Hochschule Darmstadt
Fachbereiche Mathematik und Naturwissenschaften &
Informatik

Vorhersage des peripheren Sehvermögens anhand von Fundusfotografien mittels Machine Learning

Abschlussarbeit zur Erlangung des akademischen Grades
Master of Science (M. Sc.) im Studiengang Data Science

vorgelegt von
Svenja Schuder

Referent: Prof. Dr. Arnim Malcherek
Korreferent: Prof. Dr. Horst Zisgen

Ausgabedatum: 31.03.2023
Abgabedatum: 14.09.2023

Selbstständigkeitserklärung

Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die im Literaturverzeichnis angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder noch nicht veröffentlichten Quellen entnommen sind, sind als solche kenntlich gemacht. Die Zeichnungen oder Abbildungen in dieser Arbeit sind von mir selbst erstellt worden oder mit einem entsprechenden Quellennachweis versehen. Diese Arbeit ist in gleicher oder ähnlicher Form noch bei keiner anderen Prüfungsbehörde eingereicht worden.

Darmstadt, 14.09.2023

Abstract

Many eye diseases cause damage to the visual field leading to an irreversible loss of peripheral and central vision. These changes in the visual field can be found and quantified with the method of visual field testing. However, visual field testing is considered to be error-prone and tedious since the result of the examination is largely depending on the patient's behavior during testing. Another and less error-prone examination for the diagnosis of diseases in ophthalmology is the fundus photography.
This work examines if visual field damages can be predicted on fundus photography with machine learning methods. Furthermore the experiments in this work will include results from optical coherence tomography examinations. For this purpose three machine learning models were trained in order to predict, quantify and locate visual field damages.
The data basis in this work was provided by the Clinic of Ophthalmology from the University Hospital of Münster (UKM) and contains datasets of fundus photography, perimetry and optical coherence tomography.
In order to predict and quantify the visual field damage on fundus photography a CNN model was trained and evaluated on a dataset containing 6126 pairs of fundus photography and perimetry of 2265 subjects. Another dataset containing 1372 pairs of perimetry and optical coherence tomography was used to train and evaluate a regression model to predict and locate visual field sensitivity.
The experiments in this work could not prove a clinical relevance of detecting and quantifying visual field damages just on the basis of fundus photography. The evaluation of the CNN to quantify visual field damages reached an overall accuracy of 56 percent on the test dataset. Through the combination of examination data of the optical coherence tomography the overall model accuracy in quantifying visual field damages could be improved to 86 percent. In predicting and locating visual field damages on examination data of the optical coherence tomography the regression model reached on average a R^2 of 0.56 with a MAE of 4.41 dB.

Zusammenfassung

Viele Augenerkrankungen führen zu Gesichtsfeldveränderungen, die meistens mit einem irreversiblen Verlust des peripheren und zentralen Sehvermögens einhergehen. Dabei werden einzelne Ausschnitte des Blickfeldes schwächer oder gar nicht mehr wahrgenommen. Mit der Gesichtsfeldmessung werden solche Veränderungen im Gesichtsfeld festgestellt und quantifiziert. Die Gesichtsfeldmessung ist allerdings sowohl zeitaufwändig als auch fehleranfällig, da sie von der Mitarbeit des Patienten abhängt. Eine weitere weniger aufwendige und häufig verwendete Diagnosemethode für Erkrankungen ist die Fundusfotografie.
In dieser Arbeit wird untersucht, inwieweit Gesichtsfelddefekte anhand von Fundusfotos und Ergebnissen aus optischen Kohärenztomographien durch Machine Learning Methoden vorhergesagt werden können. Hierfür wurden drei Machine Learning Modelle trainiert, mit denen die Erkennung, Quantifizierung und Verortung von Gesichtsfelddefekten untersucht wurde.
Als Datengrundlage dienten Fundusfotos, Untersuchungsergebnisse aus der Perimetrie sowie der optischen Kohärenztomographie, die durch die Klinik für Augenheilkunde am Universitätsklinikum Münster (UKM) bereitgestellt wurden. Aus diesen Datenquellen wurde für das Training eines CNN, zur Erkennung und Quantifizierung der Gesichtsfelddefekte, ein Datensatz mit 6126 Paare von Fundusfotografien und Gesichtsfeldmessung von 2265 Patienten erstellt. Mit einem weiteren Datensatz aus 1372 Paare von Gesichtsfeldmessungen und der optischen Kohärenztomographie, wurde ein XGBoost Modell für die ortsaufgelöste Vorhersage von Defekten trainiert.
Durch die Experimente in dieser Arbeit konnte nicht bestätigt werden, dass sich Fundusfotografien alleine zur Erkennung und Quantifizierung von Gesichtsfelddefekten eignen. Das CNN, das zur Quantifizierung von Gesichtsfelddefekten trainiert wurde, erreicht eine Accuracy von 56 Prozent auf den Testdaten. Durch die Verknüpfung von Ergebnissen aus den optischen Kohärenztomographien konnte die Modellgüte mit Methoden des Gradient Boosting verbessert werden. Hier erreicht das Modell zur Quantifizierung von Gesichtsfelddefekten eine Accuracy von 86 Prozent.
Für die ortsaufgelöste Vorhersage von Gesichtsfelddefekten erreichte das XGBoost Regressionsmodell ein R^2 von 0.56 bei einem MAE von 4.41 dB.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Schon mit der Erfindung der ersten Computer vor über 80 Jahren, war es das Bestreben Problemlösungs- und Entscheidungsfähigkeiten nach dem menschlichen Vorbild zu lösen. Der Begriff der Künstliche Intelligenz nach heutigem Verständnis, wurde erstmalig in den 1950er Jahren verwendet, um die Simulation der menschlichen Intelligenz durch Maschinen zu beschreiben. [40] Die Grundlage, auf der künstliche Intelligenz Probleme löst oder Entscheidungen fällt, sind programmierte Abläufe oder maschinelles Lernen bei dem mathematische Algorithmen eingesetzt und trainiert werden. Trotz ihrer frühen Ursprünge, ist die künstliche Intelligenz auch heute noch ein anhaltendes und spannendes Forschungsgebiet.
Der Einsatz von künstlicher Intelligenz ist mittlerweile in vielen alltäglichen Situationen vorzufinden, mit der Intension, die Aufgaben des Menschen zu vereinfachen oder auch ganz zu übernehmen. So unterstützen uns beispielsweise Chatbots als ein persönlicher Assistent oder Fahrzeuge können im Verkehr autonom gesteuert werden. In der Industrie wird KI bei der maschinellen Fertigung eingesetzt, um die Produktion möglichst effizient zu gestalten. Auch in der Medizin wurde schon früh der Einsatz von KI erforscht und verprobt. 1972 wurde mit „MYCIN“ ein Expertensystem programmiert, das bei der Erkennung von Infektionskrankheiten unterstützen sollte.[50]
Dabei schaffen es Lernalgorithmen einen Menschen in gewissen Aufgaben teilweise zu übertreffen, wie zum Beispiel im Schachspielen. Bei anderen, für den Menschen sehr einfachen Aufgaben wiederum, scheitern Computer auch heute noch. Zum Beispiel können Lernalgorithmen keine eigene Kreativität entwickeln oder Ursache und Wirkung erkennen. Künstliche Intelligenz löst somit für uns auch nur die Aufgaben, auf die wir sie trainieren.
Aus diesem Grund ist KI ein anhaltendes aktives Forschungsgebiet, um die Lernfähigkeiten von Computern zu verbessern oder auch Aufgaben, die bereits gut lösbar sind, zu entdecken.
So wurden in den vergangene Jahren auch in der Medizin beeindruckende Erfolge mit dem Einsatz von KI erzielt. Zum Beispiel können Machine Learning Verfahren dabei unterstützen, eine Diagnose zuverlässig zu stellen oder die Wahrscheinlichkeit für ein erhöhtest Risiko einer Erkrankung berechnen. Ein aktuelles Beispiel, das auch bereits erfolgreich in der Praxis eingesetzt wird,

ist die Erkennung von Brustkrebs. Im Schnitt wird bei der Auswertung von Mammographien etwa einer aus acht Fällen nicht erkannt. [17] [37] Durch den Einsatz von KI konnte das Risiko einer fälschlicherweise durch den Arzt nicht erkannten Krebserkrankung reduziert werden.[36] Nach diesem Vorbild wird auch in dieser Arbeit untersucht, inwieweit Erkrankungen im Bereich der Ophthalmologie anhand von bildgebenden Verfahren erkannt werden können. Die Ophthalmologie ist in der Medizin die Disziplin, die sich mit dem Aufbau, der Funktionsweise sowie den Krankheitsbildern des Auges beschäftigt.

Wie auch für viele andere Fachärzte in Deutschland, steigen für Augenärzte die Herausforderungen zunehmend. So ist die augenärztliche Versorgung von dem demographischen Wandel in der deutschen Gesellschaft besonders betroffen, da mit steigendem Alter auch die Anzahl an Betroffenen mit Augenerkrankungen zunimmt. [55, S.3] Aufgrund der erhöhten Lebenswahrscheinlichkeit durch den Fortschritt in der Medizin und neuer Diagnose- und Therapieverfahren, kommt es zu einem erhöhten Bedarf ärztlicher Behandlungen. Laut der Deutschen Ophthalmologischen Gesellschaft (DOG) wird davon ausgegangen, dass sich in Deutschland bis 2030 die Anzahl der ophthalmologischen Behandlungen für Menschen von über 60 Jahren, um 35,8 % steigern wird. In absoluten Zahlen ausgedrückt bedeutet dies ein Anstieg um 7,7 Millionen Behandlungsfällen.[55, S.28ff.]
Im Jahr 2019 waren in Deutschland rund 6300 vertragsärztliche Ophthalmologen niedergelassen. Dies entsprach nur rund 4,5% aller Vertragsärzte die 2019 in Deutschland niedergelassen waren.[3] Für jeden Augenarzt wurden im gleichen Jahr durchschnittlich 5014 Behandlungsfälle verzeichnet, rund 50% mehr Fälle gegenüber dem Durchschnitt anderer Ärzte aller anderen Fachbereiche, der bei 3293 Behandlungsfällen lag.[7][4]
Um die ständig wachsende Menge an Patienten und Untersuchungen mit gleichbleibender Qualität bewältigen zu können, ist ein möglicher Lösungsansatz, die Untersuchung und Diagnose mittels computergestützer Verfahren zu unterstützen oder ganz zu automatisieren. Die meisten Untersuchungen beim Augenarzt werden heute bereits mit Hilfe von speziellen computergestützen Maschinen getätigt, die die Messung, Diagnose und fortlaufende Vorsorge und Betreuung der Patienten ermöglicht. Diese helfen dem behandelnden Arzt seine Diagnose zu bestätigen oder zu widerlegen und anhand der Ergebnisse die richtige Therapie für den Patienten zu wählen.

