DISSERTATION

Der Einfluss von Annotationsgenauigkeit auf die automatisierte Detektion von Zahnstein auf Bissflügelaufnahmen

The Impact of Label Accuracy on Dental Calculus Detection on Bitewing Radiographs using Deep Learning

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

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Martha Büttner,
geborene Duchrau

Erstbetreuung: Prof. Dr. med. dent. Falk Schwendicke

Datum der Promotion: 29.11.2024

Inhaltsverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AP	Mittlere Genauigkeit (engl. average precision)
BB	Bounding Box
CNN	Konvolutionale neuronale Netzwerke
DL	Deep Learning
FN	Falsch negativ
FP	Falsch positiv
IoU	Intersection over Union
KI	Künstliche Intelligenz
mAP	Durchschnittliche mittlere Genauigkeit (engl. mean average precision)
ML	Maschinelles Lernen
NN	Neuronale Netzwerke
P	Positiver Vorhersagewert (engl. precision)
R	Sensitivität (engl. recall)
RCNN	Region Based Convolutional Neural Network
RN	Richtig negativ
RP	Richtig positiv
SD	Standartabweichung
SHAP	SHapley Additive exPlanations
SSD	Single-Shot-Detector
XAI	Erklärbare künstliche Intelligenz, engl. explainable artificial intelligence
YOLO	You Only Look Once
YOLOv5	You Only Look Once version 5

Zusammenfassung

Künstliche Intelligenz (KI) wurde vielseitig in der Zahnmedizin angewandt, in der Parodontologie beispielsweise für die Detektion von parodontalem Knochenverlust auf Röntgenbildern. Die meisten KI-Modelle zur Bildanalytik werden durch sogenanntes überwachtes Lernen entwickelt, wobei neben Rohdaten auch Markierungen (Annotationen) gesuchter Klassen oder Pathologien zur Verfügung gestellt werden müssen. Bei der Detektion von Objekten auf Bildern werden z.B. Umrahmungen mit Boxen zur Markierung eingesetzt. Fehlt es an Zeit, Sorgfalt oder Kalibrierung der Annotator*innen können ungenaue Annotationen die Folge sein. Die vorliegende Arbeit untersuchte den Einfluss (un-)genauer Annotationen auf KI-Modelle in der Zahnmedizin anhand einer exemplarischen Aufgabe, der Detektion von Zahnstein auf Bissflügelaufnahmen. Dabei wurden zwei Szenarien betrachtet: (1) konsistent zu große oder zu kleine Annotationen, wie sie auftreten können, wenn einzelne Personen fehlerhaft annotieren; (2) inkonsistent zu große oder zu kleine Annotationen, um mehrere Personen mit fehlender Kalibrierung zu simulieren. Die Evaluation der resultierenden KI-Modelle erfolgte sowohl auf genau annotierten Testdaten als auch auf ungenau annotierten Testdaten (äquivalent zu den jeweiligen Trainingsdaten). Letzteres diente der Bestimmung einer möglichen Maskierung der zu erwartenden Modellungenauigkeit durch ungenaue Annotationen. 4837 Bissflügelaufnahmen wurden in einem zweistufigen Verfahren möglichst genau annotiert. Das Objektdetektionsmodell YOLOv5 wurde auf einem genau annotierten, 27 konsistent ungenau annotierten und 9 inkonsistent ungenau annotierten Datensätzen trainiert und evaluiert. 5-fache Kreuzvalidierung wurde durchgeführt und die mittlere durchschnittliche Genauigkeit (mAP, engl. mean average precision) ermittelt. Die Referenzgruppe für statistische Vergleiche war das Modell, das auf genau annotierten Daten trainiert wurde. Letzteres erreichte eine mAP von 0,77 (SD = 0,01). Konsistent zu kleine Annotationen führten zu einer Verringerung der Performance unabhängig davon, ob auf genau annotierten Daten (0,74 (0,01)) oder auf ungenau annotierten Daten (0,75 (0,01)) getestet wurde. Konsistent vergrößerte Annotationen in den Trainingsdaten führten zu einer Verringerung der Performance, wenn sie auf genau annotierten Daten getestet wurden (bereits bei Verdopplung der BB-Fläche). Bei ungenau annotierten Testdaten war eine solche Performanceabnahme erst bei drastischen Ungenauigkeiten (70-fache Flächenvergrößerung) detektierbar. Bei inkonsistenten Ungenauigkeiten führte die Testung sowohl auf ungenau als auch auf

genau annotierten Testdaten zu signifikanten Performanceverlusten. Ungenau annotierte Trainingsdaten können die Modellperformance negativ beeinflussen, wobei dieser Einfluss teilweise durch das Testen auf ebenso ungenau annotierten Testaten maskiert werden kann. Genau annotierte Daten waren für Training und Evaluation von KI-Modellen zur Zahnsteindetektion unabdingbar.

Abstract

Artificial Intelligence (AI) has been widely applied in dentistry for tasks such as periodontal bone loss detection on radiographs. Most AI applications are trained in a supervised manner, where labeling (e.g., marking of specific areas using bounding boxes, (BB)) is required. A lack of time, diligence or calibration between multiple annotators may result in inaccurate labels. The impact of annotation accuracies and hence inaccurate labels has not been explored in dentistry and only rarely in general. This study evaluated the impact of (in-)accurate labels on the exemplary task of dental calculus detection on bitewing radiographs. A dataset of 4837 bitewing radiographs was annotated for dental calculus using BB. Two scenarios were evaluated: (1) consistently too large or too small annotations, as might be the case when single individuals label inaccurately, and (2) inconsistently too large or too small annotations, as might results from labeling by multiple individuals lacking calibration. Models were evaluated on both accurately labeled test data and inaccurately labeled test data (the latter is relevant as test and training data usually emanate from the same label process). The object detection model YOLOv5 was trained and evaluated on one accurately labeled dataset, 27 consistently inaccurately labeled dataset and 9 inconsistently inaccurately labeled datasets. 5-fold cross-validation was performed and models were evaluated using mean average precision (mAP). The reference group was the model trained on accurately labeled data, which achieved a mAP of 0.77 (SD = 0.01). Performance decreased immediately when trained on consistently too small annotations and tested on accurately labelled data, mAP (SD) = 0.74 (0.01), or inaccurately labelled data, mAP 0.75 (0.01), respectively. When trained on too large labels, model performance did not decrease when tested on inaccurately labeled data except when BB were drastically too large (70-fold increase in area, mAP (SD) = 0.75 (0.01). Testing on accurately labeled data showed a decay in performance starting at a twofold area enlargement, mAP (SD) = 0.24 (0.05). Inconsistent label inaccuracies led to performance decreases on both inaccurately and accurately labeled test data. Training on inaccurately labeled data negatively impacts on model performance, while testing on the same inaccurately labeled data may mask this performance decrease. Accurately labeled data was critical when training and testing dental calculus detection models.

1. Einleitung

Künstliche Intelligenz (KI) findet schon heute vielseitige Anwendung in der Zahnmedizin. Maschinelle Bildverarbeitung ist der bisher häufigste Anwendungsbereich. In der Parodontologie wurden beispielsweise Modelle zur Detektion von Plaque auf intraoralen Fotos [1] oder der Detektion von parodontalem Knochenverlust auf Röntgenbildern entwickelt [2]. Besonders relevant ist die Analyse von Röntgenbildern, da ca. 40% der jährlich aufgenommenen Röntgenbilder in Deutschland aus der Zahnmedizin stammen; mithin mehr als 50 Millionen Bilder pro Jahr [3]. Die Befundung derselbigen ist zeitaufwendig und weist mitunter große Ungenauigkeiten auf. Eine Diagnoseunterstützung mittels KI könnte zur Qualitätssicherung und Zeitersparnis beitragen. Die Entwicklung solcher KI-Applikationen ist jedoch zeit- und kostenintensiv, da für die meisten KI-Modelle befundete Trainingsdaten benötigt werden. Im Gegensatz zur Anwendungen auf Alltagsbildern, auf denen Lebewesen oder Objekte auch von Laien markiert werden können, braucht es in der (Zahn-)medizin qualifiziertes Fachpersonal. Durch fehlende Zeit, Sorgfalt oder Kalibrierung kann es zu ungenauen Befunden in Datensätzen kommen. Der Einfluss dieser Ungenauigkeiten ist unzureichend verstanden und soll in dieser Arbeit an dem Anwendungsbeispiel der Detektion von Zahnstein auf Bissflügelaufnahmen untersucht werden.

1.1 Zahnstein auf zahnärztlichen Röntgenbildern

1.1.1 Definition und Einordnung

Mineralisierte Plaque wird als Zahnstein bezeichnet. Je nach Lokalisation entstammen die verkalkten Bestandteile überwiegend dem Speichel (supragingival) oder Blut bzw. Taschenexsudat (subgingival). Subgingivaler Zahnstein wird als Konkrement bezeichnet, hat eine dunklere Farbe und eine stärkere Haftung an der Zahnoberfläche als supragingivaler Zahnstein [4]. Nachfolgend werden unter dem Begriff Zahnstein sub- und supragingivale Mineralisationen zusammengefasst.

1.1.2 Diagnose und Therapie

Parodontale Erkrankungen sind neben Karies die häufigste Ursache für Zahnverlust [5,6]. Mineralisationen stellen eine ideale Grundlage für bakterielle Ansiedelung an der Zahnoberfläche dar. Neben Alter, Geschlecht, Plaque und bestehendem Attachmentverlust gehört Zahnstein zu den vorherrschenden Risikofaktoren von Parodontitis und wird mit der Progression der Krankheit assoziiert [7,8]. Ein Kernbestandteil der parodontologischen Prophylaxe und Therapie umfasst die Entfernung des Biofilms und der Zahnsteinablagerungen [8]. Die Diagnose erfolgt meist durch visuelle Inspektion (supragingival) oder Sondierung (subgingival). Mit entsprechender Größe und Lokalisation ist Zahnstein als Opazität auf Röntgenbildern darstellbar. Abbildung 1 zeigt zwei Bissflügelaufnahmen mit multiplen Zahnsteinanlagerungen. Das Erscheinungsbild ist irregulär mit teils unscharfen Begrenzungen, angelagert an Zähne im supra- sowie im subgingivalen Bereich.

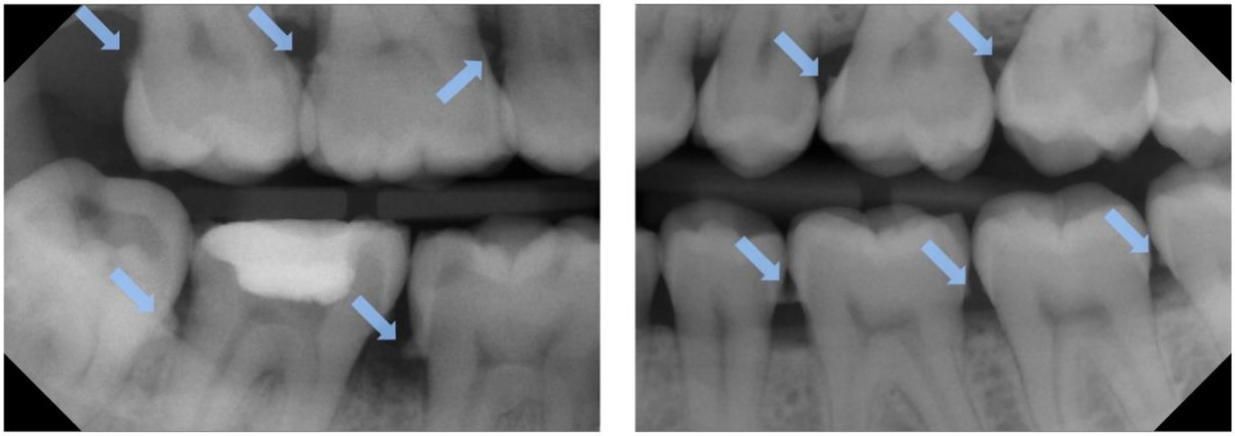

Abbildung 1: Zwei Bissflügelaufnahmen mit multiplen Zahnsteinanlagerungen. Regionen mit Zahnstein sind durch hellblaue Pfeile gekennzeichnet (eigene Darstellung).

Bissflügelaufnahmen werden, je nach Risikogruppe, im Rahmen der zahnärztlichen Vorsorgeuntersuchung zur Kariesdiagnostik alle 6 bis 48 Monate empfohlen [9]. Diese, auf die Abbildung des Approximalraums standardisierten Aufnahmen ermöglichen auch eine gute Darstellung von sub- und supragingivalen Zahnstein im Seitenzahnbereich (Abbildung 1). Eine automatisierte Detektion desselbigen durch KI-Modelle könnte eine frühzeitige Entfernung fördern und dadurch in der Behandlung von Parodontitis unterstützen.

1.2 Künstliche Intelligenz

1.2.1 Begriffseinordnung

Der lateinische Wortursprung des Wortes Intelligenz, „Intellegere“, bedeutet „erkennen“, „verstehen“ oder „zwischen etwas wählen“ [10]. Im Allgemeinen beschreibt Intelligenz menschliche Fähigkeiten wie Abstraktion oder problemlösendes Verhalten. Eine enge Definition des Begriffes ist seit über 100 Jahren Forschungsschwerpunkt der empirischen Psychologie, eine allgemein anerkannte Definition konnte bisher nicht gefunden werden [10]. Auch die Begriffe künstlichen Intelligenz (KI) und die von Teilbereichen von KI wie maschinelles Lernen (ML) und Deep Learning (DL) sind nicht einheitlich definiert und werden unterschiedlich genutzt [11]. Während Intelligenz beim Menschen mit Begabungen und besonderen Fähigkeiten assoziiert ist, verbergen sich hinter KI mathematische Modelle, die menschliche Fähigkeiten imitieren, zum Beispiel logisches Denken, Lernen, Planen und Kreativität [12]. KI umfasst die automatisierte Ausführung von Aufgaben, die ursprünglich menschlichem Denken vorbehalten waren. Ursprung dieser Entwicklung, in der ersten Hälfen des 20. Jahrhundert, waren regelbasierte Systeme.

Im Teilbereich ML erlernen mathematische Modelle in einem sogenannten Trainingsprozess aus Daten und vorgegebenen Lösungen selbstständig Regeln für die Problemlösung, die dann auf neue (ungesehene Daten) angewandt werden können. Im klassischen ML werden mögliche Merkmale (engl. features) für die Lösung eines Problems vorgegeben. Im DL sucht ein Algorithmus selbständig nach Merkmalen, die für die Lösung des Problems notwendig sind [13]. Dies ist besonders dann von Bedeutung, wenn die Regeln für ein Problem komplex und schwer beschreibbar sind, bspw. bei der Erkennung von Objekten auf Bildern. Algorithmen aus dem Bereich DL sind deutlich komplexer als einfach ML-Modelle oder regelbasierte KI, verlieren aber an Erklärbarkeit. Abbildung 2 stellt die Relationen der Bereiche zueinander schematisch dar.

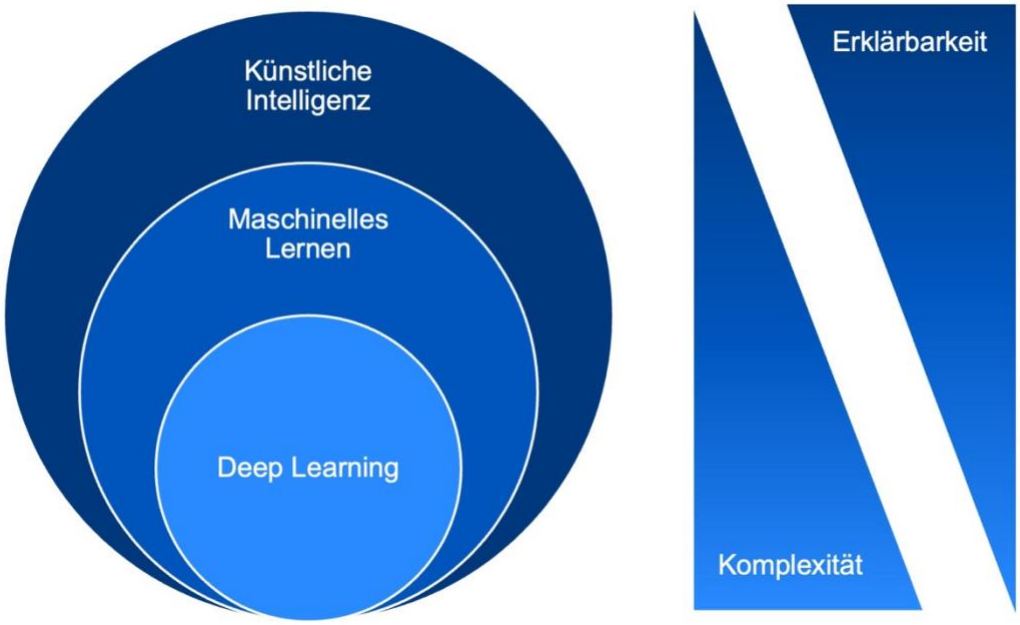

Abbildung 2: Relationen der Bereiche Künstliche Intelligenz, Maschinelles Lernen und Deep Learning. Deep Learning weist die höchste Komplexität und die geringste Erklärbarkeit auf (eigene Darstellung).

1.2.2 Neuronale Netzwerke

Neuronale Netzwerke (NN) sind das elementare Werkzeug von DL. Mathematisch wird ein NN als Abbildungsvorschrift definiert, d.h. eine Menge von Eingaben wird durch eine Funktion auf eine Menge von Ausgaben abgebildet [14]. Hierzu werden sogenannte künstliche Neuronen in komplexe Netzwerke zusammengefügt. Der Begriff Neuron verweist auf die neurophysiologische Inspiration in der Entwicklung dieser mathematischen Einheit. Hauptbestandteil des künstlichen Neurons ist eine mathematische Funktion, die Übertragungsfunktion. Sie summiert gewichtete numerische Eingaben und leitet die Summe an die Aktivierungsfunktion weiter, welche berechnet, ob und in welchem Maße ein Signal weitergegeben wird.
Die generierte Ausgabe dient meist als Eingabe für ein nächstes künstliches Neuron. Abbildung 3 stellt die Signalverarbeitung durch ein künstliches Neuron schematisch dar.

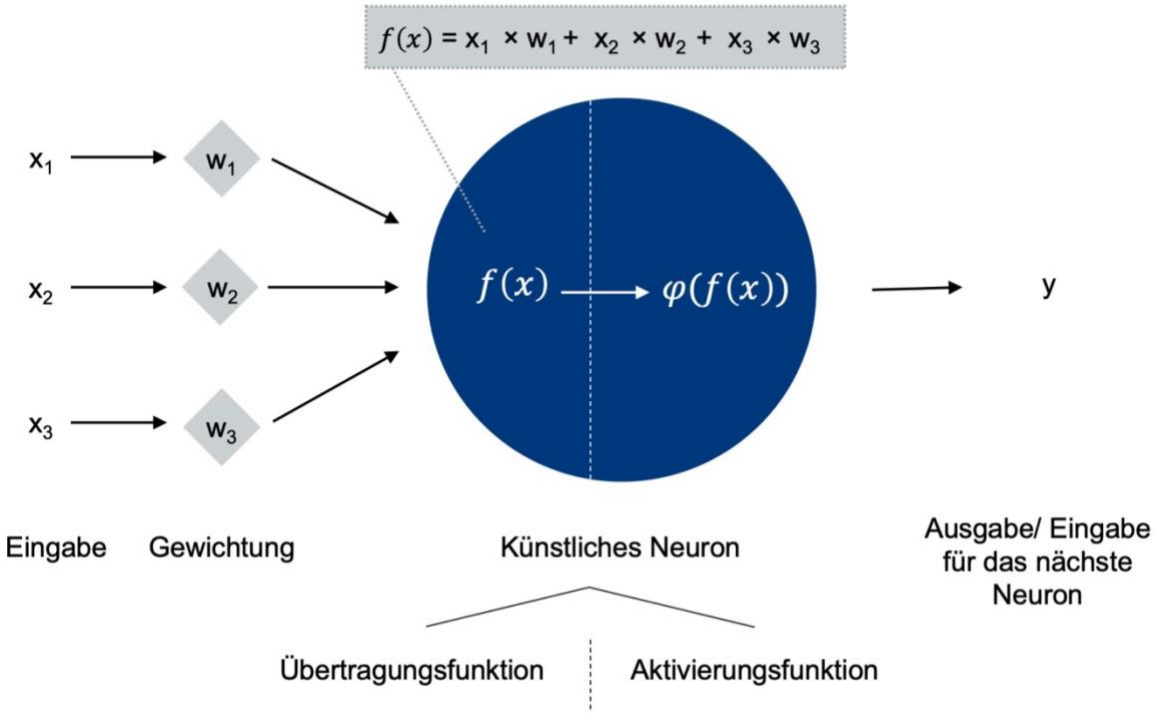

Abbildung 3: Schematische Darstellung der Signalverarbeitung durch ein künstliches Neuron (blauer Kreis), welches drei numerische Werte (x1 – x3) als Eingabe erhält. Diese werden vor der Eingabe in die Übertragungsfunktion, f(x), mit Gewichten multipliziert. Das Ergebnis der Übertragungsfunktion, welche eine Aufsummierung der gewichteten Eingaben darstellt, wird durch die Aktivierungsfunktion, φ(f(x)), prozessiert, und ein finales Signal (y) als Ausgabe weitergegeben (eigene Darstellung).

In einem NN werden einzelne künstliche Neuronen in Schichten (engl. layer) angeordnet und verknüpft. In der ersten Schicht (Eingabeschicht) werden Daten an das NN übergeben, (bspw. Graustufenwerte einzelner Pixel). In der Ausgabeschicht erfolgt die Ausgabe von Vorhersagen, z.B. anhand zuvor definierter Klassen. Befinden sich zwischen der Eingabe- und der Ausgabeschicht weitere Schichten künstlicher Neuronen, entsteht ein sogenanntes tiefes NN. Diese Schichten werden auch als versteckte Schichten (engl. hidden layers) bezeichnet. In ihnen erfolgen weitere mathematische Operationen, um relevante Merkmale (Features) zur Lösung eines Problems zu erkennen. In Abbildung 4 ist ein einfaches, tiefes NN schematisch dargestellt. Die Anordnung von Neuronen, Schichten und wie diese verbunden sind, wird als Architektur eines Netzwerkes bezeichnet.

