Aus der Klinik für Radiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Deep Learning-gestützte Klassifizierung histologisch gesicherter Lebertumoren in der Kontrastmittel-verstärkten MRT-Bildgebung

Deep learning–assisted differentiation of pathologically proven liver tumors on contrast-enhanced MRI

zur Erlangung des akademischen Grades
Doctor medicinae (Dr. med.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Paula Marie Oestmann

Datum der Promotion: 30.11.2023

Inhaltsverzeichnis

1. Abkürzungen

HCC = Hepatozelluläres Karzinom

ICC = Intrahepatisches Cholangiokarzinom

FNH = Fokalnoduläre Hyperplasie

LI-RADS = Liver Imaging Reporting and Data System

OPTN = Organ Procurement and Transplantation Network

MELD = Model for End-Stage Liver Disease

CNN = Convolutional Neural Network

IRB= Institutional Review Board

HIPAA = Health Insurance Portability and Accountability Act

PACS = Picture Archiving and Communication System

MRI = Magnetic Resonance Imaging

MRT = Magnetresonanztomographie

CT = Computertomographie

AUC = Area Under the Curve

PPV = positiv prädiktiver Wert

NPV = negativ prädiktiver Wert

NASH = nichtalkoholische Steatohepatitis

PSC = Primär Sklerosierende Cholangitis

RFA = Radiofrequenzablation

TACE = transarterielle Chemoembolisation

TAE = transarterielle Embolisation

MWA = Mikrowellenablation

SIRT = selektive interne Radiotherapie

2. Abstract:

2.1. Englische Version

Adapted and modified from „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1).

Objectives: To develop a deep learning model that discriminates pathologically proven hepatocellular carcinoma (HCC) and non-HCC lesions including lesions with atypical imaging features on MRI.

Methods: 118 patients with 150 lesions (93 HCC (62%) and 57 non-HCC (38%)) histologically confirmed via biopsy (n = 72), resections (n = 29), liver transplant (n = 46), and autopsy (n = 3) were included in this IRB-approved retrospective study. Atypical imaging features (not meeting Liver Imaging Reporting and Data System [LI-RADS] criteria for definitive HCC/LR5) were shown by 47% of HCC lesions. A 3D convolutional neural network (CNN) was trained on 140 lesions and tested for its ability to classify the 10 remaining lesions (5 HCC/5 non-HCC). Performance of the model was averaged over 150 runs with random sub-sampling to provide class-balanced test sets. The similarity between atypical HCC and non-HCC lesions prone to misclassification by the CNN was captured using a new lesion grading system.

Results: An overall accuracy of 87.3% was found for the CNN. Sensitivities/specificities for HCC and non-HCC lesions were 92.7%/82.0% and 82.0%/92.7%, respectively. The Area Under the Receiver Operating Curve was 0.912. Performance of the CNN was correlated with the lesion grading system, becoming less accurate the more atypical imaging features the lesions showed.

Conclusion: This study provides proof-of-concept for CNN-based classification of both typical- and atypical-appearing HCC on multi-phasic MRI, utilizing pathologically confirmed lesions as “ground truth”.

2.2. Deutsche Version

Hintergrund: Ziel der Studie war es, ein Deep Learning Modell zu trainieren, welches histopathologisch validierte, teils atypisch erscheinende hepatozelluläre Karzinome (HCC) und Nicht-HCC-Läsionen auf MRT-Bildern korrekt unterscheiden kann.

Methodik: Diese durch die Ethikkommission genehmigte retrospektive Studie umfasste 118 Patienten mit 150 Läsionen (93 HCC (62%) und 57 Nicht-HCC (38%)), welche mittels Biopsien (n = 72), Resektionen (n = 29), Lebertransplantationen (n = 46) und Autopsien (n = 3) histopathologisch validiert wurden. Insgesamt zeigten 47% der HCC-Läsionen atypische Bildmerkmale, welche nicht die „Liver Imaging Reporting and Data System“ [LI-RADS] Kriterien für ein definitives HCC/LR5 erfüllten. Ein 3D-Convolutional Neural Network (CNN) wurde mit 140 Läsionen trainiert. Anschließend wurde anhand der 10 verbleibenden Läsionen (5 HCC/5 Nicht-HCC) seine Fähigkeit zur korrekten Klassifizierung getestet. Die Performance des Modells wurde über 150 Runs gemittelt, wobei zufälliges Sub-Sampling eine Ausbalancierung der Testsets ermöglichte. Ein Grading-System wurde entwickelt, um Ähnlichkeiten zwischen atypischen HCC bzw. Nicht-HCC zu demonstrieren, welche zu Fehlklassifikationen des CNNs geführt haben.

Ergebnisse Das CNN zeigte eine Gesamtgenauigkeit von 87.3%. Die Sensitivität bzw. Spezifität für die Charakterisierung der HCC-Läsionen betrug 92.7% bzw. 82.0%, für Nicht-HCC 82.0% bzw. 92.7%. Die Area Under the Receiver Operating Curve (AUC) lag bei 0.912. Die Performance des CNNs wurde mit dem Grading-System korreliert und verschlechterte sich, je mehr atypische Bildmerkmale eine Läsion aufwies.

Schlussfolgerung: Diese Studie bietet eine erste Evidenz für die CNN-basierte Klassifikation von bildmorphologisch typisch sowie atypisch erscheinenden Leberläsionen in der MRT unter Verwendung der histopathologischen Diagnose als “Ground Truth”.

3. Manteltext

3.1. Disclaimer

Da es sich bei dieser Arbeit um eine publikationsbasierte Doktorarbeit handelt, sind Ähnlichkeiten zwischen dem Manteltext und der zu Grunde liegenden Publikation nicht auszuschließen. Daher weise ich hiermit explizit daraufhin, dass der nachfolgende Manteltext auf Basis der Publikation „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“ (1) verfasst wurde.

3.2. Einführung

Das hepatozelluläre Karzinom (HCC) ist die dritthäufigste Krebstodesursache weltweit und ist gleichzeitig die häufigste primäre Neoplasie in der Leber (2). Die häufigste Ursache für die Entwicklung eines HCCs ist mit über 90% die Leberzirrhose, insbesondere wenn diese auf Basis einer chronischen Hepatitis B oder Hepatitis C entstanden ist. Weitere Risikofaktoren sind unter anderem die alkoholtoxische Leberzirrhose, die chronische Virushepatitis ohne Leberzirrhose, die Hämochromatose, die nicht-alkoholische Steatohepatitis (NASH) sowie das Aflatoxin B1 des Pilzes Aspergillus flavus (3).

Klinisch zeigt sich häufig ein Druckschmerz im rechten Oberbauch sowie Kachexie. Im Rahmen der meist vorbestehenden Leberzirrhose können klassische Leberhautzeichen (Spider Naevi, Telangiektasien, Weißnägel, Palmarerythem, Bauchglatze etc.) sowie Aszites und gegebenenfalls ein Sklerenikterus nachgewiesen werden. Zum Diagnosezeitpunkt findet sich bei 50% der Patienten bereits ein multilokuläres Wachstum, bei 25% eine Pfortaderthrombose sowie bei 10% eine Infiltration von Lebervenen und Vena cava inferior (3).

Die Therapie des HCCs erfolgt in Deutschland stadiengerecht anhand der Barcelona-Clinic-Liver-Cancer (BCLC) - Klassifikation sowie der aktuellen S3-Leitlinie (4,5). Laut der im Januar 2021 publizierten BCLC-Klassifikation werden in frühen Stadien die operative Resektion, die

Radiofrequenzablation (RFA) sowie die Lebertransplantation als primär kurative Therapieoptionen empfohlen (6). Dabei konnte bisher kein Überlebensvorteil zwischen der Resektion und der RFA gezeigt werden (7). Im „intermediate stage“ wird die Durchführung einer transarteriellen Chemoembolisation (TACE) sowie eine systemische Therapie mittels Proteinkinaseinhibitoren oder monoklonalen Antikörpern (Immuncheckpoint-/ VEGF-Inhibitoren) empfohlen. Grundsätzlich können ablative Verfahren wie die TACE oder die RFA zudem auch als Bridging-Therapie angewandt werden, um eine Überbrückung bis zu einer potentiellen Lebertransplantation zu ermöglichen. Im fortgeschrittenen Stadium werden laut BCLC 2021 Systemtherapien empfohlen, wobei als erste Wahl hierbei eine Kombinationstherapie aus Atezolizumab und Bevacizumab in Frage kommt (6). In diesem Stadium kann auch eine selektive interne Radiotherapie (SIRT) erwogen werden. Trotz dieser Vielzahl an therapeutischen Maßnahmen ist die 5-Jahresüberlebensrate jedoch weiterhin schlecht. So haben Patienten mit lokalablativer Therapie oder Leberteilresektion eine 5-Jahresüberlebensrate von 20-50% (3). Im „intermediate stage“ zeigt sich bereits nur noch eine mediane Überlebensrate von 20-30 Monaten und im fortgeschrittenen Stadium lediglich eine mediane Überlebensrate von 10-16 Monaten (8–10).

Umso bedeutender ist die frühzeitige und korrekte Diagnostik eines HCCs. Diese wird dadurch erschwert, dass in der Leber auch andere Malignome wie das intrahepatische Cholangiokarzinom (ICC) und Metastasen von vor allem gastrointestinalen Neoplasien sowie eine Vielzahl an benignen Tumoren wie Hämangiomen, Zysten oder fokal noduläre Hyperplasien (FNH) auftreten. Die kontrastmittelverstärkte multiphasische Computertomographie (CT) und die Magnetresonanztomografie (MRT) spielen eine zentrale Rolle für die Diagnose und Klassifikation dieser unterschiedlichen Leberläsionen. Im klinischen Alltag bilden standardisierte Bildgebungskriterien zusammengefasst im „Organ Procurement and Transplantation Network“ (OPTN) oder in den „Liver Reporting & Data System“ (LI-RADS) Kriterien einen Rahmen für die radiologische Befundung (11,12). Darin werden klassischerweise drei Hauptkriterien beurteilt, welche bildmorphologisch den Verdacht auf ein HCC erhärten können: 1. Spätarterielle Kontrastmittelanreicherung („arterial hyperenhancement“), 2. Portalvenöse Kontrastmittelauswaschung („washout“), 3. Spätes Ring-Enhancement („enhancing rim/pseudocapsule“). Jedoch ist die Befundung mit einem hohen Zeitaufwand verbunden und unterliegt abhängig von der Erfahrung eines Radiologen einer mehr oder weniger hohen Interobserver-Variabilität, wodurch die Reproduzierbarkeit limitiert sein kann (13). Insbesondere dann, wenn die Läsionen nicht die oben beschriebenen typischen Bildgebungskriterien erfüllen,

kommt es somit häufig zu potentiell unnötigen Gewebebiopsien (14). Solche Biopsien können auf Grund ihres invasiven Charakters zu Komplikationen wie Blutungen, Sepsis, Karzinoidkrisen oder Tumorstreuung führen (15,16). Diese Komplikationen könnten wiederum eine orthotope Lebertransplantation als kurative Therapieoption für HCCs erschweren (17,18).

In den letzten Jahren hat die Anwendung von künstlicher Intelligenz („KI“) im Bereich der medizinischen Bildanalyse deutlich an Aufmerksamkeit dazugewonnen. Aufgrund der Vielzahl an verschiedenen Begrifflichkeiten wie beispielsweise „Machine Learning“, „Deep Learning“ oder „Convolutional Neural Network“ (CNN) kommt es dabei häufig zu Unklarheiten (Abbildung 1).

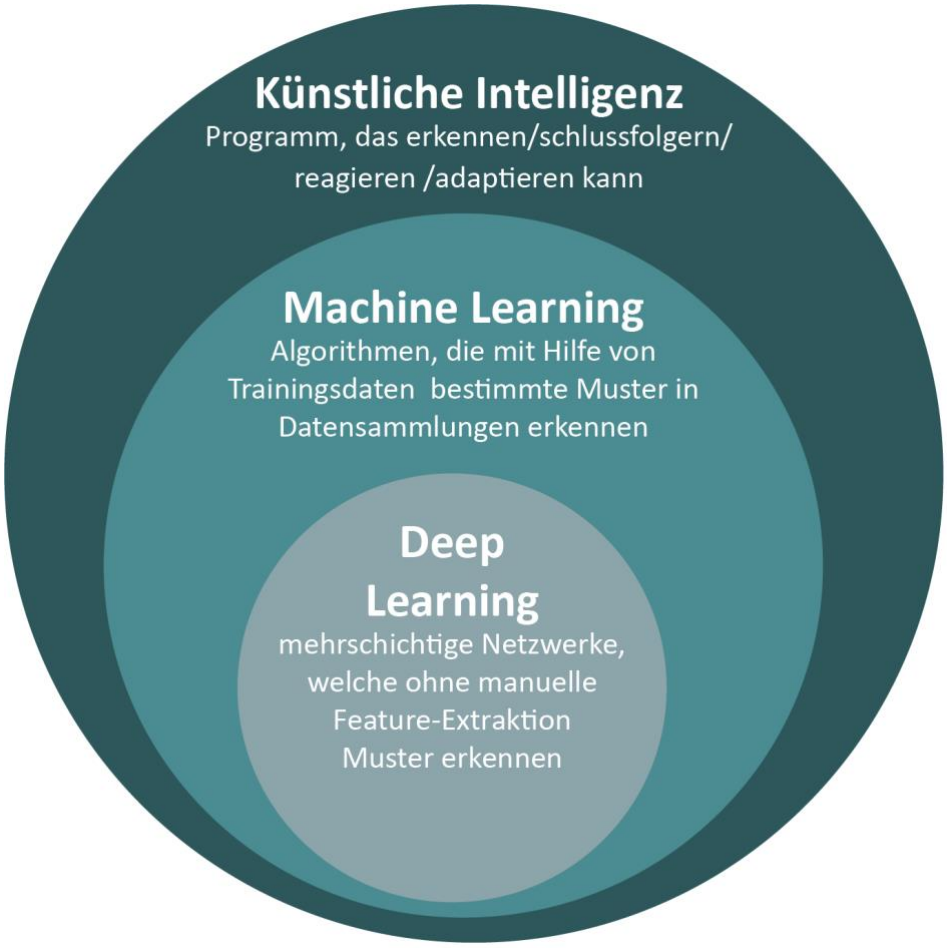

Abbildung 1: Vergleich von Künstlicher Intelligenz, Machine Learning sowie Deep Learning. Modifiziert nach (19).

Der Begriff „Künstliche Intelligenz“ ist als Überbegriff in der Literatur häufig recht unscharf definiert. So wird die künstliche Intelligenz beispielsweise als ein Programm definiert, welches

erkennen, schlussfolgern, reagieren und adaptieren kann (19). An andere Stelle wird sie als ein Teilbereich der Informatik definiert, der sich Systemen widmet, welche Aufgaben bearbeiten können, die normalerweise menschliche Intelligenz erfordern (20). Der Begriff „Machine Learning“ ist wiederum ein Teilbereich der künstlichen Intelligenz, bei dem keine explizite Programmierung durchgeführt wird, sondern ein Algorithmus trainiert wird, bestimmte Muster innerhalb einer Datensammlung zu erkennen (21). Dabei verbessert sich die Performance zunehmend mit der Menge an zur Verfügung stehenden Daten. Klassischerweise werden beim Machine Learning manuell bestimmte Bildkriterien festgelegt („Feature-Extraktion“), die der Machine Learning-Algorithmus überprüft und anschließend ein Ergebnis generiert (20). Hier liegt der entscheidende Unterschied zum „Deep Learning“ (Abbildung 2).

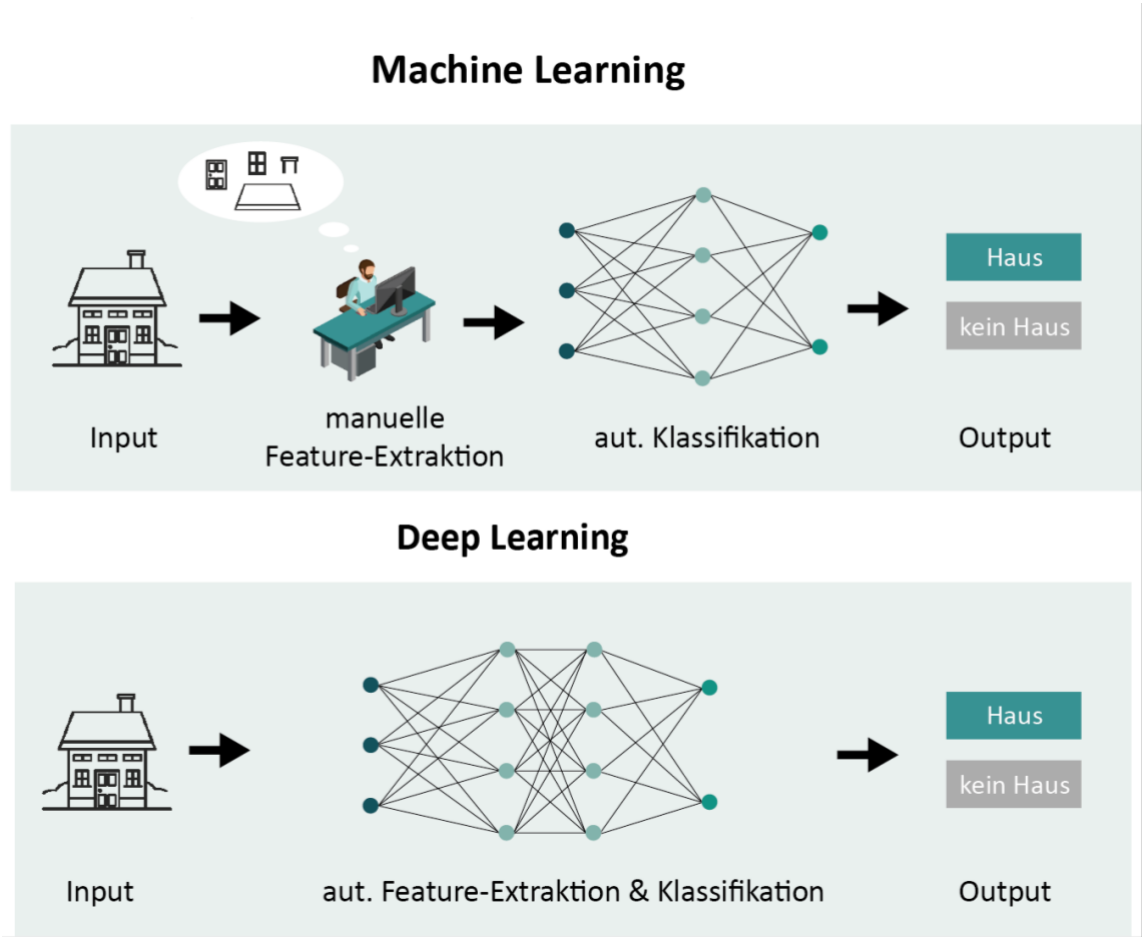

Abbildung 2: Unterschied zwischen Machine Learning und Deep Learning. aut. = automatisch. Modifiziert nach (19). In dieser Abbildung wurden Grafiken von Freepik.com verwendet.

