Aus der

Radiologischen Universitätsklinik Tübingen

Abteilung Diagnostische und Interventionelle Radiologie

Automatisierte Schätzung des biologischen Gehirnalters auf Grundlage von klinischen CT-Datensätzen mittels maschinellen Lernens

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Medizin

der Medizinischen Fakultät
der Eberhard Karls Universität
zu Tübingen

vorgelegt von

Kerber, Bjarne Jonas

2025

Dekan: Professor Dr. B. Pichler

1. Berichterstatter: Professor Dr. S. Gatidis
2. Berichterstatter: Professor Dr. S. Poli

Tag der Disputation: 25.04.2025

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Chronologisches (CA) und Biologisches Alter (BA)

Das Erfragen des Chronologischen Alters eines Patienten, also des Zeitraumes vom Moment der Geburt bis zum Tag der Untersuchung, ist Teil eines jeden Anamnesegesprächs. Das Chronologische Alter dient dabei als Proxy-Variable für die Einschätzung des körperlichen Zustands eines Patienten und beeinflusst somit jeden Behandlungsschritt, von weitergehender Diagnostik über die Therapieplanung bis hin zur Nachsorge.

Laut Rose (1991) manifestiert sich der Alterungsprozess in der fortschreitenden Abnahme von altersabhängigen Fitness-Faktoren eines Organismus. Auslöser dieser Abnahme sei ein innerer physiologischer Verfall. Diese Abnahme der Fitness wird begleitet von einer steigenden Morbidität und Mortalität, so dass das Patientenalter einen wichtigen klinischen Parameter darstellt. So ist dieses einer der wichtigsten Risikofaktoren für kardio-vaskuläre, onkologische und neurodegenerative Erkrankungen (Dhingra and Vasan, 2012, Reeve et al., 2014, White et al., 2014).

Jedoch ist der Alterungsprozess durch eine extreme interindividuelle Heterogenität geprägt. Sowohl der Beginn, als auch die Fortschrittsrate und das Ausmaß des Alterns variieren aufgrund unterschiedlicher genetischer Ausstattung, Lebensführung und einwirkenden Umweltfaktoren stark zwischen einzelnen Individuen (Fedarko, 2011). Das Chronologische Alter ist daher unzuverlässig, um den individuellen Alterungsprozess und physischen Zustand eines Patienten einschätzen zu können (Lowsky et al., 2014, Ohnishi et al., 2001).

Das Konzept des Biologischen Alters versucht diesen Prozess anhand objektiv messbarer Merkmale zu quantifizieren. Es wurden bereits verschiedene Kombinationen aus physiologischen Parametern, beispielsweise der maximalen Sauerstoffaufnahme, Einsekundenkapazität (FEV_1), Griffkraft (Jee et al., 2012), (epi-)genetischen Analysen, DNA-Methylierung oder Telomerlängen-

bestimmungen (Horvath, 2013, Hannum et al., 2013, Blackburn et al., 2015) vorgeschlagen und untersucht.

Neben der Veränderung von physiologischen Faktoren führt der Alterungsprozess zur Veränderung morphologischer Merkmale, die mittels bildgebender Verfahren dargestellt werden können, darunter Verkalkungen von Gefäßen sowie Volumenveränderungen von Gehirn-, Muskel- und Knorpelgewebe (Ohnishi et al., 2001, Narici et al., 2003, Hudelmaier et al., 2001).

Altersschätzungen von Modellen des Maschinellen Lernens, die auf Bildgebungsdaten gesunder Kohorten trainiert wurden, korrelieren mit dem Biologischen Alter und können als Bildgebungs-Biomarker eingesetzt werden (Cole and Franke, 2017).

Eine ausreichend genaue automatische Schätzung des Biologischen Alters könnte Ärzte dabei unterstützen, den tatsächlichen körperlichen Zustand eines Patienten besser einzuschätzen und damit die klinische Entscheidungsfindung verbessern. So könnte beispielsweise das geschätzte Biologische Alter das Chronologische Alter als Parameter in altersabhängigen Risiko-Scores ersetzen (Grundy, 2001), um deren Genauigkeit und klinische Effektivität zu verbessern. Außerdem könnte das Konzept des Biologischen Alters für Patienten zugänglicher sein als abstrakte Werte eines Risiko-Scores (McClelland et al., 2009).

1.2 Chancen der automatisierten Datenanalyse in der Radiologie

Die moderne Medizin ermöglicht es Ärzten, auf eine stetig wachsende Anzahl verschiedenster Daten zurückzugreifen, um informierte Entscheidungen in Diagnostik, Therapie und Nachsorge zu treffen. Bildgebende Verfahren nehmen dabei eine zentrale Rolle ein.

Die jährlich erzeugte Datenmenge im Gesundheitswesen wird in den kommenden Jahren ein exponentielles Wachstum erfahren (Rydning, 2018). Dieser Trend wird vor allem angetrieben durch medizinische Bildgebungsverfahren, die gleichzeitig häufiger (Abbildung 1) als auch in immer

besserer Auflösung angewendet werden (Rydning, 2018). Die Rate von Computertomographie (CT)-Untersuchungen pro Einwohner stieg in nahezu allen OECD-Ländern, so auch in Deutschland: Von 2004 bis 2021 erhöhte sich selbige von 90 auf ca. 160 pro 1000 Einwohner, eine Zunahme von mehr als 75% (OECD, 2023). In den USA werden heute sogar fast anderthalb mal so viele CT-Untersuchungen pro Einwohner wie in Deutschland durchgeführt.

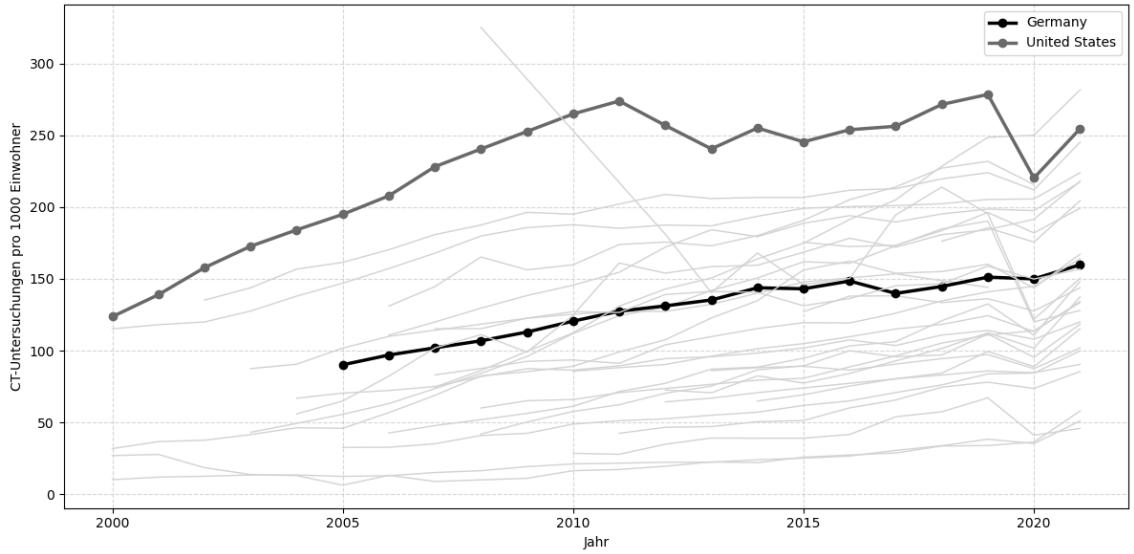


Abbildung 1: CT-Aufnahmen pro 1000 Einwohner pro Jahr in Deutschland und den USA sowie den weiteren OECD-Staaten von 2000 bis 2021 (OECD, 2023).
Die Anzahl an durchgeführten CT-Untersuchungen pro Einwohner haben sich in Deutschland von 2004 bis 2021 um mehr als 75% erhöht.

Im klinischen Alltag der Radiologie werden weltweit komplexe Bildgebungsdaten in immer größerer Menge und Geschwindigkeit erzeugt, gespeichert und müssen zeitnah verarbeitet werden.

Bruno et al. (2015) identifizierten die Interpretation von Bildgebungsdaten als eine Tätigkeit, bei der einige kognitiv anspruchsvolle Schritte vonnöten seien, die zusätzlich durch persönliche Erfahrung, Wissen und Bias des Untersuchers beeinflusst würden. Somit seien diese bislang menschlichen Untersuchern vorbehalten und auch deren Limitationen unterworfen. Heute liegt die Auswertung von medizinischen Bildgebungsdaten für den klinischen und wissenschaftlichen Gebrauch fast ausschließlich in der Hand menschlicher Spezialisten. Entsprechend hat sich bei ansteigenden Untersuchungsraten die

Arbeitslast für Radiologen bereits in den vergangenen Jahren stark erhöht (Griffith et al., 2019, Bruls and Kwee, 2020).

Mit zunehmender Arbeitsbelastung steigt die psychosoziale Belastung der befundenden Ärzte, während sie auch als Faktor für diagnostische Fehler in Frage kommt (Bruno et al., 2015).

Verfahren des Maschinellen Lernens könnten zukünftig als Werkzeuge im klinischen Alltag eingesetzt werden, um Radiologen bei der Detektion von Anomalien und Interpretationen von Befunden zur Seite stehen (Choy et al., 2018). Die Integration von Modellen des Maschinellen Lernens, die beispielsweise eine Vorbearbeitung der Bilddaten vornimmt, verspricht eine erhöhte Effizienz und Zeitersparnis für den Untersucher, während die Fehlerrate gesenkt werden könnte (Hosny et al., 2018).

1.3 Maschinelles Lernen

Maschinelles Lernen ist ein Teilgebiet der Erforschung der Künstlichen Intelligenz. Der Begriff Maschinelles Lernen umfasst Algorithmen, die anhand von bereits verarbeiteten Trainingsdaten ihre Leistung bei der Lösung eines Problems, auch auf zuvor ungesehenen Daten, verbessern, also „lernen“, können (Mohri et al., 2018, Zhou, 2021).

Maschinelles Lernen kann grob in drei Kategorien eingeteilt werden, Supervised (Überwachtes), Unsupervised (Unüberwachtes) und Reinforcement (Bestärkendes) Learning. In dieser Arbeit wurden ausschließlich Verfahren des Supervised Learnings angewandt, weshalb sich in dieser Einleitung darauf beschränkt wird.

1.3.1 Supervised Learning

Ziel des Supervised Learning ist es, eine Funktion aus dem zur Verfügung stehenden Funktionenraum zu lernen, die die gegebenen Input-Werte aus dem Feature-Raum optimal auf die zugehörigen Output-Werte des Output-Raumes abbildet. Hierfür wird ein Trainingsdatensatz verwendet, der Paare von bekannten Input- und zugehörigen Output-Werten enthält.

Während des Trainings erzeugt das Modell auf Basis der Input-Daten einen Output. Der generierte Output wird mit dem wahren, erwarteten Wert verglichen und mithilfe einer Verlustfunktion (Loss-Funktion) ein Fehler (Loss) berechnet. Dieser Fehler kann durch Feedback-Mechanismen genutzt werden, um die Parameter zur Bewertung der Input-Daten anzupassen, die generierten Outputs zu verbessern, und somit die Verlustfunktion zu minimieren (James et al., 2013).

Die beschriebenen Operationen finden nur auf dem Trainingsdatensatz statt. Allein dieser wird genutzt, um die Parameter des Modells zu verändern. Nach dem Training kann das Modell auch auf bisher ungesehene Input-Daten angewendet werden. Diese können zur Leistungsüberprüfung in einem Test-Datensatz zusammengefasst werden, um zu überprüfen, wie gut ein Algorithmus auf zuvor ungesehene Daten „generalisiert“, sich seine Leistung also auf unbekannten Datensätzen verändert. Eine strikte Trennung von Trainings- und Testdaten ist dazu vonnöten.

Supervised Learning wird eingesetzt, um Klassifikations- und Regressionsprobleme zu lösen. Ziel der Klassifikation ist die korrekte Zuordnung eines Datenpunktes zu einer spezifischen Klasse aus einer Menge von Auswahlmöglichkeiten, also ein qualitativer Output. Bei der Regression wird stattdessen versucht, die Beziehung einer abhängigen Variable und mindestens einer unabhängigen Variable in Form einer Funktion zu modellieren, um Vorhersagen über quantitative Werte treffen zu können.

Für die Minimierung der Fehlerfunktion auf dem Trainingsdatensatz existieren voneinander verschiedene Lösungen, die Funktionen unterschiedlicher Komplexität umfassen. Die Anpassung des Modells an die Trainingsdaten verbessert sich, je höher die Komplexität des Modells ist, respektive wie viele verschiedene Parameter des Modells angepasst werden können. Ist die Komplexität der zur Verfügung stehenden Funktionen zu niedrig, kann das Modell die Varianz des Datensatzes nicht ausreichend abbilden und es kommt zum „Underfitting“ mit hohem Trainings- und Testfehler (Abbildung 2).

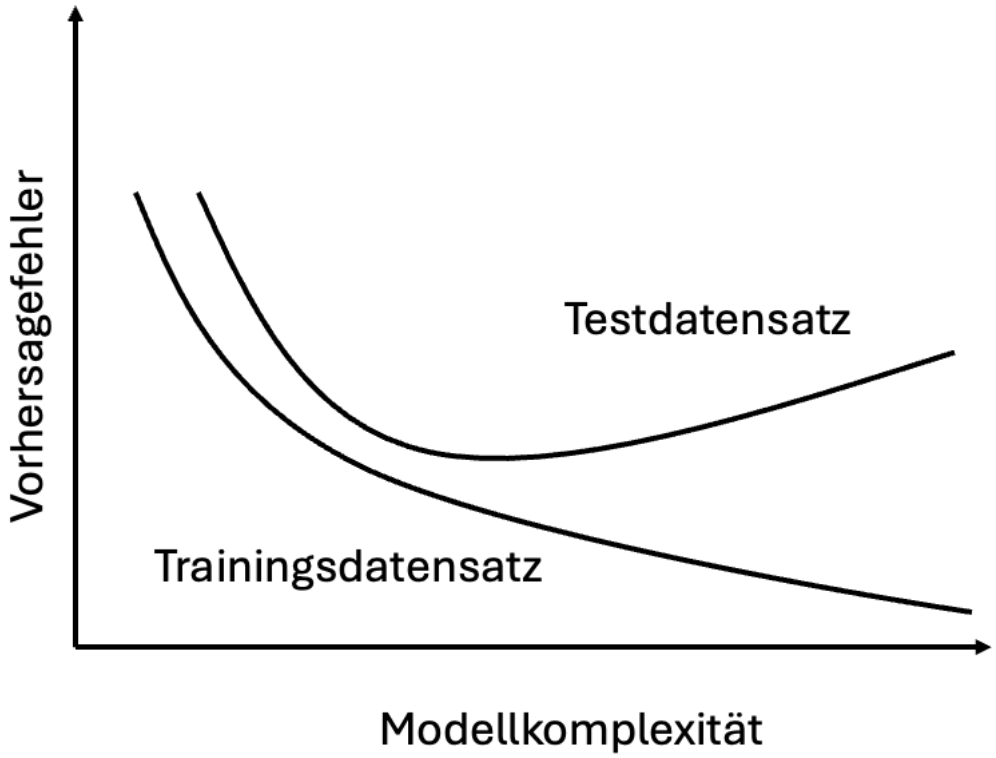


Abbildung 2: Trainings- und Test-Fehler als Funktion der Modell-Komplexität. Bei zu niedriger Modellkomplexität kommt es zu „Underfitting“, bei zu hoher zu „Overfitting“. Nach Hastie et al. (2009).

Im Gegensatz dazu ist bei hoher Komplexität des Modells die Anpassung an die spezifischen Eigenschaften der Trainingsdaten zu stark und das Modell generalisiert schlecht auf ungesehene Daten. Somit steigt der Test-Fehler bei sinkendem Trainings-Fehler, sog. „Overfitting“ (James et al., 2013). Die Komplexität des Modells muss für die Anwendung entsprechend gewählt werden.

Im Folgenden sollen vier algorithmische Ansätze kurz vorgestellt werden.

1.3.1.1 Linear Regression

Linear Regression umfasst eine Zahl verschiedener Verfahren, welche auf Basis eines mathematischen Kriteriums d-dimensionale Daten mithilfe einer linearen Funktion mit d verschiedenen Koeffizienten modellieren. Am häufigsten wird die Methode der kleinsten Quadrate genutzt. Ziel ist es hierbei, in einem linearen Modell die Koeffizienten so zu wählen, dass die Summe der Quadrate der Abweichungen zwischen Datenpunkten und Vorhersagen minimiert wird

(James et al., 2013). Abbildung 3 zeigt eine optimale lineare Anpassung (Ebene) an zweidimensionale Datenpunkte.

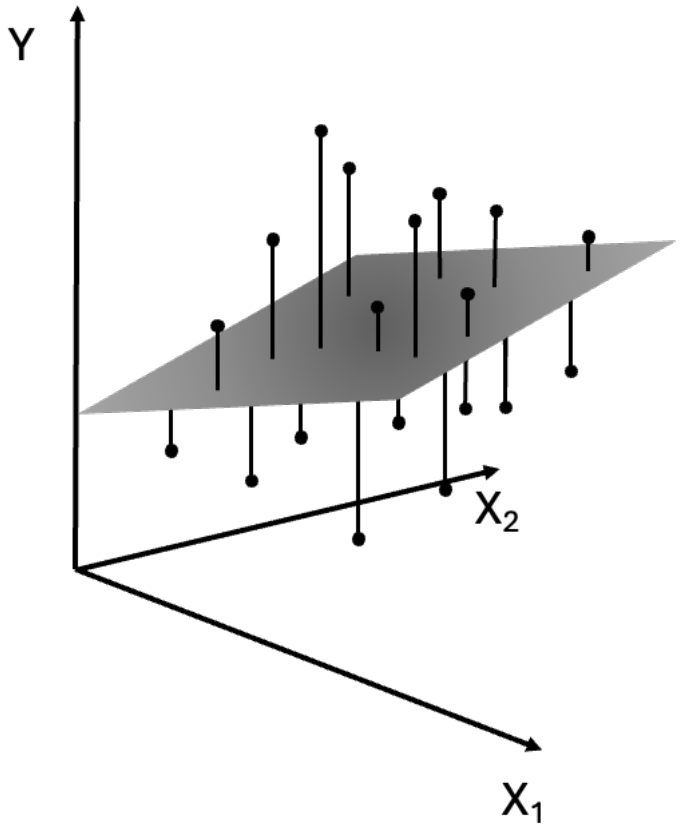


Abbildung 3: Lineare Anpassung mithilfe der Methode der kleinsten Quadrate in $\mathbb{R}^2$. Vorhersage eines quantitativen Wertes Y auf Basis der Variablen X_1 und X_2. Nach Hastie et al. (2009).

Um Overfitting zu vermeiden, können verschiedene Formen der Regularisierung eingesetzt werden, die das Modell für zu hohe Komplexität bestrafen, beispielsweise für hohe Parameter-Werte. So nutzt Ridge Regression (Hoerl and Kennard, 1970) die L2-Norm der Parameter.

1.3.1.2 Support Vector Regression

Support Vector Regression ist eine spezielle Variante des Einsatzes von Support Vector Machines (SVM). SVMs wurden in den 1990er Jahren auf Basis der Vorarbeiten von Vladimir Vapnik entwickelt (Cortes and Vapnik, 1995).

Die zugrundeliegende Idee in der Lösung eines Klassifikationsproblems mit einer Support Vector Machine besteht darin, in einem d-dimensionalen Raum

eine Hyperebene zu berechnen, die Datenpunkte verschiedener Klassen mit dem größtmöglichen Abstand (margin) zu den ihr am nächsten liegenden Datenpunkten separiert (Abbildung 4). Die Datenpunkte, die sich in genau diesem Abstand zur Hyperebene befinden, werden als support vectors (Stützvektoren) bezeichnet (James et al., 2013).

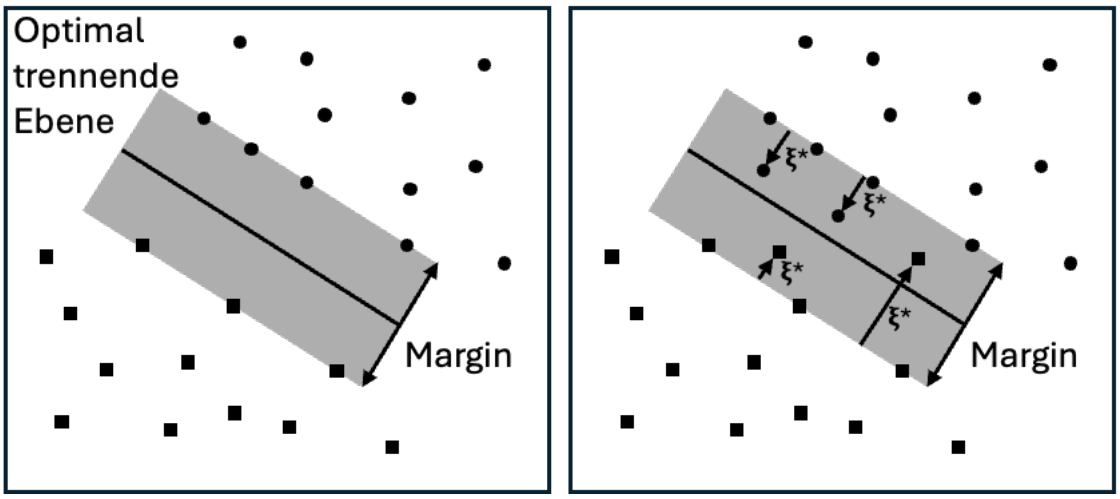


Abbildung 4: Prinzip der Support Vector Machine. Nach Hastie et al. (2009).
Links: Optimale trennende Hyperebene trennt die linear separierbaren Klassen. Die Punkte, welche auf dem Rand der margin liegen, stellen Stützvektoren dar.
Rechts: Optimale trennende Hyperebene bei nicht linear separierbaren Klassen. Einige Punkte liegen auf der falschen Seite der margin und fließen als Fehler in die Berechnung der Hyperebene mit ein.

Manche Mengen von Datenpunkten lassen sich im d-dimensionalen Raum nicht linear separieren. Ein Lösungsansatz für dieses Problem liegt darin, die Falschklassifikation einiger Punkte zu erlauben und dafür eine „Strafe“ einzuführen. In Abbildung 4 sind im rechten Bild solche falsch klassifizierten Punkte, die auf der falschen Seite der margin der Hyperebene liegen, mit den Strafen ξ versehen.

Auch kann es bei linear nicht separierbaren Klassen hilfreich sein, die Daten in einen höherdimensionalen Raum abzubilden und dort eine Hyperebene zu suchen. Dies ist von der Rechenleistung her aufwändig, weshalb sich Support Vector Machines hier des sogenannten Kernel Tricks bedienen, der eine effizientere Lösung ermöglicht. Kernel-Funktionen lassen es zu, Skalarprodukte

in einem höherdimensionalen Raum zu berechnen, ohne die spezifische Transformation in diesen Raum zu kennen (James et al., 2013).

Modifizierte Varianten können auch Regressions-Probleme lösen. Hierbei wird versucht, eine Hyperebene zu finden, in deren ε-Umgebung möglichst viele Datenpunkte liegen, also eine möglichst hohe Annäherung an die Datenpunkte zu modellieren.

1.3.1.3 Neural Networks

Heutige Neural Networks sind eine Weiterentwicklung des Perceptrons, das 1957 von Frank Rosenblatt entwickelt wurde. Dieses bestand aus einer Eingabe- und Ausgabe-Schicht von „Neuronen“ (Rosenblatt, 1957).

In Neural Networks fließen Informationen „vorwärts“ vom Input-Layer zum Output-Layer. Die Neuronen einer jeden Schicht (Layer) erhalten eine Eingabe und bilden diese auf eine Ausgabe ab. Jedes Neuron hat ein spezifisches Gewicht und einen Bias, die zusätzlich zur Eingabe zur Berechnung der Aktivierung des Neurons herangezogen werden. Mithilfe einer Aktivierungsfunktion wird die Aktivierung in eine Ausgabe transformiert, die wiederum als Input für die nächste Schicht von Neuronen eingesetzt wird. Dieser Prozess versucht, den biologischen Prozess der Aktivierung eines Neurons durch ein Aktionspotential nachzubilden (Ravì et al., 2016). Das Hinzufügen weiterer „versteckter“ („hidden“) Schichten zwischen Ein- und Ausgabeschicht erzeugt „tiefe“ („deep“) Neural Networks.

Ein Netzwerk lernt eine Abbildung von einem Input auf einen Output, so wie jeder einzelne Layer eine Abbildung von seinem Input auf seinen Output lernt. Durch die Kombination der Abbildungen und nichtlinearer Aktivierungsfunktionen können komplexe Funktionen modelliert werden (Goodfellow et al., 2016, Zell, 2023).

Die finale Ausgabe sowie das gewünschte Ergebnis werden mithilfe einer Loss-Funktion verglichen, so dass ein Fehler berechnet werden kann. Dieser ermittelte Fehler wird mittels des Verfahrens der Backpropagation genutzt, um Gewicht und Bias der Neurone anzupassen und entlang des Gradienten der

Loss-Funktion abzusteigen. Dieser Vorgang wird durch einen Optimizer vermittelt (Goodfellow et al., 2016). Dadurch wird die gelernte Abbildung besser an den Trainingsdatensatz angepasst.

1.3.1.4 Convolutional Neural Networks

Convolutional Neural Networks sind spezialisiert auf die Verarbeitung von Datensätzen, die eine gitternetzartige räumliche Struktur besitzen (Goodfellow et al., 2016) und zeigen eine überlegene Performance gegenüber anderen Methoden des Maschinellen Lernens in einer Vielzahl von Anwendungen der Computer-Vision (Selvaraju et al., 2017). Der Aufbau ist inspiriert von der Forschung am visuellen Cortex der Säugetiere (Ravì et al., 2016).

