Diplomarbeit

Künstliche Intelligenz in der Kieferorthopädie
- eine systematische Literaturrecherche

eingereicht von

Laura Daria Voss

zur Erlangung des akademischen Grades

Doktorin der Zahnmedizin
(Dr^{in}. med. dent.)

an der

Medizinischen Universität Graz

ausgeführt an der

Universitätsklinik für Zahnmedizin und Mundgesundheit
Klinische Abteilung für Orale Chirurgie und Kieferorthopädie

unter der Anleitung von

Prof. Univ. ZÄ Priv.-Doz. Dr. Brigitte Wendl

Ass.-Prof. Dr.med.univ. Margit Pichelmayer

Graz, 07.08.2024

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am 07.08.2024 Laura Daria Voss eh.

Danksagungen

Ich möchte meiner Betreuerin, Frau Professorin Wendl, meinen aufrichtigen Dank aussprechen. Trotz der schwierigen Umstände führte sie meine Betreuung weiterhin bereitwillig fort und antwortete stets unterstützend und geduldig auf Fragen aus dem Krankenstand heraus. Dies ist in solch einer Situation alles andere als selbstverständlich und ich weiß es sehr zu schätzen!

Ebenso möchte ich Frau Professorin Pichelmayer meinen Dank aussprechen, die spontan die Zweitbetreuung meiner Diplomarbeit übernahm und mich sofort sehr hilfsbereit bei der Fertigstellung unterstützte.

Vielen Dank!

Kurzfassung

Einleitung

Die künstliche Intelligenz (KI) hat in den letzten Jahren erhebliche Fortschritte gemacht und findet zunehmend Anwendung in der allgemeinen Medizin und der Zahnmedizin. Durch KI könnten diagnostische und therapeutische Verfahren in der Kieferorthopädie beschleunigt und verbessert werden. Diese Übersichtsarbeit beschäftigt sich mit aktuellen Forschungsergebnissen zu KI-Anwendungen in der Kieferorthopädie und untersucht deren zukünftige Möglichkeiten.

Material und Methoden

Eine systematische Literaturrecherche wurde nach den PRISMA-Richtlinien durchgeführt, um die jüngsten Fortschritte in der Anwendung von KI in der Kieferorthopädie zu bewerten. Der Fokus lag auf der Nutzung von maschinellem Lernen und künstlichen neuronalen Netzwerken zur Analyse und Behandlung von kieferorthopädischen Pathologien.

Ergebnisse/Diskussion

Die Ergebnisse zeigen, dass KI-gestützte Systeme mit hoher Genauigkeit diagnostische Aufgaben, wie die Identifizierung anatomischer Landmarken in Röntgenbildern, die Prognose des kraniofazialen Wachstums oder die Erstellung individueller Behandlungspläne übernehmen können. Verschiedene KI-Modelle haben eine relativ hohe Genauigkeit bei der Vorhersage der Behandlungsdauer, des Wirbelkörperreifegrades, der postoperativen Gesichtsmorphologie, des chirurgischen Bedarfs oder von Extraktionsentscheidungen gezeigt. Jedoch bestehen weiterhin Herausforderungen hinsichtlich der Datenqualität, Modellvalidierung und klinischer Integration. Die Forschung unterstreicht die Notwendigkeit großer und vielfältiger Datensätze für eine robuste KI-Entwicklung.

Konklusion

KI bietet ein vielversprechendes Potenzial zur Unterstützung der kieferorthopädischen Behandlung. Trotz der bisherigen Fortschritte erfordert die vollständige Integration von KI in die klinische Praxis weitere Forschung und kontinuierliche Anpassungen der Algorithmen, sowie weiterhin die Expertise menschlicher Fachleute.

Abstract

Introduction

Artificial intelligence (AI) has made significant progress in recent years and is increasingly being applied in general medicine and dentistry. AI has the potential to significantly improve diagnostic and therapeutic processes through machine learning and artificial neural networks. This review paper addresses current research findings on AI applications in orthodontics and explores their future possibilities.

Materials and Methods

A systematic literature review was conducted according to PRISMA guidelines to assess recent advances in the application of AI in orthodontics. The focus was on the use of machine learning and artificial neural networks for the analysis and treatment of orthodontic pathologies.

Results/Discussion

The results show that AI-supported systems can perform diagnostic tasks with high accuracy, such as identifying anatomical landmarks in X-rays, predicting craniofacial growth, or influencing the creation of individual treatment plans. Various AI models have shown relatively high accuracy in predicting treatment duration and identifying surgical needs. However, challenges remain regarding data quality, model validation, and clinical integration. Research emphasizes the need for large and diverse datasets for robust AI development.

Conclusion

AI offers promising potential to support orthodontic treatment. Despite the progress made so far, the complete integration of AI into clinical practice requires further research and continuous adaptation of algorithms, as well as the continued expertise of human professionals.

Inhaltsverzeichnis

Abkürzungen

KI	Künstliche Intelligenz
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
CVM	Cervical Vertebral Maturation
SCR	Successful Detection Rate
SDR	Successful Classification Rate
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic Curve
MRE	Mittlerer Radialer Fehler
ICC	Intraclass Correlation Coefficient
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ROI	Region of Interest
Grad-CAM	Gradient-weighted Class Activation Mapping
DVT	Digitale Volumentomographie
YOLO	You Only Look Once
YOLOv3	You Only Look Once version 3
SVM	Support Vector Machine
LR	Logistische Regression
MLP	Multi-Layer Perceptron
MTM	Massen-Tensor-Modell
CT	Computertomographie
ML	Maschinelles Lernen
n.a.	nicht angegeben
DMFR	Dentomaxillofacial Radiology
ResNet	Residual Neural Network
DL	Deep Learning

Abbildungsverzeichnis

1 Einleitung

Das Forschungsgebiet der Künstlichen Intelligenz (KI) erlangt zunehmend an Bedeutung in der aktuellen Berichterstattung und gewinnt an Relevanz im Alltag der Gesellschaft. Die Fortschritte in diesem Themenfeld sind rasant und haben eine Vielzahl von Anwendungen hervorgebracht, darunter Internetsuche, Empfehlungssysteme für Waren und Dienstleistungen, Bild- und Spracherkennungstechnologien oder kognitive Entscheidungsunterstützungssysteme (1). Prominente Beispiele, wie Siri, Alexa und ChatGPT sind global bekannt und derzeit häufig gebrauchte Anwendungen in der Bevölkerung.

Auch in der Medizin steigt der Einfluss von KI-Anwendungen.

Mit der KI-Anwendung sollen zahlreiche Möglichkeiten zur Verbesserung der Patient*innenversorgung und Forschung geboten werden (2). KI-gestützte Systeme könnten durch die schnelle Datenanalyse und das Erkennen komplexer Zusammenhänge neue medizinische Erkenntnisse und Behandlungsansätze entwickeln, Diagnosen schneller stellen und Erkrankungen früher erkennen (3). Diese Innovationen haben das Potential, die Art und Weise, wie wir Krankheiten behandeln und Gesundheitsversorgungen betreiben, grundlegend zu verändern.

KI-gestützte Systeme können in verschiedenen medizinischen Bereichen eingesetzt werden, wie zum Beispiel in der Radiologie, Endoskopie, Chirurgie oder der Telemedizin (4,5).

Sie können Aufgaben übernehmen, um Ärzt*innen mehr Zeit für die direkte Patient*innenversorgung zu gewähren und können auch als Assistenzsysteme dienen, die die Präzision bei Operationen erhöhen und Komplikationen verringern (6 – 8). Durch die Integration von KI in die medizinische Praxis könnten Mediziner*innen effizienter arbeiten, Zeit sparen, Kosten reduzieren und bessere Ergebnisse für die Patient*innen erzielen (9).

Obwohl der Einsatz von KI in der Medizin vielversprechende Möglichkeiten bietet, bringt er auch einige Herausforderungen mit sich. Dazu gehören Bedenken hinsichtlich der Datensicherheit, der Eingliederung von KI-Systemen in bereits bestehende medizinische

Abläufe und der Akzeptanz seitens der Ärzt*innen und Patient*innen. Außerdem sollte die Fairness solcher KI-Systeme nicht außer Acht gelassen werden. KI-generierte Vorurteile können zum Beispiel durch die Art der Datenrepräsentation entstehen oder durch eine ungleiche Behandlung verschiedener Bevölkerungsgruppen (3, 10, 11).

Die Hauptanwendungsbereiche in der Zahnmedizin für KI umfassen Diagnostik, Behandlungsplanung und Vorhersage der Behandlungsergebnisse.

In der Diagnostik unterstützen KI-Systeme Zahnmediziner*innen bei der Erkennung von Parodontitis, Karies, zystischen Läsionen oder anderen oralen Erkrankungen, sowie in der Planung und Durchführung von prothetischen oder implantologischen Behandlungen. Dies geschieht durch die Analyse von Röntgenbildern, CT-Scans, 3D-Scans, digitale Volumentomographie (DVT) oder anderen bildgebenden Verfahren.

In der restaurativen Zahnmedizin können neuronale Netzwerke effektiv Karies oder Restaurationen identifizieren und die Auswahl der Exkavationsmethode vereinfachen. In der Endodontie bieten KI-Systeme einen Mehrwert, indem sie periapikale Läsionen und Wurzelfrakturen zuverlässig erkennen können. Auch in der oralen Chirurgie erweisen sich neuronale Netzwerke als hilfreich. Beispielsweise bei der Planung von chirurgischen Eingriffen, der Vorhersage postoperativer Komplikationen, der Detektion von Knochenläsionen oder der Planung von Implantatbehandlungen. Darüber hinaus ermöglicht KI die automatisierte Analyse von Patient*innendaten, um individuelle Risikoprofile zu erstellen und personalisierte Behandlungspläne zu entwickeln (12 – 17).

1.1 Künstliche Intelligenz

Der Begriff „künstliche Intelligenz“ oder artifizielle Intelligenz (AI) wurde von John McCarthy im Jahr 1955 geprägt. Er führte diesen Begriff ein, um die Fähigkeit von Maschinen zu beschreiben, Aufgaben auszuführen, die als intelligent betrachtet werden. Mit anderen Worten zielt die künstliche Intelligenz darauf ab, dass eine Maschine durch Daten lernen und Probleme selbstständig lösen kann (18 – 20).

Die Dartmouth-Konferenz im Jahr 1956 gilt als ein wegweisendes Ereignis und Geburtsstunde in der Geschichte der KI. Unter der Leitung von John McCarthy kamen führende Forscher zusammen, um Konzepte und Ideen zu dem Themenfeld „Künstliche

Intelligenz“ zu sammeln. Eine bemerkenswerte Diskussion drehte sich um den „Turing-Test“, der von Alan Turing konzipiert wurde. Der Turing-Test ist ein Verfahren, das die Fähigkeit einer Maschine testet, intelligentes Verhalten zu zeigen, das von dem eines Menschen nicht zu unterscheiden ist (21).

In den 1960er und 1970er Jahren erlebte die KI-Forschung einen ersten Höhepunkt, als Wissenschaftler begannen, sich mit komplexeren Problemen, wie der Spracherkennung und der Spieltheorie zu befassen. Die Entwicklung von Expertensystemen, die auf logischen Regeln basierten, war ein wichtiger Meilenstein in dieser Zeit (22).

In den 1980er Jahren kam es zu einem Rückgang des Interesses an KI, da die technologischen Fortschritte nicht den Erwartungen entsprachen und viele Probleme ungelöst blieben. Dies änderte sich jedoch in den 1990er Jahren mit dem Aufkommen des Internets und dem Zugang zu großen Datenmengen (Big Data), die die Entwicklung von datengetriebenen KI-Systemen ermöglichten (22 – 24).

Seitdem hat die KI-Forschung exponentiell zugenommen, angetrieben durch Fortschritte in den Bereichen Maschinelles Lernen (ML) und Deep Learning (DL, tiefes Lernen). Dabei können KI-Systeme aus großen Datenmengen lernen. Heutzutage sind KI-Anwendungen in nahezu allen Bereichen des täglichen Lebens zu finden (24, 25).

Um die Funktionsweise von KI und ihren Einfluss auf die Kieferorthopädie zu untersuchen, ist es zunächst wichtig, einige Schlüsselbegriffe im Zusammenhang mit KI zu unterscheiden (Abbildung 1).

- Artifizielle Intelligenz (Künstliche Intelligenz, KI) beschreibt die Fähigkeit von Systemen, eigenständige Intelligenz zu entwickeln und zielt darauf ab, dass durch das Lernen mit Daten Probleme eigenständig gelöst werden können (104)
- Maschinelles Lernen (ML) bildet das Kernstück von KI und basiert auf Algorithmen, die Ergebnisse basierend auf Datensätzen vorhersagen können.

  Ziel ist es, Maschinen das Lernen aus Daten zu ermöglichen, ohne dafür explizit programmiert worden zu sein (26)

- Tiefes Lernen ist ein Bestandteil von ML und nutzt algorithmische Netzwerke mit verschiedenen Schichten in tiefen neuronalen Netzen, um Eingabedaten zu analysieren. Ziel ist es, ein neuronales Netzwerk auszubauen, das automatisch Muster erkennen kann, um die Merkmalsdetektion zu verbessern (8, 26)
- Künstliche neuronale Netzwerke, sogenannte ANNs und CNNs, sind eine Reihe von Algorithmen, die Signale mithilfe künstlicher Neuronen verarbeiten und dabei versuchen, die Funktionsweise menschlicher Neurone nachzuahmen (26)

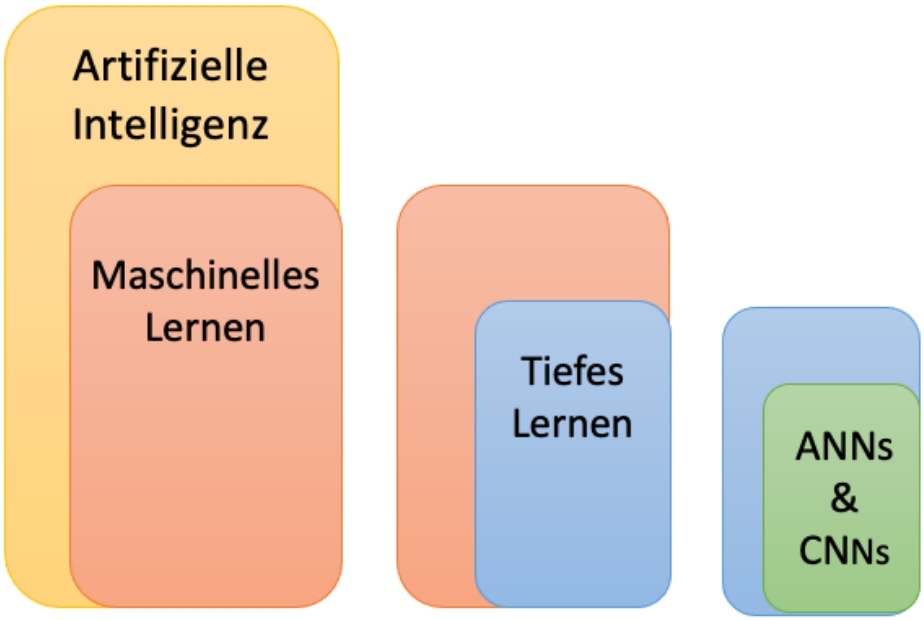


Abbildung 1: Artifizielle Intelligenz (eigene Abbildung)

1.1.1 Maschinelles Lernen

Eine klassische Programmierung besteht darin, Algorithmen für eine bestimmte Aufgabe zu entwickeln. Algorithmen stellen eine systematische Abfolge von Anweisungen dar, um ein bestimmtes Problem zu lösen. Einfach erklärt, könnte man sie sich als Gebrauchsanweisungen vorstellen.

Maschinelles Lernen hingegen beschäftigt sich mit der Entwicklung von Algorithmen, durch die Computer automatisiert aus Daten lernen können, *ohne* explizit dafür programmiert zu werden. Der Kerngedanke des maschinellen Lernens besteht darin, dass

Maschinen Muster und Strukturen in Daten erkennen und darauf basierend Vorhersagen oder Entscheidungen treffen können.

Die zwei grundlegenden Strategien des maschinellen Lernens sind *überwachtes* Lernen und *unüberwachtes* Lernen:

Überwachtes Lernen beinhaltet das Training von Modellen mithilfe von „gelabelten" Daten (Daten die bestimmten Kategorien zugeordnet sind), bei denen sowohl die Eingabe (Aufnahme von Informationsmaterial) als auch die entsprechende Ausgabe (eine durch die KI getätigte Vorhersage) bekannt sind. Das Ziel ist es, aus den vorhandenen Daten Beziehungen zu lernen, um Vorhersagen für neue, „ungelabelte" Daten treffen zu können. Ein Algorithmus wird mit Trainingsdaten gefüttert, die aus Eingabe- und Zielvariablen bestehen. Der Algorithmus lernt, indem er Muster und Beziehungen in den Daten identifiziert und übernimmt dabei Klassifizierungsaufgaben.

Ein einfaches Beispiel dafür wäre ein Pool aus Röntgenbildern von Zähnen. Dabei sind einige Zähne gesund und andere weisen kariöse Läsionen auf. Im Anschluss werden dem Computer die Röntgenbilder präsentiert und mit „kariös" oder „nicht kariös" „gelabelt". Der Computer analysiert die Röntgenbilder und filtert wichtige Merkmale zur Unterscheidung der beiden Zahnzustände heraus. Nach diesem Lernprozess ist das System in der Lage auch bei unbekannten Röntgenaufnahmen zwischen gesunden und erkrankten Zähnen zu unterscheiden, obwohl der Algorithmus nicht explizit dafür programmiert wurde.

Im Gegensatz dazu bezieht sich *unüberwachtes* Lernen auf das Trainieren von Modellen ohne Verwendung von „gelabelten" Daten. Stattdessen zielt unüberwachtes Lernen darauf ab, Muster oder Strukturen in den Daten zu identifizieren, *ohne* dass die entsprechenden Ausgaben bekannt sind. Dies ermöglicht es, verborgene Beziehungen oder Gruppierungen in den Daten zu entdecken, die möglicherweise nicht offensichtlich sind. Ein Bespiel dafür könnte die Segmentierung von Röntgenbildern sein. Ein Modell könnte darauf trainiert

werden, verschiedene Strukturen in dentalen Bildern zu identifizieren, ohne dass die spezifischen Merkmale vorher markiert werden müssen.

Zusätzlich existiert das verstärkende Lernen, bei dem die Reaktion auf Belohnung oder Bestrafung als Feedback verwendet wird, um das Verhalten des KI-Modells zu formen und anzupassen. (24, 26 – 28)

Die gängigsten *Lernalgorithmen,* die in der Medizin verwendet werden, sind die lineare Regression, logistische Regression, Entscheidungsbäume und Random Forests.

- *Lineare Regression* ist ein einfacher Algorithmus zur Vorhersage, der eine Beziehung zwischen Merkmalen und einem einzigen Ziel beschreibt. Ein Beispiel aus der Kieferorthopädie könnte die Vorhersage der Behandlungsdauer unter Einbeziehung verschiedener Faktoren wie dem Schweregrad der Zahnfehlstellung, Geschlecht, Alter und Compliance der Patient*innen sein.
- *Logistische Regression* ist ein Klassifikationsalgorithmus, der die Wahrscheinlichkeit einer bestimmten Ergebnisklasse basierend auf den Merkmalen schätzt. Ein Bespiel dafür könnte die Vorhersage darstellen, ob ein*e Patient*in die Kriterien für eine Kassenzahnspange erfüllt.
- *Entscheidungsbäume* sind überwachte Lernalgorithmen, die vor allem für Klassifizierungsaufgaben verwendet werden. Sie beginnen mit einem Wurzelknoten, der den ersten Entscheidungspunkt für die Aufteilung eines Datensatzes darstellt und enthält ein einzelnes Merkmal, das den Datensatz in verschiedene Klassen aufteilt. Jede Aufteilung kann zu einem neuen Entscheidungsknoten führen, der ein weiteres Merkmal enthält oder zu einem Terminalknoten, der die Klasse vorhersagt. Ein kieferorthopädisches Beispiel dafür wäre die Frage nach der Art der Behandlung. Der Baum könnte mit der Frage beginnen, ob die Zahnfehlstellung schwerwiegend ist und basierend auf der Antwort weitere Fragen stellen, um zu der optimalen Behandlungsstrategie zu gelangen. Diese könnte eine Alignertherapie, chirurgische Behandlungsoptionen oder den Gebrauch einer klassischen festsitzenden Zahnspange beinhalten.

- Eine Erweiterung der Entscheidungsbäume sind *Random Forests*, die mehrere Entscheidungsbäume erstellen und eine Mehrheitsabstimmung verwenden, um die endgültige Klassenprognose zu treffen (26, 29).

1.1.2 Tiefes Lernen, Künstliche Neuronale Netzwerke und Faltungsneuronale Netzwerke

Tiefes Lernen (Deep Learning) ist eine Kathegorie des maschinellen Lernens, die auf die Verwendung von künstlichen neuronalen Netzwerken mit vielen Schichten, sogenannten tiefen Netzwerken aufbaut. Diese Algorithmen-Netzwerke ahmen die Struktur von Gehirnzellen nach, die jeweils verschiedene Aspekte der Daten verarbeiten, und die miteinander verknüpft sind. Ziel ist es dabei komplexe Zusammenhänge in Daten zu erkennen und aus ihnen zu lernen. Im Gegensatz zu flacheren Netzwerkarchitekturen können tiefe neuronale Netze abstraktere Merkmale aus den Daten extrahieren. Die Abgrenzung zum Maschinellen Lernen zeichnet sich durch die höhere Komplexität aus (24).