Diese Masterarbeit befasst sich mit der Frage, inwieweit mit Methoden aus dem Bereich des Machine Learnings das periphere Sehen des menschlichen

Auges anhand von Fundusfotos vorhergesagt werden kann.
Das periphere Sehvermögen wird mittel der Gesichtsfeldmessung gemessen und gibt Aufschluss über eventuelle Ausfälle oder Einschränkungen im Gesichtsfeld. Diese Untersuchung ist jedoch aufwändig und fehleranfällig, was es häufig erschwert, die Ergebnisse richtig zu interpretieren oder weitere Untersuchungen notwendig macht.
Die Funduskopie hingegen ist eine häufig genutzte nichtinvasive Methode zur Untersuchung von Merkmalen, die bei verschiedener Augenerkrankungen auftreten. Diese Untersuchungsform ist deutlich schneller und weniger aufwändig sowohl für Arzt als auch Patient und liefert dennoch wertvolle Informationen zur Diagnostik von Augenerkrankungen. [27]

Methoden des Machine Learning, die eine genaue Vorhersage über den Zustand des Gesichtsfeld treffen können, würden in der Praxis dabei unterstützen, die Vorsorge und Kontrolle zu vereinfachen und zeitaufwändige Untersuchen zu reduzieren.
Das Ziel dieser Arbeit ist es, ein Machine Learning Modell zur Erkennung und Quantifizierung des Gesichtsfelds zu entwickeln und auszuwerten. Hierfür werden verschieden Machine Learning Modelle anhand von Fundusfotos, den Ergebnissen der Gesichtsfeldmessung und den der optischen Kohärenztomographie trainiert. Die Daten stammen von der Klinik für Augenheilkunde des Universitätsklinikums Münster (UKM).
Neben der quantitativen Erkennung von Gesichtsfelddefekten, soll in einem zweiten Schritt die Lage von Gesichtsfelddefekten bestimmt werden.

Der Aufbau der Arbeit kann in drei Bereiche untergegliedert werden. Im ersten Teil wird nach der Einleitung mit dem Einstieg in das Thema, in Kapitel 2 die relevanten Hintergründe und theoretischen Grundlagen gegeben. Hier werden die medizinisch Untersuchungen und Krankheitsbilder erläutert, die für die Untersuchungen in dieser Arbeit verwendet wurden. Zusätzlich wird grundlegendes zur Funktionsweise von Deep Learning Modellen erklärt.
Der zweite Themenbereich umfasst die Beschreibung der technologischen Umsetzung. Darunter fallen das Vorgehen zur Durchführung der Experimente, die notwendigen Datenvorverarbeitungsschritte und eine Beschreibung der verwendeten Machine Learning Algorithmen.
Im dritten Themenbereich der Arbeit werden die Ergebnisse der Experimente vorgestellt und zum Schluss das Fazit gezogen.

2. Grundlagen

In diesem Kapitel werden die theoretischen Grundlagen beschrieben.
Zunächst wird auf die Gesichtsfeldmessung eingegangen, durch die das periphere Sehvermögen gemessen wird. Dabei wird insbesondere auf die Ergebnisse der Untersuchung eingegangen, da diese von hoher Bedeutung für die Analysen und Experimente in dieser Arbeit sind. Zusätzlich wird vertiefend auf die Ursache von Gesichtsfelddefekten und die Durchführung einer Gesichtsfeldmessung eingegangen, da diese Aspekte insbesondere zur Motivation der Arbeit beitragen.
Anschließend werden zwei weitere, für die Arbeit relevante medizinische Untersuchungsformen erläutert: die Fundusfotografie und die Messungen der optischen Kohärenztomographie.
Neben den medizinischen Grundlagen wird auch ein Überblick über die relevanten Methoden des Machine Learning gegeben.

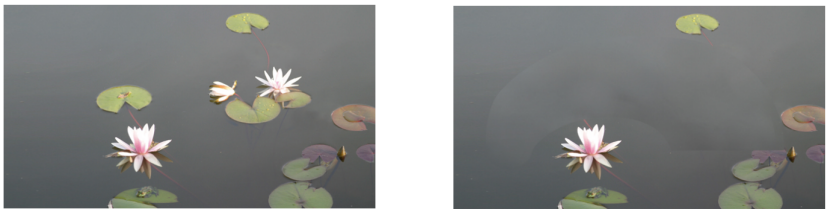
2.1. Gesichtsfeldmessung

Die Gesichtsfeldmessung, auch Perimetrie genannt, ist in der Ophthalmologie eine Untersuchung, bei der das Gesichtsfeld eines Menschen bestimmt wird. Unter Gesichtsfeld versteht man den Bereich des Sehens, der optisch wahrgenommen wird, während das Auge geradeaus einen festen Punkt fixiert.[25] Dies wird auch als das periphere Sehen bezeichnet und ist für die optische Wahrnehmung des Menschen essenziell.
Mit der Perimetrie kann gemessen werden, ob das Gesichtsfeld eingeschränkt ist oder ob Bereiche einen Ausfall aufweisen. Wird in einem Bereich des Auges keine Empfindlichkeit mehr wahrgenommen, so spricht man von einem Gesichtsfelddefekt und das Auge ist an dieser Stelle blind.
Eine Gesichtsfeldmessung ist im Vergleich zu anderen Untersuchungsmethoden recht zeitaufwändig und das Ergebnis und dessen Aussagekraft der Untersuchung hängt sehr stark von der Mitarbeit des Patienten ab. Ist der Patient auf die Untersuchung nicht konzentriert oder hat Verständnisprobleme zum Ablauf der Untersuchung, ist eine falsche Analyse und somit eine fehlerbehaftete Auswertung die Folge. Daher gilt die Untersuchung für den

Patienten häufig als ermüdend und zeitaufwendig. Aktuell ist sie allerdings die einzige Untersuchungsform, die es ermöglicht, die periphere Sehfunktion zu messen.

2.1.1. Ursache von Gesichtsfelddefekten

Die Ursache für Ausfälle im Gesichtsfeld sind meistens degenerative Augenerkrankungen, bei denen Zellen im Auge absterben. Dabei stellt das Glaukom die am häufigsten auftretende degenerative Augenerkrankung dar, die zu einer Einschränkung oder den Verlust des peripheren Sehens führt.
Das Glaukom, im Volksmund häufig auch als „Grüner Star“ bezeichnet, ist die zweithäufigste Erblindungsursache in Deutschland und weltweit.[23, S.11] Fundierten Schätzungen nach gab es im Jahr 2020 weltweit etwa 11,2 Millionen Glaukom-Blinde und etwa 79,6 Millionen an Glaukom erkrankten Menschen, bei denen bereits Gesichtsfelddefekte eingetreten sind. Die Anzahl an Menschen, die an Glaukom erkranken, nimmt mit steigendem Lebensalter zu und beträgt bei über 65 jährigen ca. 2-4%. In Deutschland kommt es durch die Erkrankungen an Glaukom jährlich zu etwa 1.000 Neuerblindungen.[1, S.4]
Bei Glaukom handelt es sich um eine chronisch fortschreitende Augenerkrankung, bei der der Sehnerv irreversibel geschädigt wird. In der Folge, leiden die Betroffenen unter Sehstörungen, die das Gesichtsfeld einschränken. Typischerweise treten die Ausfälle zuerst außerhalb des Fixierpunkts, also dem Ort des schärfsten Sehens, auf.
In der Abbildung 2.1 wird gezeigt, wie ein Schaden im Gesichtsfeld für jemand betroffenen aussehen kann.



(a) Normales Gesichtsfeld (b) Blick mit Gesichtsfelddefekt

Abbildung 2.1.: Vergleich des peripheren Sehvermögens eines Auges mit und ohne Gesichtsfelddefekt. Quelle: [8]

Die visuellen „Lücken“, die dabei entstehen, nimmt die betroffene Person

dabei zu Beginn in der Regel nicht wahr, da diese vom Gehirn „eingefüllt“ werden („filling-in“ Phänomen). [2, S.2] Das tückische bei dieser Erkrankung ist also, dass die Betroffenen selber erst spät diese Gesichtsfeldausfälle wahrnehmen, wenn die Erkrankungen schon weit fortgeschritten ist.
So kann es auch trotz guter Sehschärfe zu Behinderungen im Alltag kommen, was Stürze oder auch eingeschränkte Fahrtauglichkeit zur Folge haben kann. Es ist also als ein erheblicher Einschnitt in den Alltag und in die Lebensqualität der betroffenen Menschen anzusehen. Eine zu späte oder ungenügende Behandlung kann schließlich bis zur Erblindung führen.
Eine Heilung oder Regeneration ist bei dieser Erkrankung nicht möglich, da die Zellen des Sehnervs keine Regenerationsfähigkeit besitzen. Die richtige Behandlung aber durch Medikamente oder auch ein operativer Eingriff ist möglich, um den Sehverlust zumindest aufhalten zu können. [59]
Daher ist eine Früherkennung eben besonders wichtig, um den Fortschritt der Erkrankung quantitativ zu erfassen. Aktuell ist die Gesichtsfeldmessung allerdings nicht Teil von Routineuntersuchungen bei der Vorsorge.[1]

2.1.2. Ablauf und Metriken der Gesichtsfeldmessung

Für die Gesichtsfeldmessung gibt es verschiedene Methoden, wobei die automatische statische Perimetrie die am häufigsten angewandte Methode ist. Diese Art der Untersuchung wird mit einem computergesteuerten Perimeter durchgeführt. Genauer gesagt handelt es sich bei dem Gerät um den *Humphrey Visual Field Analyser*. Bei der Untersuchung wird immer nur ein Auge einzeln untersucht. Für die Messung wird das Kinn und die Stirn des Patienten an dem Gerät fixiert, während er mit seinem Blick einen festen vorgegebenen Punkt fixiert. Über die gesamte Dauer der Untersuchung darf der Blick von diesem Punkt nicht abgewendet werden. Nun erscheinen für den Patienten Lichtsignale auf einem Bildschirm mit variierender Intensität und an unterschiedlichen Stellen. Sobald der Patient ein Lichtsignal wahrnimmt, muss dieser einen Signalknopf drücken um den wahrgenommenen Lichtreiz zu bestätigen. Wird ein Lichtsignal nicht wahrgenommen, so wird die Intensität des Lichtsignal von dem Gerät erhöht. Anhand dieser Methode kann für das Auge die Reizschwelle an verschiedenen Punkten der Netzhaut bestimmt werden.