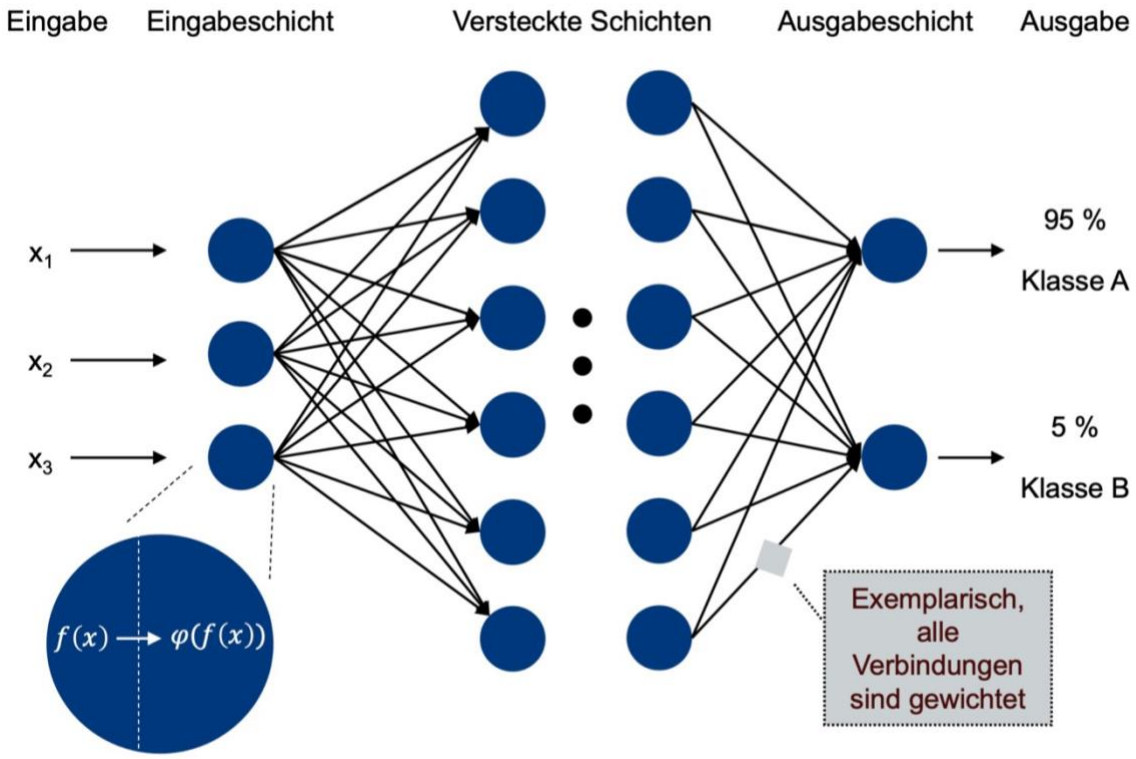

Abbildung 4: Struktur eines neuronalen Netzwerkes. Künstliche Neuronen sind durch blaue Kreise dargestellt und bestehen aus einer Übertragungsfunktion f(x) und einer Aktivierungsfunktion φ(f(x)). Eine nummerische Eingabe wird gewichtet über die Eingabeschicht an das NN übergeben, in versteckten Schichten prozessiert und auf eine Ausgabeschicht definierter Klassen, hier A und B, abgebildet (eigene Darstellung).

1.2.3 Der Trainingsprozess neuronaler Netzwerke

Der Lernprozess, bei dem sich ein NN Wissen aneignet, wird als Training bezeichnet. Die häufigste Methode, NN zu trainieren, ist das sogenannte überwachte Lernen (engl. supervised learning). Dafür iteriert das NN über große Datenmengen, trifft Vorhersagen und vergleicht diese mit einer zur Verfügung gestellten Lösung. Dies geschieht über eine sogenannte Verlustfunktion (engl. loss function). Das Ergebnis dieses Vergleiches wird als Verlust (engl. loss) bezeichnet; große Unterschiede zwischen Vorhersage und Annotation führen zu großen Verlusten. Ziel des Trainingsprozesses ist es, Modellparameter zu finden, die den Verlust minimieren. Wie in Abbildung 5 dargestellt, wird die Ausgabe der Verlustfunktion (der Verlust) an die Optimierungsfunktion übergeben, welche berechnet, in welche Richtung die mathematischen Parameter des

NN angepasst werden müssen, um den Verlust zu verringern. Die Schrittgröße, mit der die Parameter angepasst werden, wird über die Lernrate (engl. learning rate) definiert. Das NN mit den angepassten Parametern wird genutzt, um neue Vorhersagen auf den Trainingsdaten zu treffen. Es wird erneut der Verlust berechnet, an die Optimierungsfunktion übergeben und die Parameter angepasst. Wurden alle Trainingsbilder einmal für diesen Prozess genutzt, gilt eine sogenannte Epoche als beendet. Das Training wird so lange durchgeführt, bis eine zuvor definierte Anzahl von Epochen oder ein anders vordefiniertes Kriterium eingetreten ist, bspw. eine Zahl von Epochen, während der keine Verbesserung eingetreten ist (engl. early stopping). Ein trainiertes NN wird als Modell bezeichnet.

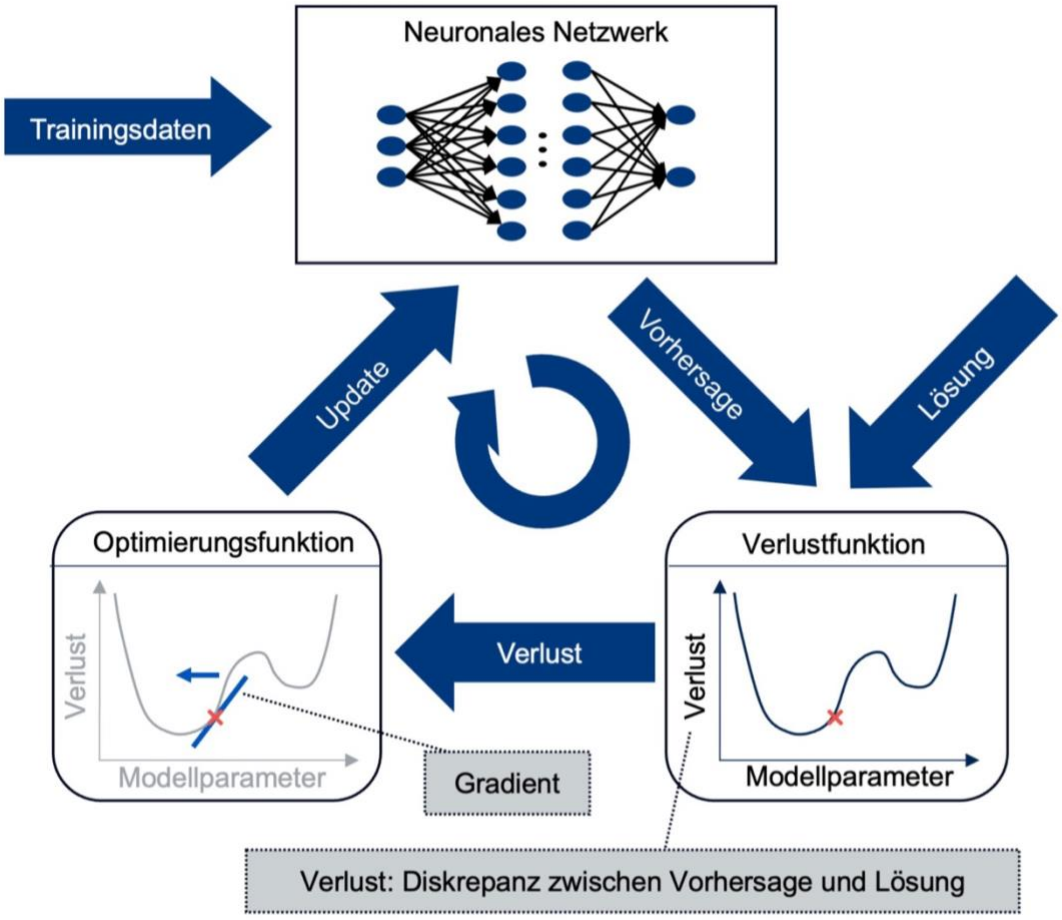

Abbildung 5: Schematische Darstellung eines überwachten Lernprozesses eines neuronalen Netzwerks (NN). Durch das NN werden Vorhersagen auf den Trainingsdaten getätigt und über eine Verlustfunktion (hier stark abstrahiert) mit der Lösung verglichen. Eine Optimierungsfunktion bestimmt dann, in welche Richtung die Parameter des NN geändert werden müssen, um den Verlust zu minimieren. Hier ist dies anhand der Gradientenmethode exemplarisch dargestellt. Die Optierungsfunktion berechnet den Gradienten und bestimmt anhand der Steigung die Richtung

des Updates (blauer Pfeil in Optimierungsfunktion). Die Parameter des Modelles werden aktualisiert und der Vorgang iterativ wiederholt (eigene Darstellung).

Die mathematischen Parameter, die im Lernprozess adaptiert werden, sind unter anderem die Gewichtungen der Verknüpfungen der einzelnen Neuronen. Sie bestimmen über die Signalstärke, mit dem eine Eingabe an ein Neuron übergegeben wird, d.h. wie hoch die Relevanz der Information ist. Die Gewichte, d.h., die Parameter eines NN, speichern somit das erlernte Wissen zur Lösung einer Aufgabe.
Der Startwert der Gewichte, die sogenannte Initialisierung des Trainings, kann unter Einhaltung gewisser Verteilungen zufällig gewählt werden oder von anderen bereits trainierten Modellen übernommen werden. Letzteres wird als transferiertes Lernen bezeichnet und basiert auf der Idee, dass sinnvolle Parameterkonfigurationen einer bestimmten Aufgabe auch eine gute Ausganglage für neue Aufgaben sein können. Auch in der Zahnmedizin konnte gezeigt werden, dass transferiertes Lernen die Performance für neue Aufgaben verbessert [15]. Weniger wichtig ist dabei, von welcher Aufgabe die Gewichte transferiert werden, sondern dass grundlegende Elemente, wie das Erkennen von Ecken oder Formen auf Bildern, bereits erlernt wurde. Ein NN, welches mit Parametern einer anderen Aufgabe initialisiert wird, wird als vortrainiertes Netzwerk bezeichnet.

1.2.4 Maschinelles Sehen und konvolutionale neuronale Netzwerke

Im Bereich des maschinellen Sehens (engl. computer vision) mittels DL werden Bilder durch NN verarbeitet und Vorhersagen getroffen. Je nach Aufgabe unterscheidet sich die Art und Weise der Vorhersage:

- Klassifizierung: Jedem Bild wird ein Label/eine Klasse zugeordnet.
- Segmentierung: Jedem Pixel eines Bildes wird eine Klasse zugeordnet.
- Objektdetektion: Gesuchten Objekte werden auf dem Bild lokalisiert und mit Koordinaten und dem Label ihrer Klasse beschrieben.

Abbildung 6 illustriert beispielhaft den Unterschied in den Vorhersagen der Modelle.

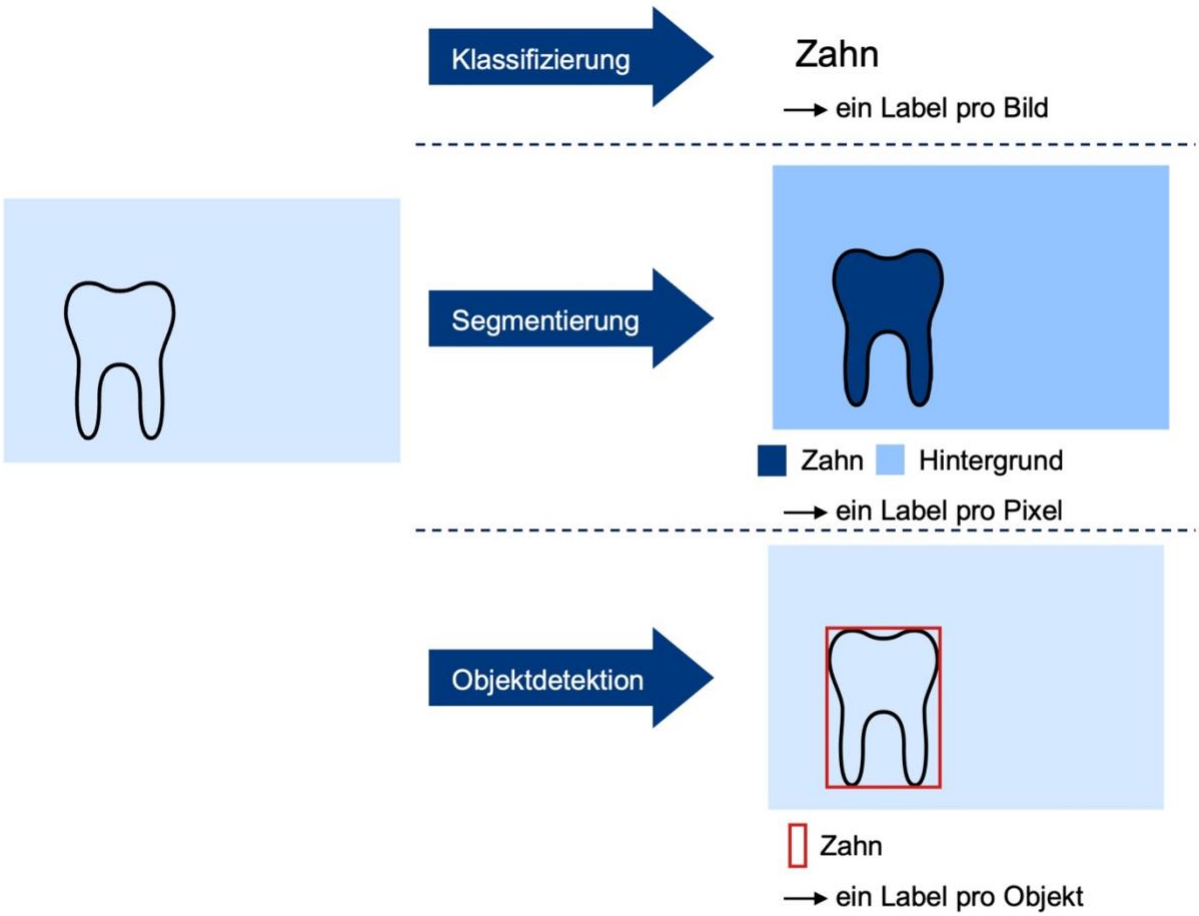

Abbildung 6: Illustration der Ausgaben von Klassifizierungs-, Segmentierungs- und Objektdetektionsmodellen (eigene Darstellung).

Für die Verarbeitung von Bildern stellte sich eine besondere Form vom NN, sogenannte konvolutionale NN (engl. convolutional neural network, CNN), als geeignet heraus. In CNN werden Daten durch Filteroperationen prozessiert. Dies geschieht in sogenannten konvolutionalen, versteckten Schichten in Form von Matrizenmultiplikationen. Über anschließende Vereinfachungsoperationen, wie bspw. Pooling, wird die Dimension des Bildes währenddessen schrittweise reduziert. Bei minimalem Pooling wird der kleinste Wert des Ergebnisses einer Matrixmultiplikation genutzt, bei maximalem Pooling der größte Wert. Letzteres sorgt bspw. für einen Fokus auf die prominentesten Bildmerkmale und eine Ausblendung von Hintergrundmerkmalen. CNN können so schrittweise Linien, Formen und Konturen vereinfacht erkennen und diese zu Objekten zusammenführen.

1.3 Detektion von Objekten

Objektdetektion (engl. object detection) ist, wie in Abbildung 7 dargestellt, ein Teilgebiet des maschinellen Sehens aus dem Bereich DL. Wie in 1.2.4 beschrieben, wird hier einem Bild nicht nur ein Label zugeordnet, sondern einzelne Objekte auf Bildern identifiziert und lokalisiert. Die genaue Lage der gesuchten Objekte wird durch umrahmende Vierecke, sogenannte Bounding Boxen (BB) beschrieben. Um diese beschreiben zu können werden bspw. die Pixelkoordinaten des Zentrums, die Höhe und die Breite der BB angegeben. Jedem Objekt wird außerdem eine Klasse zugeordnet und Vorhersagen häufig mit einem sogenannten Objektivitätsscore beschrieben. Letzteres kann als Wahrscheinlichkeit interpretiert werden, dass das detektierte Objekt existiert.

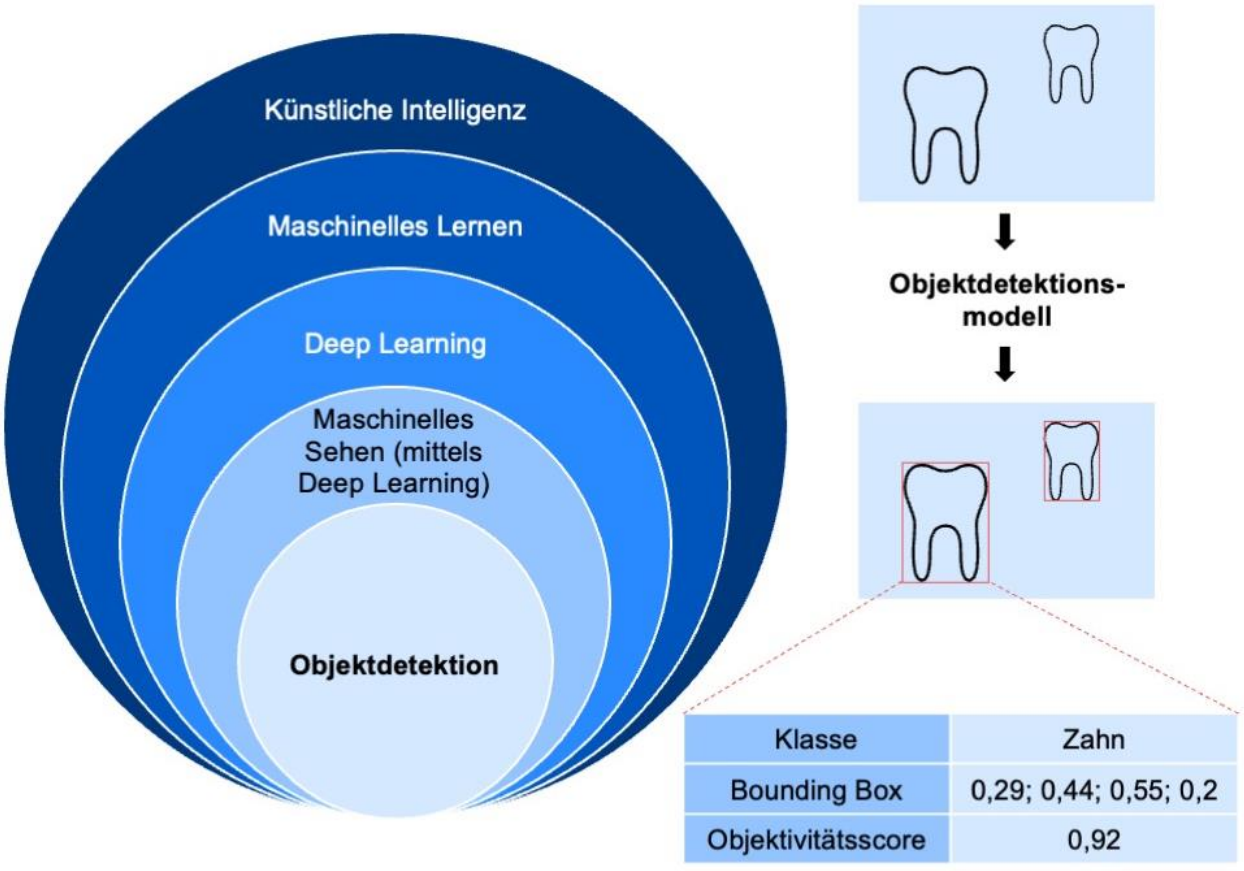

Abbildung 7: Einordnung der Objektdetektion in Teilbereiche der künstlichen Intelligenz. Rechts ist die Ein- und Ausgabe eines Objektdetektionsmodelles dargestellt. Die Ausgabe besteht hier pro Objekt aus drei Komponenten: der Klasse, der Lokalisation und dem Objektivitätsscore. Die Lokalisation wird durch die Pixelkoordinaten der Bounding Box beschrieben (eigene Darstellung).

1.3.1 Architekturen zur Objektdetektion

Modellarchitekturen zur Detektion von Objekten können in ein- und zwei phasige Netzwerke unterteilt werden (engl. one/two-stage detectors). Bei zweiphasigen Netzwerken erfolgt zunächst die Identifikation relevanter Regionen (engl. region proposal) und erst im zweiten Schritt die Detektion von Objekten auf den zuvor identifizierten Bildregionen. Bekannte Vertreter sind „R-CNN“ [16], „Fast R-CNN“ [17] und „Faster R-CNN“ [18]. Bei einphasigen Netzwerken erfolgt die Detektion von Objekten in einem Durchlauf (ohne vorherige Identifikation relevanter Regionen). Im Falle der Architektur „You Only Look Once“ (YOLO) [19] erfolgt dies beispielsweise durch eine Einteilung des Bildes in Gitterzellen, in welchen Objektmittelpunkte gesucht werden. Dies beschleunigt die Detektion erheblich und macht das Netzwerk besonders geeignet für den Echtzeiteinsatz. YOLO ist neben „SSD“ [20] und „RetinaNet“ [21] der bekannteste Vertreter der einphasigen Objektdetektoren (Architekturdetails siehe unten) und ist für die Detektion von kleinen Objekten besonders gut geeignet, weshalb es in der vorliegenden Arbeit gewählt wurde.

1.3.2 „You Only Look Once“

Das CNN YOLO Version 5 (YOLOv5) ist eine moderne CNN-Architektur zur Detektion von Objekten [22]. Sie zählt wie beschrieben zu den einphasigen Detektoren und steht in unterschiedlichen Größen zur Verfügung. Die größte Version YOLOv5x besteht aus 8,6 * 10^7 Parametern. Alle Größen haben grundlegend den gleichen Aufbau: Im sogenannten Rückgrat (engl. backbone) des CNN werden die Bilder mittels verschiedener Filter (engl. convolutions) prozessiert, um so die relevanten Bildmerkmale zu identifizieren. Im nächsten Abschnitt (engl. neck) wird diese Merkmalserkennung präzisiert: in sogenannten Featurepyramiden werden Darstellungen relevanter Merkmale in unterschiedlichen Auflösungen erstellt. Diese werden in den letzten Teil (engl. head) des Netzwerkes eingespeist, in dem die eigentliche Detektion der gesuchten Objekte stattfindet. Die Ausgabe besteht pro Objekt aus den drei in 1.3 beschriebenen Komponenten; der Klasse, der BB und dem Objektivitätsscore des detektierten Objektes.

1.4 Evaluation von Objektdetektionsmodellen

Ziel des Trainings ist es, Modelle zu erhalten, die nicht nur auf den Trainingsdaten gute Ergebnisse liefern, sondern auch auf neuen, bisher ungesehenen Daten zuverlässig Vorhersagen treffen. Bei fehlender Diversität in den Trainingsdaten kann es dazu kommen, dass Repräsentationen des Netzwerkes keine allgemeine Lösung darstellen, sondern auswendig gelernt wurden oder sich auf Objekteigenschaften fokussieren, die nur in den Trainingsdaten vorhanden sind (engl. overfitting). Ein Modell, welches sehr gute Performance auf den Trainingsdaten erreicht, kann möglicherweise auf andere (ungesehenen) Daten trotzdem scheitern. In diesem Falle fehlt dem Modell die Fähigkeit der Generalisierbarkeit. Die Evaluierung von Modellen sollte deshalb immer auf einem separatem Testdatensatz erfolgen, also Daten, die während des Trainings nicht genutzt wurden. Die Aussagekraft einer Modellevaluation hängt von der Qualität der verwendeten Testdatensatzes ab.
Bei der Entwicklung neuer Architekturen werden die Perfomance bspw. auf multiplen öffentlich publizierten Datensätzen, sogenannten Benchmarking Datensätzen, angegeben. Solche sind in der Zahnmedizin bisher nicht vorhanden, weshalb oft auf Testdaten aus der gleichen Distribution (z.B. in Bezug auf Population, Aufnahmetechnik etc.) und dem gleichen Annotationsvorgang wie die Trainingsdaten zurückgegriffen wird.