Man unterscheidet zwei Typen von neuronalen Netzen, ANNs (Artificial Neural Networks) und CNNs (Convolutional Neural Networks):

- *ANNs* bestehen aus einer Reihe miteinander verbundenen künstlichen Neuronen, die in Schichten angeordnet sind. Ein ANN besteht typischerweise aus drei Arten von Schichten: der Eingabeschicht, der Ausgabeschicht und einer oder mehreren versteckten Schichten dazwischen. Diese Netzwerke können komplexe Muster in Daten erkennen oder Klassifizierungen vornehmen, je nach ihrer Architektur und dem Zweck, für den sie trainiert werden (26, 30). Ein Beispiel für die Funktion könnten Röntgenbilder von Zähnen darstellen, die als Eingabedaten fungieren. Jedes Bild hat eine Größe von 32x32 Pixel und ist mit „kariös“ oder „nicht kariös“ „gelabelt“. Die Eingabeschicht müsste demnach 1024 Neuronen aufweisen, eines für jeden Pixel. Die Ausgabeschicht würde in diesem Beispiel zwei Neuronen beinhalten, eines für „kariös“ und eines für „nicht kariös“. Zwischen diesen beiden Schichten liegen die versteckten Schichten und die Anzahl dieser wird durch die

Komplexität der Aufgabe bestimmt. In den versteckten Schichten findet der eigentliche Entscheidungsprozess statt. Die erste Schicht könnte in diesem fiktiven Beispiel 128 Neuronen beinhalten, die zweite 64. Die Eingabeneuronen leiten ihre Bildwerte an die erste versteckte Schicht weiter, dort werden vor allem Kanten und Linien der Strukturen auf den Röntgenbildern erkannt. Im Anschluss findet eine Weiterleitung der Daten an die zweite versteckte Schicht statt. Hier werden komplexere Formen und Muster gelernt und ihre Neuronen leiten die Signale an die Ausgabeschicht weiter. Dort werden die Wahrscheinlichkeiten für die beiden Fälle „kariös“ und „nicht kariös“ berechnet und eine Diagnose gestellt. Das System ist durch wiederholtes Training in der Lage die Gewichtung verschiedener Bildparameter für die Diagnosestellung so anzupassen, dass die Vorhersage immer genauer wird.

- *CNNs* (Convolutional Neural Networks, Faltungsneuronale Netzwerke) sind eine spezielle Art von künstlichen neuronalen Netzwerken, die besonders gut für die Verarbeitung von Bildern geeignet sind. Auch sie funktionieren, indem sie spezifische Merkmale aus Bildern extrahieren und diese Merkmale zur Klassifizierung oder Erkennung von Objekten verwenden. Wie herkömmliche ANNs besitzen CNNs eine Eingabeschicht und eine Ausgabeschicht. Im Gegensatz zu herkömmlichen ANNs nutzen CNNs zusätzliche spezielle Schichten:

  Diesen sind *Faltungsschichten*, *Pooling-Schichten* und *voll verbundene Schichten* zwischengeschaltet.

  Zu Beginn verläuft das Prozedere beim Beispiel mit den Röntgenbildern der Zähne auf ähnliche Weise. Die Neurone der Eingabeschicht leiten Informationen zu den Faltungsschichten weiter. Die *Faltungsschicht* verwendet bestimmte Filter, die kleine Teilbereiche des Bildes durchlaufen und gewisse Merkmalskarten (Feature Maps) erzeugen. Darunter versteht man Karten, die bestimmte Merkmale des Bildes hervorheben und anzeigen, wo sich diese befinden. In unserem Beispiel könnten dies zum Beispiel Kanten sein, wie die Außenkonturen eines Zahnes oder abgegrenzte Aufhellungen, die Karies darstellen.

  In der nachfolgenden Schicht, der *Pooling-Schicht*, werden diese Daten komprimiert. Dies geschieht durch die Reduzierung der Größe der Feature Maps und dient dazu, den Prozess zu beschleunigen und effizienter zu machen. Einfach

erklärt verwendet die Pooling-Schicht die Informationen der Faltungsschicht und komprimiert diese auf relevante Merkmale in kleineren Bildbereichen, die in diesem Beispiel entscheidend für die Frage „kariös“ und “nicht kariös“ sind.

Die *voll verbundenen Schichten* kombinieren im Anschluss alle reduzierten Daten der vorangegangenen Schichten und lernen wie diese zusammenhängen. Man kann es sich so vorstellen, dass hier die vorab aufbereiteten Daten endgültig ausgewertet werden. Dabei werden alle Informationen miteinander kombiniert, wobei jedes Neuron, mit allen Neuronen der vorangestellten Schicht verbunden ist. Die Anzahl der Neuronen nehmen typischerweise beim Voranschreiten der Schichten ab, bis sich eine finale Entscheidungstendenz abzeichnet. Durch sie wird in unserem Beispiel eine finale Klassifikation für den Zahnzustand getroffen, der durch die Ausgabeschicht als „kariös“ oder „nicht kariös“ definiert wird, indem alle Informationen, die für die Entscheidungsfindung relevant sind, mit einbezogen werden (31, 32).

Durch das Training mit annotierten Röntgenbildern lernt das CNN automatisch relevante Merkmale zu identifizieren und zu klassifizieren und verbessert seine Genauigkeit mit zunehmender Anzahl von Trainingsdaten (24, 33).

1.2 Kieferorthopädische Diagnose und Therapie

1.2.1 Fernröntgenanalyse

Die Fernröntgenanalyse spielt eine entscheidende Rolle in der Diagnostik und Behandlungsplanung in der Kieferorthopädie und gehört mit der klinischen Untersuchung, der Anfertigung kieferorthopädischer Modelle, dem Panoramaröntgen und der intra- und extraoralen Fotodokumentation zu den Hauptdiagnostika, auf die sich die endgültige Behandlungsplanung stützt.

Das Fernröntgen, auch Schädelröntgenseitenbild, ist eine seitliche Röntgenaufnahme des Kopfes, der durch einen Kephalostaten (Kopfhalter) fixiert wird (34).

Die Analyse des Fernröntgens erfolgt durch die Identifikation anatomischer Landmarken, sogenannter Fernröntgenpunkte, welche als Referenzpunkte für eine Vielzahl von linearen

und angulären Messungen dienen. Diese Messungen ermöglichen es Kieferorthopäd*innen Abweichungen von der physiologischen kraniofazialen Anatomie zu erkennen.

Zunächst ermöglicht die Fernröntgenanalyse eine präzise Beurteilung der skelettalen Beziehungen zwischen Ober- und Unterkiefer, sowie deren Position zum restlichen Schädel. Darüber hinaus liefert sie Aufschluss über das individuelle Wachstumsmuster der Patient*innen, was bei jungen Patient*innen zur Vorhersage der künftigen Entwicklung des Kiefers und des Gesichts von Bedeutung ist. Auch die Ausrichtung (Inklination) der Frontzähne wird analysiert, sowie der Abstand zur Facialebene. Des Weiteren wird das Weichteilprofil, also die nicht knöchernen Strukturen der Patient*innen beurteilt.

Es gibt eine Vielzahl von Fernröntgenpunkten und Analyseverfahren, häufig verwendete Landmarken sind (34 – 36, 101, 106):

1. Sella (S): Mittelpunkt der Sella turcica
2. Nasion (N): ventralster Punkt der Sutura nasi frontalis am Übergang zwischen Os frontale und Os nasale
3. Orbitale (Or): Der tiefste Punkt im unteren Rand der Augenhöhle
4. Porion (Po): Der höchste Punkt des äußeren Gehörgangs
5. A-Punkt: Dorsalster Punkt an der vorderen Kontur der Maxilla
6. B-Punkt: Dorsalster Punkt an der vorderen Kontur der Mandibular
7. Pogonion (Pog): Ventralster Punkt der Unterkiefersymphyse
8. Menton (Me): kaudalster Punkt der Unterkiefersymphyse
9. Gonion (Go): Schnittpunkt der Winkelhalbierenden zwischen den zwei Tangenten des aufsteigenden Astes, dem Ramus mandibulae und dem Kieferwinkel
10. Artikulare (Ar): Schnittpunkt von der Schädelbasis kaudal mit dem Ramus mandibulae dorsal
11. Gnathion (Gn): Schnittpunkt der Mittelsenkrechten zwischen Pogonion und Menton und der ventralen Symphyse

    (34 – 36, 101, 106)

Anhand dieser Landmarken werden verschiedene anguläre und lineare Messungen vorgenommen.

Häufig verwendete anguläre Messungen sind:

1. SNA-Winkel: Zeigt die Position des Oberkiefers relativ zur Schädelbasis
2. SNB-Winkel: Gibt Aufschluss über die Position des Unterkiefers relativ zur Schädelbasis
3. ANB-Winkel: Misst die skelettale Beziehung zwischen Ober- und Unterkiefer
4. Summenwinkel nach Björk: N-S-Ar + S-Ar-Go + Ar-Go-Me

Häufig verwendete lineare Messungen sind:

1. S-Go: Hintere Gesichtshöhe
2. N-Me: Vordere Gesichtshöhe
3. Frankfurter Horizontale: Verbindung Orbitale und Porion

(34 – 36, 101, 106)

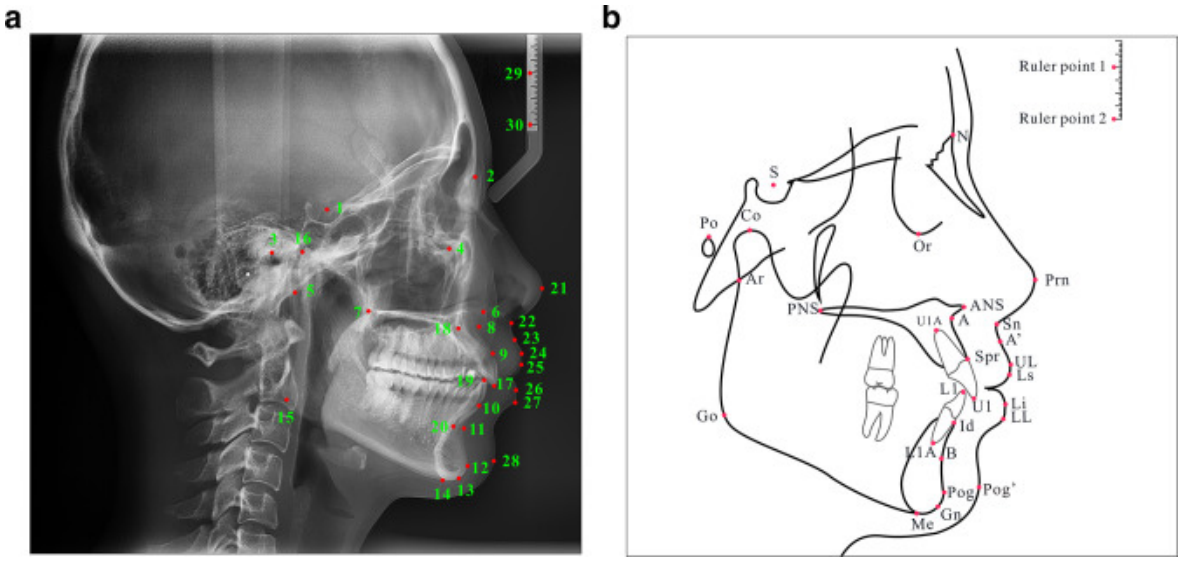


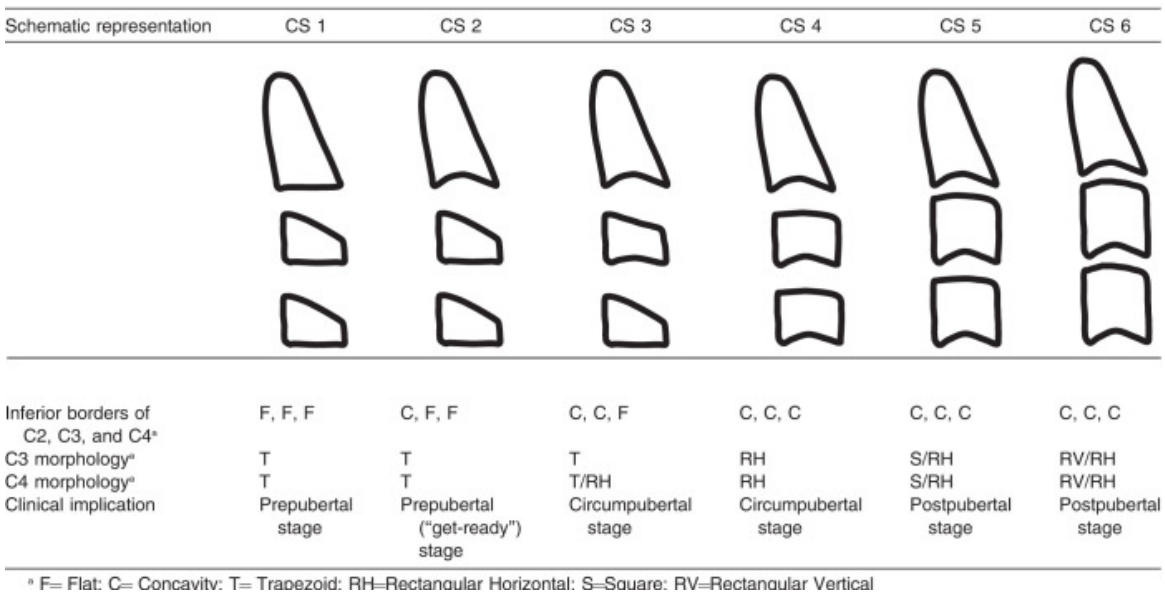
Abbildung 2: a) Fernröntgenbild mit annotierten Landmarken, b) Durchzeichnung eines Fernröntgenbildes (63)

1.2.2 Wirbelkörperreifungsgrade (Cephalometric Vertebral Maturation)

Die Analyse der Wirbelkörperreifungsgrade im Fernröntgen (Cephalometric Vertebral Maturation, CVM) ist ein entscheidendes Instrument in der kieferorthopädischen Diagnostik und Behandlungsplanung. Durch sie wird der optimale Zeitpunkt für bestimmte kieferorthopädische Behandlungen festgelegt. Die CVM-Methode ermöglicht Rückschlüsse auf die skelettale Reifephase bei Kindern und Jugendlichen und basiert auf der Analyse der Halswirbelsäule. Analysiert werden dabei der zweite (C2), dritte (C3) und vierte (C4) Halswirbel. Es können sechs Reifestadien dieser drei Halswirbel bestimmt werden, basierend auf der Morphologie der Wirbelkörper.

Diese Methode folgt einem zweistufigen Prozess: Zuerst wird die untere Fläche der Wirbelkörper auf Flachheit oder Konkavität untersucht. Im Anschluss wird die Form C3 und C4 analysiert, die typischerweise von trapezoidal über rechteckig horizontal, quadratisch, bis hin zu rechteckig vertikal fortschreitet. Diese Formveränderungen korrelieren mit spezifischen zervikalen Stadien (CS), wobei CS 1 und CS 2 als präpubertär, CS 3 und CS 4 als circumpubertär und CS 5 und CS 6 als postpubertär klassifiziert werden (Abbildung 3) (37).

Ein bedeutender Vorteil der CVM-Analyse liegt darin, dass sie keiner zusätzlichen Strahlenexposition bedarf, da die Anfertigung eines Fernröntgens bereits ein fester Bestandteil der standardisierten Diagnostik in der kieferorthopädischen Praxis darstellt.



Schematic representation	CS 1	CS 2	CS 3	CS 4	CS 5	CS 6
Inferior borders of C2, C3, and C4[a]	F, F, F	C, F, F	C, C, F	C, C, C	C, C, C	C, C, C
C3 morphology[a]	T	T	T	RH	S/RH	RV/RH
C4 morphology[a]	T	T	T/RH	RH	S/RH	RV/RH
Clinical implication	Prepubertal stage	Prepubertal ("get-ready") stage	Circumpubertal stage	Circumpubertal stage	Postpubertal stage	Postpubertal stage

[a] F= Flat; C= Concavity; T= Trapezoid; RH=Rectangular Horizontal; S=Square; RV=Rectangular Vertical

Abbildung 3: CVM-Stadieneinteilung (37)

1. Bewertung der unteren Grenze der Wirbelkörper:

- CS 1: Die untere Oberfläche von C2, C3 und C4 ist flach
- CS 2: Eine Konkavität am unteren Rand von C2 wird sichtbar
- CS 3: Konkavitäten an C2 und C3 sind zu beobachten
- CS 4, CS 5, CS 6: Konkavitäten sind an den unteren Rändern aller drei Wirbel (C2-C4) sichtbar

2. Bewertung der Form von C3 und C4:

- Bis CS 3: trapezoide Form von C3 und C4 (Ausnahmefall: einer der Wirbelkörper kann eine rechteckig horizontale Form aufweisen)
- CS 4: C3 und C4 haben eine rechteckige Form
- CS 5: C3 und C4 sind quadratisch,
- CS 6: C3 und/ oder C4 haben eine rechteckig vertikale Form, wobei die Länge des hinteren Randes länger als die des unteren Randes ist (37).

1.2.3 Strategische Herausforderungen der kieferorthopädischen Extraktion

Die kieferorthopädische Extraktionstherapie stellt eine komplexe Behandlungsstrategie dar, die vorrangig bei Platzmangel oder einem Missverhältnis zwischen Zahn- und Kiefergröße zum Einsatz kommt. Ziel ist es, eine achsengerechte Einordnung aller Zähne zu ermöglichen. Die Entscheidung zur Extraktion ist oft eine Folge eines nicht anderweitig korrigierbaren Engstands, der die Einordnung der Zähne in den Zahnbogen beeinträchtigt und somit funktionelle und ästhetische Auswirkungen hat (38).

Die Komplexität der Extraktionstherapie liegt nicht nur in der Entscheidung zur Extraktion selbst, sondern auch in der Bewältigung der daraus resultierenden Probleme, wie der Resorptionsgefahr durch lange Bewegungsstrecken, der Verbleib von Restlücken oder das Auftreten von Gingivaduplikaturen (38).

Die Diagnose und Planung der Extraktionen ist entscheidend, wobei die Bolton-Analyse, das Wachstumsmuster der Patient*innen und das Vorhandensein von genügend Knochenangebot wesentliche Faktoren darstellen (38).

1.2.4 Beurteilung der chirurgischen Notwendigkeit

In der Behandlung von Dysgnathien (fehlerhafte Kieferrelationen), steht die Entscheidung zwischen einer rein kieferorthopädischen Behandlung und einer zusätzlichen chirurgischen Behandlung im Fokus.

Während rein kieferorthopädische Methoden hauptsächlich dentoalveoläre Fehlstellungen korrigieren, ist für skelettale Dysgnathien oftmals ein chirurgischer Ansatz erforderlich, um die Kieferrelationen in eine physiologische Position zu bringen und funktionelle, sowie ästhetische Ergebnisse zu optimieren.

Die gängigen Dysgnathien, die chirurgische Korrekturen erfordern könnten, umfassen die maxilläre sowie mandibuläre Retrognathie und Prognathie, als auch andere Anomalien wie offene und tiefe Bisse. Dabei ist nicht nur die Okklusion beeinträchtigt, sondern auch das Gesichtsprofil und die Ästhetik der Patient*innen.

Die Herausforderung in der chirurgischen Entscheidungsfindung besteht darin, die Fehlbisse korrekt zu klassifizieren und die Behandlung entsprechend zu planen.

Besonders bei Patient*innen mit mandibulärer Prognathie kommt es im Regelfall zu Wachstumsschüben nach dem 16. Lebensjahr, welche in die Behandlungsplanung mit einbezogen werden sollten. Eine finale Entscheidung bezüglich einer Umstellungsosteotomie wird erst am Ende des Wachstums endgültig getroffen.

Zusammenfassend ist die Abwägung von Risiken und Vorteilen eines chirurgischen Eingriffs gegenüber einer rein kieferorthopädischen Behandlung ein wichtiger Schritt in der chirurgischen Entscheidungsfindung, der eine individuelle Behandlungsplanung und eine interdisziplinäre Abstimmung erfordert, um das optimale Ergebnis für die Patient*innen zu erzielen (39).

1.2.5 Einfluss kraniofazialer Wachstumsmuster auf orthodontische Behandlungen

Die kraniofaziale Entwicklung beeinflusst maßgeblich die orthodontischen Behandlungsstrategien und -ergebnisse. Das kraniofaziale Wachstum betrifft sowohl den Oberkiefer als auch den Unterkiefer.

Die Kenntnis der Wachstumstendenz ist entscheidend, um den optimalen Zeitpunkt und die Art der kieferorthopädischen Behandlung zu bestimmen, insbesondere bei Heranwachsenden. Eine frühzeitige Diagnose und Intervention können beispielsweise dabei helfen, ungünstige Wachstumstendenzen zu korrigieren und die Notwendigkeit einer späteren umfangreicheren Behandlung zu verringern. Zu den Einflussfaktoren, die das kraniofaziale Wachstum beeinflussen, gehören maßgeblich genetische Faktoren. Darüber hinaus spielen Umweltfaktoren, Zungenposition, Schluckmuster oder Atemgewohnheiten eine Rolle (40 – 43).

