Menschen an Magenkarzinomen, was ungefähr dem Durchschnitt der EU entspricht, im Vergleich zur Mortalität vor 40 Jahren jedoch einen Rückgang auf etwa ein Drittel bedeutet (4). Seit den 1970ern sind weltweit erhebliche Verbesserungen in der relativen 5-Jahresüberlebensrate für Magenkarzinome festzustellen: So beispielsweise von 15% in 1975 auf 29% in 2009 in den USA (6). Dennoch bleiben die Überlebensraten schlecht: Aufgrund fehlender Frühsymptome (vgl. Abschnitt 2.1.3) werden über 50% der Diagnosen in fortgeschrittenem Stadium ohne kurative Möglichkeit gestellt (7). Während in frühen Stadien ein 5-Jahresüberleben von 90-95% beschrieben wird, sinkt dieses mit Fortschreiten der Erkrankung auf einen Bereich von 5-30% (7, 8). Für 2013 wurde in Deutschland eine Angabe von durchschnittlich 32% für die 5-Jahresüberlebensrate gemacht (4). Weiterhin ist das Magenkarzinom auch unter den Karzinomformen mit der höchsten Belastung der Lebensqualität, gemessen an den DALY (disability-adjusted life years), den sogenannten behinderungsbereinigten Lebensjahren (9).

2.1.2 **Risikofaktoren**

Die Tumorentstehung des Magens ist multifaktoriell und es spielen sowohl Umwelt- als auch genetische Faktoren ätiologisch eine Rolle (7). Eine Infektion mit Helicobacter pylori gilt als wichtigster Risikofaktor für Magenkarzinome (10). Es wird geschätzt, dass H.p. 65% bis 80% aller Magenkarzinomfälle verursacht (11, 12). Bezüglich des zugrundeliegenden Mechanismus werden verschiedene Wege beobachtet: indirekte Interaktion von H.p. mit Epithelzellen des Magens durch Verursachung von Entzündung und direkte Interaktion mit den Epithelzellen, wobei H.p. die Funktion der Epithelzellen durch bakterielle Agenzien beeinflussen kann (13).

Einen weiteren Risikofaktor stellt fortgeschrittenes Alter dar; im Zeitraum von 2005 bis 2009 lag das mittlere Diagnosealter bei 70 Jahren (14). Auch das männliche Geschlecht erhöht das Risiko, die Inzidenzen sind bei Männern zwei- bis dreifach höher als bei Frauen, vor allem bezüglich der Lokalisation im Bereich der Kardia (1, 5). Die Ursachen sind nicht völlig aufgeklärt, möglicherweise sind unterschiedliches Gesundheitsverhalten oder protektive Eigenschaften von Östrogenen beteiligt (5). Rauchen wird als weiterer Risikofaktor gehandhabt (15); das Risiko soll bei Männern um 60%, bei Frauen um 20% erhöht sein (16). Zudem wird eine Assoziation der Ethnie mit Magenkarzinomen beschrieben, die am ehesten umweltbedingten Faktoren und verschiedenen kulturellen Lebensarten zuzuschreiben ist (5). Das Risiko ist (in den USA) am höchsten für Asiaten, gefolgt von Schwarzen, Hispanics und zuletzt Weißen.

Karzinome der Kardia treten bei Weißen hingegen gehäuft auf (17). (Die Formulierungen dienen in diesem Kontext der Abgrenzung ethnischer Gruppen in Bezug auf ihr biologisches Risikoprofil, es wird keine Aussage zu gesellschaftspolitischen Zugehörigkeiten getroffen.) Ein Beispiel ist der unterschiedliche Konsum salziger und geräucherter Nahrung, was wie bereits beschrieben als weiterer Risikofaktor gilt. Grund sind Benzopyrene und andere polycyclische aromatische Kohlenwasserstoffe (PAH), die sich in geräucherter Nahrung bilden (18). Bezüglich des Ernährungsverhaltens wurde weiterhin der Effekt von Obst und Gemüse untersucht. Es besteht wahrscheinlich ein protektiver Effekt dieser Nahrungsmittel: Sie beinhalten Vitamin C, Folsäure, Carotinoide und Phytochemikalien, die die Karzinogenese bremsen können (5). Eine unzureichende Aufnahme von Obst und Gemüse wird demnach ebenfalls als Risikofaktor für Magenkarzinome beschrieben (5).

Es ist bekannt, dass ein niedriger sozioökonomischer Status allgemein mit einem höheren Risiko für totale und ursachenspezifische Mortalität assoziiert ist, einschließlich dem Tod an den meisten Krebsarten (19, 20). Dies schließt auch das Magenkarzinom ein, wobei höhere Raten von H.p. Infektion, das Ernährungs- und allgemeine Gesundheitsverhalten Gründe für diesen Sachverhalt sein könnten (5). Des Weiteren werden eine positive Familienanamnese, hereditäre Erkrankungen wie beispielsweise das HNPCC (hereditäres nicht-polypöses Kolonkarzinom) oder das Peutz-Jeghers-Syndrom, die Blutgruppe A, geringe körperliche Aktivität, vorangegangene Bestrahlung des Abdomenbereiches, Alkoholkonsum sowie niedrige Aufnahme von Ballaststoffen in der Literatur als Risikofaktoren aufgeführt (5, 7).

Gesondert für Karzinome des Bereiches des gastroösophagealen Überganges ist der Risikofaktor Übergewicht zu nennen, der selbst als begünstigender Faktor für GERD (gastroösophageale Refluxkrankheit) gilt. Reflux erhöht ebenfalls das Risiko für Tumoren im Übergangsbereich von Ösophagus zu Magen (5).

Die Einnahme von Aspirin und Statinen können als protektive Faktoren wirken, wobei eine Risikoreduktion durch Statine um bis zu 15% beschrieben wird (21, 22).

2.1.3 **Symptome, Diagnose, Metastasierung und Therapie**

Frühe Karzinome des Magens sind meist symptomarm bis symptomlos (23, 24). Sie machen sich oftmals erst in fortgeschrittenen Stadien bemerkbar, wobei Symptome wie Dysphagie, rezidivierendes Erbrechen, Inappetenz, neu aufgetretene Abneigung gegen Fleisch oder unklarer Gewichtsverlust auftreten können (23-25). Auch

Anzeichen einer oberen gastrointestinalen Blutung mit eventuell einhergehendem Teerstuhl und unklare chronische Eisenmangelanämien sollten an ein Magenkarzinom denken lassen (7, 23, 25). Es können zudem verschiedene paraneoplastische Syndrome auftreten, wobei hier vor allem kutane Erscheinungen beobachtet werden (26). Im metastasierten Stadium können Symptome wie ein tastbarer Oberbauchtumor, Hepatomegalie und Aszites oder eine tastbare Virchow-Drüse (Lymphknoten links supraklavikulär) auffallen (7). Die Metastasierung erfolgt beim Magenkarzinom früh auf verschiedenen Wegen: 70% der Magenkarzinome sind bei Diagnosestellung bereits lymphogen metastasiert (7). Außerdem erfolgt eine hämatogene Ausbreitung zu Leber, Lunge, Knochen und Gehirn (7). Ebenso breitet sich der Tumor kontinuierlich in seine unmittelbare Umgebung – Ösophagus, Duodenum, Kolon, Pankreas – und auch auf das Bauchfell aus (7). Eine Besonderheit sind Abtropfmetastasen in den Ovarien oder im Douglas-Raum, die als Krukenberg-Tumore bezeichnet werden (7).

Bei Patienten mit mindestens einem der genannten Alarmsymptome sollte frühzeitig eine Ösophagogastroduodenoskopie (ÖGD), eine endoskopische Untersuchung mit Biopsiegewinnung erfolgen (24). Bei histologischer Diagnosestellung einer Dysplasie sollten die Biopsien durch mindestens zwei Pathologen beurteilt worden sein (27).

Für das Staging (onkologische Stadienbestimmung) werden Sonographie, Computertomographie des Thorax und des Abdomens inklusive Becken angewendet (27). Bei kurativer Intention kann auch die Endosonographie Bestandteil des Stagings sein, bei lokal fortgeschrittenen Tumoren eine Staging-Laparoskopie vor neoadjuvanter Therapie (27). Die histologische Klassifikation und anschließende Stadieneinteilung als Richtungsweiser für die zu wählende Therapie erfolgt nach WHO-Angaben, anhand der TNM-Klassifikation und nach UICC-Stadieneinteilung (25) (vgl. Abschnitt 2.1.4).

Die Entscheidung über eine individuell patientenangepasste Therapie bei positiver Diagnose sollte im Rahmen eines multidisziplinären Tumorboards getroffen werden, wobei die Anwesenheit von Viszeralchirurgie, Onkologie, Radiologie, Gastroenterologie, Pathologie empfohlen wird (24, 27). Eine Behandlung des Magenkarzinoms erfolgt generell Stadien-adaptiert (24). So kann die Therapie von einfacher endoskopischer Resektion im Frühstadium über chirurgische Tumorresektion einschließlich umgebender Lymphknoten bis zu einer Eskalation mit Anwendung einer perioperativen Chemotherapie (Platin, Docetaxel, 5-Fluorouracil) zusätzlich zur subtotalen oder totalen Gastrektomie reichen (24, 27). Bei nicht

resektablen Tumoren verbleiben medikamentöse Therapieversuche oder palliative Betreuung (24).

2.1.4 **Subtypen und Klassifikationen**

Circa 90% der Magentumoren sind Adenokarzinome, die aus den Drüsen der Schleimhaut des Magens entstehen (5). Darüber hinaus gibt es andere Tumorentitäten wie Lymphome (mucosa associated lymphoid tissue lymphomas), ausgehend vom lymphatischen System oder Leiomyosarkome, die aus der umgebenden glatten Muskulatur ihren Ursprung nehmen (5).

Es bestehen verschiedene Klassifikationssysteme für die Unterteilung von Magenkarzinomen. Detaillierte Ausführungen der wichtigsten Klassifikationen befinden sich im Anhang. Adenokarzinome des ösophagogastralen Überganges (AEG) unterliegen einer eigenen anatomisch basierten Klassifikation (7). Nach dieser Siewert-Klassifikation zählen nur AEG III-Tumoren, welche über 2cm distal der Kardia liegen, als Magenkarzinome, AEG I und II werden den Ösophaguskarzinomen zugeordnet (7). Die TNM Klassifikation, die häufig für die Klassifikation von Malignomen herangezogen wird (28), wurde auch für das Magenkarzinom adaptiert (25). Allgemein werden das Verhalten und die Ausdehnung des Primärtumors, das Fehlen oder Vorhandensein von regionären Lymphknotenmetastasen sowie von Fernmetastasen beurteilt (24). Die Einteilung erlaubt prognostische Aussagen und nimmt Einfluss auf die Therapieentscheidung (25). Die Einteilung der UICC legt die TNM-Klassifikation für die Einteilung verschiedener Stadien mit therapeutischen und prognostischen Intentionen zugrunde (8). Die Laurén Klassifikation nimmt eine Einteilung nach histopathologischen Gesichtspunkten vor (24). Es wird unterschieden zwischen einem intestinalen Typ (polypöses, drüsig differenziertes, klar begrenztes Wachstum), der mit circa 54% den Großteil der Fälle ausmacht, sowie einem diffusen Typ (infiltratives Wachstum mit diffuser, schlecht begrenzter Ausbreitung in der Magendwand) mit einem Anteil von ungefähr 32%, der zu früherer Metastasierung neigt und einen größeren Sicherheitsabstand in der Resektion fordert (7, 24). Ein unbestimmbarer Mischtyp wird bei 15% angegeben (29). Die Laurén-Klassifikation gibt so Anhaltspunkte für die Wahl des Resektionsausmaßes (7). Nach der WHO erfolgt die Klassifikation anhand des prädominant vorliegenden histologischen Muster, da sich oftmals verschiedene Formen parallel manifestieren (30). Unterschieden werden die mit 90% am häufigsten vorliegenden Adenokarzinome, die ein papilläres, tubuläres, muzinöses oder siegelringzelliges Wachstum zeigen können von

adenosquamösen, squamösen (Plattenepithelkarzinomen), kleinzelligen und undifferenzierten Karzinomen (7, 24). Auf Basis der Daten des ‚The Cancer Genome Atlas' (TCGA), hat Bass et al. ein Klassifikationssystem vorgeschlagen, das auf der Untersuchung molekularer Subtypen nach Unterschieden in Genom, Transkriptom, Epigenom und Proteom basiert (24). Beschrieben werden chromosomal instabile, Eppstein-Barr-Virus-assoziierte, mikrosatelliteninstabile und genomisch stabile Typen, wobei diese Unterscheidung bis heute noch keine Auswirkungen auf die Behandlung hat (31). Darüber hinaus haben Ming und Nakamura weitere bekannte Klassifikationssysteme etabliert (29, 32, 33). Insbesondere die diffus-infiltrierenden Karzinome zeichnen sich durch ein aggressives tumorbiologisches Verhalten aus (34) und stellen darüber hinaus eine große diagnostische Herausforderung dar. So genannte Siegelringzellkarzinome machen den überwiegenden Anteil dieser Unterform aus (2).

2.2 Siegelringzellkarzinome (SRCC)

Seit der Veröffentlichung der WHO Klassifikation 1990 stellt das SRCC eine eigenständige histologische Entität dar. Zuvor wurde es als „diffuser Typ" nach Laurén (29), „infiltrativer Typ" nach Ming (32), „undifferenzierter Typ" nach Nakamura (35) und als „high grade" nach der UICC (36) eingestuft. Jedoch weisen dabei nicht alle als undifferenziert oder diffus klassifizierten Magenkarzinome auch wirklich Siegelringzellen auf. Nach der WHO sind SRCC definiert als wenig zusammenhängende Karzinome, die sich überwiegend aus Tumorzellen mit prominentem, zytoplasmatischem Muzin und einem halbmondförmigen, randständigen Zellkern zusammensetzen (37) (vgl. Abb. 1).

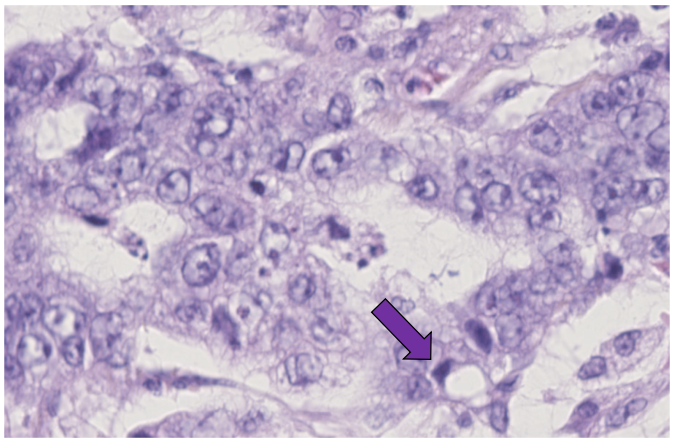


Abbildung 1: Histologie des Siegelringzellkarzinomes
Beispielhafte Darstellung eines Gewebeausschnittes eines SRCC in H.E. Färbung in 40facher Vergrößerung. Das charakteristische zytoplasmatische Muzin, das die Zellkerne an den Zellrand verdrängt, ist zu erkennen. Abbildung vom Autor aus dem annotierten Material.

2.2.1 **Inzidenz, Risikofaktoren und Prognose des SRCC**

Im Gegensatz zu der insgesamt rückläufigen Inzidenz von Magenkarzinomen allgemein, steigt die Zahl der Siegelringzellkarzinome des Magens weltweit (34). Studien beschreiben Anteile zwischen 35% bis 45% an den Magenkarzinomfällen bezogen auf Asien, die USA und Europa (38, 39). Die Inzidenz hat sich zwischen 1970 und 2000 bereits verzehnfacht (40). Gründe könnte man in Unterschieden bezüglich Risikofaktoren und epidemiologischen Faktoren im Vergleich zu anderen Magenkarzinomsubtypen sehen (34). So sank -wie zuvor beschrieben- die Inzidenz von Magenkarzinomen insgesamt seit der Einführung von Möglichkeiten zur Behandlung von Helicobacter pylori Infektionen (vgl. Abschnitt 2.1.1), die von SRCC jedoch nicht, was H.p. als mindestens fraglichen Risikofaktor identifiziert (34). Die Rolle weiterer Risikofaktoren wie Konsum eingesalzter Lebensmittel, Rauchen, Autoimmungastritis oder Adipositas ist aber noch nicht eingehend für den Subtyp des SRCC untersucht (34). Ein weiterer Unterschied zu Magenkarzinomen anderer Subtypen zeigt sich in der Betrachtung der Geschlechterverteilung: So hat sich beim SRCC eine relative Häufung von Frauen durch ein ungefähr ausgeglichenes Geschlechterverhältnis bestätigt, während bei Magenkarzinomen anderer Subtypen wie zuvor beschrieben vermehrt Männer betroffen sind (34) (vgl. Abschnitt 2.1.1). Auch das Alter der Patienten stellt einen Unterschied dar: SRCC präsentieren sich vermehrt in jüngeren Patienten, wobei das mittlere Erkrankungsalter zwischen 55 und 61 Jahren liegt (39, 41). Damit erkranken die Patienten im Durchschnitt sieben Jahre früher am SRCC Subtyp. Bezüglich der Risikoverteilung für verschiedene ethnische Gruppen legen verschiedene Berichte ebenfalls eine Häufung für Asiaten nahe (37, 39, 42). In Anbetracht von variierenden Ergebnissen in asiatischen Studien, bleibt dieser Aspekt aber bis dato unklar (41, 43-45).

Das SRCC kann in zwei Unterformen mit verschiedenen Implikationen aufgeteilt werden: frühe und fortgeschrittene Magenkarzinome (34). Der SRCC-Subtyp ist im Allgemeinen mit einem fortgeschrittenen Stadium assoziiert: Am häufigsten wird er in Stadium 4, T3/T4 und N2 Karzinomen diagnostiziert (34). Allerdings gibt es auch einige dazu widersprüchliche Angaben, die eine höhere Auftrittsrate in frühen Magenkarzinomen beschreiben (41). Einen großen Unterschied zwischen den frühen und fortgeschrittenen Stadien stellt die Prognose dar (34). Bei Betrachtung dieses Aspektes ist zu berücksichtigen, dass sich die Studienlage zwar einig über die schlechte Prognose des Magenkarzinoms des diffusen Typs nach Laurén (welches

die Untergruppe der SRCC mit umfasst) ist, die Prognose von Siegelringzellkarzinomen an sich jedoch immer noch kontrovers diskutiert wird (34). Für frühe Stadien – nach der Japanese Endoscopy Society definiert als die Submucosa nicht überschreitend – berichtet die Mehrzahl der Studien eine mindestens gleiche, überwiegend sogar bessere Prognose für das SRCC (41, 46-50) mit einer den anderen Subtypen des Magenkarzinoms ähnlichen Häufigkeit für Lymphknotenmetastasen (34). Das bessere Gesamtüberleben in den meisten Studien könnte nach Gronnier et al. auch durch das jüngere Alter der SRCC-Patienten mitbegründet sein (50). Die Prognose für fortgeschrittene Stadien ist unklar, wird jedoch im Allgemeinen als schlecht angesehen (34). Diese Ansicht kam durch retrospektive Studien auf, in denen SRCC-Patienten eine signifikant schlechtere 5-Jahresüberlebensrate zeigten als ‚Nicht-SRCC'-Patienten (51, 52). In anderen Studien konnte dies auch für das Gesamtüberleben bestätigt werden (41, 44, 53), während gleichzeitig wiederum andere Arbeiten keine signifikant schlechtere Prognose für Siegelringzellkarzinome feststellen konnten (47-49, 54-56). Auch wenn es insgesamt keinen eindeutigen Beleg für ein schlechteres Überleben gibt, verbinden die meisten Studien SRCC mit größerer Tumorausdehnung, aggressiverem Phänotyp, tieferer Invasion, ausgeprägter Lymphknotenmetastasierung und geringerer R0-Resektionsrate (51, 53), was eine schlechtere Prognose erklären könnte.

2.2.2 **Therapie, Chemosensitivität und Karzinogenese des SRCC**

Sowohl die Diagnose des Subtyps des SRCC als auch die Unterteilung in frühe und fortgeschrittene Tumore mit entsprechender Prognose nimmt Einfluss auf die Therapieentscheidung (34). Die in Abschnitt 2.1.3 beschriebenen Therapiestandards des Magenkarzinoms gelten auch für das SRCC, wobei die histologische Einordnung mit Neigung zur frühzeitigen Lymphknotenmetastasierung eine radikalere Resektion mit größerem Sicherheitsabstand fordert (7, 34). Im fortgeschrittenen Stadium ist die Peritonealkarzinose die häufigste Form der Metastasierung (57) des SRCC, welche meist erst während der chirurgischen Resektion entdeckt wird. Daher empfehlen einige Autoren hier eine routinemäßige laparoskopische Evaluation vor Behandlungsbeginn (34). Palliative Resektionen sollten aufgrund eines erhöhten postoperativen Mortalitätsrisikos nicht vorgenommen werden (58). Eine andere Kontroverse bezüglich des SRCC besteht in der Chemosensitivität dieses Subtypes, die gemeinhin als geringfügiger angesehen wird (34). Der Standard ist hier orientiert an nicht Subtyp-differenzierten Studien (vgl. Abschnitt 2.1.3). Für das SRCC in allen Stadien

erscheinen im Speziellen Taxan-basierte Therapien in Studien effizienter (34). Heutzutage wird das Konzept einer mehrstufigen Karzinogenese weitgehend als wahrscheinlich angesehen; es ist ein mehrschrittiger Prozess verschiedener Typen von Mutationen und epigenetischer Veränderungen in multiplen Genen, der letztlich zur Entwicklung von Malignität führt (59). Die zwei hauptsächlichen Veränderungen auf zellulärer Ebene für das SRCC sind - abweichend von anderen Subtypen des Magenkarzinoms - der Verlust der Zell-Zell-Adhäsionsmoleküle und die Akkumulation von Muzin in großen Vakuolen (34). So ist das SRCC schon in frühen Stadien der Karzinogenese (60) mit spezifischen Keimbahnmutationen im CDH1-Tumorsuppressorgen assoziiert, welches für ein epitheliales Zelladhäsionsprotein codiert (34). CDH1 Mutationen stellen unter der Vielzahl auftretender Mutationen die häufigste zu SRCC führende Veränderung dar (61). Die von anderen Subtypen abweichenden Mutationen könnten nicht nur die veränderte Chemosensitivität erklären, sondern auch verschiedene Angriffspunkte für zielgerichtete Krebstherapie, „targeted therapy“, bieten (34). Weiterführende Evidenz dazu ist noch ausstehend (62). Auch bezüglich der Immuntherapie stehen noch Studien aus; so wird PDL1 in ungefähr 23% der SRCC Fälle überexprimiert und eine dort ansetzende Antikörper-Therapie könnte eine vielversprechende Therapiemöglichkeit darstellen (63).
Insgesamt ist eine prompte und korrekte Diagnose dieses Subtypes des Magenkarzinoms äußerst entscheidend, hängen doch Überlegungen und Entscheidungen bezüglich der Therapie wie die Radikalität einer eventuell angestrebten Operation, die Wahl der Chemotherapeutika und zukünftig vielleicht der Einsatz zielgerichteter Krebstherapien oder Immuntherapien von dieser Entscheidung ab. Der Verbesserung der Prognose liegt demnach die Optimierung der Lösung der diagnostischen Herausforderung des SRCC zugrunde.

2.3 Die Rolle der Pathologie in der Karzinomdiagnostik

Die Pathologie übernimmt sowohl in der Diagnosestellung des SRCC als auch bei Therapieentscheidungen sowie prognostischen Aussagen eine entscheidende Rolle (64). Hierbei werden analoge Technologien wie Glasobjektträger, Lichtmikroskope und schriftliche Befunde genutzt. Die mikroskopische Analyse von Hämatoxylin und Eosin (H.E.) gefärbten Schnitten war die Basis für Krebsdiagnosen und deren ‚Grading‘ im letzten Jahrhundert (65). ‚Grading‘, aus dem Englischen für Einteilung, bezeichnet dabei eine durch Pathologen vorgenommene Beurteilung des Tumors bezüglich dessen Differenzierungsgrad, das heißt dessen mit zunehmender Malignität auch

zunehmenden Grad der Abweichung vom Normalgewebe, mit Folgen für Prognose und Therapie (66). Mit zunehmender Bedeutung molekularpathologischer Kriterien für Einteilung und Prognose müssen Pathologen heutzutage oftmals neben einer großen Anzahl von H.E. gefärbten Gewebeschnitten zusätzlich immunhistochemische Färbungen oder molekularbiologische Befunde einbeziehen, um zu einer vollständigen Einschätzung zu kommen (31). Diese Zusatzuntersuchungen werden auch für die Festlegung der patientenspezifischen Behandlungsmethoden benötigt, die je nach zugrundeliegenden molekularen Markern variieren können (67).
Während die Belastung bei steigenden Inzidenzen und zunehmender Komplexität von Krebsdiagnosen steigt, sinkt zugleich die Zahl der Nachwuchspathologen: Große Fachgesellschaften wie das ‚Royal College of Pathologists' in Großbritannien warnen bereits vor Versorgungsproblemen im Bereich der Pathologie (68).

2.3.1 **Digitalisierung in der Pathologie**

Abhilfe könnte hier eine Digitalisierung der Pathologie bringen: Hinweise mehren sich, dass zum einen eine Automatisierung der relevanten Arbeitsabläufe im Labor, zum anderen eine Digitalisierung des Untersuchungsguts an sich positiven Einfluss auf Arbeitsbelastung und -effizienz nehmen könnten (69, 70). Ein relevanter Schritt hierfür war die Einführung von sogenannten „Whole Slide Scanner" Systemen, die die Digitalisierung von Glasobjektträgern mit gefärbten Gewebeschnitten in hoher Auflösung ermöglichen (67). Diese Methode, erstmals 1999 durch Wetzel und Gilbertson beschrieben (71), bietet Pathologen die Möglichkeit ihre Arbeit vom Mikroskop auf den Computermonitor zu verlegen, wobei sie gleichermaßen über den gesamten Schnitt navigieren und verschiedene Vergrößerungen betrachten können. Dies hat neue Möglichkeiten eröffnet: Die Archivierung von Gewebeschnitten benötigt nicht mehr zwingend physikalischen Raum und auch die Bildqualität, die beim Glasträger unter Umwelteinflüssen wie Licht und Hitze leidet, kann ohne Verlust diagnostischer Information erhalten werden (72). Konsultationen spezialisierter Pathologen werden auch über weite Strecken erleichtert, da nicht länger der physikalische Transport der Schnitte vonnöten ist (72). Weiterhin wird unter anderem die weiterführende Integration elektronischer Arbeitsabläufe in den Alltag, ein breit gefächerter Zugang für die Lehre und eine unkompliziertere Präsentation in Tumorboards ermöglicht (72).

2.4 Maschinelles Lernen

Die Gewebeschnitte konnten so auch erstmalig einer automatisierten, computerbasierten Analyse zugänglich gemacht werden (72). Dies ermöglichen Verfahren des sogenannten maschinellen Lernens (72). Hierbei handelt es sich um ein Gebiet der Computerwissenschaften, genauer gesagt ein Teilgebiet der künstlichen Intelligenz (73). Als intelligent lässt sich ein System dann bezeichnen, wenn es die Fähigkeit besitzt sich einer verändernden Umgebung anzupassen und sich bei beobachteten Fehlern so zu verändern, dass beim nächsten Mal eine adäquatere Reaktion erfolgt (74). Maschinenlernsysteme erkennen Muster und Gesetzmäßigkeiten in eingespeisten Datensätzen, erstellen ein statistisches Modell und versetzen sich so in die Lage eigenständig Problemlösungen zu entwickeln, ohne dass dabei jeder einzelne Schritt der Problemlösung programmiert werden müsste (75). Im Prinzip generiert sich das System künstlich Wissen auf Basis seiner gesammelten Erfahrungen, indem es die aus seinen bisherigen Daten gewonnenen Erkenntnisse verallgemeinert und auf die neuen Daten anwendet, was als Generalisierung bezeichnet wird (75). So kann maschinelles Lernen beispielsweise relevante Daten finden, extrahieren und zusammenfassen, Prognosen anhand von eigenen Datenanalysen erstellen, Wahrscheinlichkeiten berechnen oder Prozessoptimierungen anhand von Mustererkennung vornehmen (76).

2.4.1 Lernformen

Die praktische Umsetzung all jener Funktionen basiert auf Algorithmen, das heißt einer Folge von definierten Anweisungen, die durchlaufen werden muss, um eine Eingabe in eine Ausgabe zu überführen (74). Mittlerweile gibt es viele verschiedene zugrundeliegende Lernregeln; die älteste stammt von Donald Hebb aus dem Jahr 1949 und wurde anhand von Überlegungen zum Verhalten von Neuronen beim menschlichen Lernen erstellt (77). Man unterteilt im Bereich des maschinellen Lernens grundsätzlich in zwei Gruppen: überwachtes Lernen (supervised learning) und unüberwachtes Lernen (unsupervised learning) (76). Das Grundprinzip dieser Lernformen wird in Abbildung 2 veranschaulicht.

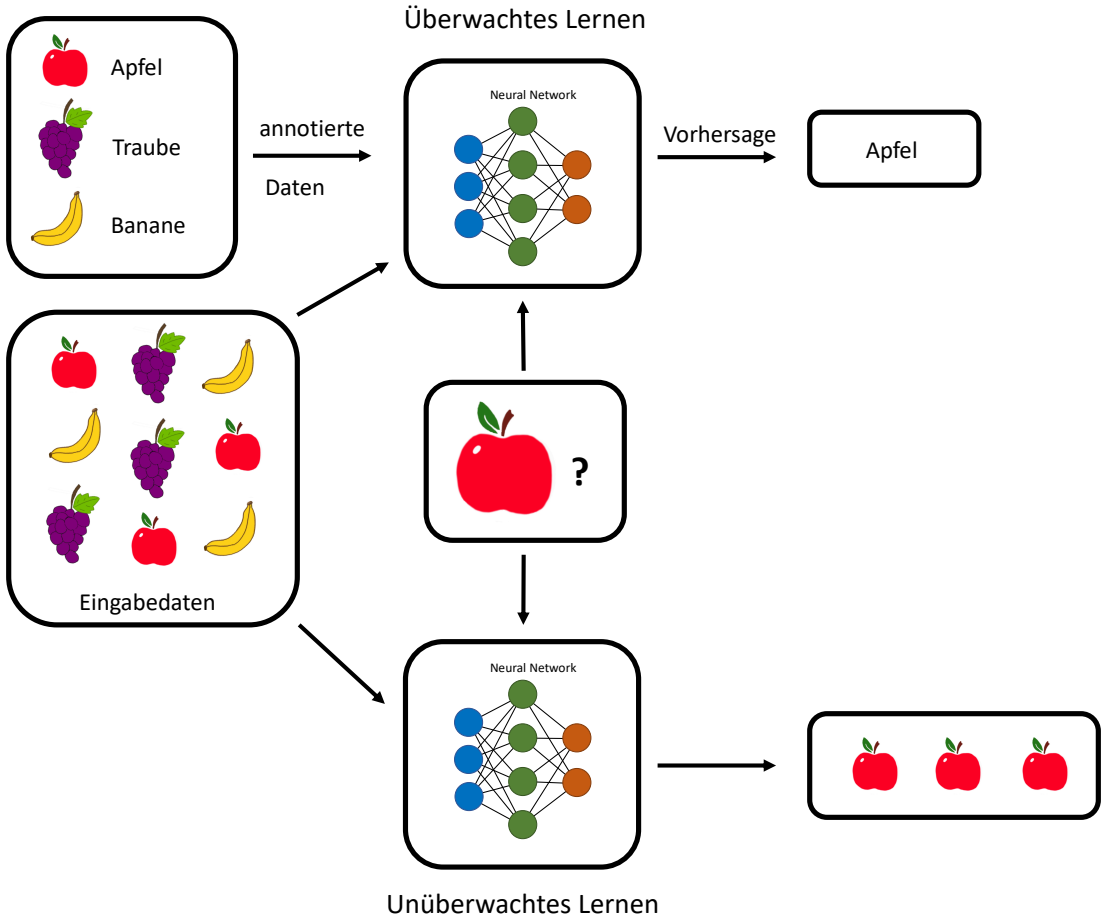


Abbildung 2: Prinzip des überwachten und unüberwachten Lernens
Die Abbildung veranschaulicht das Prinzip des überwachten und unüberwachten Lernens. Beim überwachten Lernen werden dem System zusätzlich zu den Eingabedaten Annotationen beziehungsweise Label zur Verfügung gestellt (Apfel, Traube, Banane für die jeweiligen Bilder der Früchte), sodass eine Vorhersage mit Klassifikation, das heißt die Ausgabe eines Labels (Apfel) für neue Daten (Apfelbild) erfolgen kann. Beim unüberwachten Lernen werden anhand von Mustererkennung Gruppen erstellt, in die neue Daten (Apfelbild) eingeordnet werden können (Gruppe von Apfelbildern) ohne ein Label zuzuweisen. Abbildung vom Autor.