1.5 Erklärbarkeitsanalyse

Ein weiterer Weg zu kontrollieren, ob ein Modell nur die Darstellung in den Trainingsdaten gelernt hat oder allgemeine objektbezogene Muster erkannt hat, ist eine Erklärbarkeitsanalyse. Wie in 1.2.3 beschrieben, ist das Wissen eines Modelles in den Gewichten gespeichert. Dies führt zur einer schweren Nachvollziehbarkeit der Grundlage einer Detektion durch den menschlichen Betrachter; NN wurden deshalb lange als „Black Box“ bezeichnet.
Der Bereich der erklärbaren künstlichen Intelligenz (engl. explainable artificial intelligence, XAI) versucht, diese Black Boxen zu dekodieren und die Entscheidungslogik eines NN nachvollziehbar darzustellen. Besonders relevant ist dies für Klassifikationsmodelle, die oftmals gar nicht in ihrer Logik durch Menschen überprüft werden können. Detektionsmodelle besitzen bereits ein gewisses Maß an Erklärbarkeit, da Objekte durch BB auf dem Bild gekennzeichnet werden und relevante Merkmale für eine richtige Detektion mindestens im gleichen Bereich wie das gesuchte Objekt liegen

müssen. Unklar bleibt jedoch, welche Pixel innerhalb der BB für die Detektion ausschlaggebend waren. XAI-Ansätze versuchen, hier Abhilfe zu schaffen und die Entscheidungsgrundlage detaillierter aufzuschlüsseln.

1.6 Annotationen

Ursachen für eine möglicherweise ungenügende Performance von NN sind oft die Größe, aber auch die Qualität der zur Verfügung stehenden Datensätze. Im überwachten Lernen bestehen die Datensätze aus Rohdaten und dazugehörigen Lösungen der Aufgabenstellung (Annotationen).
Im Bereich des maschinellen Sehens ist eine Annotation eine Markierung auf bzw. ein Label zu einem Bild, wodurch eine Information maschinenlesbar zur Verfügung gestellt wird. Die Annotation stellt die Lösung dar, gegen die die Vorhersagen des Netzwerkes im Trainingsprozess und in der Evaluation verglichen werden. Je nach Aufgabe entspricht sie der gewünschten Ausgabe des Netzwerkes (siehe 1.2.4), z.B. einer Markierung und Klassifizierung gesuchter Objekte mittels BB.
Annotationen für die Entwicklung von KI stellen im medizinischen Bereich eine Herausforderung dar. Während Bilder für Alltagsanwendungen (z.B. das Vorhandensein von Hunden oder Katzen auf einem Bild) von Laien annotiert werden können, braucht es für medizinische Anwendungen Fachpersonal. Dies macht den Prozess zeitaufwendig und kostenintensiv. Fehlen Zeit oder Sorgfalt, kann es zu Annotationsfehlern kommen.

1.6.1 Annotationsfehler in der Objektdetektion

Richtlinien zur Annotation für die Objektdetektion empfehlen den Einsatz kleinstmöglicher BB, die gesuchte Objekte vollständig einschließen (siehe erstes Bild Abbildung 8) [23]. Während der Erstellung von Annotationen können verschiedene Arten von Fehlern entstehen, die in der Literatur beschrieben werden [24]: Es wird zwischen Fehlern von Einzelpersonen und fehlender Übereinstimmung mehrerer Annotator*innen unterschieden; letztere entsteht z.B. durch fehlende Kalibrierung und resultiert in Inkonsistenz der Annotationen. Des Weiteren werden Fehlklassifizierungen (ein Objekt wurde genau markiert, aber falsch benannt) und klassenunabhängigen Fehlern unterschieden; zu Letzteren gehören fehlende Objekte, zusätzliche Objekte und ungenaue Annotationen (z.B. zu große, zu kleine oder verschobene Annotationen).

Abbildung 8 zeigt eine Zusammenfassung unterschiedlicher Arten von Annotationsfehlern. In der vorliegenden Arbeit wurden ungenaue klassenunabhängige Annotationen (konsistent zu große und zu kleine BB) und fehlende Kalibrierung zwischen Annotator*innen (inkonsistente Annotationsfehler) untersucht.

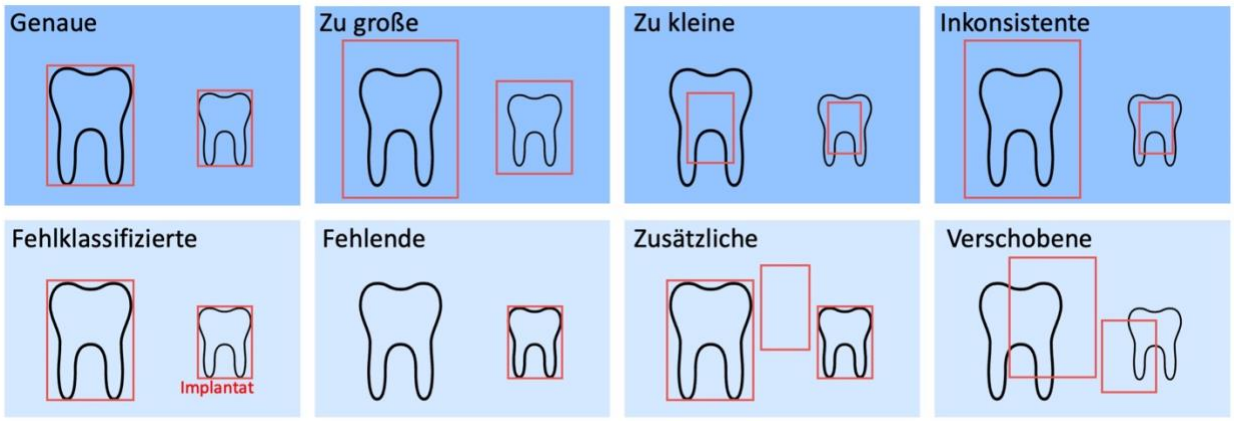

Abbildung 8: Unterschiedliche Arten von Annotationsfehlern mit Bounding Boxen (BB) für Objektdetektionsmodelle. Rote Vierecke stellen Annotationen mittels BB dar. Die Annotationsfehler, welche hier in der ersten Zeile dargestellt wurden, sind Gegenstand dieser Arbeit (eigene Darstellung).

1.6.2 Einfluss von Annotationsfehlern

Im überwachten Lernen lernt ein Modell durch die Analyse der Trainingsdaten, die Aufgabe (bspw. die Detektion von Zahnstein) zu lösen. Ausgehend von zufälligen Detektionen wird dabei, wie beschrieben, der Fehler zwischen gegebenen Annotationen und Vorhersagen minimiert. Der Einfluss von Annotationsfehlern auf diesen Prozess und die Modellperformance wurde dabei bisher nur in wenigen Studien untersucht.

So wurde im Bereich der Objektdetektion der Einfluss von zusätzlichen, verschobenen und fehlenden BB um Fahrzeuge herum mit den Netzwerkarchitekturen SSD [20] und YOLOv3 [25] untersucht. Dabei wurde eine signifikante Verschlechterung der Modellperformance bei einer Fehlklassifizierung der Fahrzeuge und einer Kombination aus Fehlerarten (verschobene, zusätzliche, fehlende und fehlklassifizierte BB) nachgewiesen [26]. Bei der Detektion von Drohnen in Videoaufnahmen mittels YOLOv3 wurden ebenfalls zusätzliche, fehlende und verschobene BB untersucht und insbesondere der Einfluss von fehlenden und zusätzlichen BB demonstriert [27]. Das in der vorliegenden Arbeit verwendete CNN YOLOv5 [22] wurde auf Fehlertoleranz bei der Detektion maritimer Fahrzeuge, wie z.B. Containerschiffe oder Segelboote, getestet.

Nach der Korrektur eines unsauberen Datensatzes in Bezug auf Fehlklassifizierungen, verschobener, fehlender, zusätzlicher und ungenauer Annotationen konnte eine signifikant verbesserte Performance des Modelles gezeigt werden [28].
Die meisten CNN Architekturen wurden an großen Datensätzen mit Fotografien aus Alltagszenen entwickelt. Für die Objektdetektion stehen z.B. Fahrzeuge, Tiere und andere Alltagsobjekte im Datensatz „Microsoft Common Objects in Context“ zur Verfügung [29]. Röntgenbilder unterscheiden sich drastisch von diesen Alltagsbildern, u.a. in ihrer Graudarstellung und ihrem Aufnahmeverfahren [30]. Medizinische Fragestellungen wurden in der Entwicklung der vorhandenen Modellarchitekturen bisher selten betrachtet und müssen getestet werden. In der medizinischen Bildgebung wurde der Einfluss von Annotationsfehlern so gut wie überhaupt nicht untersucht. Im Bereich der Segmentierung wurden zu große, zu kleine, fehlende und falsch klassifizierte Annotationen von Zellen auf histologischen Schnittbildern untersucht [31]. Während eine Falschklassifizierung nur geringen Einfluss zeigte, sorgte vor allem die zu große Segmentierung für eine geringe Modell Performance.

1.7 Fragestellung

Der Einfluss von Annotationsungenauigkeiten auf die Detektion von Objekten auf Röntgenbildern wurde nach aktuellem Kenntnisstand bisher nicht untersucht. In Arbeiten aus anderen Fachgebieten wurden vor allem fehlende, zusätzliche, verschobene BB und gemischte Annotationsfehler analysiert.
Ziel der vorliegenden Arbeit war, den Einfluss von zu großen und zu kleinen BB auf die Performance von Objektdetektionsmodellen zu untersuchen. Hierzu wurde die Modellarchitektur YOLOv5 exemplarisch anhand der Detektion von Zahnstein auf Bissflügelaufnahmen untersucht. Neben systematischen (konsistenten) Fehlern wurde ein weiteres Problem der medizinischen Annotation, die fehlende Kalibrierung zwischen mehreren Annotator*innen, simuliert (inkonsistenter Fehler). Ausgangshypothese war, dass Modelle, die anhand ungenauer Annotationen trainiert wurden, signifikant schlechter Zahnstein detektieren, als Modelle, welche anhand genauer Annotationen trainiert wurden. Hierbei sollte auch eine mögliche Maskierung durch ebenso ungenaue Testdaten untersucht werden. Außerdem sollte die Machbarkeit der Detektion kleiner Objekte mit verschwimmenden Grenzen (wie Zahnstein) auf Bissflügelaufnahmen mittels DL untersucht werden.

2. Methodik

2.1 Studiendesign

In der vorliegenden Arbeit wurde eine moderne CNN-Architektur, YOLOv5, auf Robustheit in Bezug auf Annotationsfehler untersucht. Als exemplarische Aufgabe diente die Detektion von Zahnstein auf Röntgenbildern. Abbildung 9 zeigt schematisch den Ablauf der durchgeführten Experimente. In einem ersten Experiment wurden konsistent vergrößerte und verkleinerte BB untersucht, um die ungenaue Annotation einer Einzelperson zu simulieren. In einem zweiten Experiment wurden BB inkonsistenter Größe untersucht und so fehlende Kalibrierung zwischen Annotator*innen simuliert. Die künstlich manipulierten Daten wurde genutzt, um Objektdetektionsmodelle (YOLOv5) zu trainieren und die Performance der Zahnsteindetektion auf zwei unterschiedlichen Testdatensätzen evaluiert. Die Relevanz einzelner Bildregionen für die Entscheidung einzelner Modelle wurde exemplarisch an einzelnen Bissflügelaufnahmen mittels Methoden der Erklärbarkeit untersucht. Die Experimente wurden durch die Ethikkomission der Charité – Universitätsmedizin Berlin unter der Antragsnummer EA4/080/18 genehmigt. Teile der Methoden und Ergebnisse sind in gekürzter Form in der Publikation „Impact of Noisy Labels on Dental Deep learning – Calculus Detection on Bitewing Radiographs“ veröffentlicht und die entsprechenden Abbildungen mit der Quelle gekennzeichnet [32]. Die Berichterstattung orientierte sich an der Richtlinie für Autoren für Studien zur Künstlichen Intelligenz in der Zahnmedizin [33].

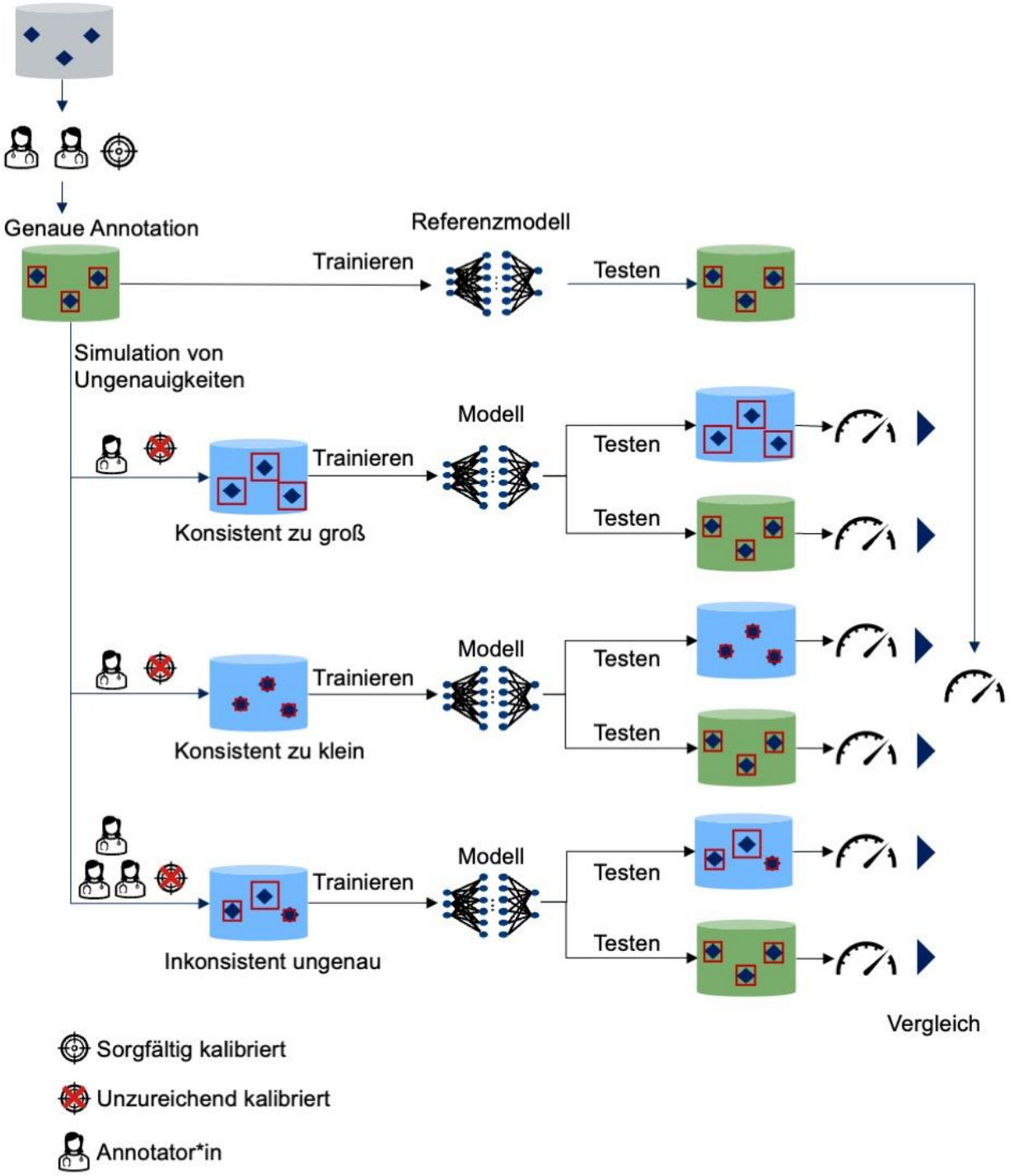

Abbildung 9: Schematische Darstellung des Studiendesigns. Ein Datensatz wurde durch zwei Annotator*innen nach Kalibrierung genau annotiert. Dieser wurde genutzt, um ein Referenzmodell zu trainieren und auf genauen Annotationen zu testen. Auf Grundlage der genauen Annotationen wurden Ungenauigkeiten simuliert: Konsistent und inkonsistent ungenaue Annotationen wurden genutzt, um Modelle zu trainieren. Diese wurden einmal auf äquivalent ungenau annotierten Daten und einmal auf genau annotierten Daten getestet und ihre Performance jeweils gegen die des Referenzmodells verglichen (eigene Darstellung).

2.2 Studienkohorte

Der eingesetzte Datensatz bestand aus 4837 Bissflügelaufnahmen, die im Rahmen der Routinebehandlung an der Charité – Universitätsmedizin Berlin (Berlin, Deutschland) generiert wurden. Die Bilder wurden mit Röntgengeräten von Dürr Dental (Bietigheim-Bissingen, Deutschland) und Dentsply Sirona (Bensheim, Deutschland) aufgenommen. Aufnahmen, die keinen Zahnstein aufzeigten, wurden exkludiert, da in der Objektdetektion alle Bereiche, auf denen keine Objekte abgebildet sind, bereits als Negativbeispiele fungieren. Die Prävalenz von Zahnstein auf Bildebene betrug 36,1%, sodass 1746 Bilder eingeschlossen werden konnten. Das Alter der Kohorte betrug im Durchschnitt 38,5 Jahre, mit einer Standardabweichung (SD) von 16 Jahren. 49% der eingeschlossenen Röntgenbilder stammte von weiblichen Patientinnen, 51% von männlichen.

2.3 Annotation

Eine genaue Annotation des Datensatzes wurde auf Grundlage der Empfehlung der Herausgeber der Netzwerkarchitektur, YOLOv5, vorgenommen: Optimale Annotationen sollen demnach möglichst kleine BB nutzen, die dennoch sicherstellen, dass das ganze Objekt vollständig eingeschlossen wurde. Dies geschah durch zwei in der Annotation von Röntgenbildern geübte Zahnärzt*innen. Nach einer Kalibrierung annotierte ein erster Zahnarzt alle Röntgenbilder streng nach dieser Regel. Eine zweite Zahnärztin überprüfte jedes Bild in einem weiteren Zyklus auf Genauigkeit, Vollständigkeit und Fehlerfreiheit und hielt bei Unstimmigkeiten Rücksprache mit dem ersten Zahnarzt.

2.4 Simulation von Annotationsfehlern

2.4.1 Konsistente Annotationsfehler

Annotiert eine Einzelperson einen Datensatz mit fehlender Sorgfalt, können konsistente Annotationsfehler entstehen. Um den Einfluss von zu großen und zu kleinen BB zu untersuchen, wurden alle BB des oben beschriebenen Datensatzes schrittweise um einen Faktor α vergrößert bzw. verkleinert, sodass gilt:

$$(1) \quad A_M = \alpha \times A_G.$$

Dabei ist A_M die Fläche der manipulierten BB und A_G die Fläche der genauen BB. Die maschinelle Verarbeitung von Annotationen in Form von BB erfolgt wie oben erläutert mittels Pixelkoordinaten. Bei YOLOv5 wird hierzu das Zentrum, die Höhe und die Breite der BB relativ zur Bildgröße genutzt. Um den Einfluss von zu großen und zu kleinen BB zu untersuchen wurde das Zentrum der BB beibehalten und nur die Höhe (h) und die Breite (b) verändert. Für die Berechnung der BB-Fläche gilt:

$$(2) \quad A_G = h \times b.$$

Aus Gleichung (1) und (2) ergibt sich für die Simulation ungenauer BB-Flächen die Multiplikation von h und b mit $\sqrt{\alpha}$:

$$(3) \quad A_M = \alpha \times A_G = \alpha \times h \times b = \sqrt{\alpha} \times h \times \sqrt{\alpha} \times b.$$

Folgende Notationen beschreiben die resultierenden originalen (genauen) und manipulierten Annotation in Koordinaten:

Original: x, y, h, b

Manipulation: x, y, $\sqrt{\alpha} \times h$, $\sqrt{\alpha} \times b$

Dabei stellen x und y die Koordinaten des Zentrums der BB dar.

Konsistent vergrößerte und verkleinerte BB wurden für folgende Faktoren α untersucht: 0,1 – 0,9 mit Steigerungen von 0,1; 1 – 9 mit Steigerungen von 1; 10 – 100 mit Steigerungen von 10. Abbildung 10 zeigt eine Annotation und beispielhafte Manipulationen an einer Bissflügelaufnahme.

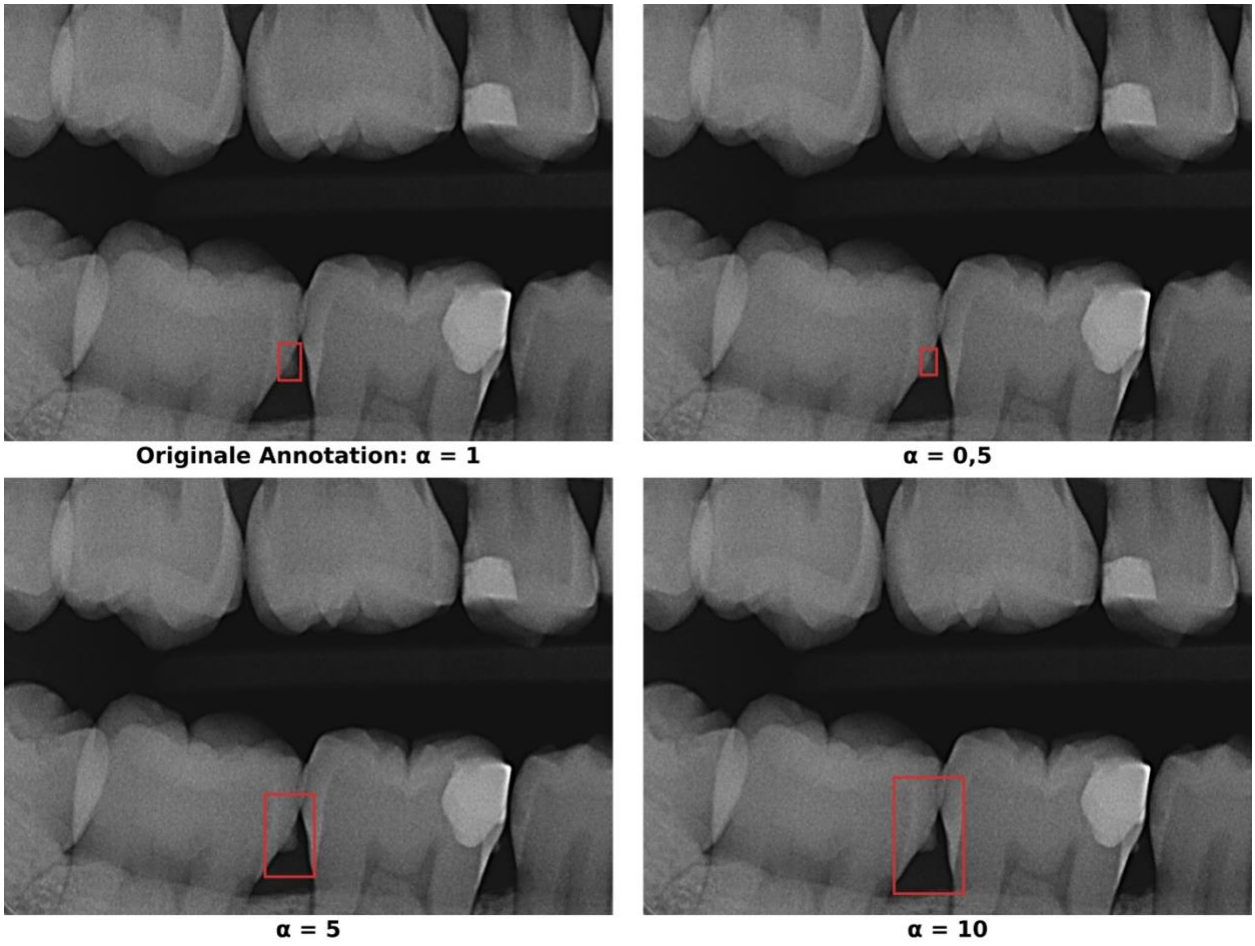

Abbildung 10: Beispielhafte Manipulation einer BB an einer Bissflügelaufnahme um den Faktor α. Die BB wurden zeilenweise von links nach rechts um den Faktor 1; 0,5; 5 und 10 verändert, wobei das erste Bild (α = 1) die genaue Annotation darstellt (modifiziert nach Büttner et al., 2023) [32].

