Aus der

Universitäts-Hautklinik Tübingen

Sektion Dermatologische Onkologie

Prognose des Gesamtüberlebens bei Patienten mit malignem Melanom anhand Bildanalyse histologischer Schnitte durch künstliche, vortrainierte, neuronale Netze

Inaugural-Dissertation
zur Erlangung des
Doktorgrades
der Medizin

der Medizinischen Fakultät
der Eberhard Karls Universität
zu Tübingen

vorgelegt von

Abu-Ghazaleh, Amar

2023

Dekan: Professor Dr. B. Pichler

1. Berichterstatter: Professor Dr. T. Eigentler
2. Berichterstatter: Privatdozent Dr. U. Vogel

Tag der Disputation: 02.06.2023

Gewidmet meinen Eltern
Für alles, was sie für mich getan haben,
und für alles, was sie für mich noch tun
werden.

Inhaltsverzeichnis

Abkürzungsverzeichnis

2YS	Two-year survival – 2-Jahres-ÜL
5YS	Five-year survival – 5-Jahres-ÜL
AJCC	American Joint Committee on Cancer
ALM	Akral-lentiginöses Melanom
AOI	Area of interest/ Area of interest-Modell
AUC	Area under the curve Fläche unter der Kurve
CNN	Convolutional neural network
ED	Erstdiagnose
GK	Gesamtkohorte
KI	Künstliche Intelligenz
LMM	Lentigo-Maligna-Melanom
NM	Noduläres Melanom

OS	Overall survival – Gesamtüberleben
RNN	Recurrent neural network
ROC	Receiver operating characteristic
SSM	Superfiziell-spreitendes Melanom
TK	Trainingskohorte
ÜB	Übersichtsbild/ Übersichtsbild-Modell
UI	User interface - Benutzeroberfläche
VK	Validierungskohorte

Zur besseren Lesbarkeit wurde in dieser Dissertation die im Sprachgebrauch gängige männliche Form verwendet. Dies soll keine Wertung darstellen und es wird darauf hingewiesen, dass - wenn nicht anders genannt - alle Geschlechterformen gemeint sind.

Eine Publikation auf der Grundlage der Daten dieser Arbeit ist erfolgt (siehe 8. Veröffentlichungen). Entsprechende Überschneidungen wurden durchgehend in dieser Arbeit zitiert.

1 Einleitung

1.1 Epidemiologische Bedeutung des malignen Melanoms

Unter einem malignen Melanom ist eine aggressive Neoplasie der pigmentproduzierenden Zellen zu verstehen - den Melanozyten [1]. Das maligne Melanom ist auch heute noch eine Tumorentität mit einer sehr hohen Mortalität. Unzählige Tumorarten nehmen aktuell in der Prävalenz ab, jedoch trifft dies nicht für das maligne Melanom zu [2]. So zeigt eine Analyse der nationalen Krebsdatenbanken in Frankreich in den Jahren 1980-2000 eine Zunahme der Inzidenz von 2,5:10.000 auf 7,5:10.000 [3]. Eine aktuellere Analyse von Ghazawi et al. aus dem Jahr 2019 zeigte auf, dass im Zeitraum von 1992-2010 in allen Provinzen Kanadas die Inzidenz des malignen Melanoms signifikant zugenommen

hat [4]. Zunächst galt das maligne Melanom primär als eine Erkrankung des globalen Nordens aufgrund der Tatsache, dass die Mehrheit der Bevölkerung den Hauttypen I und II nach Fitzpatrick zugeordnet werden können. Jedoch scheint sich nun auch in den Ländern des globalen Südens eine Zunahme der Prävalenz zu zeigen. Nach Panda et. al findet man zwar im Vergleich zu westlichen Ländern eine niedrigere Prävalenz in Indien, jedoch sei diese ansteigend [5]. Es gibt diverse Risikofaktoren für die Entstehung eines malignen Melanoms. Dazu zählen exogene Faktoren wie die UV-Exposition und endogene Faktoren wie der Hauttyp [6].

Zusammenfassend besteht in Bezug auf das Maligne Melanom eine zunehmende Relevanz auf globaler Ebene. In der Behandlung des malignen Melanoms zeigen sich hingegen vielversprechende Veränderungen.

Behandlungen mit Signaltransduktionsinhibitoren (BRAF-, c-KIT, MEK-Inhibitoren), sowie Immuntherapien wie PD-1 Antikörper gehören zu diversen Therapieschemata im klinischen Alltag [7, 8].
Eine genaue Prognoseaussage treffen zu können, ist daher für das maligne Melanom zunehmend relevant. Stetige Veränderungen der Diagnostik und neue Technologien bieten neue Anhaltspunkte, verschiedenste Faktoren der Krebserkrankung zu analysieren, und bieten Möglichkeiten zur Erstellung verschiedenster Prognosemodelle [9].

1.2 Prognose

Eine möglichst genaue Prognose abzugeben, nimmt weiterhin einen hohen Stellenwert im klinischen Alltag eines jeden Mediziners ein. Unter einer medizinischen Prognose ist eine Einschätzung des Krankheitsverlaufes zu

verstehen [10]. Eine Prognose ist keine definitive Aussage, sondern lediglich eine Vorhersage der wahrscheinlichsten Ereignisse meist anhand bekannter Prognosefaktoren, die innerhalb Studien als relevant erhoben wurden. Um einige Beispiele aufzulisten, fallen unter diese Prognosefaktoren symptomatische Faktoren wie Abgeschlagenheit oder Gewichtsverlust, histologische Faktoren wie Tumordicke oder Infiltrationstiefe oder auch epidemiologische Faktoren wie sozioökonomischer Status und je nach Krankheit diverse weitere Faktoren.

Zu einer Krebsprognose gehört auch das sogenannte „survival“. Am gebräuchlichsten in der Klinik als auch in Studien ist das „5-year overall survival“ oder Fünf-Jahres Überleben, bei der die überlebende Rate einer Patientengruppe mit einer gemeinsamen

Erkrankung nach fünf Jahren angegeben wird. Es kann jedoch auch das „one-year“- oder „two-year survival“ angeben werden.
Demgegenüber steht das „overall survival“, bei welchem der noch lebende Anteil der Patienten zu einem bestimmten Zeitpunkt meist prozentual angegeben wird. Die Angabe erfolgt grundsätzlich innerhalb von Studien und der Zeitraum und die genaue Todesursache werden dabei oftmals vernachlässigt.
Eine initiale Zielsetzung unserer Arbeitsgruppe war es, die 5-Jahres Prognose des malignen Melanoms anhand der KI zu schätzen. Start unseres Projektes war das Jahr 2020 und aus diesem Grund setzten wir den spätesten Erstdiagnosezeitpunkt unserer Patienten, die ausgewertet werden sollten, auf das Ende des Jahres 2015. So konnten wir das reale 5-Jahres Überleben wissen und es mit der KI vergleichen.

1.3 Prognosefaktoren des malignen Melanoms

1.3.1 Übersicht

Aufgrund der nachweislich zunehmenden Relevanz, der steigenden Inzidenz und der Verbesserung der Therapiemöglichkeiten ist ein wichtiger Aspekt des malignen Melanoms eine möglichst genaue Prognoseabgabe.

Will man sich der tatsächlichen Prognose des malignen Melanoms mithilfe einer KI annähern, lohnt es sich, zuerst einmal zu betrachten welche Prognosefaktoren aktuell die stärkste Aussagekraft haben und im klinischen Alltag verwendet werden. Balch et al. erfassten 2001 in einer Multivariatanalyse die zwei stärksten Prognosefaktoren als die Tumordicke nach Breslow und die Ulzeration. Weiterhin statistisch signifikant sind das Alter, die Lokalisation des Primärmelanoms, die Invasionstiefe und das Geschlecht mit

„männlich“ als schlechteren Prognosefaktor [11].

Zu ähnlichen Ergebnissen kamen auch Wisco et al. 2012. Diese nennen die Tumordicke nach Breslow als den wichtigsten Prognosefaktor. Jedoch nützlich für eine genaue Prognoseabgabe sei es auch die Mitoserate und Ulzeration hinzuzuziehen. Bei einer fortgeschrittenen Erkrankung sei die Beteiligung von Lymphknoten noch von Bedeutung. Ein weiterer Punkt, den Wisco et al. benennen sei der sozioökonomische Status. Melanomerkrankte mit einem niedrigeren sozioökonomischen Status weisen eine signifikant schlechtere Prognose auf [9, 12].
Barnhill et al. nennen 1996 zusätzlich den histologischen Subtyp (nodulär, superfiziell-spreitend etc.) als wichtigen prognostischen Marker. Diskutiert wird dementsprechend auch,

dass die besonders schlechte Prognose des nodulären Melanoms letztlich auf die Tumordicke zurückgeht, die bereits als negativer Prognosefaktor erwähnt wurde [13].

Es ist festzuhalten, dass vor allem histologische Faktoren bei der Prognose des malignen Melanoms im Vordergrund stehen, was die Frage aufwirft, inwieweit neuronale Netzwerke digitale histologische Bilder erkennen, jene histologische Faktoren auswerten und eine Prognose abgeben können.

1.3.2 Tumordicke

Die Tumordicke nach Breslow wurde ursprünglich in folgende fünf Klassen unterteilt [14]:

1. <0,75 mm
2. 0,76-1,50 mm
3. 1,51-2,25 mm

4. 2,26-3,00 mm
5. >3,00 mm

Breslow et al. beschrieben 1970, dass Patienten, die der Klasse 1 zugeordnet waren, bei weitem die beste Prognose hatten [14]. Vielmehr noch stellten sie 1975 die Überlegenheit der Tumordicke gegenüber der Invasionstiefe fest [15].

Da die untersuchten feingeweblichen Melanomschnitte dieser Arbeit in den Jahren 2012-2015 erstmals diagnostiziert wurden, soll die Einteilung des AJCC aus dem Jahr 2009 verwendet werden [16]:

1. T1 - ≤1,00 mm
2. T2 - 1,01-2,00 mm
3. T3 - 2,01-4,00mm
4. T4 - >4,00 mm

Die neueste Version des AJCC aus dem Jahr 2018 enthält eine zusätzliche Neuerung bei der Unterscheidung in T1a und T1b. T1a beschreibt

Melanome <0,8 mm ohne Ulzeration. T1b beschreibt alle Melanome von 0,8-1,0 mm mit oder ohne Ulzeration, sowie Melanome <0,8 mm mit Ulzeration [17].

1.3.3 Ulzeration

Unter der Ulzeration ist ein kompletter Substanzverlust der Epidermis oberhalb des Melanoms, meist mit einer Reaktion der betroffenen Stelle, zu verstehen [18].

Bei der Festlegung, ob eine Ulzeration vorliegt oder nicht, ist es wichtig zu differenzieren zwischen einer tatsächlichen Ulzeration oder einer lediglich einfachen Ablösung der Epidermis vom unteren Tumor [19].

Hout et al. arbeiteten 2012 heraus, dass nicht nur die Präsenz, sondern auch das Ausmaß der Ulzeration ein wesentlicher prognostischer Faktor ist [18].

1.3.4 Invasionslevel

Die Angabe des Invasionslevels wurde maßgeblich durch eine 1969 publizierte Arbeit von Clark et al. bestimmt. Hierin wurde die Invasion des Tumors in fünf Level unterteilt – den sogenannten „Clark-Levels“ [20]:

I. Alle Tumorzellen sind oberhalb des Basalmembran. Es handelt sich per definitionem um ein Melanoma in situ.
II. Die Tumorzellen haben durch die Basalmembran das Stratum papillare infiltriert. Das Stratum reticulare ist nicht infiltriert.
III. Die Tumorzellen durchsetzen das gesamte Stratum papillare und erreichen den Bereich zwischen Stratum papillare und reticulare
IV. Die Tumorzellen haben das Stratum reticulare infiltriert.
V. Die Tumorzellen reichen bis in das

Subkutangewebe.

In einer Multivariatanalyse wurde 1993 erhoben, dass das Clark-Level nicht nur ein unabhängiger, signifikanter Prognosefaktor ist, sondern auch zusätzlich zur Breslow-Dicke zur genaueren prognostischen Einschätzung herangezogen werden sollte [21].

1.3.5 Histologische Subtypen

Clark et al. haben 1969 noch zwischen drei hauptsächlichen histologischen Tumorsubtypen unterschieden [20]. Geläufiger ist heutzutage eine Unterscheidung in die vier häufigsten Subtypen: [22]

- Superfiziell-spreitendes Melanom (SSM) – Das SSM weist ein horizontales Wachstumsmuster auf. Meist wächst es pagetoid [22, 23].
- Noduläres Melanom (NM) – Das NM

weist einen überwiegend knotigen Anteil auf. Im Gegensatz zum SSM sind noduläre Melanome meist lateral scharf begrenzt [23, 24].

- Lentigo-Maligna-Melanom (LMM) – Geht per definitionem immer mit einer Elastosis solaris einher. Die Farbe des Tumors ist meist bräunlich und der Tumor ist nicht erhaben [20, 25].
- Akrolentiginöses Melanom (ALM) - Das ALM wächst hauptsächlich plantar, palmar und unter den Fingernägeln. Es zeigt sich global beim ALM eine gleiche Verteilung unter allen ethnischen Gruppen. Sie ist der häufigste Subtyp in vielen asiatischen und afrikanischen Ländern [25, 26].

In einer umfangreichen Analyse betrachteten Latanzi et al. 2018 die Datenbank

„Surveillance, Epidemiology, and End Results (SEER)“ mit über 100.000 Patienten und das eigene Institutionsregister mit über 1.600 Patienten. Es konnte der histologische Subtyp als unabhängiger Vorhersagewert für das Überleben bei Melanomerkrankten herausgearbeitet werden [27].

1.4 Künstliche neuronale Netze – convolutional neural network

Zunehmende medizinische Bedeutung gewinnen in den letzten Jahrzehnten tiefe, neuronale Netze oder besser in der englischen Form als „deep neural network“ oder „convolutional neural network“ (CNN) bekannt [28]. Hierunter ist eine künstliche Intelligenz zu verstehen, die sich an den biologischen Aufbau des menschlichen Gehirns anlehnt. Unter dem sogenannten „deep learning“ versteht man die parallele, flexible und allgemeine Lernweise,

die neuronale Netze nutzen. Gegenüberstehend ist das herkömmliche maschinelle Lernen, das strenge Algorithmen und Muster nutzt. Dieses war in der Anwendung sehr aufwendig und dazu noch ungenau, weshalb der Bedarf eines neuen künstlichen Lernens entstand. Relativ schnell wurde erkannt, dass das deep learning sich viel besser eignete als das herkömmliche maschinelle Lernen [29]. Goodfellow et al. definieren das deep learning der deep neural networks mit einem anderen vordergründigen Aspekt. Der Schwerpunkt ist hier, dass beim deep learning simple Konzepte gelernt werden, die dem CNN erlauben auf ein ähnliches, aber komplexeres Konzept zu schließen. Ein wesentlicher Vorteil ist dementsprechend, dass kein menschlicher Nutzer dauerhaft dem CNN ein Konzept durch Daten etc. beibringen muss [30].

Die Datenverarbeitung bei CNN ist nicht von Punkt A nach B aufzufassen, sondern ähnelt der parallelen Informationsverarbeitung in Synapsen. Das CNN verarbeitet meistens die Information über einen hierarchischen Aufbau, der mit Quervernetzungen zu einer tiefgründigen Struktur führt. Die Informationsausgabe auf einer Ebene führt zur Eingabe auf einer anderen Ebene. Rampasek et al. führt noch weiter das Beispiel der Bilderkennung aus. So werden auf der ersten Ebene lediglich Formen wie Kreise, Dreiecke und Linien erkannt. Auf den höheren Ebenen folgen schließlich komplexere Strukturen bis zu menschlichen Gesichtern [31]. Hinzu kommt die gewünschte Eigenschaft der neuronalen Netze, dass sie in der Lage sind, diese gelernten Daten umzusetzen und in irgendeiner Form als Daten auszugeben [32].

1.5 Google teachable machine

1.5.1 Wobei handelt es sich bei der Google teachable machine?

Die „Google teachable machine“ ist ein webbasiertes Tool, welches auf der Basis vortrainierter neuronaler Netze Modelle erstellt. Google selbst erklärt die teachable machine auf ihrer Website: „Teachable Machine nutzt ein vortrainiertes neuronales Netzwerk und die […] erstellten Klassen bilden gewissermaßen die letzte Ebene“ [33]. Ein erstelltes Modell ordnet also eine Datei meist in Form eines Bildes einer vom Nutzer vortrainierten Klasse in prozentualer Angabe zu. Noch etwas konkreter kann folgendes Beispiel herangezogen werden. Es werden zwei Klassen trainiert. In der einen Klasse werden für das Training Bilder verwendet, die hauptsächlich eine rote Farbe enthalten. In der zweiten Klasse sollen hauptsächlich blaue Bilder für das Training

hinterlegt werden. Nun wird ein Bild zur Validierung eingegeben; eine Tomate würde erwartungsgemäß der ersten Klasse zugeordnet werden [9].

Dabei nutzen die teachable machine das Grundgerüst des „tensorflow“. Es handelt es sich um ein 2015 erstmals veröffentlichtes Framework, das von Google erstellt wurde, um die künstliche Intelligenz und deren Training der Öffentlichkeit zugänglich zu machen. Die Plattform „tensorflow“ verwaltet dabei jegliche Datenverarbeitung, erlaubt das künstliche Lernen und erstellt Modelle [34]. Aufgrund der einfachen Handhabbarkeit und öffentlichen Zugänglichkeit, konnte sich die Plattform in den letzten Jahren durchsetzen [9, 35].
Eine erste Version der teachable machine, die es erlaubte drei Klassen anhand von Bildern zu erstellen wurde 2017 veröffentlicht. Die aktuelle

Version erlaubt beliebig viele Klassen und kann u.a. auch an Tönen trainiert werden [33].

1.5.2 Erstellen eines Modells

Das Erstellen eines Modells mithilfe der „Google teachable machine“ ist öffentlich zugänglich und wird dem Nutzer auf der Website mithilfe von Lehrvideos nähergebracht. Voraussetzung ist ein vorbereiteter Datensatz in Form von Bildern oder Tönen.

Öffnet man die Trainingsseite, präsentiert sich ein übersichtliches „User-Interface“ (UI). Dieses UI lässt sich in Drittel aufteilen. Das linke Drittel zeigt mindestens zwei Klassen. Hier soll das Modell anhand der Bilder trainiert werden. Es können beliebig viele weitere Klassen hinzugefügt werden. Im mittleren Drittel finden sich die „hyper-parameters“. Diese sind erweiterte Einstellungen zur Optimierung des Modells [33]:

1. Epoche – Darunter ist zu verstehen, wie oft ein Bild den Trainingszyklus durchlaufen hat. Eine höhere Epochenzahl kann das Ergebnis verändern.
2. Batchgröße – Die gesamten Daten werden in Gruppen - sogenannte „Batches“ - aufgeteilt. Wurde jeder Batch einmal eingelesen, ist eine Epoche abgeschlossen.
3. Lernrate – Hierbei handelt es sich um den wohl komplexesten „hyper-parameter“. Jeong definierte in einem Artikel 2020 die Lernrate oder „learning rate“ als Variable, die die Schrittgröße zum Errechnen des Funktionsverlusts bestimmt. Ist die Lernrate zu hoch, erfolgt kein Lernprozess innerhalb des Modells. Andersrum würde bei einer zu niedrigen

> Lernrate das Modell innerhalb eines abfallenden Gradienten verbleiben – einem Lernalgorithmus innerhalb des deep learning – und die Aussagekraft des Modells nimmt damit ab [36].

Die Einstellung der genannten „hyper-parameter“ sollte sorgfältig erfolgen, da hierdurch das Ergebnis bzw. die Aussagekraft des Modells grundlegend bestimmt wird.
Im rechten Drittel des UIs findet man abschließend die Vorschau, die ein zu validierendes Bild den vortrainierten Klassen zuordnet. Die Zuordnung erfolgt in prozentualer Angabe zur Klasse.

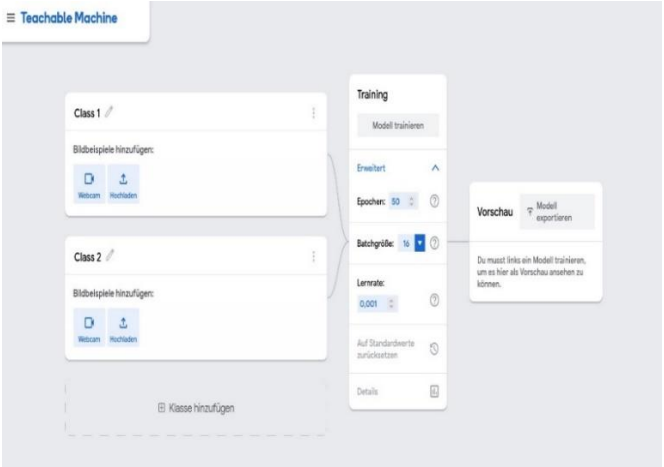

Abbildung 1: Google teachable machine: User interface der teachable machine. Hierin sieht man die erwähnten drei Abschnitte: Links lassen sich die Klassen trainieren, in der Mitte finden sich die hyper-parameters und rechts findet sich eine Vorschau der Validierung. https://teachablemachine.withgoogle.com/train/image - Stand Oktober 2021

Die Dauer des Modelltrainings hängt vor allem von der Einstellung der Hyper-parameter und des einzulesenden Datensatzes ab. So kann dies von wenigen Sekunden oder Minuten bis zu mehreren Stunden variieren. Wurde das

Modell trainiert, wird eine dauerhafte URL erstellt, die einen durchgehenden Zugriff darauf erlaubt. Zudem lässt sich das Modell exportieren.

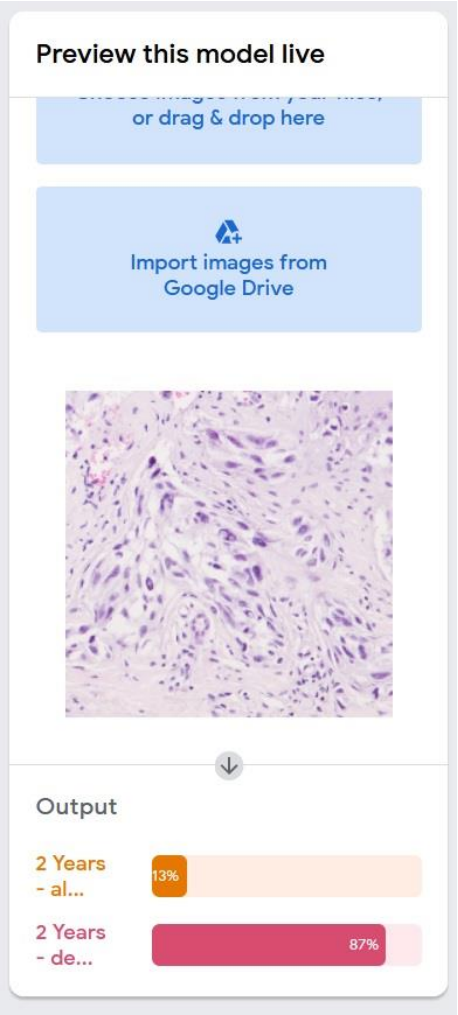

Abbildung 2: Google teachable machine: Bildvalidierung mit einer Vorschau des Bildes. Darunter sieht man die

Einteilung des Bildes und die prozentuale Zuordnung in diese Klasse. In diesem Beispiel wurde der Schnitt zu 87% "dead" zugeordnet.

Modelllink siehe Anlage - Stand Juli 2022

1.6 Zielsetzung - Melanomprognose mithilfe von deep neural networks

Ziel dieses Projekts war es, mithilfe der Google teachable machine ein Modell zu erstellen, das eine signifikante Prognose liefert bzw. in Kombination mit herkömmlichen Prognosefaktoren eine bessere Vorhersage treffen kann.

Werden die Ergebnisse validiert, kann die KI-Prognose mit dem tatsächlichen Verlauf verglichen werden. Es sollten aber auch

herkömmliche, histologische Prognosefaktoren erhoben werden. Einerseits hätte man einen direkten Vergleich mit der KI-Prognose und andererseits könnte man diese kombinieren, um eventuell eine noch zuverlässigere Prognose zu treffen, das heißt eine umfangreiche statistische Auswertung ist nötig.

2 Material und Methoden

Die gesammelten Daten wurden jederzeit datenschutzgerecht gespeichert und gelagert. Die Eintragung in das Krebsregister erfolgte mit dem Einverständnis der Patienten. Die retrospektive Studie hatte die Erlaubnis der Ethik-Kommission und lief unter der Projektnummer 874/2019BO2.

2.1 Übersicht

Es wurden feingewebliche Melanomschnitte von 836 Patienten mit der Erstdiagnose „Malignes Melanom“ zwischen dem 01.01.2012 und dem 31.12.2015 abfotografiert. 502 (ca. 60%) wurden für das Training der CNN verwendet und 334 (ca. 40%) wurden zur Validierungskohorte herangezogen. Von jedem Patienten wurden manuell semiquantitativ sechs morphologische Parameter bestimmt. Primärdaten und Verläufe (siehe Punkt 2.3)

wurden aus dem Register der deutschen dermatologischen Gesellschaft entnommen [9].

2.2 Erhobene Daten und initiale Überlegungen

Initiale Überlegungen, wie die Erstellung des Prognosemodells erfolgen sollte, und welche Arbeitsschritte hierfür nötig waren, werden folgend aufgelistet:

- Digitale Fotobibliothek von etwa 1000 Melanomschnitten.
- Erstellen zweier Google teachable machine Modelle: Das erste, das in Bezug auf einen größeren Ausschnitt des histologischen Schnittes eine Prognose abgibt, und ein zweites das bezüglich der „area of interest“ eine Prognose abgibt.
- Analyse und Vergleich der kategorialen wie auch prozentualen Zuordnung der

CNN.

- Erhebung zusätzlicher morphologischer Parameter und Erarbeiten, welche Faktoren die Prognoseaussage des Modells beeinflussen.
- Erhebung und Betrachtung herkömmlicher, histologischer Prognosefaktoren und Gegenüberstellung bzw. Kombinieren zur Prognose des Modells.

Da wir nur mit einer großen Patientenkohorte ein aussagekräftigeres Ergebnis erzielen werden, war es bereits von Anfang des Projekts Ziel, die Schnitte von 1000 Melanompatienten zu digitalisieren. Die teachable machine sollte dann mit 60% der Bilder trainiert werden und 40% validieren.

Es sollte also festgehalten werden, welche

Daten erhoben bzw. gesammelt wurden:

Tabelle 1: Überblick über die erhobenen Daten

Digitale Bilder zum Training bzw. zur Validierung	836 Melanomschnitte
Morphologie - Histologische Parameter	Cytomorphologie, Zellatypie, Mitosefiguren, Wachstumsmuster, Entzündungsinfiltrat, Pigmentierung
Prognoseabgabe der Google teachable machine in	2x 334 validierte Melanomschnitte innerhalb der zwei

Form einer Zuordnung zu einer Klasse und einer Prozentangabe, inwieweit das Melanom dieser Klasse zugeordnet wird	verschiedenen Modelle
Zentralregister-Daten (nur zur Auswertung verwendete Daten)	Tumordicke nach Breslow in mm, pT AJCC 2009, Invasionslevel nach Clark, Ulzeration, Subtyp, 2YS, 5YS, OS, Tage seit ED, Alter,

	Geschlecht, Jahr der ED

2.3 Erfassung im Melanomregister

Das Trainieren der Google teachable machine erforderte in diesem Projekt das Einlesen digitaler Bilder. Da in der Hautklinik der Universität Tübingen die histologischen Schnitte nicht in digitaler Form gespeichert wurden, war es nötig die vorhandenen histologischen Schnitte der malignen Melanome aller Patienten im Zeitraum 2012-2015 zu digitalisieren. Erstmals musste jedoch jeder Patient mit malignem Melanom, der in diesem Zeitraum an der Hautklinik der Universität Tübingen in Behandlung war, erfasst werden. Hierzu erfassten wir die Namen der Patienten, das Geburtsdatum und die Registernummer im Melanomregister der

deutschen dermatologischen Gesellschaft. Jede Melanomdiagnose wurde von zwei Dermatopathologen gestellt. Weiterhin wurde hierin das follow-up der Patienten festgehalten. Die Liste wurde zu jedem Zeitpunkt datenschutzgerecht im Laborgebäude der Hautklinik gelagert.

Anhand der Liste des Melanomregisters waren die Namen der Melanompatienten im vorgegebenen Zeitraum bekannt. Die Archivierung der histologischen Schnitte der benötigten Melanome unserer Patienten erfolgte an der Uniklinik jedoch nicht namentlich, sondern anhand von Histologie-Nummern. Das heißt, es musste in einem weiteren Schritt die Histologie-Nr. innerhalb der Befundberichte der HE-Schnitte herausgesucht werden.

Weiterhin wurden aus dem Register diverse Tumorparameter und Patienteninformationen

entnommen. Die Tumorparameter wurden nochmals im Rahmen unseres Projekts von einem zweiten Dermatopathologen überprüft:

Tabelle 2: Verwendete Daten aus dem Krebsregister der deutschen dermatologischen Gesellschaft

Alter	In Jahren
Geschlecht	Männlich oder weiblich
Jahr der ED	• 2012 • 2013 • 2014 • 2015
Tumordicke nach Breslow	In mm
pT-Einteilung nach AJCC 2009	• 1 bis 1,0 mm • 2 oder 1,01 – 2,0 mm

	• 3 oder 2,01 - 4,0 mm • 4 oder >4 mm
Ulzeration	• Ja • Nein • Unbekannt
Eindringtiefe nach Clark	• I • II • III • IV • V • Unbekannt
Nävusassoziation	• Ja • Nein • Unbekannt
Histologischer Subtyp (nur ausgewertete aufgelistet)	• SSM • NM • LMM • ALM

	• k.A. oder sonstiges

2.4 Abfrage der Histologie-Nr. und Berichte in dem Programm „Histo-DB“

Die Histologie-Nr. und die Berichte konnten im Programm „Histo-DB“ gesucht werden, das die Universitätshautklinik nutzt, um die Berichte der befundeten histologischen Schnitte namentlich zu speichern.
Das Melanomregister hat 2663 Patienten ausgeben, die von Anfang 2012 bis Ende 2015 an der Hautklinik der Universität Tübingen die Diagnose „Malignes Melanom“ erhalten haben. Davon wurden 2223 Patienten (2.10) bei der Histo-DB aufgesucht und deren Bericht und Histologie-Nr. abgespeichert.

Zugriff auf das Programm erfolgte im Laborgebäude der Hautklinik. Das Programm konnte mit einem beantragten Zugang gestartet werden und der Patient anhand seiner Daten gesucht werden. Das Programm öffnete eine Liste aller befundeten, angefertigten histologischen Feingewebsschnitte. Der Bericht des Melanomschnitts musste manuell herausgesucht werden.

War der richtige Bericht gefunden, konnte auch die Histologie-Nr. erfasst werden. Weiterhin enthält der Bericht die Lokalisation, die Entnahmetechnik, die klinische Diagnose, den Befund und die Beurteilung. Gewünschte Parameter, die im Rahmen dieses Projektes später aus dem Melanomregister erfasst wurden, wie die Tumordicke nach Breslow oder das Vorhandensein einer Ulzeration, wurden sehr unregelmäßig in den Berichten genannt.

Die Histologie-Nr. wurde in einer Excel-Tabelle einer laufenden Arbeitsnummer zugeordnet, in der aus Datenschutzgründen die weiteren Patientendaten wie Name und Geburtsdatum ausgelassen wurden. Die Arbeitsnummer konnte anhand der ursprünglichen Registernummer jederzeit im Laborgebäude auf den richtigen Patienten rückschließen lassen. Zusätzlich konnten in der Tabelle auch Besonderheiten und fehlende Berichte eingetragen werden und die genaue Zahl der gefundenen und nicht gefunden Berichte anhand von Excel-Formeln quantifiziert werden.