2.4.1.1 Überwachtes Lernen (supervised learning)

Beim überwachten Lernen (vgl. Abb. 2) werden dem System externe Trainingsdaten präsentiert, bei denen bereits korrekte Ausgabewerte mit den Eingabewerten gepaart wurden (75, 76). Ziel ist es, dass das Netzwerk erlernt eigene Assoziationen herzustellen und zu verallgemeinern, um so eine generelle Regelhaftigkeit zu erlernen mit der Input in den korrekten Output überführt werden kann (75, 76). Dabei sind Regression und Klassifikation die beiden wesentlichen Themenbereiche der Anwendung (74). Mittels Regression soll meist auf Basis von bestimmten Variablen etwas über einen Sachverhalt vorhergesagt werden, wobei die Werte hierbei aus dem kontinuierlichen Bereich stammen (74, 76). Ein Beispiel wäre die Vorhersage einer sicher rückzahlbaren Kredithöhe (74, 76). Im Gegensatz dazu zeichnen sich Klassifikationsprobleme dadurch aus, dass die Ausgabe nur wenige diskrete Werte annehmen kann, Zwischenwerte können hier meist nicht als output akzeptiert werden

(74, 76). Man unterscheidet binäre Probleme von solchen bei denen multiple Klassen vorliegen (76). Ein Beispiel wäre die Zuordnung zu einer Pflanzenunterart anhand von Messwerten für deren Blättergröße (76).

2.4.1.2 Schwach überwachtes Lernen (weakly supervised learning)

Eine Erweiterung des überwachten Lernens bildet das sogenannte schwach überwachte Lernen (weakly/ semi supervised learning) (78). Es handelt sich um ein Maschinenlernmodell, bei dem das Training anhand von Daten erfolgt, die nur teilweise mit einem Label versehen wurden (78). Ziel ist es einen Prozess zu starten, so dass anschließend weiteren Daten durch das Modell Label zugewiesen werden können (78). Damit wird versucht die zeitaufwändige Benennung von Trainingsdaten durch den Menschen zu reduzieren, da dies eine signifikante Limitation für die praktische Anwendung dieser Modelle darstellt (79).

2.4.1.3 Unüberwachtes Lernen (unsupervised learning)

Beim unüberwachten Lernen (vgl. Abb. 2) enthalten die eingespeisten Trainingsdateien keinerlei Labels oder Verknüpfungen. Ziel ist es, versteckte Strukturen in den Daten zu finden und sie auf dieser Basis zu sortieren (76).

2.4.1.4 Reinforcement Learning

Beim Reinforcement Learning arbeitet das Modell in einem dynamischen Umfeld, welches Feedback in Form von positiver oder negativer Verstärkung bietet (76). So erlernt das Modell ohne explizite Anweisungen eine Strategie zur Maximierung der Belohnungen anhand von Trial-and-Error-Simulationen (76, 80).

2.4.2 Neuronale Netzwerke

Künstliche neuronale Netzwerke sind Algorithmen des maschinellen Lernens, welche in ihrer Grundidee an der Funktionsweise des menschlichen Gehirns angelehnt sind (81). Die Methode fand ihre Anfänge um 1940 mit den Arbeiten von Walter Pitts und McCulloch und erlebte seit den 1980er Jahren eine Renaissance (82). Die Analogie zur Nervenzelle spiegelt sich in der Funktionsweise wider: Anregungen aus anderen Zellen werden aufgenommen, aufsummiert und bei Überschreitung eines festgelegten Schwellenpotentials weitergetragen (Aktivierungs-/ Schwellenwertfunktion) (81). Korrelat der Neurone sind im neuronalen Netz sogenannte Perzeptrone (81). Jede ihrer Eingaben enthält zusätzlich ein Verbindungsgewicht, auch synaptisches Gewicht

genannt, welches die Relevanz der Verknüpfung zusätzlich zur Information einbringt (81). Entsprechend der menschlichen Neurone im Gehirn können auch die neuronalen Netze in verschiedene Ebenen gegliedert sein: Eingangs- und Ausgangsneurone sowie variable Anzahlen von Schichten aus Zwischenneuronen (83). Dabei gibt es einschichtige und mehrschichtige Systeme (81). Einschichtige Netze besitzen neben der Eingabeschicht (input layer) nur eine Ausgabeschicht (output layer) und sind somit die einfachste Struktur künstlicher neuronaler Netzwerke – auch Perzeptron-Netzwerk genannt (77). Mehrschichtige Netze (vgl. Abb. 3) besitzen zusätzlich noch weitere, sogenannte verdeckte Schichten (hidden layer), deren Ausgaben außerhalb des Netzes unsichtbar bleiben (83). Sie werden ‚Multi-Layer-Perceptrons' (77), kurz MLP, genannt und besitzen eine höhere Abstraktionsfähigkeit (81, 83). Die Neuronenschichten können parallel arbeiten, die eigentliche Leistungsfähigkeit des Ansatzes resultiert jedoch wie auch beim Gehirn aus der starken Vernetzung untereinander - beim Menschen sind die Neurone über sogenannte Synapsen mit durchschnittlich 10^4 anderen Nervenzellen verbunden (81, 84). Eine höhere Anzahl von Neuronen, Schichten und Vernetzungen ist somit entscheidend dafür, dass das System auch komplexere Aufgaben zu lösen vermag (81, 83).

2.4.3 **Deep Learning**

Deep learning wiederum bezeichnet Architekturen neuronaler Netzwerke, welche besonders ‚tief' ist; vereinfacht bedeutet dies, dass besonders viele Ebenen vorliegen (81). Definitionsgemäß spricht man aber schon bei einem Netzwerk mit mehr als zwei Zwischenschichten (hidden layer) von einem Deep Network (vgl. Abb. 3) (83). Das grundlegende Konzept des Deep Learnings ist es Maschinen das Lernen an sich beizubringen, sodass sie die Fähigkeit erlangen selbstständig, das heißt unter minimalem Eingriff des Menschen in den Lernvorgang, ihre Leistungen zu verbessern (81, 83). Ermöglicht wird dies durch Extraktion von Mustern, Abhängigkeiten und Regelmäßigkeiten, die ein gewisses Maß an Verallgemeinerung erlauben (81). Zur Minimierung der Fehlerfunktion der letztlichen Ausgabe kommen für mehrschichtige Netzwerke im Bereich des überwachten Lernens Methoden wie Backpropagation zum Einsatz: Eine Eingabe durchläuft das Netz vorwärtsgerichtet, die erreichte Ausgabe wird mit der gewünschten Ausgabe (dem Label) verglichen und die Differenz als Fehler berechnet (81, 83). Dieser wird dann rückwärts zur Eingabeschicht zurückgeleitet, wobei das synaptische Gewicht je nach geschätztem Einfluss auf den Fehler adjustiert wird (81, 83). So erfolgt bei erneuter Eingabe eine Annäherung an die

gewünschte Ausgabe: Das System ist praktisch in der Lage aus seinen Fehlern zu lernen (81, 83).

2.4.3.1 Anwendungsbereiche für das Deep Learning

Deep Learning eignet sich vor allem für all jene Anwendungsfelder wo große Datenmengen auf Muster untersuchbar sind (76). Solche Systeme und Modelle liegen Innovationen wie beispielsweise Spracherkennungs- und Verarbeitungssoftware, persönlichen digitalen Assistenten in Smartphones oder autonom fahrenden Fahrzeugen zugrunde und stellen damit einen teilweise nicht mehr wegzudenkenden Teil des alltäglichen Lebens dar (85). Durch zunehmende Vernetzung über das Internet mit vereinfachtem Datenzugang und sinkende Kosten für immer größere Rechenleistungen besteht nun die Möglichkeit Lernalgorithmen an enormen Datenmengen zu trainieren und testen, was sich als entscheidender Faktor für die Nutzung von deren Potential offenbart hat (74, 85). Auch im medizinischen Sektor im Allgemeinen und dem Fachgebiet der Pathologie im Speziellen gibt es erste Proof-of-concept Studien (67) (vgl. Abschnitt 2.5.1).

2.4.4 Convolutional Neural Networks (CNN)

Wie zuvor beschrieben bestehen herkömmliche künstliche neuronale Netzwerke aus voll- oder teilvernetzten Neuronen, die in mehreren Ebenen angeordnet sind. Als man begann diese Systeme auf die Verarbeitung von Bildern zu trainieren, gerieten sie vor allem im hochauflösenden Bereich schnell an ihre Grenzen, da die Anzahl der benötigten Neurone, entstehenden Verknüpfungen und zu berechnenden Gewichte enorme Anforderungen an die Rechenleistung des Computers stellten (86). Dieses Problem können CNNs in verschiedenen Ebenen weitgehend lösen (78) (vgl. Abschnitt 2.4.4.2). Übersetzt bedeutet Convolutional Neural Network, kurz CNN oder ConvNet, „Gefaltetes Neuronales Netzwerk“ (78). Diese Form erfuhr ihre Begründung 1989 durch die Arbeiten von Yann LeCun (87). Ihren Namen erhielt sie durch Einsatz der ‚Faltung‘, einer speziellen mathematischen Operation zur Signalfilterung (vgl. Abschnitt 2.4.4.2) (78). Sobald mindestens eine Schicht in dieser Form arbeitet, handelt es sich definitionsgemäß um ein CNN (78). Grundsätzliche Aspekte wie zum Beispiel das Training – in der Regel mittels überwachter Lernalgorithmen - bleiben unverändert (78).

2.4.4.1 Aufbau eines Convolutional Neural Networks

ConvNets setzen sich ebenfalls aus verschiedenen Schichten zusammen und entsprechen prinzipiell einem partiell lokal, partiell vollständig verknüpften neuronalen Netzwerk (81, 88).

Die verschiedenen Schichten des CNN, wie in Abbildung 3 veranschaulicht, sind (78, 86):

- die Convolutional-Schicht
- die Pooling-Schicht
- die vollständig verknüpfte Schicht (fully connected layer)

Die Pooling-Schicht folgt auf die Convolutional-Schicht, kann aber als Kombination prinzipiell beliebig oft hintereinander vorkommen (78, 86).
Da sowohl Pooling- als auch Convolutional-Schicht nur lokal verknüpfte Teilnetze sind, ist die Zahl der Verknüpfungen hier selbst bei großen Eingabemengen begrenzt und somit für die Rechenleistung des Computers leichter zugänglich (86). Am Schluss sitzt immer eine vollständig verknüpfte Schicht (78, 89).

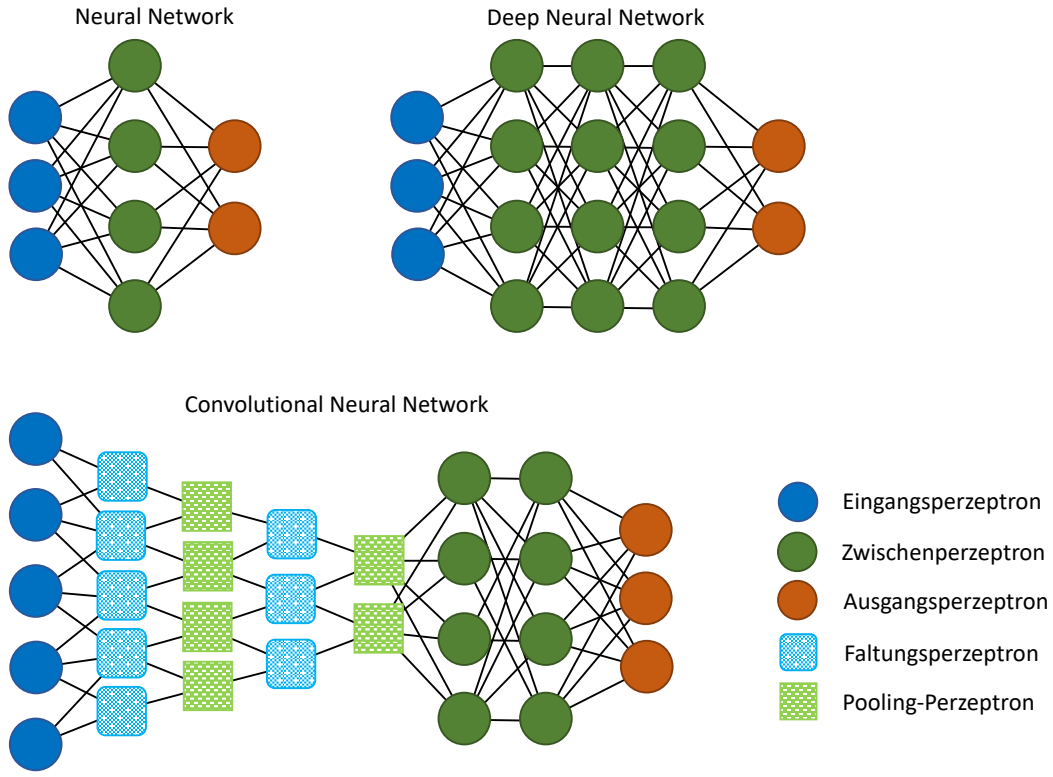


Abbildung 3: Aufbau verschiedener Neuronaler Netzwerke
Die Abbildung zeigt den Aufbau verschiedener Formen von Neuronalen Netzwerken. Jedes Arbeitselement wird als Perzeptron bezeichnet, wobei diese verschiedene Funktionen erfüllen können (vergleiche Legende der Perzeptronen). Einfache neuronale Netze bestehen aus mindestens einer Reihe von Eingangsperzeptronen (input layer)) und den Ausgabeperzeptronen (output layer), bei mehrschichtigen Systemen kommt zudem eine Reihe von Zwischenperzeptronen (hidden layer) hinzu (vergleiche Darstellung Neural Network). Besitzt das Modell

mindestens zwei hidden layer bezeichnet man es als Deep Learning Network (vgl. Darstellung Deep Neural Network). Convolutional Neural Networks besitzen zusätzlich teilverknüpfte Schichten aus Faltungsperzeptronen und Pooling-Perzeptronen (vgl. Darstellung Convolutional Neural Network). Abbildung vom Autor.

2.4.4.2 Funktionen der einzelnen Schichten eines CNN

Convolutional-Schicht

Die Convolutional-Schicht stellt dem Namen entsprechend die eigentliche Faltungsebene dar (78). Aufgabe der Convolutional-Schicht ist das Erkennen und Extrahieren von Merkmalen aus den eingespeisten Daten (90). Dies erfolgt als mathematische Funktion mittels Aufsummieren beziehungsweise Aufintegrieren von gewichteten Werten (78). Es entsteht eine Filterfunktion, welche jedoch an jeder Bildstelle abhängig vom Umfeld unterschiedlich filtert (90). Auch hier lässt sich ein Vergleich zum menschlichen Gehirn ziehen: Analog zum visuellen Cortex steigt in tiefer gelegenen Schichten sowohl die Größe der rezeptiven Felder als auch die Komplexität der erkannten Merkmale (90). Das heißt in der ersten Ebene wird zunächst auf ganz einfache Strukturen wie zum Beispiel Linien, Farbtupfer oder Kanten untersucht (81). Mit jeder nachgeschalteten Ebene kann die Komplexität der identifizierten Strukturen steigen, von einfachen Formen und Kurven bis zu aufwändigeren Gebilden (81). Dabei werden die Daten für jede Schicht immer wieder neu untersucht und ‚gefiltert' (78). Auf diese Weise gehen durch andere ‚Filter' ausgeschlossene Informationen nicht insgesamt für das System verloren (78). Welche ‚Faltungen' letztlich am vorteilhaftesten für das Ergebnis sind, kann das System selbst erlernen (78).

Pooling-Schicht

Die Pooling-Schicht wird auch als Subsampling-Schicht bezeichnet und bildet das biologische Pendant zur lateralen Hemmung im visuellen Cortex (Sehrinde des Gehirns), einem Verschaltungsprinzip bei dem aktive Nervenzellen die benachbarten Zellen in ihrer Aktivität hemmen (91). Sie bearbeitet die zuvor erkannten Merkmale indem sie diese verdichtet und die Auflösung reduziert (78). Das Pooling reduziert also die Datenmenge durch Aussortieren überschüssiger Informationen ohne dabei die Leistungsfähigkeit des Systems zu beeinträchtigen (86). Durch das geringere Datenaufkommen steigt die Berechnungsgeschwindigkeit (86). Ermöglicht wird dies durch Methoden wie dem Maximal-Pooling (max-pooling) oder Mittelwert-Pooling (average pooling) (78). Beim Maximal-Pooling werden aus einem Quadrat von ‚Neuronen' die Aktivsten gefunden und nur diese für weitere Berechnungsschritte

genutzt, beim Mittelwert-Pooling wird alternativ die mittlere Aktivität herangezogen (78).

Vollverknüpfte Schicht

Die vollverknüpfte Schicht schließt sich den wiederholten Abfolgen von Convolutional- und Pooling-Schichten an und erhält ihre Eingaben aus der letzten Schicht dieser Abfolge (78). Als Abschluss des Netzwerkes eint sie demnach sämtliche Merkmale und extrahierte Eigenschaften der vorgelagerten Schichten (78). Ihre vollständig verknüpften Neurone können wiederum in mehreren Schichten angeordnet sein (86). Es erfolgt in einem letzten Schritt die Zuordnung der Ergebnisse zu den vorgegebenen vorhandenen Ausgabemöglichkeiten, wobei die Anzahl an ‚Neuronen' von der Aufgabe, beispielsweise der Anzahl der zu unterscheidenden Klassen, abhängt (78). Die Ausgabe dieser letzten Schicht wird oftmals über eine Normalisierungsfunktion in die Ausgabe einer Wahrscheinlichkeitsverteilung überführt (89).

2.4.4.3 Anwendungsbereiche des Convolutional Neural Networks

CNNs bieten sich für Anwendungen der künstlichen Intelligenz an, die mit einer großen Menge von unstrukturierten Eingabedaten einhergehen (74). Daher sind sie für die Bilderkennung die ‚State-of-the-Art-Methode'; im Alltag sind sie beispielsweise in Software für Gesichtserkennung zu finden (74, 86). CNN arbeiten robust und sind gegenüber Verzerrungen und anderen optischen Abwandlungen bis zu einem gewissen Maße unempfindlich (90). So können auch aus Bildern unterschiedlicher Lichtverhältnisse oder Perspektiven die typischen Merkmale extrahiert werden (74). Aber auch Spracherkennung und Textverarbeitung bilden Einsatzgebiete (78).

2.4.5 Transfer Learning

Menschliches Lernen hat Wege gefunden Wissen auf verschiedenste Anwendungen zu verallgemeinern (92). Das bedeutet relevantes Wissen aus früheren Lernerfahrungen wird erkannt und -sofern es relevant erscheint- auf neue Aufgaben übertragen (92). Traditionelle Maschinenlernalgorithmen hingegen sind in der Regel auf isolierte Aufgaben fokussiert (78). Transferlernen versucht das traditionelle Maschinenlernen zu erweitern und zu verbessern, indem ebenjene menschliche Fähigkeit des Wissenstransfers den Computern zugänglich gemacht werden soll (93). Es handelt sich um eine Technik aus dem Bereich des Deep Learnings: Hier wird ein bereits vortrainiertes künstliches neuronales Netz genommen und für die Lösung einer

anderen Problemstellung verwendet (78). Der antrainierte Lernfortschritt des Modells wird dabei transferiert (78). Damit werden große Teile des intensiven Trainings gespart, welches sich bei komplexen Deep Learning Modellen selbst auf spezieller Hardware über Wochen erstrecken kann (94). Daraus ergeben sich viele Vorteile wie schnellere Erstellung, höhere Startqualität, schnellere Verbesserung, bessere abschließende Modellqualität und weniger Ressourceneinsatz (95). Die Methode findet aufgrund der nötigen Komplexität der Modelle vor allen in den Bereichen der Bild- und Textverarbeitung Anwendung (94). Eines der einfachsten Beispiele von Transferlernen ist die Verarbeitung von Bilddaten zur Objekterkennung (94). Hier zum Einsatz kommende vortrainierte Modelle sind zum Beispiel ResNet (96) und GoogLeNet (97), die bereits viele unterschiedliche Objekte erkennen können. Sie stehen kostenlos im Internet zum Download bereit. Allerdings ist das Trainingsmodell nicht auf den eigenen Anwendungsfall wie z.B. die Erkennung maligner Zellen spezialisiert. Durch das Transferlernen ist jedoch der Großteil des Trainings bereits erfolgt, lediglich die letzte Ebene, die sogenannte Klassifikationsschicht (fully connected layer) muss neu angefügt und trainiert werden, dann kann das Vorhandensein von Tumor vorhergesagt werden (78). Vor allem für all solche Fälle in denen die Datenmenge begrenzt ist, kann Transferlernen dabei helfen trotzdem Modelle mit hoher Performance zu erschaffen, die mit traditionellen Methoden schlichtweg nicht erreichbar würde (78).

2.4.6 **Data Augmentation**

Eine andere Methode bei vorliegenden Problemen bezüglich geringer Datenmenge ist die sogenannte ‚Data Augmentation', im Deutschen Datenerhöhung (78). Hier werden bereits vorhandene Dateien, zum Beispiel Bilder, in Form von Drehungen, Spiegelungen, Scherungen oder Farbvariationen verändert (78). Eine weitere Möglichkeit sind sogenannte ‚progressive sprinkles' (98). Bei diesen handelt es sich um kleine Ausschnitte – ‚cutouts' – aus dem Bild in verschiedener Anzahl variabel über das Bild verteilt (98). Dies soll das Modelles zu einer Verstärkung des Detaillernens zwingen (98). Nach Augmentation werden die variierten Daten zusätzlich als „neue" Beispiele in die Datenbank eingespeist (78). Vergleiche hierzu auch Abbildung 8, Abschnitt 3.6.

2.4.7 **Overfitting/Underfitting**

Eine adäquate Datenmenge ist wichtig, um das Problem der Überanpassung des Modelles, des sogenannten ‚overfitting' zu vermeiden (83). Overfitting beschreibt das Phänomen, das entsteht, wenn das Training eines Modelles zu lange fortgesetzt wird: Mit steigender Zahl von Trainingsepochen nimmt der Fehler am Trainingsdatensatz immer weiter ab, gleichzeitig wird der Fehler an einem Validierungsdatensatz ab einem bestimmten Zeitpunkt immer höher (99). Zu Beginn des Trainings sind alle Verbindungsgewichte zwischen den Perzeptronen nahe 0 (vgl. Abschnitt 2.4.2), mit dem Training werden dann zunächst die relevantesten Gewichte der Aufgabe angepasst (83). Mit zunehmender Trainingsdauer werden jedoch auch für die Aufgabe weniger relevante Gewichte zur Spezialisierung auf den Trainingsdatensatz angepasst (81). Das Modell wird zunehmend komplexer, während es die Trainingsdaten regelrecht auswendig lernt (81). Folge ist eine schlechte Generalisierung bei Anwendung des Modelles auf nicht im Training enthaltene Daten (81, 83). Um dies zu vermeiden, muss das Training rechtzeitig gestoppt werden (83). Der richtige Zeitpunkt wird dabei durch Kreuzvalidierung bestimmt, welche Testungen des Netzwerkes an nicht im Training enthaltenen Validierungsdaten vornimmt (99, 100). Würde das Training zu früh gestoppt, das heißt läge das Modell in seiner Komplexität weit hinter der Komplexität der zu lösenden Aufgabe, wäre die Generalisierbarkeit auf Validierungsdaten ebenfalls schlecht. Dies bezeichnet man als Unteranpassung beziehungsweise underfitting (100).

Für ein optimales Modell sollte demnach wie in Abbildung 4 veranschaulicht die Komplexität so gering wie möglich, jedoch so hoch wie nötig trainiert werden, wobei ein Kompromiss zwischen der Komplexität des Modelles, der möglichen Menge an einbringbaren Daten und dem resultierenden Generalisierungsfehler geschlossen werden muss (99).

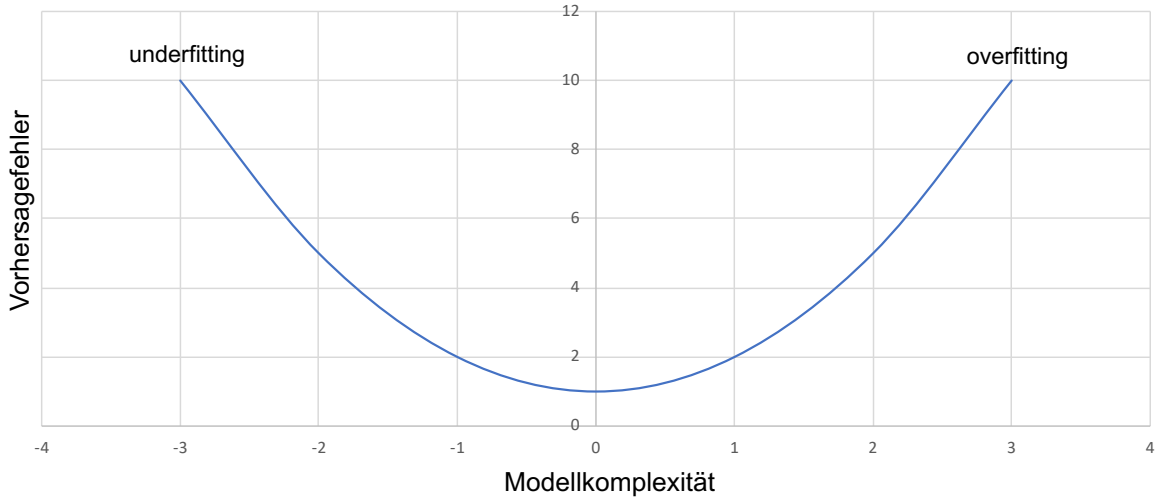


Abbildung 4: Zusammenhang zwischen Modellkomplexität und Vorhersagefehler
Sowohl bei zu geringer als auch bei zu höher Komplexität des Modelles durch zu wenig beziehungsweise zu viel Training steigt der Vorhersagefehler, was als underfitting beziehungsweise overfitting bezeichnet wird. Das optimale Modell mit dem niedrigsten Vorhersagefehler liegt im Bereich einer mittleren, der Komplexität der Aufgabe angepassten Modellkomplexität. Abbildung vom Autor.

2.5 Entwicklung der Forschung zur digitalen Pathologie

Dass die Digitalisierung in der Pathologie zunehmend relevanter und die Implementierung in die Realität nahe rückt, lässt sich unschwer an der steigenden Anzahl von Publikationen zu eben jenem Thema erkennen. Im Jahr 2010 überschritt die Zahl der Veröffentlichungen zur digitalen Pathologie auf PubMed erstmals 1000, 2020 waren es schon über 2000 Publikationen (101).

2.5.1 Beispielhafte Publikationen in der Entwicklung digitaler Pathologie

Beck et al. konnten erstmalig den erfolgreichen Einsatz maschinellen Lernens für eine prognostische Vorhersage bei Mammakarzinompatientinnen anhand von Bildeigenschaften von H.E. gefärbten Gewebeschnitten zeigen (102). Allerdings mussten die Bildeigenschaften in dieser ersten Arbeit noch durch einen menschlichen Befunder erhoben werden (102). Zudem kamen hierbei noch „oberflächliche“ statistische Verfahren, sogenanntes „shallow learning“ zum Einsatz (102). Neuere, deutlich leistungsfähigere Ansätze beruhen auf künstlichen neuronalen Netzwerken: In ersten Arbeiten, zum Beispiel zur Detektion von Lymphknotenmetastasen bei Mammakarzinompatientinnen konnten den beteiligten Pathologen und Pathologinnen mindestens ebenbürtige Genauigkeitswerte erzielt werden (103). Litjens et al. beschrieben zudem die Möglichkeit der Verbesserung von Effizienz und Genauigkeit histopathologischer Diagnosen in Bezug auf Untersuchungen von Biopsien auf Prostatakarzinome sowie Sentinellymphknoten auf Mammakarzinome durch Einsatz

von maschinellem Lernen als Screening-Instrument (67). Durch automatisierte Aussortierung benigner Gewebeproben konnte die Arbeitslast des Pathologen reduziert werden ohne dabei die Rate an falsch negativen Diagnosen zu erhöhen (67). In einer im Oktober 2018 veröffentlichten Arbeit wurden zudem erste Untersuchungen zur Veränderung der Leistung von Pathologen bei Unterstützung durch ein neuronales Netz bei der Diagnostik von Lymphknotenmetastasen bei Patienten mit Mammakarzinomen angestellt: Durch das System unterstützte Pathologen zeigten insgesamt höhere Genauigkeitswerte als ohne Hilfe, aber auch als der Algorithmus allein, was die Diskussion über eine mögliche komplementäre Arbeit von Mensch und Maschine aufwarf (104). Auch die durchschnittliche Bearbeitungszeit der einzelnen Schnitte fiel unter Nutzung des neuronalen Netzwerkes geringer aus, die subjektive Schwierigkeit der Beurteilung sank (104). Auch für den gastrointestinalen Bereich wurden bereits Forschungen angestellt, 2020 wurde eine Arbeit zur histopathologischen Klassifikation von epithelialen Tumoren in Magen und Kolon veröffentlicht (105). Hier wurden Modelle des maschinellen Lernens erfolgreich auf die Klassifikationsentscheidung zwischen Adenokarzinom, Adenom und benignem Gewebe trainiert (105). Es wird ein vielversprechendes Potential für die Implementation solcher Modelle in klinische Arbeitsabläufe beschrieben (105). Selbiges beschreiben Song et al. 2020 in einer Arbeit zur histopathologischen Diagnose von Magenkarzinomen mittels deep learning (106). Unter Zusammenarbeit von Modell und Pathologe wurden die Genauigkeit der Diagnosen erhöht sowie die Rate an Fehldiagnosen vermindert, was Hinweise auf den potentiellen Nutzen von maschinellem Lernen als Assistenzsystem im klinischen Alltag gibt (106). Allen Publikationen gemein ist die Forderung nach weiteren Untersuchungen zu diesem Thema.

3 Material und Methoden

3.1 Studienkollektiv

Zunächst wurde ein Kollektiv aus Magenkarzinompatienten der Universitätsmedizin Mainz, die eines von sechs zertifizierten Referenzzentren für Magen- und Speiseröhrenchirurgie der Deutschen Gesellschaft für Allgemein- und Viszeralchirurgie darstellt, generiert. Einschlusskriterien für diese Kohorte beinhalteten die Diagnose eines siegelringzellig differenzierten Adenokarzinoms des gastroösophagealen Übergangs oder des Magens (ICD-10 C15 und C16) mit Erstdiagnose zwischen dem 01.01.2008 und dem 31.12.2018. Anhand einer systematischen Analyse der Patientendatenbank PathoPro mit den Suchtermini „Adenokarzinom“ und „Magen*“ beziehungsweise „Ösophagus*“ erfolgte die Erstellung einer Fallliste. Dabei wurden die erweiterten Suchkriterien so gewählt, dass wenigstens eine Subpopulation der Tumorzellen als siegelringzellig definiert wurde. Die zugehörigen H.E. gefärbten Schnittpräparate und ihre Befunde wurden gesichtet, pathologisch reevaluiert und mindestens ein Objektträger ausgewählt, auf welchem sich ein möglichst großes Tumorareal abgrenzen ließ. Weiterhin wurde mindestens ein repräsentativer Normalbefund mit tumorfreien Magenanteil oder falls nicht anderweitig vorhanden ein Normalbefund aus oralem oder aboralem Absetzungsrand im Falle einer R0-Resektion ausgewählt. Ausgeschlossen wurden all jene Fälle und Schnittpräparate die den oben genannten Kriterien nicht entsprachen, beispielsweise aufgrund gänzlich fehlender siegelringzelliger Differenzierung oder Lokalisation außerhalb der Vorgaben. Weitere Ausschlusskriterien waren eine überwiegend muzinöse Differenzierung ohne erkennbare Tumorzellkerne und Sektionsfälle.

3.2 Klinische Hintergrunddaten

Zusätzlich wurden folgende klinisch-pathologische Daten der Patienten erhoben: Fallart (Schnellschnitt oder normale Einsendung), Fallnummer, Jahr, Einsendedatum, Geburtsdatum, Geschlecht, Tumorgröße, Tumorlokalisation, TNM-Stadium, Anzahl betroffener, gesunder und untersuchter Lymphknoten, Resektionsstatus, Grading, Perineuralscheideninfiltration, venöse oder lymphatische Ausbreitung sowie Bemerkungen zu Helicobacter pylori Besiedlung, intestinaler Metaplasie, Mutationen und Vorliegen als Biopsie. Nach Abschluss der Datenerhebung wurden aus Gründen des Datenschutzes identifizierende Merkmale entfernt.

3.3 Digitalisierung der Gewebeschnitte

Die ausgewählten Objektträger wurden nach Säuberung der Glasträger und Kalibrierung des Geräts mittels des Hamamatsu Nanozoomer Series Whole-Slide-Scanners (Hamamatsu Photonics, Hamamatsu, Japan) in 40-facher Vergrößerung digitalisiert.

3.4 Annotation der Gewebeschnitte und weitere Fallselektion

In einem nächsten Schritt wurden die Gewebeschnitte digital mittels der Open Source Software QuPath (107) in der Version 0.2.0-m2 annotiert. Annotation beschreibt den Prozess des „händischen“ Markierens verschiedener Bildbereiche durch Umfahren wie in Abbildung 5 veranschaulicht wird.