2.4.2 Inkonsistente Annotationsfehler

In einem zweiten Experiment wurde eine fehlende Kalibrierung zwischen mehreren Personen in der Annotation untersucht. Hierzu wurde der Datensatz in drei Teile geteilt. Im ersten Drittel wurden die BB verkleinert, im zweiten vergrößert und im letzten Drittel unverändert (genau) belassen. Die Manipulation erfolgte wie in 2.4.1 beschrieben durch Multiplikation der BB-Fläche mit einem Faktor α. Bei der Simulation inkonsistenter Annotationsfehler beschreibt die Deviation δ die Abweichung der manipulierten BB-Fläche zur genau annotieren BB-Fläche, wobei $\alpha = 1 \pm \delta$. Eine Deviation von 0,3 steht demnach für eine Manipulation der BB-Fläche um den Faktor 0,7 bei der Verkleinerung

und um den Faktor 1,3 bei der Vergrößerung. Diese Manipulation erfolgte in Schritten von 0,1, was in 9 inkonsistent ungenau annotierten Datensätzen mit Deviationen von 0,1 – 0,9 resultierte. Abbildung 11 stellt die Simulation inkonsistenter Ungenauigkeiten schematisch dar.

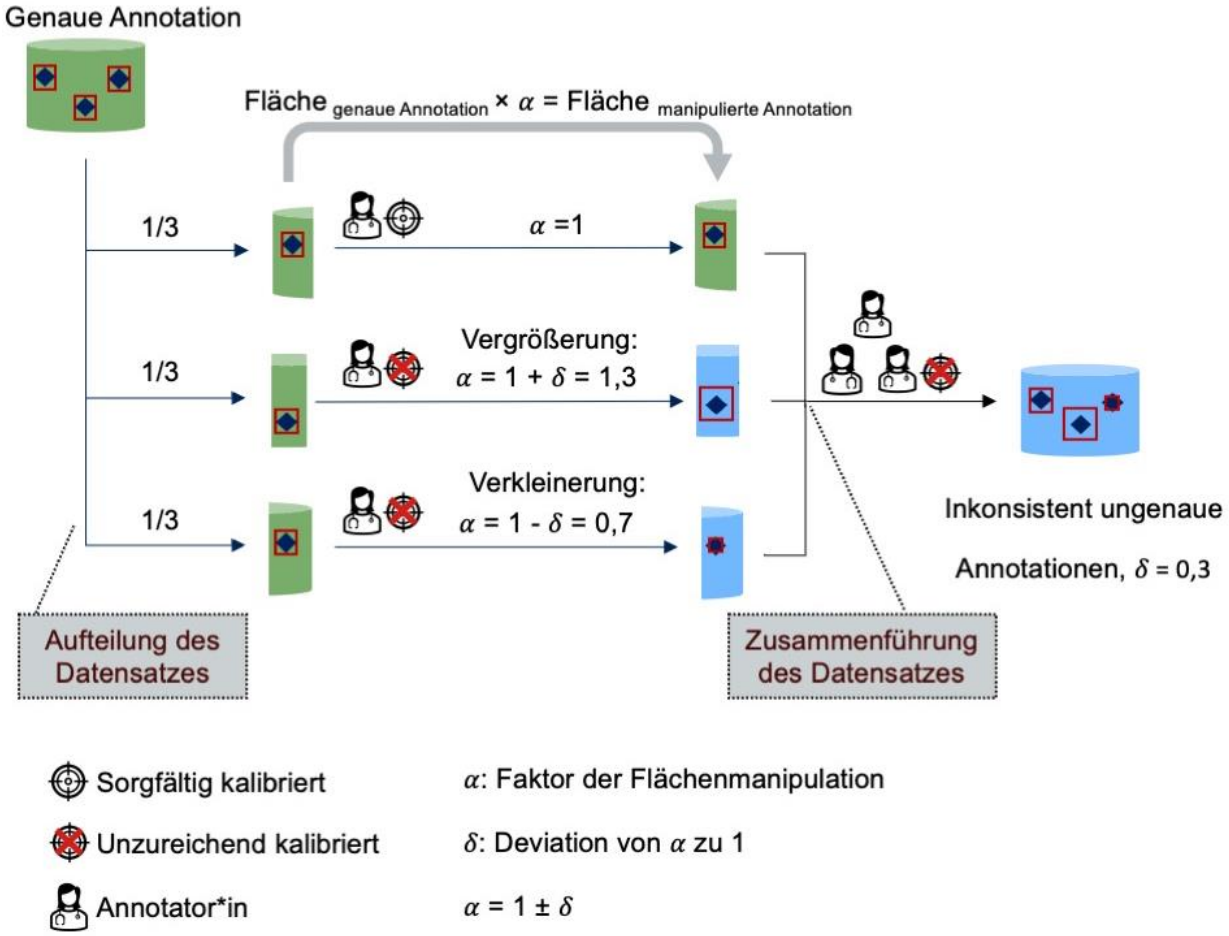

Abbildung 11: Schematische Darstellung der Manipulation der genau annotierten Daten, um inkonsistent ungenau annotierte Daten zu simulieren. Die genau annotierten Daten werden in drei Teile geteilt. Ein Teil wird genau belassen, der zweite konsistent vergrößert und der dritte konsistent verkleinert. Hierzu wird die Fläche der genauen Annotation mit einem Faktor α multipliziert. Die Abweichung zur genauen Annotation wird durch die Deviation δ beschrieben, wobei $\alpha = 1 \pm \delta$ (eigene Darstellung).

2.5 Neuronales Netzwerk zur Objektdetektion

2.5.1 Netzwerkarchitektur

Die einphasige Netzwerkarchitektur YOLOv5 steht in unterschiedlichen Größen zur Verfügung. Die Größe beschreibt die Anzahl der trainierbaren Parameter. Die größte Version YOLOv5x, bestehend aus 8,6 * 10^7 Parametern, wird für die Detektion kleiner Objekte empfohlen, weshalb sie für die vorliegende Arbeit gewählt wurde. Die Berechnung der Verlustrate erfolgt durch eine architekturspezifische Funktion, die Verbundverlustfunktion (engl. compound loss). Diese besteht aus den Komponenten Klassenverlust, Lokalisationsverlust und Objektivitätsverlust. Sie beruht auf einer Kombination aus den Funktionen „Binary Cross Entropy“ und „Complete Intersection over Union“, wobei letzteres für die Optimierung der Lokalisation und Größe der BB verwendet wird. Es wurde eine auf dem Datensatz „Microsoft Common Objects in Context“ vortrainierte Version des Modelles verwendet [29].

2.5.2 Trainingsprozess und Netzwerkparameter

Abbildung 12 illustriert schematisch den Trainingsprozess des in dieser Arbeit verwendeten Netzwerkes YOLOv5. Er entsprach dem in 1.2.3 beschriebenen Vorgang. Bissflügelaufnahmen wurden mit einer Auflösung von 640 × 640 als Eingabe in das CNN übergeben. Nach Prozessierung der Bilder innerhalb des Netzwerkes resultierte eine Ausgabe, die alle detektierten Objekte des Röntgenbildes durch drei Komponenten (siehe oben) beschrieb. Jede dieser Komponenten wurde durch die Verbundverlustfunktion evaluiert und so im Laufe des Trainings optimiert. Hierzu wurde der sogenannte stochastische Gradientenabstieg (engl. stochastic gradient descent (SGD)) als Optimierungsfunktion genutzt. Um die Richtung des notwendigen Parameterupdates zu identifizieren, wird beim SGD der Gradient der Verlustfunktion gebildet, aus dem abgeleitet werden kann in welche Richtung die Verlustfunktion am stärksten fällt und das Parameterupdate erfolgen sollte. Die Schrittgröße dieses Updates entsprach 0,01.

Jedes Modell wurde für 300 Epochen trainiert. Das Training wurde unterbrochen, wenn nach 100 Epochen keine Verbesserung mehr stattgefunden hat (early stopping). Um die Menge der Trainingsdaten künstlich zu vergrößern (Datenaugmentation), wurden die in

YOLOv5 integrierten Methoden der horizontalen Spieglung und der Mosaikerstellung angewandt. In jeder Iteration des Modelles durch die Daten wurden 16 Bilder betrachtet (engl. batch size). Training und Auswertung des Netzwerkes erfolgte auf Nvidia A100 40GB Grafikkarten.

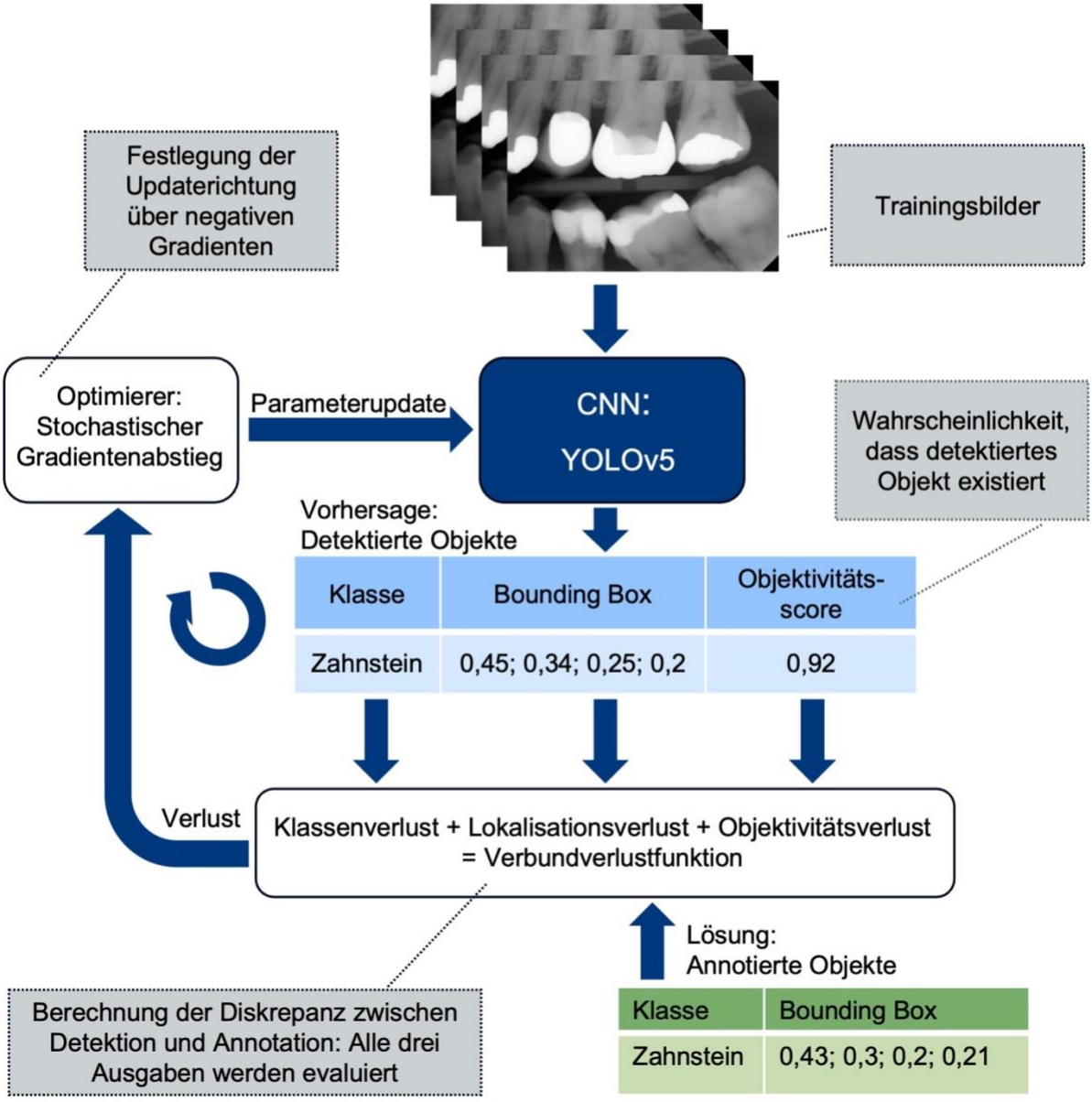

Abbildung 12: Schematische Illustration des durchgeführten Trainingsprozesses des konvolutionalen neuronalen Netzwerkes (CNN) YOLOv5 zur Objektdetektion von Zahnstein. CNN-Eingabe: Bissflügelaufnahmen; CNN-Ausgabe: Klasse, Koordinaten und Objektivitätsscore detektierter Objekte. Über eine Verbundverlustfunktion wurde die Detektion mit der Lösung verglichen und der Verlust berechnet. Dieser wurde an eine Optimierungsfunktion, dem

stochastischen Gradientenabstieg, übergeben. Der Optimierer berechnete das Update der Modellparameter, wonach die CNN-Parameter angepasst wurden und ein neuer Durchlauf beginnen konnte (eigene Darstellung).

2.6 Auswertung

2.6.1 Testdaten

Die Performance der trainierten Modelle wurde anhand separater Testdatensätze gemessen. Alle trainierten Modelle wurden anhand zwei verschiedener Testdatensätze evaluiert. Testdatensätze mit genauen Annotationen wurden genutzt, um die Genauigkeit der Zahnsteindetektion zu evaluieren und eine Referenzgruppe definieren. Eine zweite Testung mittels ungenauer Annotationen simulierte die reale Situation, in der davon ausgegangen werden kann, dass die Testdaten dem gleichen Annotationsprozess entstammen wie die Trainingsdaten. Hierfür wurde die in 2.4. beschriebenen Manipulationen der Annotationen jeweils auf dem gesamten Datensatz (einschließlich des Testdatensatzes) angewandt.

2.6.2 Kreuzvalidierung

Das einmalig zufällige Teilen der Daten in Trainings-, Validierungs- und Testdaten (welches üblich ist) kann dafür sorgen, dass der Testdatensatz zufällig Bilder enthält, bei dem das Modell besonders gut oder besonders schlecht Objekte detektiert. Um diesen Fehler zu minimieren, wurde eine sogenannte Kreuzvalidierung durchgeführt.
Ein vollständiger Datensatz wurde in 5 Teile geteilt. Für die Auswertung wurde jeder Teil einmal als Testdatensatz benutzt, wobei die übrigen 4 Teile für das Training und die Validierung während des Trainings zur Verfügung standen (engl. 5-fold cross validation). Diese Teilung erfolgte äquivalent für den manipulierten und den genau annotierten Datensatz, wobei bei jedem Trainingsdurchlauf aus beiden Datensätzen ein Testdatensatz entnommen wurde. Abbildung 13 zeigt schematisch die Aufteilung der Daten zur Kreuzvalidierung auf genau und ungenau annotierten Daten.

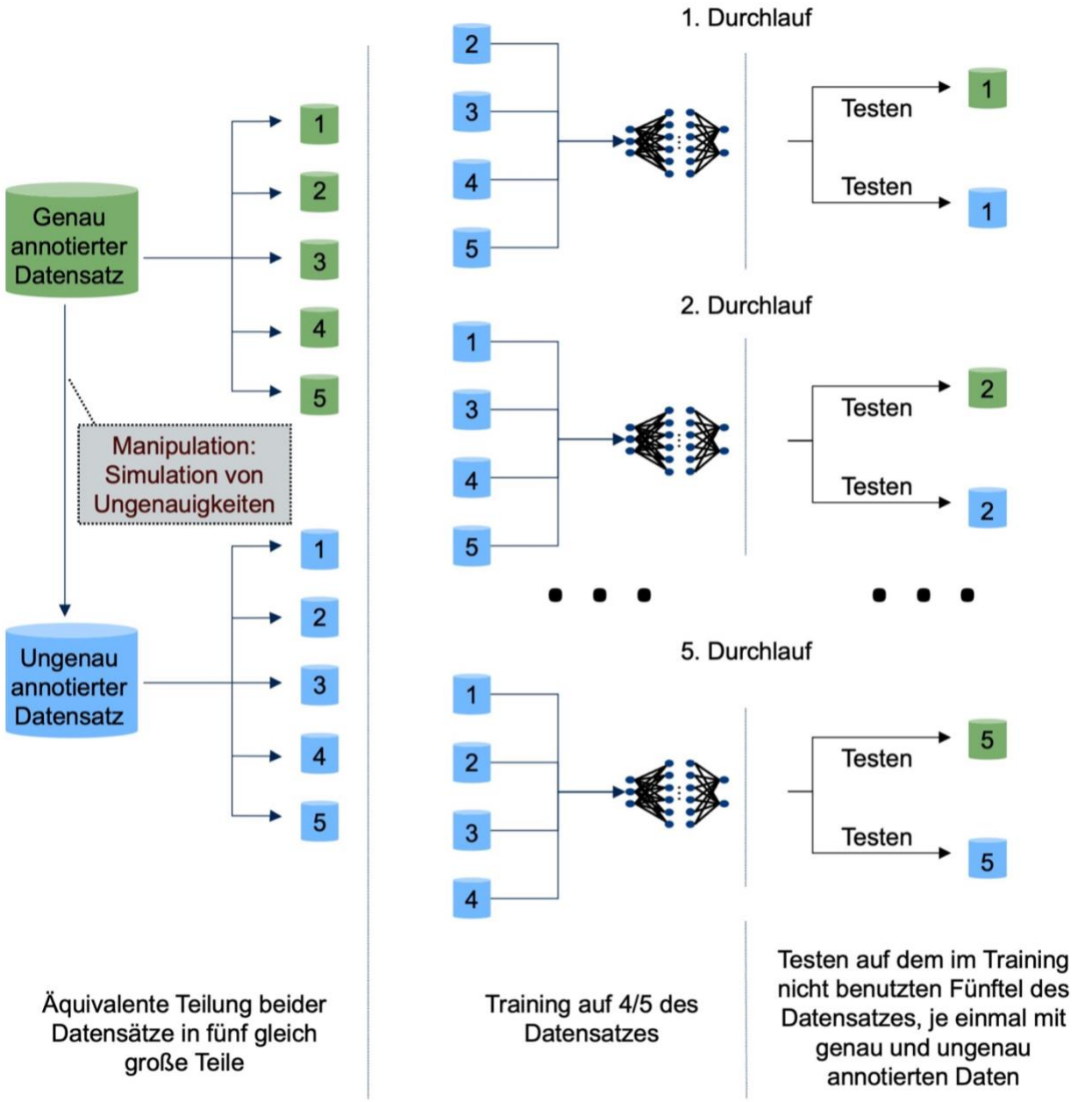

Abbildung 13: Schematische Darstellung der Testung der Modelle mittels fünffacher Kreuzvalidierung. Durch Manipulation der genau annotierten Daten wird ein ungenau annotierter Datensatz simuliert. Beide Datensätze werden in 5 Teile geteilt. In 5 Durchläufen wurden je vier Teile des ungenau annotierten Datensatzes für das Training und ein Teil zum Testen genutzt. Für eine zweite Testung wird das äquivalente Fünftel aus dem genau annotierten Datensatz entnommen (eigene Darstellung).

2.6.3 Metriken

Die Auswertung von Objektdetektionsmodellen erfolgt auf Objektebene. Da die Menge möglicher Objekte in der Objektdetektion nicht definiert ist (ein Bissflügel kann beliebig

viele Zahnsteinobjekte enthalten), können richtig negative (RN) Objekte nicht definiert werden. Zur Auswertung eines Objektdetektionsmodelles muss betrachtet werden, ob die detektieren Objekte existieren (richtig positiv, RP) oder nicht (falsch positiv, FP). Außerdem muss evaluiert werden, ob alle Objekte eines Bildes gefunden wurden oder Objekte vergessen wurden (falsch negativ, FP). Die gelingt durch die Kombination aus den Metriken Sensitivität und Genauigkeit.

Die Performance der trainierten Modelle dieser Studie wurde daher durch die in der Objektdetektion übliche Metrik, die durchschnittliche mittlere Genauigkeit (engl. mean average precision, mAP), ausgewertet. Sie ergibt sich aus der Berechnung des gewichteten Mittels der Genauigkeit (engl. precision, P) bei unterschiedlicher Sensitivität (engl. recall, R) des Modells. Die Berechnung von P und R ist in Gleichung (4) und Gleichung (5) dargestellt.

$$(4) \quad P = \frac{RP}{RP + FP}$$

$$(5) \quad R = \frac{RP}{RP + FN}$$

Um Ergebnisse mit unterschiedlicher Sensitivität zu berechnen, werden Vorhersagen des Modelles bei variierender Detektionssicherheit ausgewertet. Ein Grenzwert (engl. confidence threshold) bestimmt, welche Detektionen berücksichtigt werden. Ein hoher Grenzwert sorgt für eine Wertung derer Ergebnisse, bei dem sich das Modell sehr sicher war; dies führt in der Regel zu weniger RP-Detektionen, aber auch zu einer Verringerung der FP. Ein kleiner Grenzwert führt dementsprechend auch zu einer Berücksichtigung unsicherer Detektionen in der Auswertung.

Wird der Durchschnitt der AP über alle definierten Klassen gebildet, ergibt sich mAP. In der vorliegenden Arbeit wurde nur eine einzelne Klasse separat evaluiert; die mAP ist demnach der AP gleichzusetzen. Die Ergebnisse der Studie wurden dennoch mittels der mAP beschrieben, um die Vergleichbarkeit zu anderen Studien der Objektdetektion zu erleichtern.

Eine Detektion wurde als RP gewertet, sofern der Betrag der Überschneidung einer detektierten BB und einer annotierten BB im Verhältnis zu ihrer Gesamtfläche (engl. Intersection over Union, IoU) größer 0,5 betrug. Abbildung 14 verdeutlicht dies schematisch.