Zu jedem Zeitpunkt wurde bei der Übertragung von Nummern, Namen etc. eine mindestens doppelte Überprüfung durchgeführt, um Fehler zu vermindern.

2.5 Heraussuchen der Melanomschnitte im Archiv

Der Zutritt in das Archiv musste beantragt werden, der Zugangsschlüssel wurde an der Pforte der Hautklinik gelagert. In der Hautklinik der Universität gab es mehrere histologische Archive; das für dieses Projekt relevante Archiv war im Untergeschoß des Hauptgebäudes. Dieses enthielt einerseits alle fertigen, feingeweblichen HE-Schnitte im Zeitraum 2012-2016 und zusätzlich auch die größeren rohen Gewebsentnahmen, die in Paraffinblöcken aufgearbeitet wurden.

Innerhalb des Archivs waren die Objektträger nach Jahr und Histologie-Nr. gelagert. Die Histologie-Nr. enthielt das Jahr im Beispielformat „1234/ (20)12“. Die Gewebeprobe eines Patienten war immer zu einem feingeweblichen Schnitt aufgearbeitet und mit Hämatoxylin-Eosin (HE) gefärbt

worden. Der größte Teil der Patienten, die auffindbar waren, hatten zwei HE-gefärbte Schnitte.

Um die Objektträger nicht mit Fingerabdrücken oder ähnlichem zu verdrecken, wurden zu jedem Zeitpunkt beim Umgang mit den Objektträgern Latex-Handschuhe getragen. Die Objektträger wurden herausgesucht und in eine entsprechende Box, in der Platz für 100 Schnitte war, gelegt. Die Anzahl der entnommenen Objektträger wurde auf der ausgedruckten Zentralregisterliste aufgeschrieben, die auch die abgefragte Histologie-Nr. enthielten. Anschließend wurde sie zusätzlich in einer Excel-Tabelle vermerkt.

Eine komplettierte Box wurde an den Dermatopathologen übergeben, der parallel die Feingewebsschnitte fotografierte und die histologischen Parameter erfasste.

2.6 Herstellung und Färbung feingeweblicher Melanomschnitte

Da die Dateneingabe innerhalb der tiefen neuronalen Netze anhand digitalisierter, feingeweblicher Melanomschnitte erfolgte, sollte zur Anfertigung dieser Melanomschnitte im Folgenden ein Überblick gegeben werden.

Die Probengewinnung erfolgte in den meisten Fällen im Rahmen von diagnostischen Biopsien oder auch operativen Komplettexzisionen. Das gewonnene Material wird in Formalin fixiert und in einem weiteren Schritt in Paraffin eingebettet. Anschließend wird mit einem Mikrotom vom Paraffinblock ein dünner Schnitt erstellt. [37]. Der Gewebsschnitt wird nun auf einen Objektträger aufgezogen und mit Hämatoxylin-Eosin gefärbt. Wichtig ist es hierbei, bei jedem Schritt auf die Sauberkeit zu achten, da sonst Verschmutzungen das

Ergebnis beeinträchtigen [38]. Die HE-Färbung hat als großen Vorteil möglichst viele zelluläre Strukturen hervorzubringen, weshalb sie sich u.a. als Standardfärbung etabliert hat. Nukleinsäurehaltige Strukturen weisen in der HE-Färbung eine blaue Farbe auf. Dazu gehört u.a. der Nucleus und Ribosomen. In einem roten Farbton erscheinen die Zellmembran, Mitochondrien und Lysosomen [39]. Melanozyten hingegen erscheinen als kleine Zellen mit dunklen Zellkernen und oftmals umgebenden hellem Hof [38].

2.7 Fotografieren der histologischen Schnitte

Das Fotografieren der Melanomschnitte und Eintragen der Parameter erfolgte durch einen Dermatopathologen. Dabei wurde darauf geachtet, dass es sich um eindeutige Melanomschnitte handelte. Histologische Schnitte, die nicht eindeutig waren, wurden aussortiert. Aus den 1060 herausgesuchten Schnitten wurden mit dieser Begründung zehn heraussortiert; es wurden also 1050 Melanomschnitte abfotografiert. Parallel erfolgte die semiquantitative Bestimmung morphologischer Besonderheiten:

Tabelle 3: Erfasste semiquantitative, morphologische Besonderheiten

Zytomorphologie	■ Epitheloidzellig ■ Spindelzellig

Zell- und Kernatypie	■ Fehlend bis mild ■ Mäßig (+) ■ Ausgeprägt (++) ■ Massiv/ Anaplastisch (+++)
Mitosefiguren	■ Fehlend ■ Einzelne (+) ■ Viele (++)
Entzündungsinfiltrat	■ Fehlend ■ Leicht (+) ■ Mäßig (++) ■ Dicht (+++)
Dermales Wachstumsmuster	■ Kleine Nester ■ Große Nester ■ Zellrasen ■ Einzelzellen ■ K.A.

Pigmentierung	■ Fehlend ■ Leicht (+) ■ Stark (++)

Anschließend wurde ein Übersichtsbild ebenfalls von einem Dermatopathologen digital erstellt und im JPG-Format in einem Rechner im Laborgebäude gespeichert. Das Übersichtsbild wurde an der Stelle der höchsten Tumordicke nach Breslow in 100-facher Vergrößerung fotografiert. Es sollte festgehalten werden, dass es sich hier nicht um whole slide images (WSI) handelt. Es wurde als Mikroskop und Kamera eine Nikon Digital Sight DS-FI2 und Nikon Eclipse 80i verwendet. Die Firmware war die Digital Sight DS U3 und die Software NIS Elements D Version 4.13.04. Die Belichtung lag bei 3 ms und das Bild wurde in der Auflösung 2560x1920 Pixel abgespeichert.

Ein weiteres Ziel dieses Projekts war es zu verstehen, inwieweit wir uns der Prognose annähern, wenn wir nur die sogenannte „area of interest“ betrachten. Darunter ist das melanombestimmende Areal innerhalb des gesamten histologischen Schnittes zu verstehen. Dieses wurde aus dem bereits vorhandenen Übersichtsbild erstellt. Ein Dermatopathologe wählte melanombestimmende, repräsentative Areale von etwa 80x80 µm. Zum Ausgleich der „dead“ und „alive“ Gruppen (siehe Punkt 2.9) wurden in der „dead“-Gruppe pro Bild jeweils sechs krankheitsbestimmende Areale erstellt [9].

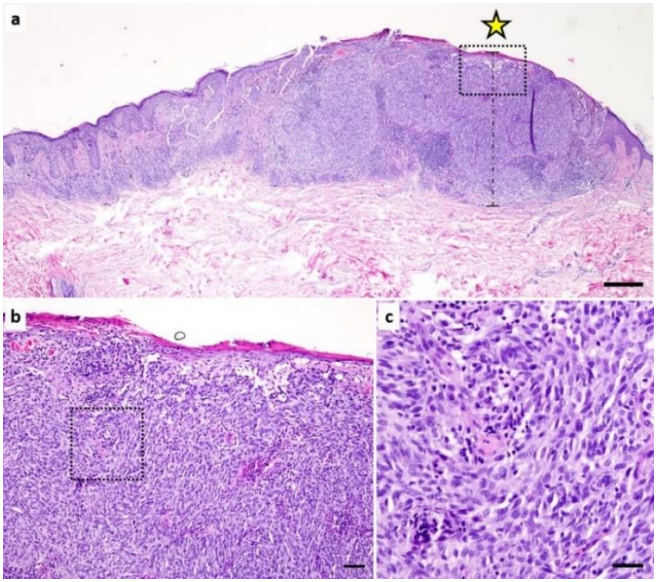

Abbildung 3: Das Bild a zeigt das Melanom in einer 25-fachen Vergrößerung. Bild b ist eine Vergrößerung aus Bild a in 100-facher Vergrößerung. Diese Bilder wurden in dieser Arbeit als „Übersichtsbild“ bezeichnet. Das Bild c ist das etwa 80x80 µm große Areal – die „area of interest“ (AOI) welche die Zytomorphologie der Melanozyten erkennen lässt. Die beiden Google teachable machine Modelle dieser

Arbeit wurden mit den Bildern b und c, also „Übersichtsbildern“ und „area of interest“-Bildern erstellt. Das hier dargestellte Melanom wurde in dieser Arbeit digitalisiert [9].

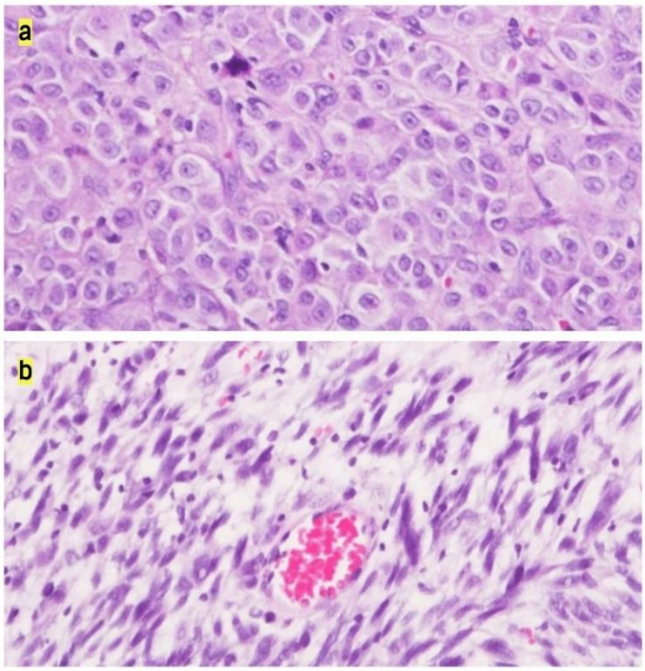

Abbildung 4: Morphologischer Parameter 1: Cytomorphologie. Bild a: Epitheloidzellig, Bild b: Spindelzellig. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

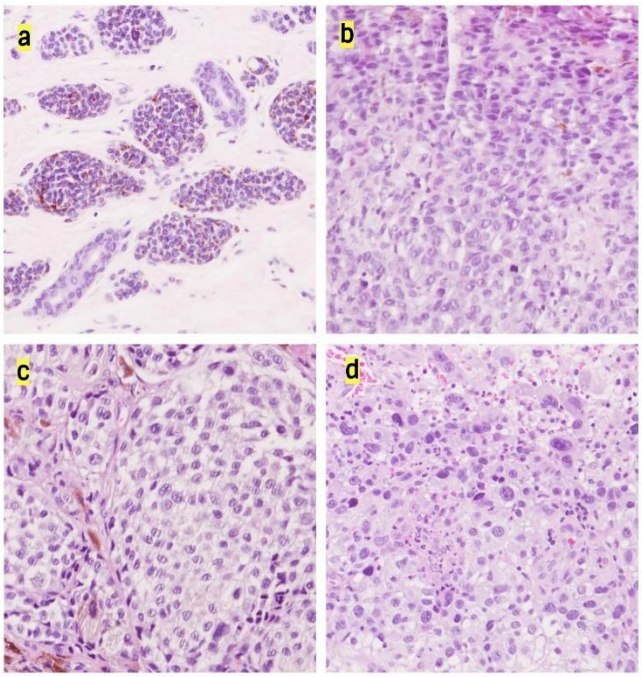

Abbildung 5: Morphologischer Parameter 2: Zellatypie. Bild a: Fehlend bis mild, Bild b: Mäßig, Bild c: Ausgeprägt, Bild d: Massiv/anaplastisch. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

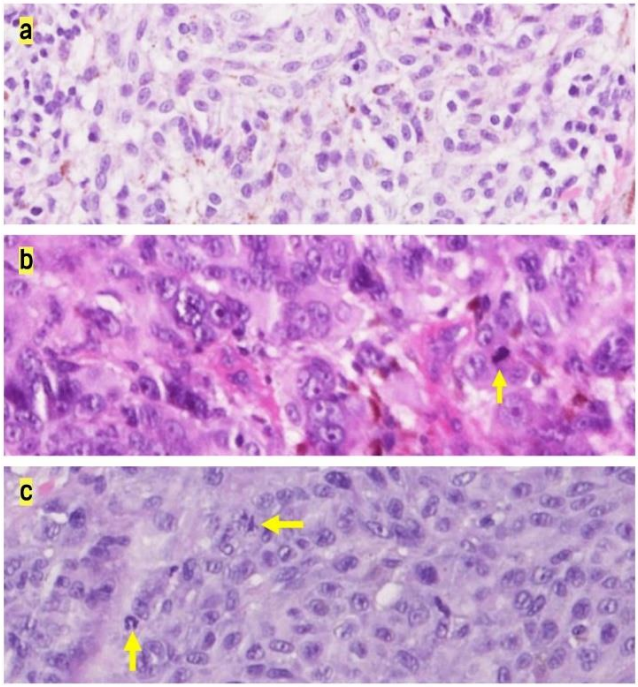

Abbildung 6: Morphologischer Parameter 3: Mitosen. Bild a: Fehlend, Bild b: Einzelne, Bild c: Viele. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

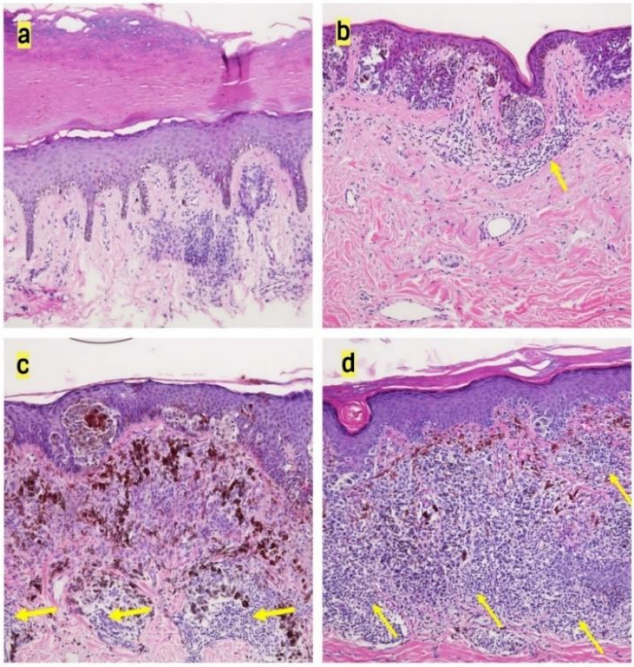

Abbildung 7: Morphologischer Parameter 4: Entzündungsinfiltrat. Bild a: Fehlend, Bild b: Leicht, Bild c: Mäßig, Bild d: Dicht. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

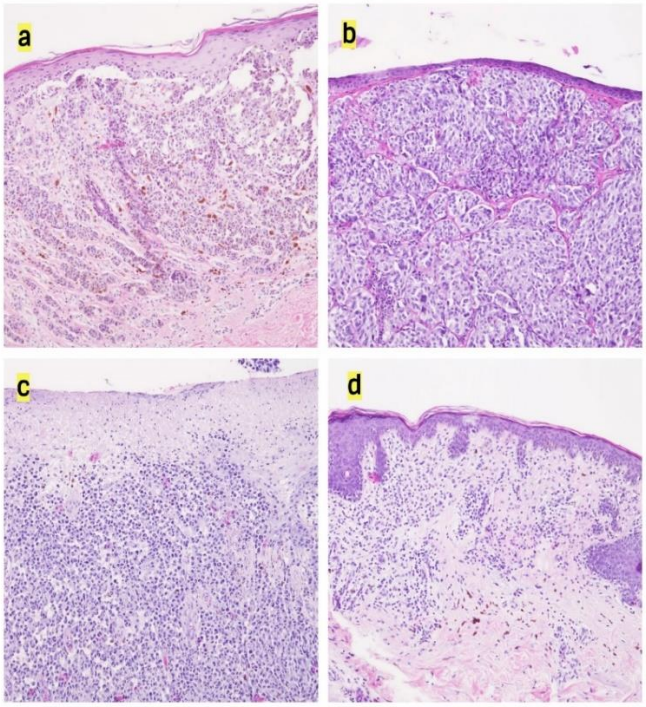

Abbildung 8: Morphologischer Parameter 5: Wachstumsmuster. Bild a: Kleine Nester, Bild b: Große Nester, Bild c: Zellrasen, Bild d: Einzelzellen. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

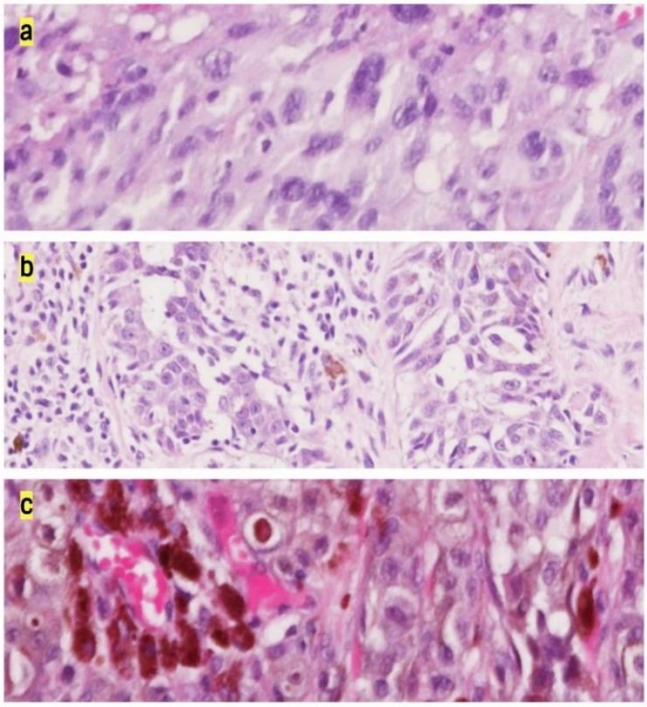

Abbildung 9: Morphologischer Parameter 6: Pigmentierung. Bild a: Fehlend, Bild b: Leicht, Bild c: Stark. Die hier dargestellten Melanome wurden in dieser Arbeit digitalisiert [9].

2.8 Follow-up

Im Rahmen der Auswertung des Modells konnten nur Patienten miteinbezogen werden, bei denen mindestens das 2-Jahres follow-up bekannt war. Auch dieses konnte aus dem Register erfasst werden. Bei Patienten, deren Nachbehandlung unbekannt war, wurde versucht, anhand der klinikinternen Arztbriefe und von öffentlichen Todesanzeigen ein follow-up zu erstellen. Von den 1050 Patienten mit nun digitalisiertem Melanomschnitt konnte mindestens das 2-Jahres-follow-up bei 836 Patienten ermittelt werden.

2.9 Erstellen des Modells

Mithilfe der digitalen Melanomschnitte konnten die „Google teachable machine“-Modelle trainiert werden. Es wurde der gesamte Zeitraum seit Erstdiagnose (frühestens 01.01.2012) bis zur Abfrage des

Todeszeitpunktes (Februar 2021) im Zentralregister betrachtet (Overall-Survival). Es sollten zwei Modelle erstellt werden: Das erste validiert die Übersichtsbilder der histologischen Schnitte, das zweite die erstellten AOI-Bilder.
In beiden Modellen wurden zwei Klassen erstellt – „dead“ und „alive“. 502 der 836 Patienten und dementsprechend deren Melanomschnitte (etwa 60%) wurden zum Training des Modells verwendet. Die Zuteilung in die Trainings- und Validierungsgruppe erfolgte dabei zufällig. Es wurden 431 für das Training der Klasse I „alive“ und 71 für das Training der Klasse II „dead“ verwendet. Betrachtet man die jeweiligen Anteile stößt man im Rahmen des „deep learning“ auf ein Problem. Die „dead“-Klasse unterliegt mit ihren 71 Schnitte deutlich der „alive“-Klasse mit 431 Schnitten. Um diesem Problem entgegenzuwirken, wurde deshalb für das

Training der „dead“-Klasse jeder Schnitt in sechsfacher Kopie eingelesen. Dadurch konnte eine Gesamtzahl von 426 für das Training der „dead“-Klasse erreicht werden. Dass ein Schnitt in sechsfacher Kopie eingelesen wurde, wurde in der weiteren Auswertung berücksichtigt.
Bei dem Modelltraining der AOI wurde dieses Problem (siehe Punkt 2.7) anders angegangen. Aus den vorherigen Übersichtsbildern wurden für die dead-Gruppe jeweils sechs unterschiedliche, repräsentative Areale herausgesucht.

Weiterhin wurden an beiden Modellen die gleichen, folgenden erweiterten Einstellungen („hyper-parameter“) vorgenommen (Erklärung siehe Punkt 1.5.2)

1. Epoche – Die Epochenanzahl wurde auf 1000 festgelegt. Mit dieser hohen

Epochenzahl konnte ein sorgfältiges Training des Modells sichergestellt werden.

2. Batchgröße – Die Batchgröße betrug 16.
3. Lernrate – Die Lernrate wurde unverändert bei 0,001 belassen. Dadurch war sie weder zu hoch noch zu niedrig, wodurch die erwähnten unerwünschten Veränderungen (siehe Punkt 1.5.2) vermieden werden konnten.

Das gesamte Training der beiden Modelle dauerte wenige Stunden. Nach Beendigung konnten die Modelle unter einer festen URL (siehe Anhang) aufgerufen werden und die Validierung war bereit. War das zu validierende Bild in der Vorschau angezeigt, wurde auf einer Excel-Tabelle und in einer Online-Datenbank auf „redcap“ die Zuordnungsklasse und zugehörige Prozentangabe notiert.

2.10 Gesammelte Patienten und Ausscheidungskriterien

Aus den identifizierten 2663 Patienten wurden 2223 in der Histo-DB aufgesucht. Von diesen Patienten war bei 631 kein Bericht auffindbar. Von weiteren 194 Patienten war kein Melanomschnitt mehr im Archiv zu finden. Zusätzliche 106 Patienten wurden nicht mehr innerhalb des Archivs aufgesucht. Aus den überbleibenden 1060 Schnitten sind 224 ausgeschieden, weil sie auf Grund schlechter Bildqualität nicht als geeignet für die KI galten (10) oder mindestens das 2-Jahres-follow-up nicht ermittelt werden konnte (214). Die nachfolgende Flowchart soll den Datenerhebungsprozess darstellen.

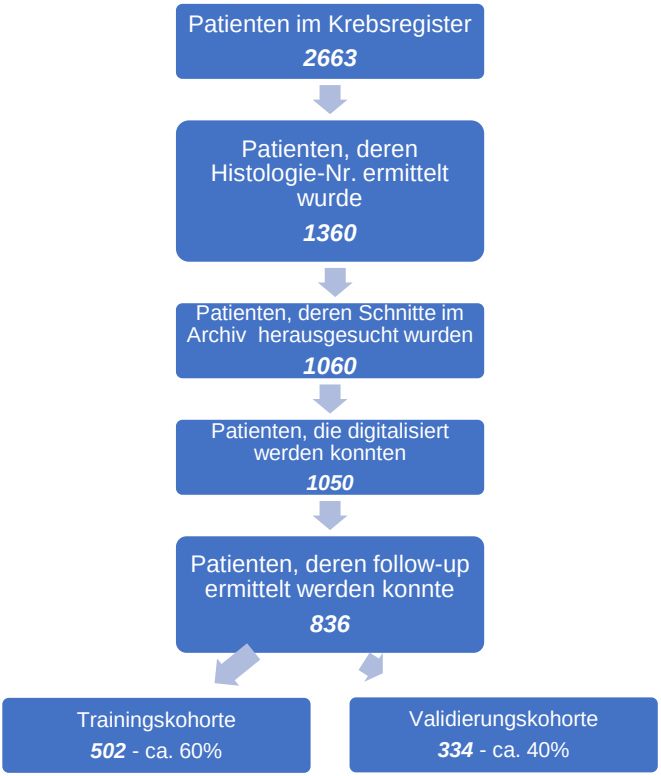

Abbildung 10: Flowchart: Prozess der Patientenauswahl und Ausscheidungskriterien

2.11 Statistische Auswertung

Im Rahmen der Auswertung wurde das Signifikanzniveau auf α=0,05 und das Konfidenzniveau auf 95% festgelegt.

Die statistische Auswertung wurde mit kommerzieller Software durchgeführt: IBM Corp. Released 2020. IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp und Microsoft Excel (Microsoft Office 365, 2021). Tabellen wurden mit Microsoft Word angefertigt. Grafiken wurden mit Microsoft Word und SPSS erstellt. Es wurden verschiedenste Analysen, meistens univariat, bezüglich Korrelation, Überleben und Vorhersage anhand der Kaplan-Meier-Schätzmethode, der logistischen Regression und ROC-Analysen durchgeführt. Testungen erfolgten u.a. anhand von Log-Rank, Omnibustests der Koeffizienten (Wald- und Chi-Quadrat-Tests) und Hosmer-

Lemeshow.

3 Ergebnisse

3.1 Gesammelte Schnitte und morphologische Parameter

Innerhalb dieser Arbeit wurden 1050 Melanomschnitte in Boxen systematisch gesammelt und abfotografiert. Zusätzlich erfolgte erstmalig eine aufwendige Bestimmung sechs verschiedener, morphologischer Parameter. Die 1050 Schnitte, wie auch erstmalig in digitaler Form sollen für zukünftige Projekte verwendet werden können. Auch mit den morphologischen Parametern kann i.R. verschiedenster Projekte gearbeitet werden. Zusätzlich sind die erstellten Google teachable machine-Modelle (siehe Anhang) weiterhin öffentlich zugänglich und zur Verwendung bereitstehend. Dies stellt bereits einen erheblichen Teil der Arbeit dar.

3.2 Deskriptive Statistik

Im Folgenden soll ein Überblick über die gesamte Kohorte gegeben werden. Die Eigenschaften der Kohorte werden in den nachfolgenden Seiten anhand von Tabellen und Diagrammen aufgearbeitet. Die wichtigsten Punkte werden als Text erwähnt. In den folgenden Kapiteln wird die Bezeichnung „Gesamtkohorte“ für die Patientengesamtheit verwendet. Die 502 Patienten, deren Schnitte für das Training der CNN-Modelle genutzt wurden, werden als „Trainingskohorte“ bezeichnet und die 334 Patienten, deren Schnitte validiert wurden, werden „Validierungskohorte“ genannt.

Nach der Betrachtung einer Eigenschaft bei der Gesamtkohorte werden die Trainings- und Validierungskohorte betrachtet.

3.2.1 Patienteneigenschaften

3.2.1.1 Geschlecht

Die 836 Patienten werden folgend anhand ihrer Geschlechterverteilung betrachtet:

463 Patienten bzw. 55,4% sind männlich und 372 Patienten bzw. 44,5% sind weiblich. Bei einem Patienten gab es keine Geschlechtsangabe im Zentralregister.

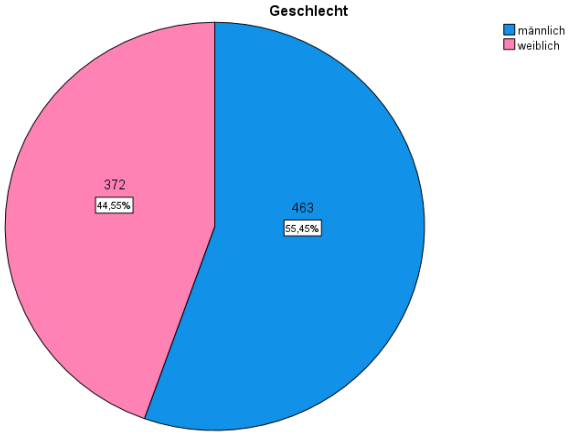

Abbildung 11: Kreisdiagramm: Gesamtkohorte – Geschlechterverteilung

Tabelle 4: Trainings- und Validierungskohorte – Geschlechterverteilung

	Trainingskohorte	**Validierungskohorte**
männlich	56,89% (n=285)	53,29% (n=178)
weiblich	43,11% (n=216)	46,71% (n=156)

Tendenziell zeigt sich eine ähnliche Verteilung wie in der Gesamtkohorte. Durch die Zufallsverteilung sind mit 56,89% (n=285) Männer noch einmal geringgradig mehr innerhalb der Trainingskohorte vertreten.

Mit männlich:weiblich 53,29%:46,71% zeigt sich bei der Validierungskohorte eine Verteilung welche näher an einer 50:50 Verteilung ist als bei der Gesamt- und

Trainingskohorte.

3.2.1.2 Alter

Das Alter des gesamten Patientenkollektivs zeigt folgende Verteilung:

Tabelle 5: Gesamtkohorte - Altersverteilung in Jahren

Mittelwert	59,89
Median	62,00
Modus	72
Std.-Abweichung	15,372
Varianz	236,306
Minimum	7
Maximum	93
Perzentile	25. – 49,00 50. – 62,00 75. – 72,00

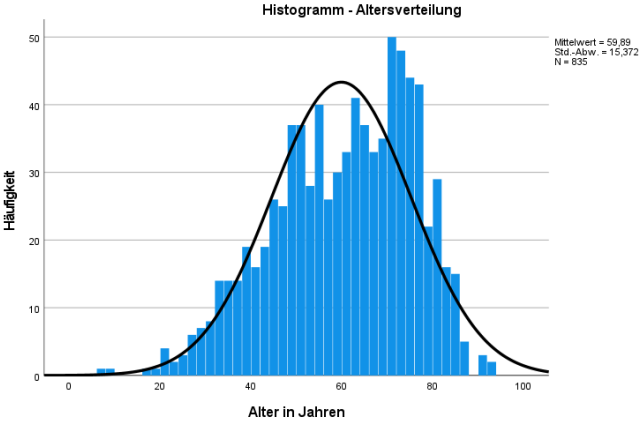

Abbildung 12: Histogramm: Gesamtkohorte - Altersverteilung in Jahren

Ein Patient hatte keine hinterlegte Altersangabe. Die meisten Patienten waren 72 Jahre alt; mehr als die Hälfte waren älter als 62 (Median). Im Durchschnitt waren die Patienten 59,89 Jahre alt. Die Altersspannweite der gesamten Patientengruppe betrug 86 Jahre.

Tabelle 6: Trainings- und Validierungskohorte - Altersverteilung in Jahren

	Trainingskohorte	**Validierungskohorte**
Mittelwert	61,09	58,08
Median	63,00	59,00
Modus	72	48
Std.-Abweichung	14,995	15,771
Varianz	224,861	248,727
Minimum	9	7
Maximum	93	91
Perzentile	25. – 50,00 50. – 63,00 75. – 73,00	25. – 48,00 50. – 59,00 75. – 71,00

Das Alter der Trainingskohorte weist einen

arithmetischen Mittelwert von 61,09 und ist 1,2 Jahre über dem der Gesamtkohorte. Das häufigste vorkommende Alter (Modus) war 72 Jahre und ist damit dasselbe wie bei der Gesamtkohorte.

Als arithmetischen Mittelwert des Alters der Validierungskohorte ergibt sich 58,08 Jahre, und liegt somit 1,81 Jahre unter dem Mittelwert der Gesamtkohorte. Der Modus zeigt sich mit 48 hier deutlich unter dem der Trainings- und Gesamtkohorte.