1.2.6 Prognose der Behandlungsdauer von kieferorthopädischen Behandlungen

Die Vorhersage der Behandlungsdauer in der Kieferorthopädie ist von einer Vielzahl von Faktoren abhängig. Darunter das Alter der Patient*innen bei Behandlungsbeginn, das Geschlecht der Patient*innen, der Schweregrad der Malokklusion, die Komplexität der Behandlung, sowie die Compliance der Patient*innen.

Im Rahmen der Behandlungsplanung ist es üblich, auf Grundlage dieser Faktoren und klinischen Erfahrungen Prognosen zu stellen, um den Patient*innen einen zeitlichen Rahmen zu geben, auf den sie sich einstellen können (44 – 46).

1.3 Metriken

Um die in dieser Literaturübersicht verglichenen Studien und ihre Ergebnisse zu verstehen, ist es entscheidend, spezielle statistische Metriken, Begriffsbezeichnungen und Verfahren im Kontext von KI-Modellen zu kennen. Die folgenden Aufzählungen stellen Definitionen dar.

Die Sensitivität ist die Fähigkeit eines Klassifikationsmodells tatsächlich positive Fälle korrekt als positiv zu identifizieren. Sie wird berechnet als der Anteil der korrekt identifizierten positiven Fälle an der Gesamtzahl der tatsächlich positiven Fälle im Datensatz. Eine hohe Sensitivität bedeutet, dass das Modell effektiv darin ist, positive Fälle zu erkennen und somit die Wahrscheinlichkeit von falsch negativen Ergebnissen minimiert (26).

Die Spezifität ist die Fähigkeit eines Klassifikationsmodells, tatsächlich negative Fälle korrekt als negativ zu identifizieren. Sie wird berechnet als der Anteil der korrekt identifizierten negativen Fälle an der Gesamtzahl der tatsächlichen negativen Fälle im Datensatz. Eine *hohe* Spezifität bedeutet, dass das Modell effektiv darin ist, negative Fälle zu erkennen und somit die Wahrscheinlichkeit von falsch positiven Ergebnissen minimiert (26).

Die Genauigkeit ist ein Maß für die Gesamtleistung eines Klassifikationsmodells, das den Anteil aller korrekt klassifizierten Fälle (sowohl wahre positive als auch wahre negative)

im Verhältnis zur Gesamtzahl aller Fälle im Datensatz angibt. Sie spiegelt die Fähigkeit des Modells wider, sowohl positive als auch negative Fälle korrekt zu identifizieren und zu klassifizieren. Eine hohe Genauigkeit deutet darauf hin, dass das Modell eine starke Vorhersagekraft besitzt (26).

Die SDR (Successful Detection Rate, Erfolgsdetektionsrate) wird bei der Landmarkendetektion gebraucht. Sie gibt an wie viele Landmarken korrekt annotiert wurden im Verhältnis zur gesamten Landmarkenzahl (47).

Die SCR (Successful Classification Rate, Erfolgsklassifizierungsrate) gibt die korrekte Klassifizierung der Messungen, die sich aus den detektierten Landmarkenpunkten ergeben, an. Dies könnten anguläre oder lineare Messungen in einer Fernröntgenanalyse sein (47).

Die Intra-Rater-Reliabilität gibt an, wie konsistent die Bewertungen von einzelnen Prüfer*innen bei Wiederholungen sind. Also, ob Prüfer*innen zu verschiedenen Zeitpunkten immer zu gleichen Bewertungen oder Messergebnissen gelangen.

Die Inter-Rater-Reliabilität gibt an, wie konsistent die Bewertungen oder Messungen zwischen verschiedenen Beurteiler*innen sind. Sie bewertet, inwiefern diese zu gleichen Ergebnissen gelangen (103).

Die AUC (Area Under the Curve) zeigt, wie gut ein Klassifikationsmodell ist. Sie stellt die Fläche unter der ROC-Kurve (Receiver Operating Characteristic) dar, die die Fähigkeit eines Modells abbildet, zwischen zwei Klassen zu unterscheiden. Zum Beispiel „kariös“ und „nicht kariös“. Die ROC-Kurve vergleicht die Rate der richtigen Vorhersagen (True Positive Rate) mit der Rate der falschen Vorhersagen (False Positiv Rate). Die AUC wird durch eine einfache Zahl ausgedrückt, von 0 bis 1. Je größer die Fläche unter der ROC-Kurve ist, desto näher liegt der AUC-Wert an 1. Ein AUC-Wert von 0,5 deutet auf ein Modell hin, dass nicht besser als zufällig klassifiziert. In der Abbildung dargestellt durch die gestrichelte Linie. Ein Wert von 1,0 deutet auf eine perfekte Klassifikation hin. Mit dem AUC-Wert lassen sich verschiedene Klassifikationsmodelle einfach untereinander vergleichen (26).

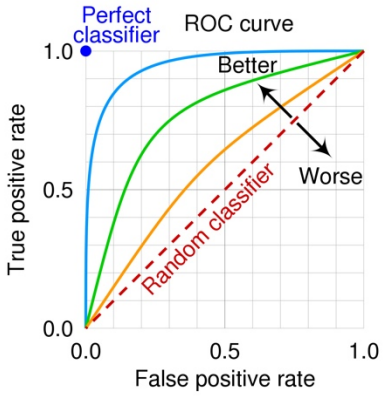


Abbildung 4: AUC- und ROC-Kurve (105)

Der ICC (Intraklassen-Korrelationskoeffizent) ist ein statistisches Maß, das verwendet wird, um die Reliabilität oder die Übereinstimmung von Bewertungen innerhalb spezifischer Gruppe oder Klassen zu beurteilen. Der ICC nimmt Werte zwischen 0 und 1 an. Wobei ein Wert von 0 darauf hinweist, dass keine Übereinstimmung zwischen den Bewertungen besteht, ein Wert von 1 würde eine perfekte Übereinstimmung darstellen (47).

Der MRE (Mean radial Error, Mittlerer radialer Fehler) ist der durchschnittliche radiale Fehler zwischen den vorhergesagten und tatsächlichen Positionen der Landmarken (48).

Ein niedriger MRE bedeutet, dass das Modell genaue Vorhersagen macht, wohingegen ein hoher MRE darauf hinweist, dass die Vorhersagen des Modells oft falsch sind (48).

Der F1-Wert fasst die Genauigkeit eines Klassifikationsmodells zusammen. Er bezieht sowohl die Präzision (Verhältnis der korrekt vorhergesagten positiven Fälle zur Gesamtzahl der vorhergesagten positiven Fälle), als auch die Sensitivität (Anteil der korrekt identifizierten positiven Fälle an der Gesamtzahl der tatsächlich positiven Fälle) mit ein (49). Oft wird er verwendet, wenn Klassen ungleichmäßig verteilt sind. Der F1-Wert wird von 0 bis 1 angegeben, wobei 1 eine perfekte Präzision, sowie Sensitivität eines Klassifikators ausdrücken würde (49).

Das gewichtete Kappa wird verwendet, um die Übereinstimmung zwischen zwei Beurteilungen zu messen. Dabei wird das Ausmaß der Abweichung mitberücksichtigt. Es kann Werte zwischen -1 und 1 annehmen. Wobei 1 eine perfekte Übereinstimmung und -1 ein perfektes Nicht-Übereinstimmen ausdrücken würde.

1.4 Ziel der Literaturrecherche

Diese Arbeit zielt darauf ab, ein umfassendes Verständnis der aktuellen KI-Anwendungen in der Kieferorthopädie zu vermitteln und die Auswirkungen auf Diagnostik, klinische Praxis und Behandlungsplanung zu erforschen. Durch die Analyse verschiedener KI-Modelle und deren Anwendungsbereiche in der Kieferorthopädie, sollen die Potentiale und Grenzen dieser Technologien in Bezug auf die kieferorthopädische Praxis aufgezeigt werden.

2 Material und Methoden

2.1 Suchstrategie und Durchsuchung der Datenbanken

Diese Literaturrecherche umfasst eine systematische Suche zum Thema „Künstliche Intelligenz in der Kieferorthopädie“, ausgeführt in drei Online-Datenbanken.

Eine sorgfältige Durchsicht umfasste die Datenbanken PubMed, Google Scholar und Mendeley, wobei der Fokus auf Veröffentlichungen im Zeitraum von 01.01.2014-01.01.2024 lag. Für die methodische Durchführung wurde sich an den PRISMA-Leitlinien (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) orientiert, welche als Leitfaden zur Erstellung systematischer Übersichtarbeiten dienen (50, 51).

Verwendete Schlüsselwörter waren „Artificial Intelligence“, „machine learning“, „deep learning“, „neural network“, „CNN“ in Kombination mit „orthodontics“, um Studien mit kieferorthopädischem Bezug zu selektieren.

Publikationen, die nicht als zuverlässig eingestuft wurden, unzureichende Informationen bereitstellten, nur geringfügigen Inhalt boten oder nicht peer-reviewt wurden, wurden von der Einbeziehung ausgenommen. Ebenso wurden kostenpflichtige Studien nicht berücksichtigt. Die Recherche beinhaltet ausschließlich Studien in englischer Sprache. Des Weiteren ist anzumerken, dass nach Abschluss der gesichteten Literatur sich die Erkenntnis ergab, dass das anfänglich definierte Zehnjahresfenster für die Sammlung der Studien eine zu großzügige Spanne darstellte. Infolge der rapiden Entwicklung im Bereich

der künstlichen Intelligenz verloren einige Studien im Hinblick auf ihren Forschungsstand an Relevanz und wurden daher nicht in die Analyse mit einbezogen. Dies spiegelt die Dynamik des Fortschritts in der KI-Forschung wider, welcher eine kontinuierliche Aktualisierung des Wissensstandes unerlässlich macht.

2.2 Datenerhebung und Analyse

Alle Studien, deren Titel potenziell den relevanten Themengebieten entsprachen, wurden mit einer Tabelle verwaltet und auf kostenfreie Zugänglichkeit geprüft.

Zur Einschätzung ihrer thematischen Eignung wurde zunächst das Abstract jedes Artikels geprüft. Von geeigneten Studien wurden die Volltexte für eine detailliertere Analyse beschafft.

Der Prozess der Studienauswahl wurde gemäß den PRISMA-Richtlinien in einem Flussdiagramm (siehe Abbildung 5) dokumentiert.

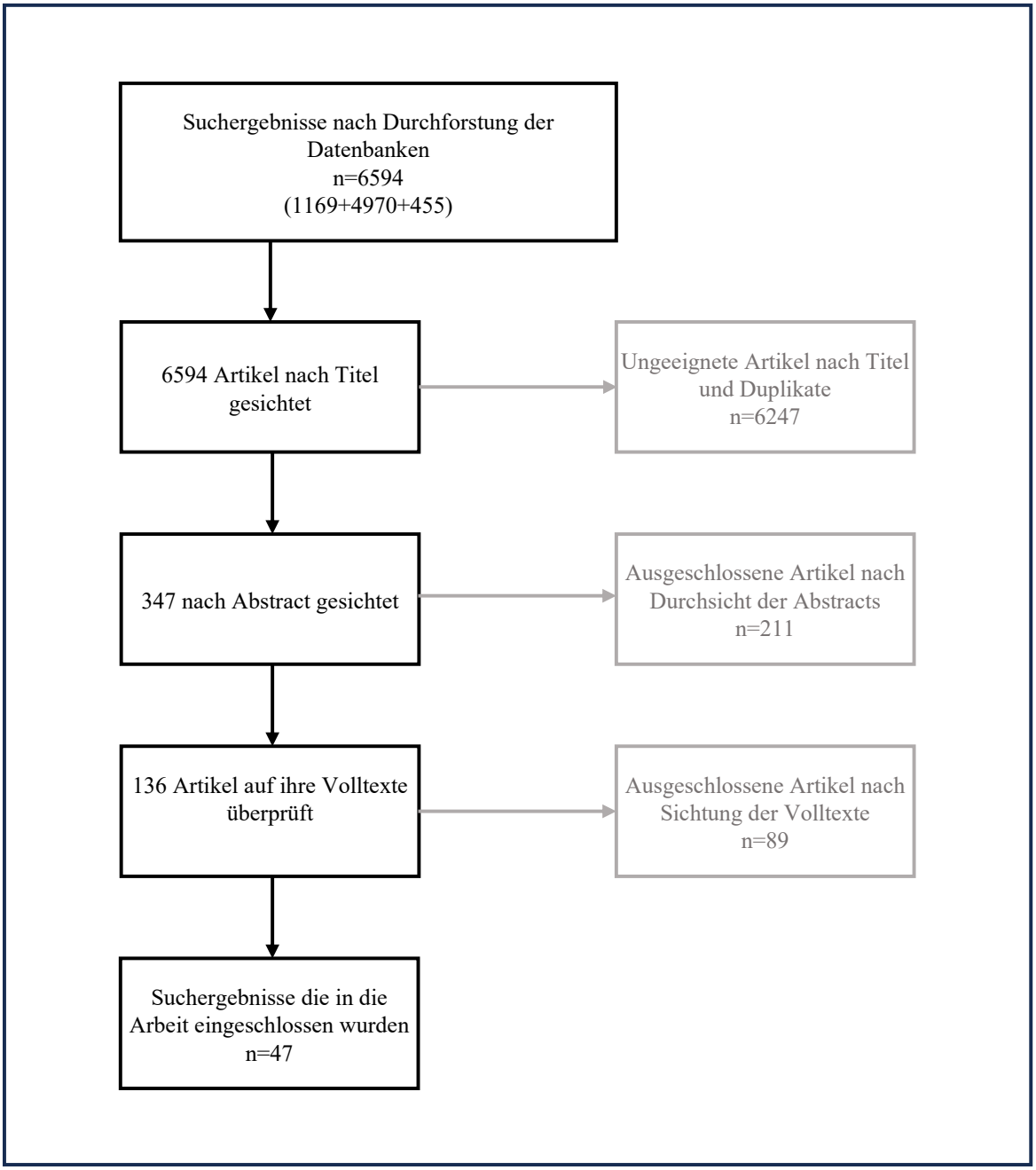


Abbildung 5: Literaturauswahlverfahren (eigene Abbildung)

3 Ergebnisse

In die Literaturrecherche wurden 47 Artikel inkludiert, welche in den Jahren 2018 bis 2023 publiziert wurden. Dabei fanden die meisten Veröffentlichungen in den Jahren 2021 bis 2023 statt. Die Studien wurden in 13 verschiedenen Ländern durchgeführt: China (n = 9), Nepal (n = 1), USA (n = 10), Griechenland (n = 1), Südkorea (n = 14), Japan (n = 5), Deutschland (n = 1), Taiwan (n = 1), Israel (1), Brasilien (n = 2), Indien (n = 1), Niederlande (n = 1). Diese Studien wurden in 20 unterschiedlichen Journalen veröffentlicht: DMFR (Dentomaxillofacial Radiology) (n = 4), BMC oral health (n = 7), Orhtodontics & Craniofacial Research (n = 6), Clinical Medicine (n = 2), Sensors basel (n = 1), Heliyon (n = 1), Applied Science (n = 4), The Angle Orthodontist (n = 3), Journal of Dental Science (n = 1), Scientific Reports (n = 4), Medicine & Biology (n = 1), Progress in Orthodontics (n = 1), Diagnostics basel (n = 6), Journal of Clinical Medicine (n = 1), Bioengineering (n = 1), European Journal of Orthodontics (n = 1), Healthcare Informatics Research (n = 2), Dentistry Journal (n = 1), Journal of personalied Medicine (n = 1), Journal of Cranio-Maxillofacial Surgery (n = 1).

3.1 Anwendungsgebiete künstlicher Intelligenz in der Kieferorthopädie

Die Themenbereiche dieser Literaturrecherche zum Thema KI in der Kieferorthopädie umfassen acht Anwendungsgebiete:

- Durchführung einer Fernröntgenanalyse (n = 21)
- Bewertung der kraniofazialen Reifung, CVM (n = 6)
- Treffen von Extraktionsentscheidungen (n = 4)
- Erstellung eines Behandlungsplans (n = 4)
- Beurteilung der chirurgischen Notwendigkeit (n = 5)
- Vorhersage der Behandlungsdauer (n = 1)
- Vorhersage des kraniofazialen Wachstums (n = 4)
- Vorhersage der postoperativen Weichgewebsmorphologie (n = 2)

3.2 Studienaufbau

Der experimentelle Aufbau in der Mehrheit der untersuchten Studien folgte einem konsistenten Schema. Bereits existierende KI-Modelle wurden entweder adaptiert, modifiziert oder neue Modelle wurden speziell für die Studienzwecke konzipiert. Diese wurden anschließend mit Datensätzen trainiert, die aus klinischen Einrichtungen oder bestehenden Sammlungen stammten.

Datensätze waren beispielsweise Röntgenbilder, Fotos, Abdrücke oder Daten von Patient*innen. Nach der Trainingsphase erfolgte eine umfassende Leistungsbeurteilung der Modelle durch einen Testdatensatz. Dabei stellte der Trainingsdatensatz immer den größten Datenpool dar. Manche KI-Strukturen waren bereits im Vorfeld trainiert worden und wurden in den jeweiligen Studien nur noch getestet. Einige Studien validierten vor der Testung die verwendeten KI-Modelle. Dies ist ein Schritt, um die Zuverlässigkeit und Allgemeingültigkeit während des Entwicklungsprozesses von KI-Systemen zu überprüfen. Bei der Methode der Kreuzvalidierung wird der Datensatz in mehrere Untergruppen aufgeteilt und das Modell wird mehrmals trainiert, wobei jeweils eine der Untergruppen als Testdatensatz verwendet wird, während die anderen zum Training dienen.

3.3 Definition des Goldstandards

Um einen Vergleichsmaßstab (Goldstandard, Ground Trouth) für die künstliche Intelligenz zu schaffen, wurden oftmals Standards durch Gruppen von Expert*innen festgelegt. Expert*innen waren häufig Kieferorthopäd*innen mit langer Berufserfahrung. Aber auch Zahnärzt*innen und Radiolog*innen definierten den Goldstandard. Zur Ermittlung des Goldstandards wurden Werte unterschiedlicher Untersucher*innen gemittelt, es wurden Konsensentscheidungen getroffen, Unstimmigkeiten wurden durch eine dritte Instanz gelöst, Expert*innen wurden auf Gleichheit trainiert oder ein*e einzelne*r Prüfer*in definierte den Goldstandard.

Um die Zuverlässigkeit des Goldstandards zu ermitteln, wurden oft die Intra- und Inter-Rater-Reabilitäten von den Prüfer*innen erhoben. Dies bedeutet, dass die Reproduzierbarkeit der Bewertungen einzelner Expert*innen zu verschiedenen

Zeitpunkten oder die Übereinstimmung der Bewertungen im Vergleich zwischen unterschiedlichen Prüfer*innen überprüft wurde.

3.4 Strategien zur Verbesserung der KI-Leistung

3.4.1 Datenaugmentation und Overfitting

Eine Datenaugmentation wird durchgeführt, um die Vielfalt und Quantität der Trainingsdaten künstlich zu erhöhen, ohne zusätzlich originale Daten sammeln zu müssen. Dies wird erreicht, indem bestehende Daten modifiziert oder transformiert werden, um neue, abgeleitete Datensätze zu erzeugen. Sind die Trainingsdaten zum Beispiel Röntgenbilder, können diese augmentiert werden, indem man sie rotiert oder spiegelt. Die Hauptziele der Datenaugmentation sind die Verbesserung der Modellgeneralisierung und die Verringerung von Overfitting (52).

Beim Overfitting einer KI-Struktur kommt es zur übermäßigen Anpassung an die Trainingsdaten. Dies geschieht auf Kosten der Fähigkeit zur Generalisierung auf neue, unbekannte Daten.

Dabei ist die Leistung eines KI-Modells bei Trainingsdaten deutlich höher als bei einem separaten Validierungs- oder Testdatensatz.

Das KI-Modell lernt dabei die Trainingsdaten auswendig, was bedeutet, dass es spezifische Eigenschaften der Trainingsdaten erfasst, die nicht allgemein auf andere Daten übertragbar sind (26, 47).

3.4.2 Region of Interest (ROI)

Einige Studien verwendeten sogenannte ROIs vor der Landmarkendetektion, um die Analyse auf relevante Bildbereiche zu begrenzen. Dabei werden im ersten Schritt wesentliche Bildbereiche detektiert, um im Anschluss die definitive Landmarke in diesem Bereich zu markieren. Ziel ist dabei eine schnellere und gezieltere Arbeit der Algorithmen zu ermöglichen und Verzerrungen zu vermeiden.

3.4.3 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-Cam stellt eine Technik dar, die verwendet wird, um Entscheidungsfindungen von CNNs visuell zu interpretieren. Diese Methode nutzten einige Studien während des Validierungs- oder Testungsprozesses. Es bietet Aufschluss darüber, welche Bereiche des Röntgenbildes dazu beigetragen haben, dass das Netzwerk eine bestimmte Diagnose oder Klassifikation getroffen hat. Durch die Analyse der Heatmaps, die Grad-CAM generiert, können Entwickler*innen Algorithmen optimieren und Schwächen identifizieren oder relevante Inputparameter herausfiltern. Bei dem Beispiel mit den kariösen Zähnen würde die Heatmap die Bildbereiche auf den Röntgenbildern, die für die Entscheidungsfindung der KI wichtig waren, rot markieren. Dies könnten zum Beispiel dunkle Bildbereiche mit unscharfen Konturen sein, wie sich eine Karies klassisch im Röntgenbild darstellt (102).

