Aus der Klinik für Zahn-, Mund- und Kieferheilkunde
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Generalisierbarkeit von Deep-Learning-Modellen zur Detektion kariöser Läsionen

Generalizability of deep learning models for the detection of carious lesions

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

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Agnes Holtkamp

Datum der Promotion: 25. Juni 2023

Burkhard!
Auch wenn Du hier nicht stehen möchtest,
ich widme Dir die Arbeit von ganzem Herzen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

Nah-Infrarot-Licht-Transillumination	NILT
Künstlichen Intelligenz	KI
Residual Neural Network	ResNet
Gradient-weighted Class Activation Mapping	GradCAM
Convolutional Neural Network	CNN
Near-Infrared Light Transillumination	NILT
Potentia hydrogenii	pH-Wert
Calcium	Ca^{2+}
Dimensional	D
zum Beispiel	z. B.
oben genannt	o. g.
Micro-Computer-Tomographie	µ-CT
unter anderem	u. a.
sogenannt	sog.
et cetera	etc.
oder ähnliches	o. ä.
Artificial Intelligence	AI
siehe unten	s. u.
siehe oben	s. o.
Integrated Desktop/Version	KID
Operation	OP
True Positive	TP
True Negative	TN
False Positive	FP
False Negative	FN
Positiver Vorhersagewert	PVW
Negativer Vorhersagewert	NVW
Area-under-the Operating Characteristic Curve	AUC
Receiver Operating Characteristic	ROC
Rot Grün Blau	RGB
Abbildung	Abb.
Rectified Linear Unit	ReLu

Digitale Volumentomographie	DVT
Explainable AI	XAI
STAndards for Reporting of Diagnostic accuracy	STARD

Zusammenfassung

Ziele: Die Nah-Infrarot-Licht-Transillumination (NILT) ist ein alternatives Verfahren zur radiologischen Kariesdetektion und vor allem geeignet zur Detektion früher approximaler Kariesläsionen. Für die Analyse von NILT-Bildern stehen erste Modelle aus dem Bereich der Künstlichen Intelligenz (KI) zur Verfügung, die auf Bildmaterial trainiert wurden, das entweder in vivo (Routinedaten) oder in vitro (extrahierte Zähne) gewonnen wurde. Die vorliegende Studie untersuchte die Generalisierbarkeit dieser KI-Modelle auf in vivo und in vitro gewonnenen NILT-Bilddaten.

Methoden: Das zugrundeliegende Datenmaterial umfasste 1319 NILT-Segmente (von 508 Prämolaren, 811 bleibende Molaren) aus 56 erwachsenen Patienten, die klinisch untersucht worden waren (in vivo). In vitro wurden 226 extrahierte Zähne (113 Prämolaren, 113 bleibende Molaren) in einem standardisierten Simulationsmodell eingebettet und NILT-Bildsegmente generiert. Die genutzte NILT-Technologie basierte auf der DIAGNOcam (DIAGNOcam, Kavo). Auf allen Bildsegmenten bewerteten drei unabhängige, erfahrene Zahnärzt*innen das Vorhandensein einer Approximalkaries, ein vierter Zahnarzt überprüfte diese Bildbewertungen („Masterannotator“). Es wurden Convolutional Neural Networks (Res-Net) zur Klassifikation (Karies auf NILT-Bild vorhanden ja/nein) trainiert und mittels k-facher Kreuzvalidierung mit jeweils 10 Trainings-, Validierungs- und Test-Splits validiert. Dabei wurde vor allem die Generalisierbarkeit von in vivo oder in vitro trainierten Daten auf dem jeweiligen anderen Datenmaterial überprüft. Um die Klassifikationsentscheidungen der KI-Modelle nachvollziehbar zu machen, wurden mittels GradCAM-Visualisierung entscheidungsrelevante Bereiche in den Bildern dargestellt.

Ergebnisse: Die Prävalenz kariöser Läsionen betrug 41 % in vitro und 49 % in vivo. Die mittlere (± Standardabweichung) Genauigkeit war signifikant höher für KI-Modelle, die an In-vivo-Daten trainiert und getestet wurden (0.78±0.04). Modelle, die an In-vitro-Daten trainiert und getestet wurden, zeigten signifikant niedrigere Genauigkeiten (0.64±0.15; $p<0.05$). Auch Modelle, die in vitro getestet und in vivo trainiert wurden, zeigten signifikant geringere Genauigkeiten (0.70±0.01; $p<0.01$), ebenso wie Modelle, die in vitro trainiert und in vivo getestet wurden (0.61±0.04; $p<0.05$). Grund dafür war die Abnahme der Sensitivität (-10 % für in vitro trainierte Modelle und -27 % für in vivo trainierte Modelle). Falsch-positive Erkennungen wurden oft mit Restaurationen in Verbindung gebracht; bei

falsch-negativen Erkennungen wurden häufig Areale als relevant erachtet, die nicht kariös waren (Aufmerksamkeitsproblem).

Schlussfolgerung: Eine Generalisierbarkeit der entwickelten KI-Modelle war nicht gegeben.

Klinische Relevanz: Für den klinischen Einsatz vorgesehene Modelle sollten auf in vivo gewonnenen Daten trainiert werden.

Abstract

Objectives: We trained deep convolutional neural networks (CNNs) on Near-Infrared Light Transillumination (NILT) images that were taken in vivo or in vitro to detect proximal caries lesions to generate generalizability of the models.

Methods: NILT images of 226 extracted posterior human teeth (DIAGNOcam, KaVo, Biberach) were taken in vitro after assembling them in a dummy head. In vivo, 1319 teeth from 56 patients were obtained and segmented similarly. Proximal caries lesions were annotated independently by three experienced dentists and reviewed by a fourth. The segments were transformed into binary labels. ResNet classification models were trained on both in vivo and in vitro datasets and 10-fold cross-validated. Generalizability and explainability were explored. We used GradCAM to increase explainability.

Results: In vitro and in vivo data showed a prevalence of caries lesions of 41 % and 49 %, respectively. Models trained and tested in vivo performed significantly better (mean ± SD accuracy: 0.78 ± 0.04) than those trained and tested in vitro (accuracy: 0.64 ± 0.15: $p < 0.05$). Using in vivo models on in vitro data led to significantly lower accuracy (0.70 ± 0.01; $p < 0.01$). Similarly, when tested in vivo, models trained in vitro showed significantly lower accuracy (0.61 ± 0.04; $p < 0.05$). In both cases, this was due to decreases in sensitivity (-10 to −27 %).

Conclusions: Deep learning models showed limited generalizability and low accuracy for imagery from in vitro versus in vivo settings.

Clinical significance: Using in vitro imagery to create deep learning models should proofed for generalizability. Acceptable Deep learning models for NILT imagery are supposed to be trained on in vivo data.

1. Einleitung

1.1 Karies und Kariesmanagement

Die Behandlung kariöser Läsionen ist die häufigste Tätigkeit in der zahnärztlichen Praxis; Karies ist die prävalenteste Erkrankung der Menschheit: Über 90% der Erwachsenen weisen Karieserfahrung, also mindestens eine kariöse Läsion, eine Füllung oder einen Zahnverlust aufgrund von Karies auf [1]. Klassischerweise wurde Karies als Infektionserkrankungen verstanden – hervorgerufen durch einen oder wenige bakterielle Erreger, zuvorderst Streptococcus mutans. Ausgehend von diesem Verständnis und der eingeschränkten Auswahl an therapeutisch verfügbaren Dentalprodukten wurde über fast ein Jahrhundert, Karies vor allem restaurativ therapiert: Kariöses Zahnhartgewebe wurde invasiv entfernt und durch alloplastisches Material, vor allem Amalgam, ersetzt. Die zahnärztliche Tätigkeit bestand zum großen Teil aus der Versorgung ausgedehnter kavitierter kariöser Läsionen mit einem solchen restaurativen Vorgehen [2].

Dieses Verständnis und die daraus resultierenden Therapieansätze haben sich über die letzten Jahrzehnte dramatisch gewandelt: Das Paradigma von Karies als Infektionserkrankungen ist nicht länger haltbar. Stattdessen wird Karies heute als das Ergebnis eines Ungleichgewichtes in der Biofilmzusammensetzung und -aktivität auf dem Zahnhartgewebe verstanden. Der physiologische dentale Biofilm wird durch äußere Rahmenbedingungen, vor allem eine hochfrequente Zufuhr fermentierbarer Kohlenhydrate, schrittweise in einen pathogenen Biofilm umgewandelt. Hierbei spielen vor allem säurebildende (azidogene) und säuretolerante (azidurische) Bakterien eine entscheidende Rolle. Sie können Kohlenhydrate zu organischen Säuren verstoffwechseln, hierdurch den pH-Wert innerhalb des Biofilms senken und somit andere, nicht azidogene und azidurische Bakterien verdrängen (Wettbewerbsvorteil durch Bildung einer ökologischen Nische; ökologische Plaquehypothese). Der so veränderte Biofilm ist in der Lage, relativ große Mengen organischer Säuren zu bilden und hiermit die Zahnhartgewebe substantiell zu demineralisieren (Nettomineralverlust).

Aus diesem veränderten Verständnis ergeben sich eine Reihe neuer Therapieansätze, wobei nicht restaurative, sondern kausal orientierte non-restaurative Maßnahmen im Vor-

dergrund stehen. So können zum Beispiel die Kontrolle der Biofilmmaturation (mechanische oder chemische Mundhygienemaßnahmen) oder Zusammensetzung (Pro- und Präbiotika), die Kontrolle der Kohlenhydratzufuhr (Zuckerrestriktion, Zuckerersatzstoffe) oder die Beeinflussung der Imbalance von Mineralverlust und -gewinn (Fluoride, Mineralliefe-ranten) eingesetzt werden. Neben diesen lang erprobten, sogenannten non-invasiven Maßnahmen wird zudem ein mikro-invasives Vorgehen empfohlen; hierbei werden bei dem Einsatz von bestimmten Säuren zur Konditionierung oder Entfernung der oberflächlichen Zahnhartsubstanz einige Mikrometer derselben abgetragen und die Karies anschließend versiegelt oder infiltriert (Abbildung 1).

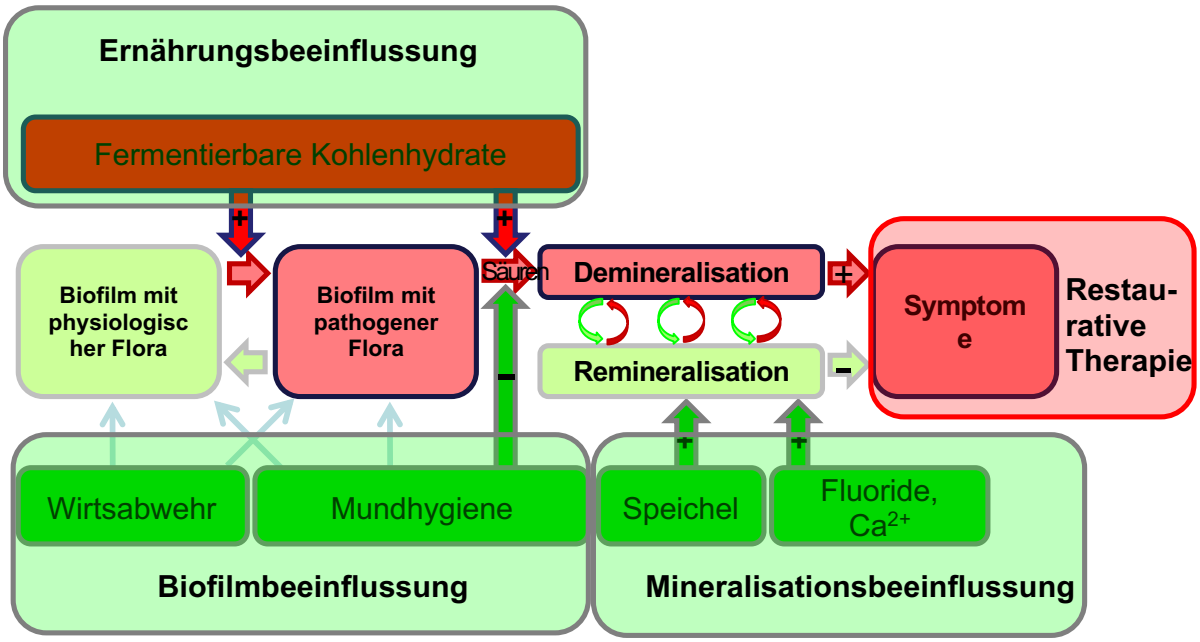

Abbildung 1: Kariespathogenese auf Basis der ökologischen Plaquehypothese und zugeordnete therapeutische Optionen. Modernes Kariesmanagement fokussiert auf der Beeinflussung pathogener oder protektiver Faktoren im Kariesprozess statt einer restaurativen, symptomatischen Therapie. Quelle: Paris S, Meyer-Lückel H. Kim R. Ekstrand 2012. Karies. Wissenschaft und Klinische Praxis. Kapitel 4. Seite 73 [3]. Thieme Verlag Stuttgart (2012). Bildrechte durch Lizenznummer 5165910287801 abgegolten.

Restaurative Maßnahmen werden hingegen zunehmend einem eingeschränkten Indikationsspektrum vorbehalten, vor allem der Versorgung kavitierter kariöser Läsionen. Eine kürzlich erschienene, internationale Konsensusempfehlung hat dies noch einmal untermauert [2]: Zahnärzt*innen sollten demnach danach streben, kariöse Läsionen frühzeitig

unter Nutzung von non- und mikro-invasiven Ansätzen zu therapieren und restaurative Therapien nur für vorangeschrittene, eingebrochene Läsionen einzusetzen.

Der versuchte Verzicht oder zumindest das zeitliche Verschieben restaurativer Maßnahmen erfolgte dabei auf ausgehend von der Erkenntnis, dass Restaurationen – auch wenn sie sorgfältig platziert werden – statistisch gesehen eine begrenzte Lebenserwartung haben. In kontrollierten Bedingungen (z. B. Studien) werden jährliche Versagensraten von 1 - 4 % berichtet, in Routinedaten (z. B. Versicherungsdaten) sind auch deutlich höheren Versagensraten pro Jahr (>5 %) nicht unüblich [4] [5]. Gerade für die Versorgung früher kariöser Läsionen bei jungen Patient*innen ist dies bedeutsam: Wird eine frühe Läsion mit einer Füllung bei einem 15-Jährigen versorgt, so muss sie möglicherweise 10 - 20 Jahre später erneuert werden, wobei stets weitere Zahnhartsubstanz geopfert werden muss (unter anderem, weil Restaurationen durch Karies oder Frakturen versagen, aber auch weil bei der Entfernung des alten Restaurationsmaterials häufig intakte Zahnhartsubtanz verloren geht). Auch die neu platzierte Restauration wird dann jedoch nach einer gewissen Zeit erneuert werden müssen; die initial restaurative Therapie leitet eine unumkehrbare Kaskade eskalierender und teurer werdender Interventionen ein – die sogenannte „Todesspirale des Zahnes“ [6] [7]. Um diese zu vermeiden oder zumindest möglichst in ein höheres Lebensalter zu verschieben, ist der Fokus der modernen Zahnmedizin eindeutig auf die Kariesprävention und -arretierung ausgelegt [2]. Um nun aber Kariesläsionen frühzeitig therapieren zu können, muss eine zentrale Voraussetzung erfüllt sein: Die Detektion der Läsion in diesem frühen Stadium.

1.2 Kariesdetektionsmethoden

Die klassische Detektionsmethode für Karies, die visuelle taktile Detektion, ist zur Erkennung früher kariöser Läsionen in vielen Fällen nicht ausreichend. Die Sensitivität, also das Maß dafür, kariöse Flächen als solche zu erkennen, ist bei dieser Methode begrenzt - vor allem zur Detektion von Approximalkaries in geschlossenen Zahnreihen. Hier werden circa 80 – 90 % (vor allem früher) kariöser Läsionen übersehen. Einzig vorangeschrittene, kavitierte Läsionen können erfolgreich mit diesem Verfahren auch im Approximalraum aufgefunden werden, dann ist jedoch eine non- oder mikro-invasive Therapie nur selten noch möglich und eine restaurative Behandlung angezeigt [8] [9].