Während bei konventionellen Verfahren des Maschinellen Lernens zuerst durch den Untersucher Features aus den Rohdaten extrahiert werden müssen, bevor sie verwendet werden können, findet bei Convolutional Neural Networks die Feature-Extraktion durch das Modell selbst statt (Yamashita et al., 2018). Die Feature-Extraktion ist ebenfalls einem Lernprozess unterworfen, so dass neue Features gelernt und ihre Erkennung während des Trainings verbessert werden können (Ravì et al., 2016).

Convolutional Neural Networks bestehen aus einem Abschnitt, der der Feature-Extraktion aus den Input-Daten dient und einem Neural Network, das aus den extrahierten Features einen Output generiert (LeCun et al., 2015). Der schematische Aufbau eines Convolutional Neural Networks mit den verschiedenen Funktionseinheiten ist in Abbildung 5 gezeigt.

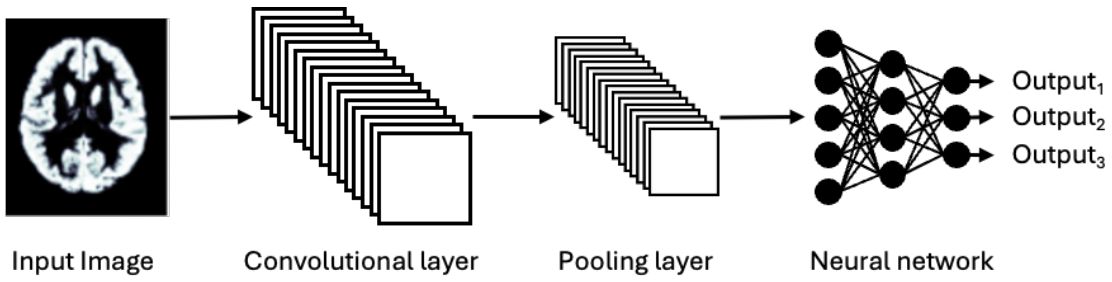


Abbildung 5: Schematischer Aufbau eines Convolutional Neural Networks. Nach Maeda-Gutiérrez et al. (2020).

Convolutional Layer erzeugen mithilfe eines Faltungskerns durch die lineare Operation der diskreten Faltung eine Feature-Map der Input-Daten (Goodfellow et al., 2016). Ein Faltungskern hat dieselbe Dimension wie die Input-Datensätze. Er „wandert“ über das Input-Bild (siehe Abbildung 6).

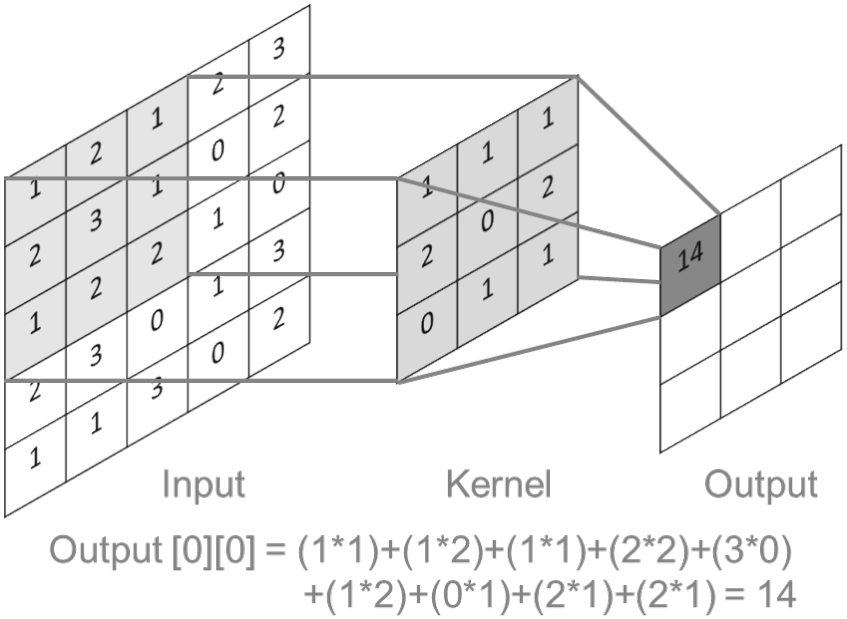


Abbildung 6: Funktion eines Faltungs-Kerns, hier Kernel genannt.

Die Summe der Produkte jedes Elementes des Kernels mit jedem Element des überlappenden Bereiches wird berechnet und als Wert an der entsprechenden Stelle der Output-Feature-Map verwendet (Dumoulin and Visin, 2016). Der Faltungskern ist kleiner als das Input-Bild, so dass kleine, lokale Features erkannt werden können. Die Fokussierung auf diese kleinen, bedeutungsvollen Features senkt zudem den Berechnungsaufwand (Goodfellow et al., 2016). Die Faltungskerne können im Lernprozess angepasst werden, um Features zu erkennen, die bedeutsam für die spezifische Anwendung sind (Yamashita et al., 2018).

Nach dem Convolutional Layer wird eine nicht-lineare Aktivierungs-Funktion genutzt (Yamashita et al., 2018). Daran schließt ein Pooling-Layer an.

Pooling hat die Eigenschaft, die Werte einer Subregion einer Feature-Map zusammenzufassen. Max-Pooling gibt die maximalen Intensitäts-Werte einer

Subregion als Output zurück, während Average-Pooling den Durchschnitt der Werte einer Subregion als Output erzeugt (Goodfellow et al., 2016).

Der Einsatz des Poolings hat zwei Vorzüge: Einerseits wird eine Invarianz gegenüber Translationen, Verschiebungen und Verzerrungen erzeugt (Yamashita et al., 2018, Goodfellow et al., 2016). Dies ist sinnvoll, wenn auf einem Bild ein Feature erkannt werden soll, allerdings unerheblich ist, an welcher Stelle des Bildes es sich befindet (Goodfellow et al., 2016). Andererseits wird die Menge an Parametern reduziert, was den erforderlichen Rechenaufwand senkt.

Durch eine Hintereinanderschaltung von Convolutional Layers, nicht-linearen Aktivierungsfunktionen und Pooling Layers kann eine Hierarchie von Features aus dem Input-Bild extrahiert werden. In den ersten Schichten werden Features niedrigerer Ordnung, wie Kanten und Kurven identifiziert, die auf den späteren Schichten zu komplexeren Features höherer Ordnung zusammengesetzt werden (Goodfellow et al., 2016). So können die tieferen Schichten eines Convolutional Neural Networks hochgradige visuelle Konstrukte erfassen (Bengio et al., 2013). Schließlich wird aus der Output-Feature-Map des letzten Pooling Layers ein eindimensionaler Feature-Vektor erzeugt, der als Input für ein Neural Network verwendet wird (Yamashita et al., 2018). Hier wird der Feature-Vektor auf den finalen Output des Netzwerkes abgebildet.

Der Output wird mit dem erwarteten Ergebnis verglichen und der aufgetretene Fehler mittels des Verfahrens der Backpropagation zurück durch das Netzwerk gesendet, um dessen Parameter anzupassen. Diese Anpassung vermittelt ein spezieller Optimierer, der auf der Fehlerfunktion den Weg des optimalen Abstieges berechnet, um den Fehler bei den nächsten Vorhersagen zu reduzieren. Dadurch können mithilfe des Feature Extractors neue Features, die für gute Vorhersagen nützlich sind, gelernt und die Parameter des Neural Networks besser an die Features im erzeugten Feature Vektor angepasst werden, um die Performance des Netzwerkes zu verbessern.

1.3.2 Herausforderungen beim Einsatz von Modellen des Maschinellen Lernens im Gesundheitssystem

Evidenzbasierte Medizin umfasst den verantwortungsvollen Einsatz der aktuell besten Evidenz bei der Behandlung von Patienten (Sackett, 1997).

Modelle des Maschinellen Lernens können in eng umrissenen Anwendungen bereits heute vergleichbar gute oder sogar bessere Vorhersagen treffen als menschliche Ärzte (Esteva et al., 2017, Gulshan et al., 2016, Langner et al., 2019, Ehteshami Bejnordi et al., 2017). Aus besseren diagnostischen Vorhersagen kann ein Nutzen für den Patienten abgeleitet werden. Gleichzeitig könnte durch den Einsatz von Modellen des Maschinellen Lernens die Arbeitslast effizienter bewältigt werden, weshalb ihr Einsatz im klinischen Alltag höchst attraktiv ist.

Das Gesundheitswesen ist eine kritische Infrastruktur, in der Fehler teilweise bedeutende Folgen nach sich ziehen. Ärzte tragen daher eine Verantwortung für das Wohlergehen der Patienten, die sich in ihre Fürsorge begeben. Dementsprechend notwendig ist es, dass Ärzte das Werkzeug verstehen können, das ihnen beim Treffen wichtiger klinischer Entscheidungen behilflich sein soll. Die Intransparenz einiger Verfahren des Maschinellen Lernens ist hier ein limitierender Faktor ihres Einsatzes.

Vorhersagen von Modellen des Maschinellen Lernens werden automatisiert aus den präsentierten Daten ohne explizites medizinisches Fachwissen abgeleitet und stellen nur eine Annäherung an eine optimale Modellierungsfunktion dar. Somit ist eine kausale Bedeutung nicht unbedingt gegeben (Prosperi et al., 2020). Stattdessen können die getroffenen Vorhersagen eines Modells auf der Basis konfundierender Faktoren oder sogar des zufälligen Rauschens entstehen. Fehlen nun diese konfundierenden Faktoren in zuvor ungesehenen Daten, so kann das trainierte Modell auf selbigen nur noch eingeschränkt sinnvolle Vorhersagen treffen. Dies stellt ein Sicherheitsrisiko für die großflächige Verwendung von Verfahren des Maschinellen Lernens in der kritischen Infrastruktur des Gesundheitswesens dar.

Ein interessantes Beispiel für ein solches Phänomen ist die automatisierte Analyse von Thorax-Röntgen-Aufnahmen, um Pneumonien zu diagnostizieren (Zech et al., 2018). In dieser Arbeit wurde ein Convolutional Neural Network auf Trainingsdatensätzen aus drei verschiedenen Krankenhäusern trainiert. Hierbei wurde festgestellt, dass die Prävalenz von Pneumonien stark unterschiedlich war und der Algorithmus allein durch die Identifizierung des Krankenhauses, in dem die Aufnahme gemacht wurde, eine gute Vorhersage-Performance erreichen konnte.

In diesem Zusammenhang kamen die Autoren zu dem Schluss, dass nicht nur die zu erkennende Pathologie zur Einschätzung des Netzwerkes beitrug, sondern auch viele konfundierende Variablen, zum Beispiel ob eine Aufnahme mit einem mobilen Gerät auf einer Station gemacht werden musste oder auch kleine Metall-Token, die bei bettlägerigen Patienten die Seite markierten, und beide ein Indiz für eine schwerere Krankheit des Patienten geben konnten. Die Unterscheidung der Krankenhäuser, in denen die Bilddaten erzeugt wurden, sei anhand subtiler einrichtungsspezifischer Charakteristika möglich gewesen, die durch spezifische verwendete Bildgebungs-Protokolle und -Parameter beeinflusst worden seien.

Diese konfundierenden Variablen fänden sich nicht unbedingt außerhalb der vorbereiteten Trainings- und Testdatensätze der Arbeit und helfen nicht bei der Erkennung der Pathologie, so dass die Performance auf ungesehenen Datensätzen der echten Welt vermutlich überschätzt würde. Ohne Möglichkeit, die inneren Abläufe des trainierten Modells zu visualisieren, wäre verborgen geblieben, dass das Modell nur sehr eingeschränkt in der Lage war, tatsächlich die bildmorphologischen Korrelate der Pneumonie zu erkennen, obwohl es in Tests gute Ergebnisse gezeigt hatte.

1.3.3 Explainable AI und dessen Bedeutung für die Medizin

Ein Anwender in einer kritischen Infrastruktur muss sich bei der Arbeit auf seine Hilfsmittel verlassen können. Dazu gehört auch ein Verständnis für ihre Funktionsweise, um oben beschriebene Fehlfunktionen ausschließen zu

können. Erklärbarkeit fördert Vertrauen in die eingesetzte Technologie und ermöglicht eine Nachprüfbarkeit von Vorhersagen.

Dass die Funktionsweise von Modellen des Maschinellen Lernens nicht direkt von Menschen nachvollzogen werden können, stellt allerdings nicht nur ein Problem, sondern gleichzeitig ein großes Potential für die medizinische Forschung dar.

Da beispielsweise Convolutional Neural Networks selbst die Features auswählen, die für die Lösung der gestellten Aufgabe am hilfreichsten sind, können hier zum Beispiel in der Analyse medizinischer Bildgebungsdaten Regionen und Strukturen, automatisiert und in größeren Datensätzen, identifiziert werden, die der Entdeckung durch den menschlichen Untersucher verborgen bleiben. Auch hierzu ist es vonnöten, diese Features und ihren Beitrag zu den Vorhersagen des Netzwerks für Menschen verständlich darzustellen.

Es existieren bereits Ansätze zur Lösung des in 1.3.2 ausgeführten Problems, die sich unter dem Schlagwort „Explainability“ oder auch „explainable AI“ zusammenfassen lassen.

Bei linearen Verfahren (Ridge Regression, Support Vector Regression mit linearem Kernel) ist es recht einfach möglich, Einblicke zu gewinnen, da direkt die Koeffizienten des trainierten Modells untersucht werden können, um festzustellen, welche Features der eingesetzten Daten die Vorhersage beeinflusst. Dagegen mussten für Convolutional Neural Networks, die ein komplexeres Verfahren darstellen, neue Verfahren entwickelt werden.

Ein Beispiel ist Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Mithilfe dieses Algorithmus können Regionen im Input-Bild identifiziert werden, die für die Vorhersage eine große Rolle gespielt haben. Hierdurch können Muster gefunden werden, die unter Umständen Rückschlüsse darauf zulassen, anhand welcher Merkmale das trainierte Modell Vorhersagen trifft. Abbildung 7 zeigt die Anwendung von Grad-Cam auf einem Netzwerk, dass für die Klassifizierung von Strukturen auf Bildern trainiert wurde.

Wie in 1.3.1.4 beschrieben, können die letzten Convolutional Layer eines Convolutional Neural Networks hochgradige visuelle Konstrukte erfassen (Bengio et al., 2013), während gleichzeitig die räumlichen Informationen erhalten bleiben, die im Feature Vector verloren gehen. Grad-CAM verwendet die Gradienten-Informationen des letzten Convolutional Layer, um die Wichtigkeit eines Neurons für eine bestimmte Entscheidung zu berechnen (Selvaraju et al., 2017). Diese kann visuell als Karte hervorstechender Merkmale („Saliency Map“) dargestellt werden (Abbildung 7).

Zusammenfassend lässt sich feststellen, dass die Anwendung von explainable AI und die Entwicklung von Ansätzen, um Erklärbarkeit zu erreichen, zweierlei Vorteile aufweist: Einerseits kann es das Verständnis dieser neuen Technologie verbessern und den Einsatz in der kritischen Infrastruktur des Gesundheitssystems rechtfertigen; Andererseits können mithilfe von Modellen des Maschinellen Lernens Eigenschaften von großen, komplexen Datenmengen identifiziert werden, die für Menschen schwer zugänglich sind.

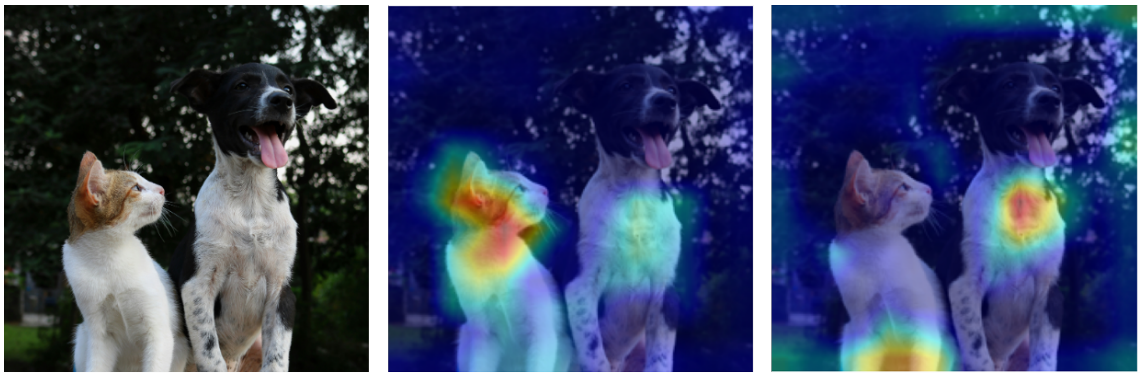


Abbildung 7: Hervorhebung der Regionen, auf deren Basis ein Convolutional Neural Network eine Struktur auf dem Originalbild der Klasse „Katze“ oder „Hund“ zuteilt. Nach Selvaraju et al. (2017).

1.4 Aktuelle Arbeiten im Bereich der automatisierten Gehirn-Altersschätzung

Einige morphologische Merkmale eines alternden Gehirns sind in medizinischen Bilddaten sichtbar. Dazu gehören beispielsweise Abnahmen im Volumen der Grauen Substanz und der kortikalen Parenchymdicke, während

das Volumen des Liquors zunimmt (Pfefferbaum et al., 1994, Sullivan and Pfefferbaum, 2007). Diese Effekte beginnen bereits ab einem Alter von nur 20 Jahren. Der Anteil des Volumens der grauen Substanz hat seinen Höhepunkt sogar im Alter von nur vier Jahren (Pfefferbaum et al., 1994).

In T2-gewichteten Magnetresonanztomographie (MRT)-Aufnahmen lassen sich mit steigendem Alter zunehmend White Matter Hyperintensities darstellen. Es besteht eine Beziehung zwischen der Belastung mit diesen Läsionen und dem Vorliegen von Demenz und kognitiven Einschränkungen (Habes et al., 2016).

Hoagey et al. (2019) versuchten, die Gehirnregionen zu identifizieren, die durch „gesundes“ Altern am stärksten beeinflusst wurden. Die Patienten wurden entsprechend danach selektioniert, keine neurologischen, kardiovaskulären, metabolischen oder psychiatrischen Erkrankungen aufzuweisen, sowie keine Traumata des Kopfes mit Bewusstlosigkeit erlitten und keine Behandlung mit Medikamenten mit kognitiv verändernder Wirkung erhalten zu haben. Der Zustand der Grauen Substanz wurde mithilfe von Messungen der kortikalen Parenchymdicke, dem Volumen in verschiedenen Regionen sowie der Oberfläche, die Weiße Substanz mithilfe der fractional anisotropy und mean diffusivity quantifiziert. Die altersabhängigen Veränderungen der Grauen und Weißen Substanz waren am höchsten in den frontalen Regionen des Gehirns, während sie im temporalen und okzipitalen Gehirn kleiner waren. Somit könnte durch Vermessung spezieller, durch das Altern besonders beeinflusster Hirnregionen eine morphometrische Altersschätzung erfolgen.

Jedoch lassen sich auch weitere morphologische Merkmale einsetzten. In einer Arbeit von Liem et al. (2017) wurden zerebrale Konnektivitäts-Matrizen, kortikale Parenchymdicke, kortikale Oberflächen- sowie subcortikale Volumenmessungen für jeweils eine Support Vector Regression genutzt. Die Vorhersagen der fünf einzelnen Modelle wurden kombiniert und von einem multi-source Random Forest Modell ausgewertet. Damit konnte ein Mean Absolute Error (MAE) von 4,29 Jahren erreicht werden.

Die meistgenutzte Modalität der Gehirnaltersschätzung ist bislang das MRT. Jedoch wurden auch Bildgebungsdaten anderer Modalitäten und sogar

Elektroenzephalographie (EEG)-Aufnahmen verwendet, um das Gehirn-Alter von Patienten zu schätzen (Cole et al., 2019).

Huang et al. (2017) nutzten eine CNN-Architektur auf der Basis des VGG16-Netzwerks (Huang et al., 2017). Von 600 MRT-Untersuchungen von Patienten zwischen 20 und 80 Jahren wurden jeweils 15 Schichten zum Training ausgewählt. Das trainierte Netzwerk war in der Lage, das Patientenalter mit einem MAE von 4,0 Jahren einzuschätzen.
Brudfors (2020) nutzte ein Gaussian Process Model für Gehirn-Altersschätzung auf CT-Bildern. Das Modell erreichte einen Root Mean Squared Error von 5,21 Jahren auf einem Datensatz von Patienten zwischen 60 und 100 Jahren (Brudfors, 2020b).

Auch Armanious et al. (2021) stellten ein Verfahren für automatisierte Schätzung des Chronologischen Alters vor. Das trainierte Modell wies einen MAE von 2,0 Jahren auf dem selbst akquirierten Datensatz und einen MAE von 3,6 Jahren auf dem OASIS-3 Datensatz (LaMontagne et al., 2019) auf. Auch konnte eine Korrelation zwischen einem erhöhten Gehirnalter und einer erwarteten Verschlechterung der Gehirnfunktion festgestellt werden.

Hepp et al. (2021) trainierten ein 3D-Convolutional Neural Network auf T1-gewichteten MRT-Gehirndatensätzen aus der NAKO-Studie (Bamberg et al., 2015). Hier konnte bei einer Altersspanne zwischen 20 und 72 Jahren, wobei die meisten Patienten allerdings zwischen 40 und 70 Jahren alt waren, ein MAE von 3,2 Jahren erreicht werden. Zudem wurden mittels Grad-Cam für die Altersschätzung wichtige Regionen identifiziert. Diese umfassten zentrale Regionen des Gehirns im Bereich des dritten Ventrikels, der Basalganglien und der Seitenventrikel. Zudem wurden Einschätzungen der Modell-Unsicherheit in der erwähnten Arbeit untersucht.

Mithilfe von Messungen des „biologischen“ Gehirnalters können sowohl der körperliche Zustand insgesamt besser als durch das Chronologische Alter allein eingeschätzt, als auch Hinweise auf das Vorliegen systemischer Erkrankungen gefunden und Vorhersagen über zukünftige Erkrankungsrisiken getroffen werden. Dementsprechend könnte es dem Patientenwohl zuträglich sein,

Patienten mit einem erhöhten biologischen Gehirnalter frühzeitig zu identifizieren, um präventiv tätig werden zu können.

So zeigte sich ein gegenüber dem Chronologischen Alter erhöhtes Gehirn-Alter in der Literatur assoziiert mit Anzeichen körperlichen Alterns, darunter einer schwächeren Griff-Stärke, schlechteren Lungenfunktion, langsameren Geh-Geschwindigkeiten, einer niedrigeren „fluiden“ Intelligenz und sogar einem erhöhten Mortalitätsrisiko (Cole et al., 2018). Patienten mit einem erhöhten biologischen Gehirnalter waren mit neurodegenerativen Erkrankungen assoziiert (Armanious et al., 2021, Cole and Franke, 2017). Zudem konnte gezeigt werden, dass verschiedene Erkrankungen, darunter HIV (Chang et al., 2008), Schizophrenie (Hajek et al., 2019) und auch Diabetes (Franke et al., 2013) zu einem höheren messbaren Gehirn-Alter führen.

1.5 Wissenschaftliche Zielsetzung

Für diese Arbeit wurden folgende Ziele definiert:

1. Training, Validierung und Testung verschiedener Methoden des Maschinellen Lernens für die Aufgabe der automatisierten Altersschätzung anhand von axialen Weichteil-CT-Scans des Gehirns die nach klinischen Standards erzeugt wurden.
2. Die Identifikation der für die Gehirn-Altersschätzung wichtigsten morphologischen Features und Gehirn-Regionen in den verwendeten Bilddaten, um Bildgebungs-Biomarker für das „gesunde“ Altern des Gehirns zu finden.

2 Material und Methoden

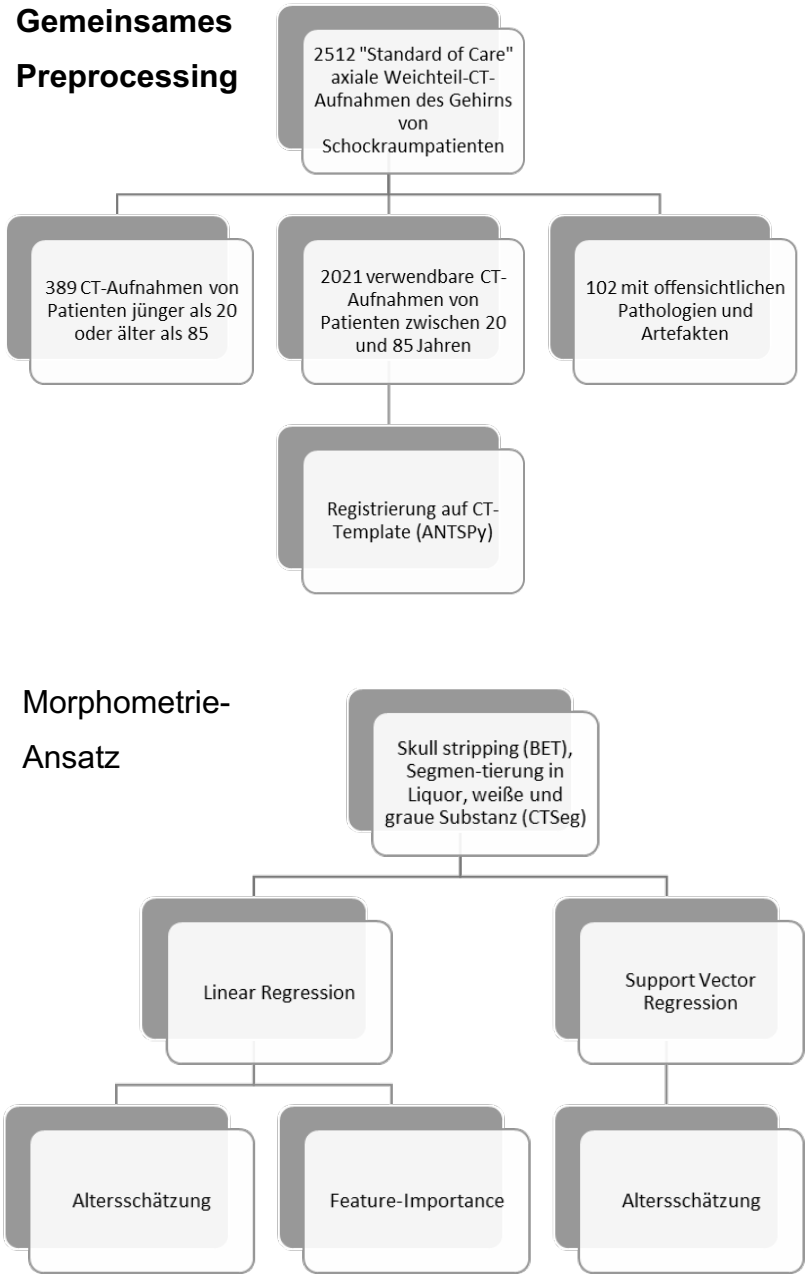
Ein feature-basierter Morphometrie-Ansatz und ein auf Deep Learning basierender Ansatz zur automatisierten Gehirn-Altersschätzung auf der Basis von dem Behandlungsstandard entsprechenden klinischen Bildgebungsdaten wurden eingesetzt.

