

Diplomarbeit

**KI-basierte Vorhersage der Tiefe der Trochlea
ossis femoris anhand seitlicher
Kniegelenksröntgen und MRT-Korrelation - eine
retrospektive Studie**

eingereicht von

Jan Lauritsch

zur Erlangung des akademischen Grades

Doktor der gesamten Heilkunde

(Dr. med. univ.)

an der

Medizinischen Universität Graz

ausgeführt an der

Klinischen Abteilung für Kinderradiologie

Universitätsklinik für Radiologie

unter der Anleitung von

Univ.-Prof. Priv.-Doz. Dr. med. univ. Dr. scient. med. Sebastian Tschauner

Dr. med. univ. Nikolaus Stranger

Graz, 21. Januar 2026

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.

Des Weiteren erkläre ich hiermit, dass, sofern bei der Erstellung dieser Arbeit Künstliche Intelligenz (KI) Werkzeuge zur Generierung und/oder Korrektur bestimmter Textpassagen verwendet wurden, dieser Einsatz unter Einhaltung ethischer Grundsätze, akademischer Integrität und den Vorgaben meiner Universität erfolgte, sowie in Folge dies transparent gemacht und in angemessener Weise gekennzeichnet wurde.

Graz, 21. Januar 2026

Jan Lauritsch eh.

Danksagungen

An dieser Stelle möchte ich mich bei allen großartigen Menschen bedanken, die mich am Weg durch das Medizinstudium begleitet haben.

Das Bearbeiten dieses Diplomarbeitsthemas hat mir besondere Freude bereitet, da sowohl medizinische als auch informatische Aspekte vereint werden. In dieser Hinsicht ist es mir eine Ehre Univ.-Prof. Priv.-Doz. DDr. Sebastian Tschauner als Mentor an meiner Seite zu haben. Seine Expertise im Bereich der künstlichen Intelligenz sowie seine geduldige und humorvolle Art der Diplomarbeitsbetreuung haben wesentlich zum Gelingen dieser Arbeit beigetragen. Insbesondere beim Programmieren und Trainieren der verwendeten KI-Modelle hat er mir wertvolle Unterstützung gegeben. Ein weiteres Dankeschön gilt meinem Zweitbetreuer Dr. med. univ. Nikolaus Stranger für die stets raschen Rückmeldungen und das hilfreiche Feedback.

Ich schätze mich sehr glücklich, dass ich bisher und vor allem im Rahmen meiner Ausbildung immer auf die liebevolle Unterstützung und den Rückhalt meiner Familie zählen konnte. Von ganzem Herzen danke ich hierfür meinen Eltern, meinen Großeltern sowie meiner lieben Schwester Kira. Ein besonderer Dank gilt Lara, auf die ich mich immer verlassen kann und die mir in allen Phasen des Studiums Mut machte und den Rücken stärkte.

Zusammenfassung in Deutsch

Hintergrund:

Die patellofemorale Instabilität stellt ein häufiges Krankheitsbild bei Kindern und Jugendlichen dar, wobei eine dysplastische Trochlea ossis femoris eine wesentliche ätiologische Rolle spielt. Die Tiefe dieser Gelenksrolle gilt als zentraler diagnostischer Parameter und wird für gewöhnlich mit der Magnetresonanztomographie präzise ermittelt. Eine flächendeckend verfügbare und vergleichsweise weniger zeit- und kostenintensive Modalität ist das konventionelle Röntgenbild des Kniegelenks. Im Rahmen der zunehmenden Bedeutung von Künstlicher Intelligenz in der medizinischen Bildverarbeitung war es Ziel dieser Arbeit, zu prüfen, ob Convolutional Neural Networks (CNNs) anhand lateraler Kniegelenksröntgenbilder die Trochleatiefe automatisiert und zuverlässig vorhersagen können.

Methodik:

Im retrospektiven Studiendesign wurden 906 Patientenfälle mit insgesamt 1084 lateralen Röntgenaufnahmen und korrelierenden MRT-Messungen der Trochleatiefe ausgewertet. Acht Varianten der EfficientNet-Architektur (B0 bis B7) wurden in einer dichotomen Klassifikationsaufgabe (Trochleatiefe unter oder über 4 mm) trainiert und getestet. Die Modellleistungen wurden mittels typischer Kennzahlen der Deep-Learning-Modellanalyse beurteilt, wie zum Beispiel Accuracy, F1-Score oder ROC- und PR-Analysen.

Ergebnis:

Die besten Ergebnisse erzielte das Modell EfficientNet-B3 mit einer Accuracy von 67,4% sowie AUC-Werten von 0,724 (ROC) und 0,738 (PR). Die erzielte Genauigkeit reicht zwar noch nicht für eine klinisch verlässliche Anwendung, beweist jedoch die grundsätzliche technische Machbarkeit einer KI-gestützten Beurteilung der Trochleatiefe. Die vorliegende Arbeit liefert zusätzlich auch Impulse für zukünftige Forschungen, um die Modellleistung zu optimieren.

Abstract in English

Background:

Patellofemoral instability is a common pathology in children and adolescents, with dysplastic femoral trochlea playing a significant aetiological role. The depth of this trochlea is considered a key diagnostic parameter and is usually determined precisely using magnetic resonance imaging. A widely available and comparatively less time-consuming and cost-intensive modality is conventional X-ray imaging of the knee joint. In light of the increasing importance of artificial intelligence in medical image processing, the aim of this study was to examine whether convolutional neural networks (CNNs) can automatically and reliably predict trochlear depth based on lateral knee joint X-ray images.

Methods:

In a retrospective study design, 906 patient cases with a total of 1,084 lateral X-rays and correlating MRI measurements of trochlear depth were evaluated. Eight variants of the EfficientNet architecture (B0 to B7) were trained and tested in a dichotomous classification task (trochlear depth below or above 4 mm). Model performance was assessed using typical deep learning model analysis metrics such as accuracy, F1 score, or ROC and PR analyses.

Results:

The EfficientNet-B3 model achieved the best results with an accuracy of 67.4% and AUC values of 0.724 (ROC) and 0.738 (PR). Although the accuracy achieved is not yet sufficient for clinically reliable application, it does prove the fundamental technical possibility of AI-supported assessment of trochlear depth. This work also provides impulses for future research to optimise model performance.

Inhaltsverzeichnis

Eidesstattliche Erklärung	II
Danksagungen	III
Zusammenfassung in Deutsch	IV
Abstract in English.....	V
Inhaltsverzeichnis.....	VI
Abbildungsverzeichnis.....	VIII
Tabellenverzeichnis	IX
1 Einleitung	1
2 Patellofemorale Instabilität	3
2.1 Aufbau und Biomechanik des Kniegelenks	3
2.1.1 Anatomie des Patellofemoralgelenks	3
2.1.2 Biomechanik der Patellaführung	4
2.2 Patellofemorale Instabilität: Definition, Epidemiologie und Ursachen	6
2.3 Trochleadysplasie als ätiologischer Faktor der PFI	7
2.3.1 Sulcuswinkel	7
2.3.2 Trochleatiefe	8
2.3.3 Klassifikation nach Dejour.....	9
3 Künstliche Intelligenz in der medizinischen Bildverarbeitung	11
3.1 Aufbau und Funktionsweise neuronaler Netze.....	11
3.1.1 Das künstliche Neuron.....	11
3.1.2 Netzstrukturen.....	12
3.1.3 Lernen neuronaler Netzwerke.....	13
3.2 Convolutional Neural Networks (CNNs)	14
3.3 EfficientNet als moderne CNN-Architektur	14
3.4 Trainingsstrategien für Deep Learning in der Medizin	15
3.4.1 Transfer Learning.....	15

3.4.2	Datenaugmentierung	15
4	Material und Methoden.....	17
4.1	Auswahl der Studienpopulation.....	17
4.1.1	Einschlusskriterien	17
4.1.2	Ausschlusskriterien	18
4.1.3	Größe der Studienpopulation.....	18
4.2	Datenerhebung	18
4.2.1	MRT-Messung der Trochleatiefe	18
4.2.2	Selektion des lateralen Röntgenbilds.....	19
4.3	Bildverarbeitung	20
4.4	Datenaufbereitung.....	21
4.5	Maschinenlernen mit EfficientNet-Modellen	22
4.6	Ethikvotum	23
5	Resultate	24
5.1	Konfusionsmatrizen.....	24
5.2	ROC-Analyse	26
5.3	Precision-Recall-Analyse	28
5.4	Leistungskennzahlen	30
6	Diskussion.....	34
6.1	Klinische Anwendbarkeit	34
6.2	Vergleich mit ähnlicher Literatur	35
6.3	Limitationen.....	37
6.4	Ausblick.....	38
6.5	Conclusio	40
	Literaturverzeichnis	41

Abbildungsverzeichnis

Abbildung 1: Auf die Patella wirkende Kräfte bei Kontraktion des M. quadriceps femoris in Frontalebene und Sagittalebene	5
Abbildung 2: Messung des Sulcuswinkels an der knöchernen und knorpeligen Trochlea	8
Abbildung 3: Tiefenmessungen an der knöchernen sowie knorpeligen Trochlea	9
Abbildung 4: Dejour-Klassifikation A-D in axialen und lateralen Projektionen	10
Abbildung 5: Schematische Darstellung eines künstlichen Neurons	12
Abbildung 6: Schematische Darstellung eines einfachen Feedforward-Neuronalen Netzwerks	13
Abbildung 7: Ausschnitt aus dem Excel-Datenblatt mit den erhobenen Daten	21
Abbildung 8: aug_transforms-Funktion zur Datenaugmentierung mit Parametern	22
Abbildung 9: Konfusionsmatrizen der getesteten EfficientNet-Modelle B0 bis B7	25
Abbildung 10: ROC-Kurven und AUC-Werte der EfficientNet-Modelle B0 bis B7	27
Abbildung 11: Precision-Recall-Kurven der EfficientNet-Modelle B0 bis B7	29

Tabellenverzeichnis

Tabelle 1: Technische Parameter der Trainingsumgebung	22
Tabelle 2: Sensitivität, Spezifität und errechneter Youden-Index der Klassifikationsaufgabe in den EfficientNet-Modellen B0 bis B7	26
Tabelle 3: AUC-Werte inklusive ihrer 95%-Konfidenzintervalle in der ROC-Analyse der EfficientNet- Modelle B0-B7	28
Tabelle 4: AUC-Werte inklusive ihrer 95%-Konfidenzintervalle in der PR-Analyse der EfficientNet- Modelle B0-B7	30
Tabelle 5: Precision und Recall der EfficientNet-Modelle mit Weighted-Mittelung.....	31
Tabelle 6: Precision und Recall der EfficientNet-Modelle mit Macro-Mittelung.....	32
Tabelle 7: Accuracy-Werte der EfficientNet-Modelle B0-B7	33

1 Einleitung

Die patellofemorale Instabilität ist ein typisches Krankheitsbild der Kinderorthopädie und tritt besonders häufig bei jungen und sportlich aktiven Menschen auf. Neben rezidivierenden Subluxationen oder Luxationen der Kniescheibe kann eine patellofemorale Instabilität erhebliche funktionelle Einschränkungen sowie Schmerzen verursachen. Eine der wichtigsten Ursachen für das Auftreten einer solchen Instabilität ist die Trochleadysplasie, bei der die Trochlea ossis femoris unzureichend oder nicht stark genug ausgeprägt ist, um ihrer Funktion als stabile Führungs- und Gleitrinne für die Patella nachzukommen.

Ein zentraler Parameter zur radiologischen Befundung einer Trochlea ossis femoris ist die Trochleatiefe, die den Höhenunterschied zwischen den vorderen Konturen der Femurkondylen und dem tiefsten Punkt der Gelenkrinne darstellt. Im Rahmen einer Trochleadysplasie fällt diese Tiefe zu gering aus.

In der klinischen Praxis wird die Trochleatiefe mithilfe von Schnittbildverfahren, wie die Magnetresonanztomographie oder Computertomographie, bestimmt. Trotz der hohen diagnostischen Genauigkeit sind diese Verfahren jedoch kosten- und personalintensiv, zeitaufwendig und nicht in jeder klinischen Einrichtung verfügbar. Im Gegensatz dazu stellt die konventionelle Röntgenaufnahme des Kniegelenks eine vergleichsweise einfache, kostengünstige und verbreitete Modalität dar. Sollte es möglich sein, die Trochleatiefe anhand lateraler Kniegelenksröntgen zuverlässig zu bestimmen, so kann dies zur Vereinfachung und Beschleunigung der diagnostischen Abklärung führen.

Eine neue Möglichkeit in diesem Zusammenhang eröffnet der Einsatz von künstlicher Intelligenz, welche in der medizinischen Bildgebung und insbesondere in der muskuloskelettalen Radiologie beachtliche Fortschritte erzielt. Ziel der vorliegenden Arbeit ist es, zu untersuchen, ob KI-Algorithmen in der Lage sind, anhand lateraler Röntgenaufnahmen des Kniegelenks die Trochleatiefe automatisiert zu bestimmen und zwischen flachen und normalen Trochleae zu unterscheiden.

Die Arbeit gliedert sich in einen theoretischen Teil, welcher die patellofemorale Instabilität sowie Grundlagen der künstlichen Intelligenz erarbeitet, sowie einen

empirischen Teil, der die Studienmethodik sowie die erzielten Ergebnisse beschreibt. Abschließend folgen eine Diskussion und Schlussfolgerung.

2 Patellofemorale Instabilität

Die Stabilität des patellofemorale Gelenks ist eine wichtige Voraussetzung für eine schmerzfreie und korrekte Bewegung des Kniegelenks. Abweichungen in der anatomischen Struktur können die biomechanische Führung beeinträchtigen und verschiedene Beschwerden verursachen, darunter auch eine vollständige Luxation der Kniescheibe. Eine solche patellofemorale Instabilität stellt bei jungen Patientinnen und Patienten eine häufige Ursache für Kniebeschwerden dar und kann die Lebensqualität erheblich einschränken.

2.1 Aufbau und Biomechanik des Kniegelenks

Das Kniegelenk gilt als das größte Gelenk des menschlichen Körpers und wird von Oberschenkelknochen (*Femur*), Kniescheibe (*Patella*) und Schienbein (*Tibia*) gebildet. Es zählt zu den zusammengesetzten Gelenken und stellt einen sogenannten Trochoginglymus dar, da es sowohl eine Dreh- als auch eine Scharnierkomponente beinhaltet. Zu den zwei Einzelgelenken des Kniegelenks gehören die *Articulatio femorotibialis* sowie die *Articulatio femoropatellaris*.