Die Auswertung mit den Ergebnissen der Untersuchung erhält der Arzt durch die Software des Geräts automatisch als Dokument. In der Abbildung 2.2 wird exemplarisch ein Befund aus einer Gesichtsfeldmessung gezeigt:

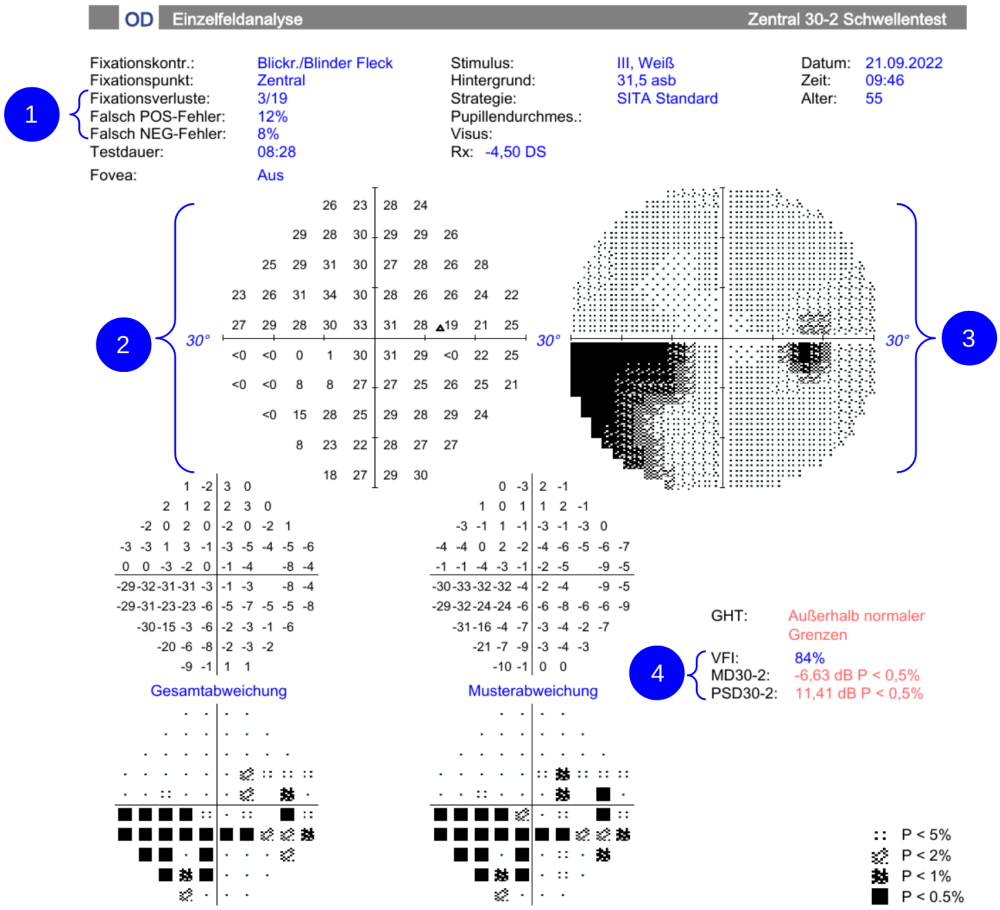


Abbildung 2.2.: Das Ergebnis einer Gesichtsfeldmessung.
1 - Zuverlässigkeitsindizes
2 - Numerische Sensitivitätsgrafik
3 - Graustufen Karte
4 - Gesichtsfeldindizes

Die für diese Arbeit relevanten Metriken und Ergebnisse aus der Perimetrie werden im Folgenden erklärt.

Bei der Messung mit dem *Humphrey Visual Field Analyser* werden in der Ausgabe der Ergebnisse zuoberst drei Metriken ausgegeben, anhand denen die Zuverlässigkeit der Untersuchungsergebnisse ausgewertet werden können. Dies sind der Fixationsverlust, die falsch-positiv Fehlerrate und die falsch-negativ Fehlerrate (vgl. 1 in Abbildung 2.2). Wie in der Einleitung

bereits beschrieben wurde, gilt die Gesichtsfeldmessung als aufwendig und fehleranfällig, da sie stark von der Mitarbeit des Patienten abhängig ist. Mit Hilfe dieser drei Metriken kann der Arzt feststellen, inwieweit die Ergebnisse aussagekräftig sind oder ob die Messergebnisse möglicherweise fehlerbehaftet sind.
Dennoch geben allein die Kennwerte nicht automatisch Aufschluss darüber, ob die Messung und deren Ergebnisse fehlerhaft ist. Daher sind neben den drei Werten auch die vorliegende Erkrankung und deren Auswirkung auf den Patienten in die Auswertung des Ergebnises einzubeziehen. Grundsätzlich aber werden Messungen mit einer falsch-positiven Fehlerrate von über 15% als kritisch betrachtet. Ebenso indiziert ein Ergebnis mit einem Fixationsverlust von über 20% eine wahrscheinlich fehlerhafte Untersuchung.[13] Bei modernen Geräten wird in der Ausgabe des Ergebnises bereits eine Aussage über die Testzuverlässigkeit mit angegeben.

Mit der numerischen Sensitivitätsgrafik wird in der Perimetrie die retinale Sensitivität des Auges an dieser gemessenen Stelle festgestellt (vgl. 2 in Abbildung 2.2). Die Sensitivitätsgrafik bildet eine Karte des Auges ab, durch die anhand der Schwellenwerte aus der Messung festgestellt werden kann, wo sich im Auge die Defekte befinden. Die retinale Sensitivität wird in der Einheit Dezibel (dB) angegeben. Dabei steht Dezibel hier in der Messskala für die Lichtempfindlichkeit. Die Skala reicht von 0 dB, das hellste nicht wahrgenommene Signal, bis zu 51 dB, das dunkelste wahrgenommene Signal. Wenn der Patient also den hellsten Lichtreiz nicht wahrnehmen kann, so wird dies mit <0 dB gewertet. [31] Normalerweise aber reichen die Werte in der Messung der Sensitivität nur bis zu 40 dB. Eine Messung von 40 dB oder größer bedeutet das der Patient „trigger-happy“ ist. Das heißt der Patient gibt an einen Reiz wahrzunehmen, bevor dieser überhaupt auf das Auge ausgeführt wurde. Ein normaler Wert in der Sensitivität liegt eher um einen Wert von 30 dB.[39] Liegt der Wert unter 30 dB indiziert dies einen möglichen Defekt des Gesichtsfelds an dieser Stelle. Die numerischen Werte in der Grafik sind jeweils die Differenz in dB zwischen dem Testergebnis des Patienten und dem alterskorrigierten Normalwert für jeden geprüften Punkt. Somit lässt sich anhand des Werts ablesen, wie intakt oder wie sensitiv das Gesichtsfeld an dieser Stelle ist.

Neben der Sensitivitätsgrafik wird das Gesichtsfeld in einer Graustufenkarte abgebildet (vgl. 3 in Abbildung 2.2). Durch die Schraffierung wird die Schwere des Defekts an der gemessenen Stelle erkenntlich gemacht. Je dunkler die Schraffierung, desto größer ist der Defekt an dieser Stelle. Durch

die Graustufenkarte wird optisch leicht erkennbar, wo und wie groß die betroffenen Areale des Defekts im Auge sind. In dem abgebildeten Befund ist der Defekt also im linken unteren Areal des Auges festzustellen. Neben vorliegenden Defekten, wird in dem untersuchten Augenausschnitt auch die Lage des Sehnervenkopfs deutlich. Dieser ist erkennbar durch die rundliche schwarz schraffierte Fläche im mittleren rechten Augenausschnitt (entsprechend für das rechte Auge). Wäre hier ein linkes Auge untersucht worden, so befände sich die Lage des Sehnervenkopfs auf der linken Seite. Da Lichtreize nur von der Netzhaut wahrgenommen werden können, ist an der Stelle, an der der Sehnervenkopf liegt, keine Reaktion zu erwarten. Daher ist hier von keinem Defekt die Rede.

Als Zusammenfassung des Befunds werden drei Indizes ausgegeben (vgl. 4 in Abbildung 2.2).
Die Mean Deviation (MD) ist dabei für die Untersuchung der Fragestellung in dieser Arbeit von zentraler Bedeutung. Sie wird als Metrik verwendet, um den Fortschritt einer Glaukomerkrankung und dem damit einhergehenden bereits eingetretenen Gesichtsfeldverlust zu messen. Der Mean Deviation Wert wird in Dezibel (dB) angegeben und berechnet sich aus der durchschnittlichen Abweichung des normal zu erwartenden dB-Wert einer Person der gleichen Altersgruppe. Dabei werden bei der Berechnung Punkte im Zentrum des Sehens stärker gewichtet, als Regionen die im peripheren Bereich liegen.[31] Ein MD-Wert von 0 dB gilt als normal, während Werte im positiven Bereich über dem Durchschnitt liegen. Negative MD-Werte bedeuten einen Verlust des Gesichtsfelds.
Bei der Mean Deviation handelt es sich also um einen Durchschnittswert für das gesamte Auge. Dieser Wert beinhaltet daher keine Aussagekraft darüber, wo im Auge die Defekte auftreten. Der Wert erlaubt jedoch eine Vergleichbarkeit und wird in der Regel von den Ärzten als Referenzwert für die Vorsorge verwendet. Bei regelmäßigen Untersuchungen kann über die Mean Deviation am besten festgestellt werden, ob eine Verschlechterung des Gesichtsfelds zu beobachten ist.

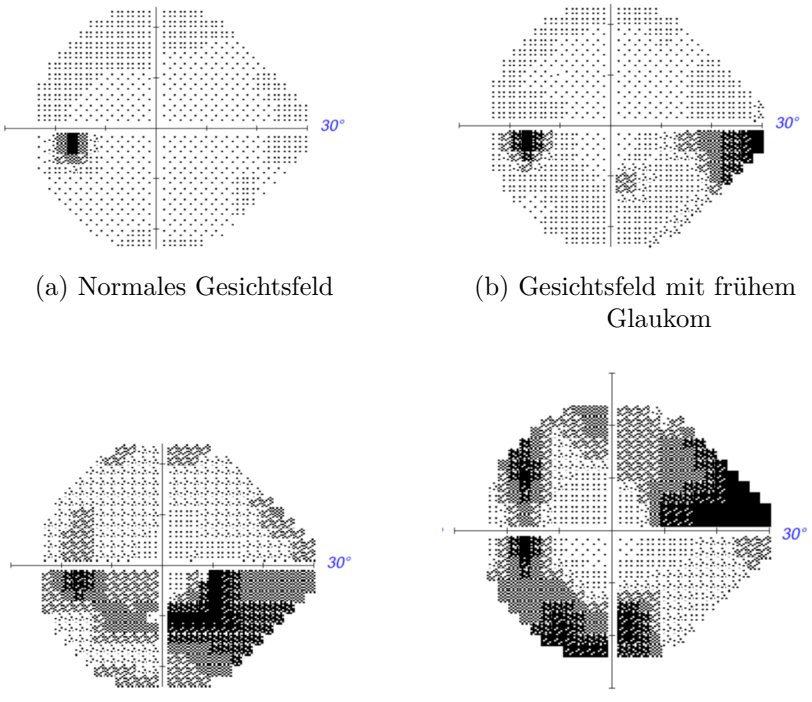
2.1.3. Einstufung in Hodapp-Stadien

In der Augenheilkunde wird die Schwere eines vorliegenden Gesichtsfeldverlusts aufgrund von Glaukom in Stadien nach Hodapp-Anderson-Parrish (Hodapp-Stadien) eingestuft. Diese Einteilung in Hodapp-Stadien orientiert sich an der Mean Deviation, die in der Gesichtsfeldmessung festgestellt wird.