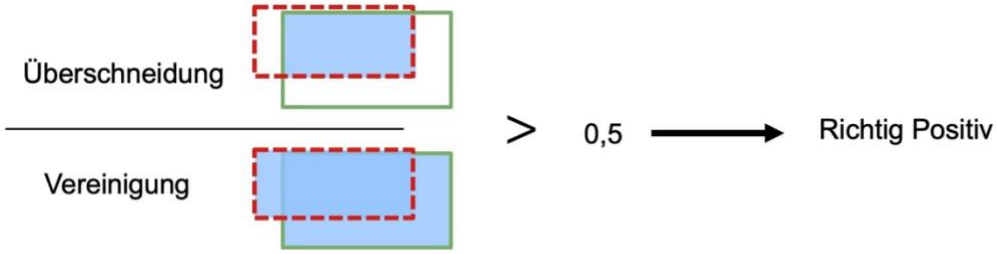

Abbildung 14: Auswertung einer richtig positiven Detektion. Das rote, gestrichelte Viereck stellt dabei die vom Modell vorhergesagte Bounding Box dar, das grüne die Annotation, gegen die das Modell geprüft wird. Ist der Anteil der Überschneidung größer als 50% der Gesamtfläche nach Vereinigung beider Boxen, wird die Detektion als richtig positiv gezählt (eigene Darstellung).

2.6.4 Statistische Analyse

Jede Annotationsart dieser Studie wurde mittels fünffacher Kreuzvalidierung untersucht (2.6.2.). So ergeben sich durch das Trainieren anhand genauer Annotationen fünf Referenzmodelle. Ebenso resultieren für jeden untersuchten Annotationsfehler fünf Ergebnisse je Testart (genaue und ungenaue Annotationen). Diese wurden gegen die Ergebnisse der Referenzmodelle auf statistisch signifikante Unterschiede untersucht. Da die Ergebnisse keiner Normalverteilung unterlagen, wurde der nicht parametrische Mann-Whitney-U-Test genutzt. Er ist geeignet, zwei unabhängige Gruppen (Modelle) in beide Richtungen zu untersuchen (die untersuchten Modell können schlechter oder besser sein als die Referenzmodelle). Die Analyse wurde mittels Python SciPy Version 1.9.0 durchgeführt [34].

2.6.5 Erklärbarkeit der Modellentscheidung

Die Modelle dieser Studie wurden exemplarisch mittels XAI untersucht. Die gewählte XAI Methode, SHapley Additive exPlanations (SHAP), basiert auf sogenannten Shapley Werten. Dabei wird der Beitrag einzelner Datenkomponenten zur Vorhersage des Modells untersucht. Bei der Objektdetektion können hierfür Bildbereiche in sogenannten Superpixel gruppiert werden. Für jeden Superpixel wird ein Wert berechnet, der den Beitrag zur Vorhersage beschreibt. Dafür werden iterativ Superpixel im Bild ausgeblendet und die Modellvorhersage auf dem Bild geprüft. Ist die Vorhersage unverändert, hat der

Superpixel einen geringen Einfluss; ist die Vorhersage nicht mehr vorhanden oder hat die Sicherheit des Modelles sich verändert, hat der Superpixel einen Einfluss auf die Vorhersage. In der vorliegenden Arbeit wurde jedes Bild in 400 Superpixel unterteilt (20 × 20). Jedes Bild wurde mit 25 zufälligen Ein- und Ausblendungen ausgewertet. Eine Ausblendung stellte dabei eine Graufärbung mit dem durchschnittlichen Grauwert des Bildes dar. Die resultierenden Shapley Werte der Superpixel wurden als Heatmap über dem Bild dargestellt. Rote Bereiche stehen für einen positiven Beitrag zu einer Detektion, während blaue Bereiche einen negativen Beitrag darstellen. Dies ermöglichte eine Einschätzung, welche Bereiche innerhalb einer BB für die Detektion entscheidend waren, insbesondere bei starken Vergrößerungen der Trainingsannotation. Abbildung 15 zeigt beispielhaft Bissflügelaufnahmen mit zufällig ausgeblendeten Pixelgruppen, wie sie dem Modell zur Detektion vorgelegt wurden.

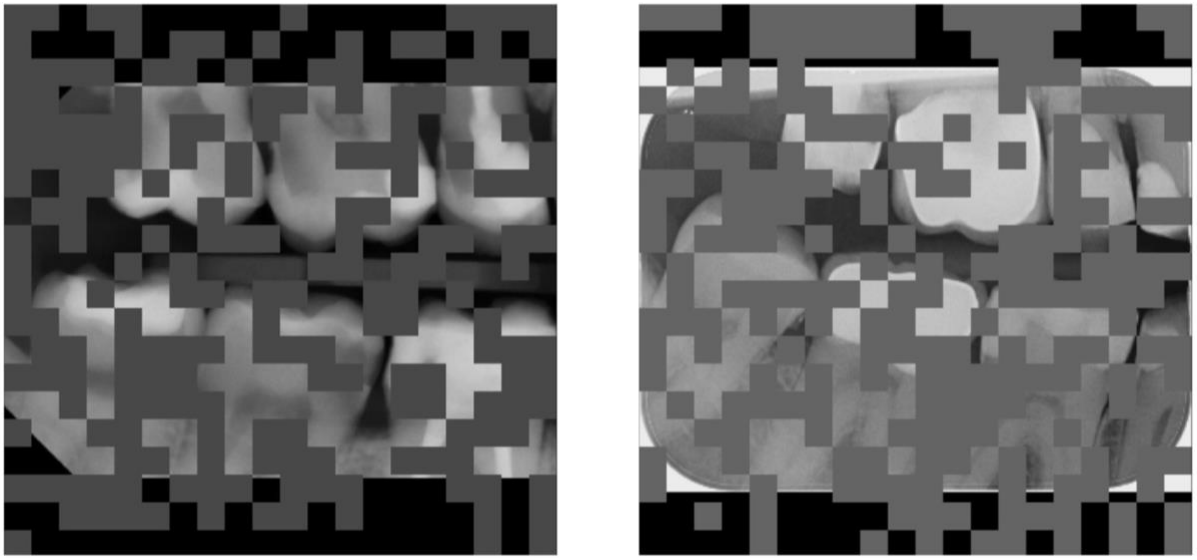

Abbildung 15: Beispiel zweier Bissflügelaufnahmen mit zufällig ausblendeten Pixelgruppen für die Evaluation mittels Erklärbarkeitsanalyse (eigene Darstellung).

3. Ergebnisse

3.1 Genaue Annotationen

Die Verteilung und Größe der genauen Annotationen ist relativ zur Bildhöhe und Breite der Bissflügelaufnahmen in Abbildung 16 dargestellt. Mit einer zur Bildgröße normalisierten Höhe × Breite von 0,005 × 0,01 – 0,15 × 0,25 liegt die Objektgröße von Zahnstein genau innerhalb diskutierten Grenzen für kleine Objekte [35], jedoch noch deutlich über der Größe, die vom Entwickler von YOLOv5 als klein diskutiert wird [22]. Die Referenzmodelle, trainiert und getestet mittels genauer Annotationen, erreichten eine durchschnittliche mAP (SD) von 0,77 (0,01).

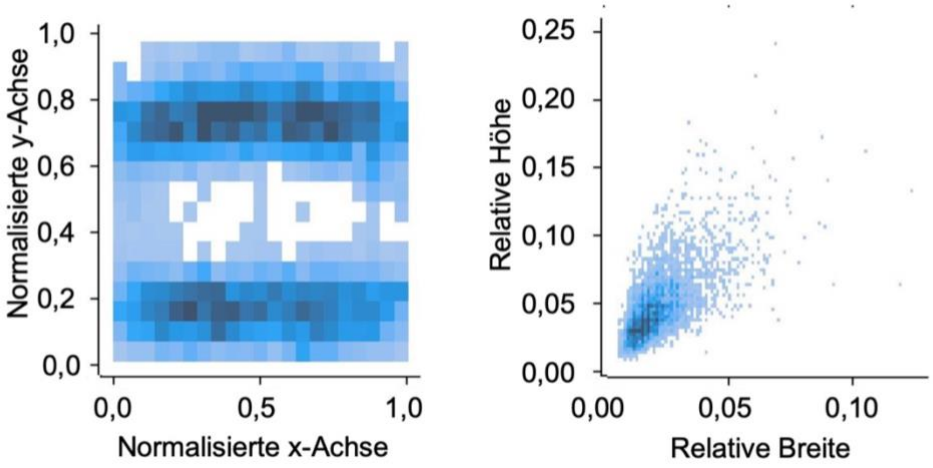

Abbildung 16: Verteilung und Größe der genauen Annotationen eines Trainingsdatensatzes. Im ersten Streudiagramm ist die Verteilung der Bouding Boxen (BB) anhand normalisierter Bildachsen der Bissflügelaufnahmen dargestellt. Auf der rechten Seite ist die Größe der BB relativ zur Höhe und Breite der Bissflügelaufnahmen dargestellt (eigene Darstellung, unter Nutzung der Funktionen der YOLOv5 Repository [22]).

3.2 Konsistente Annotationsfehler

Abbildung 17 zeigt die Ergebnisse der Modelle, die durch Datensätze mit konsistenten Annotationsfehlern trainiert wurden. Ergebnisse, basierend auf Testdatensätzen mit äquivalent fehlerhaften Annotationen (blauer gestrichelter Graph, Abbildung 17), zeigten keine signifikante Minderung der Modellperformance bis zu einem Faktor $\alpha = 60$ ($p = 0,15$). Erst bei Vergrößerungsfaktoren $\alpha > 60$ kam es zu einer signifikanten Verminderung der Modellperformance, wobei die Performance bis $\alpha = 100$ mit

mAP50 (SD) = 0,75 (0,01) auf einem hohen Niveau blieb. Einzelne Vergrößerungen sorgten sogar für eine statistisch signifikante Verbesserung der Performance (z.B. $\alpha = 7$, $p < 0{,}01$). Bei der Verkleinerung der BB zeigte sich der Effekt der Annotationsfehler hingegen umgehend; schon bei einer geringfügigen Verkleinerung der BB zeigte sich eine signifikante Verminderung der Performance ($\alpha \leq 0{,}8$; $p < 0{,}01$).

Das Testen auf genau annotierten Daten (grüner gepunkteter Graph, Abbildung 17) zeigte ebenfalls einen signifikanten Einbruch der Performance bei verkleinerten Annotationen ($\alpha \leq 0{,}8$; $p < 0{,}05$). Bei diesem Testverfahren wurde außerdem auch eine drastische Verminderung der Modellperformance bei vergrößerten BB gezeigt. Schon eine Vergrößerung um den Faktor $\alpha = 2$ führte zu einer signifikanten verschlechterten Perfomance im Vergleich zum Basismodell ($p < 0{,}01$).

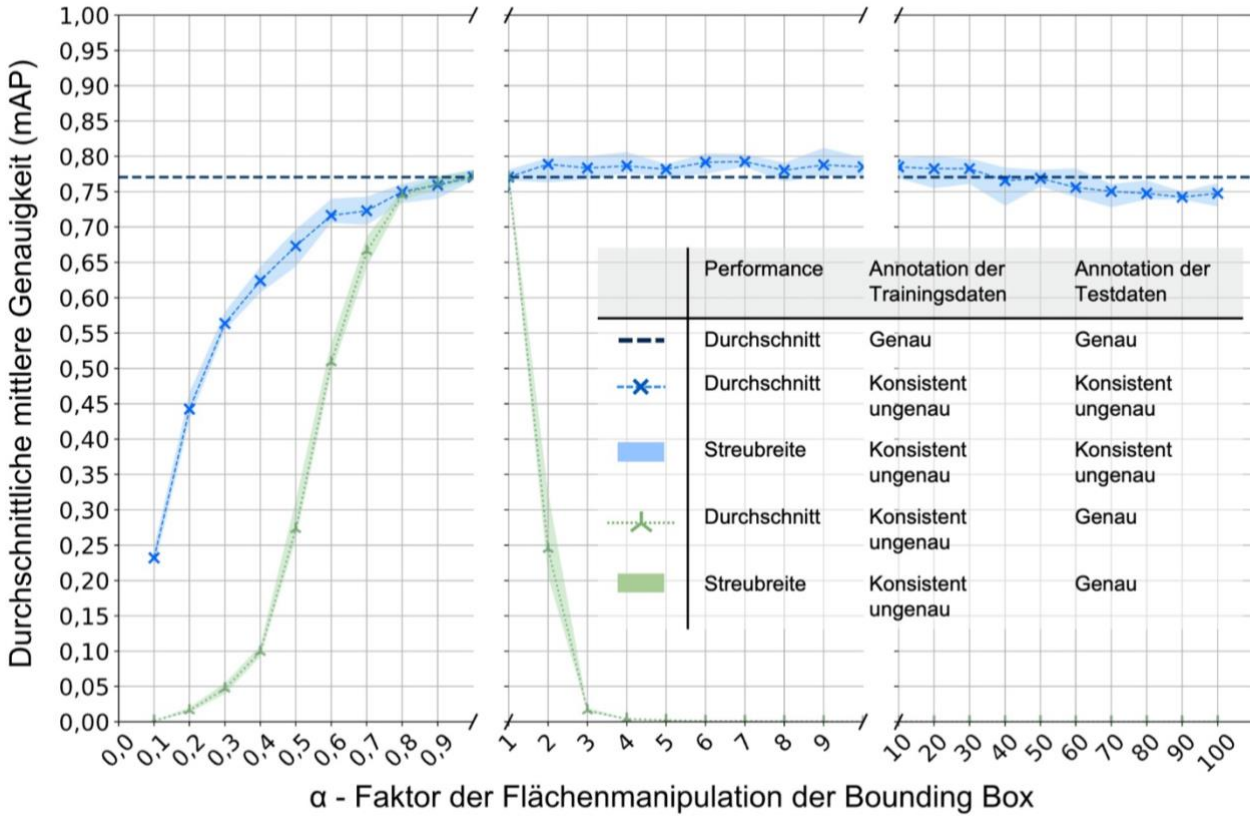

Abbildung 17: Performance der Modelle, die an Daten mit konsistenten Annotationsfehlern trainiert worden sind. Der Faktor der Flächenmanipulation ist auf einer nicht linearen x-Achse dargestellt. Die Grenzen der Streubreite beschreiben das jeweils beste bzw. schlechteste Modell der fünffacher Kreuzvalidierung (modifiziert nach Büttner et al., 2023) [32].

3.3 Inkonsistente Annotationsfehler

Die Einführung inkonsistenter Annotationsfehler (Simulation multipler Annotator*innen ohne Kalibrierung) führte sowohl bei der Testung auf äquivalent ungenauen Annotationen als auch auf genauen Annotationen zu einem negativen Effekt auf die Modellperformance (Abbildung 18). Dabei führte die Testung auf ungenau annotierten Daten (blauer gepunkteter Graph, Abbildung 18) schneller zu einer signifikant verminderten Modellperformance ($\delta = 0{,}2$, $p < 0{,}01$) als die Testung auf genau annotierten Daten ($\delta = 0{,}3$, $p < 0{,}05$) (grüner gestrichelter Graph, Abbildung 18).

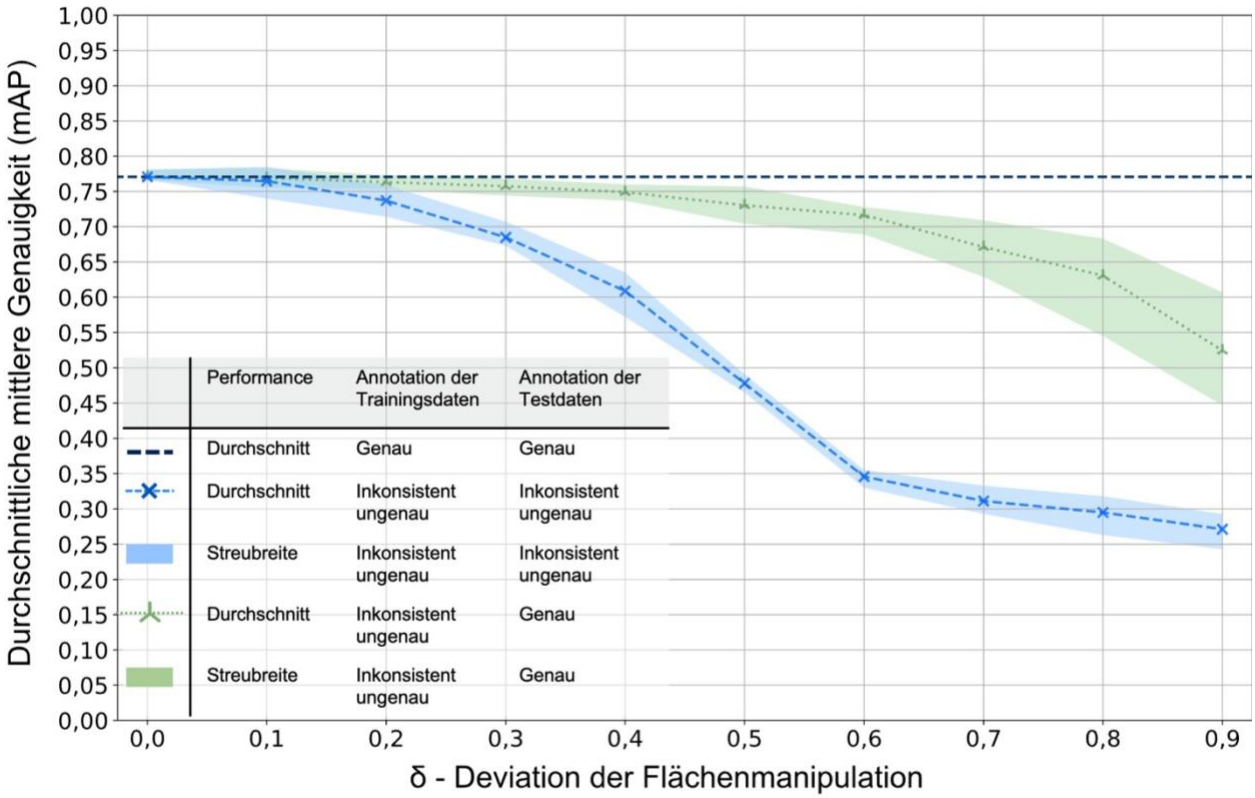

Abbildung 18: Performance der Modelle, die an Daten mit inkonsistenten Annotationsfehlern trainiert worden sind. Die Grenzen der Streubreite beschreiben das jeweils beste bzw. schlechteste Modell der fünffacher Kreuzvalidierung (modifiziert nach Büttner et al., 2023) [32].

3.4 Erklärbarkeitsanalyse

In der exemplarischen Analyse einzelner Bilder mittels XAI (SHAP) wurde dargestellt, welche Bildbereiche für eine Detektion besonders relevant waren. Abbildung 19 zeigt die Auswertung der Detektionen von drei Modellen, die mit genauen bzw. konsistent vergrößerten Annotationen trainiert wurden. Die roten Bildbereiche zeigen Superpixel mit einem positiven Beitrag zur Detektion. Die Superpixel, in denen Zahnstein abgebildet ist,

zeigen sich hierbei dunkelrot, selbst bei einer Vergrößerung der BB um α = 100. Die äußeren Bereiche der BB sind von geringerer Relevanz (hellrot und blau dargestellt).

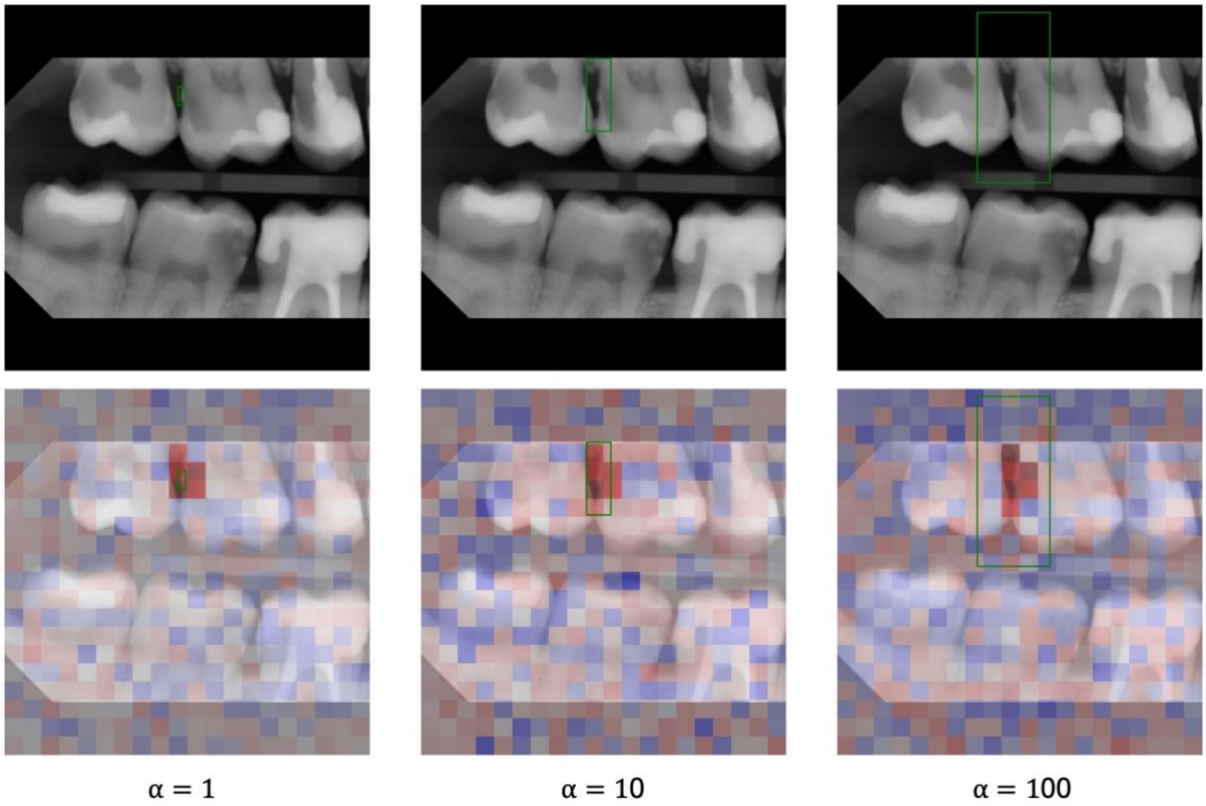

Abbildung 19: Exemplarische Evaluation von Detektionen auf einer Bissflügelaufnahme durch die Erklärbarkeitsmethode SHapley Additive exPlanations (SHAP). Eine rote Färbung visualisiert einen positiven Einfluss auf die Detektion, eine blaue Färbung einen negativen. Von links nach rechts: Detektion durch Modell trainiert mit genauen Annotationen, mit konsistent ungenauen Annotationen 10-facher (α = 10) und 100-facher (α = 100) Flächenvergrößerung (modifiziert nach Büttner et al., 2023) [32].

Abbildung 20 zeigt weitere exemplarische Untersuchungen von Detektionen eines Modelles, welches mittels konsistent vergrößerter Annotationen (α = 100) trainiert wurde. Auch hier zeigen sich die Superpixel, auf denen Zahnstein abgebildet ist, dunkelrot. Die Intensität der anderen Superpixel unterschied sich hingegen zwischen einzelnen Bildern.