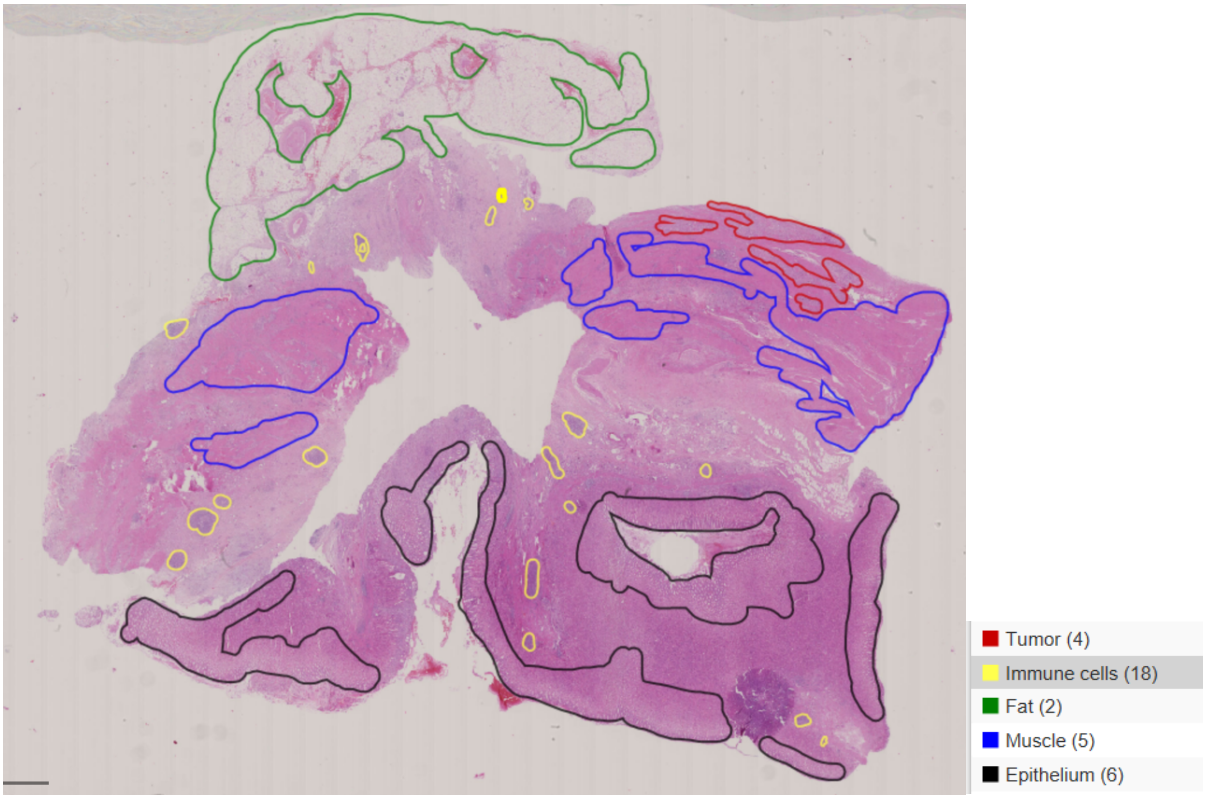


Abbildung 5: Annotation mittels QuPath
Beispielhafte Darstellung einer Annotation mittels der Software QuPath an einem gescannten Gewebeschnitt. Im oberen rechten Bildteil sind Tumorareale in rot annotiert, tiefer liegendes Muskelgewebe wurde in blau dargestellt umfahren. Außerdem Annotation von Fettgewebe in grün, Immunzellen in gelb und epithelialem Gewebe in schwarz. Abbildung vom Autor auf Basis der Annotationssoftware QuPath.

Unterschieden wurden folgende Gewebekomponenten: gesunde Magenschleimhaut (Epithel), gesundes Fettgewebe, Immunzellen, gesundes Muskelgewebe und Karzinominfiltrate (vgl. Abb. 5 und 6). Die jeweilige Zuordnung zu den einzelnen Klassen wurde stichprobenartig durch einen weiteren Befunder überprüft. Unklare Fälle wurden gemeinsam digital mikroskopiert und ein Konsens gefunden.

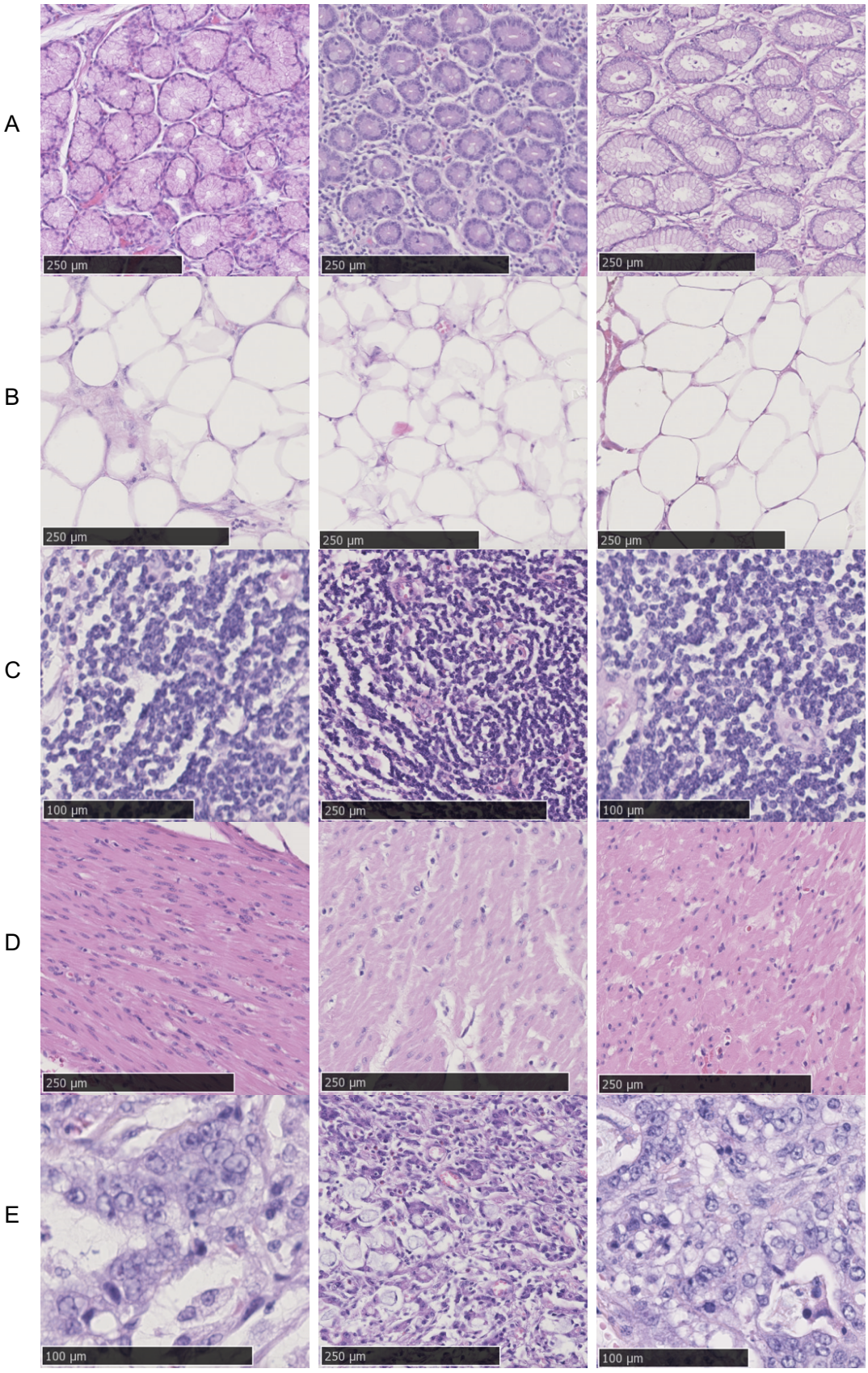


Abbildung 6: Histologische Beispielbilder für jede der fünf annotierten Gewebeklassen
Die Beispielbilder veranschaulichen die Variationen in Helligkeit, Färbungsintensität und Gewebetexturen in den histopathologischen Schnitten. A Epithel, B Fett, C Immunzellen, D Muskel, E Karzinom.

An dieser Stelle wurden noch einmal Schnitte mit schlechter Qualität nach Digitalisierung (Glasrisse, Lufteinschlüsse, zu hohe Gewebedicke mit optischer Gewebeüberlagerung) ausgeschlossen.

3.5 Kachelerstellung

Die annotierten Gebiete wurden für jeden Gewebetyp (Epithel, Fett, Immunzellen, Muskel und Karzinom) mittels eines entsprechenden Skriptes von QuPath in Kacheln mit einer Kantenlänge von 1024x1024 Pixel unterteilt. Zusätzlich wurden die Koordinaten der jeweiligen Kacheln auf dem digitalen Ganzgewebeschnitt mitcodiert, um später eine räumliche Zuordnung des Klassifikationsergebnisses zu ermöglichen.

3.6 Präprozessierung

Um robustere Ergebnisse zu erreichen, wurden alle Kacheln an einem externen Referenzbild (vgl. Abb. 7) normalisiert. Das bedeutet, dass zum Ausgleich von zum Beispiel unterschiedlichen Färbungsintensitäten oder Helligkeiten (vgl. Abb. 6) die Kacheln an den Parametern dieses einen Bildes ausgerichtet werden. Hier wurde die Methode nach Reinhard verwendet (108, 109).

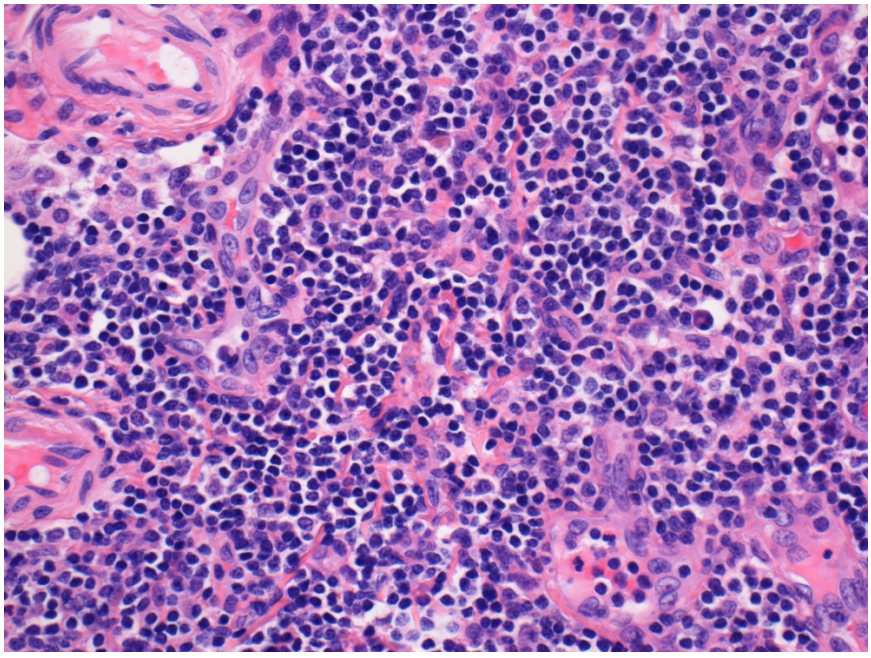


Abbildung 7: Referenzbild
Externes Referenzbild, an dem alle Kacheln normalisiert wurden. Dabei werden zum Ausgleich unterschiedlicher Färbungsintensitäten und anderer Bildmerkmale die Kacheln nach den Parametern des Referenzbildes ausgerichtet. Ziel ist die Erhöhung der Robustheit der Ergebnisse.

Zur Datenaugmentation (vgl. Abschnitt 2.4.6) wurden Drehungen um jeweils 90°, Abwandlungen von Belichtung bzw. Helligkeit und ‚progressive sprinkles' eingesetzt. Diese wurden hier mit einer Größe von 16x16 Pixel in einer Anzahl von 60-80 pro

Kachel genutzt. Abbildung 8 veranschaulicht hierbei beispielhaft die dabei aus einer Kachel entstehenden Variationen.

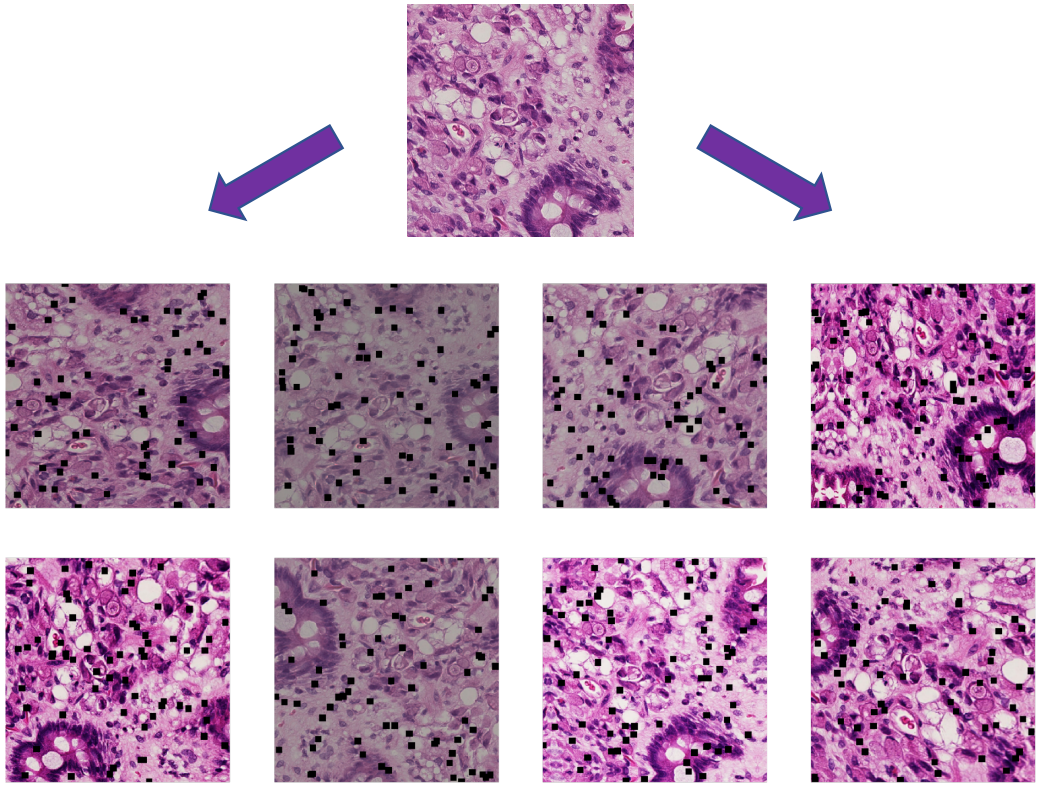


Abbildung 8: Augmentierung der Kacheln
Die Abbildung veranschaulicht am Beispiel einer Tumorkachel die Variationen, die mittels der beschriebenen Augmentationsmethoden (Drehung, Veränderung von Helligkeit bzw. Belichtung und progressive sprinkles) aus einer Eingabekachel (Kachel ganz oben) entstehen können. Abbildung vom Autor.

3.7 Training und Validierung des Neuronalen Netzwerkes

Für die Klassifikationsexperimente wurde ein CNN mittels überwachten Lernens trainiert. Dafür wurde ein standardisiertes Densely Connected Convolutional Network (DenseNet121) (110) genutzt. Das Trainingsskript befindet sich im Anhang. Als Programmiersprache wurde Python (Version >=3.6) unter Nutzung von PyTorch/fast.ai verwendet. Die Kacheln wurden für die Eingabe in das Modell auf 512x512 Pixel reduziert. Die verwendete batch size für das Training lag bei 27, die batch size für die Validierung ebenfalls bei 27. Die Anzahl der verwendeten Kerne (number of workers) lag bei 16. Als loss function wurde cross-entropy loss genutzt. Die Lernrate wurde auf 0,0001 festgelegt, der weight decay auf 0,1. Eine Erklärung der Bedeutung dieser Parameter befindet sich im Anhang. Aufgrund der vorliegenden Klassenungleichheit (vgl. Abschnitt 4.2) wurde eine Funktion zum Ausgleich eingefügt. Durch Auffüllen der Differenz zur Kachelzahl anderer Klassen mittels augmentierter Kacheln (vgl. Abschnitt 2.4.6. und 3.6) konnte für jede Klasse die jeweils gleiche Anzahl von Kacheln

herangezogen werden. Das Training wurde anhand der Metriken accuracy und error rate überwacht (vgl. Abschnitt 3.8). Die Anzahl der Trainingsepochen betrug 30. Wie in Abbildung 9 veranschaulicht, wurde eine „5-fold cross validation“ durchgeführt. Dies bedeutet, dass die Kohorte zunächst in fünf gleich große Teile zerlegt wurde. Dann wurden Teil eins bis vier für das Training des Systems genutzt und der fünfte Teil für die Validierung herangezogen. Im nächsten Schritt wurde Teil vier nach Training mit den Teilen 1,2,3,5 für die Validierung verwendet. Dies wiederholte sich, wobei jeweils ein anderer Teil für die Validierung herangezogen wurde. Nach fünf Durchgängen des Vorgangs ist somit in fünf Teilexperimenten an jedem Patienten trainiert und validiert worden. Die einzelnen Durchgänge werden dabei im Folgenden als Folds bezeichnet.

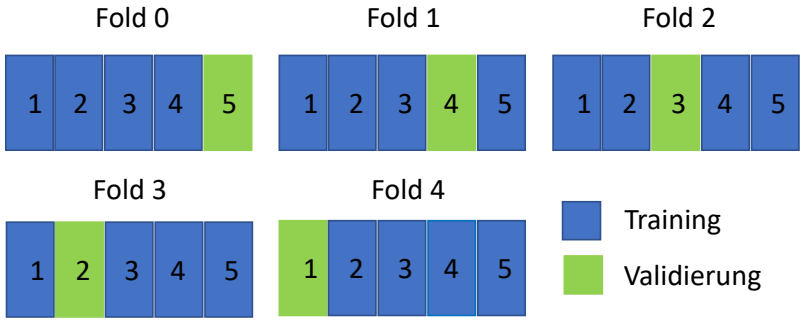


Abbildung 9: Prinzip einer 5-fold-cross-validation
Nach Aufteilung der Kohorte in fünf gleich große Teile erfolgen fünf Trainings- und Validierungseinheiten (Fold 0-4), wobei jeweils andere Teilmengen für Training und Validierung genutzt werden. Insgesamt wird so an allen Patienten trainiert und validiert. Abbildung vom Autor.

3.8 Statistische Analyse

Zur Beurteilung der Klassifizierungsexperimente wurden Leistungsmetriken basierend auf ‚scikit-learn‘ (111) berechnet. Hierzu zählen Sensitivität (recall), Spezifität (specificity), Präzision (precision), Genauigkeit (accuracy), Fehlerhäufigkeit (error rate), F-Maß (F1-Score), Präzision-Sensitivität-Kurven (Precision-Recall curves, PR-curves) und Isosensitivitätskurven (area under the receiver operating characteristic, AUROC)). Wie in Abbildung 10 dargestellt, berechnen sich Sensitivität, Spezifität, Präzision und Genauigkeit aus unterschiedlichen Kombinationen von richtig und falsch positiven und negativen Ergebnissen. Diese ergeben sich aus dem Zusammenspiel von den positiven und negativen Vorhersagen des Modelles im Vergleich zur wahren, das heißt annotierten Klasse. Die Sensitivität, welche die richtig positiven und falsch negativen Ergebnisse einbezieht, gibt dabei Auskunft über den Anteil der positiven Vorhersagen an allen wahr positiven Ergebnissen. Die Berechnung der Spezifität wiederum bezieht die richtig negativen und falsch positiven Ergebnisse mit ein, so kann eine Aussage über

die Korrektheit der negativen Vorhersagen getroffen werden. Die Korrektheit der positiven Vorhersagen wird über die Präzision beschrieben, die die richtig positiven und falsch negativen Ergebnisse einbezieht. Die Genauigkeit beschreibt die Anzahl der korrekten Vorhersagen über alle Vorhersagen. Dazu werden die richtig positiven und richtig negativen sowie falsch positiven und falsch negativen Ergebnisse mit einbezogen. Hieraus lässt sich auch die Fehlerhäufigkeit über alle Vorhersagen berechnen (1 – Genauigkeit).

		vorhergesagte Klasse		
		positiv	negativ	
wahre Klasse	positiv	richtig positiv (TP)	falsch negativ (FN)	Sensitivität $\frac{TP}{TP+FN}$
	negativ	falsch positiv (FP)	richtig negativ (TN)	Spezifität $\frac{TN}{TN+FP}$
		Präzision $\frac{TP}{TP+FP}$	negativer prädiktiver Wert $\frac{TN}{TN+FN}$	Genauigkeit $\frac{TP+TN}{TP+TN+FP+FN}$

Abbildung 10: Berechnung verschiedener Leistungsmetriken
Das Zusammenspiel aus positiven und negativen Aussagen aus der vorhergesagten Klasse und der wahren/annotierten Klasse ergibt die richtig positiven, richtig negativen, falsch positiven und falsch negativen Ergebnisse. Aus diesen wiederum lassen sich anhand der dargestellten Formeln Sensitivität, Spezifität, Genauigkeit, Präzision und der negative prädiktive Wert berechnen. Abbildung vom Autor.

Das außerdem genutzte F-Maß beschreibt die harmonische Mittelung von Präzision und Sensitivität. Bei den Präzision-Sensitivität-Kurven werden dem Namen entsprechend Präzision und Sensitivität gegeneinander aufgetragen, bei den Isosensitivitätskurven die Präzision gegen 1 – Spezifität.

In Anpassung an die Ausgabesprache des Deep Learning Modells werden im Folgenden die englischen Begriffe verwendet (vgl. Tabelle 1). Mittelwerte von F1-Score, precision, recall sowie Precision-Recall-curves und AUROC von entweder mehreren Klassen oder als eine Zusammenfassung der einzelnen Durchgänge der cross validation (Kreuzvalidierungsverfahren) für jede individuelle Klasse wurde mit micro, macro, oder weighted averaging berechnet (Mikro-, Makro-, gewichtete Mittelung) (112). Die standard deviation (Standardabweichung) für gemittelte Precision-Recall curves und AUROC wurde berechnet. Weitere Angaben zu den Bedeutungen und Berechnungen dieser Metriken befinden sich im Anhang.

englisch	deutsch
recall	Sensitivität
specificity	Spezifität
precision	Präzision
accuracy	Genauigkeit
error rate	Fehlerhäufigkeit
F1-Score	F-Maß
PR-curves, Precision-Recall curves	Präzision-Sensitivität-Kurven
AUROC, area under the receiver operating characteristic	Isosensitivitätskurven
cross validation	Kreuzvalidierungsverfahren
micro averaging	Mikromittelung
macro averaging	Makromittelung
weighted averaging	gewichtete Mittelung
standard deviation	Standardabweichung

Tabelle 1: Englische und deutsche statistische Begriffsbezeichnungen
Die im Text verwendeten englischen Begriffsbezeichnungen (linksseitig) werden den deutschen Bezeichnungen gegenübergestellt (rechtsseitig).

3.9 Visualisierung

Die durch das Modell ausgegebene Klassifizierung auf Kachelebene wäre für sich alleine klinisch nicht nutzbar. Es wurden Möglichkeiten für eine Ausgabe des Modelles in einer für den Pathologen verwertbaren visuellen Form implementiert.

3.9.1 **„Intelligente Mikroskopie“**

Mit „intelligenter Mikroskopie“ kann man die Möglichkeit der Verarbeitung von Bildausschnitten, die dem Modell während des Mikroskopierens von einer Kamera zur Verfügung gestellt werden, beschreiben. Ziel ist die Ausgabe einer Klassifikation für diese Bildbereiche mit Wahrscheinlichkeitsverteilung für die verschiedenen Klassen zeitgleich zur analogen Mikroskopie.

Hierfür wurde im Mikroskop eine Kamera installiert, welche parallel zur Mikroskopie den vom Untersucher gesehenen Bildausschnitt filmt. Die aufgezeichneten ‚videoframes‘ werden dem Modell zur Verfügung gestellt. Durch dynamische Klassifikation der Einzelbilder des Videos wird so eine kontinuierliche Ausgabe der Klassifikationswahrscheinlichkeiten für die einzelnen Klassen erstellt werden. Diese Wahrscheinlichkeiten werden dem Untersucher als „erweiterte Realität“ neben dem zu

mikroskopierenden Bildausschnitt zur Verfügung gestellt. Mit Veränderung des Bildausschnittes und entsprechend der dargestellten Gewebeklasse, wird durch die dynamische Klassifikation der ‚videoframes' eine ebenso dynamische Anpassung der Klassifikationsausgaben ermöglicht.

3.9.2 **Klassifikationskarte**

Eine weitere Möglichkeit ist die Visualisierung in Form einer Klassifikationskarte. Hier wird schon auf whole slide Ebene eine farblich codierte Klassifikation der Gewebetypen ausgegeben. Realisiert wurde dies durch eine Klassifikation und Einfärbung auf Kachelebene mittels Anwendung eines sliding-window Ansatzes. Durch Hinterlegung der Kachelkoordinaten innerhalb des whole slides konnte so die Zusammensetzung einer vollständigen Klassifikationskarte erreicht werden.

3.10 Literaturrecherche

Begleitend wurde eine Literatursuche in den elektronischen Datenbanken von PubMed, GoogleScholar, Journal of Pathology Informatics und arxiv durchgeführt. Suchparameter waren zum einen Kombinationen der Schlagworte gastric signet ring cell carcinoma, signet ring cell carcinoma, gastric cancer. Zum anderen wurde mit Verknüpfungen der Schlagworte deep learning, pathology, digital pathology, artificial intelligence, machine learning, convolutional neural network, image analysis gesucht. Diese wurden in einem weiteren Rechercheprozess mit den Suchwörtern clinical application, application, clinical use, diagnostic assistance, histopathologic diagnosis, histopathologic classification verbunden. Es wurden weiterführend jeweils auch einige der in den gefundenen Publikationen herangezogenen Referenzen genutzt. Des Weiteren wurden Veröffentlichungen aus der Online-Bibliothek der Universitätsbibliothek Mainz (ins Besondere zum Thema maschinellen Lernen) und Veröffentlichungen von Krebsstatistiken des Robert-Koch-Instituts herangezogen.

4 Ergebnisse

4.1 Zusammensetzung der Kohorte

Insgesamt wurden nach den Kriterien aus Abschnitt 3.1 96 Patienten eingeschlossen. Die Kohorte bestand aus 56 Männern und 40 Frauen (vgl. Abb. 11), das Alter der Patienten lag zwischen 31 und 93 Jahren mit einem Mittelwert von 67 Jahren. Abbildung 11 veranschaulicht dabei die Aufteilung der Fälle in Alterskategorien zwischen 30 und 100. Das durchschnittliche Alter der Frauen wie auch das der Männer lag dabei bei 67 Jahren. Die Tumorausdehnung laut pathologischer TNM-Klassifikation (pT) lag am häufigsten bei pT3 (43%), am seltensten bei pT4 (10%). Für die Lymphknotenmetastasierung (pN) lag am häufigsten der Fall pN0 vor (41%), am seltensten pN2 (15%). Auch für diese Klassifikationen zeigt Abbildung 11 die Aufteilung in die Stadien pT1-pT4 beziehungsweise pN0-pN3.

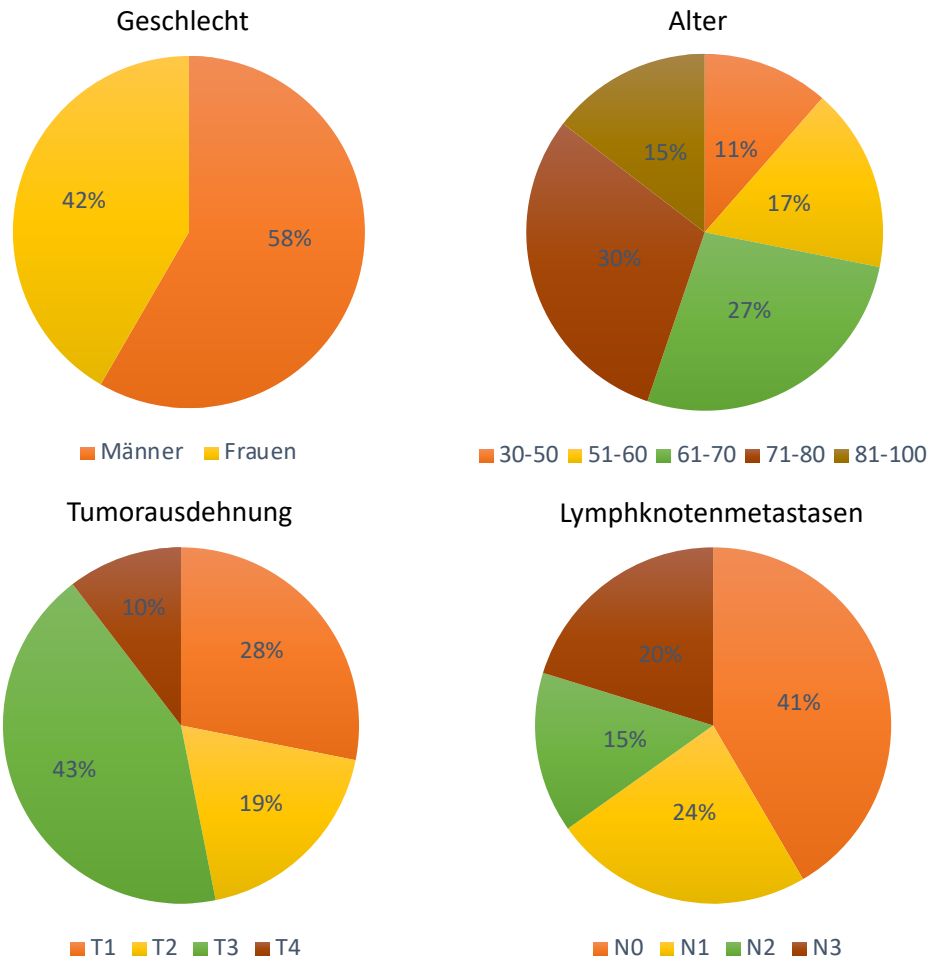


Abbildung 11: Zusammensetzung der Kohorte
Die Abbildung veranschaulicht die Zusammensetzung der Kohorte bezüglich Geschlecht, Alter, Tumorausdehnung (T1-4) und Ausprägung der Lymphknotenmetastasierung (N0-3). Abbildung vom Autor.

Metastasen beziehungsweise Fernmetastasen (pM) lagen in den meisten Fällen keine (pM0) vor. Für das Grading (pG) wurde am häufigsten pG3 beschrieben. Die

Ausdehnung der Tumoren lag zwischen 0,5 und 26cm mit einem durchschnittlichem Wert von 5,6 cm. In den meisten Fällen wurde der Resektionsstatus pR0, das heißt kein Vorliegen von Residualtumor nach der Operation, erreicht. In einigen Fällen wurde eine H.p. Gastritis beschrieben, in über einem Drittel der Fälle eine intestinale Metaplasie.

4.2 Zusammensetzung des Bilddatensatzes

Insgesamt 21501 Kacheln wurden aus der Annotation der Objektträger der 96 Patienten generiert, was einer mittleren Kachelanzahl von circa 224 pro Patient entspricht. Dabei wurden die meisten Bildkacheln für die Klasse Karzinom exportiert (8568 Kacheln), für die Klasse Immunzellen mit 222 Kacheln die wenigsten (vgl. Tabelle 2)

Gewebeklasse	Anzahl eingebrachter Kacheln
Epithel	7148
Fett	1074
Immunzellen	222
Muskel	4489
Karzinom	8568
gesamt	21501

Tabelle 2: Anzahl der eingebrachten Kacheln für die verschiedenen Gewebeklassen
Die Tabelle zeigt die Anzahl der eingebrachten Kacheln für die jeweilige Gewebeklasse sowie die Anzahl der insgesamt eingebrachten Kacheln für alle Klassen zusammen. Tabelle vom Autor.

4.3 Trainingsverlauf

Abbildung 12 veranschaulicht beispielhaft an Fold 4 die Trainingskurve des Deep Learning Modelles über die 30 Epochen hinweg. Die den Kurvenverläufen zugrunde liegenden Rohdaten befinden sich im Anhang. Der Kurvenverlauf für die accuracy ist beginnend mit 44,7% für Epoche 1 initial stark steigend, wobei ungefähr ab Epoche 7 (93,2%) eine Abflachung der Kurve mit einer nur noch geringen Verbesserung zu beobachten ist (Maximum 96,2% in Epoche 23). Die Kurven für train loss und validation loss (valid loss), die die Werte der loss function für Training und Validierung beschreiben, zeigen einen einander ähnlichen Verlauf. Nach einem initial starken Abfall bei Beginn mit Werten im Bereich von 1,3-1,4, flachen die Kurven ab ungefähr

Epoche 6 ab und schwanken nach weiterer Annäherung an 0 dann um Werte im Bereich von 0,06 beim train loss und 0,12 beim validation loss. Minimale Werte sind 0,053 für train loss und 0,116 für valid loss.

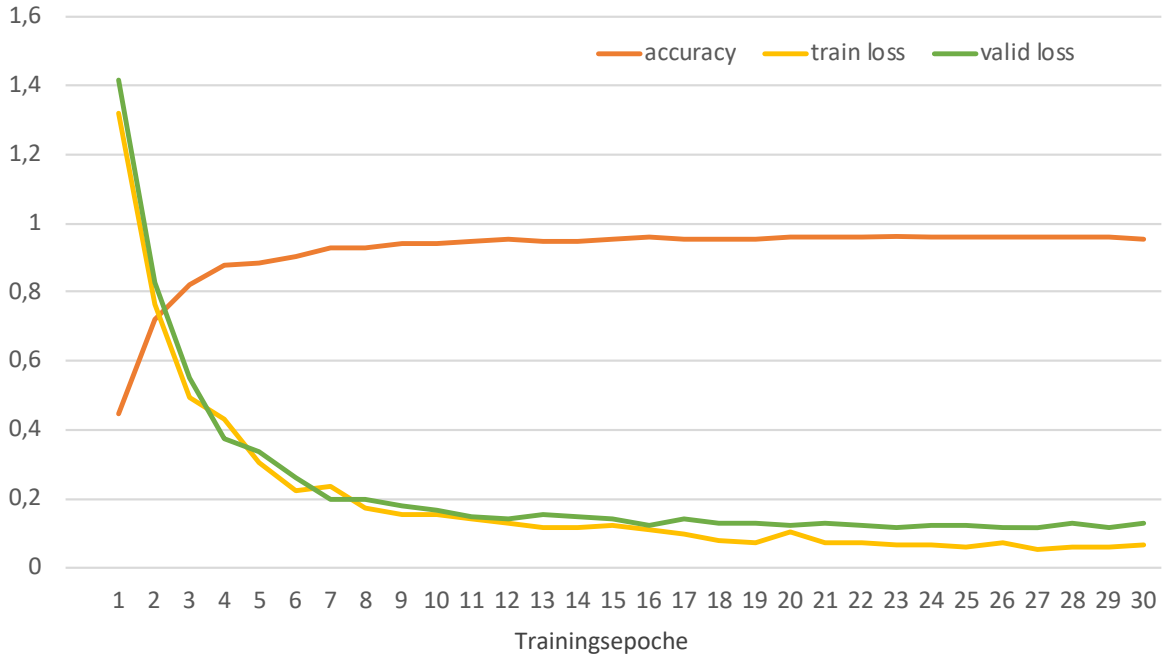


Abbildung 12: Trainingskurve des Deep Learning Modelles
Die Abbildung veranschaulicht den Verlauf von accuracy (im Rahmen der Validierung), train loss und validation loss (valid loss) innerhalb des Verlaufes der 30 Trainingsepochen beispielhaft für Fold 4. Abbildung vom Autor.