[19] Es werden dabei drei Stadien definiert:

- MD zwischen 0 dB und -6 dB: milde Form (Stadium 1)
- MD zwischen -6 dB und -12 dB: moderate Form (Stadium 2)
- MD von mehr als -12: schwere Form (Stadium 3)

Diese Einteilung wird für diese Arbeit bei der Durchführung der Experimente relevant. Die Stadien dienen zur Einteilung in Klassen, mit denen das neuronale Netz zur Klassifikation trainiert wird.
Zur Veranschaulichung ist in Abbildung 2.3 dargestellt, wie ein Defekt des Gesichtsfelds je nach Stadium aussieht.



(a) Normales Gesichtsfeld

(b) Gesichtsfeld mit frühem Glaukom

(c) Gesichtsfeld mit moderatem Glaukom

(d) Gesichtsfeld mit schwerem Glaukom

Abbildung 2.3.: Die verschiedenen Stadien eines Gesichtsfelddefekts bei einer Glaukomerkrankung. Quelle: [59]

2.2. Fundusfotografie

Bei der Fundusfotografie handelt es sich um eine bildgebende Untersuchungsmethode, bei der der Hintergrund des Auges aufgenommen wird. In der Fachsprache wird die Aufnahme des Augenhintergrunds auch Ophthalmoskopie genannt. Die Aufnahme wird mit einem speziellen Gerät, einer Funduskamera oder auch Ophthalmoskop, gemacht. Mit der Funduskopie können durch den Arzt die Netzhaut und der Sehnerv genauer untersucht werden. [18]
Die Netzhaut, auch Retina genannt, sitzt an der Rückseite des Augapfels und ist die innere Oberfläche des Auges. Sie ist ein sehr komplexes Nervengewebe und kann die einfallenden Lichtstrahlen in Nervensignale umwandeln. Diese Signale werden schließlich als Informationen über den Sehnerv an das Gehirn weitergeleitet. Über den Sehnervenkopf, auch als Papille bezeichnet, verlassen die vielen Nervenfasern das Auge und führen zum Gehirn.

Zur Veranschaulichung ist in Abbildung 2.4 der anatomische Aufbau des Auges gezeigt, an dem die Lage der Netzhaut, des Sehnervenkopfs sowie der Sehnerven betrachtet werden können.

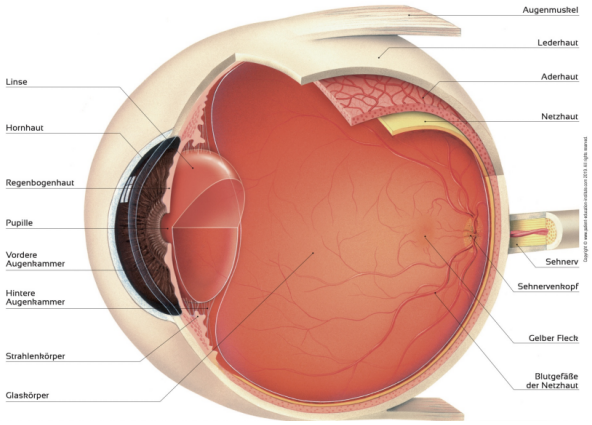


Abbildung 2.4.: Anatomisches Schema des Auges. Quelle: [6]

Die Fundusfotografie ist nicht aufwändig und zählt daher zu den häufig durchgeführten Untersuchungen im Praxisalltag. Die Aufnahme verschafft eine Momentaufnahme des Gesundheitszustands des Auges und hilft bei regelmäßig durchgeführten Kontrolluntersuchungen, die Veränderungen am Auge festzustellen. Denn bei den meisten degenerativen Augenerkrankungen sind zuerst der Sehnerv oder auch die Netzhaut betroffen. Somit verfügen die Ärzte in den meisten Fällen über eine Vielzahl von Fundusfotografien

eines Patienten, die den Verlauf der Erkrankung über viele Jahre hinweg dokumentiert. [22]

In der Abbildung 2.5 wird im linken Bild ein Fundusbild gezeigt. Im rechten Bild ist der Ausschnitt um den Sehnervenkopf zu sehen. Letzteres wird später für das Training des neuronalen Netzes verwendet.

(a) Fundusfotografie

(b) Ausschnitt um den Sehnervenkopf

Abbildung 2.5.: Beispiel einer Fundusfotografie

2.3. Retinale Nervenfaserschicht

Die retinale Nervenfaserschicht, kurz RNFL (aus dem Englischen für *Retinal Nerve Fiber Layer*), ist die innerste Schicht der Netzhaut. Durch die Messung der Nervenfaserschicht, erhält man schon frühe Erkenntnisse über Defekte und Erkrankungen des Auges. Daher ist sie ebenfalls eine sehr häufig durchgeführte und wichtige Untersuchung für die Früherkennung und Verlaufskontrolle.
Die Messung der RNFL wird mittels der optischen Kohärenztomographie, kurz OCT (aus dem Englischen für *Optical Coherence Tomography*),vorgenommen. Bei einem OCT wird ein Querschnitt der Netzhaut erstellt, anhand dessen die Dicke der Nervenfaserschicht gemessen wird.
Mit zunehmenden Alter nimmt die Dicke der Nervenfaserschicht ab. Aber auch durch Erkrankungen kann es zu einem Verlust von Nervenfasern kommen. Diese sind anhand eines OCT-Scans deutlich detaillierter und früher zu erkennen, als auf einem Fundusbild.[9, S.8]

In Abbildung 2.6 sind Auszüge einer Untersuchung der Nervenfaserschicht durch eine OCT-Aufnahme gezeigt.

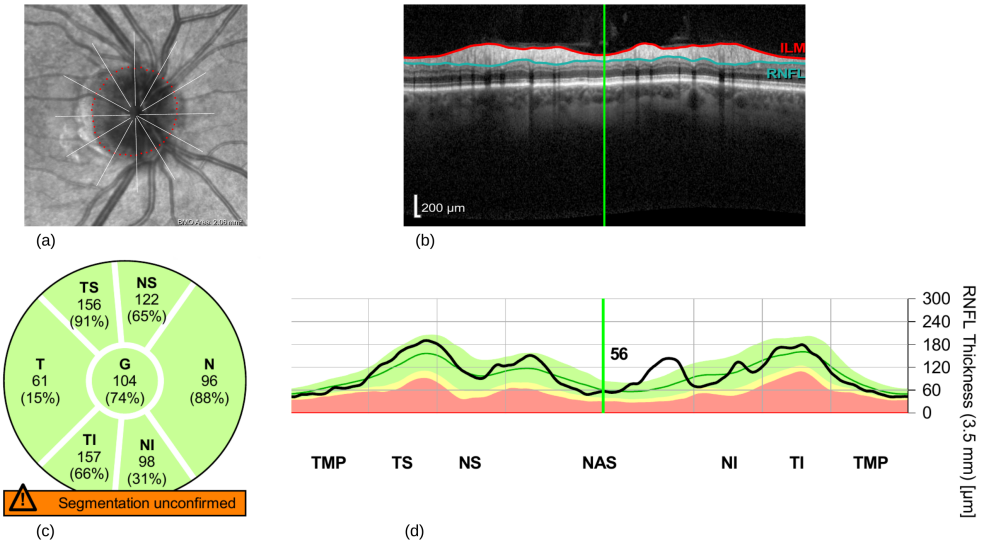


Abbildung 2.6.: OCT-Scan der Nervenfaserschicht. (Auge: OD)
(a) Fundusfoto mit Ausschnitt um den Sehnervenkopf (Papille). Der rot gepunktete Kreis zeigt die Messstellen der Nervenfasern um die Papille herum. (b) OCT-Scan mit dem Querschnitt der Nervenfaserschicht um den Sehnervenkopf gemessen (wird durch den Algorithmus automatisch erkannt und in blau eingezeichnet). (c) Durchschnittliche RNFL Dicke für die jeweiligen peripapillären Sektoren und dem globalen Mittelwert. (d) Profil der gemessenen RNFL Schicht um den Sehnervenkopf.

Bei der Auswertung der Untersuchung sind das Profil und die Werte der durchschnittlichen RNFL Dicke (vgl. (c) und (d) in Abbildung 2.6) von besonderem Interesse. Anhand dieser Werte kann eindeutung festgestellt werden, ob die gemessenen Nervenfasern innerhalb normaler Grenzen liegen oder ob eine Abweichung zum Normwert vorliegt. Ob eine Messung im Bereich normaler Grenzen, grenzwertig oder außerhalb normaler Grenzen liegt, wird durch das System mit einem Algorithmus festgestellt. Dabei werden die Werte mit einer normativen Datenbank altersabhängig abgeglichen. [34] Veranschaulicht wird dies in der Auswertung auch über eine Farbskala. Grün entspricht eine RNFL Dicke innerhalb normaler Grenzen, gelb gilt als grenzwertig und rot für außerhalb normaler Grenzen.
Bei der durchschnittlichen RNFL Dicke um die Papille, werden die gemessenen Werte als Kreisdiagramm dargestellt. In der Mitte wird der Gesamt-

durchschnitt Global (G) angezeigt. Darum herum erfolgt die Einteilung der peripapillären Sektoren in Temporal (T), Temporal-Superior (TS), Temporal-Inferior (TI), Nasal (N), Nasal-Superior(NS) und Nasal-Inferior (NI). In den jeweiligen Sektoren, werden die durchschnittlichen Dicken der RNFL für diesen peripapillären Sektor in der Einheit Mikrometer angebenden.

2.4. Minimale Randsaumbreite der Bruch'schen Membranöffnung

Neben der RNFL kann mittels der optischen Kohärenztomografie die minimale Randsaumbreite der Bruchschen Membranöffnung (engl. BMO-MRW) bestimmt werden. Die Messung der BMO-MRW ist ebenfalls ein fester Bestandteil in der klinischen Routine. [38] Die BMO-MRW ist ein Parameter in der Diagnostik von Glaukomerkrankungen und drückt die Dicke des retinalen Nervenfaserpolsters an der Bruch'schen Membranöffnung aus.[24]

In Abbildung 2.7 wird der Auszug einer OCT-Aufnahme zur Vermessung der BMO-MRW gezeigt.
Zur Bestimmung der BMO-MRW (ff. als MRW bezeichnet) wird der minimale Abstand zwischen der innerer limitierender Membran und der Bruch'schen Membranöffnung zur Papille gemessen.[54]