Im Gegensatz zu anderen Machine Learning-Methoden, wie z.B. „Support Vector Machines“, „k-Nearest Neighbors“ oder Entscheidungsbäumen benötigt Deep Learning keine manuell vorgegebenen Kriterien, um Aufgaben zu lösen. Beim Deep Learning erlernt ein Neural Network selbstständig die besten Kriterien für die Klassifizierung der zur Verfügung stehenden Daten. Dies

kann insbesondere dann von Vorteil sein, wenn Kriterien für eine Klassifizierung nur schwer manuell definierbar sind, bespielweise in der Sprach-, Bild- oder Objekterkennung (20).

Das geläufigste Deep Learning Modell, um Läsionen auf radiologischen Bildgebungen zu klassifizieren, ist das Convolutional Neural Network (CNN) (22). Seine Entwicklung begann mit der Erfindung des Neocognitrons, welches das visuelle System von Vertebraten nachahmt und den Weg bereitete für komplexe Bildanalysemodelle, welche eine hohe Performance in der Detektion und Klassifikation von Krankheiten wie Hirnblutungen, Pneumonien und Alzheimer-Demenz gezeigt haben (23–27).

CNNs benötigen, ganz im Sinne eines Deep Learning Modells, keine manuelle Definition von radiologischen Bildkriterien („Feature Extraktion“), sondern lernen selbstständig, wie sie die Bilder korrekt interpretieren. Nachdem dem CNN eine Vielzahl an Bildbeispielen mit und ohne der jeweiligen Krankheit gezeigt werden („Training“), lernt das CNN automatisch radiologische Bildkriterien (28). Anschließend werden dem CNN neue, noch nicht klassifizierte Bilder gezeigt und deren Performance bestimmt („Testing“) (20). Klassischerweise besteht ein CNN aus verschiedenen Schichten („Layers“) von Neuronen („Neurons/Nodes“) mit einem Input am Anfang und einem Output am Ende. Die Verbindungen zwischen den einzelnen Neuronen werden mittels „Weights“ gewichtet. Die Weights aller Verbindungen auf ein nachfolgendes Neuron werden summiert und in eine Aktivierungsfunktion eingespeist, welche wiederum den Output des nachfolgenden Neurons generiert. Weiterhin fungiert diese als Selektierungsinstrument, indem bestimmte Features in den Output übernommen werden und andere wiederum vernachlässigt werden. Eine häufig genutzte Aktivierungsfunktion ist z. B. das „Rectified Linear Unit (ReLU)“ (20).

Werden dem CNN nun Trainingsdaten als Input präsentiert, werden diese Layers bei der sogenannten „Forward Propagation“ der Reihe nach von Input Richtung Output aktiviert (20). Der Output, also die Klassifizierung der Trainingsdaten, wird anschließend mittels einer sogenannten „loss function“ automatisch auf seine Richtigkeit überprüft und damit die Ungenauigkeit im System erkannt. Um diese Ungenauigkeiten in der Performance zu reduzieren werden nun bei der „Back-Propagation“ verschiedene Parameter (sogenannte „Weights“ und „Biases“) jedes einzelnen Neurons adaptiert (20). Dieser Vorgang wird so lange durchgeführt, bis ein optimaler Algorithmus mit der bestmöglichen Performance entstanden ist - das CNN „hat nun seine eigenen Bildkriterien erschaffen“.

Ein CNN besteht klassischerweise aus „Convolutional Layers“, „Pooling Layers“ und ggf. einer oder mehreren „Fully Connected Layers“ (Abbildung 3). Convolutional Layers sind hierbei die namensgebenden und charakteristischen Schichten eines CNNs, da diese sogenannte „Filter“ enthalten, welche die Mustererkennung in Bildern überhaupt erst ermöglichen (20). Die Filter bestehen aus kleinen Rastern (z.B. 3x3), sogenannten „Filtermatrizen“, welche sukzessive über die einzelnen Bildpunkte des Input-Bildes bewegt werden und anschließend einen Output berechnen. Häufig werden diese Convolutional Layers gefolgt von einer Pooling/Downsampling Layer, deren Ziel es ist, eine niedrigere Auflösung mit gleichzeitig höherem Anteil an Informationen für die „Structure of Interest“ zu generieren. Je komplexere Aufgaben das CNN bearbeiten soll, desto mehr Convolutional Layers und Pooling Layers werden nachgeschaltet. Den Abschluss bildet dann meistens eine Fully Connected Layer, bei der jedes Neuron mit allen Neuronen der vorherigen Schicht verbunden ist und so Aussagen über den gesamten Inhalt des Bildes getroffen werden können (20).

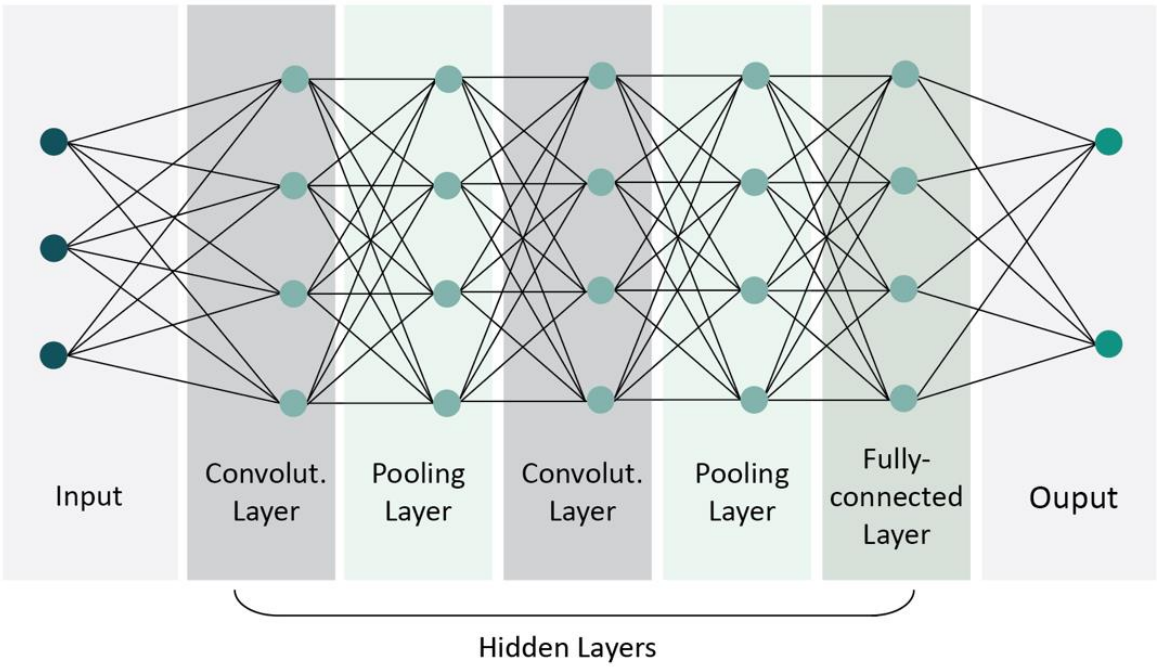

Abbildung 3: Schematischer Aufbau eines Convolutional Neural Networks (CNN). Convolut. = Convolutional. Die Abbildung wurde selbstständig erstellt.

In der letzten Zeit sind mehrere Studien erschienen, in denen CNNs genutzt wurden um Leberläsionen im CT und MRT zu diagnostizieren. Diese vorherigen Studien fokussierten sich jedoch auf Läsionen mit typischem Erscheinungsbild, welche eine eindeutige bildgebungsbasierte Diagnose anhand der oben genannten standardisierten Kriterien ermöglichten (29–31). Um jedoch

im klinischen Alltag bestehen zu können, sollten CNNs auch Läsionen erkennen, die nicht in etablierte Klassifikationssysteme passen. Mit steigender Anzahl an heterogenen Beispielbildern haben CNNs das Potential, atypische Läsionen zu erkennen und dementsprechend den Bedarf an Biopsien und deren Komplikationen zu reduzieren.

Um allerdings den Einschluss von bildmorphologisch atypisch erscheinenden Läsionen überhaupt erst zu ermöglichen, ist die Wahl einer qualitativ hochwertigen „Ground Truth" entscheidend. Der Begriff „Ground Truth" bezeichnet im Bereich des Machine Learnings den Referenzstandard, welcher für das Labeling der Trainings- und Testdaten gewählt wird. Sobald klassische Bildgebungskriterien nicht mehr für eine korrekte Klassifizierung von Trainings- und Testdaten angewendet werden können, ist eine histopathologisch validierte Ground Truth unausweichlich. Doch auch bei bildmorphologisch klassisch erscheinenden Datensets, wäre die Einführung einer histopathologischen Ground Truth ein entscheidendes Instrument, um die Validität von CNNs zu fördern und damit das Vertrauen gegenüber CNNs in radiologischen Fachkreisen zu stärken.

In unsere Arbeitsgruppe sind bereits zwei Studien zum Thema Deep Learning durchgeführt worden. Hierbei wurde zum einen ein CNN untersucht, welches basierend auf kontrastmittelverstärkten MRT-Bildern zwischen sechs verschiedenen Klassen von Leberläsionen mit einer Genauigkeit von 92% differenzieren konnte (29). Zum anderen wurde untersucht, inwiefern dieses CNN seine Entscheidung durch das Identifizieren von klassischen Bildgebungskriterien begründen und somit für den Radiologen interpretierbarer machen konnte (30). Die histopathologisch validierte Ground Truth fehlte bei diesen Untersuchungen jedoch bisher, sodass ein Vorhaben der hier vorliegenden Arbeit darin bestand, diese entscheidende Lücke zu schließen.

Diese Studie soll die Fähigkeit von CNNs prüfen, auf Basis von histopathologisch geprüften Leberläsionen als Ground Truth ein weiteres Spektrum an HCC und Nicht-HCC Läsionen mittels kontrastmittelverstärkten MRT-Bildern korrekt zu diagnostizieren.

3.3. Materialien und Methodik

Diese retrospektive Single-Center Studie wurde durch das Institutional Review Board (IRB) und den Health Insurance Portability and Accountability Act (HIPAA) genehmigt. Sie wurde

durchgeführt gemäß der „Standards for Report of Diagnostic Accuracy“- Richtlinien. Auf eine informierte Einwilligung (Informed Consent) konnte verzichtet werden.

Auswahl der Studienkohorte

Zunächst wurden histopathologisch gesicherte HCC und Nicht-HCC Läsionen von Patienten älter als 18 Jahre, diagnostiziert zwischen 2010 bis 2018 mittels einer elektronischen Patientenakte durch eine Radiologie-Doktorandin (PMO) gesammelt. Die histopathologischen Diagnosen wurden aus elektronisch dokumentierten Befunden von Biopsien (n=72), Resektionen (n=29), Lebertransplantationen (n=46) und Autopsien (n=3) entnommen. Die histopathologische Befundung bestand hierbei aus einer groben makroskopischen Analyse sowie einer detaillierten histologischen Analyse der jeweiligen Läsion. Im Anschluss wurde eine HE- Färbung durchgeführt sowie histologische Oberflächenmarker appliziert, um eine exakte histopathologische Diagnose stellen zu können. Zusätzlich zu dem histopathologischen Befund, musste ein kontrastmittelunterstütztes multiphasisches T1-gewichtetes MRT-Datenset einschließlich einer spätarteriellen, portalvenösen Phase und dem Equilibrium vorliegen, um die Einschlusskriterien zu erfüllen. Die im Pathologiebefund beschriebenen Läsionen wurden mit Hilfe des PACS sowie der elektronischen Patientenakte durch eine Radiologie-Doktorandin (PMO) in den MRT-Bildern identifiziert. Dabei wurde die in den Pathologiebefunden angebende Größe und intrahepatische Lokalisation der Läsion genutzt sowie die im PACS gespeicherten Radiologie-Befunde/Markierungen, um eine klare Zuordnung zwischen Pathologie und Bildgebung herzustellen. Dies geschah in einer kollaborativen Arbeit zusammen mit Pathologen und unter stetiger Supervision durch eine board-zertifizierten Radiologen spezialisiert in abdominaler Bildgebung mit über 25 Jahren Erfahrung. Wenn mehr als eine Läsion im MRT-Bild innerhalb des durch den Pathologen beschriebenen Segmentes sichtbar waren, wurden Bilder von CT-gesteuerten Biopsien hinzugezogen, um die biopsierte Läsion zu identifizieren. Wenn diese Bilder nicht verfügbar waren, wurden alle Läsionen des Segments aus der Studie ausgeschlossen.

Läsionen, welche vor dem MRT-Scan biopsiert wurden, mussten ausgeschlossen werden, wenn Biopsie-bedingte Blutungen zu einer signifikanten Änderung des T1 Signals führten. Bis zu vier Läsionen pro Patienten wurden genutzt. In die Nicht-HCC Gruppe wurden nur primäre Lebertumore eingeschlossen. HCC-Läsionen, welche zwischen der Bildgebung und der Resektion/Transplantation eine lokoregionale Therapie (z.B. TACE/RFA) erhielten, wurden nur eingeschlossen, wenn Reste überlebender Zellen in der Histologie gefunden werden konnten und

diese eine histopathologische Befundung ermöglichten. Tumore mit einer kompletten Nekrose wurden ausgeschlossen.

MRT-Protokoll

Die MRT-Untersuchungen wurden mit 1.5T oder 3T MRT-Geräten, u.a. Signa Excite®, GE Discovery®, Siemens Aera®, Espree®, Verio®, Avanto®, Skyra®, und Trio Tim® Geräten durchgeführt. Alle Patienten erhielten ein natives T1-gewichtetes Bild, bevor intravenös Kontrastmittel appliziert wurde. Dabei wurden Gadolinium-basierte Kontrastmittel genutzt, u.a. Gadavist® (Bayer), Dotarem® (Guerbet), Magnevist® (Bayer), ProHance® (Bracco Diagnostics) und Optimark® (Covidien), in einer Dosierung von 0.1 mmol/kg. Drei T1-gewichtete dreidimensionale (3D) gradient-echo (GRE) breath-hold-Sequenzen mit Fettunterdrückung und Akquisitionszeiten von 12-18s wurden gemäß der CT/MRI LI-RADS Empfehlungen durchgeführt: (1) spätarterielle Phase (2) portalvenöse Phase (3) Äquilibrium. Bolus Tracking wurde in einem großen Teil der Patientenkohorte angewandt. Die Bildgebungsparameter variierten auf Grund des retrospektiven Studienmodells: Repetitionszeit (TR) = 3-5 ms, Echozeit (TE) = 1-2 ms, Kippwinkel von 9-13 Grad, Bandbreite = 300-500 Hz, Schichtdicke = 3-4 mm, Bildmatrix = 256 x 132 bis 320 x 216, Bildfeld = 300 x 200 mm bis 500 x 400 mm. Wenn Patienten mehrere MRT-Aufnahmen erhielten, wurde die MRT-Aufnahme genutzt, bei der zwischen Aufnahme und histopathologischen Befundung das kürzeste Zeitintervall lag.

Bildverarbeitung

Zunächst wurden die ausgewählten MRT-Bilder von der institutionellen Datenbank heruntergeladen. Anschließend wurden von einer Radiologie-Doktorandin (PMO) manuell unter Anwendung des DICOM Viewer Radiant® x, y, und z-Koordinaten für jede einzelne Läsion definiert. Mit diesen Koordinaten war es möglich, automatisch eine dreidimensionale Begrenzungsbox („Bounding Box“) um die Läsion herum zu erstellen (Abbildung 4).