Zur Identifikation von Bildgebungs-Biomarkern wurden Modelle des Maschinellen Lernens darauf trainiert, das Chronologische Alter von „gesunden“ Patienten vorherzusagen. Bei einer ausreichend großen Stichprobe gesunder

Patienten konvergiert das geschätzte Chronologische Alter gegen das Biologische Alter. Somit kann das Biologische Alter von Patienten auf Basis dieser Stichprobe eingeschätzt werden.

2.1 Übersicht

Abb. 8 zeigt eine Übersicht der in dieser Arbeit verfolgten Ansätze.



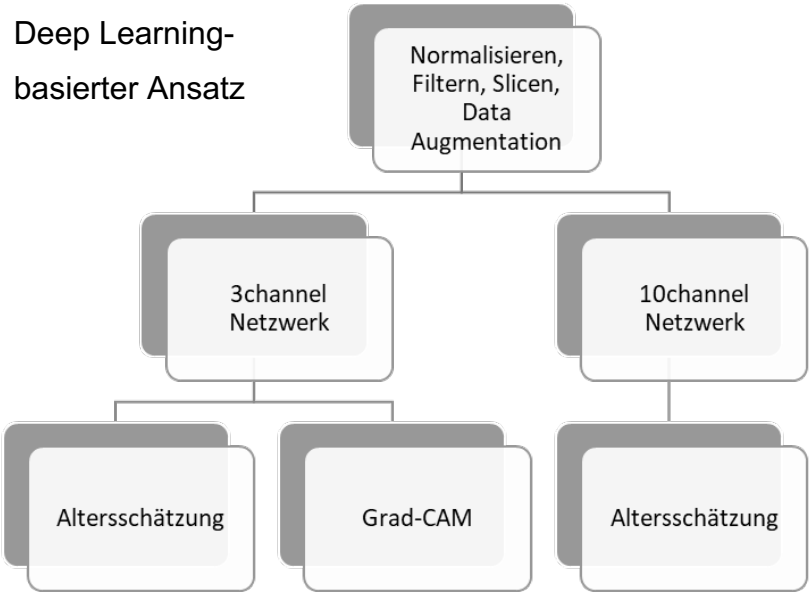


Abbildung 8: Übersicht über das gemeinsame Preprocessing, den verfolgten Morphometrie- und den Deep Learning-Ansatz

2.2 Patientenkollektiv

In dieser Arbeit wurde ein Patientenkollektiv retrospektiv untersucht. Die verantwortliche Ethik-Kommission gab ihr Einverständnis für die retrospektive Analyse der persönlichen und Bilddaten der ausgewählten Patienten (Projekt-Nr. 387/2020BO, Votum vom 3. Juni 2020).

Patienten zwischen 20 und 85 Jahren, bei denen zwischen 2010 und 2019 im Schockraum der Universität Tübingen ein axiales Weichteil-CT des Gehirns durchgeführt wurde, wurden in diese Arbeit eingeschlossen.

Ausschlusskriterien umfassten offensichtliche Pathologien und Artefakte in den Bildgebungsdaten, die für eine automatisierte Auswertung nicht geeignet waren.

Die Gehirn-Bildgebungs-Datensätze von 2512 Patienten wurden aus dem PACS heruntergeladen. 102 Patienten wurden ausgeschlossen, da ihre Gehirn-Scans offensichtliche Pathologien oder Artefakte enthielten. Zudem wurden Patienten aus dem Datensatz entfernt, die jünger als 20 Jahre oder älter als 85 Jahre waren, dies betraf insgesamt 389. Somit wurden 2021 Patienten für den gesamten Datensatz selektiert.

2.3 Eigenschaften des Gesamtdatensatzes: Alters- und Geschlechtsverteilung, verwendete Scanner, Reconstruction-Kernel und Schichtdicken

Die Altersverteilung im Patientenkollektiv ist in Abbildung 9 gezeigt.

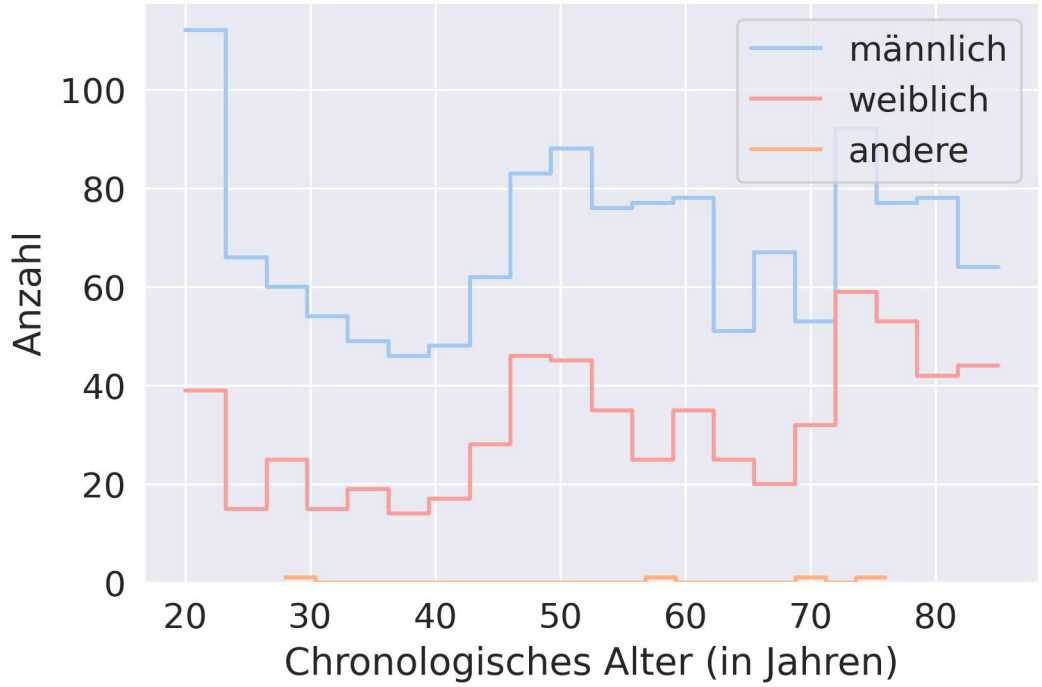


Abbildung 9: Altersverteilung der männlichen (blau) und weiblichen (rot) Patienten im Datensatz.

In Orange sind die Patienten dargestellt, deren Geschlecht nicht bestimmt war.

Von den eingeschlossenen Patienten waren 1381 (68,33%) männlichen und 633 (31,32%) weiblichen Geschlechts, während von 7 (0,35%) Patienten das Geschlecht nicht bestimmt war. Das Geschlechterverhältnis betrug 2,18 männliche Patienten auf jeden weiblichen Patienten. Das mittlere Alter der männlichen Patienten betrug 52,7 Jahre, während das mittlere Alter der weiblichen Patienten 56,9 Jahre betrug.

Abbildung 10 zeigt die Verteilung der verwendeten CT-Scanner, Kernel, die zur Rekonstruktion der CT-Daten verwendet wurden sowie Schichtdicken. Der am häufigsten genutzte CT-Scanner war der Siemens Sensation 64, gefolgt vom Siemens Somatom Definition AS+ und Sensation 16.

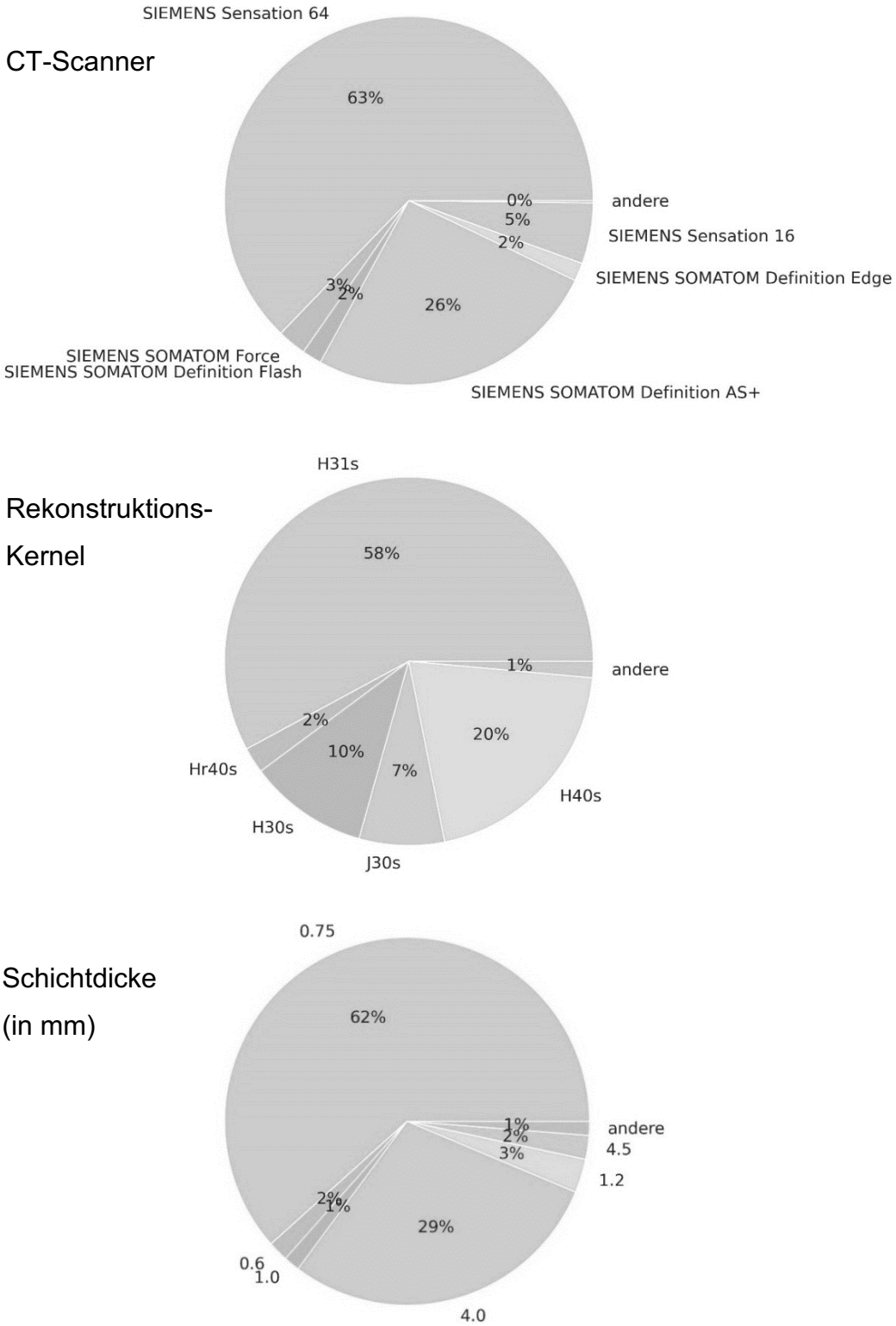


Abbildung 10: Verwendete Scanner, Kernel und Schichtdicken (in mm) im Datensatz.

Zudem wurden der Siemens Somatom Force, Somatom Definition Edge und Flash verwendet. Die am häufigsten genutzte Schichtdicke betrug 0,75 mm, die

zweitmeiste 4,0 mm. 0,6 mm, 1,0 mm, 1,2 mm und 4,5 mm wurden ebenfalls verwendet. Der am häufigsten genutzte Rekonstruktions-Kernel war H31s, gefolgt von H40s, H30s und J30s.

2.4 Eigenschaften des Testdatensatzes: Alters- und Geschlechtsverteilung, verwendete Scanner, Reconstruction-Kernel und Schichtdicken

Die Altersverteilung im Testdatensatz ist in Abbildung 11 dargestellt.

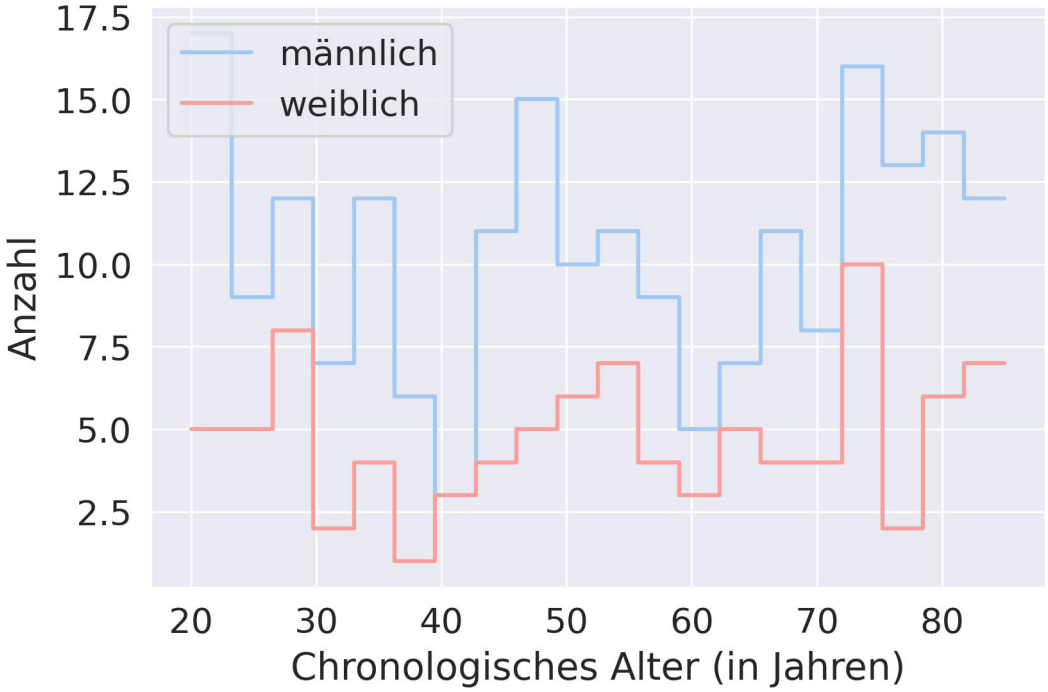


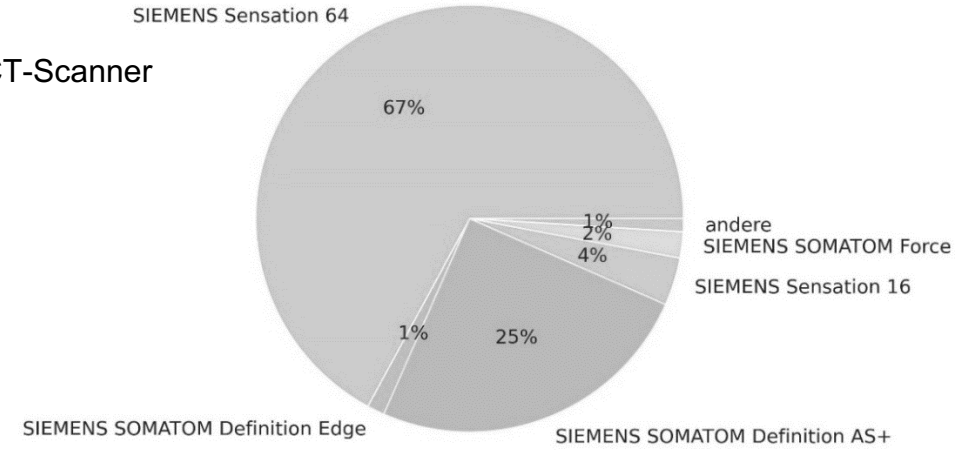
Abbildung 11: Alters- und Geschlechtsverteilung im Testdatensatz.

304 Patientendatensätze wurden zufällig dem Test-Datensatz zugeteilt, auf dessen Basis die Performance der verschiedenen Ansätze bewertet wurde. Die Zuteilung zum Testdatensatz wurde zufällig mit der Funktion train_test_split der Python-Bibliothek scikit-learn vorgenommen. Die Altersverteilung entsprach weitestgehend dem Trainings-Datensatz.

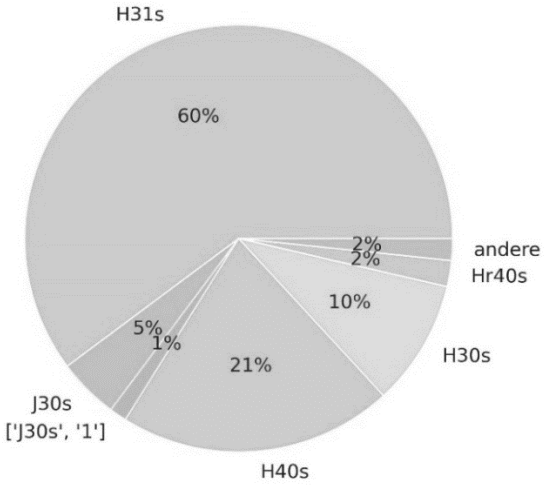
Der Anteil an männlichen und weiblichen Patienten war ebenfalls vergleichbar, auch wenn das Durchschnittsalter der weiblichen Patienten niedriger war als im gesamten Datensatz (53,9 Jahre im Testdatensatz gegenüber 56,9 Jahren im Gesamtdatensatz). 68,65% der Patienten waren männlichen Geschlechtes

während 31,35% weiblichen Geschlechtes waren. Abbildung 12 bildet in Kreisdiagrammen die verwendeten CT-Scanner, Reconstruction Kernel und Schichtdicken im Testdatensatz ab.

CT-Scanner



Rekonstruktions-Kernel



Schichtdicke (in mm)

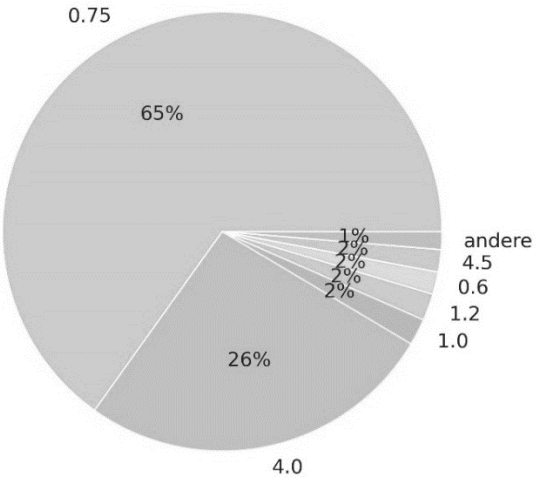


Abbildung 12: Verwendete CT-Scanner, Reconstruction Kernel und Schichtdicken (in mm) im Testdatensatz

Die meisten Aufnahmen wurden mit einem Siemens Sensation 64 CT-Scanner aufgenommen. Am häufigsten wurde der H31s Reconstruction Kernel verwendet. Die häufigste verwendete Schichtdicke waren 0,75 mm.

2.5 Bildakquisition, klinische und technische Datenerhebung

Es wurden axiale Weichteil-CT-Aufnahmen des Schädels aus dem PACS der Universitätsklinik Tübingen verwendet. Die Aufnahmen wurden mit verschiedenen CT-Scannern der Firma Siemens im Schockraum der Universitätsklinik Tübingen erzeugt. Dabei wurden verschiedene Rekonstruktions-Kernel, Schichtdicken und Röhrenspannungen verwendet.

Die Namen, Geburtstage, Patienten-IDs, Altersangaben und der Zeitpunkt der Bild-akquisition der eingeschlossenen Patienten wurden zusammengetragen.

Die Bilddatensätze wurden im DICOM-Format aus dem PACS bezogen und die DICOM-Metadaten ausgelesen, um Informationen über die zur Bildakquise verwendeten Geräte, Rekonstruktions-Kernel und Schichtdicken (in mm) sowie die Exposure (in mAs) und Röhrenspannung (in kV) zu erhalten.

2.6 Verwendete Software

Die Software für Preprocessing, Model-Training und Performance-Tests wurde in der Programmiersprache Python, Version 3.7.4, geschrieben. Die verwendeten Python-Bibliotheken und ihre Anwendungen sind in Tabelle 1 in alphabetischer Reihenfolge zusammengetragen.

Tabelle 1: Verwendete Python-Bibliotheken und ihre Anwendung im Rahmen der Arbeit.

Bibliothek	Anwendung	Referenz
antspy (v.0.2.0)	Registrierung	(Avants et al., 2009)
dicom2nii (v.2.2.1)	Konvertierung der DICOM-Daten-sätze in das Nifti-Format	(Li et al., 2016)

h5py (v.3.1.0)	Interface zur Verwendung des hdf5-Datenformates	(Collette, 2017)
jupyter (v.1.0.0)	Infrastruktur für Programmierarbeit	(Kluyver et al., 2016)
matplotlib (v.3.3.3)	Visualisierung von Ergebnissen	(Hunter, 2007)
nibabel (v.3.2.0)	Transformation von Datensätzen	(Brett et al., 2020)
nilearn (v.0.7.0)	Resampling	Siehe scikit-learn.
nipype	Python-Interface für BET (Brain Extraction Tool)	(Gorgolewski et al., 2011)
numpy (v.1.19.4)	Array-Operationen	(Harris et al., 2020)
pandas (v.1.1.4)	Ablage von Tabellendaten	(McKinney, 2010)
pydicom	Auslesen von DICOM-Metadaten	(Mason, 2011)
pytorch (v.1.7.0)	Deep Learning	(Paszke et al., 2019)
scikit-learn (v.0.0)	Feature-basiertes Maschinelles Lernen	(Pedregosa et al., 2011)
scipy (v.1.5.4)	Bildoperationen	(Virtanen et al., 2020)

Das Preprocessing für den feature-basierten Morphometrie-Ansatz wurde im Jupyter Notebook durchgeführt (Kluyver et al., 2016). Zudem wurde MATLAB Version R2020a für die Segmentierung der Bilddatensätze genutzt. Hier wurde der CTSeg-Algorithmus verwendet (Brudfors, 2020a).

2.7 Preprocessing der Bilddaten

Die Datensätze wurden mithilfe des PACS-Interfaces der Nora Imaging Plattform (Anastasopoulos et al., 2017) aus dem PACS des Universitätsklinikums Tübingen im DICOM-Format exportiert. Die DICOM-

Dateien wurden in eine Projekt-Struktur überführt und mithilfe der Python-Bibliothek dicom2nifti (Li et al., 2016) in das Nifti-Format konvertiert.

Aus den Nifti-Datensätzen wurden einzelne Schichten zur Überprüfung auf Pathologien ausgewählt. Nach den Ausschlusskriterien selektierte Datensätze wurden aus der Projektstruktur entfernt.

Die Patientennamen, sowie Informationen über Geschlecht und Alter wurden mithilfe der Python-Bibliothek pydicom (Mason, 2011) den DICOM-Metadaten entnommen und mit dem errechneten Alter zwischen Geburtstag und Untersuchungsdatum verglichen, um die korrekten Altersinformationen sicher zu stellen. Zudem wurden die Informationen über die zur Bildakquise verwendeten Geräte, Rekonstruktions-Kernel und Schichtdicken (in mm) sowie die Exposure (in mAs) und Röhrenspannung (in kV) ausgelesen. Die Patientennamen in den verbleibenden Datensätzen wurden mithilfe von sha256-hashes der Python-Bibliothek hashlib pseudonymisiert.

Die Nifti-Datensätze wurden daraufhin auf ein CT-Template registriert, das aus dem CTseg-Projekt entnommen wurde (Brudfors, 2020a, Brudfors et al., 2020). Die Registrierung wurde mithilfe der „TRSAA“-Registrierungsmethode der Python-Bibliothek ANTsPy (Avants et al., 2009) durchgeführt, die eine hochqualitative Registrierung ermöglicht.

Die registrierten Datensätze dienten als gemeinsame Ausgangsdaten für das Preprocessing des Gehirn-Morphometrie- und des Deep Learning-Ansatzes.

Die Datensätze wurden zufällig anhand der Hashes in jeweils einen Trainings-, Validierungs- und Testdatensatz aufgeteilt. Der Trainingsdatensatz umfasste 1414 (70%) der verfügbaren Studien, der Validierungs- und Testdatensatz jeweils 303, bzw. 304 (jeweils 15%). Dazu wurde die Funktion test_train_split der Python-Bibliothek scikit-learn verwendet.

Für den Morphometrie-Ansatz wurden für Training und Validierung der Trainings- und Validierungsdatensatz gemeinsam verwendet, während der Testdatensatz vor dem Test nicht durch das Modell gesehen wurde. Für den Deep Learning-Ansatz wurde ausschließlich der Trainingsdatensatz zum

Training verwendet, während der Validierungsdatensatz und der Testdatensatz nicht eingesetzt wurden, um das Modell zu verändern, sondern nur um seine Leistung auf ungesehenen Daten zu überprüfen.