3.5 Datendimension der unterschiedlichen Anwendungsbereiche

Bei der KI-gesteuerten Analyse von Fernröntgen betrug die Anzahl der detektierten Landmarken 16 bis 80 Punkte. Es wurden sowohl Hartgewebs-, als auch Weichgewebslandmarken verwendet.

Die Anzahl der verwendeten Fernröntgen lag zwischen 400 - 9870 zu Trainingszwecken und 100 - 259 als Testdatensatz.

Für die Stadieneinteilung der Cervical Vertebral Maturation beinhaltete der größte verwendete Datenpool 11745 Fernröntgenaufnahmen, der kleinste 600.

Vor der Klassifizierung wurde zum Teil ein automatischer Objektdetektor verwendet, um nur die relevanten Bereiche der Röntgenbilder zu extrahieren.

Bei den Studien die sich mit Extraktionsentscheidungen befassten beinhalteten die Datensätze 287 - 3136 Fälle von Patient*innen. Inputdaten stellten intraorale Fotografien, Fernröntgenbilder, Panoramaröntgen, zephalometrische und demografische Merkmale dar.

In Studien zu KI-generierten Behandlungsplanungen wurden Datensätze in der Größenordnung von 302 bis 1020 Fällen von Patient*innen verwendet. Die Inputparameter waren sehr vielfältig und stellten sich durch Modelle, Fotos, Röntgenbilder, Problemlisten und Behandlungsstrategien dar.

Für die Entscheidungsfindung bezüglich einer chirurgischen Notwendigkeit wurden Datensätzen im Bereich von 333 - 960 in Form von Fernröntgenaufnahmen und Fotografien genutzt.

Zur Prognose der kraniofazialen Wachstumstendenz wurden 114 - 123 Fälle von Patient*innen mit je 3 Fernröntgenbildern zu unterschiedlichen Zeitpunkten verwendet.

Die beiden Studien, die eine Vorhersage des Weichteilprofils nach chirurgischen oder kieferorthopädischen Behandlungen prognostizierten nutzten prä- und postoperative DVT-Aufnahmen und 3D-Fotografien.

3.6 Verwendete Metriken der unterschiedlichen Anwendungsgebiete

Bei der Landmarkendetektion auf Fernröntgen wurde eine Abweichung von > 2 mm als klinisch relevant definiert. Im Umkehrschluss bedeutet dies, dass eine Abweichung innerhalb der 2 mm als klinisch akzeptabel gilt (53, 54). Aus diesem Grund beschreiben zahlreiche Studien innerhalb dieser Analyse die Effizienz von KI-Systemen mittels der Erfolgsdetektionsrate (SDR), ausgedrückt als Prozentsatz innerhalb eines Radius von 2 mm und der Erfolgsklassifizierungsrate (SCR).

Die verwendeten Metriken zur Überprüfung der Leistung der KI-Systeme bei der CVM-Stadieneinteilung gestalteten sich vielfältiger als bei der Landmarkendetektion. Die häufigsten Metriken waren Genauigkeit, AUC-Wert und F1-Wert.

Bei den KI-generierten Extraktionsentscheidungen wurden die Genauigkeit, der AUC-Wert und der gewichtete Kappa-Wert zur Darstellung der Leistung verwendet.

Bei den Studien zur Behandlungsplanung durch künstliche Intelligenz gebrauchten die Autor*innen die Genauigkeit, den AUC-Wert, die Sensitivität und Spezifität zur Beschreibung der KI-Leistung.

Die Beurteilung der chirurgischen Notwendigkeit wurden mit den Metriken Genauigkeit, Spezifität, Sensitivität und dem AUC-Wert interpretiert.

Die Prognose des kraniofazialen Wachstums wurde durch die Genauigkeit beschrieben und die Leistung der Vorhersage der Gesichtsmorphologie wurde durch den mittleren absoluten Fehler und der Erfolgsdetektionsrate erfasst.

4 Diskussion

4.1 KI-generierte Fernröntgenanalyse

Die systematische Untersuchung zur Anwendung von künstlicher Intelligenz in der Kieferorthopädie zeigt, dass sich die Mehrheit der Studien auf die Erkennung anatomischer Landmarken in Fernröntgenbildern und deren Auswertung fokussierten.

Fünf Studien untersuchten die Leistung eines sogenannten YOLOv3-Algorithmus (You Only Look Once, Version 3) oder dessen Modifikationen (55 – 59).

Bulatova et al. (55) überprüfte die Leistung des klassischen YOLOv3-Algorithmus. Bei der Gegenüberstellung der manuellen Detektion und der KI-ermittelten Landmarkenpunkte generierte die KI eine Erfolgsdetektionsrate innerhalb eines 2 mm Radius mit einer Streuung von 0,13 mm für 75 % der Landmarkenmarkenpunkte. Nur die Detektion der Punkte U1 Apex, L1 Apex, Basion, Gonion und Orbitale lag außerhalb des klinisch akzeptablen Bereichs von 2 mm. Laut der Autor*innen, lässt dies darauf schließen, dass die KI in der Lage ist mit einer relativ hohen Zuverlässigkeit und geringer Abweichung Landmarken in Fernröntgen zu detektieren.

In der Studie von Zhao et al. (56) wurde eine Modifikation des YOLOv3-Algorithmus namens Multiple-Scale YOLOv3 (MS-YOLOv3) entwickelt und getestet. Die experimentellen Ergebnisse zeigten, dass MS-YOLOv3 eine robuste Fähigkeit zur Identifizierung von Landmarken besitzt, mit einer höheren Erfolgsdetektionsrate (80,84% innerhalb von 2 mm) und geringeren mittleren radialen Fehlern im Vergleich zu YOLOv3 in der Studie von Bulatova et al. (55). Herausfordernd für die automatische Detektion der KI waren der A-Punkt, Articulare, Gonion, Porion und Weichteil-Pogonion. Auffällig ist hierbei, dass sich mit Außnahme von Gonion die Punkte gänzlich in den beiden Studien unterschieden. Ob dies wirklich an der Funktionsweise und Modifikation der KI-Systeme oder durch die verschiedenen Datensätze und deren Quantität bedingt ist, müsste durch

eine direkte Gegenüberstellung beider Algorithmen mit einem gleichen Trainings- und Testdatensatz eruiert werden.

Die Studie von Hwang et al. (59) untersuchte ebenfalls den klassischen YOLOv3-Algorithmus, wie auch Bulatova et al. (55). Die Ergebnisse zeigten, dass die KI bei wiederholten Versuchen immer identische Positionen für jede Landmarke erkannte, während die menschliche Intra-Rater-Reabilität einen Detektionsfehler von 0,97 ± 1,03 mm aufwies. Der MRE (Mittlerer radialer Fehler) betrug 0,9 mm mit Ausnahme der Wurzelspitze des unteren Inzisors. Die Ergebnisse zeigen für diese KI eine vergleichbare Genauigkeit, wie für menschliche Prüfer*innen, ihre Detektionen sind jedoch konstanter. Die Genauigkeit der KI wird nicht verschlechtert durch Parameter wie Geschlecht, Bildqualität und metallische Artefakte. Da die Studie nicht explizit die Erfolgsdetektionsrate in einem Radius von 2 mm benennt, ist der direkte Vergleich mit den anderen Studien schwierig, jedoch zeigt sich, dass eine sehr gute Leistung trotz einer hohen Landmarkenzahl von 80 erzielt werden konnte. Laut der Studie von Moon et al. (61) hat die Anzahl der Landmarken direkten Einfluss auf die Leistung eines Systems.

Ein Jahr später wurde eine nachfolgende Untersuchung von Hwang et al. (60) veröffentlicht. Dabei wurde eine modifizierte Form des YOLOv3-Algorithmus eingesetzt. Die KI erreichte eine Erfolgsdetektionsrate (Success Detection Rate, SDR) von 75,5 % innerhalb einer Toleranz von 2 mm und eine Erfolgsklassifizierungsrate (Success Classification Rate, SCR) von 81,5 %. Dies sind vergleichbare Werte zu den Studien von Bulatova et al. und Zhao et al., welche ebenfalls 16 - 19 Landmarken detektierten, jedoch kleinere Datensätze gebrauchten. In einigen Analysemaßen zeigte die KI eine überlegenere Klassifizierungsrate im Vergleich zu menschlichen Prüfer*innen, was das große Potential der KI für den klinischen Einsatz unterstreicht.

Die Forschungsarbeit von Moon et al. (61) zielte darauf ab, die notwendige Quantität an Trainingsdaten zu ermitteln, um eine Künstliche Intelligenz für die automatische Landmarkenidentifizierung mit ausreichender Zuverlässigkeit zu trainieren. Sie war die Vorläuferstudie von Hwang et al. Die Analyse offenbarte einen linearen Anstieg der

Modellgenauigkeit in Korrelation mit steigender Anzahl von Trainingsdaten, sowie eine Zunahme der MRE-Werte bei einer steigenden Anzahl von Detektionszielen. Die Auswertung ergab, dass mindestens 2300 Trainingsbilder notwendig sind, um eine KI zu generieren, deren Genauigkeit der menschlichen Prüfer*innen gleichkommt. Keine der bisherigen beschriebenen Studien (59, 55, 56, 60) nutzten demnach genügend Trainingsbilder, um die optimale Leistung der Algorithmus generieren zu können.

Einige Studien untersuchten die Leistungsfähigkeit von ResNet (Residual neural Network) und verschiedenen Modifikationen dessen. Die Zahlenangabe die ResNet nachgestellt ist, zum Beispiel „ResNet50“, gibt die Anzahl der Schichten im Netzwerk an.

In der Studie von Yu Song et al. (62), wurde die Leistung eines CNN mit einer ResNet50-Architektur untersucht. Die SDR betrug 62 % innerhalb von 2 mm und die SCR 77,95 %.. Dies stellt die bisher schlechteste KI-Leistung dar. Auffällig ist der kleine Trainingsdatensatz, der nach Moon et al. (61) nicht ausreichend ist, um akzeptable Werte zu generieren. Die Kombination aus ROI-Extraktion und der Nutzung des ResNet50-Algorithmus stellte laut der Autor*innen dennoch zufriedenstellende Ergebnisse bezüglich der Landmarkendetektion dar. Auch die Verfasser*innen betonen die begrenzte Größe des Trainingsdatensatzes, sowie die längere Rechenzeit von durchschnittlich 6 Minuten pro KI-Analyse.

Für die Studie von Fulin Jiang et al. (63) wurde CephNet, ein zweistufiges CNN, zur Landmarkendetektion entwickelt. Es wurden 9870 Fernröntgen aus 20 verschiedenen medizinischen Einrichtungen gesammelt (9611 zum Training, 259 zur Testung). Fünf Kieferorthopäd*innen annotierten 30 Landmarken. Über 100 Kieferorthopäd*innen wurden rekrutiert, um die ersten Detektionen der KI neu zu beurteilen und CephNet damit zu verfeinern und neu zu trainieren. Die durchschnittliche Vorhersagegenauigkeit betrug 0,94 ± 0,74 mm. Die SDR innerhalb 2 mm betrug 91,73 % und die SCR 89,33 %. Damit zeigt CephNet eine hohe Genauigkeit und klinische Anwendbarkeit und übertraf die Leistung aller vorigen Studien. Dies lässt sich auf die Verwendung eines extrem großen und heterogenen Trainingsdatensatz mit 9870 Fernröntgen zurückführen und übertrifft die

Empfehlung von Moon et al. (61). Anzumerken ist ebenfalls, dass ca. 10 Landmarkenpunkte mehr als bei den vorigen Studien detektiert wurden. Nach Moon et al (61) müsste auch dies einen negativen Einfluss auf die Detektionsrate haben, umso imponierender ist die hohe SDR von 91,73 %. Eine weitere Besonderheit in der Durchführung dieser Studie ist die Verfeinerung des ersten KI-Outputs zu neuen Trainingszwecken. Möglicherweise führte dieses Vorgehen zu den sehr guten Ergebnissen und könnte in weiteren Studien erprobt und etabliert werden.

In der Studie von Ravi Kumar Mahto et al. (64) wurden Fernröntgenanalysen auf der KI-gesteuerten Plattform „WebCeph“ durchgeführt. Alle Messungen zeigten einen ICC über 0,75 und für 7 Parameter war der ICC > 0,9. Laut der Autor*innen deutet dies darauf hin, dass WebCeph zuverlässige Messungen liefert.

Die Studie von Huayu Ye et al. (65) untersuchte die Genauigkeit von drei KI-gestützten Programme: MyOrthoX, Angelalign und Digident. Die Detektionsraten innerhalb von 2 mm lagen zwischen 87,53 % und 93,09 % für alle Systeme. Die besten Werte für die Erfolgsdetektionsrate erzielte Angelalign mit einem durchschnittlichen Detektionsfehler von 0,80 ± 0,26 mm. Damit liefert Angelalign vergleichbare Werte wie CephNet in der Studie von Fulin Jiang et al. (63), bei einer ähnlichen Zahl von Landmarkenpunkten. Laut der Autor*innen liefern alle KI-Systeme klinisch akzeptable Werte, betont wurde des Weiteren die deutliche Zeitersparnis bei der vollautomatisierten Landmarkendetektion im Vergleich zur manuellen Methode.

In der Studie von Ioannis A Tsolakis et al. (66) wurden Fernröntgen mit der KI-unterstützen Software CS Imaging V8-Software analysiert. Es zeigt sich eine gute Übereinstimmung zwischen menschlicher Auswertung und KI mit einem ICC-Wert von 0,70 - 0,92, was der Auswertung von Ravi Kumar Mahto et al. (64) mit WebCeph entspricht. Laut der Studie werden ICC-Werte über 0,75 als gute und über 0,9 als ausgezeichnete Übereinstimmungen interpretiert. Statistisch signifikante Unterschiede zwischen den beiden Detektiermethoden zeigte sich bei den Messungen für FMA, L1-MP,

ANS-PNS/GoGn und U1-L1. Die automatische Fernröntgenanalyse durch CS Imaging V8-Software scheint zuverlässig und genau zu sein.

In der Studie von Jeong-Hoon Lee et al. (67) wurde ein neues Framework zur Landmarkendetektion entwickelt, unter Verwendung von Bayesian Convolutional Neural Networks. Die Inter-Rater-Reabilität der Expert*innen, die den Goldstandard definierten, wurde überprüft und lag bei 2,02 ± 1,53 mm. Dies dürfte als Schwachstelle dieser Studie auffallen, denn schon die gesetzten Landmarkenpunkte, der beiden Prüfer*innen, die den Goldstandard definierten lagen im Mittel über 2 mm auseinander, was nicht mehr als klinisch akzeptabel gelten würde. Die SDR in einem Bereich von 2 mm lag bei 82,11 %, was im Vergleich zu den anderen Studien ein solides Ergebnis darstellt. Laut der Autor*innen lässt dies darauf schließen, dass die KI Potential als computergestütztes Diagnosetool in der klinischen Praxis besitzt.

In der Studie von Min-Jung Kim et al. (68) wurde ein System basierend auf mehrstufigen CNNs (6 Schichten) entwickelt und überprüft. Der durchschnittliche Lokalisierungsfehler (MRE) der KI-Analyse lag bei 1,03 ± 1,29 mm und die SDR lag bei 87,13 % innerhalb von 2 mm. Es wurden Fernröntgenbilder auf zwei verschiedene Arten aus DVT-Aufnahmen generiert, dabei ergab die Art der Bildgebung keinen signifikanten Einfluss auf die Genauigkeit der KI-Detektionen. Nur Gonion zeigte MRE-Werte von über 2 mm. Eine SDR von 87,13 % stellt eine verhältnismäßig hohe Detektionsraten dar. Auffallend ist auch ein überdurchschnittlich großer Trainingsdatensatz, wobei dieser immer noch unter dem von Moon et al. (61) empfohlenen 2300 Fernröntgenbildern liegt. Diese Studie war die einzige, die Fernröntgen aus DVT Bildern generierte und die Landmarkendetektion an solchen erprobte. Dies könnte strahlungssparend genutzt werden, wenn bei Patien*innen ein DVT im Vorfeld vorliegt.

Jaerong Kim et al. (69) verwendeten eine mehrstufige CNN-Struktur zur Landmarkendetektion. Die MRE der CNN-Struktur betrug 1,36 ± 0,98 mm und liefert im direkten Vergleich zu den anderen Studien ein mittelmäßiges Ergebnis. Die Autor*innen betonen, dass die Leistung der KI bessere Werte erzielte, als Kieferorthopäd*innen ohne

große Erfahrungswerte. Zu diesem Schluss, dass die KI die Leistung von menschlichen Untersucher*innen übertreffen kann, kamen auch schon die Studien von Hwang et al. (59, 60). Die Wahl des Röntgengerätes und des Sensortyps hatten signifikanten Einfluss auf die Landmarkenidentifizierungsgenauigkeit. Im Gegensatz zu den anderen Studien haben hier die Forscher*innen Wert auf Vielfalt bei der Auswahl der Röntgengeräte gelegt und diese Studie zeichneten sich im Vergleich durch einen großen Trainingsdatensatz aus.

In der Studie von Chihiro Tanikawa et al. (70) wurden 2 vorab entwickelte CNN-Strukturen verwendet (CNN-PE und CNN-PC). Anders als bei den anderen Studien gab es keine strengen Einschlusskriterien der Patient*innen. Es wurden auch Bilder von Patient*innen mit Deformitäten, Apparaturen und Lippen-Kiefer-Gaumenspalten inkludiert. Die SDR-Werte lagen zwischen 85 % und 91 % innerhalb 2 mm, je nach Gruppe. Es stellte sich heraus, dass das Vorhandensein einer Lippen-Kiefer-Gaumen-Spalte die Landmarkenidentifikation am meisten beeinflusste. Andere Faktoren, wie das Zahnalter, das Tragen von kieferorthopädischen Apparaturen und der Überbiss zeigten keinen signifikanten Einfluss auf die Identifizierung. Dies zeigt Forschungsschwerpunkte für weitere Modifizierungen an den Algorithmen auf. Zahnbezogene Landmarken zeigten ebenfalls eine niedrigere Erfolgsrate bei der Detektion. Auffallend ist, dass in der Studie von Tanikawa et a. (69) trotz der hohen Landmarkenanzahl (n = 26) und der Inkludierung verschiedener Pathologien bei der Fernröntgenauswahl sehr gute Genauigkeitswerte erzielt werden konnten. Dies könnte auf einem relativ hohen Trainingsdatensatz von 1755 Bildern zurückzuführen sein.

Auch die Studie von Teodora Popova et al. (48) untersuchte die Einflussfaktoren auf die Landmarkendetektion mit einer vorab entwickelten CNN-Struktur. Die SDR betrug insgesamt 84,73 % innerhalb von 2 mm und erzielte damit im Vergleich mittelmäßige Werte. Es ist möglich, dass der Einschluss verschiedener Pathologien oder Merkmale zu einer Verschlechterung der Ergebnisse geführt haben könnte. Der MRE lag bei 1,47 ± 1,06 mm. Die Studie kommt zu dem Schluss, dass Wachstumsstrukturen und Entwicklungsstadien die Leistung der CNN nicht beeinflussen, während festsitzende

kieferorthopädische Geräte wie Brackets oder Bänder einen signifikanten Einfluss auf die Leistung des Modells haben.

Die Studie von Sumer Panesar et al. (71) zielte darauf ab, den Einfluss künstlicher Intelligenz auf die Verbesserung der Präzision und Genauigkeit der Fernröntgenanalysen bei Kieferorthopäd*innen verschiedener Erfahrung zu untersuchen.

Die KI wies den höchsten ICC-Wert auf (0,97), dies spiegelt das Ergebnis voriger Studien wider, bei welchen wiederholt festgestellt wurde, dass die KI eine höhere Konsistenz als menschliche Prüfer*innen aufwies (60, 61, 69). Die Genauigkeit der Expert*innen (kieferorthopädisches, assistenzärztliches, zahnärztliches und studentisches Fachpersonal) erhöhte sich KI-unterstützt jeweils um, 12,74%; 19,10%; 35,69% und 33,69% (durchschnittlich: 27,27%). Die Genauigkeit wurde signifikant von weniger erfahrenen zahnmedizinischen Fachkräften auf das Niveau erfahrener Kieferorthopäd*innen verbessert.

Die Studie von Huang-Ting Lee et al. (72) nutzte eine CNN-Struktur unterstützt durch MobileNetV2 und U-Net. Zwei Modelle wurden mit Daten unterschiedlicher Einschlusskriterien trainiert. Modell 2 generierte bessere Werte, so lag die SDR innerhalb von 2 mm bei 83,14 %. Die KI lieferte zuverlässige Werte, es scheint als könnten bestimmte Parameter, wie übermäßige konservierende Restaurationen, Einfluss auf die Trainingsleistung einer KI haben. Dies ist die einzige Studie, die den Einfluss von Bildauswahlkriterien im direkten Vergleich untersuchte. Es scheint, als würde eine strengere Auswahl beim Trainingsdatensatz zu einer Verbesserung der Leistung des Systems führen.

