FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Optimierung von Vorverarbeitungsparametern bei der Bestimmung des biologischen Alters des Gehirns

Bachelorarbeit

Zur Erlangung des akademischen Grades

Bachelor of Science (B. Sc.)

im Studiengang Angewandte Informatik

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

Fakultät für Mathematik und Informatik

eingereicht von Robin Witte

geb. am 06.09.1992 in Bayreuth

Themenverantwortlicher: Prof. Dr. Joachim Denzler

Betreuer: Prof. Dr. Christian Gaser

Jena, 27. September 2019

Zusammenfassung

Die Altersvorhersage aus neurologischen Abbildungen mithilfe von maschinellen Lernverfahren stellt einen Biomarker für das biologische Alter des Gehirns dar und weist einen Zusammenhang mit altersbedingten kognitiven und funktionellen Einschränkungen sowie mit dem Verlauf von neurodegenerativen Erkrankungen auf. Für die Altersbestimmung werden meist T1-gewichtete MRT-Aufnahmen verwendet, wobei die Vorverarbeitung ein Resampling auf eine bestimmte Voxelgröße und eine Glättungsfilterung beinhaltet. Die Parameter dieser Vorverarbeitungsschritte werden dabei meist willkürlich gewählt. Da auch führende Algorithmen zur Altersbestimmung einen gewissen Fehler enthalten, wurde in dieser Arbeit der Einfluss der Vorverarbeitungsparameter auf die Genauigkeit der Altersbestimmung und die Möglichkeit der Optimierung dieser Parameter untersucht. Dafür wurden drei Strategien zur Optimierung angewandt und miteinander verglichen: Eine Rastersuche, eine Bayessche Optimierung und eine Ensemblemethode. Als Vorhersagemodell für das Alter diente dabei jeweils eine *relevance vector regression*, die mithilfe einer vierfachen Kreuzvalidierung trainiert und getestet wurde. Der dafür verwendete Datensatz umfasste 547 T1-gewichtete MRT-Aufnahmen von gesunden Proband*innen im Alter von 19 bis 86 Jahren. Die Ergebnisse zeigen, dass die Wahl der Vorverarbeitungsparameter einen großen Einfluss auf die Genauigkeit der Altersbestimmung hat und sich die Bayessche Optimierung aufgrund der Komplexität des Parameterraumes am besten zur Optimierung dieser Vorverarbeitungsparameter eignet. Darüber hinaus wird eine Bestätigung der Einsetzbarkeit von Ensemblemethoden für die Optimierung dieser Vorverarbeitungsparameter deutlich, wobei eine abschließende Aussage über die Robustheit in diesem Fall nicht getroffen werden kann.

Inhaltsverzeichnis

Abkürzungsverzeichnis

ARD	automatic relevance determination
CSF	Cerebrospinalflüssigkeit
EI	expected improvement
FWHM	full width at half maximum
GM	graue Substanz
MAE	mean absolut error
MRF	Markov random field
MRT	Magnetresonanztomographie
MSE	mean squared error
PCA	principal component analysis
PVE	partial volume estimation
RMSE	root mean squared error
RVM	relevance vector machine
RVR	relevance vector regression
SVM	support vector machine
SVR	support vector regression
WM	weiße Substanz

1. Einleitung

Im Zuge der fortschreitenden Alterung von Menschen nehmen Belastungen durch altersbedingte kognitive und funktionelle Einschränkungen sowie durch altersbedingte Krankheiten zu [34]. Da der Alterungsprozess jedoch biologisch sehr komplex ist, gibt es große individuelle Unterschiede bezüglich der Zeitpunkte zu denen sich Alterungseffekte manifestieren oder altersbedingte Krankheiten auftreten [7]. Um dennoch Methoden entwickeln zu können, die möglichst genau die individuellen Risiken für altersbedingte Einschränkungen oder individuelle Verläufe von Krankheiten beschreiben, wird daher oft das biologische Alter einer Person betrachtet. Als biologisches Alter wird das hypothetische Alter eines Organismus bezeichnet, das durch die Messung von biologischen Aspekten des Organismus bestimmt wird [7]. Dabei kann das biologische Alter einer Person von dem chronologischen Alter abweichen und ein besserer Indikator für individuelle Risiken für altersbedingte Veränderungen sein. Zur Bestimmung dieses biologischen Alters können verschiedene biologische Messungen verwendet werden, die mit unterschiedlichen Biomarkern das biologische Alter eines Organs, eines Gewebes oder spezieller Zellen bestimmen [7]. Ein Organ, das starken altersbedingten Veränderungen unterworfen ist, ist das Gehirn [24]. Aus den physiologischen Alterungsprozessen des Gehirns resultieren viele der altersbedingten Belastungen, wie kognitive Einschränkungen oder erhöhte Risiken für neurodegenerative Erkrankungen [7]. Da die Alterungsprozesse auch morphologische Änderungen im Gehirn hervorrufen, ist es möglich, das Alter einer Person mithilfe von Aufnahmen einer Magnetresonanztomographie (MRT) und einem maschinellen Lernverfahren zu schätzen. Wird dieses Lernverfahren mit Daten von gesunden Personen angelernt, können damit im Anschluss Altersvorhersagen für andere Personen getroffen werden. Dabei weist das Ausmaß einer Überschätzung gegenüber dem chronologischen Alter, einen Zusammenhang mit dem Maß der Verstärkung von altersbedingten kognitiven und funktionellen Einschränkungen auf [6]. Ist das aus den MRT-Aufnahmen geschätzte Alter also höher als das chronologische Alter, kann das auf ein erhöhtes Risiko für neurodegenerative Erkrankungen und auf ein beschleunigtes Auftreten von altersbedingten Einschränkungen hindeuten [7]. Die Bestimmung des Alters aus MRT-Daten stellt daher einen Biomarker für

das biologische Alter des Gehirns dar und ist eine Möglichkeit frühzeitig individuelle Risiken für altersbedingte Veränderungen zu bemessen. Darüber hinaus liefert diese Herangehensweise auch Informationen über das Zusammenspiel von Erkrankungen des Gehirns mit Alterungsprozessen [7].

Es gibt verschiedene Methoden zur Bestimmung des biologischen Alters des Gehirns aus MRT-Daten [7]. Trotz vielversprechender bisheriger Ergebnisse, enthalten die Methoden jedoch weiterhin Messfehler, die unter Nutzung einer großen Anzahl an Lerndaten zwischen 4 und 5 Jahren liegen [20]. Diese Fehlerangabe bezieht sich auf den *mean absolut error* (MAE), einzelne Personen können dabei jedoch Fehler bis zu 25 Jahren aufweisen [20]. Auch wenn ein Teil dieser Variationen vermutlich die zugrundeliegende Variabilität der Bevölkerung darstellt, verbleibt eine Ungenauigkeit der Methoden, deren Verringerung für die klinische Anwendung unerlässlich ist. Das bedeutet also, dass obwohl vermutlich keine Methode mit einem MAE von 0 entwickelt werden kann, die Grenze der Genauigkeit noch nicht erreicht ist [20]. Für die Genauigkeit und die Fähigkeit der Verallgemeinerung auf ungesehene Daten ist in diesem Fall von entscheidender Bedeutung, mit welcher Herangehensweise die Merkmale aus den MRT-Aufnahmen extrahiert werden und welches Lernverfahren verwendet wird. Die Mehrheit der Methoden nutzt T1-gewichtete MRT-Aufnahmen [7] und arbeitet entweder mit voxelweisen Karten des Hirnvolumens oder mit Messungen der kortikalen und subkortikalen Dicke des Gehirns [20]. Für alle Verfahren ist dabei eine komplexe Vorverarbeitung der MRT-Aufnahmen nötig. Die verwendeten Parameter bei dieser Vorverarbeitung sind in der Regel vordefinierte Standardwerte der Softwareentwickler oder basieren auf vorher durchgeführten Studien. Aufgrund der Komplexität des Problems der Bestimmung des biologischen Alters des Gehirns, ist es jedoch nicht sinnvoll allgemeine Parameterwerte zu verwenden, sondern diese für ein optimales Ergebnis von Fall zu Fall speziell festzulegen. Außerdem kann die Wahl der Vorverarbeitungsparameter einen starken Einfluss auf das Ergebnis haben und beeinflusst daher die Genauigkeit der Methode [20].

In dieser Arbeit sollen drei verschiedene Strategien zur Optimierung der Vorverarbeitungsparameter bei der Bestimmung des biologischen Alters des Gehirns erarbeitet und miteinander verglichen werden. Als Strategien werden dafür eine Rastersuche, eine Bayessche Optimierung sowie eine Ensemblemethode verwendet. Während es sich bei der Rastersuche und der Bayesschen Optimierung um klassische Optimierungsmethoden handelt, bestimmt die Ensemblemethode kein Parameteroptimum, sondern ermittelt eine möglichst zielführende Kombination aus verschiedenen Parameterwerten. Durch die Verwendung dieser drei sehr unterschiedlichen Strategien soll ein umfassendes Verständnis über den Einfluss der Vorverarbeitungsparame-

ter auf die Bestimmung des biologischen Alters des Gehirns generiert werden. Der Fokus liegt dabei nicht auf der Erstellung einer Empfehlung von gewissen Parameterwerten, sondern auf der Bewertung der Optimierungsmethoden hinsichtlich der Verwendbarkeit in der Praxis und der jeweils erzielten Genauigkeit.

Da in dieser Arbeit sowohl Aspekte aus dem Bereich der Medizin als auch aus den Bereichen Mathematik und Informatik eine Rolle spielen, ist für die Verständlichkeit der Methoden eine umfassende Erläuterung der Grundlagen entscheidend. Daher werden diese Grundlagen im Folgenden umfassend dargestellt, wodurch es in den restlichen Abschnitten keiner zusätzlichen erläuternden Beschreibungen bedarf.

2. Theoretische Grundlagen

2.1. Alterungsprozesse im Gehirn

Das Gehirn eines Menschen liegt geschützt in der Schädelhöhle, wird von Hirnhäuten umhüllt und besteht hauptsächlich aus Nervengewebe. Es ist Teil des zentralen Nervensystems und wird anatomisch in verschiedene Strukturen eingeteilt. Neben dem Großhirn, das in zwei Hemisphären unterteilt ist, unterscheidet man noch das Zwischenhirn, das Kleinhirn und den Hirnstamm. Strukturell lässt sich das Gehirn in graue und weiße Substanz einteilen. Die graue Substanz (GM), die vor allem aus Nervenzellkörpern aber auch aus zahlreichen Verbindungen der Nervenzellen untereinander besteht, befindet sich unter anderem in der 2-4 mm dicken Oberflächenschicht des Großhirns, die Großhirnrinde genannt wird. Die weiße Substanz (WM), die im wesentlichen aus Nervenfaserbündeln besteht, liegt darunter und verbindet weiter voneinander entfernte Neurone miteinander. Die Nervenfasern der WM umgibt eine Myelinschicht, die eine beschleunigte Erregungsweiterleitung ermöglicht und das Gewebe weiß erscheinen lässt. Zusätzlich zu den beiden genannten neurologischen Strukturen gibt es noch die Cerebrospinalflüssigkeit (CSF), die sich sowohl in einer Schicht um das Gehirn, als auch in vier Hohlräumen innerhalb des Gehirns (Ventrikel) befindet. Auf der Großhirnrinde lassen sich verschiedene Rindenfelder lokalisieren, es ist also möglich einzelne Funktionen bestimmten Bereichen zuzuordnen. Dabei ermöglicht jedoch erst das konkrete Zusammenspiel verschiedener Felder eine Funktion. Das wird vor allem dadurch deutlich, dass es zusätzlich zu primären Feldern, die beispielsweise bestimmte Wahrnehmungen verarbeiten, auch Assoziationsfelder gibt, die für die Abstimmung verschiedener Funktionen verantwortlich sind. [28]

Die Veränderung des Gehirns in Abhängigkeit des Alters des Menschen läuft nach einem spezifischen Muster ab. Die sichtbaren Prozesse beruhen dabei auf unterschiedlichen neurologischen Prozessen, die in Ihrem Zusammenspiel bestimmte Veränderungen des Gehirns hervorrufen. In den ersten Lebensjahren entwickelt sich das Gehirn durch die Bildung neuer Neuronen und neuer Verbindungen. Ab einem Alter von 20 ist die Entwicklung jedoch weitgehend abgeschlossen und die Bildung neuer

Neuronen ist nur noch sehr begrenzt möglich. Nur in wenigen Regionen sind dann noch undifferenzierte neurale Vorläuferzellen vorhanden, die sich weiterhin teilen können und fähig sind, Neuroblasten und junge Neuronen zu bilden [28]. Die Regenerationsfähigkeit des Nervengewebes ist in adulten Gehirnen also sehr beschränkt, jedoch sind die vorhandenen Neuronen weiterhin in der Lage neue Verbindungen zu knüpfen und ermöglichen auf diese Art Leistungssteigerungen. Diesen Leistungssteigerungen stehen Alterungsprozesse gegenüber, die mit zunehmendem Alter eine größere Rolle spielen und eine Abnahme der Leistungsfähigkeit bewirken können. Im Allgemeinen lässt sich sagen, dass eine Alterung des Gehirns die Moleküle, die Zellen, das Gefäßsystem und auch die gesamte Morphologie beeinflusst [24]. Die Effekte einer Alterung sind also sowohl mikroskopisch, als auch makroskopisch in jeder Betrachtungsebene von Bedeutung. Die auffälligsten strukturellen Veränderungen sind dabei die Volumenänderungen von GM,WM und CSF. Grundlegend unterliegt die GM einer linearen Volumenabnahme und die CSF einer linearen Volumenzunahme, wohingegen die WM im Allgemeinen ein eher gleichbleibendes Volumen in Abhängigkeit zum Alter aufweist [14, 25]. Die Volumenabnahme der GM, die vor allem durch das Absterben von Neuronen bedingt ist, ist dabei der dominierende Prozess. Daher unterliegt das Gehirn im Gesamten einer Atrophie, das Volumen des Gehirns nimmt also insgesamt betrachtet mit zunehmendem Alter ab. Im Detail betrachtet, ist die Stärke der Volumenabnahme der GM jedoch nicht gleichmäßig verteilt sondern regionenspezifisch [14]. Die Volumenänderungen ergeben also ein heterogenes Muster. Dabei legen Untersuchungen nahe, dass die phylogenetisch jüngeren Bereiche des Gehirns stärker von einer Atrophie betroffen sind als phylogenetisch ältere [30, 12]. Auch wenn das Volumen der WM im Allgemeinen konstant ist, ist die WM dennoch von spezifischen Alterungsprozessen betroffen. Hierbei spielen vor allem altersbedingte Gefäßveränderungen eine Rolle. Besonders die kleinsten Blutgefäße sind Veränderungen unterworfen, welche dazu führen, dass der aktivitätsbezogene hohe Bedarf der Nervenzellen an Energie und Nährstoffen in zunehmendem Maße nur noch unzureichend gedeckt werden kann. Weiterhin erhöht sich durch die Gefäßveränderungen, in Verbindung mit im fortschreitenden Alter häufiger auftretenden Bluthochdruck, die Wahrscheinlichkeit, dass sich Risse in kleinen Blutgefäßen bilden können, die zu Mikro-Einblutungen führen [24]. Diese Veränderungen führen in Summe zu Demyelinisierung und dem Verlust von Nervenzellfasern und Nervenzellen, wodurch sich die Fähigkeit der weißen Substanz verschlechtert, die integrierte Informationsverarbeitung verteilter Strukturen durch Informationsübertragung und Synchronisation zu gewährleisten.