Die *Articulatio femorotibialis* bildet die Verbindung zwischen den Femurkondylen und dem Tibiaplateau. Dieses Gelenk ist verantwortlich für die Flexion und Extension des Kniegelenks und ermöglicht auch eine gewisse Rotationsbewegung beim gebeugten Knie.

Im Fokus der vorliegenden Arbeit steht die Instabilität der *Articulatio femoropatellaris*. Im Folgenden werden die Anatomie und Biomechanik dieses Gelenks im Detail ausgeführt.

2.1.1 Anatomie des Patellofemoralgelenks

Im Patellofemoralgelenk (*Articulatio femoropatellaris*) artikulieren die *Facies patellaris* des Femur sowie die *Facies articularis* der Patella miteinander, welche beide mit hyalinem Knorpel überzogen sind. Die Gelenksfläche der Patella weist einen vertikalen First auf, welcher als *Crista patellae* bezeichnet wird. Dieser teilt die Rückseite der Patella in eine laterale und mediale Gelenksfacette auf, wobei der mediale Anteil schmaler ist. [1]

Das distale Ende des Femurs weist zwei Verdickungen auf, welche den *Condylus medialis* und *lateralis* darstellen. An der Hinterseite bildet sich zwischen den Kondylen die *Fossa intercondylaris*. Diese Grube beheimatet das vordere und hintere Kreuzband und wird nach proximal durch die horizontal verlaufende *Linea intercondylaris* begrenzt. Nach vorne laufen die Flächen zur *Facies patellaris* zusammen, in der die Patella zu liegen kommt. [2]

Insgesamt weisen diese beschriebenen Strukturen des distalen Femurs eine Rollenform auf, die als *Trochlea ossis femoris* bezeichnet wird. Die Tiefe der Trochlea wird üblicherweise im zentralen eingesunkenen Abschnitt gemessen, der in der Literatur auch als *Sulcus femoris* bezeichnet wird. [3]

Die Patella ist in die Sehne des *Musculus quadriceps femoris* eingebettet und gilt als das größte Sesambein des menschlichen Körpers. Die Fasern der unterschiedlichen Muskelköpfe strahlen im oberen Pol der Kniescheibe (*Basis patellae*) ein. An der unteren Spitze der Patella (*Apex patellae*) entspringt das *Ligamentum patellae*, welches in weiterer Folge an der *Tuberositas tibiae* ansetzt. [1]

Neben der aktiven muskulären Stabilisation verfügt das Patellofemoralgelenk auch über passive Stabilisatoren. Hierzu zählen seitlich der Kniescheibe stabilisierende Verstärkungen der Kniegelenkscapsel, welche als *Retinacula patellae* bezeichnet werden. Eine besondere Rolle in der Weichteilstabilisation der Kniescheibe spielt das *Ligamentum patellofemorale mediale* (engl. *medial patellofemoral ligament* = *MPFL*). Diese dünne Bandstruktur verläuft vom proximal-medialen Rand der Patella zum medialen *Epicondylus femoris* und wirkt einer Luxation der Patella nach lateral entgegen. [4]

2.1.2 Biomechanik der Patellaführung

Die Patella spielt als Sesambein eine wichtige Rolle in der Extensionsbewegung des Kniegelenks. Diese wird durch eine Kontraktion des *Musculus quadriceps femoris* mit seinen einzelnen Muskelköpfen ermöglicht. Abbildung 1 aus Amis 2007 veranschaulicht die bei einer Kontraktion wirkenden Kräfte in zwei Ebenen. Die obere Skizze stellt die Koronalebene dar und zeigt, dass die verschiedenen Anteile des Muskels aus unterschiedlichen Richtungen Zugkräfte auf die Patella übertragen. Werden diese eingezeichneten Kraftvektoren summiert, so ergibt sich

ein Vektor, der annähernd parallel zur anatomischen Achse des Femurs verläuft und diese Zugkraft über das *Ligamentum patellae* auf die Tibia überträgt. In der unteren Skizze wird die Kraftwirkung in sagittaler Ebene dargestellt. Hier gelangt zur Darstellung, dass zusätzlich Kräfte nach posterior auf die Patella wirken, um einen stabilen Kontakt mit der Trochlea ossis femoris zu unterstützen. [4]

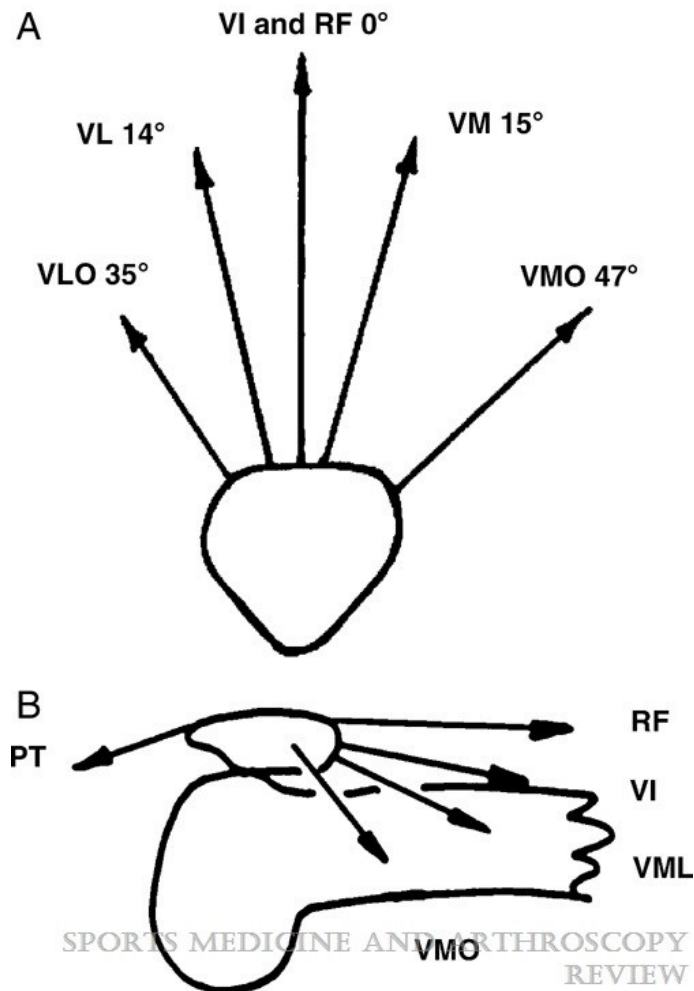


Abbildung 1: Auf die Patella wirkende Kräfte bei Kontraktion des M. quadriceps femoris in Frontalebene (A) und Sagittalebene (B). Reproduced with permission from: Amis, Andrew A. "Current concepts on anatomy and biomechanics of patellar stability." *Sports medicine and arthroscopy review* vol. 15,2 (2007): 48-56. doi:10.1097/JSA.0b013e318053eb74. © 2007 Wolters Kluwer Health, Inc.

Die Stabilisierung der Patella bei Bewegung wird je nach Flexionsgrad durch verschiedene anatomische Strukturen ermöglicht. Bei vollständiger Streckung im Kniegelenk befindet sich die Patella kranial der Trochlea und berührt diese lediglich mit der distalen lateralen Gelenkfacetten. In diesem Zustand wird die Kniescheibe durch die in Kapitel 2.1.1 beschriebenen Weichteilstrukturen stabilisiert. Bei beginnender Beugung wird die Patella in die Trochlea gezogen und erhält somit eine knöcherne Stabilität. Initial tritt die laterale Gelenkfacetten der Patella in die Trochlea

ein, was zu einer leichten medialen Translationsbewegung führt. Ab einer Flexion von 30° kommt auch die mediale Gelenkfacetten in der Gelenksrinne zu liegen. Bei zunehmender Beugung zeigt die Patella sowohl eine laterale Translationsbewegung sowie eine Kippbewegung (Tilt) in die gleiche Richtung. [5]

2.2 Patellofemorale Instabilität: Definition, Epidemiologie und Ursachen

Mit patellofemoraler Instabilität (PFI) wird ein Zustand beschrieben, bei dem die Patella in ihrer physiologischen Führung in der Trochlea ossis femoris gestört ist. Klinisch äußert sich das meist mit einem subjektiven Instabilitätsgefühl im Kniegelenk, wiederholten Subluxationen oder Luxationen der Patella und Kniegelenksergüssen. Die PFI tritt überwiegend im jungen Alter auf, während vor allem Mädchen eine erhöhte Inzidenz aufweisen (29/100.000 bei Frauen zwischen 10 und 17 Jahren). In der Allgemeinbevölkerung liegt die Inzidenz bei 5,8 von 100.000 Personen. Wenn eine erstmalige Patellaluxation auftritt, beträgt das Risiko für eine erneute Luxation 17-33%. Bei wiederkehrenden Luxationen beträgt die Reluxationsrate über 50%. Es zeigt sich auch ein signifikanter Zusammenhang zwischen Patellaluxationen und auftretender Osteoarthritis im Patellofemoralgelenk. [6]

In Pauyo et al. werden vier Hauptrisikofaktoren für das Auftreten einer PFI näher beschrieben:

1. **Trochleadysplasie:** Bei einer dysplastischen Trochlea ist die Gelenksrinne flacher oder schmaler ausgeprägt als im Normalfall. Die verschiedenen Parameter einer Trochleadysplasie werden aufgrund des Fokus dieser Arbeit im Kapitel 2.3 ausführlich behandelt.
2. **Patella alta:** Im Falle einer hochstehenden Patella tritt diese erst verspätet in die Trochlea ein und ist aufgrund der geringer ausfallenden knöchernen Stabilisierung somit vulnerabler für Luxationen nach lateral.
3. **Erhöhter TT-TG-Abstand:** Mit diesem Parameter wird der Abstand zwischen der Tuberositas tibiae (TT) und dem tiefsten Punkt der Trochlea (TG für „trochlear groove“) bezeichnet, welcher üblicherweise in einer MR-Bildgebung gemessen wird. Der Wert gibt an, wie weit lateral die TT zur TG steht und soll im Normalfall unter 20 Millimeter betragen.

4. **Torsionsdeformitäten:** Fehlstellungen wie beispielsweise eine Antetorsion des Femurs oder die Außentorsion der Tibia können dazu führen, dass die Patella mehr Zugkraft nach lateral erfährt als im Normalfall.

Als weitere Risikofaktoren gelten ein abnormaler lateraler Tilt der Patella, eine Patelladysplasie, ein Genu valgum, eine Atrophie oder Schwäche des *Musculus vastus medialis obliquus* (VMO), Insuffizienzen des medialen patellofemoralen Ligaments (MPFL) und einer generalisierte Hyperlaxizität. [6]

2.3 Trochleadysplasie als ätiologischer Faktor der PFI

Die Trochleadysplasie gilt als einer der Hauptrisikofaktoren für die patellofemorale Instabilität und beschreibt eine zu flache oder sogar konvexe Morphologie der Trochlea. Die Diagnostik erfolgt mittels Bildgebung wie Röntgen, Computertomographie und Magnetresonanztomographie, wobei verschiedene Messungen zur Beurteilung der Trochleageometrie erhoben werden. Als besonders zuverlässig gelten hierbei die Messungen des Sulcuswinkel, die Trochleatiefe sowie die Klassifikation nach Dejour. Im Folgenden werden diese drei Parameter näher erläutert.

2.3.1 Sulcuswinkel

Der Sulcuswinkel ist ein Parameter, um die Flachheit der Trochlea zu quantifizieren. Gemessen wird jener Winkel, der zwischen zwei Linien gebildet wird, die vom tiefsten Punkt der Trochlea bis zum vordersten Punkt des medialen bzw. lateralen Femurkondylen verlaufen. Die Abbildung 2 stellt dar, dass die Messung sowohl an den knöchernen als auch an den knorpeligen Anteilen vorgenommen werden kann.

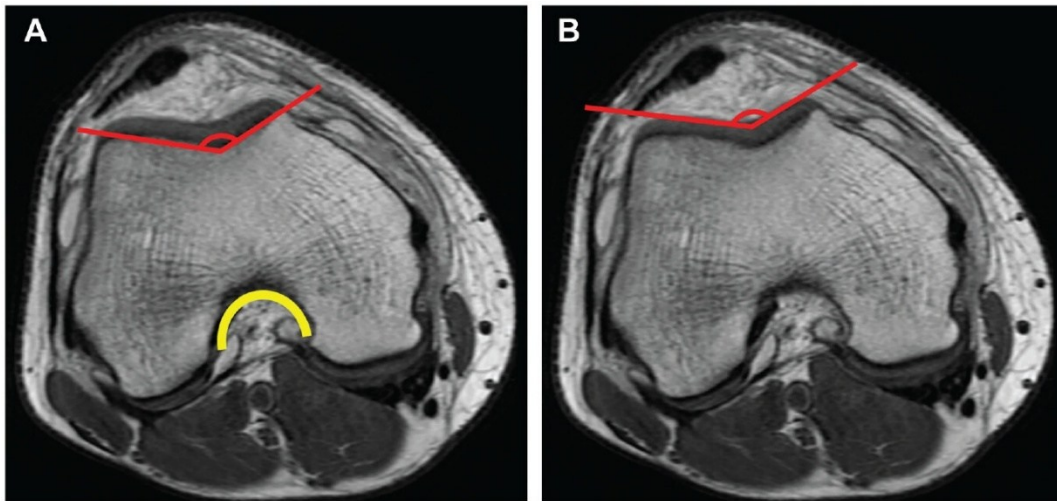


Abbildung 2: Messung des Sulcuswinkels an der knöchernen (A) und knorpeligen (B) Trochlea. Darstellung aus Tanaka MJ, Sodhi A, Wadhavkar I, Kane K, et al. Redefining Trochlear Dysplasia: Normal Thresholds Vary by Measurement Technique, Landmarks, and Sex. Am J Sports Med. 2023; 51(5): [1-9]. Copyright © 2023 by SAGE Publications. Reprinted by permission of SAGE Publications.

Die Studie von Tanaka et al. zeigt, dass die Ergebnisse der Sulcuswinkelmessung der knorpeligen Trochlea größer ausfallen als bei der knöchernen Trochlea. Als Cutoff-Werte zur Diagnose einer Trochleadysplasie werden Winkel über 145° bei knöchernen und über 154° bei knorpeligen Strukturen angegeben. Weiters zeigt sich auch ein geschlechtsspezifischer Unterschied. Sowohl die Messungen und der ermittelte Cutoff-Wert der knöchernen und knorpeligen Trochlea fallen bei weiblichen Patienten insgesamt höher aus als bei männlichen Patienten. [7]

2.3.2 Trochleatiefe

Mit der Trochleatiefe wird ein weiterer Parameter angegeben, der die Flachheit der Trochlea beschreibt. Hierbei wird zunächst der Mittelwert der Abstände der vordersten Punkte des medialen und lateralen Femurkondyls zur hinteren Kondylenlinie ermittelt. Von diesem Wert wird dann der Abstand des tiefsten Punkts der Trochlea zur hinteren Kondylenlinie abgezogen, wodurch sich die Trochleatiefe ergibt. Mit Abbildung 3 gelangt diese Messmethodik zur Darstellung.