3.2.1.3 Erkrankungsjahr

Es wurden Patienten mit dem Jahr der Erstdiagnose (ED) 2012-2015 ausgewählt.
Ein Patient hatte in der Auswertungstabelle kein Jahr der ED angegeben. Der Großteil der ausgewerteten Melanomschnitte wurde 2012 erstdiagnostiziert. Der geringe Anteil der

Patienten mit 2015 als Jahr der ED kann darauf zurückgeführt werden, dass 2223 der 2663 ausgegebenen Patienten aufgesucht wurden. Da die Patienten innerhalb der Krebsregisterliste nach Erkrankungsjahr sortiert waren, sind die fehlenden 440 Patienten alle im Jahr 2015 erstdiagnostiziert worden. Ein übersichtliches Kreisdiagramm zeigt die Verteilung:

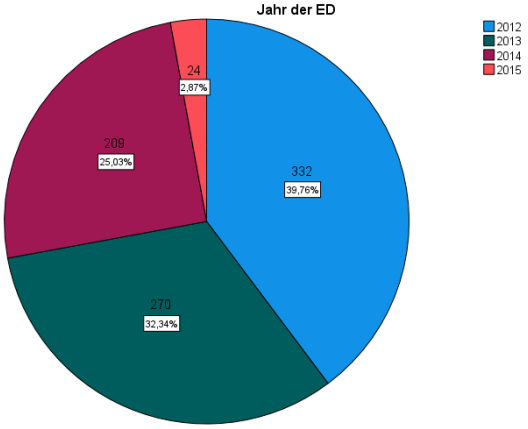

Abbildung 13: Kreisdiagramm: Gesamtkohorte - Jahr der ED

Tabelle 7: Trainings- und Validierungskohorte - Häufigkeiten: Jahr der ED

Jahr der ED	Trainingskohorte	Validierungskohorte
2012	39,2% (n=197)	40,4% (n=135)
2013	32,1% (n=161)	32,6% (n=109)
2014	25,3% (n=127)	24,6% (n=82)
2015	3,2% (n=16)	2,4% (n=8)

Insgesamt zeigt sich, dass die Trainings- und Validierungskohorte keine wesentlichen Abweichungen aufweisen. 2012 (TK 39,2% und VK 40,4%) bleibt das häufigste Jahr der ED.

3.2.1.4 Nävusassoziation

Ein möglicher Effekt einer Nävusassoziation auf die Prognoseabgabe der CNN soll analysiert werden. Die Gesamtkohorte wird anhand einer Nävusassoziation aufgeteilt:

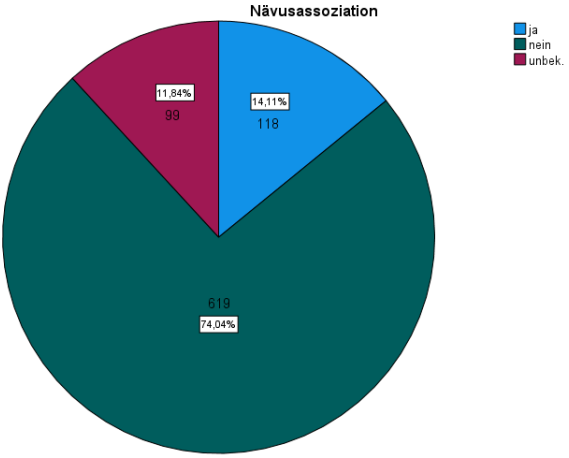

Abbildung 14: Kreisdiagramm: Gesamtkohorte - Häufigkeiten: Nävusassoziation

14,1% (n=118) der Melanome wiesen eine Nävusassoziation auf, 74,0% (n=619) wiesen keine Assoziation auf. Bei 11,8% (n=99) war sie unbekannt.

Tabelle 8: Trainings- und Validierungskohorte - Häufigkeiten: Nävusassoziation

Nävusassoziation	Trainingskohorte	Validierungskohorte
ja	13,15% (n=66)	15,57% (n=52)
nein	74,7% (n=375)	73,05% (n=244)
unbekannt	12,15% (n=61)	11,38% (n=38)

Gegenüber der Patientengesamtheit zeigen sich in beiden Gruppen keine wesentlichen Unterschiede.

3.2.2 Morphologische Parameter

Die semiquantitativ bestimmten morphologischen Parameter, die erstmalig im Rahmen dieses Projekts erhoben wurden, werden Punkt für Punkt in ihren Häufigkeiten

betrachtet:

3.2.2.1 Cytomorphologie

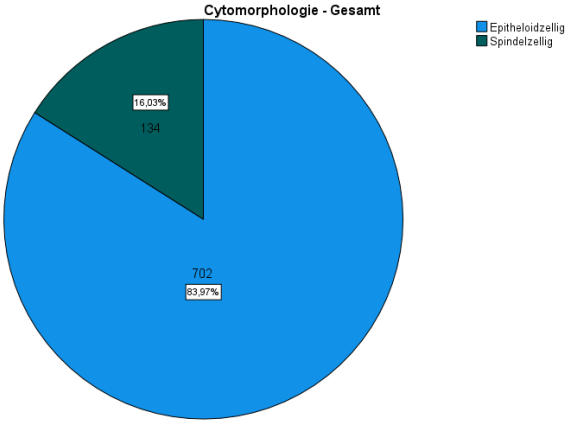

Abbildung 15: Kreisdiagramm: Gesamtkohorte – Häufigkeiten: Cytomorphologie

Den absolut größten Teil bildeten cytomorphologisch epitheloidzellige Melanome mit fast 83,97% (n=702).

Tabelle 9: Trainings- und Validierungskohorte - Häufigkeiten: Cytomorphologie

Cytomorphologie	**Trainingskohorte**	**Validierungskohorte**
Epitheloidzellig	83,27% (n=418)	85,03% (n=284)
Spindelzellig	16,73% (n=84)	14,97% (n=50)

3.2.2.2 Zellatypie

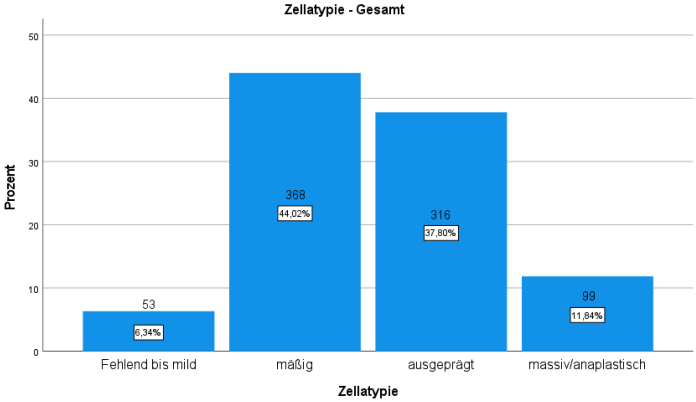

Abbildung 16: Balkendiagramm: Gesamtkohorte – Häufigkeiten: Zellatypie

Ganze 81,8% (n=686) zeigen eine mäßige bis ausgeprägte Zellatypie. Die kleinste Gruppe mit 6,34% (n=53) der insgesamt 836 Schnitte weist eine nur fehlende bis milde Zellatypie vor.

Tabelle 10: Trainings- und Validierungskohorte - Häufigkeiten: Zellatypie

Zellatypie	**Trainingskohorte**	**Validierungskohorte**
Fehlend bis mild	6,97% (n=35)	5,39% (n=18)
Mäßig	45,62% (n=229)	41,92% (n=140)
Ausgeprägt	35,86% (n=180)	40,42% (n=135)
Massiv/anaplastisch	11,55% (n=58)	12,28% (n=41)

3.2.2.3 Mitosen

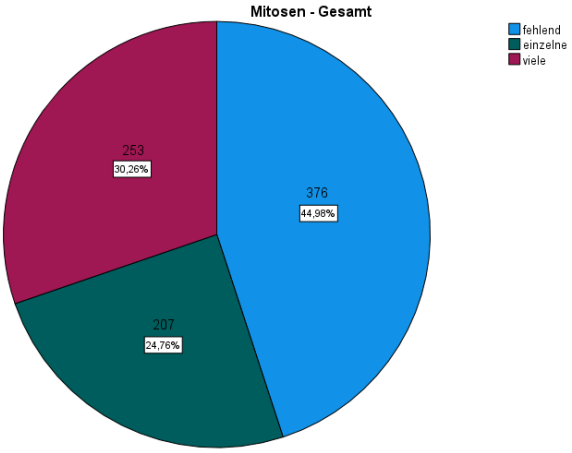

Abbildung 17: Balkendiagramm: Gesamtkohorte – Häufigkeiten: Mitosen

Bei 44,98% (n=376) der 836 Melanomschnitte, also nahezu der Hälfte, waren keine Mitosen bzw. Mitosefiguren erkennbar. Bei fast 30,3% (n=253) hingegen waren viele erkennbar. Interessanterweise sind also die zwei Extremen („fehlend“ und „viele“) stärker repräsentiert als die Zwischenkategorie, sowohl in der

Patientengesamtheit als auch in den beiden Untergruppen.

Tabelle 11: Trainings- und Validierungskohorte - Häufigkeiten: Mitosen

Mitosen	Trainingskohorte	Validierungskohorte
Fehlend	46,61% (n=234)	42,22% (n=141)
einzelne	23,51% (n=118)	26,95% (n=90)
viele	29,88% (n=150)	30,84% (n=103)

Die Validierungskohorte weist insgesamt mehr Schnitte mit Mitosen auf („einzelne“ 26,95% und viele „30,84%) gegenüber der Trainingskohorte („einzelne“ 23,51% und „viele“ 26,95%).

3.2.2.4 Entzündungsinfiltrat

Unter einem Entzündungsinfiltrat ist die

Einwanderung von u.a. Lymphozyten und Makrophagen zu verstehen. 94,6% (n=791) der Schnitte zeigten ein Entzündungsinfiltrat. Dabei wiesen 44,5% (n=372) ein leichtes auf.

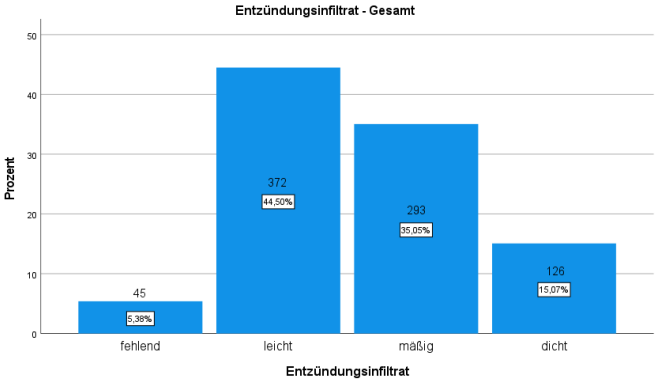

Abbildung 18: Balkendiagramm: Gesamtkohorte-Häufigkeiten: Entzündungsinfiltrat

Tabelle 12: Trainings- und Validierungskohorte - Häufigkeiten: Entzündungsinfiltrat

Entzündungsinfiltrat	Trainingskohorte	Validierungskohorte
fehlend	5,78% (n=29)	4,79% (n=16)
leicht	45,42% (n=228)	43,11% (n=144)
mäßig	34,46% (n=173)	35,93% (n=120)
dicht	14,34% (n=72)	16,17% (n=54)

3.2.2.5 Wachstumsmuster

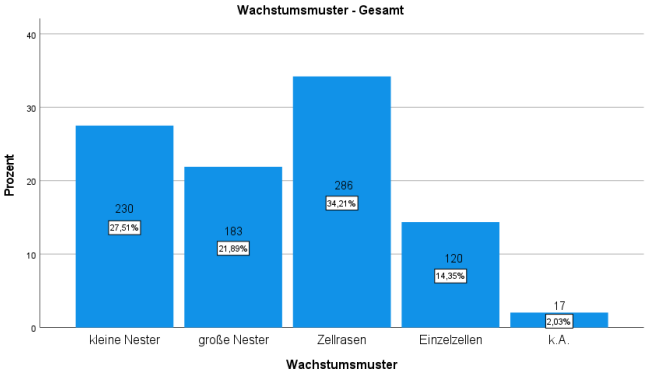

Abbildung 19: Balkendiagramm: Gesamtkohorte – Häufigkeiten: Wachstumsmuster

Die meisten Melanome wiesen in ihrem Wachstumsmuster „Zellrasen“ auf mit 34,2% (n=286). An zweiter Stelle stehen die „kleinen Nester“ (27,51%; n=230), dann „große Nester“ (21,89%; n=183) und „Einzelzellen“ (14,35%; n=120).

Tabelle 13: Trainings- und Validierungskohorte - Häufigkeiten: Wachstumsmuster

Wachstumsmuster	**Trainingskohorte**	**Validierungskohorte**
Kleine Nester	27,69% (n=139)	27,54% (n=92)
Große Nester	20,92% (n=105)	23,05% (n=77)
Zellrasen	34,06% (n=171)	34,43% (n=115)
Einzelzellen	15,14% (n=76)	13,17% (n=44)
k.A.	2,19% (n=11)	1,80% (n=6)

3.2.2.6 Pigmentierung

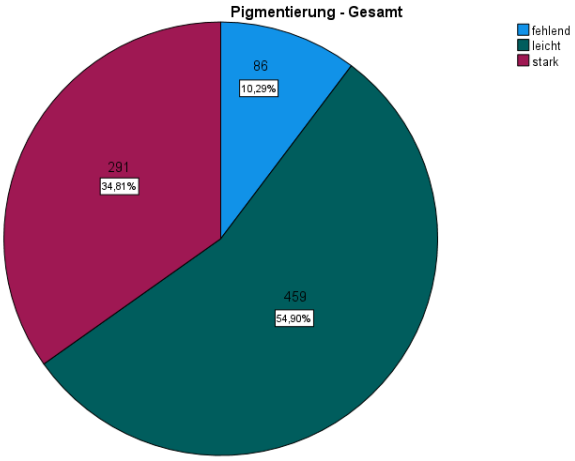

Abbildung 20: Kreisdiagramm: Gesamtkohorte – Häufigkeiten: Pigmentierung

Nur 10,29% (n=86) zeigen keine Pigmentierung. Die meisten Melanome zeigten eine leichte Pigmentierung mit 54,9% (n=459). 34,81% (n=291) zeigten eine starke Pigmentierung.

Tabelle 14: Trainings- und Validierungskohorte - Häufigkeiten: Pigmentierung

Pigmentierung	**Trainingskohorte**	**Validierungskohorte**
fehlend	10,16% (n=51)	10,78% (n=36)
leicht	52,39% (n=263)	58,68% (n=196)
stark	37,45% (n=188)	30,54% (n=102)

Die Kategorie „stark“ ist innerhalb der Trainingskohorte mit 37,45% vs. 30,54% stärker vertreten, dafür findet sich mehr von der Kategorie „leicht“ innerhalb der Validierungskohorte (TK 52,39% vs. VK 58,68%).

3.2.3 Herkömmliche histologische Parameter

3.2.3.1 Tumordicke nach Breslow

Die Tumordicke nach Breslow wird einerseits als verhältnisskalierte Einheit in mm betrachtet und andererseits die pT-Einteilung des AJCC 2009:

Tabelle 15: Gesamtkohorte - Verteilung: Tumordicke nach Breslow in mm

Mittelwert	1,9014
Median	1,0500
Modus	,30
Std.-Abweichung	2,48359
Varianz	6,168
Minimum	,15
Maximum	30,00
Perzentile	25. – ,50 50. – 1,05

	75. – 2,37

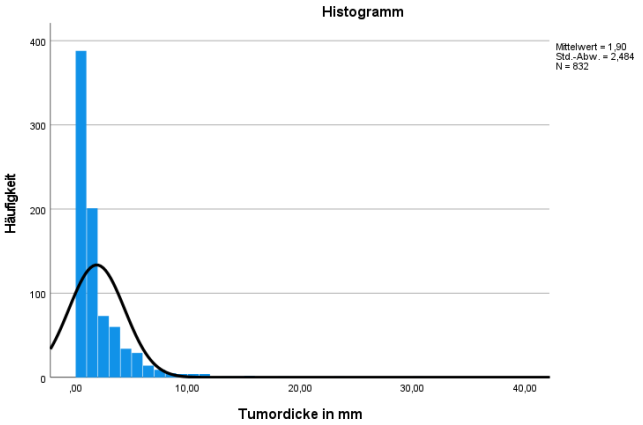

Abbildung 21: Histogramm: Gesamtkohorte – Verteilung: Tumordicke nach Breslow in mm

Bei vier Melanomschnitten war kein auswertbarer Wert hinterlegt. Der arithmetische Mittelwert aller 832 Tumordicken lag bei 1,9 mm. Die 50. Perzentile bzw. der Median liegt bei 1,05 mm. Dementsprechend liegt die Hälfte aller Tumordicken darüber oder darunter. Das bedeutet allerdings auch das durch Ausreißer

in den oberen Perzentilen die meisten Tumordicken unter dem arithmetischen Mittelwert liegen. Die am häufigsten vorgekommene Tumordicke bei den 832 ausgewerteten Melanomschnitten war 0,30 mm.

Tabelle 16: Trainings- und Validierungskohorte – Verteilung: Tumordicke nach Breslow in mm

	Trainingskohorte	**Validierungskohorte**
Mittelwert	1,8879	1,9216
Median	1,0000	1,1000
Modus	,30	1,20
Std.-Abweichung	2,61388	2,27812
Varianz	6,832	5,190

Minimum	,15	,15
Maximum	30,00	18,00
Perzentile	25. – ,45 50. – 1,00 75. – 2,20	25. – ,55 50. – 1,10 75. – 2,50

Innerhalb der Trainingskohorte waren drei Fälle mit nichtangegebener Tumordicke. Der Mittelwert ist bei 1,89 mm. Bei der Gesamtkohorte lag er bei 1,90 mm. Die häufigste angegebene Tumordicke (Modus) liegt auch bei der Trainingskohorte bei 0,30 mm.

Die Validierungskohorte hatte einen Fall ohne Angabe. Der arithmetische Mittelwert mit 1,92 mm unterscheidet sich nicht wesentlich von dem der Gesamt- und Trainingskohorte (GK 1,9 und TK 1,89). Die häufigste vorkommende Tumordicke (Modus) mit 1,2 mm unterscheidet

sich jedoch doch noch einmal hier deutlich von der der Gesamt- und Validierungskohorte (0,30 mm).

Die pT-Einteilung des AJCC 2009 zeigt folgende Verteilung:
832 gültig, 4 fehlend. Es zeigt sich, dass mit höherer Klasse die Häufigkeit der Schnitte, die dazu zugeordnet wird, abnimmt. Knapp die Hälfte aller Schnitte mit 49,28% werden der ersten Klasse (bis 1,0 mm) zugeordnet. Etwa ein Viertel mit 22,96% ist der Klasse II (1,01 mm - 2,0 mm) zugehörig. Das letzte Viertel des Gesamtanteils aller Melanomschnitte wird unter der Klasse III und Klasse IV aufgeteilt. Dabei hat Klasse IV mit 12,26% weniger zugehörige Schnitte als Klasse III mit 15,5%.

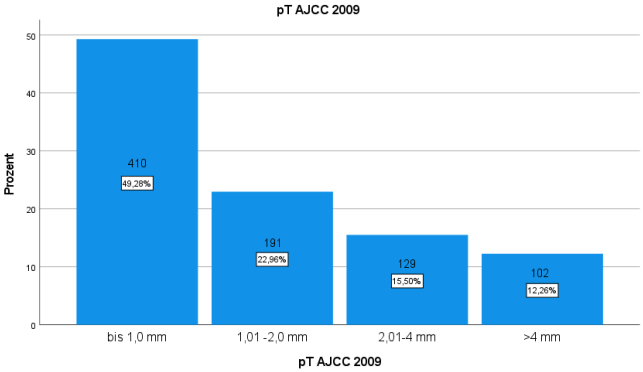

Abbildung 22: Balkendiagramm: Gesamtkohorte – Häufigkeiten: pT-Einteilung des AJCC 2009

Tabelle 17: Trainings- und Validierungskohorte - Häufigkeiten: pT-Einteilung des AJCC 2009

pT-Einteilung des AJCC 2009	**Trainingskohorte (3 fehlend)**	**Validierungskohorte (1 fehlend)**
Bis 1,0 mm	50,10% (n=250)	48,05% (n=160)
1,01-2,0	22,65%	23,42% (n=78)

mm	(n=113)	
2,01-4,0 mm	15,43% (n=77)	15,62% (n=52)
>4,0 mm	11,82% (n=59)	12,91% (n=43)

Innerhalb der beiden Gruppe zeigt sich genau wie bei der Gesamtkohorte, dass mit zunehmender Klasse die Anzahl der zugehörigen Schnitte abnimmt. Klasse I ist auch hier bei beiden die häufigste vorkommende Klasse mit ca. der Hälfte aller Schnitte (TK 50,10% und VK 48,05%).

3.2.3.2 Ulzeration

Es wird das Vorhandensein einer Ulzeration an der Patientengesamtheit betrachtet. Bei 21,3% (n=178) der Patienten war eine Ulzeration vorhanden. Der Großteil mit 77,3% (n=646) hatte keine; bei 1,4% (n=12) war unbekannt, ob eine Ulzeration vorlag.

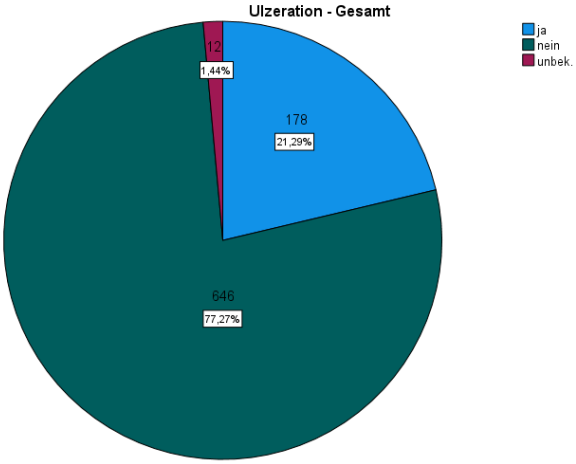

Abbildung 23: Kreisdiagramm: Gesamtkohorte – Häufigkeiten: Vorhandensein einer Ulzeration

Tabelle 18: Trainings- und Validierungskohorte: Häufigkeiten: Vorhandensein einer Ulzeration

Ulzeration	**Trainingskohorte**	**Validierungskohorte**
ja	20,52% (n=103)	22,46% (n=75)
nein	77,69% (n=390)	76,65% (n=256)
Unbekannt	1,79% (n=9)	0,90% (n=3)

3.2.3.3 Invasionslevel nach Clark

Eine Zuordnung aller Melanomschnitte zu den entsprechenden Clark-Leveln war erfolgt:

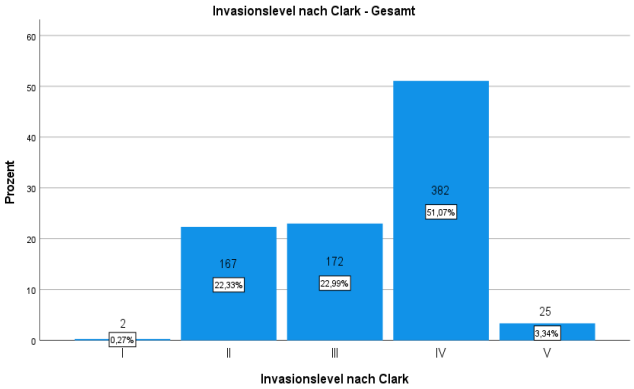

Abbildung 24: Balkendiagramm Gesamtkohorte – Häufigkeiten: Invasionslevel nach Clark

88 Melanome hatten kein hinterlegtes Clark-Level. Nur zwei Melanome fallen in das erste Clark-Level und sind somit Melanome-in-situ. Diese Einteilung in das Level I im Melanomregister ist jedoch eindeutig als fehlerhaft festzuhalten. Bei der Erhebung der morphologischen Parameter, worunter auch diese beiden Schnitte fallen, wurde von einem

Dermatopathologen darauf geachtet, dass nur invasive Melanome in die Auswertung miteinbezogen werden. Die beiden Fälle wurden bei der späteren Regressionsanalyse und Überlebensanalyse des Invasionslevels nach Clark ausgeschlossen.

3,34% (n=25) sind der Klasse V zuzuordnen und haben die Subkutis infiltriert. Der Großteil aller Melanomschnitte mit 51,07% (n=382) und damit fast die Hälfte ist dem vierten Clark-Level zuzuordnen, d.h. sie haben das Stratum reticulare infiltriert. Eine Betrachtung der Häufigkeiten zeigt, dass sich dies nicht mit den Zuordnungen bei der Tumordicke deckt, bei denen mit höherer Klasse die Häufigkeit abgenommen hatte. Beim Invasionslevel ist das Level IV mit 382 (51,07%) der 836 Melanomschnitte die häufigste vertretene Klasse.

Tabelle 19: Trainings- und Validierungskohorte – Häufigkeiten: Invasionslevel nach Clark

Invasionslevel nach Clark	Trainingskohorte (53 fehlend)	Validierungskohorte (35 fehlend)
I	0,22% (n=1)	0,33% (n=1)
II	24,01% (n=108)	19,73% (n=59)
III	22,94% (n=103)	23,08% (n=69)
IV	48,78% (n=219)	54,52% (n=163)
V	4,01% (n=18)	2,34% (n=7)

Invasionslevel IV nach Clark ist mit 219 und damit 48,78% das häufigste vorkommende Clark-Level der Trainingskohorte. Klasse II ist im Vergleich zur Gesamtkohorte häufiger als

Klasse III vertreten gewesen, allerdings zeigen sich beide Klassen in beiden Kohorten sehr nah beieinander.

In der Validierungskohorte ist das Invasionslevel IV nach Clark mit 54,42% (n=163) und damit annähernd der Hälfte auch das häufigste Invasionslevel.

3.2.3.4 Subtypen

Eine Zuordnung anhand der vier häufigsten, histologischen Subtypen (SSM, NM, LMM, ALM) ist erfolgt und im Zentralregister angegeben worden. Melanome, die nicht zu den vier Subtypen zuzuordnen waren, wurden gemeinsam mit fehlenden Zuordnungen unter einer Variable zusammengefasst.

Den absolut größten Anteil mit 59,42% machte das superfiziell-spreitende Melanom aus. An zweiter Stelle steht das noduläre Melanom mit

16,09%. 9,24% der Melanome waren einem anderen der vier Typen zugehörig oder es war keine Angabe hinterlegt.

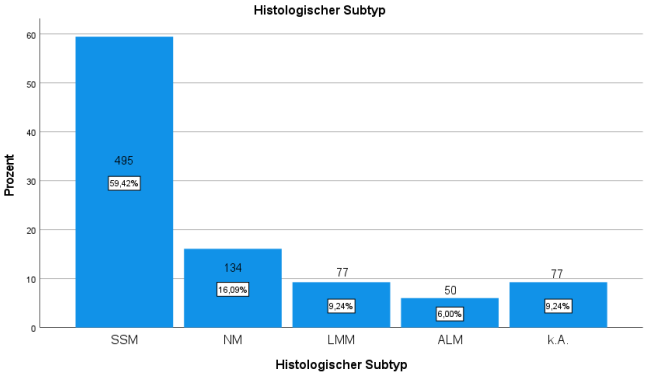

Abbildung 25: Balkendiagramm: Gesamtkohorte - Häufigkeit: Histologische Subtypen

Tabelle 20: Trainings- und Validierungskohorte - Häufigkeit: Histologische Subtypen

Hist. Subtyp	Trainingskohorte	Validierungskohorte
SSM	60,4% (n=303)	57,5% (n=192)
NM	14,9% (n=75)	17,7% (n=59)
LMM	10,6% (n=53)	7,2% (n=24)
ALM	5,4% (n=27)	6,9% (n=23)
k.A. oder sonstiges	8,8% (n=44)	10,8% (n=36)

Das SSM mit etwa 60,4% (n=303) ist unverändert gegenüber der Gesamtkohorte der häufigste Subtyp innerhalb der Trainingskohorte.

Mit 57,49% (n=192) ist das SSM auch der häufigste Subtyp der histologischen Schnitte

der Validierungskohorte, jedoch 1,71% unter dem der Gesamtkohorte. Mit 17,7% ist der Anteil der nodulären Melanome innerhalb der Validierungskohorte größer als in der Gesamtkohorte (16%) und in der Trainingskohorte (14,9%).

3.2.4 Zusammenfassung

Es findet sich ein annähernd gleicher Anteil an männlichen und weiblichen Patienten. Das Durchschnittsalter der Patienten beträgt 68 Jahre und das häufigste Jahr der ED bei den gesammelten Schnitten war 2012. Fast 60% waren dem histologischen Subtyp des SSM zuzuordnen. Lediglich ca. 14% aller Melanomschnitte waren nävusassoziiert. Etwa die Hälfte aller Melanome waren pT1 nach der Einteilung des AJCC 2009. Die durchschnittliche Tumordicke liegt bei 1,9 mm. Das Clark-Invasionslevel IV war mit knapp 46%

das häufigste vertretene Level. Eine Ulzeration fand sich bei etwa 21% aller Melanome.

Die Trainings- und Validierungskohorten zeigen in ihren Eigenschaften keine wesentlichen Ausreißer, die im Rahmen der zufälligen Einteilung hätten auftreten können.

3.3 Überlebensanalyse herkömmlicher und morphologischer Faktoren

Bei den folgenden Kaplan-Meier-Analysen wurde die zeitliche Achse als beobachteter Zeitraum in Tagen gewählt. Bei Patienten, die verstorben sind, endete der Beobachtungszeitraum am Tage des Todes. Dabei wurde als „event“ das Versterben im overall-survival definiert. Die Signifikanz wird mithilfe des Log-Rank-Tests ermittelt. Hierbei werden die einzelnen Kategorien in ihrer Äquivalenz getestet mit der durchgehend selben zeitlichen Gewichtung. Hervorgehoben wird innerhalb der Texte oftmals das Überleben nach 2000 Tagen, da man hier nach ca. 5,5 Jahren einen schlüssigen, begrenzten Zeitraum hat.

3.3.1 Herkömmliche Prognosefaktoren

Die herkömmlichen histologischen Faktoren

(Tumordicke nach Breslow – pT-Einteilung des AJCC 2009, Invasionslevel nach Clark, Ulzeration, histologischer Subtyp) wurden anhand der Kaplan-Meier-Schätzmethode ausgewertet. Es wurden alle 836 Schnitte zur Auswertung genutzt. Einzelne fehlende Fälle wurden in der deskriptiven Statistik erwähnt:

3.3.1.1 pT AJCC 2009

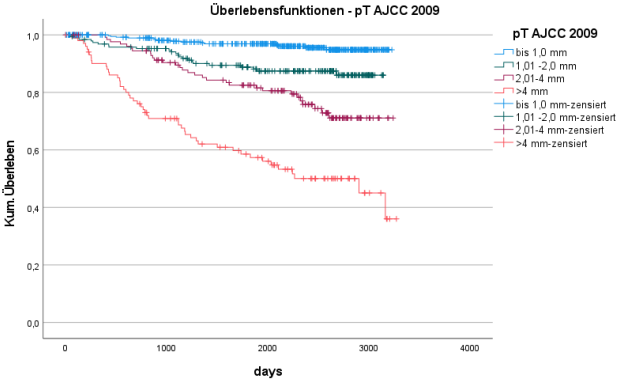

Abbildung 26: Kaplan-Meier-Kurve: pT AJCC 2009

Es wurde anhand des Log-Rank getestet und ist mit p<0,001 statistisch hochsignifikant. Die Kaplan-Meier-Kurve zeigt einen signifikanten Unterschied zwischen einzelnen pT-Klassen.

Mit zunehmender Breslow-Klasse verschlechtert sich das Gesamtüberleben deutlich.

3.3.1.2 Invasionslevel nach Clark

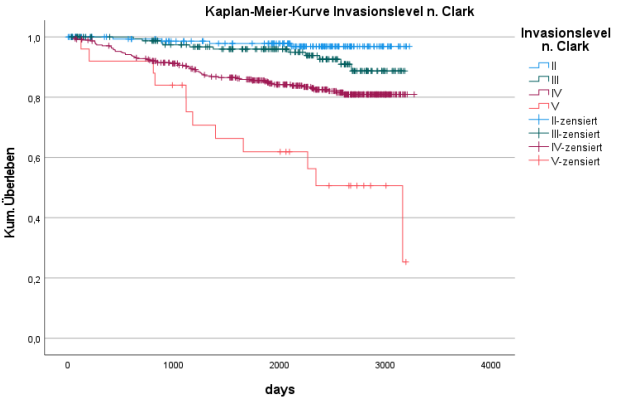

Abbildung 27: Kaplan-Meier-Kurve: Invasionslevel n. Clark

Es wurde anhand des Log-Rank getestet und ist mit p<0,001 statistisch hochsignifikant. Die Level II-IV zeigen sich relativ nah beieinander. Es ist kein so großer Unterschied im Überleben wie an der pT-Einteilung des AJCC 2009 festzustellen. Zwischen Invasionslevel IV und V

zeigte sich der größte Unterschied am zu erwartenden Überleben (IV nach 2000 Tagen ca. bei 0,84 und V nach 2000 Tagen ca. bei 0,62).