Die beschriebenen physiologischen Änderungen haben großen Einfluss auf die ko-

gnitiven Fähigkeiten des alternden Menschen. So ist mit zunehmendem Alter in vielen Bereichen ein Abbau der Leistungsfähigkeit zu beobachten. Besonders betroffen sind dabei das Gedächtnis, logisches Schlussfolgern, räumliche Orientierung sowie numerische Fähigkeiten, deren Abbau weitgehend synchron abläuft. Der Wortschatz stellt im Gegensatz dazu eine Ausnahme dar, da dieser auch bei fortschreitendem Alter konstant bleibt oder sich sogar vergrößern kann. Zurückzuführen sind die Leistungsabnahmen der verschiedensten Bereiche im Wesentlichen auf Verschlechterungen in drei weitgehend unabhängigen kognitiven Prozessen. Diese sind das Arbeitsgedächtnis, die Informationsverarbeitungsgeschwindigkeit sowie die sensorischen Funktionen. [23]

Es gibt verschiedene Krankheiten, die Veränderungen im Gehirn hervorrufen und damit die Struktur des Gehirns beeinflussen. Einige dieser Krankheiten bewirken ähnliche physiologische Veränderungen, wie sie bei der Alterung des Gehirns entstehen und beschleunigen somit diese Prozesse. Ein Beispiel dafür, das durch die hohe Zahl der Betroffenen eine große Relevanz aufweist, ist die Alzheimer-Krankheit, die eine neurodegenerative Erkrankung ist. Physiologisch ist diese Krankheit durch vermehrtes Absterben von Neuronen im Gehirn gekennzeichnet, das neben weiteren Effekten vor allem zu einer zunehmenden Demenz führt. Die Alzheimer-Krankheit ist somit als ein beschleunigtes Altern des Gehirns beschreibbar [16]. Ein weiteres Beispiel für eine solche Erkrankung ist Schizophrenie, da auch hier Bestimmungen des biologischen Alters des Gehirns nahe legen, dass Schizophrenie ein beschleunigtes Altern des Gehirns bewirkt [22].

2.2. Magnetresonanztomographie

Die MRT ist ein bildgebendes Verfahren, das vor allem in der medizinischen Diagnostik zur Darstellung von Struktur und Funktion der Gewebe und Organe im Körper eingesetzt wird. Mit diesem Verfahren ist es möglich hochauflösende Schnittbilder des menschlichen Körpers zu erzeugen. Physikalisch macht sich die MRT die Prinzipien der Kernspinresonanz zu nutze, im Fall der medizinischen Diagnostik im besonderen die Kernspinresonanz der Wasserstoff-Atomkerne die sich im menschlichen Körper befinden. Diese Atomkerne besitzen einen Kernspin und dadurch ein kleines magentisches Dipolmoment, sie weisen also eine Magnetisierung auf. Im Normalfall unterliegen die Atomkerne einer zufälligen Orientierung, wodurch sich die Momente gegenseitig ausgleichen und menschliches Gewebe nicht magnetisch ist. Wird jedoch ein starkes und homogenes Magnetfeld erzeugt, in das der Körper eingebracht wird, richten sich die Dipolmomente der Atomkerne parallel zu diesem aus. Wird zusätzlich

ein senkrecht stehendes hochfrequentes Wechselfeld induziert, werden bei einer spezifischen Frequenz die nun parallel stehenden Drehachsen der Kerne in eine Präzessionsbewegung versetzt, die anschaulich mit der Rotation eines Spielzeugkreisel mit nicht senkrecht stehender Drehachse vergleichbar ist. Die Frequenz der Präzessionsbewegung (Lamorfrequenz) und damit die erforderliche Frequenz des Wechselfeldes ist dabei für Atomkerne verschiedener Stoffe unterschiedlich, wodurch es beispielsweise möglich ist, allein die Wasserstoffatome in eine Präzessionsbewegung zu versetzen. Durch die Auslenkung der Drehachsen der Atomkerne werden natürlich auch die Dipolmomente ausgelenkt und es entsteht eine wechselnde Magnetisierung senkrecht zu dem homogenen Magnetfeld. Nach der Abschaltung des Hochfrequenzfeldes, kann dann dieses durch die rotierenden Wasserstoffatome erzeugte Wechselfeld gemessen werden. Optimal für die Signalstärke ist es, wenn am Ende des Anregungspulses die Magnetisierung im Gesamten senkrecht zu dem starken homogenen Magnetfeld steht. Man spricht in diesem Fall auch von einem 90°-Puls. Da mit der Abschaltung des Hochfrequenzfeldes keine Anregung mehr durchgeführt wird, nimmt die senkrecht stehende Magnetisierung ab Beginn der Messung über eine definierte Zeit wieder ab und kehrt in ihre Ursprungsrichtung zurück. Dieser im Allgemeinen zeitlich exponentiell verlaufende Prozess heißt Längsrelaxation oder auch Spin-Gitter-Relaxation. Die zugehörige Zeitkonstante wird als T1 bezeichnet. Ein gleichzeitig ablaufender Prozess ist die Querrelaxation oder auch Spin-Spin-Relaxation. Dieser Begriff beschreibt die Dephasierung, die durch leicht unterschiedliche Frequenzen der Präzessionsbewegung zustande kommt. Die auch hierdurch entstehende exponentielle Abnahme der Magnetisierung in senkrechter Ebene wird durch die Konstante T2 beschrieben. Dabei gilt $T2 \leq T1$. Diese Relaxationskonstanten sind von der chemischen Verbindung und der molekularen Umgebung abhängig, in der sich der präzedierende Wasserstoffkern befindet. Daher unterscheiden sich die verschiedenen Gewebearten charakteristisch in ihrem Signal, was zu verschiedenen Signalstärken und damit zu unterschiedlichen Helligkeiten im resultierenden Bild führt. [18, 17]

Um die Signale den einzelnen Volumenelementen (Voxeln) zuordnen zu können, wird mit linear ortsabhängigen Magnetfeldern (Gradientenfeldern) eine Ortskodierung erzeugt [18]. Dabei wird ausgenutzt, dass für den Atomkern eines bestimmten Stoffes die Larmorfrequenz von der magnetischen Flussdichte abhängt. Liegt bei der Anregung also zusätzlich zu dem homogenen Magnetfeld ein Gradientenfeld vor, ist es möglich nur eine spezifische Schicht anzuregen. Hierbei ist zu beachten, dass eine gleichzeitige Induktion von drei Gradientenfeldern für die drei Raumrichtungen nicht zu einer dreidimensionalen Ortsabhängigkeit führt. Durch eine Überlagerung würde in diesem Fall ein gemeinsames Gradientenfeld erzeugt werden, welches wiederum

nur die Abhängigkeit in einer Raumrichtung ermöglicht. Neben dem Gradientenfeld, das während der Anregung induziert wird und eine Schichtselektion in einer Raumrichtung ermöglicht, werden die zwei Gradientenfelder die für die verbleibenden Raumrichtungen nötig sind daher zu anderen Zeitpunkten induziert. Das erste wird nach der Anregung für kurze Zeit eingeschaltet und erzeugt so eine ortsabhängige Dephasierung. Das zweite wird bei der Messung induziert und bewirkt dadurch ortsabhängige Präzessionfrequenzen. Durch die senkrechte Ausrichtung aller drei Gradientenfelder ist im Gesamten dann eine dreidimensionale Erfassung des Signals möglich [18].

Eine Datenaufnahme mit einem MRT-Gerät besteht nicht aus einer einmaligen Anregung mit anschließender Messung, sondern aus einer komplexen Sequenz an Anregungen und Messungen. Verwendete Sequenzen können sich dabei nicht nur in der Art des Anregungspulses, sondern auch in den Zeitabschnitten zwischen Anregungen und Messungen unterscheiden. Unterschiedliche Sequenzen liefern dabei vollkommen verschiedene Ergebnisse und sind so spezialisiert für bestimmte Fragestellungen [18]. Zwei grundlegende Konzepte für Sequenzen sind dabei T2- und T1-Messungen [17]. Bei T2-Messungen werden lange Zeitabstände verwendet, wodurch das Signal vor allem von Substanzen erzeugt wird, die lange Relaxationszeiten aufweisen. Hierbei ist vor allem Wasser die entscheidende Substanz und Gewebe mit geringem Wasseranteil liefern nur schwache Signale. Der kontrastgebende Parameter ist bei dieser Messart die T2-Konstante, wodurch der Name der T2-Messung zustande kommt. Bei einer T1-Messung werden hingegen vor allem kurze Zeitabstände verwendet. Hierdurch sind bei der erneuten Anregung Substanzen mit langer Relaxationszeit noch nicht wieder in der Ausgangslage und werden so nicht erneut angeregt. Gewebe mit hohem Wasseranteil wird hierbei also eher unterdrückt und der kontrastgebende Parameter ist die T1-Konstante. Auch wenn es viele unterschiedliche Sequenzen gibt, ist im Allgemeinen das Ergebnis einer MRT eine bestimmte Anzahl an Volumenelementen, die jeweils einem bestimmten Ort zugeordnet werden können und jeweils eine eigene Signalstärke aufweisen. Wird nur eine MRT-Aufnahme gemacht, spricht man von strukturellen MRT-Daten. Werden mehrere Aufnahmen hintereinander gemacht, um zeitliche Prozesse sichtbar zu machen, spricht man von funktionellen MRT-Daten. Je nach Wahl der Sequenz haben die Voxel eine unterschiedliche Auflösung und das Signal unterschiedliche Qualität. Über die Sequenz hinaus hängt das Ergebnis jedoch auch von dem verwendeten MRT-Gerät ab. Die größten Unterschiede resultieren hierbei aus den unterschiedlich starken homogenen Magnetfeldern die dabei verwendet werden, sowie der Homogenität dieser Felder. Für MRT-Geräte liegen typische magnetische Flussdichten zwischen 1,5 und 3 Tesla, jedoch gibt es auch

Geräte, die mit Flussdichten von 7 Tesla oder mehr arbeiten [18, 35, 43]. Hierbei gilt zwar im Allgemeinen, dass eine höhere Flussdichte bessere Ergebnisse liefert, jedoch spielt dabei die Homogenität eine entscheidende Rolle. Nur bei einem möglichst homogenen Magnetfeld können gute Ergebnisse erzielt werden, die Homogenität ist bei höheren Flussdichten jedoch deutlich schwerer zu erreichen [18].

2.3. Vorverarbeitung von MRT-Daten

Fragestellungen, die durch strukturelle MRT-Daten beantwortet werden können, benötigen immer eine morphologische Betrachtung. Das Ziel ist also immer die Charakterisierung von Volumengrößen verschiedener Gewebearten, die Betrachtung der Form von spezifischen Strukturen oder die Kennzeichnung von Konzentrationsunterschieden verschiedener Gewebearten in bestimmten Bereichen. Neben der regionenorientierten und der oberflächenbasierten Morphometrie, ist die voxelbasierte Morphometrie eine viel verwendete Herangehensweise um strukturelle MRT-Aufnahmen von Gehirnen zu analysieren. Im Gegensatz zu den vorher genannten Ansätzen, bei denen Volumengrößen von Regionen oder die Ausformungen der Oberfläche miteinander verglichen werden, werden bei der voxelbasierten Morphometrie alle Voxel einzeln in Gewebearten klassifiziert und dann auf der Voxelebene Vergleiche zwischen verschiedenen Gehirnen durchgeführt. Diese Art der Morphometrie kann für die Bestimmung des biologischen Alters des Gehirns verwendet werden, da dabei Konzentrationsunterschiede von Gewebearten miteinander verglichen werden können. Um einzelne Voxel von verschiedenen Proband*innen mit statistischen Verfahren vergleichen zu können, ist jedoch eine komplexe Vorverarbeitung der Daten wichtig. Neben der Segmentierung, also der Einteilung der Voxel in die Gewebearten, ist dabei vor allem eine Normalisierung von entscheidender Bedeutung. Mit Normalisierung ist in diesem Fall gemeint, dass die unterschiedlich geformten Gehirne in die Form einer Vorlage gebracht werden. Erst dann ist es möglich einzelne Voxel verschiedener Proband*innen miteinander zu vergleichen, da erst dann sichergestellt ist, dass entsprechende Voxel auch das gleiche Hirnareal darstellen.

Die Vorverarbeitung ist kein zu entwickelnder Teil dieser Arbeit, sondern orientiert sich vollständig an dem von Franke et al. [10] vorgestellten Vorgehen. Dieses basiert auf einer Methode von Ashburner und Friston [3], die *unified segmentation* heißt und einer von Gaser [11] vorgestellten Erweiterung dieser Methode. In einem letzten Schritt wird dann noch eine Filterung und ein Resampling durchgeführt. Um die Bestimmung des biologischen Alters des Gehirns und den Einfluss der zu optimierenden Vorverarbeitungsparameter nachvollziehen zu können, ist ein grund-

legendes Verständnis der Vorverarbeitung wichtig. Deswegen werden im Folgenden die Funktionsweisen der erforderlichen Vorverarbeitungsschritte kurz erläutert.

2.3.1. Unified segmentation

Die Segmentierung und die Normalisierung können als zwei separate Vorverarbeitungsschritte durchgeführt werden, jedoch beeinflussen sich die beiden Schritte gegenseitig. Die von Ashburner und Friston [3] vorgestellte *unified segmentation* fasst beide Schritte in einem gemeinsamen probabilistischen Modell zusammen und ermöglicht so eine gemeinsame Optimierung von Parametern. Zusätzlich zu der Segmentierung und der Normalisierung werden dabei gleichzeitig auch noch Störungen im homogenen Magnetfeld des MRT-Gerätes ausgeglichen. Diese Korrektur von Inhomogenitäten wird *bias-field correction* genannt und kann durch verschiedene Methoden auch unabhängig von der *unified segmentation* durchgeführt werden. Eine Beseitigung der Inhomogenitäten ist wichtig, da sie zu Intensitätsänderungen von Voxeln führen und damit dazu, dass gleiches Gewebe über das Bild verteilt unterschiedliche Intensitäten aufweisen kann. Das führt vor allem bei der Segmentierung zu Problemen, wodurch deutlich wird, dass eine Modellierung in einem gemeinsamen Modell sinnvoll ist.

Die Grundlage für das probabilistische Modell ist eine Mischverteilung aus Normalverteilungen, ein sogenanntes *Gaussian mixture model*. Dabei werden die Helligkeitsverteilungen der Gewebearten durch Normalverteilungen modelliert. Die zusammengesetzte Verteilung ergibt dann die Helligkeitsverteilung des Bildes und durch die Intensität die ein Voxel aufweist, kann dann bestimmt werden wie hoch die Wahrscheinlichkeiten für den Voxel sind, dass er zu den verschiedenen Gewebearten gehört. Hierbei ist es sehr einfach möglich auch den Hintergrund durch eine Normalverteilung zu modellieren und so automatisch auch die Voxel, die nicht das Gehirn abbilden, zu klassifizieren. Da eine Segmentierung allein auf der Grundlage der Helligkeitsinformation des Voxels sehr rauschanfällig ist, werden zusätzlich noch Wahrscheinlichkeitskarten verwendet. Diese Wahrscheinlichkeitskarten liegen in der selben Auflösung wie die MRT-Aufnahmen vor und enthalten für jeden Voxel die Wahrscheinlichkeiten, dass dieser Voxel zu GM, WM, oder CSF gehört. Sie entsprechen einer vordefinierten Form und werden im Vorfeld bestimmt. Kombiniert werden die Wahrscheinlichkeiten aus der Helligkeitsinformation und die Wahrscheinlichkeiten aus den Karten dann über einen Bayes-Schätzer. In diesem Fall stellen die Wahrscheinlichkeitskarten also eine a priori Information dar, mit denen die a posteriori Wahrscheinlichkeit eines Voxels zu einem bestimmten Gewebe zu gehören geschätzt werden kann. Die Steuerung von Deformationen des Bildes für eine Nor-

malisierung und die Korrektur der Inhomogenitäten wird in der Zielfunktion durch weitere Parameter und Terme erreicht. Durch eine Optimierung der Zielfunkion kann dann eine möglichst genaue Lösung erzielt werden. Die Zielfunktion, partielle Ableitungen und Anmerkungen zur Optimierung, sowie eine Evaluation der Methode ist in [3] zu finden.