Aus diesem Grund nutzen Zahnärzt*innen seit fast 100 Jahren weitere Verfahren ergänzend zum visuell taktilen Vorgehen, allen voran das Bissflügelröntgen [10]. Hierbei werden auch die Zahnzwischenräume mittels einer intraoralen Röntgenaufnahme dargestellt und somit die Detektion früher Kariesläsionen ermöglicht. Allerdings hat dieses Verfahren eine Reihe von Nachteilen:

1. Es generiert ionisierende Strahlung. Hierdurch ist ein engmaschiger, wiederholter Einsatz nur bedingt möglich. Auch bei Hochrisikopatienten wird selten häufiger als alle 18 - 24 Monate geröntgt. Ebenso ist die Indikationsstellung bei Kindern sehr streng; gerade in dieser Gruppe wäre jedoch eine frühzeitige Kariesdetektion wünschenswert.

2. Bissflügelröntgen bedarf einer entsprechenden Ausrüstung (Strahlungsquelle, Strahlenschutz) und ist nur selten portabel erhältlich. Der Einsatz in Kindergärten, Schulen oder Arbeitsstätten ist demnach so gut wie unmöglich.

Ausgehend von diesen Limitationen haben sich eine Reihe alternativer, ergänzender Detektionsverfahren etabliert. Das in der vorliegenden Arbeit im Fokus stehende Verfahren ist die Nahinfrarottransillumination (NILT). Bei diesem Verfahren wird Zahnhartgewebe mit Nahinfrarotlicht durchstrahlt. Im kariösen Zahnhartgewebe (vor allem Schmelz) kommt es zu einem abweichenden Lichtbrechungsverhalten im Vergleich mit gesundem Gewebe, was wiederum mittels einer digitalen Kamera aufgezeichnet und auf einem Bildschirm visualisiert werden kann (Abb. 2). NILT erlaubt eine diagnostische Genauigkeit zur Kariesdetektion, die dem Bissflügelröntgen ähnlich ist [11] [12] [13] [14]. Vorteile dieses Verfahrens sind der Verzicht auf ionisierende Strahlung, die Portabilität der nötigen Hardware und die Möglichkeit, durch Videoaufnahmen bzw. Bewegung des Sensorkopfes im Raum auch die räumliche dreidimensionale Ausdehnung der Läsion abschätzen zu können. Im Vergleich zum Bissflügelröntgen ist jedoch die Einschätzung der Tiefenausdehnung im Dentin nur eingeschränkt möglich (vor allem durch im Vergleich zum Schmelz unterschiedlichen Brechungseigenschaften des Dentins). Auch ist die Detektion von Sekundärkreis ist nur unzureichend untersucht [15]. NILT wird häufig als Alternative zum Bissflügelröntgen, gerade bei Kindern und Schwangeren oder zur engmaschigen Evaluation von Hochrisikopatienten, diskutiert. Ebenso ist der Einsatz in Lebenswelten

außerhalb der zahnärztlichen Praxis, zum Beispiel in Kindertagesstätten, Schulen oder Betrieben möglich.

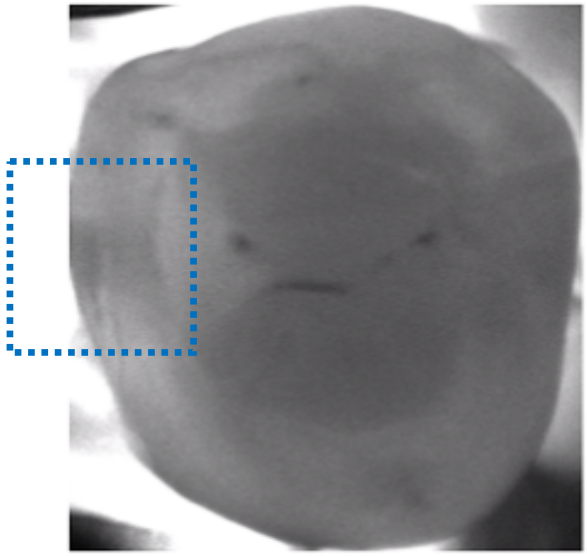

Abbildung 2: Transilluminationsbild eines Prämolaren. Zahnhartsubstanzspezifische Lichtbrechungsphänomene erlauben die Abgrenzung von Schmelz und Dentin, aber auch von kariösem Gewebe (blauer Kasten) und gesundem Gewebe. Quelle: Eigene Darstellung.

1.3 KI zur medizinischen Bildanalytik

Die Analyse medizinischer Bilder ist auch für erfahrene Expert*innen herausfordernd. So ist die Interpretation von Fotos, aber auch 2-D- oder 3-D-Röntgenbildaufnahmen durch große Variabilität der Untersucher*innen gekennzeichnet; je nach Erfahrungsgrad und individuelle Situation (z. B. Zeitdruck) schwankt die Genauigkeit der Untersucher*innen bei der Erkennung von Pathologien auf Bildmaterialien enorm. Dies trifft auch auf NILT-Bilder zu.

Eine Reihe von Studien konnten zeigen, dass gerade die Interpretation von NILT-Bildern für zahnärztliche Untersucher*innen eine Herausforderung ist. Möglicherweise, weil dieses Bildmaterial relativ neu ist und eine Interpretation für viele Zahnärzt*innen weder im grundständigen Studium gelehrt noch in der Routine geübt worden ist. Um einen Einsatz

in der Fläche (also gerade auch in o. g. Lebenswelten), zum Beispiel durch den zahnärztlichen Dienst der Länder im Rahmen der Gruppenprophylaxe zu ermöglichen, sind qualitätssichernde Elemente bei der Interpretation von NILT-Bildern wünschenswert.

Zur medizinischen Bildanalytik haben sich seit einigen Jahren Technologien aus dem Bereich der Künstlichen Intelligenz (KI) als hilfreich erwiesen, vor allem unter Rückgriff auf Methoden des sogenannten Maschinellen Lernens. Bei der häufigsten Form des Maschinellen Lernens, dem „überwachten Lernen“ (supervised learning) wird eine Maschine durch oft wiederholte Assoziation von Bildern mit dazugehörigen Bildinformation (zum Beispiel Karies ja/nein) in die Lage versetzt, eine mathematische, modellbasierte Repräsentation von Kariesläsionen auf diesem Bildmaterial zu entwickeln. Dies erlaubt schließlich auch auf neuem, ungesehen Bildmaterial diese Klassifikation. Zum Training einer solchen Repräsentation (des KI-Modells) werden oft Tausende von Bildern eingesetzt. Die zur Bildanalytik üblichsten Modelle stammen aus dem Subfeld des „tiefen Lernens“ (deep learning), beide, hochkomplexe neuronale Netzwerke, vor allem sog. Convolutional Neural Networks (CNNs) werden eingesetzt. CNNs bestehen aus gestapelten Schichten von linearen Modellen, Gewichten, die über die Verbindung, Signalstärke und die Ausgabe der Informationen entscheiden, nichtlinearen Funktionen und einem Bias-Term. CNNs unterscheiden sich von anderen neuronalen Netzwerken durch Faltungsoperationen zur Extraktion bestimmter Merkmale bei Bildern (Kanten, Flecken und makroskopische Muster).

1.4 Daten

Um nun ein neuronales Netzwerk trainieren zu können, sind, wie dargelegt Bildinformationen und Bilder notwendig. In der Zahnmedizin haben sich dabei zwei unterschiedliche Strategien zur Generierung von Bildmaterial und -information etabliert. In diagnostischen Genauigkeitsstudien zur Kariesdetektion sind sogenannte In-vitro-Ansätze üblich, bei denen extrahierte Zähne zunächst mit dem sogenannten Indextest (also z. B. Röntgen, NILT, etc.) untersucht und danach histologisch, mittels transversaler Mikroradiografie oder µ-CT analysiert werden. Letzteres dient der Definition eines sog. Goldstandards (Referenztests) und ist in allen drei Beispielfällen relativ zuverlässig möglich. Umgekehrt ist die Zahl der verfügbaren extrahierten Zähne endlich und die Nutzung dieses Probenmaterials (u. a. durch ethische Bedenken etc.) zunehmend komplex. Zudem leiden die

so aufgestellten Indextests (also z. B. Röntgen- oder NILT-Bilder) an Verzerrungen, u.a. durch mangelnde Repräsentativität (Selektionsverzerrung durch Einschluss nur extrahierter Zähne) und technischen Einschränkungen (Generierung nicht am Patienten, sondern im Labor).

Auf der anderen Seite können in der Routine (in vivo, also am Patienten gewonnene Bilder eingesetzt werden). Hierzu ist der Zugriff auf Routinedaten (also nicht für einen Forschungs-, sondern Krankenversorgungszweck routinemäßig gewonnener Daten) nötig. Jedoch erlaubt dieses Vorgehen, eine große Zahl von Bildern zu nutzen, die zudem heterogen, repräsentativ und deutlich näher an der klinischen „Grundwahrheit“ liegen. Allerdings ist hier wiederum die Herstellung eines Referenztests deutlich schwieriger, da z. B. eine histologische Aufbereitung und Wahrheitsfindung nicht möglich ist; nur selten stehen weitere bildgebende Materialien (zum Beispiel Digitale Volumentomographien o. ä.) oder weitere klinische Tests (visuelle Inspektion) zur Sicherung der Diagnose zur Verfügung. Stattdessen muss oft auf dem vorhandenen Bildmaterial durch Expert*innen ein Goldstandard festgelegt werden. Ausgehend von den oben genannten Schwierigkeiten und der mangelnden Reliabilität der Expert*innen muss dann jedes Bild häufig von mehreren Expert*innen bewertet werden und die Bewertungen z. B. durch ein Mehrheitsvotum o. ä. vereinigt werden.

2. Ziel der Studie

In der Zahnmedizin wurden CNNs zur Erkennung von Kariesläsionen, parodontalem Knochenverlust und apikalen Läsionen auf Röntgenbildern eingesetzt. Auch die Analyse von Fotografien oder die Detektion und Klassifikation anatomischer Landmarken, beispielsweise auf Fernröntgenseitenbildern oder Oberflächenscans, ist mittlerweile üblich [16].

Zwei erste Publikationen zeigten, dass CNNs auch zur Karieserkennung auf NILT Bildern angewendet werden können [17] [18]. Die Modelle in beiden Studien erzielten nützliche Genauigkeiten auch auf relativ kleinen Trainingsdatensätzen, jedoch ist unklar, inwieweit eine Generalisierbarkeit der CNNs gegeben ist: Gerade die Übertragbarkeit von KI-Modellen ist in der Medizin nur ungenügend untersucht. Neuere Studien zeigen beispielsweise für sog. Shallow Machine Learning Verfahren, dass eine Generalisierbarkeit von Modellen zur Zahnverlustvorhersage aus einer Kohorte in eine andere Kohorte nicht zwingend gegeben ist [19]. Ebenso konnte für die CNN-basierte Bildanalyse, beispielsweise für die Detektion apikaler Läsionen auf Panoramaschichtaufnahmen, gezeigt werden, dass KI-Modelle, die an einem Trainingsdatensatz aus einem europäischen Zentrum trainiert worden sind, nicht generalisierend auf Bildern aus einem indischen Zentrum angewandt werden können [20]. Demnach ist es relevant zu wissen, ob CNNs, die auf In-vitro-Daten trainiert worden sind, auch in vivo angewandt werden können oder umgekehrt in vivo trainierte CNNs sinnvoll auf In-vitro-Daten getestet werden können. Beide Szenarien sind durchaus denkbar, wobei gerade das Testen auf In-vitro-Daten vielversprechend ist, da die Herstellung eines harten Referenztests wie oben beschrieben, möglich ist. Das Training auf In-vivo-Daten wiederum wird es vermutlich ermöglichen, eine ausreichend große Fallzahl zu erreichen, die in vitro (wie dargestellt) nur selten verfügbar ist. Die Anwendung einer KI-basierten Software zur NILT-Analyse wird dann wiederum in vivo geschehen.

Ausgehend von diesen Überlegungen zielte die vorliegende Studie darauf ab, die Generalisierbarkeit von CNNs, die in vivo bzw. in vitro trainiert worden waren, auf dem jeweilig anderen Bildmaterial zu untersuchen. Unsere Hypothese lautete, dass die Detektionsgenauigkeiten der CNNs signifikant sinken, wenn sie auf einem anderen Datenmaterial angewandt werden als dem, auf dem sie entwickelt worden sind. Ein zweites Ziel der Studie war es, zu verstehen, welche Bildeigenschaften der Klassifikationsentscheidung

des Modells zugrunde liegen. Diese „Erklärbarkeit“ ist eine zunehmend relevante Eigenschaft von KI (explainable AI), die es erlaubt, die technische Logik des KI-Modells mit der medizinischen Logik des Experten zu vergleichen und so etwaige Verzerrungen aufzuzeigen und das Vertrauen der Nutzer*innen in die KI zu erhöhen (Abbildung 3).

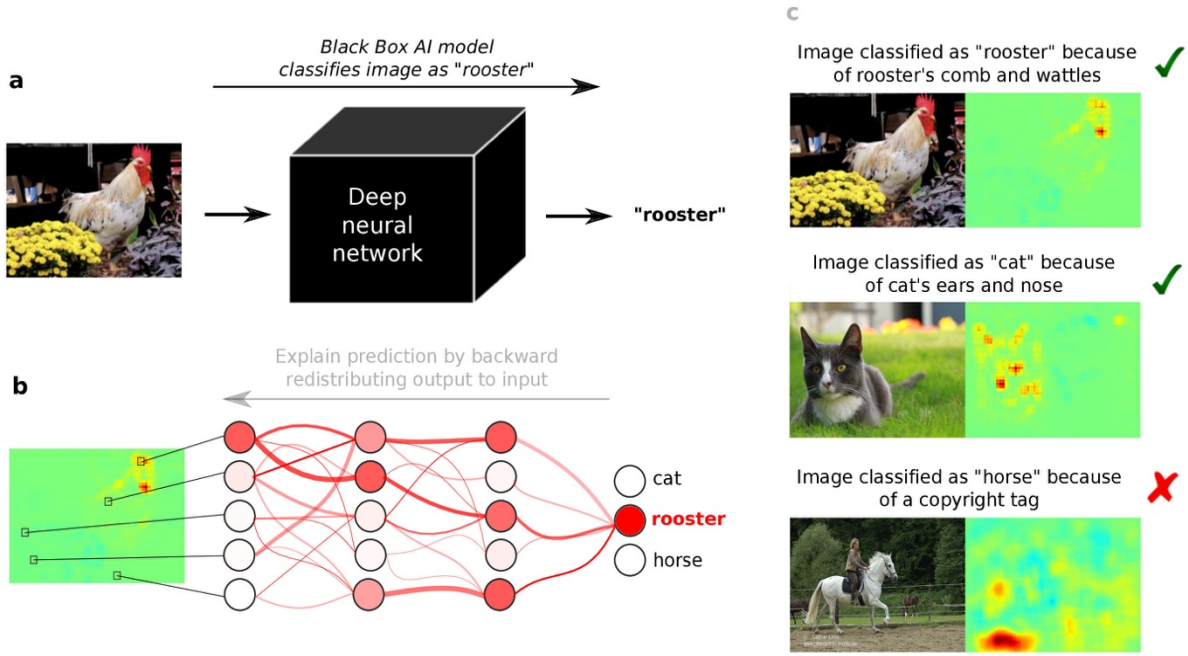

Abbildung 3: KI-Klassifikationsmodelle sind aufgrund ihrer komplexen statistischen Struktur oft nicht erklärbar („Black Box Modelle“) (a). Durch Methoden der „Erklärbarkeit“ (explainable AI) können für die Entscheidung relevante Areale im Bild sichtbar gemacht (re-identifziert) und die Modellentscheidung auf Konsistenz mit dem menschlichen Wissen überprüft werden (b). In diesem Beispiel nutzt das Modell zur Erkennung von Hähnen und Katzen für Menschen nachvollziehbare Bildareale zur Entscheidung. Für die Klassifikation „Pferd“ zieht die KI jedoch ein Copyright-Label am unteren Bildende zu Rate, vermutlich weil der Trainingsdatensatz viele Pferdebilder mit einem solchen Copyright-Label (vermutlich desselben Fotografen) enthielt: Das Modell hat gelernt, dass ein solches Label mit dem Vorhandensein eines Pferdes zusammenhängt. Medizinische KI-Modelle, u. a. zur Kariesdetektion auf NILT-Bildern, sollten auf diese innere Logik und Erklärbarkeit hin geprüft werden. Quelle: Schwendicke F, Samek W, Krois J. 2020. Artificial Intelligence in Dentistry: Chances and Challenges. Seite 773. Journal of Dental Research 99(7) [21]. Sage Journals, Bildrechte nach CC-BY NC 4.0. beim Verlag eingeholt.