4.4 Klassifikationsexperimente

4.4.1 Deep Learning für Gewebeklassifikation

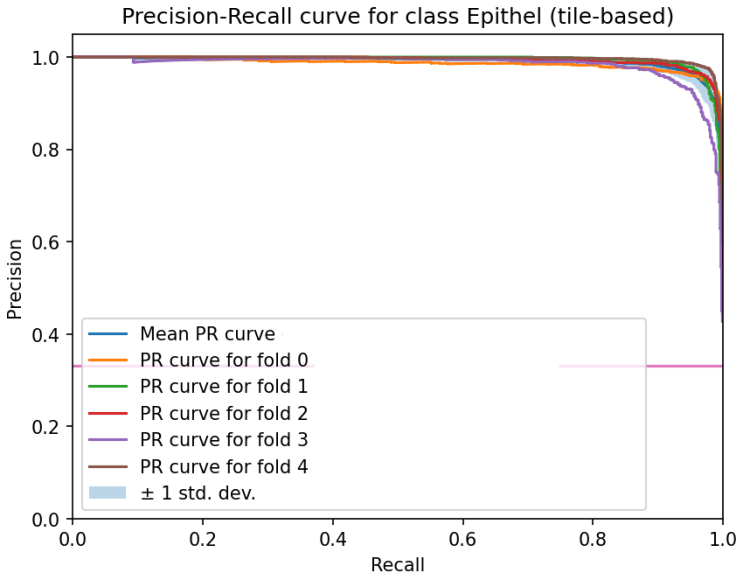
Im Rahmen dieser Studie sollte zunächst festgestellt werden, ob ein Deep Learning Modell dazu verwendet werden kann verschiedene Gewebetypen im Bereich des Magens (epitheliales Gewebe, Fettgewebe, Immunzellen, Muskelgewebe) beziehungsweise Magenkarzinomgewebe zu erkennen und zu klassifizieren. Es werden zunächst die Ergebnisse für die Klassifikation der einzelnen Gewebetypen betrachtet.

4.4.2 Klassifikation von epithelialem Gewebe

Wie in Abbildung 28 gezeigt, standen für die Klassifikation von epithelialem Gewebe 7148 wahre (als epitheliales Gewebe annotierte) Kacheln zur Verfügung. 7281 Kacheln wurden vom System insgesamt als epitheliales Gewebe klassifiziert, davon 6951 richtig, das heißt dem annotierten Label entsprechend. Mit 154 Kacheln das häufigste den annotierten Epithelkacheln falsch zugeordnete Label war das

Karzinomgewebe. Fälschlich als epitheliales Gewebe bezeichnet wurden am häufigsten als Karzinomgewebe annotierte Kacheln (322 Kacheln).

Hieraus ergab sich eine accuracy von 97,5%, ein recall von 97,24% sowie eine precision von 95,47%. Dies erbrachte einen F1-Score von 0,9635. Die PR-curves für das Label epitheliales Gewebe insgesamt sowie für die einzelnen Folds zeigt Abbildung 13. Tabelle 3 zeigt die zugehörigen durchschnittlichen AUC für die PR-curves. In Fold 4 wurde mit einer durchschnittlichen AUC von 0,996 die höchste, in Fold 3 mit 0,982 die niedrigste AUC erreicht. Die Mittelung über alle Folds lag bei 0,99 +/- 0,005.



fold	average AUC
	0,990 +/- 0,005
0	0,986
1	0,994
2	0,993
3	0,982
4	0,996

Abbildung 13: PR-curves für die Klassifikation Epithel
Die Abbildung zeigt farblich codiert die Precision-Recall-Kurven für alle Folds (0-4) für die Klassifikation des Gewebetyps Epithel sowie eine gemittelte Kurve über alle Folds.

Tabelle 3: Durchschnittliche AUC für die PR-curves der Klasse Epithel
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der Klasse Epithel. Die erste Zeile zeigt den Wert für die gemittelte Kurve über alle Folds mit Standardabweichung, die darauffolgenden für die Folds (0-4)

In Abbildung 14 sind die AUROC für die Klasse des epithelialen Gewebes dargestellt. Die mittlere AUC lag bei 0,996 +/- 0,001 mit der höchsten AUC für Fold 1 (0,997) und der niedrigsten für Fold 0 und 3 (0,994).

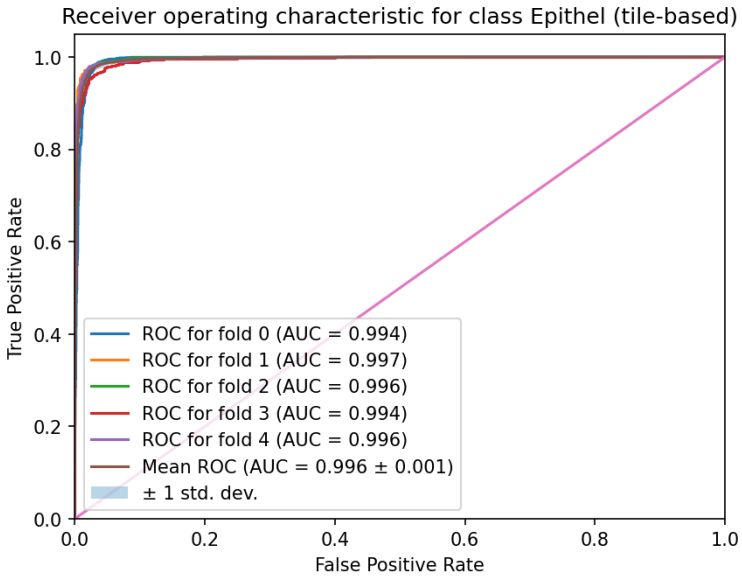


Abbildung 14: AUROC-curves für die Klassifikation von epithelialem Gewebe
Die Abbildung zeigt farblich codiert die AUROC-curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Epithel sowie eine gemittelte Kurve über alle Folds mit Angabe einer Standardabweichung. Die durchschnittlichen AUC für die einzelnen Kurven werden angegeben.

Abbildung 15 veranschaulicht die Sicherheit, die den Klassifikationsentscheidungen für das Label des epithelialen Gewebes zugrunde liegt. Hier zeigte sich eine maximale Sicherheit von 0,95 bis 1,0 für die absolute Mehrheit der Fälle.

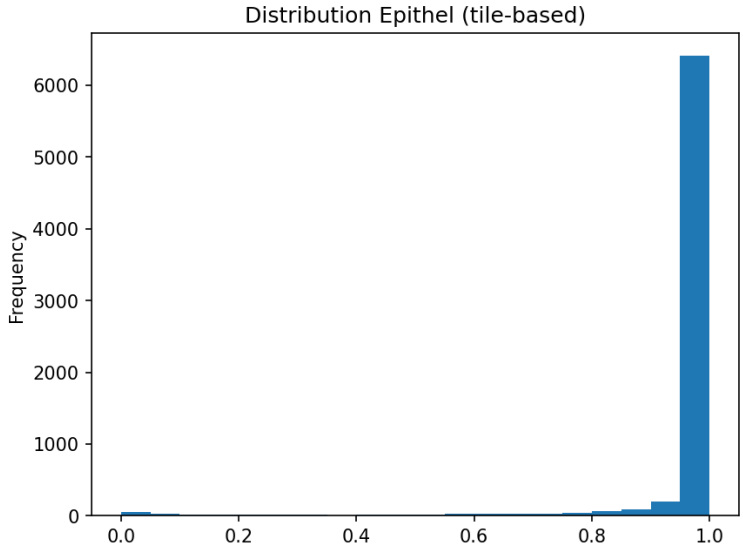
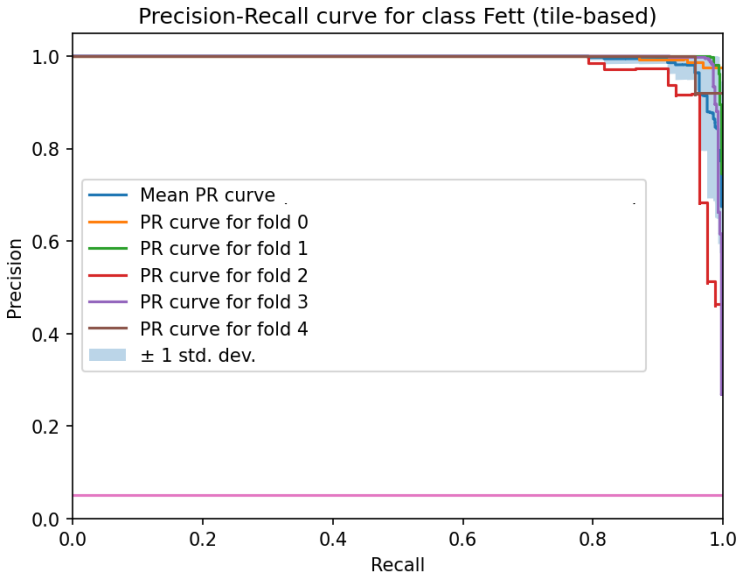


Abbildung 15: Balkendiagramm für die Verteilung der Sicherheit für die Klassifikationsentscheidung für epitheliales Gewebe
Die Abbildung zeigt die Sicherheit, die den Klassifikationsentscheidungen für das Label Epithel gemittelt über alle Folds zugrunde liegt. Die x-Achse codiert dabei in Balkenform von 0,05 Breite die Sicherheitsangabe mit Werten zwischen 0 und 1, die y-Achse die Anzahl der getroffenen Entscheidungen.

4.4.3 **Klassifikation von Fettgewebe**

Wie in Abbildung 28 dargestellt lagen für das Label Fettgewebe 1074 annotierte Kacheln vor, 1070 Mal wurde das Label Fettgewebe durch das System klassifiziert, in 1052 Fällen dem annotierten Label entsprechend. Am häufigsten der annotierten Klasse Fettgewebe fehlerhaft zugeordnet wurde das Label Muskelgewebe (13 Kacheln). Fälschlich als Fettgewebe bezeichnet wurde am häufigsten das annotierte Label Muskelgewebe (9 Kacheln). Hieraus ergab sich eine accuracy von 99,8%. Der recall für die Klasse Fettgewebe lag bei 97,95%, die precision bei 98,32%. Es ergab sich ein F1-Score von 0,9813. Abbildung 16 zeigt die PR-curves für die Klasse Fettgewebe. Tabelle 4 zeigt die zugehörigen durchschnittlichen AUC für die PR-curves. Für diese Klasse zeigte sich eine mittlere AUC von 0,993 +/- 0,008. In Fold 1 wurde mit einer durchschnittlichen AUC von 0,999 die höchste, in Fold 3 mit 0,995 die niedrigste AUC erreicht.



fold	average AUC
	0,993 +/- 0,008
0	0,998
1	0,999
2	0,977
3	0,995
4	0,997

Abbildung 16: PR-curves für die Klassifikation von Fettgewebe
Die Abbildung zeigt farblich codiert die Precision-Recall curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Fett sowie eine gemittelte Kurve über alle Folds.

Tabelle 4: Durchschnittliche AUC für die PR-curves der Klasse Fettgewebe
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der Klasse Fettgewebe. Die erste Zeile zeigt den Wert für die gemittelte Kurve über alle Folds mit Standardabweichung, die darauffolgenden für die einzelnen Folds (0-4).

Die AUROC für die Klasse Fettgewebe (vgl. Abb. 17) zeigten eine mittlere AUC von 0,999 +/- 0,000 bei den höchsten Werten von jeweils 1,00 in Fold 0,1 und 4 und den niedrigsten Werten von jeweils 0,999 in Fold 2 und 3.

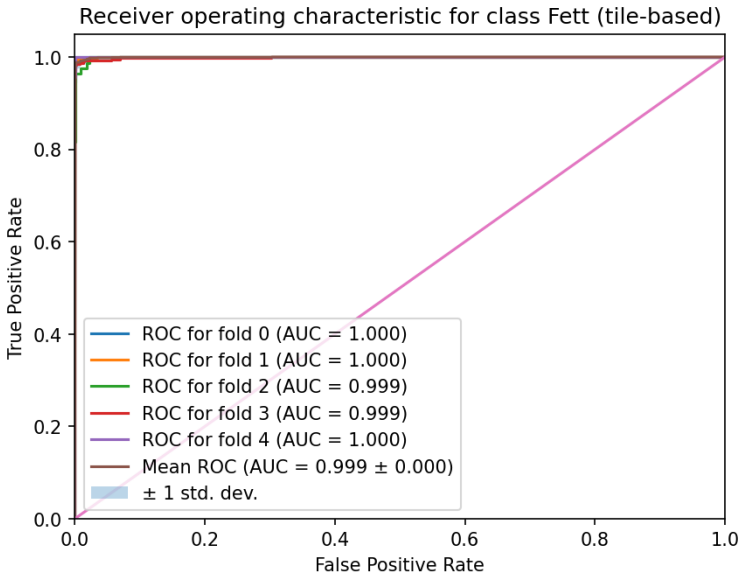


Abbildung 17: AUROC für die Klassifikation von Fettgewebe
Die Abbildung zeigt farblich codiert die AUROC für alle Folds (0-4) für die Klassifikation des Gewebetyps Fett sowie eine gemittelte Kurve über alle Folds mit Angabe einer Standardabweichung. Die durchschnittlichen AUC für die einzelnen Kurven werden angegeben.

Die Klassifikationsentscheidung Fettgewebe wurde in der absoluten Mehrheit der Fälle mit einer Sicherheit von 0,95 bis 1,0 getroffen (vgl. Abb. 18).

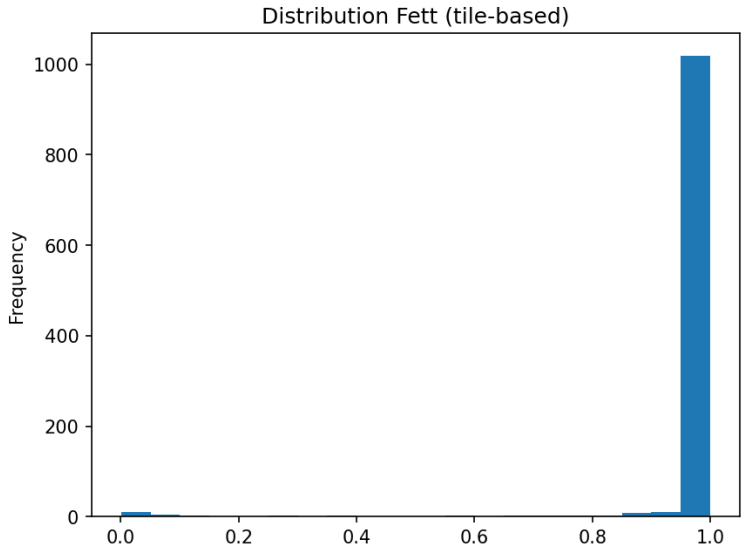
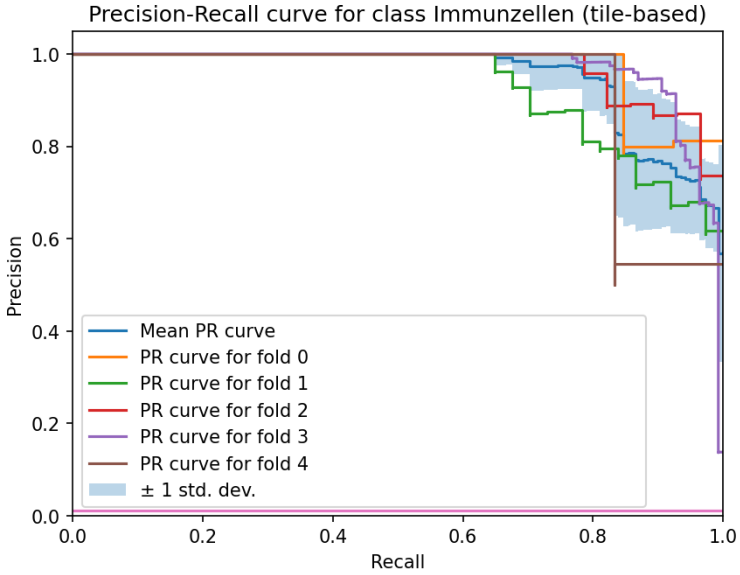


Abbildung 18: Balkendiagramm für die Verteilung der Sicherheit für die Klassifikationsentscheidung für Fettgewebe
Die Abbildung zeigt die Sicherheit, die den Klassifikationsentscheidungen für das Label Fettgewebe gemittelt über alle Folds zugrunde liegt. Die x-Achse codiert dabei in Balkenform von 0,05 Breite die Sicherheitsangabe mit Werten zwischen 0 und 1, die y-Achse die Anzahl der getroffenen Entscheidungen.

4.4.4 **Klassifikation von Immunzellen**

222 Kacheln wurden mit dem Label Immunzellen annotiert. Die Klassifikation Immunzellen gab das System bei 343 Kacheln aus, in 215 Fällen dem annotierten Label entsprechend. Am häufigsten fälschlicherweise statt des annotierten Labels Immunzellen ausgegeben wurde das Karzinom (4 Kacheln); 103 Kacheln von annotiertem Karzinom wurden falsch mit dem Label Immunzellen versehen (vgl. Abb. 28). Hieraus ergab sich eine accuracy von 99,4%. Der recall für die Klasse Immunzellen lag bei 96,85%, die precision bei 62,68%. Es ergab sich ein F1-Score von 0,7611. Die Precision-Recall curves der Klasse (vgl. Abb. 19) ergaben eine mittlere AUC von 0,953 +/- 0,022 in der Zusammenfassung aller Folds. Die höchste AUC wurde in Fold 2 mit 0,972 erreicht, die niedrigste mit 0,924 in Fold 4 (vgl. Tabelle 5).



fold	average AUC
	0,953 +/- 0,022
0	0,970
1	0,927
2	0,972
3	0,970
4	0,924

Abbildung 19: PR-curves für die Klassifikation von Immunzellen
Die Abbildung zeigt farblich codiert die Precision-Recall curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Immunzellen sowie eine gemittelte Kurve über alle Folds.

Tabelle 5: Durchschnittliche AUC für die PR-curves der Klasse Immunzellen
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der Klasse Immunzellen. Die erste Zeile zeigt den Wert für die gemittelte Kurve über alle Folds mit Standardabweichung, die darauffolgenden für die einzelnen Folds (0-4).

Die AUROC für die Klasse Immunzellen (vgl. Abb. 20) ergab 0,999 +/- 0,001 für alle Folds zusammen. Die höchste AUC lieferten Fold 0,2 und 4 mit 1,0, die niedrigste Fold 3 mit 0.997.

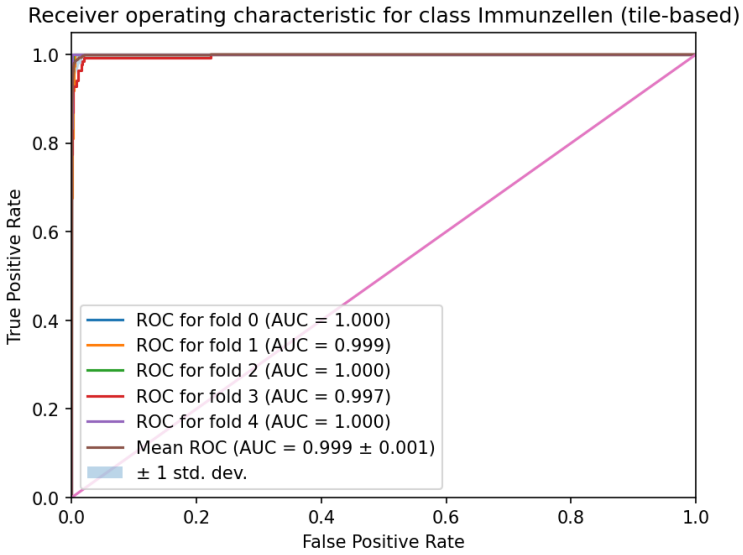


Abbildung 20: AUROC für die Klassifikation von Immunzellen
Die Abbildung zeigt farblich codiert die AUROC für alle Folds (0-4) für die Klassifikation des Gewebetyps Immunzellen sowie eine gemittelte Kurve über alle Folds mit Angabe einer Standardabweichung. Die durchschnittlichen AUC für die einzelnen Kurven werden angegeben.

In der Darstellung der Sicherheit mit welcher die Klassifikationsentscheidung Immunzellen durch das System getroffen wurde, zeigte sich eine Sicherheit von 0,95 bis 1,0 in der absoluten Mehrheit der Fälle (vgl. Abb. 21).

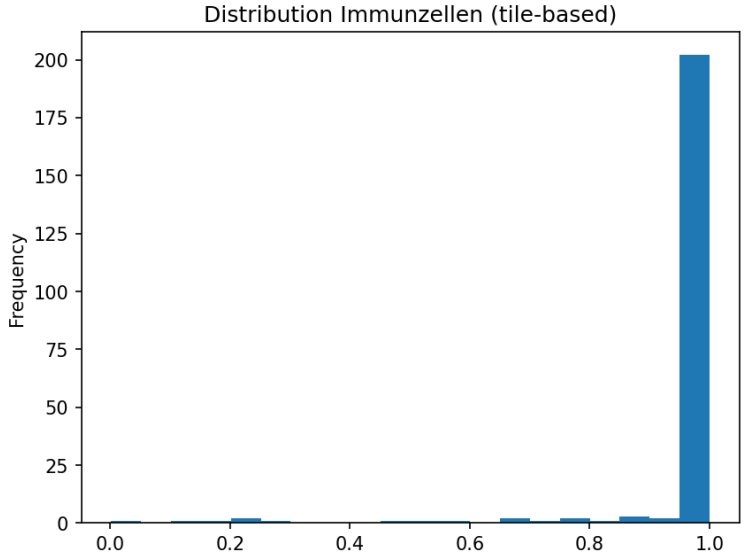
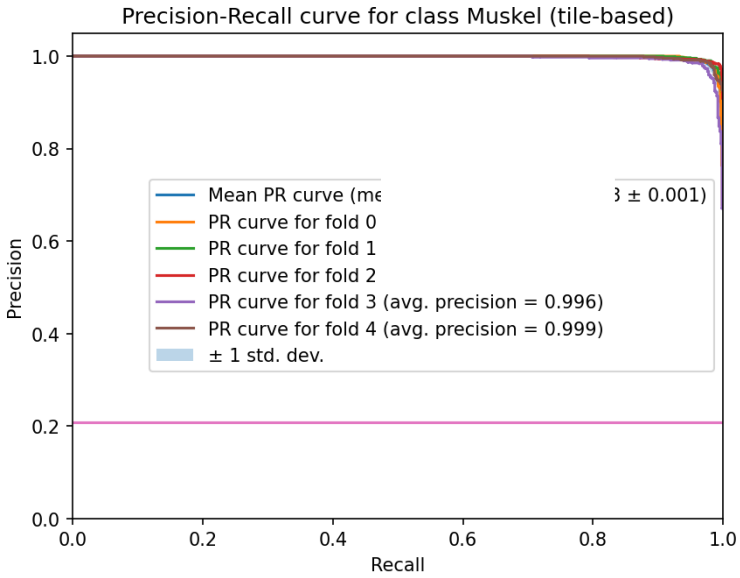


Abbildung 21: Balkendiagramm für die Verteilung der Sicherheit für die Klassifikationsentscheidung für Immunzellen
Die Abbildung zeigt die Sicherheit, die den Klassifikationsentscheidungen für das Label Immunzellen gemittelt über alle Folds zugrunde liegt. Die x-Achse codiert dabei in Balkenform von 0,05 Breite die Sicherheitsangabe mit Werten zwischen 0 und 1, die y-Achse die Anzahl der getroffenen Entscheidungen.

4.4.5 **Klassifikation Muskelgewebe**

Für die Klassifikation Muskelgewebe standen 4489 annotierte Kacheln zur Verfügung. 4504 Mal wurde die Klassifikation Muskelgewebe durch das System ausgegeben, 4408 Mal dem wahren Label entsprechend. Anstelle des richtigen Labels Muskelgewebe wurde am häufigsten das Karzinomgewebe angegeben (67 Kacheln). Fälschlicherweise als Muskelgewebe vorhergesagt wurde mit 64 Kacheln am häufigsten das annotierte Label Karzinomgewebe (vgl. Abb. 28). Hieraus ergab sich eine accuracy von 99,2%. Der recall für die Klasse Muskelgewebe lag bei 98,2%, die precision bei 97,87%. Es ergab sich ein F1-Score von 0,9803. Die Precision-Recall curve für die Klasse Muskelgewebe (vgl. Abb. 22) zeigte eine AUC von 0.998 +/- 0,001 als Mittelwert für alle Folds. Die höchste AUC ergab sich in Fold 1,2 und 3 mit jeweils 0,999, die niedrigste mit 0,996 in Fold 3 (vgl. Tabelle 6).



fold	average AUC
	0,998 +/- 0,001
0	0,998
1	0,999
2	0,999
3	0,996
4	0,999

Abbildung 22: PR-curves für die Klassifikation von Muskelgewebe
Die Abbildung zeigt farblich codiert die Precision-Recall curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Muskel sowie eine gemittelte Kurve über alle Folds.

Tabelle 6: Durchschnittliche AUC für die PR-curves der Klasse Muskelgewebe
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der Klasse Muskelgewebe. Die erste Zeile zeigt den Wert für die gemittelte Kurve über alle Folds mit Standardabweichung, die darauffolgenden für die einzelnen Folds (0-4).

Die AUROC für die Klasse Muskelgewebe ergab 0,999 +/- 0,000 als Mittelwert für alle Folds. Die höchste AUC für diese Klasse ergab sich in Fold 1 und 4 mit jeweils 1, 0, die niedrigste in Fold 0, 2 und 4 mit jeweils 0,999 (vgl. Abb. 23).

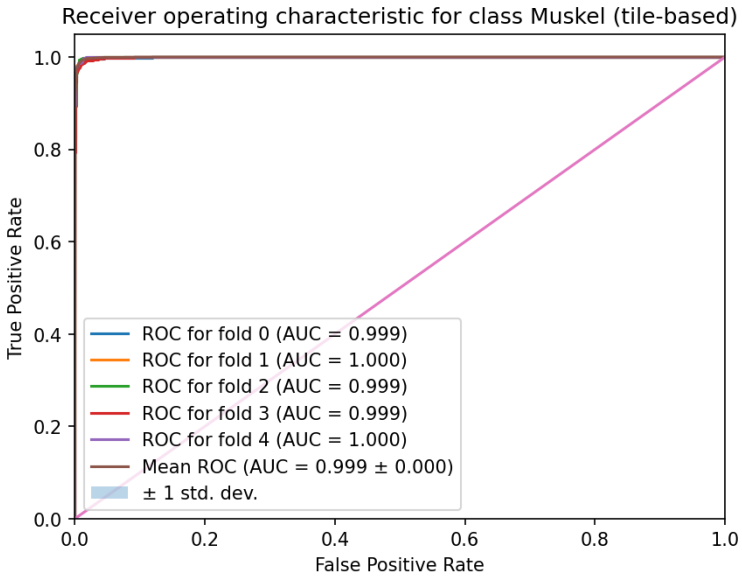


Abbildung 23: AUROC für die Klassifikation von Muskelgewebe
Die Abbildung zeigt farblich codiert die AUROC für alle Folds (0-4) für die Klassifikation des Gewebetyps Muskel sowie eine gemittelte Kurve über alle Folds mit Angabe einer Standardabweichung. Die durchschnittlichen AUC für die einzelnen Kurven werden angegeben.

In der absoluten Mehrheit der Fälle zeigte sich eine Sicherheit von 0,95 bis 1,0 in der Klassifikationssicherheit des Systems für das Label Muskelgewebe (vgl. Abb. 24).

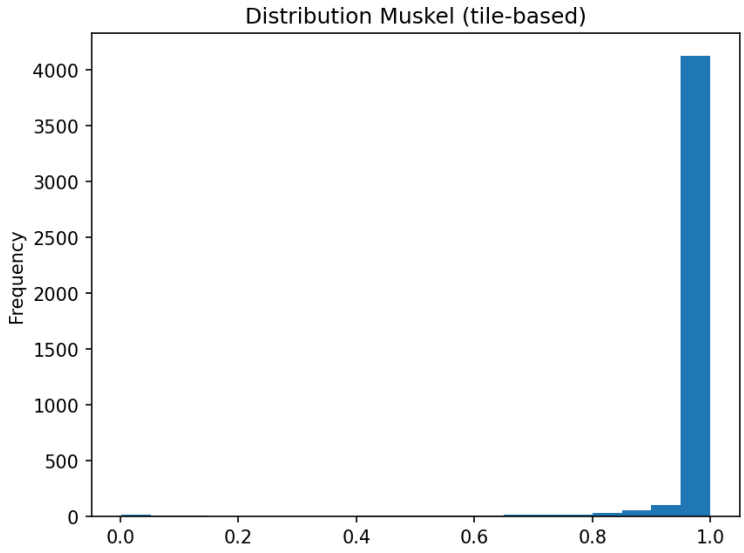
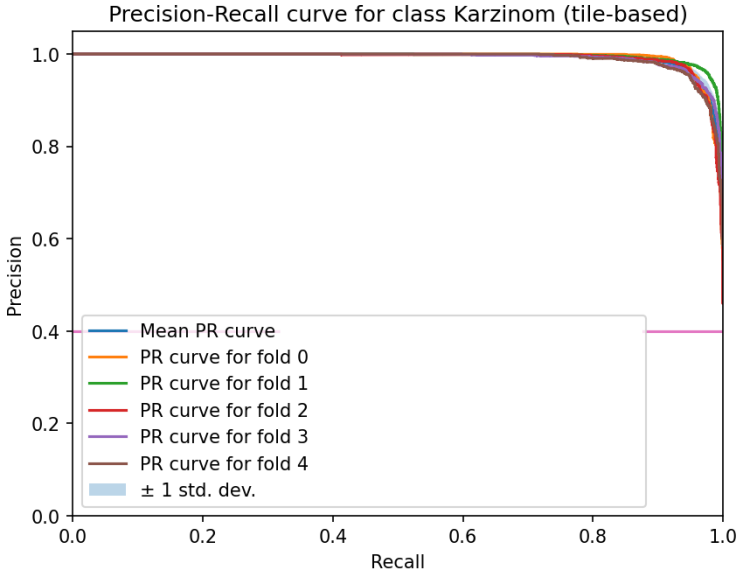


Abbildung 24: Balkendiagramm für die Verteilung der Sicherheit für die Klassifikationsentscheidung für Muskelgewebe
Die Abbildung zeigt die Sicherheit, die den Klassifikationsentscheidungen für das Label Muskel gemittelt über alle Folds zugrunde liegt. Die x-Achse codiert dabei in Balkenform von 0,05 Breite die Sicherheitsangabe mit Werten zwischen 0 und 1, die y-Achse die Anzahl der getroffenen Entscheidungen.

4.4.6 **Klassifikation Karzinomgewebe**

8568 Kacheln wurden mit dem Label Karzinomgewebe annotiert. Die Ausgabe der Klassifikation Karzinomgewebe erfolgte 8303 Mal, bei 8073 Kacheln davon dem richtigen Label entsprechend. Mit 322 Kacheln wurde das annotierte Label Karzinomgewebe am häufigsten falsch als epitheliales Gewebe ausgegeben. Fälschlicherweise als Karzinomgewebe angegeben wurde mit 154 Kacheln vor allem die annotierte Klasse Epithel (vgl. Abb. 28). Hieraus ergab sich eine accuracy von 96,6%. Der recall für die Klasse Karzinomgewebe lag bei 94,22%, die precision bei 97,23%. Es ergab sich ein F1-Score von 0,957. Die Precision-Recall curves für die Klasse Karzinom (vgl. Abb. 25) zeigte eine mittlere AUC von 0,992 +/- 0,002 zusammenfassend für alle Folds. Fold 1 erbrachte mit 0,995 die höchste AUC, Fold 4 mit 0,99 die niedrigste (vgl. Tabelle 7).



fold	average AUC
	0,992 +/- 0,002
0	0,992
1	0,995
2	0,991
3	0,992
4	0,990

Abbildung 25: PR-curves für die Klassifikation von Karzinomgewebe
Die Abbildung zeigt farblich codiert die Precision-Recall curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Karzinom sowie eine gemittelte Kurve über alle Folds.