2.3.2. Partial volume estimation

Die beschriebene *unified segmentation* betrachtet einen Voxel entweder als GM, WM, CSF oder Hintergrund. Hierbei wird jedoch nicht beachtet, dass ein Voxel auch mehrere Gewebetypen gleichzeitig enthalten kann, was bedingt durch die begrenzte Auflösung der Voxel in der Realität häufig zutrifft. Um eine genauere Segmentierung zu erhalten, ist es daher von Vorteil diese sogenannten *partial volume effects* zu berücksichtigen. Gaser [11] beschreibt hierfür ein Verfahren, bei dem die Lösung der *unified segmentation* als Startwert dient und in einem zweiten Schritt zusätzlich eine *partial volume estimation* (PVE) durchgeführt wird, die auf einer Arbeit von Tohka, Zijdenbos und Evans [32] basiert. Bei diesem sogenannten AMAP-Algorithmus werden für jeden Voxel die Mischklassen GM-WM, GM-CSF und Hintergrund-CSF geschätzt und aus der Gesamtinformation Schätzungen für die Menge der Gewebetypen in einem Voxel abgeleitet. Da durch die Berücksichtigung von *partial volume effects* die Methode eine größere Rauschanfälligkeit aufweist, wird in dem Algorithmus die PVE mit einem *Markov random field* (MRF) Modell kombiniert. Durch das MRF Modell wird bei der Klassifizierung eines Voxels zusätzlich die Beziehung zu den Nachbarvoxeln betrachtet und so die Anzahl der durch Rauschen verursachten Fehlklassifikationen verringert. Die Idee ist hierbei, dass wenn viele der 26 umliegenden Nachbarvoxel eines Voxels eine ähnliche Gewebeart aufweisen, die Wahrscheinlichkeit groß ist, dass auch der Voxel selbst diese Gewebeart enthält und umgekehrt. Im Gesamten ergibt sich durch den AMAP-Algorithmus für jeden Voxel eine Angabe, wie viel Prozent des Voxelvolumens einem bestimmten Gewebe entspricht, wobei das Modell immer nur von der Mischung zweier Klassen in einem Voxel ausgeht. Da das Verfahren jedoch ohne Vorinformationen wie beispielsweise Wahrscheinlichkeitskarten auskommt, findet die Klassifikation hierbei im naiven Raum statt, die Ergebnisse sind also nicht normalisiert. Daher wird im Anschluss noch eine affine Registrierung durchgeführt, die die Ergebnisse in die Form einer Vorlage bringt und so einen Vergleich von Probanden ermöglicht.

2.3.3. Filterung und Resampling

Um mit den MRT-Daten effizient arbeiten zu können, findet als letzter Vorverarbeitungsschritt noch eine Filterung und ein Resampling statt. Durch die Filterung werden Rauscheinflüsse, die auf der Voxelebene entscheidend sind, verringert. Hierfür wird ein gaußscher Glättungsfilter verwendet, dessen Größe als *full width at half maximum* (FWHM) angegeben wird [10]. Durch das anschließende Resampling wird dann noch eine Datenreduktion erreicht. Diese ist bei der hohen Anzahl der Voxel die eine einzelne MRT-Aufnahme enthält für alle voxelbasierten Verfahren wichtig.

2.4. Bestimmung des biologischen Alters des Gehirns

Um das biologische Alter des Gehirns aus einer MRT-Aufnahme zu bestimmen, muss eine Abbildung erzeugt werden, die aus der Menge an Voxelwerten das Alter vorhersagt. Es handelt sich hierbei also um ein typisches Regressionsproblem. Bei einem Regressionsproblem ist immer ein Eingaberaum X gegeben, mit dem Ziel eine Funktion $\hat{f} : X \to \mathbb{R}$ zu erzeugen, die eine möglichst gute Approximation einer unbekannten wahren Funktion f darstellt. Dabei wird die Funktion $\hat{f}$ durch einen Lernalgorithmus bestimmt, dem ein Lerndatensatz zur Verfügung steht. Dieser Lerndatensatz ist eine endliche Menge von n Beispielen der Form $\{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$. $\hat{f}$ wird dann Modell oder auch Hypothese genannt. Die Güte von $\hat{f}$ kann dabei durch verschiedene Fehlermaße bestimmt werden. Hierbei wird der unbekannte wahre Fehler durch einen weiteren Datensatz mit n_{test} Beispielen approximiert. Häufig verwendete Fehlermaße sind der MAE, der *mean squared error* (MSE) und der *root mean squared error* (RMSE). Sie sind definiert durch

$$\text{MAE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\hat{f}(x_i) - f(x_i) \right] , \tag{2.1}$$

$$\text{MSE} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\hat{f}(x_i) - f(x_i) \right]^2 , \tag{2.2}$$

$$\text{RMSE} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \left[\hat{f}(x_i) - f(x_i) \right]^2} . \tag{2.3}$$

Ein Regressions-Problem ist durch verschiedenste Verfahren aus dem Bereich des überwachten Lernens zu lösen. Daher gibt es auch für die Bestimmung des biologischen Alters des Gehirns verschiedene Ansätze, wie zum Beispiel *deep learning*-Verfahren mit der Nutzung von künstlichen neuronalen Netzen, oder maschinelle Lernverfahren wie *support vector regression* (SVR) oder *Gaussian process regression* [8, 7]. Um die Leistungsfähigkeit eines Verfahrens zur Bildregistrierung zu zeigen,

nutzte Ashburner [1] eine *relevance vector regression* (RVR) um das biologische Alter des Gehirns zu bestimmen und erreichte damit einen RMSE von 6.5 Jahren. Basierend auf dieser Idee entwickelten Franke et al. [10] ein allgemeines Verfahren zur Bestimmung des biologischen Alters des Gehirns, welches eine RVR verwendet und die Grundlage für die Altersbestimmung in dieser Arbeit darstellt. Die Vorteile dieses Verfahrens sind neben einer guten Genauigkeit auch die einfache Umsetzung und die Robustheit [10]. Die RVR ist eine Generalisierung der von Tipping [31] vorgestellten *relevance vector machine* (RVM) auf Regressionsprobleme, wobei die RVM wiederum eine Abwandlung der *support vector machine* (SVM) darstellt. Die Grundidee hinter sowohl der SVM, als auch der RVM ist, die Trainingsdaten aus dem Eingaberaum durch eine Funktion in einen höher-dimensionalen Raum abzubilden und in diesem Raum dann die bestmögliche trennende Hyperebene zu finden. Diese Hyperebene stellt dann im Eingaberaum eine nichtlineare Trennebene dar. Die beste Trennebene wird dabei gefunden indem der Abstand (*margin*) zwischen den beiden Gruppen erhöht wird. Im Falle der SVM werden zur Optimierung lediglich die *support vectors* verwendet, also nur die Datenpunkte, die in einem definierten Bereich um die Trennebene liegen. Bei der RVM hingegen werden zur Optimierung die *relevance vectors* verwendet. Das sind Datenpunkte, die eine Art typisches Beispiel für die jeweiligen Gruppen darstellen. Im Allgemeinen lässt sich sagen, dass die RVM eine Bayessche Alternative zur SVM ist und insgesamt weniger festzulegende Parameter als diese besitzt [10]. Die Möglichkeit der Anwendung für Regressionen wird dann dadurch erreicht, dass anstatt einer Trennebene eine Hyperebene gefunden wird, die die Datenpunkte bestmöglich beschreibt. Genauere Beschreibungen der mathematischen Hintergründe dieser Erweiterung sind in [27] zu finden.

2.5. Bayessche Optimierung

Bayessche Optimierung ist eine globale Optimierungsmethode, also eine Methode mit der es möglich ist das globale Optimum einer Funktion zu bestimmen beziehungsweise zu approximieren. Sie ist für Zielfunktionen geeignet, zu denen keine geschlossene Form vorliegt, es jedoch möglich ist einzelne Datenpunkte gegebenenfalls mit Rauscheinfluss zu berechnen. Im Gegensatz zu vielen anderen Optimierungsverfahren ist die Bayessche Optimierung auch anwendbar, wenn es keine Möglichkeit gibt für diese sogenannten Black-Box-Funktionen Ableitungen zu bestimmen oder keine Konvexität vorliegt. Dabei gehört die Bayessche Optimierung zu den effizientesten Optimierungsverfahren in Bezug auf die benötigten Datenpunkte der Zielfunktion, wodurch sich dieses Verfahren besonders für Funktionen eignet, die eine

zeitintensive Berechnung aufweisen [4]. Diese Tatsache bedeutet insbesondere, dass sich die Bayessche Optimierung besonders für die Hyperparameteroptimierung von maschinellen Lernverfahren eignet, da die Evaluation einzelner Hyperparameterwerte das gesamte Training des Lernverfahrens beinhaltet. Im Gegensatz zu anderen Optimierungsverfahren wird bei der Bayesschen Optimierung ein probabilistisches Modell für die Zielfunktion erstellt und dieses dann in einem iterativen Verfahren durch Hinzunahme weiterer Datenpunkte aktualisiert und verbessert. Hierbei ist der Grundgedanke, dass die gesamte Information aller berechneten Datenpunkte genutzt werden kann und nicht nur der lokale Gradient des letzten bestimmten Datenpunktes. Für die Erzeugung und Aktualisierung dieses probabilistischen Modells wird der Satz von Bayes verwendet. Dieser besagt in der grundlegenden Form, dass die A-posteriori Wahrscheinlichkeit eines Modells M bei gegebenen Daten D proportional zu der Likelihood von D gegeben M multipliziert mit der A-priori Wahrscheinlichkeit von M ist [4]:

$$P(M|D) \propto P(D|M)P(M) \tag{2.4}$$

Die A-priori Wahrscheinlichkeit beschreibt bei der Bayesschen Optimierung das Vorwissen über den Raum der möglichen Zielfunktionen. Liegt beispielsweise die Annahme vor, dass die Zielfunktion eher glatt verläuft, sollten Zielfunktionen mit hoher Varianz eine geringe A-priori Wahrscheinlichkeit aufweisen. Die A-posteriori Wahrscheinlichkeit beschreibt dann entsprechend das aktualisierte probabilistische Modell der unbekannten Zielgerade. Ein weiterer entscheidender Aspekt der Bayesschen Optimierung ist die Verwendung einer sogenannten *acquisition function* zur Bestimmung des jeweils nächsten zu verwendenden Datenpunktes. Diese *acquisition function*, die auf unterschiedlichste Weise definiert werden kann, beschreibt eine Form der Nützlichkeit der Datenpunkte für das Modell. Durch eine Bestimmung des Maximums dieser Funktion wird der nächste zu verwendende Datenpunkt ausgewählt, wodurch eine geringe Anzahl benötigter Datenpunkte bei gleichzeitig gutem Ergebnis erreicht werden kann. Zur Veranschaulichung ist dieses Prinzip in Abbildung 2.1 dargestellt.

Um eine Bayessche Optimierung durchführen zu können, sind also zwei wichtige Bereiche zu beachten. Einerseits die Art der Modellierung der A-priori Wahrscheinlichkeit, für die meist Gaußprozesse verwendet werden und andererseits die Wahl der *acquisition function*. Diese beiden Aspekte sollen im Folgenden näher beleuchtet werden.

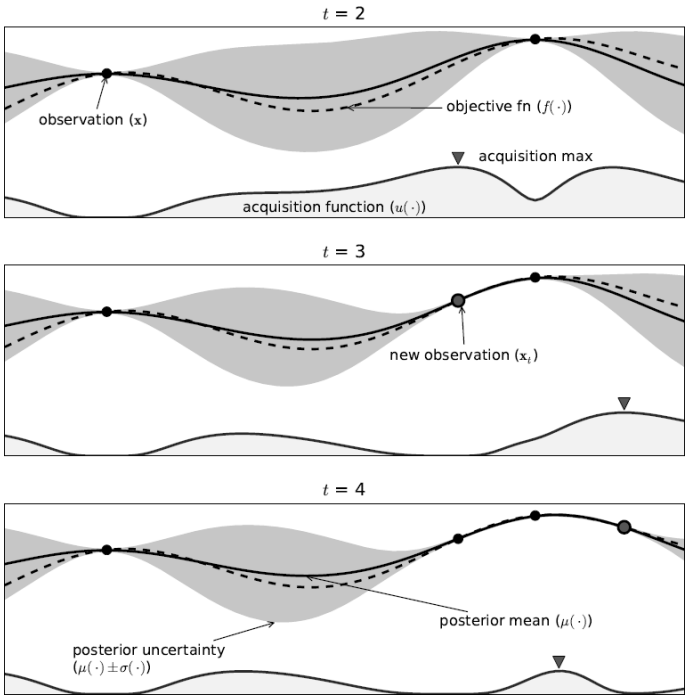

Abbildung 2.1.: Schematische Beispieldarstellung einer Bayesschen Optimierung eines einzelnen Parameters (Maximierung). Die Abbildung zeigt eine Gaußprozess-Approximation des Modells über vier Iterationen. Jeweils im unteren Bereich ist die zugehörige *acquisition function* dargestellt. [4]

2.5.1. Gaußprozesse

Ein Gaußprozess ist ein stochastischer Prozess, bei dem jede endliche Teilmenge von Zufallsvariablen mehrdimensional normalverteilt ist. Damit sind Gaußprozesse eine natürliche Generalisierung der mehrdimensionalen Normalverteilung. Analog zu einer mehrdimensionalen Normalverteilung, die durch den Erwartungswert und die Kovarianzmatrix vollständig und eindeutig bestimmt ist, ist ein Gaußprozess durch eine Erwartungswertfunktion $m(x)$ und eine Kovarianzfunktion $k(x, x')$ vollständig und eindeutig bestimmt [26]. Definiert werden sie mit

$$m(x) = \mathrm{E}[(f(x)] \tag{2.5}$$

und

$$k(x, x') = \mathrm{Cov}(f(x), f(x')) = \mathrm{E}\left[(f(x) - m(x))(f(x') - m(x'))\right] . \tag{2.6}$$