Die Studie von Ho-Jin Kim et al. (73) beschäftigte sich mit der Klassifizierung der sagittalen skelettalen Beziehung durch ein CNN-Modell auf Fernröntgenbildern. Im Vergleich mit einer automatisierten Tracing-KI-Software erlangte die CNN-Struktur bessere Werte mit einer Genauigkeit von 93 %. Dies war die einzige Studie, die keine umfassende Fernröntgenanalyse durchführte und sich nur auf eine sagittale Klassifizierung

konzentrierte. Die sehr guten Werte sind durch die Reduktion der Komplexität dieser Studie erklärbar. Ein klinischer Nutzen ist jedoch fraglich, da in der Praxis die Beurteilung der sagittalen Relation in Form einer Blickdiagnose direkt an Patient*innen stattfindet.

In dem Paper von Moshe Davidovitch et al. (74) wird die Plattform Algoceph (CephX) verwendet, um die Landmarkendetektion durchzuführen. Nur eine Landmarke (SoftpogY) lag außerhalb eines akzeptablen Fehlers von 2 mm. Da nicht bekannt ist, wie viele Daten Algoceph zu Trainingszwecken verwendete, ist es schwer diese Leistung einzuordnen. Jedoch stellt eine Detektionsrate von 97,6 % den bisher besten Wert dar. SoftpogY ist eine Weichgewebslandmarke, weshalb es nicht überrascht, dass jene am schwersten zu detektieren war. Hartgewebe, wie Knochen oder Zähne, sind in Röntgenaufnahmen deutlicher zu erkennen als weiche Gewebe.

In der Studie von Thaisa Silvia et al. (75) wurde die Software Cephbot zur Landmarkendetektion benutzt. Der ICC-Wert der KI war >0,94 und unterschied sich nicht statistisch signifikant zu menschlichen Prüfer*innen. Glabella zu Subnasale konnte von Cephbot nicht gemessen werden, wobei es im Vorfeld auch nicht auf diese Messung trainiert wurde. Die in der Studie erzielten Ergebnisse weisen eine herausragende Übereinstimmung auf, mit einem ICC-Wert, der nahe an 1 liegt und damit eine fast perfekte Übereinstimmung anzeigt. Dies ist besonders bemerkenswert, wenn man die große Anzahl von detektierten Landmarkenpunkten (n = 66) in Betracht zieht, die zu den höchsten unter den analysierten Studien zählt. Jedoch stellt die Studie keine Hintergrundinformationen zum Trainingsverfahren von Cefbot bereit, was die Kontextualisierung und Analyse der Ergebnisse im Vergleich zu anderen Studien schwierig gestaltet.

Die Studie von Liciane dos Santos Menezes et al. (76) verwendet ebenfalls die Cefbot-Software und ist die nachfolgende Studie von Thaisa Silvia et al. (75). Hierbei wurde die Leistung von Cephbot überprüft unter Berücksichtigung von verschiedenen Helligkeits- und Kontrasteinstellungen. Die Zuverlässigkeit wurde überprüft, und sowohl für die menschlichen Prüfer*innen als auch für die KI als ausgezeichnet befunden (ICC>0,91).

Dies bedeutet, dass eine konstante Positionierung der Landmarken in verschiedenen Durchläufen möglich war. Jedoch war die Reproduzierbarkeit der Positionierungen der Landmarken durch verschiedene Bildanpassungen beeinträchtigt. Das Cefbot-System hatte wesentlich größere Schwierigkeiten als menschliche Prüfer*innen sich auf die Bildanpassungen einzustellen. Vor allem eine niedrige Helligkeit und ein hoher Kontrast beeinflussten Cefbots Reproduzierbarkeit in dieser Studie stark. Möglicherweise sind KI-Systeme weniger anpassungsfähig auf veränderte Umstände. Auch schon Chihiro Tanikwaka et al. (70) fand einen signifikanten Unterschied der Landmarkendetektion bei der Verwendung verschiedener Röntgengeräte in seiner Studie. Diesen Einflussfaktor sollten zukünftige Studien in Betracht ziehen.

In allen Studien hat sich gezeigt, dass die durch künstliche Intelligenz gesteuerte Detektion von Landmarken in Fernröntgenbildern eine vielversprechende Entwicklung darstellt, die das Potential besitzt, die Analyse von Fernröntgenbildern erheblich zu bereichern. Einige Studien zeigten sogar, dass die KI in der Lage dazu ist, menschliche Leistung zu übertreffen oder jene von weniger erfahrenen Kieferorthopäd*innen zu steigern und auf das Niveau von erfahrenen Kieferorthopäd*innen zu bringen (55, 59, 60, 69, 62, 68, 71). Des Weiteren wies die KI in vielen Studien eine hohe Konsistenz in der identischen Positionierung der Landmarken in mehreren Durchläufen auf (Intra-Reater-Reabilität oder ICC), die teilweise die der Expert*innen übertraf (58, 59, 60, 71).

Ein wesentlicher Aspekt in der Validierung von KI-Systemen ist die Festlegung eines Goldstandards. Einige Studien verwendeten für die Definition dessen den Mittelwert der Landmarkenposition mehrerer Expert*innen (61,66, 68), Konsensentscheidungen zwischen Untersucher*innen (70), oder nur ein*e einzelne*r Kieferorthopäd*in diente als Referenz (63).

Die Studie von Hwag et al. (58) zeigt, dass menschliche Untersucher*innen und das KI-System ca. gleichweit vom definierten Goldstandard entfernt sind. Dies spiegelt die Schwierigkeit mit der Benennung des Goldstandards wider. Es ist faktisch nicht möglich einen allgemeingültigen Goldstandard zu definieren, da Kieferorthopäd*innen nie die exakt gleichen Landmarkenpunkte markieren würden.

Die Erkennungsgenauigkeit künstlicher Intelligenz kann durch Variationen in der Bildqualität, sowie durch Anpassungen von Helligkeit und Kontrast signifikant beeinträchtigt werden (76). Es wird darauf hingewiesen, dass solche Anpassungen die Performance von KI-Systemen stärker negativ beeinflussen könnten, als die Analysefähigkeit menschlicher Prüfer*innen (76). Dies lässt vermuten, dass KI-Systeme möglicherweise weniger flexibel auf variierende Bedingungen reagieren. Kim Jungs et al. (69) Untersuchung zeigte, dass die Bildgebungstechnik keinen merklichen Einfluss auf die KI-Resultate hatte. Dies steht im Kontrast zu den Ergebnissen der Studie von Chihiro Tanikawa et al. (70). Anzumerken ist auch, dass die meisten der Studien Ausschlusskriterien für die Röntgenbilder im Trainings- und Testdatensatz definierten. Dies spiegelt die klinische Praxis nur bedingt wider. Chihiro Tanikawa (70) et al. hingegen bezog in seiner Studie bewusst auch Röntgenbilder von Patient*innen mit Deformitäten, kieferorthpädischen Apparaturen oder Lippen-Kiefer-Gaumen-Spalten ein. Dies könnte sich auf die Leistung der KI niederschlagen. Des Weiteren stammte der Datensatz der meisten Studien aus einer homogenen Population, dadurch bedingt, dass sie rein aus einzelnen Kliniken in bestimmten Ländern stammten. Damit wird nicht nur eine gewisse Homogenität der Daten der Patient*innen hervorgerufen, sondern auch die Verwendung gleicher Röntgengeräte und Behandlungsanwendungen. Aus diesem Grund ist die allgemeingültige Anwendbarkeit der Studienergebnisse kritisch zu betrachten. In der Studie von Chihiro Tanikawa et al. (70) wurde speziell darauf geachtet Bilder von verschiedenen Röntgengeräten mit verschiedenen Sensortypen zu verwenden. Die Ergebnisse zeigten, dass dies einen Einfluss auf die Leistung der KI-Struktur haben könnte.

Mit der Frage nach der Größe des Trainingsdatensatzes hat sich Moon et al. (60) in seiner Studie beschäftigt. So hatten einige Studien einen deutlich geringeren Datensatz als die von ihm empfohlenen 2300 Röntgenbilder. Außerdem zeigte sich ein Abfall in der Trainingsleistung mit steigender Landmarkenzahl, dies sollte in zukünftigen Studien mit in Betracht gezogen werden.

Abschließend ist noch anzumerken, dass die meisten Studien eine positive Landmarkendetektion innerhalb von 2 mm definierten, jedoch wurde der Einfluss des Abweichens einzelner Landmarken und der Einfluss der Stärke der Abweichung, auf die Behandlungsplanung nicht untersucht.

Einige Forscher*innen präsentierten schon ein Jahr später Folgestudien, mit modifizierten KI-Systemen oder angepassten Versuchsdurchführungen. Dies spiegelt den rasanten Fortschritt in diesem Themenfeld wider und lässt auf bedeutende Entwicklungsschritte in naher Zukunft hoffen.

4.2 KI-generierte Klassifizierung der Cervical Vertebral Maturation

Die Klassifizierung des Zervikalen Reifestadiums behandelten 6 Studien. Dabei stellten sich die Versuchsaufbauten ähnlich denen der Landmarkendetektion dar.

In der Studie von Jing Zhou et al. (77) wurde eine CNN-Struktur vorab entwickelt.

Die Inter-Rater-Reabilität bei der Landmarkendetektion war zwischen KI und Mensch leicht besser als zwischen den menschlichen Prüfer*innen untereinander. Der ICC der KI zum Goldstandard bei der Landmarkendetektierung an den Wirbelkörpern betrug 98 %, die Genauigkeit der CVM-Stadieneinstufung lag bei 71 %. Dies spiegelt wider, dass die KI durchaus in der Lage dazu ist, wichtige anatomische Punkte an den Wirbelkörpern zu erkennen und zu markieren, jedoch Schwierigkeiten hat diese Informationen weiter zu interpretieren. Das CS6-Stadium erreichte die höchste Genauigkeit mit 85 %. CS3 erwies sich als am schwierigsten klassifizierbar (31 %), ist gemeinsam mit CS4 jedoch am relevantesten für die Behandlungsplanung, da dieses Stadium den maximalen Wachstumsschub darstellt. Die Autor*innen kamen zu dem Schluss, dass die KI ein nützliches Werkzeug bei der Bewertung der zervikalen Wirbelreifung darstellt.

Die Studie von Haizhen Li et al. (78) verglich 4 Algorithmen (VGG16, GoogLeNet, DenseNet161, ResNet152) zur Bestimmung der CVM-Stadien. ResNet152 zeigte die besten Ergebnisse mit einer Gesamtgenauigkeit von 86,06 % und übertraf damit die KI-Struktur von Jing Zhou et al. (77). Jedoch gebrauchte Haizhen Li et al. (78) einen deutlich größeren Datensatz, was zu der Leistungssteigerung geführt haben könnte. Die F1-Werte für die Stadien waren: CS6 > CS1 > CS4 > CS5 > CS3 > CS2, was bedeutet, dass die KI-Strukturen CS6 am besten klassifizieren konnten. Wie in der Studie von Jing Zhou et al.

(77) stellte sich CS6 am besten klassifizierbar dar und CS3 bildete gemeinsam mit CS2 das Schlusslicht.

Die Studie von Hyejun Seo et al. (79) verglich die Leistungen von 6 vortrainierten KI-Strukturen (ResNet-18, MobileNetv2, ResNet50, ResNet101, Inceptionv3, Inception-Resnetv2). Alle Modelle zeigten eine Genauigkeit von > 90 %, wobei Inception-ResNetv2 (94 %) die besten Ergebnisse lieferte. Die Kombination aus beiden Netzwerkarchitekturen scheint deren Vorteile zu kombinieren und besonders effektiv in der CVM-Klassifizierung zu sein. Im Vergleich zu den anderen Studien generierte diese Studie die genauste Klassifizierungsleistung. Bei der Grad-CAM-Analyse zeigte Inception-ResNetv2 Aktivitäten bei mehreren Wirbelkörpern gleichzeitig zur Klassifizierung eines Stadiums. Dies scheint effektiver zu sein, als nur einzelne Wirbelbereiche zu fokussieren. Die AUC-Werte für CS1-CS6 lagen zwischen 0,935 und 0,994. Der höchste AUC-Wert zeigte sich bei CS1, der niedrigste bei CS3, was darauf hindeutet, dass die Klassifizierung des CS1-Stadiums der KI am leichtesten und des CS3-Stadiums am schwersten viel. Dieser Trend spiegelte sich schon in den anderen Studien wider. Es wird spekuliert, dass aufgrund des aktiven Wachstumsmusters während des CS3-Stadiums, vermehrt Variationen in der Morphologie der Halswirbel auftreten können und deshalb die Klassifizierung erschwert sein könnte.

Die Studie von Salih Furkan Atici et al. (80) verwendete AggregateNet zur Ermittlung der Wirbelreifegrade. Es wurden Kantenhervorhebungsfilter verwendet und der Einfluss auf die KI-Genauigkeit geprüft. Die Ergebnisse der Studie zeigten, dass die höchste Klassifizierungsgenauigkeit mit der Verwendung von Datenaugmentationen, Kantenhervorhebungsfiltern und der Altersbekanntgabe erreicht wurden. Die größte Verbesserung wurde durch Datenaugmentationen erreicht (12 - 15 %). Diese Studie war die einzige, die auf ein ausgewogenes Geschlechterverhältnis achtete. Die Genauigkeitswerte befinden sich jedoch im Vergleich mit den anderen Studien im unteren Bereich. Interessant wäre, ob der Gebrauch eines Kantenhervorhebungsfilters auch mit anderen KI-Strukturen und Versuchssettings zu einer Verbesserung der Genauigkeit führen würde und generell in die CVM-Klassifizierung etabliert werden sollte.

Die Studie von Hairui Li et al. (49) verwendet ein Bewertungssystem für die zervikale Wirbelreifung namens psc-CVM-Assessment. Es besteht aus einem Positionierungs-, Formerkennungs- und einem CVM-Bewertungsnetzwerk. Es wurde der mit Abstand größte Datensatz unter den Studien verwendet. Der AUC-Wert betrug 0,94 und eine Gesamtgenauigkeit von 70,42 % wurde erreicht. Dies bedeutet, dass das System eine gute Unterscheidungsfähigkeit zwischen den unterschiedlichen CVM-Stadien besitzt. Jedoch bedeutet dies nicht zwangsläufig, dass auch die richtigen CVM-Klassen zugeordnet werden, was den Wert der Genauigkeit wiederspiegelt. Der ICC zwischen dem System und des Expertengremiums betrug 0,946, was eine hohe Konsistenz in der Bewertung der KI ausdrückt. Die F1-Werte stellten sich in der Reihenfolge

CVS6 > CVS1 > CVS4 > CVS5 > CVS3 > CVS2 dar, was die Ergebnisse aller anderen Studien wiederspiegelt. Laut der Autor*innen generierte die KI gute Werte und kann als diagnostische Hilfe verwendet werden.

Die Studie von Eun-Gyeong Kim et al. (81) nutzt eine für die Studie entwickelte CNN-Struktur und überprüft drei verschiedene Algorithmusvarianten. Das dritte und komplexeste 3-Schritt-Modell erreichte die besten Ergebnisse mit einer Genauigkeit von 62,5 %. Damit wurden in dieser Studie die schlechtesten Klassifizierungswerte generiert. Dies könnte durch einen relativ kleinen Datensatz bedingt sein.

Ähnlich wie bei der Landmarkendetektion finden sich potentielle Schwächen der Studien bei der Generalisierbarkeit des Goldstandards, sowie in der Auswahl, Vielfalt und Größe der Datensätze. Die Klassifizierung des CVM-Stadiums lässt selbst einen Spielraum zur Interpretation, wenn die Stadien sich im Übergang befinden und unterliegt deshalb zum Teil einer subjektiven Bewertung. Die automatische Klassifizierung des CS2 und CS3-Stadiums stellte sich als größte Herausforderung dar, wohingegen CS6 und CS1 leicht Stadien zugeordnet werden konnten. CS3 repräsentiert oft den Zeitpunkt eines intensiven Wachstumsschubes, wodurch die morphologischen Ausprägungen der Halswirbel variabel sein können. Des Weiteren ist für CS3 als einziges Stadium eine Ausnahme definiert, was wiederum die Vielfalt der möglichen Morphologien widerspiegelt. Ein weiterer

Erklärungsansatz könnten die subtilen morphologischen Veränderungen zwischen CS2 und CS3 darstellen. Wohingegen bei CS6 deutliche Veränderungen und Merkmalsausprägungen für die KI leichter zu erfassen sein könnten. Abschließend ist anzumerken, dass CS2 und CS3 Übergangsphasen darstellen, die Elemente der vorherigen und nachfolgenden Stadien enthalten. Möglicherweise erschwert dies die Zuordnung der KI.

Zusammenfassend birgt die Klassifizierung des CVM-Stadiums durch KI einen vielversprechenden Ansatz, auch wenn die Systeme noch weit von einer einwandfreien Performance entfernt sind.

4.3 KI-generierte Extraktionsentscheidungen

Vier Studien untersuchten die strategische Extraktionsnotwendigkeit und generierten mögliche Extraktionsmuster durch KI-Strukturen.

In der Studie von Jiho Ryu et al. (82) wurden 4 KI-Strukturen verwendet: ResNet50, Resnet101, VGG16 und VGG19. Dabei überprüften die CNN-Modelle Okklusionsfotografien nach Zahnengständen und Extraktionsnotwendigkeiten und klassifizierten die Engstände in 3 Stadien. VGG19 lieferte die besten Ergebnisse mit einer Genauigkeit von 0,91 und einem AUC-Wert von 0,952. Damit erzielt die KI sehr gute Werte, was angesichts einer geringen Komplexität von binären Entscheidungsfragen nicht überrascht. Für die Klassifizierung des Crowdings zeigte ebenfalls VGG19 die besten Werte mit einem gewichteten Kappa-Wert von 0,73 und damit eine substanzielle Übereinstimmung zwischen Goldstandard und KI. Des Weiteren ist anzumerken, dass in der klinischen Praxis Extraktionsentscheidungen nicht rein durch Fotografien getroffen werden und die Ausschlusskriterien der involvierten Fälle der Patient*innen sehr strickt waren, was eine Homogenität des Datenpools zur Folge hatte. Dies könnte die Allgemeingültigkeit der KI-Leistung in Frage stellen und wäre unter anderen Bedingungen womöglich nicht reproduzierbar.

Die Studie von Lily Etemad et al. (83) nutzte zwei Algorithmen (RF, Random Forrest und Multilayer Preceptron) die bereits in vorangestellten Studien erprobt wurden. Es wurde eine binäre Entscheidung für oder gegen die Notwendigkeit einer Extraktion getroffen. Die Genauigkeit der Fälle betrug 75 - 79 %, der AUC-Wert betrug 79 - 82 %, wobei Durchgänge mit weniger Eingabemerkmalen die besseren Werte erzielten. Laut der Autor*innen deuten die Ergebnisse darauf hin, dass die Leistung bestehender Modelle verbessert werden kann, wenn inkongruente Datenmuster erkannt und für die Schulung von Modellen separat behandelt werden. Überraschenderweise erzielten CNN-Strukturen mit weniger Informationseingaben bessere Ergebnisse. Vergleicht man die beiden Studien miteinander, so ist die Leistung der KI von Jiho Ryu et al. (83) um 12 % besser, wobei der Trainingsdatensatz auch deutlich größer ist. Obwohl bei Lily Etmad et al. (84) auch das Ausmaß des Engstands Teil des KI-Inputs war, schien die Komplexität der Eingabemerkmale die KI eher zu verwirren. Allerdings stellten diese echte Patient*innenfälle dar und damit auch die klinische Komplexität von Extraktionsfällen. Wohingegen in Jiho Ryus et al. (83) Studie die Expert*innen ihre hypothetische Entscheidung rein auf intraoralen Fotografien stützten. Dadurch war der KI-Input stark vereinfacht, was möglicherweise zu einer höheren Genauigkeit in den Ergebnissen führte.

In der Studie von Landon Leavitt et al. (84) wurden 3 Algorithmen zur Vorhersage von Extraktionsmustern verwendet. Random Forrest (RF), Logistische Regression (LR) und Support Vector Machine (SVM). RF lieferte die besten Ergebnisse mit einer geringen Gesamtgenauigkeit von 54,55 %. Die Klassengenauigkeit für Extraktionsmuster war am höchsten für die Extraktion der 1. Prämolaren im Unterkiefer und Oberkiefer (81,63 - 63,27 %). Gefolgt von dem Extraktionsmuster der oberen 1. Prämolaren (72,22 - 61,11 %). Alle anderen Extraktionsmuster zeigten Genauigkeiten von 0 - 36 %. Die Ergebnisse der Studie verdeutlichen, dass trotz einer guten Vorhersage bestimmter Extraktionsmuster, die KI-Systeme Schwierigkeiten hatten, alle Muster korrekt zu identifizieren, was etwa zur Hälfte der Patient*innen zutraf. Eine deutliche Schwäche dieser Studie war die ungleiche Verteilung verschiedener Extraktionsmuster im Trainingsdatensatz. Dies könnte sich auf die Klassifizierungsgenauigkeit niederschlagen und wäre eine Erklärungsmöglichkeit für die unterschiedliche Leistung bei verschiedenen Extraktionsmustern. Trotz der zahlreichen

und verschiedenen Inputparameter konnten vergleichsweise nur schlechte KI-Leistungen erbracht werden.

Auch die Studie von Suhail Yasir et al. (85) beschäftigte sich mit der Generierung von Extraktionsmustern. Es wurde Random Forrest verwendet. Die Übereinstimmung zwischen den Expert*innen lag zwischen 65 % und 71 %, was die Komplexität und Vielfalt in der Extraktionstherapie wiederspiegelt. Die Genauigkeit der KI betrug 75 % und liegt somit nahe am Grad der Übereinstimmung zwischen verschiedenen Expert*innen und könnte demnach als unterstützendes Diagnosetool betrachtet werden. Auch wenn die Genauigkeit deutlich höher ist als die von Leavitt et al. (85), weisen beide Studien einen relativ geringen Datenpool auf und die Ergebnisse müssten in weiteren Studien bestätigt werden.