2.8 Feature-basierter Ansatz: Gehirn-Morphometrie

2.8.1 Preprocessing

In diesem Ansatz wurden die Bilddaten zuerst mit dem Brain Extraction Tool (BET) (Jenkinson et al., 2005) mithilfe des nipype-Interface (Gorgolewski et al., 2011) bearbeitet, um das Gehirn zu segmentieren und sicherzustellen, dass die Vermessungen der Gehirnregionen nicht durch Schädelknochen und andere Strukturen beeinflusst wurden.

Die segmentierten Gehirn-Datensätze wurden daraufhin mithilfe des CTseg-Algorithmus des Wellcome Center for Human Neuroscience (Brudfors, 2020a) in Matlab in Graue Substanz, Weiße Substanz und Liquor segmentiert. Dieser verwendet ein verbessertes Atlas-basiertes Verfahren, das eine robuste Erweiterung der SPM12-Segmentierungs-Routine darstellt (Brudfors, 2020b). Für jeden Patienten wurden somit segmentierungsspezifische Wahrscheinlichkeitskarten erzeugt.

Auf Basis der erzeugten Wahrscheinlichkeitskarten wurden die Volumina verschiedener Hirnregionen automatisiert vermessen. Die Gehirnregionen und ihre Ausdehnung wurden mithilfe eines modifizierten dreidimensionalen digitalen Atlas des Gehirns spezifiziert. Der verwendete Atlas ist Teil des Matlab Software-Paketes SPM12. Der Atlas wurde im Rahmen der „MICCAI 2012 Grand Challenge and Workshop on Multi-Atlas Labeling“ auf der Basis von 30 Patientendatensätzen aus der OASIS Datenbank (Marcus et al., 2007) entwickelt. Der Atlas spezifiziert 138 manuell annotierte kortikale und subkortikale Regionen. Die annotierten Daten wurden durch die Firma Neuromorphometrics, Inc. zur Verfügung gestellt. In diesem Atlas wird die Zugehörigkeit eines Voxels zu einer gelabelten Region mittels des Voxelwertes dargestellt.

Auf Basis des Atlas wurden Segmentierungs-spezifische Atlanten für Graue und Weiße Substanz erzeugt, die jedem Voxel im dreidimensionalen Bildraum eine Region des Atlas zuweist. Dazu wurden Voxel mit dem Intensitätswert 0, die keiner Region zugeordnet waren, das Label des von der Entfernung her nächsten Voxels einer gelabelten Region zugeordnet. Dieser Vorgang ist schematisch in Abbildung 13 dargestellt.

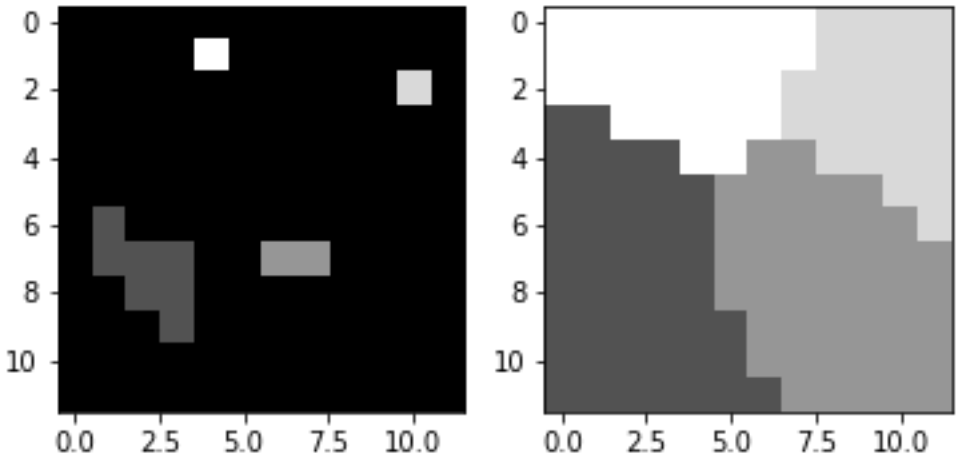


Abbildung 13: Schematische Darstellung der Zuordnung ungelabelter Regionen (schwarz) zu gelabelten Regionen (grau bis weiß) mittels nearest-neighbour-search auf k-d-Bäumen.

Links: Vor der Bearbeitung. Rechts: Nach Bearbeitung.

Technisch wurde dies durch eine nearest-neighbour-Suche auf dreidimensionalen k-d-Bäumen gelöst, die auf Basis der Voxel-Koordinaten des originalen Atlas konstruiert wurden. Dazu wurde die Implementation der Python-Bibliothek sklearn (Pedregosa et al., 2011) verwendet. Schnitte aus den modifizierten Atlanten sind in Abbildung 14 zu sehen.

Diese Modifikation war notwendig, da der Atlas lediglich eine aus MRT-Bildern der OASIS-Studie ermittelte Standard-Anatomie darstellt. Somit ergaben sich in der Überlagerung mit Patienten-Datensätzen größere anatomische Abweichungen, die zu einer Verfälschung der Regionsvolumina geführt hätten.

Abbildung 14: Zuordnung ungelabelter Regionen zu segmentierungsspezifischen gelabelten Regionen im verwendeten Gehirnatlas.
Links: Zweidimensionale Schicht aus dem dreidimensionalen Neuromorphometrics Atlas. Mitte: Zuordnung ungelabelter Regionen zu segmentierungsspezifischen gelabelten Regionen der Grauen Substanz mittels nearest-neighbour-search auf k-d-Bäumen. Rechts: Zuordnung ungelabelter Regionen zu segmentierungsspezifischen gelabelten Regionen der Weißen Substanz mittels nearest-neighbour-search auf k-d-Bäumen.

Die Wahrscheinlichkeitskarten für Graue und Weiße Substanz sowie Liquor, die durch die Segmentierung erzeugt wurden, wurden daraufhin mit einem Schwellenwert von 0,1 Einheiten gefiltert, so dass alle Intensitätswerte darunter mit dem Wert von 0 Einheiten versehen wurden, alle darüber mit dem Wert 1, um einen binären Datensatz zu erhalten. Durch voxelweise Multiplikation dieser binären Wahrscheinlichkeitskarten mit den segmentierungspezifischen Atlanten wurde jedem Voxel der Wahrscheinlichkeitskarten das Label der korrespondierenden Atlas-Region zugewiesen (Abbildung 15). Die Voxel mit einem bestimmten Label wurden ausgezählt, um das Voxelvolumen zu bestimmen.

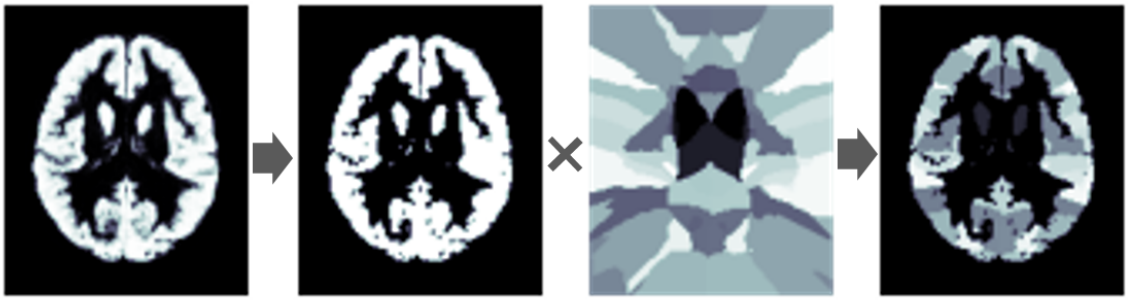


Abbildung 15: Beispiel der voxelweisen Multiplikation der Wahrscheinlichkeitskarten mit einer Segmentierung.
Aus der Wahrscheinlichkeitskarte der Segmentierung der grauen Substanz (ganz links) wird ein binärer Datensatz (mitte-links). Dieser wird voxelweise mit dem spezifischen

Atlas der grauen Substanz (mitte-rechts) multipliziert, so dass jedem Voxel der Wahrscheinlichkeitskarte oberhalb der Schwellenwertes das entsprechende Label als Intensitätswert zugewiesen wird (rechts).

Somit wurden die Voxelvolumina von 127 Regionen vermessen. Regionen, die jeweils in linke und rechte Regionen geteilt waren, wurden zu einem gemeinsamen Volumen addiert und das Ergebnis als Input-Feature-Vektor für Algorithmen des feature-basierten Maschinellen Lernens verwendet.

Auf Basis dieses Datensatzes wurde ein zweiter Datensatz erzeugt, in welchem die Volumina der Subregionen zu Regionen höherer Ordnung zusammengefasst wurden (Tabelle 2).

Die Modelle wurden sowohl auf dem Subregionen-Datensatz als auch auf dem Datensatz der Regionen höherer Ordnung trainiert.

Tabelle 2: Regionen höherer Ordnung und zugewiesene Subregionen.

Region höherer Ordnung	**Subregionen**
Liquor	CSF (Cerebrospinal Fluid)
Weiße Substanz	Right Cerebellum White Matter Left Cerebellum White Matter Right Cerebral White Matter Left Cerebral White Matter
Basalganglien	Right Accumbens Area Left Accumbens Area Right Caudate Left Caudate Right Pallidum Left Pallidum Right Basal Forebrain Left Basal Forebrain Right ACgG anterior cingulate gyrus Left ACgG anterior cingulate gyrus Right MCgG middle cingulate gyrus Left MCgG middle cingulate gyrus Right PCgG posterior cingulate gyrus Left PCgG posterior cingulate gyrus Right SCA subcallosal area Left SCA subcallosal area
Limbisches System	Right Amygdala Left Amygdala Right Hippocampus Left Hippocampus Right Putamen

	Left Putamen Right AIns anterior insula Left AIns anterior insula Right Ent entorhinal area Left Ent entorhinal area Right PHG parahippocampal gyrus Left PHG parahippocampal gyrus Right PIns posterior insula Left PIns posterior insula
Frontaler Cortex	Right AOrG anterior orbital gyrus Left AOrG anterior orbital gyrus Right FO frontal operculum Left FO frontal operculum Right FRP frontal pole Left FRP frontal pole Right GRe gyrus rectus Left GRe gyrus rectus Right LOrG lateral orbital gyrus Left LOrG lateral orbital gyrus Right MFC medial frontal cortex Left MFC medial frontal cortex Right MFG middle frontal gyrus Left MFG middle frontal gyrus Right MOrG medial orbital gyrus Left MOrG medial orbital gyrus Right MPrG precentral gyrus medial segment Left MPrG precentral gyrus medial segment Right MSFG superior frontal gyrus medial segment Left MSFG superior frontal gyrus medial segment Right OpIFG opercular part of the inferior frontal gyrus Left OpIFG opercular part of the inferior frontal gyrus Right OrIFG orbital part of the inferior frontal gyrus Left OrIFG orbital part of the inferior frontal gyrus Right POrG posterior orbital gyrus Left POrG posterior orbital gyrus Right PrG precentral gyrus Left PrG precentral gyrus Right SFG superior frontal gyrus Left SFG superior frontal gyrus Right SMC supplementary motor

	cortex Left SMC supplementary motor cortex Right TrIFG triangular part of the inferior frontal gyrus Left TrIFG triangular part of the inferior frontal gyrus
Parietaler Cortex	Right AnG angular gyrus Left AnG angular gyrus Right PCu precuneus Left PCu precuneus Right PO parietal operculum Left PO parietal operculum Right PoG postcentral gyrus Left PoG postcentral gyrus Right SMG supramarginal gyrus Left SMG supramarginal gyrus Right SPL superior parietal lobule Left SPL superior parietal lobule
Temporaler Cortex	Right CO central operculum Left CO central operculum Right FuG fusiform gyrus Left FuG fusiform gyrus Right ITG inferior temporal gyrus Left ITG inferior temporal gyrus Right MPoG postcentral gyrus medial segment Left MPoG postcentral gyrus medial segment Right MTG middle temporal gyrus Left MTG middle temporal gyrus Right PP planum polare Left PP planum polare Right PT planum temporale Left PT planum temporale Right STG superior temporal gyrus Left STG superior temporal gyrus Right TMP temporal pole Left TMP temporal pole Right TTG transverse temporal gyrus Left TTG transverse temporal gyrus
Occipitaler Cortex	Right Calc calcarine cortex Left Calc calcarine cortex Right Cun cuneus Left Cun cuneus Right IOG inferior occipital gyrus Left IOG inferior occipital gyrus Right LiG lingual gyrus Left LiG lingual gyrus

	Right MOG middle occipital gyrus Left MOG middle occipital gyrus Right OCP occipital pole Left OCP occipital pole Right OFuG occipital fusiform gyrus Left OFuG occipital fusiform gyrus Right SOG superior occipital gyrus Left SOG superior occipital gyrus
Cerebellum	Right Cerebellum Exterior Left Cerebellum Exterior Cerebellar Vermal Lobules I-V Cerebellar Vermal Lobules VI-VII Cerebellar Vermal Lobules VIII-X
Diencephalon	Right Ventral DC Left Ventral DC Right Thalamus Proper Left Thalamus Proper

2.8.2 Feature-basiertes Maschinelles Lernen

Die erzeugten Feature-Vektoren wurden für das Training und die Validierung jeweils eines Ridge Regression und Support Vector Regression Modells verwendet.

2.8.2.1 Ridge Regression

Ridge Regression minimiert den Quadratfehler zwischen wahrem Ergebnis Y und der Vorhersage $f(X)$ auf Basis der Eingabe X, indem die lineare Funktion f des Grades d mit Feature-Gewichten w und der Verschiebeparameter b angepasst werden. Zur Regularisierung wird die L2-Norm der Gewichte $\|w\|^2$ verwendet, während mittels α die Bedeutung der Regularisierung angepasst wird.

Somit wird das folgende Problem gelöst:

$$w_{n,\alpha} = argmin_{w \in R^d} \frac{1}{n} \|Y - Xw\|^2 + \ \alpha \|w\|^{\ 2}$$

Die Ridge Regression Implementation der Python-Bibliothek sklearn wurde verwendet (Pedregosa et al., 2011).

2.8.2.2 Support Vector Regression

Das Ziel der Support Vector Regression ist es, eine Funktion $f(x)$ zu finden, die für die Trainingsdaten (x_i, y_i) aus dem Input-Raum $X \in \mathbb{R}^d$ höchstens eine Abweichung ε vom wahren Ziel y aufweist (Smola and Schölkopf, 2004).

Eine lineare Funktion kann zu diesem Zweck beschrieben werden durch

$$f(x) = \langle w, x \rangle + b, w \in X, b \in \mathbb{R}$$

mit $\langle w, x \rangle$ als Skalarprodukt in X. Es soll nun eine optimale Lösung für den Datensatz mit möglichst kleinem w gefunden werden, beispielsweise durch Minimierung der Norm $\|w\|^2 = \langle w, w \rangle$. Zudem können Slack-Variablen ξ_i eingeführt werden, um in einigen Fällen größere Abweichungen von y als ε zu erlauben. Der Regularisierungsfaktor C bestimmt den Einfluss der Abweichungen. Somit kann das Optimierungs-Problem der linearen Support Vector Regression (Vapnik, 1999) formuliert werden als

$$minimiere \ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$unter\ der\ Nebenbedingung \begin{cases} y_i - \langle w, x_i \rangle - \mathrm{b} \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + \mathrm{b} - y_i \leq \varepsilon + \xi_i{}^* \\ \xi, \xi_i{}^* \geq 0 \end{cases}$$

Die Support Vector Regression lässt sich auch auf den nicht-linearen Fall erweitern. Hierzu können die Daten mittels einer Abbildung $\Phi: X \to F$ in einen Feature-Space F abgebildet werden. Dies wird jedoch für höherdimensionale Daten rechenintensiv, weshalb eine implizite Abbildung mittels Kernel-Funktion durchgeführt wird. In dieser Arbeit wurde der radial basis function (RBF-) Kernel verwendet.

$$K(x_i, x_j) = \exp\left(\frac{|x_i - x_j|^2}{2\sigma^2}\right)$$

Das Optimierungsproblem kann mittels verschiedener Methoden numerisch gelöst werden.

In dieser Arbeit wurde die Support Vector Regression Implementation der Python-Bibliothek sklearn mit einem RBF-Kernel verwendet (Pedregosa et al., 2011).

2.8.3 Training und Hyperparameter-Optimierung

Die Hyperparameter der Ridge Regression und Support Vector Regression wurden mittels Gridsearch-Cross-Validation optimal eingestellt. Dazu wurde die Implementation der Python-Bibliothek sklearn (sklearn,GridsearchCV) (Pedregosa et al., 2011) verwendet.

Fünffache Kreuzvalidierung wurde eingesetzt, um die besten Parameter zu finden. Mithilfe dieser gewählten Parameter wurde dann auf einem ungesehen Test-Datensatz die Performance der Modelle überprüft.

Für Ridge Regression wurde der Hyperparameter „alpha“ eingestellt, der mit dem Parameter α in obiger Gleichung korrespondiert und den Einfluss der Regularisierung variiert. Für die Support Vector Regression wurden die Parameter „C“, „epsilon“ und „gamma“ eingestellt. C beeinflusst die Strafe für Abweichungen von der ε-Röhre, epsilon definiert die margin der ε-Röhre während gamma den Kernel-Koeffizienten des RBF-Kernels beeinflusst und proportional zu $\frac{1}{2\sigma^2}$ ist.

2.8.4 Feature Importance

Um Einsichten in die Gewichte der einzelnen Features, respektive Gehirnregionen zu gewinnen, wurden sowohl die Gewichte des Ridge Regression Modells, das auf dem Subregionen-Datensatz trainiert wurde, als auch die des Ridge Regression Modells, das auf dem Datensatz der Regionen höherer Ordnung trainiert wurde, aus dem Modell ausgelesen und die Koeffizienten graphisch dargestellt.

Aufgrund der nichtlinearen Eigenschaften des Radial Basis Function Kernels, der bei der Support Vector Regression eingesetzt wurde, können die Parameter dieses Modells nicht analysiert und interpretiert werden.

2.9 Deep Learning

2.9.1 Preprocessing

Das verwendete VGG16-Netzwerk verwendet Eingabedaten der Dimension 224x224x224 Voxel. Daher wurden die registrierten Gehirn-Datensätze zuerst auf die Größe von 224x224x224 mit einem Voxel-Spacing von 1x1x1 mm resampled. Dazu wurde die Funktion resample_image der Python-Bibliothek sklearn verwendet. Die Intensitätswerte wurden normalisiert. Dazu wurden Werte zwischen 0 und 100 Hounsfield Einheiten geclippt und danach durch 100 geteilt, so dass die Datensätze Voxel-Intensitätswerte zwischen 0 und 1 aufwiesen. Dieser Vorgang dient dazu, zu hohen Parameterwerten im trainierten Modell entgegenzuwirken.

Die normalisierten Datensätze wurden mithilfe eines Gauss-Filters mit einem Sigma-Wert von 1 gefiltert, um Rauschen in den Datensätzen zu reduzieren.

Aus den dreidimensionalen Datensätzen wurden einzelne Schichten ausgewählt, die als Input für die zu trainierenden Netzwerke genutzt wurden.

Für den Deep-Learning-Ansatz wurden ein 10-channel-Netzwerk sowie ein 3-channel-Netzwerk erzeugt, trainiert und getestet. Das 10-channel-Netzwerk verwendete die Schichten 40, 50, 60, 70, 80, 90, 100, 110, 120 und 130 aus den dreidimensionalen Datensätzen. Jeder Channel erhielt als Input eine andere Schicht. Das 3-channel-Netzwerk nutzte für jeden seiner drei Channel die Schicht 90 aus dem dreidimensionalen Datensatz, um Grad-CAM darauf anwenden zu können.

Einzelne Schichten der dreidimensionalen Gehirndatensätze reichen aus, damit ein VGG16-Netzwerk eine hohe Performance erreichen kann. Dies liegt in der Tatsache begründet, dass nebeneinanderliegende Schichten in Gehirn-Scans nahezu gleiche Informationen enthalten, so dass weniger Schichten, die weiter auseinander liegen, fast die gleiche Menge an Informationen enthalten wie der gesamte Datensatz (Huang et al., 2017).

Danach wurde Daten-Augmentation durchgeführt, um die Anzahl an Trainingsdatensätzen zu erhöhen. Dazu wurden die bearbeiteten Datensätze

mit 5 verschiedenen Winkeln rotiert (-20 bis 20 Grad, in 10 Grad-Schritten), da Rotationen zwischen -20 und 20 Grad bei manchen Bilderkennungsaufgaben hilfreich sein können (Shorten and Khoshgoftaar, 2019).

Für das 10-channel-Netzwerk wurde somit für jeden Patienten ein Datensatz der Dimension 5x10x224x224, für das 3-channel-Netzwerk ein Datensatz der Dimension 5x3x224x224 erzeugt. Diese Datensätze wurden jeweils in einem gemeinsamen Array zusammengeführt und mithilfe des hdf5-Formates gespeichert, um schnell auf die Daten zugreifen zu können. Zusätzlich wurde das Alter jedes Patienten in der Datei abgelegt.

2.9.2 Netzwerk-Architektur

Die Anwendung von Convolutional Neural Networks hat sich in den vergangenen Jahren besonders in der Analyse von Bilddaten hervorgetan.

In dieser Arbeit wurde ein modifiziertes und auf dem ImageNet-Datensatz vortrainiertes VGG16-Netzwerk mit Batch-Normalization in der Implementation der Python-Bibliothek Pytorch (Paszke et al., 2019) erzeugt, trainiert und getestet. Da das verwendete VGG16-Netz ursprünglich für die Analyse von zweidimensionalen RGB-Daten mit 3 Input-Channels entwickelt wurde, musste es für die Regressions-Aufgabe der Altersschätzung modifiziert werden.

Für das 10-channel-Netzwerk wurde die Anzahl der Input-Channel auf 10 angepasst. Zudem wurden die Fully-connected-layer beider Netzwerke durch Schichten mit weniger Neuronen ersetzt. Der softmax-Layer wurde ebenfalls entfernt. Die Architektur des verwendeten Netzwerkes ist in Abbildung 16 gezeigt. LeakyReLU mit einer slope von 0,01 wurde als Aktivierungsfunktion verwendet.

Für eine verbesserte Regularisierung wurden Dropout-Layer hinzugefügt. Diese ignorieren während des Trainings eine vorher eingestellte Prozentzahl zufälliger Neurone, sodass die Vorhersage immer von zufälligen Neuronen abhängt. Dies verhindert, dass einzelne Neurone zu viel Gewicht bei der Vorhersage entwickeln, und ermöglicht eine verbesserte Kontrolle des overfittings.

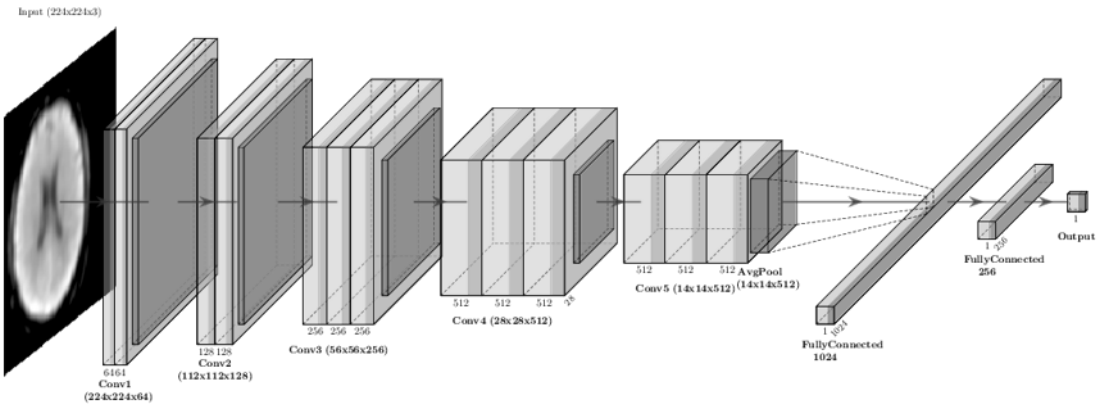


Abbildung 16: Konfiguration des VGG16-Netzwerkes. Als Input dient eine definierte Anzahl Schichten aus einem axialen Weichteil-CT des Schädels. Als Output wird das vorhergesagte Chronologische Patientenalter angegeben.

2.9.3 Training und Hyperparameter-Optimierung

Beide Netzwerke wurden auf die gleiche Art für die automatische Altersschätzung trainiert. Das 10-channel-Netzwerk wurde mit dem 10-channel-Datensatz, aufgeteilt in einen Datensatz für Training, Validierung und Test, trainiert, validiert und getestet. Für das 3-channel-Netzwerk wurde der 3-channel-Datensatz verwendet.

Das Training, die Validierung und Testung wurden auf der NVIDIA Tesla V100 32GB Graphikkarte durchgeführt. Die Datensätze wurden aus den hdf5-Dateien gelesen und in Pytorch Tensor-Datenstrukturen überführt, um Tensor-Datensätze zu erzeugen, die für das Training des Netzwerkes genutzt werden können.

Es wurde jeweils ein Trainings-, Validierungs- und Testdatensatz erzeugt. Für das Training wurden alle Rotationen der Originaldaten verwendet, während für Validierung und Test nur die originalen, nicht-rotierten Datensätze eingesetzt wurden.

Das Training erfolgte in Batches, deren Größe manuell angepasst wurde. Der von PyTorch implementierte AdamW-Optimierer wurde eingesetzt, um während des Trainings die Netzwerk-Parameter zu aktualisieren. Als Loss-Funktion

wurde Pytorchs MSE-Loss mit L1-Regularisierung eingesetzt. Dazu wurden zusätzlich sämtliche Parameterwerte addiert, mit einem konstanten Faktor (10^{-3} bzw. 10^{-4}) multipliziert und dieses Ergebnis zum Trainingsfehler addiert.