Der Gaußprozess wird dann notiert als

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) . \tag{2.7}$$

Hierbei ist zu beachten, dass ein Gaußprozess im Gegensatz zu mehrdimensionalen Normalverteilungen keine Verteilung von Zufallsvariablen, sondern eine Verteilung von Funktionen darstellt. Für ein Verständnis dieses Konzepts ist es hilfreich, den Gaußprozess als Funktion zu betrachten, die für ein zufälliges x keinen Skalar, sondern den Erwartungswert und die Varianz einer Normalverteilung zurückgibt. Diese Normalverteilung beschreibt dann die Verteilung der möglichen Werte von f an der Stelle x. Ein Gaußprozess kann als A-priori Wahrscheinlichkeit des probabilistischen Modells der Bayesschen Optimierung verwendet werden. Mithilfe von Datenpunkten, zu denen die Funktionswerte bekannt sind, ist es dann möglich die A-posteriori Wahrscheinlichkeit des Modells zu berechnen und somit das Modell zu aktualisieren. Seien x_i, $i = 1, ..., n$ die bekannten Datenpunkte, $\boldsymbol{f}$ ein Vektor aller zugehörigen Funktionswerte $f(x_i)$ und $\boldsymbol{m}$ ein Vektor aller Erwartungswerte $m(x_i)$, dann ist der A-posteriori Gaußprozess gegeben durch [26]

$$\begin{aligned} f(x)|D \sim \mathcal{GP}(m_D(x), k_D(x, x'))\,, \\ m_D(x) = m(x) + \Sigma(X, x)^T \Sigma^{-1}(\boldsymbol{f} - \boldsymbol{m}) \\ k_D(x, x') = k(x, x') - \Sigma(X, x)^T \Sigma^{-1} \Sigma(X, x')\,, \end{aligned} \tag{2.8}$$

wobei $\Sigma(X, x)$ ein Vektor der Kovarianzen zwischen jedem bekannten Datenpunkt und x ist. Zur Erzeugung eines Gaußprozesses, der als A-priori Wahrscheinlichkeit genutzt werden soll, ist es also nötig eine Erwartungswertfunktion und eine Kovarianzfunktion zu definieren. Für die Erwartungswertfunktion wird dabei in den meisten Fällen eine Konstante gewählt [4]. Falls ein bekannter Trend vorliegt, ist es auch möglich diese anwendungsspezifische Struktur durch die Wahl eines niederdimensionalen Polynoms als Erwartungswertfunktion in dem Gaußprozess zu berücksichtigen. Für die Kovarianzfunktion wird in der Regel eine vordefinierte Kernelfunktion verwendet. Ein grundlegender Kernel ist der *automatic relevance determination* (ARD) *squared exponential* Kernel, der definiert ist durch [29]

$$k_{SE}(x, x') = \theta_0 \exp\left\{-\frac{1}{2} r^2(x, x')\right\} \tag{2.9}$$

mit

$$r^2(x, x') = \sum_{d=1}^{d'} \frac{(x_d - x'_d)^2}{\theta_d^2}\,. \tag{2.10}$$

Dabei ist d' die Anzahl der Dimensionen und θ_0 ein Parameter für die Amplitude der Varianz. Je niedriger der Wert von θ_0 ist, desto stärker ist die Begünstigung von Funktionen die nah an der Erwartungsfunktion liegen und umgekehrt. Die Parame-

ter θ_1 bis $\theta_{d'}$ sind ein Maß für die Glätte der Funktionen, wobei jeder Parameter das Verhalten in einer Dimension beeinflusst. Da dieser Kernel in praktischen Optimierungsverfahren teilweise unrealistisch glatte Funktionen generiert, ist ein weiterer oft verwendeter Kernel der ARD Matérn-Kernel, bei dem dieses Problem einen geringeren Stellenwert einnimmt [29]. Der ARD Matérn-Kernel ist definiert durch [4]

$$k_M(x, x') = \theta_0 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu r^2}\right)^{\nu} K_\nu \left(\sqrt{2\nu r^2}\right) . \tag{2.11}$$

$\Gamma(\cdot)$ beschreibt dabei die Eulersche Gammafunktion und $K_\nu(\cdot)$ steht für die modifizierte Bessel-Funktion. ν stellt einen zusätzlichen Parameter dar, der als allgemeiner Glattheitsparameter interpretiert werden kann. Um eine leicht berechenbare Form des ARD Matérn-Kernels zu erzeugen, werden hierfür meist Standardwerte wie zum Beispiel $\nu = 3/2$ oder $\nu = 5/2$ gewählt.

2.5.2. Acquisition function

Bei dem Einsatz von Bayesscher Optimierung für die Hyperparameteroptimierung eines Lernverfahrens ist das Ziel immer eine Minimierung, da dabei stets versucht wird den Fehler des Lernverfahrens zu minimieren. Daher soll die *acquisition function* an Punkten hohe Werte aufweisen, an denen sehr wahrscheinlich ein niedriger Wert der zu optimierenden Funktion zu finden ist. Aus diesem Grund sind typische *acquisition functions* so formuliert, dass sich ein hoher Wert ergibt, wenn das geschätzte Modell an dieser Position einen geringen Wert aufweist, eine große Ungenauigkeit an dieser Position vorliegt, oder wenn beides zutrifft [4]. Sehr häufig wird für die *acquisition function* das *expected improvement* (EI) verwendet. Neben einem besonders wünschenswerten Verhalten bei Optimierungen, ist ein weiterer großer Vorteil des EI gegenüber anderen möglichen Funktionen, dass für die Definition keine festzulegenden Parameter benötigt werden [29]. Außerdem umfasst die Definition des EI nur wenige Schritte. Ist $f' = \min \boldsymbol{f}$ der bisher niedrigste beobachtete Wert, lässt sich die Verbesserung an einem Punkt x mit der Funktion

$$u(x) = \max\ (0, f' - f(x)) \tag{2.12}$$

beschreiben. Das EI ist dann die erwartete Verbesserung als eine Funktion von x und lässt sich notieren als [15, 38]

$$\begin{aligned} \mathrm{EI}(x)|D = \mathrm{E}\left[u(x)\right] &= \int_{-\infty}^{f'} (f' - f)\, \mathcal{N}\left(f; m_D(x), k_D(x,x)\right)\, \mathrm{d}f \\ &= (f' - m_D(x))\, \Phi(f'; m_D(x), k_D(x,x)) \\ &\quad + k_D(x,x)\, \mathcal{N}(f'; m_D(x), k_D(x,x))\,. \end{aligned} \tag{2.13}$$

Diese Notation veranschaulicht auch die Funktionsweise des EI. Der erste Term wird durch eine Verringerung der Erwartungswertfunktion $m_D(x)$ vergrößert und der zweite Term durch eine Erhöhung der Varianz $k_D(x,x)$. Hierbei wird also der Trade-off zwischen *exploitation* (Auswahl von Punkten mit einem niedrigen geschätzten Wert) und *exploration* (Auswahl von Punkten mit einer hohen Ungenauigkeit) explizit dargestellt. Dadurch wird deutlich, dass dieser Trade-off bei der Verwendung des EI automatisch berücksichtigt wird und keiner Einstellung durch einen Parameter bedarf. Der Punkt mit dem höchsten EI wird dann als nächster zu beobachtender Punkt ausgewählt. Um dieses Maximum der *acquisition function* zu bestimmen gibt es verschiedene effizient durchzuführende Methoden.

2.6. Ensemble learning

Maschinelle Lernverfahren versuchen mithilfe von Trainingsdaten ein Modell zu generieren, das die Daten bestmöglich beschreibt und dabei gleichzeitig eine möglichst hohe Generalisierung für ungesehene Daten aufweist (vgl. Definition des Regressionsproblems in Abschnitt 2.4). *Ensemble learning* stellt eine besondere Klasse von maschinellen Lernverfahren dar und kann sowohl für eine Klassifikation, als auch für eine Regression verwendet werden. Ensemblemethoden zeichnet aus, dass sie eine Menge von Modellen verwenden, die jeweils durch die Anwendung eines Lernverfahrens generiert wurden. Diese Menge von Modellen (Ensemble) wird zu einem Gesamtmodell kombiniert, das dann eine Vorhersage ermöglicht [21]. Die Idee ist also nicht den Raum aller möglichen Modelle nach dem besten Modell zu durchsuchen, sondern mehrere geeignete Modelle aus diesem Raum zu einem neuen Modell zu kombinieren. Dieses finale Modell muss dabei nicht zu dem Raum gehören, der durch die im ersten Schritt eingesetzten Lernverfahren definiert wurde. Die möglichen Einsatzfelder von *ensemble learning* sind nicht eingeschränkt, wobei jedoch zu beachten ist, dass der zu erbringende Aufwand gegenüber anderen Lernverfahren deutlich erhöht ist. Laut Dietterich [9] können aber bei Lernverfahren die keine Ensemblemethoden nutzen drei Probleme auftreten, bei denen die Verwendung von

ensemble learnig zu einer Verbesserung führen kann. Er nennt diese Probleme *statistical problem*, *computational problem* und *representational problem*. Das *statistical problem* kann auftreten, wenn der zu durchsuchende Raum zu groß für die Anzahl der Trainingsdaten ist. In diesem Fall ist die Wahrscheinlichkeit groß, dass es mehrere verschiedene Lösungen mit gleicher Genauigkeit gibt und das Verfahren sich dann für eines der Modelle entscheiden muss. Da die verschiedenen Modelle unterschiedliche Generalisierungseigenschaften aufweisen können, ist das Risiko dabei hoch ein Modell zu wählen, dass auf ungesehenen Daten schlechte Ergebnisse liefert. Eine Kombination mehrerer Modelle verringert dann das Risiko sich von einem einzelnen Modell abhängig zu machen, auch wenn die Kombination nicht unbedingt zu einer Verbesserung auf den zur Verfügung stehenden Daten führen muss. Das *computational problem* beschreibt die Schwierigkeit die viele Lernverfahren haben ein lokales Optimum von einem globalen Optimum zu unterscheiden. Ist es möglich, dass ein verwendetes Lernverfahren in einem lokalen Optimum endet, kann eine mehrmalige Durchführung und anschließende Wichtung der Ergebnisse zu einem verbesserten Modell führen. Als drittes und letztes Problem beschreibt Dietterich [9] das *representational problem*. Dieses tritt auf, wenn der Raum der Modelle zu klein ist und keines der zur Verfügung stehenden Modelle eine gute Approximation des unbekannten Systems darstellt. In manchen Fällen kann diese zu geringe Komplexität durch die Verwendung von Ensemblemethoden kompensiert werden. Eine Empfehlung zur Verwendung von *ensemble learning* in diesen drei Fällen ist jedoch rein theoretischer Natur, da die Frage ob eines dieser Probleme vorliegt in der Praxis oft nicht gesichert beantwortet werden kann. Die Formulierung dieser Probleme dient eher dem Verständnis von Ensemblemethoden und soll zeigen, was sie leisten können und was nicht. In der Praxis hat sich jedoch im Allgemeinen gezeigt, dass die Berechnung mehrerer Lernverfahren mit geringer Komplexität und eine anschließende Kombination der Ergebnisse oft effizienter durchzuführen ist, als die Berechnung eines einzelnen Lernverfahrens mit hoher Komplexität [9]. Die zu erreichende Verbesserung durch den Einsatz von *ensemble learning* ist aber selbstverständlich von der jeweiligen Situation und vor allem von der jeweiligen Implementierung abhängig.

Eine Ensemblemethode besteht im Wesentlichen aus zwei Schritten, die separat voneinander betrachtet werden können. Zunächst die Erzeugung der Menge von Modellen (*ensemble generation*) und danach die Kombination der Modelle in ein finales Modell (*ensemble integration*). Für beide Schritte gibt es verschiedene Ansätze die jeweils Vor- und Nachteile aufweisen. Um ein grundlegendes Verständnis für die Umsetzung von Ensemblemethoden zu liefern, werden beide Bereiche im Folgenden näher beleuchtet.

2.6.1. Ensemble generation

Wenn ein perfektes Lernverfahren zur Verfügung steht, das für ein gegebenes Problem keine Fehler aufweist, ist der Einsatz von *ensemble learning* überflüssig. Weist ein Lernverfahren jedoch Ungenauigkeiten oder Fehlklassifikationen auf, liegt die Idee nahe das durch ein weiteres Lernverfahren zu kompensieren, dessen Fehler bei anderen Objekten oder anderen Merkmalen auftreten. Neben der leicht ersichtlichen Tatsache, dass die verwendeten Lernverfahren möglichst genau sein sollten, ist eine weitere wichtige Eigenschaft also, dass die verwendeten Lernverfahren möglichst unterschiedliche Modelle erzeugen sollten [19]. Nur auf diese Weise können sich die Fehler gegenseitig ausgleichen. Diese wünschenswerte Eigenschaft des zu erzeugenden Ensembles nennt sich *diversity* und ist ein grundlegender Aspekt von Ensemblemethoden. Bei der Erzeugung eines Ensembles kann *diversity* durch eine Vielzahl von Strategien erreicht werden. Werden grundlegend verschiedene Ansätze für Lernverfahren verwendet, nennt man das Ensemble heterogen [21]. Hierbei ist meist implizit gesichert, dass sich die erzeugten Modelle deutlich unterscheiden, da die unterschiedlichen Annahmen der Lernverfahren meist unterschiedliche Suchräume für die Modelle zur Folge haben. Dieser Ansatz wird oft dazu verwendet eine automatisierte Bewertung und Wichtung verschiedener Lernverfahren für ein spezifisches Problem durchzuführen. Ist das zugrundeliegende Lernverfahren für alle zu erzeugenden Modelle gleich (homogenes Ensemble), muss die Verschiedenheit durch andere Strategien erreicht werden. Möglichkeiten sind hierfür die Variation der Trainingsdaten, die Manipulation des Merkmalraumes oder die Manipulation des Lernalgorithmus [9]. Um jedem Lernverfahren unterschiedliche Trainingsdaten zur Verfügung stellen zu können, ist es möglich den Gesamtdatensatz in disjunkte Teildatensätze aufzuteilen, oder die verschiedenen Datensätze durch ziehen und zurücklegen zu erzeugen. Den zuletzt genannten Ansatz verfolgen die in der Praxis weit verbreiteten Verfahren *bagging* und *boosting*, die verschiedene Algorithmen besitzen um die Teildatensätze möglichst zielführend zu erzeugen. Eine Manipulation des Merkmalraumes ist durch eine Aufteilung der Merkmale in verschiedene Unterräume erreichbar. Ebenso ist es möglich hierfür verschiedene Repräsentationen der Trainingsdaten zu verwenden. Dies kann entweder durch die Veränderung der Vorverarbeitung beziehungsweise der Merkmalsextraktion, oder durch die Verwendung einer grundlegend verschiedenen Repräsentation erreicht werden. Die als drittes genannte Möglichkeit der Manipulation des Lernalgorithmus ist zusätzlich ein guter Ansatz um trotz des Vorliegens eines homogenen Ensembles eine Verschiedenheit zwischen den Lernverfahren zu erzeugen. Hierfür bietet es sich an in den Algorithmus Zufallselemente einzuführen oder bei iterativ ablaufenden Verfahren

unterschiedliche Startkonfiguration zu wählen.

Allen diesen Strategien liegt die Annahme zugrunde, dass sie für eine Verschiedenheit der Modelle im Ensemble sorgen und diese Verschiedenheit zu besseren und robusteren Ergebnissen führt. Wird *ensemble learning* für eine Klassifikation eingesetzt, gibt es darüber hinaus keine Möglichkeit genauere Aussagen über den gewünschten Grad der Verschiedenheit zu treffen. Im Bereich der Regression, der für diese Arbeit relevant ist, ist dies jedoch anders. Hier gibt es durch die *bias-variance-covariance decomposition* eine Möglichkeit, die Verschiedenheit eines Ensembles analytisch auszudrücken und dadurch den Zusammenhang zwischen der Verschiedenheit und der Qualität des finalen Modells zu beschreiben. Die *bias-variance-covariance decomposition* ist eine Zerlegung des Generalisierungsfehlers, also des Fehlers, der durch den Testdatensatz approximiert wird. Das Ziel einer solchen Zerlegung ist es immer diesen Fehler durch mehrere Terme auszudrücken. Diese können dann Aufschluss darüber geben, welche Eigenschaften ein Modell aufweisen sollte, damit der Fehler minimal wird. Grundlage für die *bias-variance-covariance decomposition* ist die *bias-variance-decomposition*, die 1992 von Geman, Bienenstock und Doursat [13] formuliert wurde. Diese wird für überwachte Lernverfahren eingesetzt, die keine Ensemblemethode verwenden. Um eine bessere Lesbarkeit zu ermöglichen, wird im Folgenden für die Definitionen der Zerlegungen abkürzend $f = f(x)$ und $\hat{f} = \hat{f}(x)$ geschrieben.