3.3.1.3 Ulzeration

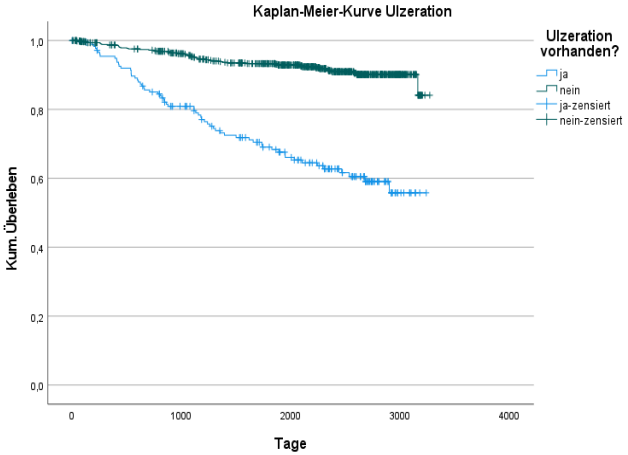

Abbildung 28: Kaplan-Meier-Kurve: Vorhandensein einer Ulzeration

Es wurde anhand des Log-Rank getestet und ist mit $p<0,001$ statistisch hochsignifikant. Der signifikante Unterschied im Überleben bei Vorhandensein oder Nicht-Vorhandensein

einer Ulzeration wurde an der obigen Kaplan-Meier-Kurve bestätigt. So ist nach 2000 Tagen bei fehlender Ulzeration das Überleben bei ca. 0,92, bei einer vorhandenen Ulzeration hingegen ist es bei ca. 0,64 und damit erheblich schlechter.

3.3.1.4 Histologischer Subtyp

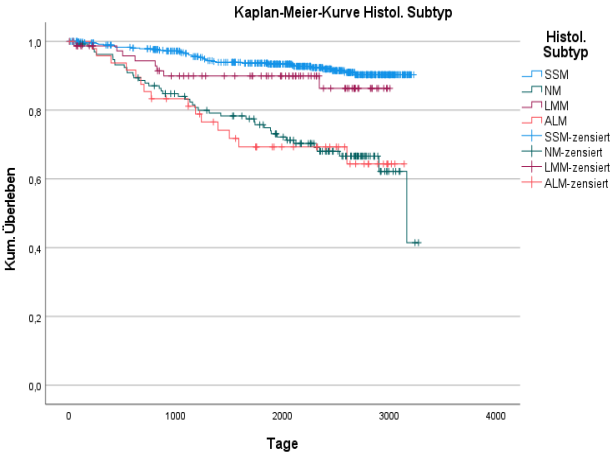

Abbildung 29: Kaplan-Meier-Kurve: Histologischer Subtyp

Es wurde anhand des Log-Rank getestet und ist mit p<0,001 statistisch hochsignifikant. Die Kurve zeigt, dass das SSM mit der besten Prognose einhergeht. Weiterhin ist zu erkennen, dass nach 2000 Tagen das Gesamtüberleben des NM und des ALM sehr ähnlich sind, aber im späten Verlauf das Überleben des NM deutlich schlechter wird.

3.3.2 Morphologische Faktoren

3.3.2.1 Cytomorphologie

Die Kaplan-Meier-Analyse der Cytomorphologie weist einen Log-Rank-Test mit p=0,553>0,05 auf und ist als nicht signifikant zu werten.

3.3.2.2 Zellatypie

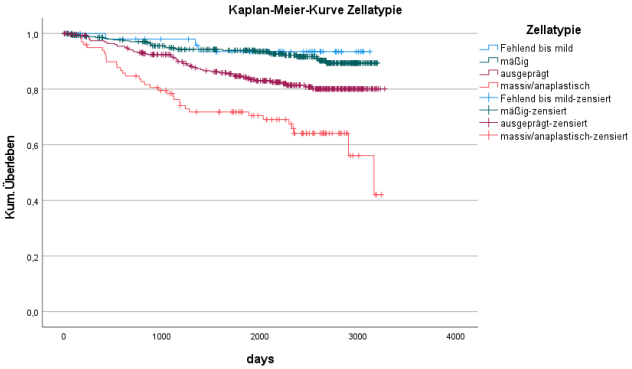

Abbildung 30: Kaplan-Meier-Kurve: Zellatypie

Der Log-Rank-Test ist mit p<0,001 hochsignifikant. Eine zunehmende Zellatypie ist ein Faktor, der das Überleben verschlechtert. Die Kategorien „fehlend bis mild“ und „mäßig“ zeigen sich in ihrem Überleben relativ nah

beieinander. Deutlich schlechter wird das Überleben bei den zwei folgenden Kategorien „ausgeprägt“ und „massiv/anaplastisch“.

3.3.2.3 Mitosen

Abbildung 31: Kaplan-Meier-Kurve: Mitosen

Der Log-Rank-Test ist mit $p<0,001$ hochsignifikant. Die Klassen zeigen einen signifikanten Unterschied bezüglich des Überlebens. Vor allem die ersten beiden Kategorien „fehlend“ und „einzelne“, die sich auch nach 2000 Tagen relativ nah beieinander bewegen, zeigen einen deutlichen Unterschied

gegenüber der letzten Kategorie „viele“.

3.3.2.4 Entzündungsinfiltrat

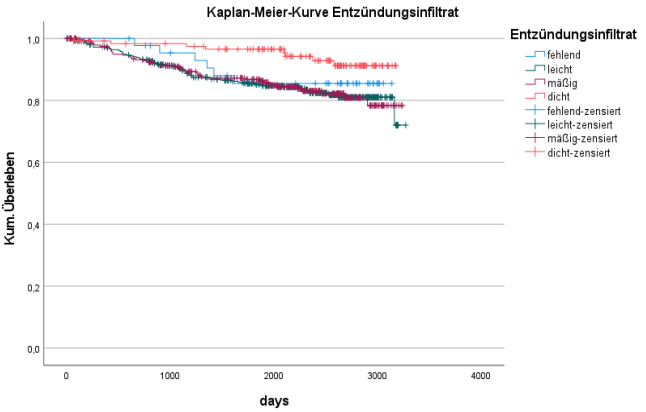

Abbildung 32: Kaplan-Meier-Kurve: Entzündungsinfiltrat

Der Log-Rank-Test ist mit p=0,041<0,05 signifikant. Die Kaplan-Meier-Kurve des Entzündungsinfiltrats zeigt ein nicht ganz schlüssiges Ergebnis. Wohingegen bis auf die Kategorie „dicht“ nach 2000 Tagen die drei weiteren Kategorien beinahe das gleiche kumulative Überleben hervorweisen, zeigen die beiden Extremen „fehlend“ und „dicht“ nach 3000 Tagen interessanterweise das beste

Gesamtüberleben. Die Zwischenkategorie „leicht“ und „mäßig“ bewegen sich bezüglich ihren Überlebens nach 3000 Tagen bei etwa 0,8 und sind damit ähnlich schlechter gegenüber den beiden vorher genannten Kategorien. Ein dichtes Entzündungsinfiltrat geht mit dem besten Überleben einher (2000 Tage ca. 0,93).

3.3.2.5 Wachstumsmuster

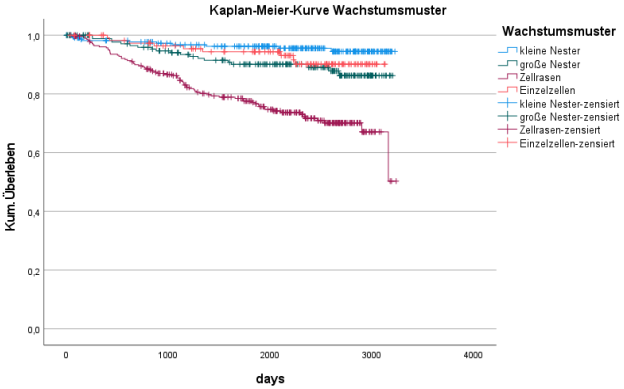

Abbildung 33: Kaplan-Meier-Kurve: Wachstumsmuster

Der Log-Rank-Test ist mit $p<0,001$ hochsignifikant. Die Kategorie „Zellrasen“ zeigt ein deutlich schlechteres Gesamtüberleben bei

etwa 0,7 nach 2000 Tagen, wohingegen sich die anderen Kategorien nach 2000 Tagen zwischen 0,9 und 0,97 bewegen. Die Kategorie „Einzelzellen“ hat nach 2000 Tagen ein schlechteres Gesamtüberleben als die Kategorie „kleine Nester“.

3.3.2.6 Pigmentierung

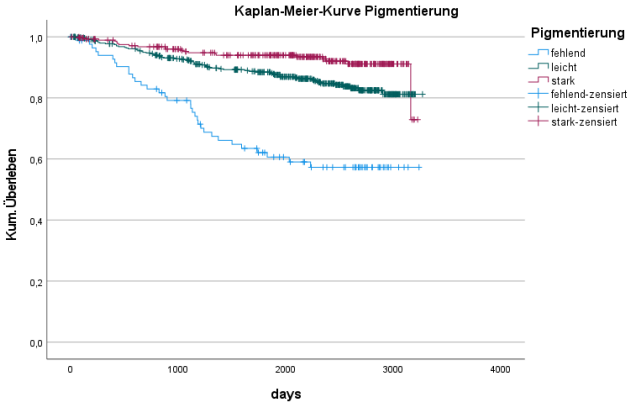

Abbildung 34: Kaplan-Meier-Kurve: Pigmentierung

Der Log-Rank-Test ist mit $p<0,001$ hochsignifikant. Eine fehlende Pigmentierung zeigt bereits nach wenigen Tagen ein deutlich schlechteres Gesamtüberleben als die beiden anderen Kategorien. Nach 2000 Tagen hat ein

Melanomerkrankter mit zunehmender Pigmentierung ein deutlich besseres Überleben. Eine leichte Pigmentierung verbessert das Überleben gegenüber einer fehlenden erheblicher als eine starke gegenüber einer leichten.

3.4 Auswertung der CNN-Prognose

3.4.1 Prognose der beiden erstellten Modelle

Bevor die Einordnung der CNN dargestellt wird, sollte zunächst das reale Überleben angegeben werden - folgend anhand des 2 Jahres-, 5 Jahres- und Gesamtüberleben. Es ist noch einmal zu betonen, dass die beiden Google teachable machine Modelle über das overall-survival (OS) trainiert wurden, d.h. dass die Klassen „dead“ und „alive“ darauf gebildet wurden. Das Modell, das an den Übersichtsbildern trainiert wurde, wird nachfolgend Übersichtsbild-Modell (ÜB-Modell) genannt, das der „area of interest“ AOI-Modell:

Tabelle 21: Zwei-Jahres-, Fünf-Jahres- und Gesamtüberlebensrate der Gesamt-, Trainings-, und Validierungskohorte

	Gesamtkohorte (n=836)	**Trainingskohorte (n=502)**	**Validierungskohorte (n=334)**
Two-year survival (2YS)	- Alive – 94,7% (n=792) - Dead – 5,3% (n=44)	- Alive – 95,0% (n=477) - Dead – 5,0% (n=25)	- Alive – 94,3% (n=315) - Dead – 5,7% (n=19)
Five-year survival (5YS)	- Alive – 88,6% (n=741) - Dead – 11,4% (n=95)	- Alive – 89,0% (n=447) - Dead – 11,0% (n=55)	- Alive – 88,0% (n=294) - Dead – 12,0% (n=40)

Overall-survival (OS)	- Alive – 85,9% (n=718) - Dead – 14,1% (n=118)	- Alive – 86,3% (n=433) - Dead - 13,7% (n=69)	- Alive – 85,3% (n=285) - Dead – 14,7% (n=49)

Die unterschiedlichen „survival“-Kategorien zeigen, dass wie erwartet mit längerem Zeitraum auch die Sterblichkeit zunimmt. Bei einer anfangs noch niedrigen Sterblichkeit von 5,3% (2YS, Gesamtkohorte), liegt sie über dem gesamten Beobachtungszeitraum bei 14,1% (OS, Gesamtkohorte). Es ist festzustellen, dass sich zwischen den Gruppen keine wesentlichen, relativen Unterschiede feststellen lassen. Die Validierungskohorte hat mit 14,7% den relativ gesehen höchsten Anteil der „dead“-Gruppe innerhalb des OS.

Nun also die Betrachtung der Prognoseabgabe der CNN:

Tabelle 22: Gesamtprognose der CNN anhand der Validierungskohorte

	Validierungskohorte
Übersichtsbild-Modell (ÜB)	- Alive – 92,5% (n=309) - Dead – 7,5% (n=25)
„Area of interest“-Modell (AOI)	- Alive – 69,8% (n=233) - Dead – 30,2% (n=101)

Da die KI anhand des OS trainiert wurde, wird es in der Auswertung auch daran verglichen. Zuerst ist festzuhalten, dass es einen deutlichen Unterschied zwischen dem Modell der Übersichtsbilder (ÜB) und dem Modell der „area of interest“ (AOI) gibt (Alive: 92,5% vs. 69,8%). Bei den Übersichtsbildern hat die CNN das OS besser als das reale eingeordnet (92,5% vs. 85,3%). Die CNN, welche mit den

AOI-Schnitten trainiert wurde, schätzt die Prognose schlechter als das reale Überleben (69,8% vs. 85,3%) ein. Die angegebenen Häufigkeiten sagen auch noch nicht aus, ob auch wirklich bei individuellen Schnitten die CNN-Prognose mit dem tatsächlichen Gesamtüberleben übereinstimmt:

Tabelle 23: Übersichtsbild-Modell: Übereinstimmung der CNN-Prognose mit dem tatsächlichen Überleben

	Übereinstimmung (Stimmt die Prognose der CNN mit dem realen Überleben überein?)
Gesamt – Alive und Dead	- Ja – 84,13% (n=281) - Nein – 15,87% (n=53)
Alive	- Ja – 94,76% (n=271) - Nein – 5,24% (n=15)
Dead	- Ja – 20,83% (n=10) - Nein – 79,17% (n=38)

Tabelle 24: <u>Area of interest-Modell</u>: Übereinstimmung der CNN-Prognose mit dem tatsächlichen Überleben

	Übereinstimmung (Stimmt die Prognose der CNN mit dem realen Überleben überein?)
Gesamt – Alive und Dead	- Ja – 70,36% (n=235) - Nein – 29,64% (n=99)
Alive	- Ja – 73,43% (n=210) - Nein – 26,57% (n=76)
Dead	- Ja – 52,08% (n=25) - Nein – 47,92% (n=23)

Es lässt sich sagen, dass die Google teachable machine Modelle sowohl im Übersichtsbild als auch in den AOI-Schnitten das Gesamtüberleben eher richtig als falsch einordnen (Übersichtsbild 84,13% und AOI 70,36%). Das Übersichtsbild-Modell schätzt bei den alive-Schnitten der Validierungskohorte

das tatsächliche Überleben zu 94,76% (n=271) richtig ein. Deutlich schlechter ist die Prognose jedoch bei den dead-Schnitten. Hier liegt das Übersichts-Modell gerade einmal bei 20,83% (n=10) richtig. Demgegenüber steht das AOI-Modell. Dies hat bei den dead-Schnitten eine deutlich bessere Trefferquote mit 52,08% (n=25). Dafür jedoch gibt es eine Einbuße bei der alive-Gruppe. Hier werden nur noch 73,43% (n=210) richtig eingestuft, damit also 21,33% schlechter als bei dem Übersichtsbild-Modell.

Im Methodik-Teil wurde erwähnt, dass die Zuordnung des Modells zusätzlich anhand einer Prozentzahl erfolgt. Dabei wird ein Schnitt entweder der „alive“ oder „dead“-Klasse in Prozent zugeordnet. Bei zwei vorhandenen Klassen liegt der cut-off Wert bei 50%, d.h. wird ein Schnitt der Kategorie „dead“ oder „alive“ zu 51% oder mehr zugeordnet, ist er dieser

zugehörig. Im Rahmen dieser statistischen Auswertung hat zur besseren Handhabung eine Transformation dieser Prozentzahlen stattgefunden. Dabei wurde 0% auf „alive“ und 100% auf „dead“ festgelegt, d.h. Schnitte ab 51% gehören zur „dead“-Gruppe:

Tabelle 25: <u>Übersichtsbild-Modell</u>: Prognose der CNN anhand der Prozenteinordnung von 0% (alive) bis 100% (dead)

Prozentzuordnung innerhalb des Modells	Häufigkeiten	Prozent
0	275	82,3
1	7	2,1
2	6	1,8

3	3	,9
4	1	,3
6	2	,6
7	1	,3
8	1	,3
9	2	,6
11	1	,3
12	1	,3
14	1	,3
15	3	,9
16	1	,3
22	1	,3
23	1	,3
35	1	,3
44	1	,3
67	1	,3
73	1	,3
77	1	,3
80	1	,3
84	1	,3

85	1	,3
86	1	,3
88	1	,3
91	1	,3
95	3	,9
98	1	,3
99	4	1,2
100	8	2,4
Gesamt	334	100,0

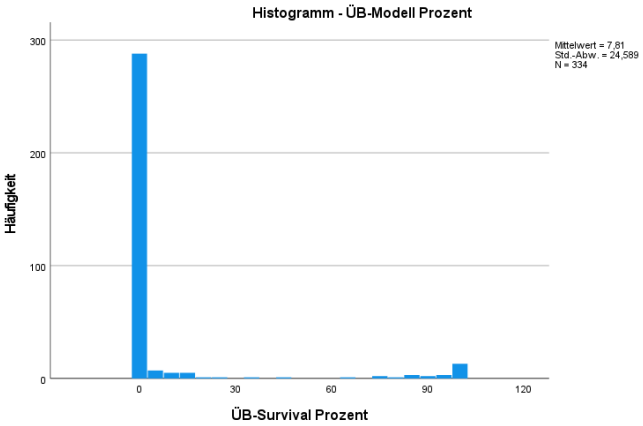

Abbildung 35: Histogramm: Übersichtsbild-Modell: Prognose der CNN anhand der Prozenteinordnung von

0% (alive) bis 100% (dead)

Mit 275 von 334 Schnitten ist 0%, d.h. eine vollständige Zuordnung zur „alive“-Gruppe am häufigsten. Die zweithäufigste Variable ist 100%, d.h. die vollständige Zuordnung zur „dead“-Gruppe mit 8 Schnitten. 51 Schnitte waren nicht komplett der „alive“- oder „dead“-Gruppe zugehörig.

Tabelle 26: AOI-Modell: Prognose der CNN anhand der Prozenteinordnung von 0% (alive) bis 100% (dead)

Prozentzuordnung innerhalb des Modells	Häufigkeiten	Prozent
0	161	48,2
1	13	3,9
2	10	3,0
3	7	2,1

4	3	,9
5	2	,6
6	4	1,2
7	5	1,5
9	1	,3
10	1	,3
11	1	,3
13	3	,9
14	1	,3
15	2	,6
16	1	,3
18	1	,3
19	1	,3
20	1	,3
21	1	,3
22	1	,3
25	1	,3
27	1	,3
29	2	,6
30	1	,3

34	1	,3
39	2	,6
40	2	,6
41	1	,3
44	1	,3
47	1	,3
52	1	,3
57	1	,3
58	2	,6
64	1	,3
69	1	,3
70	1	,3
71	2	,6
74	1	,3
76	1	,3
78	1	,3
81	1	,3
83	1	,3
85	1	,3
87	2	,6

88	1	,3
89	2	,6
90	1	,3
92	2	,6
93	1	,3
94	1	,3
96	2	,6
97	1	,3
98	5	1,5
99	9	2,7
100	59	17,7
Gesamt	334	100,0

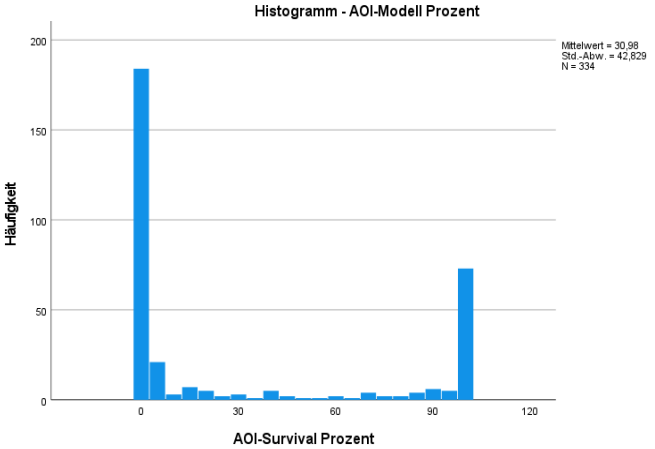

Abbildung 36: Histogramm: AOI-Modell: Prognose der CNN anhand der Prozenteinordnung von 0% (alive) bis 100% (dead)

Die Verteilung des AOI-Modells zeigt deutliche Unterschiede gegenüber dem ÜB-Modell. Auf die beiden Extremen 0% und 100% fallen hier nur noch 230 Schnitte und 114 Schnitte (vs. 51 beim ÜB-Modell) sind hier nicht komplett einer der beiden Gruppen zugehörig.

3.4.2 ROC-Analyse der beiden Modelle

Eine Klassifizierung innerhalb derart erstellter Modelle erfolgt oftmals anhand des Wertes 0,5. Konkreter möchte man die Güte des Modells prüfen, indem man schaut, ob das Modell eher richtig als falsch liegt. Übersichtlich lässt sich dies mithilfe von ROC-Kurven darstellen und analysieren:

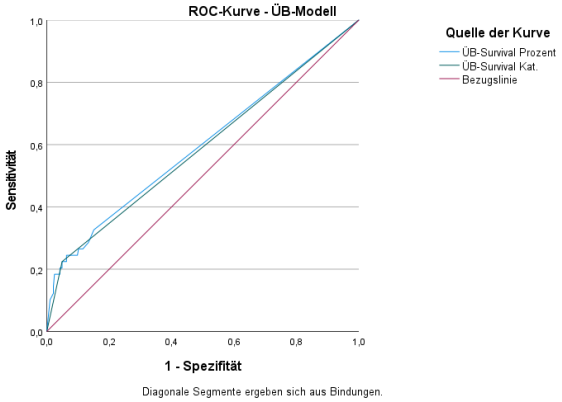

Abbildung 37: ROC-Kurve: ÜB-Modell

Tabelle 27: ROC-Analyse: Prognose des ÜB-Modells kategorial in „dead“ und „alive“, wie auch prozentual

	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
ÜB-Survival Prozent	,597	,048	,031	,503-,691
ÜB-Survival kat.	,588	,048	,05	,494-,682

Die Flächen unter der Kurve mit 0,588 und 0,597 lassen eine Güte des Modells feststellen, wobei die prozentuale Zuordnung überlegen ist.

Auch die Kurve über der Diagonalen zeigt, dass die Modelle mehr richtig als falsch liegen. Mit p=0,031 ist das ÜB-Survival Prozent-Modell signifikant; der p-Wert bei 0,05 des ÜB-Survival kat.-Modells ist als nicht-signifikant zu werten.

Wie sich die AOI-Prognose im Vergleich einordnen lässt, soll folgend aufgeführt werden:

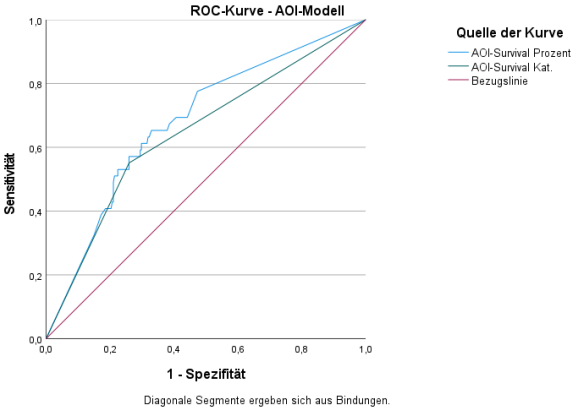

Abbildung 38: ROC-Kurve: AOI-Modell

Tabelle 28: ROC-Analyse: Prognose des AOI-Modells kategorial in „dead“ und „alive“, wie auch prozentual

	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
AOI-Survival Prozent	,681	,041	,000	,601-,762
AOI-Survival kat.	,646	,045	,001	,558-,733

Ein deutlich besseres Ergebnis erhält man bei der ROC-Analyse des AOI-Modells. Hier sind beide Zuordnungsmodelle mit der „area under the curve“ (Fläche unter der Kurve) mit 0,646

und 0,681 noch einmal besser als beim ÜB-Modell. Wieder lässt sich die prozentuale Zuordnung als überlegen feststellen. Auch lassen sich hier bei beiden Zuordnungen signifikantere p-Werte mit 0,001 und <0,001 nachweisen und die 95%-Konfidenzintervalle liegen über 0,5 (Kriterium für die Güte des Modells).

Zusammenfassend kann also gesagt werden, dass das AOI-Modell die Prognose in beiden Gruppen, die dead-Schnitte und die alive-Schnitte, eher richtig als falsch einschätzt. Das Modell der Übersichtsbilder hat dahingegen eine hohe Trefferquote bei den alive-Schnitten. Anhand der ROC-Analysen und Kurven ist zudem festzuhalten, dass das CNN AOI-Modell dem CNN ÜB-Modell überlegen ist.

3.5 Binäre logistische Regressionsanalyse

Bei dem vorhergesagten Survival der Modelle handelt es sich um eine dichotome, nominale, abhängige Variable, weshalb sich die binäre logistische Regressionsanalyse hervorragend eignet, um eben dieses Survival der Modelle auf eine Abhängigkeit von den unterschiedlichen morphologischen und histologischen Parametern zu untersuchen. Die Auswertung der Prozentangabe der CNN soll in der Regressionsanalyse vernachlässigt werden, da die primäre Ergebnisausgabe der CNN kategorial in „dead“ und „alive“ erfolgt. Eine Abwägung zwischen einer Überlebensanalyse anhand von Log-Rank-Tests und Kaplan-Meier-Schätzmethoden ist erfolgt. Letztlich fiel die Entscheidung auf die logistische Regression, da sie in diesem Rahmen ein umfangreicheres Ergebnis liefert.

Weiterhin soll innerhalb dieser Auswertung die zeitliche Variable vernachlässigt werden, was innerhalb einer Überlebensanalyse schwierig bzw. nicht möglich ist. Als abschließender Punkt ist zusätzlich aufzuführen, dass zwar die Prognose „dead“ oder „alive“ innerhalb des Modells ist, es sich jedoch nicht um das tatsächliche Überleben handelt, sondern nur um eine Einordnung des Modells. Deshalb ist die logistische Regression in diesem Falle zur Analyse der einzelnen Variablen i.R. der Prognose des Modells gewählt worden.

Eine binär logistische Regression analysiert dies, indem sie selbst Blöcke (oder Modelle) erstellt. Hierbei wird im ersten Schritt nur ein Block erstellt, der nicht die unabhängige Variable enthält, sondern eine vom System ausgewählte Konstante. Anschließend wird ein zweiter Block erstellt mit der gewünschten zu analysierenden unabhängigen Variablen. Die

beiden Modelle werden nun verglichen und analysiert, um hervorzubringen, ob sich ein signifikanter Unterschied darstellen lässt. Dabei werden Omnibustests der Modellkoeffizienten (beinhaltet Chi-Quadrat-Test) durchgeführt, bei dem die herausgegebene Signifikanz $p<0,05$ liegen muss. Gleichzeitig wird der Hosmer-Lemeshow-Test durchgeführt, der die erwarteten Ereignisse mit den realen vergleicht. Wichtig ist zu verstehen, dass zur Annahme des Modells die Signifikanz des Hosmer-Lemeshow-Tests **über** 0,05 liegen muss. Innerhalb des erstellten Regressionsmodells wird bei kategorialen, unabhängigen Variablen immer mit einer Referenzkategorie verglichen und anschließend ausgewertet, ob ein signifikanter Unterschied in der Abhängigkeit dazu besteht (Wald-Test). Dabei wurde bei nominal skalierten Variablen die „niedrigste“ Klasse als Referenz ausgewählt, z.B. die

Klasse I bei der pT-Einteilung nach AJCC 2009. Alle Variablen wurden einzeln anhand der logistischen Regression betrachtet. Zwar hat dies den Nachteil eines möglich verbleibenden Confounder-Effekts, jedoch soll keine Kontrolle einzelner möglicher Störvariablen erfolgen, da ggf. eine fehlerhaft schwache Odds Ratio auftreten kann. Die Klassifizierung innerhalb der Modelle erfolgte anhand des Wertes 0,5. Die maximale Fallzahl liegt bei n=334 (Validierungskohorte). Folgend die Regressionsanalyse der herkömmlichen Prognoseparameter:

3.5.1 Etablierte Prognosefaktoren

3.5.1.1 Tumordicke nach Breslow (verhältnisskaliert)

3.5.1.1.1 Übersichtsbild-Modell:

Tabelle 29: Logistische Regression: Übersichtsbild-Modell - Tumordicke nach Breslow verhältnisskaliert

	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Tumordicke	,299	,069	18,528	,000	1,348	1,177-1,545

Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 0,116 (muss >0,05 zur Annahme des Regressionsmodells). Die Odds Ratio ist bei 1,348 und damit innerhalb des 95%

Konfidenzintervalls (1,177; 1,545).

Die eindeutige Abhängigkeit der Survivaleinordung des KI-Modells an der Tumordicke ist festzustellen. An dem über eins liegenden Konfidenzintervall und der Odds Ratio (OR) schätzt das Modell bei der Zunahme von 1 mm der Tumordicke eine Zunahme der relativen Sterblichkeit um 34,8%.

3.5.1.1.2 AOI-Modell:

Tabelle 30: Logistische Regression: AOI-Modell - Tumordicke nach Breslow verhältnisskaliert

	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Tumordicke	,169	,054	9,803	,002	1,184	1,065-1,317

Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,002 ein signifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 0,116 (muss >0,05 zur Annahme des Regressionsmodells). Die Odds Ratio ist mit 1,184 innerhalb des 95% Konfidenzintervalls (1,065-1,317). Bei der Zunahme der Tumordicke von 1 mm ist eine Zunahme der relativen Sterblichkeit von 18,4% innerhalb dieses Modells zu erwarten. Auch im AOI-Modell lässt sich also eine Abhängigkeit der Prognoseabgabe der KI zur Tumordicke feststellen.

3.5.1.2 Tumordicke nach Breslow (pT AJCC 2009)

Wie es sich mit den vier Klassen der pT-Einteilung des AJCC 2009 verhält, und ob eine signifikante Abhängigkeit der Prognosefähigkeit der beiden Modelle vorliegt, soll folgend auch mit einer binär logistischen

Regressionsanalyse bewiesen werden:

3.5.1.2.1 Übersichtsbild-Modell:

Tabelle 31: Logistische Regression: Übersichtsbild-Modell - Tumordicke nach Breslow (pT AJCC 2009)

Tumordicken. Breslow - pT AJCC 2009	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Bis 1 mm			14,920	,002		
1,0 mm-2,0 mm	2,752	1,079	6,511	,011	15,676	1,893-129,803
2,0 mm – 4,0 mm	3,208	1,082	8,787	,003	24,733	2,965-206,318
>4,0 mm	3,875	1,066	13,21	,00	48,182	5,962-389,366

			1	0		

333 Fälle wurden analysiert, 1 Fall fehlt. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00 (muss >0,05 zur Annahme des Regressionsmodells). Die Referenzkategorie ist „bis 1,0 mm“. Das Prognoseergebnis des ÜB-Modell ist deutlich abhängig von der Tumordicke nach Breslow, wie in der obigen Regressionsanalyse hervorgebracht wurde. Alle drei Klassen (II, III und IV) zeigen sich gegenüber der Referenzkategorie (Klasse I) als signifikanter, negativer Prognosefaktor (II: p-Wert bei 0,011<0,05 und OR=15,676>1 mit 95%-KI 1,893-129,803; III: p-Wert bei 0,003<0,05 und OR=24,733>1 mit 95%-KI 2,965-206,31; IV: p-Wert bei <0,001 und OR=48,182>1 mit 95%-KI 5,962-389,366). Zusätzlich lässt sich auch

sagen, dass mit höherer Klassifikation eine negativere Prognose innerhalb des ÜB-Modells einhergeht.