3. Material und Methoden

3.1 Studiendesign

In dieser Studie wurden Datensätze von NILT-Bildern aus zwei verschiedenen Quellen verwendet. Der eine Datensatz wurde in vitro für eine frühere Studie [18] und der andere in vivo innerhalb der klinischen Routine generiert (s. u.). Jedes Bild wurde zunächst pixelweise von drei unabhängigen Zahnärzt*innen bewertet und abschließend noch von einem Master-Reviewer. Die Pixelsegmentierungen wurden dann in Klassifkationslabel (Karies ja/nein) umgewandelt. Es wurden CNNs vom Typ ResNet zur binären Bild-Klassifikation jeweils auf einem Trainingsdatensatz aus einer Quelle trainiert und auf Hold-out-Datensätzen derselben oder der anderen Quelle getestet (Abb. 2.). Die Berichterstattung der Studie folgt der STARD-Richtlinie [22] und der Checkliste für künstliche Intelligenz in der medizinischen Bildgebung, CLAIM [23].

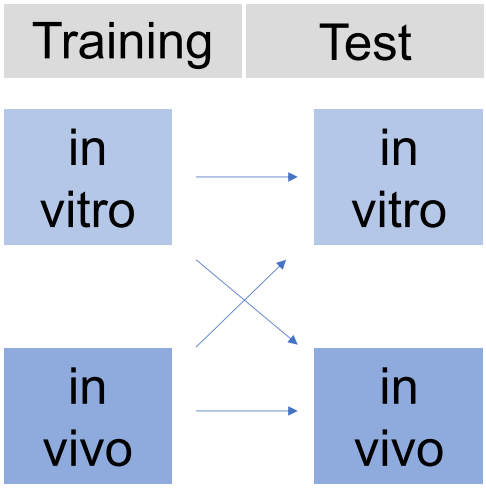

Abbildung 4: NILT-Daten, die in vitro oder in vivo generiert worden waren, wurden zum Training und Testen von KI-Modellen verwendet, wobei sowohl auf Daten derselben als auch der jeweils anderen Quelle trainiert und getestet wurde. Quelle: Eigene Darstellung.

3.2 Datengenerierung

Die Funktionsweise von NILT zur Kariesdetektion mittels DIAGNOcam ist in Abbildung 5 zusammengefasst. Zur Generierung des In-vitro-Datensatzes wurden 226 extrahierte Seitenzähne (113 Prämolaren und 113 Molaren) eingesetzt; die Nutzung der Zähne wurde durch die Ethikkommission der Charité, EA4/102/14, genehmigt. Die Zähne wurden in transparentes Epoxidharz (Epo-Thin 2, Buehler, Lake Bluff, USA) eingebettet. Die Modelle wurden in einen Phantomkopf (P-6, Frasaco, Tettnang, Deutschland) montiert und NILT-Bilder mit der DIAGNOcam von einem erfahrenen Zahnarzt aufgenommen, wobei die Kamera senkrecht zur Okklusalfläche über jeden Zahn bewegt wurde. Das Licht der Behandlungseinheit war während dieser Untersuchung ausgeschaltet. Die Bilder wurden mit der KID-Software (KaVo Integrated Desktop/Version 2.4.1.6821, KaVo) aufgenommen, wobei jedes Bild auf einen Zahn fokussiert war.

Zur Generierung des In-vivo-Datensatzes wurden von Juli 2019 bis März 2020 innerhalb der klinischen Routinebehandlung am Centrum 3 der Charité – Universitätsmedizin Berlin, zusätzlich zur herkömmlichen visuellen Kariesdiagnostik sowie radiologischer Bildgebung 56 volljährige Patient*innen auch mittels NILT untersucht (1319 bleibende Molaren und Prämolaren). Eine Einwilligung der Proband*innen sowie die Zustimmung der Ethikkommission der Charité - Universitätsmedizin Berlin lag vor (EA4/080/18). Beurteilt wurden ausschließlich Zähne, die keine anatomischen Anomalien zeigten. Zähne mit folgenden klinischen Befunden wurden innerhalb der vorliegenden Studie nicht berücksichtigt: Direkte oder indirekte Restauration (Füllungen, Kronen, Teilkronen, Inlays, Brückenanker), Amelogenesis imperfecta, Dentinogenesis imperfecta, Hypoplasien. Zahnfehlstellungen oder nicht vorhandene Approximalkontakte wurden nicht als Ausschlusskriterium eingesetzt. Nach Lufttrocknung der entsprechenden okklusal-approximal Flächen mittels der 3-Wege-Spritze und Abschaltung von OP- oder Kopflampen wurde der Aufsatz der DIAGNOcam senkrecht auf die Zähne gesetzt und bewegt. Um möglichst aussagekräftige Bilder der einzelnen Zähne zu generieren, wurden die Bilder aus leicht unterschiedlichen Perspektiven aufgenommen. Ziel war es, die Approximalkontakte mesial und distal sowie eine Einzelaufnahme des Zahnes von okklusal zu erhalten. Die Datenerhebung erfolgte hierbei immer durch dieselbe Untersucherin und ebenfalls mit der KID-Software.

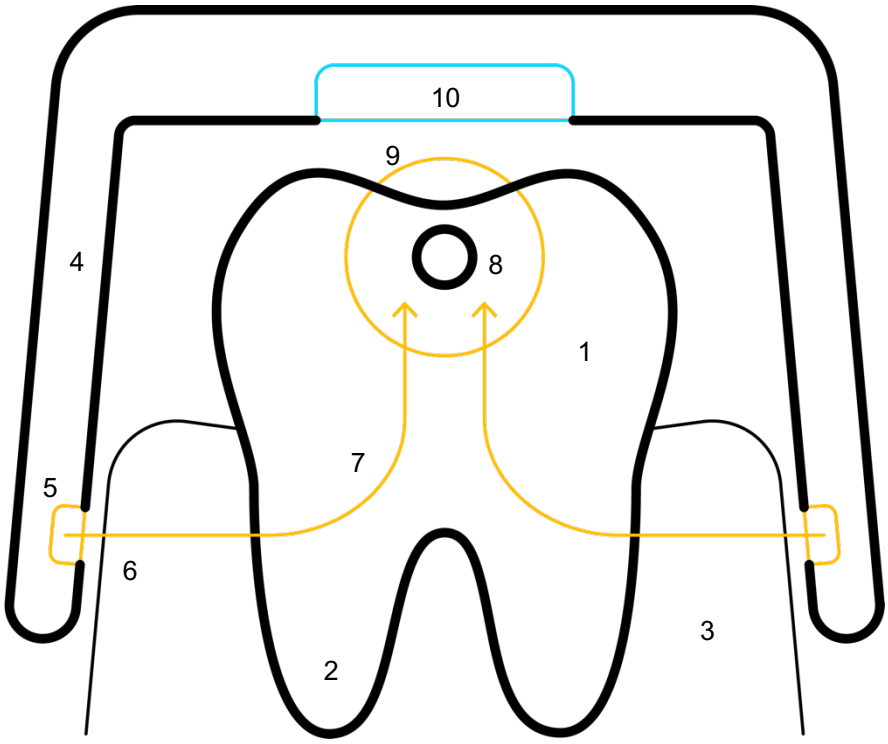

Abbildung 5: Funktionsweise der DIAGNOcam. (1) Zahn, (2) Wurzel, (3) Alveolarknochen, (4) Flexible Arme lateral des Zahnes, (5) Austrittsstelle Nahinfrarotlicht, (6) Illumination auf Wurzelhöhe, (7) Transillumination durch Zahn, (8) Kariöse Läsion, (9) Austritt des Nahinfrarotlichts nach Durchleuchtung des Zahnes, (10) Digitale Kamera. Am Kopf des Gerätes ist die Einheit zur Bildgebung lokalisiert. Sie setzt sich aus Kamera und zwei Lichtquellen zusammen. Führt man die DiagnoCam senkrecht über den Zahn, befinden sich die Lichtquellen, die aus zwei flexiblen Armen austreten, oral und vestibulär des Zahnes. Die Biegsamkeit des Gerätekopfes erlaubt die individuelle Anpassung an jeden Zahn. Das Nahinfrarotlicht wird mittels Glasfaserkabel weitergeleitet. Es durchleuchtet den Zahn beginnend von apikal nach koronal. Die unterschiedlichen Absorptionsspektren werden mit der entsprechenden Kamera im Aufsatz des Gerätes aufgefangen und können digital an einem Bildschirm visualisiert und gespeichert werden. Approximale Schatten können als kariöse Läsion gedeutet werden. Quelle: Eigene Darstellung.

3.3 Referenz-Test

Zur Erstellung eines Referenztests erfolgte auf beiden Datensätzen zunächst eine pixelweise Bewertung der Bilder (Segmentierung). Drei unabhängige, erfahrene Zahnärzt*innen (klinische Erfahrung: 8-11 Jahre) markierten die von approximaler Primärkaries betroffenen Bildareale in einer für diesen Zweck erstellten und validierten Software [24]. Ein Master-Reviewer, der sämtliche Bildbewertungen sehen konnte, überprüfte alle annotierten Bilder und überarbeitete sie bei Bedarf. Nach Vereinigung und Überprüfung der verbleibenden Pixelbewertungen eines jeden Bildes wurde das endgültige Klassifikationsreferenzset durch Binärisierung (Karies vorhanden ja/nein) hergestellt. Eine Instruktion der bewertenden Zahnärzt*innen erfolgte durch die Studienleitung, zudem stand ein Handbuch zur Verfügung, das den Annotationsvorgang erklärte. Eine Kalibrierung fand auf einem separaten Testdatensatz von 15 NILT Bildern statt.

3.4 Leistungsmetriken

Zur Messung der Leistungen der KI-Modelle auf den jeweiligen Trainingsdatensätzen wurden sechs Metriken verwendet:

F1-Score: F-Maß gibt Auskunft über Genauigkeit eines Tests bei statistischer Analyse binärer Klassifizierungen. F1 = 2TP / (2TP + FP + FN)

Sensitivität: Wahr-positiv Rate, richtig als kariöse Läsion erkannt.

Spezifität: Fasch-positiv Rate, fälschlich als kariöse Läsion erkannt.

Positiver Vorhersagewert (PVW): Anzahl der richtig positiven Treffer dividiert durch die Summe aller positiven Treffer. PVW = TP / (TP + FP)

Negativer Vorhersagewert (NVW): Anzahl der richtig negativen Treffer dividiert durch die Summe aller negativen Treffer. NVW = TN / (TN + FN)

Area-under-the Operating Characteristic Curve (AUC): Zweidimensionale Fläche unter der ROC-Kurve.

3.5 Stichprobengrößen

Für beide generierten Datensätze wurde keine Stichprobenziehung durchgeführt, sondern der gesamte Datensatz verwendet. Eine Berechnung der Stichprobengröße wurde daher nicht durchgeführt.

3.6 Datenaufbereitung, Modell und Training

Für die Klassifizierung wurde ein Residuals CNN (ResNet-34) verwendet. Dieses CNN nutzt RGB-Bilder als Eingabe und gibt Wahrscheinlichkeiten für binäre Klassen aus (hier: gesunde und kariöse Zähne). Die Architektur von ResNet besteht aus Stapeln von ResNet-Blöcken, einem CNN als Merkmalsextraktor und ein voll verbundener Klassifikationskopf mit binärem Ausgang. ResNet ist ein residuales Netz; im Unterschied zu traditionellen neuronalen Netzen werden Blöcke verwendet [25]: In herkömmlichen neuronalen Netzen mündet jede Schicht in die nächste Schicht. In einem Netz mit Residualblöcken speist jede Schicht in die nächste Schicht und dringt direkt in die etwa 2-3 Schritte entfernten Schichten ein (Abbildung 6). Wir verwendeten die Merkmalsextraktionsblöcke aus einem auf dem ImageNet Datensatz der Pytorch-Bibliothek vortrainierten Modell.

Zur Augmentation wurden während des Trainings zufällige Rotationen, vertikales und horizontales Flipping, Shifting und Zooming angewendet. Die Bilder wurden auf 224 × 224 × 3 Tensoren verkleinert, was sich als ausreichende Auflösung erwies. Die Leistung des Modells auf den In-vivo- und In-vitro-Datensätzen wurde mittels k-facher Kreuzvalidierung mit jeweils 10 Trainings-, Validierungs- und Test-Splits bewertet. Wenn die Trainings- und Testdaten aus unterschiedlichen Datensätzen stammten (z. B. Training auf In-vivo- und Test auf In-vitro-Daten), wurden alle Bilder aus der Trainingsquelle für das Training des Modells verwendet. Die Generalisierbarkeit des Modells wurde dann durch Validierungs- und Test-Splits bewertet. Die Validierungs-Splits wurden während des Trainings ausgewertet, die Test-Splits wurden ausgewertet, nachdem die Modelle konvergiert hatten. Dieser Prozess wurde für 10 verschiedene Splits wiederholt.

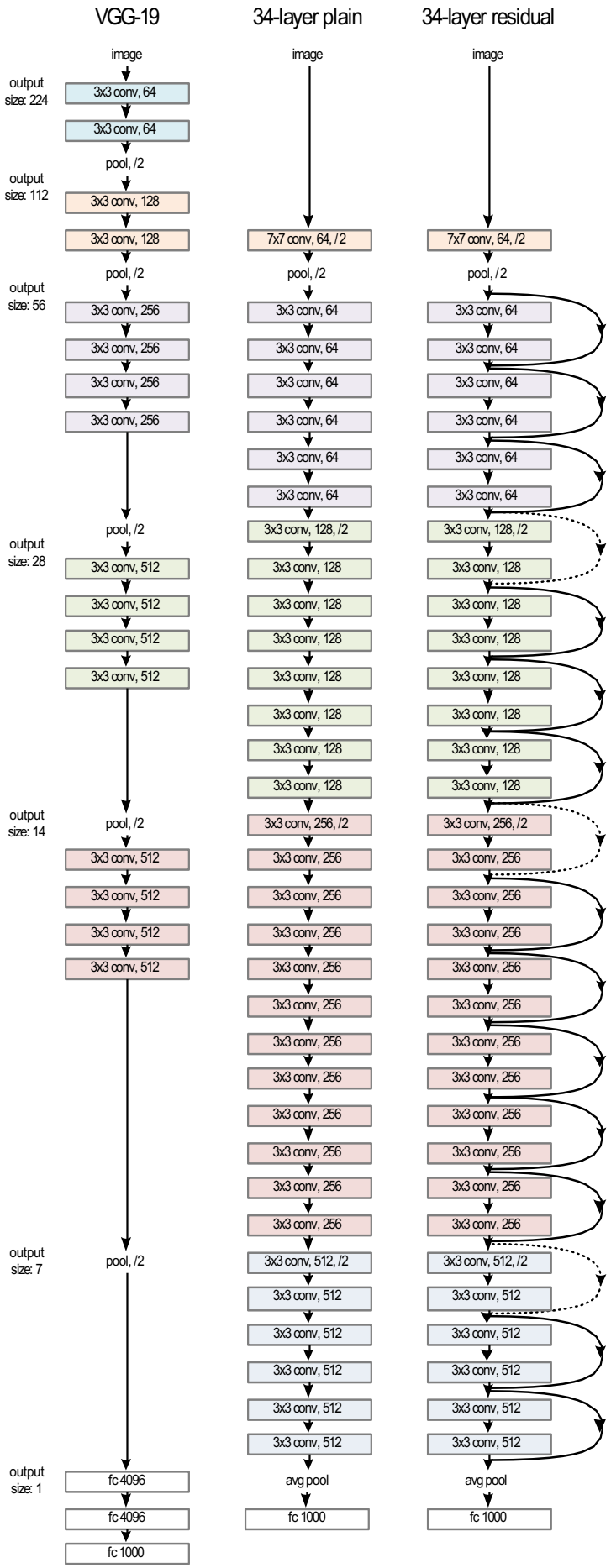

Abbildung 6: Schematische Darstellung von ResNet-34 (rechts), einem 34 Schichten tiefen Residualnetzwerk, im Vergleich mit einem weniger tiefen, älteren Netzwerk VGG-19 (links) sowie einem nicht residualen, 34-Schichten tiefem Netzwerk (Mitte). Bei Residualnetzwerken werden Schichten in den ersten Trainingsphasen übersprungen (Pfeile), wodurch das Training weniger ressourcenintensiv ist. Quelle: He K, Zhang X, Ren S, Sun J. 2016. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). arXiv:1512.03385 [25]. Bildrechte beim IEEE. Verwendung des Bildes bedarf keiner formellen Lizenz.

