Aus der Abteilung für Zahnerhaltung und Präventivzahnmedizin der
Charité – Centrum 03 für Zahn-, Mund- und Kieferheilkunde der
Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

Machine Learning Techniques for Computer Aided Classification of Dental Radiographic Images

Machine Learning Techniken zur Computer-gestützten Klassifizierung von zahnmedizinischen Röntgenbildern

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Thomas Ekert

aus Neckarsulm, Deutschland

Datum der Promotion: 06. März 2020

1 Zusammenfassung

Ziele: Entwicklung eines Convolutional Neural Networks (CNN) zur Anwendung von Machine Learning zur Computer-gestützten Klassifizierung apikaler Läsionen (AL) auf zahnärztlichen Panoramaröntgenscans.

Methoden: Ein annotierter Datensatz von 2001 Zahnsegmenten aus 85 zahnärztlichen Panoramaschichtaufnahmen wurden für das Training des CNN eingesetzt. Die Bilder waren von sechs Experten auf einer ordinal skalierten Skala (0: kein AL; 1: erweitertes Parodontalligament/unsichere AL; 2: eindeutig nachweisbare Läsion/sichere AL) bewertet worden. Zur Festlegung, ob AL vorhanden war, mussten vier Experten sich einig sein. Für die Bildklassifizierung (apikale Läsion vorhanden ja/nein) mittels Machine Learning wurde eine eigene CNN-Architektur entwickelt. Bei einer Hyperparametersuche in Form einer grid search wurden zudem batch size, learning rate, batch normalization, Augmentierung und dropout variiert. Die Klassifikationsleistungsfähigkeit des CNNs wurde in Bezug auf die Metriken AUC („area-under-the-receiver-operating-characteristics-curve“), Sensitivität, Spezifität und negativem und positivem Vorhersagewert (NPV/PPV) hin betrachtet und in einem automatisierten Verfahren dokumentiert. Die Validierung erfolgte über eine zehnfache Kreuzvalidierung. Mit Hilfe eines group shufflings wurde sichergestellt, dass Zahnsegmente eines Gebisses jeweils entweder ausschließlich im Trainings- oder im Validierungsset lagen, um Wechselwirkungen auszuschließen. In Subgruppenanalysen wurde die Applikation des CNNs auf verschiedene Zahntypen untersucht. Ebenso wurden verschiedenen Übereinstimmungsgrade der Mehrheitsentscheidung der Experten variiert und zwei Szenarien („sowohl unsichere als auch sichere AL“ versus „nur sichere AL“) miteinander verglichen

Ergebnisse: Ein siebenschichtiges feed forward CNN mit 4.299.651 trainierbaren Gewichten wurde entwickelt. Der mittlere (Standardabweichung) AUC des CNN für gleichzeitig sichere und unsicher AL lag bei 0,85 (0,04), die Sensitivität und Spezifität bei 0,65 (0,12) bzw. 0,87 (0,04). Der resultierende PPV betrug 0,49 (0,10), der NPV 0,93 (0,03) bei einer Prävalenz von 0,16 (0,03). Bei Molaren war die Sensitivität signifikant höher als bei anderen Zahntypen, während die Spezifität geringer war. Für

ausschließlich „sichere AL“ lag der AUC bei 0,89 (0,04). Wurde der Grad der Übereinstimmung der Experten auf sechs erhöht, stieg der AUC signifikant auf 0,95 (0,02) und die Sensitivität auf 0,74 (0,19) stieg.

Schlussfolgerung: Mit Hilfe eines CNNs, das mit grid search, Augmentierung und dynamischer Architekturvariation optimiert wurde, konnte auf Basis eines relativ kleinen Bilddatensatzes eine Computer-gestützte Klassifizierung von AL mit zufriedenstellender Klassifizierungs-Genauigkeit entwickelt werden.

2 Abstract

Objectives: We developed a convolutional neural network (CNN) to apply machine learning for computer-aided classification of apical lesions (AL) on dental radiographs.

Methods: An annotated dataset of 2001 tooth segments from 85 dental panoramic images was used for CNN training. The images were evaluated by six experts on an ordinally scaled scale (0: no AL; 1: extended periodontal ligament/insecure AL; 2: clearly detectable lesion/secure AL). To determine whether AL was present four experts had to agree. A CNN architecture was developed for image classification (apical lesion present yes/no) using machine learning. To optimize the CNN batch size, learning rate, batch normalization, augmentation, and dropout were tuned in a hyperparameter grid search. The classification performance of the CNN was assessed with respect to the metrics AUC ("area-under-the-receiver-operating-characteristics-curve"), sensitivity, specificity and negative and positive predictive value (NPV/PPV). The validation results for each training and validation run were documented in an automated process. A ten-cross-fold-validation was performed. The application of group shuffling ensured that tooth segments of each dentition were either exclusively in the training set or in the validation set to avoid unwanted correlations between training and validation. The CNN was applied to different tooth types for subgroup analyses. Likewise, different levels of agreement were varied when computing the majority vote of the experts. Two scenarios ("both unsafe and safe AL" versus "only safe AL") were compared.

Results: A seven-layer feed-forward CNN with 4,299,651 trainable weights was developed. The mean (standard deviation) AUC of CNN for both certain and uncertain AL was 0.85 (0.04), sensitivity and specificity resulted in 0.65 (0.12) and 0.87 (0.04), respectively. The prevalence in the base-case was 0.16 (0.03) and the PPV was 0.49 (0.10) while the NPV was 0.93 (0.03). Compared to other tooth types sensitivity for molars was significantly higher while specificity was lower. For clearly detectable AL only, the AUC was 0.89 (0.04). When the level of agreement between the votes of the experts was increased to six, the AUC increased significantly to 0.95 (0.02), and the sensitivity increased to 0.74 (0.19).

Conclusion: A CNN developed based on a relatively small image data set and optimized through grid search, augmentation and dynamic architecture variation can be used for computer-aided classification of AL with satisfactory discrimination ability.

3 Einführung

Bildgebende Verfahren sind in der Medizin von besonderer Bedeutung und werden bei der Beantwortung vieler diagnostischer Fragestellungen herangezogen. Hierbei spielt insbesondere die korrekte Interpretation der durch die bildgebenden Verfahren erstellten Bilder und der darin vorgefundenen Strukturen während der Befundung durch das medizinische Fachpersonal eine herausragende Rolle. Ziel ist in der Regel eine (möglichst korrekte) Klassifizierung mit dem Ziel zu erkennen, ob eine bestimmte Pathologie vorliegt oder nicht. Diese Klassifizierung dient dann als Grundlage für die erfolgreiche Beantwortung der anfänglichen medizinischen Fragestellung und der hieraus abgeleiteten Therapie (Dössel 2016).

3.1 Klassifizierungsleistung bei klassischer Begutachtung

Radiographische Untersuchungen gehören zu den wichtigsten bildgebenden Verfahren in der Zahnmedizin. In dieser Studie wurde die Erkennung apikaler Parodontitis anhand radiographischer Bilder durch Zahnärzte und ein möglicher Einsatz Computer-gestützter Verfahren zur Klassifizierung zur Detektion von apikaler Parodontitis untersucht. Apikale Parodontitis ist ein entzündlicher Prozess um die Spitze der Zahnwurzel herum, der durch eine bakterielle Infektion der Zahnmarkhöhle (Pulpahöhle) hervorgerufen wird und mit einer Auflösung des periapikalen Knochengewebes einhergeht (Segura-Egea et al. 2015; Huumonen et al. 2017; Connert et al. 2018). Bei radiographischen Verfahren wird sie durch periapikale Aufhellung (erweitertes Parodontalligament oder deutlich erkennbare apikale Läsionen (AL)) sichtbar.

Für die Diagnose von AL sind periapikale Röntgenaufnahmen zwar üblich, aber verglichen mit einem Goldstandard (z.B. in Schädelstudien) nur begrenzt genau (Kanagasingam et al. 2017). Umgekehrt sind AL in digitalen Volumentomogrammen (DVT, auch Kegelstrahl-CT; CBCT) gut nachweisbar, jedoch führt dieses Verfahren (noch) zu hohen Kosten und einer erhöhten Strahlendosis, weshalb es bisher seltener eingesetzt wird (Leonardi et al. 2016).

Im Vergleich mit beiden Verfahren erfolgt der Nachweis von AL auf Panoramaröntgenaufnahmen, die zu den häufigsten zahnmedizinischen Röntgenverfahren zählen, zwar mit begrenzter Sensitivität und negativem Vorhersagewert (NPV), jedoch hoher Spezifität und positivem Vorhersagewert (PPV). Gleichzeitig ist die Strahlendosis gering und es können, anders als auf Einzelbildern, alle Zähne gleichzeitig befundet werden. Insgesamt ergibt sich eine gute diagnostische Genauigkeit (Nardi et al. 2018; Nardi et al. 2017; Ahlqwist et al. 1986), jedoch hängt die Zuverlässigkeit AL zu diagnostizieren von der Erfahrung des Untersuchers ab (Parker et al. 2017).

Generell ist eine begrenzte Zuverlässigkeit im Hinblick auf die Übereinstimmung der Ergebnisse mehrere Begutachter bei der Beurteilung radiographischer Bilder zu beobachten (Nardi et al. 2018; Nardi et al. 2017). Als Maßstab für den Grad der Übereinstimmung eignet sich das Fleiss‘ Kappa (siehe Kapitel 0). Dieser Wert ist ein statistisches Maß für die Übereinstimmung mehrerer Bewerter untereinander und liegt bei 1 für vollständige Übereinstimmung und hat einen Wert kleiner oder gleich Null für keine Übereinstimmung. Das Fleiss‘ Kappa liegt beispielsweise bei der Erkennung apikaler Läsionen (AL) auf DVT zwischen κ=0,28 für Studenten und κ=0,49 für Endodontie-Spezialisten (Parker et al. 2017), kann also lediglich als nur angemessen oder bestenfalls als moderat bewertet werden; verschiedene und vor allem verschieden erfahrene Bewerter werden stark unterschiedliche Detektionsergebnisse erzielen (siehe Kapitel 4.3).

3.2 Machine Learning und automatische Klassifizierung

Maschinelles Lernen („machine learning“) ist ein Oberbegriff für Computer-gestützte Verfahren, anhand derer Computersysteme durch den Einsatz statistischer Modelle oder mathematischer Verfahren befähigt werden, bestimmte Aufgaben ohne ausdrückliche Anweisungen auszuführen. Ein Teilbereich ist das so genannte „supervised learning“ bei dem Computer anhand von Paaren aus vorgegebenen Ein- und Ausgangsdaten („Trainings-Daten“) Strukturen lernt und in die Lage versetzt wird, den Zusammenhang zwischen Ein- und Ausgangsdaten zu verallgemeinern. Hierfür werden Künstliche Neuronale Netze eingesetzt. Für supervised learning von Bilddaten

kommen spezielle Künstliche Neuronale Netze, die Convolutional Neural Networks (CNN – „gefaltetes neuronales Netz“) zum Einsatz. Diese sind mathematische Verfahren (vgl. Kapitel 0), anhand deren in mehreren Stufen (die so genannten Ebenen) Strukturen in Bildern gefiltert und anschließend nach im Vorfeld festgelegten Kriterien klassifiziert werden können (Goodfellow et al. 2016).

Eingangsdaten können z.B. Bilder sein, für die zuvor eine bestimmte Diagnose ermittelt wurde („Annotation“). Nach dem Training kann ein solches System im Idealfall die im Training gezeigten Strukturen verallgemeinern und nun auch für unbekannte Bilder eine gültige Klassifizierung vornehmen, d.h. eine Wahrscheinlichkeit für die Zugehörigkeit zu einer bestimmten Klasse (z.B. ob eine Pathologie vorliegt oder nicht) berechnen.

3.3 Entwicklung eines Machine Learning Ansatzes zur automatischen Erkennung apikaler Läsionen

Maschinelles Lernen wurde bereits in der Vergangenheit erfolgreich für die automatische Klassifizierung medizinischer Bilder eingesetzt (Mazurowski et al. 2018; Litjens et al. 2017), unter anderem zur Beurteilung von Brustkrebs in Mammographien (Becker et al. 2017), von Hautkrebs in klinischen Hautuntersuchungen (Esteva et al. 2017) oder von diabetischer Retinopathie in Augenuntersuchungen (Gulshan et al. 2016). In der Zahnmedizin wurden CNNs bereits eingesetzt, um kariöse Läsionen auf Bissflügelröntgenaufnahmen (Lee et al. 2018a) oder parodontalen Knochenverlust auf periapikalen Röntgenaufnahmen (Lee et al. 2018b) zu erkennen.

Ziel dieser Dissertation war es, ein CNN zu entwickeln und für die Beurteilung von AL auf Panoramaröntgenaufnahmen anzuwenden. Um das CNN trainieren zu können wurde zunächst ein Bilddatensatz anhand von zufällig ausgewählten anonymisierten Zahnsegmenten erstellt und durch erfahrene Zahnärzte auf das Vorliegen von AL hin beurteilt (siehe Kapitel 4.1). Diese Daten dienten als Ein- und Ausgangsdaten im oben beschriebenen „supervised learning“-Ansatz. Die Aufteilung in einen Trainings- und Validierungsdatensatz (80/20% der Daten) ermöglichte, zunächst das CNN zu trainieren und dann an dem CNN noch unbekannten Daten zu validieren. Idealerweise sollte ein CNN gefunden werden, das geeignet ist, AL zuverlässig zu diagnostizieren,

um so als automatisiertes Assistenzsystem Zahnärzte in der Diagnostik zu unterstützen und die diagnostische Genauigkeit zu erhöhen.