Die Hyperparameter-Optimierung wurde manuell durchgeführt. Die Batchgröße (16, 24, 32, 48, 64, 96), Lernrate (10^{-3}, 10^{-4}, 10^{-5}), der konstante Faktor der L1-Regularisierung (10^{-3}, 10^{-4}) und die Dropout-Rate der Dropout-Layer (0,3, 0,4, 0,5) wurden eingestellt und überprüft.

Die Hyperparameter, die für beide VGG16-Netzwerke mit AdamW-Optimizer schlussendlich ausgewählt und verwendet wurden, waren ein Momentum von 0,9, eine Lernrate von 10^{-5} und eine Batch-Größe von 24. Die Dropout-Rate wurde mit 0,4 für alle Dropout-Layer gewählt. Der Koeffizient der L1-Regularisierung war 10^{-3}.

2.9.4 Visualisierung der Feature-Repräsentation

Um tiefgreifendere Einblicke in die „Blackbox“-Modelle zu erhalten, die Convolutional Neural Networks darstellen, wurde der Grad-CAM-Algorithmus eingesetzt. Ein wichtiger Vorteil von Grad-CAM besteht darin, dass der Algorithmus angewendet werden kann, wenn ein Modell bereits trainiert wurde und nicht während des Trainings implementiert werden muss.

Nachdem das 3-channel-Netzwerk trainiert wurde, wurde eine Grad-CAM-Implementation für Pytorch auf den Bildern des Test-Datensatzes eingesetzt (Früh et al., 2021). Als Ziel-Layer wurde der letzte Convolutional Layer des Feature Extraktors ausgewählt, da in den tiefen Schichten des Feature Extraktors komplexe, hochgradige visuelle Konstrukte der Features repräsentiert werden, während gleichzeitig räumliche Informationen erhalten bleiben (Selvaraju et al., 2017, Bengio et al., 2013).

Die Grundzüge von Grad-CAM sollen kurz erläutert werden. Zuerst macht das trainierte Netzwerk auf Basis des Input-Bildes eine Vorhersage. Diese Vorhersage wird mithilfe der Backpropagation durch das Modell zurücktransportiert. Die Gradienten werden berechnet und gepoolt. Die gepoolten Gradienten werden dann verwendet, um die Feature Maps des

gewählten Convolutional Layers und damit Regionen zu identifizieren, die die Vorhersage der automatisierten Alterseinschätzung am stärksten beeinflusst haben. Damit kann eine spezifische Saliency Map (auch Aufmerksamkeitskarte), die angibt, welche Bildregionen für die Vorhersage am wichtigsten waren, für die anhand des gegebenen Input-Bildes getroffene Vorhersage interpoliert werden.

In dieser Arbeit wurde sich bei der Analyse der erzeugten Saliency Maps auf die Interpretation wiederkehrender Muster konzentriert.

2.10 Statistische Analyse des Test-Datensatzes

304 Patientendatensätze wurden zufällig dem Test-Datensatz zugeteilt, auf dessen Basis die Performance der verschiedenen Ansätze bewertet wurde.

Hierzu wurden die Vorhersagen der Ansätze (Ridge Regression, Support Vector Regression, 3-channel-Netzwerk, 10-channel-Netzwerk) mit dem wahren Alter und den Vorhersagen der anderen Ansätze verglichen und korreliert. Um festzustellen, ob sich die absoluten Vorhersagefehler zwischen den Modellen signifikant unterschieden, wurde der t-Test für gepaarte Stichproben angewendet.

Zudem wurden die Auswirkungen der unabhängigen Variablen des Geschlechts, Altersgruppe, Scanners, Reconstruction-Kernels, der Schichtdicke, der Röhrenspannung und der Strahlenexposition auf die absoluten Vorhersagefehler der jeweiligen Modelle untersucht.

Für die Untersuchung der Unterschiede innerhalb der Modelle für die unabhängigen Variablen des Geschlechts und der Röhrenspannung wurde der Mann-Whitney-U-Test verwendet, da hierbei nur zwei unterschiedliche Gruppen verglichen wurden (Männlich vs. weiblich, respektive 100 kV vs. 120 kV).

Für die Untersuchung der Unterschiede der übrigen Variablen wurde der Kruskal-Wallis-Test verwendet. Für die Untersuchung der Verteilung der absoluten Vorhersagefehler in verschiedenen Altersgruppen wurden die Patienten in 7 Altersgruppen eingeteilt (20-29, 30-39, 40-49, 50-59, 60-69, 70-79, >79).

3 Ergebnisse

3.1 Feature-basiertes Maschinelles Lernen

3.1.1 Ridge Regression

Auf dem Datensatz der Regionen höherer Ordnung (Modell 1) erreichte das beste Ridge Regression Modell (alpha=0,1) auf dem Testdatensatz einen MAE von 10,785 Jahren mit einem R^2-Wert von 0,497.

Das beste Modell, das auf dem Subregionen-Datensatz trainiert wurde (Modell 2), erreichte einen MAE auf dem Test-Datensatz von 9,840 Jahren mit einem R^2-Wert von 0,585. Es wurde ein alpha von 1 als optimaler Parameter bestimmt.

3.1.1.1 Statistische Analyse der Altersschätzung

Abbildung 17 zeigt die Altersschätzung des Ridge Regression Modells 2 aufgetragen zum tatsächlichen Alter. Die Vorhersagen des Modells zeigten eine Korrelation mit einem R^2-Wert von 0,585 mit dem tatsächlichen Alter.

Im Folgenden wurden die Vorhersagefehler der Altersschätzungen von Modell 2 statistisch auf Einflüsse der Altersgruppen, des Geschlechts, der verwendeten Scanner, Reconstruction Kernel sowie Schichtdicken untersucht.

Die jüngeren und älteren Altersgruppen hatten einen signifikant höheren Vorhersagefehler als die mittelalten, die jüngsten und ältesten unterschieden sich jedoch nicht signifikant untereinander.

Das Geschlecht des Patienten, der verwendete CT-Scanner, Reconstruction Kernel, sowie die verwendete Schichtdicke hatten keinen signifikanten Einfluss auf den absoluten Vorhersagefehler.

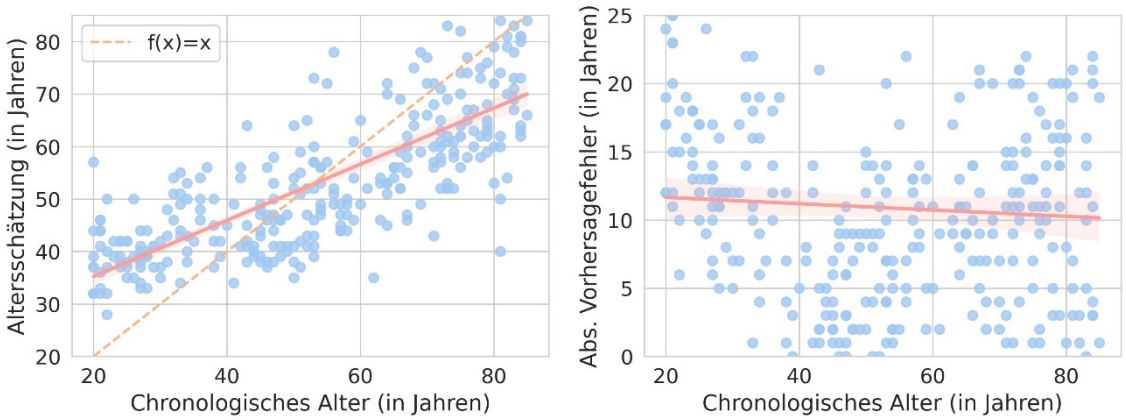


Abbildung 17: Streudiagramm der Altersschätzung und des absoluten Vorhersagefehlers des Ridge Regression Modells 2 aufgetragen zum tatsächlichen Alter.

Links: In Orange ist die Regressionsgerade gezeigt, während die gestrichelte Linie die Einheitsgerade darstellt. Jüngere Patienten scheinen älter geschätzt zu werden, ältere jünger. Rechts: Absoluter Vorhersagefehler (in Jahren) aufgetragen zum Chronologischen Alter. In Orange ist die Regressionsgerade dargestellt. Der Vorhersagefehler für jüngere und ältere Patienten war signifikant höher.

3.1.1.2 Feature Importance

Die Koeffizienten des Ridge Regression Modells 1, das auf dem Datensatz der Regionen höherer Ordnung trainiert wurde, sind in Abbildung 18 dargestellt, die Koeffizienten des Modells 2, das auf dem Subregionen-Datensatz trainiert wurde, in Abbildung 19.

Das Liquorvolumen, das Volumen der Weißen Substanz, des temporalen sowie des parietalen Cortex erhielten positive Koeffizienten. Der Koeffizient des Liquor-Volumens war mit Abstand am stärksten positiv. Die Volumina des limbischen Systems, des Cerebellums, des frontalen Cortex, des occipitalen Cortex, der Basalganglien sowie des Diencephalons erhielten negative Koeffizienten, trugen also zu niedrigeren Altersschätzungen bei.

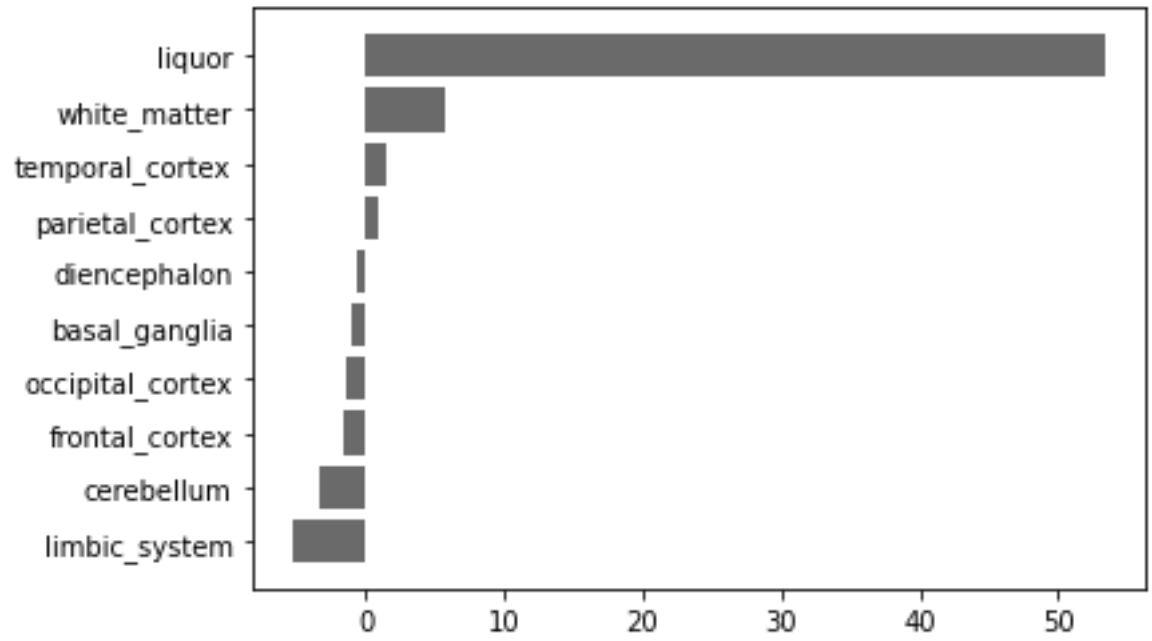


Abbildung 18: Koeffizienten der Regionen höherer Ordnung (Modell 1)

Die Gehirnregionen mit den höchsten positiven Koeffizienten waren der Supramarginale Gyrus sowie das Liquor-System. Diese wiesen entsprechend ein höheres (relatives) Volumen bei älteren Patienten auf. Höhere Volumina im Precuneus, Thalamus proper, sowie Lobus caudatus waren hingegen mit niedrigeren Altersschätzungen assoziiert. Andere Regionen der Basalganglien, wie der Gyrus Cinguli erhielten für die medialen und posterioren Abschnitte positive Koeffizienten, für den anterioren jedoch negative Koeffizienten.

Die weiße Substanz war in Modell 1 mit einem positiven Koeffizienten versehen. Im Modell 2 zeigte jedoch nur das Volumen der weißen Substanz des Cerebellums einen positiven Koeffizienten, während das Volumen der cerebralen Weißen Substanz einen negativen Koeffizienten zeigte, also mit niedrigeren Altersschätzungen assoziiert war. Generell ließen sich in jeder übergeordneten Region Subregionen mit positiven und negativen Koeffizienten finden.

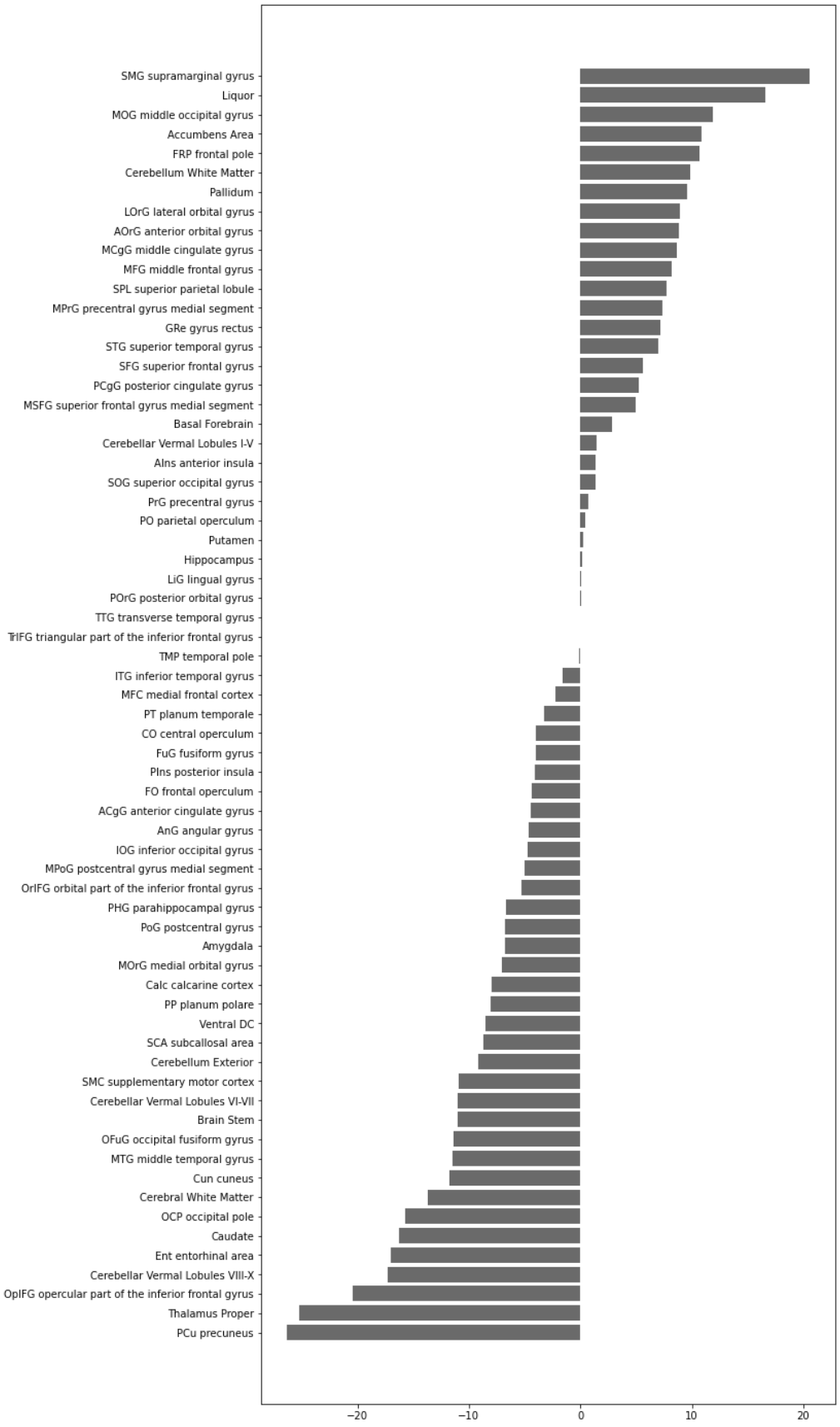


Abbildung 19: Koeffizienten der Subregionen des Ridge Regression Modells 2.

3.1.2 Support Vector Regression

Ein Support Vector Regression Modell mit Radial Basis Function Kernel wurde auf dem Subregionen-Datensatz trainiert. Die besten Hyperparameter, welche mittels 5-fold Gridsearch-Cross-Validation bestimmt wurden, waren C=100, epsilon=3 und gamma=1. Ein MAE von 8,317 Jahren wurde erreicht, mit einem R^2-Wert von 0,743.

3.1.2.1 Statistische Analyse der Altersschätzung

Abbildung 20 zeigt die Altersschätzung des Support Vector Regression Modells aufgetragen zum tatsächlichen Alter.

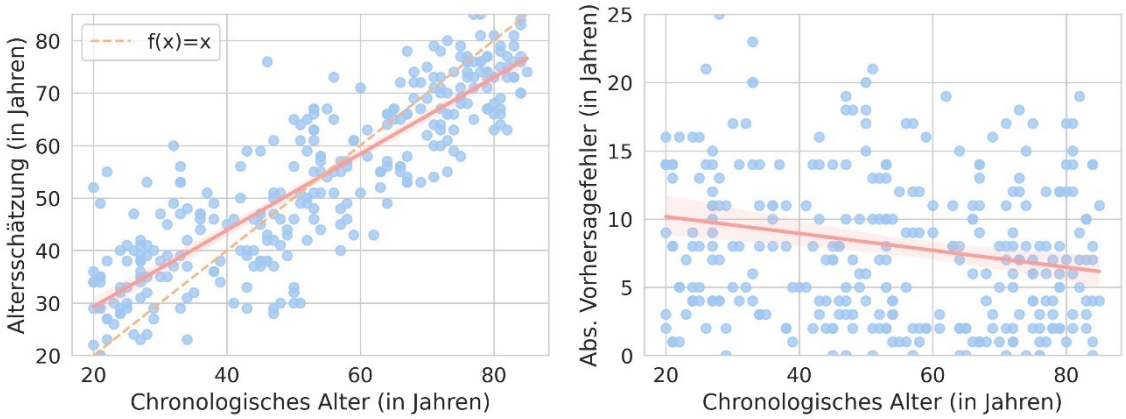


Abbildung 20: Streudiagramm der Altersschätzung und des absoluten Vorhersagefehlers des Support Vector Regression Modells aufgetragen zum tatsächlichen Alter.
Links: In Orange ist die Regressionsgerade gezeigt, während die gestrichelte Linie die Einheitsgerade darstellt. Jüngere Patienten scheinen älter geschätzt zu werden, ältere jünger. Rechts: Absoluter Vorhersagefehler (in Jahren) aufgetragen zum Chronologischen Alter. In Orange ist die Regressionsgerade dargestellt. Der absolute Vorhersagefehler scheint insgesamt recht stabil, jedoch für jüngere Patienten höher zu sein.

Die Vorhersagen des Modells zeigten eine Korrelation mit einem R^2-Wert von 0,743 mit dem tatsächlichen Alter.

Bei der Untersuchung der Vorhersagefehler nach Altersgruppen zeigten sich signifikante Unterschiede. So wurden jüngere Patienten älter geschätzt und ältere Patienten jünger, während mittelalte Patienten keine Abweichungsrichtung zeigen. Die absoluten Vorhersagefehler unterschieden sich in den definierten Altersgruppen jedoch nicht signifikant.

Ebenso hatten weder das Geschlecht des Patienten, der verwendete CT-Scanner, Reconstruction-Kernel noch die gewählte Schichtdicke einen signifikanten Einfluss auf den Vorhersagefehler.

3.2 Deep Learning

3.2.1 Automatisierte Altersschätzung

Abbildung 21 zeigt den Verlauf des Trainings für das 3-channel-Netzwerk.

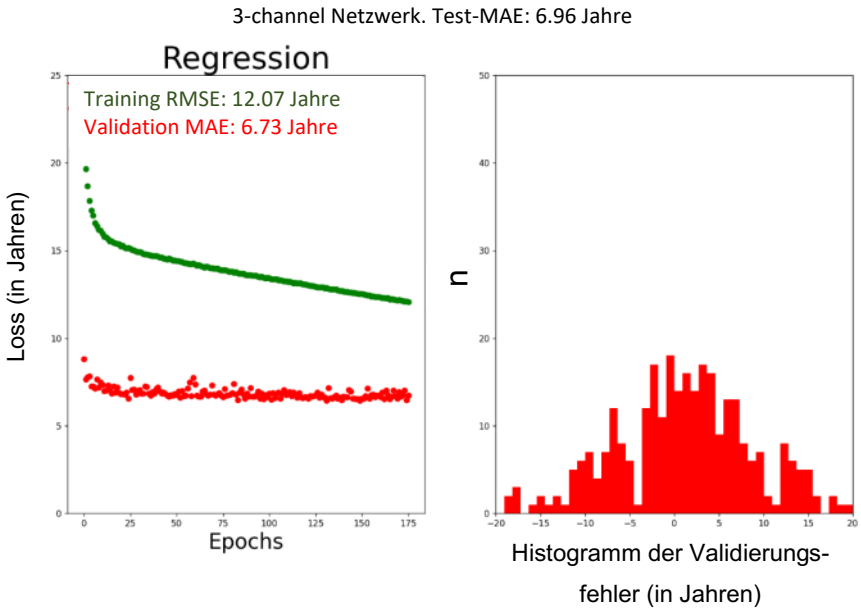


Abbildung 21: Verlauf des Trainings für das 3-channel-Netzwerk.
Links: Der Validierungsfehler (rot) fällt erst stark und nach 150 Epochen fast nicht mehr, weshalb das Training beendet wird. Der Trainings-RMSE umfasst zusätzlich den Regularisierungswert und ist nicht als tatsächlicher Altersschätzungs-Fehler anzusehen. Rechts ist das Histogramm der Vorhersagefehler des Validierungsdatensatzes abgebildet.

Das 3-channel-Netzwerk erreichte nach 175 Epochen einen Validierungs-MAE von 6,44 Jahren mit einem korrespondierenden Test-MAE von 6,96 Jahren.

Abbildung 22 zeigt die Altersschätzung des 3-channel-Netzwerks aufgetragen zum tatsächlichen Alter. Die gestrichelte Gerade gibt die Funktion $f(x) = x$ an, auf der die Vorhersagen im Falle optimaler Vorhersagen lägen. Die Vorhersagen des Modells zeigten eine Korrelation mit einem R^2-Wert von 0,813 mit dem tatsächlichen Alter.

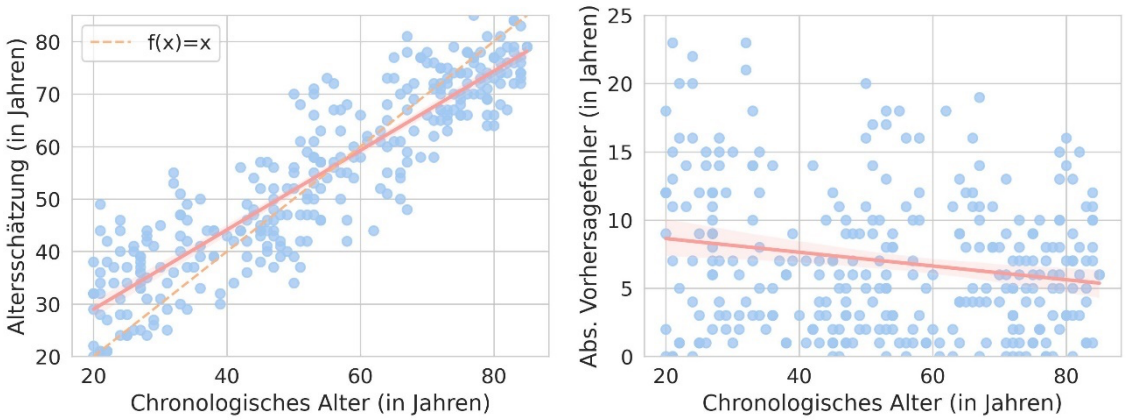


Abbildung 22: Streudiagramm der Altersschätzung und des absoluten Vorhersagefehlers des 3-channel-Netzwerks aufgetragen zum tatsächlichen Alter. Links: In Orange ist die Regressionsgerade gezeigt, während die gestrichelte Linie die Einheitsgerade darstellt. Jüngere Patienten scheinen älter geschätzt zu werden, ältere jünger. Die Regressionsgerade nähert sich der Einheitsgerade an. Rechts: Absoluter Vorhersagefehler (in Jahren) aufgetragen zum Chronologischen Alter. In Orange ist die Regressionsgerade dargestellt. Der absolute Vorhersagefehler scheint für jüngere Patienten höher zu sein.

Bei der Untersuchung der Vorhersagefehler nach Altersgruppen zeigten sich signifikante Unterschiede. So wurden jüngere Patienten älter geschätzt und ältere Patienten jünger, während Patienten mittleren Alters keine Abweichungsrichtung zeigen. Ein signifikant höherer absoluter Vorhersagefehler der jüngsten Altersgruppe (20-29 Jahre) gegenüber der mittleren (40-49) und ältesten (70-85) wurde beobachtet.

Weder das Geschlecht des Patienten, der verwendete CT-Scanner, Reconstruction-Kernel noch die gewählte Schichtdicke hatten einen signifikanten Einfluss auf den Vorhersagefehler.

Abbildung 23 zeigt den Verlauf des Trainings für das 10-channel Netzwerk. Das 10-channel-Netzwerk erreichte nach 250 Epochen des Trainings einen Validierungs-MAE von 5,62 Jahren mit einem korrespondierenden Test-MAE von 5,73 Jahren. Dies ergab einen „weighted MAE“ $\left(\frac{MAE}{\Delta\ Alter}\right)$ von 0,88. Die Vorhersagen des Modells zeigten eine Korrelation mit einem R^2-Wert von 0.859 mit dem tatsächlichen Alter.