Jeder Split wurde für 200 Epochen mit dem Adam-Optimierer trainiert. Der binäre Kreuzentropieverlust auf den Validierungsdaten wurde während des Trainings mit einem Parameter von 20 Epochen ermittelt, nachdem ein frühes Stoppen angewandt wurde. Der Mittelwert und die Standardabweichung der oben genannten Leistungsmetriken wurde über alle Testsätze hinweg ausgewertet.

Aufgrund der geringen Größe des In-vitro-Datensatzes war es eine Herausforderung, eine geeignete Kombination aus Chargengröße und Lernrate festzulegen, die ein stabiles Verhalten beim Training mit diesem Datensatz ergab. Es wurden Lernraten von 5×10-7, 5×10-6, 5×10-5 und 5×10-4 und Stapelgrößen von 4, 8, 16, 32 bzw. 64 angewandt. Für jede mögliche Kombination wurde eine Gittersuche durchgeführt. Die beste Kombination lautete 16x5×10-6 (Training/Test in vitro) sowie 8x5×10-5 für alle anderen Datensätze. Die Modelle wurden auf einer NVIDIA Quadro RTX 6000 Grafikkarte (NVIDIA, Santa Clara, USA) trainiert.

3.7 Erklärbarkeit

Ein weiteres Ziel war es, die Modellentscheidung nachvollziehbar zu machen und hierdurch potentielle Fehlerquellen identifizieren zu können. Zu diesem Zweck wurden verschiedene Visualisierungsalgorithmen entwickelt, u. a. Grad-CAM [26]. Dieser Algorithmus erstellt eine Visualisierung der entscheidungsrelevanten Bildareale, eine sogenannte „saliency map“. Diese wird als gewichtete Kombination der Feature Maps einer bestimmten Schicht berechnet, gefolgt von einer ReLU-Aktivierung. Die Koeffizienten der Kombination werden als Durchschnitt der Gradienten der Ausgangsklasse in Bezug auf die Feature Maps errechnet.

3.8 Statistische Auswertung

Zum Vergleich der Modellleistungen wurden unabhängige 2-seitige t-Tests eingesetzt, wobei $p < 0,05$ als statistisch signifikant definiert wurde. Zur Analyse wurde die Python-Bibliothek SciPy 1.5.2 genutzt.

4. Ergebnisse

Die Prävalenz von Kariesläsionen auf Zahnebene betrug 41 % in vitro und 49 % in vivo. Die auf In-vivo-Daten trainierten Modelle schnitten signifikant besser ab, wenn sie auf den Bildern desselben Datensatzes getestet wurden (mittlere ± Standardabweichung: Genauigkeit und AUC: 0,78 ± 0,04), als die in vitro trainierten Modelle (Genauigkeit: 0,64 ± 0,15, AUC: 0,65 ± 0,12, $p < 0,05$). Innerhalb jeder Analyse waren Sensitivität und Spezifität ähnlich, während der NVW signifikant höher als der PVW war.

Tabelle 1: Genauigkeitsmetriken (Mittelwerte ± Standardabweichung) von Modellen, die in vitro und/oder in vivo trainiert und getestet wurden.

Training→ Test	***Genauigkeit***	***F1-Score***	***AUC***	***Sensitivität***	***Spezifität***	***PVW***	***NVW***
in vitro	*0,64 ± 0,15*	*0,57 ± 0,16*	*0,65 ± 0,12*	*0,66 ± 0,12*	*0,64 ± 0,21*	*0,55 ± 0,22*	*0,76 ± 0,09*
in vivo	*0,78 ± 0,04*	*0,73 ± 0,04*	*0,78 ± 0,04*	*0,76 ± 0,06*	*0,79 ± 0,05*	*0,70 ± 0,05*	*0,84 ± 0,03*
Training in vitro, Test in vivo	*0,61 ± 0,04*	*0,52 ± 0,03*	*0,60 ± 0,04*	*0,55 ± 0,03*	*0,65 ± 0,07*	*0,49 ± 0,05*	*0,70 ± 0,03*
Training in vivo, Test in vitro	*0,70 ± 0,01*	*0,56 ± 0,03*	*0,66 ± 0,01*	*0,49 ± 0,06*	*0,83 ± 0,04*	*0,67 ± 0,04*	*0,71 ± 0,03*

AUC Area-under-the-curve. PVW/NVW: Positiver/negativer Vorhersagewert. Quelle: Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. 2021. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images, J Clin Med 10(5):961 [27]. Bildrechte bei den Autoren der Publikation, u.a. Holtkamp A, nach Lizenz CC BY 4.0.

Beim Test in vitro zeigten die in vivo trainierten Modelle eine signifikant niedrigere Genauigkeit (0,70 ± 0,01) und AUC (0,66 ± 0,01) ($p < 0,01$). In ähnlicher Weise zeigten die in vitro trainierten Modelle beim Test in vivo eine signifikant niedrigere Genauigkeit (0,61 ± 0,04) und AUC (0,60 ± 0,04) ($p < 0,05$). In beiden Fällen war dies auf eine Abnahme der Sensitivität zurückzuführen (um -27 % für in vivo trainierte Modelle und -10 % für in vitro trainierte Modelle). Die Spezifitätswerte änderten sich nicht oder stiegen sogar leicht an beim Testen auf dem jeweils anderen Datenmaterial. In vitro trainierte und in vivo getestete Modelle wiesen keine nützliche Sensitivität mehr auf (Abbildung 7).

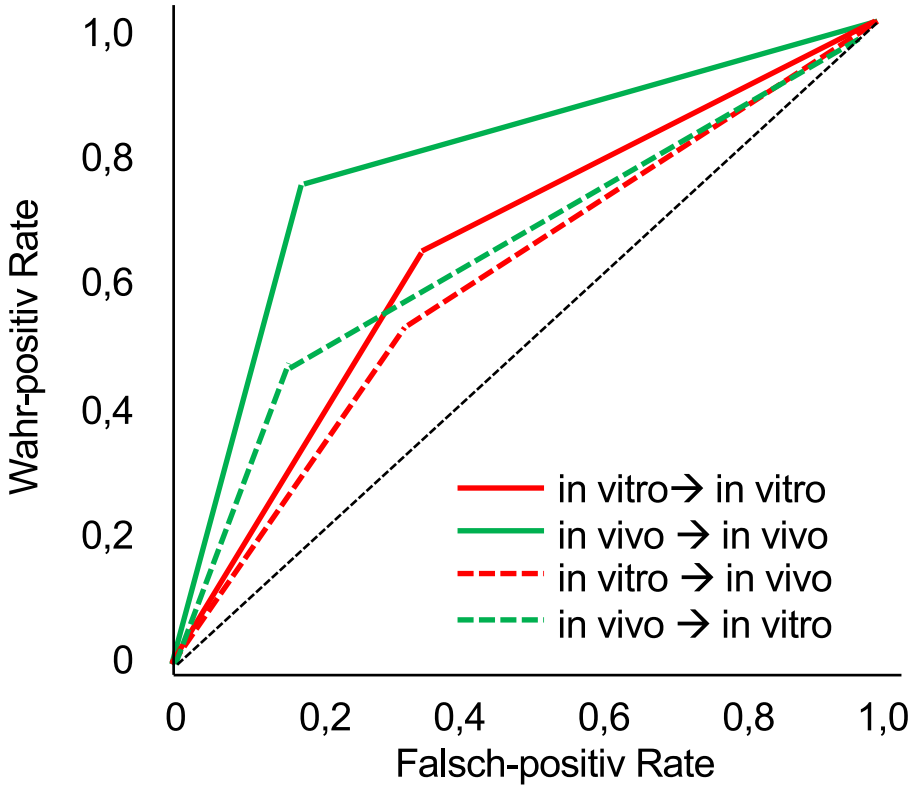

Abbildung 7: ROC-(Receiver Operating Characteristic) Kurven für Modelle, die in vivo bzw. in vitro trainiert und an Daten aus derselben oder der anderen Datenquelle getestet wurden. Die sich daraus ergebenden AUC-Werte sind in Tabelle 1 zu finden. Quelle: Eigene Darstellung, modifiziert nach Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. 2021. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images, Journal of Clinical Medicine 10(5):961 [27]. Bildrechte bei den Autoren der Publikation, u. a. Holtkamp A, nach Lizenz CC BY 4.0.

Bei der Bewertung der für die Klassifikationsentscheidung relevanter Bereiche zeigte sich, dass für wahr-positive Erkennungen die relevantesten Pixel auch diejenigen waren, die von Zahnärzt*innen ebenso als Karies deklariert worden waren. Falsch-positive Erkennungen auf In-vivo-Bildern wurden oft mit Restaurationen in Verbindung gebracht, die ein ähnliches Aussehen wie kariöse Läsionen hatten. In vitro war es nicht immer möglich, die Gründe für die Entscheidung der Modelle zu identifizieren. Bei falsch-negativen Erkennungen wurde deutlich, dass die Modelle Bereiche berücksichtigten, die Zahnärzt*innen weitgehend für irrelevant bei der Kariesdiagnostik halten würden, was auf ein Aufmerksamkeitsproblem hindeutet (Abbildung 8).

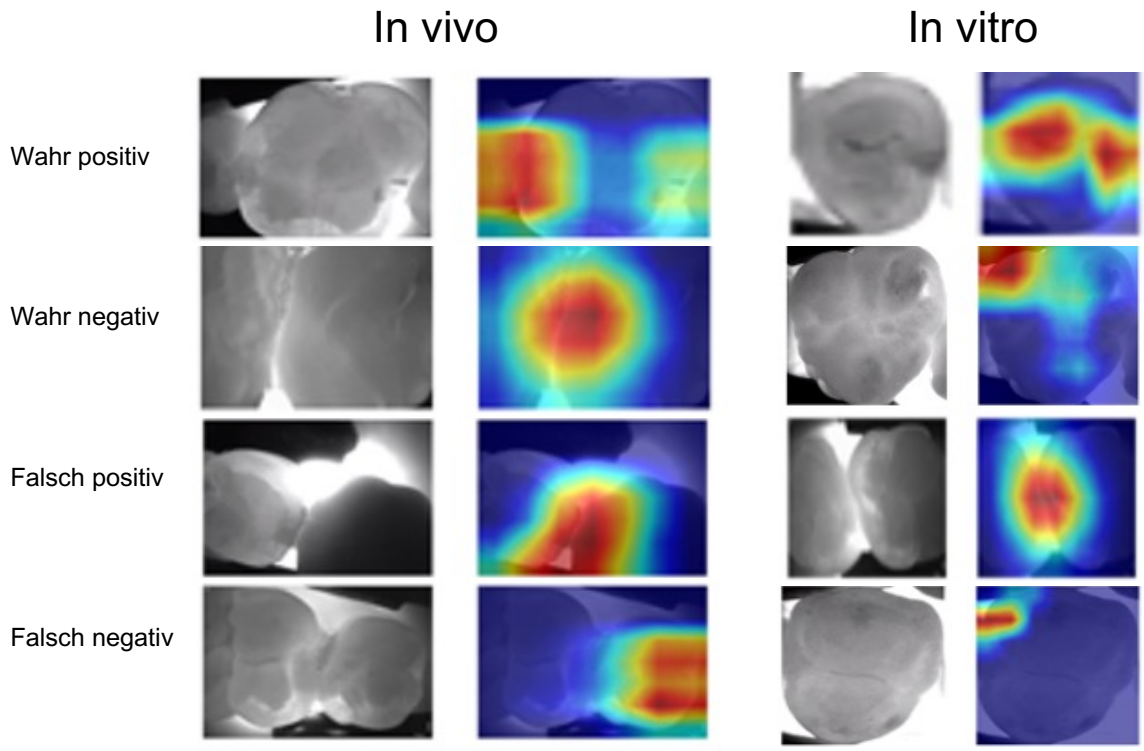

Abbildung 8: Visualisierung der für die Klassifikationsentscheidung relevanter Bildareale. Dargestellt sind die Originalbilder (in vivo, in vitro) und die für die KI-Entscheidung relevanten Bildareale als Heatmaps (gelb bis rot hervorgehoben). Quelle: Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. 2021. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images. Journal of Clinical Medicine. 10(5):961 [27]. Bildrechte bei den Autoren der Publikation, u.a. Holtkamp A, nach Lizenz CC BY 4.0.

5. Diskussion

5.1 Zusammenfassung der Ergebnisse

Karies ist die weltweit häufigste chronische Erkrankung [28]. Das ist bemerkenswert, ist sie doch weitgehend vermeidbar. Karies kann in frühen Stadien wenig invasiv und mit begrenztem Aufwand behandelt werden. Diesbezüglich hat in den letzten Jahrzehnten ein Paradigmenwechsel stattgefunden. Statt der vergleichsweise teuren und aufwändigeren invasiv-restaurativen Therapie fortgeschrittener Karies, steht nun die Verhinderung und frühe, weniger bzw. minimal-invasive Behandlung, die so genannte Primär- bzw. Sekundärprävention im Vordergrund. Die Präventionserfolge sind in epidemiologischen Erhebungen wie den Deutschen Mundgesundheitsstudien u. a. im Rückgang fortgeschrittener (kavitierter) Karies bei Kindern und jungen Erwachsenen messbar. Allerdings zeigt sich eine überproportionale Verschiebung der Karieslast in soziale Risikogruppen (Polarisierung) mit niedrigem sozioökonomischem Hintergrund und Bildungsstatus. Eine zusätzliche Beobachtung ist, dass bei dem Großteil der Kinder und jungen Erwachsenen, Karies nicht vollständig verhindert, sondern eher verlangsamt wird. Das bedeutet, dass eine Mehrheit der Kinder und jungen Erwachsenen heute zwar weniger fortgeschrittene (kavitierte), dafür aber mehr frühe (nicht-kavitierte) Läsionen zeigen [29] [30]. Ausgehend von den genannten Aspekten, Polarisierung und vermehrt auftretenden frühen kariösen Läsionen, gilt es, die Hochrisikoindividuen zu identifizieren, frühe Kariesstadien zu detektieren und anschließend adäquat zu therapieren. Folglich kommt der Detektion früher kariöser Läsionen eine große Bedeutung zu. In der zahnärztlichen Routine wird bisher hierfür die beschriebene röntgenologische Diagnostik, in der Regel mit so genannten Bissflügel-Röntgenaufnahmen genutzt [8]. Bissflügelröntgen steht wie dargelegt, nur begrenzt häufig, in bestimmten Altersgruppen und nur in einem klar umschriebenen Setting einer Zahnarztpraxis zur Verfügung. Es wird ionisierende Strahlung durch nicht oder schwer transportable Ausstattung benötigt. Bei jüngeren Individuen steht zur Untersuchung in engeren zeitlichen Abständen oder zum Einsatz in Lebenswelten wie Kindergärten, Schulen und Betrieben, eine Röntgeneinrichtung nicht zur Verfügung. Gerade der Einsatz in Lebenswelten ist allerdings notwendig, um frühe Läsionen bei Hochrisikoindividuen zu identifizieren, da diese Personen seltener oder gar nicht präventiv eine zahnärztliche Praxis aufsuchen [29].