4 Materialien und Methoden

4.1 Studiendesign

Für die Entwicklung eines CNN zur Klassifizierung von Bildern sind mehrere Schritte notwendig. Für diese Studie wurde das oben beschriebene Verfahren des supervised learning gewählt. Die Dokumentation der Studie folgt der STARD-Richtlinie (Bossuyt et al. 2015).

Im ersten Schritt wurden aus 85 zufällig ausgewählten anonymisierten digitalen Panoramaröntgenaufnahmen zunächst einzelne Zähne manuell ausgeschnitten. So wurden nur Röntgenaufnahmen einzelner Zähne einbezogen (insgesamt 2001 Zahnsegmente). Ein Experte hatte diese zuvor auf ihre Qualität (Beurteilbarkeit) überprüft, wobei 249 Zahnsegmente (hauptsächlich Frontzähne, meist aufgrund der Überlagerung der Aufnahme mit der Wirbelsäule) ausgeschlossen wurden. Es wurden keine weiteren Ein- oder Ausschlusskriterien (wie z.B. Ausschluss wg. Kontrast, Trübung, Positionierung etc.) angewandt.

Die Segmente wurden von sechs unabhängigen Experten auf das Vorliegen von apikalen Läsionen anhand einer ordinal skalierten Annotationsskala (0: kein AL; 1: erweitertes Parodontalligament/unsichere AL; 2: eindeutig nachweisbare Läsion/sichere AL) annotiert. Anhand einer Mehrheitsentscheidung (Referenztest siehe Kapitel 4.3) wurde ein binärer Wert ermittelt (1 | 0 – AL liegt vor | AL liegt nicht vor). Die Marge der Mehrheitsentscheidung wurde in Sensitivitätsanalyse variiert, wobei hierbei solche Zahnsegmente ausgeschlossen wurden, bei denen die geforderte Marge nicht vorlag. Für eine weitere Sensitivitätsanalyse wurde die Ermittlung des binären Wertes im Basis-Szenario zunächst anhand unsicherer und sicherer AL gebildet und diese dann mit dem alternativen Szenario (nur sichere AL) verglichen (siehe Sensitivitätsanalyse Kapitel 0).

Der so erstellte und annotierte Bilddatensatz wurde nun im zweiten Schritt für das Training eines CNNs herangezogen. Hierfür wurde der Datensatz in Trainings- und Validierungsdaten aufgeteilt. 20 % des Datensatzes wurden als Validierungsdatensatz für die Validierung zurückbehalten, an 80% der Daten wurden trainiert. Durch den

Vergleich mit dem Referenztest (Annotationen) konnten Performance-Metriken ermittelt werden (vgl. Kapitel 0). Anhand der Ergebnisse erfolgt eine Optimierung des CNN. Die einzelnen Schritte dieses Prozesses sind in Abbildung 1 dargestellt.

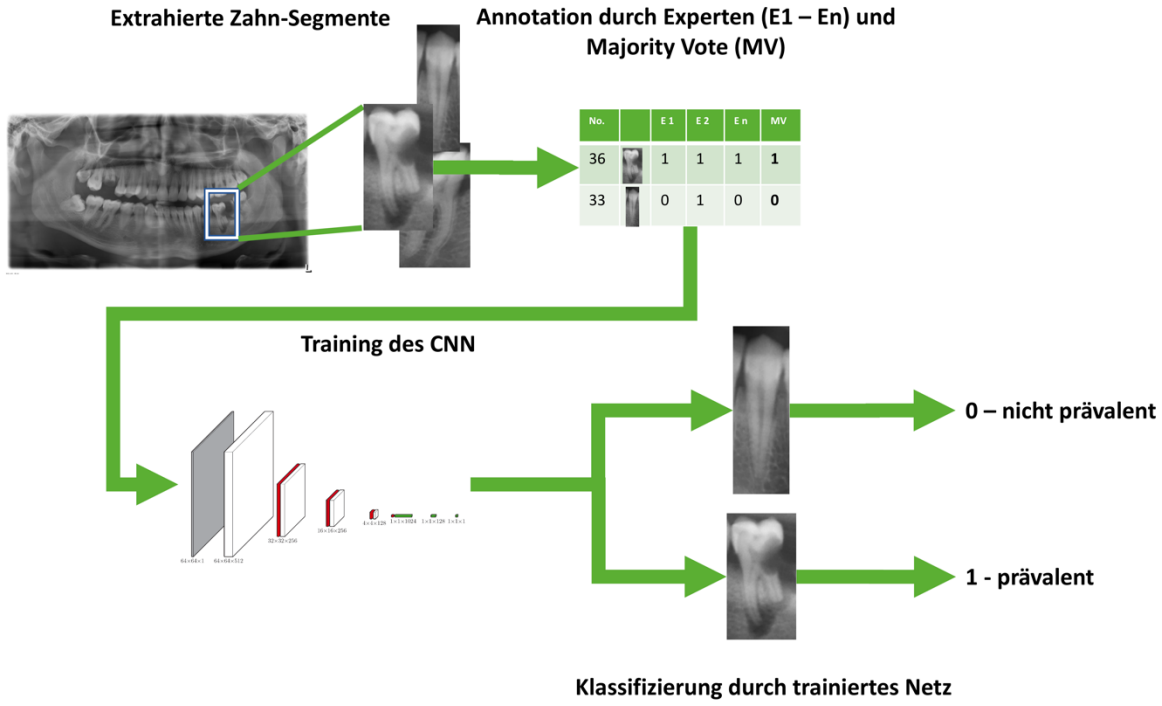


Abbildung 1 *– Trainings- und Klassifizierungspipeline*

Das Training eines CNNs ist ein Optimierungsproblem, bei dem einerseits die Gewichte der Ebenen des CNN und andererseits der Hyperparameter (z.B. die Netzwerkarchitektur, d.h. die Anordnung, Art und Parametrisierung der einzelnen Schichten) optimiert werden. Beim Training variiert der so genannte Optimizer (siehe Kapitel 0) die Gewichte des CNN so lange, bis ein idealerweise globales Maximum in Bezug auf eine zuvor festgelegte Performance-Metrik erreicht wird. Dies lässt sich anhand einer dreidimensionalen Funktion visualisieren. Zwei der drei Dimensionen stellen in der vereinfachenden und beispielhaften Darstellung in Abbildung 2 zwei der Gewichte dar, während die dritte Dimension die Performance-Metrik darstellt. Für jeden gewählten Hyperparameter ergibt sich jeweils ein individuelle n-dimensionales Optimierungsproblem. Bei der Optimierung des CNNs werden die so erhaltenen Optima miteinander verglichen und so die beste Kombination der Hyperparameter gefunden.

Für die Studie wurde eine Computersoftware entwickelt, um auch bei der Durchführung mehrerer tausend Trainingsläufe eine standardisierte Durchführung, Dokumentation und Auswertung zu ermöglichen, mit dem Ziel, eine Vielzahl von Hyperparametern durchsuchen zu können und so jene Architektur eines CNN zu finden, das mit der gegebenen Zahl und Qualität von Ausgangsdaten eine optimale automatisierte Klassifizierung von AL ermöglicht.

Zunächst wurde mit einer einfachen Architektur mit wenigen Schichten gestartet. In der Hyperparametersuche wurde diese Architektur erweitert und die Parameter in jeder einzelnen Schicht variiert. Jeder Trainingslauf wurde standardisiert dokumentiert – sowohl in Bezug auf die verwendete Architektur und Parameter als auch in Bezug auf die so erzielten Ergebnisse, d.h. die jeweils erreichte Performance-Metrik der Klassifizierungsleistung des CNN gegenüber dem Validierungsdatensatz wurde erfasst. Ebenso wurden diejenigen Gewichte des CNN archiviert, bei der die Performance-Metrik optimal war, und in der Dokumentation verlinkt, um später die Ergebnisse reproduzieren zu können. Die Dokumentation erfolgte in automatisch generierten Tabellen einer Tabellenverarbeitung, in denen sowohl die verwendete Parametrisierung als auch die Ergebnisse jeder einzelnen Epoche des Trainings enthalten waren. Jedes erstellte Modell wurde gegen den gesamten

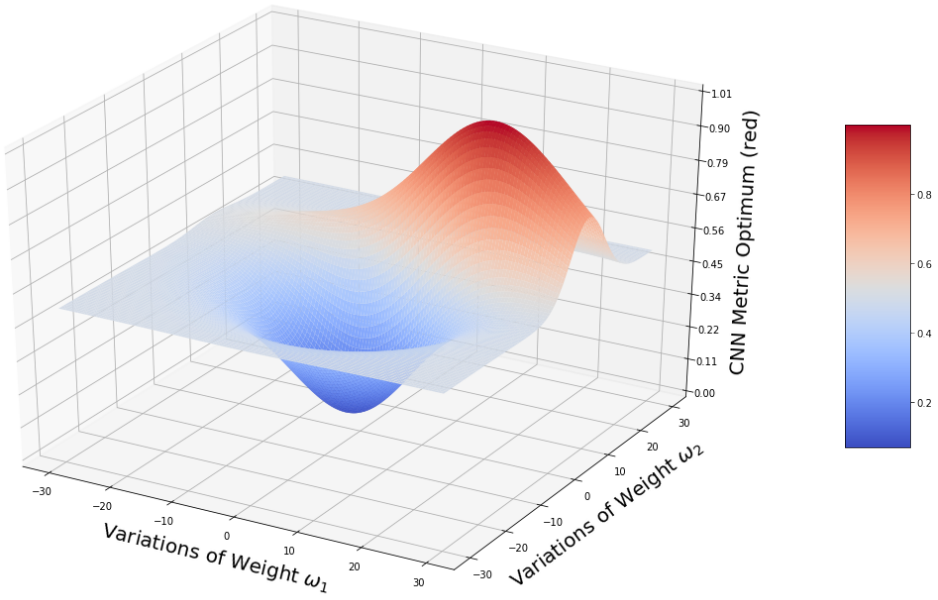


***Abbildung 2** – Vereinfachte Darstellung einer CNN Parameter-Suche*

Validierungsdatensatz und zusätzlich in der Sensitivitätsanalyse gegenüber einzelnen Zahngruppen validiert und die Performance-Metriken berechnet. Die Ergebnisse wurden im TEX Dateiformat gespeichert und in PDF Dateien umgewandelt. Diese zeigen sämtliche ermittelten Performance-Metriken in Tabellen und ROC-Kurven („receiver operating characteristics“) als Grafiken. Die Fläche unter der ROC-Kurve (area-under-the-curve, AUC) Metrik dient dazu, die Qualität einer Klassifizierung zu bewerten. ROC-Kurven bilden die Sensitivität (auch „true positive rate“) gegen 1 - Spezifität (auch „false positive rate“) ab; ein optimaler Test liegt hierbei links oben (Sensitivität = 1; 1 - Spezifität = 0, d.h. Spezifität = 1). Je steiler die Kurve ist (je mehr sie sich in die linke obere Ecke „schmiegt“) desto besser ist die Klassifikation (und desto höher die AUC). Die Diagonale von links unten nach rechts oben entspricht einem zufälligen Ergebnis (AUC=0.5). Tests mit einer AUC unter 0.5 sind schlechter als zufälliges Raten.

Da für die Studie nur ein relativ begrenzter Datensatz (2001 Bildsegmente) vorlag (siehe unten), wurden die Ausgangsbilder vor dem Training augmentiert, d.h. durch Drehen, Scheren, Zoomen etc. verändert, um künstlich den Datensatz zu erweitern. Das CNN wurde durch eine zehnfache Kreuzvalidierung („ten-fold cross-validation“) überprüft, wobei ein so genanntes „group shuffling“ zum Einsatz kam, bei dem sichergestellt wird, dass die Bilder von Zähnen desselben Patienten jeweils ausschließlich entweder im Trainings- oder im Validierungsdatensatz liegen, um eine Übertragung von im Training gelernten Mustern in die Validierung auszuschließen und so die Generalisierungsleistung des CNN sicher überprüfen zu können. Die Leistungsfähigkeit des CNN wurde am Referenztest gemessen und durch eine Sensitivitätsanalyse überprüft.

4.2 Bilddatensatz

Für diese Studie wurden 85 anonymisierte digitale Panoramaröntgenaufnahmen zufällig ausgewählt, die mit Orthophos XG (Sirona, Bensheim, Deutschland) nach Herstellerangaben (unter Berücksichtigung von Geschlecht und Alter der Patienten usw.) aufgenommen wurden. Hieraus wurden 2001 manuell zugeschnittene Bildsegmente entnommen, die jeweils einen einzelnen Zahn abbilden. Hierbei wurde Zahnposition und Zugehörigkeit zur jeweiligen Panoramaröntgenaufnahme festgehalten. Die Panoramaröntgenaufnahmen wurden in der zentralen zahnmedizinischen Röntgenabteilung der Charité - Universitätsmedizin erstellt. Die Positionierung erfolgte nach Herstellerangaben durch erfahrene medizinisch-technische Assistenten (Fachrichtung Zahnröntgen). Das mittlere (Mindest-/Maximal) Alter der 85 Patienten betrug 51 (15/91) Jahre. Die mittlere (Median, Minimum/Maximum) Anzahl von Zähnen je Patienten (d.h. je Panoramaröntgenaufnahme) betrug 19,5 (20, 1/30). Es gab 30,2, 15,5, 27,8 und 26,5% Schneidezähne, Eckzähne, Prämolaren und Molaren. Es wurden zusätzlich Sensitivitätsanalysen zur Zahnposition und in Bezug auf die Eindeutigkeit der Annotation durchgeführt (siehe unten). Für die Datenerhebung lag eine Genehmigung der Charité-Ethikkommission vor (EA4/080/18).