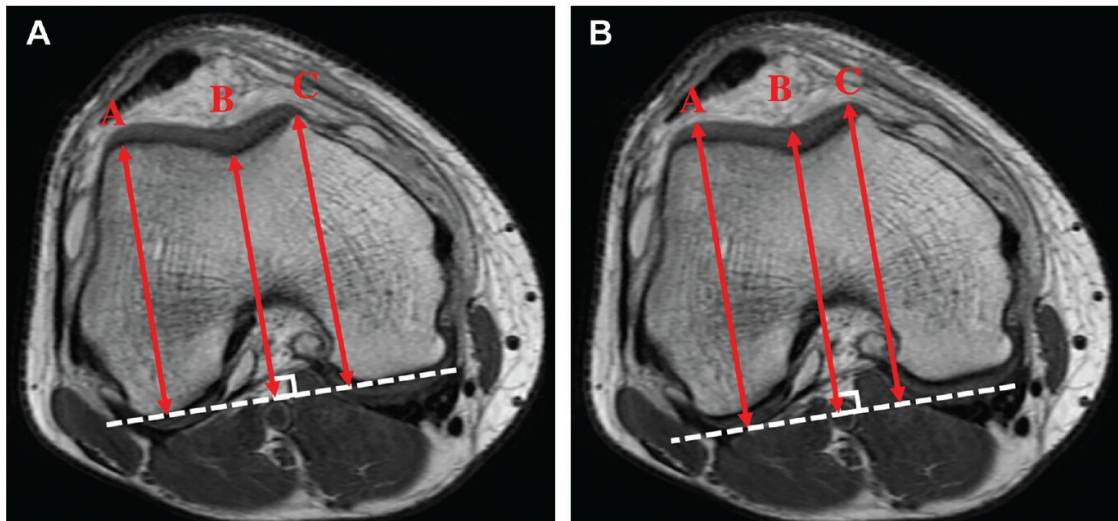


Abbildung 3: Tiefenmessungen an der knöchernen (A) sowie knorpeligen (B) Trochlea. Die ermittelten Abstände werden in die Formel $((A+C)/2)-B$ eingesetzt, um die Trochleatiefe zu erhalten. Darstellung aus Tanaka MJ, Sodhi A, Wadhavkar I, Kane K, et al. Redefining Trochlear Dysplasia: Normal Thresholds Vary by Measurement Technique, Landmarks, and Sex. Am J Sports Med. 2023; 51(5): [1-9]. Copyright © 2023 by SAGE Publications. Reprinted by permission of SAGE Publications.

Die Tiefenmessung der knöchernen Trochlea fällt im Schnitt höher aus als die der knorpeligen Trochlea. Als Cutoff-Werte zur Diagnose einer Trochleadysplasie wurden in Tanaka et al. Werte unter 4 mm für die knöchernen und unter 3 mm für die knorpelige Trochlea ermittelt. Während die Cutoff-Werte der knorpeligen Trochlea unabhängig vom Geschlecht sind, zeigen sich bei der Vermessung der knöchernen Trochlea geschlechtsspezifische Unterschiede. Hierbei gilt bei männlichen Patienten eine Tiefe unter 5 mm, bei weiblichen Patienten eine Tiefe unter 4 mm als pathologisch.

2.3.3 Klassifikation nach Dejour

Die von Henri Dejour entwickelte Klassifikation ermöglicht eine Einteilung der Trochleadysplasie in vier Schweregrade (A-D). Diese erfolgt durch drei radiologische Zeichen, die im lateralen Kniegelenksröntgen aber auch in CT oder MR-Schichtbildern darstellbar sind. Das sogenannte „Crossing Sign“ im lateralen Röntgen spielt hierbei eine entscheidende Rolle und gilt als positiv, wenn der tiefste Punkt des Sulcus femoris den Vorderrand eines Femurkondyls kreuzt. Je weiter distal diese Kreuzung auftritt, desto schwergradiger ist die Trochleadysplasie einzustufen. Das Vorhandensein eines supratrochleären Sporns, ein Knochenvorsprung über der Trochlea, ist ein weiteres relevantes Zeichen in der Dejour-Klassifikation. Bei einer Doppelkontur handelt es sich um ein radiologisches Zeichen, das auf eine Hypoplasie einer Trochleafacette hinweist und im lateralen

Röntgenbild als zweite Linie hinter der Kontur der gegenüberliegenden Facette sichtbar wird. [8]

Anhand dieser Zeichen erfolgt die Einteilung der Dejour-Schweregrade:

Dejour Typ A: Crossing Sign

Dejour Typ B: Crossing Sign und supratrochleärer Sporn

Dejour Typ C: Crossing Sign und Doppelkontur

Dejour Typ D: Crossing Sign, supratrochleärer Sporn und Doppelkontur

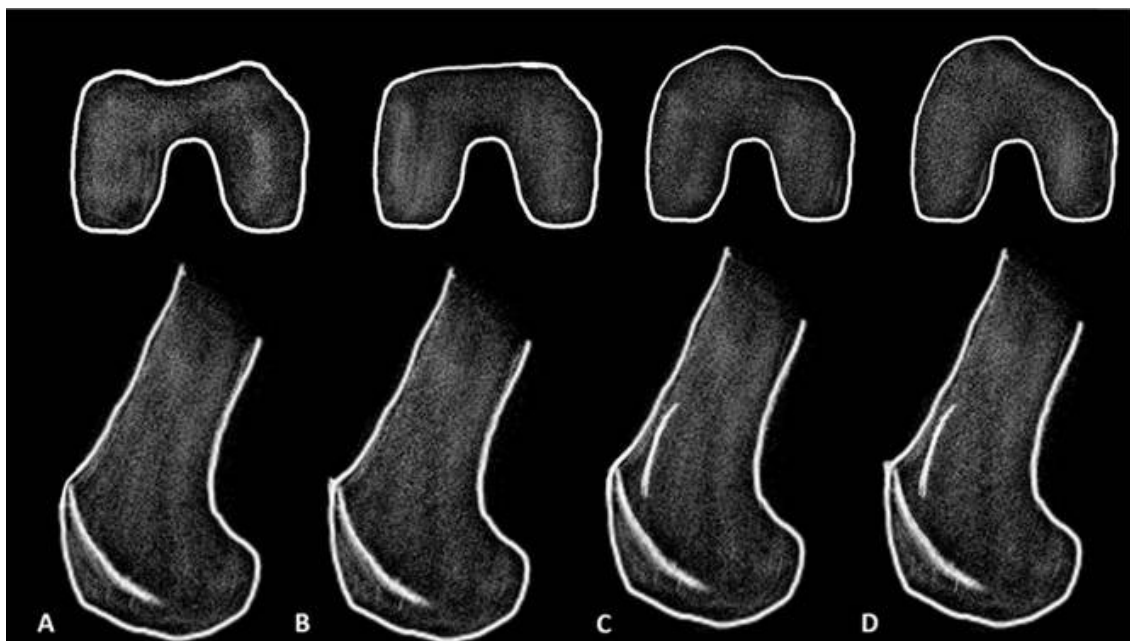


Abbildung 4: Dejour-Klassifikation A-D in axialen und lateralen Projektionen. Reproduced with permission from: Kazley JM, Banerjee S. Classifications in Brief: The Dejour Classification of Trochlear Dysplasia. Clin Orthop Relat Res. 2019; 477(10): 2236–2240. © 2019 Wolters Kluwer Health, Inc.

Die Abbildung 4 aus der Publikation von Kazley und Banerjee illustriert die einzelnen Grade der Dejour-Klassifikation sowohl in lateraler als auch in axialer Projektion.

3 Künstliche Intelligenz in der medizinischen Bildverarbeitung

Das Ziel von Computerprogrammen, die sich künstlicher Intelligenz (KI) bedienen, ist es, menschliche kognitive Prozesse nachzuahmen. Eine Methode der KI ist das Maschinlernen (engl. Machine Learning, ML), bei welchem ein Computer für das Lösen eines Problems eigenständig Muster erkennt und Algorithmen erlernt ohne hierfür explizit programmiert werden zu müssen [9].

Deep Learning beschreibt eine spezielle Variante des maschinellen Lernens. Hierbei kommen sogenannte „tiefe“ neuronale Netze zur Anwendung, die aus hintereinander aufgebauten Schichten bestehen. Ziel dieses Aufbaus ist es, nach einem Trainingsprozess mit annotierten oder klassifizierten Beispieldaten, für das Programm unbekannte Eingabedaten in sinnvolle Ausgaben zu transformieren. Diese Technik wird als end-to-end-Ansatz verstanden, da innerhalb dieses Prozesses keine manuellen Schritte nötig sind. Typisches Anwendungsbeispiel in der medizinischen Bildverarbeitung wäre die automatisierte Erkennung von Neoplasien in einer Computertomographie der Lunge. [10]

3.1 Aufbau und Funktionsweise neuronaler Netze

Die Funktion des menschlichen Gehirns basiert auf vielen Neuronen, die Signale über Dendriten empfangen und dann im Sinne des sogenannten Alles-oder-Nichts-Gesetzes bei Erreichen eines bestimmten Schwellenwertes ein Signal über das Axon weitergeben. Dieses biologische Prinzip dient als Inspiration für künstliche neuronale Netze.

3.1.1 Das künstliche Neuron

Die kleinste Einheit eines neuronalen Netzes ist das künstliche Neuron. Angelehnt an das biologische Neuron erhält dieses mehrere Eingangssignale und erzeugt daraus nach einer Verarbeitung ein Ausgangssignal. Die Eingangswerte x_i werden zunächst mit den dazugehörigen Gewichten w_i multipliziert und anschließend zur sogenannten Präaktivierung z summiert. Auf diese gewichtete Summe wird eine Aktivierungsfunktion φ angewendet und letztendlich das Ausgangssignal des

Neurons y erzeugt. Aktivierungsfunktionen sind in der Regel nicht-lineare mathematische Funktionen wie beispielsweise die Sigmoid-Funktion.

Zusätzlich zu den gewichteten Eingabesignalen besitzt jedes Neuron einen Bias-Wert b , der zur Präaktivierung addiert wird. Dieser kann als weiteres Gewicht mit einem konstanten Eingabewert betrachtet werden und dient als Schwellenwertverschiebung.

In Abbildung 5 wird der Aufbau eines künstlichen Neurons schematisch dargestellt.

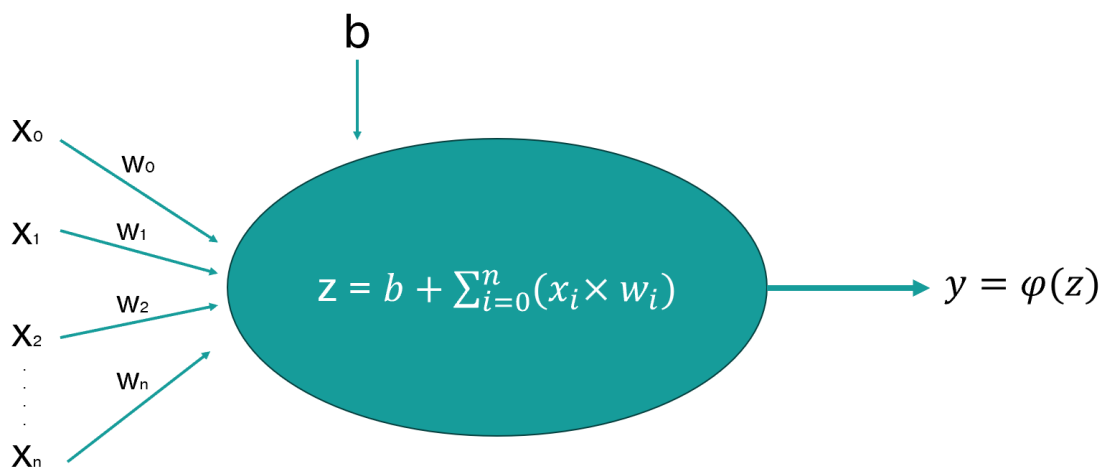


Abbildung 5: Schematische Darstellung eines künstlichen Neurons. (Eigene Darstellung)

3.1.2 Netzstrukturen

In neuronalen Netzen sind die zuvor beschriebenen künstlichen Neuronen in Schichten (engl. *layer*) angeordnet. Man unterscheidet grundsätzlich zwischen drei Arten von Schichten: Die Eingabeschicht (*input layer*), eine oder mehrere versteckte Schichten (*hidden layer*) sowie einer Ausgabeschicht (*output layer*). In der Eingabeschicht werden die Rohdaten, wie beispielsweise Bildwerte, an das neuronale Netz übergeben. Diese Werte gelangen dann durch die versteckten Schichten und liefern in der Form der Ausgabeschicht das Ergebnis des Rechenprozesses.