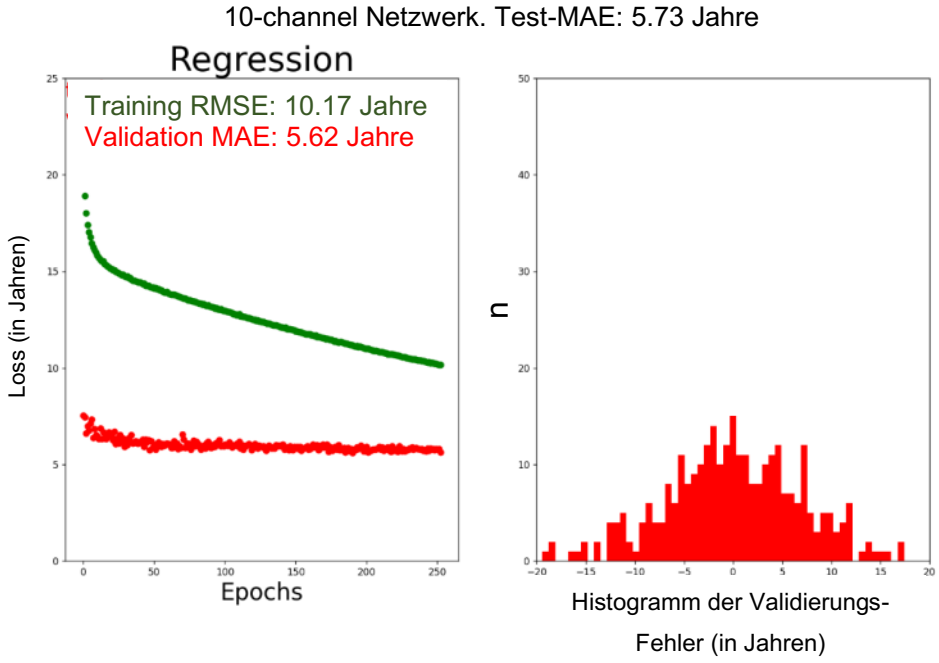


Abbildung 23: Verlauf des Trainings für das 10-channel-Netzwerk.

Links: Der Validierungsfehler (rot) fällt erst stark und nach 200 Epochen fast nicht mehr, weshalb das Training beendet wird. Der Trainings-RMSE umfasst zusätzlich den Regularisierungswert und ist nicht als tatsächlicher Altersschätzungs-Fehler anzusehen. Rechts ist das Histogramm der Vorhersagefehler des Validierungsdatensatzes abgebildet.

Abbildung 24 zeigt die Altersschätzung des 10-channel-Netzwerks aufgetragen zum tatsächlichen Alter.

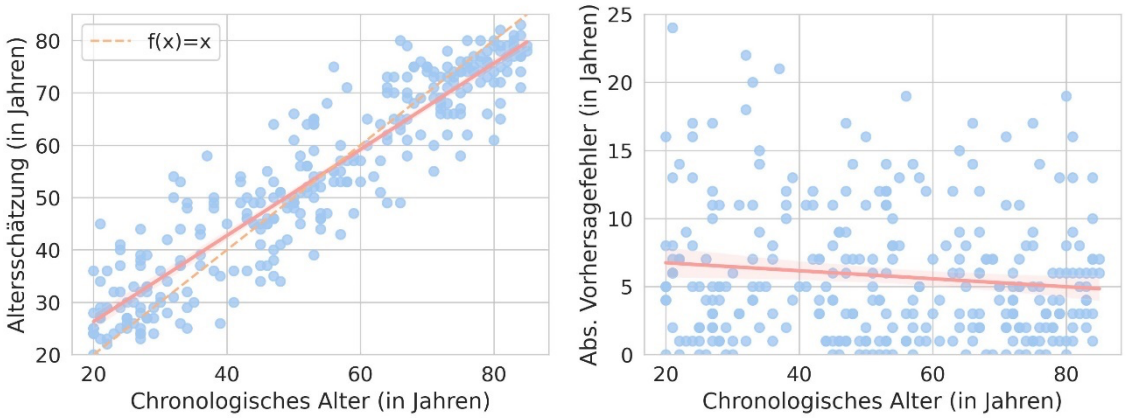


Abbildung 24: Streudiagramm der Altersschätzung und des absoluten Vorhersagefehlers des 10-channel-Netzwerks aufgetragen zum tatsächlichen Alter. Links: In Orange ist die Regressionsgerade gezeigt, während die gestrichelte Linie die Einheitsgerade darstellt. Jüngere Patienten scheinen älter geschätzt zu werden, ältere jünger. Die Regressionsgerade nähert sich der Einheitsgerade an. Rechts: Absoluter Vorhersagefehler (in Jahren) aufgetragen zum Chronologischen Alter. In Orange ist die

Regressionsgerade dargestellt. Der absolute Vorhersagefehler scheint für jüngere Patienten höher zu sein, jedoch relativ stabil.

Bei der Untersuchung der Vorhersagefehler nach Altersgruppen zeigten sich signifikante Unterschiede. So wurden jüngere Patienten älter geschätzt und ältere Patienten jünger, während mittelalte Patienten keine klare Abweichungsrichtung zeigten. Die absoluten Vorhersagefehler unterschieden sich in den definierten Altersgruppen jedoch nicht signifikant.

Das Geschlecht der Patienten hatte einen signifikanten Einfluss auf den Vorhersagefehler. So wurden weibliche Patienten signifikant älter geschätzt als männliche Patienten, was in einem stärker positiven Vorhersagefehler zur Darstellung kam. Die absoluten Vorhersagefehler für männliche und weibliche Patienten unterschieden sich nicht signifikant.

Die technischen Parameter, also der verwendete CT-Scanner, Reconstruction-Kernel oder die gewählte Schichtdicke, hatten keinen signifikanten Einfluss auf den Vorhersagefehler.

3.2.2 Visualisierung der Feature-Repräsentation

Grad-CAM wurde auf das trainierte 3-channel-Netzwerk angewendet. Spezifische Saliency Maps wurden für jedes Bild aus dem Test-Datensatz erzeugt, sowie eine gemittelte Saliency Map über alle erzeugten berechnet. Die Saliency Maps wiesen eine Auflösung von 14x14 Pixeln auf. Die Darstellung wurde mithilfe eines Gauss-Filters interpoliert.

Abbildung 25 zeigt die gemittelte Grad-CAM Saliency Map als Overlay über die gemittelten verwendeten Gehirnschnitte aller Patienten des Testdatensatzes.

Durchschnittlich war eine Region links frontal wichtig für die Vorhersage des Netzwerkes. Diese Region umfasste Teile des Vorderhorns des linken Ventrikels sowie der frontalen Grauen und Weißen Substanz.

Ebenfalls wurde eine Region im occipitalen rechten Gehirn, welche das Hinterhorn des rechten Ventrikels miteinschloss, als einflussreich auf die Vorhersage hervorgehoben. Auch zentrale Anteile, die den rechten Seitenventrikel umfassten, wurden konstant hervorgehoben. Geringeren

Einfluss hatten occipitale Regionen des linken Gehirns, die das Hinterhorn des linken Ventrikels miteinschlossen, sowie die Insula-Region.

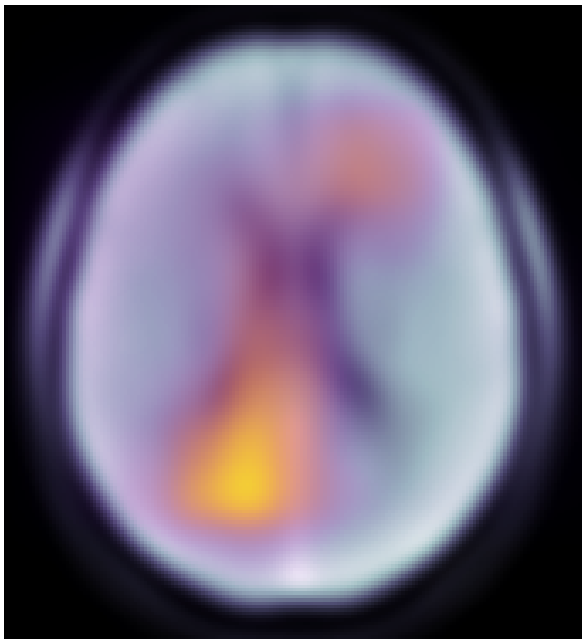


Abbildung 25: Gemittelte Grad-CAM Saliency Map als Overlay über die gemittelten verwendeten Gehirnschnitte aller Patienten des Testdatensatzes.
Die Regionen um das Vorderhorn des linken Seitenventrikels sowie um das Hinterhorn des rechten Ventrikels wurden hervorgehoben.

Saliency Maps ausgewählter Patienten sind in Abbildung 26 dargestellt. Im Vergleich zu den Saliency Maps der mittelalten und älteren Patienten sowie zur gemittelten Saliency Map stellten sich die für die Vorhersagen wichtigen Regionen bei jüngeren Patienten deutlich anders dar. Statt der hervorgehobenen Regionen im rechten occipitalen und linken frontalen Hirn fand sich der Aufmerksamkeitsfokus deutlich häufiger in anderen Hirnregionen, beispielsweise rechts frontal.

Dabei wurde neben dem Hinterhorn des linken Ventrikels auch Regionen des limbischen Systems, der occipitalen weißen Substanz und der Insula hervorgehoben.

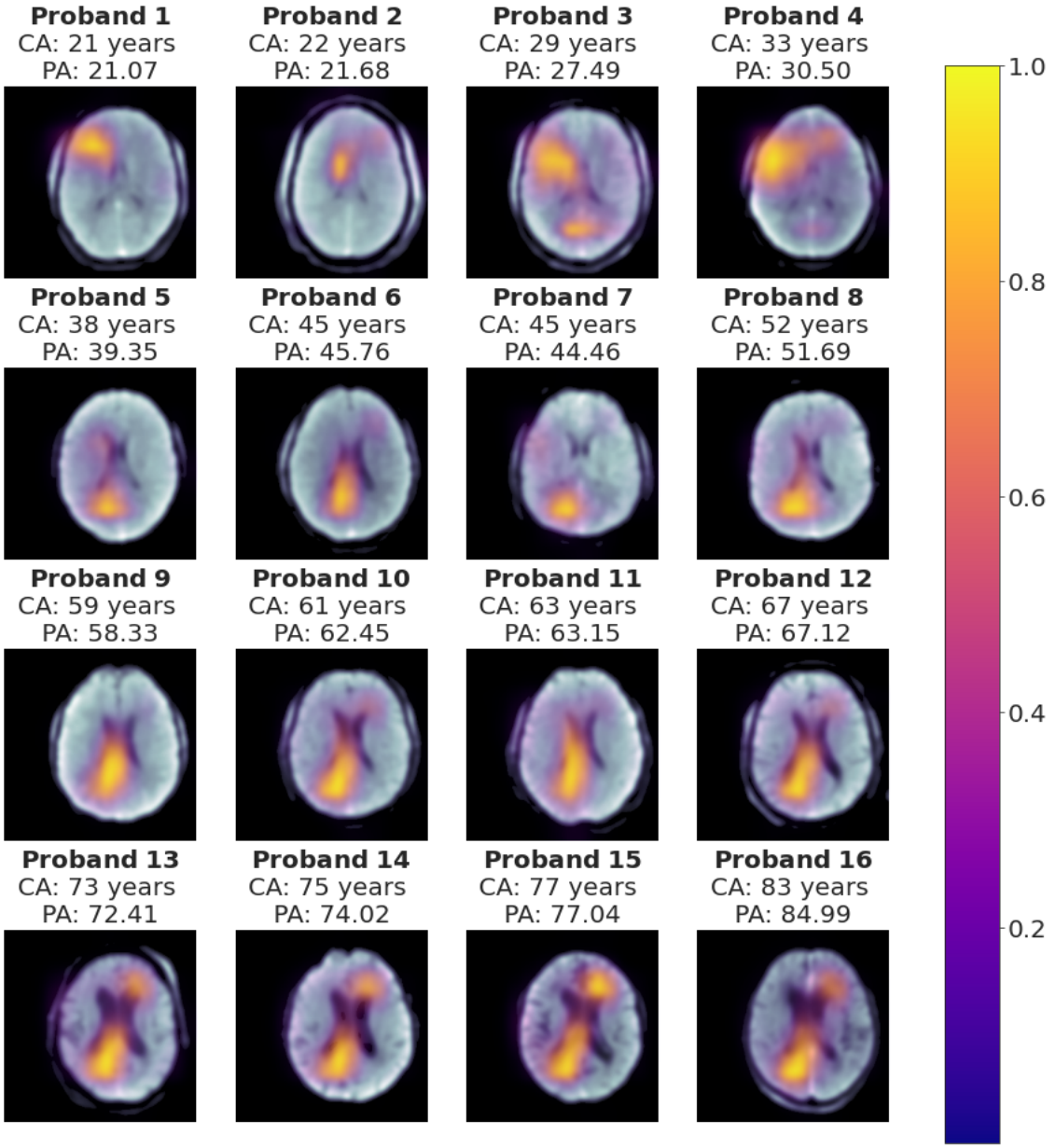


Abbildung 26: Grad-CAM Saliency Maps als Overlay über den verwendeten Gehirnschnitt ausgewählter Probanden.
Zudem ist das Chronologische Alter (CA) und das vorhergesagte Alter (PA) in Jahren angegeben. Mit steigender Intensität der Saliency Map (siehe Legende) ist eine erhöhte Bedeutung der entsprechenden Region für die Altersschätzung vergesellschaftet.

Die Saliency Maps mittelalter Patienten wiesen keine großen Abweichungen von der durchschnittlichen Saliency Map auf. Tendenziell zeigte sich der Bereich im linken Frontalhirn als weniger wichtig für die Vorhersage als der im rechten Occipitalhirn.

Bei den Saliency Maps älterer Patienten schienen die Regionen, die Teile des Ventrikelsystems umfassten, im Vergleich zu den anderen Gruppen noch wichtiger zu sein. So wurden auch Regionen im Verlauf der Seitenventrikel verstärkt hervorgehoben.

3.3 Vergleich der trainierten Modelle

Im Vergleich der absoluten Vorhersagefehler auf dem Testdatensatz zeigte das 10-channel-Netzwerk einen signifikant niedrigeren Vorhersagefehler als das 3-channel-Netzwerk, das Support Vector Regression und das Ridge Regression Modell. Ebenso wies das 3-channel-Netzwerk einen signifikant niedrigeren Testfehler als das Support Vector Regression und Ridge Regression Modell auf. Das Support Vector Regression Modell erreichte einen signifikant niedrigeren Vorhersagefehler als das Ridge Regression Modell.

Die beste Korrelation mit dem Chronologischen Alter erreichte das 10-channel-Netzwerk mit einer Pearson-r von 0,927. Der mittlere absolute Vorhersagefehler und Pearson-r der trainierten Modelle ist in Tabelle 3 dargestellt.

Tabelle 3: Mittlerer absoluter Vorhersagefehler (MAE) und Pearson-r der Altersschätzungen der Modelle mit dem Chronologischen Alter auf dem Testdatensatz

Modell	Test-MAE (Jahre)	Pearson-r
Ridge Regression (Modell 2)	9,84	0,765
Support Vector Regression	8,32	0,862
3-channel-Netzwerk	6,96	0,902
10-channel-Netzwerk	5,73	0,927

4 Diskussion

4.1 Einordnung der Performance

In dieser Arbeit wurde gezeigt, dass automatisierte Altersschätzung auf Basis von klinisch erzeugten Bildgebungsdaten, die auf einer Vielzahl verschiedener CT-Scannern mit unterschiedlichen Schichtdicken und Rekonstruktions-Kernel über einen Zeitraum von 9 Jahren erzeugt wurden, sowohl über einen Feature-basierten Ansatz wie auch einen Deep Learning Ansatz möglich ist. Es wurde eine weite Altersspanne zwischen 20 und 85 Jahren untersucht.

Im Feature-basierten Morphometrie-Ansatz wurden die Volumina von 127 Subregionen des Gehirns von 2021 Patienten automatisiert vermessen und auf dieser Basis sowohl ein Datensatz mit selbigen Subregionen-Volumina als Features als auch ein Datensatz, der einzelne Subregionen zu Regionen höherer Ordnung zusammenfasste und die kombinierten Volumina als Features verwendete, erzeugt. Auf dem Subregionen-Datensatz wurden jeweils ein Ridge Regression und ein nicht-lineares Support Vector Regression Modell trainiert, während ein weiteres Ridge Regression Modell auf dem Datensatz der Regionen höherer Ordnung trainiert wurde.

Im Deep Learning Ansatz wurden 2021 axiale Weichteil-CT Gehirndatensätze automatisiert standardisiert und für das Training vorbereitet. Hierbei wurde ein Datensatz mit 10 verschiedenen axialen Schichten des jeweils untersuchten Gehirns sowie ein Datensatz mit der zentralen axialen Schicht erzeugt. Auf ersterem Datensatz wurde ein Netzwerk mit 10 Input-Channels (10-channel-Netzwerk), auf letzterem eines mit 3 Input-Channels (3-channel-Netzwerk), bei dem in jedem Input-Channel dieselbe Schicht verwendet wurde, trainiert. Zudem wurden visuelle Erklärungen der Vorhersagen mittels Grad-CAM erzeugt.

Von den 5 trainierten Modellen zeigte das 10-channel-Netzwerk mit einem Test-MAE von 5,73 Jahren und R^2 von 0,86 die beste Performance, gefolgt vom 3-channel-Netzwerk mit einem Test-MAE von 6,96 Jahren und R^2 von 0,81 und der Support Vector Regression auf dem Subregionen-Datensatz mit einem Test-MAE von 8,3 Jahren und einem R^2 von 0,74. Das Ridge Regression

Verfahren, das als Benchmark eingesetzt wurde, erreichte auf dem Subregionen Datensatz lediglich einen Test-MAE von 9,840 Jahren und einen R^2 von 0,585. Auf dem Datensatz der Regionen höherer Ordnung konnte das beste Ridge Regression Modell einen Test-MAE von 10,785 Jahren mit einem R^2-Wert von 0,497 erreichen.

Das 10-channel Netzwerk zeigte sich kompetitiv mit in der Literatur beschriebenen Modellen. Diese wiesen zwar nativ betrachtet einen niedrigeren MAE auf, jedoch wurden in den meisten Fällen deutlich weniger breite Altersspannen auf besser kontrollierten und weniger diversen Datensätzen untersucht.

Neben den grundsätzlich unterschiedlichen Studienpopulationen verwendeten die in 1.4 erwähnten Arbeiten größtenteils Aufnahmen des Gehirns aus standardisierten MRT-Studien, während in dieser Arbeit CT-Datensätze verwendet wurden. Aufgrund eines besseren Weichteilkontrastes (Kilcoyne et al., 1988, Chang et al., 1987) kann zudem das MRT die Anatomie des Gehirnes in höherem Detailreichtum abbilden und somit auch kleinere und weniger auffällige Veränderungen spezifischer und sensitiver abbilden als das CT.

Die einzige Arbeit, in der CT-Datensätze für Altersschätzung verwendet wurden, erreichte einen Root Mean Squared Error (RMSE) von 5,21 Jahren, allerdings auf einer deutlich enger umrissenen Altersspanne von 60 bis 100 Jahren mit geringer Standardabweichung von 7 Jahren (Brudfors, 2020b) im Gegensatz zum hier erreichten MAE von 5,73 Jahren bei einer deutlich größeren Altersspanne von 20 bis 85 Jahren mit einer Standardabweichung von 19,4 Jahren.

4.2 Auswirkungen des Datensatzes auf den Vorhersagefehler

4.2.1 Geschlecht

Ein Einfluss des Geschlechts auf den Vorhersagefehler der Altersschätzung zeigte sich nur im 10-channel-Netzwerk. Hier wurden weibliche Patienten signifikant älter eingeschätzt, als sie tatsächlich waren. Das mittlere Alter der

weiblichen Patienten im Trainingsdatensatz war höher als das der Männer und auch höher als das der weiblichen Patienten im Testdatensatz.

In diesem Fall ist es möglich, dass das Netzwerk während des Trainings gelernt hat, weibliche Patienten zu identifizieren und aufgrund ihrer Altersverteilung im Trainingsdatensatz älter einzuschätzen. Hier könnte also das Geschlecht und das damit verbundene höhere Alter im Trainingsdatensatz als konfundierende Variable gelernt worden sein.

Auf der Basis der absoluten Vorhersagefehler konnte kein signifikanter Effekt des Geschlechts in einem der Modelle gezeigt werden.

4.2.2 Alter

Im Ridge Regression Modell wurde das Alter jüngerer und älterer Patientengruppen signifikant schlechter geschätzt als das der mittelalten nahe dem Altersdurchschnitt. In diesem Ansatz erklärt vermutlich die globale, nicht auf einzelne Altersgruppen beschränkte Tendenz, das Patientenalter zum Altersdurchschnitt hin zu schätzen, diesen Umstand. Bei nicht ausreichender Anpassungsmöglichkeit an den Datensatz zeigt ein Modell einen niedrigeren Fehler, wenn es extreme Schätzungen vermeidet.

Überraschenderweise zeigte sich auch ein signifikant höherer absoluter Vorhersagefehler der jüngsten Altersgruppe (20-29 Jahre) gegenüber der mittleren (40-49) und ältesten (70-85) im 3-channel-Netzwerk. Hier scheint das Netzwerk tatsächlich weniger gut in der Lage gewesen zu sein, das Alter junger Patienten einzuschätzen. Möglicherweise wurden Features gelernt, die sich erst im mittleren und höheren Alter manifestieren. Tatsächlich zeigten die Saliency Maps jüngerer Patienten auch eine deutlich unterschiedliche Verteilung für die Vorhersage bedeutender Regionen, in der das Liquor-System eine geringere Rolle zu spielen schien.

Die Zunahme des Ventrikelvolumens, die auch im Morphometrie-Ansatz die Vorhersagen stark beeinflusste, beschleunigt sich beispielsweise erst ab 30 Jahren (Pfefferbaum et al., 1994), so dass diese in den jüngeren Altersgruppen möglicherweise nur eingeschränkt Informationen liefern konnte.

In einer Arbeit von Hepp et al. (2021) wurden Vorhersagefehler und Unsicherheit (Uncertainty) eines Netzwerkes bei der Altersschätzung auf dreidimensionalen MRT-Datensätze aus der NAKO-Studie (Bamberg et al., 2015) berechnet. Bei jungen Patienten zeigten sich sowohl ein signifikant höherer Vorhersagefehler als auch eine erhöhte aleatorische Unsicherheit bei der Vorhersage. Dies könnte darin begründet liegen, dass sich in dieser Altersgruppe die Morphologie bei unterschiedlichem Chronologischen Alter weniger stark variiert, so dass eine Altersschätzung schwerer fällt.

4.2.3 Akquisitionsparameter

Das untersuchte Bildmaterial wurde mithilfe nicht-standardisierter Akquisitonsparameter erzeugt. Verschiedene Arbeiten zeigten, dass der verwendete CT-Scanner, Reconstruction Kernel, die gewählte Schichtdicke sowie weitere Parameter wie die Röhrenspannung und Stromfluss die Bildqualität und -eigenschaften beeinflussen (Lu et al., 2016).

Werden größere Schichtdicken gewählt, ergeben sich stärkere Teilvolumenartefakte (Monnin et al., 2017, Lu et al., 2016), während bei geringerer Schichtdicke mehr Rauschen festzustellen ist (Alshipli and Kabir, 2017).

Die Verwendung unterschiedlicher Reconstruction Kernel kann Bildeigenschaften verändern. So variiert die Dichte von atherosklerotischen Plaques signifikant mit dem gewählten Reconstruction Kernel (Achenbach et al., 2010), Auch verringert sich die Reproduzierbarkeit von verschiedenen Radiomics-Features, die Tumordensität und -textur beschrieben (Choe et al., 2019).

Für die beiden Feature-basierten Ansätzen wurden morphometrische Features berechnet, die die Voxelvolumina verschiedener Gehirnregionen repräsentierten. Dazu war jedoch zuvor eine Atlas-basierte Segmentierung notwendig, die als Ergebnis segmentierungsspezifische Wahrscheinlichkeitskarten erzeugte. Diese wurden auf der Basis der Korrelation eines Voxels mit der Atlasposition und des Dichtewertes berechnet. Hier können Akquisitionsparameter die Segmentierung beeinflusst haben.

Convolutional Neural Networks lernen Dichtewert-basierte Features. Hier kann es aufgrund unterschiedlicher Akquisitionsparameter zu Vorhersagefehlern kommen. Mit variierender Schichtdicke können unterschiedlich starke Teilvolumenartefakte entstehen, auch das Rauschen wird durch Wahl der Schichtdicke verändert. Dies könnte sich negativ auf die Performance von Deep Learning Algorithmen auswirken.

Jedoch schien nach dem gemeinsamen, standardisierten Preprocessing (Registrierung auf ein gemeinsames Template und Vermessung der Gehirnregionen, Resampling), weder das verwendete Gerät, noch Reconstruction-Kernel, Schichtdicke, Röhrenspannung oder die Menge an verwendeter Strahlung einen signifikanten Einfluss auf den absoluten Messfehler zu haben, weder im Feature-basierten noch im Deep Learning Ansatz. Zudem profitierte die Altersschätzung nicht von einer erhöhten Strahlendosis mit erhöhter Exposition der Patienten.

Somit zeigten alle Ansätze eine robuste Performance über alle CT-Geräte, Reconstruction-Kernel, Schichtdicken und Röhrenspannungen hinweg, die ebenso unabhängig von der verwendeten Strahlendosis war.

Die Vielfältigkeit und fehlende Standardisierung im Datensatz kann darüber hinaus auch als seine Stärke angesehen werden, vor allem im Bereich des Deep Learning. Convolutional Neural Networks können subtile Eigenschaften ihres Trainingsdatensatzes lernen, die auf anderen Datensätzen wahrscheinlich nicht vorhanden sind. In diesem Fall lässt sich in manchen Fällen, vor allem wenn die gelernten Eigenschaften die Vorhersageaufgabe konfundieren, eine schlechtere Generalisierung auf ungesehene Datensätze feststellen (Zech et al., 2018). Insofern kann eine geringere Standardisierung des Trainingsdatensatzes eine nützliche Regularisierung darstellen und zu einer besseren Generalisierung auf ungesehen Daten beitragen, auch wenn unter Umständen der erreichte Vorhersagefehler höher ist.