Die vorliegenden Studien deuten darauf hin, dass die Verwendung von künstlicher Intelligenz in der Entscheidungsfindung für oder gegen Zahnextraktionen in der Kieferorthopädie eine wertvolle Ergänzung darstellen kann. Ähnlich wie in den anderen Untersuchungsthemen muss auch bei der strategischen Extraktionsentscheidung die Frage nach der Generalisierbarkeit, Definition des Goldstandards und Art und Vielfalt des Datenpools kritisch gestellt werden. Des Weiteren unterscheiden sich die Studien in der Merkmalseingabe. In der Studie von Jiho Ryu et al. (84) wurden ausschließlich fotografische Aufnahmen zur Unterstützung des Entscheidungsprozesses herangezogen, wobei das KI-System auf eine binäre Entscheidungsfindung reduziert wurde. Landon Leavitt et al. (85) hingegen integrierte eine Vielzahl von Datenpunkten und es flossen Fernröntgenaufnahmen, Modelle und Fotografien in die Analyse mit ein. Die KI generierte zudem ein spezifisches Extraktionsmuster. Bei allen Studien wurden im Vorfeld Ausschlusskriterien für den verwendeten Datenpool getroffen und Patient*innenfälle mit gewissen Pathologien und Anomalien wurden ausgeschlossen.

4.4 Erstellung eines KI-generierten Behandlungsplans

Vier Studien befassten sich mit der Generierung eines Behandlungsplans.

Die Studie von Peilin Li et al. (86) verwendete ein KI-System aus drei neuronalen Netzwerken. Dabei entscheidet das erste Netzwerk über eine Extraktion, die beiden weiteren sagen im Anschluss das Verankerungsmuster oder Extraktionsmuster voraus. Die Genauigkeit für die Extraktions-Nichtextraktions-Entscheidung für die KI betrug 94 % mit einem AUC-Wert von 0,982, einer Sensitivität von 94,6 % und einer Spezifität von 93,8 %. Dies spiegelt ähnliche Werte, wie die der Studie von Jiho Ryu et al. (83) wider. Die Genauigkeit des Extraktions- und Verankerungsmusters betrug 84,2 % und 92,8 %. Diese Werte sind überraschend gut, so konnten in der Studie von Landon Leavitt et al. (85) in etwa die Hälfte der Verankerungsmuster korrekt klassifiziert werden. Der Studienaufbau ist vergleichbar, nur verwendete Leavitt et al. (85) in etwa doppelt so viele Input-Parameter. Die wichtigsten Vorhersageparameter waren Verengung des Oberkiefers, ANB-Wert und Spee-Kurve. Die Autor*innen betonen, dass für weniger erfahrene Kieferorthopäd*innen die KI als Orientierungshilfe genutzt werden könnte. Ein weiterer Vorteil ist, dass verschiedene Behandlungsmodelle vorgeschlagen werden. Die Reproduzierbarkeit dieser exzellenten Ergebnisse sollte in zukünftigen Studien überprüft werden.

In der Studie von Yuujin Shimizu et al. (87) wurden zwei KI-Systeme entwickelt und bewertet, die jeweils eine priorisierte Problem- und Behandlungsliste erstellten. Keine weitere Studie folgte einem vergleichbaren Studienaufbau. Das KI-System erlangte eine Präzision von 65 % für das Identifizieren von Problemlisten (Subtask 1) und 48 % bei der Erstellung von Behandlungsplänen (Subtask 2). Die Erkennung von Problemen viel dem System also leichter, als die Erstellung von konkreten Behandlungsplanungen. Das System erzielte eine mittlere Rangfolge in der Peer-Bewertung für Subtaks 1, was bedeutet, dass es im menschlichen Vergleich eine durchschnittliche Leistung erbrachte. Für den Subtask 2 erbrachte die KI vergleichbare Behandlungspläne zu*r niedrigstbewerteten Kieferorthopäd*in. Dies deutet darauf hin, dass die Fähigkeit von KI-Systemen komplexe Behandlungspläne autonom zu erstellen noch verbessert werden muss, auch wenn sie in dieser Studie mit dem Niveau menschlicher Prüfer*innen mithalten konnte.

Das Ziel der Studie von Bhornsawan Thanathornwong et al. (88) war die Entwicklung eines Bayesianisches Netzwerkes zur Bewertung des Bedarfs an kieferorthopädischer Behandlung bei Patient*innen mit bleibendem Gebiss. Die KI-Struktur erreichte im Vergleich zu den Expert*innen sehr gute und damit klinisch akzeptable Werte: Genauigkeit 96 %, Sensitivität 95 %, Spezifität: 100 %. Diese Werte spiegeln eine sehr hohe Übereinstimmung mit dem Goldstandard wider. 95 % der Patient*innen wurden korrekt als solche mit Behandlungsbedarf identifiziert. Alle Patient*innen, die das System als nicht behandlungsbedürftig einstufte, hatten tatsächlich auch keine Behandlungsnotwendigkeit. Dies könnte bei der Vorabentscheidung von Zahnärzt*innen über die kieferorthopädische Notwendigkeit genutzt werden, um die Zuweisungsentscheidung zu erleichtern.

Die Studie von Jahnavi Prasad et al. (89) hatte zur Absicht 7 KI-Modelle auf ihre Genauigkeit in der kieferorthopädischen Behandlungsplanung zu überprüfen. Die KI-Systeme stellten eine Diagnose und generierten einen groben Behandlungsplan. Die durchschnittliche Genauigkeit im Vergleich zu den Behandlungsplänen von Expert*innen betrug 84 %, wobei Decision Tree, Random Forest und XGB die höchste Genauigkeit mit 87 - 93 % aufwiesen. Diese hohen Genauigkeitswerte könnten mit der Reduzierung des Behandlungsplans und der Aufteilung der Generierung des Plans in Teilschritten erklärt werden. Die Studien lassen sich untereinander schwer vergleichen, da sie sich stark in Studiendesign und Art des KI-Inputs unterscheiden. So wählte Yuujin Shimizu et al. (87) einen viel komplexeren Studienaufbau und verlangte differenziertere Ergebnisse. Peilin Li et al. (86) hingegen generierte konkrete Extraktions- und Verankerungsmuster.

Die kieferorthopädische Behandlungsplanung ist sehr komplex und vielfältig. Dies spiegelt sich auch in den Ergebnissen der Studien wider. Es scheint für KI-Systeme möglich zu sein, eine Entscheidung zur Indikation für oder gegen eine Kieferorthopädie mit einer relativ hohen Genauigkeit treffen zu können. Die Entwicklung komplexer Behandlungspläne durch künstliche Intelligenz hingegen stellt sich als herausfordernd dar und es bedarf weiterer Forschung, um die entscheidenden Inputparameter, sowie adäquate Versuchsaufbauten zu identifizieren. Obwohl einige Autor*innen bereits die

Einsatzfähigkeit von KI-Systemen als Entscheidungshilfe beschreiben, ist die Generalisierbarkeit dieser Aussage in weiteren Studien zu überprüfen.

4.5 KI-generierte Prognose der Behandlungsdauer

Die Studie von James Volovic et al. (90) hatte zum Ziel, ein KI-Modell zur Vorhersage der kieferorthopädischen Behandlungsdauer zu entwickeln. Ausschlusskriterien waren unter anderem eine chirurgische Notwendigkeit und Extraktionsfälle. Die Modelle waren Random Forest, Lasso und Elastic Net, wobei der absolute durchschnittliche Fehler bei 7,27 Monaten lag. Laut der Autor*innen waren alle KI-Modelle in der Lage die Behandlungsdauer in einem klinisch akzeptablen Bereich vorherzusagen. Die Behandlungsdauer in der Kieferorthopädie wird von vielen Faktoren beeinflusst. Es gibt keine allgemeingültigen Zeitrahmen, die eine Prognose als klinisch akzeptabel definieren würde. Viele Kieferorthopäd*innen setzen einen Zeitrahmen von etwa 2 Jahren für eine einfache und komplikationslose kieferorthopädische Behandlung fest, mit einer Abweichung von 7,27 Monaten scheint dies in Relation eine erhebliche Diskrepanz zu sein.

4.6 KI-generierte Diagnose der kieferorthopädischen chirurgischen Notwendigkeit

Vier Studien beschäftigten sich mit der Diagnose der kieferorthopädischen chirurgischen Notwendigkeit.

In der retroperspektiven Studie von WooSang Shin et al. (91) wurde ein KI-Modell entwickelt, um den Bedarf an kieferorthopädischer Chirurgie anhand von Röntgenbildern vorherzusagen. Das KI-System war in der Lage den Bedarf an kieferorthopädischer Chirurgie mit einer Genauigkeit von 0,954, einer Sensivität von 0,844 und einer Spezifität von 0,993 vorherzusagen. Es kommen mehr Patient*innen zu einer falsch-negativen Diagnose als zu einer falsch-positiven. Dennoch stellen sich die Werte als vielversprechend dar.

Die Studie von Ye-Hyun et al. (92) hatte zum Ziel die Leistung von ResNet-18, -34, -50 und -101 zur Diagnose von kieferorthopädischer Chirurgie zu überprüfen und damit den Einfluss der Tiefe eines Netzwerkes. Überraschenderweise nimmt die Leistung des KI-Modells mit der Tiefe der Netzwerke ab. So erzielte ResNet-18 die besten Ergebnisse mit einer Genauigkeit von 93,8 %, einem AUC-Wert von 0,979, einer Sensitivität von 0,882 % und einer Spezifität von 0,966 %. Tiefe Netzwerkstrukturen haben die Tendenz zu „overfitten“, sich also zu stark an Trainingsdaten anzupassen und dadurch die Allgemeingültigkeit zu verlieren. In der vorliegenden Studie wurden jedoch mehrere Maßnahmen unternommen, um Overfitting zu vermeiden. Weitere Erklärungsansätze der Autor*innen liegen in der Architektur der KI-Strukturen. In der Studie wird darauf hingewiesen, dass eine geringe Modelltiefe in einigen klinischen Anwendungen, vorteilhaft sein könnten, da die Modelle eine bessere Balance zwischen Lernfähigkeit und Generalisierung bieten. Im Vergleich zu der Studie von Woosang Shin et al. (92) sind die Genauigkeitswerte vergleichbar.

Die Studie von Natkritta Chaiprasittikul et al. (93) beschäftigt sich ebenfalls mit der Vorhersage einer OP-Entscheidung für kieferorthopädische Patient*innen mit Hilfe eines Multi-Layer Perceptrons (MLP). Die Genauigkeit der MLP lag bei 96 %. Potentiell könnte die Entscheidungsfindung bezüglich der kieferorthopädischen Chirurgie durch KI-Systeme unterstützt werden, jedoch merken die Autor*innen an, dass weitere Studien mit einem größeren Datenpool und mehr Eingabeparameter nötig wären. Des Weiteren gab es viele Ausschlusskriterien der Patient*innenfälle, was die Generalisierbarkeit in Frage stellen lässt. Die Werte sind vergleichbar mit denen der beiden anderen Studien.

Die Studie von Ki-Sun Lee et al. (94) wurden 3 CNNs (Modified-Alexnet, MobileNet, Resnet50) zur Diagnose der Indikation für kieferorthopädische Chirurgie genutzt. Die Genauigkeiten für die CNNs lag zwischen 83,8 % und 91,9 %, wobei Modified Alexnet die besten Ergebnisse erzielte. Auch diese Werte sind vergleichbar mit denen der anderen Studien.

Die Studien zeigen ähnliche Studiendesigns und Datensatzgrößen mit der Verwendung unterschiedlicher Algorithmen und erzielten vergleichbare Leistungen. Aus diesem Grund lässt sich vermuten, dass KI-Modelle durchaus das Potential haben, kieferorthopädische chirurgische Entscheidungen mit hoher Genauigkeit vorherzusagen. Dies überrascht nicht, da diese Entscheidungen weniger komplex als beispielsweise das Erstellen eines Behandlungsplans sind. Jedoch gelten für diese Studien die gleichen potenziellen Verzerrungsquellen, die bei der Erforschung von KI-Algorithmen generell gelten. Homogene Patient*innengruppen, die Definition des Goldstandards, Ausschlusskriterien für Testdatensätze und die Größe des Datenpools können die Testergebnisse beeinflussen und müssen bei der Generalisierbarkeit mit in Betracht gezogen werden.

4.7 KI-generierte Vorhersage des Unterkieferwachstums

Fünf Studien untersuchten die KI-generierte Vorhersagegenauigkeit des Unterkieferwachstums.

Die Studie von Jia-Nan Zhang et al. (95) zielte darauf ab, das Unterkieferwachstum bei Kindern mit einer Klasse III-Malokklusion und einem vorderen Kreuzbiss mit Hilfe einer CNN-Struktur auf Basis des ResNet50-Algorithmus vorherzusagen. Die Ergebnisse zeigten eine Genauigkeit von 85 %. Auffällig war, dass die KI-Leistung die von menschlichen Prüfer*innen mit ca. 30 % übertraf. Die Vorhersageentscheidung traf die KI hauptsächlich durch Identifizierung der Merkmale Kinn, unterer Rand des UK, Schneidezähne, Atemwege und Kondylus.

Die Absicht der Studie von Tyler Wood et al. (96) war es, die postpubertäre Länge und die Y-Achse des Wachstums des Unterkiefers bei Männern mit einer Angle Klasse I zu ermitteln. Alle Algorithmen zeigten eine Genauigkeit zwischen 96,6 % und 98,34 % für die Vorhersage der Y-Achse und Genauigkeiten zwischen 95,80 % und 97,64 % bei der Vorhersage der postpubertären Unterkieferlage. Laut der Autor*innen kann die KI-unterstützte Wachstumsprognose als Hilfestellung für die Behandlungsplanung angesehen werden, wobei diese nur Patienten mit physiologischer Klasse I-Verzahnung involvierte,

was nicht die klinische Bandbreite wiederspiegelt. Die Genauigkeit lag bei 95,26 % und 98,35 % für T1und T2 und war nicht signifikant unterschiedlich zu nur T1. Im Vergleich zu Jia-Nan Zhang et al. (95) waren die Genauigkeitswerte um ca 10 % höher. Erklärbar wäre dies durch die anfängliche Überprüfung der relevanten Inputvariablen.

Die Studie von Grant Zakhar et al. (96) verfolgte ein ähnliches Forschungsdesign, wie die Untersuchung von Tyler Wood et al. (96). Jedoch wurde das Unterkieferwachstum bei männlichen Patienten mit einer Klasse-II-Malokklusion untersucht. Beide Studien zeigten ähnliche Vorhersagegenauigkeiten, wobei in beiden Studien nicht idente Schlüsselprädiktoren für die Vorhersage des Wachstums von den KI-Technologien genutzt wurden. Die Anfangslänge des Unterkiefers und die Gesichtshöhen erwiesen sich in beiden Studien als wichtige Prädiktoren. Des Weiteren wird das Potential in diesem Themengebiet von KI-Methoden betont, jedoch weisen die Autor*innen in beiden Studien darauf hin, dass weitere Studien mit größeren Stichproben nötig sind, um die Ergebnisse verbessern und verallgemeinern zu können.

In der Studie von Matthew Parrish et al. (97), die ebenfalls dem gleichen Aufbau wie Zakhars et al. (95) und Woods et al. (96) folgte, wurde die KI-unterstütze Vorhersage der postpubertären Länge und des Y-Achsen-Winkels des Unterkiefers bei weiblichen Personen mit einer Angle-Klasse I durchgeführt. Auch die Genauigkeitswerte dieser Studie glichen denen der anderen beiden Studien. Die wichtigsten Prädiktoren für die Mandibularlänge waren unter anderem die Anfangslänge des Unterkiefers, das Alter und die vordere und hintere Gesichtshöhe. Die ML-Methoden wurden als fähig bewertet, die mandibuläre Länge innerhalb von 3 mm und die Y-Achse innerhalb von einem Grad vorherzusagen.

Die Absicht der Studie von Eun-Gyeong Kim et al. (98) ist die Identifikation des genausten Modells zur Vorhersage des longitudinalen kraniofazialen Wachstums einer japanischen Bevölkerungsgruppe. Lasso generierte die höchste Vorhersagegenauigkeit mit 97,87 %-94,45 %. Die Studie generierte exzellente Genauigkeitswerte, jedoch wurde ein Datenpool mit einer sehr geringen Fallzahl verwendet.

Zusammenfassend ist anzumerken, dass KI-Systeme eine sehr gute Leistung bei der Vorhersage des Mandibularwachstums erbringen. Es wurden ähnliche Studienaufbauten in den verschiedenen Untersuchungen verwendet, was einen direkten Vergleich ermöglicht. Auffällig ist, dass dabei die Art der Verzahnung, das Geschlecht oder Malokklusion keinen Einfluss auf die Genauigkeitswerte auszuüben scheinen. Jedoch gewichten in den unterschiedlichen Studien die KI-Systeme verschiedene Parameter je nach Angle-Klasse unterschiedlich. Schwächen der Studien stellen die verhältnismäßig kleinen Datensätze dar. Ob eine Allgemeingültigkeit der einzelnen KI-Systeme für einen homogenen Patient*innenpool besteht, sollte in weiteren Untersuchungen mit größeren und heterogenen Datensätzen überprüft werden.

4.8 KI-generierte Weichgewebsvorhersage nach chirurgischen oder kieferorthopädischen Eingriffen

Die Studie von Rutger ter Horst et al. (99) entwickelte eine CNN-Struktur zur Vorhersage des virtuellen Weichteilprofils nach einer mandibulären Osteotomie und verglich die Ergebnisse mit dem Massen-Tensor-Modell (MTM). Die durchschnittliche absolute Fehlerquote der KI-generierten Simulation im Bereich des unteren Gesichts betrug im Vergleich zu realen postoperativen 3D-Fotografien 1,0 ± 0,6 mm und war signifikant geringer als die der MTM-basierten Simulationen. Für die Position der Lippe und des Kinns waren mittlere Fehler leicht höher. Die Autor*innen weisen darauf hin, dass KI-unterstütze kieferorthopädische Planung und Simulation als fortschrittliche Methode betrachtet werden kann.

Die Absicht der Studie von Chihiro Tanikawa et al. (100) war die Entwicklung von KI-Systemen zur Vorhersage der dreidimensionalen Gesichtsmorphologie nach kieferorthopädischen Operationen oder orthodontischen Behandlungen bei einer Population japanischer Patient*innen. Es wurden Daten von 137 Patient*innen verwendet, die entweder eine kieferorthopädische Operation oder eine orthodontische Behandlung mit der Extraktion von vier Prämolaren durchlaufen hatten. Beide Systeme konnten eine

Erfolgsrate von 100 % für einen Systemfehler kleiner 2 mm erreichen, wobei bei Extraktionsfällen bessere Werte generiert werden konnten. Die bessere Erfolgsrate für Patient*innen mit Extraktionsfällen überrascht nicht, da dabei keine Kieferbasenveränderung stattfindet. Bei Umstellungsosteotomien hingegen wird die Lage des Ober- oder Unterkiefers oder beider gleichzeitig operativ verändert. Auffallend ist, dass die Studien von Rutger ter Horst et al. (101) und Chihiro Tanikawa et al. (100) vergleichbare durchschnittliche Systemfehler und Standardabweichungen für die Vorhersage nach Gesichtsoperationen zeigten. In beiden Studien wurde jedoch ein relativ kleiner und homogener Datensatz verwendet. Die KI-Systeme wurden von den Autor*innen als klinisch akzeptabel eingestuft und das hohe Potenzial für Vorhersagen von posttherapeutischen Gesichtsmorphologien durch KI-Systeme wurde damit unterstrichen.

Die Autor*innen kritisieren die Annahme vieler kommerzieller Softwareprogramme, dass die Bewegung von hartem und weichem Gewebe proportional sei. Die Studien können als Grundstein in diesem Forschungsfeld angesehen werden und eine weitere Überprüfung der Generalisierbarkeit und Reproduktion der Ergebnisse ist durchzuführen.

5 Konklusion

Die in dieser Literaturrecherche analysierten Studien zeigen das vielfältige Potenzial der künstlichen Intelligenz zur Verbesserung verschiedener Aspekte der kieferorthopädischen Behandlung. Es wird demonstriert, dass KI-basierte Systeme in der Lage sind, mit hoher Genauigkeit diagnostische Aufgaben auszuführen. Diese reichen von der Erkennung anatomischer Landmarken auf Röntgenbildern über die Unterstützung bei der Behandlungsplanung bis hin zur Erstellung von Wachstumsprognosen und der Identifizierung von kieferorthopädischen oder chirurgischen Notwendigkeiten.

Es wurde deutlich, dass die Qualität der KI-Leistung stark von der Größe und Vielfalt der Trainingsdaten abhängt. Die Studien legen nahe, dass ein umfangreicher und diverser Trainingsdatensatz essentiell ist, um eine robuste CNN-Struktur zu entwickeln und die KI-Leistung zu optimieren. Dabei ist ein demografisch vielfältiger Patient*innenpool mit heterogenen Morphologien und Pathologien, sowie die Nutzung und Anwendung verschiedener Röntgengeräte und Behandlungsabläufen essentiell.