4.3 Referenztest

Für den Referenztest wurden sechs Experten (Zahnärzte mit 3 - 10 Jahren klinischer Erfahrung) für die Annotation der Zahnsegmente herangezogen. Die Annotation erfolgte anhand Panoramaröntgenaufnahmen, die die Experten auf entsprechend für die Diagnose zugelassenen Bildschirmen beurteilten. Dies erfolgte unter standardisierten Bedingungen (gedimmte Beleuchtung, Einsatz von Kontrast- und Helligkeitskorrektur). Hierbei wurde durch die Zahnärzte systematisch bewertet und dokumentiert, ob auf einer ordinalen Skala AL vorlag (0: kein AL; 1: erweitertes Parodontalligament/unsichere AL; 2: eindeutig nachweisbare Läsion/sichere AL). Aus

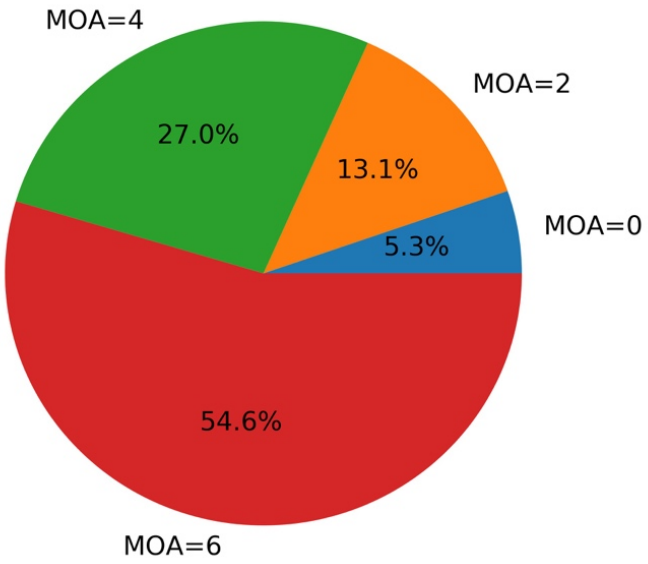


Abbildung 3 – Marge der Übereinstimmung in der Mehrheitsentscheidung aller sechs Experten.

Die Grafik zeigt den Übereinstimmungsgrad (MOA = „margin of agreement“) der Annotationen. Es wird die Differenz zwischen den unterschiedlichen Annotationen berechnet. Stimmen z.B. fünf Experten für „AL liegt vor“ und einer erkennt auf „AL liegt nicht vor“, dann beträgt der MOA vier (5 minus 1). Stimmen vier Experten überein (und die verbleibenden zwei teilen die gegenteilige Meinung) beträgt der MOA zwei. Ein MOA von 6 bedeutet, dass alle 6 Experten in ihrer Bewertung (auf das Vorliegen von AL hin) übereinstimmten. Bei insgesamt 94,7 Prozent der Bewertungen stimmten mindestens vier oder mehr Experten in Ihrer Bewertung überein.

diesen Ergebnissen wurde eine Mehrheitsentscheidung gebildet. Für die Ergebnisse wurde das Fleiss‘ Kappa (Fleiss et al. 1971) berechnet, um die Übereinstimmung der Zahnärzte untereinander zu bewerten, wobei ein Wert von 0-0,20 als geringe, 0,21-0,40 als angemessen, 0,41-0,60 als moderate, 0,61-0,80 als stichhaltige und 0,81-1 als nahezu perfekte Übereinstimmung bewertet wurde.

Für die Mehrheitsentscheidung wurde im Basis-Szenario einem Bildsegment ein positives bzw. negatives Label zugeordnet, wenn sich mindestens vier von sechs Untersuchern auf die Diagnose geeinigt haben (vgl. Abbildung 3 – Übereinstimmungsmarge von zwei), andernfalls wurde das Bildsegment aus dem Datensatz entfernt. In einer Sensitivitätsanalyse wurde diese Marge auf sechs erhöht, d.h. es wurde eine Übereinstimmung der Annotation aller Zahnärzte vorausgesetzt. Bei einer Übereinstimmungsmarge von zwei stimmten die Experten moderat in ihren Annotationen überein (Fleiss‘ Kappa von 0,48), bei einer Marge von vier war die Übereinstimmung stichhaltig (0,61). Das Histogramm in Abbildung 4 zeigt die Verteilung der Annotationen für die 6 Experten jeweils für kein AL („No AL“), unsichere AL („uncertain AL“) und eindeutig nachweisbare Läsion/sichere AL („certain AL“).

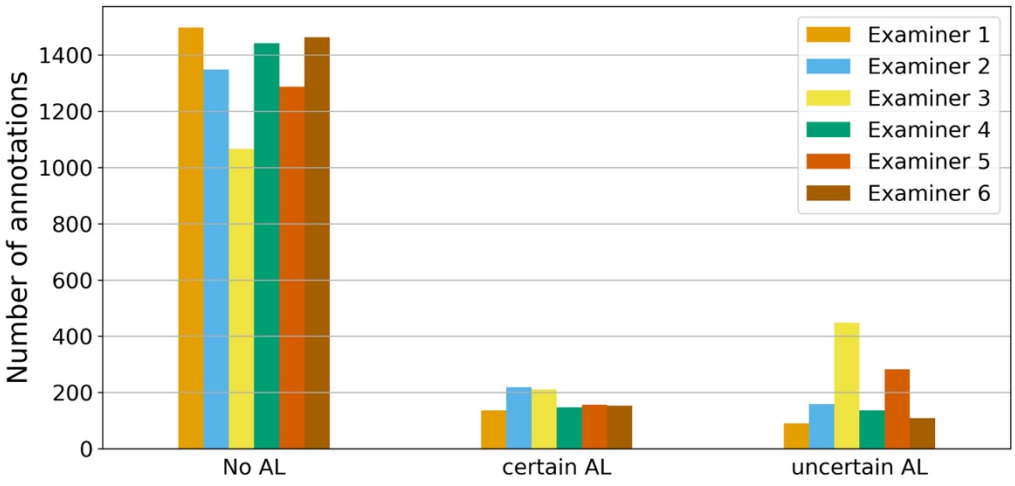


Abbildung 4 – Histogramm für die Annotation von AL *entsprechend der Bewertung durch die sechs unabhängigen Experten.*

4.4 Modellierung durch CNNs

Neuronale Netze bestehen aus Schichten mathematischer Funktionen; das dahinterliegende Konzept geht auf die Arbeiten von McCulloch, Pitts und Rosenblatt aus den Jahren 1943 bzw. 1958 zurück (McCulloch, Pitts 1943; Rosenblatt 1958). Neuronale Netze sind eine Technik des maschinellen Lernens und haben sich als eigenständiges Feld in den späten 1990er Jahren innerhalb des Gebiets der künstlichen Intelligenz (AI) etabliert. Durch die rasant wachsende Leistungsfähigkeit von Computer Hardware – unterstützt durch die Entwicklungen auf dem Gebiet spezialisierter Hardware wie GPU (Graphic Processing Unit) und TPU (Tensor Processing Unit), die in der Lage sind, die spezifischen Berechnungen neuronaler Netze besonders schnell und effizient auszuführen – aber auch durch die explosionsartig wachsende Verfügbarkeit von Daten, die zum Training neuronaler Netze verwendet werden können, haben sich neuronale Netze zu einer weit verbreiteten Technologie in vielen digitalen Services und Anwendungen entwickelt. Für die Studie wurde ein mehrschichtiges Feed-Forward-Netzwerk (s.u.) modelliert, um eine binäre Klassifizierung durchzuführen, d.h. um die Entscheidung „AL liegt vor“ oder „AL liegt nicht vor“ treffen zu können.

Um ein Klassifizierungsproblem durch maschinelles Lernen zu lösen, wird eine mathematische Funktion gesucht, die einem beliebigen Eingangs-Datensatz einer bestimmten Klasse (z.B. digitales Röntgenbild eines Zahns mit AL) einen

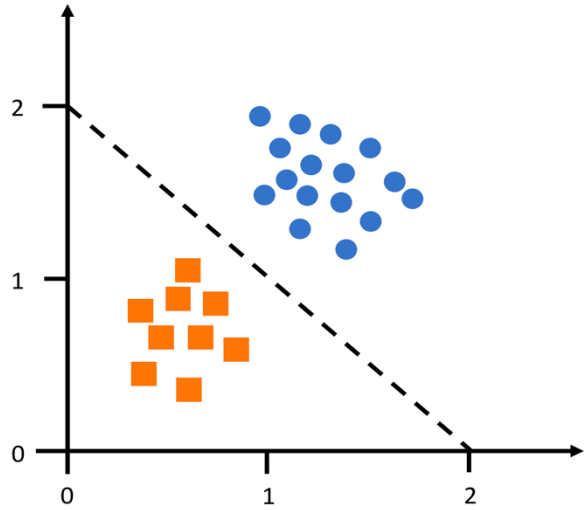
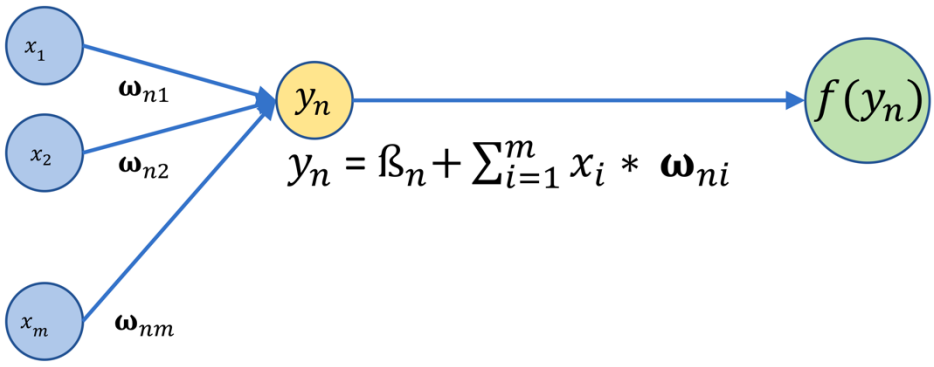


***Abbildung 5** – Binäre Klassifikation von Daten mit linearer Teilbarkeit*

Ausgangswert (im Beispiel also 1 für „AL liegt vor“) zuordnet. In der Regel wird für die Zugehörigkeit zu einer Klasse ein Wahrscheinlichkeitswert berechnet.

Abbildung 5 zeigt ein Beispiel, in dem der Eingangsdatensatz lediglich aus Punkten im zweidimensionalen Raum (markiert durch Kreise und Quadrate, die jeweils eine von zwei Klassen bezeichnen) besteht. Die Funktion der Klassifizierung ist in diesem Beispiel die einfache mathematische Funktion einer Geraden (Klasse der Kreise: y > 2 - x; Klasse der Quadrate: y < 2 - x), man spricht von linearer Teilbarkeit (Gupta 2017).

Neuronale Netze werden durch Schichten („Layer“) dargestellt, die jeweils durch eine mathematische Funktion beschrieben werden. Ein typisches Beispiel sind so genannte „dense layer“. Sie bestehen aus so genannten Neuronen (vgl. Abbildung 6), wobei die Neuronen im Wesentlichen durch mathematische Matrizen beschrieben werden, deren Werte beim Training durch Vektormultiplikation mit Gewichten variiert werden.



Detailansicht – Neuron n

y_n -> Matrixprodukt der Gewichte ω_{ni} des Neuron n und der Eingangswerte x_1 bis x_m
$f(y_n)$ -> Aktivierungsfunktion berechnet Ergebniswert des Neurons n = Prozentwert oder 1|0
$ß_n$ -> Bias-wert für dieses Neuron

***Abbildung 6** – Detailansicht eines Neurons*

Typischerweise besteht ein Neuronales Netz aus einer Eingangs-Schicht („Input Layer“), mehreren Zwischen-Schichten („Hidden Layer“) und einer Ausgangs-Schicht. Man spricht dann auch vom Deep Learning. Im Training eines Neuronalen Netzes werden die numerischen Werte der Neuronen und Gewichte durch ein weiteres mathematisches Verfahren (durch den so genannten „Optimizer“) so lange variiert, bis

sich aus den Eingangsdaten (z.B. den Pixeln eines digitalen Röntgenbildes) die erwarteten Ausgangsdaten ergeben. Im Falle einer Klassifizierung ergibt sich ein Wahrscheinlichkeitswert, der die Eingangsdaten der korrekten Klasse zuordnet. Ziel ist es, die Optimierung so durchzuführen, dass eine Generalisierung erreicht wird und eine Klassifikation auch für zuvor unbekannte Eingangsdaten gelingt.

In einem neuronalen Feed-Forward-Netzwerk werden Informationen durch das Netzwerk von der ersten (Eingangs-) Schicht zur letzten (Ausgangs-) Schicht (LeCun et al. 2015) geleitet (vgl. Abbildung 7) und in jeder Schicht durch die jeweilige mathematische Funktion verarbeitet. Das Ergebnis dieser numerischen Operationen wird einer Aktivierungsfunktion f zugeführt, die dieses entweder in einen binären (0 | 1) oder prozentualen (0...1) Wert umwandelt, um so eine Klassifizierung zu ermöglichen.