Zwischen benachbarten Schichten ist in klassischen Netzarchitekturen jedes Neuron mit allen Neuronen der folgenden Schicht verbunden, was als *fully connected* bezeichnet wird. Wenn Informationen ausschließlich vorwärts und in

einer Richtung durch das Netz fließen, spricht man von einem Feedforward-Neuronalen Netzwerk (FNN), welches in Abbildung 6 schematisch dargestellt wird. Neben diesen Netzen existieren auch noch alternative Architekturen, die unter anderem zyklische neuronale Verbindungen aufweisen. Hier lassen sich beispielhaft rekurrente neuronale Netze (RNNs) nennen. [10]

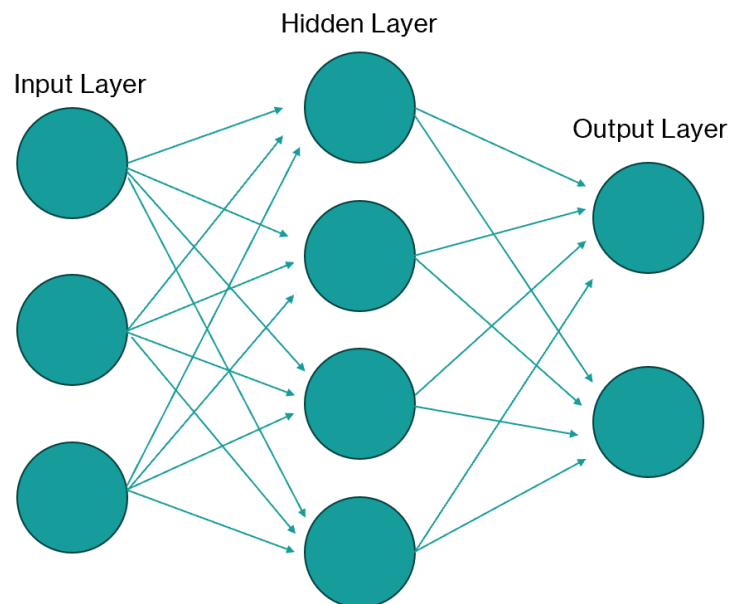


Abbildung 6: Schematische Darstellung eines einfachen Feedforward-Neuronalen Netzwerks (Eigene Darstellung)

3.1.3 Lernen neuronaler Netzwerke

Beim Trainingsprozess von neuronalen Netzen in der medizinischen Bildverarbeitung wird im Regelfall die Methode des überwachten Lernens (*supervised learning*) verwendet. Ein einfaches FNN wird hierbei mit initial zufällig gewählten Gewichtsparametern auf einen Beispieldatensatz trainiert, der menschlich annotierte oder klassifizierte Beispiele enthält. Auf Basis dieser Daten erzeugt das Netz eine Ausgabe, die dann mit der menschlich annotierten Referenz (*ground truth*) verglichen wird. Der Unterschied zwischen den beiden Ausgaben wird mit einer Verlustfunktion (*loss function*) quantifiziert. Anschließend erfolgt der Backpropagation-Algorithmus, bei dem der ermittelte Fehler rückwärts durch das Netzwerk gerechnet wird und die Gewichte in die korrekte Richtung angepasst werden. Der numerische Parameter der Lernrate (*learning rate*) bestimmt, wie groß diese Anpassung ausfallen soll. Beim Training wird der Datensatz mehrmals

durchlaufen und die Parameter iterativ optimiert. Ein solcher Durchlauf wird als Epoche bezeichnet. [11]

3.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) sind künstliche neuronale Netze mit einer speziellen Architektur, die für die Verarbeitung von Daten mit Gitterstruktur insbesondere Bilddaten ausgelegt sind. Bei diesen CNNs kommen drei Arten von Schichten zur Anwendung:

- **Convolution Layers:** Hierbei wird an jeder Position ein lokales Feld des Bildes mit einem Filter multipliziert und die Ergebnisse summiert. Es entstehen folglich sogenannte Feature Maps, in denen markante Strukturen wie beispielsweise Kanten sichtbar gemacht werden.
- **Pooling Layers:** Dieser reduziert die räumliche Dimension dieser Feature Maps, was zwei Vorteile mit sich bringt. Das Netzwerk wird robuster gegenüber kleinen Bildverschiebungen und die Anzahl der lernenden Parameter sowie der benötigten Rechenleistung verringert sich.
- **Fully Connected Layers:** Die Struktur dieser Schicht wird in Kapitel 3.1.2 beschrieben und verwendet die in den vorherigen Schichten extrahierten Merkmale zur Entscheidungsfindung beispielsweise zur Klassifikation eines Bildes in nicht-pathologisch/pathologisch.

Auf alle beschriebenen Schichten folgt eine nicht-lineare Aktivierung. Das Training eines CNNs erfolgt durch die Methode der Backpropagation. [12]

3.3 EfficientNet als moderne CNN-Architektur

EfficientNet ist eine spezielle CNN-Architektur, die im Jahr 2019 von Google AI entwickelt wurde. Ziel dieses Netzes ist eine hohe Genauigkeit bei optimaler Recheneffizienz. Grundlage dieses Ansatzes ist das sogenannte Compound Scaling, bei welchem mehrere Netzwerkdimensionen gleichzeitig skaliert werden. Vor der Entwicklung von EfficientNet wurden bei CNNs nur die Tiefe, die Breite oder die Eingabebildauflösung einzeln verändert. Die EfficientNet-Architektur kombiniert diese drei Aspekte.

Mit EfficientNet-B0 bis EfficientNet-B7 existieren acht Modelle, die auf diesem Prinzip basieren, wobei das B7-Modell in Tests die größte Effizienz aufweist. Tests

wurden unter anderem auf dem standardisierten ImageNet-Datensatz durchgeführt, welcher 14 Millionen gelabelte Bilder in 1.000 Klassen beinhaltet. EfficientNet-B7 erreicht hierbei eine beste Genauigkeit von 84,3% und benötigt ungefähr acht mal weniger Parameter als das vergleichbare Modell GPipe. [13]

3.4 Trainingsstrategien für Deep Learning in der Medizin

Deep Learning kann einen wichtigen Beitrag in der medizinischen Bildanalyse liefern. Hierbei lässt sich vor allem die Anwendungsmöglichkeit der Bildklassifikation nennen, in der auf Basis von Bildaufnahmen automatisiert entschieden wird, ob eine bestimmte Diagnose, beispielsweise eine Fraktur, vorliegt oder nicht. Da die Trainingsdatensätze in der medizinischen Bildverarbeitung vergleichsweise klein und limitiert sind, reicht eine hohe Datenqualität allein nicht aus. Im Folgenden werden zwei gängige Trainingsstrategien zur Modelloptimierung näher erläutert. [10]

3.4.1 Transfer Learning

Das Transfer Learning beschreibt einen Prozess, bei welchem bereits vortrainierte neuronale Netze auf eine andere Fragestellung feinjustiert werden. In der medizinischen Bildanalyse werden üblicherweise CNNs mit dem im Kapitel 3.3 erwähnten ImageNet-Datensatzes vortrainiert, um sie dann im Anschluss auf die jeweilige medizinische Aufgabenstellung anzupassen. In der Studie von Shin et al. wurde Transfer Learning anhand von zwei klinischen Anwendungsfällen untersucht. Hierbei zeigte sich, dass feinjustierte Modelle auf Basis vortrainierter CNNs einen Vorteil gegenüber Modellen mit zufälliger Initialisierung liefern. [14]

3.4.2 Datenaugmentierung

Beim Training von neuronalen Netzwerken gilt es ein „Overfitting“ zu verhindern. Man spricht davon, wenn sich das Modell zu stark an den Trainingsdatensatz anpasst und für neue Testdaten schlechtere Ergebnisse erzielt. Eine bewährte Methode, um einen solchen Zustand zu vermeiden, ist die sogenannte Datenaugmentierung, bei der die Bilddaten im Trainingsprozess künstlich verändert werden, um die Vielfalt des Datensatzes zu erhöhen. [15]

Die im Rahmen dieser Studie vorliegende Python-Bibliothek *FastAI* stellt die Funktion *aug_transforms* bereit, welche eine solche Datenaugmentierung

ermöglicht. Zu den automatischen Bildtransformationen, die mit dieser Funktion durchgeführt werden können, zählen beispielsweise horizontales und vertikales Spiegeln, zufällige Drehungen und Skalierungen, Änderungen der Bildhelligkeit oder Verzerrungen der Bildgeometrie.

4 Material und Methoden

Um die Zuverlässigkeit von Machine-Learning-Algorithmen bei der Vorhersage der Trochleatiefe anhand von lateralen Röntgenaufnahmen zu überprüfen, wird ein retrospektives Studiendesign gewählt. Zunächst wird ein Kollektiv aus Patientinnen und Patienten der Klinischen Abteilung für Kinderradiologie am Universitätsklinikum Graz gewonnen. Für die in der Studie untersuchten Fälle existiert jeweils eine MRT-Untersuchung des Kniegelenks sowie ein oder mehrere korrelierende Röntgenbilder des Kniegelenks in lateraler Projektion. Es erfolgt eine manuelle Messung der Tiefe der Trochlea ossis femoris in den MRT-Schnittbildern und die Zusammenführung der erhobenen Werte mit den exportierten Röntgenaufnahmen. Anschließend erfolgt das Training von CNNs mit dem Ziel, die Trochleatiefe aus den Röntgenbildern vorherzusagen.

Die einzelnen methodischen Schritte werden im Folgenden näher ausgeführt.

4.1 Auswahl der Studienpopulation

Für diese Studie wurden Patientinnen und Patienten aus der Klinischen Abteilung für Kinderradiologie am Universitätsklinikum Graz ausgewählt, bei denen zwischen den Jahren 1995 und 2024 bildgebende Untersuchungen des Kniegelenks durchgeführt wurden. Eine aus den Untersuchungen resultierende Diagnose sowie potenziell vorliegende Risikofaktoren einer PFI spielen für die Datenerhebung keine Rolle. Die Selektion der Patienten- und Bilddaten erfolgte über die an der Abteilung verwendete PACS-Software *syngo.plaza*.

4.1.1 Einschlusskriterien

Folgende Einschlusskriterien kamen zur Anwendung:

- Vorliegen einer MRT-Untersuchung des Kniegelenks mit ausreichender Bildqualität und dem Vorhandensein axialer Schichten zur Messung der Trochleatiefe
- Verfügbarkeit mindestens einer lateralen Röntgenaufnahme des zur MRT-Untersuchung passenden Kniegelenks

4.1.2 Ausschlusskriterien

Patientinnen und Patienten wurden von der Studie ausgeschlossen, für die zumindest eines der unten angeführten Kriterien zutrifft:

- Unzureichende Qualität der lateralen Röntgenaufnahme: Bilder mit schlechter Kontrastierung oder einer starken Abweichung von der korrekten lateralen Projektion
- Vorliegen von postoperativen Veränderungen, die die Messung der Trochleattiefe verunmöglichen: Metallimplantate, Osteosynthesematerial oder andere postoperative Veränderungen, die die anatomische Darstellung der Trochlea ossis femoris in den bildgebenden Untersuchungen stark einschränken.

4.1.3 Größe der Studienpopulation

Nach Sichtung aller verfügbaren Patientendaten und Anwendung der Einschlusskriterien wurden 1000 Patientenfälle zur Messung der Trochleattiefe herangezogen. Für 94 Patientinnen und Patienten traf mindestens eines der Ausschlusskriterien zu. Das ergibt eine Studienpopulation von 906 Patientenfällen (davon 519 männlich und 387 weiblich). Das Alter der untersuchten Patientinnen und Patienten lag zwischen 1 und 24 Jahren (Median 14 Jahre, Mittelwert 13,7 Jahre, Standardabweichung 2,9 Jahre). Der überwiegende Teil der Studienpopulation befand sich somit im Schul- und Jugendalter.

Für einen Teil der erfassten Patientinnen und Patienten existieren mehrere verwendbare Röntgenaufnahmen. Dadurch stehen für das KI-Training insgesamt 1084 Röntgenaufnahmen mit den zugehörigen Trochleattiefen zur Verfügung.

4.2 Datenerhebung

Die für die Studie erforderlichen Daten wurden aus dem Bildarchiv der Klinischen Abteilung für Kinderradiologie des Universitätsklinikums Graz akquiriert, auf einem lokalen Rechner verarbeitet und ebendort in ein Excel-Arbeitsblatt übertragen.

4.2.1 MRT-Messung der Trochleattiefe

Die Erfassung der Tiefe der Trochlea ossis femoris erfolgt radiologisch in den axialen Schichten einer Magnetresonanztomographie des Kniegelenks. Dabei kann

die Messung sowohl an den knöchernen als auch an den knorpeligen Strukturen erfolgen. In der vorliegenden Arbeit liegt der Fokus ausschließlich auf der knöchernen Trochlea, ohne Berücksichtigung des knorpeligen Überzugs.

Bei der Messung der Trochlealtiefe können unterschiedliche Untersucher zu abweichenden Ergebnissen gelangen. In der Publikation von Tanaka et al. werden diese Abweichungen näher untersucht. Hierbei zeigt sich, dass die Konsistenz der einzelnen Messergebnisse der knöchernen Trochlealtiefe beim einzelnen Untersucher (Intrater-Reliability: 0,90) deutlich höher ist als die Übereinstimmung bei verschiedenen Untersuchern (Interrater-Reliability: 0,76) [7]. Um diese Abweichungen auf ein Minimum zu reduzieren und eine möglichst hohe Konsistenz in den erhobenen Daten herzustellen, wurden alle Messungen dieser Studie von einer einzelnen Person durchgeführt.

Die höchste Genauigkeit der Trochlealtiefen-Messung in der MR-Untersuchung wird erreicht, wenn man diese in einer axialen Schicht drei Zentimeter über dem femorotibialen Gelenkspalt durchführt [16]. Im PACS-Viewer werden zunächst die hinteren Femurkondylen mit einer Linie verbunden. Danach werden mit dem virtuellen Lineal drei Messstrecken gezogen, die an der zuvor gezogenen Linie im 90° Winkel beginnen. Zwei dieser Messstrecken enden dann am Vorderrand des medialen (m) und lateralen (l) Femurkondyls. Der Endpunkt der Strecke t markiert den tiefsten Punkt der Trochlea. Die Trochlealtiefe wird dann mit der Formel $TD = \frac{m+l}{2} - t$ errechnet und in der Einheit Millimeter angegeben.

4.2.2 Selektion des lateralen Röntgenbilds

Für jeden Patientenfall, bei dem eine Messung der Trochlealtiefe vorgenommen wurde, wurde mindestens ein Röntgenbild des entsprechenden Kniegelenks in lateraler Projektion ausgewählt. Hierfür wurde darauf geachtet, dass die Röntgenaufnahme in einem engen zeitlichen Zusammenhang zur zuvor beschriebenen MR-Untersuchung steht. Das dient zur Vermeidung des Auftretens potenzieller anatomischer Veränderungen, die physiologisch durch das Größenwachstum des Kindes oder iatrogen durch zwischenzeitlich durchgeführte operative Eingriffe an der Trochlea bedingt sein könnten. Wenn mehrere passende Röntgenaufnahmen vorlagen, wurde all jene Aufnahmen mit guter Bildqualität hinsichtlich Kontrastierung und Projektion ausgewählt, was dazu führt, dass in

einzelnen Fällen mehrere laterale Aufnahmen desselben Kniegelenks in die Studie einbezogen wurden.

Ungeeignete Röntgenaufnahmen wurden gemäß der in Punkt 4.1.2 formulierten Ausschlusskriterien nicht weiterverwendet.

4.3 Bildverarbeitung

Die selektierten Röntgenbilder wurden im DICOM-Format exportiert und an der Klinischen Abteilung für Kinderradiologie mittels Python-Skript in ein PNG-Format umgewandelt. Ebenso wurden aus dem DICOM-Header die Metadaten der Untersuchung in einer CSV-Datei ausgegeben. Diese beinhalten persönliche Informationen wie Alter und Geschlecht sowie relevante Parameter der Bilddatei wie beispielsweise der Pixelspacing-Parameter.