3.5.1.2.2 AOI-Modell:

Tabelle 32: Logistische Regression: AOI-Modell - Tumordicke nach Breslow (pT AJCC 2009)

Tumordicken. Breslow - pT AJCC 2009	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Bis 1 mm			18,643	,000		
1,0 mm-2,0 mm	,596	,312	3,664	,056	1,815	,986-3,343
2,0 mm – 4,0	1,425	,339	17,616	,000	4,156	2,137-8,084

mm						
>4,0 mm	,824	,371	4,935	,026	2,281	1,102-4,720

333 Fälle wurden analysiert, 1 Fall fehlt. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00 (muss >0,05 zur Annahme des Regressionsmodells). Die Referenzkategorie ist „bis 1,0 mm“ (Klasse I). Es zeigt sich im Vergleich zum ÜB-Modell beim AOI-Modell kein signifikanter Unterschied (p=0,056>0,05 und OR=1,815 mit 95%-KI 0,986-3,343) der Klasse II zur Klasse I hinsichtlich des Überlebens bei der Tumordicke nach Breslow. Gleich jedoch ist das auch im AOI-Modell ein signifikanter Unterschied bei den Klassen III und IV festgestellt werden kann (III: p-Wert bei <0,001 und OR=4,156>1 mit

95%-KI 2,137-8,084; IV: p-Wert bei 0,026<0,001 und OR=2,281>1 mit 95%-KI 1,102-4,720). Eine Zunahme der Klasse ist ein negativer Prognosefaktor (pos. Regressionskoeffizient; 95%-KI>1).

3.5.1.3 Invasionslevel nach Clark

3.5.1.3.1 Übersichtsbild-Modell:

Tabelle 33: Logistische Regression: Übersichtsbild-Modell - Invasionslevel nach Clark

Invasionslevel nach Clark	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
II			3,846	,279		
III	-17,142	4838,665	,000	,99	,000	,000

				7		
IV	2,035	1,038	3,846	,050	7,653	1,001-58,494
V	-17,142	15191,515	,000	,999	,000	,000

298 Fälle wurden analysiert, 36 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00 (muss >0,05 zur Annahme des Regressionsmodells). Die Referenzkategorie ist „Invasionslevel II“. Im Rahmen der Auswertung wurde die Signifikanz bei p<0,05 festgelegt und beim Invasionslevel IV liegt der p-Wert des Unterschieds zu Invasionslevel II bei genau 0,05 und wird daher als nicht-signifikant gewertet. Es lässt sich allerdings ein negativer prognostischer Effekt

des Invasionslevels IV vermuten (pos. Regressionskoeffizient; OR>1).

3.5.1.3.2 AOI-Modell:

Tabelle 34: Logistische Regression: AOI-Modell - Invasionslevel nach Clark

Invasionslevel nach Clark	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
II			16,440	,001		
III	-,073	,478	,024	,878	,929	,364-2,372
IV	1,127	,382	8,686	,003	3,087	1,459-6,533
V	,673	,906	,552	,458	1,960	,332-11,568

298 Fälle wurden analysiert, 36 fehlen. Die Omnibustests der Modellkoeffizienten ergeben mit $p<0,001$ ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00 (muss >0,05 zur Annahme des Regressionsmodells). Die Referenzkategorie ist „Invasionslevel II“. Im AOI-Modell lässt sich nun ein negativer Effekt der Gruppe IV im Vergleich zur Gruppe II (p-Wert bei 0,003<0,05 und OR=3,087> mit 95%-KI 1,459-6,533) auf das Überleben nachweisen. Die Invasionslevel III und V zeigen keinen signifikanten Unterschied (p_1=0,878>0,05 und p_2=0,458>0,05) zur Referenzkategorie.

3.5.1.4 Ulzeration

3.5.1.4.1 Übersichtsbild-Modell:

Tabelle 35: Logistische Regression: Übersichtsbild-Modell - Ulzeration

	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Ulzeration	2,007	,441	20,703	,000	7,443	3,135-17,670

331 Fälle wurden analysiert, 3 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit $p<0,001$ ein hochsignifikantes Ergebnis. Der Hosmer-Lemeshow-Test war nicht durchführbar. Referenzkategorie ist „keine Ulzeration“. Die Regressionsanalyse der Ulzeration bringt eine signifikante Abhängigkeit des Übersichtsbild-Modells hervor

(OR=7,443>1 KI 3,135;17,670; p=,0,000). Liegt eine Ulzeration vor, verschlechtert (pos. Regressionskoeffizient; OR>1) sich das Überleben innerhalb des Modells. Das Vorliegen einer Ulzeration erhöht die Wahrscheinlichkeit zu „dead“ zugeordnet zu werden, um das 6,44-fache (OR=7,443).

3.5.1.4.2 AOI-Modell:

Tabelle 36: Logistische Regression: AOI-Modell - Ulzeration

	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Ulzeration	1,146	,273	17,660	,000	3,146	1,843-5,370

331 Fälle wurden analysiert, 3 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes

Ergebnis. Der Hosmer-Lemeshow-Test war nicht durchführbar. Referenzkategorie ist „keine Ulzeration“. Auch im AOI-Modell besteht eine positive Abhängigkeit zur Ulzeration (OR=3,146; KI 1,843-5,370>1 und p-Wert bei 0,000; pos. Regressionskoeffzient). Hier erhöht die Ulzeration die Wahrscheinlichkeit der Zugehörigkeit zur „dead“-Kategorie um das 2,15-fache (OR=3,146).

3.5.1.5 Histologische Subtypen

3.5.1.5.1 Übersichtsbild-Modell:

Tabelle 37: Logistische Regression: Übersichtsbild-Modell - Histologische Subtypen

Subtyp	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR

SSM			10,620	,014		
NM	1,539	,478	10,373	,001	4,660	1,827-11,887
LMM	-18,191	8204,356	,000	,998	,000	,000
ALM	1,115	,707	2,488	,115	3,050	,763-12,194

298 Fälle wurden analysiert, 36 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,002 ein signifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Referenzkategorie ist das „SSM“. Beim ÜB-Modell besteht ein signifikanter Unterschied bzgl. des Überlebens zwischen dem NM und dem SSM (p=0,001<0,05 mit OR=4,660>1 und 95%-KI 1,827-11,887). Das Überleben ist schlechter beim NM (pos. Regressionskoeffizient). Vielmehr noch ist bei

einem NM eine 3,66-fache relative Sterblichkeitszunahme innerhalb des ÜB-Modells zu erwarten.

3.5.1.5.2 AOI-Modell:

Tabelle 38: Logistische Regression: AOI-Modell – Histologische Subtypen

	Chi-Quadrat-Test	Signifikanz
Modell	6,773	,08

Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,08>0,05 ein nicht-signifikantes Ergebnis, d.h. es lässt sich keine signifikante Abhängigkeit des Survivals zum hist. Subtyp innerhalb des AOI-Modells nachweisen.

3.5.1.6 Nävusassoziation

3.5.1.6.1 Übersichtsbild-Modell:

Tabelle 39: Logistische Regression: Übersichtsbild-Modell - Nävusassoziation

	Chi-Quadrat	Sig.
Modell	1,367	,242

296 Fälle wurden analysiert, 38 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,242>0,05 ein nicht-signifikantes Ergebnis. Die Prognoseaussage des ÜB-Modells zeigt keine Abhängigkeit zur Nävusassoziation.

3.5.1.6.2 AOI-Modell:

Tabelle 40: Logistische Regression: AOI-Modell - Nävusassoziation

	Chi-Quadrat	Sig.
Modell	,294	,587

296 Fälle wurden analysiert, 38 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,587>0,05 ein nicht-signifikantes Ergebnis. Auch die Prognoseaussage des AOI-Modells zeigt keine Abhängigkeit zur Nävusassoziation.

3.5.2 Morphologische Parameter

Folgend soll die Abhängigkeit der Modelle zu den morphologisch erhobenen Parametern binär logistisch analysiert werden:

3.5.2.1 Cytomorphologie

3.5.2.1.1 Übersichtsbild-Modell:

Tabelle 41: Logistische Regression: Übersichtsbild-Modell - Cytomorphologie

	Chi-Quadrat	Sig.
Modell	3,373	,66

Alle 334 Fälle wurden analysiert. Bei den

Omnibustests der Modellkoeffizienten erhält man mit p=0,66>0,05 ein nicht-signifikantes Ergebnis. Die Prognoseaussage des ÜB-Modells zeigt keine Abhängigkeit zur Cytomorphologie.

3.5.2.1.2 AOI-Modell:

Tabelle 42: Logistische Regression: AOI-Modell - Cytomorphologie

	Chi-Quadrat	Sig.
Modell	1,622	,203

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,203>0,05 ein nicht-signifikantes Ergebnis. Die Prognoseaussage des AOI-Modells zeigt keine Abhängigkeit zur Cytomorphologie.

3.5.2.2 Zellatypie

3.5.2.2.1 Übersichtsbild-Modell:

Tabelle 43: Logistische Regression: Übersichtsbild-Modell - Zellatypie

Zellatypie	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.
Fehlend bis mild			7,830	,050
Mäßig	17,907	9473,609	,000	,998
Ausgeprägt	18,964	9473,609	,000	,998
Massiv/ anaplastisch	19,622	9473,609	,000	,998

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,009 ein signifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Das Regressionsmodell zeigt jedoch keinen Unterschied der weiteren Klassen der

Zellatypie zur Referenzkategorie „Fehlend bis mild“ (p_{1-3}=0,998>0,05).

3.5.2.2.2 AOI-Modell:

Tabelle 44: Logistische Regression: AOI-Modell - Zellatypie

Zellatypie	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Fehlend bis mild			16,549	,001		
Mäßig	,223	,667	,112	,738	1,250	,338-4,619
Ausgeprägt	1,110	,657	2,858	,091	3,036	,838-11,000
Massiv/ anapl	1,463	,706	4,296	,038	4,318	1,083-17,220

astisch						

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,001 ein signifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist „Fehlend bis mild“. Ein anderes Ergebnis findet man innerhalb des AOI-Modells. Ein signifikanter negativer (pos. Regressionskoeffizient) Unterschied bei der Zellatypie von „Fehlend bis mild“ zu „Massiv/ anaplastisch“ (p-Wert bei 0,038<0,05 und OR=4,318>1 mit 95%-KI 1,083-17,220) lässt sich mithilfe der log. Regression darstellen. Die relative Sterblichkeitszunahme liegt hier bei 2,04 (OR=3,036). Die Kategorien „mäßig“ und „ausgeprägt“ haben keinen signifikanten Unterschied zur Referenzkategorie (p_1=0,738>0,05 und p_2=0,09>0,05).

3.5.2.3 Mitosen

3.5.2.3.1 Übersichtsbild-Modell:

Tabelle 45: Logistische Regression: Übersichtsbild-Modell - Mitosen

Mitosen	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Fehlend			19,239	,000		
Einzelne	,461	,828	,311	,577	1,586	,313-8.036
Viele	2,342	,636	13,542	,000	10,405	2,988-36,225

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-

Tests liegt bei 1,00. Die Referenzkategorie ist „Fehlend“. Die Kategorie „Viele“ Mitosen zeigt sich als negativer Prognosefaktor (pos. Regressionskoeffizient; OR>1) im Vergleich zur Kategorie „Fehlend“ (p-Wert bei <0,001 mit OR=10,405 und 95%-KI 2,988-36,225). Die Kategorie „Einzelne“ zeigt keinen signifikanten Unterschied zur Referenzkategorie (p=0,577>0,05).

3.5.2.3.2 AOI-Modell:

Tabelle 46: Logistische Regression: AOI-Modell - Mitosen

Mitosen	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
Fehlend			17,666	,000		

Einzelne	,645	,312	4,266	,039	1,907	1,034-3,518
Viele	1,226	,292	17,663	,000	3,407	1,924-6,036

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist „Fehlend“. Die Regressionsanalyse der Mitosen innerhalb des AOI-Modells zeigt einen signifikanten Unterschied (I: p-Wert bei 0,039<0,05 mit OR=1,907>1 und 95%-KI 1,034-3,518; II: p-Wert bei <0,001 mit OR=3,407>1 und 95%-KI 1,924-6,036) beider Kategorien („Einzelne“, „viele“) zur Referenzkategorie. Eine Zuordnung zur Kategorie „Einzelne“ erhöht (pos.

Regressionskoeffizient) bereits die relative Sterblichkeit innerhalb des Modells um 90,7%.

3.5.2.4 Entzündungsinfiltrat

3.5.2.4.1 Übersichtsbild-Modell:

Tabelle 47: Logistische Regression: Übersichtsbild-Modell - Entzündungsinfiltrat

	Chi-Quadrat	Sig.
Modell	2,616	,455

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,455>0,05 ein nicht-signifikantes Ergebnis. Die Prognoseaussage des ÜB-Modells zeigt keine Abhängigkeit zum Entzündungsinfiltrat.

3.5.2.4.2 AOI-Modell:

Tabelle 48: Logistische Regression: AOI-Modell - Entzündungsinfiltrat

	Chi-Quadrat	Sig.
Modell	1,360	,715

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p=0,715>0,05 ein nicht-signifikantes Ergebnis. Auch die Prognoseaussage des AOI-Modells zeigt keine Abhängigkeit zum Entzündungsinfiltrat.

3.5.2.5 Wachstumsmuster

3.5.2.5.1 Übersichtsbild-Modell:

Tabelle 49: Logistische Regression: Übersichtsbild-Modell - Wachstumsmuster

Wachstumsmuster	Regressionskoeffizient B	Standardfehler	Wald-Te	Sig.	Exp(B)/OR	95% Konfidenzintervall für

			st			Exp(B)/ OR
Kleine Nester			8,972	,030		
Große Nester	2,040	1,092	3,492	,062	7,690	,905-65,338
Zellrasen	2,827	1,038	7,419	,006	16,887	2,209-129,077
Einzelzellen	-16,692	6059,318	,000	,998	,000	,000

328 Fälle wurden analysiert, 6 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit $p<0,001$ ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist „kleine Nester“, da sie im Rahmen der Kaplan-Meier-Schätzmethode das beste Überleben hatte. Die Kategorie „Zellrasen“ lässt sich als Faktor innerhalb des AOI-Modells mit

signifikant schlechterer Prognose herausarbeiten (p-Wert bei 0,006<0,05 mit OR=16,887>1 und 95%-KI 2,209-129,077).

3.5.2.5.2 AOI-Modell:

Tabelle 50: Logistische Regression: AOI-Modell - Wachstumsmuster

Wachstumsmuster	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/OR	95% Konfidenzintervall für Exp(B)/OR
Kleine Nester			19,682	,000		
Große Nester	,430	,361	1,421	,233	1,537	,758-3,115
Zellrasen	1,119	,319	12,328	,000	3,062	1,639-5,718
Einzelzellen	-,500	,509	,963	,326	,607	,224-1,646

328 Fälle wurden analysiert, 6 fehlen. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist „kleine Nester“, da sie im Rahmen der Kaplan-Meier-Schätzmethode das beste Überleben hatte. „Zellrasen“ lässt sich als Faktor auch innerhalb des AOI-Modells mit signifikant schlechterer Prognose herausarbeiten (p-Wert bei <0,001 mit OR=3,062>1 und 95%-KI 1,639-5,718). Die anderen Kategorien zeigen wie auch beim ÜB-Modell keinen signifikanten Unterschied im prognostischen Überleben gegenüber der Referenzkategorie.

3.5.2.6 Pigmentierung

3.5.2.6.1 Übersichtsbild-Modell:

Tabelle 51: Logistische Regression: Übersichtsbild-Modell - Pigmentierung

Pigmentierung	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
fehlend			15,646	,000		
leicht	-1,466	,474	9,553	,002	,231	,091-,585
stark	-2,813	,811	12,027	,001	,060	,012-,294

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist

„fehlend“. Die Kategorien „leicht“ (p-Wert bei 0,002<0,05 mit OR=0,231<1 und 95%-KI 0,091-0,585) und „stark“ (p-Wert bei 0,001<0,05 mit OR=0,06<1 und 95%-KI 0,012-0,294) lassen sich als Faktoren innerhalb des ÜB-Modells herausarbeiten, die mit einem signifikant besseren Überleben einhergehen (neg. Regressionskoeffizient; OR<1).

3.5.2.6.2 AOI-Modell:

Tabelle 52: Logistische Regression: AOI-Modell - Pigmentierung

Pigmentierung	Regressionskoeffizient B	Standardfehler	Wald-Test	Sig.	Exp(B)/ OR	95% Konfidenzintervall für Exp(B)/ OR
fehlend			16,404	,000		

leicht	-1,179	,372	10,032	,002	,308	,148-,638
stark	-1,686	,417	16,321	,000	,185	,082-,420

Alle 334 Fälle wurden analysiert. Bei den Omnibustests der Modellkoeffizienten erhält man mit p<0,001 ein hochsignifikantes Ergebnis. Der p-Wert des Hosmer-Lemeshow-Tests liegt bei 1,00. Die Referenzkategorie ist „fehlend“. Das AOI-Modell zeigt in der logistischen Regression das gleiche Ergebnis wie das ÜB-Modell. Die Kategorien „leicht“ (p-Wert bei 0,002<0,05 mit OR=0,308<1 und 95%-KI 0,148-0,638) und „stark“ (p-Wert bei 0,001<0,05 mit OR=0,185<1 und 95%-KI 0,082-0,420) sind Faktoren mit einem signifikant besseren Überleben (neg. Regressionskoeffizient; OR<1).

3.6 Prognosemodellerstellung

Mithilfe der logistischen Regression sollen nun Prognosemodelle erstellt und verglichen werden. Die zentrale Frage und einer der wesentlichen Ziele dieser Arbeit, nämlich ob eine bessere Modellerstellung mit der Prognoseaussage der KI möglich ist, wird zudem in diesem Punkt behandelt. Zum Vergleich der Modelle wird weiterhin die ROC-Analyse und Kurve verwendet. Die Zustandsvariable der ROC-Kurven ist das Gesamtüberleben (OS) mit dem Event „dead“.
Nachdem eine Überlebensanalyse nach der Kaplan-Meier-Schätzmethode erfolgt ist, soll zunächst in einem ersten Schritt betrachtet werden, inwieweit sich die in diesem Projekt erhobenen, morphologischen Faktoren als alleinstehende Prognosemarker eignen. Hierfür können alle 836 Schnitte ausgewertet werden. Vereinzelt fehlende Fälle wurden in der

deskriptiven Statistik gelistet:

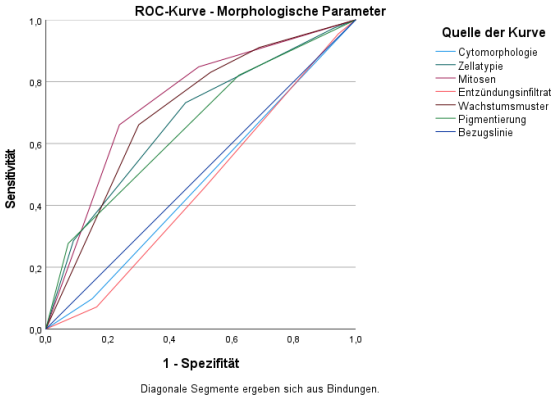

Abbildung 39: ROC-Kurve: Alle sechs erhobene, morphologische Parameter

Tabelle 53: ROC-Analyse: AUC, Standard-Fehler, Signifikanz und 95%-KI der sechs morphologischen Parameter

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Cytomorphologie	,474	,028	,379	,418-,530
Zellatypie	,674	,028	,000	,620-,729
Mitosen	,740	,025	,000	,691-,788
Entzündungsinfiltrat	,457	,027	,147	,404-,515
Wachstumsmuster	,706	,025	,000	,657-,755
Pigmentierung	,656	,029	,000	,599-,712

Die Cytomorphologie und das Entzündungsinfiltrat zeigen sich mit ihren p-Werten als nicht signifikant (p_1=0,379>0,05 und p_2=0,147>0,05). Die vier weiteren, signifikanten Faktoren zeigen sich als Prognosemarker geeignet. „Mitosen“ mit einer „area under the curve“ (AUC) von 0,74 zeigt unter den morphologischen Parametern das beste Ergebnis für die Prognosevorhersage.

Eine selbige, alleinstehende Betrachtung soll nun anhand der etablierten Faktoren erfolgen (836 Fälle, vereinzelt fehlende; siehe Punkt deskriptive Statistik):

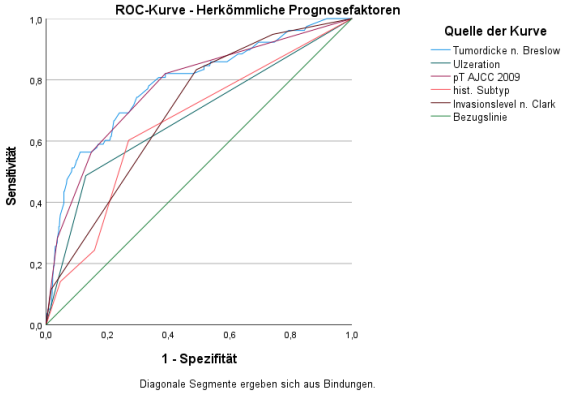

Abbildung 40: ROC-Kurve: Die fünf etablierten Prognosefaktoren

Tabelle 54: ROC-Analyse: AUC, Standard-Fehler, Signifikanz und 95%-KI der fünf etablierten Prognosefaktoren

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Tumordicke	,784	,030	,000	,726-,842
Ulzeration	,679	,036	,000	,608-,750
pT AJCC 2009	,775	,030	,000	,716-,834
hist. Subtyp	,657	,034	,000	,591-,723

Invasion slevel	,701	,029	,000	,645-,758

Alle herkömmlichen Prognosefaktoren sind hochsignifikant innerhalb der ROC-Analyse (p_{1-5}<0,001). Die verhältnisskalierte Tumordicke nach Breslow (AUC 0,784), wie auch die pT Einteilung des AJCC 2009 (AUC 0,775) zeigt unter allen Prognosefaktoren das beste Ergebnis. Auch das Invasionslevel mit einer AUC von 0,701 eignet sich hervorragend als Prognosefaktor.

Nun sollen die vier, innerhalb der ROC-Analyse signifikanten morphologischen Faktoren (Zellatypie, Mitosen, Wachstumsmuster, Pigmentierung) und die herkömmlichen Faktoren mithilfe der log. Regression in jeweils zwei Modelle zusammengefasst und verglichen werden. In den nächsten Analysen wird zusätzlich immer die verhältnisskalierte

Tumordicke nach Breslow, als ein etablierter Prognosefaktor, zur Referenz angegeben. Die ggf. unterschiedlichen AUCs der Tumordicke sind u.a. auf fehlende Fälle innerhalb der anderen Parameter zurückzuführen. Vordergründig ist die Relation der anderen Modelle dazu (836 Fälle, vereinzelt fehlende; siehe Punkt 3.2):

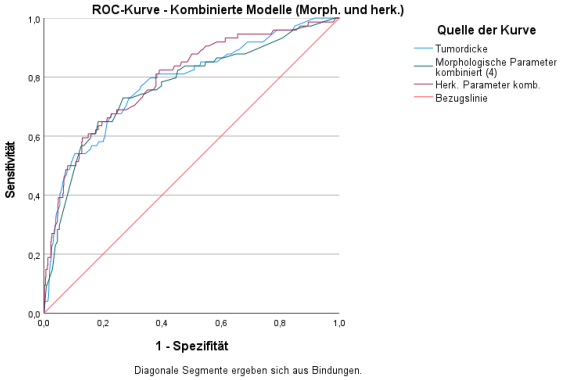

Abbildung 41: ROC-Kurve: Kombinierte morphologische Faktoren und kombinierte etablierte Faktoren mit der Tumordicke nach Breslow als Referenz.

Tabelle 55: ROC-Analyse: Kombinierte morphologische Parameter und kombinierte herkömmliche Parameter mit der Tumordicke nach Breslow als Referenz.

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Tumordicke	,779	,031	,000	,719-,839
Morphologische Parameter kombiniert (4)	,768	,032	,000	,705-,831

Herk. Parameter komb.	,794	,029	,000	,737-,851

Beide Modelle sind mit p<0,001 innerhalb der ROC-Analyse hochsignifikant. Nicht überraschend zeigt sich das Modell der herkömmlichen Faktoren mit einer AUC von 0,794 besser als das der morphologischen Parameter mit einer AUC von 0,768. Dennoch ist eine Güte des letzten genannten Modells festzustellen. Tatsächlich nähert sich deren AUC der der alleinstehenden Tumordicke nach Breslow mit 0,779 an.

Der zentrale Teil der Modellerstellung folgt in den nächsten Schritten. Einerseits sollen nun alle als signifikant herausgearbeiteten Faktoren aus den vorherigen ROC-Analysen kombiniert als Modell betrachtet werden, einmal mit der CNN-Prognose und einmal ohne. Für die

Prognoseaussage des CNNs wurde das AOI-Modell gewählt, da es in den vorherigen Punkten als überlegeneres Modell gegenüber dem ÜB-Modell hervorgebracht wurde. Die Fallzahl muss hier auf die der Validierungskohorte mit 334 verringert werden, da nur innerhalb dieser Gruppe eine Prognoseaussage des AOI-Modells erfolgen konnte. Aus Gründen einer zugänglicheren Analyse soll in den ersten Schritten nur die kategoriale Zuordnung in „dead“ oder „alive“ des AOI-Modells betrachtet werden. Die prozentuale Zuordnung wird zunächst vernachlässigt. Vereinzelt fehlende Fälle innerhalb der Parameter können in der deskriptiven Statistik gefunden werden:

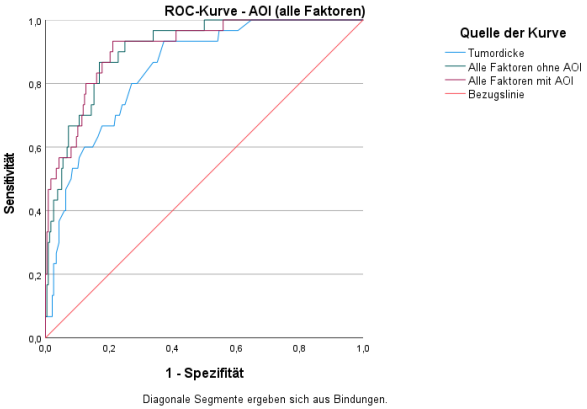

Abbildung 42: ROC-Kurve: Alle als signifikant bewerteten morphologischen und etablierten Prognosefaktoren ohne die AOI-Prognose „kategorial“ kombiniert und alle signifikant erhobenen morphologischen Parameter mit der AOI-Prognose „kategorial“ kombiniert. Die Tumordicke nach Breslow ist als Referenz angegeben.

Tabelle 56: ROC-Analyse: Alle signifikant erhobenen morphologischen und etablierten Prognosefaktoren ohne die AOI-Prognose „kategorial“ kombiniert und alle signifikant erhobenen etablierten und morphologischen Parameter mit der AOI-Prognose „kategorial“ kombiniert. Die Tumordicke nach Breslow dient als Referenz.

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Tumordicke	,845	,033	,000	,781-,909
Alle Faktore	,908	,024	,000	,862-,955

n ohne AOI				
Alle Faktoren mit AOI	,912	,025	,000	,863-,962

Die vier signifikanten morphologischen Faktoren (Zellatypie, Mitosen, Wachstumsmuster, Pigmentierung) und die herkömmlichen Faktoren (Tumordicke nach Breslow verhältnisskaliert, pT-Einteilung AJCC 2009, Subtyp, Invasionslevel, Ulzeration) wurden in ein einziges Modell zusammengefasst. Ein weiteres wurde zusätzlich mit der AOI-Vorhersage erstellt. Beide Modelle zeigen hochsignifikante p-Werte <0,001 innerhalb der ROC-Analyse. Mit einer AUC von >0,9 lassen sich zusätzlich beide Modelle als hervorragend feststellen. Das Modell mit der AOI-Vorhersage lässt sich

gegenüber dem, das die Vorhersage nicht enthält, als überlegen festhalten (AUC 0,912 vs. 0,908).

Im klinischen Alltag erfolgt jedoch oftmals nur eine Erhebung der herkömmlichen Faktoren, weshalb sie in einem vorletzten Schritt ohne die morphologischen Parameter analysiert werden (334 Fälle, vereinzelt fehlende; siehe Punkt 3.2):

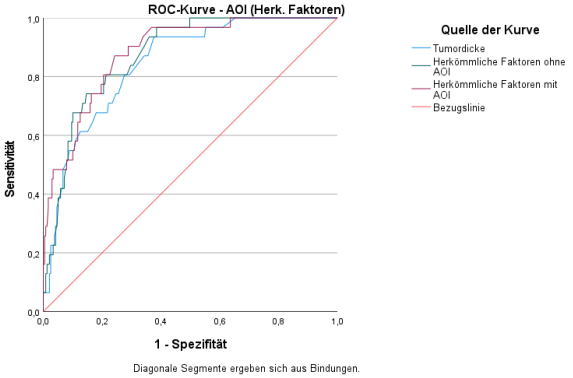

Abbildung 43: ROC-Kurve: Alle fünf etablierten Prognosefaktoren ohne die AOI-Prognose „kategorial“ kombiniert und alle fünf etablierten Prognosefaktoren mit der AOI-Prognose „kategorial“ kombiniert. Die Tumordicke nach Breslow dient als Referenz.

Tabelle 57: ROC-Kurve: Alle fünf etablierten Prognosefaktoren ohne die AOI-Prognose „kategorial“ kombiniert und alle fünf etablierten Prognosefaktoren mit der AOI-Prognose „kategorial“ kombiniert. Die Tumordicke nach Breslow dient als Referenz.

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Tumordicke	,845	,032	,000	,782-,909
Herkömmliche Faktoren	,873	,027	,000	,820-,926

ohne AOI				
Herköm mliche Faktoren mit AOI	,884	,028	,000	,829-,939

Die Ergebnisse decken sich mit denen der vorherigen ROC-Analyse. Auch hier sind beide Modelle mit $p<0,001$ hochsignifikant innerhalb der Analyse. Das Modell, das die AOI-Vorhersage miteinschließt, bleibt gegenüber dem, das die Vorhersage nicht beinhaltet, überlegen.

Abschließend ist noch zu klären, inwieweit die Prozentangabe der CNN in einer Modellerstellung miteinbezogen werden kann. Einerseits soll sie noch einmal isoliert mit allen erarbeiteten, signifikanten Faktoren betrachtet werden und mit der kategorialen Zuordnung

verglichen werden. Zusätzlich sollen beide kombiniert mit allen weiteren Faktoren in einem Modell zusammenkommen.

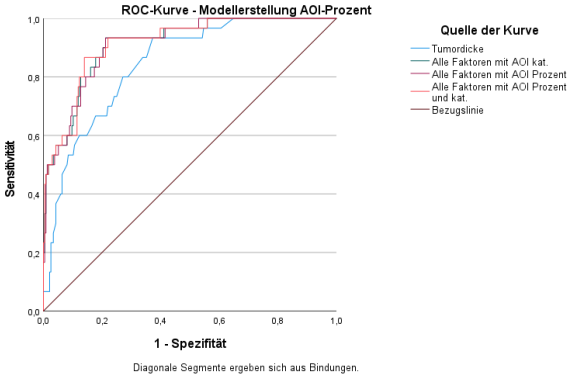

Abbildung 44: ROC-Kurve: Das erste Modell mit allen signifikanten etablierten und morphologischen Faktoren und der kategoriellen Einteilung des AOI-Modells; das zweite Modell mit allen signifikanten etablierten und morphologischen Faktoren und der prozentualen Einteilung des AOI-Modells; das dritte Modell mit allen signifikanten etablierten und morphologischen Faktoren und sowohl der kategoriellen als auch der prozentualen Einteilung. Die Tumordicke nach Breslow dient als Referenz.