Gegeben sei ein Regressionsproblem (siehe Abschnitt 2.4), bei dem alle Einträge des Lerndatensatzes $z = \{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$ durch eine Zufallsvariable Z mit unbekannter Wahrscheinlichkeitsfunktion $p(x, f(x))$ erzeugt wurden. Für die mittlere quadratische Abweichung lautet dann die *bias-variance-decomposition* [13]

$$\begin{aligned} \mathrm{E}\left[\left(\hat{f} - \mathrm{E}[f]\right)^2\right] &= \left(\mathrm{E}[\hat{f}] - \mathrm{E}[f]\right)^2 + \mathrm{E}\left[\left(\hat{f} - \mathrm{E}[\hat{f}]\right)^2\right] \\ &= \mathrm{Bias}\left[\hat{f}\right]^2 \qquad + \mathrm{Var}\left[\hat{f}\right] . \end{aligned} \tag{2.14}$$

Unmittelbar ersichtlich ist hierbei, dass der Generalisierungsfehler in zwei Terme aufgeteilt wird und ein minimaler Fehler durch eine Minimierung beider Terme erreicht werden kann. Der erste Term der Zerlegung wird *bias* oder auch Verzerrung genannt. Er beschreibt den Unterschied zwischen dem Erwartungswert der Schätzfunktion $\hat{f}$ und dem unbekannten Erwartungswert der wahren Funktion f. Anschaulich betrachtet ist die Verzerrung der Fehler ausgehend von falschen Annahmen im Lernalgorithmus. Eine hohe Verzerrung kann einen Algorithmus dazu veranlassen, nicht die vorliegenden Beziehungen zwischen Ein- und Ausgabe zu modellieren. Dies entspricht einer Unteranpassung an das Problem. Der zweite Term, Varianz genannt,

beschreibt hingegen die Schwankungsbreite von $\hat{f}$. Die Varianz ist also der Fehler ausgehend von der Empfindlichkeit auf kleinere Schwankungen in den Trainingsdaten. Bei einer hohen Varianz wird das Rauschen der Trainingsdaten anstelle der vorgesehenen Ausgabe modelliert, es findet also eine Überanpassung statt. Für ein geeignetes Modell ist es daher wichtig, dass sowohl die Verzerrung, als auch die Varianz minimal ist. Dies entspricht der intuitiven Annahme, dass ein überwachtes Lernverfahren idealerweise ein Modell wählen sollte, das sowohl die Gesetzmäßigkeiten in den Trainingsdaten genau erfasst, als auch sich auf ungesehene Testdaten generalisieren lässt. Die Formulierung in der Gleichung 2.14 zeigt jedoch, dass es sich hierbei um gegensätzliche Aspekte handelt und es nicht möglich ist gleichzeitig sowohl die Verzerrung als auch die Varianz zu minimieren. Es handelt sich hierbei also um einen Trade-off, bei dem eine Abwägung beider Bereiche stattfinden muss. Ueda und Nakano [33] erweiterten die *bias-variance-decomposition*, um eine Zerlegung für Ensemblemethoden zu erzeugen. Da bei *ensemble learning* mehrere Modelle erzeugt werden, liegt in diesem Fall nicht eine einzelne Zufallsvariable, sondern eine Menge von Zufallsvariablen $Z = (Z_1, ..., Z_K)$ vor, wobei K die Anzahl der verwendeten Lernverfahren beschreibt. Das i-te Lernverfahren wird dann mit einem Lerndatensatz z_i trainiert, der durch die Zufallsvariable Z_i erzeugt wurde. Dabei ist es natürlich möglich, dass alle Z_i identisch sind, falls für jedes Lernverfahren der gleiche Datensatz verwendet wird. Ueda und Nakano [33] nehmen zusätzlich an, dass die Kombination der Modelle durch eine einfache Mittelwertbildung stattfindet, also das finale Modell $\hat{f}_{\mathcal{F}}(x)$ beschrieben wird durch

$$\hat{f}_{\mathcal{F}}(x) = \frac{1}{K} \cdot \sum_{i=1}^{K} \hat{f}_i(x)\,. \tag{2.15}$$

Auch wenn in der Praxis oft andere Kombinationsmethoden angewendet werden (siehe Abschnitt 2.6.2), verursacht diese Annahme keine Einschränkungen, da die Erkenntnisse aus der Zerlegung auf alle Kombinationsmethoden übertragbar sind [5]. Die durch Ueda und Nakano [33] definierte *bias-variance-covariance decomposition* für die mittlere quadratische Abweichung lautet

$$\mathrm{E}\left[\left(\hat{f}_{\mathcal{F}} - \mathrm{E}[f]\right)^2\right] = \overline{\mathrm{Bias}}^2 + \frac{1}{K} \cdot \overline{\mathrm{Var}} + \left(1 - \frac{1}{K}\right) \cdot \overline{\mathrm{Covar}}\,, \tag{2.16}$$

mit

$$\overline{\text{Bias}} = \frac{1}{K} \cdot \sum_{i=1}^{K} \left(\text{E}_i[\hat{f}_i] - f \right) , \tag{2.17}$$

$$\overline{\text{Var}} = \frac{1}{K} \cdot \sum_{i=1}^{K} \text{E}_i \left[\left(\hat{f}_i - \text{E}_i[\hat{f}_i] \right)^2 \right] , \tag{2.18}$$

$$\overline{\text{Covar}} = \frac{1}{K \cdot (K-1)} \cdot \sum_{i=1}^{K} \sum_{j=1, j \neq i}^{K} \text{E}_{i,j} \left[\left(\hat{f}_i - \text{E}_i[\hat{f}_i] \right) \left(\hat{f}_j - \text{E}_i[\hat{f}_j] \right) \right] . \tag{2.19}$$

Zusätzlich zu der Verzerrung und der Varianz, gibt es bei Ensemblemethoden also noch einen dritten Term der den Generalisierungsfehler bestimmt. Diesen Term nennt man Kovarianz und er ist ein Maß für den Zusammenhang der einzelnen Modelle. Damit beschreibt dieser Term genau das Konzept der Verschiedenheit, das zuvor erläutert wurde. Hier wird nun also auch analytisch deutlich, dass ein kleiner Generalisierungsfehler durch eine minimale Kovarianz und damit eine maximale Verschiedenheit der Modelle erreicht werden kann. Die *bias-variance-covariance decomposition* zeigt jedoch auch, dass eine Verringerung der Kovarianz zu einer Erhöhung der anderen beiden Terme führt [5]. Es handelt sich hierbei also ebenso um einen Trade-off, da eine gleichzeitige Minimierung aller drei Terme unmöglich ist. Ein minimaler Generalisierungsfehler wird demnach durch die richtige Balance zwischen der Genauigkeit der einzelnen Modelle, die die Verzerrung und die Varianz beeinflussen, und der Verschiedenheit der Modelle erreicht.

Abschließend sei noch erwähnt, dass in anderen Beschreibungen der vorgestellten Zerlegungen oft zusätzlich ein irreduzibler Fehler additiv hinzugefügt wird, der das Rauschen in den Daten repräsentiert. Erst mit diesem Term ergeben die Gleichungen 2.14 und 2.16 vollständige Zerlegungen des Generalisierungsfehlers. Da der Einfluss des Rauschens jedoch nicht durch eine aktive Gestaltung des Lernverfahrens beeinflusst werden kann, spielt er für die Betrachtung von gewünschten Eigenschaften des Lernverfahrens keine Rolle. Aus diesem Grund wurde in diesem Abschnitt das Rauschen vernachlässigt und ein Datensatz angenommen, der kein Rauschen enthält.

2.6.2. Ensemble integration

Bei Regressionsproblemen wird die *ensemble integration* durch eine lineare Kombination durchgeführt. Die kann formuliert werden als [21]

$$\hat{f}_{\mathcal{F}}(x) = \sum_{i=1}^{K} \left(h_i(x) \cdot \hat{f}_i(x) \right) , \tag{2.20}$$

wobei $h_i(x)$ die Gewichtsfunktionen darstellen. Eingeteilt werden die verschiedenen Umsetzungen dann danach, ob die Gewichtsfunktionen konstant oder nicht konstant sind [21]. Im Fall von konstanten Gewichtsfunkionen werden lediglich Konstanten als Gewichte gewählt, dabei gilt also $h_i(x) = \alpha_i$. Bei nicht konstanten Gewichtsfunktionen hängen die jeweiligen Gewichte von der Eingabe x ab. Die grundlegendste Methode hierfür ist die Bildung des Mittelwertes, die in Gleichung 2.15 abgebildet ist. Dabei sind die Gewichte definiert als $h_i = 1/K$. Diese Methode basiert nicht auf den Modellen oder den Daten und nimmt an, dass die Fehler der Modelle $(f(x) - \hat{f}_i(x))$ einen Mittelwert von Null haben und unabhängig voneinander sind [21]. Eine etwas komplexere Methode ist die *generalised ensemble method*. Bei dieser Methode werden zunächst die Genauigkeiten der Modelle durch einen separaten Validierungsdatensatz bestimmt und im Anschluss daran die Gewichte α_i umgekehrt proportional zu den Fehlern gewählt. Dabei gilt $\sum_{i=1}^{K} \alpha_i = 1$. Durch diese Herangehensweise werden bessere Modelle automatisch stärker gewichtet, weswegen dabei meist eine genauere Lösung als durch die einfache Mittelwertbildung erreicht werden kann. Eine dritte oft angewendete Möglichkeit für eine Kombination der Modelle mit nicht konstanten Gewichtsfunktionen ist die Nutzung einer linearen Regression. Dieser Ansatz entspricht der *generalised ensemble method*, wobei in diesem Fall die Summe der Gewichte nicht Eins ergeben muss. Ob dabei eine Formulierung der linearen Regression mit zusätzlicher Konstante oder ohne gewählt wird hat in der Umsetzung keine Relevanz, da $\mathrm{E}[\hat{f}_i(x)] \approx \mathrm{E}[f(x)]$ [21]. Über die genannten Ansätze hinaus, gibt es noch viele weitere Methoden zur Kombination der Modelle. Eine umfassende Übersicht ist in [21] zu finden.

3. Methodik

3.1. Datensatz

Für diese Arbeit wurden die MRT-Aufnahmen des öffentlich zugänglichen IXI-Datensatzes [39] verwendet. Dieser Datensatz enthält T1-gewichtete Aufnahmen der Gehirne von 547 gesunden Proband*innen im Alter von 19 bis 86 Jahren (242 männlich, 305 weiblich). Die Verteilung des Alters ist in Abbildung 3.1 dargestellt. Die Aufnahmen des IXI-Datensatzes wurden insgesamt an drei verschiedenen Standorten mit jeweils unterschiedlichen MRT-Geräten erstellt (Philips 1.5 T, General Electric 1.5 T, Philips 3 T). Genaue Informationen zu den MRT-Geräten, sowie den verwendeten Parametern bei den Messungen sind unter [39] zu finden. Um bei allen Verfahren eine vierfache Kreuzvalidierung durchführen zu können, wurde der Datensatz mithilfe der cvpartition-Funktion [37] in Matlab zufällig in vier nahezu gleich große Teildatensätze geteilt. Davon dienten dann jeweils drei für das Training und einer für den Test.

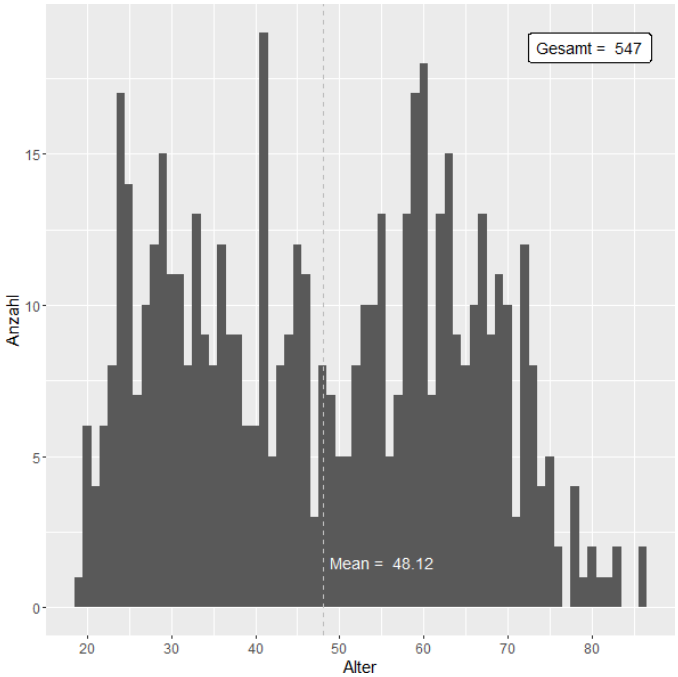

Abbildung 3.1.: Verteilung des Alters der Proband*innen

3.2. Vorverarbeitung

Die Vorverarbeitung der MRT-Aufnahmen wurde mit dem Paket SPM12 [40] und der Toolbox CAT12 [36] durchgeführt. Dafür wurden zunächst alle T1-gewichteten Bilder mit einem Verfahren von Ashburner und Friston [3] segmentiert, normalisiert und für die Entfernung von Inhomogenitäten korrigiert. Zur Berücksichtigung von PVEs wurde das Verfahren dann wie von Gaser [11] beschrieben erweitert und zusätzlich mit einer anschließenden affinen Registrierung ergänzt. Der letzte Schritt der Vorverarbeitung umfasste dann jeweils noch die Anwendung eines Glättungsfilters und die Durchführung eines Resamplings, wobei die Größe des Glättungsfilters und die Auflösung nach dem Resampling die Vorverarbeitungsparameter sind, die in dieser Arbeit betrachtet wurden. Lancaster et al. [20] beschreiben als möglichen Parameterbereich für das FWHM des Glättungsfilters 1-10 mm und als Parameterbereich für die Auflösung der Voxel nach dem Resampling 1-15 mm. In der praktischen Anwendung werden jedoch meist keine sehr kleinen Auflösungen sowie keine sehr geringen Filtergrößen verwendet [10]. Dies folgt aus der Tatsache, dass kleine Auflösungen zu umfangreichen und nicht effizient verwendbaren Datenmengen führen und kleine Filtergrößen den Einfluss von Rauschfaktoren auf die Daten nur wenig verringern. Andererseits liegen gegensätzliche Effekte vor, wenn die Parameter zu groß gewählt werden. Typische Werte für beide Parameter liegen daher zwischen 4 und 8 mm [10]. Um eine umfangreiche Untersuchung zu gewährleisten wurde für beide Parameter jeweils der Bereich zwischen 3 und 9 mm festgelegt, innerhalb dessen die optimalen Werte gesucht werden sollten. Von den resultierenden Schätzungen der Anteile der verschiedenen Gewebearten in einem Voxel wurden jeweils nur die Schätzungen der GM verwendet. Erklärungen und Hintergrundinformationen zu der Vorverarbeitung sind in Abschnitt 2.3 zu finden.

3.3. Modell zur Vorhersage des Alters

In der Regel weisen viele Voxel einer strukturellen MRT-Aufnahme eines Gehirns hohe Korrelationen zueinander auf. Die daraus resultierenden redundanten Voxel führen zu einer Erhöhung der Dimensionalität der Daten ohne eine Erhöhung der Menge an Information zu bewirken. Da dies eine Verschlechterung des Modells zur Folge haben kann, ist es wichtig vor der Anwendung von Mustererkennungsverfahren auf MRT-Daten eine Dimensionsreduktion durchzuführen [2]. In dieser Arbeit fand dafür eine *principal component analysis* (PCA) Verwendung, deren Transformation mithilfe der Trainingsdaten bestimmt und dann mit einer Ergebnisgröße von 410 Komponenten auf alle Daten angewendet wurde. Wie von Franke et al. [10]

beschrieben, wurde als Vorhersagemodell für das biologische Alter des Gehirns eine RVR verwendet, die mit der Toolbox The Spider (Version 1.71) [42] berechnet wurde. Das Ziel der Erstellung des Modells war dabei eine möglichst genaue Vorhersage des wahren Alters der gesunden Proband*innen. Um die Genauigkeit der verschiedenen Methoden zur Optimierung der Vorverarbeitungsparameter vergleichen zu können, wurde jeweils der MAE und der RMSE in Jahren bestimmt. Wie in [10] empfohlen, wurde als *kernel* für die RVR ein Polynom vom Grad eins gewählt. Hintergründe zu der Verwendung einer RVR für die Bestimmung des biologischen Alters des Gehirns sind in Abschnitt 2.4 zu finden.