Eine Anonymisierung der Patientendaten erfolgte ebenfalls über das Python-Skript. Hierbei wurde aus der Patienten-ID und der Institution mit Hashfunktionen eine Zeichenkette generiert, die dann als Teil des Dateinamens der PNG-Bilder verwendet wurde. Dadurch sind keine Rückschlüsse auf die ursprünglichen Patientendaten möglich.

Die exportierten PNG-Dateien lagen initial in unterschiedlichen Farbtiefen zwischen 8 und 15 Bit vor. Um alle Bilder in ein einheitliches Format zu überführen, wurden diese in ein 8-Bit-Format konvertiert. Eine Farbtiefe von 8 Bit bedeutet, dass pro Pixel 256 Graustufen ($2^8 = 256$) gespeichert werden, während beispielsweise eine 12-Bit-Darstellung 4096 Graustufen ($2^{12} = 4096$) umfasst. Dementsprechend konnte auch eine Reduzierung der Dateigrößen erzielt werden.

Letztlich erfolgte eine manuelle Sichtung aller exportierten Röntgenbilder, um deren Eignung für das Machine-Learning zu überprüfen. Dabei wurden auch vorhandene Übersichtsaufnahmen der gesamten unteren Extremität auf den relevanten Kniebereich zugeschnitten. Im Rahmen dieser Sichtung konnte auch festgestellt werden, dass keine relevanten Qualitätsverluste durch die Konvertierung aufgetreten sind.

4.4 Datenaufbereitung

Die konvertierten und geprüften Röntgenbilder wurden in einem Ordner gesammelt gespeichert. Zur Durchführung des KI-Trainings wurde eine Datenbank in Form eines Excel-Arbeitsblattes erstellt. Wie in Abbildung 7 dargestellt, werden für jeden Fall sieben Merkmale beschrieben.

Internal ID	Filename	Gender	Age	Pixelspacing	TD	GROUP
G0001	4674741c7a1c7f1896623eb9576de058a01300065100e2e18d46e89f7bc F	F	016-7	'0.144'		5,7 D
G0002	56720d7cc14dafcc3f0aa32f526e92ad980b66eb2e0ab2fdbd7076464647 M	M	015-3	'0.15'		3,6 B
G0004	3602e16027341d1322e68419ec5974ab50fa7f6e09349e9acfc24ca8381f M	M	015-4	'0.15'		3,4 B
G0004	3602e16027341d1322e68419ec5974ab50fa7f6e09349e9acfc24ca8381f M	M	015-4	'0.143'		3,4 B
G0004	3602e16027341d1322e68419ec5974ab50fa7f6e09349e9acfc24ca8381f M	M	015-8	'0.1000'		3,4 B
G0005	5a4e3e222bdbe9442d290fb8d6a5a0f54287a4b7ea7d68e2e6cb8de6ad9 M	M	015-3	'0.144'		2,6 A
G0006	a2faf225c6ef9fc2261a9b0c3f560a22769f4f65ac5f8d174a6c1f8a84d926 M	M	014-7	'0.1'		6,1 D
G0006	a2faf225c6ef9fc2261a9b0c3f560a22769f4f65ac5f8d174a6c1f8a84d926 M	M	015-2	'0.1'		6,1 D
G0007	1b1065afed8e84708b71d37c9f6b81725e64a3d8b70332c9a6ffd4b6533c M	M	019-4	'0.144'		5,9 D
G0008	8e735e2212cdcf92af2fb37819528cc208159edcfc3bd4befbfbf8398dd3 F	F	014-4	'0.144'		4,4 C
G0009	9917240c946594f8c113b4a4cf679f3623228fd7bddcb715e10af4a37554 F	F	016-2	'0.144'		6,6 D
G0011	116a24e5035de85576b5ee7f56e3ac54c3286774afb1880836c616331cb M	M	013-5	'0.144'		4,3 C
G0012	c95c545d3278bde754a464f0c9225a9e63a46bc0f7948e56e814046f765f M	M	013-5	'0.1000'		4,1 C
G0012	c95c545d3278bde754a464f0c9225a9e63a46bc0f7948e56e814046f765f M	M	013-8	'0.143'		4,1 C
G0013	bbb7dba9aa4278235e1e1ac764f87ea1a8af77007290b65fdd8b9f650b9d F	F	016-9	'0.144'		4,2 C
G0013	bbb7dba9aa4278235e1e1ac764f87ea1a8af77007290b65fdd8b9f650b9d F	F	017-8	'0.144'		4,2 C
G0014	d9de734410b0ae0866c46fe86cc1c9fd18d8e06a248212f7568ce0f3c4db M	M	017-0	'0.144'		3,1 A
G0015	73197a96350235012d074678774cc33a7b83d46e4127d751feec108a74f F	F	013-8	'0.144'		4,5 C
G0016	f03a02a7fee2102568ac8d0704799cfb49a1ca6e04c80777809fb28df3fd1 M	M	017-1	'0.144'		5,6 D
G0016	f03a02a7fee2102568ac8d0704799cfb49a1ca6e04c80777809fb28df3fd1 M	M	017-9	'0.144'		5,6 D
G0017	1cdacc6fd4025279b1d86058b5c9575b193058719ec7feff4e3f614bfe42c F	F	015-4	'0.144'		5,3 D
G0018	cfb9a5a432d886d96b356b3600b37c0e180864e22a4665a8d706ba0bd7 M	M	014-4	'0.10'		2,8 A
G0018	cfb9a5a432d886d96b356b3600b37c0e180864e22a4665a8d706ba0bd7 M	M	014-9	'0.10'		2,8 A
G0018	cfb9a5a432d886d96b356b3600b37c0e180864e22a4665a8d706ba0bd7 M	M	015-5	'0.10'		2,8 A
G0018	cfb9a5a432d886d96b356b3600b37c0e180864e22a4665a8d706ba0bd7 M	M	016-8	'0.144'		2,8 A

Abbildung 7: Ausschnitt aus dem Excel-Datenblatt mit den erhobenen Daten

- Internal ID:** Eine fortlaufende Identifikationsnummer, die nur innerhalb der Studie verwendet wird und keine Rückschlüsse auf die Patientenidentität zulässt.
- Filename:** Der eindeutige Dateiname des Röntgenbilds im PNG-Format. Der Name beinhaltet kryptographisch anonymisierte Patientendaten, wie in Kapitel 4.3 beschrieben.
- Gender:** Das Geschlecht der untersuchten Person mit den Merkmalsausprägungen M (männlich) sowie W (weiblich).
- Age:** Das Alter der Patientin bzw. des Patienten zum Zeitpunkt der Röntgenuntersuchung. Jahr und Monat werden durch einen Bindestrich getrennt (z. B. 016-7 steht für 16 Jahre und 7 Monate).
- Pixelspacing:** Bezeichnet den realen Abstand zwischen benachbarten Pixeln in einer Bildaufnahme. Der Wert wird in Millimetern angegeben.
- TD:** Die gemessene Trochlealtiefe, angegeben in der Einheit Millimeter.

- g. **GROUP**: Die gemessenen Trochlealtiefen in gleich große Gruppen eingeteilt. Die Gruppen sind mit Buchstaben gekennzeichnet (z. B. A, B, C, D).

4.5 Maschinenlernen mit EfficientNet-Modellen

Für die KI-basierte Bildklassifikation der vorliegenden lateralen Röntgenaufnahmen wurden verschiedene CNN-Modelle eingesetzt, die auf der EfficientNet-Architektur (B0 bis B7) basieren. Die technischen Parameter der Trainingsumgebung sind folgender Tabelle zu entnehmen:

Systemkomponente	Spezifikation
Betriebssystem	Linux Ubuntu 22.04 LTS
Prozessor	Intel Core i7-13700K
Arbeitsspeicher	64 GB RAM
Grafikkarten	2x NVIDIA RTX 4090, jeweils 24 GB Videospeicher
Programmierschnittstelle	CUDA-Framework Version 12.4
Python	Version 3.10
Bildgrößen	für alle Modellvarianten (EfficientNet B0 bis B7) die Standardbildgröße (224x224px bis 600x600px)
Deep-Learning-Framework	FastAI 2.7.15

Tabelle 1: Technische Parameter der Trainingsumgebung

Zur Bewertung der Modellleistung wurde eine 16-fache Kreuzvalidierung (k-fold = 16) gewählt. Bei dieser Methode wird der Datensatz in 16 gleich große Teile unterteilt, wobei in jedem Durchlauf ein Teil als Testmenge und die anderen 15 Teile als Trainingsmenge verwendet werden. Jedes Bild des Trainingsdatensatzes wird also sowohl zur Modelloptimierung als auch zur Testung herangezogen.

Während des Trainings wurden die folgenden Augmentierungsparameter in der Funktion `aug_transforms` verwendet:

```

385 aug_transforms(
386     mult=1, do_flip=True, flip_vert=True, max_rotate=20,
387     min_zoom=0.8, max_zoom=1.2, max_lighting=0.8, max_warp=0.2,
388     p_affine=0.8, p_lighting=1, size=None, mode='bilinear',
389     pad_mode='reflection', align_corners=False, batch=False
390 )

```

Abbildung 8: aug_transforms-Funktion zur Datenaugmentierung mit Parametern

Die Zielvariable für jedes Bild war die manuell erhobene Trochleatiefe, die zunächst in vier gleich große Gruppen (A-D) eingeteilt wurde. In einem zweiten Klassifikationsversuch erfolgte eine Zusammenfassung in zwei gleich große Gruppen (A-B). Die Zuordnung der PNG-Dateien zu den jeweiligen Gruppen erfolgte automatisiert auf Basis des im Kapitel 4.4 beschriebenen Excel-Datensatzes. Die daraus resultierende Zweierklassifikation wurde im anschließenden KI-Training endgültig verwendet, wobei sich der Cutoff-Wert von 4 mm zur Trennung der beiden Klassen ergab.

Nach Training und Test der EfficientNet-Modelle wurden verschiedene Modellleistungsparameter exportiert und werden im folgenden Kapitel analysiert.

4.6 Ethikvotum

Für die Datenauswertung für diese Arbeit wurde um ein Votum der Ethikkommission der Medizinischen Universität Graz (IRB00002556) angesucht. Die Referenznummer des positiven Votums lautet 34-119 ex 21/22.

5 Resultate

Dieses Kapitel dient dazu, die Ergebnisse des Trainings und der Testung der eingesetzten KI-Modelle darzustellen. Es werden die Leistungen der verschiedenen EfficientNet-Architekturen anhand statistischer Kennzahlen sowie grafischer Darstellungen beschrieben und analysiert.

Die getestete Klassifikationsaufgabe bestand in der dichotomen Einteilung der Trochleatiefen auf Basis der lateralen Röntgenbilder in zwei gleich große Gruppen (A und B). Als Grenzwert für diese Einteilung wurde eine Trochleatiefe von 4 mm gewählt. Wenn im Folgenden von positiven/kranken Fällen gesprochen wird, so werden Trochleatiefen von über 4 mm (Gruppe B) gemeint.

5.1 Konfusionsmatrizen

Konfusionsmatrizen eignen sich gut, um Deep-Learning-Modelle zu analysieren, die dichotome Klassifikationsaufgaben durchführen. Für jedes Modell wird eine Tabelle ausgegeben, die vier statistische Kennzahlen beinhaltet:

- True Positive (TP): Die positive Klasse wurde korrekt vorhergesagt.
- True Negative (TN): Die negative Klasse wurde korrekt vorhergesagt.
- False Positive (FP): Die negative Klasse wurde fälschlicherweise als positiv vorhergesagt.
- False Negative (FN): Die positive Klasse wurde fälschlicherweise als negativ vorhergesagt.

Im vorliegenden Fall steht die negative Klasse (True 0) für die Gruppe A (Trochleatiefe höchstens 4 mm) und die positive Klasse (True 1) für die Gruppe B (Trochleatiefe mehr als 4 mm). Die Matrizen zeigen, wie die einzelnen EfficientNet-Modelle die Fälle beider Gruppen klassifiziert haben. Die angegebenen Werte sind als relative Häufigkeiten zu verstehen und zeigen zeilenweise, wie häufig das Modell die jeweilige Klasse korrekt oder falsch vorhergesagt hat.

In der folgenden Abbildung gelangen die Konfusionsmatrizen für die EfficientNet-Modelle B0 bis B7 zur Darstellung:

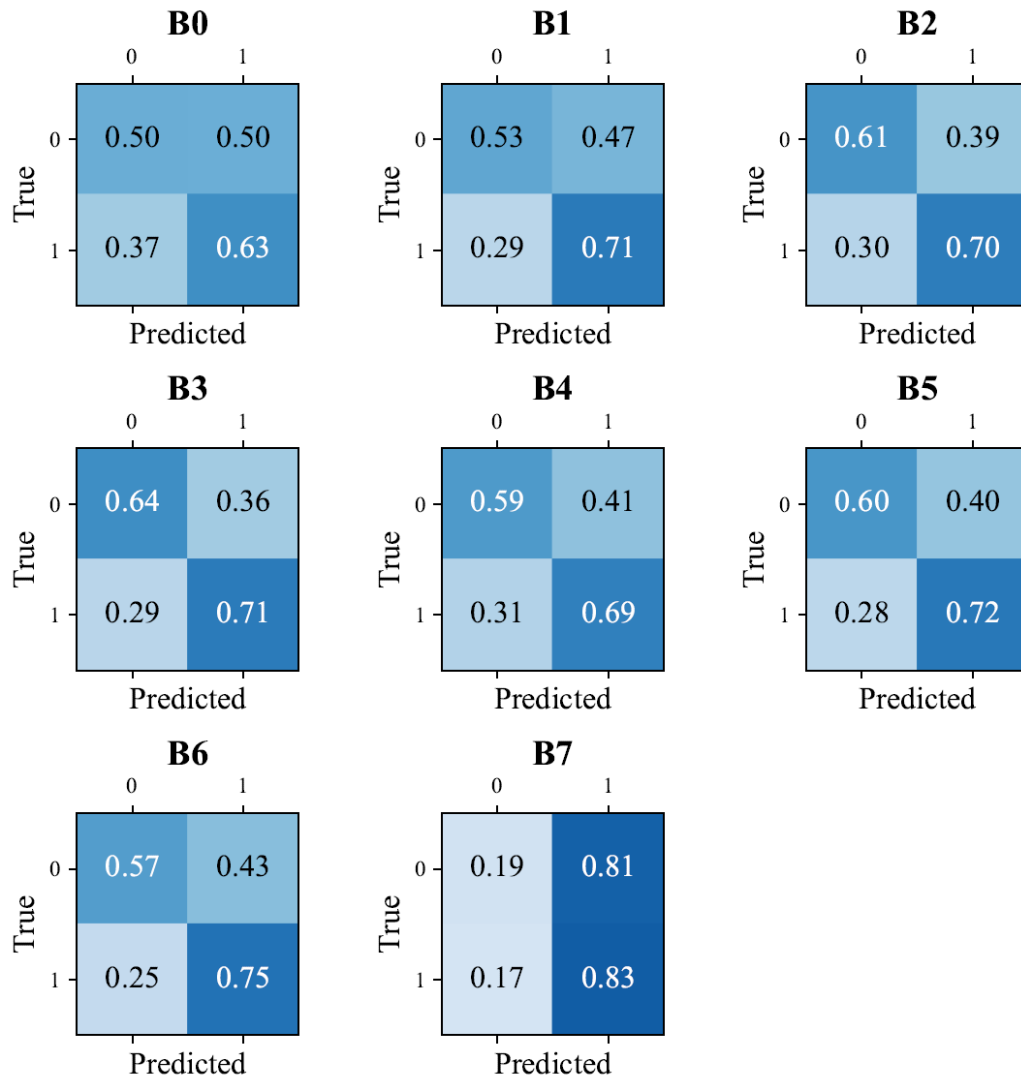


Abbildung 9: Konfusionsmatrizen der getesteten EfficientNet-Modelle B0 bis B7

In der medizinischen Diagnostik sind die Kennzahlen der Spezifität und Sensitivität von großer Bedeutung. Diese lassen sich aus den Konfusionsmatrizen ohne weitere Berechnung einfach ablesen, da es sich bei den angegebenen Werten bereits um relative Häufigkeiten handelt. Die Zahl links oben (True Negative Rate) entspricht der Spezifität und gibt daher an, wie zuverlässig das jeweilige Modell negative Fälle korrekt als negativ erkennt. Dahingegen spiegelt die Zahl rechts unten (True Positive Rate) die Sensitivität wider, was den Anteil der real positiven Fälle, die vom Modell als positiv klassifiziert wurden, entspricht.