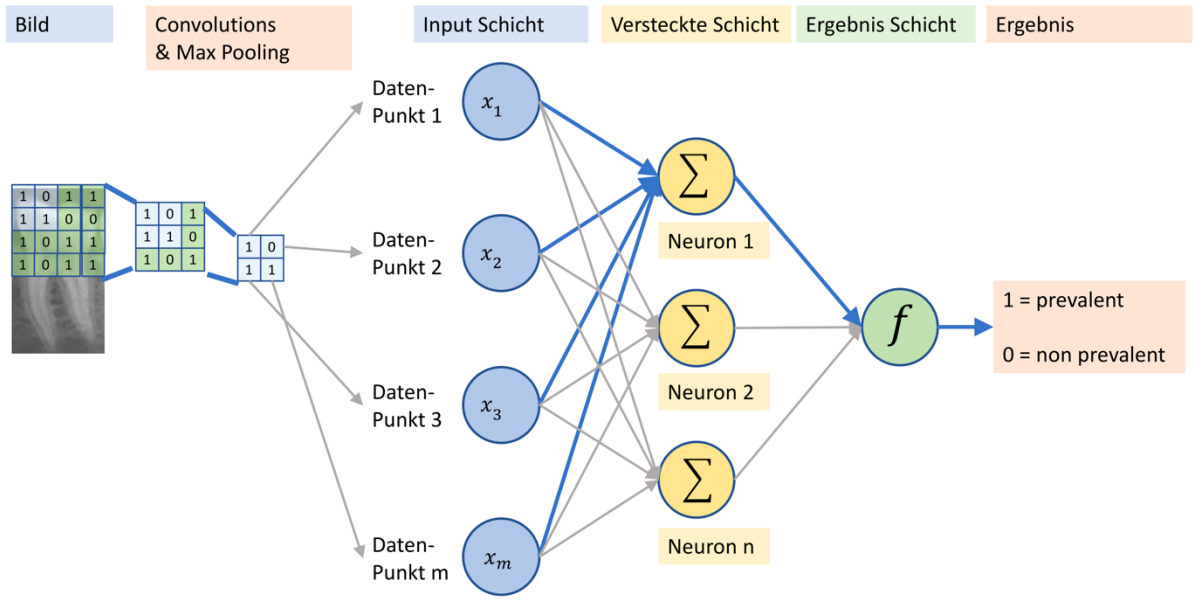


Abbildung 7 *– Convolutional Neural Network (CNN)*
CNN bestehend aus Faltungsschicht („Convolutional Layer“) und nachgeschaltetem Neuronalen Netz mit Input Schicht ("Input Layer"), versteckter Schicht ("Hidden Layer") und Ergebnis Schicht ("Output Layer")

CNNs sind eine spezielle Art neuronaler Netze. In einem CNN sind den beschriebenen, auch „dense layer“ genannten Schichten so genannte Convolutional

Layer („Faltungsschichten“) vorgeschaltet. Sie sind besonders nützlich, um hierarchische Merkmale aus Bilddaten durch mathematische Faltungsoperationen („convolution“) zu extrahieren. Hierdurch können diese Modelle verschiedene Merkmale eines Eingabebildes wie Kanten, Ecken und Punkte oder zunehmend komplexere Merkmale wie Formen und makroskopische Strukturen und Muster extrahieren. Ein einem Convolutional Layer wird anhand einer Matrix-Multiplikation mit einem so genannten Kernel (auch Filter genannt), der in Pixel-Schritten über das Original verschoben wird, eine Feature Map erstellt (Abbildung 8), die den Output dieses Layers darstellt.

Anhand von Abbildung 8 kann das Verfahren erklärt werden: Die Eingangsdaten bestehen aus der digitalen Repräsentation des Röntgenbildes, das aus einzelnen Bildpunkten („Pixel“) besteht. Jeder Bildpunkt hat einen Wert zwischen 0 (Schwarz) und 1 (Weiß). Zwischenwerte sind graue Pixel. In der Abbildung ist dies schematisch im linken blauen Quadrat („Digitale Darstellung des Bildes“) als 6x6 Pixel dargestellt. Der Einfachheit halber wurden keine Grauwerte, sondern nur schwarze (0) und weiße (1) Pixel angenommen. Der Kernel ist in Abbildung 8 als gelbe 3x3 Matrix („Kernel (Filter)“) dargestellt. Die Feature-Map wird durch Anwendung des Kernels wie folgt

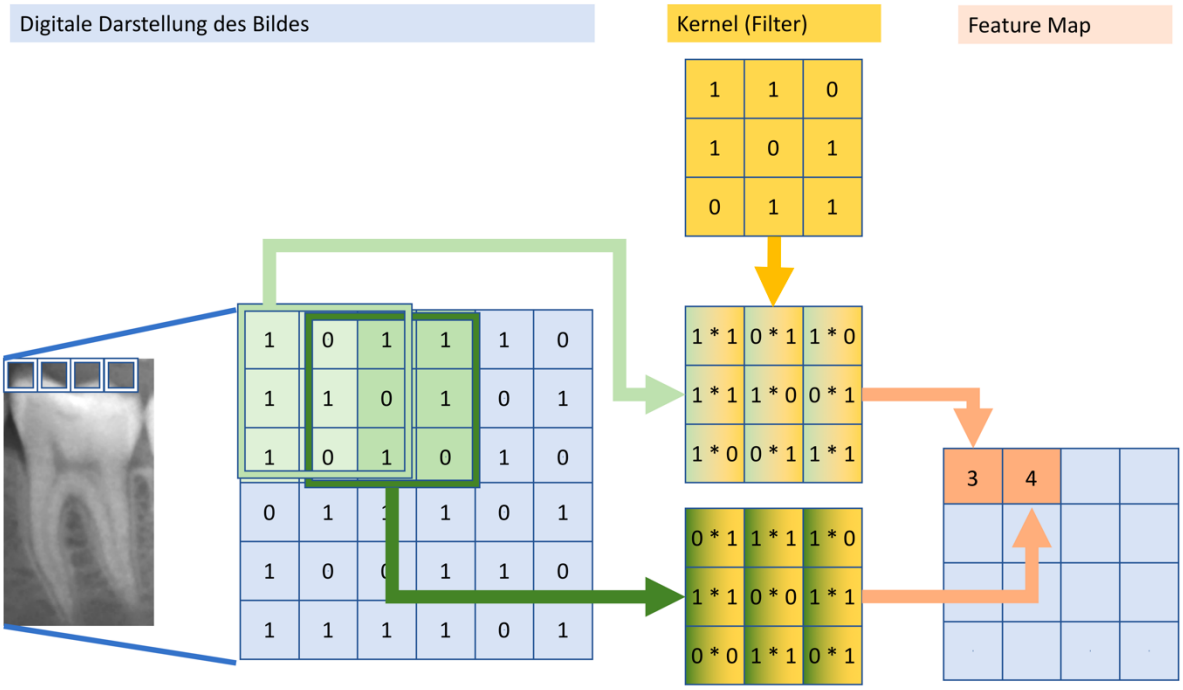


Abbildung 8 *– Convolutional Layer*

berechnet: Zunächst wird eine Kernelgröße festgelegt (in der Abbildung 3x3). Die Kernel-Größe kann ausschlaggebend für die Leistungsfähigkeit des CNNs sein und stellt einen Hyperparameter dar, denn sie wird nicht während des Trainings verändert, sondern zu Beginn bei der Festlegung der Architektur des CNN angenommen. Danach wird der Kernel mit zufälligen Werten belegt. Beginnend mit der linken oberen Ecke der Bilddaten wird nun der Kernel über die Pixel des Bildes gelegt (hellgrünes Quadrat links) und die Werte des Bildes mit den Werten des Kernel multipliziert und die Ergebnisse summiert. In der Abbildung ist dies in der mittleren grün-gelben Matrix unterhalb des Kernels dargestellt. Die Berechnung im Beispiel ergibt den Wert 3 für den ersten Datenpunkt des Filters. Nun wird der Kernel um einen Pixel nach rechts verschoben (dunkel-grünes Quadrat links) und die Berechnung für die nächste Gruppe von Pixeln des Eingangsbildes berechnet (unteres grün-gelbes Quadrat).

In der Regel werden mehrere dieser Convolutional Layer hintereinandergeschaltet und jeweils mit einem so genannten Max Pooling – eine weitere mathematische Operation, die auf die Werte der Feature Map angewendet wird – kombiniert. Hierbei wird erneut ein Kernel pixelweise über die Daten verschoben, jedoch jeweils der Maximal-Wert als Ergebnis verwendet. Die letzten Netzwerkschichten eines CNN sind „dense layer“ (s.o.) und übersetzen die nach ihren Merkmalen gefilterten Bilder typischerweise in Klassenwerte, die als Wahrscheinlichkeiten für eine bestimmte Klasse (z.B. krank oder nicht) interpretierbar sind. Dies wird durch die Klassifizierungsfunktion (siehe Abbildung 7) in der Ergebnisschicht des Neuronalen Netzwerkes erreicht. Das Ergebnis der verwendeten sigmoid Funktion ($\frac{1}{1+e^{-x}}$) wird als eine Wahrscheinlichkeit (0...1) interpretiert, die angibt, ob der Input (das Bild) zur prävalenten Klasse gehört (bei einer binären Klassifikation). In unserem Fall ist dies die Wahrscheinlichkeit, ob AL vorliegt. Der Ausgabewert liegt zwischen 0 und 1. Es wurden Werte kleiner als 0,5 der Klasse „AL liegt nicht vor“ zugeordnet. Werte größer oder gleich 0,5 wurden der Klasse „AL liegt vor“ zugeordnet.

Für die Modellentwicklung kam das Keras- und das Tensorflow-Framework zum Einsatz. Im verwendeten CNN wurden „convolutional layer“ mit „rectified linear units“ (ReLUs) und „max-pooling layer“ kombiniert. „Batch-normalization layer“ (Ioffe, Szegedy 2015) und „dropout layer“ kamen für die Modellregulation zum Einsatz. Eine

Reihe von „fully connected dense layer“ und ein „sigmoid classifier“ bildeten die letzten Modellschichten. Details zur Modell-Architektur befinden sich in Kapitel 5.1 und Tabelle 1.

Eine weitere Schwierigkeit neben der Optimierung der Gewichte während des Trainings eines neuronalen Netzes ist das Auffinden einer optimalen Architektur. Außerdem ist die Optimierung weiterer Parameter – z.B. Trainingsparameter wie Batchsize oder Art des „Optimizer“ – erfolgskritisch. Das systematische Optimieren dieser Parameter wird auch als Hyperparameter-Suche bezeichnet. Hyperparameter sind Parameter, die – anders als die eingangs beschriebenen Gewichte des Neuronalen Netzes – nicht während des Trainings variiert werden, sondern die Architektur beeinflussen. Sie werden vor dem Training festgelegt, dann wird das Training durchgeführt. Wird der Hyperparameter verändert, muss das Netzwerk neu trainiert werden. Zwei so trainierte Netzwerke können anhand der ermittelten Performance-Metriken miteinander verglichen werden. Bei der Hyperparameter-Suche werden derartige Parameter systematische variiert und die Auswirkung auf die Leistung des Netzwerkes ermittelt, um die Leistungsfähigkeit zu optimieren. Für das vorliegende Modell kam zum Hyperparametertuning eine Rastersuche (Claesen, De Moor 2015) zur Anwendung. Es wurden die Anzahl und Reihenfolge der Netzwerk-Schichten, die Dimensionen der Zwischenschichten (d.h. unter anderem die Anzahl der Neuronen), die Anzahl der Filter innerhalb eines „convolutional layer“, deren Kernelgrößen und das verwendete Padding (alle drei sind Parameter, die die Wirkungsweise eines „convolutional layer“ beeinflussen), Pooling-Größe, Bias- und Aktivierungsfunktionen variiert. Des Weiteren wurden unterschiedliche Ansätze und Parametrisierungen in der Bildvorverarbeitung und verschiedene Arten von „Optimizer“ untersucht, ebenso die Lernraten, die Verwendung von Dropout- und Batch-Normalisierungs-Schichten und deren Parametrisierung und Positionen und die Batch-Größe. Auch die optimalen Parameter für die eingesetzte „image augmentation“ – hierbei werden die Eingangsbilder durch einfache Bildoperationen wie Drehen und Zoom verändert, um einen größeren Eingangs-Datensatz zu erhalten – wurden durch eine Hyperparameter-Suche gefunden.

Zur Vorbereitung für das Training im CNN wurden die Bilddaten digital bearbeitet: (1) Jedes Bildsegment wurde in Graustufen umgewandelt. (2) Segmente aus dem Oberkiefer wurden um 180° gedreht, so dass in allen Bildern die Kronen nach oben und die Wurzeln nach unten gerichtet waren. (3) alle Pixelwerte jedes Bildsegments wurden auf einen festen Bereich [0, 1] normiert. (4) alle Bildsegmente wurden auf 64 × 64 Pixel skaliert. Für die „image augmentation“ kamen „feature-wise center“, “zca-whitening”, Skalierung, Scheren, Zoomen und Rotation zum Einsatz sowie das horizontale und vertikale Spiegeln der Bilder (Hussain et al. 2018). Am besten waren Bildrotation (± 20 Grad), Scheren (von 0,8 bis 1,2) und Zoomen (von 0,8 bis 1,2) geeignet. Die Programmierung erfolgte mit Python und Drittanbieter-Bibliotheken wie NumPy, Pandas, scikit-image und scikit-learn.

Für die Aufteilung des Datensatzes in Trainings- und Validierungsdatensatz kam ein so genanntes „group shuffling“ zum Einsatz. Hierbei werden die Daten einer bestimmten Gruppe zusammengehalten und sichergestellt, dass sich die gesamte Gruppe entweder ausschließlich im Trainings- oder ausschließlich im Validierungsdatensatz befindet. So wird verhindert, dass gelernte Strukturen sowohl trainiert und gleichzeitig später validiert werden, wodurch die Generalisierungsleistung des CNN nicht mehr überprüfbar wäre (Müller, Guido 2017). Im vorliegenden Fall besteht eine „Gruppe“ aus allen einzelnen Zahnsegmenten eines Panoramaröntgenbildes eines Patienten. Das „group shuffling“ hat im vorliegenden Fall zur Folge, dass in der Kreuzvalidierung (s.u.) nicht in jedem Durchlauf („fold“) die exakt selbe Anzahl einzelner Zahnsegmenten im Trainings- und Validierungssatz sind, da nicht jedes Gebiss (Panoramaröntgenaufnahme) die exakt selbe Anzahl von Zähnen hat.