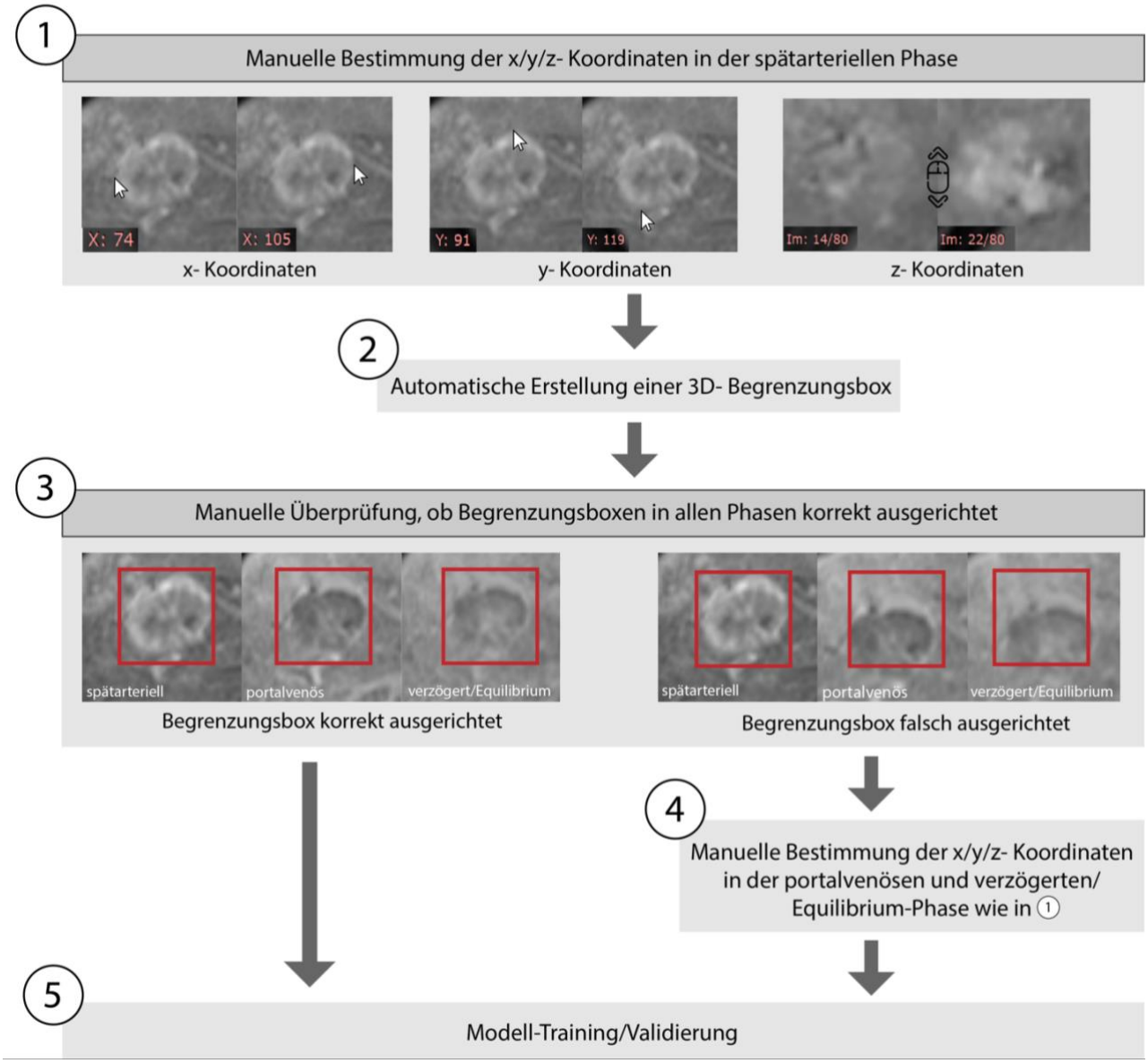

Abbildung 4: Erstellung der Koordinaten und Begrenzungsboxen. 1): Alle Koordinaten wurden manuell in der spätarteriellen Phase unter Einsatz eines DICOM-Viewers (Radiant®) bestimmt. Die maximale Ausbreitung der Läsion innerhalb einer Achse wurde durch zwei Koordinaten definiert. 2): Die 3D-Begrenzungsboxen wurden automatisch auf Grundlage der vorher definierten Koordinaten erstellt. 3): Anschließend wurden die Begrenzungsboxen manuell überprüft, um sicher zu stellen, dass diese in allen Phasen korrekt über der jeweiligen Läsion ausgerichtet sind. 4): In den wenigen Fällen, bei denen die Begrenzungsboxen beispielsweise durch atmungsassoziierte Bewegungsartefakte in späteren Phasen nicht mehr korrekt ausgerichtet waren, wurden die Koordinaten manuell für die portalvenöse bzw. verzögerte/Equilibrium-Phase modifiziert. Dabei wurde wie in Schritt 1) beschrieben vorgegangen. 5): Anschließend wurde das Modell-Training bzw. die Validierung durchgeführt wie in Abbildung 5 beschrieben.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

Nur das Bildvolumen innerhalb dieser Begrenzungsbox wurde vom Modell analysiert. Die Bilder wurden in der Programmiersprache Python 3.5 (Python Software Foundation) verarbeitet. Die MRT-Sequenzen wurden mittels Mutual-Information automatisch registriert, um die portalvenöse Phase sowie das Äquilibrium bestmöglich in Übereinstimmung mit der spätarteriellen Phase zu bringen. Die Bilder wurden dann auf die oben beschriebenen 3D- Begrenzungsboxen zugeschnitten. Zudem wurde im Rahmen der Bildverarbeitung eine Normalisierung hin zu einem Intensitätsbereich von -1 zu 1 durchgeführt, um Bias Field Effekte zu reduzieren. Anschließend wurden die Bilder auf eine Größe von 36x36x12 Voxel überführt.

Um die Anzahl an Training-Bildproben zu erhöhen, wurde das Training Set nach aktuellen Standards um den Faktor 100 (n=14000) augmentiert (Abbildung 5). Dabei wurden die Bilder nach dem Zufallsprinzip rotiert, verschoben, skaliert, umgedreht, innerhalb der Phasen verschoben und eine Intensitäts-Skalierung bzw. Verschiebung durchgeführt. Dies ermöglichte dem Modell bestimmte Bildmerkmale zu lernen, die sich invariant gegenüber Rotation oder Translation zeigen (32).

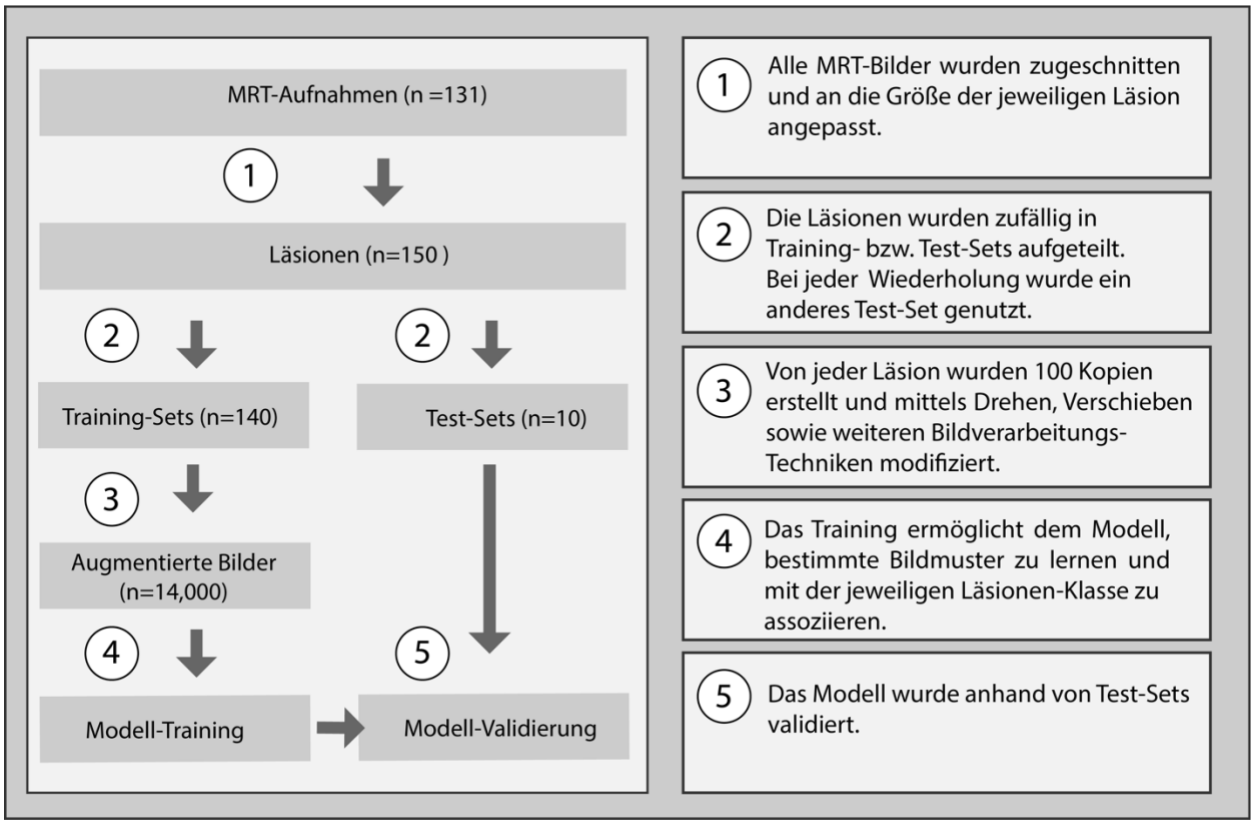

Abbildung 5: Übersicht der einzelnen Arbeitsschritte einschließlich Modell-Training und Validierung.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

Architektur des Convolutional Neural Network (CNN)

Das Modell wurde in der Programmiersprache Python 3.5 und Keras 2.2 (https://keras.io/) auf einem „Tensorflow backend“ (Google, Mountain View, https://www.tensorflow.org/) erstellt und auf einem GeForce GTX 1060 (NVIDIA) Grafikprozessor trainiert. Bei dem CNN handelt es sich um ein Vanilla CNN, welches aus drei Convolutional-Layers (64, 128 bzw. 128 Channels, Kernel-Size 3x3x2), zwei Maximum Pooling Layers (Size 2x2x2 bzw. 2x2x1) und zwei Fully Connected Layers (100 bzw. 1 Neuron) besteht. Das hier genutzte CNN errechnet einen sigmoidalen Output, welcher mit der Wahrscheinlichkeit einer Läsion, ein HCC zu sein, übereinstimmt. Das CNN nutzt Rectified Linear Units, Batch Normalization und einen 10% Dropout.

Training und Bewertung

Das CNN wurde trainiert mittels 70 HCC und 70 Nicht-HCC-Bildbeispielen, welche zufällig aus dem augmentierten Datenset gewählt wurden. Es wurde ein Adam Optimizer mit einer Minibatch Size von 20 und einer Lernrate von 0.01 genutzt. Anschließend wurde das Modell anhand eines Testsets aus 10 bisher ungesehen Läsionen in seiner Fähigkeit, Läsionen korrekt zu klassifizieren (Performance), getestet. Dabei wurde das Testset per Zufallsprinzips aus 5 HCC-Läsionen und 5 Nicht-HCC-Läsionen erstellt. Insgesamt wurden 150 unabhängige Runs mit verschiedenen Aufteilungen der Training- und Testdaten-Sets durchgeführt, um die Performance des CNNs zu ermitteln. Hierbei wurde eine Monte Carlo Cross-Validation genutzt, um die Anzahl an HCC- und Nicht-HCC Läsionen innerhalb eines Sets auszubalancieren. Diese Herangehensweise sowie die 14:1 Training/Test-Ratio stimmen mit der aktuellen „Best Practice“ des Machine Learning überein (21,33).

Läsionen-Grading

Da das Datenset Läsionen enthielt, die atypische Bildmerkmale aufwiesen, wurde ein Läsionen-Grading-System entwickelt basierend auf den LI-RADS Major Imaging Features. Diese umfassen die drei typischen Bildmerkmale einer HCC-Läsion: 1. spätarterielle Kontrastmittelanreicherung

2. Portalvenöse Kontrasmittelauswaschung 3. spätes Ring-Enhancement (Abbildung 6). Für jedes anwendbare Bildmerkmal wurde von einer Radiologie-Doktorandin (PMO) unter Supervision jeweils ein Punkt vergeben, sodass eine Läsion zwischen 0 und 3 Punkten erhalten konnte.
Nach diesem Grading-System wurden sowohl die HCC-Läsionen als auch die Nicht-HCC Läsionen bewertet, um zwischen diesen beiden Klassen die Ähnlichkeit zu demonstrieren, welche zu Fehlklassifikationen des CNNs geführt haben könnten. Eine Läsion mit 3 Punkten konnte demnach sowohl eine typische LI-RADS-konforme HCC-Läsion sein als auch eine histopathologisch geprüfte Nicht-HCC Läsion, die sich jedoch wie ein HCC in der Bildgebung präsentierte. Eine Läsion mit einem Punkt wiederum, konnte zum einen eine Nicht-HCC Läsion sein, zum anderen eine HCC-Läsion mit atypischen Bildmerkmalen, welche nicht dem typischen Kontrastmittelverlauf der LI-RADS Kriterien entsprach.

Zwischen den gut (>90% Genauigkeit) und schlecht (<90% Genauigkeit) klassifizierten Läsionen wurden die Unterschiede im Grading-Score analysiert, um mögliche Erklärungsansätze für die Fehlklassifikationen des CNNs zu erhalten.

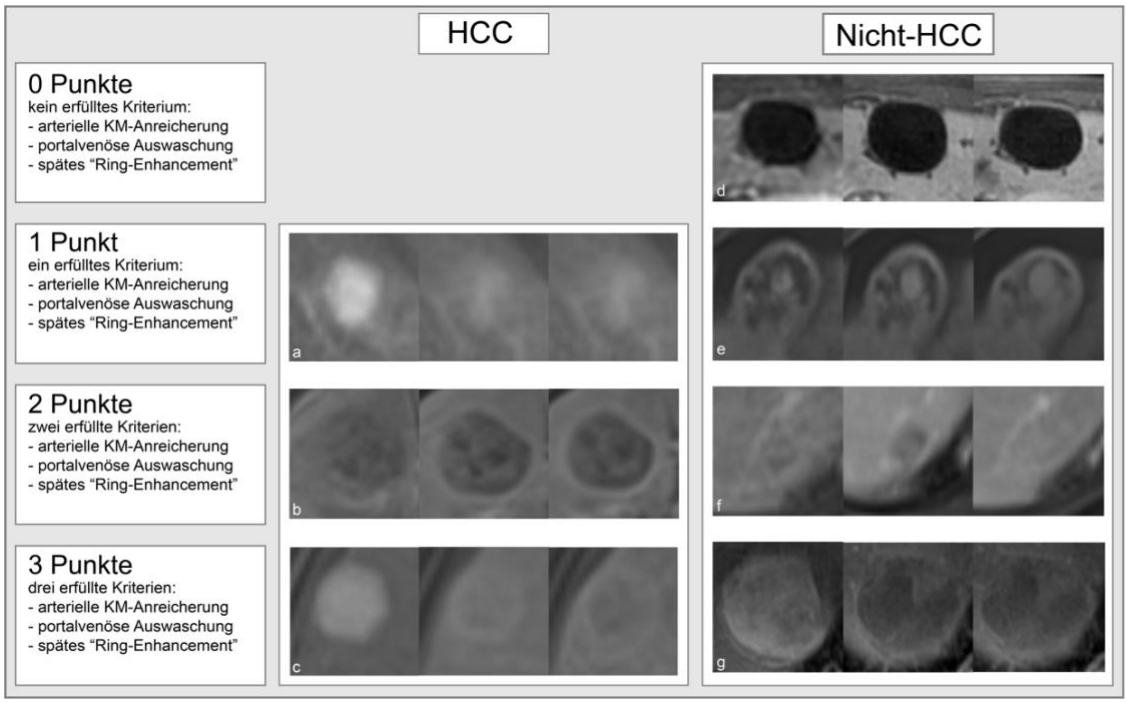

Abbildung 6: HCC und Nicht-HCC-Läsionen wurden mit 0-3 Punkten bewertet, um die Ähnlichkeit zwischen HCC und Nicht-HCC-Läsionen als mögliche Ursache für Fehlklassifikationen durch das CNN darzustellen. KM = Kontrastmittel. a: HCC mit arterieller KM-Anreicherung, b: HCC mit portalvenöser Auswaschung und Ring-Enhancement, c: HCC mit arterieller KM-Anreicherung, mit portalvenöser Auswaschung und Ring-Enhancement d: Zyste mit keinem erfüllten Kriterium, e: Hämangiom mit Ring-Enhancement, f: Hämangiom mit Auswaschung und Ring-Enhancement, g: Zyste mit arterieller KM-Anreicherung, portalvenöser Auswaschung und Ring-Enhancement.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

Statistik

Sensitivität, Spezifizität und die Gesamtgenauigkeit wurden über 150 Runs gemittelt, um die Performance des Deep Learning Modells zu validieren. Dabei wurde mit zufälligem Sub-Sampling gearbeitet, um ausbalancierte Testsets zu erhalten. Zudem wurde einer Receiver Operating Characteristic -Kurve ermittelt und die Area Under the Curve (AUC) berechnet (Abbildung 7).

3.4. Ergebnisse

Studienpopulation

In dieser Studie wurden 118 Patienten mit HCC (n=73, 62%) und Nicht-HCC Läsionen (n=45, 38%) eingeschlossen. Die HCC-Kohorte bestand aus 57 (78%) Männern und 16 (22%) Frauen, während die Nicht-HCC Kohorte 23 (51%) Männer und 22 (49%) Frauen umfasste.
Das Durchschnittsalter der HCC-Patienten lag bei 61±8 (Durchschnitt, Standardabweichung) und das Durchschnittsalter der Nicht-HCC Patienten bei 59±13 Jahren. Die Studienkohorte enthielt 87 Patienten mit Leberzirrhose, davon 73 (84%) in der HCC-Gruppe und 14 (16%) in der Nicht-HCC-Gruppe. Die Mehrheit dieser Patienten wurde mit einem Child-Turcotte-Pugh-Score A (n=50, 57%) klassifiziert und die häufigste Ursache (n=61, 59%) für die Leberzirrhose stellte eine Hepatitis-C-Infektion dar. Der Mittelwert des „Model for End-Stage Liver Disease“ (MELD) Scores betrug 9. Die exakten Zahlen können der Tabelle 1 entnommen werden.