Durch die farbliche Gestaltung der Konfusionsmatrizen ist am ersten Blick erkennbar, dass die Ergebnisse im Modell B7 von den anderen Modellen abweichen. EfficientNet B7 weist eine besonders hohe Sensitivität von 83% auf, was bedeutet, dass die positiven Fälle verlässlich erkannt werden. Auffällig ist

jedoch die im Vergleich sehr niedrige Spezifität von 19%. Das zeigt, dass viele negative Fälle fälschlicherweise als positiv klassifiziert werden. Außerdem wird hierdurch verständlich, dass eine hohe Sensitivität alleine nicht ausreicht und es von großer Bedeutung ist, Sensitivität und Spezifität immer gemeinsam zu betrachten, um die Aussagekraft eines diagnostischen Tests zu beurteilen. Um beide Kennzahlen zu einem Gesamtmaß zusammenzufassen und so die Güte eines Modells fair zu bewerten, bieten sich errechnete Kennzahlen wie der Youden-Index an. Diese Größe wird mit der Formel $\text{Sensitivität} + \text{Spezifität} - 1$ berechnet und kann maximal den Wert 1 ergeben. Ein Ergebnis von 1 entspricht einem perfekten Test ohne Fehlklassifikationen. Ein Wert von 0 bedeutet hingegen, dass der Test nicht besser als zufälliges Raten ist. [17]

EfficientNet-Modell	Sensitivität	Spezifität	Youden-Index
B0	0,63	0,50	0,13
B1	0,71	0,53	0,24
B2	0,70	0,61	0,31
B3	0,71	0,64	0,35
B4	0,69	0,59	0,28
B5	0,72	0,60	0,32
B6	0,75	0,57	0,32
B7	0,83	0,19	0,02

Tabelle 2: Sensitivität, Spezifität und errechneter Youden-Index der Klassifikationsaufgabe in den EfficientNet-Modellen B0 bis B7

Die Tabelle 2 führt die Sensitivität, Spezifität und den daraus berechneten Youden-Index der getesteten EfficientNet-Modelle B0 bis B7 auf. Hier gelangt zur Darstellung, dass das Modell B3 die beste kombinierte Leistung aus Sensitivität und Spezifität erbringt und damit den höchsten Youden-Index aufweist. Durch die hohe Diskrepanz von Sensitivität und Spezifität erreicht das Modell B7 einen Youden-Index von annähernd 0, was die diagnostische Unbrauchbarkeit des Modells statistisch unterstreicht.

5.2 ROC-Analyse

ROC-Analysen (Receiver Operating Characteristic) sind eine weitere Möglichkeit, um Klassifikationsmodelle zu bewerten. Die visuelle Darstellung zeigt eine Kurve, welche den Zusammenhang zwischen der True-Positive-Rate (Sensitivität) und False-Positive-Rate (1 - Spezifität) über verschiedene Schwellenwerte anzeigt. Eine ROC-Kurve beginnt immer im Punkt (0,0), was bedeutet, dass bei einem sehr hohen

Schwellenwert weder positive Fälle erkannt werden noch Fehllarme auftreten. Sie endet im Punkt (1,1), was den Fall beschreibt, dass bei einem sehr niedrigen Schwellenwert alle positiven Fälle erkannt werden, aber auch die negativen Fälle fälschlicherweise als positiv klassifiziert werden. Jeder weitere Punkt auf der ROC-Kurve entspricht einem theoretisch möglichen Schwellenwert, womit ermöglicht wird, einen optimalen Schwellenwert zu finden, der Sensitivität und Spezifität in ein brauchbares Verhältnis bringt. Die Fläche unter der ROC-Kurve wird als AUC (Area Under The Curve) bezeichnet, mit der man die diagnostische Güte eines Tests beschreiben kann. Eine AUC von 1 entspricht einem perfekten Model, das die beiden zu klassifizierenden Gruppen ohne Fehler trennt, während eine AUC von 0,5 zufälligem Raten durch das Modell entspricht. [17]

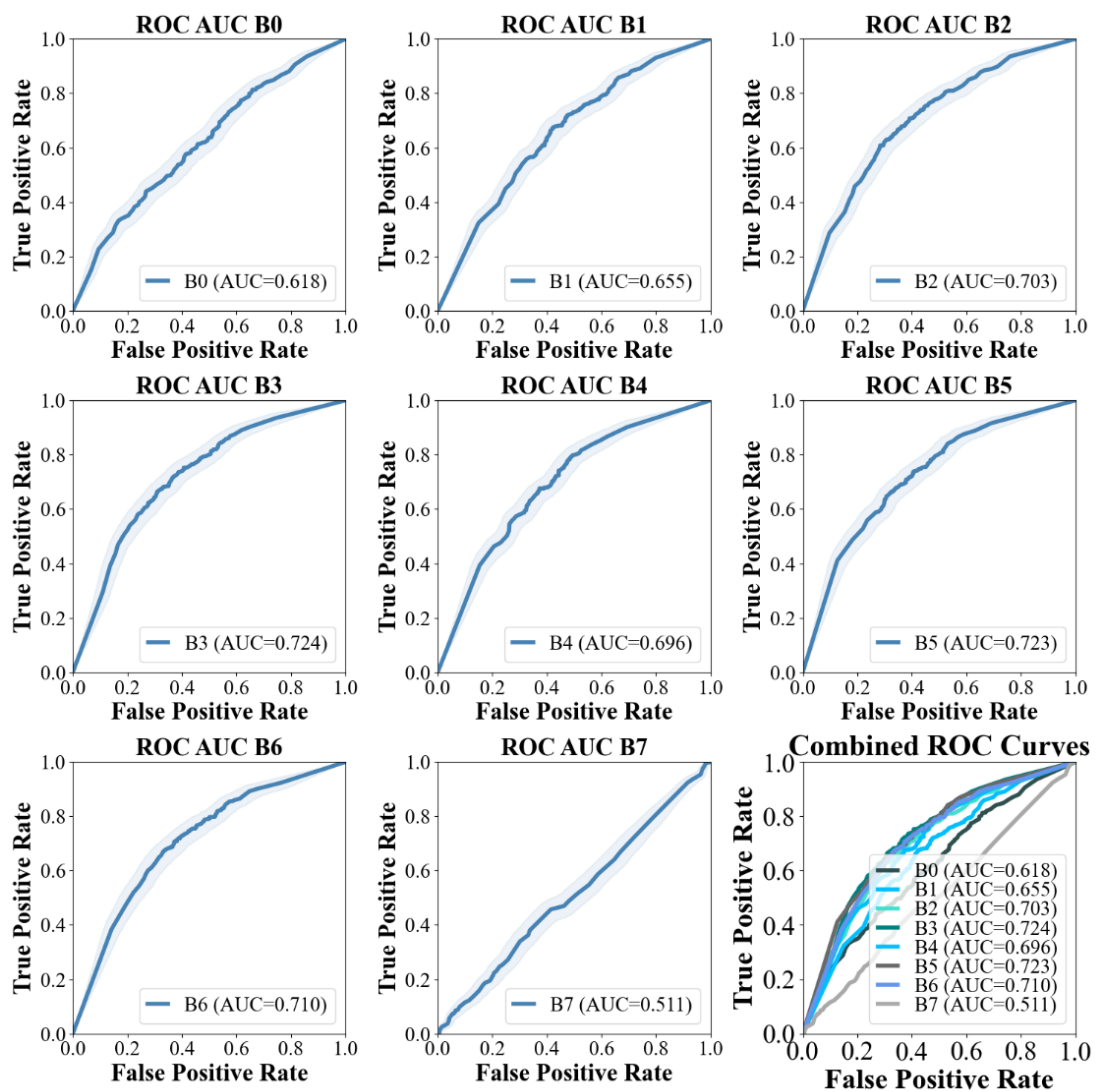


Abbildung 10: ROC-Kurven und AUC-Werte der EfficientNet-Modelle B0 bis B7

Die ROC-Kurven der verschiedenen EfficientNet-Modelle für die Trochleatiefen-Klassifikationsaufgabe werden in Abbildung 10 dargestellt. Zur Darstellung gelangen acht Kurven für die Modelle B0 bis B7 sowie ein kombinierter Graph. Wie bereits bei den Konfusionsmatrizen zeigt auch das Modell B7 in der ROC-Analyse ein auffälliges Muster. Die ROC-Kurve von B7 verläuft annähernd linear, was sich in einem AUC-Wert von 0,511 widerspiegelt und darauf hinweist, dass dieses Modell die beiden Gruppen nicht besser als der Zufall trennt. Grundsätzlich lässt sich sagen, dass ein Modell eine bessere Performance anzeigt, je stärker sich die ROC-Kurve in den linken oberen Quadranten erstreckt. Diese Krümmung ist bei den Modellen B0 bis B6 erkennbar, was darauf hinweist, dass die KI-Modelle in der Lage sind, Muster zu erkennen und eine Unterscheidung zwischen den beiden Gruppen zu treffen. In der ROC-Analyse ist erkennbar, dass die beste Performance bei den EfficientNet-Modellen B3 (AUC=0,724) und B5 (AUC=0,723) zu finden ist.

Eine übersichtliche Darstellung der AUC-Werte inklusive ihrer 95%-Konfidenzintervalle sind der Tabelle 3 zu entnehmen.

EfficientNet-Modell	ROC AUC	95% Konfidenzintervall
B0	0,618	0,584–0,651
B1	0,655	0,624–0,685
B2	0,703	0,670–0,732
B3	0,724	0,694–0,755
B4	0,696	0,664–0,726
B5	0,723	0,692–0,753
B6	0,710	0,677–0,740
B7	0,511	0,476–0,548

Tabelle 3: AUC-Werte inklusive ihrer 95%-Konfidenzintervalle in der ROC-Analyse der EfficientNet-Modelle B0-B7

5.3 Precision-Recall-Analyse

Zur Bewertung eines Klassifikationsmodell können auch die statistischen Kennzahlen Precision und Recall herangezogen werden. Die Precision beschreibt die Wahrscheinlichkeit, dass bei positiver Klassifikation wirklich ein positiver Fall getroffen wurde. Diese Kennzahl kann auch als positiv prädiktiver Wert (PPV) bezeichnet werden. Der Recall-Wert (Synonym: Sensitivität) gibt den Anteil der

korrekt positiv klassifizierten Ergebnisse bezogen auf die Gesamtheit der tatsächlich positiven Ergebnisse an.

Im Rahmen der Trochleatiefen-Klassifikation würde eine hohe Precision bedeuten, dass unter den positiven Vorhersagen durch das Modell nur wenige falsch-positive Einschätzungen vorkommen. Ein hoher Recall würde bedeuten, dass möglichst alle tatsächlich positiven Fälle korrekt erkannt werden und nur wenige Positive übersehen werden.

Mit Precision-Recall-Kurven (PR-Kurven) werden diese statistischen Kenngrößen in Beziehung gebracht, wobei die Darstellung wie bei den ROC-Kurven über verschiedene Schwellenwerte erfolgt. Ebenfalls analog zu den ROC-Kurven spielt die AUC eine wichtige Rolle und beschreibt die Gesamtgüte des Modells. Im theoretischen Falle eines idealen Klassifikationsmodells bliebe die Precision bei zunehmendem Recall gleich hoch.