Der Bilddatensatz war schlecht balanciert, d.h. es gab erheblich weniger als prävalent eingestufte Zahnsegmente als solche, die als nicht betroffen eingestuft worden waren (AL war relativ selten). Für das Training eines Neuronalen Netzes ist es aber wichtig, dass es die Strukturen aller zu klassifizierenden Klassen lernen kann, mithin also für jede Klasse (AL liegt vor / liegt nicht vor) die möglichst gleiche Menge an Trainingsdaten erhält, da sonst im ungünstigsten Fall (nur sehr wenige prävalente Daten unter einer erheblichen Überzahl nicht prävalenter Daten) „lernen“ könnte, dass

die ausschließliche Klassifizierung in „nicht prävalent“ stets hervorragende Ergebnisse liefert. Um diese negativen Auswirkungen des Klassenungleichgewichts auf die Modellleistung zu reduzieren (Buda et al. 2017), wurden vor dem Training Bildinstanzen aus der Minderheitenklasse vervielfältigt, also ein so genanntes „over sampling“ durchgeführt. Hierdurch vergrößert sich die Gesamtzahl der Zahnsegmente. Die Validierung erfolgte stets anhand des unveränderten Validierungsdatensatzes. Jedes Training wurde durch Kreuzvalidierung („k-fold-cross -validation“) zehnfach wiederholt und überprüft, um die Robustheit der CNN-Leistung zu bewerten. Das Training erfolgte unter Ubuntu Linux 16.04 LTS und auf einer Nvidia GTX 1080 TI GPU.

4.5 Performance-Metriken

Die AUC wurde als primäre Performance-Metrik verwendet. Sie zeigt, wie gut ein Test (hier: ein Modell) in der Lage ist, korrekte Klassifizierungen (gesund/krank) vorzunehmen. Weitere Metriken waren die Sensitivität und Spezifität sowie die positiven und negativen Vorhersagewerte (PPV/NPV). Man beachte, dass die PPV/NPV stark von der Prävalenz der Krankheit (hier: AL) abhängen. Sie eignen sich zwar, um den diagnostischen Wert eines Tests (eines Modells) in einer bestimmten Population zu beschreiben, können aber je nach Population stark variieren. Wie gut die Experten in ihrer Annotation übereinstimmten wurde durch die Berechnung des Fleiss‘ Kappa (A) berechnet, das sich aus dem Scott’s pi (Scott 1955) ableitet. Hierbei wurde die Zuverlässigkeit der Übereinstimmung der Bewertungen der Nominalskale jeweils zwischen mehr als zwei Experten zugrunde gelegt (Fleiss 1971). Das Fleiss‘ Kappa ist ein Maßstab für die Übereinstimmung innerhalb einer Gruppe.

$$\kappa_{Fleiss} = \frac{P^* - P_e^*}{1 - P_e^*} \qquad \text{(A)}$$

Die Ausdrücke P^* und P_e^* sind die Wahrscheinlichkeit der vollständigen Übereinstimmung und die Wahrscheinlichkeit der Übereinstimmung bei Zufälligkeit.

Die Sensitivität (B, auch als recall bezeichnet), die Spezifität (C), der positive Vorhersagewert (PPV, auch als precision bezeichnet, D) und der negative Vorhersagewert (NPV, (E) (Sokolova, Lapalme 2009) werden im Folgenden abgeleitet:

$$sensitivity\ (recall) = \frac{TP}{TP+FN} \qquad \text{(B)}$$

$$specificity = \frac{TN}{TN+FP} \qquad \text{(C)}$$

$$PPV\ (precision) = \frac{TP}{TP+FP} \qquad \text{(D)}$$

$$NPV = \frac{TN}{TN+FN} \qquad \text{(E)}$$

Hierbei bezeichnen TP und TN die Werte “true positive” (tatsächlich betroffen) und „true negative“ (tatsächlich nicht betroffen). FP und FN sind entsprechend die Werte „false positive“ (fälschlich als betroffen eingestuft) und „false negative“ (fälschlich als nicht betroffen eingestuft). Die AUC zeigt die Fähigkeit einer Klassifizierung, eine

derartige falsche Klassifizierung zu vermeiden. Die Sensitivität (recall) zeigt die Fähigkeit einer Klassifizierung, positive Werte korrekt zu erkennen, die Spezifität zeigt die Fähigkeit, negative Werte korrekt zu erkennen. Der PPV (precision) erklärt die Übereinstimmung der positiven Annotationen mit den tatsächlich positiven Werten der Klasse, der NPV erklärt die entsprechende Übereinstimmung zwischen den negativen Klassifikationen und den tatsächlich negativen Werten (Sokolova, Lapalme 2009).

4.6 Sensitivitätsanalysen

Für das Basis-Szenario war das CNN auf Grundlage von Annotationen trainiert und validiert worden, die sowohl unsichere (erweitertes Parodontalligament) als auch sichere (eindeutig nachweisbare Läsion) AL an bewertbaren Zähnen einbeziehen. Referenztest war die Mehrheitsentscheidung der annotierenden Zahnärzte mit einer Übereinstimmungsmarge von zwei (siehe oben). Zur Validierung des CNN wurde überprüft, ob das CNN die Zahnsegmente des Validierungsdatensatzes – also Zähne, die dem CNN nicht durchs Training bekannt waren – übereinstimmend klassifiziert.

In einem weiteren Szenario wurde die Klassifizierungsqualität des Modells nur für sichere AL überprüft, also ein Zahnsegment nur dann als betroffen klassifiziert, wenn vier aus sechs Zahnärzten (ausschließlich) sichere AL annotiert hatten. Darüber hinaus wurde die Klassifizierungsfähigkeit des Modells auch in Bezug auf die verschiedenen Zahngruppen (Schneidezähne, Eckzähne, Prämolaren, Molaren) trainiert und validiert. Hintergrund dieser Analysen ist, dass sich die diagnostische Unsicherheit (Bewertung in Bezug auf unsichere und sichere AL versus nur in Bezug auf sichere AL) auch in der Leistungsfähigkeit des CNN widerspiegeln kann (oder auch nicht), und dass aufgrund des radiographischen Bilderstellung verschiedene Zahngruppen unterschiedlich schwer zu beurteilen sind.

Weiterhin war Gegenstand der Untersuchung, wie sich die diagnostische Übereinstimmung der jeweiligen Annotation der einzelnen Zahnärzte – Maßstab für die Interpretierbarkeit eines Zahnbildes und die Einfachheit der Diagnose – auf die Qualität der Klassifizierungen des CNN auswirken würde, indem die Übereinstimmungsmarge im Referenztest in einer weiteren Untersuchung auf sechs

erhöht wurde (d.h. alle Zahnärzte müssen in ihrer Klassifizierung übereinstimmen, ansonsten wurde das Bild verworfen).

5 Ergebnisse

5.1 Modellarchitektur

Das Ergebnis der Hyperparametersuche und der Architektur-Optimierung war ein Feed-Forward-CNN zur Klassifizierung von AL bestehend aus sieben Schichten und einer Gesamtzahl von 4.299.651 trainierbaren Gewichten (Abbildung 9). Im Basis-Szenario wurde das CNN im Durchschnitt (SD) auf 2238 (56) Bildern trainiert und auf 341 (24) Bildern validiert (cave: die Zahl der Bilder ist durch over-sampling höher als die Zahl der tatsächlich vorhandenen Bilder).

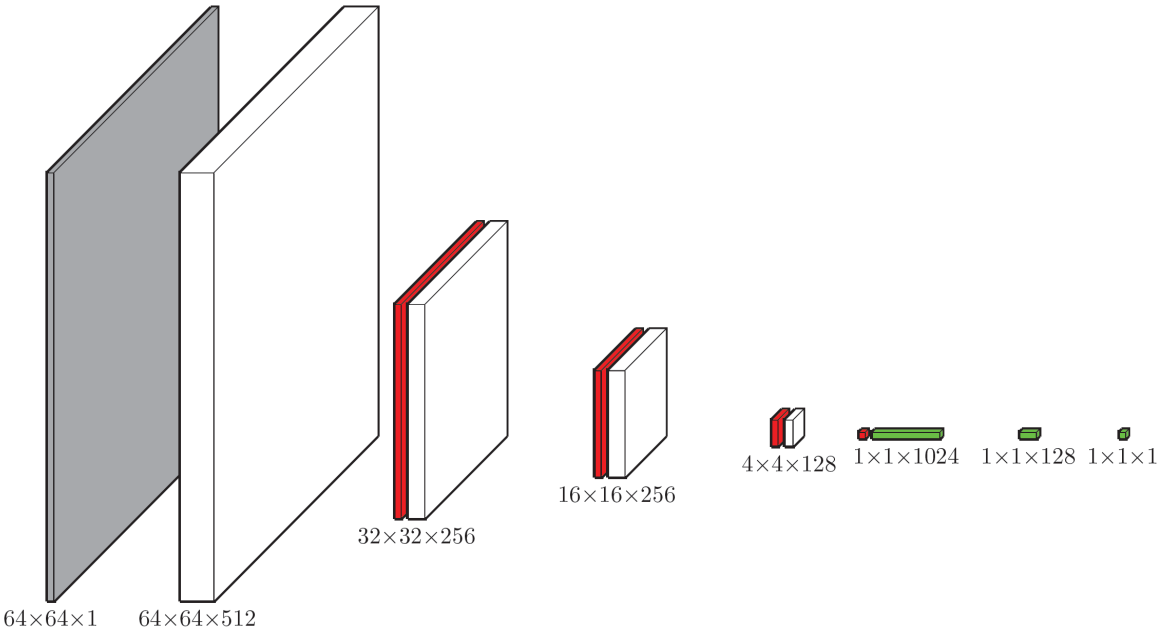


Abbildung 9 *– Modell Architektur.*

Das Modell besteht aus einer Kette mehrerer verbundener Schichten („convolutional layer“ – weiß; „max-pooling layer“ – rot; „dense layer“ – grün). Die Zahlen zeigen die Breite, Höhe und Tiefe jeder einzelnen Schicht. Die Rohdaten des Bildes (graue Schicht) werden von links nach rechts durch das Neuronale Netz verarbeitet. Durch die Aufeinanderschichtung mehrerer „convolutional layer“ und durch das „max-pooling“ werden die extrahierten Strukturen immer komplexer. Die „dense-layer“ (grün) auf der rechten Seite dienen dazu die so extrahierten Strukturen in eine binäre Klassifizierung umzuwandeln (AL vorhanden oder nicht).

Detailinformationen zum gefundenen architektonischen Aufbau sind in Tabelle 1 dargestellt. Für die Entwicklung des Modells wurde jeweils die Performance des CNNs in Bezug auf die AUC zugrunde gelegt. Für die Optimierung der Architektur und Hyperparameter stellten sich folgende Hyperparameterbereiche als zielführend heraus: Es wurden verschiedene Anzahlen von Neuronen (16 bis 2048 in Exponenten von 2) und die Anzahl der Filter (16 bis 2048 in Exponenten von 29 jeder Convolution Schicht variiert. Verschiedene Kernel-Größen (2x2 bis 5x5) und verschiedene Konfigurationen der max-pooling Schichten (2x2 bis 4x4) wurden untersucht.

Tabelle 1 *- Schichten des Modells und Hyperparameter*

Nummer der Schicht	Art der Schicht	Ausgabe-Größe im finalen CNN	Kernel- und Pooling-Größe im finalen Modell	Anzahl der trainierbaren Gewichte im finalen Modell
	Input	64, 64, 1		0
1	Conv2D	64, 64, 512	3, 3	5120
	ReLU	64, 64, 512	-	0
	Max Pooling	32, 32, 512	2, 2	0
2	Conv2D	32, 32, 256	3, 3	1179904
	ReLU	32, 32, 256	-	0
	Max Pooling	16, 16, 256	2, 2	0
3	Conv2D	16, 16, 256	3, 3	590080
	ReLU	16, 16, 256	-	0
	Max Pooling	8, 8, 256	2, 2	0
4	Conv2D	8, 8, 128	3, 3	295040
	ReLU	8, 8 128	-	0
	Max Pooling	4, 4, 128	2, 2	0
	Flatten	2048	-	0
5	Dense	1024		2098176
	ReLU	1024		0
	DropOut (0.5)	1024	-	0
6	Dense	128		131200
	ReLU	128		0
	DropOut (0.7)	128	-	0
7	Dense	1		129
	Batch normalization	1	-	4
	sigmoid	1	-	0
	DropOut (0.5)	1	-	0
	Output	1	-	0

Als Aktivierungsfunktion wurden ReLUs und sigmoid verwendet. Dropout Schichten wurden auf Ihre Wirkung mit Parametern (Wirkungsgrad des dropout: 0.0 = keine Wirkung bis 1.0 = Alle Gewichte werden verworfen) von 0.1 bis 0.9 in Schritten von 0.1

hin untersucht. Weiterhin wurde die Klassifizierungsqualität des CNNs im Zusammenspiel mit oder ohne den Einsatz von batch-normalization Schichten untersucht, um over-fitting zu vermeiden und die Konvergenz des Neuronalen Netzes zu optimieren. Over-fitting ist ein Phänomen, das das „Auswendiglernen“ der Trainingsdaten unter Verlust der Urteilsfähigkeit für unbekannte Validierungsdaten durch Modelle bezeichnet.

Für das endgültige Modell ergaben sich optimale Werte beim Einsatz des „Adam Optimizer“, einer „learning rate“ von 0,0001, einer Bild-Größe von 64x64 und einer Batch-Größe von 32 über 100 Epochen. Als Loss-Funktion kam so genannte „cross entropy“ zum Einsatz. Für die Ermittlung dieses Ergebnisses wurden als Optimizer sowohl der „Adam Optimizer“ als auch der RMSprop Optimizer untersucht. Die learning-rate wurde mit verschiedenen Werten variiert (0,0001; 0,0002; 0,002) ebenso die Bild-Größen (64x64 und 128x128) und Batch-Größen (1 bis 128 in Exponenten von 2). Für die „image augmentation“ waren Bildrotation (± 20 Grad), Scheren (von 0,8 bis 1,2) und Zoomen (von 0,8 bis 1,2) am besten geeignet (vgl. Kapitel 0).