Tabelle 1: Patientencharakteristika

	HCC	Nicht - HCC						
		ICC	Regenerat-knoten	Dysplast. Knoten	Hämangiom	Zyste	FNH	Gallengangs-adenom
Anzahl der Patienten	73	12	2	2	16	10	2	1
Geschlecht								
- **Männlich**	57 (78)	9 (75)	1 (50)	1 (50)	7 (44)	3 (30)	1 (50)	1 (100)
- **Weiblich**	16 (22)	3 (25)	1 (50)	1 (50)	9 (56)	7 (70)	1 (50)	0 (0)
Alter bei Bildgebung	61±8	69±13	37*	61*	57±10	56±9	42*	53*
Ethnie								
- **Kaukasisch**	53 (73)	9 (75)	1 (50)	2 (100)	11 (69)	8 (80)	1 (50)	1 (100)
- **Afroamerikanisch**	9 (12)	2 (17)	0 (0)	0 (0)	2 (13)	0 (0)	1 (50)	0 (0)
- **Asiatisch**	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
- **Andere**	10 (14)	1 (8)	1 (50)	0 (0)	3 (19)	2 (20)	0 (0)	0 (0)
MELD	10*	13±6	20*	10*	8±2	6*	10*	10*
Zirrhose	73	1	2	2	6	2	0	1
- **Child-Pugh**								
o **A**	44 (60)	0 (0)	0 (0)	1 (50)	4 (67)	0 (0)	0 (0)	1 (100)
o **B**	26 (36)	1 (100)	1 (50)	1 (50)	2 (33)	2 (100)	0 (0)	0 (0)
o **C**	3 (4)	0 (0)	1 (50)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
- **Ätiologie**								
o **Hepatitis B**	2 (3)	0 (0)	1 (50)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
o **Hepatitis C**	53 (62)	1 (50)	0 (0)	2 (100)	2 (33)	2 (100)	0 (0)	1 (100)
o **Alkohol**	21 (25)	0 (0)	0 (0)	1 (50)	2 (33)	1 (50)	1(100)	0 (0)
o **NASH**	8 (9)	0 (0)	1 (50)	0 (0)	1 (17)	0 (0)	0 (0)	1 (100)
o **PSC**	1(1)	1 (50)	0 (0)	0 (0)	1 (17)	0 (0)	0 (0)	0 (0)
Bei malignen Läsionen								
- **ECOG**								
o **0**	55 (75)	3 (25)						
o **1**	16 (22)	4 (33)						
o **2**	1 (1)	2 (17)						
o **3**	1 (1)	1 (8)						
o **unbekannt**	0 (0)	2 (17)						
- **extrahep. Metastasen**	1 (14)	0 (0)						
HCC related								
- **BCLC**								
o **0**	12 (16)							
o **A**	45 (62)							
o **B**	0 (0)							
o **C**	13 (18)							
o **D**	3 (4)							
- **HKLC**								
o **1**	43 (58)							
o **2**	26 (36)							
o **3**	1 (1)							
o **4**	0 (0)							
o **5**	3 (4)							

Tabelle 1: Patientencharakteristika. Numerische Daten wurden als Durchschnitt + Standardabweichung oder Median (*) dargestellt, kategorischen Daten als prozentuale Häufigkeit. HCC = Hepatozelluläres Karzinom, ICC = Intrahepatisches Cholangiokarzinom, FNH = Fokal noduläre Hyperplasie, MELD = Model for End-Stage Liver Disease, Child-Pugh= Child-Turcotte-Pugh-Score, NASH= Nicht-alkoholische Steatohepatitis, PSC = Primär Sklerosierende Cholangitis, ECOG = Eastern Cooperative Oncology Group,

BCLC = Barcelona Clinic Liver Cancer, HKLC = Hong Kong Liver Cancer classification system. Dysplast. = Dysplastisch.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

Insgesamt wurden 93 (62%) HCC-Läsionen und 57 (38%) Nicht-HCC Läsionen analysiert. Die Nicht-HCC Gruppe bestand aus 19 (33%) ICCs, 16 (28%) Hämangiomen, 15 (26%) Zysten, 2 (4%) regenerativen Knoten, 2 (4%) dysplastischen Knoten, 2 (4%) FNHs und 1 (2%) Gallengangsadenom. Die mittlere Läsionsgröße betrug 2,3cm. Für HCC-Läsionen lag das mittlere Zeitintervall zwischen MRT-Scan und histopathologischer Befundung bei 1.6 Monate (Zeitraum, 0-25 Monate), sofern die Bildgebung vor der pathologischen Befundung stattfand. Sofern der MRT-Scan erst nach der histopathologischen Befundung durchgeführt wurde, geschah dies innerhalb eines Tages. Für Nicht-HCC Läsionen lag das mittlere Zeitintervall zwischen MRT und pathologischer Befundung bei 1.4 Monaten (Zeitraum, 0-73 Monate), sofern die Bildgebung vor der pathologischen Befundung stattfand. Wurde die Bildgebung nach der histopathologischen Befundung durchgeführt, geschah dies in einem mittleren Zeitraum von 5.5 Monaten (Zeitraum, 0-24 Monate) (Tabelle 2). Es wurden pro Patienten eine bis vier Läsionen (Mittelwert=1) und ein bis drei MRT-Scans (median=1) eingeschlossen (Tabelle 3).

Tabelle 2: Charakteristika der Läsionen

	HCC	Nicht - HCC						
		ICC	Regenerat-knoten	Dysplast. Knoten	Hämangiom	Zyste	FNH	Gallengangs-adenom
Anzahl der Läsionen	93	19	2	2	16	15	2	1
Histopathologie								
- **Biopsie**	47 (50)	15 (79)	1 (50)	1 (50)	6 (37)	0 (0)	2 (100)	0 (0)
- **Resektion**	10 (11)	4 (21)	0 (0)	0 (0)	5 (31)	10 (67)	0 (0)	0 (0)
- **Explantat**	36 (39)	0 (0)	1 (50)	1 (50)	3 (19)	4 (27)	0 (0)	1 (100)
- **Autopsie**	0 (0)	0 (0)	0 (0)	0 (0)	2 (13)	1 (7)	0 (0)	0 (0)
Zirrhose	93	1	2	2	6	4	0	1
Zeitintervall in Tagen (Median)								
- **MRT vor Histologie**	49	22	42	68	104	181	509	27
- **MRT nach Histologie**	1	295	0	0	143	0	0	0
Durchmesser in cm	2,0*	4.2±1.4	3,7*	1.1*	5.0±4.0	4.9±3.5	4,46*	1.4*
Resttumor	8	0						
Vorbehandelt	29	0						
- **TACE**	22 (76)							
- **TAE**	3 (10)							
- **Ethanolablation**	2 (7)							
- **MWA**	6 (21)							
- **RFA**	3 (10)							
LI-RADS								
- **LR5**	49 (53)							
- **< LR5**	44 (47)							

Tabelle 2: Charakteristika der Läsionen. Numerische Daten wurden als Durchschnitt + Standardabweichung oder Median (*) dargestellt, kategorischen Daten als prozentuale Häufigkeit. HCC = Hepatozelluläres Karzinom, ICC = Intrahepatisches Cholangiokarzinom, FNH = Fokal noduläre Hyperplasie, TACE = transarterielle Chemoembolisation, TAE = transarterielle Embolisation, MWA = Mikrowellenablation, RFA = Radiofrequenzablation, Dysplast. = Dysplastisch.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

Tabelle 3: Bildcharakteristika

	HCC	Nicht - HCC						
		ICC	Regenerat-knoten	Dysplast. Knoten	Hämangiom	Zyste	FNH	Gallengansadenom
Anzahl Patienten	73	12	2	2	16	10	2	1
Anzahl MRT-Scans	80	17	2	2	16	11	2	1
Anzahl Läsionen	93	19	2	2	16	15	2	1

Tabelle 3: Bildcharakteristika. HCC = Hepatozelluläres Karzinom, ICC = Intrahepatisches Cholangiokarzinom, FNH = Fokal noduläre Hyperplasie, Dysplast. = Dysplastisch.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver"(1) mit freundlicher Genehmigung des „Springer Nature"-Verlags.)*

Deep Learning Modell-Performance

Das Deep Learning Modell zeigte eine Trainingsgenauigkeit von 94.1% ± 2.0 (19766/21000 Volumetric Samples). Die Performance wurde mittels eines Test-Sets nach 30 Wiederholungen validiert. Dabei zeigte das CNN eine Gesamtgenauigkeit von 87.3% ± 10.5 (1310/1500). Die Sensitivität, HCC bzw. Nicht-HCC-Läsionen zu klassifizieren lag bei 92.7%, bzw. 82.0% und die Spezifität für HCC bzw. Nicht-HCC Gruppe lag bei 82.0% bzw. 92.7% (Tabelle 4).

Tabelle 4: Performance des Neural Networks bei der HCC-Klassifikation

	HCC	**Nicht - HCC**	**Insgesamt**
Training Läsionen	88	52	140
Test Läsionen	5	5	10
Sensitivität	92.7%	82.0%	87.3%
Spezifität	82.0%	92.7%	87.3%

Tabelle 4: Performance des Neural Networks bei der HCC-Klassifikation. Die Performance wurde über 150 Runs mit zufälligen Stichproben gemittelt, um Klassen-balancierte Test-Sets zu erhalten. HCC = Hepatozelluläres Karzinom.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver"(1) mit freundlicher Genehmigung des „Springer Nature"-Verlags.)*

Die Receiver Operating Characteristic Curve zeigte eine AUC von 0.912 (Abbildung 7). Das CNN wurde innerhalb von 3.2 Minuten ± 0.9 trainiert und die Berechnungszeit für die Klassifizierung einer Läsion im Testset lag bei 2.9 Millisekunden ± 1.7.

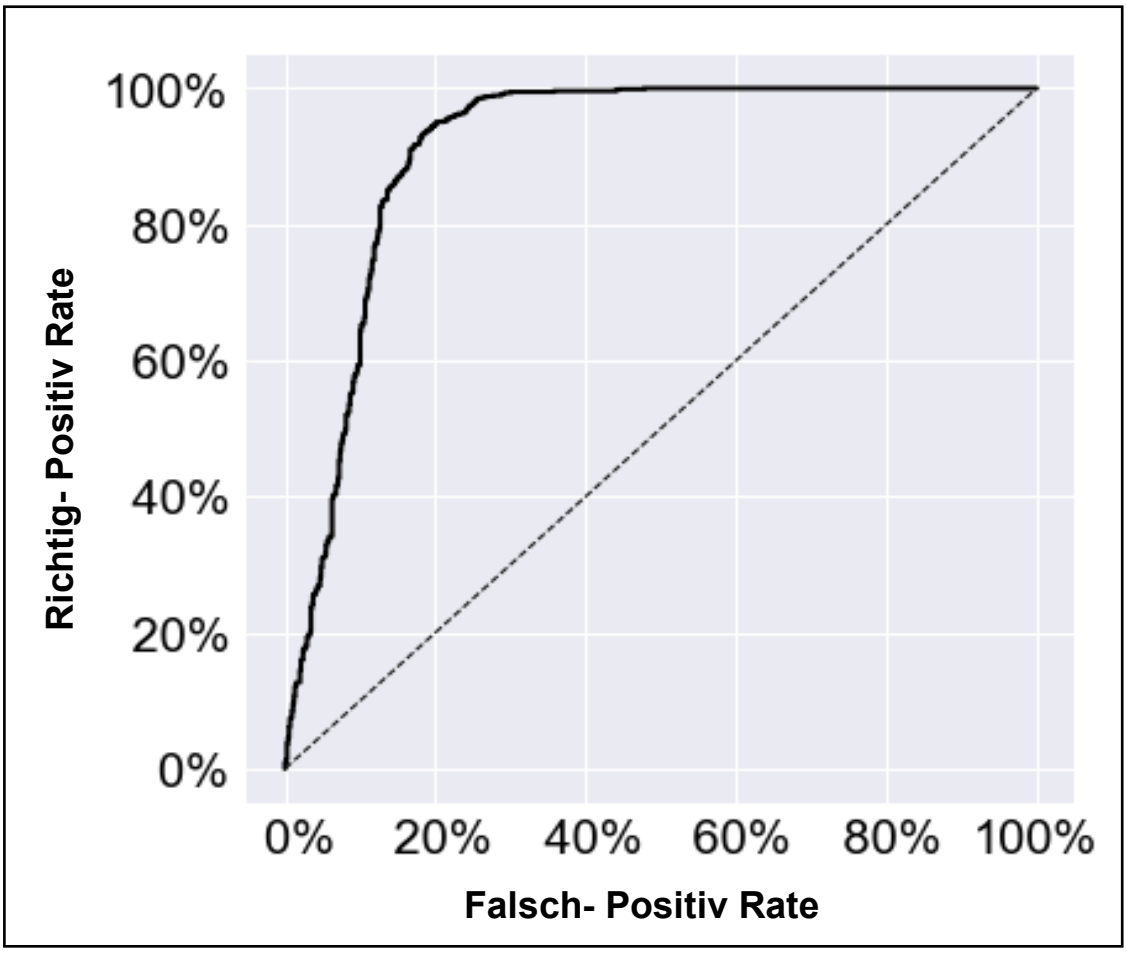

Abbildung 7: Receiver Operating Characteristic (ROC)-Kurve für die Unterscheidung zwischen hepatozellulärem Karzinom (HCC) und Nicht-HCC Läsionen durch das CNN. AUC = Area Under the Curve.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver"(1) mit freundlicher Genehmigung des „Springer Nature"-Verlags.)*

Bewertung des Läsionen-Grading-Scores

Gemäß des entwickelten Grading Systems wurden 23 (25%) der HCC-Läsionen mit einem Punkt bewertet, 28 (30%) mit 2 Punkten und 42 (45%) mit 3 Punkten (Abbildung 8). In der Nicht-HCC-Klasse wurden 16 (28%) Läsionen mit 0 Punkten, 24 (42%) mit einem Punkt, 11 (19%) mit 2 Punkten und 6 (11%) mit 3 Punkten kreditiert. The Kruskal-Wallis Test zeigte eine signifikant positive Korrelation zwischen der steigenden Punktzahl des Grading-Scores und einer verbesserten Klassifikationsgenauigkeit bei HCC-Läsionen (p=0.012), sowie einer reduzierten Klassifikationsgenauigkeit für Nicht-HCC- Läsionen (p < .001). In der HCC-Klasse wurden 1 von 42 (2%) 3-Punkte-Läsionen, 4 von 28 2-Punkte-Läsionen (14%) und 5 von 23 (22%) 1-Punkt-Läsionen schlecht vom CNN klassifiziert (≤90% Genauigkeit in 150 Runs). Die eine schlecht klassifizierte 3-Punkte-HCC-Läsion sowie drei der vier schlecht klassifizierten 2-Punkte-HCC-Läsionen zeigten eine schlechte Bildqualität. Zudem waren zwei der vier schlecht klassifizierten 2-Punkt- HCC-Läsionen nahe am Leberrand gelegen. In der Nicht-HCC-Klasse wurden keine der 0-Punkte Läsionen, 2 von 24 (8%) 1-Punkte-Läsionen, 3 von 11 (27%) 2-Punkte-Läsionen und 6 von 6 (100%) 3-Punkte-Läsionen schlecht vom CNN klassifiziert.

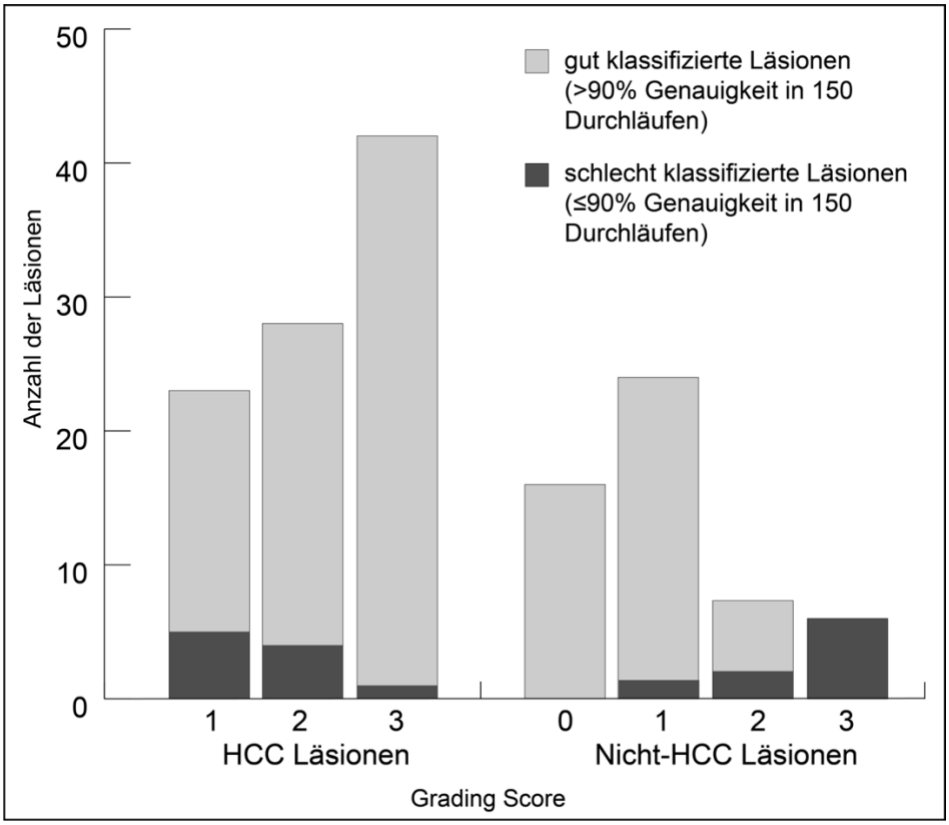

Abbildung 8: Anzahl der Läsionen gemäß des Grading Scores. HCC = Hepatozelluläres Karzinom.