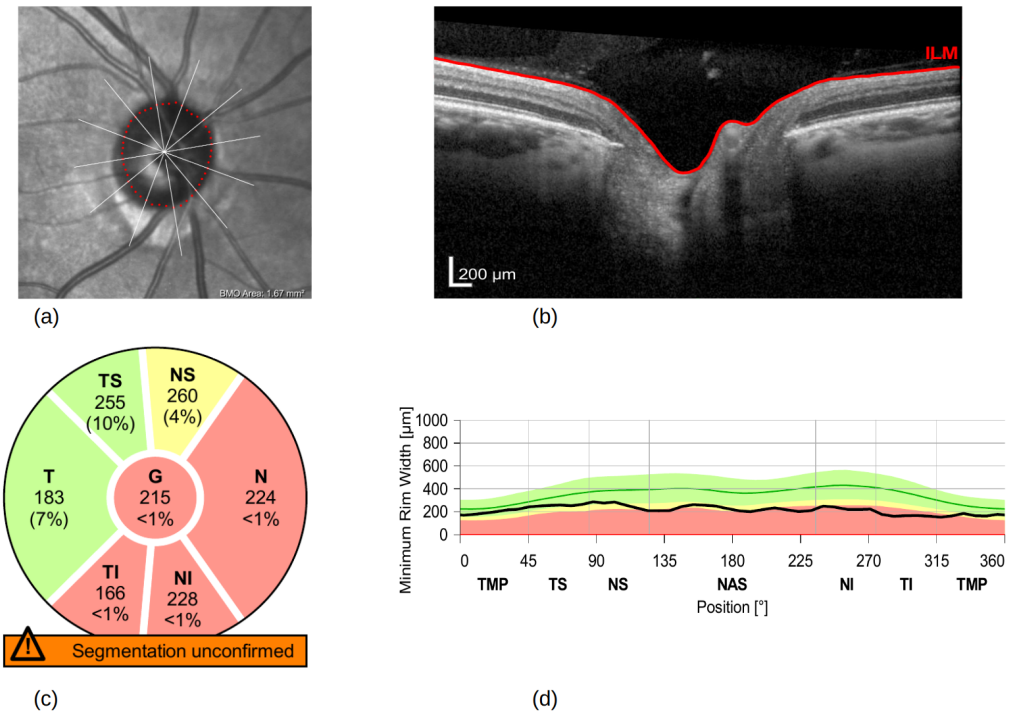


Abbildung 2.7.: OCT-Scan der BMO-MRW. (Auge: OD)
(a) Fundusfoto mit Ausschnitt um den Sehnervenkopf (Papille). Der rot gepunktete Kreis zeigt die Messstellen der Nervenfasern um die Papille herum. (b) OCT-Scan mit dem Querschnitt der Bruch'schen Membranöffnung und der Membrana limitans interna (ILM). (c) Gemessene BMO-MRW in den vordefinierten Sektoren. (d) Profil der gemessenen BMO-MRW.

In mehreren Studien konnte gezeigt werden, dass die MRW zur Erkennung und Diagnostik von Glaukom beitragen.[57][20] Aus diesem Grund werden auch die Werte der MRW in die Experimente der Arbeit einbezogen und untersucht, inwieweit sie eine Auswirkung auf die Vorhersage der Gesichtsfelddefekte haben. Die MRW wird, wie auch bei der RNFL, in den sechs Sektoren Temporal (T), Temporal-Superior (TS), Temporal-Inferior (TI), Nasal (N), Nasal-Superior(NS) und Nasal-Inferior (NI) angegeben (vgl. (c) in Abbildung 2.6). Zusätzlich wird ein globaler Mittelwert der MRW angegeben. Die Abstände zur Membran werden in der Einheit Mikrometer angegeben.

2.5. Deep Learning

Das Deep Learning ist eine Methode innerhalb des Machine Learnings, bei der große Datenmengen durch den Einsatz künstlicher neuronaler Netze verarbeitet werden. Neuronale Netze fallen in die Kategorie des *Supervised Learnings*. Diese Gruppe von Machine Learning Algorithmen, umfasst jene, bei der die Eingabevektoren der Trainingsdaten einem Zielvektor zugeordnet werden können. Der Algorithmus lernt während des Trainings sozusagen die Regeln, nach denen Muster in Daten erkannt werden. [11, S.3]
Ein berühmtes Beispiel für die Anwendung eines *Supervised Learning* Algorithmus ist die Erkennung von handschriftlichen Zahlen. Hierbei werden Bilder von handschriftlichen Zahlen im Eingabevektor an ein neuronales Netz übergeben, welches eine Zuordnung jeden Bildes zu eine Zahl zwischen Null und Neun treffen soll. Die Zuordnung einer Zahl in eine Kategorie ist im Vorfeld schon bekannt. Dem Lernalgorithmus wird also neben dem Bild im Eingabevektor auch die Information übergeben, um welche Zahl es sich auf dem Bild handelt.
Die Aufgabe der Bilderkennung durch den Lernalgorithmus kann als Funktion $y(x)$ ausgedrückt werden, die es zu approximieren gilt. Der Eingabevektor in Form der Bilder ist dabei x und generiert wird ein Ausgabevektor y.
Nach Abschluss des Trainings, kann der Lernalgorithmus auf ihm unbekannten Testdaten die Zahlen zuordnen. Diese Eigenschaft, gelerntes anhand von Trainingsdaten auf neuen, noch unbekannten Daten zu abstrahieren und anzuwenden, nennt man *Generalisierung* und ist ein zentrales Ziel bei der Mustererkennung. [11, S.2 ff.]
Das Approximieren der Funktion $y(x)$, nach der ein Algorithmus die Entscheidung zur Zuordnung in bestimmte Kategorien oder Klassen trifft, ist die Komponente die beim *Supervised Learning* zu Beginn unbekannt ist und bestimmt werden soll.
Im folgenden Abschnitt wird ein Überblick über die Methodik, die Architektur und den grundlegenden mathematischen Funktionen von neuronalen Netzen gegeben. Zur Erklärung wird ein simples Feedforward Netz mit zwei Schichten verwendet.

2.5.1. Grundlagen von Neural Networks

Neuronale Netze sind nichtlineare statistische Modelle, die typischerweise in einem Netzwerk-Diagramm dargestellt werden, wie in Abbildung 2.8. Das Netz besteht aus einer Vielzahl von miteinander verbundenen Schichten und

Knoten, auch Neuronen genannt.
Die grundlegende und simpelste Form eines neuronalen Netz ist das mehrschichtige Feed-Forward-Netz. Es besteht aus einer Eingabeschicht (engl. *Input Layer*), einer oder mehreren versteckten Schichten (engl. *Hidden Layer*) und einer Ausgabeschicht (engl. *Output Layer*). Jede Schicht kann aus einer unterschiedlichen Anzahl von Neuronen bestehen und ist vollständig mit der benachbarten Schicht verbunden.
Die erste Schicht eines neuronalen Netzwerks ist die Eingabeschicht und nimmt die Eingabewerte auf. Dies sind die unabhängigen Variablen oder auch Merkmale des Modells. Daher bestimmt auch die Anzahl an unabhängigen Variablen des Datensatzes die Anzahl an Neuronen im Input Layer.
Darauf folgt eine oder mehrere Hidden Layer, bestehend aus einer bestimmten Anzahl von Neuronen. Die Hidden Layer liegen zwischen Input und Output Layer und transportieren die Informationen durch das Netzwerk. Über die Verbindungen zum Input Layer extrahieren die Hidden Units die Rohdaten aus dem Input Layer.[46, Kap.2]
In der Ausgabeschicht liefert das Modell die Vorhersage des zu lösenden Problems. Bei einem Klassifikationsproblem wäre dies die Wahrscheinlichkeit für die Zugehörigkeit zu einer Klasse oder ein kontinuierlicher reeler Wert bei einem Regressionsproblem. Die Anzahl der Neuronen in der Ausgabeschicht entspricht bei einem Klassifikationsproblem der Anzahl an Klassen, in die eingeteilt werden sollen.

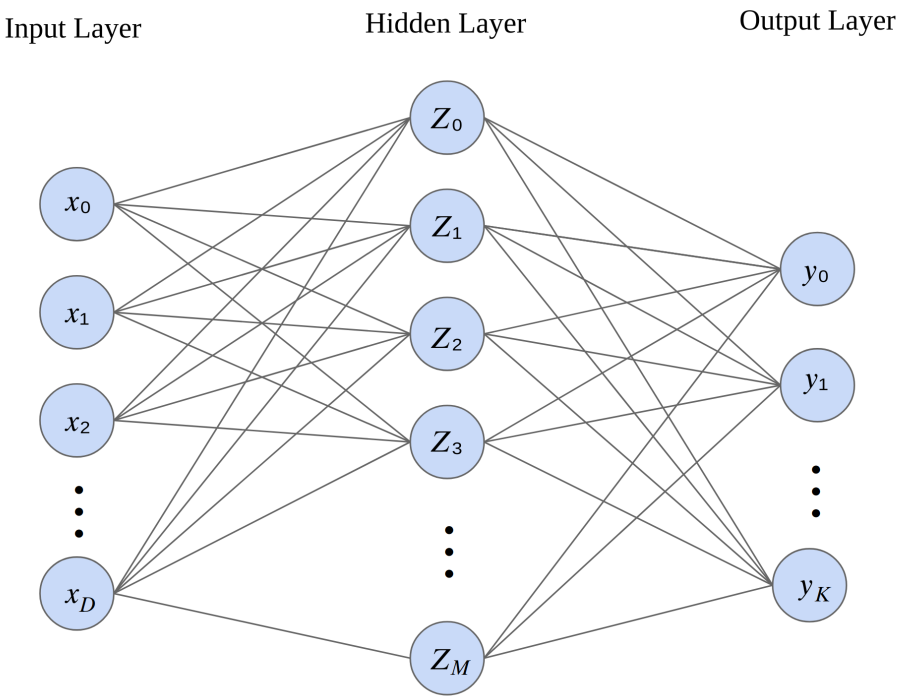


Abbildung 2.8.: Architektur eines simplen Feed-Forward Neural Network mit einem Hidden Layer.

Vereinfacht formuliert, ist das Ziel eines neuronalen Netzwerks die Approximation einer nichtlinearen Funktion, die das vorliegende Problem bestmöglich beschreibt. Dieser iterative Lernprozess setzt sich aus fünf aufeinanderfolgenden Schritten zusammen:

- Initialisierung der Gewichte und der Bias Parameter
- Forward Propagation
- Berechnung der Verlustfunktion
- Backpropagation
- Update der Gewichte und der Bias Parameter

Zu Beginn des Lernalgorithmus, werden die Gewichte w zwischen den Neuronen zufällig initialisiert. Die Gewichte an den Verbindungen der einzelnen Neuronen sind sozusagen ein Informationsspeicher, über die das Netzwerk relevantes Wissen erkennt und merkt. Sie können als Koeffizienten verstanden

werden, die das Eingangssignal für ein bestimmtes Neuron im Netz so skalieren, dass eine möglichst genau Vorhersage auf Grundlage der Trainingsdaten getroffen wird. Über das iterative Anpassen der Gewichte lernt das Netz neue Informationen.[46, Kap.2]