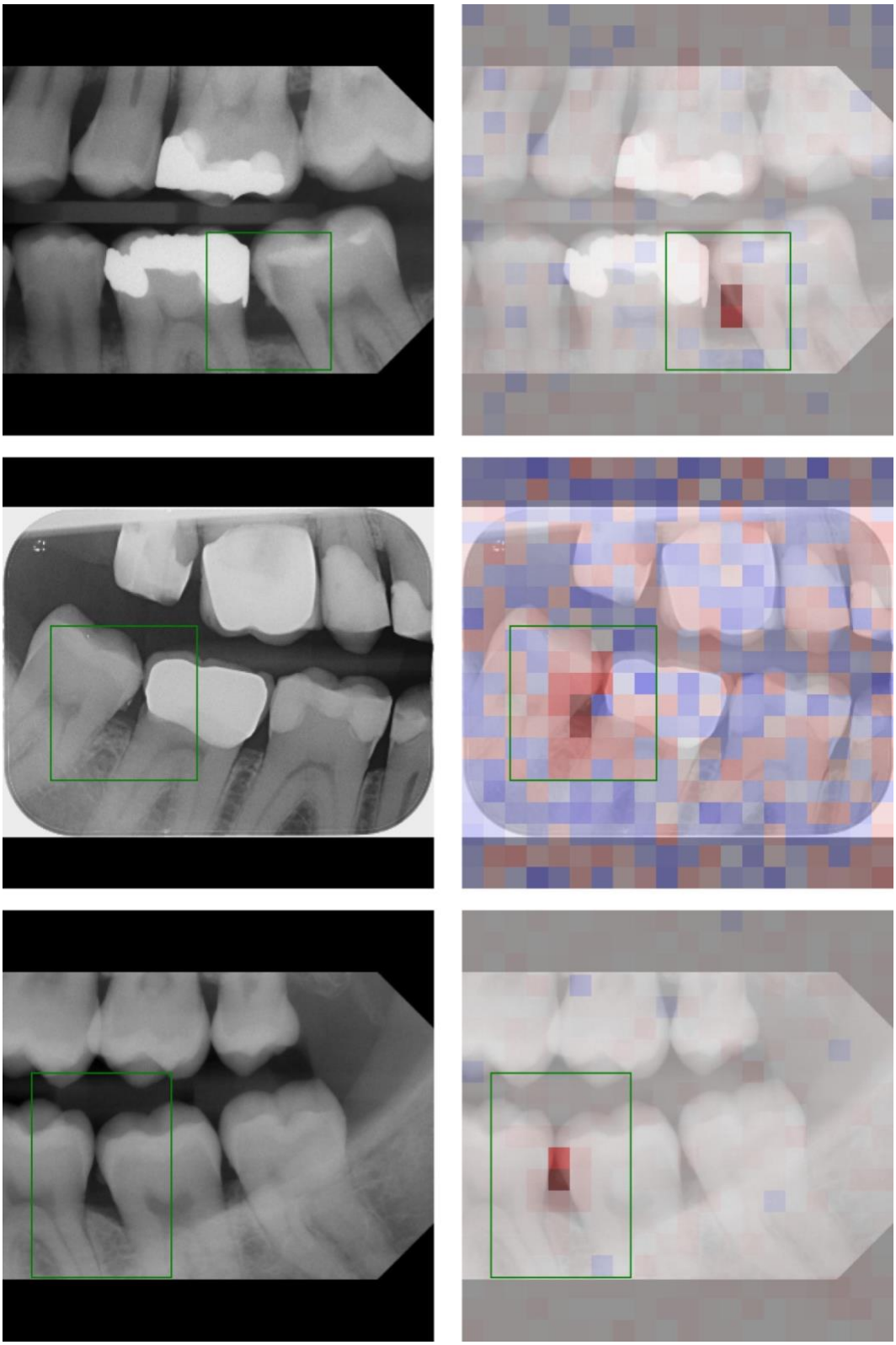

Abbildung 20: Exemplarische Evaluation von Detektionen durch ein Modell trainiert mittels Annotationen 100-facher Flächenvergrößerung durch die Erklärbarkeitsmethode SHapley Additive exPlanations (SHAP). Eine rote Färbung visualisiert einen positiven Einfluss auf die Detektion, eine blaue Färbung einen negativen (eigene Darstellung).

4. Diskussion

4.1 Zusammenfassung und Interpretation der Ergebnisse

In der vorliegenden Arbeit wurde der Einfluss von Annotationsfehlern auf die Performance von YOLOv5x an einem exemplarischen Anwendungsanfall, der Detektion von Zahnstein auf Bissflügelaufnahmen, untersucht. Das Training mit genau annotierten Daten führte zu einem Modell mit einer hohen Detektionsfähigkeit für Zahnstein (mAP (SD) = 0,77 (0,01)). Es kann davon ausgegangen werden, dass notwendige Merkmale zur Erkennung von Zahnstein erlernt wurden.

4.1.1 Konsistente Verkleinerungen

Konsistent verkleinerte BB im Trainingsdatensatz, welche die Annotation durch Einzelpersonen simulierten, führten zu einer umgehenden Verminderung der Fähigkeit der Modelle Zahnstein zu detektieren. Dieser Effekt war sowohl bei der Testung auf genau annotierten Daten als auch auf äquivalent verkleinerten BB zu beobachten. Das CNN war folglich nicht in der Lage, die Detektion von Zahnstein anhand von verkleinerten Annotationen zu erlernen.

4.1.2 Konsistente Vergrößerungen

Auch konsistent vergrößerte Annotationen führten zu einer drastischen Verminderung der Performance, wenn sie auf genau annotierten Daten getestet wurden. Die Testung auf äquivalent zu großen BB führte hingegen zu guten Testergebnissen, selbst bei massiven Vergrößerungen ($\alpha \leq 60$). Erst die Testung auf genau annotierten Daten zeigte die Ungenauigkeit des Modelles. Die Evaluation mittels XAI zeigte eine deutliche rote Färbung der Pixelgruppen, auf denen der Zahnstein abgebildet war (Abbildung 19, Abbildung 20).

4.1.3. Inkonsistente Annotationsfehler

Inkonsistente Ungenauigkeiten, wie sie beispielsweise entstehen, wenn mehrere Personen ohne Kalibrierung annotieren, führten zu einer Verminderung der

Modelfähigkeit. Die Testung auf äquivalent fehlerhaften Daten suggerierte eine schnellere Verminderung als die Testung auf genau annotierten Daten. Die inkonsistenten BB-Größen im Testdatensatz verschleierten die Fähigkeit des Modelles Zahnstein zu detektieren. Dies war auch bei fortschreitender Fehlerschwere zu beobachten.

4.1.4 Gesamtbetrachtung

Modelle, die anhand ungenauer Annotationen trainiert wurden, zeigten eine signifikant schlechtere Performance in der Detektion von Zahnstein als Modelle, welche anhand genauer Annotationen trainiert wurden. Die Ausgangshypothese kann somit angenommen werden. Die Relevanz des Testdatensatz auf die Interpretation der Modellfähigkeiten wurde demonstriert. Auch wurde die Machbarkeit der Detektion von Zahnstein auf Bissflügelaufnahmen gezeigt. Abbildung 21 fast die Interpretation der Teilergebnisse zusammen.

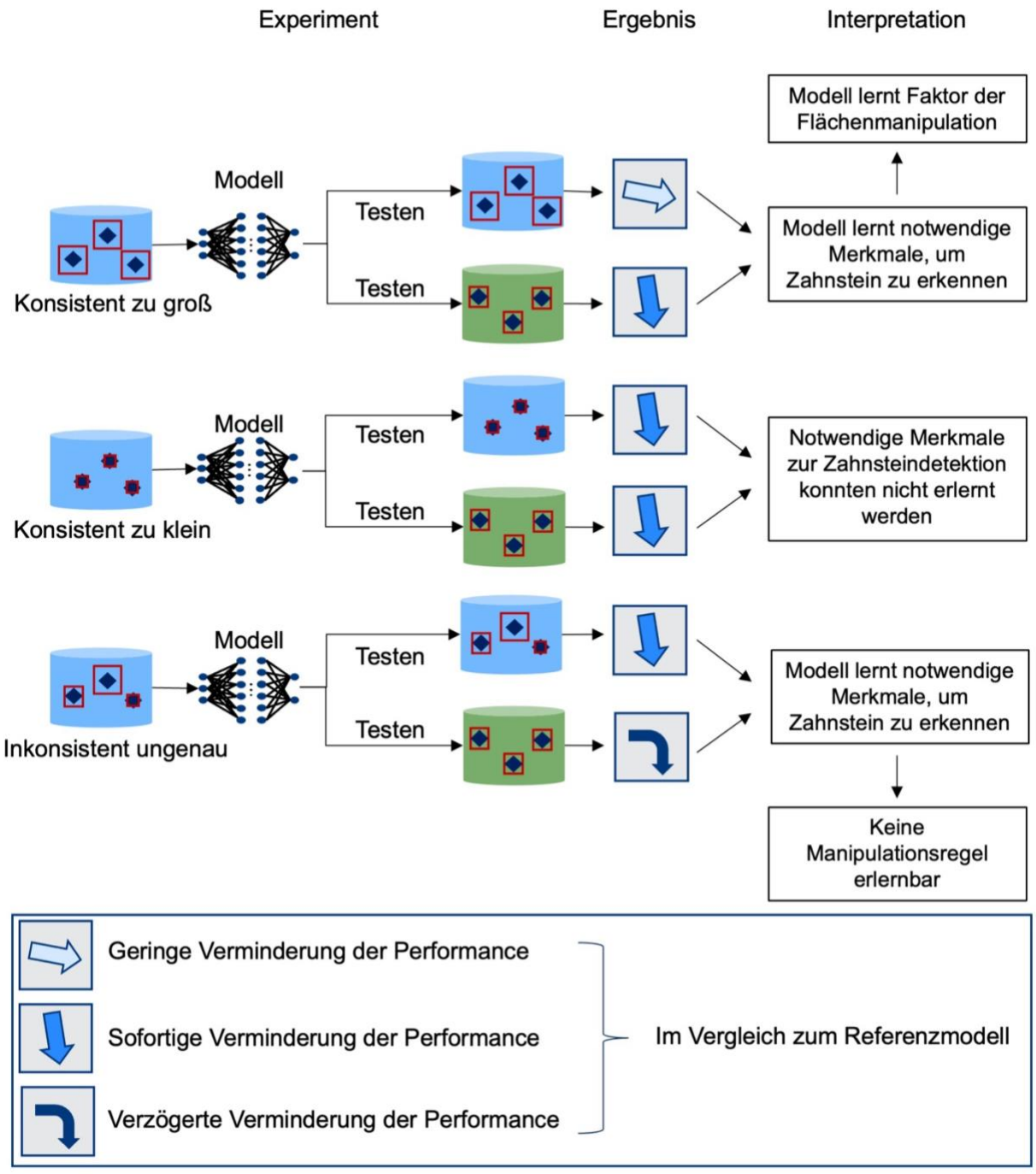

Abbildung 21: Zusammenfassung und Interpretation der Ergebnisse und schematische Darstellung der Schlussfolgerungen (eigene Darstellung).

4.2 Schlussfolgerungen unter Betrachtung der Verlustfunktion

Die Ergebnisse müssen im Detail diskutiert werden, wobei der Lernvorgang NN betrachtet werden muss. Elementarer Bestandteil des Lernprozesses ist die Minimierung der Verlustfunktion. Das verwendete Modell YOLOv5 generiert pro Detektion drei Ausgaben: Eine BB, definiert über Zentrum, Höhe und Breite, die zugehörige Klasse des

Objektes und einen Objektivitätsscore. Letzteres kann als Wahrscheinlichkeit der Existenz des detektierten Objektes interpretiert werden. Alle drei Komponenten werden im Trainingsprozess optimiert. Hierzu wird der Unterschied zwischen der Annotation und der Detektion über eine Verbundverlustfunktion gemessen. Sie besteht ,wie oben beschrieben, aus drei Teilen: Lokalisationsverlust, Klassenverlust und Objektivitätsverlust. Der Klassenverlust misst, ob das Objekt korrekt klassifiziert wurde und ist (wie auch der Objektivitätsverlust) nicht von Größenmanipulationen der Annotation betroffen. Die für den Lokalisationsverlust verwendete Funktion („Complete Intersection over Union“) evaluiert jedoch, wie exakt das Zentrum des Objektes getroffen wurde, ob die Seitenverhältnisse der detektierten BB korrekt waren und wie gut das Objekt durch die BB umschlossen wurde (Höhe und Breite der BB). Zu einer Minimierung dieses Teils der Verlustfunktion muss somit eine maximale IoU zwischen detektierter und annotierter BB erreicht werden. Ausgehend von der Annahme, dass das Netzwerk die notwendigen Merkmale zur Detektion von Zahnstein bei vergrößerten BB erlernt hat, ist somit zu vermuten, dass auch der Faktor der Flächenmanipulation erlernt wurde, da nur so der Lokalisationsverlust minimiert werden konnte. Da dies bei zu kleinen Annotationen nicht der Fall war, ist zu schlussfolgern, dass das Modell die notwendigen Merkmale zur Zahnsteinerkennung hier nicht erlernen konnte. Die erlernte Flächenmanipulation zeigt sich auch in den Evaluierungen mittels XAI: Die Pixelgruppen, auf denen der Zahnstein zu erkennen ist, blieben selbst bei massiver Vergrößerung der BB wichtig für die Detektion. Bei inkonsistenten Annotationsfehlern war keine allgemeine Manipulationsregel gegeben, welche das Netzwerk erlernen konnte: eine BB könnte genau, vergrößert oder verkleinert sein. Das Netzwerk war somit nicht in der Lage, eine Regel zu erlernen, die den Lokalisationsverlust minimiert, was den Abfall der Performance beider Testverfahren erklären könnte Abbildung 21.

4.3 Einbettung der Ergebnisse in den bisherigen Forschungsstand

Die automatisierte Diagnostik von Röntgenbildern gewann in der zahnmedizinischen Forschung zuletzt vermehrt an Aufmerksamkeit. Ein 2023 veröffentlichtes Review identifizierte 186 Studien aus dem Bereich DL in der Zahnmedizin, 22 davon aus dem Bereich der Objektdetektion [36]. In der Parodontologie wurden 50 Studien aus dem Bereich KI in ein ebenfalls 2023 veröffentlichtes Review eingeschlossen, 67% nutzten KI auf Bilddaten [37]. Entwickelt wurden Anwendungen auf Fotografien (bspw. für die

Segmentierung und Klassifizierung von Gingivitis oder die Detektion von parodontal geschädigten Zähnen), Panoramaschichtaufnahmen (bspw. für die Segmentierung von parodontologisch bedingten Knochenverlust) und Einzelzahnfilmen (bspw. für die Vorhersage einer Extraktion in Form einer Bildklassifizierung) [37]. Die Detektion von Zahnstein wurde zum aktuellen Kenntnisstand noch nicht untersucht und ihre Machbarkeit in dieser Arbeit gezeigt.

Der Einfluss von menschlichen Fehlern in den Annotationen auf die Performance und die Evaluation der Netzwerke wurde bisher in der Zahnmedizin noch nicht betrachtet. Außerhalb der Zahnmedizin existieren erste Studien, allerdings bleibt dabei die Genauigkeit der Testdaten meist unberücksichtigt. Zudem sind durch fehlende isolierte Betrachtung einzelner Annotationsfehler oft keine Schlussfolgerungen über den Einfluss der jeweiligen Fehler (bspw. zu großer BB) möglich.

Im medizinischen Bereich gibt es noch weniger Untersuchen zum Einfluss von Annotationsungenauigkeiten. Im Anwendungsbereich histopathologischer Schnittbilder wurde der Einfluss von ungenauen Annotationen auf Segmentierungsmodelle untersucht [31]. Auch hier wurde der negative Einfluss von zu großen und zu kleinen Segmentierungen gezeigt. Zur Testung der Modellperformance wurden synthetisch generierte Bilder und Annotationen benutzt, welches eine genaue Testung zuließ.

Die Ergebnisse der vorliegenden Arbeit ergänzen die bisherigen Studien mit der isolierten Untersuchung von zu großen und zu kleinen BB in zwei unterschiedlichen Szenarien. Eine Vergleichbarkeit ist aufgrund unterschiedlicher Anwendungen und Annotationsfehler anderer Studien nur bedingt gegeben.

4.4 Stärken und Schwächen der Studie

Nach dem bisherigen Kenntnisstand ist die vorliegende Arbeit die erste, die den Einfluss von Annotationsfehlern auf Modelle zur Objektdetektion in der Medizin untersucht; sie ist ebenso die erste, die den Einfluss von Annotationsungenauigkeiten im Bereich der zahnmedizinischen Bildanalytik betrachtet. Die zwei untersuchten Szenarien (konsistente und inkonsistente Größenfehler) gewährleisten eine vielseitige Interpretation der Ergebnisse (siehe 4.1). Durch die zweifache Auswertung aller Modelle auf genau und ungenau annotierten Testdaten konnte die Studie die Relevanz genauer Testdaten und den möglichen Maskierungseffekt ungenauer Testdaten demonstrieren. Dies ist

insbesondere für die Zahnmedizin relevant: Während in anderen Domänen standardisierte Testdatensätze (sogenannte Benchmarking-Daten) genutzt werden, um die Modellperformance zu testen, fehlt es an diesen in der Zahnmedizin. Infolgedessen stammen Testdatensätze meist aus der gleichen Datenquelle wie die Trainingsdaten; fehlerhafte Annotationen würden demnach sowohl in Test- and als auch Trainingsdaten vorkommen.

Die Verifizierung mittels Kreuzvalidierung stellte robuste Ergebnisse sicher und minimierte den zufälligen Einfluss einzelner Bilder. Als weitere Neuheit dieser Studie ist die Evaluation mittels einer XAI-Methode zu nennen, welche eine Interpretation der relevanten Merkmale bei massiven BB-Vergrößerungen ermöglichte.

Die Studie wies jedoch auch eine Reihe von Limitationen auf. (1) Der Einfluss von Fehlern in der Größe von BB wurde exemplarisch an einer einzelnen zahnmedizinischen Anwendung und einer Netzwerkarchitektur demonstriert; die Transferierbarkeit auf andere Probleme, Modellarchitekturen und Verlustfunktionen ist demnach nur bedingt gewährleistet. Verlustfunktionen, in denen die Größe der detektierten BB einen geringeren Einfluss hat, würden höchstwahrscheinlich zu einer geringeren Adaption an die BB-Größe führen.

(2) Die untersuchten Modelle wurden nicht mittels Hyperparametersuche (engl. hyperparameter tuning) optimiert. Das Suchen von optimalen Hyperparametern ist sehr ressourcen-intensiv und hätte für jedes der 185 Modelle (5-fache Kreuzvalidierung von 37 Datensätzen) durchgeführt werden müssen. Da die Fragestellung der Studie auch ohne die Entwicklung optimierter Modelle beantwortet werden konnte, wurde auf eine Hyperparameteroptimierung verzichtet. Es muss davon ausgegangen werden, dass eine Optimierung von Hyperparametern auf Basis der Verbundverlustfunktion von YOLOv5 das Modell weiter an die konsistent ungenauen Annotationen angepasst hätte. (3) Die trainierten Modelle wurden nicht auf Generalisierbarkeit überprüft, sondern nur auf Daten einer Population, Patient*innen der Charité – Universitätsklinik Berlin, trainiert. Die Anwendung des Modelles auf Daten einer anderen Population oder anderen Röntgengeräten kann nicht gewährleistet werden. Auch hier ist zu betonen, dass das Ziel dieser Studie nicht die Entwicklung eines Modelles zur klinischen Anwendung war und die Modellgeneralisierbarkeit nicht im Vordergrund stand.

4.5 Implikationen für Praxis und Forschung

Die automatisierte Detektion von Zahnstein als Teil des Mundgesundheitsscreenings, u.a. mittels Bissflügelröntgen, kann in der Mundhygieneaufklärung unterstützen und fördert möglicherweise eine frühzeitige Entfernung des Zahnsteins. Nutzer*innen etwaiger Anwendungen in diesem Bereich sollten die Relevanz von genau und ungenau annotierten Trainings- und Testdaten kennen. Die Ergebnisse der vorliegenden Arbeit zeigen, wie Fehler in Trainings- und Testdaten die Fähigkeiten bzw. die Testergebnisse eines Modelles beeinflussen. Zur Bewertung etwaiger Modelle sollte ein besonderes Augenmerk auf den Testdaten liegen, da diese etwaig geminderte Modellperformances maskieren könnten.

Zudem untermauert die Studie die Relevanz kalibrierter und genauer Annotationen. Annotationsvorgänge sollten stets mit einer sorgfältigen Kalibrierung beginnen. Sollten Ressourcen nur eingeschränkt zur Verfügung stehen, sollte besonderes Augenmerk auf die genaue Annotation der Testdaten gelegt werden. Erst dies sichert die Interpretierbarkeit der Ergebnisse und lässt Schlussfolgerungen auf die Genauigkeit der Fähigkeiten eines Modelles zu.

Wie oben diskutiert, hängt das Lernverhalten eines NN stark von der gewählten Verlustfunktion ab. Für NN aus dem Bereich der Bildklassifizierung haben sich Verlustfunktion des mittleren absoluten Fehlers (engl. mean absolute error) und des mittleren quadratischen Fehlers (engl. mean squared error) als robust gegenüber Annotationsfehlern erwiesen [38,39]. Sofern die Korrektur von Annotationsfehlern nicht möglich ist, sollte die Verlustfunktion den Fehlern entsprechend gewählt werden. In der Klassifizierung von Bildern kann der Fehler nur in der gewählten Klasse liegen, bei der Detektion von Objekten sind Fehlermöglichkeiten vielseitiger (Abbildung 8) und fehlerrobuste Verlustfunktionen sind wenig beschrieben. Wie oben diskutiert, hat in der vorliegenden Arbeit der Lokalisationsverlust als Teil der Verbundverlustfunktion die Überlappung der BB bewertet. Eine geringere Gewichtung dieses Teiles der Funktion würde möglicherweise zu einem robusteren Modell beim Vorliegen von Größenfehlern in der Annotation beitragen. Die weitere Exploration von fehlerrobusten Verlustfunktionen für die Objektdetektion ist erstrebenswert.

5. Schlussfolgerungen

Ungenauigkeiten in der Annotation hatten einen signifikanten Einfluss auf die Objektdetektion von Zahnstein mit YOLOv5. Während konsistente Annotationsfehler zu einer Anpassung des Modelles an die Fehler führten, sorgten übermäßig kleine Annotationen zu einem Scheitern des Lernprozesses. Sowohl konsistent als auch inkonsistent ungenaue Testdaten verschleierten die Fähigkeiten der Modelle Zahnstein zu detektieren. Bei der Evaluation von KI-Modellen sollte daher ein besonderes Augenmerk auf die Testdaten gelegt werden. Mediziner*Innen die Modelle in der Behandlung einsetzen, sollten sich dem möglichen Einfluss von ungenau annotierten Daten bewusst sein. Datenwissenschaftler*innen sollten das Problem durch Entwicklung robuster Netzwerkarchitekturen und Verlustfunktionen adressieren. Untersuchungen weiterer Anwendungsfälle, Netzwerkarchitekturen und Verlustfunktionen sollten durchgeführt werden.