***(Entnommen aus** „Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver“(1) mit freundlicher Genehmigung des „Springer Nature“-Verlags.)*

3.5. Diskussion

In dieser Studie wurde die Fähigkeit eines Deep Learning Modells, ein breiteres Spektrum an histopathologisch validierten Leberläsionen mit zum Teil atypischen Bildmerkmalen korrekt in die Klassen HCC und Nicht-HCC zu unterteilen auf Basis von kontrastmittelverstärkten MRT-Bildern. Das Modell erreichte eine Gesamtgenauigkeit von 87.3% mit einer hohen Sensitivität (92.7%) und einer moderaten Spezifität (82.0%) für HCC. Die kurze Rechenzeit des CNNs könnte eine praxisnahe Integration in den klinischen Alltag eines Radiologen ermöglichen ohne Verzögerungen hervorzurufen.

In letzter Zeit haben sich bereits mehrere Studien auf die Klassifikation verschiedener Leberläsionen mittels eines Deep Learning-Ansatzes fokussiert. So wurde zum Beispiel in einer vorherigen Studie (29) ein CNN mit sechs verschiedenen Arten von Leberläsionen trainiert (HCC, ICC, Zysten, Hämangiome, FNH und kolorektale Metastasen) und zeigte bei der Klassifikation dieser Läsionen eine Gesamtgenauigkeit von ungefähr 90%. Diese Proof-of-Concept Studie nutzte jedoch nur Leberläsionen mit typischen Bildmerkmalen. Das Einschließen von atypischen Läsionen in Training als auch Testsets ermöglicht jedoch ein repräsentativeres Datenset, dessen Testergebnisse sich eher in den klinischen Alltag übersetzen ließe.

Eine andere Studie, welche sich mit Deep Learning-basierter Lebertumorklassifikation befasst, schloss auch atypische bzw. unklare Läsionen mit ein. Jedoch wurden hier alle unklaren Läsionen vom CNN in eine Klasse gruppiert ohne eine weitere Subklassifikation (31). Das CNN in unserer Studie wurde auf einer Mehrheit von atypischen Läsionen trainiert, um diese als HCC bzw. Nicht-HCC Läsion zu klassifizieren. Diese binäre Unterscheidung ist ein signifikanter Schritt hin zu der nicht-invasiven Diagnosestellung bildmorphologisch unklarer Läsionen im klinischen Alltag. Die Entscheidung HCC vs. Nicht-HCC ist insbesondere deshalb entscheidend, weil das HCC als maligne Entität eine hohe Letalität aufweist, jedoch kurativ therapiert werden kann, wenn eine Diagnose frühzeitig erfolgt. Zudem wurden in der eben genannten Studie CT-Bilder genutzt, während in unserer Studie MRT-Bilder als Training- bzw. Testset genutzt wurden. Auf Grund des

wesentlich besseren Weichteilkontrastes können MRT-Bilder dem CNN eine größere Bandbreite an Bildmerkmalen zur Verfügung stellen.

In unserer Studie entsprachen 47% der HCC-Läsionen nicht den LI-RADS Kriterien für ein definitives HCC (LR5) und 48% der Läsionen wurden biopsiert, was generell ein unklares Erscheinungsbild der Läsionen nahelegt. Zudem wurde ein Grading-System angewandt, um die Repräsentation von atypischen Läsionen einzuschätzen. Hierbei wurde jeweils ein Punkt für jedes klassische Bildmerkmal eines HCCs (spätarterielle Kontrastmittelanreicherung, venöses Auswaschen, spätes Ring-Enhancement) vergeben. Gemäß diesem Grading-System wurden 25% der HCC-Läsionen auf Grund ihres atypischen Erscheinungsbildes mit einem Punkt versehen und 30% der Nicht-HCC Läsionen mit 2 oder mehr Punkten bewertet, da ihr Erscheinungsbild dem einer klassischen HCC-Läsionen ähnelte. Obwohl unsere Studie eine etwas geringere Gesamtgenauigkeit als die obengenannte Studie mit den klassisch-erscheinenden Läsionen zeigt, legen die Ergebnisse nahe, dass ein CNN, welches mit pathologisch-validierten teils atypischen Läsionen trainiert wurde, immer noch eine relativ hohe Genauigkeit aufweist. Im Allgemeinen zeigten Läsionen mit typischem Erscheinungsbild eine höhere Klassifikationsgenauigkeit. Die erniedrigte Spezifität der HCC-Klassifikation ist vermutlich durch Nicht-HCC-Läsionen bedingt, welche in der Bildgebung klassische HCC-Kriterien erfüllten. Jedoch wurde auch eine kleine Anzahl an HCC-Läsionen mit 2 und 3 Punkten schlecht klassifiziert, was vermutlich durch schlechte Bildqualität oder die Nähe zum Leberrand bedingt war.

Bei dem hier genutzten CNN handelt es sich um ein Vanilla CNN, welches eine angemessene Wahl für die hier genutzten kleinen zugeschnittenen 3D Bilder war. Andere anspruchsvollere Architekturen wie ResNet (34) oder das DenseNet (35) sind für größere Datensets und 2D High Resolution-Bilder vorgesehen. Die scheinbar hohe Standardabweichung kommt durch die Zahl von Validierungs-Bildern pro Fold zu Stande.

Diese Studie hat mehrere Limitationen. Die relativ kleine Studienkohorte ist einerseits bedingt durch das obligate Einschlusskriterium eines histopathologischen Referenzstandards, andererseits bedingt durch die monozentrische Datenakquisition. Da die Mehrheit der Nicht-HCC Läsionen in der Leber gutartig ist und keine chirurgische Therapie benötigt, waren im Vergleich zu HCC-Läsionen weniger histopathologisch validierte Nicht-HCC-Läsionen verfügbar. Dementsprechend wurden Nicht-HCC-Läsionen häufig als Zufallsbefund im Rahmen einer Transplantation auf Grund von Leberversagen oder eines parallel bestehenden HCCs histopathologisch untersucht.

Auf Grund dieser begrenzten Verfügbarkeit wurden die Nicht-HCC Läsionen in eine einzelne Gruppe zusammengefasst. Metastatische Läsionen wurden ausgeschlossen, da bei diesen ein histopathologischer Befund häufig nicht zur Verfügung stand, was dem Umstand geschuldet war, dass sekundäre Malignitäten häufig nicht chirurgisch therapiert werden. Die histopathologische Validierung erfolgte im Rahmen von Biopsien, Resektionen, Transplantaten und Autopsien. Zudem war das Zeitintervall zwischen MRT und pathologischer Validierung relativ lang. Jedoch ist die Wahrscheinlichkeit der malignen Transformation eines definitiv benignen Befundes äußerst gering (36). Außerdem war das Zeitintervall in dieser Studie weniger relevant, da die Histopathologie nur genutzt wurde, um die korrekte Diagnose der Läsion zu erhalten. Auf Grund der begrenzten Anzahl an Läsionen wurde zudem eine große Anzahl an HCC-Läsionen mit Leberzirrhose eingeschlossen. Jedoch wurden alle Läsionen zugeschnitten und damit der Einfluss von Hintergrundgewebe auf die Bildanalyse reduziert. Als eine weitere Limitation erscheint zunächst der Gebrauch heterogener Bildgebungsquellen. Hierdurch wird jedoch die Robustheit des CNNs gegenüber verschiedenen MRT-Scannern bzw. Protokollen deutlich. Der Algorithmus erfasst keine Variabilität bezüglich Kontrastmittel-Arten, MRT Scan-Zeiten oder der Bildqualität. Diese Aspekte wären ein interessanter Ansatz für prospektive Studien. Zudem wäre in zukünftigen Studien eine Gegenüberstellung der diagnostischen Performance des CNNs vs. eines CNN-assistierten Radiologen bzw. die Performance des CNNs vs. eines nicht-assistierten Radiologen denkbar, um die klinische Anwendbarkeit zu prüfen. Bezüglich des Grading-Systems könnten mögliche Bias entstanden sein, da dieses durch eine einzelne Person (PMO) durchgeführt wurde. Diese wurden jedoch durch konstante Supervision minimiert.

Diese Studie ist nur ein kleiner Teil einer rapide zunehmenden Anzahl von Forschungsarbeiten, welche sich mit den Nutzen und Chancen von Machine Learning im Bereich der radiologischen Bildgebung befassen. Das finale Ziel soll dabei jedoch keineswegs der Ersatz der Radiolog*innen als Berufsgruppe sein, sondern die klinische Integration von sogenannten „Decision Support Tools“. Dabei handelt es sich um computergestützte Diagnosesysteme, die Radiolog*innen einen ersten Diagnosevorschlag übermitteln und damit die Entscheidung erleichtern könnten. Die häufig jahrzehntelange Erfahrung von Radiolog*innen ist von immenser Bedeutung, wenn es darum geht, Bildbefunde in komplexe klinischen Patientenbilder einzuordnen, und dabei irrelevante Informationen von relevanten unterscheiden zu können. Diese effektive jedoch gleichzeitig subjektive Art der Befundung könnte durch die Arbeit eines objektiven, quantifizierbaren sowie reproduzierbaren Decision Support Tools optimal ergänzt werden. So zeigte sich beispielsweise, dass die klinische Anwendung eines Decision Support Tools in der Befundung von Hirn-MRT-

Bildern die Befundungs-Perfomance insbesondere von weniger erfahrenen Radiolog*innen deutlich verbessert werden konnte (37) . Auch im Bereich der MRT-basierten Prostatadiagnostik konnte in einer Multi-Reader-Studie gezeigt werden, dass der Einsatz von Decision Support Tools zu einer verbesserten Interreader-Übereinstimmung führen kann (38,39).

Trotz dieser vielversprechenden Forschungsergebnisse bestehen jedoch weiterhin diverse Hürden, die es zu überwinden gilt, bevor eine klinische Anwendung von Decision Support Tools möglich sein wird. Eine wichtige Maßnahme wäre hierbei die Einführung eines einheitlichen Referenzstandards für die Erstellung von Datensets sowie feste Kriterien und Leitlinien für das methodische Vorgehen innerhalb von Machine-Learning-Studien. Zudem fehlen weiterhin große qualitativ hochwertige Datensätze aus vorzugsweise prospektiven multizentrischen Studien, um die Performance der Algorithmen zu optimieren und gleichzeitig die Akzeptanz in radiologischen Fachkreisen zu stärken. Dabei sollte auch die Angst vor nicht menschlich nachvollziehbaren Entscheidungen („Black-Box“) eines Deep Learning Systems adressiert werden. Hierbei gibt es bereits erste Ansätze, die zeigen, dass ein CNN menschlich definierte Bildkriterien erkennen und seine Diagnoseentscheidung durch die prozentuale Anteilhabe jedes Bildkriteriums begründen kann (29). Da es sich hierbei um eine Proof-of-Concept-Studie handelt, wird auch hier weitere Forschung notwendig sein. Sind jedoch all diese Hürden erst einmal überwunden, so werden zukünftig Deep Learning Modelle ihr volles Potential als schnelle und leistungsstarke Decision Support Tools im klinischen Alltag entfalten können.

Zusammenfassend, zeigt diese Studie den erfolgreichen Einsatz von Deep Learning bei der Klassifikation von sowohl typischen als auch atypischen histopathologisch validierten Leberläsionen mittels multiphasischen MRT-Bildern. Bisher basieren nur wenige Deep Learning Systeme auf sowohl radiologisch als auch pathologisch validierten Training-Sets. Indem ausschließlich histopathologisch bestätigte Läsionen eingeschlossen werden, kann die zugrunde liegende biologische Validität eines Deep Learning Systems optimiert und damit der Weg hin zu einer klinischen Integration von Decision Support Tools erleichtert werden. Des Weiteren ermöglicht dies die Bewertung von Läsionen mit atypischem Erscheinungsbild und erweitert damit die Grenzen der nicht-invasiven bildgebungsbasierten Diagnostik. Infolgedessen haben CNNs das Potential, den Bedarf an Biopsien sowie die damit verbundenen Komplikationen zu minimieren und die Patientenversorgung zu verbessern. Die kurze Rechenzeit unseres CNNs könnte die Eingliederung in den klinischen Alltag erleichtern.

3.6. Referenzliste

1. Oestmann PM, Wang CJ, Savic LJ, Hamm CA, Stark S, Schobert I, Gebauer B, Schlachter T, Lin M, Weinreb JC, Batra R, Mulligan D, Zhang X, Duncan JS, Chapiro J. Deep learning-assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. Eur Radiol. 2021 Jan 6;
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209–49.
3. Herold G. Innere Medizin 2021. Köln;
4. Llovet JM, Brú C, Bruix J. Prognosis of hepatocellular carcinoma: the BCLC staging classification. Semin Liver Dis. 1999;19(3):329–38.
5. Deutsche Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten e.V. (DGVS). S3-Leitlinie Hepatozelluläres Karzinom und biliäre Karzinome [Internet]. [cited 2021 Jul 8]. Available from: https://www.awmf.org/leitlinien/detail/ll/032-053OL.html
6. Trial Design and Endpoints in Hepatocellular Carcinoma: AASLD Consensus Conference. [cited 2021 Aug 9]; Available from: https://aasldpubs.onlinelibrary.wiley.com/doi/10.1002/hep.31327
7. Cho YK, Kim JK, Kim WT, Chung JW. Hepatic resection versus radiofrequency ablation for very early stage hepatocellular carcinoma: A Markov model analysis. Hepatology. 2010;51(4):1284–90.
8. Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul J-L, Schirmacher P, Vilgrain V. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. Journal of Hepatology. 2018 Jul 1;69(1):182–236.
9. Marrero JA, Kulik LM, Sirlin CB, Zhu AX, Finn RS, Abecassis MM, Roberts LR, Heimbach JK. Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. Hepatology. 2018;68(2):723–50.
10. Llovet JM, Montal R, Sia D, Finn RS. Molecular therapies and precision medicine for hepatocellular carcinoma. Nat Rev Clin Oncol. 2018 Oct;15(10):599–616.
11. Wald C, Russo MW, Heimbach JK, Hussain HK, Pomfret EA, Bruix J. New OPTN/UNOS Policy for Liver Transplant Allocation: Standardization of Liver Imaging, Diagnosis,

Classification, and Reporting of Hepatocellular Carcinoma. Radiology. 2013 Feb 1;266(2):376–82.

12. CT/MRI LI-RADS v2018 [Internet]. [cited 2020 Aug 31]. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS/CT-MRI-LI-RADS-v2018
13. Kang JH, Choi SH, Lee JS, Park SH, Kim KW, Kim SY, Lee SS, Byun JH. Interreader Agreement of Liver Imaging Reporting and Data System on MRI: A Systematic Review and Meta-Analysis. Journal of Magnetic Resonance Imaging. 2020;52(3):795–804.
14. Davenport MS, Khalatbari S, Liu PSC, Maturen KE, Kaza RK, Wasnik AP, Al-Hawary MM, Glazer DI, Stein EB, Patel J, Somashekar DK, Viglianti BL, Hussain HK. Repeatability of Diagnostic Features and Scoring Systems for Hepatocellular Carcinoma by Using MR Imaging. Radiology. 2014 Feb 18;272(1):132–42.
15. Smith EH. Complications of percutaneous abdominal fine-needle biopsy. Review. Radiology. 1991 Jan 1;178(1):253–8.
16. Seehofer D, Öllinger R, Denecke T, Schmelzle M, Andreou A, Schott E, Pratschke J. Blood Transfusions and Tumor Biopsy May Increase HCC Recurrence Rates after Liver Transplantation. J Transplant [Internet]. 2017 [cited 2018 Mar 5];2017. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5244021/
17. Quaia E, De Paoli L, Angileri R, Cabibbo B, Cova MA. Indeterminate solid hepatic lesions identified on non-diagnostic contrast-enhanced computed tomography: Assessment of the additional diagnostic value of contrast-enhanced ultrasound in the non-cirrhotic liver. European Journal of Radiology. 2014 Mar;83(3):456–62.
18. Pérez Saborido B, Menéu Díaz JC, Jiménez de los Galanes S, Loinaz Segurola C, Abradelo de Usera M, Donat Garrido M, Moreno Elola-Olaso A, Gómez Sánz R, Jiménez Romero C, García García I, Moreno González E. Does Preoperative Fine Needle Aspiration-Biopsy Produce Tumor Recurrence in Patients Following Liver Transplantation for Hepatocellular Carcinoma? Transplantation Proceedings. 2005 Nov 1;37(9):3874–7.
19. Machine Learning vs. Deep Learning: Wo ist der Unterschied? [Internet]. datasolut GmbH. 2021 [cited 2021 Jul 8]. Available from: https://datasolut.com/machine-learning-vs-deep-learning/
20. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A. Deep Learning: A Primer for Radiologists. RadioGraphics. 2017 Nov 1;37(7):2113–31.
21. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine Learning for Medical Imaging. Radiographics. 2017 Mar;37(2):505–15.

22. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights into Imaging [Internet]. 2018 Jun 22 [cited 2018 Aug 1]; Available from: http://link.springer.com/10.1007/s13244-018-0639-9
23. Fukushima K, Miyake S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. Pattern Recognition. 1982 Jan 1;15(6):455–69.
24. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol. 1962 Jan;160(1):106-154.2.
25. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P. Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head CT Scans. arXiv:180305854 [cs] [Internet]. 2018 Mar 13 [cited 2018 Jul 14]; Available from: http://arxiv.org/abs/1803.05854
26. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:171105225 [cs, stat] [Internet]. 2017 Nov 14 [cited 2018 Jul 14]; Available from: http://arxiv.org/abs/1711.05225
27. Kloppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, Mader I, Mitchell LA, Patel AC, Roberts CC, Fox NC, Jack CR, Ashburner J, Frackowiak RSJ. Accuracy of dementia diagnosis--a direct comparison between radiologists and a computerized method. Brain. 2008 Jun 21;131(11):2969–74.
28. Greenspan H, van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging. 2016 May;35(5):1153–9.
29. Hamm CA, Wang CJ, Savic LJ, Ferrante M, Schobert I, Schlachter T, Lin M, Duncan JS, Weinreb JC, Chapiro J, Letzen B. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol [Internet]. 2019 Apr 23 [cited 2019 May 9]; Available from: https://doi.org/10.1007/s00330-019-06205-9
30. Wang CJ, Hamm CA, Savic LJ, Ferrante M, Schobert I, Schlachter T, Lin M, Weinreb JC, Duncan JS, Chapiro J, Letzen B. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol. 2019 Jul 1;29(7):3348–57.
31. Yasaka K, Akai H, Abe O, Kiryu S. Deep Learning with Convolutional Neural Network for Differentiation of Liver Masses at Dynamic Contrast-enhanced CT: A Preliminary Study. Radiology. 2018 Mar;286(3):887–96.

32. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM. 2017 May 24;60(6):84–90.
33. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statist Surv. 2010;4:40–79.
34. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. arXiv:151203385 [cs] [Internet]. 2015 Dec 10 [cited 2020 Aug 31]; Available from: http://arxiv.org/abs/1512.03385
35. Huang G, Liu Z, Maaten LVD, Weinberger KQ. Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Jul;2261–9.
36. Fodor M, Primavesi F, Braunwarth E, Cardini B, Resch T, Bale R, Putzer D, Henninger B, Oberhuber R, Maglione M, Margreiter C, Schneeberger S, Öfner D, Stättner S. Indications for liver surgery in benign tumours. Eur Surg. 2018;50(3):125–31.
37. Rudie JD, Duda J, Duong MT, Chen P-H, Xie L, Kurtz R, Ware JB, Choi J, Mattay RR, Botzolakis EJ, Gee JC, Bryan RN, Cook TS, Mohan S, Nasrallah IM, Rauschecker AM. Brain MRI Deep Learning and Bayesian Inference System Augments Radiology Resident Performance. J Digit Imaging [Internet]. 2021 Jun 15 [cited 2021 Aug 16]; Available from: https://doi.org/10.1007/s10278-021-00470-1
38. Hamm CA, Beetz NL, Savic LJ, Penzkofer T. Künstliche Intelligenz und Radiomics in der MRT-basierten Prostatadiagnostik. Radiologe. 2020 Jan 1;60(1):48–55.
39. Greer MD, Lay N, Shih JH, Barrett T, Bittencourt LK, Borofsky S, Kabakus I, Law YM, Marko J, Shebel H, Mertan FV, Merino MJ, Wood BJ, Pinto PA, Summers RM, Choyke PL, Turkbey B. Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study. Eur Radiol. 2018 Oct;28(10):4407–17.

4. Eidesstattliche Versicherung

„Ich, Paula Marie Oestmann, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Deep Learning-gestützte Klassifizierung histologisch gesicherter Lebertumoren in der Kontrastmittel-verstärkten MRT-Bildgebung/ Deep learning–assisted differentiation of pathologically proven liver tumors on contrast-enhanced MRI“ selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet. Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem/der Erstbetreuer/in, angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe. Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Düsseldorf, den 7.9.21

5. Anteilserklärung

Paula Marie Oestmann hatte Anteil an der folgenden Publikation:

Paula M. Oestmann, Clinton J. Wang, Lynn J. Savic, Charlie A. Hamm, Sophie Stark, Isabel Schobert, Bernhard Gebauer, Todd Schlachter, MingDe Lin, Jeffrey C. Weinreb, Ramesh Batra, David Mulligan, Xuchen Zhang, James S. Duncan, Julius Chapiro
Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver European Radiology, Volume 31, Ausgabe 7, Juli 2021 (S. 4981 – 4990)

- Eigener Beitrag:
    - Beteiligung bei der Erarbeitung des Studiendesigns und der Arbeitshypothese
    - Wöchentliche Präsentationen der vorläufigen Daten bei Laborbesprechungen
    - Eigenständige Erhebung der Primärdaten: Patientenselektion; Patientenrekrutierung (zusammen mit Study Nurse); Auswertung aller Labordaten, Patientendaten und anamnestischer Angaben; Aufbau und Aktualisierung der kompletten Datenbank
    - Auswertung aller MRT- und CT-Bilder unter Supervision
    - Mitwirkung bei der Auswertung der histopathologischen Daten
    - Durchführung der statistischen Tests in beratender Zusammenarbeit mit einer Statistikerin
    - Auswahl geeigneter Visualisierungsmethoden; eigenständige Erstellung aller Tabellen und Grafiken
    - Eigenständige Literaturrecherche und -auswahl, Erstellung und Korrektur des Manuskriptes
    - Einreichung bei Journal und anschließende Revision des Manuskriptes nach Peer-Review
    - Präsentation bei wissenschaftlicher Konferenz (RSNA 2018)

_____

Unterschrift des Doktoranden/der Doktorandin

6. Auszug aus Journal Summary List

Journal Data Filtered By: **Selected JCR Year: 2019** Selected Editions: SCIE,SSCI Selected Categories: **"RADIOLOGY, NUCLEAR MEDICINE and MEDICAL IMAGING"** Selected Category Scheme: WoS
Gesamtanzahl: 133 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	JACC-Cardiovascular Imaging	10,110	12.740	0.027550
2	MEDICAL IMAGE ANALYSIS	9,028	11.148	0.017100
3	RADIOLOGY	52,731	7.931	0.057130
4	JOURNAL OF NUCLEAR MEDICINE	26,844	7.887	0.032990
5	EUROPEAN JOURNAL OF NUCLEAR MEDICINE AND MOLECULAR IMAGING	15,787	7.081	0.023630
6	IEEE TRANSACTIONS ON MEDICAL IMAGING	21,657	6.685	0.030060
7	CLINICAL NUCLEAR MEDICINE	5,042	6.587	0.006200
8	NEUROIMAGE	102,632	5.902	0.125360
9	Photoacoustics	715	5.870	0.001760
10	INTERNATIONAL JOURNAL OF RADIATION ONCOLOGY BIOLOGY PHYSICS	44,197	5.859	0.042160
11	Circulation-Cardiovascular Imaging	5,574	5.691	0.016320
12	ULTRASOUND IN OBSTETRICS & GYNECOLOGY	13,078	5.571	0.018050
13	JOURNAL OF CARDIOVASCULAR MAGNETIC RESONANCE	5,205	5.361	0.011120
14	INVESTIGATIVE RADIOLOGY	6,136	5.156	0.008830
15	RADIOGRAPHICS	12,418	4.967	0.010750
16	ULTRASCHALL IN DER MEDIZIN	2,185	4.966	0.002530
17	RADIOTHERAPY AND ONCOLOGY	17,774	4.856	0.026510
18	European Heart Journal-Cardiovascular Imaging	6,359	4.841	0.023110
19	HUMAN BRAIN MAPPING	23,094	4.421	0.042760
20	Journal of the American College of Radiology	4,409	4.268	0.010730

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
21	EUROPEAN RADIOLOGY	20,761	4.101	0.033260
22	SEMINARS IN RADIATION ONCOLOGY	2,531	4.076	0.003540
23	JOURNAL OF MAGNETIC RESONANCE IMAGING	17,046	3.954	0.024900
24	Biomedical Optics Express	11,090	3.921	0.025030
25	COMPUTERIZED MEDICAL IMAGING AND GRAPHICS	2,656	3.750	0.002940
26	JOURNAL OF DIGITAL IMAGING	2,494	3.697	0.003790
27	MAGNETIC RESONANCE IN MEDICINE	32,159	3.635	0.029700
28	Insights into Imaging	1,948	3.579	0.003260
29	INTERNATIONAL JOURNAL OF HYPERTHERMIA	4,397	3.574	0.004880
30	SEMINARS IN NUCLEAR MEDICINE	2,194	3.544	0.002420
31	AMERICAN JOURNAL OF NEURORADIOLOGY	23,135	3.381	0.027120
32	JOURNAL OF NUCLEAR CARDIOLOGY	3,600	3.366	0.004570
33	MEDICAL PHYSICS	26,445	3.317	0.027280
34	Quantitative Imaging in Medicine and Surgery	1,335	3.226	0.002800
35	NMR IN BIOMEDICINE	7,537	3.221	0.011610
36	Clinical Neuroradiology	935	3.183	0.002710
37	KOREAN JOURNAL OF RADIOLOGY	2,967	3.179	0.004490
38	Ultrasonography	618	3.075	0.001710
39	ULTRASONICS	7,808	3.065	0.008930
40	JOURNAL OF VASCULAR AND INTERVENTIONAL RADIOLOGY	9,045	3.037	0.009790
41	AMERICAN JOURNAL OF ROENTGENOLOGY	32,209	3.013	0.024770
42	Practical Radiation Oncology	1,879	2.948	0.005780

Selected JCR Year: 2019; Selected Categories: "RADIOLOGY, NUCLEAR MEDICINE and MEDICAL IMAGING"

7. Publikation

Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver

European Radiology
https://doi.org/10.1007/s00330-020-07559-1

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver

Paula M. Oestmann[1,2,3] · Clinton J. Wang[1,4] · Lynn J. Savic[1,2] · Charlie A. Hamm[1,2] · Sophie Stark[1,2,5] · Isabel Schobert[1,2] · Bernhard Gebauer[2] · Todd Schlachter[1] · MingDe Lin[1] · Jeffrey C. Weinreb[1] · Ramesh Batra[6] · David Mulligan[6] · Xuchen Zhang[7] · James S. Duncan[1,4] · Julius Chapiro[1]

Received: 30 September 2020 / Revised: 6 November 2020 / Accepted: 23 November 2020

Abstract

Objectives To train a deep learning model to differentiate between pathologically proven hepatocellular carcinoma (HCC) and non-HCC lesions including lesions with atypical imaging features on MRI.

Methods This IRB-approved retrospective study included 118 patients with 150 lesions (93 (62%) HCC and 57 (38%) non-HCC) pathologically confirmed through biopsies ($n = 72$), resections ($n = 29$), liver transplants ($n = 46$), and autopsies ($n = 3$). Forty-seven percent of HCC lesions showed atypical imaging features (not meeting Liver Imaging Reporting and Data System [LI-RADS] criteria for definitive HCC/LR5). A 3D convolutional neural network (CNN) was trained on 140 lesions and tested for its ability to classify the 10 remaining lesions (5 HCC/5 non-HCC). Performance of the model was averaged over 150 runs with random sub-sampling to provide class-balanced test sets. A lesion grading system was developed to demonstrate the similarity between atypical HCC and non-HCC lesions prone to misclassification by the CNN.

Results The CNN demonstrated an overall accuracy of 87.3%. Sensitivities/specificities for HCC and non-HCC lesions were 92.7%/82.0% and 82.0%/92.7%, respectively. The area under the receiver operating curve was 0.912. CNN's performance was correlated with the lesion grading system, becoming less accurate the more atypical imaging features the lesions showed.

Conclusion This study provides proof-of-concept for CNN-based classification of both typical- and atypical-appearing HCC lesions on multi-phasic MRI, utilizing pathologically confirmed lesions as "ground truth."

Key Points

- *A CNN trained on atypical appearing pathologically proven HCC lesions not meeting LI-RADS criteria for definitive HCC (LR5) can correctly differentiate HCC lesions from other liver malignancies, potentially expanding the role of image-based diagnosis in primary liver cancer with atypical features.*
- *The trained CNN demonstrated an overall accuracy of 87.3% and a computational time of < 3 ms which paves the way for clinical application as a decision support instrument.*

Keywords Carcinoma, hepatocellular · Liver neoplasms · Deep learning · Magnetic resonance imaging · Neural networks, computer

✉ Julius Chapiro
julius.chapiro@yale.edu

[1] Department of Radiology and Biomedical Imaging, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06520, USA

[2] Institute of Radiology, Berlin Institute of Health, Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt-Universität, 10117 Berlin, Germany

[3] Faculty of Medicine, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[4] Department of Biomedical Engineering, Yale School of Engineering and Applied Science, New Haven, CT 06520, USA

[5] Faculty of Medicine, Albert-Ludwigs-University Freiburg, Freiburg, Germany

[6] Department of Transplantation and Immunology, 333 Cedar Street, New Haven, CT 06520, USA

[7] Department of Pathology, Yale School of Medicine, 310 Cedar Street, New Haven, CT 06520, USA

Published online: 06 January 2021

Abbreviations

AUC	Area under the curve
CNN	Convolutional neural network
FNH	Focal nodular hyperplasia
HCC	Hepatocellular carcinoma
HIPAA	Health Insurance Portability and Accountability Act
ICC	Intrahepatic cholangiocarcinoma
LI-RADS	Liver Imaging Reporting and Data System
MELD	Model for End-Stage Liver Disease
NPV	Negative predictive value
PACS	Picture archiving and communication system
PPV	Positive predictive value

Introduction

Hepatocellular carcinoma (HCC), the fourth most common cause of malignancy-related death worldwide, represents the most frequent primary liver cancer and its incidence rates continue to rise [1]. Other liver lesions to be differentiated on diagnostic imaging include intrahepatic cholangiocarcinoma (ICC), metastases, and various types of benign lesions. Contrast-enhanced multi-phasic computed tomography (CT) and magnetic resonance imaging (MRI) play a central role for diagnosis and classification of these lesions. Standardized imaging features of HCC summarized in Organ Procurement and Transplantation Network (OPTN) or Liver Imaging Reporting and Data System (LI-RADS) criteria provide the framework for clinical diagnostic workup [2, 3]. In lesions not meeting typical imaging criteria, the diagnosis can be challenging. High inter-reader variability depending on the radiologist's experience may lead to unnecessary tissue biopsies [4] prone to complications such as hemorrhage, sepsis, carcinoid crisis [5], or tumor seeding [6]. These may compromise orthotopic liver transplantation which is the only established curative therapy for HCC [7, 8].

In recent years, deep learning has gained considerable traction in the field of medical image analysis. The most common tool to classify lesions on radiologic imaging is the convolutional neural network (CNN) [9]. Unlike other machine learning methods, CNNs do not require definition of specific radiological features to learn how to interpret images. After being shown imaging examples with and without the disease, the CNN automatically learns features through backpropagation using multiple layers [10].

Recently, several studies used CNNs on CT/MRI focusing on liver lesions with typical appearances, allowing for distinctive image-based diagnosis according to the standardized criteria [11–13]. However, in order to be used in clinical management, CNNs should also correctly diagnose lesions that do not fit into established classification systems. As the number of heterogeneous input samples grows, CNNs have the potential to recognize atypical lesions, thus reducing the need for biopsies and subsequent post-biopsy complications.

The aim of this study was to prove the capability of CNNs to handle a wider spectrum of HCC and non-HCC lesions on multi-phasic contrast-enhanced MRI, using pathologically proven liver lesions as the "ground truth."

Materials and methods

This retrospective, single-center study was approved by the Institutional Review Board and Health Insurance Portability and Accountability Act (HIPAA). It was conducted according to the Standards for Report of Diagnostic Accuracy guidelines. Informed consent was waived.

Study cohort selection

HCC and non-HCC lesions from patients older than 18 years diagnosed between 2010 and 2018 were identified using a picture archiving and communication system (PACS) as well as the electronic medical record. Only patients with histopathological diagnosis were included. Pathological proof was established for all through biopsies ($n = 72$), resections ($n = 29$), liver transplants ($n = 46$), and autopsies ($n = 3$). In case of transplants/autopsies, the liver was subject to gross pathological/histopathological analysis including full histological assessment of the HCC lesion. H&E staining was used to assess lesions and additional histopathological surface markers were applied. Lesions indicated in pathology reports were identified by a radiology trainee supervised by a board-certified radiologist sub-specialized in abdominal imaging with approximately 25 years of experience in body imaging. The lesions were qualified regarding size and intrahepatic localization. A multi-phasic T1-weighted MRI dataset including contrast-enhanced late arterial, portal venous, and delayed/equilibrium phases had to be present to meet inclusion criteria. Clear correspondence between pathology and imaging was achieved collaboratively with a pathologist and side-by-side review of location for each tumor. If more than one lesion was visible on MRI in the segment described by the pathologist, images of CT-guided biopsy were used to ascertain the biopsied lesion. If these were unavailable, all lesions in the segment were excluded. Lesions that were biopsied before the MRI scan were excluded if procedure-related hemorrhage was leading to significant alteration of T1 signal. Up to 4 lesions per patient were used. In the non-HCC class, only primary liver neoplasms were included. HCC lesions with loco-regional therapy performed between MR imaging and resection/transplantation were included only if residual viable tumor was present on histology that would allow confirmation of etiology. Tumors with complete necrosis were excluded.

MRI acquisition protocol

MRI examinations were conducted on 1.5-T or 3-T MRI scanners including Signa Excite®, GE Discovery®, Siemens Aera®, Espree®, Verio®, Avanto®, Skyra®, and Trio Tim® scanners. Non-contrast T1 images were acquired in all patients prior to administration of intravenous contrast. After the administration of intravenous gadolinium-based contrast agent (including Gadavist® (Bayer), Dotarem® (Guerbet), Magnevist® (Bayer), ProHance® (Bracco Diagnostics), and Optimark® (Covidien), dosed at 0.1 mmol/kg), three T1-weighted three-dimensional (3D) gradient-echo (GRE) breath-hold imaging series (acquisition times of 12–18 s, with fat suppression) were acquired reflecting CT/MRI LI-RADS recommendations: (1) late arterial, (2) portal venous, and (3) delayed or equilibrium phase. Bolus tracking was applied in a large proportion of patients. Imaging parameters were in the range of TR 3–5 ms, TE 1–2 ms, flip angle 9–13°, bandwidth 300–500 Hz, slice thickness 3–4 mm, image matrix 256 × 132 to 320 × 216, and field of view 300 × 200 to 500 × 400 mm. If a patient received multiple MRI scans, then the MRI performed closest to the date of pathological confirmation was used.