Eine mögliche Alternative zur Detektion früher Karies ist NILT. Dieses Verfahren bedient sich nicht ionisierender Strahlung. Die notwendigen Geräte sind portabel und die Diagnostik beliebig oft wiederholbar. Es können sowohl Fotos als auch Videos erzeugt werden. Zudem erfolgt die Evaluation unmittelbar unter Bewegung des Detektionskopfes. In (Abbildung 9) ist ein Anwender dargestellt, der entsprechende Daten am Patienten erhebt und die folgende Diagnostik, unterstützt durch KI erstellt. Das Ergebnis der Auswertung hilft bei der Diagnosefindung und dem Therapieentscheid, wobei letzterer in Rücksprache mit dem Patienten erfolgt. Die Validität von NILT zur Detektion von Primär- und Sekundärkaries im Milch- und bleibenden Gebiss ist gut dokumentiert. Die Sensitivität und Spezifität ist ähnlich wie die von Bissflügelröntgenaufnahmen [15] [31] [11]. Demnach ist Kariesdetektion mittels NILT gerade für den Einsatz in verschiedenen Lebenswelten prädestiniert und vielversprechend. Die Evaluation des gewonnenen NILT-Bildmaterials ist jedoch Untersucher- und erfahrungsabhängig und teilweise wenig zuverlässig. KI kann genau hier möglicherweise Abhilfe schaffen.

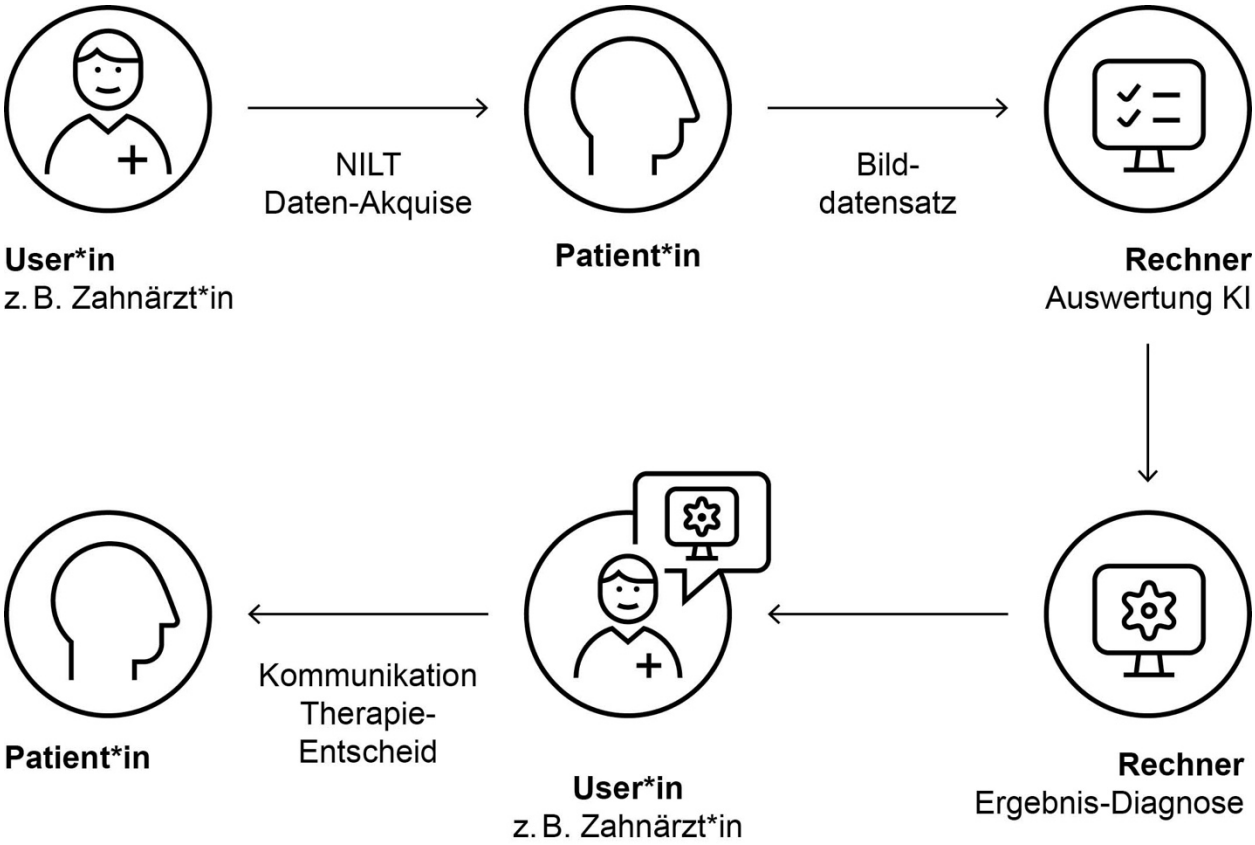

Abbildung 9: Darstellung einer Userin, eines Users bei der Anwendung von NILT an der Patientin, am Patienten, mit anschließender Auswertung durch KI. Quelle: Eigene Darstellung

Es ist allerdings noch unklar, ob die Kariesdetektion mittels NILT in Lebenswelten praktikabel ist. So ist in Schulen oder Kindergärten nur wenig Zeit für die routinemäßig visuell-taktil durchgeführte Kariesdiagnostik. Würde NILT bei der Diagnostik als bildgebendes Verfahren Anwendung finden, hätten Zahnärzt*innen im öffentlichen Dienst folglich den doppelten Arbeitsaufwand für ihre Befunderhebung. Durch den strahlenfreien Einsatz von NILT, kann die Diagnostik allerdings auch von fachfremdem Personal angewandt werden. Ob eine erweiterte Kariesdetektion sinnvoll ist und Einfluss auf das Verhalten der Eltern in Bezug auf die Zahnpflege ihrer Kinder oder das Intervall der Zahnärzt*innenbesuche nimmt, sollte in weiteren Studien untersucht werden.

5.2 Interpretation der Ergebnisse

KI-Modelle werden zunehmend zur Analyse und Entscheidungsunterstützung in medizinischen Bereichen eingesetzt. Hierbei kommen vor allem, wie auch in der vorliegenden Studie, neuronale Netzwerke zum Einsatz. Diese erlauben durch ihre hochkomplexe intrinsische Struktur eine Repräsentation des gelernten Bildmaterials. Sie sind in der Lage, Bildmaterial auswendig zu lernen und werden als „universelle Approximationsmaschinen“ angesehen [32]. Wichtiger als das Lernen des gesehenen Bildmaterials ist jedoch die Anwendung auf bisher unbekanntem, neuen Bildmaterial. Nur KI-Modelle, die auch eine hohe Genauigkeit auf ungesehenem Bildmaterial aufweisen, wie beispielsweise auf einem unabhängigen Testdatensatz, sollten auch klinisch eingesetzt werden. Die Wichtigkeit eines solchen Testens auf unabhängigem Datenmaterial wird in verschiedenen Standards und Normen, u. a. zur Durchführung und zum Berichtswesen von Studien im Bereich der KI in der Medizin und Zahnmedizin, betont [23] [33].

Die Generalisierbarkeit von KI-Modellen in der Medizin und Zahnmedizin ist hingegen weniger untersucht. Die mangelnde Generalisierbarkeit von statistischen Modellen zur Vorhersage von Zahnverlusten ist bereits diskutiert worden. Erste Pilotstudien zur Bildanalytik der Zahnmedizin deuten darauf hin, dass eine Generalisierbarkeit von Modellen, die auf Daten einer Kohorte trainiert worden sind und nun auf einer anderen Kohorte getestet werden, nicht zwingend gegeben ist. Verschiedene Faktoren werden dafür verantwortlich gemacht. Exemplarisch sind die Bildqualität und -eigenschaften, genutzte Hardware zur Generierung der Daten, soziodemografische Unterschiede zwischen den Kohorten und Prävalenz der Pathologie zu nennen. Eine kürzlich publizierte Studie zeigt

zudem, dass der Zahnstatus und die Zahl und Art der Restaurationen einen Einfluss haben kann, da Restaurationen als Marker vergangener Behandlungen mit dem Vorhandensein von Pathologien korrelieren [20].

In der vorliegenden Studie wurde die Frage dieser Generalisierbarkeit aus einer anderen Perspektive beleuchtet. Im Speziellen wurde untersucht, ob Modelle, die an Bilddaten aus In-vitro-Datensätzen gewonnen wurden, auch auf In-vivo-Bilddaten anwendbar sind und vice versa. Diese Frage stellt sich insbesondere bei Studien zur Kariesdiagnostik, bei der ein harter Goldstandard oft nur durch laborexperimentelle Verfahren wie Histologie, Mikroradiografie oder µ-CT hergestellt werden kann [34] [35]. Gewinnt man die Daten aus der zahnärztlichen Routine, also in vivo, ist dies de facto nicht möglich. Stattdessen könnten visuell-taktile Befunde oder ein zweites Bildmaterial auf Basis eines DVT oder Bissflügelröntgenaufnahmen verwendet werden. Durch Triangulation kann die Robustheit des Goldstandards so unter Umständen erhöht werden. Alternativ erfolgt üblicherweise eine Mehrfachbewertung desselben Bildes durch unterschiedliche Behandler*innen, um die mangelnde inter- bzw. intraindividuelle Reliabilität der Untersucher*innen zu kompensieren.

5.3 Stärken und Schwächen der Studie

Auf Basis der gewonnenen Daten in der vorliegenden Arbeit zeigte sich, dass eine Generalisierbarkeit von KI-Modellen auf sowohl In-vitro- als auch In-vivo-Daten nicht gegeben ist. Unsere Hypothese müssen wir ablehnen. Darüber hinaus muss festgestellt werden, dass beim Transport der Modelle auf das jeweils andere Bildmaterial der Verlust an Genauigkeit, vor allem an Sensitivität, teils dramatisch war. Bei einer Sensitivität unter 50 % war mitunter eine Nützlichkeit des Modells nicht mehr gegeben, d.h. sie war folglich schlechter als einfaches Raten. Zudem konnten wir zeigen, dass KI-Modelle auf In-vivo-Daten generell höhere Genauigkeiten aufwiesen als jene, die in vitro trainiert wurden. Es kann allerdings nicht ausgeschlossen werden, dass dies durchaus ein Effekt der unterschiedlich zur Verfügung stehenden Fallzahlen sein kann, denn der vorliegende In-vivo- war deutlich größer als der vorliegende In-vitro-Datensatz. Weitere Experimente, bei denen Modelle z. B. auf standardisiert großen Datensätzen trainiert werden, könnten hier neue Einsichten geben.

Es bleibt bisher unklar, warum eine Generalisierbarkeit im Rahmen dieser Studie nicht gegeben war. Zwar konnten unsere Ansätze helfen zu erklären, warum KI-Modelle teilweise unter- oder überdetektieren, eine Aufklärung der Limitation der Generalisierbarkeit war jedoch nicht möglich. Es ist denkbar, dass beispielsweise die Bildqualität der unterschiedlich generierten Bildmaterialien einen Einfluss auf die Genauigkeit der Modelle hatte. Dies scheint insbesondere für NILT-Bilder plausibel: In vitro wurden die Daten unter Rückgriff auf ein Simulationssystem erzeugt, bei dem der Alveolarfortsatz und seine Durchstrahlungseigenschaften durch ein transparentes Akrylharz modelliert waren. Dadurch wird wiederum eine Transillumination des Zahnes von apikal nach okklusal möglich. Obwohl dieses Verfahren in einer vorhergehenden Studie aus Sicht menschlicher Untersucher geeignet war Bilder zu erzeugen, die klinischen Bildern ähnlich sind, ist es doch wahrscheinlich, dass aus Sicht des Maschinellen Lernens und der Bildanalyse mit CNNs relevante Bildeigenschaften von In-vitro-Bildern fehlen. Das ist bedauerlich, da das Testen von KI-Modellen zur NILT-Analyse auf In-vitro-Daten durchaus sinnvoll wäre: Gerade auf diesen Daten kann ein harter Goldstandard basieren. Eine „Kreuzkombination“ eines Trainings in vivo und Testens in vitro könnte so durchgeführt werden. Ausgehend von der vorliegenden Studie sollte hingegen auch weiterhin auf in vivo generiertes Datenmaterial gesetzt werden. Dieses steht, sofern NILT routinemäßig eingesetzt wird, in großer Zahl zur Verfügung. Es sollte dann jedoch möglichst mit einer zweiten Bildquelle oder einem Test kombiniert werden. Auch in der vorliegenden Studie standen z.B. Bissflügelaufnahmen zur Verfügung. Diese werden in zukünftigen Untersuchungen für einen solchen Triangulationsansatz eingesetzt.

Zusätzlich zur Generalisierbarkeit wurde in der vorliegenden Untersuchung die Erklärbarkeit der Modelle beleuchtet. Dies ist hilfreich, um Verzerrungen (Bias) zu identifizieren. Diese Verzerrungen können wie bereits dargelegt, teilweise gewünscht sein. Beispielsweise verbessert die Ausnutzung der oben genannten Korrelationsstrukturen zwischen Pathologien und restaurativen Status vermutlich die Modellgenauigkeit. Andererseits ist zu beachten, dass die so verzerrten Modelle nur in jenen Kohorten eingesetzt werden sollten, in denen der restaurative Status ähnlich ist. In anderen Kohorten würde die Genauigkeit geringer und eine Generalisierbarkeit nicht gegeben sein. In den meisten Fällen werden Verzerrungen jedoch unerwünscht sein, vor allem, wenn bildanalytische Modelle Klassifikationsentscheidungen treffen, die nicht auf medizinisch logischen Eigenschaften,

sondern Artefakten beruhen. Dies wurde bereits klar belegt, weshalb „erklärbare KI“ (explainable AI/XAI) zunehmend auch in den Standards und Normen gefordert wird [21]. Erklärbarkeit kann dann wiederum auch genutzt werden, um das Vertrauen des anwendenden Personenkreises in die KI zu unterstützen. Dazu muss gezeigt werden, dass die Entscheidungsgrundlagen der KI mit dem medizinischen Wissen konsistent sind. Die durchgeführten Versuche zu XAI zeigten, dass falsch-negative Erkennungen oftmals das Resultat von „Aufmerksamkeitsdefiziten“ war. Die Modelle berücksichtigten hierbei oft Bereiche, in denen nur selten Karies vorkommt und Zahnärzt*innen diese nicht vermuten würden. Bei falsch-positiven Erkennungen stellten wir fest, dass das Vorhandensein von Füllungen die Entscheidungen der Modelle beeinflusste. Dies kann auf teilweise ähnliche Brechungsphänomene in kariösem Schmelz und Füllungsmaterial zurückgeführt werden. Ähnliches findet sich auch in der Röntgenbilddiagnostik, wenn radioluzente, zahnärztliche Restaurationsmaterialien benutzt werden. Inwieweit eine Erhöhung der Fallzahl oder die Berücksichtigung der klinischen Behandlungshistorie hier Abhilfe schaffen könnten, bleibt bisher jedoch noch unklar.