Literaturverzeichnis

1. Revilla-León M, Gómez-Polo M, Barmak AB, Inam W, Kan JYK, Kois JC, Akal O. Artificial intelligence models for diagnosing gingivitis and periodontal disease: A systematic review. J Prosthet Dent. 2022 Mar 14;S0022-3913(22)00075-0.

2. Mohammad-Rahimi H, Motamedian SR, Pirayesh Z, Haiat A, Zahedrozegar S, Mahmoudinia E, Rohban MH, Krois J, Lee JH, Schwendicke F. Deep learning in periodontology and oral implantology: A scoping review. J Periodontal Res. 2022 Oct;57(5):942–51.

3. Röntgendiagnostik: Häufigkeit und Strahlenexposition für die deutsche Bevölkerung [Internet]. Bundesamt für Strahlenschutz. BfS; [cited 2023 May 14]. Available from: https://www.bfs.de/DE/themen/ion/anwendung-medizin/diagnostik/roentgen/haeufigkeit-exposition.html

4. Weber T. Memorix Zahnmedizin. 3. Auflage. Georg Thieme Verlag; 2010. p. 45 of 616.

5. Phipps KR, Stevens VJ. Relative contribution of caries and periodontal disease in adult tooth loss for an HMO dental population. Journal of public health dentistry. 1995;55(4):250–2.

6. Montandon A, Zuza E, Toledo BE. Prevalence and reasons for tooth loss in a sample from a dental clinic in Brazil. Int J Dent. 2012;2012:719750.

7. Timmerman MF, van der Weijden GA. Risk factors for periodontitis. Int J Dent Hyg. 2006 Feb;4(1):2–7.

8. Archana V. Calculus detection technologies: where do we stand now? J Med Life. 2014;7 Spec No. 2(Spec Iss 2):18–23.

9. Goodwin T, Devlin H, Glenny A, O'Malley L, Horner K. Guidelines on the timing and frequency of bitewing radiography: a systematic review. British dental journal. 2017;222(7):519–26.

10. Intelligenz (Psychologie) - Enzyklopädie - Brockhaus.de [Internet]. [cited 2023 May 13]. Available from: https://brockhaus.de/ecs/enzy/article/intelligenz-psychologie

11. Zhang W, Yang G, Lin Y, Ji C, Gupta MM. On definition of deep learning. In: 2018 World automation congress (WAC). IEEE; 2018. p. 1–5.

12. Was ist künstliche Intelligenz und wie wird sie genutzt? | Aktuelles | Europäisches Parlament [Internet]. 2020 [cited 2023 Jan 5]. Available from: https://www.europarl.europa.eu/news/de/headlines/society/20200827STO85804/was-ist-kunstliche-intelligenz-und-wie-wird-sie-genutzt

13. Chassagnon G, Vakalopolou M, Paragios N, Revel MP. Deep learning: definition and perspectives for thoracic imaging. European radiology. 2020;30(4):2021–30.

14. Scherer A. Neuronale Netze: Grundlagen und Anwendungen. E-Book. Vieweg+Teubner Verlag; 2013. p. 4 of 259.

15. Schneider L, Arsiwala-Scheppach L, Krois J, Meyer-Lueckel H, Bressem KK, Niehues SM, Schwendicke F. Benchmarking Deep Learning Models for Tooth Structure Segmentation. J Dent Res. 2022 Oct 1;101(11):1343–9.

16. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation [Internet]. arXiv; 2014 [cited 2023 Jun 18]. Available from: http://arxiv.org/abs/1311.2524

17. Girshick R. Fast R-CNN [Internet]. arXiv; 2015 [cited 2023 Jun 18]. Available from: http://arxiv.org/abs/1504.08083

18. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks [Internet]. arXiv; 2016 [cited 2023 Jun 18]. Available from: http://arxiv.org/abs/1506.01497

19. Redmon J, Divvala S, Girshick R, Farhadi A. You Only Look Once: Unified, Real-Time Object Detection [Internet]. arXiv; 2016 [cited 2023 Jun 18]. Available from: http://arxiv.org/abs/1506.02640

20. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single Shot MultiBox Detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision – ECCV 2016. Cham: Springer International Publishing; 2016. p. 21–37. (Lecture Notes in Computer Science; vol. 9905).

21. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal Loss for Dense Object Detection [Internet]. arXiv; 2017 [cited 2022 Apr 28]. Available from: https://arxiv.org/abs/1708.02002

22. GitHub - ultralytics/yolov5 at v6.0 [Internet]. GitHub. [cited 2022 Apr 3]. Available from: https://github.com/ultralytics/yolov5

23. Annotation Best Practices for Object Detection · Nanonets [Internet]. [cited 2023 Mar 3]. Available from: https://nanonets.github.io/tutorials-page/index.html

24. Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis. 2020;65:101759.

25. Redmon J, Farhadi A. YOLOv3: An Incremental Improvement [Internet]. arXiv; 2018 [cited 2022 Apr 28]. Available from: https://arxiv.org/abs/1804.02767

26. Chadwick S, Newman P. Training object detectors with noisy data. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE; 2019. p. 1319–25.

27. Koksal A, Ince KG, Aydin Alatan A. Effect of Annotation Errors on Drone Detection with YOLOv3. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2020. p. 4439–47.

28. Kim JH, Kim N, Park YW, Won CS. Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset. Journal of Marine Science and Engineering. 2022;10(3):377.

29. Lin TY, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P. Microsoft COCO: Common Objects in Context [Internet]. In arXiv; 2014. Available from: https://arxiv.org/abs/1405.0312

30. Ke A, Ellsworth W, Banerjee O, Ng AY, Rajpurkar P. CheXtransfer: Performance and Parameter Efficiency of ImageNet Models for Chest X-Ray Interpretation. Proceedings of the Conference on Health, Inference, and Learning. 2021 Apr 8;116–24.

31. Vădineanu Ş, Pelt DM, Dzyubachyk O, Batenburg KJ. An Analysis of the Impact of Annotation Errors on the Accuracy of Deep Learning for Cell Segmentation. Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. 2022 Dec 4;1251–67.

32. Büttner M, Schneider L, Krasowski A, Krois J, Feldberg B, Schwendicke F. Impact of Noisy Labels on Dental Deep Learning—Calculus Detection on Bitewing Radiographs. Journal of Clinical Medicine. 2023 Jan;12(9):3058.

33. Schwendicke F, Singh T, Lee JH, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J, others. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. Journal of dentistry. 2021;107:103610.

34. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261–72.

35. Nguyen ND, Do T, Ngo TD, Le DD. An Evaluation of Deep Learning Methods for Small Object Detection. Journal of Electrical and Computer Engineering. 2020 Apr 27;2020:1–18.

36. Arsiwala-Scheppach LT, Chaurasia A, Müller A, Krois J, Schwendicke F. Machine Learning in Dentistry: A Scoping Review. Journal of Clinical Medicine. 2023 Jan;12(3):937.

37. Scott J, Biancardi AM, Jones O, Andrew D. Artificial Intelligence in Periodontology: A Scoping Review. Dentistry Journal. 2023 Feb;11(2):43.

38. Ghosh A, Kumar H, Sastry PS. Robust Loss Functions under Label Noise for Deep Neural Networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

39. Rusiecki A. Trimmed Robust Loss Function for Training Deep Neural Networks with Label Noise. In: Rutkowski L, Scherer R, Korytkowski M, Pedrycz W, Tadeusiewicz

R, Zurada JM, editors. Artificial Intelligence and Soft Computing. Cham: Springer International Publishing; 2019. p. 215–22.

Eidesstattliche Versicherung

„Ich, Martha Büttner, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema:
Der Einfluss von Annotationsgenauigkeit auf die automatisierte Detektion von Zahnstein auf Bissflügelaufnahmen (The Impact of Label Accuracy on Dental Calculus Detection on Bitewing Radiographs using Deep Learning)
selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.
Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren*innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

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

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Anteilserklärung an den erfolgten Publikationen

Martha Büttner hatte folgenden Anteil an den folgenden Publikationen:

Publikation 1: Büttner M, Schneider L, Krasowski A, Krois J, Feldberg B, Schwendicke F, Impact of Noisy Labels on Dental Deep Learning—Calculus Detection on Bitewing Radiographs, Journal of Clinical Medicine, 23.04.2023

Beitrag im Einzelnen:

Die Fragestellung dieser Arbeit wurde eigenhändig von mir entwickelt und gemeinsam mit dem Data Science Team der Abteilung und meinen Betreuern Schwendicke F und Krois J diskutiert und präzisiert. Ich konzipierte die Annotation inklusive der Erstellung einer schriftlichen Annotationsanleitung zur Kalibrierung zwischen beiden Annotator*innen. Die Zusammenstellung der Bilddaten und Administration der Annotation erfolgte durch Krois J. Die Annotation der Daten erfolgte gemeinsam durch Feldberg B und mich. Die Simulation von Ungenauigkeiten in den Annotationen zur Erstellung der zu untersuchenden Datensätze wurde von mir mittels Python durchgeführt. Das Resultat wurde durch mich exemplarisch in der Abbildung 1 der Publikation dargestellt. Ich bereitete die Rohdaten und Annotationen für die Objektdetektionsmodelle mittels Python vor. Das Training und die Evaluierung aller Objektdetektionsmodelle erfolgten durch mich unter technischer Beratung und Unterstützung der Data Scientists Krasowski A und Schneider L. Die Ergebnisse wurden durch mich in Abbildung 2 und 3 der Publikation dargestellt. Ich wendete selbstständig einen Algorithmus zur Erklärbarkeit der Modellentscheidungen an und stellte das Ergebnis exemplarisch in Abbildung 4 der Publikation dar. Die statistische Auswertung der Ergebnisse erfolgte nach Beratung mit Krois J und Pitchika V (Statistiker) durch mich. Ich verfasste den ersten Entwurf des Manuskriptes in Zusammenarbeit mit Schwendicke F, welcher durch alle Autor*innen geprüft und überarbeitet wurde. Die Einreichung des Manuskripts und dessen Revision erfolgte durch mich in Unterstützung aller Autor*innen der Publikation.

_____

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

_____

Unterschrift des Doktoranden/der Doktorandin

Druckexemplar der Publikation

Journal of *Clinical Medicine*

MDPI

Article

Impact of Noisy Labels on Dental Deep Learning—Calculus Detection on Bitewing Radiographs

Martha Büttner [1,2,*], Lisa Schneider [1,2], Aleksander Krasowski [1], Joachim Krois [2], Ben Feldberg [1] and Falk Schwendicke [1,2]

[1] Department of Oral Diagnostics, Digital Health and Health Services Research, Charité—Universitätsmedizin Berlin, 14197 Berlin, Germany
[2] ITU/WHO Focus Group AI4Health, Topic Group Dental Diagnostics and Digital Dentistry, CH-1211 Geneva 20, Switzerland
* Correspondence: martha.buettner@charite.de

Abstract: Supervised deep learning requires labelled data. On medical images, data is often labelled inconsistently (e.g., too large) with varying accuracies. We aimed to assess the impact of such label noise on dental calculus detection on bitewing radiographs. On 2584 bitewings calculus was accurately labeled using bounding boxes (BBs) and artificially increased and decreased stepwise, resulting in 30 consistently and 9 inconsistently noisy datasets. An object detection network (YOLOv5) was trained on each dataset and evaluated on noisy and accurate test data. Training on accurately labeled data yielded an mAP50: 0.77 (SD: 0.01). When trained on consistently too small BBs model performance significantly decreased on accurate and noisy test data. Model performance trained on consistently too large BBs decreased immediately on accurate test data (e.g., 200% BBs: mAP50: 0.24; SD: 0.05; $p < 0.05$), but only after drastically increasing BBs on noisy test data (e.g., 70,000%: mAP50: 0.75; SD: 0.01; $p < 0.05$). Models trained on inconsistent BB sizes showed a significant decrease of performance when deviating 20% or more from the original when tested on noisy data (mAP50: 0.74; SD: 0.02; $p < 0.05$), or 30% or more when tested on accurate data (mAP50: 0.76; SD: 0.01; $p < 0.05$). In conclusion, accurate predictions need accurate labeled data in the training process. Testing on noisy data may disguise the effects of noisy training data. Researchers should be aware of the relevance of accurately annotated data, especially when testing model performances.

Keywords: artificial intelligence; machine learning; deep learning; computer vision; convolutional neural networks; calculus; digital imaging; radiology

Citation: Büttner, M.; Schneider, L.; Krasowski, A.; Krois, J.; Feldberg, B.; Schwendicke, F. Impact of Noisy Labels on Dental Deep Learning—Calculus Detection on Bitewing Radiographs. *J. Clin. Med.* **2023**, *12*, 3058. https://doi.org/10.3390/jcm12093058

Academic Editor: Takeyasu Maeda

Received: 15 March 2023
Revised: 14 April 2023
Accepted: 19 April 2023
Published: 23 April 2023

1. Introduction

Object detection is a computer vision technique which seeks to identify and label object instances within images by means of outlining rectangles, also called bounding boxes (BBs). Often, object detection models are trained in a supervised manner, with annotators labeling objects of interest by drawing BBs over them. In natural scenes of images this seems to be a relatively easy task for annotators as humans process such visual data without actively thinking about it. However, object detection has long moved beyond natural image scenes and has become an important component of medical image analysis [1,2].

Providing labels for medical data remains challenging. Besides datasets being smaller in comparison to open datasets of common objects (like cats or dogs) [3] experts are needed to label medical data. As labeling medical images is time consuming and as individual experts may miss certain findings often multiple experts are needed to label a dataset. These experts—even if not missing a finding—may introduce noise to a dataset during labeling: They may, if calibrated imperfectly, consistently label certain objects as too small or too large, for example, or be inconsistent in their labeling, e.g., one expert labels too small, one too large, and the third one perfectly right. The present study focuses on this type of label noise (other noise emanates from the discussed variability in accuracy, for instance).

Noisy labels in supervised learning are challenging researchers and data scientists. Supervised object detection models are trained by iterating through data. During the training process the model is being optimized to minimize the difference between the model predictions and the provided labels. Given noisy labels, the model may learn incorrect features to recognize an object. Different types of label noise have been described, e.g., inter-observer variability (inconsistent labels) and class-independent errors (consistent, e.g., too small or too big labels) [4]. The present study inspects the effects of on both aforementioned types of label noise.

In recent years first proposals have been developed to reduce the negative impact of noisy labels in machine learning, e.g., choice of loss function, data weighting or filtering of noisy labels [4,5]. Even though some approaches led to better model performance the impact of noisy labels remains not well described. Few studies have been conducted to investigate the influence of noisy labels on deep learning based object detection. The effect of additional, missing and shifted BBs was examined using SSD [6,7] and YOLOv3 [8,9] model architectures for tasks such as drone detection, demonstrating that especially missing BBs negatively impact model performance [9]. The beneficial effect of relabeling a noisy dataset with accurate labels was proven for tasks as maritime object detection using YOLOv5 [10,11]. In medicine noisy labels been explored even less often, e.g., for histopathological image analysis, demonstrating that too large labels negatively affect model performance. Further, inconsistent label sizes have been shown to be disadvantageous [12].

Demonstrating that noise is detrimental would underpin the relevance of accurate, consistent labeling. Automated detection of dental calculus on radiographs has so far not been studied but is relevant in the context of this study as calculus is represented by small irregular objects with oftentimes blurred boundaries, i.e., objects which are hard to label accurately. Bitewings are a type of dental radiograph that is used to visualize the coronal part of the posterior teeth. The main indication for bitewing radiographs is caries diagnosis. Clinically, the automated detection of calculus on radiographs could warrant further clinical examination and trigger certain therapies like professional tooth cleaning or scaling and root planning.

Our objective was to assess the impact of label noise on the performance of a state-of-the-art deep learning based object detection model for one particular problem: detection of dental calculus on bitewing radiographs. Our hypothesis was that both consistent and inconsistent noise significantly affects model performance. We further investigate the object detection models with explainable artificial intelligence (XAI) methods to visualize the impact of noisy labels on an exemplary model prediction.

2. Materials and Methods

2.1. Study Design

This study employed a commonly used deep learning based single-shot object detector: YOLOv5, which demands accurate labels for optimal performance. We employed a dataset of bitewing radiographs labelled for dental calculus by two calibrated experts using optimal (accurate) BBs. To simulate noise we first consistently increased or decreased the BBs sizes to generate consistently too small or too large labels. In a second step we increased and decreased only parts of the dataset, i.e., generated an inconsistently labeled dataset. We then explored the performance of YOLOv5 to detect calculus on these datasets and further employed methods of XAI to assess which image features were particularly relevant for the model's decision when trained on differently noisy datasets. The waiver for informed consent is approved by ethics committee Charité—Universitätsmedizin Berlin. Reporting of this study follows the checklist for authors for artificial intelligence in dental research [13].

2.2. Dataset

Our dataset contained 4837 bitewings collected during routine care at a public university clinic in Berlin, Germany with radiographic machines from Dürr Dental SE (Bietigheim-Bissingen, Germany) and Sirona Densply Inc. (Bensheim, Germany). The prevalence of

for informed consent is approved by ethics committee Charité—Universitätsmedizin Berlin. Reporting of this study follows the checklist for authors for artificial intelligence in dental research [13].

2.2. Dataset

Our dataset contained 4837 bitewings collected during routine care at a public university clinic in Berlin, Germany with radiographic machines from Dürr Dental SE (Bietigheim-Bissingen, Germany) and Sirona Dentsply Inc. (Bensheim, Germany). The prevalence of calculus on image level was 20.1%. Bitewings without calculus were excluded, resulting in 1746 included images from a German subpopulation with a mean [SD, min, max] age of 38.5 years [16.0, 4.8, 83.5], 51.0% males and 49.0% females. Two calibrated dentists, experienced in image analysis performed the labeling process. In the first cycle one expert labeled dental calculus using BBs with the aim of achieving the most accurate label (as small as possible but enclosing the whole object). A second expert checked all images in a second pass and controlled them once more resulting in the "accurate" base-case dataset. A comprehensive sample of available bitewings was used.

2.3. Simulating Noise

2.3.1. Consistent Noise

To assess the impact of consistent label noise, the area of each BB was stepwise increased or decreased by a factor α. For this purpose, the center of the BBs was kept and the height and width were each multiplied by the square root of α. These labels will be referred to as manipulated labels below:

$$\text{Original: } x, y, h, w$$

$$\text{Manipulated: } x, y, \sqrt{\alpha} \times h, \sqrt{\alpha} \times w,$$

where x and y are the coordinates of the BB center, h the height and w the width of the BB. The experiments were conducted to the following α values: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100 (Figure 1).

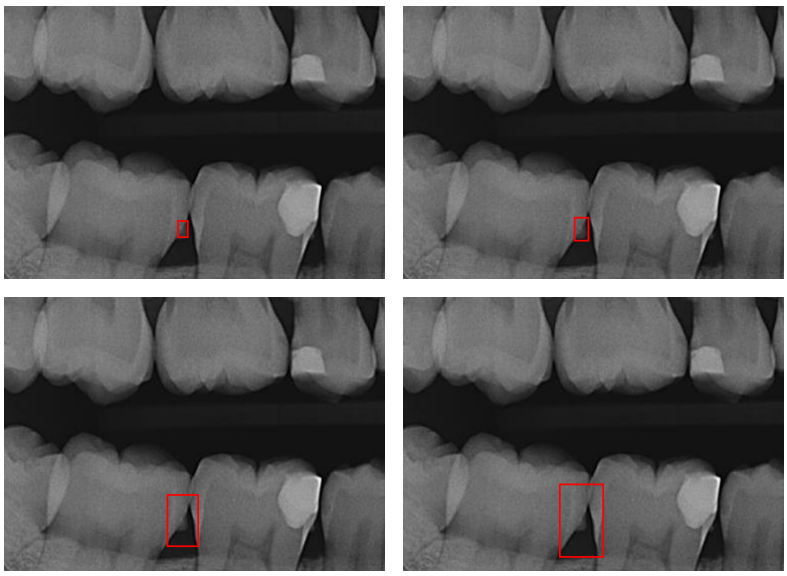

Figure 1. Example of simulated label noise: Correctly placed Bounding Boxes (BBs) outlining dental calculus were artificially increased and decreased to generate noise by multiplying the BB rectangle area with a factor α, resulting in consistently too small or too large BBs. Here, from left to right: $\alpha = 0.5$, $\alpha = 1$ (original), $\alpha = 5$, $\alpha = 10$.

2.3.2. Inconsistent Noise

A further experiment was performed to assess the impact of inconsistent noise. We simulated the behavior of three hypothetical experts, each labeling one third of the dataset (which may be the case for large datasets where it is impossible to have one expert label the

full dataset). For the first third we kept the original BBs to represent an accurate labeler. For the second and third parts we increased and decreased the BBs respectively, to simulate too small and too large labeling. The manipulation of the labels was performed as described above with different deviations of too large and too small BBs. The deviation δ from the original annotation of the BB area was systematically increased from 0.1 to 0.9 in 0.1 steps. A δ of 0.1 means an α of 1.1 for manipulating the third of the too large labeler and an α of 0.9 for manipulating the third of the too small labeler, etc.

Noise was introduced to the overall dataset (including the validation and test dataset), while for the evaluation (see below) performance testing was performed on both noisy test data (as it can be expected that the test data would usually be drawn from an overall noisily labelled dataset as well) and accurate (non-noisy) test data to evaluate the "true" effect of noise, which may be disguised by testing on noisy data but also to gauge if only paying specific attention to labeling the test dataset would be an option.

Data preprocessing and manipulation was done with Pythons pandas library version 1.4.1 [14].

2.4. Model

In this study the state-of-the-art object detection model architecture "You Only Look Once" version 5 (YOLOv5) was employed [10]. YOLOv5 is a one-stage object detector architecturally similar to the single shot detector (SSD) [7] and RetinaNet [15]. YOLOv5 provides different sub architectures such as YOLOv5x, which is recommended for small objects. The model was pretrained on the Microsoft Common Objects in Context (COCO) dataset [3].

Each model was trained for up to 300 epochs, referring to a complete pass through the entire training dataset. The training process was stopped after 100 epochs without improvement on validation data (early stopping). Mosaic and right-left-flip data augmentation, techniques to increase the dataset with image modifications to build a more robust model, were applied. The model was optimized using stochastic gradient descent, an optimization algorithm to minimize the difference between the model predictions and the true labels (loss function). The number of images used in each iteration (batch size) was set to 16, while the step size to updates the model parameters (learning rate) was set to 0.01. Training was performed with an image resolution of 640 $\times$ 640. Five-fold cross validation was performed. Data was randomly split into separate training, validation and test sets of 60%, 20% and 20% respectively. For each fold the model for evaluation on the test set was selected based on the epoch with best performance on the validation set. All computations were performed on Nvidia A100 40 GB GPU.