Tabelle 58: ROC-Analyse: Das erste Modell mit allen signifikanten etablierten und morphologischen Faktoren und der kategoriellen Einteilung des AOI-Modells; das zweite Modell mit allen signifikanten etablierten und morphologischen Faktoren und der prozentualen Einteilung des AOI-Modells; das dritte Modell mit allen signifikanten etablierten und morphologischen Faktoren und sowohl der kategoriellen als auch der prozentualen Einteilung. Die Tumordicke nach Breslow dient als Referenz.

Modell	Fläche unter der Kurve	Std.-Fehler	Asymptotische Signifikanz	Asymptotisches 95% Konfidenzintervall
Tumordicke	,845	,033	,000	,781-,909
Alle Faktoren mit AOI kat.	,912	,025	,000	,863-,962
Alle Faktoren mit AOI Prozent	,912	,025	,000	,863-,960
Alle Faktore	,915	,025	,000	,866-,964

n mit AOI Prozent und kat.				

Die Modelle sind hochsignifikant mit p-Werten <0,001. Bis auf minimale Unterschiede an der Obergrenze des 95%-Konfidenzintervalls zeigen sich das Modell mit der kat. AOI-Zuordnung und allen Faktoren und die Prozentzuordnung und allen Faktoren nicht verschieden (AUC bei beiden 0,912). Nimmt man sowohl die kat. Zuordnung als auch die Prozentzuordnung mit allen weiteren Faktoren in ein Modell, erhält man aber das beste erstellte Modell innerhalb dieses Projekts (p<0,001) mit einer AUC von 0,915 und einem 95%-Konfidenzintervall von 0,866-0,964.

Es ist festzuhalten, dass damit das wesentliche Ziel dieses Projekts, nämlich ein Modell zu

erstellen, das den herkömmlichen Prognoseparametern überlegen ist, erfolgreich war.

4 Diskussion

4.1 Deskriptive Statistik

Nachfolgend sollen die erhobenen Ergebnisse der deskriptiven Statistik diskutiert werden.

Bzgl. der Geschlechterverteilung zeigte sich ein Verhältnis von männlich:weiblich bei 55,45%:44,55%. Diese Verteilung deckt sich auch mit der im epidemiologischen Krebsregister 2017 angegebenen Inzidenz in Deutschland im Verhältnis von 52,53%:47,47% bzw. 12.022:10.863, d.h. mit einem geringgradig höheren Anteil an Männern, die am malignen Melanom erkranken [40].

Das Erkrankungsalter des analysierten Patientenkollektivs in diesem Projekt entspricht dem aus größeren Beobachtungsstudien und retrospektiven Analysen. In Ali et al. 2013 wird erwähnt, dass der Großteil der Patienten über 65 diagnostiziert wird - mit einem Altersmedian von 57 [41, 42]. Ähnlich war die Altersverteilung

auch in diesem Projekt. Hier war die Hälfte der diagnostizierten Patienten über 62.

Die meisten Erstdiagnosen wurden im Jahr 2012 erhoben. Dies ist als positiv zu werten. Über diesen längeren Beobachtungszeitraum kann auch ein besseres Ergebnis geliefert werden. Einerseits ist er nicht zu lange, wodurch es zu einer Verzerrung des Gesamtüberlebens und follow-ups kommt. Andererseits ist der Zeitraum auch nicht zu kurz, wodurch ein aussagekräftiges follow-up bzw. „events“ innerhalb des OS möglich ist. Ein Verdacht, dass im Jahr 2012 vergleichsweise mehr Menschen die Erstdiagnose „Malignes Melanom“ erhalten haben, bestätigt sich nicht. Vielmehr noch ist der Trend im Vergleich zu den anderen betrachteten Jahren der ED (2013, 2014 und 2015) steigend, d.h. die Inzidenz im Jahr 2012 war am niedrigsten [43–45].

Die Häufigkeiten der Subtypen innerhalb der

Literatur bezüglich verschiedener Populationsstudien variieren sehr stark. Das SSM lag in dieser Arbeit bei ca. 60% aller Melanome. Wissenschaftliche Arbeiten geben oftmals sogar höhere Anteile an, mit 65-70% [46, 47].

Die cytomorphologische Einteilung in „epitheloidzellig“ und „spindelzellig“ ist erfolgt. Die oftmals angegebenen Häufigkeiten des spindelzelligen Melanoms scheinen zu variieren mit Angaben von 3-14%, jedoch kann eine zu große Divergenz davon mit 16,03% nicht festgestellt werden [48–51].
Bezüglich der Zellatypie sind vergleichbare Daten innerhalb der Literatur derzeit nicht auffindbar.

In Ghasemi et al. 2018 wurde die Mitoserate analysiert, und zwar nicht wie in dieser Arbeit in

„fehlend“, „einzeln“ und „viele“ eingeteilt, sondern vielmehr wurde betrachtet, wie viele Mitosen pro mm^2 zu sehen sind. Auf dieser Basis wurden die drei Kategorien „0-1 Mitose“, „2-3 Mitosen“ und „≥4 Mitosen“ gebildet [52]. Die 50 betrachteten Patienten zeigten eine Verteilung in Kategorie 1-3 von 40%, 42% und 18%. Zum Vergleich liegt in dieser Arbeit die Verteilung bei 44,98%, 30,26% und 24,76%. Obwohl es sich hier um andere Kategorien handelt, scheint sich zumindest eine gewisse Ähnlichkeit der Verteilung abzuzeichnen, wobei bei Ghasemi et al. die zweite Kategorie am häufigsten vertreten ist, jedoch mit einer gesamten Patientenanzahl von nur 50 [52].

Bezüglich des Entzündungsinfiltrates und der Häufigkeiten seiner Kategorien kann nur schwierig eine Diskussion stattfinden, da sich zumindest in dieser überbegrifflichen Form keine Literaturangaben finden lassen. Hierfür

hätte eine weitere Aufteilung der genaueren Bestandteile des Infiltrats stattfinden müssen.
Auch das Wiederfinden des Wachstumsmusters, wie es innerhalb dieser Arbeit erhoben wurde, ist in der Literatur nicht möglich.
Der wissenschaftliche Diskurs bezüglich der Pigmentation maligner Melanome ist dafür gegenwärtiger. In einem wissenschaftlichen Kommentar von Halaban wird die Bedeutung der Pigmentierung des Melanoms im Rahmen der Tyrosinase erläutert, nämlich, dass die Pigmentierung mit der Tyrosinaseaktivität zusammenhängt. Bei 20% der getesteten Zellreihen konnte keine Tyrosinase-mRNA nachgewiesen werden. Vielmehr noch tritt eben dieser Verlust deutlich häufiger bei spindelzelligen Melanomen im Gegensatz zu epitheloidzelligen auf [53–55]. Inwieweit sich diese Häufigkeiten auf die drei erhobenen

Kategorien der Pigmentierung hier übertragen lassen, sei dahingestellt. Allerdings sind die Zusammenhänge doch interessant, vor allem kann mit den hier erhobenen morphologischen Parametern in einer zukünftigen Arbeit der Zusammenhang zwischen der Zytomorphologie und der Pigmentierung dargestellt werden.
Insgesamt lassen sich bezüglich der deskriptiven Statistikanalyse der morphologischen Parameter nur schwierig vergleichbare Ergebnisse im Rahmen der Literaturrecherche wiederfinden. Jedoch wurde in dieser Arbeit diesbezüglich ein wesentlicher Grundstein gelegt und zukünftige Erarbeitungen könnten dann mit dieser Arbeit verglichen werden.

Eine Analyse bezüglich der Tumordicke nach Breslow von 1610 Patienten i.R. eines Projekts,

das den Zusammenhang zwischen Tumorgröße und Breslow-Dicke verstehen will, zeigt einen Mittelwert der Tumordicke nach Breslow von 2 mm [56]. Der hier erhobene Mittelwert der Tumordicke war bei 1,9 mm und damit fast identisch mit dem der erwähnten Arbeit. Auch die Klassen werden innerhalb einer Population oftmals in ihrer Häufigkeit ähnlich wie hier angegeben, d.h. am häufigsten findet sich die erste Klasse wieder und die anderen Klassen sind abnehmend in ihrer Häufigkeit [14, 57].

Innerhalb dieser Dissertation wurde die Tumordicke zahlreich als der „beste“ Prognosefaktor festgehalten, allerdings wurde noch nicht diskutiert, worauf dieser Zusammenhang zurückgeführt werden könnte. Warum hat die Dicke des Melanoms - nicht festgemacht an Schichten oder ähnlichen Marken innerhalb der Haut - einen so

erheblichen Effekt auf das Überleben? In Rousseau et al. 2003 wird der Zusammenhang positiver Wächterlymphknoten mit der Tumordicke und der Ulzerationen betrachtet. Hierbei konnte nachgewiesen werden, dass eine größere Tumordicke und damit auch eine tiefere Invasion mit positiven Lymphknoten einhergeht. Damit geht auch ein schlechteres Überleben einher [57]. Das schlechte Überleben bei einer zunehmenden Tumordicke kann nicht allein an dieser Tatsache festgemacht werden, jedoch findet sich hier zumindest ein Aspekt, der dazu beiträgt. Aber auch das Invasionslevel und das Vorhandensein einer Ulzeration zeigten sich innerhalb dieser Studie univariat analysiert als signifikante Prädiktoren positiver Wächterlymphknoten. Dazu hatte in der betrachteten Studienkohorte von Rousseau et al. 21% eine Ulzeration am Tumor. In dieser

Arbeit hatten 21,29% eine Ulzeration und damit sind die Ergebnisse wieder fast identisch [57].
Um einen Überblick über die Verteilung der Invasionslevel zu erhalten, eignet sich die Arbeit „Malignes Melanom, Invasionstiefe und Melanomtyp“ von Hermanek et al. hervorragend [58]. Hierin wird u.a. zwischen diversen Arbeiten die Häufigkeit des vorkommenden Invasionslevels verglichen, damals auch als Mikrostadium bezeichnet. In fünf der neun betrachteten Gruppen (Projekt Erlangen drei Mal gelistet, jedoch in unterschiedlichen Lokalisationen angegeben) ist das Invasionslevel IV, das am meisten auftretende und in allen Arbeiten ist es unter den häufigsten beiden Leveln [14, 20, 58–62]. Auch hier zeigen die Ergebnisse der statistischen Analyse dieser Arbeit keine zu große Abweichung.

Im Prozess der zufälligen Verteilung der 836 Schnitte auf die Trainings- und Validierungskohorte wurde festgehalten, dass keine wesentlichen Abweichungen aufgetreten sind. Eine Auffälligkeit lässt sich jedoch wiederfinden: Die Kategorien mit der schlechteren Prognose innerhalb der bestimmten Faktoren (Ulzeration, Pigmentierung, Invasionslevel etc.) finden sich oftmals geringgradig stärker in der Validierungskohorte vertreten. Ein möglicher Effekt dieser Tatsache ist, dass hierdurch möglicherweise das Training anhand der „besseren“ Kategorien stattgefunden hat und damit auch die Prognose der KI (nur beim ÜB-Modell) grundsätzlich besser ausgefallen ist. Die Unterschiede in den Kategorien sind jedoch so gering, dass ein erheblicher Effekt nicht aufgetreten sein sollte.

Zusammenfassend lässt sich sagen, dass die 836 Melanomerkrankten in ihren weiteren Eigenschaften keine größeren Abweichungen von vergleichbaren Häufigkeiten zeigen, die in anderen Studien erhoben wurden. Damit wurde in dieser Arbeit eine repräsentative Patientengesamtheit betrachtet. Bezüglich der morphologischen Parameter ließen sich oftmals nur schwierig Vergleichswerte im Rahmen einer Literaturrecherche wiederfinden.

4.2 Überlebensanalyse

Nachdem in der Einleitung diverse wissenschaftliche Arbeiten aufgeführt wurden, die Prognose und Überleben verschiedenster Faktoren thematisieren, war ein weiteres wichtiges Ziel dieser Arbeit, zu betrachten, inwieweit diese Ergebnisse bestätigt werden können. Vielmehr noch erfolgte eine Überlebensanalyse der morphologischen Faktoren, die im wissenschaftlichen Diskurs nicht sehr präsent sind, um zu betrachten, inwieweit diese Faktoren relevant sein könnten, und um diesen ggf. vorhandenen Einfluss zu diskutieren.

Die Kaplan-Meier-Schätzmethode der Tumordicke nach Breslow lässt ein deutliches Ergebnis herausarbeiten. Mit zunehmender Tumordicke, bzw. zunehmender Klasse innerhalb der pT-Einteilung nach AJCC 2009 verschlechtert sich das Überleben deutlich. Vor

allem interessant ist das Überleben der Klasse IV gegenüber den anderen drei Klassen, da es bedeutend schlechter scheint. Allerdings sollte ein wichtiger Punkt aufgeführt werden, der von Baade et al. 2020 in „long-term deaths from melanoma according to tumor thickness at diagnosis“ erarbeitet wurde. Hierin wurde i.R. einer retrospektiven Studie die Tumordicke nach Breslow und das Überleben von 44.531 Patienten, die an einem singulären, invasiven Melanom erkrankt sind, analysiert. Baade et al. präsentierten in ihren Ergebnissen, dass bei Tumordicken <1,0 mm die Fallsterblichkeit im 5-20 Jahre follow-up höher war als im follow-up <5 Jahre. Gegenteilig war dies bei „dicken“ Melanomen (>4,0 mm). [63]

Innerhalb der in dieser Arbeit angefertigten Kaplan-Meier-Kurven bewegt man sich bevorzugt innerhalb der ersten 5 Jahre, weshalb ein Anstieg bei den ersten drei

Klassen nach dem Ende der Kurve durchaus möglich wäre. Letztlich bleibt aber eindeutig, dass die Klasse IV insgesamt die schlechtere Prognose behält [14].

Eine vorhandene Ulzeration geht ebenfalls mit einem schlechteren Überleben einher. Es verbleibt die Frage, wieso dieser Zusammenhang besteht. In Callender et al. 2011 [64] wird eine Studie von Ellerhorst et. al 2011 [65] diskutiert. Hierin betrachtete man das Vorhandensein von NRAS und BRAF-Mutationen. Die Arbeitsgruppe Ellerhorst et al. fand heraus, dass NRAS-Mutationen grundsätzlich mit einer höheren Tumordicke einhergehen, BRAF-Mutationen eher mit dem Vorhandensein einer Ulzeration. Interessanterweise zeigte sich kein Unterschied bzgl. des Überlebens zwischen beiden Gruppen, evtl. bedingt dadurch, dass beide mit ihren eigenen

negativen Prognosefaktoren einhergehen.
Das Invasionslevel wurde in nur wenigen wissenschaftlichen Arbeiten als vordergründiger Prognosefaktor genannt und ist auch in der TNM-Klassifikation des AJCC nicht mehr enthalten [66]. Zwar zeigt sich in den Ergebnissen dieses Projekts ein geringerer Unterschied der Level II-IV untereinander wie im Vergleich zum fünften Level, jedoch sind die Ergebnisse eindeutig: Mit zunehmendem Invasionslevel nach Clark verschlechtert sich das Gesamtüberleben. Deshalb fiel auch die Entscheidung, das Invasionslevel innerhalb der Modellerstellung beizubehalten.
Bezüglich des Subtyps konnte, wie auch in der Einleitung erwähnt, das NM als schlechterer Prognosefaktor in der Kaplan-Meier-Kurve herausgearbeitet werden [27]. Interessanterweise geht innerhalb dieser Überlebensanalyse das ALM mit einem ähnlich

schlechten Gesamtüberleben einher. In einer Arbeit von Goydos et. al 2016 wird erwähnt, dass in diversen Studien das Überleben des ALM analysiert wurde [67, 68]. Hierbei konnte eine insgesamt schlechtere 5- und 10-Jahresprognose des ALM gegenüber dem SSM und NM nachgewiesen werden.
Zusammenfassend deckt sich die Überlebensanalyse der einzelnen herkömmlichen Faktoren in dieser Arbeit mit denen vorheriger wissenschaftlicher Erkenntnisse.

Interessant ist nun die Überlebensanalyse der morphologischen Faktoren, die nur sehr wenig innerhalb der Literatur diskutiert werden. Fünf der sechs morphologischen Parameter zeigen sich nach Log-Rank getestet als signifikant.

Die Cytomorphologie und deren Einteilung in

„Epitheloidzellig“ und „Spindelzellig“ hat keinen Einfluss im Rahmen dieser Arbeit auf das Gesamtüberleben. In der Literatur zeichnen sich diesbezüglich jedoch andere Erkenntnisse ab. Hier wird das Überleben des spindelzelligen Melanoms im Rahmen von oftmals Aderhautmelanomen als besser eingestuft, gegenüber dem des epitheloidzelligen Melanoms [69, 70]. Der Fokus der genannten Arbeit lag jedoch auf Aderhautmelanome und inwieweit eine Übertragung auf andere Melanome erfolgen kann, bleibt Gegenstand aktueller Diskussionen.

Die Analyse der Zellatypie bringt ein gewissermaßen zu erwartendes Ergebnis hervor. Mit zunehmender Atypie der Zellen verschlechtert sich das Gesamtüberleben. Tatsächlich lässt sich diese rationale Vermutung jedoch nur schwierig im Rahmen einer Literaturrecherche bestätigen. In einer

wissenschaftlichen Arbeit von 2003 wurde der Atypiegrad von Patienten mit Nävi angeschaut und die Korrelation mit dem Melanomrisiko. Die Arbeitsgruppe Arumi-Uria et al. erarbeitete ein 4,08-fach höheres Risiko eine Melanomvorgeschichte zu haben bei Patienten mit ausgeprägter „Atypie“ (NAD) gegenüber Patienten mit milder „Atypie“ [71]. Wenn bereits eine Atypie innerhalb der Nävi ein solch starken Effekt haben kann, lässt sich die Zellatypie innerhalb des Melanoms auch als starker Prognosefaktor vermuten. Ein eindeutiges Ergebnis ist allerdings nicht festzustellen.

In der Einleitung wurden Mitosen bzw. die Mitoserate, wie von Wisco et al. 2012 herausgearbeitet, als signifikante Prognosefaktoren genannt [12]. Dementsprechend kann diese Erkenntnis hier im Rahmen der Überlebensanalyse bestätigt werden. In Ghasemi et al 2018 wird vielmehr

noch dargestellt, dass bei höheren Clark-Leveln und T-Stadien die Mitoserate auch erhöht war. Es gibt eine signifikante, positive Korrelation zwischen der Breslow-Tumordicke und der Mitoserate, so Ghasemi et al. [52].

Weniger liefert eine Literaturrecherche bezüglich des dermalen Wachstumsmusters. Es lässt sich vermuten, dass im Rahmen des neoplastischen Entwicklungsprozesses eine höhere Zellzahl im Wachstumsmuster auch für eine aktivere Tumorentität spricht und dementsprechend für eine schlechtere Prognose. Die Kategorie „Einzelzellen“ geht mit einem schlechteren Überleben einher als die Kategorie „kleine Nester“. Eine Vermutung wäre, dass hier die Tumoraktivität in gewisser Weise „begrenzt“ ist auf die kleinen Nester. Weitere Arbeiten, die das Wachstumsmuster des malignen Melanoms der Haut analysieren, bleibt abzuwarten.

Deutlich mehr scheint über das Vorhandensein bzw. Fehlen einer Pigmentierung des malignen Melanoms bekannt zu sein. Eine Arbeit über Konjunktivalmelanome von Brouwer et al. 2018 ergab, dass eine geringe Tumorpigmentierung mit mehr Metastasen, häufigerem melanomassoziiertem Tod und Rezidiven einhergeht [72]. Inwieweit sich diese Erkenntnis von Konjunktivalmelanomen übertragen lässt, sei dahingestellt. Die Überlebensanalyse dieser Dissertation konnte ein signifikant schlechteres Überleben bei Tumoren mit fehlender Pigmentierung nachweisen. Klar ist, dass es Aufgabe der Melanozyten ist, Melanin zu produzieren und damit zu einer Pigmentierung beizutragen. Folglich lässt auch eine fehlende Pigmentierung eine so stärke Veränderung der Melanozyten vermuten, dass sie ihrer Aufgabe nicht mehr nachkommen, womit auch die schlechtere Prognose bei starker Entartung der

Melanozyten zusätzlich begründet werden kann. Ein weiterer möglicher Grund wäre, dass dem Betrachter vor allem pigmentierte Melanome imponieren. Die Folge wäre eine spätere Erkennung der Melanome mit fehlender Pigmentierung und damit auch eine schlechtere Prognose.

In Sarna et al. 2019 wurden unter anderem Mäuse mit Melanomzellen verschiedener Pigmentierungsgrade inokuliert. In diesem in vivo Experiment konnte beobachtet werden, dass mäßiggradig pigmentierte Tumoren ein deutlich höheres Metastasierungspotenzial gegenüber hochgradig pigmentierten Tumoren zeigten [73]. Diese vermutete Schutzfunktion des Melanins außerhalb des Schutzes gegenüber der UV-Strahlung könnte auch ein möglicher Grund des schlechteren Survivals bei fehlender Pigmentierung im Rahmen des malignen Melanoms des Menschen sein und

die Ergebnisse dieser Arbeit mitbegründen.
Die Deutung der Ergebnisse der Kaplan-Meier-Schätzmethode des Entzündungsinfiltrats zeigen sich als eher schwierig. Ein Fazit, dass mit z.B. zunehmendem oder abnehmendem Entzündungsinfiltrat das Überleben schlechter wird, lässt sich aus der Analyse nicht schließen. Vielmehr zeigt sich, dass ein leicht bis mäßiges Entzündungsinfiltrat mit dem schlechtesten Gesamtüberleben einhergeht und dass gar kein oder ein starkes Entzündungsinfiltrat ein besseres Überleben haben. Das starke Entzündungsinfiltrat hat das beste Gesamtüberleben gegenüber den anderen Kategorien.
In „the role of inflammation in skin cancer“ von Maru et al. 2014 wird der Entzündungsprozess im Rahmen von verschiedenen Hautkrebsen betrachtet. Man beachte, dass es hier nicht nur um Melanome geht. Allerdings wurde

herausgearbeitet, dass es einen starken Verdacht bezüglich eines Einflusses des Entzündungsprozesses im Rahmen der Tumorgenese gibt [74]. Ein Übertragen der Ergebnisse ist verständlich nicht möglich. Bei Maru et al. wurden verschiedenste Hautkrebsarten betrachtet und es ging vordergründig um die Tumorgenese, im Rahmen dieser Arbeit ging es hauptsächlich um kutane Melanome und das Überleben. Dennoch ist ein Vergleich interessant, denn diese Arbeit zeigt ein starkes Entzündungsinfiltrat als positiven Faktor bezüglich des Gesamtüberlebens, wohingegen Maru et al. den Entzündungsprozess als Faktor in der Tumorgenese verdächtigt.

Eine weitere aufschlussreiche Arbeit kam von Pan et al. 2018. Hierin wurden Entzündungsmarker betrachtet und inwieweit sie sich verändern bei einem Progress von

metastasierten Melanompatienten. Diese Beobachtung wurde an Patienten durchgeführt, die mit PD-1 checkpoint inhibitoren wie Nivolumab und Pembrolizumab behandelt wurden. Es wurde beobachtet, dass die Zunahme der Plättchen und neutrophilen Granulozyten und die Abnahme der Leukozyten mit einem Progress einhergeht [75]. Ein Vergleich mit dieser Arbeit soll an diesem Punkt nicht erfolgen, da die Behandlung mit PD-1 checkpoint Inhibitoren den Entzündungsprozess wahrscheinlich beeinflusst hat und eine Gegenüberstellung deshalb nicht sinnvoll erscheint. Dennoch ist eine vereinzelte Betrachtung der Entzündungsmarker interessant, weshalb auch diese Arbeit hier diskutiert wurde.

Eine mögliche Vermutung, wieso sich das Gesamtüberleben des Entzündungsinfiltrats so

präsentiert, wie es hier der Fall ist, wäre, dass eine starke Entzündungsreaktion ein starkes Immunsystem des Körpers widerspiegelt und d.h. auch, dass Therapien etc. besser anschlagen und somit das Gesamtüberleben besser ist. Dahingegen steht ein komplett fehlendes Entzündungsinfiltrat ggf. für einen nur wenig aktiven und schwach progredienten Tumor und deshalb ist auch hier das Gesamtüberleben besser. Diese Vermutungen können jedoch anhand einer Literaturrecherche nicht bestätigt werden. Zukünftige Analysen könnten evtl. weiter auf die einzelnen Entzündungsmarker eingehen und auch den genauen biochemischen Prozess, der den Einfluss auf das Überleben begründet, aufarbeiten.

Schlussfolgernd lässt sich der in der Einleitung erwähnte Effekt der etablierten

Prognosefaktoren auf das Überleben in der Überlebensanalyse dieser Arbeit bestätigen. Auch zeigen sich fünf der sechs morphologischen Parameter als signifikant in der Überlebensanalyse und eine weitere Entwicklung bezüglich ihrer Relevanz in der Prognose des malignen Melanoms bleibt abzuwarten.

4.3 Regressionsanalyse der Modelle

Die Regressionsanalyse der Tumordicke nach Breslow liefert kein überraschendes Ergebnis. Als wichtigster bekannter Prognosefaktor des malignen Melanoms wäre ein nicht signifikanter Einfluss auf die Prognose der Modelle sehr überraschend gewesen. Interessanterweise lässt sich in einer Regressionsanalyse der Tumordicke mit dem tatsächlichen OS ein stärkerer Einfluss auf das Überleben als

innerhalb der Modelle herausarbeiten. Vielmehr noch scheint die Tumordicke auf das AOI-Modell den geringeren Einfluss zu haben, dennoch ist das Ergebnis des AOI-Modells besser als das des ÜB-Modells.
Die Klassen der pT-Einteilung nach AJCC 2009 zeigen eine ähnliche Tendenz. Hier wurde innerhalb des ÜB-Modells ein deutlich höherer Einfluss der Klassen als im AOI-Modell erkannt. Tatsächlich sollte beachtet werden, dass aufgrund der wenigen Schnitte die als „dead“ innerhalb des Übersichtsbild-Modells eingestuft wurden, jeglicher Effekt als übermäßig bewertet worden sein kann.

Beim Invasionslevel zeigt das Clark-Level IV einen signifikanten Unterschied im Überleben gegenüber dem Clark-Level II. Ein wesentlicher Grund, warum das Clark-Level nicht mehr vordergründig in der Prognostik ist, wurde u.a.

von Colloby et al. erarbeitet. Hierin wurde der Unterschied in der Tumordicke nach Breslow und des Clark-Levels bei individuellen Betrachtern analysiert. Der Mittelwert der Variation der Ergebnisse wurde mit 0,25 mm bei der Tumordicke angegeben und sei damit vergleichsweise besser als beim Clark-Level. Hier gab es nämlich zwischen den Betrachtern gerade mal eine Übereinstimmung von 60%. Dies sei laut Colloby et al. „little better than random“ – also nur ein wenig besser als der Zufall bzw. eine zufällige Einteilung [76].

In beiden Modellen zeigt sich das Vorhandensein einer Ulzeration als ein sehr starker Einflussfaktor auf das Überleben und deckt sich damit mit den Erkenntnissen von Balch et al. [11].
Im Jahr 2000 gab es eine umfangreiche Revision von Seiten des AJCC bezüglich der

TNM-Klassifikation mit Einfluss der Prognosefaktoren des malignen Melanoms. An erster Stelle wird die Tumordicke mit dem Vorhandensein einer Ulzeration genannt; das Invasionslevel solle vernachlässigt werden [66]. In gewisser Weise liefert die Regressionsanalyse der CNN innerhalb dieser Arbeit dasselbe Ergebnis. Mit einer OR (innerhalb des AOI-Modells) von 2,28 der Klasse IV der Tumordicke nach Breslow und 3,15 bei der Ulzeration ist eben jener Effekt nur zu unterstreichen.
Die Regressionsanalyse der Subtypen hingegen brachten das noduläre Melanom als signifikanter negativer Prognosefaktor im ÜB-Modell hervor, was mit den Ergebnissen von Barnhill et al. 1996 übereinstimmte [13].

Da die Modelle an histologischen Bildern trainiert wurden, reichten uns die

herkömmlichen Prognosefaktoren noch nicht vollständig aus, um die unterschiedlichen Einflussfaktoren auf die Prognose der CNN zu analysieren. Unter anderem deshalb erfolgte auch die Erhebung der semiquantitativen, morphologischen Parameter. Es zeigten sich die Zellatypie, Mitosen, das Wachstumsmuster und die Pigmentierung als signifikante Einflussfaktoren. Mitosen zeigt zudem noch in der Regressionsanalyse des AOI-Modells alle Kategorien als signifikant unterschiedlich. Als Faktor, der zumindest von den morphologischen Parametern mit am häufigsten analysiert und als wichtig hervorgebracht wurde (siehe Überlebensanalyse) war jedoch dieser starke Einfluss auf die Prognoseaussage der Modelle zu erwarten.

Die Zellatypie und das Wachstumsmuster scheinen dahingegen einen unterschiedlichen,

aber untereinander vergleichbaren Einfluss zu haben. Beide Modelle zeigen sich als signifikant, wohingegen sich innerhalb der Regressionsanalyse nur die Kategorie, die mit dem schlechtesten Überleben einhergeht, signifikant ist. Etwas anders waren die Ergebnisse der Kaplan-Meier-Analyse dieser beiden Faktoren. Hier waren nach Log-Rank getestet beide Faktoren hochsignifikant und damit konnten Unterschiede bezüglich des Gesamtüberlebens der einzelnen Kategorien nachgewiesen werden.

Ein sehr eindrückliches Ergebnis der Regressionsanalyse der morphologischen Parameter lieferte jedoch die Pigmentierung. Hier zeigten sich auch die Kategorien untereinander in beiden Modellen signifikant. Die logistische Regressionsanalyse hier bestätigt zumindest die schlechtere Prognose innerhalb der Modelle.

Anhand der Regressionsanalysen wurde versucht der Prognoseaussage der beiden Google teachable machine Modelle in gewisser Weise „auf den Grund zu gehen“ und es wurde versucht zu verstehen, welche Faktoren signifikant in der Einordung der CNN in „dead“ oder „alive“ sind. Allerdings muss dem entgegengehalten werden, dass nur analysiert werden konnte, was ein Mensch unter dem Mikroskop, oder konkreter ein Dermatopathologe an einem Melanomschnitt sieht. Weitere Einflussfaktoren, die das CNN registriert, aber von Untersuchern nicht gesehen werden können, können dementsprechend nicht analysiert werden.

4.4 Google teachable machine Modelle

4.4.1 Ergebnisse dieser Arbeit

Bei der Betrachtung des 2YS und 5YS fällt auf, dass die „dead“-Gruppe der „alive“-Gruppe deutlich unterlegen ist. Um die CNN besser trainieren zu können, war ein möglichst großer Anteil beider Gruppen von unserer Arbeitsgruppe angestrebt worden. Dies konnte mit der Auswertung des Gesamtüberlebens (overall survival) gewährleistet werden. Hier war der größte Anteil an „dead“-Schnitten enthalten. Ein Kritikpunkt, der diesem Vorteil entgegensteht, ist, dass beim OS der geeignetste Zeitraum nicht präzise festgelegt werden kann.