Tabelle 7: Durchschnittliche AUC für die PR-curves der Klasse Karzinomgewebe
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der Klasse Karzinomgewebe. Die erste Zeile zeigt den Wert für die gemittelte Kurve über alle Folds mit Standardabweichung, die darauffolgenden für die einzelnen Folds (0-4).

Abbildung 26 zeigt die AUROC-curves für die Klasse Karzinomgewebe. Es ergab sich eine mittlere AUC von 0,994 +/- 0,001 für die Mittelung über alle Folds. Die höchste AUC für diese Klasse lag bei 0,996 in Fold 1, die niedrigste bei 0,992 in Fold 3.

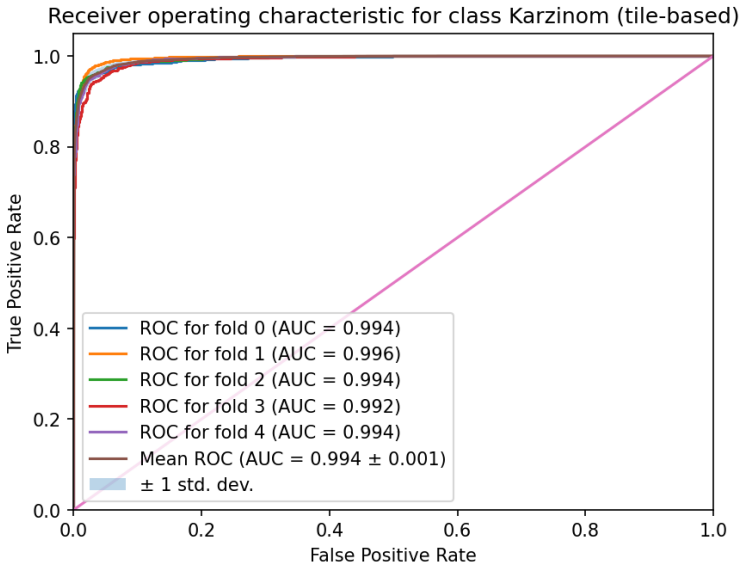


Abbildung 26: AUROC für die Klassifikation von Karzinomgewebe
Die Abbildung zeigt farblich codiert die AUROC-curves für alle Folds (0-4) für die Klassifikation des Gewebetyps Karzinom sowie eine gemittelte Kurve über alle Folds mit Angabe einer Standardabweichung. Die durchschnittlichen AUC für die einzelnen Kurven werden angegeben.

In der Darstellung der Sicherheit mit welcher die Klassifikationsentscheidung Karzinomgewebe durch das System getroffen wurde, zeigte sich eine Sicherheit von 0,95 bis 1,0 in der absoluten Mehrheit der Fälle (vgl. Abb. 27).

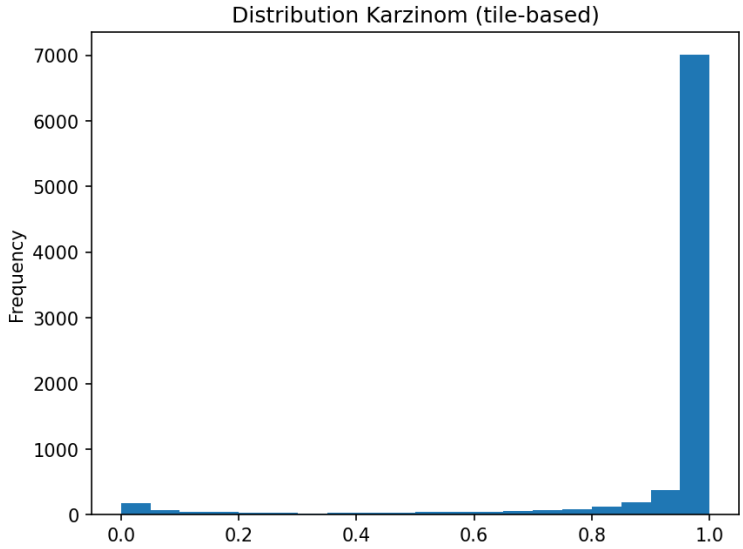


Abbildung 27: Balkendiagramm für die Verteilung der Sicherheit für die Klassifikationsentscheidung für Karzinomgewebe
Die Abbildung zeigt die Sicherheit, die den Klassifikationsentscheidungen für das Label Karzinomgewebe gemittelt über alle Folds zugrunde liegt. Die x-Achse codiert dabei in Balkenform von 0,05 Breite die Sicherheitsangabe mit Werten zwischen 0 und 1, die y-Achse die Anzahl der getroffenen Entscheidungen.

4.4.7 **Gesamtsystemleistung für die Gewebeklassifikation**

Die folgenden Ergebnisse zeigen nun die Gesamtleistung des Systems im Klassifikationsexperiment gemittelt über alle Gewebetypen.

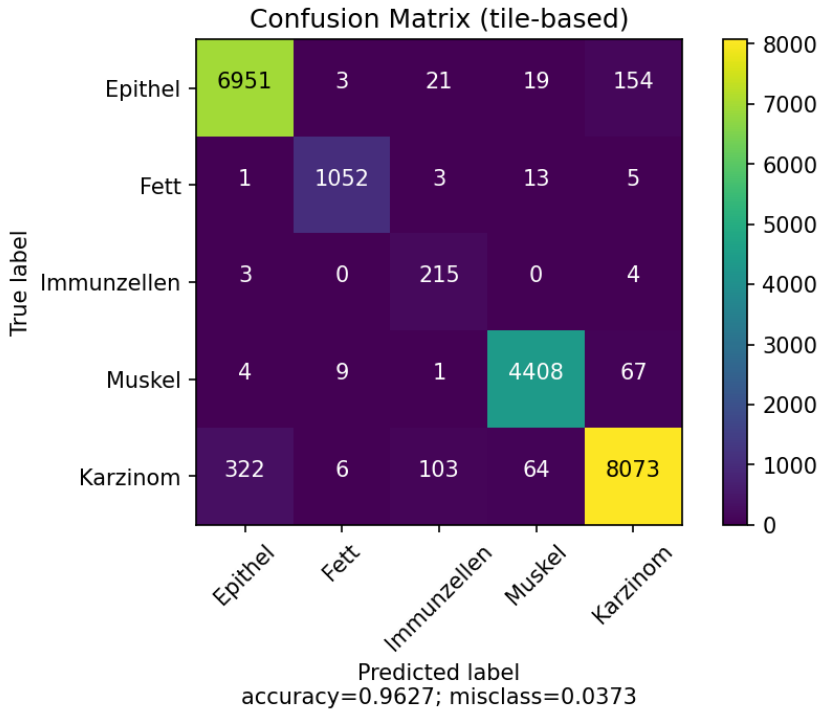


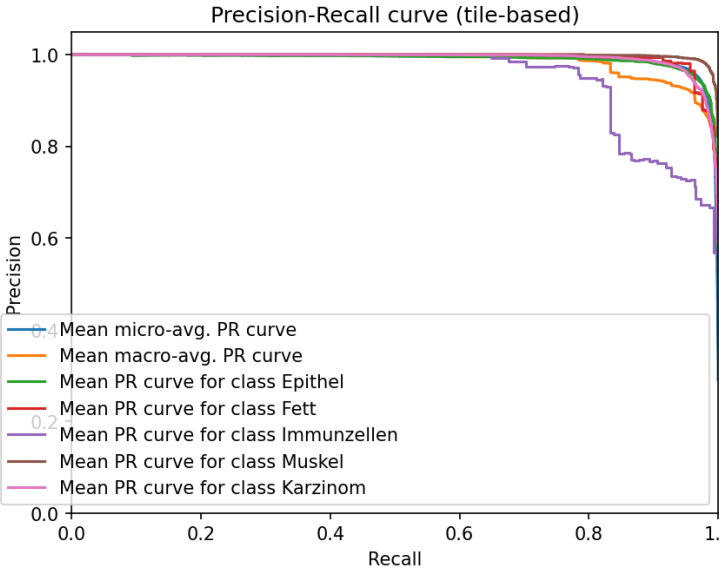
Abbildung 28: Gemittelte confusion matrix über alle Folds
Die Matrix stellt den annotierten Labeln die durch das System ausgegebenen Label für die einzelnen Gewebetypen Epithel, Fett, Immunzellen, Muskel und Karzinom farblich nach Höhe der Kachelzahl codiert gegenüber. In der Diagonalen von links oben nach rechts unten liegen die passend zum annotierten Label ausgegebenen Kacheln, daneben die fehlerhaft klassifizierten Kacheln mit jeweiligem Bezug zum eigentlich wahren Label.

Insgesamt wurde an 21501 Kacheln validiert. 20699 Kacheln wurden dabei richtig dem wahren Label zugeordnet, 802 Kacheln wurden falsch zugeordnet. Die häufigsten Fehler lagen hier bei der fälschlichen Bezeichnung von annotiertem Karzinomgewebe als epitheliales Gewebe (322 Fälle), annotiertem epithelialem Gewebe als Karzinomgewebe (154 Fälle) und annotiertem Karzinomgewebe als Immunzellen (103 Fälle). Vergleiche hierzu Abbildung 28. Hieraus ergab sich eine accuracy von 96,27% mit einem Anteil von Fehlklassifikationen von 3,73% (vgl. Tabelle 8). Bei einer Teilbetrachtung der einzelnen Folds ergab sich die höchste accuracy von 96,89% und eine Fehlklassifikation von 3,11% für Fold 2 und die niedrigste für Fold 3 mit 95,3% und einer Fehlklassifikation von 4,7% (vgl. Tabelle 8).

Fold	accuracy	misclass
0	96,17%	3,83%
1	96,71%	3,29%
2	96,89%	3,11%
3	95,3%	4,7%
4	96,21%	3,79%
gesamt	96,27%	3,73%

Tabelle 8: accuracy und misclass für die einzelnen Folds und gemittelt über alle Folds

Der recall der Gesamtklassifikation lag damit bei Mittelung nach macro average bei 96,89%, bei Mittelung nach weighted average bei 96,27%. Für die precision ergaben sich nach macro average 90,31%, nach weighted average 96,48%. Hieraus ergaben sich F1-Scores von 0,9286 bei Mittelung mit macro average und von 0,9632 bei Mittelung mit weighted average. Die Precision-Recall curves in Abbildung 29 zeigen die Kurvenverläufe für alle Klassen gemittelt über alle Folds. Die zugehörige Tabelle 9 zeigt die entsprechenden durchschnittlichen AUC für die PR-curves. Die höchste AUC lag mit 0,998 +/- 0,001 bei der Klasse Muskelgewebe, die niedrigste mit 0,953 +/- 0,022 bei der Klasse Immunzellen. Außerdem dargestellt ist die Mittelung der Werte aller Klassen mittels micro und macro average. Hier zeigte sich für micro average eine mittlere AUC von 0,991 +/- 0,002, für macro average eine mittlere AUC von 0,985 +/- 0,003. Bei Teilbetrachtung der micro average PR-curves für die einzelnen Folds (Abb. siehe Anhang) lag die höchste gemittelte AUC bei 0,994 in Fold 1, die niedrigste mit 0,987 in Fold 3. Bei macro average Mittelung lag der höchste Wert mit 0,989 in Fold 0, der niedrigste mit 0,981 in Fold 4 (Abb. siehe Anhang).



	average AUC
micro average	0,991 +/- 0,002
macro average	0,985 +/- 0,003
Epithel	0,990 +/- 0,005
Fett	0,993 +/- 0,008
Immunzellen	0,953 +/- 0,022
Muskel	0,998 +/- 0,001
Karzinom	0,992 +/- 0,002

Abbildung 29: Gemittelte PR-curves für jede Klasse und für Mittelwerte über alle Klassen
Es werden farblich codiert die gemittelten PR-curves über alle Folds für jede gezeigt. Weiterhin sind PR-curves für den Mittelwert über alle Klassen nach micro und macro average Mittelung dargestellt.

Tabelle 9: Durchschnittliche AUC für die PR-curves für jede Klasse und für Mittelwerte über alle Klassen
Die Tabelle zeigt die durchschnittliche AUC für die PR-curves der einzelnen Gewebeklassen sowie für die gemittelten PR-curves nach micro und macro average mit jeweiliger Angabe der Standardabweichung.

Abbildung 30 zeigt die AUROC für die verschiedenen Klassen gemittelt über alle Folds. Die höchste lag mit 0,999 in den Klassen Fettgewebe (+/- 0,000) Immunzellen (+/- 0,001) und Muskelgewebe (+/- 0,000), die niedrigste bei dem Label Karzinomgewebe mit 0,994 +/- 0,001. Außerdem dargestellt ist die Mittelung der Werte aller Klassen mittels micro und macro average. Hier ergab sich jeweils (für micro und macro average) eine AUC von 0,997 +/- 0,001. Bei Teilbetrachtung der micro average AUC der AUROC-curves für die einzelnen Folds (Abb. siehe Anhang) lag die höchste gemittelte AUC in Folds 1 und 4 mit jeweils 0,998, die niedrigste mit 0,996 in Folds 0 und 3. Bei macro average Mittelung lag der höchste Wert in Fold 1 mit 0,999, der niedrigste in Fold 0,996 in Fold 3 (Abb. siehe Anhang).

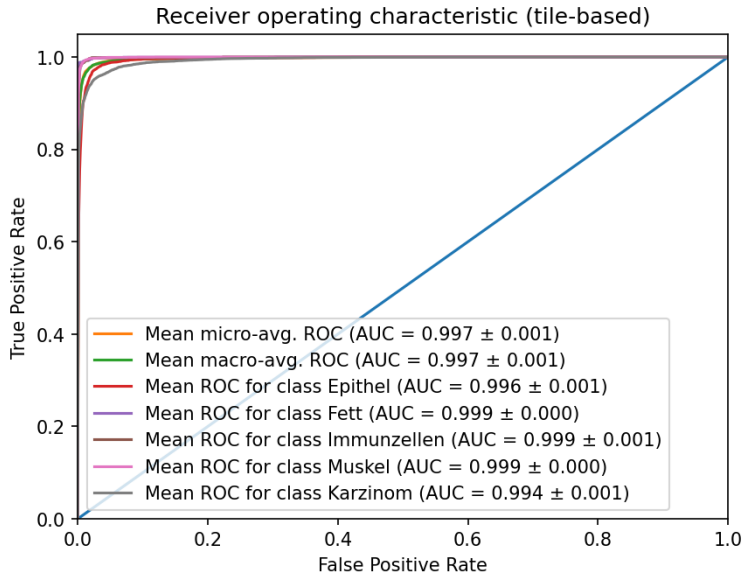


Abbildung 30: Gemittelte AUROC für jede Klasse und für Mittelwerte über alle Klassen
Es werden farblich codiert die gemittelten AUROC über alle Folds für jede Klasse mit Angabe der gemittelten AUC und deren Standardabweichung gezeigt. Weiterhin sind AUROC-curves für den Mittelwert über alle Klassen nach micro und macro average Mittelung dargestellt.

4.4.8 Ergebnisse der Visualisierung

4.4.8.1 Ergebnisse zur „Intelligenten Mikroskopie“

Abbildung 31 veranschaulicht die Ergebnisse des Visualisierungsexperimentes zur „Echtzeitanalyse“ von Bildausschnitten unter dem Mikroskop. Es konnte eine annähernde Reproduzierbarkeit der Genauigkeitswerte für die Analyse bei klinisch kaum bemerkbarer Zeitverzögerung der Ausgabe der Klassifikationen erreicht werden.

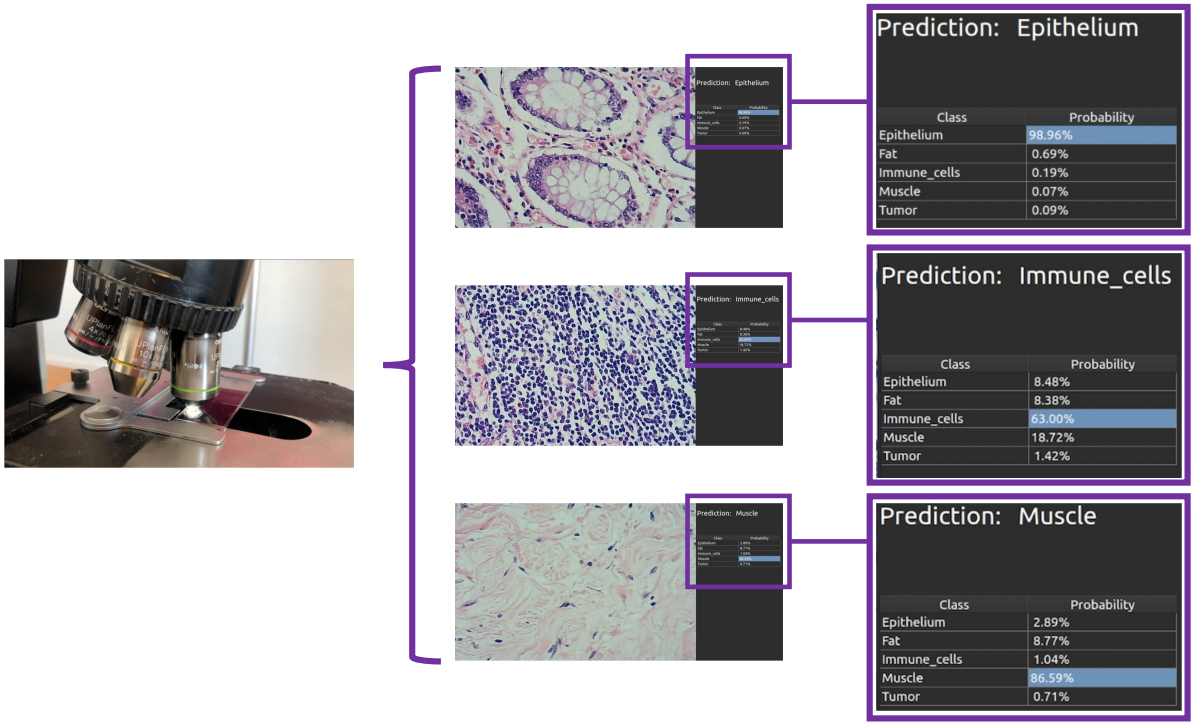


Abbildung 31: Beispielausschnitte aus der „Echtzeitanalyse“ von Gewebeschnitten unter dem Mikroskop

H.E. gefärbte Gewebeschnitte werden unter dem Mikroskop betrachtet, wobei die Bildausschnitte nahezu zeitgleich durch das Modell klassifiziert werden. Die ausgegebene Klassifikation wird angezeigt, daneben die Wahrscheinlichkeit in Prozent (%), die das Modell den fünf Klassifikationsoptionen zuschreibt (Epithel, Fett, Immunzellen, Muskel, Karzinom). Hier werden beispielhaft Bildausschnitte von Epithel, Immunzellen und Muskel dargestellt. Bilder aus der Arbeitsgruppe Foersch et al.

4.4.8.2 Ergebnisse zur Klassifikationskarte

Abbildung 32 veranschaulicht die Ergebnisse des Visualisierungsexperimentes zur Erstellung einer Klassifikationskarte.

Hierbei wurde für jede der einzelnen Kacheln innerhalb des digitalisierten Ganztumorschnittes eine farbliche Markierung entsprechend der Klassifikationswahrscheinlichkeit für Karzinomgewebe vorgenommen. Die Farbkodierung von blau bis rot entspricht dabei der Vorhersagesicherheit für Karzinomgewebe zwischen 0 und 1 beziehungsweise zwischen 0 und 100%. Es entsteht eine Karte, die Bereiche mit hoher Klassifikationswahrscheinlichkeit für Karzinomgewebe farblich hervorhebt.

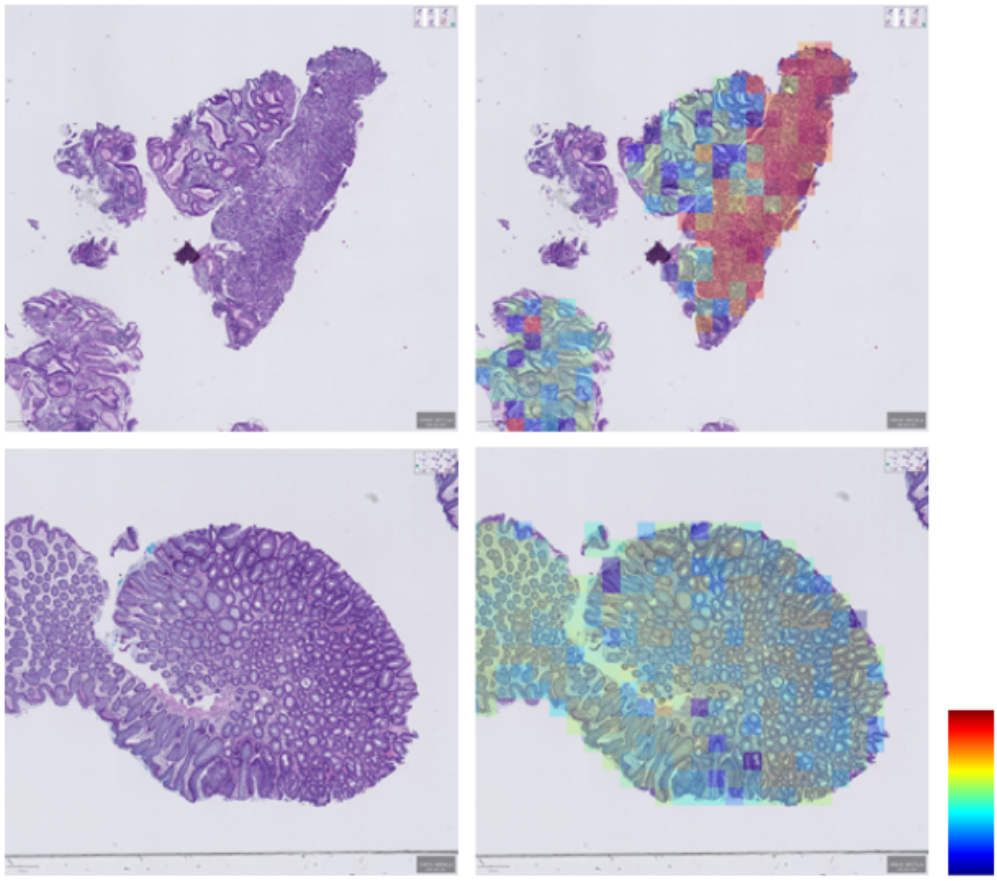


Abbildung 32: Klassifikationskarte für Karzinomgewebe
Visualisierung der Vorhersagesicherheit für Karzinomgewebe für die einzelnen Kacheln projiziert auf den Ganzgewebeschnitt anhand von deren Koordinaten. Es entsteht eine Klassifikationskarte mit farblicher Codierung der Vorhersagesicherheit von 0-100% auf den dargestellten Farbbalken von blau bis rot. Im oberen Beispielbild von nebeneinandergestelltem Original und Klassifikationskarte ist dabei farblich ein Areal mit hoher Vorhersagesicherheit für Karzinomgewebe abzugrenzen, im unteren hingegen bei niedrigerer Sicherheit Darstellung einer geringen Wahrscheinlichkeit für Karzinomgewebe.

4.4.9 Einsatzmöglichkeiten eines Deep Learning Klassifikationsmodelles in der Pathologie

Weitere Zielsetzung der Studie war die Recherche und Abwägung verschiedener Einsatzmöglichkeiten eines Deep Learning Modelles für Gewebeklassifikationen im Magenbereich in der klinischen Pathologie (vgl. Abb. 33). Nachfolgend werden die Ergebnisse der Literaturrecherche für die möglichen Verwendungsoptionen dargestellt.

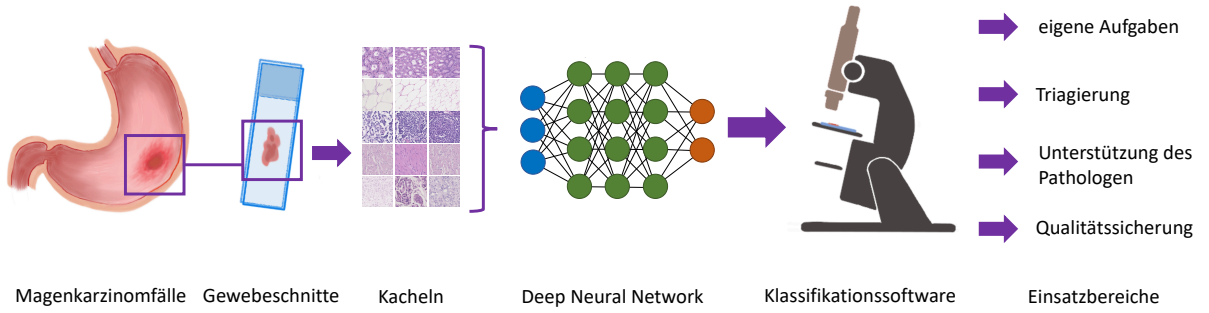


Abbildung 33: Entstehung und Einsatzmöglichkeiten eines Deep Learning Klassifikationsmodelles
Nach Auswahl der Magenkarzinomfälle und der zugehörigen Gewebeschnitte werden die daraus erstellten Kacheln einem Deep Neural Network zum Training zur Verfügung gestellt: Eine Klassifikationssoftware entsteht. Diese kann in verschiedenen klinischen Bereichen zum Einsatz kommen (eigenständige Übernahme von Aufgaben, Triagierung, Unterstützung der pathologischen Kompetenz, Qualitätssicherung). Abbildung vom Autor.

4.4.9.1 Eigenständige Übernahme von Aufgaben

Derzeitiges Einsatzgebiet für Verfahren der künstlichen Intelligenz beziehungsweise des maschinellen Lernens sind vor allem weniger komplexe, repetitive Aufgaben (113). Damit könnten unkompliziertere Arbeitsvorgänge, die in der Praxis jedoch zeit- und arbeitsintensiv sind, automatisiert werden und den Medizinern mehr Arbeitszeit für die komplexeren Aufgaben lassen, was eine verbesserte Nutzung des begrenzten menschlichen Kapitals darstellen würde (114). Zudem können die Problemstellungen maschinell mit höherer Geschwindigkeit und Reproduzierbarkeit durchgeführt werden, was den allgemeinen Arbeitsablauf beschleunigen und qualitativ verbessern würde (114). Beispiele aus der pathologischen Routinediagnostik sind hier die hohe Zahl von Falleinsendungen mit ähnlichen Gewebeschnitten aus großen Screeningprogrammen wie Dickdarmkrebs- oder Prostatakrebsvorsorge (113). Durch Übernahme von repetitiver Sichtung und Selektion durch ein Deep Learning Modell, könnte der Pathologe seine Zeit auf die anspruchsvollen weiterführenden Arbeiten wie die Interpretation von molekularpathologischen Tests verschieben (113). Dass Deep Learning Modelle gerade unter Zeitbeschränkung dem Menschen ebenbürtige oder sogar überlegene Klassifikationsleistungen erbringen können, belegt beispielhaft die

Veröffentlichung von Iizuka et al. aus dem Jahr 2020, bei der Pathologen unter Zeitdruck bei der Klassifikation von Magen- und Darmgewebe (Unterscheidung nicht-neoplastisch, Adenom, Adenokarzinom) eine durchschnittliche Genauigkeit von circa 86%, das trainierte Modell hingegen eine Genauigkeit von 96% erreichte (105). In der Arbeit von Litjens et al. aus dem Jahr 2016 wurde ein Deep Learning Modell als ein „Karzinom-Screeninginstrument“ für Prostatabiopsien und Brustkrebsmetastasen in Sentinel-Lymphknoten eingesetzt (67). 30-40% der Schnitte mit gutartigem Gewebe konnten durch das Modell vor Beurteilung durch einen Pathologen bereits aussortiert werden, ohne dass dadurch Fälle von Karzinom übersehen worden wären, was in Anbetracht der hohen Fallzahlen in der täglichen Routine eine große Reduktion der Arbeitsbelastung bedeutet (67, 103). In einer weiteren Veröffentlichung zu diesem Thema von Campanella et al. konnten im Screening sogar 65-75% aller Schnitte unter Beibehaltung einer Sensitivität von 100% aussortiert werden (115).

4.4.9.2 Triagierung

Ähnlich dem Einsatzgebietes des Screenings, könnten Deep Learning Modelle für eine Triagierung genutzt werden. Hierbei würde das Modell nicht eigenständig Fälle auf Basis eigener Klassifikationsentscheidungen aussortieren, sondern diese lediglich nach Relevanz und Dringlichkeit für den Pathologen vorsortieren (114). Damit würden die Ressourcen auf die Patienten dirigiert, die die größte Wahrscheinlichkeit für einen Bedarf an medizinischer Betreuung haben (114). Gleichzeitig könnte die hieraus gewonnene Zeitersparnis zu einem früheren Erkennen von Erkrankungen führen, was in einigen Fällen positive Konsequenzen bezüglich Behandlung und Prognose haben könnte (114). Dies wird in der Arbeit von Titano et al. für den Bereich der Radiologie beschrieben, wo eine automatisierte Triagierung von kranialen Bildgebungen zu einer beschleunigten Diagnose akuter neurologischer Ereignisse und somit beschleunigter Behandlung mit verbessertem Outcome führte (116). Aber auch für den Bereich der Pathologie liegen ‚proof-of-concept‘ Studien wie die Arbeit von Gehrung et al. aus dem Jahr 2021 vor, wo eine Triagierung in verschiedene Prioritätsklassen (definiert nach Sicherheit des Modells bezüglich Qualität des Schnittes und Sicherheit für die Diagnose) für die Bearbeitung durch den Pathologen mit einer automatisierten Aussortierung von Fällen in den niedrigsten Prioritätsklassen zur Reduktion der Arbeitsbelastung kombiniert wurde (117).

4.4.9.3 Unterstützung und Erweiterung der menschlichen Kompetenz

Eine mögliche wichtige zukünftige Rolle von Deep Learning Modellen stellt der Einsatz als Unterstützungssystem des Pathologen dar. Hierbei befundet der Pathologe selbst die Gewebeschnitte, erhält aber durch das Modell zusätzliche Informationen wie einen Klassifikationsvorschlag oder eine Klassifikationskarte, die durch farbliche Markierung auf potentiell relevante Gebiete des Gewebeschnittes (zum Beispiel Karzinomzellen) aufmerksam macht (106). In Abschnitt 3.9 beziehungsweise 4.4.8 sind diese Visualisierungsmöglichkeiten näher beschrieben. Ein Beispiel für die Anwendung solcher Visualisierungsmethoden ist die Arbeit von Wang et al., wo durch die Bereitstellung einer Heat map durch das Modell die diagnostische Genauigkeit der Pathologen bei der Erkennung von Brustkrebsmetastasen gesteigert und die menschliche Fehlerrate um 85% reduziert werden konnte (118). Die Arbeit von Steiner et al. aus dem Jahr 2018 beschreibt einen synergistischen Effekt der Zusammenarbeit von Mensch und Maschine, das heißt, dass eine bessere Gesamtgenauigkeit erzielt wird, als Modell oder Pathologe alleine erreicht hätten. Außerdem wurden eine Reduktion der benötigten Arbeitszeit und größere Sicherheit in der Entscheidung, wenn diese mit maschineller Unterstützung getroffen wurde, festgestellt (104).

Algorithmen der künstlichen Intelligenz zur Gewebeklassifikation lernen durch Identifizierung von wiederkehrenden Strukturmerkmalen (113). Dabei werden Charakteristika erkannt, die mit dem menschlichen Auge nicht erfassbar wären (113). Hier liegt die Basis für eine weitere Einsatzmöglichkeit als erweitertes Unterstützungssystem: die Abschätzung der Prognose eines Krebspatienten anhand von gewebsmorphologischen Markern anstelle der bisherigen immunhistochemischen und molekularbiologischen Zusatzuntersuchungen (113). Dies gelang zum Beispiel bereits in einer retrospektiven Studie, bei welcher maschinelles Lernen zur Prognoseabschätzung anhand von Gewebeschnitten bei Brustkrebspatientinnen eingesetzt wurde (102). Der vom Modell generierte prognostische Score zeigte eine hochsignifikante Assoziation zum Gesamtüberleben der Patientinnen (102). Auch die Feststellung der einer phänotypischen Veränderung zugrundeliegende Mutation beziehungsweise des molekularen Subtyps eines Karzinoms, die potenziell Einfluss auf die Therapieplanung nehmen, wäre so denkbar (119). Dies belegt beispielhaft die Arbeit von Woerl et al., wo mittels eines Deep Learning Modells anhand eines histopathologischen Gewebeschnittes der molekulare Subtyp von muskelinvasiven Blasenkarzinomen festgestellt werden konnte (120).

4.4.9.4 Qualitätssicherung

In der Befundung pathologischer Präparate kann aufgrund der teilweise recht subjektiven Beurteilungskriterien eine Interobserver-Variabilität entstehen (121). Daher ist die Teilnahme an Maßnahmen der Qualitätssicherung wie die unabhängige Beurteilung von Fällen durch eine Zweitperson umso wichtiger (121). Hier können relevante Fehler identifiziert und bei Bedarf Korrekturen - oftmals mit klinischer Relevanz - vorgenommen werden (121). Da die Fehlerrate in der Pathologie auf den Bereich zwischen 1% und 5% mit signifikanten klinischen Auswirkungen in 1% der Fälle geschätzt wird, werden ausgedehnte Qualitätssicherungsprogramme angewandt (122). Auch hier könnte die Künstliche Intelligenz als Assistenz des Mediziners oder alleinige Instanz bei der Erstellung einer Zweitmeinung zum Einsatz gebracht werden; Probleme wie der potenzielle Bias durch das Wissen um die Diagnose durch den ursprünglichen Befunder oder die Identität und Erfahrung dieses Befunders würden ebenfalls ausgeklammert (121). Im Bereich der Radiologie wurde hierzu bereits 2004 eine Studie veröffentlicht, die zeigt, dass computerassistierte Detektion – hier im Kontext des Mammographiescreenings – das Potential hat, in einer Zweitbefundung die falsch negative Rate bedeutend (um mehr als ein Drittel) zu senken (123). Das heißt, die diagnostische Genauigkeit könnte durch den Einsatz neuronaler Netzwerke erhöht werden (123).