2.5. Model Evaluation

Performance was evaluated using mean average precision with the intersection over union (*IoU*) threshold set to 50% (mAP50). *IoU* describes the overlap of the predicted BB (*pBB*) with the ground truth BB (*gBB*) in relation to the total area of unified BBs:

$$IoU = \frac{pBB \cap gBB}{pBB \cup gBB}$$

With the *IoU* being higher than the given threshold the object is counted as correctly detected (true positive—*tp*) or not (false positive—*fp*). Average precision (AP) is the weighted mean of Precision in the Precision-Recall-Curve—calculating Precision and Recall (sensitivity) with different model confidence thresholds.

Precision (*P*) describes what proportion of the detected calculus is truly dental calculus:

$$P = \frac{tp}{tp + fp}$$

Recall (R), in medical domain better known as sensitivity, describes how many of all existing concretions are detected:

$$R = \frac{tp}{tp + fn}$$

where fn are false negative/not detected BB. Mean average precision (mAP) is the mean of AP over all classes:

$$\mathrm{mAP} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{AP_i}$$

where N is the number of classes. Since in this experiment only one class was considered, mAP is equivalent to AP. We nevertheless refer to mAP because it is the common metric to compare object detection models.

All models were examined for significant differences compared to the base-case model using non-parametric Mann-Whitney-U-test. p-values below 0.05 were considered statistically significant. The statistical analysis was performed with Pythons SciPy library version 1.9.0 [16].

2.6. Explainability

In order to interpret the model results an XAI method, namely SHapley Additive exPlanations (SHAP) based on Shapley Values was applied [17]. Shapley values capture the contribution of each feature to the prediction in comparison to the average prediction. Within object detection tasks these features are created by grouping pixels in the input images to form super pixels. The super pixels were subsequently included and excluded, and it was evaluated how much this affected the output of the model. The results were represented as heatmaps overlayed on the input image, i.e., the contribution of each super pixels to the model prediction was represented. Each image was divided into 400 (20×20) super pixels. The evaluation was performed 25 times per detection.

3. Results

3.1. Base-Case Model

The model trained and tested on the base-case (accurate) dataset resulted in a mean [SD, min, max] mAP50 of 0.77 [0.01, 0.77, 0.78]. The base-case model is used as reference model and all models trained on noisy labels were compared against it (blue dashed line in Figures 2 and 3).

3.2. Consistent Noise

Models trained on consistently smaller BBs showed significantly and escalatingly decreased performance when tested on noisy test data (orange graph in Figure 2; $p < 0.05$/ Mann-Whitney). In contrast, increasing BB size up to $\alpha = 60$ did not lead to significant changes in mAP50 ($p = 0.15$); moderate increase (e.g., BB size increased by $\alpha = 6$) even led to a significant improvement ($p = 0.02$). However, if tested on accurate test data (green graph in Figure 2) the effect of consistently smaller BB was considerable once more while this time also consistently increased BB sizes detrimentally affected the model. Decreased BB sizes to $\alpha = 0.8$ and increased to $\alpha = 2$ already led to significant performance drops ($p = 0.008$ for both). Performance of models trained on consistent noisy labels were listed in Table S1 and Table S2 of the supplementary material.

3.3. Inconsistent Noise

When inconsistently noisy data (to simulate different annotators) was used, increasing noise had escalating detrimental effects when tested on inconsistently noisy data (orange graph in Figure 3). A δ of 0.2 or more caused a significant decrease to mean [SD, min, max] mAP50 0.74 [0.02, 0.71, 0.76] ($p = 0.008$). When the model was tested on accurate data (green graph in Figure 3) this effect was slightly attenuated; a δ of 0.3 or more resulted in a significant deterioration ($p = 0.03$), mean [SD, min, max] mAP50 of 0.76 [0.01, 0.74, 0.76].

Performance of models trained on inconsistent noisy labels were listed in Tables S3 and S4 of the supplementary material.

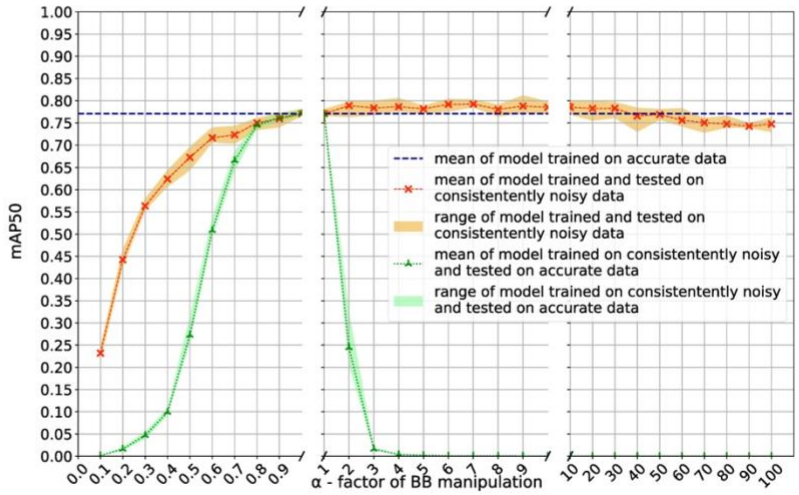

Figure 2. Performance (mean average precision with an intersection over union threshold of 50% (mAP50)) of calculus detection using deep learning models trained on consistently too small or too large BBs (α) when tested on consistently noisy test data (orange graph; mean and range of cross-validation mAP50 values) or accurate test data (green graph) compared with the model trained on accurate data (dashed blue line).

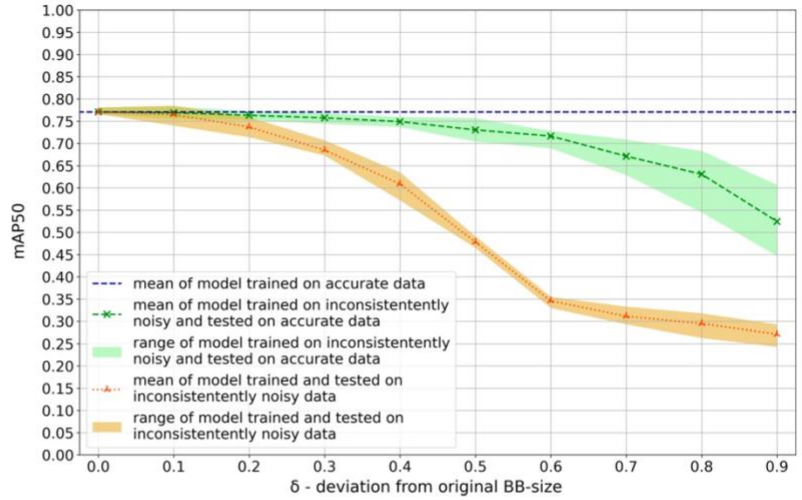

Figure 3. Model performance (mAP50) of models trained on inconsistently noisy data to simulate different annotators. Data was split into thirds. In one third of the data, the bounding boxes (BB) area was decreased, in one third kept constant, and in one third increased, respectively. δ specifies the deviation from the original size in both directions. Graphs show the mean and range values of models tested on such noisy data (orange graph) and tested on accurate data (green graph) compared with the model trained on accurate data (dashed blue line).

Figure 3. Model performance (mAP50) of models trained on inconsistently noisy data to simulate different annotators. Data was split into thirds. In one third of the data, the bounding boxes (BB) area was decreased, in one third kept constant, and in one third increased, respectively. δ specifies the deviation from the original size in both directions. Graphs show the mean and range values of models tested on such noisy data (orange graph) and tested on accurate data (green graph) compared with the model trained on accurate data (dashed blue line).

3.4. Explainability

Figure 4 shows an evaluation of a detection using SHAP. Red values represent a positive contribution of the super pixel to the detection and blue values a contribution against it. Even with an enlargement by a factor of 100, visible calculus still played an important role for the detection.

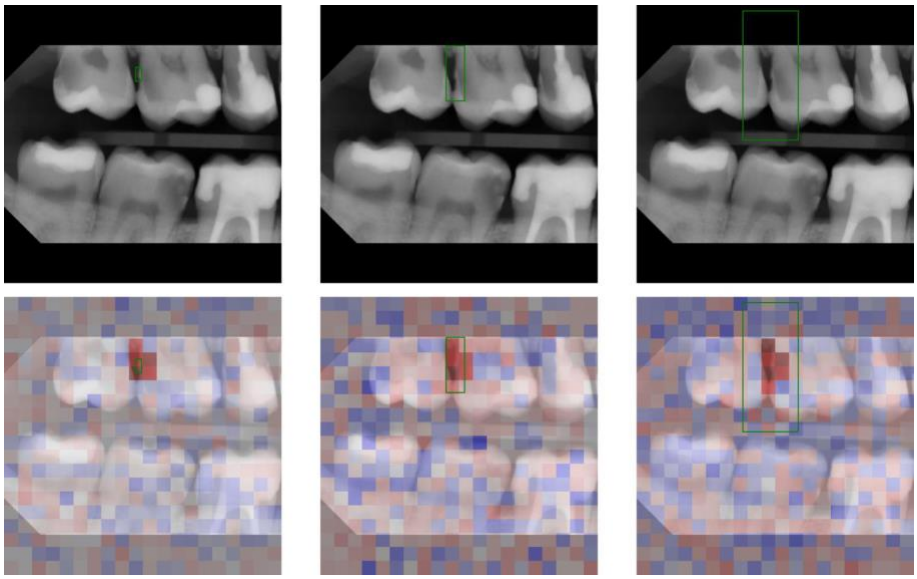

Figure 4. XAI evaluation. **First row**: Model prediction (green rectangle) trained on annotation with α = 1, α = 10, α = 100 (from **left** to **right**). **Second row**: SHapley Additive exPlanations (SHAP) heatmap where red values represent a positive contribution of the super pixel to the detection and blue values a negative contribution.

4. Discussion

Labeling medical data for deep learning is a complex and understudied aspect. Many studies published over the last years focus on developing models while basic research into understanding the impact of the labeling process and how to optimize it is scarce [2,18,19]. The present study assessed if consistent or inconsistent label noise detrimentally affects the performance of a deep learning model for a specific problem, calculus detection on bitewings. This exemplary task was chosen as calculus is highly prevalent and detecting it would be clinically relevant, but more so as labeling images for calculus is a task which may lead to noise given the discussed challenges. Our hypothesis that both consistent and inconsistent noise significantly affects the model performance needs to be accepted. Moreover, we demonstrated that the true performance of models trained on noisy data may not be reflected if the test data is similarly noisy; only on accurate test data the impact of label noise was fully reflected.

Our findings need to be discussed in detail. If the model was both trained and tested on consistently noisy data its performance remained moderately high for most datasets, indicating that inaccurate labeling and its impact on performance may be disguised if testing is performed on the resulting noisy data too. This was particularly true for larger BBs: When tested on noisy data the effect of enlarged BBs during training was absent for considerable time, even for massively increased sizes ($\alpha \leq 60$).

Evaluating the models with XAI methods as shown in Figure 4 gives us insights into possible causes: calculus remained most relevant for models prediction even when trained using large BBs; the super-pixels used to decide where calculus is present were near identical regardless of the BB size. Additional pixels included in the enlarged BB (α = 100) played a subordinate role for the prediction, likely as the standardization of the image (bitewings are similar to each other, which is different in comparison with natural scene images) allowed the model to learn the background and consequently ignore it successfully even if BBs were too large. The illustration of the predictions further show that the object (calculus) remained center of the predicted BBs. This reinforces the assumption that the model has learned the annotation error (consistently too large BBs). This phenomenon might be further explained by the learning process of neural networks: the model is optimized to predict BBs as close to the ground truth as possible. To increase the performance the objective thus becomes: (1) identification of the object of interest (calculus); (2) fit BB with dimensions to maximize area overlap with ground truth (e.g., consistently too large).

In contrast, if the BB were too small, we observed poor performance regardless of the chosen test (noisy or accurate). It can be assumed that given that calculus is already a very small object, models trained on too small BBs simply did not receive enough information to allow learning the features of calculus. Furthermore, with small objects, a small error already leads to the required *IoU* threshold not being reached. When tested on accurate data, the detrimental effect of noise on training success was demonstrated; researchers should pay special attention to consistent and accurate labeling of their test set to allow providing reliable information about the true accuracy of their model. This aspect of the results should be emphasized, since other studies dealing with noisy data tested their models solely on accurate data [4,9,12]. Testing on noisy data highlights the risk of performance obfuscation. It is especially relevant for medical applications where results are difficult to interpret and to compare due to a lack of standardized testing approaches and benchmarking datasets. In the medical domain deep learning models trained on noisy data are likely to be tested on a subset of the same distribution of data and therefore carry the same amount of noise.

In contrast, testing on noisy data may even lead to false conclusions, for example if a model was trained on accurate data and the (correctly) predicted BBs would not fit to the provided test data, leading to low performance metrics. A similar effect was demonstrated when dealing with inconsistent noise: Testing on accurate data showed no significant performance deterioration while $\delta \leq 0.3$, indicating that the model was capable of learning from inconsistent noisy labels. However, testing on noisy labels disguises this capability and already suggests a decrease in model performance when $\delta \geq 0.2$.

As discussed, several methods have been proposed to handle label noise through technical methods: a different loss function, measuring the difference between the model predictions and the desired output showed promising results when the underlying data set for a neural network contained noisy labels [4,20,21]. However, technical developments that aim to reduce the influence of label noise are often tested on clean benchmarking data sets. These data sets are currently not available in dentistry due to sensitive nature of the data and the associated data protection making the generation of a particularly clean test data set indispensable as demonstrated by the discussed results.

This study has a number of limitations. First, our goal was not to develop the best model for our specific problem; we did not aim to optimize performance by employing, for example, hyperparameter tuning but to understand the impact of noise in principle. Secondly, our trained models were not tested for generalizability while being developed using data from one German subpopulation only. Again, we accept this caveat in the context of our studys aims. Third, we assessed the impact of noise for one particular task, detecting calculus on bitewings. The effects on other modeling tasks like segmentation (where noisy labels are also likely) or other clinical problems (e.g., caries, apical lesion, periodontal bone loss detection) or images (other radiographs, photographs, histological data) may differ; we hence cannot claim transferability of our findings. Similarly only one

model—YOLOv5—was employed; the effect of noise on other models may differ to some degree, as shown on histopathological images for cell segmentation [12].

5. Conclusions

Accurate predictions need accurate labeled data in the training process. Testing on noisy data may disguise the effects of noisy training data. Modelers should be aware of the relevance of accurately annotated data, especially when testing model performance, and users should scrutinize models accordingly for labeling quality.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jcm12093058/s1, Table S1: Performance of calculus detection trained on consistently too large ($\alpha > 1$) or too small ($\alpha < 1$) bounding boxes when tested on accurate data; Table S2: Performance of calculus detection trained on consistently too large ($\alpha > 1$) or too small ($\alpha < 1$) bounding boxes when tested on consistent noisy data; Table S3: Model performance of models trained on inconsistently noisy data tested on inconsistent noisy data; Table S4: Model performance of models trained on inconsistently noisy data tested on accurate data.

Author Contributions: Conceptualization, M.B., L.S., A.K., J.K. and F.S.; methodology, M.B.; formal analysis, M.B.; data preparation and curation, J.K., M.B. and B.F.; writing—original draft preparation, M.B.; writing—review and editing, F.S., L.S. and A.K.; visualization, M.B.; supervision, F.S. and J.K.; All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge financial support from the Open Access Publication Fund of Charité—Universitätsmedizin Berlin and the German Research Foundation (DFG).

Institutional Review Board Statement: All experiments were carried out in accordance with relevant guidelines and regulations. Data collection was ethically approved by the ethics committee of the Charité (EA4/080/18).

Informed Consent Statement: Patient consent was waived not needed as data was only used in an anonymized way.

Data Availability Statement: The weights of the trained models can be provided on request. Medical image data cannot be made available given data privacy reasons.

Conflicts of Interest: F.S. and J.K. are co-founders of a Charité startup on dental image analysis. The conduct, analysis and interpretation of this study and its findings was unrelated to this.

References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
2. Arsiwala-Scheppach, L.T.; Chaurasia, A.; Müller, A.; Krois, J.; Schwendicke, F. Machine Learning in Dentistry: A Scoping Review. *J. Clin. Med.* **2023**, *12*, 937. [CrossRef] [PubMed]
3. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312.
4. Karimi, D.; Dou, H.; Warfield, S.K.; Gholipour, A. Deep Learning with Noisy Labels: Exploring Techniques and Remedies in Medical Image Analysis. *Med. Image Anal.* **2020**, *65*, 101759. [CrossRef] [PubMed]
5. Hu, Z.; Gao, K.; Zhang, X.; Wang, J.; Wang, H.; Han, J. Probability Differential-Based Class Label Noise Purification for Object Detection in Aerial Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6509705. [CrossRef]
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016; pp. 21–37.
7. Chadwick, S.; Newman, P. Training Object Detectors with Noisy Data. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 1319–1325.
8. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
9. Koksal, A.; Ince, K.G.; Alatan, A.A. Effect of Annotation Errors on Drone Detection with YOLOv3. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020.
10. Jocher, G.; Stoken, A.; Chaurasia, A.; Borovec, J.; NanoCode012; Xie, T.; Kwon, Y.; Michael, K.; Liu, C.; Fang, J.; et al. Ultralytics/Yolov5: V6.0—YOLOv5n "Nano" Models, Roboflow Integration, TensorFlow Export, OpenCV DNN Support. 2021. Available online: https://github.com/ultralytics/yolov5 (accessed on 4 May 2022).
11. Kim, J.-H.; Kim, N.; Park, Y.W.; Won, C.S. Object Detection and Classification Based on YOLO-V5 with Improved Maritime Dataset. *J. Mar. Sci. Eng.* **2022**, *10*, 377. [CrossRef]

12. Vădineanu, Ş.; Pelt, D.M.; Dzyubachyk, O.; Batenburg, K.J. An Analysis of the Impact of Annotation Errors on the Accuracy of Deep Learning for Cell Segmentation. In Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, PMLR, Zurich, Switzerland, 6–8 July 2022; pp. 1251–1267.
13. Schwendicke, F.; Singh, T.; Lee, J.-H.; Gaudin, R.; Chaurasia, A.; Wiegand, T.; Uribe, S.; Krois, J.; IADR e-Oral Health Network and the ITU WHO Focus Group AI for Health. Artificial Intelligence in Dental Research: Checklist for Authors, Reviewers, Readers. *J. Dent.* **2021**, *107*, 103610. [CrossRef] [PubMed]
14. McKinney, W. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010; Volume 445, pp. 51–56.
15. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
16. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
17. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.
18. 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]
19. Mohammad-Rahimi, H.; Motamedian, S.R.; Rohban, M.H.; Krois, J.; Uribe, S.E.; Mahmoudinia, E.; Rokhshad, R.; Nadimi, M.; Schwendicke, F. Deep Learning for Caries Detection: A Systematic Review. *J. Dent.* **2022**, *122*, 104115. [CrossRef] [PubMed]
20. Ma, X.; Huang, H.; Wang, Y.; Romano, S.; Erfani, S.; Bailey, J. Normalized Loss Functions for Deep Learning with Noisy Labels. In Proceedings of the 37th International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 6543–6553.
21. Zhou, X.; Liu, X.; Jiang, J.; Gao, X.; Ji, X. Asymmetric Loss Functions for Learning with Noisy Labels. In Proceedings of the 38th International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12846–12856.

Lebenslauf

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

Komplette Publikationsliste

Schneider, L., Rischke, R., Krois, J., Krasowski, A., **Büttner, M**., Mohammad-Rahimi, H., Chaurasia, A., Pereira, N. S., Lee, J. H., Uribe, S. E., Shahab, S., Koca-Ünsal, R. B., Ünsal, G., Martinez-Beneyto, Y., Brinz, J., Tryfonos, O., & Schwendicke, F. (2023). Federated vs Local vs Central Deep Learning of Tooth Segmentation on Panoramic Radiographs. *Journal of dentistry*, *135*, 104556.

Impact-Faktor: 4,379

Büttner, M., & Schwendicke, F. (2023). Natural language processing in dentistry. *British dental journal*, *234*(10), 753.
Impact Faktor:2,6

Schwendicke, F., & **Büttner, M.** (2023). Artificial intelligence: advances and pitfalls. *British dental journal*, *234*(10), 749–750.

Impact Faktor:2,6

Büttner, M., Schneider, L., Krasowski, A., Krois, J., Feldberg, B., & Schwendicke, F. (2023). Impact of Noisy Labels on Dental Deep Learning-Calculus Detection on Bitewing Radiographs. *Journal of clinical medicine*, *12*(9), 3058.

Impact Faktor: 4,96

Pfänder, L., Schneider, L., **Büttner, M.**, Krois, J., Meyer-Lueckel, H., & Schwendicke, F. (2023). Multi-modal deep learning for automated assembly of periapical radiographs. *Journal of dentistry*, *135*, 104588.
Impact-Faktor: 4,379

Danek, S., **Büttner, M.**, Krois, J., & Schwendicke, F. (2023). How Do Users Respond to Mass Vaccination Centers? A Cross-Sectional Study Using Natural Language Processing on Online Reviews to Explore User Experience and Satisfaction with COVID-19 Vaccination Centers. *Vaccines*, *11*(1), 144.

Impact Faktor: 7,8

Ma, J., Schneider, L., Lapuschkin, S., Achtibat, R., **Duchrau, M**., Krois, J., Schwendicke, F., & Samek, W. (2022). Towards Trustworthy AI in Dentistry. *Journal of dental research*, *101*(11), 1263–1268.

Impact Faktor: 8,924

Bergner, B., Rohrer, C., Taleb, A., **Duchrau, M.,** De Leon, G., Rodrigues, J., Schwendicke, F., Krois, J., Lippert, C., (2022) Proceedings of The 5th International Conference on Medical Imaging with Deep Learning, PMLR 172:130-149, 2022

Konferenzpublikation

Danksagung

Mein Dank gilt insbesondere meinen Erstbetreuer Prof. Falk Schwendicke für die umfangreiche Unterstützung von der Entwicklung der Idee bis hin zur fertigen Dissertation. Auch für die vielseitige fachliche Förderung und das Vertrauen in meine Arbeit möchte ich mich bedanken.

Meinem Zweitbetreuer Dr. Joachim Krois möchte ich für die Motivation zur Vertiefung meiner Programmierkenntnisse und der Entwicklung meiner Fragestellung danken.

Meinen Kolleg*innen des Data Science Teams danke ich von Herzen für ihre fachliche und freundschaftliche Unterstützung während und nach der Arbeitszeit. Ich danke insbesondere Aleksander Krasowski für die Unterstützung bei mühsamen Fehlerbehebungen zu jeder Uhrzeit und umfangreichen Tipps im Umgang mit dem Terminal, die meine Freude am Programmieren weiter steigerten. Lisa Schneider danke ich besonders für die vielen methodischen und beruflich wegweisenden Diskussionen und ihre aufmunternde Unterstützung in den letzten Zügen. Bei José Eduardo Cejudo Grano de Oro möchte ich mich für die Einführung in die Objektdetektion, bei Dr. Vinay Pitchika für die umfassende statistische Beratung und bei Dr. Noah Nordblom für die Motivation und hilfreichen Tipps in den Entzügen bedanken.

Zuletzt möchte ich auch meiner Familie für die Förderung und Begleitung auf meinem Weg und meinen Freunden und meinen Mann für ihre private Unterstützung danken.