Image processing

After MR imaging studies were retrieved from an institutional database, the x, y, and z coordinates of each lesion were manually recorded to define a 3D bounding box around the lesion (Fig. 1). Only the image volume within this bounding box was analyzed by the model. Images were processed using code written in Python 3.5 (Python Software Foundation). Affine registration with a mutual information metric was used to register portal venous and delayed phase MRI sequences to the late arterial phase. The images were cropped to the bounding box defined above and normalized to an intensity range of − 1 to 1 to reduce bias field effects. The images were further resampled to 36 × 36 × 12 voxels.

To increase the number of training samples, the training set was augmented by a factor of 100 (n = 14,000) in standard fashion (Fig. 2). Briefly, images were randomly rotated, shifted, scaled, flipped, shifted between phases, and scaled or shifted in intensity. This allows for the model to learn imaging features that are invariant to rotation or translation [14].

Neural network architecture

The model was trained on a GeForce GTX 1060 (NVIDIA) graphics processing unit. It was built using Python 3.5 and Keras 2.2 (https://keras.io/) on a Tensorflow backend (Google, https://www.tensorflow.org/). The CNN consisted of three convolutional layers (64, 128, and 128 channels, respectively; kernel size 3 × 3 × 2), two maximum pooling layers (size 2 × 2 × 2 and 2 × 2 × 1, respectively), and two fully connected layers (100 and 1 neurons, respectively), with a sigmoid output corresponding to the probability of a lesion being HCC. The CNN used rectified linear units, batch normalization, and 10% dropout.

Training and evaluation

The CNN was trained on 70 HCC examples and 70 non-HCC examples, drawn randomly from the augmented dataset. An Adam optimizer was used with a minibatch size of 20 and learning rate of 0.01. The model was tested on its ability to correctly classify ten lesions in the test dataset, which was created by randomly selecting 5 HCC lesions and 5 non-HCC lesions. In total, 150 independent runs with different splits of training and test datasets (i.e., Monte Carlo cross-validation rather than k-fold cross-validation in order to balance HCC/non-HCC cases within each set) were used to estimate the model's performance. This approach in conjunction with a 14:1 training:test ratio is consistent with machine learning best practice [15, 16].

Lesion grading

As the dataset contained lesions with atypical appearances on MRI, a lesion grading system was developed based on the established LI-RADS major imaging criteria [17] using imaging features typical of HCC: arterial hyperenhancement, washout, and enhancing rim/pseudocapsule (Fig. 3). A supervised radiology trainee credited lesions 1 point for every applicable imaging feature so that a lesion could be graded on a scale of 0 to 3 points. According to this grading system, both HCC and non-HCC lesions were staged to demonstrate the similarity between HCC and non-HCC lesions prone to misclassification by the CNN. On the one hand, lesions receiving 3 points could either be typical LI-RADS-applicable HCC or pathologically proven non-HCC lesions that presented like HCC on imaging. On the other hand, HCC lesions graded with 1 point showed atypical contrast dynamics with only one of these features. The differences of the grading scores between the well (> 90% accuracy) and poorly (< 90% accuracy) classified lesions were analyzed to provide possible explanations for misclassifications of lesions by the CNN.

Statistics

Sensitivity, specificity, and overall accuracy were calculated in order to validate the performance of the deep learning model. These metrics were averaged over 150 runs with random sub-sampling to yield class-balanced test sets. The receiver operating characteristic curve was obtained and the area under the curve (AUC) was calculated (Fig. 4).

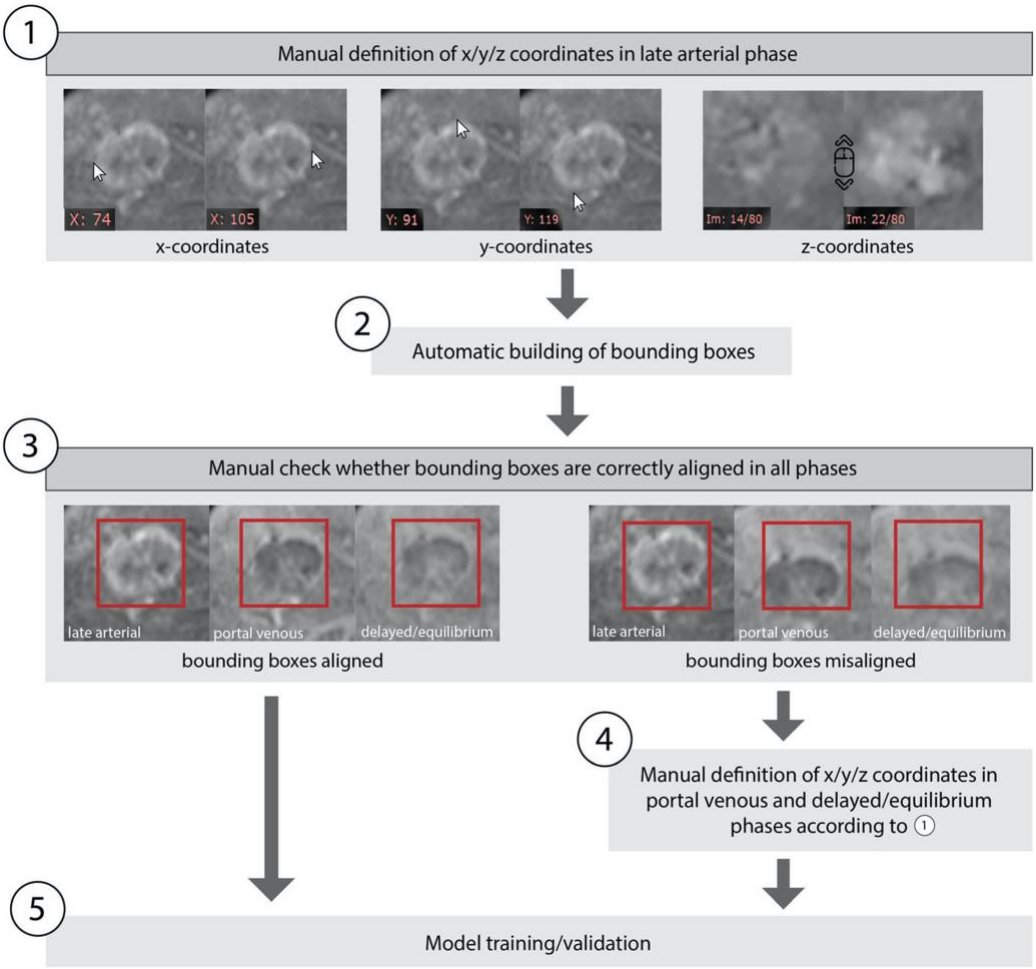

Fig. 1 Determination of coordinates and bounding boxes. (1) All coordinates were determined manually in the late arterial phase using a DICOM viewer (Radiant®). The maximum extent of each lesion within an axis was determined using 2 coordinates. (2) Bounding boxes were automatically built according to the defined coordinates. (3) Bounding boxes were checked manually to ensure that they are aligned correctly in all phases. (4) In the few cases where bounding boxes were misaligned due to breathing motion artifact, coordinates were manually specified separately for the portal venous and delayed/equilibrium phases according to step 1. (5) After all bounding boxes were correctly aligned, model training/validation was conducted according to Fig. 2

Results

Study population

This study included 118 patients with HCC ($n = 73$, 62%) and non-HCC lesions ($n = 45$, 38%). The HCC cohort contained 57 (78%) men and 16 (22%) women, whereas 23 (51%) men and 22 (49%) women were included in the non-HCC cohort. The mean age of the HCC patients was 61 ± 8 (mean, standard deviation), and the mean age of the non-HCC patients was 59 ± 13 years. The cohort contained 87 patients with cirrhosis, including 73 (84%) in the HCC class and 14 (16%) in the non-HCC class. The majority of these patients were classified as Child-Turcotte-Pugh-Score A ($n = 50$, 57%) and the most common etiology was hepatitis C infection ($n = 61$, 59%). The median Model for End-Stage Liver Disease (MELD) score for all patients was 9. The exact values can be obtained in Table 1.

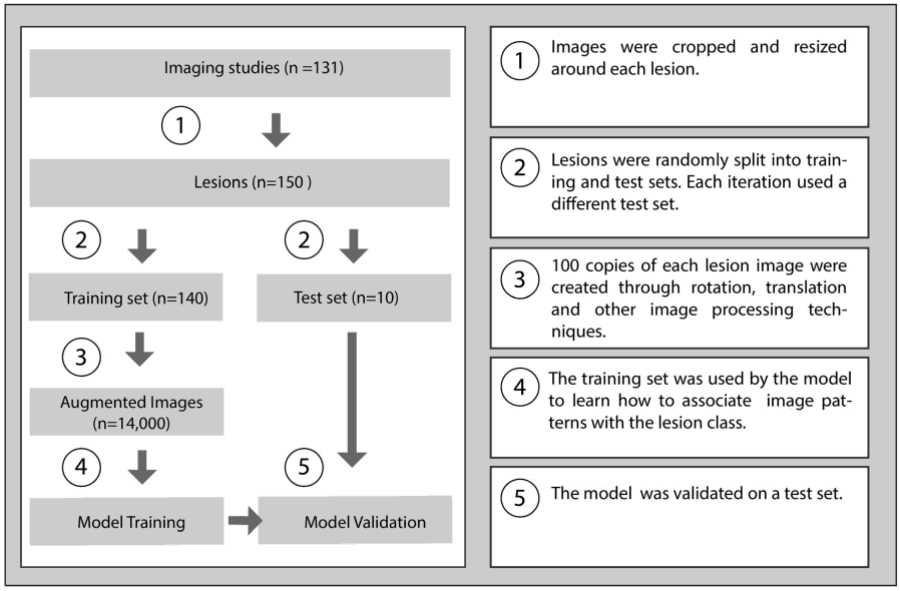

Fig. 2 Flowchart of the lesion classification approach, including model training and testing

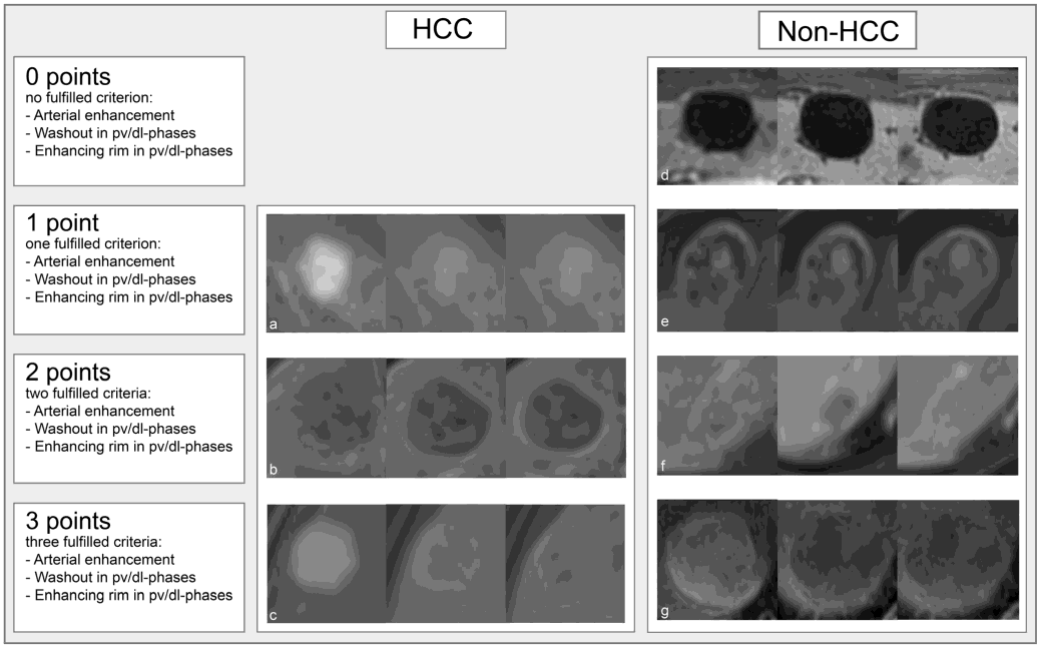

Fig. 3 HCC as well as non-HCC lesions were graded with 0 to 3 points in order to demonstrate the similarity between HCC and non-HCC lesions prone to misclassification of lesions by the CNN. HCC hepatocellular carcinoma, pv/dl portal venous/delayed. (**a**) HCC with arterial enhancement, (**b**) HCC with washout and enhancing rim, (**c**) HCC with arterial enhancement, washout, and enhancing rim, (**d**) cyst with no fulfilled criterion, (**e**) hemangioma with enhancing rim, (**f**) hemangioma with enhancing rim and washout, and (**g**) cyst with arterial enhancement, washout, and enhancing rim

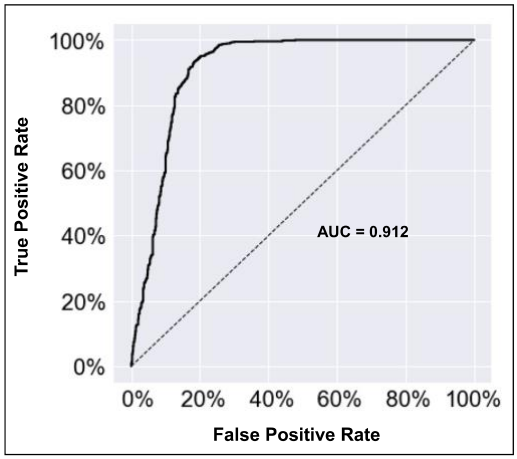

Fig. 4 Model receiver operating characteristic curve for distinguishing hepatocellular carcinoma (HCC) from non-HCC lesions. AUC area under the curve

A total of 93 (62%) HCC lesions and 57 (38%) non-HCC lesions were analyzed. The non-HCC group consisted of 19 (33%) ICCs, 16 (28%) hemangiomas, 15 (26%) cysts, 2 (4%) regenerative nodules, 2 (4%) dysplastic nodules, 2 (4%) FNHs, and 1 (2%) bile duct adenoma.

The median diameter for all lesions was 2.3 cm. The median timeframe between the MRI study and pathological proof was 1.6 months (range, 0–25 months) for HCC lesions if imaging was obtained prior to the pathological confirmation. Imaging after pathological confirmation was performed within 1 day. For non-HCC lesions, the median time between the MRI study and pathological confirmation was 1.4 months (range, 0–73 months), if imaging was obtained prior to the pathological confirmation. Imaging after pathological confirmation was performed within a median time of 5.5 months (0–24 months) (Table 2). One to four lesions per patient (median = 1) and one to three lesions per imaging set (median = 1) were included (Table 3).

Deep learning model performance

The deep learning model demonstrated a training accuracy of 94.1% ± 2.0 (19,766/21,000 volumetric samples). The performance was validated on a test set after 30 iterations, where the CNN demonstrated an overall accuracy of 87.3% ± 10.5 (1310/1500). The sensitivity to classify HCC and the non-HCC class was 92.7% and 82.0%, respectively, and the specificity for HCC and the non-HCC class was 82.0% and 92.7%, respectively (Table 4). The receiver operating characteristic curve demonstrated an AUC of 0.912 (Fig. 4). The CNN was trained in 3.2 min ± 0.9, and the computing time to classify each lesion in the test dataset was 2.9 ms ± 1.7.

Evaluation of lesion grading

According to the grading system, 23 (25%) of the HCC lesions were scored with 1 point, 28 (30%) with 2 points, and 42 (45%) with 3 points (Fig. 5). In the non-HCC class, 16 (28%) lesions were scored with 0, 24 (42%) with 1, 11 (19%) with 2, and 6 (11%) with 3 points. The Kruskal-Wallis test showed a significant positive correlation of the grading score with improved classification accuracy in HCC lesions ($p = 0.012$) and reduced classification accuracy in non-HCCs ($p < 0.001$). Specifically, in the HCC class, 1 of 42 (2%) lesions graded with 3 points, 4 of 28 (14%) lesions with 2 points, and 5 of 23 (22%) lesions graded with 1 point were poorly classified (≤ 90% accuracy in 150 runs) by the CNN. The one poorly classified 3-point HCC lesion as well as 3 of 4 poorly classified 2-point HCC lesions showed poor image quality. Moreover, 2 of the 4 poorly classified 2-point HCC lesions were in close proximity to the liver margin. In the non-HCC class, none of the lesions with 0 point, 2 of 24 (8%) lesions graded with 1 point, 3 of 11 (27%) lesions with 2 points, and 6 of 6 (100%) lesions graded with 3 points (100%) (6/6) were poorly classified.

Discussion

This study establishes a histopathologically validated deep learning approach capable of differentiating between HCC and non-HCC lesions on multi-phasic contrast-enhanced MRI. The model achieved an overall accuracy of 87.3%, with high sensitivity (92.7%) and moderate specificity (82.0%) for HCC. The CNN's short computation time could allow for practical integration into a radiologist's workflow without producing delays.

A few recent studies have focused on classifying different types of liver lesions using a deep learning approach. A previous study [11] utilized a CNN trained to differentiate between six different types of liver lesions with an overall accuracy of approximately 90%. This proof-of-concept study only used lesions with typical imaging features. However, inclusion of atypical lesions may provide a more representative dataset and increased translatability to clinical practice. Another study investigating deep learning–based liver tumor classification also included atypical/indeterminate lesions. However, all indeterminate lesions were grouped into one class without further sub-classification [13]. The CNN developed in the current study was trained on a majority of atypical lesions to further classify those lesions as HCC or non-HCC as verified by pathology. This binary differentiation is a significant step towards classifying indeterminant lesions non-invasively in clinical practice. The decision HCC versus non-HCC is particularly important since HCC is a malignant disease which can be treated curatively if diagnosed early.