4.3 Bildgebungs-Biomarker für den Alterungsprozess

4.3.1 Feature-basiertes Maschinelles Lernen

Die Koeffizienten der trainierten Ridge Regression Modell wurden untersucht. Hierbei sei explizit darauf hingewiesen, dass die gewählten Koeffizienten nur die gewählte Fehlerfunktion minimieren. Ein kausaler Zusammenhang kann nicht direkt geschlussfolgert werden.

Das Ridge Regression Modell, das auf dem Datensatz der Regionen höherer Ordnung trainiert wurde (Modell 1), zeigte, dass ein steigendes Liquor-Volumen der stärkste Prädiktor für einen Patienten höheren Alters ist, gefolgt vom Volumen der Weißen Substanz. Auch die Koeffizienten des Volumens des temporalen Cortex und parietalen Cortex waren positiv, wenn auch nur gering. Das Volumen des Diencephalon hatte kaum Auswirkungen auf die Altersvorhersage. Ein höheres Volumen des limbischen Systems, Cerebellums, frontalen und occipitalen Cortex war stattdessen ein Prädiktor für ein jüngeres Alter.

Die Feature-Koeffizienten des Ridge Regression Modells, das auf dem Subregionen Datensatz (Modell 2) trainiert wurde, sind schwerer zu interpretieren und vermitteln kein eindeutiges Bild.

So weist ein erhöhtes Liquor-Volumen auch in diesem Modell auf einen älteren Patienten hin. Während das Volumen der Weißen Substanz in Modell 1 eher Prädiktor für einen älteren Patienten war, ist hier das Volumen der cerebralen Weißen Substanz der stärkste Prädiktor für einen jüngeren Patienten, im Gegensatz zur cerebellären Weißen Substanz.

Die Zunahme des Liquorvolumens und Abnahme des Volumens der Grauen Substanz im Alter sind in der Literatur gut belegt (Sowell et al., 2003, Taki et al., 2011, Pfefferbaum et al., 1994).

Überraschenderweise scheint auch das Volumen der Weißen Substanz eher ein Prädiktor für einen älteren Menschen zu sein, obwohl ihr Volumen recht stabil bleibt (Pfefferbaum et al., 1994) oder bis zu einem mittleren Alter sogar ansteigt, um erst dann abzufallen (Sowell et al., 2003). Dieser Fakt ließe sich

damit erklären, dass vermutlich die Abnahme der Grauen Substanz entscheidender für die Altersvorhersage ist, während die Abnahme der Weißen Substanz nicht im selben Maße erfolgt und daher relativ zum Volumen der Grauen Substanz zunimmt. Dies wurde auch in einer Arbeit gezeigt, in der das Verhältnis von Grauer Substanz, Weißer Substanz und Liquor jeweils zum gesamten Hirnvolumen vermessen wurde (Taki et al., 2011). Während das Verhältnis der Grauen Substanz mit dem Alter ab- und das des Liquors zunimmt, veränderte sich das Volumenverhältnis der weißen Substanz nicht signifikant mit dem Alter (Taki et al., 2011).

In einigen Studien konnte ein signifikantes Absinken des zerebellaren Volumens festgestellt werden (Raz et al., 2005, Koppelmans et al., 2015), während andere vor allem eine Abnahme der weißen Substanz des Cerebellums beschrieben (Hoogendam et al., 2012) und andere keinen Einfluss des Alters auf das Volumen des Cerebellums feststellen konnten (Bergfield et al., 2010). In einer Arbeit wurde ein Zusammenhang von Diabetes und geringerem Volumen des Cerebellums beschrieben (Hoogendam et al., 2012). Des Weiteren wurde geschlussfolgert, dass Faktoren, die zum Altern und zur Volumenabnahme beitragen, nicht vollständig mit denen für Altern und Volumenabnahme des Cerebrums kongruent sind. Möglicherweise ist dies der Grund, warum das zerebellare Volumen nach dem limbischen System den am stärksten negativen Koeffizienten im Ridge Regression Modell 1 trägt. Andererseits handelt es sich bei der grauen Substanz des Cerebellums um eine feine Struktur, so dass Messungenauigkeiten einen größeren Einfluss als auf andere Hirnregionen gehabt haben können.

Auch das Volumen des Hippocampus, Teil des limbischen Systems, sinkt mit dem Alter, die Abnahme beschleunigt sich dabei sogar (Raz et al., 2005). Insgesamt ist die Geschwindigkeit der Volumenabnahme ab einem gewissen Alter größer als die der gesamten Grauen Substanz und auch schneller als die der benachbarten Areale des Temporalen Cortex (Nobis et al., 2019). Dies könnte erklären, warum das Volumen des limbischen Systems den am stärksten negativen Koeffizienten erhielt.

Das Volumen zentraler Strukturen der Basalganglien (Striatum, Globus Pallidus sowie des Thalamus) sinkt ebenfalls mit dem Alter (Tullo et al., 2019).

Für das Volumen des frontalen Cortex wurde ein negativer Koeffizient bestimmt, für die des parietalen und temporalen Cortex jedoch ein positiver.

Eine Arbeit zeigte, dass sowohl der frontale, als auch der parietale und temporale Cortex im Alter an Volumen verlieren (Resnick et al., 2003). Die Abnahme des Volumens der grauen Substanz des frontalen Cortex verläuft dabei am schnellsten, während dies des parietalen und temporalen Cortex etwas langsamer vonstattengeht. Auch hier könnte eine ähnliche Erklärung greifen wie in der Fragestellung der unterschiedlichen Koeffizienten der Volumina Weißen und Grauen Substanz. So ist vermutlich ein hohes Volumen des frontalen Cortex aussagekräftiger für ein jüngeres Testsubjekt und die Volumina des temporalen und parietalen Cortex liefern weniger Zusatzinformationen.

4.3.2 Deep Learning

Grad-CAM wurde auf das beste trainierte 3-channel-Netzwerk angewendet. Die gemittelte Saliency Map (Abbildung 25) wies ein klares Muster auf. Eine Region im frontalen linken Gehirn, die Teile des Vorderhorns des linken Seitenventrikels umfassten, sowie auch die Regionen der Basalganglien und den Forceps minor des Corpus Callosum wurde hervorgehoben. Zudem wurde ein Bereich des occipitalen rechten Gehirns, der vor allem im Bereich der weißen Substanz lag, aber auch Teile des Hinterhorns des rechten Ventrikels und des Areals des Cuneus und Precuneus enthielt, durch Grad-CAM hervorgehoben.

Es muss allerdings darauf hingewiesen werden, dass die gezeigten Saliency Maps Interpolationen einer im Raum des letzten Convolutional Layers berechneten Matrix der Dimension 14x14 Pixel auf den Raum des Eingabebildes darstellen. Somit ist fraglich, inwiefern die erwähnten Sub-Regionen sinnvoll abgegrenzt werden können.

Wie auch im Feature-basierten Ansatz scheint das Liquor-System für die Altersschätzung eine wichtige Rolle zu spielen. In nahezu jeder berechneten Saliency Map sind Anteile des Ventrikelsystems hervorgehoben. Hier fand sich eine interessante Übereinstimmung zwischen dem Deep Learning Ansatz und dem Feature-basierten Ansatz, in dem das Liquor-Volumen ebenfalls eine wichtige Rolle bei der Vorhersage spielte.

Es wurden auch spezifische Saliency Maps für jeden Bilddatensatz des Test-Datensatzes erstellt, insgesamt also 304 einzelne Saliency Maps. In diesen zeigte sich das vorbeschriebene Muster, das sich aus der Mittelung aller Saliency Maps ergab, vor allem bei älteren Patienten. Die Saliency Maps jüngerer Patienten variierten dabei deutlich stärker. Bei jüngeren Patienten fehlte zudem häufig der Fokus auf dem rechten Hinterhorn und occipitalen Gehirn, während variierende Anteile vom frontalen Cortex hervorgehoben wurden. Hier scheint das Netzwerk grundlegend unterschiedliche Regionen zur Vorhersage des Alters verwendet zu haben.

Die mittels Grad-CAM erzeugten Saliency Maps zeigten ähnliche Ergebnisse wie bereits in der Literatur beschrieben (Hepp et al., 2021). Auch hier waren, wie in dieser Arbeit, die Regionen um die Vorherhörner der Seitenventrikel von großer Bedeutung für die Altersschätzung. Patches aus diesem Bereich wurden gesondert analysiert. Zudem wurden zentrale Hirnregionen sowie Teile der Insula hervorgehoben.

Einige in dieser Arbeit generierte sowie die gemittelte Saliency Map zeigten ebenfalls Aufmerksamkeit für zentrale Hirnregionen und Regionen der Insula, ersteres vor allem bei älteren Patienten, letzteres eher bei jüngeren.

Diese Übereinstimmungen sind vor dem Hintergrund von großem Interesse, dass sich sowohl Netzwerkarchitektur, als auch verwendete Modalität, Preprocessing und Patientenkollektiv unterschieden. Hier scheinen sich tatsächlich wichtige anatomische Regionen für die Einschätzung des Patientenalters zu finden.

4.4 Limitationen

Die genutzten Datensätze wurden im klinischen Alltag im Schockraum der Universität Tübingen über einen Zeitraum von fast 10 Jahren auf unterschiedlichen Geräten und unter Verwendung von unterschiedlichen Aufnahmeparametern (Schichtdicke und Rekonstruktionskernel) aufgenommen.

Im Rahmen der Schockraum-Diagnostik ist ein Schädel-CT absolut indiziert, wenn neurologische Auffälligkeiten oder der Verdacht auf eine Schädelfraktur, beispielsweise auch aufgrund des Unfallmechanismus, bestehen. Fakultative Indikationen umfassen jedoch auch unklare Unfälle, starke Kopfschmerzen oder Intoxikationen (Firsching et al., 2015).

Die häufigsten Ursachen für schwere Verletzungen und Polytrauma sind Unfälle, darunter vor allem Verkehrsunfälle, aber auch Stürze aus größeren Höhen und Suizidversuche. Jede Altersgruppe ist in solche Unfälle verwickelt.

Aufgrund dieser möglichen Unfallmechanismen und Indikationen sind viele im Schockraum durchgeführte Schädel-CTs unauffällig und bilden eine weite Altersspanne ab.

Allerdings ist die Altersverteilung der Patienten im Schockraum unterschiedlich von der Gesamtbevölkerung. In dieser Arbeit waren Männer, vor allem in den jüngeren Altersgruppen, überrepräsentiert.

Männer sind häufiger Opfer von Verletzungen. So neigen junge Männer eher zu risiko-freudigem Verhalten, zum Beispiel beim Autofahren, und gehen häufiger gefährlicheren Arbeiten nach. Zudem sind sie eher in tätliche Auseinandersetzungen verwickelt.

Bei älteren Patienten dominieren Verletzungen aufgrund abnehmender Fitness (Lecky et al., 2010). In jeder Altersgruppe sind in dieser Arbeit mehr Männer als Frauen repräsentiert, auch wenn der Frauenanteil im höheren Alter zunimmt. Dies könnte möglicherweise mit einer höheren Frailty von Frauen in diesem Alter erklärt werden (Gordon and Hubbard, 2020).

Die ungleiche Geschlechterverteilung hatte jedoch nur im 10-channel-Netzwerk Auswirkungen auf den absoluten Vorhersagefehler, während sich bei den

anderen trainierten Modellen keine signifikanten Unterschiede zwischen den Geschlechtern bezüglich des Vorhersagefehlers zeigten.

Ziel dieser Arbeit war es, Biomarker für „gesundes“ Altern zu identifizieren. Aus diesen Gründen wurden offensichtliche Pathologien und Artefakte möglichst ausgeschlossen.

Einerseits gelang dies nicht vollständig. Bei Patienten, die Pathologien aufwiesen, zeigten Saliency Maps jedoch keine besondere Aufmerksamkeit für diese Regionen, was allerdings nicht bedeutet, dass diese keine Auswirkung auf die Altersschätzung hatten.

Allerdings erhöht sich die Wahrscheinlichkeit für Pathologien mit steigendem Alter, so dass es möglicherweise sinnvoll gewesen wäre, wenn das Netzwerk diese als Hinweis auf höheres Patientenalter gelernt hätte. Andererseits wiesen auch einige junge Patienten aufgrund des erlittenen Unfalls Schädelverletzungen und intrakranielle Blutungen auf.

Solche großen Pathologien sowie Artefakte können dazu führen, dass ein Bild sich grundlegend von der Menge an Bildern unterscheidet, auf denen das Netzwerk die Vorhersagen lernt. Dies kann im Lernprozess zu schlechteren Vorhersagen führen.

Im Feature-basierten Ansatz ergaben sich weiterhin Probleme dadurch, dass die automatisierte Vermessung der verschiedenen Hirnregionen teilweise ungenau war. Dies lässt sich dadurch erklären, dass der verwendete Atlas, auf dessen Basis den segmentierungsspezifischen Wahrscheinlichkeitskarten Regionen-Label zugewiesen wurden, trotz des Preprocessings nicht optimal übereinstimmte. Die Patienten-Datensätze wurden auf ein Template registriert, resampled, und segmentiert, um eine möglichst hohe Standardisierung zu erreichen. Trotzdem waren die Köpfe mancher Patienten in den Aufnahmen leicht gedreht oder gekippt, so dass einzelne Gehirnregionen verzerrt wurden oder auch kleine Regionen nicht im entsprechenden Atlas-Gebiet lagen. Somit sind die Volumen-Messungen nicht vollkommen zuverlässig und könnten gerade für kleinere Regionen schwanken. Systematische Fehler können so nicht ausgeschlossen werden.

Die verwendeten Verfahren des klassischen Maschinellen Lernens lernen eine optimale Abbildung der Features auf die Vorhersagewerte. Daher lassen sich nicht zuverlässig kausale Beziehungen in die gewählten Koeffizienten interpretieren. Gerade im Modell 2, das auf dem Subregionen-Datensatz trainiert wurde, zeigte sich die Interpretation schwierig. Möglicherweise wurde in diesem Zusammenhang auch viel zusammenhangloses Rauschen in den Trainingsdaten interpretiert. Modell 1, das auf dem Datensatz der Regionen höherer Ordnung trainiert wurde, sollte diesem Problem gegenüber robuster sein. Jedoch ist auch hier die Interpretation fraglich.

Es zeigten sich in dieser Arbeit bei der Analyse der Auswirkungen der Akquisitionsparameter auf den Vorhersagefehler keine signifikanten Unterschiede zwischen den Gruppen. Allerdings waren die Gruppengrößen in den Kategorien der verwendeten Scanner, Reconstruction Kernel und Schichtdicken stark unterschiedlich groß, mit einigen Gruppen von geringer Stichprobenzahl. Hier könnten möglicherweise signifikante Unterschiede durch die geringe Stichprobenzahl verborgen geblieben sein.

Overfitting stellt ein großes Problem des Maschinellen Lernens dar. Convolutional Neural Networks wie das VGG16 haben Millionen von veränderlichen Parametern. Vor allem, wenn der Trainingsdatensatz klein ist, können neben sinnvollen, generalisierbaren Features auch spezifische Eigenschaften des Datensatzes gelernt werden, im schlimmsten Fall irrelevantes Rauschen, das nur zufällig mit der zu lernenden optimalen Funktion korreliert. Größere Trainingsdatensätze können in dieser Hinsicht zu einer besseren Generalisierung auf ungesehene Daten beitragen.

In dieser Arbeit wurden 1414 Trainings-Datenpunkte verwendet (mit 303 Datenpunkten im Validierungs- und 304 im Testdatensatz). Die NAKO-Studie, die beispielsweise in der Arbeit Hepp et al. (2021) als Grundlage für das Training verwendet wurde, umfasst 10691 Datensätze. Die meisten von Cole et al. untersuchten Studien hatten jedoch deutlich weniger Trainingsdaten. Auch Huang et al. (2017) verwendeten lediglich 600 Trainingsdatenpunkte. Somit scheint die Anzahl an Trainingsdatensätzen zwar angemessen, trotzdem

könnte die Vorhersagegenauigkeit durch das Training auf einem besser kontrollierten und größeren Datensatz verbessert werden.

Mithilfe von Grad-CAM lassen sich Einblicke in das Innenleben eines Convolutional Neural Networks gewinnen. So können Vorhersagen besser verstanden und neue Hypothesen generiert werden. Allerdings sind die Ergebnisse für Menschen häufig schwer zu interpretieren und lassen nicht unbedingt sinnvolle Schlüsse auf die Prozesse zu, die zur entsprechenden Vorhersage führen (Rudin, 2019).

Zudem war die Auflösung der berechneten Saliency Maps mit 14x14 Pixeln in Bezug auf das Originalbild mit 224x224 Pixel recht niedrig. Es können somit auch keine genauen Rückschlüsse auf die ausgewählten Features vorgenommen werden.

Nichtsdestotrotz erschienen die in dieser Arbeit erzeugten Saliency Maps in ihrer räumlichen Ausbreitung sinnvoll und valide und differenzierten in ihrer Darstellung zwischen jungen und alten Patienten.

4.5 Ausblick

Diese Arbeit liefert vielversprechende Ergebnisse für die tiefergehende Anwendung von Feature- sowie Deep Learning-basierten Anwendungen auf im klinischen Kontext erzeugten CT-Datensätzen mit nicht-standardisierten Akquisitionsparametern.

In dieser Arbeit wurden die Volumina verschiedener Hirnregionen automatisiert vermessen. Die Volumina von Gehirnregionen lassen jedoch nicht nur eine Einschätzung des Alters zu, sondern haben auch einen Zusammenhang mit physiologischen Parametern, beispielsweise kognitiven Fähigkeiten. So wiesen beispielsweise Footballspieler, die wiederholten Kopf-Zusammenstößen ausgesetzt waren, verminderte Volumina im limbischen System auf. Diese waren auch mit niedrigerer Aufmerksamkeit, psychomotorischer Geschwindigkeit und visuellen Gedächtnisleistungen vergesellschaftet (Lepage et al., 2019). Bei älteren Patienten besteht eine Beziehung zwischen dem Volumen der grauen Substanz im Cerebellum und den kognitiven Fähigkeiten

(Hogan et al., 2011), während das Volumen des Hippocampus einen Zusammenhang mit regionalen Pathologien und abnormalen Alterungsprozessen hat (De Marco et al., 2019) und auch einen sensitiven Marker für neurodegenerative Prozesse darstellt (Frisoni et al., 2010). Hier bestehen weitere interessante Anwendungsfälle.

Es wurde in vorhergehenden Arbeiten bereits gezeigt, dass genaue Vorhersagen des Chronologischen Alters durch auf gesunden Kohorten trainierte Modelle des Maschinellen Lernens als Bildgebungs-Biomarker eingesetzt werden können (Cole et al., 2019). Entsprechend wäre die Korrelation der Vorhersagen der trainierten Modelle mit klinischen Daten der untersuchten Patienten interessant.

Die bisher beschriebenen Ansätze basieren zumeist auf MRT-Daten. Jedoch werden im Rahmen vieler klinischer Untersuchungen auch CT-Datensätze des Gehirns erzeugt, die im Rahmen der Früherkennung neurologischer Krankheiten ähnlich nützlich sein könnten. In dieser Stoßrichtung ist weitere Forschung an größeren Patientenkollektiven mit mehr klinischen Daten notwendig.

Die Erstellung von visuellen Erklärungen mithilfe von Saliency Maps ist nützlich, um neues Wissen aus Bilddaten und zusätzliches Vertrauen in die trainierten Modelle durch erhöhte Nachvollziehbarkeit der Vorhersagen zu generieren.

Zudem ist die Feststellung, dass eine ausreichend genaue Altersschätzung auf einem kleinen und heterogenen Datensatz möglich ist, wertvoll. Die Integration von Modellen des Maschinellen Lernens in sicherheitskritische Infrastruktur wie das Gesundheitssystem erfordert, dass diese Modelle verlässlich und robust arbeiten. Dies ist eine zentrale Herausforderung beim Einsatz in einem reellen Szenario (Taori et al., 2020). Eine zentrale Annahme des Maschinellen Lernens ist, dass die Daten, die für Training und Test eingesetzt werden, aus demselben Feature-Raum stammen und dieselbe Verteilung aufweisen (Pan and Yang, 2009, Lu et al., 2015, Weiss et al., 2016). Dies ist jedoch bei den meisten Anwendungen in der echten Welt meist nicht gegeben (Pan and Yang, 2009).

Wie in 1.3.2 ausgeführt, konnten Zech et al. (2018) zeigen, dass das spezifische Krankenhaus, indem eine Röntgenaufnahme gemacht wurde, verlässlich mithilfe eines Convolutional Neural Network bestimmt werden konnte. Es wurde angenommen, dass subtile krankenhausspezifische Charakteristika in den erzeugten Bilddaten, die beeinflusst wurden durch unterschiedliche Standardpraktiken, genutzte Geräte, Akquisitions-Protokolle, Bildverarbeitungsschritte, etc. diese Identifizierung möglich machten. Diese Charakteristika betreffen auch die zur Altersschätzung verwendeten Features, so dass ein Distributions-Shift auftritt. Ein solcher kann die Funktion eines Modells beeinträchtigen (Taori et al., 2020).

Während sich einige Charakteristika der Bildgebungsdaten verändern, bleiben die zugrundeliegenden Features, die die Altersschätzung ermöglichen, relativ stabil. Es ist entsprechend von Vorteil, sich an neue Daten mit anderen Charakteristika anzupassen, während das zuvor gelernte Wissen beibehalten wird. Transferlernen, das auf die Übertragung von erlerntem Wissen aus einer bereits bearbeiteten Aufgabe auf eine verwandte abzielt (Torrey and Shavlik, 2010), stellt eine mögliche Lösung dar. Beispielsweise könnte es zukünftig für ein Krankenhaus, das sich für die Nutzung von Künstlicher Intelligenz entscheidet, möglich sein, ein bereits vortrainiertes Netzwerk auf ihren spezifischen Daten, die im klinischen Alltag erzeugt werden, feinabzustimmen. Hierzu könnte ein Vorgehen, wie es in dieser Arbeit gewählt wurde, sinnvoll sein.

Nachdem eine gute Performance von Algorithmen des Maschinellen Lernens auf CT-Gehirndatensätzen erreicht werden konnte, ist dieses auch auf Thorax/Abdomen- oder Ganzkörperdatensätzen anwendbar.

In dieser Arbeit wurde die Aufgabe der Altersschätzung als Beispiel gewählt, um zu zeigen, dass mithilfe von Algorithmen des Maschinellen Lernens auf der Basis von klinischen Bildgebungsdaten abstrakte Fragestellung automatisiert bearbeitet werden können, so dass zukünftig auf dieser Basis weitere spezifische Anwendungen entwickelt werden können.

5 Zusammenfassung

Das Konzept des Biologischen Alters wird in der Literatur als wichtiger Prognosemarker diskutiert. So konnte gezeigt werden, dass ein höheres geschätztes biologisches Gehirnalter beispielsweise mit neurodegenerativen Erkrankungen assoziiert ist. Die Erfassung des Biologischen Alters eines Menschen ist herausfordernd, weshalb verlässliche Indikatoren für eine genaue Einschätzung benötigt werden. Altersschätzungen von Modellen des Maschinellen Lernens, die auf Bildgebungsdaten gesunder Kohorten trainiert wurden, können zu diesem Zweck als Bildgebungs-Biomarker eingesetzt werden. Allerdings wurde das Biologische Gehirnalter vorwiegend im Forschungskontext untersucht, wobei vor allem MRT-Datensätze basierend auf standardisierten Studien verwendet wurden. Die Ergebnisse sind also nicht direkt auf die klinische Realität übertragbar.

In dieser Arbeit wurde daher die automatisierte Schätzung des Gehirnalters auf CT-Bildgebungsdaten, die im Rahmen der klinischen Routine erzeugt wurden, mittels Methoden des Maschinellen Lernens durchgeführt. Als Grundlage diente ein Datensatz von 2021 Patienten im Alter zwischen 20 und 85 Jahren, die zwischen 2010 und 2019 ein axiales Weichteil-CT des Gehirns im Schockraum der Universitätsklinik Tübingen erhalten hatten, welches als unauffällig bewertet wurde. Ein Feature-basierter und ein Deep Learning-Ansatz wurden verfolgt.

Im Rahmen des Feature-basierten Ansatzes wurden die Volumina von 127 Gehirnregionen automatisiert vermessen. Auf Basis der volumetrischen Daten wurden ein Ridge Regression und ein Support Vector Regression Modell mittels 5-fold-Cross-Validation für die Gehirnaltersschätzung trainiert und validiert. Die Modelle wurden auf einem ungesehenen Testdatensatz getestet und die Modell-Parameter ausgelesen.

Im Deep Learning-Ansatz wurden Convolutional Neural Networks für die CT-basierte Gehirnaltersschätzung implementiert, trainiert und validiert. Es kam zur Konvergenz auf dem Trainings- und Validierungsdatensatz. Die trainierten Modelle wurden auf einem zuvor ungesehenen Testdatensatz getestet. Visuelle Erklärungen wurden mittels Grad-CAM Saliency Maps erzeugt.

Es zeigte sich ein Mean Average Error (MAE) von 9,8 Jahren im Ridge Regression Modell und ein MAE von 8,3 Jahren im Support Vector Regression Modell. Das Liquorvolumen stellte den stärksten Prädiktor für ein höheres Patientenalter dar.