Die Zahl der Studien, die sich mit der Generalisierbarkeit auf dem Gebiet von KI-Modellen in der Zahnmedizin auseinandersetzen ist gering. Die vorliegende Studie trägt zum Wissensgewinn signifikant bei. Die hier verwendete XAI-Methode erlaubt Rückschlüsse auf die Entscheidungsfindung der Modelle. Damit stärkt sie das Vertrauen in die Ergebnisse und gibt Einblicke in die Gründe falscher Klassifizierungen. Die geringe Stichprobengröße der Datensätze und die diskutierte, unterschiedlich hohe Fallzahl sind kritisch zu bewerten. Die Problematik der Etablierung eines Goldstandards wurde diskutiert. Eine Triangulation mit klinischen Methoden wie oben beschrieben sollte angestrebt werden. Da die Prävalenz kariöser Läsionen relativ hoch war, erleichterte das zwar einerseits das Training der Modelle, kann aber anderseits zu Verzerrungen führen. In der Realität könnte Karies hingegen nur auf 5-10 % der Flächen kommen [36] [37], was das Modell vor deutlich höhere Anforderungen stellen würde – es müsste dann auch sehr hohe Genauigkeiten aufweisen. In der vorliegenden Studie sind schon Modelle, die eine Genauigkeit von 60 % vorweisen nützlich, da sie genauer sind als das Raten der Mehrheitsklasse. Bei niedrigerer Prävalenz würde die Genauigkeit auf über 90 % steigen [38] [39]. Da die Analyse anhand von Zahnsegmentbildern erfolgte, sind Clustereffekte und damit verbundene Korrelationsstrukturen nicht berücksichtigt [40]. Zusätzlich ist der vermutlich positive Ein-

fluss von mehr Bildkontext, d. h. es sind mehr als ein Zahn abgebildet, auf die Modellgenauigkeit nicht erfasst worden [41]. Weitere Studien sollten den Nutzen von KI-Modellen für die NILT-Diagnostik, den Einfluss auf die klinische Entscheidungsfindung sowie die Kosten-Nutzen-Relation im Vergleich zu konventionellen Methoden eines solchen Einsatzes, die nicht Gegenstand der vorliegenden Untersuchungen waren, evaluieren.

Auch die Kombination von NILT mit anderen Datenquellen könnte hierbei eine Rolle spielen. So wird NILT heute in erste Intraoralscanner integriert; diese zeichnen zusätzlich Fotografien und 3-D-Oberflächenscans auf. Der regelmäßige Einsatz eines solchen Scanners würde zu longitudinalen, multimodalen Datensätzen führen. Die Beurteilung von Fotografien, Oberflächendaten und NILT in Kombination könnte nicht nur eine bessere Diagnostik, sondern auch Prognostik ermöglichen – bisher sind jedoch multimodale Prognosemodelle unter Einsatz von verschiedenen Bilddaten in der Zahnmedizin nicht erprobt.

6. Schlussfolgerungen

Die Verwendung von In-vitro-Datensätzen für die Erzeugung von NILT-Bildern und das Training von CNNs ist ungenau und wenig generalisierbar. In vivo trainierte und getestete Modelle zeigten eine höhere Genauigkeit. Sie haben aber nur eine begrenzte Verallgemeinerbarkeit, wenn sie an In-vitro-Daten getestet wurden. Studien, die In-vitro-Bildmaterial zur Entwicklung von Deep-Learning-Modellen verwenden, sollten kritisch auf ihre Generalisierbarkeit geprüft werden. Es ist empfehlenswert, geeignete Deep-Learning-Modelle zur Beurteilung von NILT-Bildern auf In-vivo-Daten zu trainieren.

Literaturverzeichnis

1. Bernabe E, Marcenes W, Hernandez CR, Bailey J, Abreu LG, Alipour V, Amini S, Arabloo J, Arefi Z, Arora A, Ayanore MA, Bärnighausen TW, Bijani A, Cho DY, Chu DT, Crowe CS, Demoz GT, Demsie DG, Dibaji Forooshani ZS, Du M, El Tantawi M, Fischer F, Folayan MO, Futran ND, Geramo YCD, Haj-Mirzaian A, Hariyani N, Hasanzadeh A, Hassanipour S, Hay SI, Hole MK, Hostiuc S, Ilic MD, James SL, Kalhor R, Kemmer L, Keramati M, Khader YS, Kisa S, Kisa A, Koyanagi A, Lalloo R, Le Nguyen Q, London SD, Manohar ND, Massenburg BB, Mathur MR, Meles HG, Mestrovic T, Mohammadian-Hafshejani A, Mohammadpourhodki R, Mokdad AH, Morrison SD, Nazari J, Nguyen TH, Nguyen CT, Nixon MR, Olagunju TO, Pakshir K, Pathak M, Rabiee N, Rafiei A, Ramezanzadeh K, Rios-Blancas MJ, Roro EM, Sabour S, Samy AM, Sawhney M, Schwendicke F, Shaahmadi F, Shaikh MA, Stein C, Tovani-Palone MR, Tran BX, Unnikrishnan B, Vu GT, Vukovic A, Warouw TSS, Zaidi Z, Zhang ZJ, Kassebaum NJ. Global, Regional, and National Levels and Trends in Burden of Oral Conditions from 1990 to 2017: A Systematic Analysis for the Global Burden of Disease 2017 Study. Journal of dental research. 2020;99(4):362-73.
2. Schwendicke F, Splieth C, Breschi L, Banerjee A, Fontana M, Paris S, Burrow MF, Crombie F, Page LF, Gatón-Hernández P, Giacaman R, Gugnani N, Hickel R, Jordan RA, Leal S, Lo E, Tassery H, Thomson WM, Manton DJ. When to intervene in the caries process? An expert Delphi consensus statement. Clinical oral investigations. 2019;23(10):3691-703.
3. Paris S M-LH, Kim R. Ekstrand. Karies. Wissenschaft und Klinische Praxis2012. 73 p.
4. Raedel M, Hartmann A, Bohm S, Priess HW, Samietz S, Konstantinidis I, Walter MH. Four-year outcomes of restored posterior tooth surfaces-a massive data analysis. Clinical oral investigations. 2017;21(9):2819-25.
5. Splieth CH, Kanzow P, Wiegand A, Schmoeckel J, Jablonski-Momeni A. How to intervene in the caries process: proximal caries in adolescents and adults-a systematic review and meta-analysis. Clinical oral investigations. 2020;24(5):1623-36.
6. Brantley CF, Bader JD, Shugars DA, Nesbit SP. Does the cycle of rerestoration lead to larger restorations? J Am Dent Assoc. 1995;126(10):1407-13.

7. Qvist V. Longevity of restorations - "the death spiral". In Dental caries - The disease and its clinical management. Dental caries - The disease and its clinical management. 2008:443-5.
8. Schwendicke F, Tzschoppe M, Paris S. Radiographic caries detection: A systematic review and meta-analysis. J Dent. 2015;43(8):924-33.
9. Walsh T, Macey R, Riley P, Glenny AM, Schwendicke F, Worthington HV, Clarkson JE, Ricketts D, Su TL, Sengupta A. Imaging modalities to inform the detection and diagnosis of early caries. Cochrane Database Syst Rev. 2021;3(3):Cd014545.
10. Wenzel A. Bitewing and digital bitewing radiography for detection of caries lesions. Journal of dental research. 2004;83 Spec No C:C72-5.
11. Kühnisch J, Söchtig F, Pitchika V, Laubender R, Neuhaus KW, Lussi A, Hickel R. In vivo validation of near-infrared light transillumination for interproximal dentin caries detection. Clinical oral investigations. 2016;20(4):821-9.
12. Künisch J, Schaefer G, Pitchika V, Garcia-Godoy F, Hickel R. Evaluation of detecting proximal caries in posterior teeth via visual inspection, digital bitewing radiography and near-infrared light transillumination. Am J Dent. 2019;32(2):74-80.
13. Ortiz MIG, de Melo Alencar C, De Paula BLF, Magno MB, Maia LC, Silva CM. Accuracy of near-infrared light transillumination (NILT) compared to bitewing radiograph for detection of interproximal caries in the permanent dentition: A systematic review and meta-analysis. J Dent. 2020;98:103351.
14. Shaya SAA, Saeed MH, Kheder W, editors. Proximal Caries Detection in Permanent Teeth by Using DIAGNOcam : An in Vivo Study2018.
15. Elhennawy K, Askar H, Jost-Brinkmann PG, Reda S, Al-Abdi A, Paris S, Schwendicke F. In vitro performance of the DIAGNOcam for detecting proximal carious lesions adjacent to composite restorations. J Dent. 2018;72:39-43.
16. Schwendicke F, Golla T, Dreher M, Krois J. Convolutional neural networks for dental image diagnostics: A scoping review. J Dent. 2019;91:103226.
17. Casalegno F, Newton T, Daher R, Abdelaziz M, Lodi-Rizzini A, Schurmann F, Krejci I, Markram H. Caries Detection with Near-Infrared Transillumination Using Deep Learning. Journal of dental research. 2019;98(11):1227-33.

18. Schwendicke F, Elhennawy K, Paris S, Friebertshäuser P, Krois J. Deep learning for caries lesion detection in near-infrared light transillumination images: A pilot study. J Dent. 2020;92:103260.
19. Krois J, Graetz C, Holtfreter B, Brinkmann P, Kocher T, Schwendicke F. Evaluating Modeling and Validation Strategies for Tooth Loss. Journal of dental research. 2019;98(10):1088-95.
20. Krois J, Garcia Cantu A, Chaurasia A, Patil R, Chaudhari PK, Gaudin R, Gehrung S, Schwendicke F. Generalizability of deep learning models for dental image analysis. Sci Rep. 2021;11(1):6102.
21. Schwendicke F, Samek W, Krois J. Artificial Intelligence in Dentistry: Chances and Challenges. Journal of dental research. 2020;99(7):769-74.
22. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HC, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Bmj. 2015;351:h5527.
23. Mongan J, Moy L, Kahn CE, Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell. 2020;2(2):e200029.
24. Thomas Ekert JK, Falk Schwendicke. Building a mass online annotation tool for dental radiographic imagery. OpenReviewnet. 2018.
25. Kaiming He XZ, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition 2016.
26. Ramprasaath R. Selvaraju MC, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision. 2017.
27. Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images. J Clin Med. 2021;10(5).
28. Kassebaum NJ, Smith AGC, Bernabé E, Fleming TD, Reynolds AE, Vos T, Murray CJL, Marcenes W. Global, Regional, and National Prevalence, Incidence, and Disability-Adjusted Life Years for Oral Conditions for 195 Countries, 1990-

2015: A Systematic Analysis for the Global Burden of Diseases, Injuries, and Risk Factors. Journal of dental research. 2017;96(4):380-7.

29. Jordan RA, Bodechtel C, Hertrampf K, Hoffmann T, Kocher T, Nitschke I, Schiffner U, Stark H, Zimmer S, Micheelis W. The Fifth German Oral Health Study (Fünfte Deutsche Mundgesundheitsstudie, DMS V) - rationale, design, and methods. BMC Oral Health. 2014;14:161.
30. Schwendicke F, Dörfer CE, Schlattmann P, Foster Page L, Thomson WM, Paris S. Socioeconomic inequality and caries: a systematic review and meta-analysis. Journal of dental research. 2015;94(1):10-8.
31. Schwendicke F, Friebertshäuser, Philipp, Krois, Joachim, Elhennawy, Karim (Caries Detection on DIAGNOcam Images Using Convolutional Neural Networks IADR; 09/20/2019; Madrid2019.
32. KurtHornik. Approximation capabilities of multilayer feedforward networks. Neural Networks. 1991;4(2):251-7.
33. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. J Dent. 2021;107:103610.
34. Kühnisch J, Janjic Rankovic M, Kapor S, Schüler I, Krause F, Michou S, Ekstrand K, Eggmann F, Neuhaus KW, Lussi A, Huysmans MC. Identifying and Avoiding Risk of Bias in Caries Diagnostic Studies. J Clin Med. 2021;10(15).
35. Janjic Rankovic M, Kapor S, Khazaei Y, Crispin A, Schüler I, Krause F, Ekstrand K, Michou S, Eggmann F, Lussi A, Huysmans MC, Neuhaus K, Kühnisch J. Systematic review and meta-analysis of diagnostic studies of proximal surface caries. Clinical oral investigations. 2021;25(11):6069-79.
36. Lillehagen M, Grindefjord M, Mejàre I. Detection of approximal caries by clinical and radiographic examination in 9-year-old Swedish children. Caries Res. 2007;41(3):177-85.
37. Mejàre I, Stenlund H, Zelezny-Holmlund C. Caries incidence and lesion progression from adolescence to young adulthood: a prospective 15-year cohort study in Sweden. Caries Res. 2004;38(2):130-41.
38. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. A comparison of deep learning performance against health-care professionals in detecting diseases from

medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e97.

39. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. Bmj. 2020;368:m689.
40. Meinhold L, Krois J, Jordan R, Nestler N, Schwendicke F. Clustering effects of oral conditions based on clinical and radiographic examinations. Clinical oral investigations. 2020;24(9):3001-8.
41. Krois J, Schneider L, Schwendicke F. Impact of Image Context on Deep Learning for Classification of Teeth on Radiographs. J Clin Med. 2021;10(8).

Eidesstattliche Versicherung

„Ich, Agnes Holtkamp, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema: „Generalisierbarkeit von Deep-Learning-Modellen zur Detektion kariöser Läsionen“ bzw. „Generalizability of deep learning models for the detection of carious lesions“ 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. 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; www.icmje.og) 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

Anteilserklärung an der erfolgten Publikation

Agnes Holtkamp hatte folgenden Anteil an den folgenden Publikationen:

Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. 2021. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images. Journal of Clinical Medicine 10(5):961.

Impact Factor 2021: 4.242

Das Konzept und Design dieser Studie wurde von mir, Krois J, Elhennawy K, und Schwendicke F. entwickelt. Mein Beitrag zu dieser Publikation umfasste u.a. die Datenakquise aus der klinischen Routineuntersuchung. Patienten, die bei mir zur dentalen Befunderhebung vorstellig wurden, im Rahmen der Kariesdiagnostik sowohl mittels Röntgenbild als auch NILT routinemäßig untersucht. Die Kuratierung der Daten habe ich einheitlich, für jeden nachvollziehbar gestaltet. Die Unterweisung der Annotatoren erfolgte durch mich persönlich und ergänzend mit einer durch mich angefertigten Arbeitsanleitung. Die gesamte Organisation und Supervision der Bildannotation erfolgte ebenfalls durch mich. Der In-vivo-Datensatz wurde dann mit dem In-vitro-Datensatz zusammengeführt. Ich unterstützte die Erstellung der Machine-Learning-Modelle in Zusammenarbeit mit den Data Scientists der Abteilung. Die Ergebnisse des Trainierens und Testens der Modelle der unterschiedlichen Datensätze habe ich in der Publikation in Abb. 1 dargestellt. Sie beinhaltet die ROC-Kurven für Deep-Learning-Modelle, die an in vivo bzw. in vitro generierten Daten trainiert und an Daten aus derselben oder der anderen Datenquelle getestet wurden. Daraus ergaben sich die Werte für die Fläche unter der Kurve (AUC).

Um die beitragenden Merkmalskarten zu visualisieren, habe ich in Abb. 2 die Originalbilder und die markanten Bereiche dargestellt, die die Modelle als am relevantesten für ihre Entscheidung erachteten (gelb bis rot hervorgehoben). Diese Abbildung zeigt die in vivo bzw. in vitro trainierten Modelle, die mit Daten aus derselben oder der anderen Datenquelle getestet wurden. Um die Logik der Klassifikationsmodelle untersuchen zu können, wurden weitere Techniken zur „erklärbaren KI“ eingesetzt.

Die Ergebnisse und Standardabweichung der Modelle, die an in vivo bzw. in vitro generierten Daten trainiert und an Daten aus der gleichen oder der anderen Datenquelle getestet wurden, habe ich in Tabelle 1 zusammengefasst. Diese Daten habe ich federführend interpretiert und ausgewertet.

Für die Publikation habe ich die weiterführende Literaturrecherche, die Erstellung der in dieser Dissertation aufgeführten Tabellen und Abbildungen sowie das Verfassen der ersten Textversion übernommen. Alle Autoren haben die finale Version gelesen und der Publikation des Manuskriptes zugestimmt. Folgende Abbildungen und Tabellen aus der Publikation fanden Verwendung im Manteltext: Abbildung 7, 8. Tabelle 1.