Table 1 Patient characteristics. The numerical data are summarized as mean ± standard deviation or median (*) and the categorical data are shown as frequency (percentage). *HCC* hepatocellular carcinoma, *ICC* intrahepatic cholangiocarcinoma, *FNH* focal nodular hyperplasia, *MELD* Model for End-Stage Liver Disease, *Child-Pugh* Child-Turcotte-Pugh Score, *NASH* non-alcoholic fatty liver disease, *PSC* primary sclerosing cholangitis, *ECOG* Eastern Cooperative Oncology Group, *BCLC* Barcelona Clinic Liver Cancer, *HKLC* Hong Kong Liver Cancer classification system

	HCC	Non-HCC						
		ICC	Regenerative nodule	Dysplastic nodule	Hemangioma	Cyst	FNH	Bile duct adenoma
Number of patients	73	12	2	2	16	10	2	1
Gender								
Male	57 (78)	9 (75)	1 (50)	1 (50)	7 (44)	3 (30)	1 (50)	1 (100)
Female	16 (22)	3 (25)	1 (50)	1 (50)	9 (56)	7 (70)	1 (50)	0 (0)
Age at imaging	61 ± 8	69 ± 13	37*	61*	57 ± 10	56 ± 9	42*	53*
Ethnic								
Caucasian	53 (73)	9 (75)	1 (50)	2 (100)	11 (69)	8 (80)	1 (50)	1 (100)
African American	9 (12)	2 (17)	0 (0)	0 (0)	2 (13)	0 (0)	1 (50)	0 (0)
Asian	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Other	10 (14)	1 (8)	1 (50)	0 (0)	3 (19)	2 (20)	0 (0)	0 (0)
MELD	10*	13 ± 6	20*	10*	8 ± 2	6*	10*	10*
Cirrhosis	73	1	2	2	6	2	0	1
Child-Pugh								
A	44 (60)	0 (0)	0 (0)	1 (50)	4 (67)	0 (0)	0 (0)	1 (100)
B	26 (36)	1 (100)	1 (50)	1 (50)	2 (33)	2 (100)	0 (0)	0 (0)
C	3 (4)	0 (0)	1 (50)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Cause								
Hepatitis B	2 (3)	0 (0)	1 (50)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hepatitis C	53 (62)	1 (50)	0 (0)	2 (100)	2 (33)	2 (100)	0 (0)	1 (100)
Alcohol	21 (25)	0 (0)	0 (0)	1 (50)	2 (33)	1 (50)	1 (100)	0 (0)
NASH	8 (9)	0 (0)	1 (50)	0 (0)	1 (17)	0 (0)	0 (0)	1 (100)
PSC	1(1)	1 (50)	0 (0)	0 (0)	1 (17)	0 (0)	0 (0)	0 (0)
Malignancy related								
ECOG								
0	55 (75)	3 (25)						
1	16 (22)	4 (33)						
2	1 (1)	2 (17)						
3	1 (1)	1 (8)						
Unknown	0 (0)	2 (17)						
Extrahepatic spread	1 (14)	0 (0)						
HCC related								
BCLC								
0	12 (16)							
A	45 (62)							
B	0 (0)							
C	13 (18)							
D	3 (4)							
HKLC								
1	43 (58)							
2	26 (36)							
3	1 (1)							
4	0 (0)							
5	3 (4)							

Moreover, the aforementioned study was based on CT whereas the current study utilized MRI, providing a wider variety of imaging features for the CNN to capture. In the present study, 47% of HCC lesions did not meet LI-RADS criteria for definitive HCC (LR5) and 48% of all lesions were biopsied, generally suggesting indeterminate appearance on imaging. A

Table 2 Lesion characteristics. The numerical data are summarized as mean ± standard deviation or median (*) and the categorical data are shown as frequency (percentage). *HCC* hepatocellular carcinoma, *ICC* intrahepatic cholangiocarcinoma, *FNH* focal nodular hyperplasia, *TACE* transcatheter arterial chemoembolization, *MWA* microwave ablation, *RFA* radiofrequency ablation

	HCC	Non-HCC						
		ICC	Regenerative nodule	Dysplastic nodule	Hemangioma	Cyst	FNH	Bile duct adenoma
Number of lesions	93	19	2	2	16	15	2	1
Pathological proof								
Biopsy	47 (50)	15 (79)	1 (50)	1 (50)	6 (37)	0 (0)	2 (100)	0 (0)
Resection	10 (11)	4 (21)	0 (0)	0 (0)	5 (31)	10 (67)	0 (0)	0 (0)
Explant	36 (39)	0 (0)	1 (50)	1 (50)	3 (19)	4 (27)	0 (0)	1 (100)
Autopsy	0 (0)	0 (0)	0 (0)	0 (0)	2 (13)	1 (7)	0 (0)	0 (0)
Cirrhosis	93	1	2	2	6	4	0	1
Timeframe in days (median)								
Imaging pre path	49	22	42	68	104	181	509	27
Imaging post path	1	295	0	0	143	0	0	0
Diameter in cm	2,0*	4.2 ± 1.4	3,7*	1.1*	5.0 ± 4.0	4.9 ± 3.5	4,46*	1.4*
Residual tumor	8	0						
Treatments	29	0						
TACE	22 (76)							
Bland embolization	3 (10)							
Ethanol ablation	2 (7)							
MWA	6 (21)							
RFA	3 (10)							
LI-RADS								
LR5	49 (53)							
< LR5	44 (47)							

grading system was used to evaluate the representation of atypical-appearing lesions, assigning 1 point for each classical imaging feature of HCC (arterial hyperenhancement, washout, and pseudocapsule). According to this grading system, 25% of the HCC lesions scored 1 point because of their atypical appearances, and 30% of non-HCC lesions scored 2 or more points, mimicking typical appearances of HCC lesions. While the present study showed a slightly lower overall accuracy than the previous study with classical-appearing lesions, the results suggest that a CNN model trained with pathologically proven atypical lesions can still provide relatively high accuracy.

Classical-appearing lesions generally demonstrated higher classification accuracy. The lower specificity of HCC classification is likely related to non-HCC lesions displaying features of HCC on imaging. However, a small number of HCC lesions graded with 2 and 3 points were poorly classified, possibly caused by poor image quality or lesions in close proximity to the liver margin. The seemingly high standard deviation is a consequence of the number of validation images in each fold. Vanilla CNNs were considered appropriate for the small cropped 3D images in our study, as sophisticated architectures such as ResNet [18] and DenseNet [19] are designed for larger datasets and 2D high-resolution images.

This study has several limitations. A relatively small cohort was used due to the single-center nature and the requirement for histopathological reference standard.

Table 3 Image characteristics. *HCC* hepatocellular carcinoma, *ICC* intrahepatic cholangiocarcinoma, *FNH* focal nodular hyperplasia

	HCC	Non-HCC						
		ICC	Regenerative nodule	Dysplastic nodule	Hemangioma	Cyst	FNH	Bile duct adenoma
Number of patients	73	12	2	2	16	10	2	1
Number of imaging studies	80	17	2	2	16	11	2	1
Number of lesions	93	19	2	2	16	15	2	1

Table 4 Performance of the neural network on HCC classification. Performance was averaged over 150 runs with random sub-sampling to yield class-balanced test sets. *HCC* hepatocellular carcinoma

	HCC	Non-HCC	Overall
Training lesions	88	52	140
Test lesions	5	5	10
Sensitivity	92.7%	82.0%	87.3%
Specificity	82.0%	92.7%	87.3%

Because the majority of non-HCC lesions in the liver were benign and did not require surgical therapy, fewer pathological-proven non-HCCs than HCCs were available with ground-truth pathological proof and were mostly acquired incidentally in the setting of transplantation for liver failure or accompanied by secondary HCC in the liver. Therefore, these non-HCC lesions were grouped into a single pooled category. Metastatic lesions were excluded because pathology proof is frequently unavailable for secondary malignancies which do not generally undergo surgical resection. Pathological confirmation from various sources was used, including biopsies, resections, explants, and autopsies. Additionally, the time interval between MRI and pathological confirmation was variable and, especially in benign lesions, relatively large. However, the probability of a malignant transformation for a definitively benign finding is exceedingly low [20]. Additionally, the time interval in this study was less relevant, since pathology was only used to provide proof of diagnosis. Due to the small sample size, a large number of non-HCC lesions without background cirrhosis were used. However, lesions were cropped which reduced the impact of background liver tissue on the image analysis. Moreover, using heterogeneous imaging sources may seem like a limiting factor, but demonstrates the robustness of the CNN in the setting of different MRI scanners and acquisition protocols. The algorithm does not account for variabilities in contrast agents/acquisition time/image quality, suggesting that prospective studies should validate those points. Additionally, the diagnostic performance of CNN versus non-assisted radiologist versus CNN-assisted radiologist should be investigated in future studies in order to prove the CNN's clinical applicability. Moreover, lesion grading was conducted by single human reader leading to possible bias, which we tried to minimize through supervision.

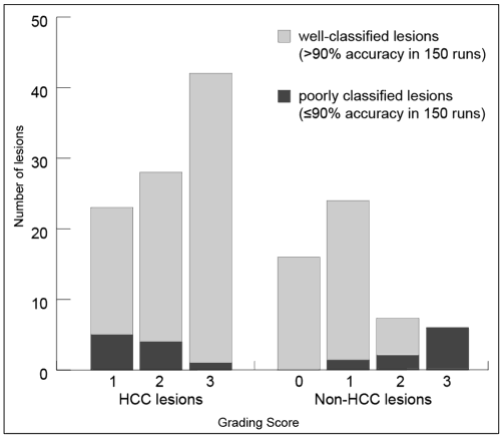

Fig. 5 Number of lesions by grading score. HCC hepatocellular carcinoma

In conclusion, this study demonstrates the use of deep learning for classification of both typical- and atypical-appearing HCC lesions on multi-phasic MRI, utilizing pathologically confirmed lesions as "ground truth." Currently most deep learning tools do not provide radiological-pathological validation in their training dataset. By strictly including only pathologically confirmed lesions, the underlying biological validity of deep learning systems can be optimized, paving the way for integration of decision support tools in clinical practice. Moreover, this allows for the evaluation of lesions with more atypical appearances, pushing the boundaries of non-invasive imaging-based diagnosis. In this manner, CNNs have the potential to eventually reduce the need for biopsies and their associated complications, resulting in improved patient care. The short computing time of our CNN will facilitate the inclusion into clinical routine.

Funding CW received funding from the Radiological Society of North America (RSNA Research Resident Grant #RR1731). JD, JC, ML, and CW received funding from the National Institutes of Health (NIH/NCI R01 CA206180).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Dr. Julius Chapiro, MD, PhD.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- retrospective
- diagnostic study
- performed at one institution

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424. https://doi.org/10.3322/caac.21492
2. Wald C, Russo MW, Heimbach JK, Hussain HK, Pomfret EA, Bruix J (2013) New OPTN/UNOS policy for liver transplant allocation: standardization of liver imaging, diagnosis, classification, and reporting of hepatocellular carcinoma. Radiology 266:376–382. https://doi.org/10.1148/radiol.12121698
3. CT/MRI LI-RADS v2018. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS/CT-MRI-LI-RADS-v2018. Accessed 31 Aug 2020
4. Davenport MS, Khalatbari S, Liu PSC et al (2014) Repeatability of diagnostic features and scoring systems for hepatocellular carcinoma by using MR imaging. Radiology 272:132–142. https://doi.org/10.1148/radiol.14131963
5. Smith EH (1991) Complications of percutaneous abdominal fine-needle biopsy. Review. Radiology 178:253–258. https://doi.org/10.1148/radiology.178.1.1984314
6. Seehofer D, Öllinger R, Denecke T et al (2017) Blood transfusions and tumor biopsy may increase HCC recurrence rates after liver transplantation. J Transplant. https://doi.org/10.1155/2017/9731095
7. Quaia E, De Paoli L, Angileri R, Cabibbo B, Cova MA (2014) Indeterminate solid hepatic lesions identified on non-diagnostic contrast-enhanced computed tomography: assessment of the additional diagnostic value of contrast-enhanced ultrasound in the non-cirrhotic liver. Eur J Radiol 83:456–462. https://doi.org/10.1016/j.ejrad.2013.12.012
8. Pérez Saborido B, Menéu Díaz JC, Jiménez de los Galanes S et al (2005) Does preoperative fine needle aspiration-biopsy produce tumor recurrence in patients following liver transplantation for hepatocellular carcinoma? Transplant Proc 37:3874–3877. https://doi.org/10.1016/j.transproceed.2005.09.169
9. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging. https://doi.org/10.1007/s13244-018-0639-9
10. Greenspan H, van Ginneken B, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 35:1153–1159. https://doi.org/10.1109/TMI.2016.2553401
11. Hamm CA, Wang CJ, Savic LJ et al (2019) Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol. https://doi.org/10.1007/s00330-019-06205-9
12. Wang CJ, Hamm CA, Savic LJ et al (2019) Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 29:3348–3357. https://doi.org/10.1007/s00330-019-06214-8
13. Yasaka K, Akai H, Abe O, Kiryu S (2018) Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology 286:887–896. https://doi.org/10.1148/radiol.2017170706
14. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90. https://doi.org/10.1145/3065386
15. Erickson BJ, Korfiatis P, Akkus Z, Kline TL (2017) Machine learning for medical imaging. Radiographics 37:505–515. https://doi.org/10.1148/rg.2017160130
16. Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. Statist Surv 4:40–79. https://doi.org/10.1214/09-SS054
17. CT/MRI LI-RADS v2017. https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS/CT-MRI-LI-RADS-v2017. Accessed 17 May 2018
18. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90
19. Huang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2261–2269. https://doi.org/10.1109/CVPR.2017.243
20. Fodor M, Primavesi F, Braunwarth E et al (2018) Indications for liver surgery in benign tumours. Eur Surg 50:125–131. https://doi.org/10.1007/s10353-018-0536-y

8. Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

9. Publikationsliste

Erstautorenschaften:

Paula M. Oestmann, Clinton J. Wang, Lynn J. Savic, Charlie A. Hamm, Sophie Stark, Isabel Schobert, Bernhard Gebauer, Todd Schlachter, MingDe Lin, Jeffrey C. Weinreb, Ramesh Batra, David Mulligan, Xuchen Zhang, James S. Duncan, Julius Chapiro
Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver
European Radiology, Volume 31, Ausgabe 7, Juli 2021 (S. 4981 – 4990)

Posterpräsentationen:

Poster-Präsentation auf dem „Annual Meeting der Radiological Society of North America“ („RSNA“ 2019) in Chicago

Koautorenschaften:

Sophie Stark, Clinton Wang, Lynn Jeanette Savic, Brian Letzen, Isabel Schobert, Milena Miszczuk, Nikitha Murali, **Paula Oestmann**, Bernhard Gebauer, MingDe Lin, James Duncan, Todd Schlachter, Julius Chapiro,
Automated feature quantification of Lipiodol as imaging biomarker to predict therapeutic efficacy of conventional transarterial chemoembolization of liver cancer
Scientific Reports, Oktober 2020 (online eingesehen am 22.9.21)

10. Danksagung

Zu allererst möchte ich mich bei meiner Doktormutter Priv.-Doz. Dr. med. Lynn Savic bedanken für die exzellente und zeitintensive Betreuung während meines Forschungssemesters im Yale Interventional Oncology Research Lab. Ihr regelmäßiges Feedback, ihre fachliche Expertise sowie ihre stets großzügige Hilfsbereitschaft habe ich sehr zu schätzen gelernt.

Außerdem möchte ich mich bei meinem Betreuer Dr. med. Julius Chapiro bedanken, der mir durch die Aufnahme in sein Forschungslabor diese Forschungsarbeit erst ermöglichte. Seine konstruktive Unterstützung bei der Datenakquisition, Methodik, Erstellung und Publikation des Manuskriptes ermöglichte schnell große Fortschritte in meinem wissenschaftlichen Vorhaben. Zudem möchte ich mich bei meinem Betreuer Prof. Dr. Gebauer bedanken für das Vertrauen in meine Forschung und seine Unterstützung. Auch möchte ich Prof. Dr. Rolf W. Günter danken, welcher mit einem Stipendium seiner Rolf W. Günther Stiftung diese Arbeit großzügig finanziell unterstützt hat.

Besonders großer Dank gilt meinem Zweitautor Clinton Wang, welcher mit der Programmierung des hierbeschriebenen CNNs diese Arbeit maßgeblich mitgestaltet hat und mir bei allen Informatik-Fragen mit Rat und Tat freundschaftlich bei Seite stand. Auch möchte ich Brian Letzen und Charlie A. Hamm, danken, welche mit ihrer Forschung wichtige Vorarbeiten für diese Dissertation geleistet haben. Zudem möchte ich Ramesh Batra und Xuchen Zhang danken für die Bereitstellung ihrer histopathologischen Datenbanken, durch welche die Patientenakquisition maßgeblich erleichtert wurde. Auch möchte ich mich bei allen Mitgliedern des Labors für ihr regelmäßiges Feedback während der wöchentlichen Lab-Meetings bedanken, durch welches sie die Entwicklung des Forschungsprojektes nachhaltig vorangetrieben haben, sowie für die vielen hilfreichen Anmerkungen während der Überarbeitung des Manuskripts. Besonderer Dank gilt hier Sophie Stark für den täglichen wissenschaftlichen Austausch sowie die vielen motivierenden und freundschaftlichen Gespräche während unseres gemeinsamen Forschungsaufenthaltes.

Zuletzt möchte ich mich ausdrücklich bei meiner Familie für ihre Unterstützung bedanken, durch die mir dieser Forschungsaufenthalt ermöglicht wurde. Meine Familie stand mir immer mit einem offenen Ohr zur Seite und ihre konstruktiven und aufbauenden Worte haben mich stets motiviert.