5 Ergebnisdiskussion

Dies ist soweit zum aktuellen Zeitpunkt bekannt trotz der großen klinischen Bedeutung und dem hohen Stellenwert in der pathologischen Routinediagnostik eine der ersten Studien, die den möglichen Einsatz eines Deep Learning Gewebeklassifikationsmodells für den Bereich des Magenkarzinoms ins Besondere mit siegelringzelliger Differenzierung untersucht und dabei eine Klassifikation multipler, auch nicht maligner Gewebetypen anstrebt. Mit der Betrachtung der potenziellen Einsatzmöglichkeiten wird die Brücke zu einer möglichen Anwendung im klinischen Alltag geschlagen.

5.1 Diskussion der Kohortenzusammensetzung

Bei Betrachtung der Kohortenzusammensetzung sollte die Fallzahl von 96 berücksichtigt werden, die in ihrer Anzahl sicherlich zu klein ist, um statistisch repräsentativ auf die Grundgesamtheit von Magenkarzinompatienten beziehungsweise SRCC-Patienten verallgemeinert werden zu können. Dennoch soll hier ein orientierender Vergleich zu den in Abschnitt 2.1 und 2.2 beschriebenen, statistischen Merkmalen der Karzinome vorgenommen werden. In der vorliegenden Kohorte wurden mehr Männer eingeschlossen, was eher der Geschlechterverteilung des Magenkarzinomes allgemein, als der des SRCC mit ausgeglichener Geschlechterverteilung entspricht. Auch bezüglich des Alters, welches in dieser Arbeit durchschnittlich bei 67 mit dem häufigsten Vorkommen der Alterskategorie 71-80 Jahre lag, ist eher eine Beziehung zu den Parametern des Magenkarzinomes allgemein mit einem Durchschnittsalter von 72 Jahren zu sehen. SRCC treten wie beschrieben statistisch früher, im Bereich von 55-61 Jahren auf. Die zweithäufigste Alterskategorie dieser Arbeit (61-70 Jahre) lässt jedoch eine Orientierung in den jüngeren Altersbereich erahnen. Die Tumorausdehnung (pT) lag in dieser Arbeit am häufigsten bei pT3, das Grading bei pG3, was der Statistik sowohl für Magenkarzinome im Allgemeinen als auch SRCC im Speziellen mit einer Verschiebung der Stadien in den höheren Bereich entspricht. Die in der Literatur beschriebene Neigung zur schnellen Ausbreitung und frühen Metastasierung konnte in der Kohorte dieser Arbeit nicht wiedergefunden werden; es lagen am häufigsten Fälle ohne Lymphknotenbefall (pN0) oder Fernmetastasierung (pM0) vor. Passend zur beschriebenen schrittweisen Entstehung von Karzinomen (vgl. Abschnitt 2.2.2) wurden bei den Patienten dieser Kohorte allerdings vielfach intestinale Metaplasien als Vorstufe des Magenkarzinomes beschrieben. Weitere klinische Hintergrunddaten

der Kohorte wie eine Besiedelung des Magens mit H.p., die Anwendung neoadjuvanter Behandlung oder spezifische vorliegende Mutationen wurden in der vorliegenden Arbeit nicht erfasst.

Für das Training und die Validierung aller Klassen standen durch die 96 Patienten insgesamt 21501 Kacheln zur Verfügung. Zum Vergleich wurden in der Arbeit von Iizuka et al. (105), die ebenfalls ein Deep Learning Modell im Bereich Magenkarzinom trainierten, 4128 Gewebeschnitte genutzt, die zu einer Kachelzahl von circa 4 Millionen führten. Die genaue Kantenlänge wurde nicht angegeben. Studien zeigen, dass die Genauigkeit der Ergebnisse eines Modelles stark mit der Menge an eingebrachten Daten korreliert (115). Obgleich unserem Modell weniger Trainingsbeispiele zur Verfügung gestellt wurden, ergibt sich letztlich eine mindestens ebenbürtige Klassifikationsleistung (vgl. Abschnitt 4.4). Dies könnte als Hinweis darauf gewertet werden, dass die Klassifikationsaufgabe eine vergleichsweise einfache Fragestellung darstellt, welche das Modell auch nach Training an weniger Kacheln mit guter Genauigkeit beantworten kann. Zur Klärung könnte trotz der bereits guten Ergebnisse des Modelles der vorliegenden Arbeit (vgl. Abschnitt 5.4) erwogen werden die Anzahl der Trainingsbeispiele weiter zu erhöhen und den Effekt auf die Klassifikationsleistung zu beobachten. Eine mögliche Verbesserung der Klassifikationsleistung durch Erhöhung der Trainingsbeispiele wird im Vergleich der Arbeiten zur Gewebeklassifikation von kolorektalem Tumorgewebe von Kather et al. aus dem Jahr 2016 (124) und 2019 (125) sichtbar: Während 2016 mit 5000 Trainingsbeispielen eine accuracy von 87,4% für die acht untersuchten Klassen erreicht werden konnte, wurde 2019 eine accuracy von circa 94% unter Einsatz von 100.000 Beispielen für die Klassifikation von neun Klassen erreicht.

5.2 Diskussion der Trainingsergebnisse

Anhand der Trainingskurve (vgl. Abschnitt 4.3) kann das Modell auf Probleme wie overfitting oder underfitting (vgl. Abschnitt 2.4.7) untersucht werden. Während des Trainings des Modelles kann der aktuelle Lernfortschritt nach jeder Epoche beurteilt werden. Die Betrachtung des training loss gibt dabei Aufschluss darüber wie gut das Modell „lernt“. Die Beurteilung des validation loss und der accuracy (in der Validierung) geben hingegen Auskunft darüber wie gut das Modell auf ungesehene Daten generalisieren kann. Dass bei dem Modell der vorliegenden Arbeit das Training zum richtigen Zeitpunkt abgeschlossen, das heißt weder deutlich overfitting noch underfitting vorliegen, erkennt man am Kurvenverlauf des training loss, der sich bis zur

Stabilisierung auf ungefähr gleichem Niveau absenkt. Ebenso verhält es sich mit dem validation loss, wobei die schmale Abweichung der Stabilitätsebene in höhere Werte aufgrund des Generalisierungsfehlers zustande kommt: Der Klassifikationsfehler ist für die Trainingsdaten immer geringer zu erwarten als für die Validierungsdaten. Zeichen für underfitting wie ein flacher Kurvenverlauf unabhängig vom Training oder ein immer noch stark abfallender Kurvenverlauf am Ende des Trainings sind nicht zu erkennen. Auch für overfitting gibt es keinen Hinweis. Dies würde sich beispielsweise an einem erneuten Anstieg des validation loss nach anfänglichem Abfall oder einem Abfall der accuracy nach anfänglichem Anstieg zeigen.

Zu beachten ist hier allerdings, dass sich diese Daten und Kurven ausschließlich auf eine Validierung an Trainingsdaten aus einem Institut beziehen. Zur weiterführenden Reduktion des Risikos für die Einflussnahme von Bias beziehungsweise over- oder underfitting auf die Leistung des Modelles sind Trainingsdaten größerer Variabilität nötig. In der vorliegenden Arbeit wurde keine Validierung an externen Trainingsdaten vorgenommen, wodurch trotz fehlender Hinweise vor allem für overfitting an den internen Daten dieses nicht gänzlich ausgeschlossen werden kann. Die Generalisierbarkeit der Daten ist somit nicht ausreichend nachgewiesen. Dies müsste vor der Erwägung einer klinischen Nutzung nachgeholt werden (126).

5.3 Diskussion der Gewebetypenklassifikationen

Das Deep Learning Modell dieser Arbeit wurde auf die Klassifikation von verschiedenen Gewebetypen im Gastrointestinaltrakt (Epithel, Fett, Immunzellen, Muskel, Karzinom) trainiert. Die Einzelbetrachtung der Ergebnisse der verschiedenen Klassen ermöglicht dabei eine dezidierte Fehler- und Problembetrachtung und gibt beispielsweise Aufschlüsse über diagnostische Schwierigkeiten mancher Klassen oder mögliche Probleme durch ungleiche Verteilung von Trainings- und Testdaten. Dies unterscheidet diese Arbeit von der Mehrzahl der aktuellen Arbeiten, die die Modelle für eine binäre Klassifikationsentscheidung zwischen Tumorgewebe und gesundem Gewebe nutzen (vgl. mit den in diesem Abschnitt genannten Arbeiten). Subtypen von gesundem Gewebe werden dort nicht unterschieden, was einen tieferen Einblick in die Tumormikroumgebung, das heißt das unmittelbare Umfeld eines Malignomes, und die allgemeine Zusammensetzung des Gewebes verhindern könnte. Diese ist von Relevanz, da sich die Architektur des Tumors und seiner Umgebung, die beispielsweise Stroma, infiltrierende Immunzellen, nekrotische Areale oder Inseln verbleibenden gesunden Gewebes enthält, mit der Progression des Karzinomes

ändert und dabei eng mit der Prognose des Patienten verbunden ist (125). Die Quantifizierung der Zusammensetzung der Tumormikroumgebung ist daher eine relevante histopathologische Aufgabe; auch bezüglich des Gradings (124). So ist es nicht verwunderlich, dass die Analyse der „Nicht-Tumor-Komponenten“ von Tumorgewebeschnitten in den Fokus der Biomarkersuche der Onkologie rückt (125).
Eine Unterscheidung multipler Klassen entgegen der Mehrzahl der aktuellen Arbeiten nehmen die Untersuchungen von Kather et al. aus dem Jahr 2016 (124) und 2019 (125) vor. Zwar werden auch hier keine Werte spezifisch für die Klassifikationsleistungen für die einzelnen Klassen angegeben, die Arbeit von 2016 nennt jedoch eine accuracy von 87,4% für die acht Klassen mit einer AUROC von 0,976, von der die einzelnen Klassen eine nur geringfügige Abweichung zeigen sollen. Die Arbeit von 2019 beschreibt eine accuracy von circa 94% ohne Nennung der AUROC. In der vorliegenden Arbeit konnten im Vergleich dazu ebenbürtige beziehungsweise überlegene Leistungen erzielt werden; die AUROC liegen im Bereich von 0,99, die accuracy schwankt je nach Klasse mit einem Mittelwert von 96,27%.
Für die Klasse Epithel als ersten Subtyp gesunden Gewebes ergab sich eine accuracy von 97,54% mit einem recall von 97,24%. Im Vergleich mit der Arbeit von Thiem et al. (127) zur Gewebeerkennung im oralen Bereich, die für die Mukosa eine accuracy von 81% und einen recall von 57% beschreiben, zeugt dies von guten Ergebnissen des Modelles in der vorliegenden Arbeit. Mukosa ist jedoch histologisch nicht gleichzusetzen mit epithelialem Gewebe, ebenso wenig Gewebe im oralen Bereich mit dem weiter aboral und auch die Anzahl von Validierungsfällen in der Arbeit von Thiem et al. war sehr gering, sodass insgesamt die Vergleichbarkeit stark eingeschränkt ist. Es fehlt an Studien, die eine bessere Vergleichbarkeit erbrächten.
Die AUROC von 0,996 +/- 0,001 sowie die hohe Sicherheit in den Klassifikationsentscheidungen bezeugen die geringe Streuung und stabile Leistung des Modelles der vorliegenden Arbeit für diese Klassifikationsentscheidung. Bei Betrachtung der Einschränkungen des Modells lag die höchste Anzahl an Fehlklassifikationen erwartungsgemäß bei der fälschlichen Ausgabe von Karzinomgewebe für das annotierte Label Epithel und epitheliales Gewebe für das annotierte Label Karzinom. Karzinomgewebe entsteht mittels Mutationen, welche sich auch phänotypisch niederschlagen und einen histologischen Unterschied begründen, aus Epithelgewebe und liefert mit der großen histologischen Ähnlichkeit der beiden Gewebetypen eine mögliche Erklärung für die Schwierigkeiten in dieser Klassifikationsentscheidung (59).

Die oben genannten Einschränkungen der Vergleichbarkeit mit der Arbeit von Thiem et al. und der Mangel an vergleichbaren Studien gilt auch für die Klassen Fettgewebe und Muskelgewebe. Für die Klasse Fettgewebe ergaben sich in dieser Arbeit eine accuracy von 99,8% und ein recall von 97,95%, während die Arbeit von Thiem et al. Werte von 95% für die accuracy und 86% für den recall ergab. Bezüglich der Klasse Muskelgewebe ergab deren Arbeit eine accuracy von 86% und einen recall von 100%, während diese Arbeit Ergebnisse von 99,2% für die accuracy und 98,2% für den recall erbrachte. Somit liegen im Vergleich in der vorliegenden Arbeit für beide Klassen gute Ergebnisse vor, die ebenfalls mit hoher Sicherheit und geringer Streuung (Fett: AUROC 0,999 +/- 0,000, Muskel: AUC 0,999 +/- 0,000) einhergehen. Es konnte, obwohl für die Klassifikationsaufgabe Fettgewebe vor Augmentation nur 1074 Kacheln für Training und Validierung und damit nur ein Bruchteil der Beispiele für die anderen Klassen vorlagen, dennoch eine vergleichsweise gute Leistung erreicht werden. Grund hierfür könnte sein, dass sich dieser Gewebetyp histologisch stark von den übrigen unterscheidet und eine Erkennung dieser Klasse somit einfacher erscheint. Weiterhin wurde im Training bezüglich der Klassenungleichheit ein Ausgleich implementiert, der mit Hilfe augmentierter Kacheln die Differenz der vorliegenden Kacheln der einzelnen Klassen auffüllt.

Für die Klasse Immunzellen lagen mit 222 Kacheln noch weniger Beispiele vor. Es ergaben sich Ergebnisse von 99,4% für die accuracy und eine AUC von 0,999 +/- 0,001. Zum Vergleich kann man die Arbeit von Turkki et al. (128) heranziehen, die sich ausschließlich mit der Unterscheidung von Gewebe mit und ohne Immunzellen bei Brustkrebspatientinnen befasst. Hier wurde eine accuracy von 98% und eine AUROC von 0,98 erreicht. Auch hier bestätigen sich gute Ergebnisse unseres Modells, wobei wiederum Einschränkungen in der Vergleichbarkeit zu berücksichtigen sind: Die Untersuchungen von Turkki et al. fanden an Brustkrebsgewebe statt, dessen diagnostische Herausforderung von der von Magenkarzinomgewebe abweicht. Außerdem bestand eine binäre Klassifikationsentscheidung, während in dieser Arbeit die Metriken unter Berücksichtigung aller fünf Klassen berechnet wurden. Auch hier fehlt es an Studien mit besserer Vergleichbarkeit. Dass die geringe Anzahl von Trainingskacheln für Immunzellen in dieser Arbeit womöglich dennoch ein Problem sein könnte, ergibt sich bei Betrachtung der precision für diese Klasse: Während für alle anderen Klassen Werte von mindestens 95% erreicht wurden, zeigt die Klasse Immunzellen einen Wert von nur 62,68%. Es wäre möglich, dass die augmentierten Kacheln aufgrund der bestehenden Ähnlichkeit zu den ursprünglichen Kacheln einen

geringeren Lerneffekt haben als gänzlich neue Beispiele. Obgleich die anderen betrachteten Metriken gute Werte ergeben, wäre es zu evaluieren, ob mit einer Erhöhung der Trainings- und Validierungsfälle, das heißt einem Ausgleich der Klassenungleichheit nicht nur durch augmentierte Kacheln, auch diese Metrik optimiert werden könnte. In der Fehlerbetrachtung der Klassifikationsentscheidungen zur Klasse Immunzellen fällt als häufigste Fehlerquelle die Fehlausgabe von Karzinomgewebe für annotierte Immunzellen und Immunzellen für annotiertes Karzinomgewebe auf. Dies erklärt sich womöglich durch das oftmals gemeinschaftliche Auftreten der beiden Klassen, da Malignome meist entzündliche Infiltrate in ihrer Umgebung hervorrufen (125). Eine genaue Abgrenzung in Klassen ist so bei gleichzeitigem Vorliegen erschwert. Auch besteht histopathologisch eine gewisse Ähnlichkeit der Gewebe bei beispielsweise bestehender hoher Zellularität des Gewebes und Hyperchromasie der Zellkerne (66).

Besondere Relevanz im Hinblick auf die möglichen Einsatzfelder des Modelles im Bereich der Tumorerkennung ist der Leistung bei der Klassifikation von Karzinomgewebe beizumessen. Bezüglich der Intention, die der Klassifikationsaufgabe zugrunde liegt, erscheint vor allen Dingen die Betrachtung des recall relevant: Bei einem Erkennungssystem für Krebs muss die Rate der falsch negativen Ergebnisse maximal abgesenkt werden, da eine ausbleibende Diagnose für den Patienten fatale Auswirkungen hätte. Hier erreichte das Modell 94,22%. Auch die Betrachtung der precision ist für die Klassifikationsaufgabe insofern relevant, als dass bei einem falsch positiven Ergebnis der Patient nicht nur einer psychischen Belastung, sondern womöglich nicht notwendigen Untersuchungen und Behandlungen ausgesetzt wird (129). Für diese Metrik erreichte das Modell 97,23%. Die AUROC bezieht sowohl falsch negative als auch falsch positive Ergebnisse mit ein, hier erzielte das Modell eine AUC von 0,994+/-0,001. Beim Vergleich mit anderen Arbeiten ist abermals die Problematik des Vorliegens von überwiegend binären Studien zur Unterscheidung von Tumor zu Normalgewebe zu berücksichtigen, welche als Ergebnisse die Gesamtleistung des Systems und nicht die für die einzelnen Klassen ausgeben. In der Arbeit von Iizuka et al. (105) aus dem Jahr 2020, die sich mit Deep Learning Modellen für die histopathologische Klassifikation von epithelialen Tumoren in Magen und Kolon beschäftigt, werden die AUROC für die Klasse Karzinomgewebe im Magenbereich gesondert aufgeführt, wobei das Modell einen maximalen Wert von 0,98 (Konfidenzintervall 0,966-0,990) erreicht, was als hohe Systemleistung gewertet wird. Andere Metriken werden nicht angegeben. Die Leistung des Modelles der

vorliegenden Arbeit ist hier demnach ebenfalls als hoch einzustufen, weiterhin war auch eine hohe Sicherheit für die einzelnen Klassifikationsentscheidungen gegeben. Die meisten Klassifikationsfehler im Bereich des Karzinomgewebes machte das Modell in der Falschausgabe von Epithel für annotiertes Karzinomgewebe und Karzinomgewebe für annotiertes Epithel. Die wahrscheinlichste Erklärung hierfür – die histologische Ähnlichkeit von Karzinom und Epithel aufgrund der Entstehung aus diesem – wurde oben bereits beschrieben.

Zusammenfassend ergab sich für alle Klassen eine hohe Klassifikationsleistung, wobei im Gegensatz zu anderen Studien durch die Aufteilung des gesunden Gewebes in seine histologischen Subtypen sowohl eine genauere Betrachtung der Tumorumgebung als auch eine verbesserte Evaluation etwaiger Fehlerquellen möglich ist. Da der Großteil der Studien einen anderen Aufbau aufweist, ist die Vergleichbarkeit der Leistungsmetriken jedoch in seinen Einschränkungen zu sehen.

5.4 Diskussion der Gesamtsystemleistung für die Gewebeklassifikation

Die accuracy lag für die Gesamtleistung des Systems unter Einbezug aller Klassen bei 96,27%. Dass hiermit ein sehr gutes Gesamtergebnis erreicht werden konnte, bestätigt sich im Vergleich mit anderen Studien: In der Arbeit von Song et al. aus dem Jahr 2020 (106), wo ein Modell ebenfalls im Bereich des Magens trainiert wurde, wurde eine accuracy von 87,3% erreicht. Unterschieden wurde hier allerdings nicht in fünf, sondern nur in die zwei Klassen maligne und benigne, was als weniger komplexe Aufgabe jedoch prinzipiell eine höhere Gesamtleistung erwarten ließe. Bei genauerer Betrachtung fällt dort ein recall von 99,6% auf, der höher liegt als der recall dieser Arbeit (macro average 96,89%, weighted average 96,27%). Die Maximierung des recall mit Absenkung der Anzahl der falsch negativen Ergebnisse bezüglich Tumorerkennung ist wegen der großen klinischen Bedeutung für die Arbeit von Song et al. durchaus entscheidend, lässt aber dadurch nur eine bedingte Vergleichbarkeit mit dieser Arbeit zu, da hier in den recall einfließende falsch negative Ergebnisse aufgrund der höheren Klassenanzahl nicht gleichzusetzen sind mit einer übersehenen Tumordiagnose. Zur allgemeinen Betrachtung der Diskriminierungsfähigkeit des Modells zwischen den vorhandenen Klassen eignet sich daher eher der Vergleich der AUROC. Hier ergibt die Arbeit von Song et al. eine AUC von 0,986, in der vorliegenden Arbeit wurde eine AUC von 0,997 (micro und macro average) über alle Klassen erreicht. Die damit gute Leistung des Modells dieser Arbeit bestätigt sich auch im Vergleich mit anderen Arbeiten. In der Arbeit von Etheshami et al. von 2017 zur

Erkennung von Lymphknotenmetastasen bei Präparaten von Brustkrebspatientinnen wurden im Rahmen eines Wettbewerbes 23 Modelle trainiert und vorgestellt, die eine AUC von 0,556 für das schlechteste Modell bis 0,994 für das Modell mit der besten Leistung aufwiesen (103). Die Diskriminierungsfähigkeit des Modells der vorliegenden Arbeit zeigt vergleichbare Ergebnisse. Kanavati et al. veröffentlichten 2021 eine Arbeit (130) für die Klassifikation bei diffusem Magenkarzinom, was dem Gewebetyp dieser Arbeit nahezu entspricht. Die AUC lag dort zwischen 0,95 und 0,99 bei Evaluierung an verschiedenen Testdaten. Demnach erbringt das Modell der vorliegenden Arbeit auch im Vergleich zu ähnlichen Gewebetypen eine ebenbürtige Leistung. Kritisch betrachten muss man allerdings die im Gegensatz zur Arbeit von F. Kanavati et al. fehlende Evaluierung an unabhängigen Testdaten anderer Kliniken, die beispielsweise durch andere Färbeprotokolle oder Whole-slide-Scanner von den genutzten Trainings- und Evaluierungsdaten abweichende Daten produzieren. Vor einer klinik- beziehungsweise flächenübergreifenden Nutzung eines Klassifikationsmodells müsste in jedem Fall eine solche Testung zur Robustheit des Modells erfolgen. Ein weiterer Unterschied zur obigen Arbeit ist die Gewebegrundlage der Trainingsdaten: Während diese Arbeit im Hauptanteil an Gewebeschnitten im Bereich des Magenkarzinoms trainiert hat, nutzten Kanavati et al. Biopsien aus endoskopischen Untersuchungen. Es wäre zu untersuchen, ob die Modellleistung der vorliegenden Arbeit auch bei Anwendung auf Biopsien abzurufen ist, da so der Nutzungsbereich auf die in der pathologischen Routinediagnostik in enormer Anzahl vorkommenden Proben dieser Form ausgeweitet werden könnte. Vergleichbarkeit bezüglich des Aspektes der Analyse multipler Klassen bieten die Arbeiten von Kather et al. aus dem Jahr 2016 (124) und 2019 (125) zum kolorektalen Karzinom. Während 2016 acht Klassen mit einem Ergebnis von 87,4% für die accuracy und 0,976 für die AUROC unterschieden wurden, waren es 2019 neun Klassen mit einer accuracy von circa 94% für die Validierung an einem externen und 98,7% für die Validierung an einem internen Datensatz. Zwar kann die vorliegende Arbeit wie oben beschrieben ebenbürtige Ergebnisse vorweisen, wiederum fehlt jedoch in der vorliegenden Arbeit eine Validierung an externen Daten für eine bessere Vergleichbarkeit der Arbeiten. Die Anzahl der unterschiedenen Klassen kann als Anregung gesehen werden, in einer zukünftigen Arbeit zum SRCC ebenfalls weitere Klassen der Tumormikroumgebung wie Stroma oder nekrotische Areale mit einzubeziehen.

Betrachtet man die Ergebnisse für die precision des Modells der vorliegenden Arbeit, fällt eine große Differenz zwischen macro average (90,31%) und weighted average

(96,48%) auf. Erklären lässt sich dies bei einer Teilbetrachtung der precision der einzelnen Klassen: Während für epitheliales Gewebe, Fettgewebe, Muskelgewebe und Karzinomgewebe Werte zwischen 95,47% und 98,32% erreicht wurden, fiel der Wert für Immunzellen mit 62,68% deutlich niedriger aus. Dies schlägt sich bei Mittelung nach macro average stark nieder, da alle Klassen mit der gleichen Gewichtung behandelt werden. Die Mittelung mit weighted average hingegen berücksichtigt die sehr geringe Anzahl von Daten in der Klasse der Immunzellen (222 Kacheln im Gegensatz zu den jeweils über 1000 Kacheln für alle anderen Klassen). Die geringe Anzahl an vorhandenen Kacheln vor Augmentierung (vgl. Abschnitt 4.2) könnte auch die Ursache für die geringere precision in dieser Klasse sein. Dies wäre durch eine erneute Validierung nach Erhöhung der Kachelzahl für diese Klasse zu überprüfen. Da die precision sich auf die Genauigkeit bezüglich der falsch positiven Fälle bezieht, bei einer Klassifikationsaufgabe im Setting der Tumorerkennung wie bereits beschrieben jedoch vor allem eine niedrige Zahl von falsch negativen, das heißt übersehenen Tumorfällen klinisch relevant ist, erscheint die Betrachtung des recall hier sinnvoller. Hier liefert das Modell Werte von 96,89% für eine Mittelung mit macro average und 96,27% für weighted average. Eine Differenz der beiden Methoden zur Mittelung ist kaum noch zu erkennen, was bedeutet, dass die Leistung des Modells über alle Kacheln und Klassen hinweg durchgehend ähnlich und in Anbetracht der Ergebnisse sehr gut ist. Keine Klasse hebt sich durch größere diagnostische Schwierigkeit hervor. Dies gilt nicht nur für die Betrachtung des recall sondern auch für die AUROC, wo der Wert für micro und macro average identisch bei 0,997 liegt. Aufgrund der klinischen Bedeutsamkeit übersehener Krebsdiagnosen wäre ein Vergleich verschiedener Studien über die Metrik recall ebenso oder vielleicht sogar sinnvoller als der Vergleich der AUROC, wo der klinisch wichtige Bereich der hohen Sensitivität nur einen Teilbereich der berechneten AUC darstellt. Viele Studien stellen jedoch nur die Metrik der AUROC zur Verfügung.

Im F1-Score als harmonische Mittelung von recall und precision schlägt sich die oben genannte Problematik der precision für Immunzellen im macro average noch einmal nieder: Es ergab sich ein Wert von 0,9286 für macro average und 0,9632 für weighted average, was aufgrund des Ausgleichs der Klassenungleichheit eher als wirkliches Maß für die Modellleistung betrachtet werden kann. Gleiche Überlegungen sind auch bezüglich der Ergebnisse für die PR-curves und die AUROC anzustellen. Bei den PR-curves ergab sich die schlechteste Leistung – wiederum vermutlich durch eine zu geringe Kachelzahl - für Immunzellen, die beste für das Label Muskel. Bei der AUROC,

die den recall mit den relevanten falsch Negativen einbezieht, waren die schlechtesten Ergebnisse für die Klassen Epithelgewebe und Karzinomgewebe zu verzeichnen, wobei die Leistung mit Werten von 0,996 (Epithel) beziehungsweise 0,994 (Karzinom) immer noch - wie im obigen Vergleich mit anderen Arbeiten gezeigt – als hoch einzustufen ist. Diese Ergebnisse passen zur Analyse der Fehlerquellen der Klassifikationsentscheidungen des Modelles. Als häufigste Fehler insgesamt ließen sich die fälschliche Ausgabe des Labels Karzinomgewebe für annotierte Epithelfälle und des Labels Epithel für annotierte Karzinomfälle feststellen. Erklären kann dies wie schon in Abschnitt 5.3 beschrieben mit großer Wahrscheinlichkeit die histologische Ähnlichkeit von Epithel und Karzinom. Gerade im Vergleich mit den anderen zu unterscheidenden Gewebetypen (Fett, Muskel, Immunzellen), die in Aufbau und Erscheinung stark unterschiedlich sind, ist die höchste Fehlerrate in der komplexen Unterscheidung von epithelialem Gewbe und Karzinomgewebe erwartungsgemäß. Betrachtet man die Standardabweichung für die oben genannten AUROC ergeben sich maximale Werte von +/- 0,001, was mit dieser geringen Streuung die Stabilität der Klassifikationsentscheidungen des Modells bezeugt.

Zusammenfassend lässt sich feststellen, dass das Modell auch im Vergleich mit anderen Studien eine sehr gute Leistung erbringt. Zur verbesserten Generalisierbarkeit und Erweiterung der Nutzbarkeit sollten allerdings noch weiterführende Untersuchungen wie eine Validierung an unabhängigen, externen Testdaten und Biopsie-Fällen erfolgen. Ein Ausgleich der Klassenungleichheit unabhängig von der Nutzung augmentierter Kacheln vor allem im Bereich der Immunzellen kann erwogen werden, nimmt jedoch vermutlich keinen großen Einfluss auf die Ergebnisse im Bereich der für diese Aufgabe wichtigen Leistungsmetriken.

5.5 Diskussion der Visualisierungsmöglichkeiten

Wie in Abschnitt 2.3 beschrieben, bedienen sich Pathologen in ihrer alltäglichen Arbeit bis heute hauptsächlich analoger Technologien; für die Diagnostik kommen hauptsächlich Glasobjektträger und Lichtmikroskope zum Einsatz. Mit Hilfe eines „intelligenten Mikroskops“ (vgl. Abschnitte 3.9.1 und 4.4.8), welches Klassifikationsvorschläge in Echtzeit während des Mikroskopierens bietet, könnten die Vorteile (vgl. Abschnitt 5.6) eines digitalen Unterstützungssystemes genutzt werden, ohne dass auf das Mikroskop verzichtet werden müsste. Dieser Hybridansatz zwischen analoger und digitaler Pathologie könnte zudem den Übergang zu einer komplett digitalen Arbeitsweise erleichtern (113). Mit zukünftig potentiell

fortschreitender Digitalisierung der Pathologie und zunehmendem Vorliegen von Gewebeschnitten in digitaler Form könnte die Visualisierungsmethode mittels Klassifikationskarte (vgl. Abschnitte 3.9.2 und 4.4.8) vorteilhafter werden. Hier wird ein schneller Überblick über relevante Bildareale schon in der Bildübersicht vor einer vergrößerten Betrachtung von Gewebeausschnitten ermöglicht, was vor allem bezüglich Aspekten wie Zeitersparnis und Erhöhung der diagnostischen Genauigkeit interessant wäre.

Selbigen Ansatz verfolgen auch Chen et al. in ihrer Arbeit aus dem Jahr 2019. Hier wurde eine durch künstliche Intelligenz erstellte Klassifikationskarte in Echtzeit auf das Bild eines Lichtmikroskopes zur Diagnostik von metastasierten Mammakarzinomen und Prostatakarzinomen projiziert. Der Vorteil der erweiterten Realität wird hier in einer verbesserten Genauigkeit und Effizienz bei gleichzeitig unverändertem Arbeitsablauf des Pathologen in gewohnter Geschwindigkeit gesehen (131). Weitere Studien in diesem Bereich, gerade zum tatsächlichen Einsatz im klinischen Arbeitsablauf werden gefordert (131).

5.6 Diskussion der Einsatzmöglichkeiten eines Deep Learning Modells

Die Literaturrecherche zu den möglichen Einsatzmöglichkeiten eines Deep Learning Modelles für Gewebeklassifikationen im Magenbereich in der klinischen Pathologie ergab vier mögliche große Teilbereiche (vgl. Abschnitt 4.4.9): Zunächst die eigenständige Übernahme wenig komplexer, aber arbeitsaufwändiger Arbeitsschritte zur Entlastung des Pathologen und Verlagerung der menschlichen Ressourcen auf komplexere Arbeitsgebiete. Weiterhin die Möglichkeit der Triagierung der vorliegenden Fälle, was die Ressourcen auf die Fälle mit der größten Wahrscheinlichkeit für einen Bedarf an medizinischer Versorgung richten würde. Außerdem könnten sie als Unterstützungssysteme des Pathologen dienen, was dessen Genauigkeit erhöhen und um Kompetenzen wie eine vereinfachte Prognoseabschätzung und Erkennung der einem Karzinom zugrundeliegenden genetischen Merkmale erweitern könnte. Als weiterer Bereich wurde die Qualitätssicherung beschrieben, wo Deep Learning Modelle in der Funktion als Zweitmeinung Fehler mit potentiell relevanten klinischen Auswirkungen reduzieren könnten.

Obgleich die Vorteile in allen genannten Einsatzfeldern klar erkennbar sind, stehen der praktischen Umsetzung im klinischen Alltag einige große Hürden im Weg.