Die beiden Ergebnisse der Modelle sind jedoch sehr interessant. Das ÜB-Modell hat eine ausgeprägte Tendenz, Schnitte als „alive“ einzuordnen. Hierfür können diverse Gründe

diskutiert werden. Einerseits wurden 431 Schnitte zum Training der alive-Gruppe vs. 71 Schnitte in der dead-Gruppe in sechsfacher Kopie verwendet. Da das Modell jedoch deutlich mehr verschiedene „alive“ Schnitte innerhalb des Trainings verwendet hat, könnte diese starke Tendenz zur Einordnung zur „alive“-Gruppe erklärt werden [9]. So könnte das Modell an den Validierungsschnitten Muster „gesehen“ haben, die es innerhalb des alive-Trainings durch die größere Diversität der Schnitte, wiedererkannt haben und dadurch mehr als „alive“ einstuft. Ein weiterer Grund, der diskutiert werden könnte, ist, dass innerhalb der Übersichtsbild-Schnitte das krankheitsbestimmende Areal einen geringeren Teil des gesamten Bildes einnimmt. Hier kommen auch gesunde Teile bei Schnitten vor, die als „dead“ hätten eingestuft werden sollen. Das Modell erkennt jedoch eben diese

gesunden Areale wieder und stuft es deshalb nicht als „dead“ ein. Auch daran könnte die übermäßige Zuordnung des ÜB-Modells zur „alive“ Gruppe erklärt werden können. Abschließend muss gesagt werden, dass bei der Komplexität der CNN kein definitiver Grund festgehalten werden kann.

Eine Trefferquote von gerade mal 20,83% innerhalb der dead-Validierungskohorte ist jedoch klar als schlecht einzuordnen, wodurch u.a. auch die nicht-statistische Signifikanz im Rahmen der ROC-Analyse zustande kam. Auch deshalb wurde dieses Modell in der Erstellung des kombinierten Modells vernachlässigt. Es bleibt die Frage, inwieweit sich das ÜB-Modell verbessern würde, wenn man den gleichen Anteil verschiedener dead- und alive-Schnitte gehabt hätte.

Einen Vorteil demgegenüber hat das AOI-Modell. Hier wurden nicht die gleichen 71 dead-

Schnitte für das Training verwendet, sondern innerhalb der dead-Schnitte jeweils sechs verschiedene krankheitsbestimmende Areale. Auch dass das krankheitsbestimmende Areal (area of interest) den Großteil des Bildes ausgemacht hat, kann dazu beigetragen haben, dass das Modell deutlich besser als das ÜB-Modell ist. Eine Einbuße der Trefferquote von ca. 25% bei der alive-Gruppe, dafür jedoch eine Verbesserung von ca. 31% innerhalb der dead-Gruppe sprechen von einem deutlich besseren Ergebnis und einer nachweisbaren Signifikanz (p_1=0,001 und p_2<0,001). Vielmehr liegt dieses Modell bei einer Einordnung von dead oder alive eher richtig als falsch (beide über 50%).

Letztlich sollte festgehalten werden, dass das AOI-Modell zwar beeindruckende Ergebnisse liefert, jedoch allein betrachtet zur Prognoseeinschätzung nicht geeignet ist.

Dasselbe gilt auch für das ÜB-Modell. Eine „area under the curve“ (AUC) innerhalb der ROC-Analyse von 0,646 ist definitiv nicht ausreichend, um Therapieschemata festzulegen. Es ist zu beobachten, welchen Weg die Google teachable machine in den kommenden Jahren geht. Derzeit können nach einem abgeschlossenen Training keine neuen Daten mehr für das Training eingelesen werden. Möglich wäre in einer neueren, zukünftigen Version ein kontinuierliches Training. Durch mehr Daten wäre es dann auch eventuell möglich, die Prognoseaussage in einem laufenden Prozess zu verbessern.

Allerdings könnte eine diversere dead-Gruppe innerhalb des Trainings auch den gegenteiligen Effekt haben. Vernachlässigt man zunächst, ob sich die Modell-Prognose mit der tatsächlichen deckt, gibt das AOI-Modell ein schlechteres

Überleben als das tatsächliche Gesamtüberleben (85,3% vs. 69,8%) her. Ein mögliches Szenario, das berücksichtigt werden sollte, wäre auch, dass mit dem gleichen Anteil an verschiedenen dead- wie alive-Schnitten die Zuordnung, ob nun richtig oder falsch, zur dead-Gruppe ansteigt und damit das vermutete Überleben des Modells zunehmend schlechter wird.

Tatsächlich ist dies jedoch nur ein mögliches Szenario und der gegenteilige Fall wäre doch wahrscheinlicher, da Modelle innerhalb eines CNNs oftmals besser mit mehr Daten werden [77]. Dies konnten Uchida et. al 2016 in einem umfangreichen Projekt bestätigen. Hierbei wurden drei große Datensets verwendet. Das erste Datenset bestand aus 822.714 handgeschriebenen Formularen. Das zweite Datenset waren 622.678 maschinengedruckte Formulare in zwei Schriftarten und das letzte

Datenset setzte sich aus 66.470 Bildern zusammen, die aus 6.647 verschiedenen Schriftarten generiert wurden. Die CNN wurde immer mit 90% der Daten trainiert und sollte die restlichen 10% validieren. Es wurden fast perfekte Ergebnisse erzielt. Im ersten Datenset konnte eine Genauigkeit von 99,88% erreicht werden. Im zweiten waren es sogar 99,99%. Für das dritte Datenset kam eine Genauigkeit von 95,7% heraus [77].

Die Prozentangaben der beiden Modelle zeigen sich ebenfalls als sehr unterschiedlich. Das ÜB-Modell neigt grundsätzlich dazu, Schnitte als 100% „alive“ (0%) einzuordnen, und scheint nur sehr selten von dieser 100% „alive“ Einordnung zu divergieren. Dahingegen ist das AOI-Modell deutlich anders. Auch hier besteht eine Neigung zu den Extremen, jedoch finden sich hier oftmals auch andere Einordnungen in

Prozent zu der jeweiligen Kategorie.

Es lohnt sich, an dieser Stelle den Begriff des „overfitting“ (Überanpassung) heranzuziehen. Mutasa et al. definieren diesen Begriff als ein Szenario, in dem ein KI-Modell auf so eine Weise anhand eines Datensets trainiert wurde, dass eine Testung oder Validation nur auf demselben Datenset anwendbar ist und eine Übertragung auf weitere Daten nicht mehr gegeben ist. Die Arbeitsgruppe Mutasa et al. konkretisieren dies mit einem Beispiel: Ein KI-Modell, dass zwischen Hund und Katze unterscheiden soll aber nur mit Bildern von Schäferhunden und Siamkatzen trainiert wird, wird auch nur in der Testung dieser Arten ein gutes Ergebnis liefern. Bei der Testung anderer Arten wird das Ergebnis deutlich schlechter sein. Die Lösung dieses Problem, sei die Vergrößerung des Trainingsdatensets [9, 78].

Dies lässt innerhalb dieser Arbeit auf den

möglichen Grund schließen, dass die sechsfache Kopie beim ÜB-Modell vs. sechs verschiedene Ausschnitte beim AOI-Modell für diversere Prozentzuordnungen beim AOI-Modell sorgte. Die ROC-Analyse zeigt bei beiden Modellen die Prozentzuordnung als überlegen. Vielmehr noch scheint dies beim AOI-Modell deutlich betonter zu sein. Konkreter hat die AOI-Prozentzuordnung eine Zunahme der AUC von 0,035 gegenüber der kategorialen Zuordnung wohingegen beim ÜB-Modell die Zunahme lediglich bei 0,009 liegt. Dies kann auf die diversere Verteilung des AOI-Modells auf die einzelnen Prozentränge zurückgeführt werden. Will man sich also dem realen Outcome mit einer Prognoseaussage anhand der erstellten Modelle nähern, eignet sich die Prozentzuordnung des AOI-Modells, wenn man nur einen einzelnen Parameter betrachtet, am besten.

4.4.2 Wissenschaftlicher Stand

Zentrale Themen dieser Dissertation sind das maligne Melanom und die CNN. Eine in dieser Hinsicht maßgebende Arbeit, die definitiv diskutiert werden sollte, ist die von Hekler et al. 2019 „Pathologist-level classification of histopathological melanoma images with deep neural networks“. Deren Arbeitsgruppe sammelten 695 Läsionen und ließen sie von einem erfahrenen Pathologen klassifizieren in Nävus oder Melanom (350 Nävi vs. 345 Melanome). Anschließend wurde, wie auch in diesem Projekt, eine CNN trainiert mit dem Ziel der anschließenden Validierung. 595 Schnitte wurden für das Training verwendet, 100 zur Validierung. Es zeigte sich eine fehlende Übereinstimmung mit dem menschlichen Pathologen bei insgesamt 19% (18% Melanom und 20% Nävi). Vielmehr noch ist eine ähnliche fehlende Übereinstimmung zwischen

menschlichen Pathologen untereinander oftmals festzustellen, so Hekler et al. [79].

Die Fragestellung von Hekler et al. bzw. das zentrale Thema ist diagnostischer Natur. Da in dieser Arbeit eine Fragestellung bzgl. der Prognose erfolgt ist, sind die Ergebnisse von Hekler et al., die diese anderen Aspekte hervorbringen sehr interessant, jedoch nicht zu vergleichen. In einer zweiten Publikation von Hekler 2019 et al. wurde die Sensitivität, Spezifizität und Genauigkeit der CNN mit elf Histopathologen verglichen [80]. Das Datenset war wieder dasselbe der vorherigen Arbeit (350 Nävi und 345 Melanome) mit 100 zu validierenden Bildern. Die CNN hatte eine durchschnittliche Sensitivität/Spezifizität/Genauigkeit von 76%/60%/68% gegenüber den elf Pathologen, die im Schnitt bei 51.8%/66.5%/59.2% lagen. Damit war die CNN signifikant besser bei der

Einordnung. Einen möglichen Punkt den Hekler et al. diskutieren im Hinblick auf warum die CNN besser ist, sei die Fähigkeit des CNN sogenannte „sub-visual image features“ zu erkennen, also Eigenschaften des Schnittes, die für die Augen der Histopathologen nicht zu erkennen sind [80, 81].

Das Projekt mit den wahrscheinlich meisten Ähnlichkeiten zu diesem findet sich bei Kulkarni et al. 2019. Der Fokus lag hier bei der Erstellung eines Biomarkes bzw. Risikoklassifikators für das disease-specific survival (DSS) des malignen Melanoms, vor allem aufgrund der Notwendigkeit, geeignete Patienten für adjuvante Immuntherapien zu identifizieren. Das Modell wurde an 108 Patienten trainiert und an zwei Validierungskohorten mit je 104 und 51 Patienten getestet. Zusätzlich wurde anhand

der „distant metastatic recurrence“ (DMR), definiert als das Auftreten jenseits der lokalen Lymphknotenpools, ein weiterer binärer Klassifikator hinzugenommen. Es wurden auch hier „areas of interest“, genannt „regions of interest“, ausgewertet, jedoch wurden diese nicht über einen Dermatopathologen, sondern über einen Softwarealgorithmus identifiziert. In der ersten Testkohorte konnte eine starke Korrelation des Biomarkers mit dem DMR nachgewiesen werden (AUC 0,905). Auch die zweite Testkohorte zeigte ähnliche Ergebnisse (AUC 0,880) [9, 82].

Ein vorsichtiger Vergleich erlaubt einige interessante Aspekte zu diskutieren. Bei Kulkarni et al. wurde das DSS bzw. DMR betrachtet. Innerhalb dieser Arbeit wollten wir den Fokus auf das OS setzen, da zwar auch unser Ziel wie bei Kulkarni et al. die Erstellung eines zusätzlichen Biomarkers war, aber auch

das sogenannte „proof of concept“, d.h. zu beweisen, dass es in der Tat möglich ist mit einer CNN unter geeigneten Voraussetzungen eine Prognose abzugeben. Bei Kulkarni et al. gab es weiterhin wahrscheinlich eine umfangreichere Vorauswahl bei der beispielsweise Schnitte mit Objektträgerbrüchen oder zu starkem Melaningehalt aussortiert wurden. Auch wurde innerhalb der Softwarebestimmung der „regions of interest“ vom Programm selbst ein Teil der Schnitte aussortiert. Ein letzter, interessanter Diskussionspunkt bei Kulkarni et al. ist die KI. Hier wurde nämlich ein „recurrent neural network“ (RNN) mit einem CNN kombiniert. Ein wesentlicher Aspekt, der RNNs ausmacht, ist, dass zu jedem Zeitpunkt der vorherige Input die weitere Vorhersage der KI mitbestimmt [83]. Weiterhin wurde versucht dem bereits diskutierten „overfitting“

entgegenzuwirken mit einer dropout procedure, d.h. es wurde zufällig ein bestimmter prozentualer Anteil des Inputs innerhalb einer Ebene des deep neural networks (DNN) auf 0 gesetzt [9, 82, 84].

Eine weitere vergleichende Arbeit mit dem Fokus auf Prognose und CNN lässt sich mit „deep learning-based classification of mesothelioma improves prediction of patient outcome“ von Courtiol et al. 2019 heranziehen [85]. Die Zielsetzung im Rahmen dieser Arbeit ähnelte der dieser. Es wurde der Fokus auf die Prognose gelegt und ob die CNN evtl. auch Aspekte berücksichtigt, die bei Pathologen keine Relevanz finden. 2.981 Mesotheliomschnitte wurden verwendet, 2.300 für das Training und 681 zur Validierung. Die Arbeitsgruppe von Courtiol et al. erstellten mit den Schnitten und einem vorhandenen

Framework ein neues CNN-Modell - genannt MesoNet. Das Modell war in der Lage eine signifikant bessere Prognose abzugeben, verglichen mit herkömmlichen pathologischen Verfahren. Vielmehr war überraschend, dass der Subtyp des Mesothelioms, der mit der schlechtesten Prognose assoziiert ist, nur sehr gering in der Gruppe mit schlechter Prognose vertreten war. Dies veranlasste die Arbeitsgruppe zu weiteren Analysen und es konnte entdeckt werden, dass ein stark berücksichtigter Aspekt das Stroma mit Entzündungsprozessen und ähnlichem war. Derartiges fand im bisherigen Prozess zur Prognosebestimmung keine große Relevanz [85].

Auch hier konnte im Vergleich mit diesem Projekt ein größerer Datensatz verwendet werden, der zu einer besseren Prognoseabgabe geführt hat. Innerhalb der

logistischen Regressionsanalyse wurde in dieser Arbeit versucht die Prognosezuordnung der CNN zu verstehen. Letztlich konnten aber nur etablierte oder bekannte Faktoren berücksichtigt werden. Der Arbeitsgruppe von Courtiol et al. ist es allerdings gelungen einen völlig neuen Parameter bzw. Areal in Bezug auf die Prognose der CNN festzumachen. Eine derartige Analyse wäre auch hier wieder für eine zukünftige Arbeitsgruppe möglich, allerdings mit der Voraussetzung, die Prognosefähigkeit der CNN zu verbessern, da sonst eine derartige Analyse nicht sinnvoll erscheint. Weiterhin diskutiert werden kann auch die Wahl der CNN. Bei Courtiol et al. erfolgte ein umfangreicherer Prozess bei der Auswahl und Erstellung des Modells. Wahrscheinlich hätte ein derartiger Prozess auch in diesem Projekt noch einmal zu besseren Ergebnissen geführt, jedoch lassen

die Ergebnisse der Modelle der Google teachable machine vermuten, dass lediglich mit einem großen Datensatz ähnliche Ergebnisse erzielt werden könnten.

In der Einleitung wurden die Hyperparameter der Google teachable machine erläutert. Eine umfangreiche Aufarbeitung, wie sie die Ergebnisse beeinflussen, fand durch Jeong 2020 statt [36]. Die Arbeit erfolgte anhand der Erkennung zahnmarkierter Zungen mit 1,250 Bilder. Davon waren 704 zahnmarkierte Zungen und 546 hatten diese Markierungen nicht. Eine zufällige Einteilung ist erfolgt mit einer Verteilung von Trainings- und Validierungskohorte von 90:10. Es konnten folgende Erkenntnisse bezüglich der drei Hyperparameter gewonnen werden: Die Epoche wurde von 20 auf 200 kontinuierlich verändert. Die höchste Genauigkeit konnte bei einer Epochenanzahl von etwa 75 erzielt

werden. Darunter gab es zu starke Fluktuationen und darüber gab es zwar auch gute Ergebnisse, jedoch hat die Genauigkeit um einige Prozentpunkte abgenommen. Bezüglich der Lernrate gab es eine kontinuierliche Zunahme von 0,00001 auf 0,01. Darüber war die CNN nicht in der Lage, trainiert zu werden. Die beste Genauigkeit konnte mit einer Lernrate von 0,0001 erzielt werden. Bei der Batchgröße gab es eine Zunahme von 16 auf 256. Unter 64 waren die Ergebnisse zu ungenau und bei 128 waren sie am besten [36]. Ein Vergleich sollte vorsichtig gezogen werden, da Jeong mit einem größeren Datensatz und einer unterschiedlichen Fragestellung arbeitete. Die Epochenanzahl wurde in diesem Projekt auf 1000 festgelegt, laut Jeong lag die optimale Epochenzahl in dessen Projekt bei 75. Allerdings wurden die Ergebnisse der Epochenzahl nur bis 200 getestet jedoch lässt

sich aus der Arbeit noch folgendes erkennen: Gegen Ende nimmt die Kurve wieder zu und die Genauigkeit steigt, so dass bei 200 Epochen die Ergebnisse fast identisch mit denen der 75 Epochen ist. Deshalb lässt sich auch bei einer Zunahme auf 1000 ein gutes Ergebnis erwarten. Die Batchgröße sei am besten bei 64, hier wurde sie auf 16 gesetzt. Auch hier lässt sich aus der Arbeit von Jeong wieder folgende Erkenntnisse ziehen: Bei 16 Batches kommen die Ergebnisse fast an die 64 Batches heran. Lediglich die Einstellungen dazwischen zeigen sich schlechter. Eine Änderung der Batchsize würde in diesem Projekt also zu keinem wesentlich unterschiedlichen Ergebnis führen. Die Lernrate hier mit 0,001 präsentiert sich in Jeong nur geringgradig schlechter als die Lernrate bei 0,0001 [36]. Zusammenfassend kann zumindest bezüglich der Hyperparameter festgestellt werden, dass eine Änderung die

Ergebnisse nicht wesentlich verbessert hätten.

In einem wissenschaftlichen Artikel, der im Dezember 2020 veröffentlicht wurde, wird von Wells et al. die Bedeutung der künstlichen Intelligenz in der Dermatopathologie diskutiert [86]. Ein Argumentationspunkt hierin ist, dass die meisten CNN-Modelle nur auf ein bestimmtes Krankheitsbild trainiert werden. Es sei also zu hinterfragen welche Probleme entstehen, wenn die CNN nicht über das differentialdiagnostische Potential von Dermatopathologen verfügen. Eine derartige Schwäche weist das Modell dieses Projekts nicht hervor. Hier werden nämlich nur Schnitte bezüglich ihrer Prognose beurteilt, die vorher von Dermatopathologen bezüglich ihrer Diagnose gesichert wurden. Weiterhin argumentieren Wells et al., dass die Rolle der CNN nicht die sei, Dermatopathologen zu

ersetzen, sondern lediglich sie bezüglich der Diagnose und Prognose zu unterstützen [86]. Dies kann zum derzeitigen Zeitpunkt nur unterstrichen werden. Die meisten CNN-Modelle, inkl. die Modelle innerhalb dieser Arbeit, sind noch nicht zuverlässig genug, Prognose oder Diagnose unabhängig abzugeben [9]. Eine absolute Aussage diesbezüglich sollte jedoch auch nicht getroffen werden. Die rasante Verbesserung der CNN ermöglicht es, zumindest ein Szenario zu erwägen, dass in Zukunft Diagnose und Prognose allein durch die Modelle erfolgen könnten.

4.5 Kombinierte Modelle

Die Überlebensanalyse der morphologischen Faktoren und die Analyse bzgl. ihrer Eignung als Prognosefaktoren zeigt ähnliche

Ergebnisse. Die Zellatypie, das Vorhandensein von Mitosefiguren, die Pigmentierung und das Wachstumsmuster eignen sich hierfür. Das Entzündungsinfiltrat, das in der Kaplan-Meier-Schätzmethode noch einen signifikanten Unterschied bzgl. des Überlebens innerhalb der einzelnen Klassen nachweisen ließ, ist in der ROC-Analyse interessanterweise nicht signifikant. Dem ist entgegenzuhalten, dass die Signifikanz bei 0,041 lag und damit knapp unter 0,05. Zusätzlich ist der Überbegriff Entzündungsinfiltrat ggf. etwas zu umfangreich definiert. Eventuell könne in einer zukünftigen Analyse die einzelnen Bestandteile dieser Entzündungsinfiltrate genauer kategorisiert und noch einmal analysiert werden.

Sogenannte „tumor infiltrating lymphocytes“ (TIL) gelten beispielsweise univariat und multivariat analysiert als unabhängiger, signifikanter Prognose des malignen Melanoms

mit vertikaler Wachstumsphase [87]. Noch weiter lassen sich „brisk“ von „non-brisk“ tumor infiltrating lymphocytes unterscheiden. Brisk TILs beschreibt Lymphozyten, die den Tumor direkt durchwandern oder umgeben; „non-brisk“ ist als ein fokales TIL-„Infiltrat“ innerhalb des Tumors zu definieren [88]. In einem hervorragenden systematischen Review und einer umfangreichen Metaanalyse von Fu et al. 2019 konnten beide Klassifikationen als signifikante, vorteilhafte Prognosefaktoren herausgearbeitet werden [89].

Die Nicht-Signifikanz in der ROC-Analyse dieses Projekts hat aber zu einem Ausschluss des Entzündungsinfiltrates in der weiteren Analyse geführt. Aber auch die Zytomorphologie zeigt sich in der ROC-Analyse als nicht signifikant. Damit ist dieser morphologische Parameter innerhalb der Überlebens-, Regressions- und ROC-Analyse

nicht signifikant und schlussfolgernd kann die Einteilung in „spindelzellig“ und „epitheloidzellig“ bezüglich der in dieser Dissertation beobachteten Aspekte als nicht aussagekräftig gewertet werden.

Dass sich vier der sechs Parameter, die im Rahmen dieser Arbeit erhoben wurden, als signifikant innerhalb der ROC-Analyse feststellen lassen, ist ein hervorragendes Ergebnis. Demgegenüber steht jedoch die äußert umständliche Erhebung dieser Parameter. Im Rahmen dieser Arbeit wurden sie von einem Dermatopathologen erhoben, jedoch ist fraglich ob derartiges Personal und Zeit im klinischen Alltag vorhanden ist, bei jedem Patienten diese morphologischen Faktoren zu erheben. Laut den Ergebnissen dieser Arbeit wäre es zumindest bezüglich der Prognose von Vorteil. Es bleibt auch die weitere Entwicklung der KI abzuwarten.

Möglicherweise ist eine derartige Erhebung gar nicht mehr in der Zukunft nötig, da CNNs bereits im Rahmen ihres „imaging“ derartige Parameter berücksichtigen.

Die herkömmlichen Prognoseparameter lassen sich nicht überraschend in der ROC-Analyse als geeignete Prognosefaktoren herausarbeiten. Innerhalb dieser Dissertation wurde mehrmals die Tumordicke und die Ulzeration [11] als die stärksten Prognosefaktoren genannt. Bezüglich der Tumordicke kann dieses Ergebnis vollständig bestätigt werden. Wäre es hypothetisch nötig, nur einen einzelnen Prognosefaktor zu verwenden so eignet sich die Tumordicke nach Breslow, ob nun verhältnisskaliert oder mithilfe der pT-Einteilung nach AJCC 2009, am besten. Sie zeigt sich in allen Analysen dieser Arbeit als alleinstehender Faktor überlegen gegenüber allen morphologischen wie auch

herkömmlichen Parametern.

Im Rahmen dieser Dissertation wurde bereits häufiger die Bedeutung des Invasionslevels bezüglich der Prognose diskutiert. Überblickend soll jedoch die Arbeit von Lyth et al. 2013 behandelt werden [90]. Hierin wurde mit der Zielsetzung der Risikostratifizierung des T1 Melanoms der Haut eine umfangreiche Betrachtung etablierter Prognosefaktoren durchgeführt, vordergründig die Ulzeration, Tumordicke und das Invasionslevel. Hierfür wurde das schwedische Melanomregister von 1990 bis 2008 analysiert und 13.026 Patienten betrachtet. Ulzeration, Invasionslevel und Tumordicke zeigten sich alle als signifikante Prognosefaktoren. Zusätzlich ließ sich im Rahmen diverser Modellerstellungen drei prognostische Untergruppen im Rahmen des T1-Melanoms feststellen [90].

Um also ein abschließendes Fazit bezüglich

des Invasionslevels zu ziehen, kann auf Basis der Ergebnisse dieser Arbeit das Invasionslevel innerhalb der Prognose des malignen Melanoms der Haut als bedeutender Prognosefaktor festgehalten werden. Die Ulzeration zeigte sich dennoch in den vorherigen Analysen als besser. Zwar wurde im Rahmen dieser Arbeit vordergründig nur vier der etablierten Prognosefaktoren analysiert, jedoch zeigten sich gerade die drei in Lyth et al. genannten als am aussagekräftigsten [90].

Der Subtyp hatte unter den etablierten Prognosefaktoren die niedrigste AUC. Diskutiert kann unter anderem das ähnlich schlechte Überleben des ALM und des NM bzw. ähnlich gute Überleben des SSM und LMM und damit eine fehlende Trennschärfe. Die Regressionsanalyse der Subtypen zeigte sich in nur wenigen Punkten signifikant, die Überlebensanalyse hingegen ließ einen

signifikanten Unterschied im Überleben nachweisen. Letztlich sprechen auch die Ergebnisse dieser Arbeit für eine Eignung des histologischen Subtyps als Prognosefaktor. Es muss aber auch festgehalten werden, dass die Ergebnisse nicht so eindrücklich sind wie bei der Tumordicke und Ulzeration.

Bei den ersten Modellerstellungen zeigt sich das kombinierte Modell der etablierten Prognosefaktoren, gegenüber dem der vier signifikanten morphologischen Parameter überlegen. Ein anderes Ergebnis wäre nicht zu erwarten. Allerdings zeigt sich das morphologische Faktoren-Modell ähnlich gut, wie die Tumordicke nach Breslow, wobei die Tumordicke nach Breslow geringgradig besser ist. Dies sollte als ein sehr gutes Ergebnis festgehalten werden. Ein Modell mit vier wenig erforschten Faktoren zu erstellen, das an den

Prognosefaktor schlechthin - der Tumordicke nach Breslow - herankommt ist ein positives Ergebnis. Demgegenüber steht leider wieder die klinische Anwendung im Alltag eines Mediziners. Bei einem histologischen Befund sollen die Tumordicke nach Breslow, das Invasionslevel nach Clark, das Vorhandensein einer Ulzeration und die Angabe des mitotischen Indexes angegeben werden, d.h. lediglich einer der vier signifikanten Prognosefaktoren der Morphologie wird routinemäßig erhoben [23]. Es kann durchaus argumentiert werden, dass die Erhebung jeder einzelnen der weiteren drei morphologischen Parameter umständlicher ist als die Messung der Tumordicke und diese eignet sich dann auch noch besser. Auch eignet sich die Tumordicke nach Breslow als ein reproduzierbarer Wert bei unterschiedlichen Betrachtern [76]. Die erwähnten Entwicklungen

der CNN in der Zukunft bleiben abzuwarten. Auch die Verwendung innerhalb eines hochwertigen, kombinierten Modells bleibt weiterhin als Grund zur Erhebung bestehen.
Anschließend folgte die Modellerstellungen mit der KI-Prognose und es konnte das gewünschte Ergebnis erzielt werden: Ein Prognosemodell, das verschiedenste Faktoren mit der kategorialen Einteilung der CNN beinhaltet, ist besser als ein selbiges Modell ohne diese Einteilung. Vielmehr noch ist auf dem hohen Level, auf dem sich derartige Prognosemodelle bewegen mit AUCs von >0,9 eine Verbesserung von 0,004 als eine erhebliche Verbesserung festzuhalten. Deutlich wird dies u.a. auch durch die AUC von 0,845 der Tumordicke gemacht. Zwar ist die Tumordicke der beste Prognosefaktor, jedoch lässt sich eine Verbesserung um 0,067 nur dadurch erzielen, indem alle anderen

Prognosefaktoren plus die CNN-Prognose hinzugenommen werden. Im weiteren Schritt wurden die morphologischen Faktoren ausgelassen und auch hierbei finden sich einige interessante Punkte. Einerseits kann nur anhand der etablierten Prognosefaktoren innerhalb dieser Studienpopulation keine AUC von >0,9 erreicht werden. Weiterhin sieht man hier noch einmal den vorher erwähnten Effekt. Nimmt man nämlich hier noch einmal die CNN-Prognose hinzu wird eine Verbesserung von 0,011 erreicht im Vergleich zur vorherigen Zunahme von 0,004. Vergleicht man das „alle Faktoren plus AOI-Modell“ mit dem der „etablierten Faktoren plus AOI“ findet sich doch ein erheblicher Unterschied (AUC 0,912 vs. 0,887). Wieder rückt die Frage in den Vordergrund inwieweit die Erhebung einerseits der weiteren morphologischen Faktoren und andererseits der CNN-Prognose im klinischen

Alltag an Relevanz hat. Wird der Fokus auf hochwertige Prognosemodelle gesetzt, dann ist diese erweiterte Erhebung verschiedenster Faktoren sinnvoll. Vielmehr ist die Erhebung der CNN-Prognose vom Aufwand deutlich geringer verglichen mit den anderen Prognosefaktoren. Ist das digitale Bild des histologischen Schnittes vorhanden, kann das CNN-Modell in Sekunden aufgerufen werden und das Bild darin eingelesen werden. Der Validierungsprozess selbst braucht 1-3 Sekunden. Dazu kommt, dass diese Erhebung durch jedwedes Personal möglich ist, nach einer kurzen Einweisung, und kein Dermatopathologe dies durchführen muss.

Es verblieb die Frage welche Relevanz die prozentuale Einteilung der CNN in die jeweilige Kategorie übernimmt. Interessanterweise zeigt sich das Modell mit allen Faktoren und der CNN-Prognose einerseits kategorial in „dead“

oder „alive“ und andererseits in prozentualer Angabe nicht unterschiedlich. Eigentlich jedoch würde sich ein anderes Ergebnis vermuten lassen. Anhand der einzelnen Prozentangaben hat man nämlich eine viel größere Trennschärfe und würde ein besseres Modell erwarten. Auch war in der einzelnen ROC-Analyse die Prozentangabe immer der kategorialen Zuordnung überlegen, sowohl beim ÜB- als auch beim AOI-Modell. Die AOI-Prozentzuordnung zeigte sich mit einer AUC die um 0,035 besser war, jedoch ist plötzlich kein Effekt bei einem kombinierten Modell festzustellen. Eine Vermutung, wieso die Modelle sich als fast gleich darstellen wäre, dass im Rahmen der Modellerstellung in der logistischen Regression eine Klassifikation der Prozentangaben anhand von 0,5 bzw. 50% stattfindet und damit noch einmal dasselbe Ergebnis wie innerhalb der kategorialen

Zuordnung entsteht. Demgegenüber steht allerdings, dass mit beiden Zuordnungen – kategorial wie auch prozentual – ein besseres Modell erzielt wird und es nicht gleich bleibt. Würde man also die Prognose eines Patienten innerhalb dieser Studienpopulation wissen wollen, würde dieses letzte erstellte Modell ($p<0,001$) mit einer AUC von 0,915 die naheliegendste Aussage zum realen Outcome treffen können. Die Erhebung der Prozentzuordnung der CNN stellt kein größeres Hindernis dar. Bei jeder Validierung wird sie mit angegeben und kann im selben Prozess dann einfach entnommen und für weitere Modelle genutzt werden.