Bei der Forward Propagation werden die Daten aus dem Input Layer von Schicht zu Schicht durch das neuronale Netz propagiert. Dabei berechnet sich das Modell als eine Linearkombination, die auf einer festen nichtlineare Basisfunktion $\phi_j(\mathbf{x})$ basiert. Im Training wird durch anpassen der Basisfunktion das Modell optimiert.[11, S.227]
Der Lernprozess ist abhängig von der *Aktivierung* der Neuronen. Dieser Mechanismus bestimmt, ob Informationen durch das Netzwerk propagiert werden. Ob ein Neuron aktiviert wird oder nicht, ist abhängig vom Ergebnis der Aktivierungsfunktion. Der Wert ergibt sich aus der Transformation der Kombination aus Input, Gewicht und Bias und liegt je nach Funktion im Wertebereich von 0 bis 1 oder -1 bis 1.[46, Kap.2]
Eine häufig genutzte Aktivierungsfunktion ist die Sigmoid-Funktion:

$$\sigma(a) = \frac{1}{1 + exp(-a)} \tag{2.5.1}$$

mit

$$a = \sum_{i=1}^{n} w_i x_i + b \tag{2.5.2}$$

wobei n als die Anzahl der Neuronen der vorherigen Schicht,
w als der Gewichtsvektor,
x als der Eingabevektor und
b als der Bias definiert ist.[11, S.228]

Eine weitere, insbesondere für Klassifikationsmodelle mit mehreren Klassen relevante Aktivierungsfunktion ist die *Softmax* Funktion. Die Softmax-Aktivierungsfunktion liefert die Wahrscheinlichkeitsverteilung für die Zuordnung zu unabhängigen Klassen in der Ausgabeschicht.[46, Kap.2]
Bei Klassifikationsproblemen wird die Softmax-Funktion üblicherweise in der letzten Schicht des neuronalen Netzes angewendet, um die Wahrscheinlichkeit zu einer Klassenzuordnung berechnen zu können.
Erst durch das Hinzufügen einer Aktivierungsfunktion kann Nichtlinearität in dem Modell abgebildet werden. Das Produkt dieser Transformation ist die Eingabe für die nächste Schicht im neuronalen Netz.[46, Kap.2]

Das Kernziel des Trainings ist die Anpassung der Gewichte und des Bias an jedem Neuron, bis ein Optimum gefunden wird. Idealerweise bewirken die Gewichte eine Verstärkung des Signals und dämpfen dagegen Rauschen in den Daten. Je größer der Wert eines Gewichts, desto größer kann die Korrelation zwischen einem Signal und dem Ergebnis des Netzwerks gedeutet werden.[46, Kap.2]
Das Finden des Optimums passiert typischerweise über das Minimieren der Fehlerfunktion. Die Fehlerfunktion ist eine Methode, um zu evaluieren, wie gut das Modell auf den Daten Vorhersagen treffen kann. Der Fehler oder auch Verlust wird ermittelt, indem die vom Netz getroffene Vorhersage in der Ausgabeschicht mit dem wahren Ergebnis verglichen wird. Für Klassifikationsprobleme mit K Klassen berechnet sich der Fehler über die Kreuzentropie-Fehlerfunktion wie in 2.5.3 gezeigt.

$$E(\mathbf{x}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\, ln\, y_k(\mathbf{x}_n, \mathbf{w}) \tag{2.5.3}$$

Dabei ist E der Fehler oder auch Verlust,
N die Anzahl der Eingabevektoren,
K die Anzahl der Klassen,
t_k die Zielwerte und
y_k die berechneten Ausgabewerte des Modells.[11, S.235]

Das Minimieren der Fehlerfunktion $E(\mathbf{x})$ wird mit dem Optimierungsalgorithmus der *Backpropagation* erreicht. Der Backpropagation Algorithmus ist eine Sonderform des Gradientenabstiegverfahrens, dessen Ziel es ist, die Fehlerfunktion durch anpassen der Gewichte und des Bias zu minimieren. Bei diesem Verfahren wird der Gradient berechnet, um ein lokales Minimum der Funktion zu finden.[28, S.396] Je kleiner der Verlust, desto besser ist das Modell. Bei der Backpropagation werden also die berechneten Fehler „rückwärts“ durch das Netz von Schicht zu Schicht transportiert, um durch Anpassung der Gewichte und Bias, die Vorhersagegüte des Modells zu verbessern. Dieser Vorgang wird für mehrere Epochen oder bis zur Konvergenz durchgeführt. Epochen drückt eine Anzahl von Wiederholungen aus.
Nach Abschluss des Trainings kann schließlich die Inferenz des Modells mit neuen unbekannten Daten erfolgen, um eine Vorhersagen oder Klassifizierungen zu treffen.

2.5.2. Convolutional Neural Networks

Convolutional Neural Networks (CNNs), sind eine spezielle Form neuronaler Netze, die besonders geeignet sind, um Muster in Bildern zu erkennen und in Klassen oder Kategorien einzuteilen.
CNNs ermöglichen die Verarbeitung einer hohen Menge an komplexen Daten, die eine rasterähnliche Topologie aufweisen. Beispiele für derartige Daten sind Zeitreihendaten, die ein 1-D-Raster mit Stichproben in regelmäßigen Zeitabständen darstellen oder Bilddaten, die ein 2-D-Raster aus Pixeln darstellen.[26, Kapitel 9]
Das erste CNN das der heutigen Anwendung noch als Vorbild dient, wurde 1989 in der Arbeit von Yann LeCun et al. vorgestellt.[35] In der Arbeit wurde ein CNN zur Bilderkennung erstmals mithilfe des Backpropagation Algorithmus trainiert. Die Bilderkennung bezog sich auf die Erkennung von handschriftlichen Zahlen, ein auch heute noch sehr beliebtes Beispiel, welches auch schon in der Einleitung dieses Kapitels angeführt wurde.
Im folgenden Abschnitt werden die elementaren Bestandteile eines CNN und die wesentlichen Erweiterungen zu einem klassischen neuronalen Netz erklärt.

Ein CNN setzt sich typischerweise aus drei Schichten zusammen: dem *Convolutional Layer*, dem *Pooling Layer* und dem *Fully Connected Layer*.[26, Kapitel 9] Der typische Aufbau eines CNN ist auch in der Abbildung 2.9 Veranschaulicht.

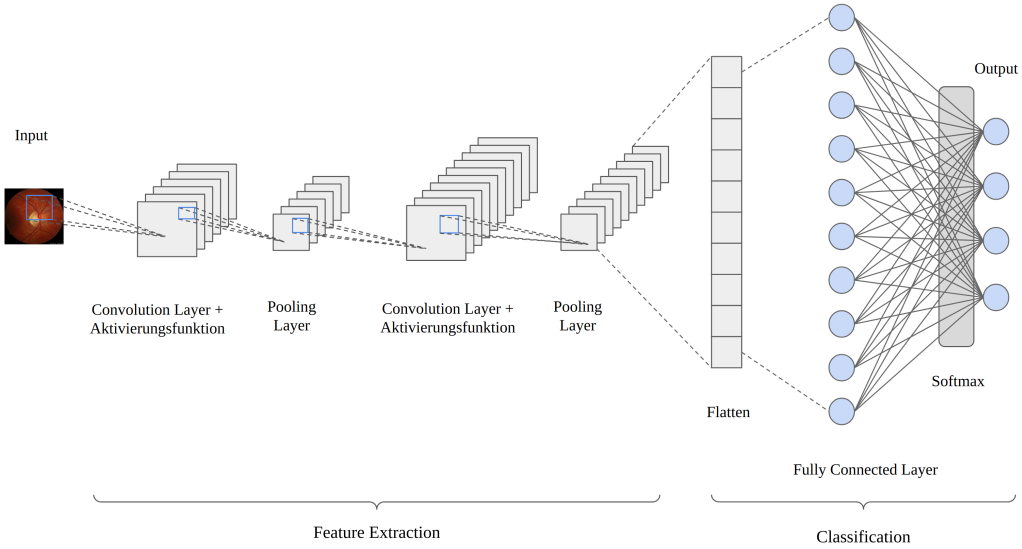


Abbildung 2.9.: Veranschaulichte Architektur eines Convolutional Neural Network.

Das besondere an CNNs im Vergleich zu neuronalen Netzen ist, dass CNNs mindestens in einer Schicht des Netzes eine sogenannte Faltung anstelle der allgemeinen Matrizenmultiplikation einsetzen. Faltung bedeutet im Englischen *convolution*, daher auch der Name Convolutional Neural Networks.
Die Faltung passiert in den Convolutional Layer und sorgt dafür, dass Merkmale aus einem Bild über den sogenannten Kernel extrahiert werden. Der Kernel kann auch als ein Filter bezeichnet werden und ist eine rechteckige Matrix einer festen Größe. Über den Kernel wird eine räumlich kleinere Abstraktion des Eingabebilds erzeugt, die aber detaillierte Informationen der Eingabedaten enthält. Diese Repräsentation wird auch *Feature Map* bezeichnet und fasst in ihr die Merkmale aus den Input Daten zusammen.[11, S.269 ff.]
Im Anschluss an die Faltung wird auf die Ausgabe des Kernels eine Aktivierungsfunktion angewendet. Hier wird häufig die ReLU (Rectified Linear Unit) Funktion verwendet, durch die negative Werte auf Null gesetzt werden und positive Werte unverändert bleiben.[26, Kapitel 9]
Darauf folgt eine weitere entscheidende Komponente in CNNs, der Pooling Layer. Er bewirkt, dass die Größe der Matrix reduziert wird, indem semantisch ähnliche Feature zusammengefasst werden. Die Pooling-Operation wird auf jedem Ausschnitt der Feature Map einzeln durchgeführt. Zur Berechnung wird am häufigsten das Max-Pooling verwendet, bei dem nur die Merkmale mit maximaler Größe einer Matrix beibehalten werden.[26, Kapitel 9] Der

Stride definiert die Schrittgröße, die je Verschiebung der Matrix über die Feature Map gegangen wird.

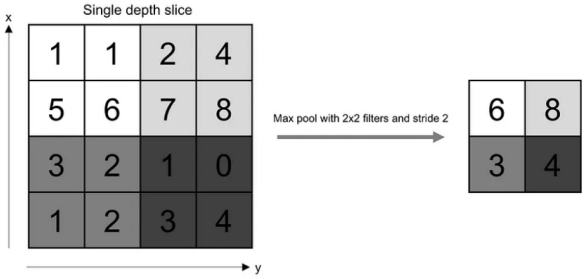


Abbildung 2.10.: Beispiel einer Max-Pooling Operation über einen Ausschnitt der Feature Map und eines 2x2 Kernels [42]

In dem Fully Connected Layer sind alle Neuronen einer Schicht vollständig mit allen Neuronen der vorangehenden und nachfolgenden Schicht verbunden. Die Fully Connected Layer in einem CNN ermöglichen die Zuordnung von detektierten Feature zu Ausgabewerten zwischen dem Input und Output Layer.
Durch den Aufbau von mehreren aufeinander folgenden Convolutional und Pooling Layer erreichen CNNs eine immer bessere Repräsentation von Feature die aus den Eingabedaten extrahiert werden. Die Anzahl von aufeinanderfolgenden Convolutional Layer und Pooling Layer kann beliebig oft wiederholt werden.[46, Kap.2]