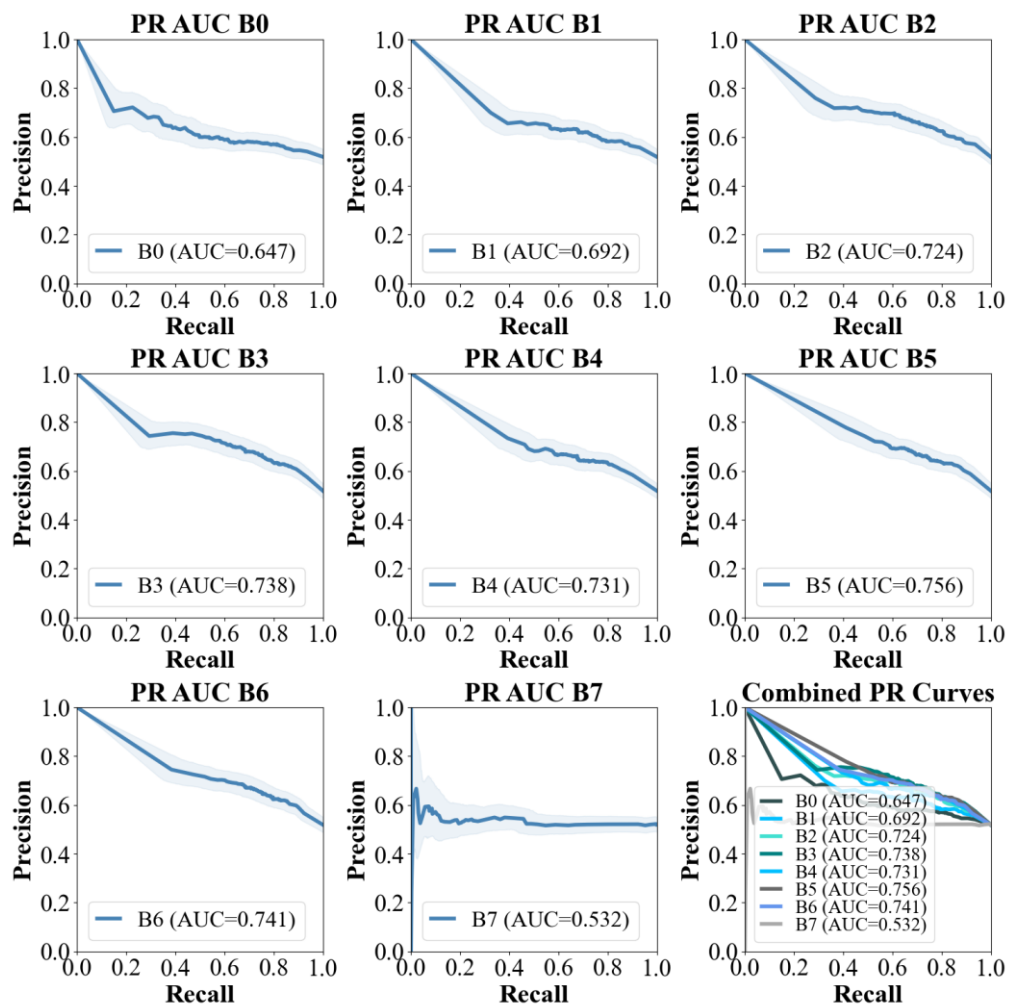


Abbildung 11: Precision-Recall-Kurven der EfficientNet-Modelle B0 bis B7

Die Precision-Recall-Kurven der Testung der EfficientNet-Modelle für die Trochleatiefen-Klassifikation sind in Abbildung 11 dargestellt. Die Modelle B0 bis B6 zeigen eine sinkende Precision bei steigendem Recall, was in Praxisbeispielen üblich ist, da bei mehr erkannten positiven Fällen realistischerweise auch falsch-positive Fälle klassifiziert werden. Das Modell B7 zeigt über alle Schwellenwerte hinweg eine annähernd gleichbleibende Precision von 50% und kommt damit, wie in den anderen statistischen Beschreibungen bewiesen, zufälligem Raten gleich (AUC = 0,532). Wird die AUC der einzelnen Modelle zur Beurteilung herangezogen, so lassen sich die EfficientNet-Modelle B5 (AUC = 0,756) und B6 (AUC = 0,741) in der Precision-Recall-Analyse als die leistungsstärksten bezeichnen.

Analog zur ROC-Analyse wurden auch die 95%-Konfidenzintervalle der AUC-Werte der Precision-Recall-Kurven bestimmt. Diese sind der Tabelle 4 zu entnehmen.

EfficientNet-Modell	PR AUC	95% Konfidenzintervall
B0	0,647	0,607–0,686
B1	0,692	0,658–0,729
B2	0,724	0,687–0,760
B3	0,738	0,702–0,774
B4	0,731	0,696–0,766
B5	0,756	0,723–0,788
B6	0,741	0,708–0,774
B7	0,532	0,492–0,579

Tabelle 4: AUC-Werte inklusive ihrer 95%-Konfidenzintervalle in der PR-Analyse der EfficientNet-Modelle B0-B7

5.4 Leistungskennzahlen

Die Kennzahlen Precision und Recall sind zentrale Elemente zur Leistungsbeurteilung eines Klassifikatormodells wie im vorliegenden Fall. Die Definitionen dieser statistischen Kennzahlen sind dem Kapitel 5.3 Precision-Recall-Analyse zu entnehmen. Um Precision und Recall mathematisch zusammenzufassen, eignet sich der sogenannte F1-Score, der dem harmonischen Mittelwert der beiden Kennzahlen entspricht. Berechnet wird der F1-Score mit der Formel $F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ und kann maximal den Wert 1 erreichen.

Während die in Kapitel 5.3 dargestellten Kurven die Entwicklung von Precision und Recall über verschiedene Schwellenwerte hinweg veranschaulichen, wurden die im Folgenden gezeigten Leistungskennzahlen für jedes Modell auf Basis eines festen Schwellenwertes berechnet. In der Literatur gilt der Schwellenwert 0,5 als Standard.

Im Rahmen dieser Arbeit wurden die Leistungskennzahlen Precision, Recall und F1-Score automatisch sowohl mit dem macro- als auch mit dem weighted-Durchschnitt ermittelt. Beim macro-Verfahren werden die Kennzahlen für alle Klassen gebildet und ungewichtet gemittelt, was dazu führt, dass jede Klasse unabhängig ihrer Größe gleich stark berücksichtigt wird. Mit dem weighted-Durchschnitt hingegen wird die Größe der jeweiligen Klassen in der Rechnung berücksichtigt und größere Klassen für die Kennzahlberechnung entsprechend höher gewichtet. [18]

Die beiden nachfolgenden Tabellen zeigen eine Übersicht der errechneten Leistungskennzahlen mit der Methode der weighted- (Tabelle 5) und macro-Mittelung (Tabelle 6).

EfficientNet-Modell	Precision (Weighted)	Recall (Weighted)	F1-Score (Weighted)
efficientnet-b0	0,569	0,570	0,568
efficientnet-b1	0,626	0,625	0,622
efficientnet-b2	0,656	0,657	0,656
efficientnet-b3	0,674	0,674	0,674
efficientnet-b4	0,640	0,640	0,639
efficientnet-b5	0,663	0,663	0,662
efficientnet-b6	0,664	0,662	0,659
efficientnet-b7	0,515	0,521	0,466

Tabelle 5: Precision und Recall der EfficientNet-Modelle mit Weighted-Mittelung

EfficientNet-Modell	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
efficientnet-b0	0,568	0,567	0,566
efficientnet-b1	0,626	0,622	0,620
efficientnet-b2	0,656	0,655	0,655
efficientnet-b3	0,674	0,673	0,673
efficientnet-b4	0,640	0,638	0,638
efficientnet-b5	0,663	0,661	0,661
efficientnet-b6	0,664	0,659	0,658
efficientnet-b7	0,515	0,509	0,459

Tabelle 6: Precision und Recall der EfficientNet-Modelle mit Macro-Mittelung

Da die beiden zu klassifizierenden Gruppen annähernd die gleiche Größe aufweisen, zeigen sich die errechneten Zahlen in der weighted- und macro-Mittelung nahezu identisch. Ein größerer Unterschied zwischen diesen Kennzahlen wäre dann zu erwarten, wenn relevante Unterschiede in der Klassengröße (beispielsweise 90:10) bestehen würden.

Zieht man den aus Precision und Recall ermittelten F1-Score zur Bewertung der Modellgüte heran, so erzielt EfficientNet-B3 die besten Ergebnisse in der Klassifikation der Trochlealtiefen (weighted F1 = 0,677; macro F1 = 0,673). Dicht dahinter reihen sich das B5-Modell (weighted F1 = 0,662; macro F1 = 0,661) sowie B6 (weighted F1 = 0,659; macro F1 = 0,658) ein. Während sich die Modelle B2, B4 und B1 in absteigender Reihenfolge im Mittelfeld befinden, erzielen EfficientNet-B0 (weighted F1 = 0,568; macro F1 = 0,566) sowie EfficientNet-B7 (weighted F1 = 0,466; macro F1 = 0,459) die schlechtesten Ergebnisse.

Als weitere geläufige Leistungskennzahl ist die Genauigkeit (engl. Accuracy) zu nennen. Diese gibt an, welcher Anteil aller Fälle vom Modell korrekt klassifiziert wurde. Hierbei spielt es keine Rolle, ob es sich um negative oder positive Fälle handelt. Berechnet wird die Accuracy mit der Formel $Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$ und kann Werte zwischen 0 und 1 annehmen.

EfficientNet-Modell	Accuracy
efficientnet-b0	0,570
efficientnet-b1	0,625
efficientnet-b2	0,657
efficientnet-b3	0,674
efficientnet-b4	0,640
efficientnet-b5	0,663
efficientnet-b6	0,662
efficientnet-b7	0,521

Tabelle 7: Accuracy-Werte der EfficientNet-Modelle B0-B7

Die Reihung der Modelle anhand der Genauigkeit deckt sich mit jener auf Grundlage des F1-Scores. Hierbei zeigt sich, dass das leistungsstärkste Modell B3 67,4 % aller Fälle korrekt klassifiziert. Dicht dahinter folgen die Modelle B5, B6 und B2 mit Genauigkeitswerten zwischen 66,3 % und 65,7 %. Am schwächsten schnitten EfficientNet-B0 mit einer Genauigkeit von 57,0 % sowie EfficientNet-B7 mit nur 52,1 % ab.

6 Diskussion

Ziel dieser Studie war es, zu untersuchen, ob KI-Modelle in Form von Convolutional Neural Networks (CNNs) in der Lage sind, anhand lateraler Röntgenaufnahmen des Kniegelenks eine zuverlässige Vorhersage der Trochleatiefe zu treffen. Hierzu wurden acht Varianten der EfficientNet-Architektur (B0 bis B7) trainiert und getestet. Die Aufgabe der Modelle bestand darin, die vorliegenden Bilddateien in eine von zwei Gruppen zu kategorisieren. Gruppe A umfasste hierbei alle Fälle mit einer Trochleatiefe von höchstens 4 mm, während Gruppe B alle Fälle mit einer Trochleatiefe von mehr als 4 mm beinhaltete. In Tanaka et al. werden Trochleatiefen von weniger als vier Millimeter als pathologisch flach beschrieben, weshalb die in dieser Studie durchgeführte KI-Analyse auch als diagnostischer Test gedeutet werden kann und mit entsprechenden Kennzahlen untersuchbar ist.

Mehrere Modelle der EfficientNet-Reihe zeigten in der Analyse eine durchaus akzeptable Leistung. Die besten Klassifikationsergebnisse erzielte hierbei das Modell B3 (Accuracy: 67,4 %, F1-Score: 0,674, Youden-Index: 0,35, ROC-AUC: 0,724, PR-AUC: 0,738). Dicht dahinter folgten die Modelle B5 und B6 und erreichten ebenfalls akzeptable Leistungen in allen Kennzahlen. Die Modelle B0, B1, B2 und B4 lagen im mittleren Bereich. Von allen acht getesteten EfficientNet-Architekturen war das Modell B7 das einzige, welches keine sinnvolle Differenzierung zwischen beiden Gruppen durchführen konnte. Mit einem ROC-AUC von 0,511, einem PR-AUC von 0,532 und einem Youden-Index von 0,02 zeigt Modell B7 kein mustererkennendes Verhalten, sondern liefert Ergebnisse, die eher einem zufälligen Raten entsprechen.

6.1 Klinische Anwendbarkeit

Die vorliegenden Ergebnisse bestätigen die Hypothese, dass eine automatisierte Vorhersage der Trochleatiefe auf Basis lateraler Röntgenbilder technisch realisierbar ist und CNN-Modelle ein gewisses Lernverhalten zeigen. Vorallem das EfficientNet-Modell B3 konnte eine erkennbare Differenzierungsfähigkeit zwischen normalen und flachen Trochleae entwickeln, erreichte mit einer Genauigkeit von 67,4 % jedoch noch keine Leistung, die für eine klinisch zuverlässige Anwendung ausreichen würde. Zwar gibt es in der Literatur bis dato keinen allgemein gültigen Schwellenwert, ab welcher Genauigkeit ein KI-Modell klinisch akzeptiert werden

kann, jedoch ist es naheliegend, dass es bei einer geringen Accuracy in der Praxis zu einer durchaus relevanten Anzahl an falschen-negativen sowie falsch-positiven Vorhersagen führen.

Bei einer falsch-negativen Befundung durch die KI würden zu flache Trochleae übersehen werden, was das Risiko für spätere Luxationen erhöht und konsekutiv eine einzuleitende Therapie verzögert. Auch eine hohe Rate an falsch-positiven Befunden wäre in diesem Zusammenhang nicht wünschenswert, da Patienten mit physiologischer Trochleamorphologie unnötiger diagnostischer Abklärung ausgesetzt würden. Neben zusätzlicher psychischer Belastung für Betroffene würde dies auch zu personellen und finanziellen Belastungen im Gesundheitssystem führen.

Klinisch betrachtet hätte die Anwendung eines solchen KI-gestützten Modells vor allem dann Sinn, wenn man vor einer automatisierten Diagnostik eine entsprechend hohe Vortestwahrscheinlichkeit gewährleistet. Dieses Prinzip ist in der klinischen Praxis durchaus üblich und findet sich beispielsweise bei der Abklärung einer Pulmonalarterienembolie. Hier wird zunächst anhand des sogenannten Wells-Scores eine klinische Wahrscheinlichkeit für das Vorliegen einer PAE ermittelt, um dann bei niedriger Wahrscheinlichkeit zunächst einen D-Dimer-Test und bei hoher Wahrscheinlichkeit direkt eine CT-Angiographie der Pulmonalarterien durchzuführen. Ähnlich verhält es sich bei der Interpretation erhöhter PSA-Werte, die ohne klinischen Verdacht eines Prostatakarzinoms nur eine geringe Aussagekraft besitzen.

6.2 Vergleich mit ähnlicher Literatur

Es existieren mit heutigem Stand zahlreiche wissenschaftliche Publikationen zum Einsatz von KI-Algorithmen in der radiologischen Diagnostik, insbesondere im Bereich der muskuloskelettalen Bildgebung. Durch Literaturrecherche in PubMed konnte keine Arbeit ermittelt werden, die KI zur Beurteilung der Trochleattiefe anhand lateraler Kniegelenksröntgen untersucht. Um die in dieser Diplomarbeit erzielten Forschungsergebnisse in Beziehung zu anderen Arbeiten zu setzen, erfolgt dies anhand einer ausgewählten Arbeit, die sich mit einer ähnlichen Fragestellung befasst.