3.4. Rastersuche

Bei einer Rastersuche werden vordefinierte Parametersätze unabhängig voneinander verwendet und im Anschluss der Parametersatz gewählt, mit dem das beste Ergebnis erzielt werden konnte. Zur Optimierung der Filtergröße und der Auflösung wurden alle ganzzahligen Kombinationen innerhalb der festgelegten Parameterbereiche ausgewählt. Um die Rastersuche durchführen zu können, wurde zunächst für jedes der 49 Parameterpaare ein Datensatz durch die Vorverarbeitung mit der passenden Filterung und dem passenden Resampling erstellt. Daraufhin wurde für jeden Datensatz das zuvor beschriebene Verfahren zur Vorhersage des Alters in einer vierfachen Kreuzvalidierung durchgeführt. Die RVR wurde also jeweils vier mal trainiert und getestet, sodass jede MRT-Aufnahme einmal Bestandteil des Testdatensatzes war. Als vergleichende Größe fand dann der gemittelte MAE, sowie der gemittelte RMSE Verwendung.

3.5. Bayessche Optimierung

Bei einer Bayesschen Optimierung werden im Gegensatz zu einer Rastersuche nicht nur die Ergebnisse einzelner Parametersätze miteinander verglichen, sondern ein gesamtes Modell für den Parameterraum erzeugt (siehe Abschnitt 2.5). Um ein initiales Modell zu erhalten, mit dem es dann möglich ist über eine *acquisition function* den jeweils nächsten zu beobachtenden Parametersatz zu bestimmen, werden dabei die ersten Parametersätze zufällig gewählt. Für die Bayessche Optimierung der Filtergröße und der Auflösung wurden 80 Iterationen durchgeführt, wovon die ersten 20 ein zufälliges Parameterpaar nutzten. Die Parameterbereiche wurden auch hier jeweils auf 3 bis 9 mm begrenzt. Als zu minimierende Zielgröße diente der gemittelte MAE des Verfahrens zur Vorhersage des Alters unter Verwendung einer vierfachen

Kreuzvalidierung. Für die Umsetzung der Bayesschen Optimierung wurde die Statistics and Machine Learning Toolbox (Version 11.5) [41] verwendet, wobei die EI aus Gleichung 2.13 als *acquisition function* gewählt wurde. Als Kovarianzfunktion wurde der ARD Matérn-Kernel mit $\nu = 5/2$ verwendet, da dieser auch ohne weitere Informationen über die zu optimierenden Parameter meist robuste Ergebnisse liefert [29]. Durch die Wahl von $\nu = 5/2$ ergibt sich aus der Definition in Gleichung 2.11 die Formulierung

$$k(x, x') = \theta_0^2 \left(1 + \sqrt{5}r + \frac{5}{3}r^2\right) \exp\left(-\sqrt{5}r\right) \tag{3.1}$$

mit

$$r = \sqrt{\sum_{d=1}^{d'} \frac{(x_d - x'_d)^2}{\theta_d^2}}\,. \tag{3.2}$$

3.6. Ensemblemethode

Das Ziel des Einsatzes einer Ensemblemethode zur Parameteroptimierung besteht darin, verschiedene Parametersätze zu kombinieren und so eine gewichtete Auswahl zu erreichen. Es handelt sich hierbei also nicht um eine Optimierung im eigentlichen Sinne, da nicht die Optimalwerte der Parameter gesucht werden, sondern eine möglichst erfolgreiche Kombination mehrerer Parameterwerte ermittelt werden soll. In dem vorliegenden Fall geht es also darum die Ergebnisse, die unter der Verwendung verschiedener Parameterpaare erzielt wurden, in einem zusammenfassenden Modell zu kombinieren. In dieser Arbeit wurde für die Kombination (vgl. *ensemble integration* in Abschnitt 2.6.2) ein lineares Regressionsmodell verwendet. Das Modell hat die Form

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon\,. \tag{3.3}$$

$X_1, X_2, ..., X_n$ beschreiben dabei die Schätzungen des Alters unter Verwendung der n verschiedenen Parameterkombiantionen, $\beta_0, \beta_1, ..., \beta_n$ die zu bestimmenden Koeffizienten des Modells, y das Alter und ϵ den Fehler durch Rauschfaktoren. Zur Bestimmung der Koeffizienten des Regressionsmodells wurde die Statistics and Machine Learning Toolbox (Version 11.5) [41] verwendet, die mit einem *least square* Verfahren arbeitet. Der algorithmische Ablauf bestand also aus zwei Schritten. Zunächst wurde für bestimmte Parameterpaare jeweils ein Datensatz durch die passende Filterung und das passende Resampling erstellt und damit das in Abschnitt 3.3 beschriebene Modell zur Vorhersage des Alters trainiert. Die durch diese Mo-

delle geschätzten Werte für das Alter der Proband*innen ergaben dann die Daten für das Regressionsmodell, das im Anschluss trainiert wurde. Dabei wurde auch hier mit einer vierfachen Kreuzvalidierung gearbeitet, wobei der Trainingsdatensatz sowohl für das Training des Modells im ersten Schritt, als auch für die Bestimmung des Regressionsmodells verwendet wurde. Aufgrund der vergleichsweise kleinen zur Verfügung stehenden Datenmenge gegenüber der großen Komplexität des Problems, wurde eine Aufteilung des Trainingsdatensatzes auf die beiden Trainingsstufen nicht implementiert.

Um sowohl Aussagen über die Einsatzmöglichkeiten des Verfahrens, als auch über den Einfluss der Anzahl der verwendeten Parameterpaare in der Regression treffen zu können, wurde die beschriebene Methode in drei verschiedenen Varianten durchgeführt:

1. Nutzung von 49 Parameterpaaren (alle ganzzahligen Paare zwischen 3 und 9) mit einer Regression als Ensemblemethode, die alle 49 Parameterpaare kombiniert.

2. Nutzung von lediglich neun Parameterpaaren (alle Paare der Werte 4,6 und 8) mit einer Regression als Ensemblemethode, die alle neun Parameterpaare kombiniert.

3. Nutzung von 49 Parameterpaaren (alle ganzzahligen Paare zwischen 3 und 9) mit einer Regression als Ensemblemethode, die jeweils nur die drei Parameterpaare mit dem geringsten MAE auf den Trainingsdaten kombiniert.

4. Ergebnisse

4.1. Rastersuche

Bei der Rastersuche erreichte das Parameterpaar aus 4 mm Filtergröße und 5 mm Auflösung (S4 - R5) mit einem MAE von 4.294 Jahren und einem RMSE von 5.466 Jahren das beste Ergebnis. Das schlechteste Ergebnis, mit einem MAE von 4.796 Jahren und einem RMSE von 6.064 Jahren, wurde mit einer Filtergröße von 4 mm und einer Auflösung von 9 mm (S4 - R9) erzielt. Von den 49 getesteten Parameterpaaren wurden die 22 geringsten MAE jeweils mit einer Auflösung von 6 mm oder geringer erreicht. Die Verteilung der Fehler der verschiedenen Parameterpaare ist in Abbildung 4.1 zu sehen. Die benötigte Zeit zur Erstellung der 49 Datensätze durch Filterung und Resampling betrug 100 Minuten und 13 Sekunden. Die Rastersuche selbst, also das Training des Modells für alle Parameterpaare, dauerte 21 Minuten und 5 Sekunden. Insgesamt ergibt sich für die Rastersuche also eine Gesamtzeit von 121 Minuten und 18 Sekunden.

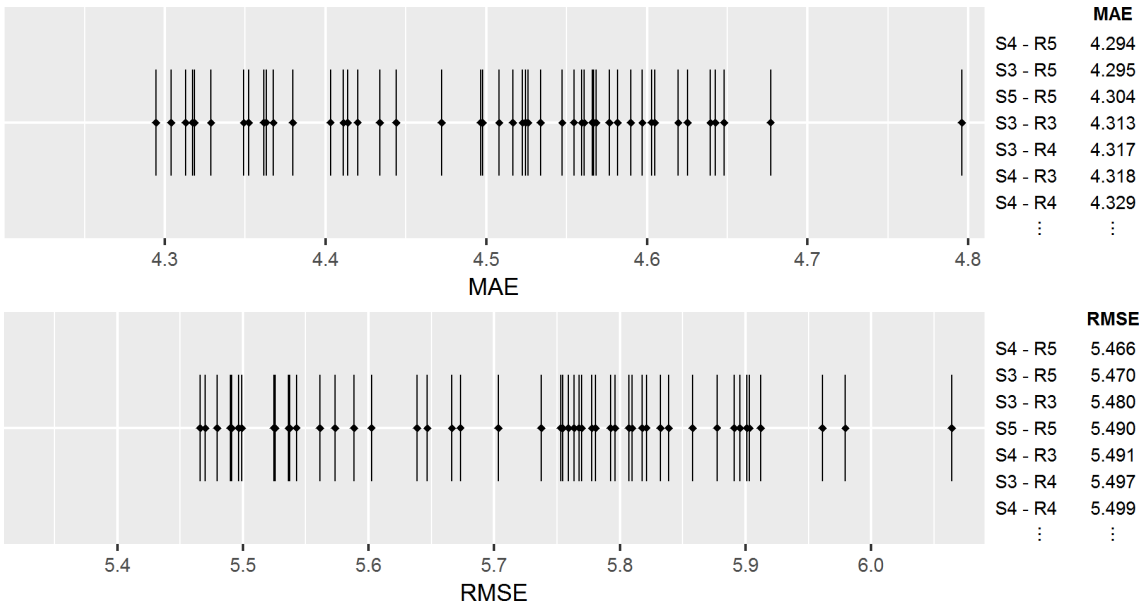

Abbildung 4.1.: MAE (in Jahren) und RMSE (in Jahren) der 49 verschiedenen Parameterkombinationen bei der Rastersuche mit Angabe der jeweils 7 besten Werte

4.2. Bayessche Optimierung

Die Bayessche Optimierung ergab ein geschätztes Optimum bei einer Filtergröße von 3.009 mm und einer Auflösung von 4.993 mm. Mit diesen Parameterwerten weist das Modell einen MAE von 4.264 Jahren und einen RMSE von 5.451 Jahren auf. Abbildung 4.2 zeigt die geschätzten Optima über alle Iterationen der Optimierung. Deutlich wird hierbei, dass nach einer anfänglich starken Verringerung des Fehlers nur noch in kleinerem Maße Verringerungen stattfinden. Die letzte erzielte Verbesserung ist in Iteration 76 erreicht worden. Ein Abbruch der Optimierung nach weniger Iterationen hätte also zu einem schlechteren Ergebnis geführt.

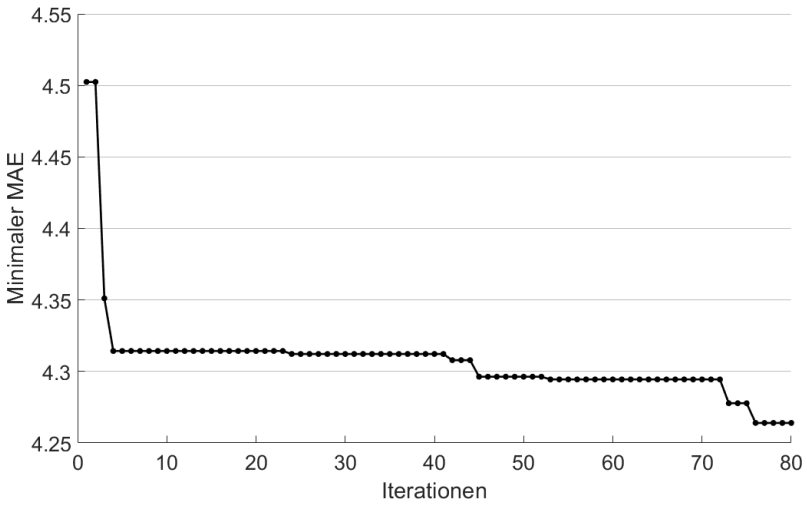

Abbildung 4.2.: Geschätztes Minimum des MAE über die 80 Iterationen der Bayesschen Optimierung

Der Erwartungswert des geschätzten Modells für den zweidimensionalen Parameterraum ist gemeinsam mit den beobachteten Parameterpaaren in Abbildung 4.3 dargestellt. Dabei wird ersichtlich, dass der Parameterraum kein unmittelbar deutlich werdendes Minimum aufweist, dessen Fehler sich drastisch von allen anderen Bereichen des Raumes unterscheidet. Darüber hinaus ist ein allgemeiner Trend zu kleineren Werten für die Parameter zu erkennen. Dieser Trend reicht jedoch zur Beschreibung des Raumes nicht aus, da das Minimum nicht bei dem Parameterpaar mit den jeweils kleinsten Werten liegt. Laut dem geschätzten Modell sind geringere Fehler bei einer Filtergröße zwischen 3 und 5 mm und einer Auflösung zwischen 3 und 6 mm zu erwarten. In diesem Bereich zeigen sich jedoch lokale Schwankungen, wobei diese insbesondere hinsichtlich der Auflösung auftreten. Die Durchführungsdauer der Bayesschen Optimierung betrug 179 Minuten und 15 Sekunden.

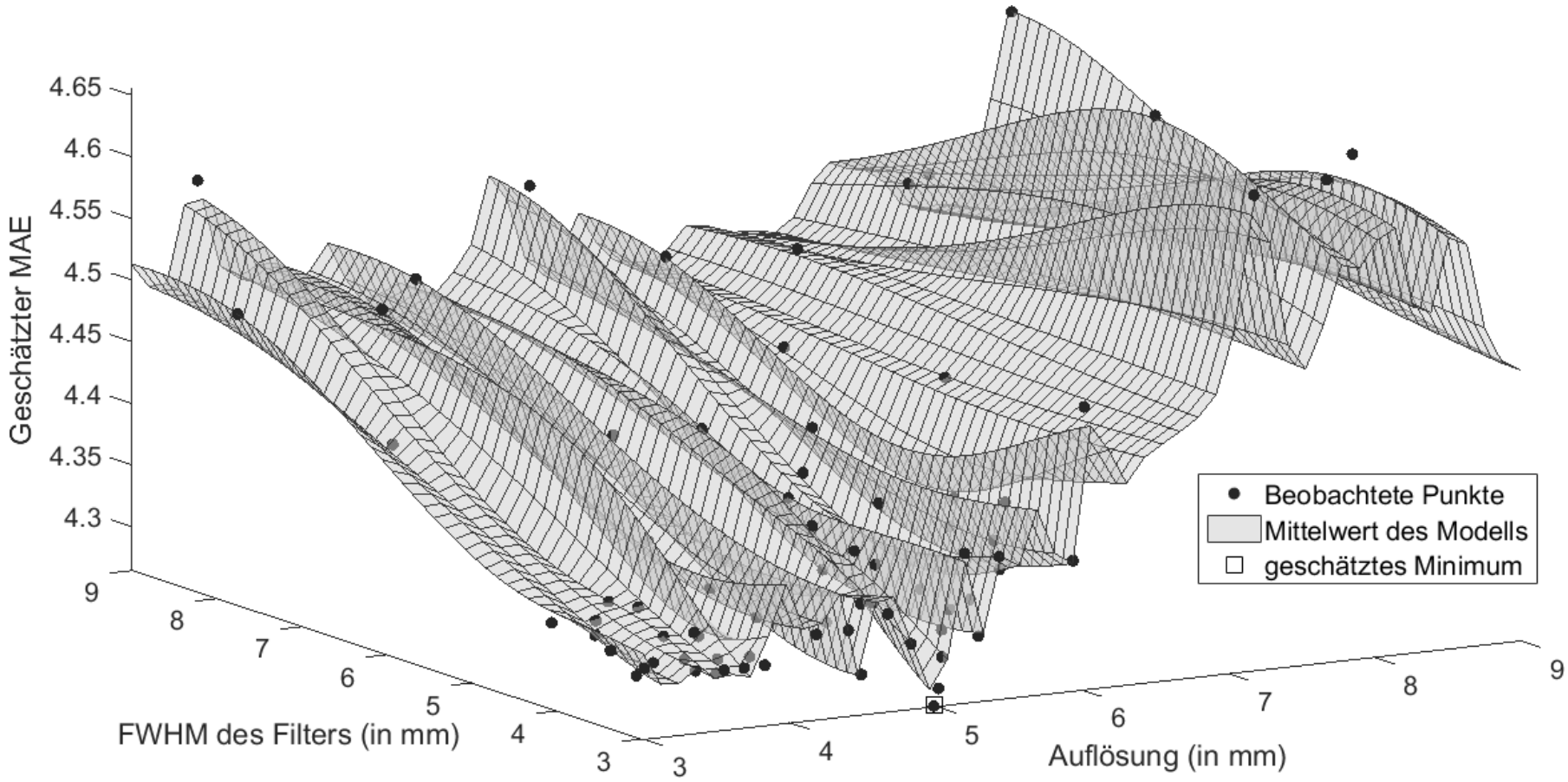

Abbildung 4.3.: Geschätztes Modell der Bayesschen Optimierung nach 80 Iterationen (geschätzter MAE in Jahren). Eingezeichnet sind zusätzlich die beobachteten Parameterpaare und das geschätzte Minimum bei einer Filtergröße von 3.009 mm und einer Auflösung von 4.993 mm.