2.5.3. Gradient Boosting

Das Gradient Boosting fällt in der Statistik unter die Kategorie des *Ensemble Learnings* und beschreibt im Machine Learning ein Verfahren, welches in Regressions- und Klassifikationsmodellen verwendet wird, um Entscheidungsbäume aufzubauen. Die Idee des Ensemble Learning besteht darin, mehrere schwache Modelle zu einem stärkeren Modell zu kombinieren, welches in seiner Genauigkeit besser als die einzelnen Modelle ist. Diese schwächeren Lernmethoden zeichnen sich typischerweise durch eine schlechtere Vorhersage aus und bauen damit sogenannte flache Entscheidungsbäume auf. Das Verfahren kombiniert durch seinen Algorithmus die einzelnen flachen Bäume zu einem starken Entscheidungsbaum, dessen Vorhersagegenauigkeit höher liegt. Generell verfolgen Boosting-Verfaren den Ansatz, schwächere Lernalgorithmen sequenziell auf einen Datensatz, mit unterschiedlichen Teilmengen

der Trainingsdaten pro Durchlauf (genannt "Stage"), zu trainieren und deren Ergebnisse zu einer resultierenden Vorhersage zu kombinieren. [28] Als "schwächer" wird hierbei ein Klassifikator bezeichnet, der nur wenig besser als zufälliges Raten ist. Der resultierende Boosted Tree ist bei diesen Verfahren die Summe der einzelnen Entscheidungsbäume. [28]
Beim Gradient Boosting wird dieses Vorgehen erweitert, indem das Resultat des Boostings nicht nur als Summe sequentieller Entscheidungsbäume, sondern als numerisches Optimierungsproblem gesehen wird. Hierbei wird von einer Verlustfunktion über alle Stages ausgegangen, welche sich analog zu den schwachen Entscheidungsbäumen als Summe der Verlustfunktionen der einzelnen Stages zusammensetzt. Jede Stage bzw. dessen Baum wird hier als Funktion $f(x)$ verstanden, die einen Wert y auf den Trainingsdaten vorhersagt und dessen Verlust $L(y, f(x))$ es zu minimieren gilt. [21] Wie der Name besagt, wird hierzu mittels des Gradientenverfahrens eine Optimierung angewendet, um über die Minimierung des Verlustes eines Klassifikations- bzw. Regressionsmodells nach jeder Stage, die Vorhersage der folgenden Stage zu verbessern.[28]

2.5.4. Metriken

Für die in Kapitel 5 vorgestellten Ergebnisse werden zur Evaluierung der durchgeführten Experimente gängige Metriken genutzt. Zum besseren Verständnis und zur Vollständigkeit werden im folgenden Abschnitt die genutzten Metriken erläutert.
Zur Auswertung von Klassifikationsproblemen wird als Grundlage eine Confusion Matrix je Modell aufgestellt. Die Confusion Matrix (siehe Abbildung 2.11) ist eine Tabelle mit Zeilen und Spalten, die die Vorhersage des Modells und die tatsächlichen Werte für eine Klassifikation darstellt. Als *True Positive* werde die Datenpunkte beschrieben, die richtig in die Klasse K zugeteilt werden. Als *True Negative* werden die Zuteilungen eingestuft, die richtigerweise einer anderen Klasse angehören. Dagegen werden als *False Positive* die Datenpunkte bezeichnet, die in eine Klasse eingeteilt wurden, tatsächlich aber einer anderen Klasse angehören. Als *False Negative* gelten die Datenpunkte, die eigentlich zu der Klasse gehören aber nicht richtig zu dieser eingeteilt wurden. [46, Kap.1]
In der Statistik wird üblicherweise ein falsch positives Ergebnis auch als „Fehler 1. Art“ und ein falsch negatives Ergebnis als „Fehler 2. Art“ bezeichnet.

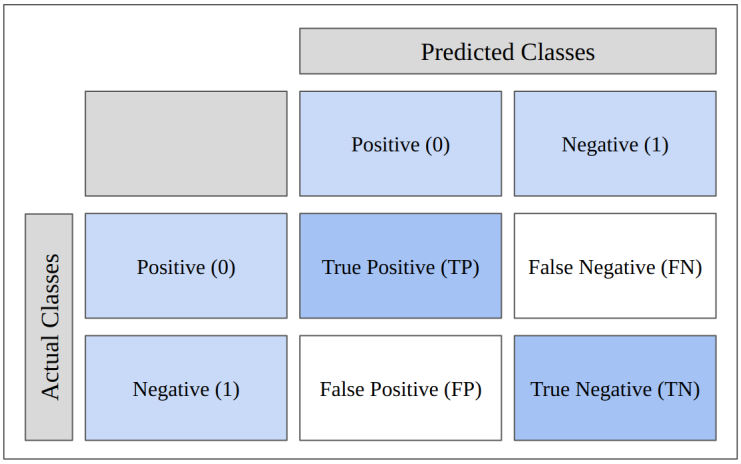


Abbildung 2.11.: Schema einer Confusion Matrix mit zwei Klassen.

Das Aufstellen der Confusion Matrix dient als Grundlage für eine Vielzahl an weiteren Metriken, die eine detaillierte Analyse und Vergleichbarkeit über die Vorhersagegüte von Modellen erlaubt.

Accuracy
Die Accuracy wird zur Interpretation der Modellgüte genutzt und stellt das Verhältnis zwischen richtig und falsch klassifizierten Datenpunkten dar. Der Wert liegt zwischen 0 und 1 und wird üblicherweise in Prozent ausgedrückt. Je höher der Wert, desto höher ist die Anzahl der richtig klassifizierten Datenpunkte im Datensatz und somit hat das Modell eine bessere Vorhersagegüte.[46, Kap.1]
Die Accuracy alleine bietet aber noch keine Aussagekraft über die Sicherheit in der Vorhersage je Klasse. Besonders bei einer Multiclass-Klassifikation, kann der Wert verzerrend wirken, wenn im Trainingsdatensatz unausgeglichene Klassen vorliegen.[12] Zum Beispiel wirkt ein Modell mit einer Accuracy von 85% auf den ersten Blick als ein guter Schätzer. Die Vorhersagegenauigkeit je Klasse eines Modells kann aber deutlich unterhalb der allgemeinen Modellgüte liegen. Daher werden weitere Metriken eingesetzt, die insbesondere bei unausgeglichenen Klassen eine genauerer Interpretation der Modellgüte zulassen.

Precision
Die Precision ist ein Maß zur Bestimmung der Genauigkeit des Modells. Sie drückt aus, wie exakt das Modell richtige Vorhersagen trifft, indem der Anteil der richtigen klassifizieren Datenpunkte ins Verhältnis zu der

Gesamttreffermenge setzt. Der Wert liegt zwischen 0 und 1. Je höher der Wert, desto genauer das Modell. Wobei ein Wert von 1 bedeuten würde das alle Datenpunkte die der Klasse zugeordnet wurden auch tatsächlich dieser Klasse angehört.[46, Kap.1]

Recall

Der Recall ist ein Maß für die Vollständigkeit der Vorhersage. Er setzt alle richtig klassifizierten Datenpunkte pro Klasse ins Verhältnis zu der Anzahl aller dieser Klasse zugehörigen Datenpunkten. Der Recall drückt aus, ob ein Klassifizierungsmodell alle Instanzen einer Klasse korrekt zuordnen kann.[46, Kap.1]
Ein hoher Recall bedeutet eine geringe Anzahl von falsch negativen Treffern durch das Modell. Dieser Wert ist insbesondere im medizinischen Kontext von hoher Bedeutung, denn im Falle von der Erkennung einer Erkrankung durch ein Modell, sollte die Anzahl von falsch negativen Vorhersagen möglichst gering oder idealerweise 0 sein. Das Vorliegen der Erkrankung sollte möglichst immer erkannt werden, wohingegen einige falsch positive Treffer des Modells eher akzeptiert werden können. Das Verhältnis zwischen Precision und Recall kann daher als ein Trade-Off verstanden werden, bei dem die Genauigkeit des Modells gegen die Vollständigkeit abgewogen werden muss.

$\mathbf{F_1}$ – Score

Der F_1-Score ist das harmonische Mittel aus Precision und Recall. Der Wert liegt auch hier zwischen 0 und 1, wobei 1 den bestmöglichen Wert darstellt. Der F_1-Score wird je Klasse berechnet. [46, Kap.1]

Micro- und Macro-Average

Für die Evaluierung eines Multiclass-Klassifizierungsproblems ist es von Vorteil, nur eine Metrik zu berechnen, mit der die allgemeine Performance des Modells bewertet werden kann. Hier eignen sich zwei Methoden, die alle Klassen in einem Score berücksichtigen. Der Micro-Average und der Macro-Average. Mit beiden Methoden lassen sich für Precision, Recall und den F_1-Score eine Metrik berechnen, die einen Durchschnitt über alle Klassen bildet.
Der Micro-Average aggregiert zunächst die Ergebnisse aller Klassen und bildet dann den Durchschnitt zu einer Metrik. Der Micro-Average eignet sich somit insbesondere bei unausgeglichenen Datensätzen.
Beim Macro Average wird zunächst die Metrik für jede Klasse einzeln berechnet und anschließend der Durchschnitt gebildet. Dadurch wird jede Klasse gleich behandelt und ausgeglichene Klassengrößen werden nicht berücksich-

tigt. [51]

ROC-Kurve
Eine weitere gängige Metrik bei der Auswertung von Klassifikationsproblemen sind die ROC-Kurve (Receiver Operating Characteristics) und der AUC-Score (Area Under The Curve). Die ROC-Kurve wird mit der True Positive Rate gegen die False Negative Rate in einem Graphen aufgetragen, wobei die True Positive Rate auf der y-Achse und die False Negative Rate auf der x-Achse liegt. Der ideale Verlauf einer ROC-Kurve verläuft eng an der oberen linken Ecke der Achse und erzeugt somit eine möglichst große Fläche unter Kurve. Ein hoher AUC-Score steht somit für ein gutes Klassifikationsmodell. Mit der ROC-Kurve wird dargestellt, wie sicher das Modell in der Einteilung der Klassen ist, indem die True Positive Rate (TPR) ins Verhältnis zur False Positive Rate (FPR) gesetzt wird.[30, S.147]

2.6. Verwandte Arbeiten

In der Einleitung wurde bereits anhand des Beispiels für die Erkennung von Brustkrebs aufgezeigt, dass Machine Learning Verfahren heute bereits im medizinischen Praxisalltag bei der Diagnostik von Erkrankungen unterstützen. Auch in der Ophthalmologie gibt es bereits eine Vielzahl an Studien die zeigen, dass Augenerkrankungen durch den Einsatz von ML erkannt und quantifiziert werden können. Einige Studien und Arbeiten, die mit dem Thema dieser Arbeit verwandt sind, werden in diesem Kapitel erläutert.

In den vergangenen Jahren hat die Anwendung von ML Verfahren auf Fundusfotos schon zu beachtlichen Erfolgen in der automatisierten Erkennung von Erkrankungen geführt. Dazu gehört zum Beispiel die Entwicklung von Modellen, die eine der häufigsten Formen von Augenerkrankungen, dem Glaukom, auf Fundusfotos erkennt. In der Studie [44] erzielte das trainierte CNN einen AUC-Score von 0.96 in der Erkennung von einer Glaukomerkrankung auf Fundusfotos.
Weitere Studien haben auch gezeigt, dass Deep Convolutional Neural Networks (DCNN) in der Lage sind, multiple Arten von Erkrankungen anhand von Fundusfotos zu erkennen. In den Studien [27] und [14] wurden DCNN mit einem Datensatz trainiert, in dem nicht nur Glaukomerkrankung vorliegen. In der Studie [14] schafft es das DCNN erfolgreich 39 verschiedene Augenerkrankungen zu klassifizieren. Das Netz wurde auf 249,620 Fundusfotos trainiert und erreicht in der Evaluierung einen AUC-Score von 0.9984.