Um solche Modelle maschinellen Lernens zu trainieren, zu validieren und zu verbessern wird nicht nur initial, sondern auch im Verlauf kontinuierliche Datenzufuhr

gebraucht (114). Studien haben gezeigt, dass die Genauigkeit des Modelles stark mit der Menge der eingesetzten Daten korreliert (115), während in der Realität nur wenig histopathologisches Material überhaupt in digitalisierter Form vorliegt. Die Menge wird in Anbetracht der Dynamik im Bereich der Digitalisierung mittelfristig steigen, dennoch besteht weiterhin das Problem, dass die Daten durch Pathologen – die auch so schon einer starken Arbeitsbelastung ausgesetzt sind – beschrieben werden müssen , hängt doch die Genauigkeit der Vorhersagen von supervised learning stark von der Genauigkeit der in den Algorithmus eingebrachten Annotationen ab (132). Für eine übergreifende Implementierung und Erweiterung der Datenmenge müssten die annotierten Daten über Institutions- und bestenfalls Ländergrenzen hinweg ausgetauscht werden (114). Es bestehen für medizinische Bilder bereits einige Datenbanken wie beispielsweise das „Cardiac Atlas Projekt“ (133). Solche Möglichkeiten müssten ausgeweitet oder in größerem Maßstab angelegt werden. Ein solcher Datentransfer bringt neue Probleme für den Datenschutz der Patienten aufgrund der Möglichkeit der unrechtmäßigen Nutzung oder Veröffentlichung patientenbezogener Daten (114). In diesem Zusammenhang wird die Datenschutzverordnung der EU vom Mai 2018 einen großen Einfluss auf den Vorgang der Einführung von künstlicher Intelligenz in den medizinischen Bereich allgemein haben (134): Es wird beispielsweise expliziter und informierter Konsens vor der Sammlung persönlicher Daten gefordert. Informierter Konsens ist zwar schon lange natürlicher Teil des medizinischen Entscheidungsprozesses, dennoch stellt die Frage nach der Möglichkeit einer verständlichen Information in einem solch komplexen Bereich eine Hürde dar (114). Inbegriffen sind weiterhin ausgiebige Regelungen bezüglich Datensammlung, -aufbewahrung und die Nutzung persönlicher Informationen (134). Kritisch zeichnet sich hier der Artikel 22 der Datenschutzverordnung ab, der das Recht eines jeden Bürgers beschreibt eine Erklärung für algorithmische Entscheidungen zu verlangen (134). Hier liegt ein weiteres Grundproblem von Deep Learning Modellen: die vermeintlich fehlende Interpretationsmöglichkeit (Black-Box-Problem) (28), bei isolierter Betrachtung von in- und output. Anders gesagt sollte es den Gesundheitsexperten möglich sein zu verstehen wie das Modell zu bestimmten Entscheidungen und Vorhersagen gekommen ist (27), um so auch überprüfen zu können, ob die Ausgabe nach aktuellem Wissensstand valide ist. Das Fehlen dieser Möglichkeit ist in einem Setting, in dem Entscheidungen über Diagnosen, Behandlungen und Prognosen von Patienten getroffen werden müssen, sicher keine zu rechtfertigende Herangehensweise und

würde nach Datenschutzverordnung den Einsatz potentiell verbieten (134). Da diese Verordnung erst vor sehr kurzer Zeit in Kraft getreten ist, sind die Langzeiteffekte noch abzuwarten. Es wird erwartet, dass die Neuerungen die Implementation von künstlicher Intelligenz im Gesundheitswesen kurzfristig betrachtet deutlich verlangsamen, langfristig jedoch erleichtern, da die strikten regulatorischen Standards das öffentliche Vertrauen und den Einbezug der Patienten fördern (114). Darüber hinaus gibt es durchaus eine Reihe von Möglichkeiten den Entscheidungsprozess von Algorithmen „sichtbar“ und transparent zu machen (135). Die Thematik der „Erklärbarkeit“ wir derzeit noch intensiv beforscht (explainable KI).

Nicht nur die Menge und Verarbeitung der Daten ist relevant, sondern auch das vorliegende Format, das standardisiert werden müsste, um für verschiedene Institutionen und deren Arbeitsmittel zugänglich zu sein (114). Hierbei handelt es sich um ein weiteres Kernproblem, da gerade im Gesundheitssystem Daten für verschiedene Zwecke mit verschiedenen Methoden gesammelt werden und es eine ganze Bandbreite verschiedener Möglichkeiten der Speicherungsformate gibt (29).

All diese Punkte führen, gerade vor dem Hintergrund, dass kaum eine pathologische Einrichtung bisher überhaupt digitalisiert arbeitet (113), zum Problem der Kostenfrage. Die Technologien benötigen neben kostspieliger Neuanschaffungen von Hardware für eine optimierte Leistung dauerhafte Instandhaltung beispielsweise in Form von kontinuierlicher Zufuhr neuer Patientendaten und Softwareaktualisierungen (114). Während die Investitionen der Regierungen, akademischen Einrichtungen und der Industrie in dieser frühe Phase, in der sich die künstliche Intelligenz befindet, weiter steigen, bleibt deren Aufrechterhaltung abzuwarten (114). Diese hängt vermutlich zu nicht unwesentlichem Teil von dem Erfolg der frühen Implementierungsversuche und einer Klärung der Vergütungsfrage für den Einsatz solcher Technologien ab (114). Theoretische Untersuchungen zum Kostenvorteil digitaler Pathologie beschreiben jedoch eine rentable Investition durch Ersparnisse bei Verbesserung der Produktivität der Pathologen bei gleichzeitiger Verbesserung der Patientenversorgung (70). Weiterhin ist noch nicht abschließend geklärt, welche Mechanismen für die Qualitätskontrolle angewandt werden sollten. Normalerweise müssen sich Medizinprodukte einem ausgiebigen und komplexen Zulassungsverfahren von mehreren Monaten unterziehen (136). Diese Vorgehensweise ist nicht an die schnellen Fortschritte und Veränderungen der Softwareentwicklung angepasst. Daher besteht Notwendigkeit einer Aktualisierung der Zulassungsverfahren für Software als Medizinprodukt. Beispielhaft hat die FDA in den USA eine erste Anpassung

vorgenommen, in der sich der Prüfungsprozess statt auf die Technologie auf den Entwickler fokussiert, in dessen Verantwortung die weitere Produktbeobachtung liegt (137). Ein weiterer mit der Patientensicherheit verknüpfter Aspekt, ist der der Schuldigkeit. Sollte ein Patient Opfer einer für seine Gesundheit nachteiligen Entscheidung werden, stellt sich die Frage der Verantwortlichkeit. Das Gefühl persönlicher Verantwortlichkeit des Arztes wird mit steigender Einflussnahme künstlicher Intelligenz abnehmen, doch es ist unklar, wohin sich die Schuldigkeit verschiebt. Hier wären verschiedene Instanzen denkbar: der Anbieter der Software, der Entwickler oder die Instanz der Bereitstellung der Trainingsdaten (114).

Da eine alleinige Einführung von Software ohne einen im Umgang damit geschulten Mitarbeiter wenig Sinn ergibt, besteht außerdem die Notwendigkeit fachkundiger Arbeitskräfte (114). Die Pathologen sollten über künstliche Intelligenz und maschinelles Lernen belehrt werden, sodass sie sowohl die Vorteile bewusst nutzen als auch Limitationen dieser Technologie zum Schutz der Patienten verstehen können. Der Ausbildungsprozess müsste langfristig gesehen neue Themen wie Gesundheitsinformatik, Computerwissenschaften und Statistik aufnehmen (114). In Anbetracht der begrenzten Ressourcen und zahlreichen klinischen Aufgaben bei sowieso beschränkter Zeit, ist ein tiefes Verständnis auch vor dem Hintergrund der großen Dynamik in diesem Feld nicht von jedem Kliniker zu erwarten. Dennoch wird künstliche Intelligenz nicht nur in der Medizin, sondern auch in unserem alltäglichen Leben immer präsenter, wodurch mit einer steigenden Anzahl von Klinikern zu rechnen ist, die sich für eine Fortbildung in diesem Bereich interessieren könnten (114). All diesen Punkten übergeordnet ist das Problem des Fehlens prospektiver, randomisierter, multizentrischer Studien, die den Nutzen des Einsatzes für Pathologen und Patienten bestätigen und so die Sinnhaftigkeit einer flächendeckenden Implementierung überhaupt erst bestätigen würden (113).

Dass die genannten Widrigkeiten und Probleme, die einer Implementation von Deep Learning Modellen noch im Wege stehen, bearbeitet und gelöst werden können und der reale Einsatz wie beschrieben möglich und erfolgsversprechend ist, zeigt ein Blick nach China (114): China weist nicht nur eine große Population auf, sondern besitzt auch ein relativ zentralisiertes Gesundheitssystem, sodass die Menge an Daten für Training und Validierung von Algorithmen enorm sind (114, 138). Es wurde ein konkreter Entwicklungsplan für künstliche Intelligenz vorgelegt, der auch starke Unterstützung im Gesundheitswesen beinhaltet (138): Er beschreibt erleichterte politische Grundlagen und finanzielle Unterstützungen für Start-ups in dieser Domäne

und Anstrengungen im Bereich der Implementation. Durch diese Maßnahmen stiegen die Zahlen der an diesem Themengebiet arbeitenden Unternehmen beträchtlich (138). Weiterhin wurden bereits einige Screening-Systeme in verschiedenen Krankenhäusern in klinischen Studien implementiert: Ein Beispiel für einen erfolgreichen Versuch stellt ein Screening- und Empfehlungssystem für die Diagnose von wichtigen Augen- und Systemerkrankungen dar, welches in einer großen Risikopopulation auf riesiger Fläche eingesetzt wird (114). Vorläufige Ergebnisse beschreiben hohe Genauigkeit der generierten Diagnosen, vergleichbar mit denen eines trainierten Augenarztes (114).

Da nur wenige pathologische Institutionen bisher vollständig digitale Arbeitsweisen implementiert haben, könnten die genannten Einsatzmöglichkeiten jede für sich als schon frühzeitig einsetzbare Zwischenschritte, die den Wert dieser neuen Technologien unter Beweis stellen, den Weg zur vollständigen Digitalisierung der Pathologie bereiten (113). Dabei werden voraussichtlich vor allem Systeme, die in Kooperation mit dem Menschen arbeiten und diesen in seiner Verantwortung der Entscheidungsfindung unterstützen, Vorreiter sein. Zur einer solchen Nutzung ermuntert auch der Leitfaden „Digitale Pathologie“ des Berufsverbandes Deutscher Pathologen e.V. , der die Wahlfreiheit des Pathologen bezüglich der Methode der Diagnosefindung betont, zugleich aber die einhergehende Verantwortung beim Pathologen belässt (139).

Gelingt es die durchaus lösbaren Hürden und Problemstellungen der Implementierung zu bewältigen, werden die beschriebenen Einsatzmöglichkeiten nutzbar und stellen der Pathologie eine Transformation und Revolutionierung mit Vorteilen wie einer Erhöhung der Präzision und Arbeitsgeschwindigkeit in Aussicht.

6 Zusammenfassung

Magenkarzinome sind häufig vorkommende Malignome und mit einer hohen Morbidität und Mortalität assoziiert. Siegelringzellkarzinome repräsentieren einen Subtyp mit steigender Inzidenz, der vor allem jüngere Patienten betrifft und mit einer schlechteren Prognose einhergeht, weshalb eine prompte und korrekte Diagnose hier besonders entscheidend ist. Die histopathologische Klassifikation von Magenkarzinomen ist eine der Routineaufgaben der Pathologie, die Erkennung von Siegelringzellkarzinomen stellt dabei eine diagnostische Herausforderung dar. In Zeiten individualisierter Diagnosen und Therapien mit zunehmender Arbeitsbelastung bei sinkender Anzahl von Pathologen, stellt die Digitalisierung der Pathologie einen möglichen Ausweg dar. Damit würde mittels maschinellem Lernen, einem Teilgebiet der künstlichen Intelligenz, auch eine automatisierte Analyse von Gewebeschnitten ermöglicht. Ziel dieser Arbeit war die Evaluation des möglichen Einsatzes von maschinellem Lernen für die Gewebeklassifizierung bei Magen- beziehungsweise Siegelringzellkarzinomen und die Ableitung der Implikationen für eine potenzielle klinische Implementation. Es wurde ein Deep Learning Modell zur Klassifikation von Gewebetypen (Epithel, Muskel, Immunzellen, Fett, Karzinom) im Bereich von Magenkarzinomgewebe ins Besondere siegelringzelliger Differenzierung trainiert. Für die einzelnen Klassen konnte eine AUROC von 0,994 bis 0,999 erreicht werden mit einer Gesamtsystemleistung von 0,997. Klinisch nutzbare Softwarevisualisierungen wie eine Anzeige der Klassifikation während der Mikroskopie oder Klassifikationskarten wurden implementiert. Als mögliche Einsatzbereiche wurden die eigenständige Übernahme wenig komplexer, aber arbeitsaufwändiger Arbeitsschritte, eine Triagierung von vorliegenden Fällen, der Einsatz als Unterstützungssystem für Pathologen sowie die Funktion als Zweitmeinung im Bereich der Qualitätssicherung diskutiert, was Vorteile wie eine optimierte Nutzung der menschlichen Ressourcen oder eine Erhöhung der diagnostischen Genauigkeit erbringen könnte. Einschränkungen für die praktische Anwendung liegen noch in Feldern wie Finanzierung, Datenschutz, Qualitätskontrolle oder unzureichender Studienlage für eine klinische Anwendung. Eine Überwindung dieser Hürden ist für die nicht allzu ferne Zukunft zu erwarten, wobei die Pathologie mit ihren bildbasierten Arbeitsmethoden Teil der Vorreiter für die Implementation von maschinellem Lernen in den klinischen Alltag werden könnte.

7 Ausblick

Bevor eine Anwendung im klinischen Alltag Realität werden kann, müssen Untersuchungen zum Nachweis eines praktischen Nutzens, zum Beispiel im Rahmen eines klinischen Simulationsexperimentes der Anwendung, erfolgen. Weitere Untersuchungen wären auch im Rahmen einer Anwendung des Modelles für eine tiefere Analyse der Gewebeschnitte zur Betrachtung der Tumormikroumgebung, um besseres Verständnis für das Verhalten maligner Zellen zu erhalten, sicherlich aufschlussreich.

Künstliche Intelligenz beziehungsweise maschinelles Lernen kann aber nicht nur für den Bereich des Magenkarzinomes, sondern auch für Untersuchungen zu anderen Karzinomen und auch Nicht-Tumorerkrankungen Vorteile bringen. Weiterführende Untersuchungen werden benötigt, um neue Algorithmen für medizinische Aufgaben zu entwickeln und bestehende zu verbessern. Es ist für die nicht allzu ferne Zukunft zu erwarten, dass solche Felder, die eine starke Fixierung auf bildbasierte Arbeitsmethoden aufweisen, als Vorreiter KI-basierte Technologien zu implementieren versuchen. Hierzu gehören beispielsweise die Radiologie, Ophthalmologie, Dermatologie, aber auch die Pathologie.

8 Literaturverzeichnis

1. Bray F RJ, Masuyer E, Ferlay J. . Global estimates of cancer prevalence for 27 sites in the adult population in 2008. 2013.
2. E. Brambilla WDT, T. Colby. WHO-Classification for Tumours of the Digestive System: IARC-Press; 2010.
3. Forman D. BV. Gastric cancer: global pattern of the disease and an overview of environmental risk factors. 2006.
4. B.Barnes JB, N.Buttmann-Schweiger. Bericht zum Krebsgeschehen in Deutschland 2016. Robert-Koch-Institut. 2016.
5. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2014;23(5):700-13.
6. Siegel R. MJ, Zou Z. . Cancer statistics. CA Cancer J Clin. 2014.
7. Herold et al.: Innere Medizin. Eigenverlag 2021.
8. AMBOSS GmbH; Kapitel: Magenkarzinom zaa, abgerufen am: 04.01.2022. https://next.amboss.com/de/article/-g0DB2?q=magenkarzinom#Z9b76bfe8d1cb208e3e39b8e6bd84d377.
9. Soerjomataram I. L-TJ, Parkin D.M. Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. 2012.
10. Kamangar F, Sheikhattari P, Mohebtash M. Helicobacter pylori and its effects on human health and disease. Arch Iran Med. 2011;14(3):192-9.
11. Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. Gut. 2001;49(3):347-53.
12. de Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol. 2012;13(6):607-15.
13. Chiba T, Marusawa H, Seno H, Watanabe N. Mechanism for gastric cancer development by Helicobacter pylori infection. J Gastroenterol Hepatol. 2008;23(8 Pt 1):1175-81.
14. Howlader N NA, Krapcho M, Neyman N, Aminou R, Waldron W, Altekruse SF, Kosary CL, Ruhl J, Tatalovich Z, Cho H, Mariotto A, Eisner MP, Lewis DR, Chen HS, Feuer EJ, Cronin KA, Edwards BK (eds). SEER Cancer Statistics Review, 1975-2008, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2008/, based on November 2010 SEER data submission, posted to the SEER web site, 2011.
15. IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins. IARC Monogr Eval Carcinog Risks Hum. 2010;94:v-vii, 1-412.
16. Ladeiras-Lopes R, Pereira AK, Nogueira A, Pinheiro-Torres T, Pinto I, Santos-Pereira R, et al. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. Cancer Causes Control. 2008;19(7):689-701.
17. Brown LM, Devesa SS. Epidemiologic trends in esophageal and gastric cancer in the United States. Surg Oncol Clin N Am. 2002;11(2):235-56.
18. Nagini S. Carcinoma of the stomach: A review of epidemiology, pathogenesis, molecular genetics and chemoprevention. World J Gastrointest Oncol. 2012;4(7):156-69.
19. Kamangar F. Socio-economic health inequalities: ever-lasting facts or amenable to change? Int J Prev Med. 2013;4(6):621-3.
20. Adler NE, Ostrove JM. Socioeconomic status and health: what we know and what we don't. Ann N Y Acad Sci. 1999;896:3-15.

21. Singh PP, Singh S. Statins are associated with reduced risk of gastric cancer: a systematic review and meta-analysis. Ann Oncol. 2013;24(7):1721-30.
22. Wu XD, Zeng K, Xue FQ, Chen JH, Chen YQ. Statins are associated with reduced risk of gastric cancer: a meta-analysis. Eur J Clin Pharmacol. 2013;69(10):1855-60.
23. R. Witzig BS, U. Fink, R. Busch, H. Gundel, A. Sendler, C. Peschel, J. R. Siewert, F. Lordick Delays in diagnosis and therapy of gastric cancer and esophageal adenocarcinoma. Georg Thieme Verlag KG2006. 1122-6 p.
24. Lordick F. AD, Borner M. et al. Magenkarzinom Onkopedia2018 [Available from: https://www.onkopedia.com/de/onkopedia/guidelines/magenkarzinom/@@guideline/html/index.html.
25. Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft DK, AWMF):S3-Leitlinie Magenkarzinom, Kurzversion 2.0, 2019 AWMF Registernummer: 032/009OL, http://www.leitlinienprogramm-onkologie.de/leitlinien/magenkarzinom/ (abgerufen am: 25.03.2020).
26. Michael Hejna EW, Philipp Tschandl,Markus Raderer. Cutaneous paraneoplastic disorders in stomach cancer: Collaboration between oncologically active dermatologists and clinical oncologists. Elsevier. 2016.
27. AWMF S3 Leitlinie: Magenkarzinom - Diagnostik und Therapie der Adenokarzinome des Magens und ösophagogastralen Übergangs 2012 [Available from: http://www.awmf.org/leitlinien/detail/ll/032-009OL.html.
28. Rosen RD, Sapra A. TNM Classification. StatPearls. Treasure Island (FL): StatPearls Publishing
Copyright © 2020, StatPearls Publishing LLC.; 2020.
29. Lauren P. THE TWO HISTOLOGICAL MAIN TYPES OF GASTRIC CARCINOMA: DIFFUSE AND SO-CALLED INTESTINAL-TYPE CARCINOMA. AN ATTEMPT AT A HISTO-CLINICAL CLASSIFICATION. Acta pathologica et microbiologica Scandinavica. 1965;64:31-49.
30. Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours: John Wiley & Sons; 2017.
31. Bass AJ, Thorsson V, Shmulevich I, Reynolds SM, Miller M, Bernard B, et al. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014;513(7517):202-9.
32. Ming SC. Gastric carcinoma. A pathobiological classification. Cancer. 1977;39(6):2475-85.
33. Takagi K, Kumakura K, Sugano H, Nakamura K. [Polypoid lesions of the stomach--with special reference to atypical epithelial lesions]. Gan no rinsho Japan journal of cancer clinics. 1967;13(10):809-17.
34. Pernot S, Voron T, Perkins G, Lagorce-Pages C, Berger A, Taieb J. Signet-ring cell carcinoma of the stomach: Impact on prognosis and specific therapeutic challenge. World J Gastroenterol. 2015;21(40):11428-38.
35. Nakamura K, Sugano H, Takagi K. Carcinoma of the stomach in incipient phase: its histogenesis and histological appearances. Gan. 1968;59(3):251-8.
36. Patel MI, Rhoads KF, Ma Y, Ford JM, Visser BC, Kunz PL, et al. Seventh edition (2010) of the AJCC/UICC staging system for gastric adenocarcinoma: is there room for improvement? Ann Surg Oncol. 2013;20(5):1631-8.
37. Lauwers G CF, Graham D, Curado M, Franceschi S. Classification of Tumours of the Digestive System. IARC Press. 2010;Zitat 9 aus Pernot, Signet ring cell carcinoma.
38. Bamboat ZM, Tang LH, Vinuela E, Kuk D, Gonen M, Shah MA, et al. Stage-stratified prognosis of signet ring cell histology in patients undergoing curative resection for gastric adenocarcinoma. Ann Surg Oncol. 2014;21(5):1678-85.

39. Taghavi S, Jayarajan SN, Davey A, Willis AI. Prognostic significance of signet ring gastric cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2012;30(28):3493-8.
40. Henson DE, Dittus C, Younes M, Nguyen H, Albores-Saavedra J. Differential trends in the intestinal and diffuse types of gastric carcinoma in the United States, 1973-2000: increase in the signet ring cell type. Archives of pathology & laboratory medicine. 2004;128(7):765-70.
41. Kwon KJ, Shim KN, Song EM, Choi JY, Kim SE, Jung HK, et al. Clinicopathological characteristics and prognosis of signet ring cell carcinoma of the stomach. Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association. 2014;17(1):43-53.
42. Gill S, Shah A, Le N, Cook EF, Yoshida EM. Asian ethnicity-related differences in gastric cancer presentation and outcome among patients treated at a canadian cancer center. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2003;21(11):2070-6.
43. Terada T. Histopathological study using computer database of 10 000 consecutive gastric specimens: (1) benign conditions. Gastroenterol Rep (Oxf). 2015;3(3):238-42.
44. Zu H, Wang H, Li C, Xue Y. Clinicopathologic characteristics and prognostic value of various histological types in advanced gastric cancer. Int J Clin Exp Pathol. 2014;7(9):5692-700.
45. Zhang M, Zhu G, Zhang H, Gao H, Xue Y. Clinicopathologic features of gastric carcinoma with signet ring cell histology. J Gastrointest Surg. 2010;14(4):601-6.
46. Ha TK, An JY, Youn HK, Noh JH, Sohn TS, Kim S. Indication for endoscopic mucosal resection in early signet ring cell gastric cancer. Ann Surg Oncol. 2008;15(2):508-13.
47. Kunisaki C, Shimada H, Nomura M, Matsuda G, Otsuka Y, Akiyama H. Therapeutic strategy for signet ring cell carcinoma of the stomach. Br J Surg. 2004;91(10):1319-24.
48. Jiang CG, Wang ZN, Sun Z, Liu FN, Yu M, Xu HM. Clinicopathologic characteristics and prognosis of signet ring cell carcinoma of the stomach: results from a Chinese mono-institutional study. J Surg Oncol. 2011;103(7):700-3.
49. Kim DY, Park YK, Joo JK, Ryu SY, Kim YJ, Kim SK, et al. Clinicopathological characteristics of signet ring cell carcinoma of the stomach. ANZ J Surg. 2004;74(12):1060-4.
50. Gronnier C, Messager M, Robb WB, Thiebot T, Louis D, Luc G, et al. Is the negative prognostic impact of signet ring cell histology maintained in early gastric adenocarcinoma? Surgery. 2013;154(5):1093-9.
51. Kim JP, Kim SC, Yang HK. Prognostic significance of signet ring cell carcinoma of the stomach. Surg Oncol. 1994;3(4):221-7.
52. Li C, Kim S, Lai JF, Hyung WJ, Choi WH, Choi SH, et al. Advanced gastric carcinoma with signet ring cell histology. Oncology. 2007;72(1-2):64-8.
53. Heger U, Blank S, Wiecha C, Langer R, Weichert W, Lordick F, et al. Is preoperative chemotherapy followed by surgery the appropriate treatment for signet ring cell containing adenocarcinomas of the esophagogastric junction and stomach? Ann Surg Oncol. 2014;21(5):1739-48.
54. Yokota T, Kunii Y, Teshima S, Yamada Y, Saito T, Kikuchi S, et al. Signet ring cell carcinoma of the stomach: a clinicopathological comparison with the other histological types. Tohoku J Exp Med. 1998;186(2):121-30.
55. Maehara Y, Sakaguchi Y, Moriguchi S, Orita H, Korenaga D, Kohnoe S, et al. Signet ring cell carcinoma of the stomach. Cancer. 1992;69(7):1645-50.

56. Otsuji E, Yamaguchi T, Sawai K, Takahashi T. Characterization of signet ring cell carcinoma of the stomach. J Surg Oncol. 1998;67(4):216-20.
57. Honore C, Goere D, Messager M, Souadka A, Dumont F, Piessen G, et al. Risk factors of peritoneal recurrence in eso-gastric signet ring cell adenocarcinoma: results of a multicentre retrospective study. Eur J Surg Oncol. 2013;39(3):235-41.
58. Mariette C, Bruyere E, Messager M, Pichot-Delahaye V, Paye F, Dumont F, et al. Palliative resection for advanced gastric and junctional adenocarcinoma: which patients will benefit from surgery? Ann Surg Oncol. 2013;20(4):1240-9.
59. Machlowska J, Puculek M, Sitarz M, Terlecki P, Maciejewski R, Sitarz R. State of the art for gastric signet ring cell carcinoma: from classification, prognosis, and genomic characteristics to specified treatments. Cancer Manag Res. 2019;11:2151-61.
60. Grady WM, Willis J, Guilford PJ, Dunbier AK, Toro TT, Lynch H, et al. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. Nat Genet. 2000;26(1):16-7.
61. Machado JC, Oliveira C, Carvalho R, Soares P, Berx G, Caldas C, et al. E-cadherin gene (CDH1) promoter methylation as the second hit in sporadic diffuse gastric carcinoma. Oncogene. 2001;20(12):1525-8.
62. Fuchs CS, Tomasek J, Yong CJ, Dumitru F, Passalacqua R, Goswami C, et al. Ramucirumab monotherapy for previously treated advanced gastric or gastro-oesophageal junction adenocarcinoma (REGARD): an international, randomised, multicentre, placebo-controlled, phase 3 trial. Lancet. 2014;383(9911):31-9.
63. Bang YJ CH, Shankaran V, Geva R, Catenacci DVT, Gupta S, Eder JP. Relationship between PD-L1 expression and clinical outcomes in patients with advanced gastric cancer treated with the anti-PD-1 monoclonal antibody pembrolizumab (MK-3475) in KEYNOTE-012. J Clin Oncol 2015. Available from: URL: http://meetinglibrary.asco.org/content/150958-156.
64. Cloetingh D, Schmidt RA, Kong CS. Comparison of Three Methods for Measuring Workload in Surgical Pathology and Cytopathology. American journal of clinical pathology. 2017;148(1):16-22.
65. Fischer AH, Jacobson KA, Rose J, Zeller R. Hematoxylin and eosin staining of tissue and cell sections. CSH Protoc. 2008;2008:pdb.prot4986.
66. Bubendorf L. FG, Obermann E., Dalquen P. (2011) Zytologische Tumorkriterien. In: Klöppel G., Kreipe H., Remmele W. (eds) Pathologie. Springer, Berlin, Heidelberg.
67. Litjens G, Sanchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports. 2016;6:26286.
68. The Royal College of Pathologists. Meeting pathology demand - Histopathology workforce census. (2018). 2018.
69. Zarbo RJ, D'Angelo R. The Henry ford production system: effective reduction of process defects and waste in surgical pathology. American journal of clinical pathology. 2007;128(6):1015-22.
70. Ho J, Ahlers SM, Stratman C, Aridor O, Pantanowitz L, Fine JL, et al. Can digital pathology result in cost savings? A financial projection for digital pathology implementation at a large integrated health care organization. Journal of pathology informatics. 2014;5(1):33.
71. Ho J, Parwani AV, Jukic DM, Yagi Y, Anthony L, Gilbertson JR. Use of whole slide imaging in surgical pathology quality assurance: design and pilot validation studies. Hum Pathol. 2006;37(3):322-31.

72. Zarella MD, Bowman D, Aeffner F, Farahani N, Xthona A, Absar SF, et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. Archives of pathology & laboratory medicine. 2019;143(2):222-34.
73. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 9-12.
74. Alpaydin EMLB, Boston: De Gruyter Oldenbourg, 2019. https://doi.org/10.1515/9783110617894 Seite 1-22.
75. Alpaydin EMLB, Boston: De Gruyter Oldenbourg, 2019. https://doi.org/10.1515/9783110617894 Seite 23-50.
76. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 13-31.
77. Otte Ralf KI, John Wiley & Sons, Incorporated, 2019, 1.Auflage, Seite 227-276.
78. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 210-250.
79. Torresani L. Weakly Supervised Learning. In: Ikeuchi K, editor. Computer Vision: A Reference Guide. Boston, MA: Springer US; 2014. p. 883-5.
80. Choy G KO, Michalski M, Do S, Samir AE, Pianykh OS, Geis JR, Pandharipande PV, Brink JA, Dreyer KJ. Current Applications and Future Impact of Machine Learning in Radiology. Radiology. 2018 Aug;288(2):318-328. doi: 10.1148/radiol.2018171820. Epub 2018 Jun 26. PMID: 29944078; PMCID: PMC6542626.
81. .
82. W.S. McCulloch WP, A logical calculus of the ideas immanent in nervousactivity,Bulletin of Mathematical Biophysics5 (1943) 115–133.
83. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 161-209.
84. Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V. M, https://www.mpg.de/gehirn, abgerufen am 03.12.2022.
85. Brock Oliver KIuRBuF, Sankt Augustin: Konrad-Adenauer-Stiftung, 2018.
86. Luber S. LN. Was ist ein Convolutional Neural Network? 2019 [Available from: https://www.bigdata-insider.de/was-ist-ein-convolutional-neural-network-a-801246/.
87. Y. LeCun LB, Y. Bengio and P. Haffner. Gradient- Based Learning Applied to Document Recognition. 1998.
88. Montana DJ, Davis L, editors. Training Feedforward Neural Networks Using Genetic Algorithms. IJCAI; 1989.
89. Kolb T. Entwicklung eines Convolutional Neural Network zur Handschrifterkennung. Angewandtes maschinelles Lernen–SS2019.28.
90. Otte Ralf KI, John Wiley & Sons, Incorporated, 2019, 1.Auflage, Seite 277-292.
91. Gajewska-Dendek E, Wróbel A, Bekisz M, Suffczynski P. Lateral Inhibition Organizes Beta Attentional Modulation in the Primary Visual Cortex. Int J Neural Syst. 2019;29(3):1850047.
92. Lerntransfer V--, Hauptautor: K. Walter, https://de.wikipedia.org/wiki/Lerntransfer.
93. Pratt LYD-btbnn, NIPS Conference: Advances in Neural Information Processing Systems 5. Morgan Kaufmann Publishers. pp. 204–211.
94. L. W. Was ist Transfer Learning? 2020 [Available from: https://datasolut.com/was-ist-transfer-learning/
95. George Karimpanal T, and Roland Bouffanais. “Self-Organizing Maps for Storage and Transfer of Knowledge in Reinforcement Learning.” Adaptive Behavior 27.2 (2018): 111–126. Crossref. Web.
96. K. He XZ, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385,2015.

97. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al., editors. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; 2015.
98. Pytel R, and Tomasz Motyka. "Data-efficient semantic segmentation via extremely perturbed data augmentation.".
99. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 275-329.
100. Frochte Jörg MLGuAiP, 2. Auflage, Carl Hanser Verlag, München, 2019, Seite 23-50. Podcast
101. https://pubmed.ncbi.nlm.nih.gov/?term=digital+pathology&filter=years.1946-2020&timeline=expanded aa.
102. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Science translational medicine. 2011;3(108):108ra13.
103. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA. 2017;318(22):2199-210.
104. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. The American journal of surgical pathology. 2018;42(12):1636-46.
105. Iizuka O, Kanavati, F., Kato, K. et al. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. Sci Rep 10, 1504 (2020). https://doi.org/10.1038/s41598-020-58467-9.
106. Song Z, Zou, S., Zhou, W. et al. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. Nat Commun 11, 4294 (2020). https://doi.org/10.1038/s41467-020-18147-8.
107. Bankhead P LM, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, McQuaid S, Gray RT, Murray LJ, Coleman HG, James JA, Salto-Tellez M, Hamilton PW. QuPath: Open source software for digital pathology image analysis. Sci Rep. 2017 Dec 4;7(1):16878. doi: 10.1038/s41598-017-17204-5. PMID: 29203879; PMCID: PMC5715110.
108. Foersch S EM, Wagner DC, Gach F, Woerl AC, Geiger J, Glasner C, Schelbert S, Schulz S, Porubsky S, Kreft A, Hartmann A, Agaimy A, Roth W. Deep learning for diagnosis and survival prediction in soft tissue sarcoma. Ann Oncol. 2021 Sep;32(9):1178-1187. doi: 10.1016/j.annonc.2021.06.007. Epub 2021 Jun 15. PMID: 34139273.
109. Schulz S WA, Jungmann F, Glasner C, Stenzel P, Strobl S, Fernandez A, Wagner DC, Haferkamp A, Mildenberger P, Roth W, Foersch S. Multimodal Deep Learning for Prognosis Prediction in Renal Cancer. Front Oncol. 2021 Nov 24;11:788740. doi: 10.3389/fonc.2021.788740. PMID: 34900744; PMCID: PMC8651560.
110. G. Huang ZL, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, 2261-2269 (2017).
111. Scikit-learn: Machine Learning in Python Pea, JMLR 12, pp. 2825-2830, 2011.
112. M. Sokolova GL, A systematic analysis of performance measures for classification tasks, Inf. Process. Manag. 45, 427-437 (2009).
113. Försch S KF, Hufnagl P, Roth W: Artificial intelligence in pathology. Dtsch Arztebl Int 2021; 118: 199-204. DOI: 10.3238/arztebl.m2021.0011.

114. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nature medicine. 2019;25(1):30-6.
115. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine. 2019;25(8):1301-9.
116. Titano JJ BM, Schefflein J, Pain M, Su A, Cai M, Swinburne N, Zech J, Kim J, Bederson J, Mocco J, Drayer B, Lehar J, Cho S, Costa A, Oermann EK. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med. 2018 Sep;24(9):1337-1341. doi: 10.1038/s41591-018-0147-y. Epub 2018 Aug 13. PMID: 30104767.
117. Gehrung M, Crispin-Ortuzar, M., Berman, A.G. et al. Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. Nat Med 27, 833–841 (2021). https://doi.org/10.1038/s41591-021-01287-9.
118. Wang D, Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.
119. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical image analysis. 2016;33:170-5.
120. Woerl AC EM, Geiger J, et al. Deep Learning Predicts Molecular Subtype of Muscle-invasive Bladder Cancer from Conventional Histopathological Slides. European Urology. 2020 Aug;78(2):256-264. DOI: 10.1016/j.eururo.2020.04.023. PMID: 32354610.
121. Thrall M, Pantanowitz, L., & Khalbuss, W. (2011). Telecytology: Clinical applications, current challenges, and future benefits. Journal of pathology informatics, 2, 51. https://doi.org/10.4103/2153-3539.91129.
122. Raab SS NR, Ruby SG. Patient safety in anatomic pathology: measuring discrepancy frequencies and causes. Arch Pathol Lab Med 2005; 129:459-66.
123. S.V. Destounis PD, W. Logan-Young, et al. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate?: Initial experience. Radiology, 232 (2004), pp. 578-584.
124. Kather J, Weis, CA., Bianconi, F. et al. Multi-class texture analysis in colorectal cancer histology. Sci Rep 6, 27988 (2016). https://doi.org/10.1038/srep27988.
125. Kather JN KJ, Charoentong P, Luedde T, Herpel E, Weis C-A, et al. (2019) Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med 16(1): e1002730. https://doi.org/10.1371/journal.pmed.1002730.
126. Kleppe A SO, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer. 2021 Mar;21(3):199-211. doi: 10.1038/s41568-020-00327-9. Epub 2021 Jan 29. PMID: 33514930.
127. Thiem D.G.E. RP, Gielisch M. et al. Hyperspectral imaging and artificial intelligence to detect oral malignancy – part 1 - automated tissue classification of oral muscle, fat and mucosa using a light-weight 6-layer deep neural network. Head Face Med 17, 38 (2021). https://doi.org/10.1186/s13005-021-00292-0.
128. Turkki R, Linder, N., Kovanen, P. E., Pellinen, T., and Lundin, J., "Identification of immune cell infiltration in hematoxylin-eosin stained breast cancer samples: texture-based classification of tissue morphologies", in Medical Imaging 2016: Digital Pathology, 2016, vol. 9791. doi:10.1117/12.2217040.
129. Eva Lykke Toft SEK, Jessica Malmqvist & John Brodersen (2019) Psychosocial consequences of receiving false-positive colorectal cancer screening results: a qualitative study, Scandinavian Journal of Primary Health Care, 37:2, 145-154, DOI: 10.1080/02813432.2019.1608040.
130. Kanavati F TMAdlmfgd-taciwsiSROd.

131. Chen PC, Gadepalli K, MacDonald R, Liu Y, Kadowaki S, Nagpal K, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nature medicine. 2019.
132. Hashimoto DA RG, Rus D, Meireles OR. Artificial Intelligence in Surgery: Promises and Perils. Ann Surg. 2018;268(1):70-76. doi:10.1097/SLA.0000000000002693.
133. Fonseca CG BM, Bluemke DA, Britten RD, Chung JD, Cowan BR, Dinov ID, Finn JP, Hunter PJ, Kadish AH, Lee DC, Lima JA, Medrano-Gracia P, Shivkumar K, Suinesiaputra A, Tao W, Young AA. The Cardiac Atlas Project--an imaging database for computational modeling and statistical atlases of the heart. Bioinformatics. 2011 Aug 15;27(16):2288-95. doi: 10.1093/bioinformatics/btr360. Epub 2011 Jul 6. PMID: 21737439; PMCID: PMC3150036.
134. Goodman B, and Seth Flaxman. "European Union regulations on algorithmic decision-making and a “right to explanation”." AI magazine 38.3 (2017): 50-57.
135. Sussillo DB, O. Opening the black box: low-dimensional dynamics in high-demensional recurrent neural networks. Neural Comput. 25, 626-649 (2013).
136. Bundesministerium für Gesundheit M, https://www.bundesgesundheitsministerium.de/themen/gesundheitswesen/medizinprodukte/marktzugangsvoraussetzungen.html, zuletzt aktualisiert: 02.08.2022, abgerufen: 26.11.2022.
137. US Food & Drug Administration. Developing Software Precertification Program: A Working Model (United States Food & Drug Administration.
138. Bundesministerium für Wirtschaft und Energie FC, https://www.medways.eu/Factsheet%20CHN%20BioTech.pdf, abgerufen am 26.11.2022.
139. Haroske G ZR, Hufnagl P, Kommission Digitale Pathologie: Leitfaden „Digitale Pathologie in der Diagnostik“: Befunderstellung an digitalen Bildern. Pathologe 2018; 39: 216-21.
140. Zeiler MD, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701 (2012).
141. Xie Z, Issei Sato, and Masashi Sugiyama. "Understanding and Scheduling Weight Decay." arXiv preprint arXiv:2011.11152 (2020).

9 Anhang

Anhangsverzeichnis

Anhang 1: Siewert-Klassifikation der Adenokarzinome des ösophagogastralen Überganges (AEG)
Tabelle orientiert an (7).

Siewert-Klassifikation	Bedeutung
I	Ösophaguskarzinom, 1-5 cm oberhalb der Z-Linie
II	Ösophaguskarzinom, 1 cm oberhalb bis 2 cm unterhalb der Z-Linie
III	Magenkardiakarzinom, 2-5 cm unterhalb der Z-Linie

Anhang 2: TNM-Klassifikation des Magenkarzinomes
Tabelle orientiert an (8).

TNM	Bedeutung
Tis	Carcinoma in situ: keine Überschreitung der Basalmembran, keine Metastasierung
T1	Auf Mukosa (T1a) und Submukosa (T1b) begrenzt
T2	Infiltration der Muscularis propria
T3	Infiltration der Subserosa
T4	Durchbruch in die Serosa (T4a) und Infiltration benachbarter Strukturen (T4b)
N1	1-2 regionäre Lymphknoten infiltriert
N2	3-6 regionäre Lymphknoten infiltriert
N3	$\geq$ 7 regionäre Lymphknoten infiltriert
M1	Fernmetastasen, Peritonealkarzinose

Anhang 3: Stadieneinteilung der UICC für das Magenkarzinom
Tabelle orientiert an (8).

UICC-Stadium	TNM-Klassifikation Magenkarzinom
0	Tis
IA	T1 N0 M0
IB	T1 N1 M0
	T2 N0 M0
II	T1 N2 M0
	T2 N1 M0
	T3 N0 M0
IIIA	T2 N2 M0

	T3 N1 M0
	T4 N0 M0
IIIB	T3 N2 M0
IV	T1-T3 N3 M0
	T4 N1-3 M0
	Tx Nx M1 (jede Fernmetastasierung)

Anhang 4: Statistische Analyse-Parameter mit Berechnung und Bedeutung
Tabelle orientiert an Scikit-Learn-Implementierung (111)

Metrik	**Formel**	**Bedeutung**
TP = true positives, TN = true negatives, FP = false positives, FN = false negatives		
recall	$\frac{TP}{TP+FN}$	Anteil der positiven Vorhersagen an allen wirklich Positiven
precision	$\frac{TP}{TP+FP}$	Korrektheit der positiven Vorhersagen
accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Anzahl der korrekten Vorhersagen über alle Vorhersagen
error rate	1 - accuracy	Fehlerhäufigkeit über alle Vorhersagen
specificity	$\frac{TN}{TN+FP}$	Korrektheit der negativen Vorhersagen
F1-Score	$\frac{2\ x\ precision\ x\ recall}{precision+recall}$	harmonische Mittelung von precision und recall
micro average	Beispiel precision: $\frac{TP1+TP2}{TP1+TP2+FP1+FP2}$	Durchschnitt durch Aggregation der Klassen, Ausgleich von Klassenungleichgewicht
macro average	Beispiel precision: $\frac{precision1+precision2}{2}$	Durchschnitt mit gleicher Relevanz aller Klassen
weighted average	Beispiel precision: $\frac{TP1+TP2}{total\ number1+total\ number2}$	Einbezug der Anzahl der Daten in jeder Klasse für die Relevanz für den Durchschnitt
precision recall curve	Auftragung von precision gegen recall	
AUROC	Auftragung von precision gegen 1-specificity	
standard deviation	$\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$	Maß für die Streuung der Werte um den Mittelwert

Anhang 5: Trainingsskript für das neuronale Netzwerk

In [1]:

```
%reload_ext autoreload
%autoreload 2
%matplotlib inline
```

In [2]:

```
import os
os.environ['http_proxy'] = "http://10.10.6.10:8080"
os.environ['https_proxy'] = "https://10.10.6.10:8080"
```

In [3]:

```
import fastai
from fastai.vision import *
#from fastai.distributed import *
import matplotlib.pyplot as plt
import pandas as pd
import torch
import torchvision
import datetime
from path import Path
#from sprinkles import *

from PIL import Image, ImageFile
ImageFile.LOAD_TRUNCATED_IMAGES = True

import sklearn.utils.class_weight as cw
```

In [4]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [5]:

```
gpu_device_number = 0
torch.cuda.set_device(gpu_device_number)
print("INFO: You are running this experiment on: "
      + torch.cuda.get_device_name(device=gpu_device_number))
```

```
INFO: You are running this experiment on: TITAN RTX
```

In [6]:

```
os.getcwd()
saving_path = os.getcwd() + "/Ergebnis_4/"
saving_path = os.path.dirname(saving_path)
if not os.path.exists(saving_path):
    os.makedirs(saving_path)
print(saving_path)
```

```
C:\Users\AGFoersch\Desktop\Siegelringzell_Training_Neu/Ergebnis_4
```

Path to Images

In [7]:

```
path_ = "D:/Datasets/Siegelringzellkarzinome_Projektkopie_fuer_Export/Tiles"
```

In [8]:

```
df = pd.read_csv(os.getcwd() + "/schluessel/4.csv", sep=',')
tmp_df = df.groupby('training_set')
training = pd.concat([tmp_df.get_group('TRAIN'), tmp_df.get_group('VALID')])
#testing = tmp_df.get_group('Test')
#test_data = vision.ImageList.from_df(testing, path_)
mod_df = df[["Path", "Label", "training_set", "is_valid"]]
```

In [9]:

```
mod_df
```

Out[9]:

	Path	Label	training_set	is_valid
0	H_2012_012639_HE_1.9 - 2019-08-01 13.38.02.ndp...	Epithelium	TRAIN	False
1	H_2012_012639_HE_1.9 - 2019-08-01 13.38.02.ndp...	Epithelium	TRAIN	False
2	H_2012_012639_HE_1.9 - 2019-08-01 13.38.02.ndp...	Epithelium	TRAIN	False
3	H_2012_012639_HE_1.9 - 2019-08-01 13.38.02.ndp...	Epithelium	TRAIN	False
4	H_2012_012639_HE_1.9 - 2019-08-01 13.38.02.ndp...	Epithelium	TRAIN	False
...	...	...	...	...
21496	SS_2017_070827_He_1.12 - 2019-08-05 20.42.02.n...	Muscle	VALID	True
21497	SS_2017_070827_He_1.12 - 2019-08-05 20.42.02.n...	Muscle	VALID	True
21498	SS_2017_070827_He_1.12 - 2019-08-05 20.42.02.n...	Muscle	VALID	True
21499	SS_2017_070827_He_1.12 - 2019-08-05 20.42.02.n...	Muscle	VALID	True
21500	SS_2017_070827_He_1.12 - 2019-08-05 20.42.02.n...	Muscle	VALID	True

21501 rows × 4 columns

In [27]:

```
tfms = get_transforms(flip_vert=True,
                      max_rotate=90.0,
                     # max_zoom=1.4,
                      max_lighting=0.3,
                      xtra_tfms=[#sprink(p=1.),
                                 cutout(n_holes=(60,80),
                                 length=(16, 16), p=1.),
                                 contrast(scale=(0.5, 2.), p=1.),
                                 #jitter(magnitude=0.005, p=.1),
                                 brightness(change=(0.4), p=1.)])
```

In [28]:

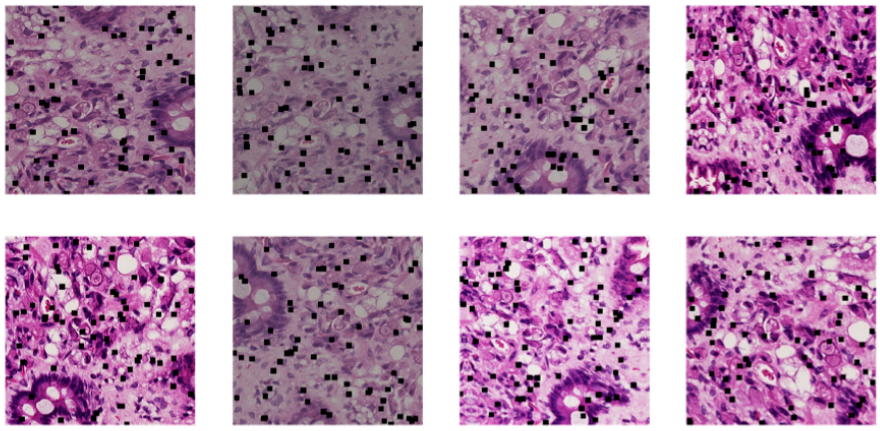
```
def get_ex(): return open_image('C:/Users/AGFoersch/Desktop/Siegelringzell_Training_Neu/H_2012_023780_He_3.7 - 2019-08-01 19.47.49.ndpi_Tumor711.jpg')
```

In [29]:

```
def plots_f(rows, cols, width, height, **kwargs):
    [get_ex().apply_tfms(tfms[0], **kwargs).show(ax=ax) for i,ax in enumerate(plt.subpl
ots(
        rows,cols,figsize=(width,height))[1].flatten())]
```

In [31]:

```
plots_f(2, 4, 24, 12, size=548)
```



In [278]:

```
data = (vision.ImageList.from_df(mod_df, path_)
          .split_from_df(col='is_valid')
          .label_from_df(cols='Label')
          #.add_test(test_data)
          .transform(tfms, size=(512, 512))
          .databunch(bs=27, val_bs=27, num_workers=16)
          .normalize(vision.imagenet_stats))
```

In [279]:

```
classweights = torch.FloatTensor(cw.compute_class_weight('balanced', np.unique(data.tra
in_ds.y.items), data.train_ds.y.items)).cuda()
print(classweights)
```

```
tensor([ 0.6681,  3.3041, 16.0769,  0.9278,  0.4853], device='cuda:0')
```

In [280]:

```
#classweights[0] = 2.8
#classweights
```

In [281]:

```
learn = vision.cnn_learner(data,
                           vision.models.densenet121,
                           metrics=[fastai.metrics.accuracy, fastai.metrics.error_rate
],
                           path=saving_path,
                           loss_func=nn.CrossEntropyLoss(weight=classweights),
                           model_dir='models')
```

In [282]:

```
Durchgang=0
Modelname = "Bestes_Model_" + str(Durchgang)

model_logger = fastai.callbacks.CSVLogger(learn,
                                          filename="Trainingsverlauf",
                                          append=True)
#save_model = fastai.callbacks.SaveModelCallback(learn,
                          #                       every="improvement",
                          #                       monitor="accuracy",
                          #                       name=Modelname)

def save_model(Durchgang, learn):
    Modelname = "Bestes_Model_" + str(Durchgang)
    save_model = fastai.callbacks.SaveModelCallback(learn,
                                                    every="improvement",
                                                    monitor="accuracy",
                                                    name=Modelname)
    return save_model

mixup = fastai.callbacks.MixUpCallback(learn)
```

In [283]:

```
Durchgang=1
```

In [119]:

```
lr_avg = 0
for i in range(3):
    learn.lr_find()
    img = learn.recorder.plot(suggestion=True, return_fig=True)
    plt.savefig(saving_path + '/first_LRF_' + str(i) + '.png')
    plt.close()
    lr_avg += learn.recorder.min_grad_lr
lr_avg /= 3
print('Avarage Learningrate: {}'.format(lr_avg))
```

0.00% [0/1 00:00<00:00]

epoch	train_loss	valid_loss	accuracy	error_rate	time

13.53% [87/643 03:02<19:28 6.3848]

```
LR Finder is complete, type {learner_name}.recorder.plot() to see the grap
h.
Min numerical gradient: 6.92E-04
Min loss divided by 10: 5.75E-03
```

0.00% [0/1 00:00<00:00]

epoch	train_loss	valid_loss	accuracy	error_rate	time

13.37% [86/643 02:35<16:50 5.9027]

```
LR Finder is complete, type {learner_name}.recorder.plot() to see the grap
h.
Min numerical gradient: 6.31E-07
Min loss divided by 10: 3.98E-03
```

0.00% [0/1 00:00<00:00]

epoch	train_loss	valid_loss	accuracy	error_rate	time

13.53% [87/643 02:37<16:46 6.5155]

```
LR Finder is complete, type {learner_name}.recorder.plot() to see the grap
h.
Min numerical gradient: 5.75E-04
Min loss divided by 10: 2.29E-03
Avarage Learningrate: 0.00042263395520019103
```

In [120]:

```
lr_avg = 0.0001
lr_avg_2 = lr_avg / 10
```

In [121]:

```
#learn.unfreeze()
learn.fit_one_cycle(30,
                    lr_avg,
                    callbacks=[save_model(Durchgang, learn), model_logger],
                    wd=0.1)
```

epoch	train_loss	valid_loss	accuracy	error_rate	time
0	1.319862	1.416155	0.447076	0.552924	11:19
1	0.762572	0.826189	0.717980	0.282020	11:19
2	0.496999	0.552789	0.822378	0.177622	11:08
3	0.428610	0.373260	0.876269	0.123731	11:16
4	0.302668	0.336265	0.883035	0.116965	11:15
5	0.225719	0.263391	0.905510	0.094490	11:08
6	0.238012	0.198998	0.931851	0.068149	11:12
7	0.174144	0.196801	0.927018	0.072982	11:10
8	0.156059	0.179163	0.939101	0.060899	11:14
9	0.156020	0.168238	0.942726	0.057274	11:07
10	0.139162	0.149204	0.949976	0.050024	11:10
11	0.130084	0.139308	0.951426	0.048574	11:03
12	0.115686	0.156185	0.947318	0.052682	11:03
13	0.114156	0.146058	0.949009	0.050991	11:08
14	0.121538	0.140035	0.953359	0.046641	11:01
15	0.110406	0.125437	0.959159	0.040841	11:21
16	0.099986	0.140946	0.952634	0.047366	11:27
17	0.081702	0.131597	0.955051	0.044949	11:00
18	0.069959	0.131485	0.955776	0.044224	10:53
19	0.103421	0.119966	0.960126	0.039874	10:57
20	0.072693	0.126876	0.957951	0.042049	11:04
21	0.072603	0.122606	0.959159	0.040841	11:00
22	0.067488	0.118079	0.962059	0.037941	10:57
23	0.067831	0.122048	0.958434	0.041566	10:54
24	0.061544	0.123526	0.957226	0.042774	10:57
25	0.071765	0.118475	0.961576	0.038424	11:03
26	0.053101	0.116314	0.960851	0.039149	11:00
27	0.062423	0.127506	0.957467	0.042533	11:04
28	0.059604	0.117871	0.960126	0.039874	11:04
29	0.065768	0.127959	0.956742	0.043258	10:56

```
Better model found at epoch 0 with accuracy value: 0.44707587361335754.
Better model found at epoch 1 with accuracy value: 0.7179797291755676.
Better model found at epoch 2 with accuracy value: 0.8223779797554016.
Better model found at epoch 3 with accuracy value: 0.876268744468689.
Better model found at epoch 4 with accuracy value: 0.8830353021621704.
Better model found at epoch 5 with accuracy value: 0.905509889125824.
Better model found at epoch 6 with accuracy value: 0.9318511486053467.
Better model found at epoch 8 with accuracy value: 0.9391010403633118.
Better model found at epoch 9 with accuracy value: 0.9427259564399719.
Better model found at epoch 10 with accuracy value: 0.949975848197937.
Better model found at epoch 11 with accuracy value: 0.9514257907867432.
Better model found at epoch 14 with accuracy value: 0.9533591270446777.
Better model found at epoch 15 with accuracy value: 0.9591590166091919.
Better model found at epoch 19 with accuracy value: 0.9601256847381592.
Better model found at epoch 22 with accuracy value: 0.962058961391449.
```

In [122]:

```
learn.load(saving_path + "/models/Bestes_Model_" + str(Durchgang))
learn.save("Zusaetzliche_Kopie_Durchgang_" + str(Durchgang))
learn.export(saving_path + "\\" + "Model_Nach_" + str(Durchgang)+"_Stufe.pkl")
```

In [285]:

```
preds = learn.get_preds()
```

In [286]:

```
class_dict = learn.data.c2i
keys = ['Patient_ID', 'Label', 'Prediction'] + [k for k in class_dict.keys()]
valid = df.groupby('training_set').get_group('VALID').reset_index(drop=True)
valid.rename(columns={'patient_ID':'Patient_ID'}, inplace=True)
values = [valid.Patient_ID, valid.Label, [learn.data.classes[x] for x in np.argmax(pred
s[0], axis=1)]] + [pd.Series(preds[0][:,class_dict.get(k)]) for k in class_dict.keys()]
```

In [287]:

```
df_ = pd.DataFrame(dict(zip(keys, values)))
```

In [288]:

```
df_
```

Out[288]:

	Patient_ID	Label	Prediction	Epithelium	Fat	Immune	Muscle	Tumor
0	H_2008_023856	Tumor	Tumor	1.396253e-01	0.000202	0.000048	0.000017	0.860108
1	H_2008_023856	Tumor	Tumor	1.712591e-01	0.000393	0.000152	0.000451	0.827745
2	H_2008_023856	Tumor	Tumor	1.023482e-02	0.000395	0.006164	0.000049	0.983157
3	H_2008_023856	Tumor	Tumor	2.415699e-03	0.000435	0.025272	0.000031	0.971845
4	H_2008_023856	Tumor	Tumor	2.070640e-02	0.000085	0.000027	0.000041	0.979140
...	...	...	...	...	...	...	...	...
4133	S_2017_070827	Muscle	Muscle	5.584193e-06	0.000245	0.000178	0.998448	0.001124
4134	S_2017_070827	Muscle	Muscle	9.922633e-08	0.000084	0.000150	0.999744	0.000022
4135	S_2017_070827	Muscle	Muscle	1.324950e-08	0.000027	0.000037	0.999808	0.000127
4136	S_2017_070827	Muscle	Muscle	1.117933e-06	0.000033	0.001299	0.998197	0.000469
4137	S_2017_070827	Muscle	Muscle	1.320192e-05	0.000201	0.001028	0.998440	0.000317

4138 rows × 8 columns

In [289]:

```
df_.to_csv(saving_path + '/Ergebnis_5.csv', index=False)
```

In [290]:

```
class_dict = {0: 'Epithelium',
              1: 'Fat',
              2: 'Immune',
              3: 'Muscle',
              4: 'Tumor'}

#class_dict = {0: 'Altered',
#               1: 'no'}
```

In [291]:

```
preds_2, y = learn.get_preds()
```

Anhang 6: Einfache Beschreibungen der angewandten Trainingsparameter
Die Erklärungen für batch size, number of workers und loss function beruhen auf Informationen von (111). Quelle für die Beschreibung der Lernrate ist (140), für weight decay (141).

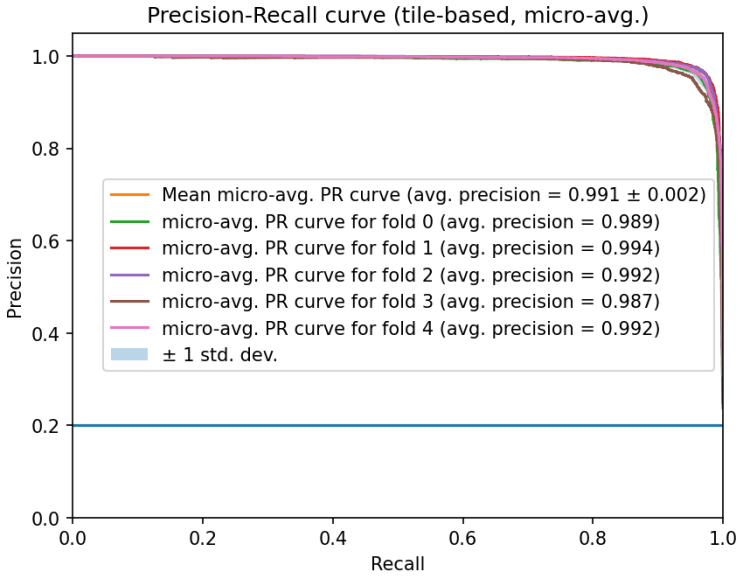
batch size	Die batch size beschreibt die Anzahl der Daten, die pro Durchlauf dem Netzwerk zur Verfügung gestellt werden, bevor eine Aktualisierung z.B. mit Neueinstellung der Gewichtungen zwischen den Perzeptronen erfolgt.
number of workers / Kerne	Die number of workers beschreibt die Anzahl der genutzten Subprozesse. Während ein Durchlauf bearbeitet wird, können weitere Subprozesse bereits die folgenden Durchgänge laden, sodass die Arbeitsgeschwindigkeit erhöht wird.
loss function	Die loss function beschreibt den Wert des Fehlers, der in der Differenz zwischen der Ausgabe des Algorithmus und dem vorgegebenen Zielwert liegt. Über eine Minimierung dieses Fehlers soll eine Maximierung der accuracy erreicht werden.
learning rate / Lernrate	Die Lernrate beschreibt wie stark das Netzwerk die Gewichtung einzelner Perzeptrone in Bezug auf erkannte Fehler nach jedem Durchlauf anpasst.
weight decay	Die Anwendung eines weight decay dient durch Nutzung einer mathematischen Operation der Vereinfachung von komplexen Modellen mit vielen Parametern. So soll overfitting vorgebeugt werden.

Anhang 7: Rohdaten der Trainingskurve des Deep Learning Modelles für Fold 4
Die Tabelle zeigt die Werte der loss function für Training und Validierung sowie die accuracy für die einzelnen Epochen von Fold 4 der 5-fold cross validation.

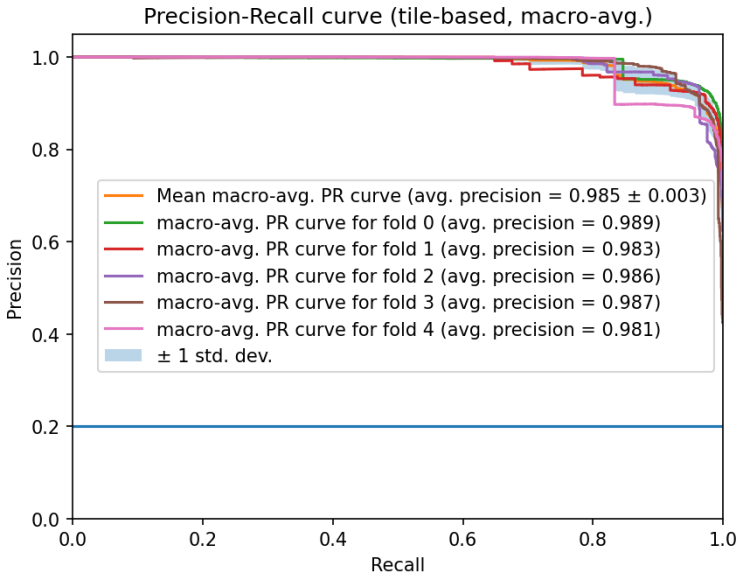
Epoche	train loss	valid loss	accuracy
1	1.319862	1.416155	0.447076
2	0.762572	0.826189	0.717980
3	0.496999	0.552789	0.822378
4	0.428610	0.373260	0.876269
5	0.302668	0.336265	0.883035
6	0.225719	0.263391	0.905510
7	0.238012	0.198998	0.931851
8	0.174144	0.196801	0.927018
9	0.156059	0.179163	0.939101
10	0.156020	0.168238	0.942726
11	0.139162	0.149204	0.949976

12	0.130084	0.139308	0.951426
13	0.115686	0.156185	0.947318
14	0.114156	0.146058	0.949009
15	0.121538	0.140035	0.953359
16	0.110406	0.125437	0.959159
17	0.099986	0.140946	0.952634
18	0.081702	0.131597	0.955051
19	0.069959	0.131485	0.955776
20	0.103421	0.119966	0.960126
21	0.072693	0.126876	0.957951
22	0.072603	0.122606	0.959159
23	0.067488	0.118079	0.962059
24	0.067831	0.122048	0.958434
25	0.061544	0.123526	0.957226
26	0.071765	0.118475	0.961576
27	0.053101	0.116314	0.960851
28	0.062423	0.127506	0.957467
29	0.059604	0.117871	0.960126
30	0.065768	0.127959	0.956742

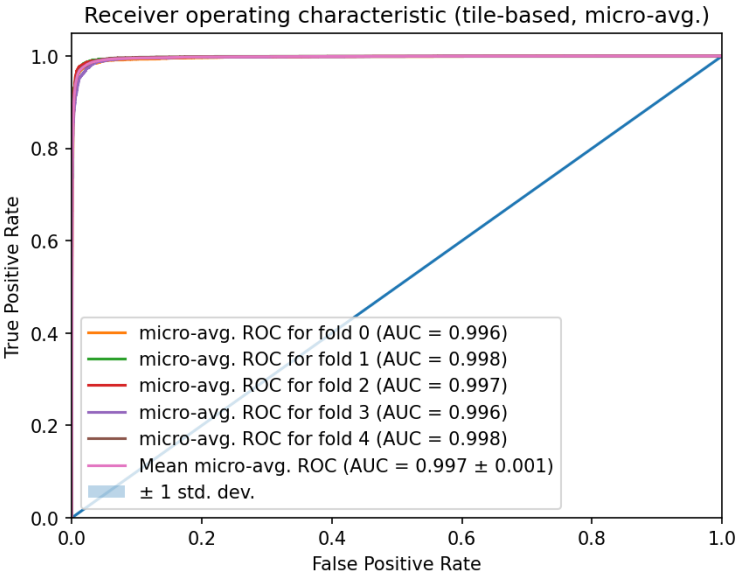
Anhang 8: micro average PR-curves für die einzelnen Folds und gemittelt über alle Folds



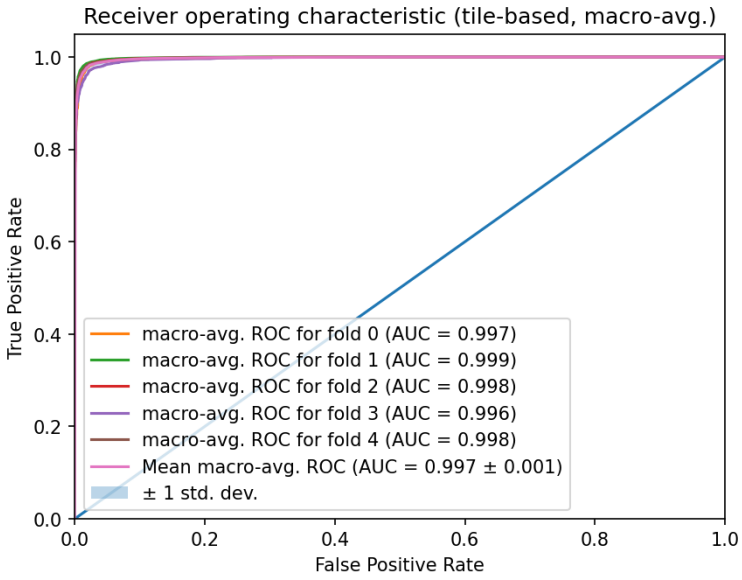
Anhang 9: macro average PR-curves für die einzelnen Folds und gemittelt über alle Folds



Anhang 10: micro average AUROC-curves für die einzelnen Folds und gemittelt über alle Folds



Anhang 11: macro average AUROC-curves für die einzelnen Folds und gemittelt über alle Folds



10 Danksagung

An dieser Stelle möchte ich allen beteiligten Personen meinen Dank aussprechen, die mir mein Studium und die Vollendung meiner Promotion ermöglicht haben.

11 Lebenslauf