Die Studie *“Deep learning-assisted diagnosis of femoral trochlear dysplasia based on magnetic resonance imaging measurements”* von Xu et al. aus dem Jahr 2023 untersuchte den Einsatz eines CNNs zur Diagnose der Trochleadysplasie anhand von MRT-Bildern des Kniegelenks. In dieser Studie wurden 464 MRT-Datensätze (202 mit Trochleadysplasie, 252 mit normaler Trochlea) ausgewertet. Das eingesetzte CNN übernahm die Aufgabe, anatomische Schlüsselpunkte in axialen Schichten zu erkennen und daraus die relevanten morphologischen Parameter der Trochlea, darunter auch die Trochleattiefe, zu ermitteln. Auf Basis der Berechnungen erfolgte dann eine Klassifikation der Aufnahmen in normal oder dysplastisch. Die Leistungsfähigkeit des CNN wurde anschließend mit den Ergebnissen von Chirurginnen und Chirurgen sowie Radiologinnen und Radiologen mit unterschiedlichen Erfahrungsstufen verglichen. Es zeigte sich, dass das CNN in seiner diagnostischen Genauigkeit den weniger erfahrenen Ärztinnen und Ärzten deutlich überlegen war und eine Leistung auf dem Niveau der erfahrenen Fachärzte erzielte. Mit 0,14 Sekunden pro Fall zeigte das CNN auch eine deutlich schneller Diagnosezeit als die klinisch tätigen Ärztinnen und Ärzte (86 bis 103 Sekunden). Insgesamt kommt die Studie zum Schluss, dass der Einsatz von Deep Learning eine zuverlässige Diagnostik der Trochleadysplasie auf Knie-MRTs ermöglicht, welche vor allem weniger erfahrenen Klinikerinnen und Klinikern als unterstützendes Werkzeug dienen kann. [19]

Xu et al. konnten mit ihrem auf U-Net basierendem CNN bessere Ergebnisse erzielen (Accuracy = 0,88) als das beste EfficientNet-Modell der vorliegenden Arbeit (Accuracy = 0,67). Dies lässt sich auf mehrere Unterschiede im Studiendesign und den zugrundeliegenden Daten zurückführen.

Erstens basiert die Arbeit von Xu et al. auf MRT-Aufnahmen, die im Gegensatz zu lateralen Knieröntgenbildern höhere Gewebekontrastierung und mehr Tiefeninformation liefern. Die Konturen der Trochlea sind im MRT scharf abgegrenzt und nicht Überlagerungseffekten und der Projektion bei Röntgenaufnahmen unterliegen. Zweitens wurden die MRT-Bilder in Xu et al. vorab manuell mit anatomischen Landmarks versehen. In der vorliegenden Arbeit mussten die EfficientNet-Modelle im Rahmen des Trainingsprozesses eigenständig lernen, welche Bildregionen für die Bestimmung der Trochleattiefe relevant sind. Auch der Erfahrungsstand der beteiligten Personen spielt in der Annotation beziehungsweise

Klassifikation vor KI-Training eine Rolle. In Xu et al. erfolgte die Referenzdiagnose und Annotation durch mehrere erfahrenen Chirurgen und Radiologen („senior doctors“), während die vorliegende Arbeit im Rahmen einer studentischen Diplomarbeit mit vergleichsweise geringer klinisch-radiologischen Erfahrung durchgeführt wurde.

Trotz der genannten methodischen Differenzen und der unterschiedlichen Genauigkeit der KI-Modelle kann sowohl in der beschriebenen Vergleichsstudie als auch in der vorliegenden Arbeit gezeigt werden, dass Methoden der künstlichen Intelligenz, insbesondere die Anwendung von Convolutional Neural Networks, zur Analyse morphologischer Parameter des Kniegelenks grundsätzlich geeignet sind. In beiden Arbeiten gelangt zur Darstellung, dass die automatisierte Beurteilung der Trochlea technisch möglich ist und dysplastische Veränderungen erkannt werden können.

6.3 Limitationen

Die vorliegende Arbeit weist mehrere Limitationen auf, die bei der Deutung der erzielten Ergebnisse berücksichtigt werden sollten. Ein wesentlicher limitierender Aspekt ist die Wahl der Studienpopulation, welche nicht repräsentativ für die gesunde Allgemeinbevölkerung ist. Hierbei handelt es sich um einen Selektionsbias, da ausschließlich Patientinnen und Patienten in die Stichprobe eingeschlossen wurden, bei denen eine MR-Untersuchung des Kniegelenks vorliegt, was bei Kinder- und Jugendlichen in der Regel nur bei entsprechender klinischer Auffälligkeit und pathologischem Verdacht durchgeführt wird. Man könnte behaupten, dass das CNN überproportional mit pathologischen Kniegelenken konfrontiert wird, was somit in einer gesunden Population vermehrt falsch-positive beziehungsweise pathologische Ergebnisse liefern würde.

Ein weiterer Faktor, welcher das Training der EfficientNet-Architekturen erschwert, ist die Altersheterogenität in der Studienpopulation. Die untersuchte Gruppe umfasst Kinder und Jugendliche in verschiedenen Altersgruppen, wodurch signifikante anatomische Unterschiede bedingt durch die unterschiedlichen Ossifikationsstadien bestehen. Für das lernende Convolutional Neural Network ergibt sich somit das Problem, dass keine einheitliche Referenzanatomie vorliegt,

sondern alle Fälle einen unterschiedlichen Grad der Epiphysenverknöcherung aufweisen, was den Lernprozess beeinflusst.

Diese genannten Einflüsse werden zusätzlich verstärkt durch eine begrenzte Fallzahl. Für KI-Anwendungen gilt ganz grundsätzlich, dass die Modellleistung von der Größe und Diversität des zugrundeliegenden Datensatzes abhängt. Mit 1048 verfügbaren Röntgenaufnahmen, die nur von einem einzigen Zentrum akquiriert wurden, steht der Datensatz dieser Arbeit eher im unteren Bereich vergleichbarer Studien. Nach Sichtung mehrerer in PubMed verfügbarer Arbeiten, die KI-Anwendungen auf Basis von zweidimensionalen Röntgenbildern untersuchten, zeigt sich, dass hierfür eher Datensätze mit mehreren tausend Aufnahmen verwendet werden, um Modelle zu trainieren.

Eine weitere Limitation betrifft die Projektion der lateralen Kniegelenksröntgenbilder. Wie bereits in Kapitel 4.1.2 beschrieben, zählt eine starke Abweichung von der korrekten lateralen Projektion zu den Ausschlusskriterien für diese Studie. Diese Beurteilung erfolgte im Wesentlichen nach subjektiver visueller Einschätzung und nicht anhand objektiver Kriterien. Auch geringe Abweichungen der lateralen Projektion verändern die zweidimensionale Darstellung der Trochlea und können somit die Modellleistung beeinflussen. Eine standardisierte Erfassung oder Eintragung des Projektionswinkels im Datensatz hätte eine bessere Bewertung und Analyse der Modellleistung bei verschiedenen Projektionen ermöglicht.

6.4 Ausblick

Die genannten Limitationen liefern eine Erklärung für die eher geringe Modellgenauigkeit und die daraus folgende eingeschränkte klinische Anwendbarkeit. Die grundsätzliche technische Machbarkeit der automatisierten Beurteilung der Trochleatiefe wird dadurch nicht infrage gestellt. Vielmehr zeigen die beschriebenen Einschränkungen konkrete Ansatzpunkte für weitere Forschung auf, welche im Folgenden ausgeführt werden.

Naheliegender für weitere Arbeiten wäre eine deutliche Erweiterung des Datensatzes auch im Sinne einer multizentrischen Studie. Bei entsprechend großer Datenbasis und Einschluss verschiedener Altersgruppen wäre auch denkbar, dass KI-Modelle gezielt auf bestimmte Altersgruppen trainiert werden, um bessere Ergebnisse zu erzielen. Eine hiermit resultierende Subgruppenanalyse kann insbesondere im

Bereich der Kinderradiologie wertvolle Erkenntnisse über die altersabhängige Leistungsfähigkeit von KI liefern.

Durchaus sinnvoll für künftige Studien wäre auch ein gezielter Einschluss von Röntgenaufnahmen gesunder Kniegelenke, das heißt ohne vorliegenden klinischen Verdacht auf eine Pathologie, um den KI-Modellen auch eine gesunde „ground truth“ zu bieten. Diese Herangehensweise wäre aus ethischer Sicht jedoch kritisch zu bewerten, da insbesondere bei Kindern der Einsatz ionisierender Strahlung auf ein Minimum begrenzt werden sollte. Das Problem könnte umgangen werden, indem man auf bereits vorhandene Datensätze verwendet oder auf spezifisch vortrainierte Modelle („pre-trained models“, siehe Kapitel 3.4.1) zurückgreift.

Neben der quantitativen Erweiterung des Datensatzes könnte auch eine qualitative Steigerung in weiterer Forschung angewandt werden. Ein Ansatzpunkt wäre die Verwendung von vorannotierten Bilddaten, um im Rahmen des Lernprozesses der Modelle den Fokus a priori auf die relevanten Strukturen zu lenken. In Anlehnung an die Studie von Xu et al. wäre auch eine Prüfung der Daten durch mehrere erfahrenen Radiologen wünschenswert, um eine höhere Genauigkeit und Konsistenz des Datensatzes zu gewährleisten.

Neben der Präannotation von Bilddaten wäre auch ein Zuschneiden der Aufnahme eine sinnvolle und effizienzsteigernde Maßnahme. Ein konventionelles laterales Kniegelenksröntgen umfasst neben der Trochlea auch beträchtliche Anteile der anderen gelenkbildenden Knochen. Für die Beurteilung der Trochleatiefe sind diese jedoch nicht von diagnostischer Relevanz, und für ein CNN vor allem irrelevante Bildinformation und zusätzliche Rechenleistung. Ein Zuschnitt auf den relevanten Bildteil, welcher auch automatisiert vonstattengehen kann, ließe das Lernverhalten optimieren und auch Ablenkungseffekte durch überflüssige Strukturen reduzieren.

Ein weiterer Ansatz zukünftiger Forschung wäre, die Vorhersage nicht als dichotome Klassifikation, sondern in Form absoluter Messwerte der Trochleatiefe auszugeben und den durchschnittlichen Fehler zu untersuchen. Die Bewertung, ob es sich dann um eine zu flache Trochlea handelt, obliegt dann dem Arzt, was vor allem bei Grenzfällen zwischen normaler und pathologisch flacher Trochlea von Vorteil wäre. Ergänzend hierzu wäre auch die automatisierte Erhebung weiterer für

die Trochleadysplasie relevante Parameter, wie im Kapitel 2.3 beschrieben, von wissenschaftlichem Interesse.

Neben all den genannten Forschungsansätzen wäre auch ein Vergleich mit anderen CNN-Architekturen von Interesse. In der vorliegenden Arbeit kamen ausschließlich Modelle der EfficientNet-Reihe zur Anwendung. Es kann aktuell nicht ausgeschlossen werden, dass alternative Architekturen bei gleicher Datengrundlage eine höhere Vorhersagegenauigkeit erzielen könnten.

6.5 Conclusio

Zusammenfassend kann im Rahmen dieser Diplomarbeit erstmals gezeigt werden, dass Convolutional Neural Networks anhand lateraler Kniegelenksröntgen die Trochleatiefe automatisiert differenzieren können. Die erzielten Resultate beweisen eine technische Machbarkeit, sind in ihrem aktuellen Zustand jedoch noch nicht flächendeckend klinisch einsetzbar. Nichtsdestotrotz können die vorliegenden Ergebnisse als richtungsweisend für weitere Forschung im Bereich der KI-gestützten Diagnostik der patellofemoralen Instabilität angesehen werden und liefern wertvolle Ansatzpunkte für zukünftige Arbeiten, wie im Ausblick zuvor beschreiben.

Literaturverzeichnis

- [1] Flandry F. Normal Anatomy and Biomechanics of the Knee 2011;19.
- [2] Anderhuber F, Pera F, Streicher J. Waldeyer: Anatomie des Menschen ; Lehrbuch und Atlas in einem Band. 19. Aufl. s.l.: Walter de Gruyter GmbH Co.KG; 2012.
- [3] Toms AP, Cahir J, Swift L, Donell ST. Imaging the femoral sulcus with ultrasound, CT, and MRI: reliability and generalizability in patients with patellar instability. Skeletal Radiol 2009;38:329–38. <https://doi.org/10.1007/s00256-008-0639-9>.
- [4] Amis AA. Current Concepts on Anatomy and Biomechanics of Patellar Stability. Sports Med Arthrosc Rev 2007;15:48–56. <https://doi.org/10.1097/JSA.0b013e318053eb74>.
- [5] Wheatley MGA, Rainbow MJ, Clouthier AL. Patellofemoral Mechanics: a Review of Pathomechanics and Research Approaches. Curr Rev Musculoskelet Med 2020;13:326–37. <https://doi.org/10.1007/s12178-020-09626-y>.
- [6] Pauyo T, Park JP, Bozzo I, Bernstein M. Patellofemoral Instability Part I: Evaluation and Nonsurgical Treatment. J Am Acad Orthop Surg 2022;30:e1431–42. <https://doi.org/10.5435/JAAOS-D-22-00254>.
- [7] Tanaka MJ, Sodhi A, Wadhavkar I, Kane K, Velasquez Hammerle MV, Mangudi Varadarajan K, et al. Redefining Trochlear Dysplasia: Normal Thresholds Vary by Measurement Technique, Landmarks, and Sex. Am J Sports Med 2023;51:1202–10. <https://doi.org/10.1177/03635465231158099>.
- [8] Kazley JM, Banerjee S. Classifications in Brief: The Dejour Classification of Trochlear Dysplasia. Clin Orthop 2019;477:2380–6. <https://doi.org/10.1097/CORR.0000000000000886>.
- [9] 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;25:1315–60. <https://doi.org/10.1007/s11030-021-10217-3>.

- [10] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88. <https://doi.org/10.1016/j.media.2017.07.005>.
- [11] Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, Mass: The MIT press; 2016.
- [12] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018;9:611–29. <https://doi.org/10.1007/s13244-018-0639-9>.
- [13] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks 2020. <https://doi.org/10.48550/arXiv.1905.11946>.
- [14] Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, et al. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging* 2016;35:1285–98. <https://doi.org/10.1109/TMI.2016.2528162>.
- [15] Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data* 2019;6:60. <https://doi.org/10.1186/s40537-019-0197-0>.
- [16] Pfirrmann CWA, Zanetti M, Romero J, Hodler J. Femoral Trochlear Dysplasia: MR Findings. *Radiology* 2000;216:858–64. <https://doi.org/10.1148/radiology.216.3.r00se38858>.
- [17] Weiß C. *Basiswissen Medizinische Statistik*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2019. <https://doi.org/10.1007/978-3-662-56588-9>.
- [18] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825–30.
- [19] Xu S-M, Dong D, Li W, Bai T, Zhu M-Z, Gu G-S. Deep learning-assisted diagnosis of femoral trochlear dysplasia based on magnetic resonance imaging measurements. *World J Clin Cases* 2023;11:1477–87. <https://doi.org/10.12998/wjcc.v11.i7.1477>.