4.3. Ensemblemethode

4.3.1. Kombination von 49 Parameterpaaren

Das Ensembleverfahren, das alle 49 Parameterpaare in einer gemeinsamen Regression verbindet, erreichte einen MAE von 4.474 Jahren und einen RMSE von 5.576 Jahren. Damit wies dieses Verfahren einen deutlich höheren Fehler auf als viele einzeln verwendeten Parameterpaare. Die Ensemblebildung hat in diesem Fall also zu einer Verringerung der Genauigkeit geführt. In Abbildung 4.4 ist die Einordnung des Fehlers dieser Ensemblemethode in die Fehler der einzelnen Parameterpaare dargestellt. Da hierbei die gleichen Parameterpaare wie bei der in Abschnitt 4.1 beschriebenen Rastersuche verwendet wurden, lag auch hier die benötigte Zeit für die Erstellung der Datensätze für alle Parameterpaare durch Filterung und Resampling bei 100 Minuten und 13 Sekunden. Die Durchführung dieser Variante der Ensemblemethode dauerte daraufhin zusätzlich noch 21 Minuten und 31 Sekunden. Es ergibt sich für diesen Verfahren also ein Gesamtzeitaufwand von 121 Minuten und 44 Sekunden.

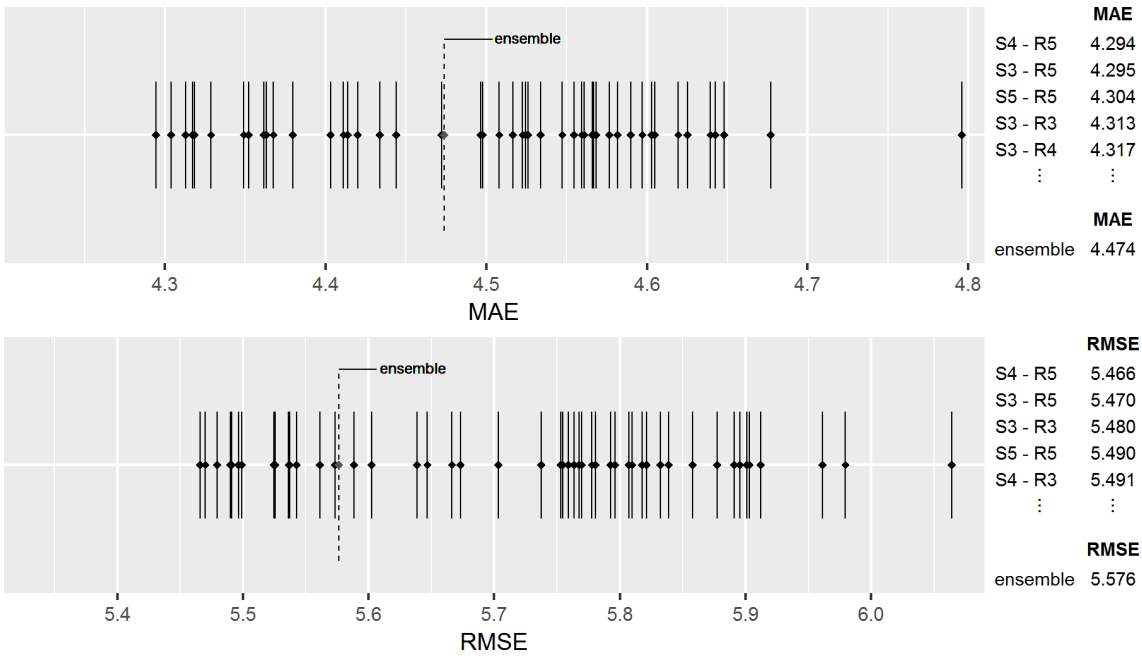

Abbildung 4.4.: MAE (in Jahren) und RMSE (in Jahren) der 49 verschiedenen Parameterpaare und der Ensemblemethode, die alle 49 Parameterpaare kombiniert. Jeweils mit Angabe der Werte der fünf besten Parameterpaare und dem Wert der Ensemblemethode

4.3.2. Kombination von neun Parameterpaaren

Die Variante der Ensemblemethode, die weniger Parameterpaare nutzt, erreichte deutlich bessere Werte. Bei der Verwendung der neun Parameterpaare aus den Werten 4, 6 und 8 ergab das Ensembleverfahren einen MAE von 4.239 Jahren und einen RMSE von 5.353 Jahren. Die geringsten Fehlerwerte der neun einzelnen Parameterpaare ohne eine Ensemblebildung wurden mit einer Filtergröße von 4 mm und einer Auflösung von 4 mm erreicht (S4 - R4). Dieses Parameterpaar erzielte einen MAE von 4.329 Jahren und einen RMSE von 5.499 Jahren. Die Ensemblebildung erbrachte also eine Verbesserung des MAE um 0.09 Jahre und eine Verbesserung des RMSE um 0.146 Jahre. Der Vergleich der Ensemblemethode gegenüber den einzelnen Parameterpaaren ist in Abbildung 4.5 dargestellt. Für die Erstellung der neun Datensätze durch Filterung und Resampling wurde eine Zeit von 17 Minuten und 48 Sekunden benötigt. Das Training der Ensemblemethode dauerte daraufhin noch 3 Minuten und 26 Sekunden. Diese Variante der Ensemblemethode benötigte also einen Gesamtzeitaufwand von 21 Minuten und 14 Sekunden.

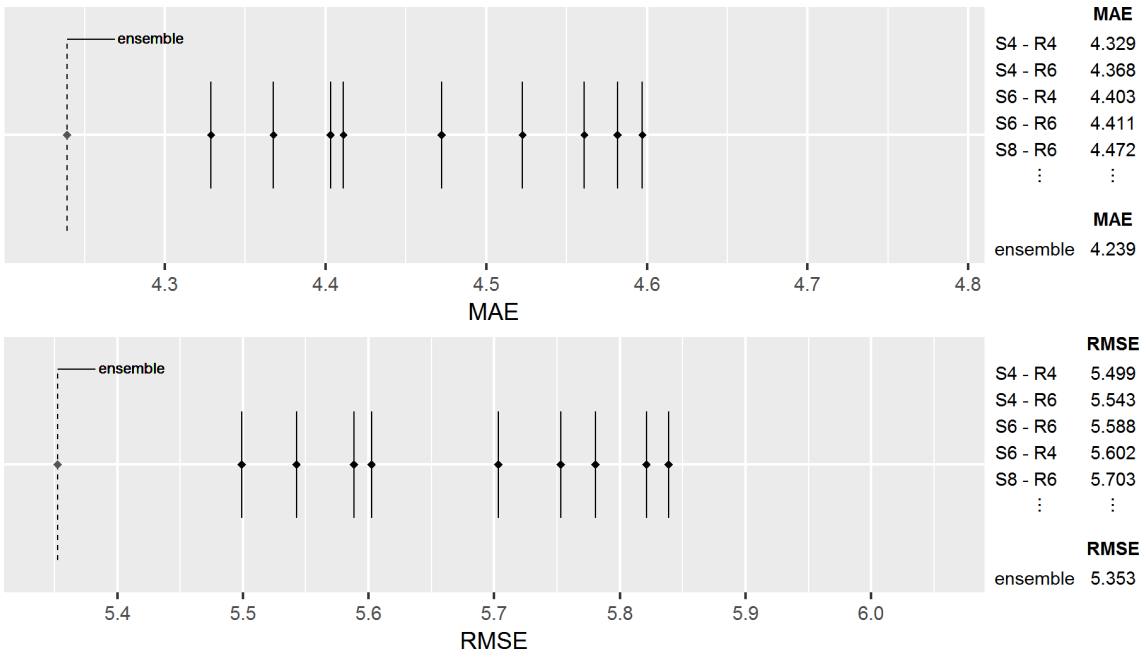

Abbildung 4.5.: MAE (in Jahren) und RMSE (in Jahren) der neun verschiedenen Parameterpaare (4,6,8) und der Ensemblemethode, die diese neun Parameterpaare kombiniert. Jeweils mit Angabe der Werte der fünf besten Parameterpaare und dem Wert der Ensemblemethode

4.3.3. Kombination der drei besten Parameterpaare

Auch die Variante der Ensemblemethode, die jeweils die drei Parameterpaare mit dem geringsten MAE auf den Trainingsdaten kombiniert, konnte die Werte der Einzelpaare übertreffen. Hierbei wurde ein MAE von 4.218 Jahren und ein RMSE von 5.332 Jahren erreicht. So fand bei dieser Variante durch die Ensemblebildung eine Verbesserung des MAE um 0.076 Jahre und eine Verbesserung des RMSE um 0.134 Jahre gegenüber dem besten einzeln verwendeten Parameterpaar statt. Der Vergleich der Ensemblemethode gegenüber den einzelnen Parameterpaaren ist in Abbildung 4.6 dargestellt. An dieser Stelle sei noch einmal erwähnt, dass die Wahl der zu kombinierenden Parameterpaare durch den Vergleich der MAEs auf den Trainingsdaten durchgeführt wurde, um eine Unabhängigkeit der Testdaten zu gewährleisten. Die in Abbildung 4.6 dargestellten Werte für die einzelnen Parameterpaare sind hingegen die am Ende bestimmten Fehler, also die MAEs auf den Testdaten. Die benötigte Zeit zur Erstellung der Datensätze durch Filterung und Resampling betrug auch hier 100 Minuten und 13 Sekunden. Das Training der Ensemblemethode dauerte daraufhin noch 21 Minuten und 9 Sekunden. Diese Variante der Ensemblemethode benötigte also einen Gesamtzeitaufwand von 121 Minuten und 22 Sekunden.

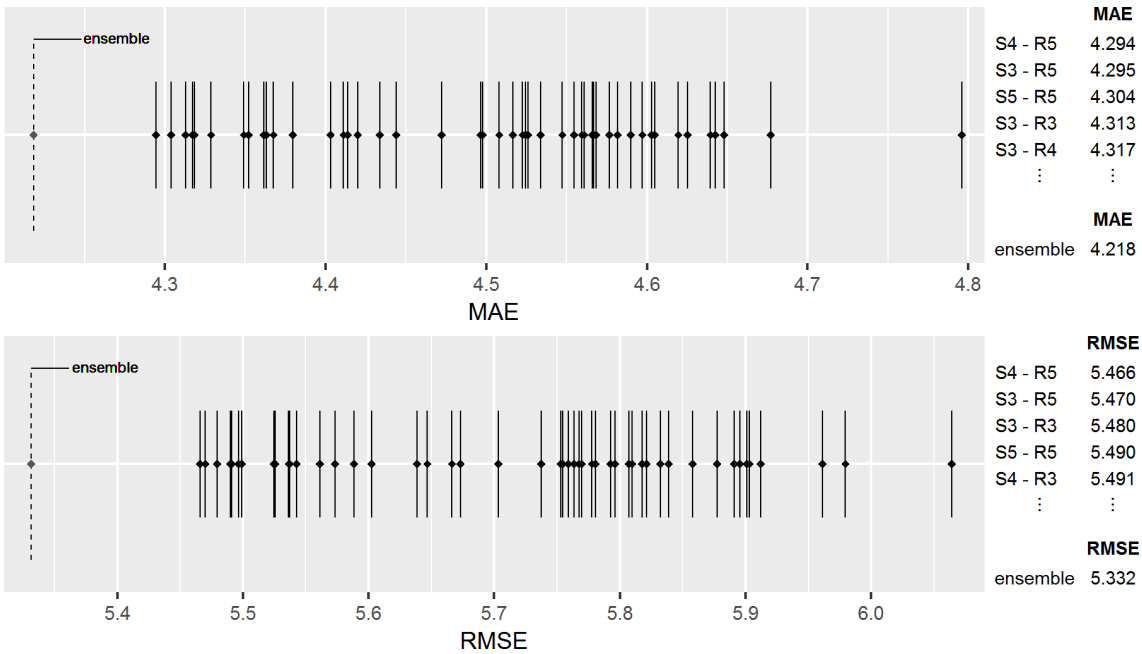

Abbildung 4.6.: MAE (in Jahren) und RMSE (in Jahren) der 49 verschiedenen Parameterpaare und der Ensemblemethode, die jeweils die besten drei Parameterpaare kombiniert. Jeweils mit Angabe der Werte der fünf besten Parameterpaare und dem Wert der Ensemblemethode

5. Diskussion

Die Ergebnisse zeigen, dass sich der MAE des Verfahrens zur Vorhersage des Alters bei verschiedenen Parameterwerten deutlich unterscheidet. Die Genauigkeit der Bestimmung des biologischen Alters des Gehirns hängt also wie erwartet von der Wahl der Parameterwerte ab. Dabei weist sowohl die Filtergröße als auch die Auflösung einen Einfluss auf die Genauigkeit des Verfahrens auf. Denn auch wenn das Minimum und das Maximum der Rastersuche jeweils mit der selben Filtergröße erreicht wurden und damit eine Unabhängigkeit der Genauigkeit gegenüber der Größe des verwendeten Filters nahe liegt, zeigt das geschätzte Modell der Bayesschen Optimierung eine klare Abhängigkeit gegenüber beiden Parametern. Diese Erkenntnis widerspricht der Analyse von Lancaster et al. [20], die die Auflösung als deutlich größeren Einflussfaktor gegenüber der Filtergröße einstufen. Darüber hinaus zeigen die Rastersuche und die Bayessche Optimierung, dass in dem vorliegenden Fall besonders mit Auflösungen unter 6 mm und Filtergrößen unter 5 mm gute Ergebnisse erzielt wurden. Durch das geschätzte Modell der Bayesschen Optimierung wird dabei jedoch auch gleichzeitig deutlich, dass die Verwendung einer kleinen Filtergröße und einer kleinen Auflösung, aufgrund der lokalen Schwankungen in diesem Bereich, nicht automatisch zu einem minimalen Fehler führt. Aus dieser Eigenschaft des Parameterraumes und dem Verlauf des minimalen MAE über die Iterationen der Bayesschen Optimierung lässt sich daher ableiten, dass das Minimum des Fehlers im Parameterraum nicht durch einfache Heuristiken gefunden werden kann, womit die Notwendigkeit der Optimierung dieser Parameter bestätigt wird. Mit den gewonnenen Erkenntnissen über den Parameterraum lassen sich die verwendeten Optimierungsmethoden hinsichtlich ihrer Eignung für die Optimierung von Vorverarbeitungsparametern bei der Bestimmung des biologischen Alters des Gehirns vergleichen.