_____

Unterschrift, Datum und Stempel des/der erstbetreuenden Hochschullehrers/in

_____

Unterschrift des Doktoranden/der Doktorandin

Auszug aus der Journal Summary List

Journal Data Filtered By: **Selected JCR Year: 2019** Selected Editions: SCIE,SSCI
Selected Categories: **"MEDICINE, GENERAL and INTERNAL"**
Selected Category Scheme: WoS
Gesamtanzahl: 165 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	NEW ENGLAND JOURNAL OF MEDICINE	347,451	74.699	0.660800
2	LANCET	256,199	60.392	0.437300
3	JAMA-JOURNAL OF THE AMERICAN MEDICAL ASSOCIATION	158,632	45.540	0.290050
4	Nature Reviews Disease Primers	7,567	40.689	0.032310
5	BMJ-British Medical Journal	118,586	30.223	0.145170
6	ANNALS OF INTERNAL MEDICINE	58,033	21.317	0.091210
7	JAMA Internal Medicine	17,260	18.652	0.086180
8	PLOS MEDICINE	32,312	10.500	0.065990
9	Journal of Cachexia Sarcopenia and Muscle	3,553	9.802	0.007860
10	Cochrane Database of Systematic Reviews	67,763	7.890	0.134360
11	CANADIAN MEDICAL ASSOCIATION JOURNAL	15,212	7.744	0.016160
12	JOURNAL OF TRAVEL MEDICINE	2,659	7.089	0.006360
13	MAYO CLINIC PROCEEDINGS	15,627	6.942	0.024990
14	JOURNAL OF INTERNAL MEDICINE	10,912	6.871	0.014180
15	BMC Medicine	15,204	6.782	0.042500
16	MEDICAL JOURNAL OF AUSTRALIA	11,075	6.112	0.011070
17	Translational Research	4,043	5.411	0.008350
18	JOURNAL OF THE ROYAL SOCIETY OF MEDICINE	4,214	5.238	0.002580
19	JAMA Network Open	2,239	5.032	0.007660

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
20	Deutsches Arzteblatt International	4,817	4.796	0.007380
21	ANNALS OF FAMILY MEDICINE	5,567	4.686	0.010880
22	JOURNAL OF GENERAL INTERNAL MEDICINE	20,229	4.597	0.026960
23	AMERICAN JOURNAL OF MEDICINE	24,975	4.529	0.024230
24	Journal of Personalized Medicine	617	4.433	0.001950
25	AMERICAN JOURNAL OF PREVENTIVE MEDICINE	23,547	4.420	0.040180
26	European Journal of Internal Medicine	4,933	4.329	0.010280
27	AMYLOID-JOURNAL OF PROTEIN FOLDING DISORDERS	1,486	4.323	0.002920
28	BRITISH JOURNAL OF GENERAL PRACTICE	6,669	4.190	0.008670
29	Frontiers in Medicine	3,034	3.900	0.009870
30	PREVENTIVE MEDICINE	17,316	3.788	0.030080
31	PALLIATIVE MEDICINE	5,413	3.739	0.008460
32	AMERICAN JOURNAL OF CHINESE MEDICINE	3,531	3.682	0.002970
33	MEDICAL CLINICS OF NORTH AMERICA	3,161	3.529	0.004080
34	EUROPEAN JOURNAL OF CLINICAL INVESTIGATION	6,344	3.481	0.006590
35	PANMINERVA MEDICA	806	3.467	0.000660
36	Journal of Clinical Medicine	5,214	3.303	0.010940
37	ANNALS OF MEDICINE	4,510	3.243	0.005190
38	CANADIAN FAMILY PHYSICIAN	3,833	3.112	0.005150

Druckexemplar der Publikation

Journal of
Clinical Medicine

MDPI

Article

Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images

Agnes Holtkamp [1,2], Karim Elhennawy [3], José E. Cejudo Grano de Oro [1], Joachim Krois [1], Sebastian Paris [2] and Falk Schwendicke [1,*]

[1] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité-Universitätsmedizin Berlin, 14197 Berlin, Germany; agnes.holtkamp@charite.de (A.H.); jose-eduardo.cejudo@charite.de (J.E.C.G.d.O.); Joachim.krois@charite.de (J.K.)
[2] Department of Operative and Preventive Dentistry, Charité-Universitätsmedizin Berlin, 14197 Berlin, Germany; sebastian.paris@charite.de
[3] Department of Orthodontics, Dentofacial Orthopedics and Pedodontics, Charité-Universitätsmedizin Berlin, 14197 Berlin, Germany; karim.elhennawy@charite.de
* Correspondence: falk.schwendicke@charite.de; Tel.: +49-30-450-562-556

Abstract: Objectives: The present study aimed to train deep convolutional neural networks (CNNs) to detect caries lesions on Near-Infrared Light Transillumination (NILT) imagery obtained either in vitro or in vivo and to assess the models' generalizability. Methods: In vitro, 226 extracted posterior permanent human teeth were mounted in a diagnostic model in a dummy head. Then, NILT images were generated (DIAGNOcam, KaVo, Biberach), and images were segmented tooth-wise. In vivo, 1319 teeth from 56 patients were obtained and segmented similarly. Proximal caries lesions were annotated pixel-wise by three experienced dentists, reviewed by a fourth dentist, and then transformed into binary labels. We trained ResNet classification models on both in vivo and in vitro datasets and used 10-fold cross-validation for estimating the performance and generalizability of the models. We used GradCAM to increase explainability. Results: The tooth-level prevalence of caries lesions was 41% in vitro and 49% in vivo, respectively. Models trained and tested on in vivo data performed significantly better (mean ± SD accuracy: 0.78 ± 0.04) than those trained and tested on in vitro data (accuracy: 0.64 ± 0.15; $p < 0.05$). When tested in vitro, the models trained in vivo showed significantly lower accuracy (0.70 ± 0.01; $p < 0.01$). Similarly, when tested in vivo, models trained in vitro showed significantly lower accuracy (0.61 ± 0.04; $p < 0.05$). In both cases, this was due to decreases in sensitivity (by −27% for models trained in vivo and −10% for models trained in vitro). Conclusions: Using in vitro setups for generating NILT imagery and training CNNs comes with low accuracy and generalizability. Clinical significance: Studies employing in vitro imagery for developing deep learning models should be critically appraised for their generalizability. Applicable deep learning models for assessing NILT imagery should be trained on in vivo data.

Keywords: artificial intelligence; caries; diagnostics; digital imaging/radiology; mathematical modeling

Citation: Holtkamp, A.; Elhennawy, K.; Cejudo Grano de Oro, J.E.; Krois, J.; Paris, S.; Schwendicke, F. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images. *J. Clin. Med.* **2021**, *10*, 961. https://doi.org/10.3390/jcm10050961

Academic Editor: Luca Testarelli

Received: 1 February 2021
Accepted: 23 February 2021
Published: 1 March 2021

1. Introduction

For detecting and assessing dental caries lesions, visual-tactile and radiographic methods, often in combination with each other, have been the standard over recent decades. Over the last years, Near-Infrared Light Transillumination (NILT), an alternative to radiography for caries lesion detection and assessment has been developed and tested, for example, DIAGNOcam (KaVo, Biberach, Germany). NILT makes it possible to assess teeth in motion (thereby providing some three-dimensional assessment) or to record videos or images. In contrast to radiography, no ionizing radiation is generated. Further, the device is portable and can be repeatedly applied in children and in short intervals, for example, in high-risk individuals. NILT has been confirmed to show similar accuracies for detecting

proximal caries lesions to radiography by both in vitro and in vivo studies [1–6], and a recent meta-analysis demonstrated the underlying evidence to be robust [7].

For assessing diagnostic imagery such as radiographs or NILT, individual examiners rely on their training and experience. A wealth of studies have shown that dentists show limited accuracy (often being associated with their experience) and inter- and intra-examiner reliability when assessing diagnostic images [5–8]. Automated assistance systems based on deep learning, for example, using Convolutional Neural Networks (CNNs), might help to overcome these limitations, reducing false detections and diagnostic and treatment variability, especially of less experienced dentists. CNNs are machine learning models consisting of stacked layers of linear models, corresponding weights, and a bias term and nonlinear functions. They differentiate from other neural networks using convolution operations to extract features, particularly from imagery, such as edges, corners, spots, and macroscopic patterns. As with all machine learning models, they learn data from pairs of datapoints, in this case, images and labels (e.g., caries lesion present yes/no, or the areas or pixels affected by a caries lesion). By repeatedly passing these data-label pairs through the network and optimizing the model weights during training, CNNs are able to identify the inherent statistical patterns and to eventually predict a label on unseen data [9]. In dentistry, CNNs have been applied to detect caries lesions, periodontal bone loss, and apical lesions on radiographs [10].

In two previous publications, it was shown that CNNs can also be used for caries detection on NILT imagery attained in vitro [11] or in vivo [12]. Imagery obtained in vitro can come with a hard ground truth (as teeth can be sectioned and histologically evaluated after the assessment via NILT) but inherent biases routed in their acquisition (extracted teeth mounted in simulation models), while in vivo imagery usually results in a fuzzy ground truth (based on the labeling by multiple human experts) but reflects the clinical situation more realistically. From a researcher's point of view, it could be useful to train a CNN on clinical data and test it on in vitro data (where a firm ground truth can be obtained). Clinicians, on the other hand, would like to know if CNNs developed in vitro are also applicable clinically. The present study aimed to train CNNs on NILT imagery attained in vitro and in vivo and aimed to test their generalizability.

2. Materials and Methods

2.1. Study Design

This study employed 2 datasets of NILT imagery, 1 generated in vitro for a previous study [11] and the other generated in vivo. Images were pixel-wise annotated for proximal caries lesions by 3 independent dental specialists followed by a master reviewer. The segmentation masks were then transformed into binary labels (caries present yes/no). ResNet-type deep CNNs for binary classification were trained on a training dataset from 1 source and tested on hold-out datasets from both sources. Reporting of this study follows the STARD guideline [13] and the Checklist for Artificial Intelligence in Medical Imaging, CLAIM [14].

2.2. Performance Metrics

We used k-fold cross-validation with 10 train, validation, and test splits for evaluating the performance of the models, which were optimized using binary cross-entropy as the loss function. For measuring performance, 6 metrics were employed: F1-score, sensitivity, specificity, predicted positive value (PPV), negative predictive value (NPV,) and area under the receiver operating characteristic curve (AUC).

2.3. Sample Size

No formal sample size calculation was performed. For both datasets, a comprehensive sample of available imagery was employed.

2.4. Dataset

For the in vitro dataset, 226 extracted posterior teeth (113 premolars and 113 molars) were obtained with informed consent under an ethics-approved protocol (ethics committee of Charité, EA4/102/14), as described before [11]. The teeth were embedded in transparent epoxy resin (Epo-Thin 2, Buehler, Lake Bluff, IL, USA), which allowed the resulting NILT imagery to be similar to clinically attained images for the human eye. The models were mounted in a dummy head (Phantomkopf P-6, Frasaco, Tettnang, Germany). NILT images were acquired using the DIAGNOcam by 1 examiner (KE), moving the camera perpendicularly to the occlusal surface over each tooth. The dental unit light was switched off during this examination. Images were captured using the KID software (KaVo Integrated Desktop/version 2.4.1.6821, KaVo), with each image focusing on 1 tooth.

For the in vivo dataset, 1319 images from routine examination data at Charité- Universitätsmedizin Berlin, obtained since 2019, were used (as approved by Charité ethics EA4/080/18) A comprehensive sample of patients aged 18 years or older receiving NILT, radiographic, and visual-tactile examination over a maximum time period of 12 months was included, resulting in 56 patients. NILT was employed by 1 experienced dentist (AH).

2.5. Reference Test

The reference test constituted the pixel-wise annotations of proximal primary caries lesions by 3 independent and experienced dentists (clinical experience: 8–11 years), followed by the review of all annotated images by 1 master reviewer who was able to evaluate and revise all existing annotations. The union of all pixel labels on each image remaining after the review was obtained, and the final classification reference set was established by transforming the segmentation masks into binary targets (caries present yes/no). The decision to first provide pixel-wise annotations was chosen to allow segmentation modeling at some point. For this study on generalizability, a classification model was used as outlined. Each annotator independently assessed each image under standardized conditions using an in-house custom-built annotation tool as described before [11]. Examiners were informed about the study and all trained annotations on a separate set of 10–50 NILT images. Note that we did not score the lesions into further classes such as incipient or advanced lesions. Although this would have been possible based on the pixel-based annotations, it was not of interest within the classification task of this study.

2.6. Data Preparation, Model, and Training

A Residual Convolutional Neural Network (ResNet) was used for classification. The model takes an RGB image as an input and outputs probabilities for binary classes corresponding to sound and carious teeth. The architecture of our residual network was a stack of residual blocks of CNNs as a feature extractor and a fully connected classification head with binary output. We used the feature extraction blocks from a pretrained model on Imagenet of the Pytorch library.

For augmentation, random rotations, vertical and horizontal flipping, shifting, and zooming were applied during training, with a probability of 0.5. The images were resized to $224 \times 224 \times 3$ tensors, which proved to be enough resolution for this task.

The performance of the model on the in vivo and in vitro datasets was assessed separately using k-fold cross-validation with 10 train, validation, and test splits, respectively. When the training and evaluation data were from different sources (e.g., training on in vivo and testing on in vitro data), all the images from the training source were used for training the model. The generalizability of the model was then assessed by splitting the data from the testing dataset into validation and test splits. The validation splits were evaluated during training, and the test splits were evaluated after the models had converged. This process was repeated for 10 different splits.

For each split, we trained for 200 epochs using the Adam optimizer. The binary cross-entropy loss on the validation data was monitored during training with a patience

parameter of 20 epochs, after which early stopping was applied and the mean and standard deviation of the outlined performance metrics were evaluated across the test sets.

Due to the small size of the in vitro data set, it was challenging to set a proper combination of batch size and learning rate that yielded a stable behavior when training with this dataset. Learning rates of 5×10^{-7}, 5×10^{-6}, 5×10^{-5}, and 5×10^{-4} and batch sizes of 4, 8,16, 32, and 64 were considered, respectively. We performed a grid search for each possible combination. The best combination can be found in the Supplementary Table S1. Our models were trained on a NVIDIA Quadro RTX 6000 graphics card (NVIDIA, Santa Clara, CA, USA).

2.7. Explainability

Another aim was to make the models interpretable. There are several visualization algorithms for deep learning models available, and we chose to use GradCAM [15]. This algorithm creates a visualization that makes it possible to distinguish the most salient areas relevant for a particular class. This saliency map is computed as a weighted combination of the feature maps of a particular layer followed by a ReLU activation. The coefficients of the combination are computed as the average of the gradients of the output class with respect to the feature maps.

2.8. Statistical Analysis

Differences in model performance were evaluated via independent 2-sided t-tests, using $p < 0.05$ as a discriminating criterion. Computations were performed using the Python library SciPy 1.5.2.

3. Results

The tooth-level prevalence of caries lesions was 41% in vitro and 49% in vivo, respectively. The models trained on in vivo data performed significantly better (mean $\pm$ SD accuracy and AUC: 0.78 ± 0.04) than those trained in vitro when tested on the imagery from the same dataset (accuracy: 0.64 ± 0.15, AUC: 0.65 ± 0.12, $p < 0.05$). Within each analysis, sensitivity and specificity were similar, while the NPV was significantly higher than the PPV.

When tested in vitro, the models trained in vivo showed significantly lower accuracy (0.70 ± 0.01) and AUC (0.66 ± 0.01) ($p < 0.01$). Similarly, when tested in vivo, models trained in vitro showed significantly lower accuracy (0.61 ± 0.04) and AUC (0.60 ± 0.04) ($p < 0.05$). In both cases, this was due to decreases in sensitivity (by -27% for models trained in vivo and -10% for models trained in vitro). Specificity values did not change or even increased slightly when cross tested. NPV values decreased by -13% and -6%, while decreases in PPV were only -5% and -3%. Models trained in vitro and tested in vivo no longer showed useful sensitivity and PPV. Figure 1 sums up the ROC curves for all four scenarios.

When assessing salient areas found relevant by the models to come to a decision (Figure 2), it became apparent that for true positive detections, the most relevant pixels were also those that dentists found to be affected by caries. False-positive detections on in vivo imagery were often associated with restorations, with a similar appearance to carious lesions. In vitro, it was not always possible to identify reasons behind the models' decision. For false-negative detections, it became clear that the models found other areas than the lesion relevant, indicating an attention problem.