Ein weiterer wichtiger Aspekt, der in mehreren Studien hervorgehoben wurde, ist die Notwendigkeit einer kontinuierlichen Validierung und Anpassung der KI-Modelle. Eine regelmäßige Überprüfung und Aktualisierung der Algorithmen ist notwendig, um die Leistungen zu verbessern und den Forschungsstand weiter zu treiben.

Ein potenzieller Kritikpunkt KI-gestützter automatisierter Prozesse in der Kieferorthopädie könnten die Auswirkungen auf das selbstständige Denken der Mediziner*innen darstellen. Es besteht die Möglichkeit, dass bei fortschreitender Einführung von KI die Gefahr besteht, dass sich Kieferorthopäd*innen zunehmend auf die KI-Leistung verlassen und dabei ihre eigenen Kompetenzen vernachlässigen oder nicht weiterentwickeln. Diese mögliche Abhängigkeit sollte bei der Einführung KI-betriebener Technologien berücksichtigt werden, um sicher zu stellen, dass die Fähigkeit zur eigenständigen Problemlösung und kritischen Beurteilung erhalten bleibt.

Zusammenfassend lässt sich feststellen, dass künstliche Intelligenz ein vielversprechendes Werkzeug für die kieferorthopädische Praxis darstellt. Die Geschwindigkeit, in der neue Studien publiziert werden und der damit verbundene wissenschaftliche Fortschritt, ist rasant. Vor allem weniger erfahrene Kieferorthopäd*innen könnten von KI-betriebenen Unterstützungssystemen profitieren und dadurch die Patient*innenversorgung verbessern. Trotz der fortschreitenden Entwicklung ist derzeit eine Überprüfung des KI-Outputs durch Expert*innen erforderlich. Besonders bei weniger typischen Fällen oder veränderten Begleitumständen scheinen sich KI-Systeme schwieriger anpassen zu können. Für eine vollständige Integration in den klinischen Alltag bedarf es weiterer Forschung, um die Möglichkeiten dieser Technologien weiter zu verbessern und zu validieren.

6 Literaturverzeichnis

1. Sonntagbauer M, Haar M, Kluge S. Artificial intelligence: How will ChatGPT and other AI applications change our everyday medical practice? Med Klin Intensivmed Notfmed. 2023 Jun 1;118(5):366–71.

2. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial Intelligence in Primary Health Care: Perceptions, Issues, and Challenges. In: Yearbook of medical informatics. NLM (Medline); 2019. p. 41–6.

3. Elendu C, Amaechi DC, Elendu TC, Jingwa KA, Okoye OK, John Okah M, et al. Ethical implications of AI and robotics in healthcare: A review. Vol. 102, Medicine (United States). Lippincott Williams and Wilkins; 2023. p. E36671.

4. Liu P ran, Lu L, Zhang J yao, Huo T tong, Liu S xiang, Ye Z wei. Application of Artificial Intelligence in Medicine: An Overview. Curr Med Sci. 2021 Dec 1;41(6):1105–15.

5. Jheng YC, Kao CL, Yarmishyn AA, Chou YB, Hsu CC, Lin TC, et al. The era of artificial intelligence-based individualized telemedicine is coming. Journal of the Chinese Medical Association. 2020 Nov 1;83(11):981–3.

6. Mishra K, Leng T. Artificial intelligence and ophthalmic surgery. Vol. 32, Current Opinion in Ophthalmology. Lippincott Williams and Wilkins; 2021. p. 425–30.

7. Bodenstedt S, Wagner M, Müller-Stich BP, Weitz J, Speidel S. Artificial intelligence-assisted surgery: Potential and challenges. Vol. 36, Visceral Medicine. S. Karger AG; 2020. p. 450–5.

8. Bellini V, Valente M, Rio P Del, Bignami E. Artificial intelligence in thoracic surgery: a narrative review. Vol. 13, Journal of Thoracic Disease. AME Publishing Company; 2021. p. 6963–75.

9. Alowais SA, Alghamdi SS, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb SN, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. Vol. 23, BMC Medical Education. BioMed Central Ltd; 2023.

10. Sharma M, Savage C, Nair M, Larsson I, Svedberg P, Nygren JM. Artificial Intelligence Applications in Health Care Practice: Scoping Review. Vol. 24, Journal of Medical Internet Research. JMIR Publications Inc.; 2022.

11. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Vol. 7, Nature biomedical engineering. NLM (Medline); 2023. p. 719–42.

12. Nguyen TT. Use of Artificial Intelligence in Dentistry: Current Clinical Trends and Research Advances. 2021;

13. Ahmed N, Abbasi MS, Zuberi F, Qamar W, Halim MS Bin, Maqsood A, et al. Artificial Intelligence Techniques: Analysis, Application, and Outcome in Dentistry - A Systematic Review. Vol. 2021, BioMed Research International. Hindawi Limited; 2021.

14. Pethani F. Promises and perils of artificial intelligence in dentistry. Vol. 66, Australian Dental Journal. Blackwell Publishing; 2021. p. 124–35.

15. Revilla-León M, Gómez-Polo M, Vyas S, Barmak BA, Galluci GO, Att W, et al. Artificial intelligence applications in implant dentistry: A systematic review. Vol. 129, Journal of Prosthetic Dentistry. Elsevier Inc.; 2023. p. 293–300.

16. Bernauer SA, Zitzmann NU, Joda T. The use and performance of artificial intelligence in prosthodontics: A systematic review. Vol. 21, Sensors. MDPI; 2021.

17. Ossowska A, Kusiak A, Świetlik D. Artificial Intelligence in Dentistry—Narrative Review. Vol. 19, International Journal of Environmental Research and Public Health. MDPI; 2022.

18. Monill-González A, Rovira-Calatayud L, d'Oliveira NG, Ustrell-Torrent JM. Artificial intelligence in orthodontics: Where are we now? A scoping review. Vol. 24, Orthodontics and Craniofacial Research. John Wiley and Sons Inc; 2021. p. 6–15.

19. Rajaraman V. John McCarthy-father of artificial intelligence.

20. Skinner Rebecca. Building the second mind : 1956 and the origins of artificial intelligence computing. University of California; 2012.

21. Lang V. Digitale Kompetenz [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2022. Available from: https://link.springer.com/10.1007/978-3-662-66285-4

22. Anusuya MA, Katti SK. Speech Recognition by Machine: A Review [Internet]. Vol. 6, IJCSIS) International Journal of Computer Science and Information Security. 2009. Available from: http://sites.google.com/site/ijcsis/

23. Waldron T, Carr T, McMullen L, Westhorp G, Duncan V, Neufeld SM, et al. Development of a program theory for shared decision-making: A realist synthesis. BMC Health Serv Res. 2020 Jan 23;20(1).

24. Lecun Y, Bengio Y, Hinton G. Deep learning. Vol. 521, Nature. Nature Publishing Group; 2015. p. 436–44.

25. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;2019(10).

26. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Peter Campbell J. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol. 2020;9(2).

27. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021 Aug 1;25(3):1315–60.

28. Deo RC. Machine learning in medicine. Circulation. 2015 Nov 17;132(20):1920–30.

29. James G (Gareth M, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning : with applications in R. 426 p.

30. Kriegeskorte N, Golan T. Neural network models and deep learning. Vol. 29, Current Biology. Cell Press; 2019. p. R231–6.

31. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015 May 18; Available from: http://arxiv.org/abs/1505.04597

32. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. 2015 Nov 26; Available from: http://arxiv.org/abs/1511.08458

33. Ghosh S, Das N, Das I, Maulik U. Understanding Deep Learning Techniques for Image Segmentation. 2019 Jul 13; Available from: http://arxiv.org/abs/1907.06119

34. KIEFERORTHOPÄDISCHE DIAGNOSTIK II. RÖNTGENDIAGNOSTIK UND FERNRÖNTGENANALYSE EINLEITUNG.

35. Harzer W. Checkliste der Zahnmedizin, Kapitel Röntgenanalyse und Bildgebende Verfahren, 2011.

36. Kieferorthopädie, Sander Franz Günter.

37. McNamara JA, Franchi L. The cervical vertebral maturation method: A user's guide. Angle Orthodontist. 2018 Mar 1;88(2):133–43.

38. Extraktionstherapie, Wichelhaus Andrea.

39. Einteilung der Dysgnathien, Sander Franz.

40. Miyajima K, McNamara JA, Kimura T, Murata S, lizuka T, Nagoya D, et al. Craniofacial structure of Japanese and European-American adults with normal and well-balanced faces occlusions. Vol. 110, Am J Orthod Dentofac Orthop. 1996.

41. Process of maturation and growth prediction.

42. Rischen RJ, Breuning KH, Bronkhorst EM, Kuijpers-Jagtman AM. Records needed for orthodontic diagnosis and treatment planning: A systematic review. Vol. 8, PLoS ONE. 2013.

43. Thribhuvanan L, Saravanakumar MS. Influence of mode of breathing on pharyngeal airway space and dento facial parameters in children: a short clinical study. Bull Natl Res Cent. 2022 Dec;46(1).

44. Bichara LM, de Aragón MLC, Brandão GAM, Normando D. Factors influencing orthodontic treatment time for non-surgical Class III malocclusion. Journal of Applied Oral Science. 2016 Sep 1;24(5):431–6.

45. Elias KG, Sivamurthy G, Bearn DR. Extraction vs nonextraction orthodontic treatment: a systematic review and meta-analysis. Angle Orthod. 2024 Jan 1;94(1):83–106.

46. Skidmore KJ, Brook KJ, Thomson WM, Harding WJ. Factors influencing treatment time in orthodontic patients. American Journal of Orthodontics and Dentofacial Orthopedics. 2006 Feb;129(2):230–8.

47. Shrout PE, Fleiss JL. Intraclass Correlations : Uses in Assessing Rater Reliability. Vol. 86, Psychological Bulletin. 1979.

48. Popova T, Stocker T, Khazaei Y, Malenova Y, Wichelhaus A, Sabbagh H. Influence of growth structures and fixed appliances on automated cephalometric landmark recognition with a customized convolutional neural network. BMC Oral Health. 2023 Dec 1;23(1).

49. Li H, Li H, Yuan L, Liu C, Xiao S, Liu Z, et al. The psc-CVM assessment system: A three-stage type system for CVM assessment based on deep learning. BMC Oral Health. 2023 Dec 1;23(1).

50. Tabelle 1. Checkliste zum Bericht einer systematischen Übersicht oder einer Meta-Analyse.

51. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Vol. 18, PLoS Medicine. Public Library of Science; 2021.

52. Kebaili A, Lapuyade-Lahorgue J, Ruan S. Deep Learning Approaches for Data Augmentation in Medical Imaging: A Review. Vol. 9, Journal of Imaging. MDPI; 2023.

53. Chen YJ, Chen SK, Jane ;, Yao CC, Chang HF. The Effects of Differences in Landmark Identification on the Cephalometric Measurements in Traditional Versus Digitized Cephalometry [Internet]. Vol. 74, Angle Orthodontist. 2004. Available from: http://meridian.allenpress.com/angle-orthodontist/article-pdf/74/2/155/1380244/0003-3219

54. Livas C, Delli K, Spijkervet FKL, Vissink A, Dijkstra PU. Concurrent validity and reliability of cephalometric analysis using smartphone apps and computer software. Angle Orthodontist. 2019;89(6):889–96.

55. Bulatova G, Kusnoto B, Grace V, Tsay TP, Avenetti DM, Sanchez FJC. Assessment of automatic cephalometric landmark identification using artificial intelligence. Orthod Craniofac Res. 2021 Dec 1;24(S2):37–42.

56. Zhao C, Yuan Z, Luo S, Wang W, Ren Z, Yao X, et al. Automatic recognition of cephalometric landmarks via multi-scale sampling strategy. Heliyon. 2023 Jun 1;9(6).

57. Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. Angle Orthodontist. 2021 May 1;91(3):329–35.

58. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? A cephalometric example. Angle Orthodontist. 2020 Nov 1;90(6):823–30.

59. Hwang HW, Park JH, Moon JH, Yu Y, Kim H, Her SB, et al. Automated identification of cephalometric landmarks: Part 2-Might it be better than human? Angle Orthodontist. 2020;90(1):69–76.

60. Hwang HW, Moon JH, Kim MG, Donatelli RE, Lee SJ. Evaluation of automated cephalometric analysis based on the latest deep learning method. Angle Orthodontist. 2021 May 1;91(3):329–35.

61. Moon JH, Hwang HW, Yu Y, Kim MG, Donatelli RE, Lee SJ. How much deep learning is enough for automatic identification to be reliable? A cephalometric example. Angle Orthodontist. 2020 Nov 1;90(6):823–30.

62. Song Y, Qiao X, Iwamoto Y, Chen YW. Automatic Cephalometric Landmark Detection on X-ray images using a deep-learning method. Applied Sciences (Switzerland). 2020 Apr 1;10(7).

63. Jiang F, Guo Y, Yang C, Zhou Y, Lin Y, Cheng F, et al. Artificial intelligence system for automated landmark localization and analysis of cephalometry. Dentomaxillofacial Radiology. 2023;52(1).

64. Mahto RK, Kafle D, Giri A, Luintel S, Karki A. Evaluation of fully automated cephalometric measurements obtained from web-based artificial intelligence driven platform. BMC Oral Health. 2022 Dec 1;22(1).

65. Ye H, Cheng Z, Ungvijanpunya N, Chen W, Cao L, Gou Y. Is automatic cephalometric software using artificial intelligence better than orthodontist experts in landmark identification? BMC Oral Health. 2023 Dec 1;23(1).

66. Tsolakis IA, Tsolakis AI, Elshebiny T, Matthaios S, Palomo JM. Comparing a Fully Automated Cephalometric Tracing Method to a Manual Tracing Method for Orthodontic Diagnosis. J Clin Med. 2022 Nov 1;11(22).

67. Lee JH, Yu HJ, Kim MJ, Kim JW, Choi J. Automated cephalometric landmark detection with confidence regions using Bayesian convolutional neural networks. BMC Oral Health. 2020 Oct 7;20(1).

68. Kim MJ, Liu Y, Oh SH, Ahn HW, Kim SH, Nelson G. Automatic cephalometric landmark identification system based on the multi-stage convolutional neural networks with cbct combination images. Sensors (Switzerland). 2021 Jan 2;21(2):1–16.

69. Kim J, Kim I, Kim YJ, Kim M, Cho JH, Hong M, et al. Accuracy of automated identification of lateral cephalometric landmarks using cascade convolutional neural networks on lateral cephalograms from nationwide multi-centres. Orthod Craniofac Res. 2021 Dec 1;24(S2):59–67.

70. Tanikawa C, Lee C, Lim J, Oka A, Yamashiro T. Clinical applicability of automated cephalometric landmark identification: Part I—Patient-related identification errors. Orthod Craniofac Res. 2021 Dec 1;24(S2):43–52.

71. Panesar S, Zhao A, Hollensbe E, Wong A, Bhamidipalli SS, Eckert G, et al. Precision and Accuracy Assessment of Cephalometric Analyses Performed by Deep Learning Artificial Intelligence with and without Human Augmentation. Applied Sciences (Switzerland). 2023 Jun 1;13(12).

72. Lee HT, Chiu PY, Yen CW, Chou ST, Tseng YC. Application of artificial intelligence in lateral cephalometric analysis. J Dent Sci. 2023;

73. Kim HJ, Kim KD, Kim DH. Deep convolutional neural network-based skeletal classification of cephalometric image compared with automated-tracing software. Sci Rep. 2022 Dec 1;12(1).

74. Davidovitch M, Sella-Tunis T, Abramovicz L, Reiter S, Matalon S, Shpack N. Verification of Convolutional Neural Network Cephalometric Landmark Identification. Applied Sciences (Switzerland). 2022 Dec 1;12(24).

75. Silva TP, Hughes MM, dos Santos Menezes L, de Melo M de FB, de Freitas PHL, Takeshita WM. Artificial intelligence-based cephalometric landmark annotation and measurements according to Arnett's analysis: can we trust a bot to do that? Dentomaxillofacial Radiology. 2022 Sep 1;51(6).

76. Santos Menezes L dos, Silva TP, Lima dos Santos MA, Hughes MM, Reis Mariano Souza S dos, Leite Ribeiro PM, et al. Assessment of landmark detection in cephalometric radiographs with different conditions of brightness and contrast using the an artificial intelligence software. Dentomaxillofacial Radiology. 2023;52(8).

77. Zhou J, Zhou H, Pu L, Gao Y, Tang Z, Yang Y, et al. Development of an artificial intelligence system for the automatic evaluation of cervical vertebral maturation status. Diagnostics. 2021 Dec 1;11(12).

78. Li H, Chen Y, Wang Q, Gong X, Lei Y, Tian J, et al. Convolutional neural network-based automatic cervical vertebral maturation classification method. Dentomaxillofacial Radiology. 2022 Sep 1;51(6).

79. Seo H, Hwang J, Jeong T, Shin J. Comparison of deep learning models for cervical vertebral maturation stage classification on lateral cephalometric radiographs. J Clin Med. 2021 Aug 2;10(16).

80. Atici SF, Ansari R, Allareddy V, Suhaym O, Cetin AE, Elnagar MH. AggregateNet: A deep learning model for automated classification of cervical vertebrae maturation stages. Orthod Craniofac Res. 2023 Dec 1;26(S1):111–7.

81. Kim EG, Oh IS, So JE, Kang J, Le VNT, Tak MK, et al. Estimating cervical vertebral maturation with a lateral cephalogram using the convolutional neural network. J Clin Med. 2021 Nov 1;10(22).

82. Ryu J, Kim YH, Kim TW, Jung SK. Evaluation of artificial intelligence model for crowding categorization and extraction diagnosis using intraoral photographs. Sci Rep. 2023 Dec 1;13(1).

83. Etemad L, Wu TH, Heiner P, Liu J, Lee S, Chao WL, et al. Machine learning from clinical data sets of a contemporary decision for orthodontic tooth extraction. Orthod Craniofac Res. 2021 Dec 1;24(S2):193–200.

84. Leavitt L, Volovic J, Steinhauer L, Mason T, Eckert G, Dean JA, et al. Can we predict orthodontic extraction patterns by using machine learning? Orthod Craniofac Res. 2023 Nov 1;26(4):552–9.

85. Suhail Y, Upadhyay M, Chhibber A, Kshitiz. Machine learning for the diagnosis of orthodontic extractions: A computational analysis using ensemble learning. Bioengineering. 2020 Jun 1;7(2):1–13.

86. Li P, Kong D, Tang T, Su D, Yang P, Wang H, et al. Orthodontic Treatment Planning based on Artificial Neural Networks. Sci Rep. 2019 Dec 1;9(1).

87. Shimizu Y, Tanikawa C, Kajiwara T, Nagahara H, Yamashiro T. The validation of orthodontic artificial intelligence systems that perform orthodontic diagnoses and treatment planning. Eur J Orthod. 2022 Aug 1;44(4):436–44.

88. Thanathornwong B. Bayesian-based decision support system for assessing the needs for orthodontic treatment. Healthc Inform Res. 2018 Jan 1;24(1):22–8.

89. Prasad J, Mallikarjunaiah DR, Shetty A, Gandedkar N, Chikkamuniswamy AB, Shivashankar PC. Machine Learning Predictive Model as Clinical Decision Support System in Orthodontic Treatment Planning. Dent J (Basel). 2023 Jan 1;11(1).

90. Volovic J, Badirli S, Ahmad S, Leavitt L, Mason T, Bhamidipalli SS, et al. A Novel Machine Learning Model for Predicting Orthodontic Treatment Duration. Diagnostics. 2023 Sep 1;13(17).

91. Shin WS, Yeom HG, Lee GH, Yun JP, Jeong SH, Lee JH, et al. Deep learning based prediction of necessity for orthognathic surgery of skeletal malocclusion using cephalogram in Korean individuals. BMC Oral Health. 2021 Dec 1;21(1).

92. Kim YH, Park JB, Chang MS, Ryu JJ, Lim WH, Jung SK. Influence of the depth of the convolutional neural networks on an artificial intelligence model for diagnosis of orthognathic surgery. J Pers Med. 2021 May 1;11(5).

93. Chaiprasittikul N, Thanathornwong B, Pornprasertsuk-Damrongsri S, Raocharernporn S, Maponthong S, Manopatanakul S. Application of a Multi-Layer

Perceptron in Preoperative Screening for Orthognathic Surgery. Healthc Inform Res. 2023 Jan 1;29(1):16–22.

94. Lee KS, Ryu JJ, Jang HS, Lee DY, Jung SK. Deep convolutional neural networks based analysis of cephalometric radiographs for differential diagnosis of orthognathic surgery indications. Applied Sciences (Switzerland). 2020 Mar 1;10(6).

95. Zhang JN, Lu HP, Hou J, Wang Q, Yu FY, Zhong C, et al. Deep learning-based prediction of mandibular growth trend in children with anterior crossbite using cephalometric radiographs. BMC Oral Health. 2023 Dec 1;23(1).

96. Wood T, Anigbo JO, Eckert G, Stewart KT, Dundar MM, Turkkahraman H. Prediction of the Post-Pubertal Mandibular Length and Y Axis of Growth by Using Various Machine Learning Techniques: A Retrospective Longitudinal Study. Diagnostics. 2023 May 1;13(9).

97. Parrish M, O'Connell E, Eckert G, Hughes J, Badirli S, Turkkahraman H. Short- and Long-Term Prediction of the Post-Pubertal Mandibular Length and Y-Axis in Females Utilizing Machine Learning. Diagnostics. 2023 Sep 1;13(17).