Im Deep Learning-Ansatz konnten signifikant bessere Altersschätzungen erreicht werden, wobei das beste trainierte Modell einen MAE von 5,73 Jahren aufwies. Die Genauigkeit der CT-basierten Altersschätzung war in allen Altersgruppen und zwischen männlichen und weiblichen Patienten weitgehend einheitlich. Vor allem zeigten die Akquisitionsparameter keinen signifikanten Einfluss auf die Genauigkeit der Altersschätzung. In den erzeugten Saliency Maps stellten sich zuverlässig Regionen um das Vorderhorn des linken Seitenventrikels sowie am Hinterhorn des rechten Seitenventrikels als von hoher Bedeutung für die Altersschätzung dar.

In dieser Arbeit konnte ich zeigen, dass mittels heterogener, klinisch erzeugter CT-Bildgebungsdaten eine hinreichend genaue Schätzung des Patientenalters erreicht werden kann. Die dargelegte Methodik kann in Zukunft nach weiterer Erforschung und klinischer Erprobung zur Unterstützung bei Entscheidungen von Diagnostik, Therapie und Nachsorge genutzt werden.

6 Literaturverzeichnis

ACHENBACH, S., BOEHMER, K., PFLEDERER, T., ROPERS, D., SELTMANN, M., LELL, M., ANDERS, K., KUETTNER, A., UDER, M., DANIEL, W. G. & MARWAN, M. 2010. Influence of slice thickness and reconstruction kernel on the computed tomographic attenuation of coronary atherosclerotic plaque. *Journal of Cardiovascular Computed Tomography,* 4**,** 110-115.

ALSHIPLI, M. & KABIR, N. A. 2017. Effect of slice thickness on image noise and diagnostic content of single-source-dual energy computed tomography. *Journal of Physics: Conference Series,* 851**,** 012005.

ANASTASOPOULOS, C., REISERT, M. & KELLNER, E. 2017. "Nora Imaging": A Web-Based Platform for Medical Imaging. *Neuropediatrics,* 48**,** P26.

ARMANIOUS, K., ABDULATIF, S., SHI, W., SALIAN, S., KüSTNER, T., WEISKOPF, D., HEPP, T., GATIDIS, S. & YANG, B. 2021. Age-Net: An MRI-Based Iterative Framework for Brain Biological Age Estimation. *IEEE Transactions on Medical Imaging,* 40**,** 1778-1791.

AVANTS, B. B., TUSTISON, N. & SONG, G. 2009. Advanced normalization tools (ANTS). *Insight j,* 2**,** 1-35.

BAMBERG, F., KAUCZOR, H.-U., WECKBACH, S., SCHLETT, C. L., FORSTING, M., LADD, S. C., GREISER, K. H., WEBER, M.-A., SCHULZ-MENGER, J. & NIENDORF, T. 2015. Whole-body MR imaging in the German National Cohort: rationale, design, and technical background. *Radiology,* 277**,** 206-220.

BENGIO, Y., COURVILLE, A. & VINCENT, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence,* 35**,** 1798-1828.

BERGFIELD, K. L., HANSON, K. D., CHEN, K., TEIPEL, S. J., HAMPEL, H., RAPOPORT, S. I., MOELLER, J. R. & ALEXANDER, G. E. 2010. Age-related networks of regional covariance in MRI gray matter: reproducible multivariate patterns in healthy aging. *NeuroImage,* 49**,** 1750-1759.

BLACKBURN, E. H., EPEL, E. S. & LIN, J. 2015. Human telomere biology: a contributory and interactive factor in aging, disease risks, and protection. *Science,* 350**,** 1193-1198.

BRETT, M., MARKIEWICZ, C. J., HANKE, M., CôTé, M.-A., CIPOLLINI, B., MCCARTHY, P., JARECKA, D., CHENG, C. P., HALCHENKO, Y. O., COTTAAR, M., LARSON, E., GHOSH, S., WASSERMANN, D., GERHARD, S., LEE, G. R., WANG, H.-T., KASTMAN, E., KACZMARZYK, J. & GUIDOTTI, R. 2020. nipy/nibabel: 3.2.1 (3.2.1). *Zenodo*.

BRUDFORS, M. 2020a. CTseg. https://github.com/WCHN/CTseg.

BRUDFORS, M. 2020b. *Generative Models for Preprocessing of Hospital Brain Scans.* UCL (University College London).

BRUDFORS, M., BALBASTRE, Y., FLANDIN, G., NACHEV, P. & ASHBURNER, J. Flexible bayesian modelling for nonlinear image registration. Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23, 2020. Springer, 253-263.

BRULS, R. & KWEE, R. 2020. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. *Insights into Imaging,* 11**,** 1-7.
BRUNO, M. A., WALKER, E. A. & ABUJUDEH, H. H. 2015. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *RadioGraphics,* 35**,** 1668-1676.
CHANG, A. E., MATORY, Y. L., DWYER, A. J., HILL, S. C., GIRTON, M. E., STEINBERG, S. M., KNOP, R. H., FRANK, J. A., HYAMS, D. & DOPPMAN, J. L. 1987. Magnetic resonance imaging versus computed tomography in the evaluation of soft tissue tumors of the extremities. *Annals of surgery,* 205**,** 340.
CHANG, L., WONG, V., NAKAMA, H., WATTERS, M., RAMONES, D., MILLER, E. N., CLOAK, C. & ERNST, T. 2008. Greater Than Age-Related Changes in Brain Diffusion of HIV Patients After 1 Year. *Journal of Neuroimmune Pharmacology,* 3**,** 265-274.
CHOE, J., LEE, S. M., DO, K.-H., LEE, G., LEE, J.-G., LEE, S. M. & SEO, J. B. 2019. Deep learning–based image conversion of CT reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology,* 292**,** 365-373.
CHOY, G., KHALILZADEH, O., MICHALSKI, M., DO, S., SAMIR, A. E., PIANYKH, O. S., GEIS, J. R., PANDHARIPANDE, P. V., BRINK, J. A. & DREYER, K. J. 2018. Current applications and future impact of machine learning in radiology. *Radiology,* 288**,** 318-328.
COLE, J. H. & FRANKE, K. 2017. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences,* 40**,** 681-690.
COLE, J. H., FRANKE, K. & CHERBUIN, N. 2019. Quantification of the biological age of the brain using neuroimaging. *Biomarkers of human aging.* Springer.
COLE, J. H., RITCHIE, S. J., BASTIN, M. E., HERNáNDEZ, M. V., MANIEGA, S. M., ROYLE, N., CORLEY, J., PATTIE, A., HARRIS, S. E. & ZHANG, Q. 2018. Brain age predicts mortality. *Molecular psychiatry,* 23**,** 1385-1392.
COLLETTE, A. 2017. H5Py/H5Py: 2.4.0.
CORTES, C. & VAPNIK, V. 1995. Support-vector networks. *Machine learning,* 20**,** 273-297.
DE MARCO, M., OURSELIN, S. & VENNERI, A. 2019. Age and hippocampal volume predict distinct parts of default mode network activity. *Scientific reports,* 9**,** 16075-16075.
DHINGRA, R. & VASAN, R. S. 2012. Age as a risk factor. *The Medical clinics of North America,* 96**,** 87-91.
DUMOULIN, V. & VISIN, F. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*.
EHTESHAMI BEJNORDI, B., VETA, M. & DIEST, P. 2017. van, Ginneken B, van, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA,* 318**,** 2199-210.

ESTEVA, A., KUPREL, B., NOVOA, R. A., KO, J., SWETTER, S. M., BLAU, H. M. & THRUN, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature,* 542**,** 115-118.

FEDARKO, N. S. 2011. The biology of aging and frailty. *Clinics in geriatric medicine,* 27**,** 27-37.

FIRSCHING, R., RICKELS, E., MAUER, U., SAKOWITZ, O., MESSING-JüNGER, M., ENGELHARD, K., SCHWENKREIS, P., LINN, J., BIBERTHALER, P. & SCHWERDTFEGER, K. 2015. Leitlinie Schädel-Hirn-Trauma im Erwachsenenalter. *Update,* 7.

FRANKE, K., GASER, C., MANOR, B. & NOVAK, V. 2013. Advanced BrainAGE in older adults with type 2 diabetes mellitus. *Frontiers in aging neuroscience,* 5**,** 90.

FRISONI, G. B., FOX, N. C., JACK, C. R., JR., SCHELTENS, P. & THOMPSON, P. M. 2010. The clinical use of structural MRI in Alzheimer disease. *Nature reviews. Neurology,* 6**,** 67-77.

FRüH, M., FISCHER, M., SCHILLING, A., GATIDIS, S. & HEPP, T. 2021. Weakly supervised segmentation of tumor lesions in PET-CT hybrid imaging. *Journal of Medical Imaging,* 8**,** 054003.

GOODFELLOW, I., BENGIO, Y. & COURVILLE, A. 2016. *Deep learning*, MIT press.

GORDON, E. H. & HUBBARD, R. E. 2020. Differences in frailty in older men and women. *Medical Journal of Australia,* 212**,** 183-188.

GORGOLEWSKI, K., BURNS, C. D., MADISON, C., CLARK, D., HALCHENKO, Y. O., WASKOM, M. L. & GHOSH, S. S. 2011. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics,* 5**,** 13.

GRIFFITH, B., KADOM, N. & STRAUS, C. M. 2019. Radiology education in the 21st century: threats and opportunities. *Journal of the American College of Radiology,* 16**,** 1482-1487.

GRUNDY, S. M. 2001. Coronary plaque as a replacement for age as a risk factor in global risk assessment. *The American Journal of Cardiology,* 88**,** 8-11.

GULSHAN, V., PENG, L., CORAM, M., STUMPE, M. C., WU, D., NARAYANASWAMY, A., VENUGOPALAN, S., WIDNER, K., MADAMS, T. & CUADROS, J. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama,* 316**,** 2402-2410.

HABES, M., ERUS, G., TOLEDO, J. B., ZHANG, T., BRYAN, N., LAUNER, L. J., ROSSEEL, Y., JANOWITZ, D., DOSHI, J. & VAN DER AUWERA, S. 2016. White matter hyperintensities and imaging patterns of brain ageing in the general population. *Brain,* 139**,** 1164-1179.

HAJEK, T., FRANKE, K., KOLENIC, M., CAPKOVA, J., MATEJKA, M., PROPPER, L., UHER, R., STOPKOVA, P., NOVAK, T. & PAUS, T. 2019. Brain age in early stages of bipolar disorders or schizophrenia. *Schizophrenia bulletin,* 45**,** 190-198.

HANNUM, G., GUINNEY, J., ZHAO, L., ZHANG, L., HUGHES, G., SADDA, S., KLOTZLE, B., BIBIKOVA, M., FAN, J.-B. & GAO, Y. 2013. Genome-wide

methylation profiles reveal quantitative views of human aging rates. *Molecular cell,* 49**,** 359-367.
HARRIS, C. R., MILLMAN, K. J., VAN DER WALT, S. J., GOMMERS, R., VIRTANEN, P., COURNAPEAU, D., WIESER, E., TAYLOR, J., BERG, S. & SMITH, N. J. 2020. Array programming with NumPy. *Nature,* 585**,** 357-362.
HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H. & FRIEDMAN, J. H. 2009. *The elements of statistical learning: data mining, inference, and prediction*, Springer.
HEPP, T., BLUM, D., ARMANIOUS, K., SCHöLKOPF, B., STERN, D., YANG, B. & GATIDIS, S. 2021. Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the German National Cohort MRI study. *Computerized Medical Imaging and Graphics,* 92**,** 101967.
HOAGEY, D. A., RIECK, J. R., RODRIGUE, K. M. & KENNEDY, K. M. 2019. Joint contributions of cortical morphometry and white matter microstructure in healthy brain aging: a partial least squares correlation analysis. *Human brain mapping,* 40**,** 5315-5329.
HOERL, A. E. & KENNARD, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics,* 12**,** 55-67.
HOGAN, M. J., STAFF, R. T., BUNTING, B. P., MURRAY, A. D., AHEARN, T. S., DEARY, I. J. & WHALLEY, L. J. 2011. Cerebellar brain volume accounts for variance in cognitive performance in older adults. *Cortex,* 47**,** 441-450.
HOOGENDAM, Y. Y., VAN DER GEEST, J. N., VAN DER LIJN, F., VAN DER LUGT, A., NIESSEN, W. J., KRESTIN, G. P., HOFMAN, A., VERNOOIJ, M. W., BRETELER, M. M. B. & IKRAM, M. A. 2012. Determinants of cerebellar and cerebral volume in the general elderly population. *Neurobiology of Aging,* 33**,** 2774-2781.
HORVATH, S. 2013. DNA methylation age of human tissues and cell types. *Genome biology,* 14**,** 1-20.
HOSNY, A., PARMAR, C., QUACKENBUSH, J., SCHWARTZ, L. H. & AERTS, H. J. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer,* 18**,** 500-510.
HUANG, T.-W., CHEN, H.-T., FUJIMOTO, R., ITO, K., WU, K., SATO, K., TAKI, Y., FUKUDA, H. & AOKI, T. Age estimation from brain MRI images using deep learning. 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), 2017. IEEE, 849-852.
HUDELMAIER, M., GLASER, C., HOHE, J., ENGLMEIER, K. H., REISER, M., PUTZ, R. & ECKSTEIN, F. 2001. Age-related changes in the morphology and deformational behavior of knee joint cartilage. *Arthritis & Rheumatism,* 44**,** 2556-2561.
HUNTER, J. D. 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering,* 9**,** 90-95.
JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2013. *An introduction to statistical learning*, Springer.
JEE, H., JEON, B. H., KIM, Y. H., KIM, H. K., CHOE, J., PARK, J. & JIN, Y. 2012. Development and Application of Biological Age Prediction Models

with Physical Fitness and Physiological Components in Korean Adults. *Gerontology,* 58**,** 344-353.
JENKINSON, M., PECHAUD, M. & SMITH, S. BET2: MR-based estimation of brain, skull and scalp surfaces. Eleventh annual meeting of the organization for human brain mapping, 2005. Toronto., 167.
KILCOYNE, R. F., RICHARDSON, M. L., PORTER, B. A., OLSON, D. O., GREENLEE, T. K. & LANZER, W. 1988. Magnetic resonance imaging of soft tissue masses. *Clinical orthopaedics and related research***,** 13-19.
KLUYVER, T., RAGAN-KELLEY, B., PéREZ, F., GRANGER, B. E., BUSSONNIER, M., FREDERIC, J., KELLEY, K., HAMRICK, J. B., GROUT, J. & CORLAY, S. 2016. *Jupyter Notebooks-a publishing format for reproducible computational workflows*.
KOPPELMANS, V., HIRSIGER, S., MéRILLAT, S., JäNCKE, L. & SEIDLER, R. D. 2015. Cerebellar gray and white matter volume and their relation with age and manual motor performance in healthy older adults. *Human brain mapping,* 36**,** 2352-2363.
LAMONTAGNE, P. J., BENZINGER, T. L., MORRIS, J. C., KEEFE, S., HORNBECK, R., XIONG, C., GRANT, E., HASSENSTAB, J., MOULDER, K. & VLASSENKO, A. G. 2019. OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *MedRxiv*.
LANGNER, T., WIKSTRöM, J., BJERNER, T., AHLSTRöM, H. & KULLBERG, J. 2019. Identifying morphological indicators of aging with neural networks on large-scale whole-body MRI. *IEEE transactions on medical imaging,* 39**,** 1430-1437.
LECKY, F. E., BOUAMRA, O., WOODFORD, M., ALEXANDRESCU, R. & O'BRIEN, S. J. 2010. Epidemiology of polytrauma. *Damage control management in the polytrauma patient.* Springer.
LECUN, Y., BENGIO, Y. & HINTON, G. 2015. Deep learning. *Nature,* 521**,** 436-444.
LEPAGE, C., MUEHLMANN, M., TRIPODIS, Y., HUFSCHMIDT, J., STAMM, J., GREEN, K., WROBEL, P., SCHULTZ, V., WEIR, I. & ALOSCO, M. L. 2019. Limbic system structure volumes and associated neurocognitive functioning in former NFL players. *Brain imaging and behavior,* 13**,** 725-734.
LI, X., MORGAN, P., ASHBURNER, J., SMITH, J. & RORDEN, C. 2016. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods,* 264.
LIEM, F., VAROQUAUX, G., KYNAST, J., BEYER, F., MASOULEH, S. K., HUNTENBURG, J. M., LAMPE, L., RAHIM, M., ABRAHAM, A. & CRADDOCK, R. C. 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage,* 148**,** 179-188.
LOWSKY, D. J., OLSHANSKY, S. J., BHATTACHARYA, J. & GOLDMAN, D. P. 2014. Heterogeneity in healthy aging. *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences,* 69**,** 640-649.
LU, J., BEHBOOD, V., HAO, P., ZUO, H., XUE, S. & ZHANG, G. 2015. Transfer learning using computational intelligence: A survey. *Knowledge-Based Systems,* 80**,** 14-23.

LU, L., EHMKE, R. C., SCHWARTZ, L. H. & ZHAO, B. 2016. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PloS one,* 11**,** e0166550.

MAEDA-GUTIéRREZ, V., GALVáN-TEJADA, C. E., ZANELLA-CALZADA, L. A., CELAYA-PADILLA, J. M., GALVáN-TEJADA, J. I., GAMBOA-ROSALES, H., LUNA-GARCíA, H., MAGALLANES-QUINTANAR, R., GUERRERO MéNDEZ, C. A. & OLVERA-OLVERA, C. A. 2020. Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases. *Applied Sciences,* 10**,** 1245.

MARCUS, D. S., WANG, T. H., PARKER, J., CSERNANSKY, J. G., MORRIS, J. C. & BUCKNER, R. L. 2007. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience,* 19**,** 1498-1507.

MASON, D. 2011. SU-E-T-33: pydicom: an open source DICOM library. *Medical Physics,* 38**,** 3493.

MCCLELLAND, R. L., NASIR, K., BUDOFF, M., BLUMENTHAL, R. S. & KRONMAL, R. A. 2009. Arterial Age as a Function of Coronary Artery Calcium (from the Multi-Ethnic Study of Atherosclerosis [MESA]). *The American Journal of Cardiology,* 103**,** 59-63.

MCKINNEY, W. Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference, 2010. Austin, TX, 51-56.

MOHRI, M., ROSTAMIZADEH, A. & TALWALKAR, A. 2018. *Foundations of machine learning*, MIT press.

MONNIN, P., SFAMENI, N., GIANOLI, A. & DING, S. 2017. Optimal slice thickness for object detection with longitudinal partial volume effects in computed tomography. *Journal of applied clinical medical physics,* 18**,** 251-259.

NARICI, M. V., MAGANARIS, C. N., REEVES, N. D. & CAPODAGLIO, P. 2003. Effect of aging on human muscle architecture. *Journal of applied physiology,* 95**,** 2229-2234.

NOBIS, L., MANOHAR, S. G., SMITH, S. M., ALFARO-ALMAGRO, F., JENKINSON, M., MACKAY, C. E. & HUSAIN, M. 2019. Hippocampal volume across age: Nomograms derived from over 19,700 people in UK Biobank. *NeuroImage: Clinical,* 23**,** 101904.

OECD 2023. Computed tomography (CT) exams (indicator). . https://data.oecd.org/healthcare/computed-tomography-ct-exams.htm.

OHNISHI, T., MATSUDA, H., TABIRA, T., ASADA, T. & UNO, M. 2001. Changes in Brain Morphology in Alzheimer Disease and Normal Aging: Is Alzheimer Disease an Exaggerated Aging Process? *American Journal of Neuroradiology,* 22**,** 1680-1685.

PAN, S. J. & YANG, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering,* 22**,** 1345-1359.

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N. & ANTIGA, L. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems,* 32**,** 8026-8037.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research,* 12**,** 2825-2830.

PFEFFERBAUM, A., MATHALON, D. H., SULLIVAN, E. V., RAWLES, J. M., ZIPURSKY, R. B. & LIM, K. O. 1994. A quantitative magnetic resonance imaging study of changes in brain morphology from infancy to late adulthood. *Archives of neurology,* 51**,** 874-887.

PROSPERI, M., GUO, Y., SPERRIN, M., KOOPMAN, J. S., MIN, J. S., HE, X., RICH, S., WANG, M., BUCHAN, I. E. & BIAN, J. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence,* 2**,** 369-375.

RAVì, D., WONG, C., DELIGIANNI, F., BERTHELOT, M., ANDREU-PEREZ, J., LO, B. & YANG, G.-Z. 2016. Deep learning for health informatics. *IEEE journal of biomedical and health informatics,* 21**,** 4-21.

RAZ, N., LINDENBERGER, U., RODRIGUE, K. M., KENNEDY, K. M., HEAD, D., WILLIAMSON, A., DAHLE, C., GERSTORF, D. & ACKER, J. D. 2005. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex,* 15**,** 1676-1689.

REEVE, A., SIMCOX, E. & TURNBULL, D. 2014. Ageing and Parkinson's disease: Why is advancing age the biggest risk factor? *Ageing Research Reviews,* 14**,** 19-30.

RESNICK, S. M., PHAM, D. L., KRAUT, M. A., ZONDERMAN, A. B. & DAVATZIKOS, C. 2003. Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *Journal of Neuroscience,* 23**,** 3295-3301.

ROSE, M. R. 1991. *Evolutionary biology of aging*, Oxford University Press on Demand.

ROSENBLATT, F. 1957. *The Perceptron, a Perceiving and Recognizing Automaton Project Para*, Cornell Aeronautical Laboratory.

RUDIN, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence,* 1**,** 206-215.

RYDNING, D. R. J. G. J. 2018. The digitization of the world from edge to core. *Framingham: International Data Corporation***,** 16.

SACKETT, D. L. Evidence-based medicine. Seminars in perinatology, 1997. Elsevier, 3-5.

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. & BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision, 2017. 618-626.

SHORTEN, C. & KHOSHGOFTAAR, T. M. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data,* 6**,** 60.

SMOLA, A. J. & SCHöLKOPF, B. 2004. A tutorial on support vector regression. *Statistics and computing,* 14**,** 199-222.

SOWELL, E. R., PETERSON, B. S., THOMPSON, P. M., WELCOME, S. E., HENKENIUS, A. L. & TOGA, A. W. 2003. Mapping cortical change across the human life span. *Nature neuroscience,* 6**,** 309-315.

SULLIVAN, E. V. & PFEFFERBAUM, A. 2007. Neuroradiological characterization of normal adult ageing. *The British Journal of Radiology,* 80**,** S99-S108.
TAKI, Y., THYREAU, B., KINOMURA, S., SATO, K., GOTO, R., KAWASHIMA, R. & FUKUDA, H. 2011. Correlations among Brain Gray Matter Volumes, Age, Gender, and Hemisphere in Healthy Individuals. *PloS one,* 6**,** e22734.
TAORI, R., DAVE, A., SHANKAR, V., CARLINI, N., RECHT, B. & SCHMIDT, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems,* 33**,** 18583-18599.
TORREY, L. & SHAVLIK, J. 2010. Transfer learning. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques.* IGI global.
TULLO, S., PATEL, R., DEVENYI, G. A., SALACIAK, A., BEDFORD, S. A., FARZIN, S., WLODARSKI, N., TARDIF, C. L., GROUP, P. A. R. & BREITNER, J. C. 2019. MR-based age-related effects on the striatum, globus pallidus, and thalamus in healthy individuals across the adult lifespan. *Human brain mapping,* 40**,** 5269-5288.
VAPNIK, V. 1999. *The nature of statistical learning theory*, Springer science & business media.
VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P., WECKESSER, W. & BRIGHT, J. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods,* 17**,** 261-272.
WEISS, K., KHOSHGOFTAAR, T. M. & WANG, D. 2016. A survey of transfer learning. *Journal of Big Data,* 3**,** 9.
WHITE, M. C., HOLMAN, D. M., BOEHM, J. E., PEIPINS, L. A., GROSSMAN, M. & JANE HENLEY, S. 2014. Age and Cancer Risk: A Potentially Modifiable Relationship. *American Journal of Preventive Medicine,* 46**,** S7-S15.
YAMASHITA, R., NISHIO, M., DO, R. K. G. & TOGASHI, K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging,* 9**,** 611-629.
ZECH, J. R., BADGELEY, M. A., LIU, M., COSTA, A. B., TITANO, J. J. & OERMANN, E. K. 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine,* 15**,** e1002683.
ZELL, A. 2023. Introduction to Neural Networks [Vorlesungsfolien].
ZHOU, Z.-H. 2021. *Machine learning*, Springer Nature.

7 Erklärung zum Eigenanteil

Die Arbeit wurde in der Diagnostischen und Interventionellen Radiologie unter Betreuung von Prof. Sergios Gatidis durchgeführt.

Die Konzeption der Studie erfolgte in Zusammenarbeit mit Prof. Gatidis.

Sämtliche Versuche wurden von mir eigenständig durchgeführt.

Die statistische Auswertung erfolgte durch mich.

Ich versichere, das Manuskript selbständig verfasst zu haben und keine weiteren als die von mir angegebenen Quellen verwendet zu haben.

Tübingen, den 14.03.2023

8 Veröffentlichungen

Kerber B, Hepp T, Küstner T, Gatidis S. Deep learning-based age estimation from clinical Computed Tomography image data of the thorax and abdomen in the adult population. PLoS One. 2023 Nov 7;18(11):e0292993. doi: 10.1371/journal.pone.0292993. PMID: 37934735; PMCID: PMC10629654.

9 Danksagung

Mein besonderer Dank gilt Herrn Prof. Dr. Gatidis für die vertrauensvolle Betreuung dieser Arbeit und seine wertvolle Unterstützung. Besonders danke ich ihm dafür, dass er mich ermutigt hat, mich auf anspruchsvolle technische Fragestellungen einzulassen und ein tiefergehendes Verständnis der mathematischen und informatischen Grundlagen zu entwickeln.

Darüber hinaus danke ich meiner (mittlerweile) Frau und meiner Familie herzlich für ihre Geduld, ihre Unterstützung und ihren beständigen Rückhalt während der Entstehung dieser Arbeit.

Mein Dank gilt außerdem dem IZKF-Studienkolleg für die großzügige Förderung.