Das Ergebnis mit dem geringsten Fehler lieferte das Ensembleverfahren, das jeweils die drei besten Parameterpaare kombiniert. Es zeigt sich hier also, dass durch eine Kombination von Parameterpaaren ein geschätztes Optimum von klassischen Optimierungsverfahren übertroffen werden kann. Da die Verbesserung in diesem Fall nur sehr gering ausfiel, stellt sich dabei jedoch die Frage, ob dies auch für das reale Optimum des Parameterraums gilt. Dennoch zeigt sich durch die Verbesserung der

Kombination gegenüber den einzeln verwendeten Parameterpaaren, dass die verwendete Ensemblemethode für den vorliegenden Fall zielführend ist. Ebenso wird dadurch deutlich, dass die Verschiedenheit der einzelnen Methoden im Ensemble grundlegend ausreichend ist und es keiner weiteren Anpassung wie beispielsweise der Aufteilung der Testdaten bedarf. Wie bei der Kombination aller 49 Parameterpaare zu sehen ist, kann die Ensemblemethode jedoch auch zu einer Verschlechterung gegenüber den Einzelergebnissen führen. Hierbei wird der Nachteil dieser Methode zur Parameteroptimierung sichtbar. Denn durch die Kombination liegen die Ergebnisse nicht mehr im regulären Parameterraum, wodurch Vorhersagen über den Erfolg einer bestimmten Kombination nur sehr schwer zu treffen sind. Eine geringe Anzahl zu verwendender Parameter scheint zur Erreichung guter Ergebnisse eine wichtige Voraussetzung zu sein. Jedoch stellt sich an dieser Stelle die Frage, ob dies auch für andere Daten zutrifft.

Die Bayessche Optimierung erbrachte nur geringfügig schlechtere Ergebnisse als die Ensemblemethode, die jeweils die besten drei Parameterpaare kombiniert. Damit zeigt sich, dass sich die Bayessche Optimierung für die Optimierung der Vorverarbeitungsparameter bei der Bestimmung des biologischen Alters des Gehirns eignet. Darüber hinaus wird deutlich, dass die Bayessche Optimierung durch die Erzeugung eines Modells, zusätzlich zu dem Ergebnis, ein erweitertes Verständnis über den Parameterraum liefern kann. Damit sind die durch diese Optimierung erzielten Ergebnisse im Gegensatz zu den Ergebnissen der Ensemblemethode problemlos in den Kontext anderer Parameterpaare einzuordnen und auf Plausibilität zu prüfen. Außerdem ermöglicht das geschätzte Modell für den Parameterraum allgemeine Aussagen über geeignete Wertebereiche für die Parameter zu treffen, wodurch weitere Erkenntnisse für die verwendete Vorverarbeitung generiert werden. Die Tatsache, dass in Iteration 76 noch eine Verbesserung stattgefunden hat, zeigt jedoch auch, dass die Bayessche Optimierung in diesem Fall viele Iterationen benötigt. Hierbei spielt vermutlich die Komplexität der realen Funktion im Parameterraum eine Rolle, deren lokale Schwankungen zu Schwierigkeiten bei der Bestimmung des Optimums führen. Das heißt, auch wenn durch die Bayessche Optimierung gute Ergebnisse erzielt wurden, wird durch die Analyse auch sichtbar, dass dieses Verfahren von vielen Einstellungen, wie beispielsweise der Wahl der *acquisition function*, abhängt. Nach der Anwendung bleibt also die Frage bestehen, ob durch eine andere Implementierung eine effizientere und bessere Lösung erreicht werden könnte.

Bei der Rastersuche bedarf es im Gegensatz zur Bayesschen Optimierung, außer der Wahl der zu verwendenden Parameterpaare, keiner weiteren Einstellungen. Die erlangten Erkenntnisse über den Parameterraum zeigen jedoch, dass hierbei die

Erreichung eines guten Ergebnisses stark vom Zufall abhängt. Durch die lokalen Schwankungen ist ein umfassender Überblick über den Parameterraum nicht durch eine uniforme Verteilung von zu beobachtenden Parameterpaaren erreichbar. Dies wird ebenso durch die Tatsache sichtbar, dass die Rastersuche das schlechteste Ergebnis der drei verwendeten Methoden aufwies. Durch den Vergleich der Verfahren wird darüber hinaus ein weiteres Mal der Einfluss der lokalen Schwankungen im Parameterraum deutlich. Denn auch wenn sich das Parameterpaar des bestimmten Optimums der Bayesschen Optimierung nur marginal von einem bei der Rastersuche verwendeten Parameterpaar unterscheidet, erzielte die Bayessche Optimierung ein deutlich besseres Ergebnis als die Rastersuche.

Die Betrachtung der Berechnungszeit zeigt, dass die Bayessche Optimierung die meiste Zeit und die Ensemblemethode, die neun Parameterpaare kombiniert, die geringste Zeit benötigte. Hierbei wird deutlich, dass bei der Optimierung der Vorverarbeitungsparameter vor allem die Filterung und das Resampling zeitintensive Schritte sind, die die Berechnungszeit der Methode beeinflussen. Dominiert wird die benötigte Zeit für ein Verfahren also davon, wie viele Parameterpaare in der Methode verwendet werden. Wichtig ist dabei, dass sowohl bei der Rastersuche als auch bei der Ensemblemethode die verwendeten Parameterpaare im voraus festgelegt werden und so die Erstellung der Datensätze durch die Filterung und das Resampling unabhängig von dem Optimierungsverfahren durchgeführt werden kann. So können diese Datensätze ohne eine Neuberechnung für verschiedene Fragestellungen verwendet werden, wodurch der zeitliche Aufwand deutlich vermindert wird.

Insgesamt zeigt die vorliegende Arbeit, dass die Optimierung der Vorverarbeitungsparameter einen großen Einfluss auf die Genauigkeit der Bestimmung des biologischen Alters des Gehirns hat. Diese Optimierung ist dabei nicht durch Heuristiken durchzuführen und sollte vom Fall zu Fall separat stattfinden. Die vorgestellten Verfahren weisen unterschiedliche Vor- und Nachteile auf, jedoch ist das zuverlässigste Ergebnis durch eine Bayessche Optimierung zu erreichen. Auch wenn diese Herangehensweise eine große Berechnungszeit aufweist, überwiegt der Vorteil der Methode, neben der robusten Bestimmung eines Optimums, auch Informationen über den Parameterraum zu liefern. Darüber hinaus ist jedoch auch hervorzuheben, dass die Ergebnisse dieser Arbeit eine Bestätigung der Einsetzbarkeit von Ensemblemethoden für die Optimierung von Vorverarbeitungsparametern liefern. Auch wenn durch die Analyse von lediglich drei verschiedenen Varianten keine Aussagen über die allgemeine Robustheit und Einsatzfähigkeit für andere Datensätze getroffen werden kann, bietet diese Methode hohes Potential durch weitere Untersuchungen ein

einsatzfähiges Verfahren zu bilden. Bei weiteren Untersuchungen ist dabei vor allem die Anwendung auf andere Datensätze wichtig, da nur so die Robustheit des Verfahrens überprüft werden kann, die für klinische Anwendungen entscheidend ist.

Literaturverzeichnis

[1] J. Ashburner. „A fast diffeomorphic image registration algorithm“. In: *NeuroImage* 38.1 (2007), S. 95–113.

[2] J. Ashburner. „Computational anatomy with the SPM software“. In: *Magnetic resonance imaging* 27.8 (2009), S. 1163–1174.

[3] J. Ashburner und K. J. Friston. „Unified segmentation“. In: *NeuroImage* 26.3 (2005), S. 839–851.

[4] E. Brochu, V. M. Cora und N. de Freitas. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. 2010.

[5] G. Brown, J. L. Wyatt und P. Tiňo. „Managing Diversity in Regression Ensembles“. In: *Journal of Machine Learning Research* 6.Sep (2005), S. 1621–1650.

[6] J. H. Cole, S. J. Ritchie, M. E. Bastin, M. C. V. Hernández, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, N. R. Wray, P. Redmond, R. E. Marioni, J. M. Starr, S. R. Cox, J. M. Wardlaw, D. J. Sharp und I. J. Deary. „Brain age predicts mortality“. In: *Molecular Psychiatry* 23.5 (2018), S. 1385–1392.

[7] J. H. Cole und K. Franke. „Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers“. In: *Trends in Neurosciences* 40.12 (2017), S. 681–690.

[8] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector und G. Montana. „Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker“. In: *NeuroImage* 163 (2017), S. 115–124.

[9] T. G. Dietterich. „Ensemble Learning“. In: *The Handbook of Brain Theory and Neural Networks*. Hrsg. von M. A. Arbib. 2. Aufl. MIT Press, 2003, S. 110–125.

[10] K. Franke, G. Ziegler, S. Klöppel und C. Gaser. „Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the

influence of various parameters“. In: *NeuroImage* 50.3 (2010), S. 883–892.

[11] C. Gaser. „Partial Volume Segmentation with Adaptive Maximum A Posteriori (MAP) Approach“. In: *NeuroImage* 47 (2009).

[12] Y. Ge, R. I. Grossman, J. S. Babb, M. L. Rabin, L. J. Mannon und D. L. Kolson. „Age-Related Total Gray Matter and White Matter Changes in Normal Adult Brain. Part I: Volumetric MR Imaging Analysis“. In: *American Journal of Neuroradiology* 23.8 (2002), S. 1327–1333.

[13] S. Geman, E. Bienenstock und R. Doursat. „Neural Networks and the Bias/Variance Dilemma“. In: *Neural Computation* 4.1 (1992), S. 1–58.

[14] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston und R. S. Frackowiak. „A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains“. In: *NeuroImage* 14.1 (2001), S. 21–36.

[15] D. R. Jones, M. Schonlau und W. J. Welch. „Efficient Global Optimization of Expensive Black-Box Functions“. In: *Journal of Global Optimization* 13.4 (1998), S. 455–492.

[16] Katja Franke und Christian Gaser. „Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease“. In: *Geropsych: The Journal of Gerontopsychology and Geriatric Psychiatry* 25.4 (2012).

[17] V. D. Köchli und B. Marincek. *Wie funktioniert MRI?: Eine Einführung in Physik und Funktionsweise der Magnetresonanzbildgebung*. Springer Berlin Heidelberg, 2013.

[18] R. Kramme. *Medizintechnik: Verfahren - Systeme - Informationsverarbeitung*. Springer Berlin Heidelberg, 2016.

[19] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley, 2004.

[20] J. Lancaster, R. Lorenz, R. Leech und J. H. Cole. „Bayesian Optimization for Neuroimaging Pre-processing in Brain Age Classification and Prediction“. In: *Frontiers in Aging Neuroscience* 10 (2018), S. 28.

[21] J. Mendes-Moreira, C. Soares, A. M. Jorge und J. F. D. Sousa. „Ensemble approaches for regression: A survey“. In: *ACM Computing Surveys (CSUR)* 45.1 (2012), S. 10.

[22] I. Nenadić, M. Dietzek, K. Langbein, H. Sauer und C. Gaser. „BrainAGE score indicates accelerated brain aging in schizophrenia, but not bipolar disorder“. In: *Psychiatry Research: Neuroimaging* 266 (2017), S. 86–89.

[23] D. C. Park und N. Schwarz, Hrsg. *Cognitive aging. A primer*. Philadelphia, PA: Psychology Press, 2000.

[24] R. Peters. „Ageing and the brain“. In: *Postgraduate Medical Journal* 82.964 (2006), S. 84–88.

[25] A. Pfefferbaum, D. H. Mathalon, E. V. Sullivan, J. M. Rawles, R. B. Zipursky und K. O. Lim. „A Quantitative Magnetic Resonance Imaging Study of Changes in Brain Morphology From Infancy to Late Adulthood“. In: *Archives of Neurology* 51.9 (1994), S. 874–887.

[26] C. E. Rasmussen und C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. Cambridge, Massachusetts und London, England: The MIT Press, 2006.

[27] B. Schölkopf, B. Smola, A. J. Smola, M. Scholkopf und F. Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[28] M. Schünke, E. Schulte, U. Schumacher, M. Voll und K. Wesker. *Prometheus Lernatlas - Kopf, Hals und Neuroanatomie*. Stuttgart: Thieme, 2018.

[29] J. Snoek, H. Larochelle und R. P. Adams. „Practical Bayesian Optimization of Machine Learning Algorithms“. In: *Advances in Neural Information Processing Systems*. Neural Information Processing Systems 2012. 2012, S. 2951–2959.

[30] D. Terribilli, M. S. Schaufelberger, F. L. Duran, M. V. Zanetti, P. K. Curiati, P. R. Menezes, M. Scazufca, E. Amaro, C. C. Leite und G. F. Busatto. „Age-related gray matter volume changes in the brain during non-elderly adulthood“. In: *Neurobiology of Aging* 32.2 (2011), S. 354–368.

[31] M. E. Tipping. „Sparse Bayesian Learning and the Relevance Vector Machine“. In: *Journal of Machine Learning Research* 1.Jun (2001), S. 211–244.

[32] J. Tohka, A. Zijdenbos und A. Evans. „Fast and robust parameter estimation for statistical partial volume models in brain MRI“. In: *NeuroImage* 23.1 (2004), S. 84–97.

[33] N. Ueda und R. Nakano. „Generalization error of ensemble estimators“. In: *The 1996 IEEE International Conference on Neural Networks*. International Conference on Neural Networks (ICNN'96) (Washington, DC, USA). New York und Piscataway, N.J: Institute of Electrical and Electronics Engineers, 1996, S. 90–95.

[34] T. Vos et al. „Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden

of Disease Study 2010“. In: *The Lancet* 380.9859 (2012), S. 2163–2196.

Online-Ressourcen

[35] *Abteilung für Hochfeld-Magnetresonanz*. Max-Planck-Institut für biologische Kybernetik. URL: https://www.kyb.tuebingen.mpg.de/hochfeld-magnetresonanz (besucht am 09. 08. 2019).

[36] *CAT - A Comptational Anatomy Toolbox for SPM*. Christian Gaser, Structural Brain Mapping Group, Universität Jena. URL: http://dbm.neuro.uni-jena.de/cat/ (besucht am 28. 08. 2019).

[37] *Create cross-validation partition for data - MATLAB*. MathWorks. URL: https://de.mathworks.com/help/stats/cvpartition.html (besucht am 27. 08. 2019).

[38] R. Garnett. *Bayesian Methods in Machine Learning – Spring 2019. Lecture notes - 06.03.2019*. Washington University in St. Louis. URL: https://www.cse.wustl.edu/~garnett/cse515t/spring_2019/files/lecture_notes/12.pdf (besucht am 16. 09. 2019).

[39] *IXI Dataset – Brain Development*. Imperial College London. URL: https://brain-development.org/ixi-dataset/ (besucht am 27. 08. 2019).

[40] *SPM - Statistical Parametric Mapping*. Wellcome Centre for Human Neuroimaging, Institute of Neurology London. URL: https://www.fil.ion.ucl.ac.uk/spm/ (besucht am 28. 08. 2019).

[41] *Statistics and Machine Learning Toolbox*. MathWorks. URL: https://de.mathworks.com/products/statistics.html (besucht am 28. 08. 2019).

[42] *The Spider*. Empirical Inference Department, Max Planck Institute for Intelligent Systems, Tübingen. URL: https://people.kyb.tuebingen.mpg.de/spider/main.html (besucht am 28. 08. 2019).

[43] *Universitätsklinik für Neurologie - Magnetresonanztomographie*. Universitätsklinik Magdeburg. 2018. URL: http://www.kneu.ovgu.de/Forschung/Forschungsgruppen/Arbeitsgruppen/Magnetresonanztomographie.html (besucht am 09. 08. 2019).

Abbildungsverzeichnis

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Seitens des Verfassers bestehen keine Einwände die vorliegende Bachelorarbeit für die öffentliche Benutzung im Universitätsarchiv zur Verfügung zu stellen.

Robin Witte
Jena, 27. September 2019