98. Kim E, Kuroda Y, Soeda Y, Koizumi S, Yamaguchi T. Validation of Machine Learning Models for Craniofacial Growth Prediction. Diagnostics. 2023 Nov 1;13(21).

99. ter Horst R, van Weert H, Loonen T, Bergé S, Vinayahalingam S, Baan F, et al. Three-dimensional virtual planning in mandibular advancement surgery: Soft tissue prediction based on deep learning. Journal of Cranio-Maxillofacial Surgery. 2021 Sep 1;49(9):775–82.

100. Tanikawa C, Yamashiro T. Development of novel artificial intelligence systems to predict facial morphology after orthognathic surgery and orthodontic treatment in Japanese patients. Sci Rep. 2021 Dec 1;11(1).

101. Nötzel Frank, Fernröntgenseitenbild-Analyse, 2007

102. DataScientitest, 2023, https://datascientest.com/de/was-ist-die-grad-cam-methode

103. Wikipedia, 2024, https://de.wikipedia.org/wiki/Interrater-Reliabilit%C3%A4t

104. Bundesminesterium, Bildung Wissenschaft und Forschung, 2024 https://www.bmbwf.gv.at/Themen/HS-Uni/Hochschulgovernance/Leitthemen/Digitalisierung/K%C3%BCnstliche-Intelligenz.html

105. Medium, 2023, https://medium.com/@ilyurek/roc-curve-and-auc-evaluating-model-performance-c2178008b02

106. DocCheck Flexikon, Bogomil Sabev et al., https://flexikon.doccheck.com/de/Fernr%C3%B6ntgenanalyse

7 Anhang

Autor*in	Fulin Jiang et al.	Ravi Kumar Mahto et al.	Galina Bulatova et al.
Jahr	2023	2022	2021
Journal	DMFR (Dentomaxillofacial Radiology)	BMC oral Health	Orthodontics & Craniofacial Research
Land	China	Nepal	USA
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	Cephnet	Webceph	YOLOv3
Landmarken (N)	30	18	16
Totaler Datensatz	9870	30	110
Trainingsdatensatz	9611	n.a.	n.a.
Testdatensatz	259	n.a.	n.a.
Statistische Auswertung	SDR: 91,73% innerhalb von 2 mm SCR: 89,33%	ICC-Wert > 0,75 für 5 Parameter ICC-Wert > 0,9 für 7 Parameter	SDR: 75% innerhalb von 2 mm
Autor*in	**Huayu Ye et al.**	**Ioannis A Tsolakis et al.**	**Jeong-Hoon Lee et al.**
Jahr	2023	2022	2020
Journal	BMC Oral Health	Clinical Medicine	BMC Oral Health
Land	China	Griechenland	Korea
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	MyorthoX Angelalign Digident	CS Imaging V8	Entwicklung eines CNN unter Verwendung von BCNN
Landmarken (N)	N 32	N 16	N 19
Totaler Datensatz	43	100	400
Trainingsdatensatz	n.a.	n.a.	150
Testdatensatz	n.a.	n.a.	250
Statistische Auswertung	Angelalign MRE: 0,80 ± 0,26 mm SDR: 93,09%	ICC-Wert: 0,7-9,2	MRE: 1,53 ± 1,74 mm SDR: 82,11%,

	innerhalb von 2 mm Digident MRE: 1,11 ± 0,48 mm SDR: 87,53% innerhalb von 2 mm MyOrthoX MRE: 0,97 ± 0,51 mm SDR: 89,99% innerhalb von 2 mm		92,28%, 95,95% Im Bereich von 2 mm, 3 mm, 4 mm Inter-Rater-Reabialität zwischen erfahrenen und jüngeren Ärtz*innen: 2,02 ± 1,53 mm
Autor*in	**Min-Jung Kim et al.**	**Jaerong Kim et al.**	**Congyi Zhao et al.**
Jahr	2021	2021	2023
Journal	Sensors (Basel)	Orthodontics and Craniofacial Research	Heliyon
Land	Südkorea	Südkorea	China
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	Entwicklung eines individuellen CNN	Entwicklung eines individuellen CNN	MS-YOLOv3
Landmarken (N)	15	20	19
Totaler Datensatz	860 Fernröntgen aus 430 DVTs ➔ 430 CBCT-LC ➔ 430 MIP-LC	3250	400 -> durch Datenaugmentationen auf 2100 erhöht
Trainingsdatensatz	690	3150	1950
Testdatensatz	170	100	150
Statistische Auswertung	MRE: 1,03 ± 1,29mm SDR: 87,13%-96,59% im Bereich von 2-4 mm	Inter-Rater-Reabilität zwischen Untersucher*innen: 1,31 ± 1,13 mm MRE: 1,36 ± 0,98 mm	SDR: 80,84%-98,14% innerhalb von 2-4 mm MRE: 1,59 ± 1,33 mm
Autor*in	**Chihiro Tanikawa et al.**	**Theodora Popova et al.**	**Sumer Pansear et al.**
Jahr	2021	2023	2023
Journal	Orthodontics and Craniofacial Research	BMC Oral Health	Applied sciences
Land	Japan	Deutschland	USA
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	2 vorab entwickelte	Vorab entwickelte	RadioCefv3

	CNN-Strukturen (CNN-PC, CNN-PE)	CNN-Struktur	
Landmarken (N)	N 26	N 16	N 31
Totaler Datensatz	1785	890	30
Trainingsdatensatz	1755	430	n.a.
Testdatensatz	30 (aus 8 Untergruppen)	460	n.a.
Statistische Auswertung	SDR gesamt: 85%-91% innerhalb von 2 mm Gruppen mit LKG-Spalten: 85%-87% innerhalb von 2 mm	SDR gesamt: 84,73%-96,48% innerhalb von 2-4 mm MRE gesamt: 1,47 ± 1,06 mm	KI-assistierte Analyse, Präzisionssteigerung/ Steigerung SDR in %: Kieferorthopäd*in: 3,26 Assistenzarzt*ärztin: 2,17 Zahnarzt*ärztin:19,75 Student*in: 23,38 Insgesamt: 10,47 Verbesserung der Genauigkeit KI-assistiert in %: Kieferorthopäd*in: 12,74 Assistenzärzt*in:19,10 Zahnärzt*in: 35,69 Student*in: 33,96 Insgesamt: 27,27
Autor*in	**Hye-Won et al.**	**Moon Jun-Ho et al.**	**Huang-Ting Lee et al.**
Jahr	2021	2020	2023
Journal	The Angle Orthodontist	The Angle Orthodontist	Journal of Dental Sciences
Land	Südkorea	Südkorea	Taiwan
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	Modifikation des YOLO-v3-Algorithmus	Modifikation des YOLO-v3-Algorithmus	MobileNetv2
Landmarken	N 19	N 19-80	N 14

(N)			
Totaler Datensatz	2183	2400	1002
Trainingsdatensatz	1983	2200	700/609
Testdatensatz	200	200	302
Statistische Auswertung	SDR: 75,5% innerhalb von 2 mm SCR: 81,5%	MRE bei einem Trainingsdatensatz von 2000 Fernröntgen und 19, 40, 80 Landmarken: 1,61 mm, 1,76 mm, 1,90 mm	Model 2: SDR: 83,14%-98,23% innerhalb von 2-4 mm MRE: < 2 mm (Ausnahme Apex des oberen Schneidezahnes)
Autor*in	**Ho-Jin Kim et al.**	**Moshe Davidovitch et al.**	**Hye-Won Hwang et al.**
Jahr	2022	2022	2020
Journal	Scientific Reports	Applied Sciences	The Angle Orthodontist
Land	Südkorea	Israel	Südkorea
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	Vorab entwickelte DCNN-Struktur	Algoceph	YOLOv3
Landmarken (N)	N 3	N 21	N 80
Totaler Datensatz	1574	n.a.	1311
Trainingsdatensatz	1334	n.a.	1028
Testdatensatz	120	10	283
Statistische Auswertung	Sesitivität:94% Spezifität:97% Präzision: 94% Genauigkeit: 96% Genauigkeit: Klasse I: 97% Klasse II:97% Klasse III: 88%	SDR: 97,6% innerhalb von 2 mm 18 von 21 Landmarken weisen eine hohe Korrelation (r>0.90) zwischen KI und Menschen auf 3 Landmarken korrelierten moderat (r=0,73-0.89)	Intra-Rater-Reliabilität Untersucher*in 2: 0,97 ± 1,03 mm ICC KI: 1,46 ± 2,97 mm Inter-Rater-Reabilität beider Prüfer*innen: 1,50 ± 1,48 mm MRE: <0,9mm Ausnahme: Wurzelspitze unterer Inzisor

Autor*in	**Yu Song et al.**	**Thaisa Pinheiro Silvia et al.**	**Liciane dos Santos Menezes et al.**
Jahr	2020	2022	2023
Journal	Applied Science	DMFR	DMFR
Land	Japan	Brasilien	Brasilien
Studienthema	Fernröntgenanalyse	Fernröntgenanalyse	Fernröntgenanalyse
KI-Struktur	ResNet50	Cefbot	Cefbot
Landmarken (N)	N 19	N 66	N 19
Totaler Datensatz	500	n.a.	n.a.
Trainingsdatensatz	150	n.a.	n.a.
Testdatensatz	150+100+100	30	150 (30x5)
Statistische Auswertung	SDR innerhalb von 2-4 mm: 62%-86,6% SCR: 77,95%	ICC >0,94 Der Glabella zu Subnasale- Punkt konnte von Cefbot nicht gemessen werden.	ICC für beide Prüfer*innen und KI: >0,91 Veränderte Detektierbarkeit bei Bildanpassungen: N, B, ANS, POG, Ll, UL, Sn
Autor*in	**Jung Zhou et al.**	**Haizhen Li et al.**	**Hyejun Seo et al.**
Jahr	2021	2022	2021
Journal	Diagnostics basel	DMFR	Clinical Medicine
Land	China	China	Südkorea
Studienthema	CVM	CVM	CVM
KI-Struktur	Vorab entwickelte KI-Struktur	VGG16 GoogLeNet DenseNet161 ResNet152	- ResNet-18 - MobileNet_v2 - ResNet-50 - ResNet101 - Inception-v3 - Inception-ResNet v2
Landmarken (N)	15	n.a.	n.a.
Totaler Datensatz	1080	6079	n.a.
Trainingsdatensatz	980	4253	n.a.
Testdatensatz	100	914	600

Statistische Auswertung	MRE: 0,36 ± 0,09 mm F1-Score-Rangfolge: CS6>CS1>CS4> CS5>CS2>CS3 Genauigkeit CVM-Stadium: 71%	F1-Score-Rangfolge: CS6>CS1>CS4> CS5>CS3>CS2 ResNet152 Gesamtgenauigkeit: 67,06%	Genauigkeit aller Modelle: >90% AUC CS1-CS6: >0,9
Autor*in	**Salih Furkan et al.**	**Eun-Gyeong Kim et al.**	**Hairui Li et al.**
Jahr	2023	2021	2023
Journal	Orthodontics & Craniofacial Research	Clinical medicine	BMC Oral Health
Land	USA	Korea	China
Studienthema	CVM	CVM	CVM
KI-Struktur	AggregateNet	Vorab entwickelte KI-Struktur	Psc-CVM-Assessment
Totaler Datensatz	1012	n.a.	10200
Trainingsdatensatz	823	600	7111
Testdatensatz	189	n.a.	1545
Statistische Auswertung	Genauigkeit mit Datenaugmentationen, Kantenhervorhebungsfiltern und Altersbekanntgabe: 82,35% Frauen 75,0% Männer Ohne Datenaugmentation: 69,41% Frauen 60,57% Männer Ohne Kantenhervorhebungsfilter: 80% Frauen 74,03% Männer Ohne Altersbekanntgabe: 74,11% Frauen	CVM-Klassifikation bei Verwendung von ROI-Detektion und zervikaler Segmentierung: 62,5% Genauigkeit	AUC: 0,94 Gesamtgenauigkeit: 70,42% ICC: 0,946 F1-Score: CVS6>CVS1>CVS4 >CVS5>CVS3>CVS2

	64,42% Männer		
Autor*in	**Jiho Ryu et al.**	**Lily Etemad et al.**	**Landon Leavitt et al.**
Jahr	2023	2021	
Journal	Scientific Reports	Orhtodontics & Craniofacial Research	Orhtodontics & Craniofacial Research
Land	Korea	USA	USA
Studienthema	Extraktions-entscheidung	Extraktions-entscheidung	Extraktions-entscheidung
KI-Struktur	ResNet50 -ResNet101 -VGG16 -VGG19	2 Algorithmen (Random Forest (RF), Multilayer Perceptron (MLP)	Random Forest (RF) Logistische Regression (LR) Support-Vektor-Maschine (SVM
Totaler Datensatz	3136	n.a.	
Trainingsdatensatz	2736		
Testdatensatz	400	838	366
Statistische Auswertung	Kategorisierung Zahnengstände: VGG19 gewichtetes Kappa: 0,73 VGG19>VGG16> ResNet101>ResNet50 Zahnextraktionen Durchschnitt: Genauigkeit: 0,91 AUC: 0.952 Sensitivität: 0,836 Spezifität: 0,929	Genauigkeit: 75%-79% AUC: 79%-82%	Gessamtgenauigkeit RF:54,55% SVM:52,73% LR: 49,09% Genauigkeits-prozentsätze für Extraktionsmuster am höchsten für 1. Prämolar OK/UK 81,63%-63,27%, 1.Prämolar OK 72,22%-61,11% Alle anderen Extraktionsmuster Genauigkeits-prozentsätze von 0-36% Die Extraktionsmuster wurden bei 55% korrekt identifiziert. Die Algorithmen

			unterschieden sich nicht wesentlich in ihrer Gesamtgenauigkeit.
Autor*in	**Yasir Suhail et al.**	**Natkritta Chaiprasittikul et al.**	**Ki-Sun Lee et al.**
Jahr	2020	2023	2020
Journal		Healthcare Informatics Research	Applied Science
Land	USA	Thailand	Korea
Studienthema	Extraktions-entscheidung	Chirurgische Notwendigkeit	Chirurgische Notwendigkeit
KI-Struktur	Random Forest	Multi-Layer Perceptron (MLP)	-Modified-Alexnet -MobileNet -Resnet50
Landmarken (N)			N 50
Totaler Datensatz	n.a.	538	333
Trainingsdatensatz	n.a.	484	220
Testdatensatz	287	54	73
Statistische Auswertung	Übereinstimmung bezüglich primären Behandlungsplans zwischen Experten*innen: 65%-71% Genauigkeit: ca 75%	Genauigkeit: 0,963 Sensitivität: 1 AUC: 0,96 Präzision: 0,93 F-Wert: 0,963 -es wurden 2 von 54 Fällen fehldiagnostiziert	Genauigkeit: Modified-Alexnet: 91,9% MobileNet: 83,8% Resnet50: 83,8% Modified-Alexnet: Sensitivität: 0,852 Spezifität: 0,973 AUC: 0,969
Autor*in	**Peilin Li et al.**	**Yuujin Shimizu et al.**	**Bhornsawan Thanathornwong et al.**
Jahr	2019	2022	2018
Journal	Scientific Reports	European Journal of Orhtodontics	Healthcare informatic research
Land	China	Japan	Thailand
Studienthema	Behandlungsplanung	Behandlungsplanung	Behandlungsplanung
KI-Struktur	KI aus 3 neuronalen Netzwerken	2 KI-Systeme	Bayesianisches Netzwerk (BN)

Totaler Datensatz	302	967	1020
Trainingsdaten satz	182	800	1000
Testdatensatz	60	100	20
Statistische Auswertung	Extraktion-Nichtextraktions-Entscheidung: Genauigkeit: 94% AUC: 0,982 Sensitivität: 94,6% Spezifität: 93,8% Genauigkeit der Extraktions- und Verankerungsmuster: 84,2%-92,8%	Präzision: Teilaufgabe 1: 65% Teilaufgabe 2: 48% Erkennungsrate: Teilaufgabe 1: 55% Teilaufgabe 2: 48%	AUC: 91% Genauigkeit: 96% Sensivität: 95% Spezifität: 100%
Autor*in	**Jahnavi Prasad et al.**	**Rutger ter Horst et al.**	**Chihiro Tanikawa et al.**
Jahr	2022	2021	2021
Journal	Dentistry Journal	Journal of Cranio-Maxillofacial Surgery	Scientific Reports
Land	Indien	Niederlande	Japan
Studienthema	Behandlungs-planung	Weichgewebs-vorhersage	Weichgewebs-vorhersage
KI-Struktur	-Decision Tree -Random Forest -XGB (eXtreme gradient Boosting) -Logistic -Regression -K-Neighbors -Linear SVM -Naive Bayes	Autoencoder inspiriertes neuronales Netzwerk	System S und System E
Totaler Datensatz	700	133	137 (72 OP, 65 Extraktionen)
Trainings-datensatz	490	119	n.a.
Testdatensatz	210	14	n.a.
Statistische Auswertung	Genauigkeit Durchschnitt: 84% XGB (höchste Genauigkeit): 87-92%	SDR: 1,0 ± 0,6 mm MTM-Algorithmus basiert: 1,5 ± 0,5 mm Für das DL-basierte	Gesamterfolgsrate (Systemfehler <1mm): System S: 54% System E:98% 100% Erfolgsrate bei einem

		Modell: •Unteres Gesicht: SDR =1,0 ± 0,6 mm •Untere Lippe: SDR =1,1 ± 0,9 mm •Kinn: SDR =1,4 ± 0,9 mm	Systemfehler von <2 mm Durchschnittliche Systemfehler: System S: 0,94 mm ± 0,43 mm System E: 0,69 mm ± 0,28 mm
Autor*in	**James Volovic et al.**	**WooSang Shin et al.**	**Ye-Hyun Kim et al.**
Jahr	2023	2021	2021
Journal	Diagnostics Basel	BMC Oral Health	Journal of personalized medicine
Land	USA	Korea	Korea
Studienthema	Behandlungsdauer	Chirurgische Notwendigkeit	Chirurgische Notwendigkeit
KI-Struktur	8 ML-Modelle 1 neuronales Netzwerk	Deep-Learning-Netzwerk	ResNet-18, -34, -50, -101
Landmarken (N)	31		
Totaler Datensatz	478	840x2	960
Trainingsdatensatz	319	371x2	810
Testdatensatz	159	413x2	150
Statistische Auswertung	Genauste Modelle: -Random Forest -Lasso -Elastic Net Absoluter Fehler im Durchschnitt: 7,27 Monate	Genauigkeit: 0,954 Sensitivität: 0,844 Spezifität: 0,993	Genauigkeit: 93,80% ResNet-18 93,60% ResNet-34 91,13% ResNet-50 91,33% ResNet-101 AUC: 0,979 ResNet-18 0,974 ResNet-34 0,945 ResNet-50 0,944 ResNet-101
Autor*in	**Jia-Nan Zhang et al.**	**Tyler Wood et al.**	**Grant Zakhar et al.**
Jahr	2023	2023	2023
Journal	BMC Oral Health	Diagnostics Basel	Diagnostics Basel
Land	China	USA	USA
Studienthema	Wachstumsvorhersage	Wachstumsvorhersage	Wachstumsvorhersage

KI-Struktur	ResNet50	7 Regressionalgorithmen: -Least Square, Ridge, Lasso, Elastic Net, XGBoost, Random Forest -ein neuronales Netzwerk	7 ML-Modelle -XGBoost -Random Forest -Lasso -Ridge -Linear Regression -Support Vector Regression -Multilayer Perceptron Regressor
Landmarken (N)	38	25	38
Totaler Datensatz	296	163	123
Trainingsdatensatz	256	114	92
Testdatensatz	40	49	31
Statistische Auswertung	Genauigkeit KI: 85% Sensitivität: 0,95 Spezifität: 0,75 AUC: 0,9775 Genauigkeit Kieferorthopäd*innen: 54,2%	Alle Methoden Genauigkeit Y-Achse: 95,8%-97,64% ICC: >0,75 T1 und T2: 96,6%-98,34% Nur T1: 97,52-97,89%	Genauigkeit; T1 und T2: 95,26%-98,35% Nur T1: 95,88%-98,19% ICC T1 und T2: >0,75 (außer Linear Regression) ICC T1: 0,67<ICCs>0,84
Autor*in	**Matthew Parrish et al.**	**Eungyeong Kim et al.**	
Jahr	2023	2023	
Journal	Diagnostics Basel	Diagnostics basel	
Land	USA	Japan	
Studienthema	Wachstumsvorhersage	Wachstumsvorhersage	
KI-Struktur	7 ML-Algorithmen: XGBoost-Regression, Random Forest-Regressor, Lasso, Ridge, Linear Regression, Support Vector Regression, Multilayer Preceptron-Regressor (MLP)	Lasso, Multilayer Perceptron (MLP), Radial Basis Funktion Network, Gradient-Boosted Decision Tree, Mehrfachregressionsanalyse (Kein KI Algorithmus)	
Landmarken (N)	25	26	
Totaler Datensatz	176	59	

Trainingsdatensatz	140	53	
Testdatensatz	36	6	
Statistische Auswertung	Genauigkeit: T1+T2: 97,83%-98,71% ICC: 0,79<ICCs<0,94 Nur T1: 97,56%-98,25% ICC: 0,87<ICCs<0,90	Genauigkeit Lasso: 97,87%-94,45%	

Studientabelle aller inkludierten Studien