4.6 Modelleignung

Letztlich ist festzuhalten, dass sich das erstellte KI-Modell durchaus für die Prognose eignet, jedoch nur in Kombination mit den anderen

bekannten Prognosefaktoren. Die alleinige Prognoseaussage eignet sich nicht. Es stellt sich die Frage, ob die Google teachable machines - KI überhaupt geeignet ist für derartige Prognosemodelle bezüglich einer Tumorentität. Tatsächlich lassen sich als User nur drei Parameter überhaupt beeinflussen. Ob mit den in den vorherigen herausgearbeiteten Schwachstellen und möglichen Verbesserungen (genauere Diskussion im nachfolgenden Abschnitt) an diesem Modell ein geeignetes Modell, das alleinstehend eine zufriedenstellende Prognoseaussage erlaubt, überhaupt in der Praxis erstellt werden kann, ist schwierig festzuhalten. Die in Punkt 4.4.2 diskutierte spezialisierte KI „Mesonet“, die beim Mesotheliom eine bessere Aussage bezüglich des Überlebens lieferte als herkömmliche Prognosefaktoren, scheint eine bessere Alternative darzustellen. Auch der

grundsätzliche Typ der KI ist zu hinterfragen. In einer Prognoseaussage wäre es von Vorteil, im Prozess der Datenverarbeitung auf bereits vorhandene Daten in vorherigen Ebenen zuzugreifen, einen Vorteil den RNNs gegenüber CNNs bieten. Diese Aussage stützt sich auch an die bereits diskutierte Arbeit von Kulkarni et al., in der eine solche RNN mit einer CNN hinzugezogen wurde [82]. Auch wurde in der gleichen Arbeit dargestellt, dass ein großes Datenset mit rigorosem Selektionsprozess bezüglich der miteinbezogenen Objektträger notwendig war. Unter diesen Umständen kann auch festgehalten werden, dass ein Modell dann durchaus in der Lage ist, signifikant gute Prognoseaussagen abzugeben.

Vergleichbare Arbeiten, die die hier verwendete KI verwenden für eine Prognoseaussage, gibt es nicht. Vielmehr findet man jedoch positive Arbeiten bzgl. einer Klassifizierung

beispielsweise zweier Tumorarten, in denen die KI der teachable machines eher ihre Stärken aufweist.

4.7 Schlussfolgerungen aus den Ergebnissen und der Diskussion

In diesem letzten Abschnitt sollen alle gewonnen Erkenntnisse des Ergebnis- und Diskussionsteils, die in ihren jeweiligen Kapiteln herausgearbeitet wurden, dargestellt werden. Die vorherigen Erkenntnisse der etablierten Prognosefaktoren bezüglich ihrer Häufigkeit und ihrer Prognose konnten in dieser Arbeit bestätigt werden. Die Tumordicke nach Breslow zeigte sich im Rahmen dieser Arbeit durchgehend als der beste Prognosefaktor und kann alleinstehend das 5-Jahres-Überleben relativ genau vorhersagen.

Ein wesentlicher Grundstein wurde in dieser Arbeit bezüglich der morphologischen Faktoren gelegt. In der Literatur findet man nur sehr wenige Überlebens- und Häufigkeitsangaben dieser morphologischen Faktoren, welche in dieser Arbeit erfolgt sind. Die Mitose und die

Pigmentierung fanden sich vergleichsweise am häufigsten wieder und es konnte bezüglich ihres Effekts auch die besten Ergebnisse erzielt werden. Die Zellatypie und das Wachstumsmuster zeigten auch signifikante Effekte auf das Überleben. Innerhalb der morphologischen Faktoren hatten die Zytomorphologie und das Entzündungsinfiltrat die schwächsten Ergebnisse. Die Zytomorphologie zeigte sich in keiner Analyse als signifikant. Diesbezüglich wäre zumindest ein anderes Ergebnis zu erwarten gewesen, da in der Literatur andere Ergebnisse bekannt sind [69, 70].

Bei der Erstellung eines Modells bei den Google teachable machine sollte versucht werden, dass möglichst der krankheitsdefinierende Bereich oder die „area of interest“ für das Training und zur Validierung verwendet werden. Das AOI-Modell zeigte sich

durchgehend als deutlich überlegeneres Modell gegenüber dem ÜB-Modell. Will man nur einen einzelnen Parameter der CNN-Modelle für die Prognoseaussage heranziehen, eignet sich die Prozentangabe des AOI-Modells am besten, d.h. schlussfolgernd die prozentuale Zuordnung ist der kategorialen bei der Google teachable machine vorzuziehen, vorausgesetzt sie wurde angemessen trainiert. Sie ist jedoch der Tumordicke nach Breslow weiterhin unterlegen und es empfiehlt sich nicht, die Prognose allein daran festzumachen. Welche Schlussfolgerungen können also zur Verbesserung des Modells herangezogen werden, um zukünftig ein Modell zu erhalten, das sogar die Tumordicke nach Breslow oder die kombinierten Modelle übertrifft? Der Datensatz sollte so groß wie möglich gewählt werden. Mehr Daten liefern ein besseres Ergebnis in Bezug auf die CNN, wie von Uchida

et al. festgestellt wurde [77]. Weiterhin sollte nicht nur der gesamte Datensatz so groß wie möglich sein, sondern er sollte auch möglichst groß und möglichst gleich innerhalb der zu trainierenden Klassen sein. Es gibt keinen Anhalt, dass eine Veränderung der Hyperparameter nötig wäre, d.h. die Einstellung im Rahmen dieser Arbeit waren zufriedenstellend.
Die logistische Regressionsanalyse zum Verstehen verschiedenster Einflüsse auf die Prognose der CNN zeigte sich bei allen etablierten und, ausgenommen von der Zytomorphologie und dem Entzündungsinfiltrat, auch bei den morphologischen Parametern als signifikant.

Will man ein hochwertiges Modell zur Prognoseaussage erstellen, sollten alle etablierten Parameter enthalten sein, inklusive

des Invasionslevels und des histologischen Subtyps. Eine Diskussion der prognostischen Relevanz dieser beiden genannten Parameter ist derzeit sehr präsent, jedoch zeigten sie sich in dieser Arbeit durchgehend als geeignete Prognosefaktoren. Zu den etablierten Prognosefaktoren lohnt es sich, die vier weiteren Parameter zu erheben: Mitosen, Zellatypie, Wachstumsmuster und Pigmentierung. Vor allem die Zellatypie und das Wachstumsmuster scheinen nicht sehr präsent im klinischen Alltag zu sein, jedoch zeigen die Ergebnisse dieser Arbeit, dass sie eindeutig als Prognosefaktoren geeignet sind. Nimmt man zu den genannten Faktoren nun noch die CNN-Prognose hinzu, entsteht ein hervorragendes Modell, dass die Prognose eines Patienten mit einer sehr hohen Trefferquote richtig einschätzt. Vielmehr noch sollte die kategoriale als auch die prozentuale

Zuordnung der CNN herangezogen werden, um das bestmögliche Modell zu erhalten.

5 Zusammenfassung

Das maligne Melanom zeichnet sich weiter als eine Erkrankung mit hohem Stellenwert im klinischen Alltag ab. Mit dem zunehmend relevanten Thema einer möglichst genauen Prognose der Erkrankung wurden in dieser Arbeit „convolutional neural network“-Modelle (CNN-Modelle) trainiert und bezüglich ihrer Prognose ausgewertet mit der Frage, welche Relevanz künstliche Intelligenz im Rahmen einer genauen Prognoseabgabe beim malignen Melanom einnehmen kann.

836 Melanomschnitte von Patienten mit dem Zeitpunkt der Erstdiagnose zwischen 2012 und 2015 wurden gesammelt und erstmalig digitalisiert. Über eine zufällige Verteilung wurden ca. 60% (502 Schnitte) für das Training der Modelle verwendet. Hierin wurden die zwei Klassen „alive“ und „dead“ innerhalb der

„Google teachable machine“ trainiert. Bei der „Google teachable machine“ handelt es sich um ein allgemein zugängliches und benutzerfreundliches Tool, das es erlaubt, Modelle innerhalb von CNN-Plattformen zu erstellen. Die restlichen 40% (334 Schnitte) wurden durch die trainierten Modelle validiert. Die Prognoseaussage der CNN erfolgt anhand der Kategorien „dead“ und „alive“ und einer Prozentzahl, die aussagt, inwieweit der Schnitt aus „Sicht“ der CNN in die Kategorie hineinpasst. Die zwei trainierten Modelle unterschieden sich in Bezug auf die verwendeten digitalen Bilder. In einem ersten Modell wurde eine größere Aufnahme, die möglichst viel des feingeweblichen Schnittes beinhaltete, angefertigt. Das zweite Modell wurde nur mit und für die „areas of interest“ (AOI) trainiert, d.h. den krankheitsbestimmenden Arealen innerhalb der

Schnitte

Das AOI-Modell zeigte sich gegenüber dem Übersichtsbild-Modell deutlich überlegen. Innerhalb des Übersichtsbild-Modells konnten 84,13% richtig als „alive“ eingeordnet werden, jedoch lag die Trefferquote bei den „dead“-Schnitten lediglich bei 20,83%. Bei dem AOI-Modell wurden 70,36% richtig als „alive“, und 52,08% richtig als „dead“ eingeordnet. In einer ROC-Analyse zeigte sich einerseits das AOI-Modell als signifikanter gegenüber dem Übersichtsbild-Modell. Weiterhin eignete sich die prozentuale Einordnung der CNN gegenüber der kategorialen in „dead“ und „alive“ besser.

In univariaten, logistischen Regressionsanalysen ließen sich die etablierten Prognosefaktoren Tumordicke nach Breslow verhältnisskaliert, die pT-Einteilung nach AJCC 2009, Ulzeration, histologischer

Subtyp und Invasionslevel nach Clark, sowie die erstmals in dieser Arbeit erhobenen morphologischen Parameter Zellatypie, Mitosen, Wachstumsmuster und Pigmentierung als signifikante Einflussfaktoren auf die Prognose der CNN nachweisen.

Mit dem weiteren Ziel, ein möglichst hochwertiges Prognosemodell zu erstellen, wurden innerhalb logistischer Regressionsanalysen diverse Modelle erstellt und in ROC-Analysen ausgewertet. Die etablierten Prognosefaktoren zeigten sich gegenüber den hier erhobenen morphologischen Faktoren als überlegen. Weiterhin ließen sich Modelle, die die CNN-Prognose beinhalteten, durchgehend als besser nachweisen. Dies galt sowohl bei Modellen, die alle Faktoren beinhalteten, als auch bei Modellen, die nur die etablierten Faktoren beinhalteten. Das beste Modell dieses

Projekts konnte mit der Einbeziehung der fünf genannten etablierten und der vier genannten morphologischen Faktoren, sowie die prozentuale und kategoriale Einteilung der CNN (AOI-Modell) erstellt werden.
In einem anderen Schritt wurden ROC- und Überlebensanalysen der etablierten Prognosefaktoren durchgeführt. Hierin konnten die Erkenntnisse des aktuellen wissenschaftlichen Standes bestätigt werden. Die Tumordicke nach Breslow und die Ulzeration sind zentral in der Prognose des malignen Melanoms. Das Invasionslevel nach Clark und der histologische Subtyp sind zwar signifikant, in der prognostischen Aussagekraft den Faktoren Tumordicke und Ulzeration jedoch unterlegen.

Aus den Ergebnissen lässt sich schließen, dass eine zunehmend relevante Rolle der KI im

Rahmen der Prognoseabgabe des malignen Melanoms zu erwarten ist. Das zweite CNN-Modell (AOI-Modell) zeigte eine zufriedenstellende Übereinstimmung bei der Prognoseabgabe und eine hohe Signifikanz und hervorragende Ergebnisse in den nachfolgenden Analysen. Für die komplette Ersetzung der etablierten Prognosefaktoren und der diagnostischen Arbeit eines Dermatopathologen reichen die Ergebnisse noch nicht aus, aber es bleibt, die zukünftigen Entwicklungen abzuwarten.

6 Literatur

1 *Kozovska Z, Gabrisova V, Kucerova L*. Malignant melanoma: diagnosis, treatment and cancer stem cells. Neoplasma 2016; 63: 510 – 517

2 Centers for Disease Control and Prevention. Melanoma Incidence and Mortality, United States 2012-2016. U.S. Cancer Statistics data brief, no 9. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2019

3 R.M. MacKie, A. Hauschild, A.M.M. Eggermont. Epidemiology of invasive cutaneous melanoma. Annals of Oncology, Volume 20 2009: Seite 2

4 *Ghazawi FM, Le M, Lagacé F et al.* Incidence, Mortality, and Spatiotemporal Distribution of Cutaneous Malignant Melanoma Cases Across Canada. Journal

of cutaneous medicine and surgery 2019; 23: 394 – 412

5 *Panda S DS,* Besra K, Samantaray S *et al.* Clinicopathological study of malignant melanoma in a regional cancer center. Indian J Cancer 2018: 55:292-6

6 *Hamm H, Höger PH*. Skin tumors in childhood. Deutsches Arzteblatt international 2011; 108: 347 – 353

7 Leitlinienprogramm Onkologie. Im Internet: https://www.leitlinienprogramm-onkologie.de/leitlinien/melanom/; Stand: 22.07.2021, 18:06 Uhr

8 *Lamos C, Hunger RE*. Checkpoint-Inhibitoren – Indikation und Verwendung bei Melanompatienten. Zeitschrift fur Rheumatologie 2020; 79: 818 – 825

9 *Forchhammer S, Abu-Ghazaleh A, Metzler G et al.* Development of an Image Analysis-Based Prognosis Score Using

Google's Teachable Machine in Melanoma. Cancers 2022; 14

10 *Overwien O, Hrsg.* EINE ANONYME VORLESUNG ÜBER DAS PROGNOSTIKON AUS DEM SPÄTANTIKEN ALEXANDRIA; 2011

11 *Balch CM, Soong SJ, Gershenwald JE et al.* Prognostic factors analysis of 17,600 melanoma patients: validation of the American Joint Committee on Cancer melanoma staging system. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2001; 19: 3622 – 3634

12 *Wisco OJ, Sober AJ*. Prognostic factors for melanoma. Dermatologic clinics 2012; 30: 469 – 485

13 *Barnhill RL, Fine JA, Roush GC et al.* Predicting five-year outcome for patients with cutaneous melanoma in a population-

based study. Cancer 1996; 78: 427 – 432

14 *Breslow A*. Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma. Annals of surgery 1970; 172: 902 – 908

15 *Breslow A*. Tumor thickness, level of invasion and node dissection in stage I cutaneous melanoma. Annals of surgery 1975; 182: 572 – 575

16 *Balch CM, Gershenwald JE, Soong S-J et al.* Final version of 2009 AJCC melanoma staging and classification. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2009; 27: 6199 – 6206

17 *Keung EZ, Gershenwald JE*. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. Expert review of anticancer therapy

2018; 18: 775 – 784

18 *Hout FEM* in 't, *Haydu LE, Murali R et al.* Prognostic importance of the extent of ulceration in patients with clinically localized cutaneous melanoma. Annals of surgery 2012; 255: 1165 – 1170

19 *Scolyer RA, Rawson RV, Gershenwald JE et al.* Melanoma pathology reporting and staging. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2020; 33: 15 – 24

20 *Clark WH, From L, Bernardino EA et al.* The histogenesis and biologic behavior of primary human malignant melanomas of the skin. Cancer research 1969; 29: 705 – 727

21 *Morton DL, Davtyan DG, Wanek LA et al.* Multivariate analysis of the relationship between survival and the microstage of

primary melanoma by clark level and breslow thickness. Cancer 1993; 71: 3737 – 3743

22 *Forman SB, Ferringer TC, Peckham SJ et al.* Is superficial spreading melanoma still the most common form of malignant melanoma? Journal of the American Academy of Dermatology 2008; 58: 1013 – 1020

23 *Kempf W, Hantschke M, Kutzner H, Burgdorf W*. Dermatopathologie. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015

24 *Smoller BR*. Histologic criteria for diagnosing primary cutaneous malignant melanoma. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2006; 19 Suppl 2: S34-40

25 *Elder DE, Bastian BC, Cree IA et al.* The

2018 World Health Organization Classification of Cutaneous, Mucosal, and Uveal Melanoma: Detailed Analysis of 9 Distinct Subtypes Defined by Their Evolutionary Pathway. Archives of pathology & laboratory medicine 2020; 144: 500 – 522

26 *Wang Y, Zhao Y, Ma S*. Racial differences in six major subtypes of melanoma: descriptive epidemiology. BMC cancer 2016; 16: 691

27 *Lattanzi M, Lee Y, Simpson D et al.* Primary Melanoma Histologic Subtype: Impact on Survival and Response to Therapy. Journal of the National Cancer Institute 2019; 111: 180 – 188

28 *Gu J, Wang Z, Kuen J et al.* Recent advances in convolutional neural networks. Pattern Recognition 2018; 77: 354 – 377

29 *Zhang L, Tan J, Han D et al.* From

machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug discovery today 2017; 22: 1680 – 1685

30 *Ian Goodfellow, Yoshua Bengio, Aaron Courville*. Deep learning. Adaptive Computation and Machine Learning series: MIT Press; 2016

31 *Rampasek L, Goldenberg A*. TensorFlow: Biology's Gateway to Deep Learning? Cell systems 2016; 2: 12 – 14

32 *Renganathan V*. Overview of artificial neural network models in the biomedical domain. Bratislavske lekarske listy 2019; 120: 536 – 540

33 *Google teachable machine*. Google teachable machine - FAQ. Im Internet: https://teachablemachine.withgoogle.com/faq; Stand: 15.07.2021, 16:52

34 *Martín Abadi, Paul Barham, Jianmin Chen,*

Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: a system for large-scale machine learning. In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16); 2016

35 *Erickson BJ, Korfiatis P, Akkus Z et al.* Toolkits and Libraries for Deep Learning. Journal of digital imaging 2017; 30: 400 – 405

36 Feasibility Study of Google’s Teachable Machine in Diagnosis of Tooth-Marked Tongue. J Dent Hyg Sci 2020; 20: 206 – 212

37 *Feldman AT, Wolfe D*. Tissue processing and hematoxylin and eosin staining. Methods in molecular biology (Clifton, N.J.) 2014; 1180: 31 – 43

38 *Cerroni L, Garbe C, Metze D, Kutzner H, Kerl H*. Histopathologie der Haut. Berlin, Heidelberg: Springer Berlin Heidelberg; 2016

39 *Chan JKC*. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. International journal of surgical pathology 2014; 22: 12 – 32

40 *Robert Koch-Institut, Gesellschaft der epidemiologischen Krebsregister in Deutschland e.V.* Krebs in Deutschland 2015/2016: Robert Koch-Institut; 2019

41 *Ali Z, Yousaf N, Larkin J*. Melanoma epidemiology, biology and prognosis. EJC supplements : EJC : official journal of EORTC, European Organization for

Research and Treatment of Cancer … [et al.] 2013; 11: 81 – 91

42 Cancer Statistics UK. Im Internet: http://info.cancerresearchuk.org/; Stand: 17.11.2011

43 *Zentrum der epidemiologischen Krebsregister in Deutschland, Zentrum für Krebsregisterdaten im Robert-Koch-Institut*. Krebs in Deutschland für 2011/2012. Im Internet: https://www.gbe-bund.de/pdf/kid_2019_c43_melanom.pdf; Stand: 12.10.2021 17:04

44 *Zentrum der epidemiologischen Krebsregister in Deutschland, Zentrum für Krebsregisterdaten im Robert-Koch-Institut*. Krebs in Deutschland für 2011/2012. Im Internet: https://www.gbe-bund.de/pdf/kid_2015_melanom.pdf; Stand: 12.10.2021 17:00

45 *Zentrum der epidemiologischen*

Krebsregister in Deutschland, Zentrum für Krebsregisterdaten im Robert-Koch-Institut. Krebs in Deutschland für 2013/2014. Im Internet: https://www.gbe-bund.de/pdf/kid_2017_melanom.pdf; Stand: 12.10.2021 17:03

46 *Rastrelli M, Tropea S, Rossi CR et al.* Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. In vivo (Athens, Greece) 2014; 28: 1005 – 1011

47 *Shaikh WR, Xiong M, Weinstock MA*. The contribution of nodular subtype to melanoma mortality in the United States, 1978 to 2007. Archives of dermatology 2012; 148: 30 – 36

48 *Xu Z, Shi P, Yibulayin F et al.* Spindle cell melanoma: Incidence and survival, 1973-2017. Oncology letters 2018; 16: 5091 – 5099

49 *Piao Y, Guo M, Gong Y*. Diagnostic challenges of metastatic spindle cell melanoma on fine-needle aspiration specimens. Cancer 2008; 114: 94 – 101

50 *Murali R, Doubrovsky A, Watson GF et al.* Diagnosis of metastatic melanoma by fine-needle biopsy: analysis of 2,204 cases. American journal of clinical pathology 2007; 127: 385 – 397

51 *Gupta SK, Rajwanshi AK, Das DK*. Fine needle aspiration cytology smear patterns of malignant melanoma. Acta cytologica 1985; 29: 983 – 988

52 *Ghasemi Basir HR, Alirezaei P, Ahovan S et al.* The relationship between mitotic rate and depth of invasion in biopsies of malignant melanoma. Clinical, cosmetic and investigational dermatology 2018; 11: 125 – 130

53 *Halaban R*. Pigmentation in melanomas:

changes manifesting underlying oncogenic and metabolic activities. Oncology research 2002; 13: 3 – 8

54 *Granter SR, Weilbaecher KN, Quigley C et al.* Role for microphthalmia transcription factor in the diagnosis of metastatic malignant melanoma. Applied immunohistochemistry & molecular morphology : AIMM 2002; 10: 47 – 51

55 *Eberle J, Garbe C, Wang N et al.* Incomplete expression of the tyrosinase gene family (tyrosinase, TRP-1, and TRP-2) in human malignant melanoma cells in vitro. Pigment cell research 1995; 8: 307 – 313

56 *Moreno-Ramírez D, Ojeda-Vila T, Ríos-Martín JJ et al.* Association between tumor size and Breslow's thickness in malignant melanoma: a cross-sectional, multicenter study. Melanoma research 2015; 25: 450 –

452

57 *Rousseau DL, Ross MI, Johnson MM et al.* Revised American Joint Committee on Cancer staging criteria accurately predict sentinel lymph node positivity in clinically node-negative melanoma patients. Annals of surgical oncology 2003; 10: 569 – 574

58 *Hermanek P, Hornstein OP, Tonak J et al.* Malignes Melanom. Invasionstiefe und Melanomtyp. Beiträge zur Pathologie 1976; 157: 269 – 282

59 *Donnellan MJ, Seemayer T, Huvos AG et al.* Clinicopathologic study of cutaneous melanoma of the head and neck. The American Journal of Surgery 1972; 124: 450 – 455

60 *Hansen MG, McCarten AB*. Tumor thickness and lymphocytic infiltration in malignant melanoma of the head and neck. The American Journal of Surgery

1974; 128: 557 – 561

61 *McGovern VJ*. The classification of melanoma and its relationship with prognosis. Pathology 1970; 2: 85 – 98

62 *Wanebo HJ, Woodruff J, Fortner JG*. Malignant melanoma of the extremities: A clinicopathologic study using levels of invasion (microstage). Cancer 1975; 35: 666 – 676

63 *Baade PD, Whiteman DC, Janda M et al.* Long-term deaths from melanoma according to tumor thickness at diagnosis. International journal of cancer 2020; 147: 1391 – 1396

64 *Callender GG, McMasters KM*. What does ulceration of a melanoma mean for prognosis? Advances in surgery 2011; 45: 225 – 236

65 *Ellerhorst JA, Greene VR, Ekmekcioglu S et al.* Clinical correlates of NRAS and

BRAF mutations in primary human melanoma. Clinical cancer research : an official journal of the American Association for Cancer Research 2011; 17: 229 – 235

66 *Balch CM, Buzaid AC, Atkins MB et al.* A new American Joint Committee on Cancer staging system for cutaneous melanoma. Cancer 2000; 88: 1484 – 1491

67 *Goydos JS, Shoen SL*. Acral Lentiginous Melanoma. Cancer treatment and research 2016; 167: 321 – 329

68 *Bradford PT, Goldstein AM, McMaster ML et al.* Acral lentiginous melanoma: incidence and survival patterns in the United States, 1986-2005. Archives of dermatology 2009; 145: 427 – 434

69 *Gilda Gudacker*. Das Melanom der Uvea: Dissertation: 10-Jahres-Follow-up und klinische Prognosefaktoren. 29.01.2010

70 *Al-Jamal RT, Kivelä T*. KI-67

immunopositivity in choroidal and ciliary body melanoma with respect to nucleolar diameter and other prognostic factors. Current eye research 2006; 31: 57 – 67

71 *Arumi-Uria M, McNutt NS, Finnerty B*. Grading of atypia in nevi: correlation with melanoma risk. Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc 2003; 16: 764 – 771

72 *Brouwer NJ, Marinkovic M, Luyten GPM et al.* Lack of tumour pigmentation in conjunctival melanoma is associated with light iris colour and worse prognosis. The British journal of ophthalmology 2019; 103: 332 – 337

73 *Sarna M, Krzykawska-Serda M, Jakubowska M et al.* Melanin presence inhibits melanoma cell spread in mice in a unique mechanical fashion. Scientific

reports 2019; 9: 9280

74 *Maru GB, Gandhi K, Ramchandani A et al.* The role of inflammation in skin cancer. Advances in experimental medicine and biology 2014; 816: 437 – 469

75 *Pan M, Alavi M, Herrinton LJ*. Association of Inflammatory Markers with Disease Progression in Patients with Metastatic Melanoma Treated with Immune Checkpoint Inhibitors. The Permanente journal 2018; 22: 17 – 149

76 *Colloby PS, West KP, Fletcher A*. Observer variation in the measurement of Breslow depth and Clark's level in thin cutaneous malignant melanoma. The Journal of pathology 1991; 163: 245 – 250

77 *Uchida S, Ide S, Iwana BK, Zhu A*. A Further Step to Perfect Accuracy by Training CNN with Larger Data. 2016 15th International Conference on Frontiers in

Handwriting Recognition (ICFHR): IEEE; 102016: 405 – 410

78 *Mutasa S, Sun S, Ha R*. Understanding artificial intelligence based radiology studies: What is overfitting? Clinical imaging 2020; 65: 96 – 99

79 *Hekler A, Utikal JS, Enk AH et al.* Pathologist-level classification of histopathological melanoma images with deep neural networks. European journal of cancer (Oxford, England : 1990) 2019; 115: 79 – 83

80 *Hekler A, Utikal JS, Enk AH et al.* Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. European journal of cancer (Oxford, England : 1990) 2019; 118: 91 – 96

81 *Madabhushi A, Lee G*. Image analysis and machine learning in digital pathology:

Challenges and opportunities. Medical image analysis 2016; 33: 170 – 175

82 *Kulkarni PM, Robinson EJ, Sarin Pradhan J et al.* Deep Learning Based on Standard H&E Images of Primary Melanoma Tumors Identifies Patients at Risk for Visceral Recurrence and Death. Clinical cancer research : an official journal of the American Association for Cancer Research 2020; 26: 1126 – 1134

83 *Schuster M, Paliwal KK*. Bidirectional recurrent neural networks. IEEE Trans. Signal Process. 1997; 45: 2673 – 2681

84 *Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov*. Dropout: A Simple Way to Prevent Neural Networks from Overfitting: Journal of Machine Learning Research 15; 2014

85 *Courtiol P, Maussion C, Moarii M et al.*

Deep learning-based classification of mesothelioma improves prediction of patient outcome. Nature medicine 2019; 25: 1519 – 1525

86 *Wells A, Patel S, Lee JB et al.* Artificial intelligence in dermatopathology: Diagnosis, education, and research. Journal of cutaneous pathology 2021; 48: 1061 – 1068

87 *Clemente CG, Mihm MC, Bufalino R et al.* Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. Cancer 1996; 77: 1303 – 1310

88 *Scolyer RA, Busam KJ*. Prognosis, Staging, and Reporting of Melanomas. Pathology of Melanocytic Tumors: Elsevier; 2019: 386 – 396

89 *Fu Q, Chen N, Ge C et al.* Prognostic value of tumor-infiltrating lymphocytes in

melanoma: a systematic review and meta-analysis. Oncoimmunology 2019; 8: 1593806

90 *Lyth J, Hansson J, Ingvar C et al.* Prognostic subclassifications of T1 cutaneous melanomas based on ulceration, tumour thickness and Clark's level of invasion: results of a population-based study from the Swedish Melanoma Register. The British journal of dermatology 2013; 168: 779 – 786

7 Erklärungen zum Eigenanteil

Die Arbeit wurde von Prof. Dr. Thomas Eigentler betreut. Durch Dr.med. Stephan Forchhammer als Mentor erfolgte eine unterstützende Anleitung im Rahmen des gesamten Projekts.

Die Konzeption der Studie erfolgte in Zusammenarbeit mit Prof. Dr. Eigentler und Dr.med. Forchhammer.

Das digitale Erfassen der histologischen Schnitte und das Auffinden innerhalb des Archivs erfolgte nach Anleitung eigenständig.
Das Erheben der morphologischen Parameter und das Abfotografieren der histologischen Schnitte erfolgten durch Dr. med. Stephan Forchhammer.

Das Erfassen der follow-up Daten erfolgte sowohl durch Dr. med. Forchhammer als auch eigenständig.

Die statistische Auswertung erfolgte nach Beratung durch das Institut für klinische Epidemiologie und angewandte Biometrie von Seiten Prof. Dr. Peter Martus und Anleitung von Dr. med. Forchhammer eigenständig.

Die Publikation auf der Grundlage dieser Arbeit erfolgte von Dr. med. Forchhammer. Von meiner Seite, als genannter Zweitautor, erfolgten als Beitrag die erhobenen Daten, Teile der Auswertung, sowie Kommentare und Verbesserungsvorschläge. Bis auf einzelne Ergänzungen wurde diese Doktorarbeit zeitlich vor der Publikation verfasst.

Ich versichere, die Doktorarbeit komplett

eigenständig verfasst zu haben und keine weiteren als die von mir angegebenen Quellen verwendet zu haben.

Tübingen, den 03.07.2022

8 Veröffentlichungen

Eine Publikation mit den Daten dieser Arbeit ist vom Mentor dieser Dissertation, Dr. med. Stephan Forchhammer, als Erstautor am 29.04.2022 erfolgt. Von meiner Seite, als genannter Zweitautor, erfolgte als Beitrag die erhobenen Daten, Teile der Auswertung, sowie Kommentare und Verbesserungsvorschläge:
Development of an Image Analysis-Based Prognosis Score Using Google's Teachable Machine in Melanoma, Cancers 2022, 14(9), 2243,
https://doi.org/10.3390/cancers14092243

9 Abbildungsverzeichnis

10 Tabellenverzeichnis

Danksagungen

Mein Dank gilt Prof. Dr. Thomas Eigentler für die Betreuung dieser Arbeit und die durchgehende Möglichkeit, ihn als Ansprechpartner zu erreichen.

Mit diesen wenigen Zeilen kann man nur schwierig meinem Dank gegenüber Dr. med. Stephan Forchhammer gerecht werden, der mir in jedem Arbeitsschritt zur Seite stand und seine Rolle als Mentor perfekt erfüllte. Auch die ständige Erreichbarkeit und die großartigen Beratungen waren keine Selbstverständlichkeit.

Mein Dank gilt auch meinen Eltern und meinen Geschwistern, die mir immer zur Seite standen.

Ein weiterer Dank gilt auch meinen Kommilitonen und Freunden, die oftmals zu Fragen bzgl. Formalien und Abläufen bei Dissertationen zur Verfügung standen.

Abschließend gebührt alles Lob und Dank dem einen Gott, dem Herrn der Welten.

Anhang

Die beiden im Rahmen dieses Projekts erstellten Modelle wurden mit 836 Schnitten umfangreich trainiert und sollen für den weiteren Prozess der Wissenschaft zur Verfügung stehen. Ein Fehler bei dem Output der Modelle sollte vernachlässigt werden: Es handelte sich nicht um das „two-year survival“, sondern um das „overall survival“.
Die beiden Modelle sind zum Stand Juli 2023 weiter öffentlich zugänglich und bereit zur Validierung weiterer Schnitte:

Übersichtsbild-Modell:
https://teachablemachine.withgoogle.com/models/q3W4kP4zk/

„Area of interest“-Modell (AOI):
https://teachablemachine.withgoogle.com/models/EWFL98pti/

Anhang

Alle weiteren erhobenen Auswertungsdaten, darunter die sechs erhobenen morphologischen Parameter und die digitalen Schnitte, sind in der Universitäts-Hautklinik Tübingen archiviert.