5.2 Ergebnisse der Klassifizierung

Im Basis-Szenario – also bei der Untersuchung sicherer und unsicherer AL bei einer Übereinstimmungsmarge der Mehrheitsentscheidung von zwei – betrug der mittlere (SD) AUC des CNN 0,85 (0,04) bei einer Sensitivität von 0,65 (0,12) und einer Spezifität von 0,87 (0,04). Der resultierende PPV betrug 0,49 (0,10), der NPV 0,93 (0,03). Die Prävalenz betrug im Durchschnitt 0,16 (0,03). AL waren bei Molaren häufiger vorhanden als bei Prämolaren, Eckzähnen oder Schneidezähnen (Tabelle 2). Die entsprechenden ROC-Kurven sind in Abbildung 10 dargestellt.

Tabelle 2 – Klassifizierung im Basis-Szenario – *Vorliegen von sicherer und unsicherer AL bei einer Marge von zwei bzw. sechs in der Mehrheitsentscheidung.*

T.	Referenz Test	-Prävalenz Valid. Set	Bilder Train. Set	Bilder Valid. Set	AUC	Sens.	Specif.	PPV	NPV
All	Mehrheit (2)	0.16 ±0.03	2238 ±56	341 ±24	0.85 ±0.04	0.65 ±0.12	0.87 ±0.04	0.49 ±0.10	0.93 ±0.03
in	Mehrheit (2)	0.08 ±0.04	2238 ±56	102 ±8	0.82 ±0.06	0.55 ±0.18	0.92 ±0.05	0.42 ±0.19	0.96 ±0.03
ca	Mehrheit (2)	0.10 ±0.03	2238 ±56	51 ±5	0.86 ±0.08	0.52 ±0.22	0.96 ±0.03	0.56 ±0.20	0.95 ±0.03
pm	Mehrheit (2)	0.16 ±0.05	2238 ±56	93 ±5	0.85 ±0.06	0.50 ±0.12	0.90 ±0.04	0.49 ±0.17	0.90 ±0.04
m	Mehrheit (2)	0.27 ±0.04	2238 ±56	94 ±9	0.84 ±0.06	0.80 ±0.14	0.70 ±0.08	0.51 ±0.07	0.90 ±0.07
All	Mehrheit (6)	0.13 ±0.04	1331 ±38	195 ±17	0.95 ±0.02	0.74 ±0.19	0.94 ±0.04	0.67 ±0.14	0.95 ±0.04
in	Mehrheit (6)	0.07 ±0.03	1331 ±38	63 ±5	0.92 ±0.08	0.64 ±0.30	0.96 ±0.04	0.64 ±0.27	0.97 ±0.03
ca	Mehrheit (6)	0.09 ±0.06	1331 ±38	37 ±3	0.96 ±0.03	0.52 ±0.31	0.98 ±0.03	0.73 ±0.33	0.95 ±0.04
pm	Mehrheit (6)	0.11 ±0.05	1331 ±38	55 ±7	0.95 ±0.04	0.63 ±0.31	0.95 ±0.03	0.55 ±0.21	0.94 ±0.05
m	Mehrheit (6)	0.31 ±0.08	1331 ±38	41 ±6	0.94 ±0.03	0.87 ±0.14	0.84 ±0.13	0.74 ±0.15	0.94 ±0.07

T.: All: Alle Zähne, in: Schneidezähne, c: Eckzähne, pm: Prämolaren, m: Molaren

Mehrheit: Referenztest ist das Ergebnis der Mehrheitsentscheidung von sechs unabhängigen Experten ob AL vorliegt oder nicht. Die Marge musste bei zwei (d.h. mindestens vier Experten stimmen überein – 4-2 = 2) bzw. sechs (alle stimmen überein) liegen.

Die Tabelle zeigt die area-under-the-curve (AUC), Sensitivität, Spezifität und positiven und negativen Vorhersagewert (PPV, NPV) als auch die jeweiligen Standardabweichungen (SD). Für das Basis-Szenario wurden alle Zähne für das Training verwendet und die Annotation anhand der Mehrheitsentscheidung auf unsichere und sichere AL bei einer Marge von 2 durchgeführt. Für die Sensitivitätsanalyse wurde die Marge auf sechs hochgesetzt und das Modell zusätzlich auf verschiedene Zahngruppen angewandt. Zu beachten ist, dass die Anzahl der Zähne im Trainings-Set aufgrund des over-samplings der Minderheitenklasse (also der prävalenten Zahnsegmente) variiert.

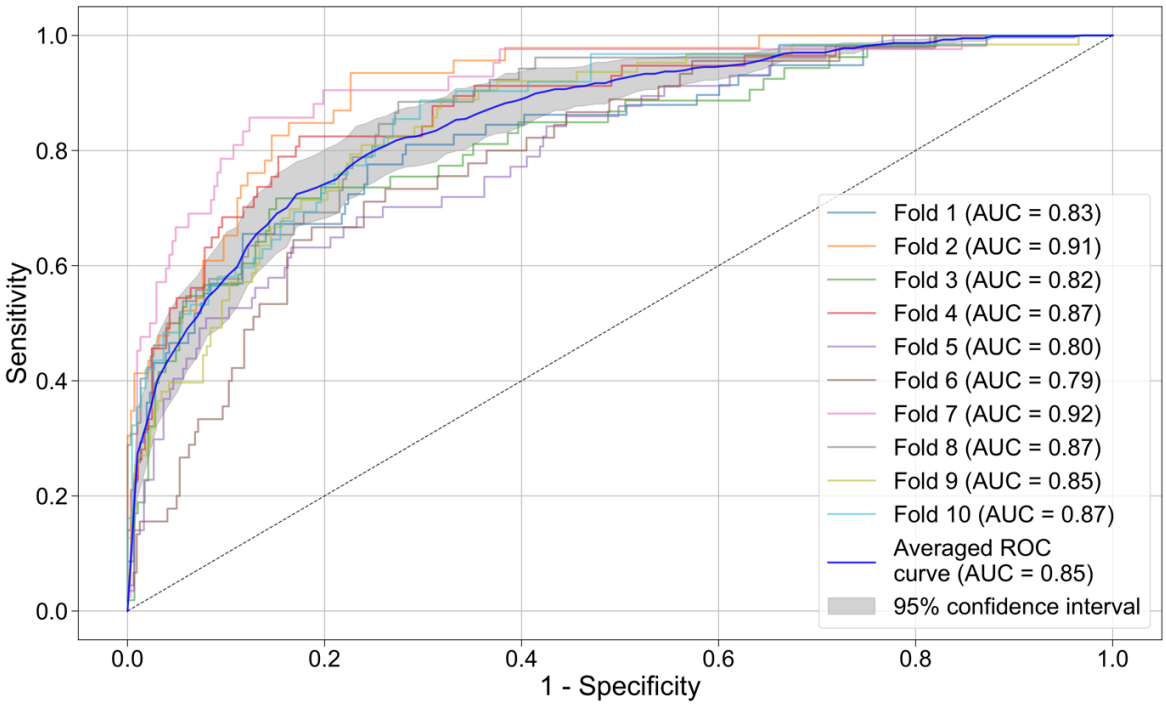


***Abbildung 10** – Receiver operation characteristic (ROC) Kurve des **Basis-Szenarios**.*

Das CNN wurde gegen den Referenztest in Bezug auf die Sensitivität (der Anteil positiver – AL ist vorhanden – Ergebnisse, die korrekt klassifiziert wurden) und die Spezifität (der Anteil der negativen – AL ist nicht vorhanden – Ergebnisse, die korrekt klassifiziert wurden) überprüft. Die farbigen Kurven zeigen die Fähigkeit jedes einzelnen Durchlaufs der zehnfachen Kreuzvalidierung an, die Klassifizierung korrekt auszuführen. Die dickere Blaue Linie zeigt den Mittelwert aller ROC Kurven. Der graue Bereich entspricht dem 95% Konfidenz-Intervall. Zusätzlich wird der AUC als Maß für die Klassifizierungs-Fähigkeit angegeben.

Nachdem das CNN mittels aller Zähne (des Trainingsdatensatzes) trainiert worden war, wurde es herangezogen, um in einer Sensitivitätsanalyse eine Klassifizierung für nur einzelne Zahngruppen vorzunehmen (Tabelle 2). Die AUC für Schneidezähne, Eckzähne, Prämolaren und Molaren betrug 0,82 (0,06), 0,86 (0,08), 0,85 (0,06) bzw. 0,84 (0,06). Bei Molaren war die Sensitivität signifikant höher als bei anderen Zahngruppen, während die Spezifität geringer war. Der resultierende PPV blieb jedoch

begrenzt und lag zwischen 0,42 (0,19) für Schneidezähne und 0,56 (0,20) für Eckzähne. Der NPV hingegen war für alle Zahngruppen hoch (>0,90).

Durch die Einschränkung auf Bilder, bei denen alle Untersucher zustimmten (Übereinstimmungsmarge von sechs), nahm die Anzahl der verfügbaren Bilder ab (Tabelle 2). Die AUC stieg signifikant auf 0,95 (0,02), ebenso die Sensitivität auf 0,74 (0,19). Infolgedessen stieg der PPV deutlich auf 0,67 (0,14). Die entsprechenden

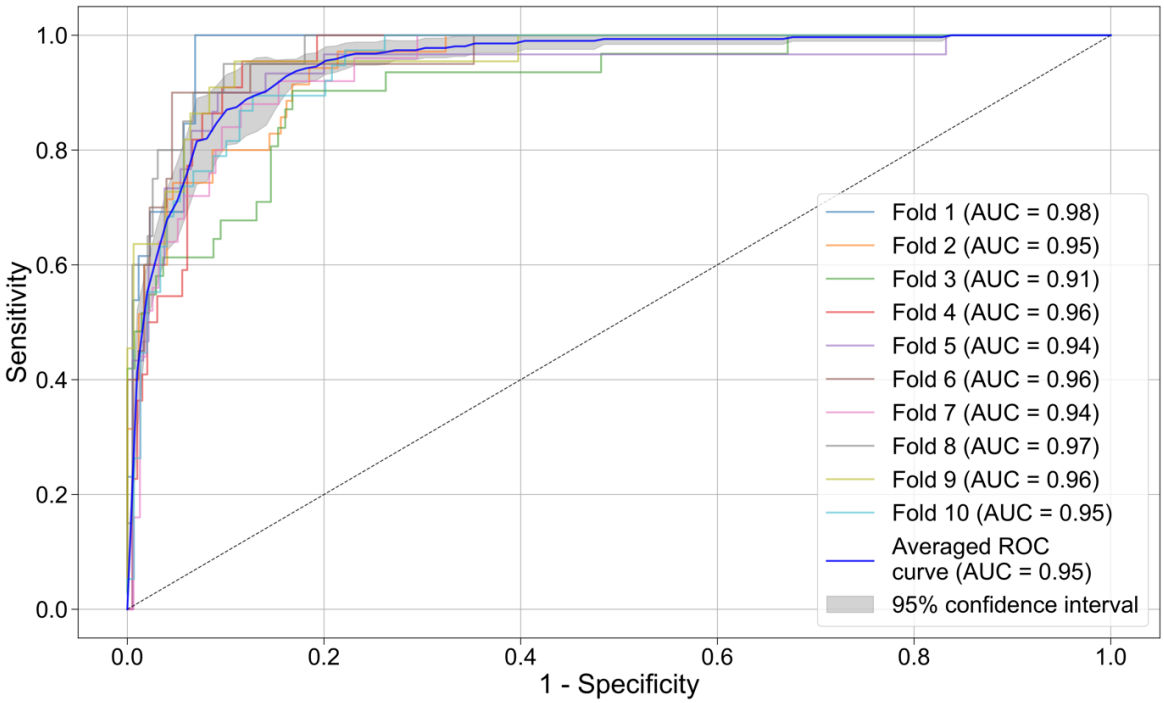


***Abbildung 11** – Receiver Operation Characteristics (ROC) Kurve **für unsichere und sichere AL gemeinsam** bei einer gleichzeitigen **Übereinstimmungs-Marge von sechs**.*

Das CNN wurde gegen den Referenztest in Bezug auf die Sensitivität (der Anteil positiver – AL ist vorhanden – Ergebnisse, die korrekt klassifiziert wurden) und die Spezifität (der Anteil der negativen – AL ist nicht vorhanden – Ergebnisse, die korrekt klassifiziert wurden) überprüft. Die farbigen Kurven zeigen die Fähigkeit jedes einzelnen Durchlaufs der zehnfachen Kreuzvalidierung an, die Klassifizierung korrekt auszuführen. Die dickere Blaue Linie zeigt den Mittelwert aller ROC Kurven. Der graue Bereich entspricht dem 95% Konfidenz-Intervall. Zusätzlich wird der AUC als Maß für die Klassifizierungs-Fähigkeit angegeben.

ROC-Kurven sind in Abbildung 11 dargestellt. Es gab nur geringe Unterschiede bei der Analyse nach Zahngruppen im Vergleich zum Basis-Szenario (d.h. auch hier war die Sensitivität bei Molaren höher, während die Spezifität geringer war).