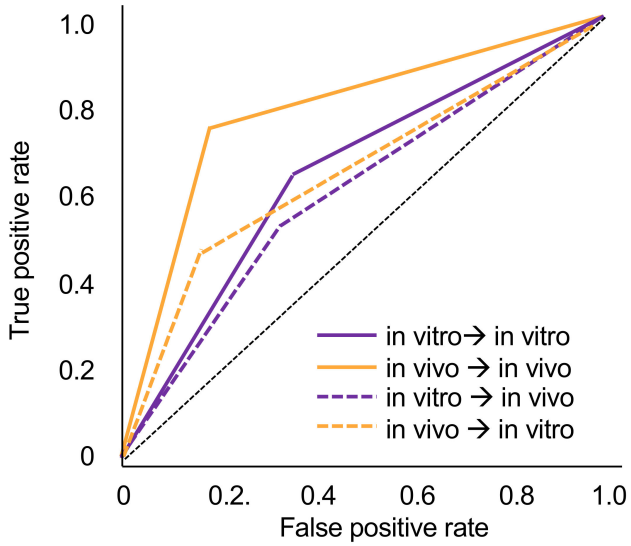

Figure 1. Receiver operating characteristic (ROC) curves for models trained in vivo and in vitro, respectively, and tested on data from the same or the other data source. The resulting area under the curve (AUC) values can be found in Table 1.

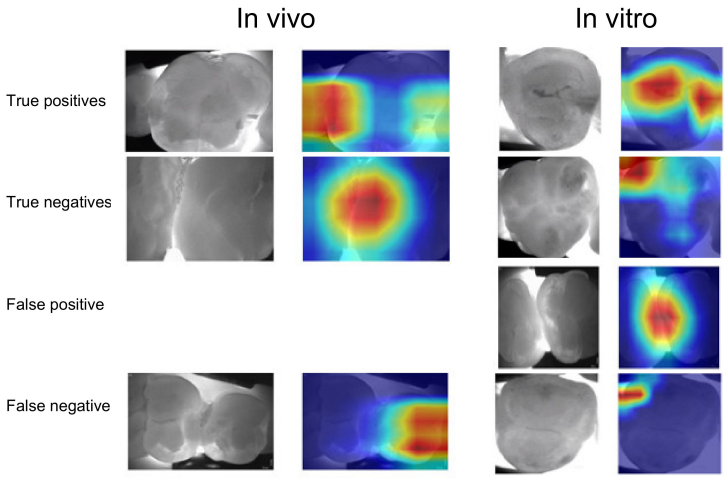

Figure 2. Visualization of contributing feature maps. The original images and the salient areas the models found most relevant for their decision (highlighted in yellow to red) are shown. Data for models trained in vivo and in vitro, respectively, and tested on data from the same or the other data source are shown.

Table 1. Mean performance ± standard deviation of the models trained in vivo and in vitro, respectively, and tested on data from the same or the other data source.

Trained → Tested	Accuracy	F1-Score	AUC	Sensitivity	Specificity	PPV	NPV
in vitro	0.64 ± 0.15	0.57 ± 0.16	0.65 ± 0.12	0.66 ± 0.12	0.64 ± 0.21	0.55 ± 0.22	0.76 ± 0.09
in vivo	0.78 ± 0.04	0.73 ± 0.04	0.78 ± 0.04	0.76 ± 0.06	0.79 ± 0.05	0.70 ± 0.05	0.84 ± 0.03
trained in vitro → tested in vivo	0.61 ± 0.04	0.52 ± 0.03	0.60 ± 0.04	0.55 ± 0.03	0.65 ± 0.07	0.49 ± 0.05	0.70 ± 0.03
trained in vivo → tested in vitro	0.70 ± 0.01	0.56 ± 0.03	0.66 ± 0.01	0.49 ± 0.06	0.83 ± 0.04	0.67 ± 0.04	0.71 ± 0.03

AUC: Area under the curve. PPV/NPV: Positive/negative predictive value.

4. Discussion

NILT is an imaging method for caries detection and assessment. As the device is portable and near-infrared light not being ionizing, it offers a range of advantages over radiography. It specifically lends itself for usage in outreach settings such as schools, care homes, or nondental clinics, for example, in the hands of dental auxiliary or nondental staff. Notably, assessing NILT imagery is challenging, with limited accuracy and low inter-examiner reliability. Using deep learning via CNNs may support NILT diagnostics. There are currently two studies available on CNNs for NILT diagnostics, one conducted in vitro-generated image material and one on in vivo-generated image material [11,12]. In vivo, it is hard to establish a solid ground truth, as histologic assessment or other means (microradiography, μCT) are not available [16], while in vitro, such validation is possible, but sufficiently large sample sizes (e.g., thousands of tooth segments) for training and testing of CNNs are hard to attain. Hence, this study aimed to evaluate models trained in vivo and tested in vitro and vice versa showed generalizability. The research question is relevant beyond specific use cases and has not been explored in dentistry so far.

Based on the findings, the generalizability of models trained on imagery from one source is not necessarily given when tested on imagery from another source. Generally, models trained on in vitro data showed limited accuracy, despite prevalence rates being similar. Notably, the dataset available for training was much smaller in vitro than in vivo, and accuracy (and possibly generalizability) can be expected to increase if larger datasets are used. Overall, in vitro trained models were at the border of being useful or, when applied on in vivo data, no longer useful. In vivo models also showed a drop in accuracy when tested for their generalizability but remained useful.

Interestingly, when tested on the same data material, both models showed similar sensitivities and specificities, but when tested on the other image source, sensitivity dropped drastically while specificity remained stable. Obviously, the learned pattern to identify carious lesions on one image material was not readily applicable on the other dataset. This was confirmed by visualization. In case of false-negative detections, models focused completely on other areas than those marked as carious by dentists. For false-positive detections, we found the presence of fillings to affect models' decisions on in vivo material, while no such pattern emerged in vitro. It should be highlighted that pixels relevant for true-positive detections were similar to those areas marked by dentists as constituting carious lesions.

This study has a number of strengths and limitations. First, it is one of few generalizability studies in the field of deep learning and, more so, deep learning in dentistry. Second, the trained models showed useful accuracies, at least when trained on in vivo imagery. Third, employing methods of explainable AI allowed inference toward the models' decision-making, strengthened confidence into the results, and allowed insights as to reasons behind false classifications. Fourth, as a limitation, the sample sizes, especially in vitro, were limited. Moreover, the in vitro ground truth was not established via histology but by the independent marking of affected pixels and review by experts. This process is obviously not without bias but has been previously used in a range of studies [10].

Fifth, the prevalence of carious lesions was rather high in both datasets compared with other reports on proximal carious lesions (e.g., a study from Sweden found 1.3 proximal lesions in molars or premolars, translating to 1.3 per 20 proximal surfaces being carious, or 6.5%) [17]. Whereas having a balanced dataset helps training CNNs, any PPVs or NPVs will be biased by spectrum bias and should be interpreted accordingly. Sixth, we only used tooth segments for training, making no use of clustering effects (and associated correlation structures) or further context [18]. Last, this study did not evaluate how using a CNN to assist NILT diagnostics impacts clinical care and decision making. This was an active decision, as our focus was a methodological one, and future studies should explore this in detail.

5. Conclusions

Using in vitro setups for generating NILT imagery and training CNNs comes with low accuracy and generalizability. Models trained and tested in vivo showed higher accuracy but limited generalizability when tested on in vitro data. Studies employing in vitro image materials for developing deep learning models should be critically appraised for their generalizability. Applicable deep learning models for assessing NILT imagery should be trained on in vivo data.

Supplementary Materials: The following are available online at https://www.mdpi.com/2077-0383/10/5/961/s1, Table S1: Batch size and learning rate resulting in the best performing models for the four scenarios discussed.

Author Contributions: A.H., J.K., K.E., F.S.: Conceived and designed the study. A.H., J.E.C.G.d.O., J.K., K.E., F.S.: Analyzed the data. All authors: Interpreted the data. A.H. and F.S. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the authors. J.K. is supported by a grant by the German Research Foundation (DFG KR 5457/1-1).

Institutional Review Board Statement: This study was ethically approved by the ethics committee of the Charité (EA4/102/14 and EA4/080/18).

Informed Consent Statement: Informed consent was provided by individuals according to the outlined ethics approval as needed, i.e., for the usage of the extracted teeth. No informed consent was needed for the usage of routinely collected data.

Data Availability Statement: Data cannot be made available given data protection reasons.

Conflicts of Interest: F.S. and J.K. are co-founders of a startup focusing on dental image analysis using artificial intelligence. This study was conceived, conducted, analyzed and reported fully independently.

References

1. Künisch, J.; Schaefer, G.; Pitchika, V.; Garcia-Godoy, F.; Hickel, R. Evaluation of detecting proximal caries in posterior teeth via visual inspection, digital bitewing radiography and near-infrared light transillumination. *Am. J. Dent.* **2019**, *32*, 74–80.
2. Elhennawy, K.; Askar, H.; Jost-Brinkmann, P.-G.; Reda, S.; Al-Abdi, A.; Paris, S.; Schwendicke, F. In vitro performance of the DIAGNOcam for detecting proximal carious lesions adjacent to composite restorations. *J. Dent.* **2018**, *72*, 39–43. [CrossRef]
3. Shaya, S.; Saeed, M.H.; Kheder, W. Proximal Caries Detection in Permanent Teeth by Using DIAGNOcam: An in Vivo Study. *J. Int. Dent. Med. Res.* **2018**, *11*, 45–50.
4. Kühnisch, J.; Söchtig, F.; Pitchika, V.; Laubender, R.P.; Neuhaus, K.W.; Lussi, A.; Hickel, R. In vivo validation of near-infrared light transillumination for interproximal dentin caries detection. *Clin. Oral Investig.* **2016**, *20*, 821–829. [CrossRef] [PubMed]
5. Litzenburger, F.; Heck, K.; Pitchika, V.; Neuhaus, K.W.; Jost, F.N.; Hickel, R.; Jablonski-Momeni, A.; Welk, A.; Lederer, A.; Kühnisch, J. Inter- and intraexaminer reliability of bitewing radiography and near-infrared light transillumination for proximal caries detection and assessment. *Dento. Maxillo. Facial Radiol.* **2018**, *47*, 20170292. [CrossRef] [PubMed]
6. Jablonski-Momeni, A.; Jablonski, B.; Lippe, N. Clinical performance of the near-infrared imaging system VistaCam iX Proxi for detection of approximal enamel lesions. *BDJ Open* **2017**, *3*, 17012. [CrossRef] [PubMed]
7. Ortiz, M.I.G.; de Melo Alencar, C.; de Paula, B.L.F.; Magno, M.B.; Maia, L.C.; Silva, C.M. Accuracy of near-infrared light trans-illumination (NILT) compared to bitewing radiograph for detection of interproximal caries in the permanent dentition: A systematic review and meta-analysis. *J. Dent.* **2020**, *98*, 103351. [CrossRef] [PubMed]

8. Schwendicke, F.; Tzschoppe, M.; Paris, S. Radiographic caries detection: A systematic review and meta-analysis. *J. Dent.* **2015**, *43*, 924–933. [CrossRef] [PubMed]
9. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
10. Schwendicke, F.; Golla, T.; Dreher, M.; Krois, J. Convolutional neural networks for dental image diagnostics: A scoping review. *J. Dent.* **2019**, *91*, 103226. [CrossRef] [PubMed]
11. Schwendicke, F.; Elhennawy, K.; Paris, S.; Friebertshäuser, P.; Krois, J. Deep learning for caries lesion detection in near-infrared light transillumination images: A pilot study. *J. Dent.* **2020**, *92*, 103260. [CrossRef] [PubMed]
12. Casalegno, F.; Newton, T.; Daher, R.; Abdelaziz, M.; Lodi-Rizzini, A.; Schürmann, F.; Krejci, I.; Markram, H. Caries Detection with Near-Infrared Transillumination Using Deep Learning. *J. Dent. Res.* **2019**, *98*, 1227–1233. [CrossRef] [PubMed]
13. Bossuyt, P.M.; Reitsma, J.B.; Bruns, D.E.; Gatsonis, C.A.; Glasziou, P.P.; Irwig, L.; Lijmer, J.G.; Moher, D.; Rennie, D.; de Vet, H.C.; et al. STARD 2015: An updated list of essen-tial items for reporting diagnostic accuracy studies. *BMJ* **2015**, *351*, h5527. [CrossRef] [PubMed]
14. Mongan, J.; Moy, L.; Kahn, C.E. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef]
15. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual EXPLANATIONS from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
16. Walsh, T. Fuzzy gold standards: Approaches to handling an imperfect reference standard. *J. Dent.* **2018**, *74*, S47–S49. [CrossRef] [PubMed]
17. Isaksson, H.; Alm, A.; Koch, G.; Birkhed, D.; Wendt, L.K. Caries Prevalence in Swedish 20-Year-Olds in Relation to Their Pre-vious Caries Experience. *Caries Res.* **2013**, *47*, 234–242. [CrossRef]
18. Meinhold, L.; Krois, J.; Jordan, R.; Nestler, N.; Schwendicke, F. Clustering effects of oral conditions based on clinical and ra-diographic examinations. *Clin. Oral Investig.* **2020**, *24*, 3001–3008. [CrossRef] [PubMed]

Lebenslauf

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

Komplette Publikationsliste

5.1. Holtkamp A, Elhennawy K, Cejudo Grano de Oro JE, Krois J, Paris S, Schwendicke F. 2021. Generalizability of Deep Learning Models for Caries Detection in Near-Infrared Light Transillumination Images. Journal of Clinical Medicine 10(5):961.
Impact Factor 2021: 4.242

5.2. Toelle S, Holtkamp A, Blunck U, Paris S, Schwendicke F. 2020.
Improving the Bond Strength of Radiographically Tagged Caries Lesions In Vitro.
Materials (Basel, Switzerland). 2020 Aug 21;13(17):3702
Impact Factor 2020: 3.623

5.3. von Laffert AJ, Holtkamp A, Strietzel FP. 2020. Dehiszenzen bei Augmentationen mit patientenindividuellen Titangittern - gibt es eine klinische Relevanz? Quintessenz Publishing Deutschland. Zeitschriften Implantologie. 04/2020. 375-383.
Impact Factor 2020: 0

5.4. Krois J, Ekert T, Meinhold L, Golla T, Kharbot B, Wittemeier A, Dörfer C, Schwendicke F. 2019. Deep Learning for the Radiographic Detection of Periodontal Bone Loss. Scientific Reports. 2019 Jun 11;9(1):8495.
Impact Factor 2019: 3.998

5.5 Wittemeier A, Paris S. 2018. Hydroxylapatit – eine Alternative zu Fluoriden? Quintessence Publishing USA. Journals. Team – Journal. 04/2018. 197-203.
Impact Factor 2020: 0

Danksagung

Mein ganz besonderer Dank gilt meinem Erstbetreuer Professor Falk Schwendicke, der bei der Erstellung eines Konzeptes, der Datenerhebung, Bildbearbeitung und schriftlichen Umsetzung meiner Dissertation immer erreichbar war und mich kontinuierlich bestärkt und unterstützt hat. Durch seine Verbindlichkeit und unermüdbare Geduld war er der perfekte Betreuer für mich.

Meinem Kollegen und Zweitbetreuer Dr. Karim Elhennawy möchte ich danken, dass er mir seine In-vitro-Publikation „In vitro performance of the DIAGNOcam for detecting proximal carious lesions adjacent to composite restorations“ von 2018 zur Verfügung gestellt hat und ich seinen erhobenen Datensatz zum Trainieren und Vergleichen meiner KI-Modelle nutzen durfte.

Professor Paris, meinem damaligen Chef, möchte ich für sein Vertrauen in mich und die jahrelange, gute Zusammenarbeit danken.

Bei den Data Scientists Dr. Joachim Krois und Jose Eduardo Cejudo Grano de Oro möchte ich mich für die dauerhafte Unterstützung in allen datenwissenschaftlichen Themen unterstützen – bei Joachim insbesondere auch für seinen unbeirrbaren Enthusiasmus beim Erklären!

Meiner Familie, meinem Mann und meinen Freunden gilt großer Dank für ihr Verständnis in schwierigen Zeiten und ihren Glauben an mich. An dieser Stelle möchte ich noch meinen Bruder Burkhard erwähnen, der mich immer unterstützt hat und es mir ermöglichte, meine beruflichen Träume zu verwirklichen. Danke für alles.