Im Szenario, für das nur sichere AL als betroffen eingestuft wurde, sank die Prävalenz auf 0,07 (0,02). In diesem Fall stieg die AUC deutlich auf 0,89 (0,04) (Tabelle 3, ROC-Kurve in Abbildung 12). Dies war auf einen deutlichen Anstieg der Sensitivität auf 0,71 (0,14) zurückzuführen, während die Spezifität leicht auf 0,84 (0,07) zurückging. Aufgrund der geringen Prävalenz sank der PPV je nach Zahngruppe auf Werte zwischen 0,21 und 0,34. Die Untersuchung nach Zahngruppen ergab hier ein vergleichbares Bild wie im Basis-Szenario.

Tabelle 3 – Sensitivitätsanalyse: *Klassifizierung auf das Vorliegen von nur sicherer AL bei einer Marge von zwei bzw. sechs in der Mehrheitsentscheidung*

T.	Referenz Test	-Prävalenz Valid. Set	Bilder Train. Set	Bilder Valid. Set	AUC	Sens.	Specif.	PPV	NPV
All	Mehrheit (2)	0.07 ±0.02	2536 ±54	353 ±24	0.89 ±0.04	0.71 ±0.14	0.84 ±0.07	0.29 ±0.09	0.97 ±0.01
in	Mehrheit (2)	0.03 ±0.02	2536 ±54	105 ±8	0.86 ±0.26	0.69 ±0.38	0.91 ±0.05	0.22 ±0.16	0.99 ±0.01
ca	Mehrheit (2)	0.04 ±0.03	2536 ±54	54 ±4	0.97 ±0.04	0.62 ±0.44	0.94 ±0.04	0.34 ±0.26	0.99 ±0.02
pm	Mehrheit (2)	0.07 ±0.02	2536 ±54	95 ±6	0.82 ±0.05	0.50 ±0.24	0.84 ±0.09	0.21 ±0.13	0.96 ±0.02
m	Mehrheit (2)	0.13 ±0.03	2536 ±54	99 ±12	0.87 ±0.04	0.82 ±0.15	0.71 ±0.13	0.33 ±0.09	0.96 ±0.03
All	Mehrheit (6)	0.05 ±0.02	2162 ±53	292 ±24	0.92 ±0.04	0.40 ±0.24	0.96 ±0.05	0.43 ±0.20	0.97 ±0.01
in	Mehrheit (6)	0.01 ±0.01	2162 ±53	91 ±10	0.93 ±0.09	0.10 ±0.30	0.98 ±0.02	0.01 ±0.04	0.99 ±0.01
ca	Mehrheit (6)	0.04 ±0.03	2162 ±53	48 ±4	0.94 ±0.05	0.15 ±0.30	0.98 ±0.02	0.12 ±0.20	0.97 ±0.03
pm	Mehrheit (6)	0.04 ±0.03	2162 ±53	81 ±5	0.89 ±0.07	0.32 ±0.33	0.96 ±0.04	0.25 ±0.31	0.97 ±0.02
m	Mehrheit (6)	0.10 ±0.03	2162 ±53	73 ±9	0.88 ±0.06	0.53 ±0.28	0.90 ±0.12	0.50 ±0.21	0.95 ±0.02

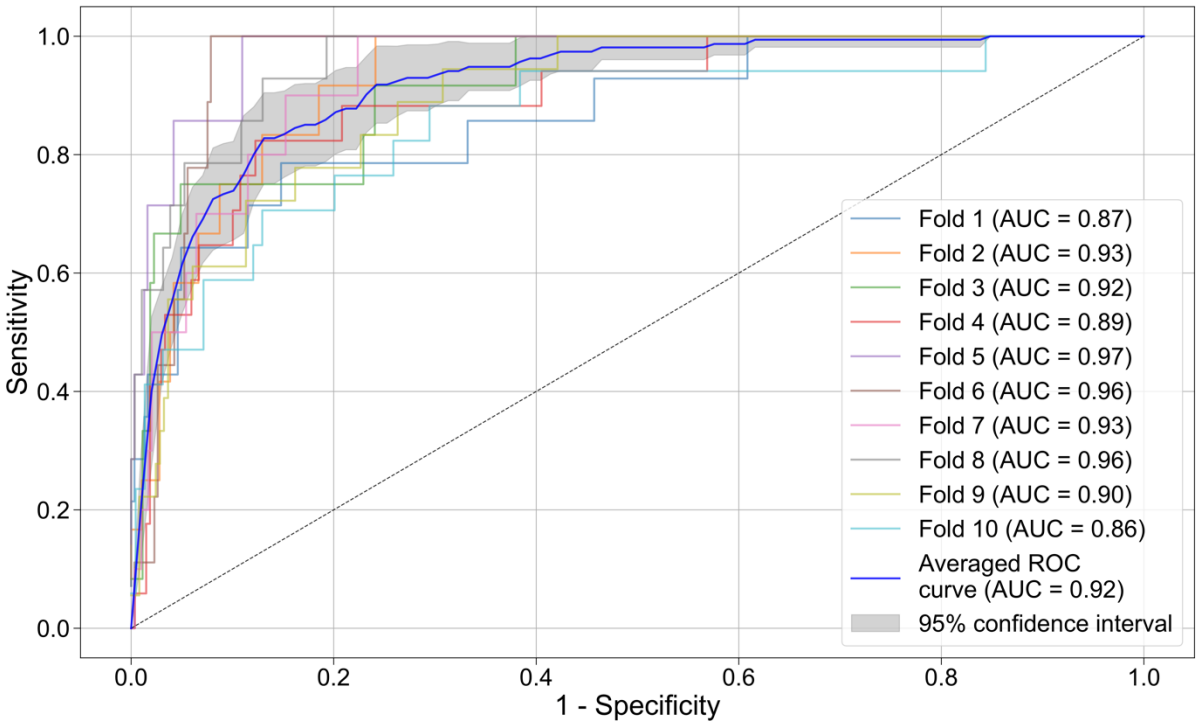


Abbildung 12 *– Receiver Operation Characteristics (ROC) Kurve für* ***nur sichere AL*** *gemeinsam bei einer gleichzeitigen* ***Übereinstimmungs-Marge von sechs****.*

Das CNN wurde gegen den Referenztest in Bezug auf die Sensitivität (der Anteil positiver – AL ist vorhanden – Ergebnisse, die korrekt klassifiziert wurden) und die Spezifität (der Anteil der negativen – AL ist nicht vorhanden – Ergebnisse, die korrekt klassifiziert wurden) überprüft. Die farbigen Kurven zeigen die Fähigkeit jedes einzelnen Durchlaufs der zehnfachen Kreuzvalidierung an, die Klassifizierung korrekt auszuführen. Die dickere Blaue Linie zeigt den Mittelwert aller ROC Kurven. Der graue Bereich entspricht dem 95% Konfidenz-Intervall. Zusätzlich wird der AUC als Maß für die Klassifizierungs-Fähigkeit angegeben.

6 Diskussion

In der vorliegenden Studie wurde ein einfaches Feed-Forward CNN entwickelt. Es bestand aus sieben Schichten und war in seiner Architektur und Parametrisierung anhand einer Hyperparametersuche optimiert und auf einzeln zugeschnittenen Bildsegmenten jeweils eines Zahns trainiert worden. Die Bildsegmente waren zuvor durch Experten auf das Vorliegen von AL hin bewertet und anhand dieser Annotationen ein binärer Wert berechnet worden. Der Bilddatensatz wurde in Trainings- und Validierungsdatensatz aufgeteilt. Das trainierte CNN wurde herangezogen, um die Bilder des Validierungsdatensatzes zu klassifizieren, das so ermittelte Ergebnis mit den vorliegenden Annotationen durch die Experten verglichen und so die Performance-Metriken ermittelt. Das CNN erwies sich trotz der verhältnismäßig geringen Anzahl von Bilddaten als geeignet, um eine zufriedenstellende Zuordnungsfähigkeit für die Erkennung von AL zu erreichen.

Positiven Einfluss auf die Klassifizierungsgenauigkeit des CNN hatte die Augmentierung der Bilddaten. In der Hyperparametersuche war neben einigen globalen Parametern auch die Architektur des CNN dynamisch in einem selbst entwickelten halb-automatischen AutoML Verfahren (vgl. Feurer et al. 2015) variiert worden. Interessant hierbei war, dass bei zu hoher Komplexität der Architektur das over-fitting zu schnell einsetzte und dadurch die Generalisierungsfähigkeit litt. Möglicherweise können bei zu hoher Komplexität der Architektur tiefere Schichten des CNN bei der vorliegenden beschränkten Anzahl von Bilddaten nicht mehr hinreichend gut trainiert werden, da nicht mehr genügend differenzierte Bild-Features zur Verfügung stehen, um hier eine Unterscheidungsfähigkeit herzustellen.

Wie in der Sensitivitätsanalyse gezeigt wurde, stieg der PPV mit steigender Prävalenz. Die Sensitivität blieb beschränkt. Diesem konnte mit einem „over-sampling“ (Vervielfältigung) der Bildinstanzen der Minderheitenklasse (Zahnsegmente mit AL) begegnet werden, dennoch ist zu erwarten, dass das vorliegende CNN zu einer zu niedrigen Erkennungsquote auf AL tendiert, wobei es gleichzeitig einen stabile NPV zeigt.

Bei der Beurteilung für die klinische Einsatzfähigkeit ist der PPV des CNN näher zu betrachten. Der PPV war niedrig, auch wenn das CNN eine recht hohe Spezifität hatte. In der Sensitivitätsanalyse konnte durch die Einschränkung auf Segmente, bei denen ausschließlich sichere AL erkannt worden war, die Klassifizierungsfähigkeit deutlich gesteigert werden (der AUC lag in diesem Fall bei 0,95). Hierbei kam dem Training des CNN zugute, dass die für die Klassifizierung notwendigen Bildstrukturen aufgrund der besser erkennbaren Prävalenz (nur eindeutige AL) deutlicher hervortraten und mithin besser für das Training und eine eindeutige Klassifizierung geeignet waren. Begünstigend kam hinzu, dass die Übereinstimmung der Experten in diesem Fall höher war (das Fleiss-Kappa betrug für die Beurteilung von nur sicherer AL 0,59, während es bei der Beurteilung von sicherer und unsicherer AL bei 0,48 lag), wodurch die für das Training eines CNN störende Ungenauigkeit der Annotationen verringert wurde. Weiterhin konnte gezeigt werden, dass die diagnostische Genauigkeit des CNN auch von den Zahntypen (die Sensitivität war bei der Klassifizierung von Molaren erhöht) – mithin also von der generellen Beurteilbarkeit des Bildmaterials – abhing. Neben der höheren Prävalenz von AL an Molaren ist dies darauf zurückzuführen, dass insbesondere bei weiter anterior liegenden Zahntypen Überlagerungen mit der Wirbelsäule auftreten, die eine Beurteilung erschweren können (Nardi et al. 2017, Nardi et al. 2018).

Weiterhin wurde in dieser Studie erkennbar, dass die Bewertung von Panoramaröntgenbildern in Bezug auf das Vorliegen von AL stark zwischen den einzelnen Experten schwankte (lediglich moderate Übereinstimmung bei einem Fleiss‘ Kappa von 0,48). Wie in der Sensitivitätsanalyse gezeigt wurde, konnte der AUC durch die Betrachtung nur von Ergebnissen mit vollständiger Übereinstimmung aller Experten signifikant gesteigert werden. Im klinischen Einsatz kann ein CNN dazu dienen, Experten während der Diagnose auf besonders schwer zu beurteilende bzw. als vom CNN als prävalent eingestufte Bilder hinzuweisen und so als Computer-gestütztes Assistenzsystem, die diagnostische Arbeit zu erleichtern oder zu verbessern.

Eine Reihe von Einschränkungen und weiterführenden Überlegungen sollen diskutiert werden. Erstens zeigte sich, dass die Hyperparametersuche anhand einer grid search

aufwändig ist und wie auch bereits in der Literatur (Bergstra, Bengio 2012) gezeigt wurde, nicht immer zum optimalen Ergebnis führt. Random Search oder Bayesian Search sollten untersucht werden, um die Hyperparametersuche weiterzuentwickeln. Gleiches gilt zweitens für die Augmentierung. Hier wurden bisher Methoden wie Drehen, Zoomen oder Scheren angewandt. Da gezeigt werden konnte, dass die zugrundeliegende Menge prävalenter Bilddatensätze relevant für das Training des CNNs ist, wäre es interessant zu untersuchen, ob der Einsatz von GAN (Generating Adversarial Networks; Mariani et al. 2018) zur Erzeugung weiterer augmentierter Bilddaten prävalenter Zahnsegmente geeignet ist. Darüber hinaus gibt es weitere Annotationstechniken und spezialisierte Neuronale Netze, die sich nicht auf die binäre Klassifizierung konzentrieren, sondern versuchen, betroffene Areale zu annotieren. Hierbei wird ebenfalls ein supervised learning angewandt, jedoch die betroffenen Areale bei der Annotation durch Experten markiert und durch ein entsprechendes Training vom Neuronalen Netz gelernt werden können (Mask R-CNN; Kaiming et al. 2018). Neben der Augmentierung der einzelnen Zahnsegmente sollte auch das gesamte Gebiss in das Training des CNNs und dessen Klassifizierung einbezogen werden, um die mögliche Korrelation von AL innerhalb desselben Patienten als Information mit nutzen zu können (Masood et al. 2015). In der Studie zeigte sich, dass mit dem vorliegenden begrenzten Bilddatensatz ein Transfer-Learning mit Hilfe von state-of-the-art Architekturen zum over-fitting und dem damit einhergehenden Verlust der Generalisierungsfähigkeit auf Kosten der Genauigkeit in der Klassifizierung führten. Es wäre also drittens zu untersuchen, ob durch die signifikante Vergrößerung des Bilddatensatzes mit besonderem Augenmerk auf das Vorhandensein prävalenter Datensätze eine Verbesserung der Klassifizierungsfähigkeit erreicht werden kann. Wie oben diskutiert und wie zu erwarten war, war viertens die Einigkeit der Bewertung der einzelnen Bildsegmente zwischen den Experten begrenzt. In der Studie wurde diesem mit einer recht großen Anzahl an Experten begegnet. In der Sensitivitätsanalyse stieg für den Fall der vollständigen Übereinstimmung der AUC auf 0.95 bei steigender Sensitivität und stabiler Prävalenz. Der hier zugrunde liegenden Schwierigkeit eines unscharfen Goldstandards könnte durch eine Triangulation der radiographischen Ergebnisse mit den klinischen Befundungen begegnet werden (Walsh 2018). Hilfreich bei der Beleuchtung dieses Zusammenhangs zwischen Genauigkeit des

Goldstandards und der Klassifizierungsfähigkeit des CNN könnte fünftens ein besseres Verständnis über die dem CNN zugrunde liegenden Mechaniken sein. Eine Visualisierung der vom CNN gelernten Bild-Features wie Kanten oder Makrostrukturen könnten zum Verständnis beitragen, inwieweit und ob sich diese algorithmische mit der medizinischen Logik in Verbindung bringen lassen. Gleiches gilt für die Überlegung, inwieweit der Einsatz von CNN im diagnostischen Prozess einen Einfluss auf die Entscheidungsfindung haben und möglicherweise eine wahrscheinlichkeitsbasierte Entscheidungsfindung flankieren kann.

7 Schlussfolgerung

Das entwickelte mehrschichtige, in Hinblick auf Architektur und Hyperparameter optimierte CNN war für die Detektion von AL auf zahnmedizinischen Röntgenbildern geeignet. Der relativ kleine Datensatz und das Klassenungleichgewicht haben möglicherweise jedoch die Klassifikationsfähigkeit des Netzes beschränkt. Vor einem klinischen Einsatz sollte die Sensitivität des CNNs verbessert werden. CNNs können die Detektion von Pathologien auf zahnmedizinischen Röntgenbildern erleichtern und möglicherweise auch verbessern.

8 Abbildungsverzeichnis

9 Tabellenverzeichnis

10 Literaturverzeichnis

Ahlqwist M, Halling A, Hollender L. (1986): Rotational panoramic radiography in epidemiological studies of dental health. Comparison between panoramic radiographs and intraoral full mouth surveys. Swedish Dental Journal 1986;10(1-2):73-84.

Becker AS, Marcon M, Ghafoor S, Wurnig MC, Frauenfelder T, Boss A. (2017): Deep Learning in Mammography: Diagnostic Accuracy of a Multipurpose Image Analysis Software in the Detection of Breast Cancer. Investigative Radiology 2017;52(7):434-440.

Bergstra J., Bengio Y. (2012): Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research 2012; 13: 281-305.

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, de Vet HCW, Kressel HY, Rifai N, Golub RM, Altman DG, Hooft L, Korevaar DA, Cohen JF (2015): STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

Buda M, Maki A, Mazurowski M (2017): A systematic study of the class imbalance problem in convolutional neural network. arXiv 2017: arXiv:1710.05381.

Claesen M, De Moor B. (2015): Hyperparameter Search in Machine Learning. arXiv:1502.02127v2 [cs.LG] 2015.

Connert T, Truckenmüller M, ElAyouti A, Eggmann F, Krastl G, Löst C, Weiger R (2018): Changes in periapical status, quality of root fillings and estimated endodontic treatment need in a similar urban German population 20 years later. Clinical Oral Investigations 2018.

Dössel O. (2016) Bildgebende Verfahren in der Medizin. Springer-Verlag Berlin Heidelberg; E-book Ausgabe; Kapitel „1.1 Bildgebende Verfahren als Bestandteil der Diagnostik und Therapie“.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017): Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115-118.

Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F (2015): Efficient and Robust Automated Machine Learning. In: Advances in Neural Information Processing Systems (28): 2962-2970, [online] http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf [2019-04-07].

Fleiss J. (1971): Measuring Nominal Scale Agreement Among Many Raters. Psychol Bull 1971;76(5):378–382.

Goodfellow I, Bengio Y, Courville A. (2016): Deep Learning. Cambridge, MA: MIT Press; 2016: 326 ff.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016): Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316(22):2402-2410.

Gupta, V (2017): Understanding Feedforward Neural Networks, [online] https://www.learnopencv.com/understanding-feedforward-neural-networks, 2017 [2019-03-05].

Hussain Z, Gimenez F, Yi D, Rubin D. (2018): Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. AMIA Annual Symposium proceedings. AMIA Symposium 2018;2017:979-984.

Huumonen S, Suominen AL, Vehkalahti MM. (2017): Prevalence of apical periodontitis in root filled teeth: findings from a nationwide survey in Finland. International Endodontic Journal 2017;50(3):229-236.

Ioffe S, Szegedy C. (2015): Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv 2015: arXiv:1502.03167v3 [cs.LG].

Kaiming He, Gkioxari G., Dollár P., Girshick R. (2018): Mask R-CNN, arXiv:1703.06870v3 [cs.CV].

Kanagasingam S, Hussaini HM, Soo I, Baharin S, Ashar A, Patel S. (2017): Accuracy of single and parallax film and digital periapical radiographs in diagnosing apical periodontitis - a cadaver study. International Endodontic Journal 2017;50(5):427-436.

LeCun Y, Bengio Y, Hinton G. (2015): Deep learning. Nature 2015;521(7553):436-444.

Lee JH, Kim DH, Jeong SN, Choi SH (2018a): Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. Journal of Dentistry 2018;77:106-111.

Lee JH, Kim DH, Jeong SN, Choi SH (2018b): Diagnosis and prediction of periodontally compromised teeth using a deep learning-based convolutional neural network algorithm. Journal of Periodontal & Implant Science 2018;48(2):114-123.

Leonardi Dutra K, Haas L, Porporatti AL, Flores-Mir C, Nascimento Santos J, Mezzomo LA, Corrêa M, De Luca Canto G (2016): Diagnostic Accuracy of Cone-beam Computed Tomography and Conventional Radiography on Apical Periodontitis: A Systematic Review and Meta-analysis. Journal of Endodontics 2016;42(3):356-364.

Lin PL, Huang PY, Huang PW. (2017): Automatic methods for alveolar bone loss degree measurement in periodontitis periapical radiographs. Computer methods and programs in biomedicine 2017;148:1-11.

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI (2017): A survey on deep learning in medical image analysis. Medical Image Analysis 2017;42:60-88.

Mariani G, Scheidegger F, Istrate R, Bekas C, Malossi C (2018): BAGAN: Data Augmentation with Balancing GAN. arXiv:1803.09655v2 [cs.CV] 5 Jun 2018.

Masood M, Masood Y, Newton JT. (2015): The clustering effects of surfaces within the tooth and teeth within individuals. Journal of Dental Research 2015;94(2):281-288.

Mazurowski M, Buda M, Saha A, Bashir M. (2018): Deep learning in radiology: an overview of the concepts and a survey of the state of the art. arXiv:1802.08717v1 2018.

McCulloch WS, Pitts W. (1943): A logical calculus of the ideas immanent in nervous activity. Bull Math. Biophys., 5, 1943, 115-133.

Müller A. C., Guido S. (2017): Introduction to Machine Learning with Python; A guide for Data Scientists. O'Reilly Media Inc., 2017, Sebastopol, CA 95472; 261-262.

Nardi C, Calistri L, Grazzini G, Desideri I, Lorini C, Occhipinti M, Mungai F, Colagrande S (2018): Is Panoramic Radiography an Accurate Imaging Technique for the Detection of Endodontically Treated Asymptomatic Apical Periodontitis? Journal of Endodontics 2018;44(10):1500-1508.

Nardi C, Calistri L, Pradella S, Desideri I, Lorini C, Colagrande S. (2017): Accuracy of Orthopantomography for Apical Periodontitis without Endodontic Treatment. Journal of Endodontics 2017;43(10):1640-1646.

Parker JM, Mol A, Rivera EM, Tawil PZ. (2017): Cone-beam Computed Tomography Uses in Clinical Endodontics: Observer Variability in Detecting Periapical Lesions. Journal of Endodontics 2017;43(2):184-187.

Rosenblatt F. (1958): The perceptron - a probabilistic model for information storage and organization in the brain. Psychological Review 65, 1958.

Scott WA. (1955): Reliability of Content Analysis: The Case of Nominal Scale Coding. Public Opin Q. 19(3):321.

Segura-Egea JJ, Martin-Gonzalez J, Castellanos-Cosano L. (2015): Endodontic medicine: connections between apical periodontitis and systemic diseases. International Endodontic Journal 2015;48(10):933-951.

Sokolova M, Lapalme G. (2009): A systematic analysis of performance measures for classification tasks. Inf Process Manag. 45(4): 427-437.

Walsh T. (2018): Fuzzy gold standards: Approaches to handling an imperfect reference standard. Journal of Dentistry 2018;74 Suppl 1:47-49.

11 Eidesstattliche Versicherung

„Ich, Thomas Ekert, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Thema „Machine Learning Techniques for Computer Aided Classification of Dental Radiographic Images – Machine Learning Techniken zur Computer-gestützten Klassifizierung von zahnmedizinischen Röntgenbildern“ 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 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.

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

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

_____

Datum Unterschrift des Doktoranden

12 Anteilserklärung an der ausgewählten Publikation

Thomas Ekert, Joachim Krois, Leonie Meinhold, Karim Elhennawy, Ramy Emara, Tatiana Golla, and Falk Schwendicke; Deep Learning for the Radiographic Detection of Apical Lesions. Journal of Endodontics 45/7 (2019) pp. 915-920.

Beitrag im Einzelnen – Thomas Ekert:

- Durchführung der Experimente (Maschinelles Lernen) auf Basis der vorgegebenen Idee, Bild-Daten und Experten-Begutachtungen. Die Durchführung der Experimente einschließlich Hyperparameter-Suche sind in Kapitel 4.1 des Manteltextes beschrieben (Abbildung 1, Abbildung 2).
- Entwicklung des Neuronalen Netzes (Convolutional Neural Network; CNN). Für die Klassifizierung von zahnärztlichen Röntgenbildern wurde ein neuronales Netz entwickelt und durch eine Hyperparameter-Suche optimiert. Die Details sind in Kapitel 4.4 des Manteltextes beschrieben (siehe Abbildung 5-8). Das im Ergebnis entwickelte CNN ist in Tabelle 1 und Kapitel 5.1 beschrieben. Grundlage waren der durch zahnärztliche Experten annotierte Bilddatensatz und Referenztest.
- Dokumentation der Experimente.
- Auswertung und Berechnung der Metriken (vgl. Kapitel 5.1; Abbildung 10, 11; Tabelle 2, 3). Grundlage waren der vorgegebene Referenztest für den die Klassifizierungsleistung des CNN bezüglich der Ziel-Metriken berechnet wurde.
- Erheblicher Beitrag zur Analyse und Interpretation der Ergebnisse (Kapitel 6)
- Erstellen der Abbildungen (1, 2, 3, 6, 7, 8, 10, 11, 12) und Tabellen (1, 2, 3).
- Literaturrecherche zu der vorliegenden Arbeit
- Schreiben des Manuskriptes

_____

Unterschrift des Doktoranden

13 Auszug aus der Journal Summary List

Journal Data Filtered By: **Selected JCR Year: 2017** Selected Editions: SCIE,SSCI
Selected Categories: **"DENTISTRY, ORAL SURGERY and MEDICINE"**
Selected Category Scheme: WoS
Gesamtanzahl: 91 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	PERIODONTOLOGY 2000	4,308	6.220	0.004430
2	JOURNAL OF DENTAL RESEARCH	19,305	5.380	0.020950
3	ORAL ONCOLOGY	8,949	4.636	0.013760
4	CLINICAL ORAL IMPLANTS RESEARCH	14,065	4.305	0.016880
5	International Journal of Oral Science	918	4.138	0.002240
6	JOURNAL OF CLINICAL PERIODONTOLOGY	13,300	4.046	0.011660
7	DENTAL MATERIALS	12,466	4.039	0.012560
8	JOURNAL OF DENTISTRY	8,247	3.770	0.012020
9	JOURNAL OF PERIODONTOLOGY	15,619	3.392	0.011420
10	Journal of Prosthodontic Research	686	3.306	0.001650
11	Clinical Implant Dentistry and Related Research	3,633	3.097	0.008520
12	INTERNATIONAL ENDODONTIC JOURNAL	7,002	3.015	0.007330
13	JOURNAL OF ENDODONTICS	16,585	2.886	0.013050
	JOURNAL OF			

14 Druckexemplar der ausgewählten Publikation

Thomas Ekert, Joachim Krois, Leonie Meinhold, Karim Elhennawy, Ramy Emara, Tatiana Golla, and Falk Schwendicke

Deep Learning for the Radiographic Detection of Apical Lesions. Journal of Endodontics 45/7 (2019) pp. 915-920.

https://doi.org/10.1016/j.joen.2019.03.016

15 Lebenslauf

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

16 Publikationsliste

1. **Thomas Ekert**, Joachim Krois, Leonie Meinhold, Karim Elhennawy, Ramy Emara, Tatiana Golla, and Falk Schwendicke
   Deep Learning for the Radiographic Detection of Apical Lesions. Journal of Endodontics 45/7 (2019) pp. 915-920.

   https://doi.org/10.1016/j.joen.2019.03.016

Impact Factor: 2,886

17 Danksagung

Mein herzlicher Dank gilt meinem Doktorvater PD Dr. Falk Schwendicke für die tolle Betreuung und Unterstützung.

Herrn Dr. Joachim Krois danke ich für seine Begeisterung für die Themen „Machine Learning“ und „Data Science“.

Ich danke im Besonderen meiner Frau Maria, die mich jeden Tag unterstützt hat und mir zur Seite stand und meinen beiden Kindern, die zu oft auf mich verzichten mussten.