Auch für das menschliche Auge vermeintlich unerkennbare Merkmale konnten anhand eines CNN bereits erfolgreich auf Fundusfotos erkannt werden. So kann das Geschlecht eines Patienten anhand einer funduskopischen Untersuchung ermittelt werden.[32]

Eine weitere interessante Fragestellung bei der Diagnose von Augenerkrankungen ist nicht nur die allgemeine Feststellung über das Vorliegen einer bestimmten Erkrankung, sondern auch die Quantifizierung über den Grad, zu dem die Erkrankungen bereits fortgeschritten ist. Damit befasst sich die Studie [41], in der die objektive Quantifizierung von glaukomatösen Defekten anhand von Fundusfotos untersucht wurde. Hier wurde eine DCNN trainiert, um anhand von Fundusfotos die durchschnittliche Dicke der retinalen Nervenfaserschicht (RNFL) vorherzusagen. Die Information zu der Dicke der RNFL stammt aus Aufnahmen der Spektral-Domänen optischen Kohärenztomografie (SD-OCT). Als Datengrundlage dienten 32,820 Paare von Fundusfotos und SD-OCT-RNFL Scans von 2,312 Augen und 1,198 Patienten. Im Ergebnis kommen die Autoren zu dem Schluss, dass die Vorhersage des DCNN auf den Testdaten sehr nahe an den Messungen aus einem OCT liegen. Die Korrelation zwischen den Werten aus der Vorhersage des Modells und der tatsächlichen RNFL Dicke entspricht einem Pearson $r = 0.832$ bzw. $R^2 = 69.3\%$ ($P < 0.001$), bei einem MAE von 7.39 μm.
Somit konnte gezeigt werden, dass die Bestimmung der durchschnittlichen RNFL Dicke anhand von Fundusfotos eine Quantifizierung über den Fortschritt von Defekten im Auge erlaubt. Dies ist ein wichtiger Aspekt bei der automatisierten Diagnostik einer Glaukomerkrankung, da diese Methodik nicht mehr von dem labeling von Daten durch Experten abhängig ist.

In einer ähnlichen Studie wurde mit einem DCNN die glaukomatösen neuroretinalen Schäden anhand der minimalen Randsaumweite der Bruch'schen Membranöffnung (BMO-MRW) aus SD-OCT Bildern quantifiziert und vorhergesagt.[53] Es wurde ein DCNN trainiert, das anhand von Fundusfotos die im SD-OCT gemessenen globalen und sektoralen Werte der MRW vorhersagen soll. Diese Werte lassen sich in der Art der Messung mit den Werten der RNFLT vergleichen. Als Datengrundlage dienten 9,282 Paare von Fundusfotos und SD-OCT MRW Scans von 927 Augen von insgesamt 490 Patienten. Die Ergebnisse der Studie zeigen eine hohe Korrelation von Pearson's $r = 0.8$ bzw. $R^2 = 77\%$ ($P < 0.001$) zwischen den vom DCNN vorhergesagten MRW Werten und den tatsächlich gemessenen Werten, bei einem MAE von 27.8 μm. Auch die Vorhersage der sektoralen Werte der MRW aufgeteilt nach Temporal inferior, Temporal Superior, Temporal, Nasal Superior, Nasal

Inferior und Nasal zeigten eine hohe Korrelation zu den wahren Werteb aus den SD-OCT-Scans.

Eine zu dem Thema dieser Arbeit eng verwandte Studie ist [16], zur Erkennung von glaukomatösen Gesichtsfelddefekten anhand von OCT und der RNFL Schichtdicke. Hier wurde ein Deep Learning Verfahren trainiert, um anhand von unterschiedlichen bildgebenden Verfahren, zum einen glaukomatöse Gesichtsfelddefekte zu erkennen und zum anderen die Defekte zu quantifizieren. Als Eingabe in das Modell dienten SD-OCT RNFL Thickness Maps, RNFL *en-face* Bilder und *Confocal Scanning Laser Ophthalmoscopy* (cSLO) Bilder. Die Quantifizierung des Defekts erfolgt anhand der Mean Deviation (MD), Pattern Standard Deviation (PSD) und sektoralen Pattern Deviation (PD) aus der Perimetrie. In der Studie wurden insgesamt 9765 Paare von SD-OCT Aufnahmen und Gesichtsfeldmessungen verwendet. Diese enthielten 1081 Augen von 665 Patienten ohne Gesichtsfelddefekten und 828 Augen von 529 Patienten, bei denen ein glaukomatöser Gesichtsfelddefekt festgestellt wurde.
Das Deep Learning Modell erreicht auf den Testdaten eine AUC-Score von 0.88, bei der Erkennung von glaukomatösen Gesichtsfelddefekten, basierend auf den RNFL *en-face* Bildern. Für die Erkennung von milden glaukomatösen Gesichtsfelddefekten erreicht das Modell einen AUC-Score von 0.82. Die Performance des Modells ist auf RNFL *en-face* Bildern dabei signifikant besser, als anhand von RNFL Schichtdicken Messungen (AUC = 0.82 und 0.73).
Bei der Vorhersage der MD aus der Perimetrie, schneidet das auf RNFL *en-face* Bildern trainierte Modell mit einem $R^2 = 0.70$ und MAE = 2.5 dB besser ab, als das anhand von RNFL Schichtdicken Messungen trainiert Modell mit einem R^2 von 0.45 und MAE von 3.7 dB.
Für die Vorhersage der sektoralen Durchschnittswerte der Pattern Deviation aus der Perimetrie, erreicht das Modell ein R^2 von 0.60 für den Inferior Nasal Sektor und ein R^2 von 0.67 für den Superior Nasal Sektor. In den Sektoren Superior ($R^2 = 0.35$), Inferior ($R^2 = 0.26$), Global ($R^2 = 0.15$) und Temporal ($R^2 = 0.12$) nimmt die Erklärbarkeit des Modells ab.

Diese Arbeit unterscheidet sich in zwei Aspekten von bisherigen Arbeiten. Ein sehr wesentlicher Unterschied ist, dass die Vorhersage und Quantifizierung von Gesichtsfelddefekten, ohne eine eindeutige Zuordnung zu einer Erkrankung erfolgt. Die Daten erhalten keine Label, somit wird keine direkte Klassifizierung auf Grundlage von Vorerkrankungen vorgenommen.
Zudem werden in dieser Arbeit mehrere Metadaten zur Vorhersage mitein-

bezogen. Der Datensatz verfügt neben den Fundusfotos und der Perimetrie auch über Informationen zu der RNFL und MRW sowie dem Alter und dem Geschlecht. Hier wird untersucht, inwieweit diese Metadaten die Vorhersage von Gesichtsfelddefekten verbessert.

3. Datengrundlage

In diesem Kapitel wird ein Überblick über die Herkunft und den Umfang der Daten gegeben. Im darauf folgenden Kapitel 4 wird dann auf die Vorverarbeitung und die Zusammensetzung der Datensätze eingegangen.

3.1. Datensatz des UKM

Die in dieser Arbeit verwendeten Daten stammen nahezu vollständig von der Klinik für Augenheilkunde des Universitätsklinikums Münster (UKM). Die Daten wurden in den ophthalmologischen Untersuchungen erhoben, die am UKM durchgeführt wurden und reichen über einen Zeitraum von 2013 bis Anfang 2023. Sie umfassen allgemeine Patientendaten, Fundusfotografien, Gesichtsfeldmessungen, OCT-Scans der RNFL und MRW sowie Daten zum Visus und zur Refraktion eines Patienten (vgl. Kapitel 2).
Die Patientendaten wurden in Form einer Excel Tabelle übergeben und enthalten lediglich die Patienten ID, das Geburtsdatum und das Geschlecht. Darüber hinaus sind keinerlei personenbezogene Daten erfasst, sodass die Daten anonym verwendet werden.
Die Fundusfotografien liegen als Bilddateien vor. Die Gesichtsfeldmessungen und OCT-Scans der RNFL und MRW liegen als PDF Dokumente vor. Neben den Dateien jeder einzelnen Untersuchung, wurden vom UKM zu den Fundusfotos, der Perimetrie und den OCT-Scans ebenfalls die Metadaten der Untersuchung in Form von Excel Tabellen mitgeliefert. In diesen Tabellen sind jeweils die Patienten ID, Datum der Aufnahme und der Dateiname des Bilds bzw. des Dokuments enthalten. Aus dem Dateinamen lässt sich auch ablesen, welches Auge untersucht wurde, bzw. im Fall der Fundusfotos und der OCT-Scans, ist eine eigene Spalte angefügt, aus der diese Information gezogen werden kann.
Anhand der Excel Tabellen, können die Bilder bzw. Untersuchungen, dem Patienten über die Patienten ID zugeordnet werden.

In Tabelle 3.1 wird eine Übersicht über die bereitgestellten Datensätze des UKM gegeben.

Datensatz	Anzahl	Dateityp
Fundusfotos	19699	JPEG
Perimetrie	15663	PDF
OCT-Scans	10668	PDF
Visus	71290	CSV
Refraktion	55827	CSV

Tabelle 3.1.: Übersicht über gesamten Datensatz des UKM

Für die Erkennung und Qunatifizierung des peripheren Sehens, sind zunächst die Fundusfotografien und die Daten aus der Gesichtsfeldmessung von Interesse. Diese werden im ersten Modell zur Klassifizierung des Gesichtsfeldverlusts verwendet. Im darauf folgenden Schritt werden die RNFL und MRW Werte mit einbezogen. Die Daten zu Visus und Refraktion stellen für die Fragestellung keine zu erwartende signifikante Vorhersagekraft dar, weshalb diese nicht einbezogen werden. Auf diese Daten wird daher nicht weiter eingegangen.

In Abschnitt 4.2 des folgenden Kapitels wird genauer auf die Datensätze für das Modelltraining eingegangen.

3.2. Open Source Daten

Wie im vorherigen Unterkapitel bereits erläutert, stammen nahezu alle Daten, die für das Trainieren der Machine Learning Verfahren verwendet werden, vom UKM. Lediglich ein frei verfügbarer Datensatz von Fundusfotos wird für das Transfer Learning des CNN verwendet. Hierbei handelt es sich um einen großen Datensatz von Fundusfotos, der im Zusammenhang einer Studie erstellt wurde. In der Studie [14] wurde ein Modell auf 249,620 Fundusfotos trainiert und getestet, in dem 39 verschiedene gelabelte Augenerkrankungen enthalten sind. Das Ziel der Studie ist es, ein multi-labeling Modell zu trainieren, das zuverlässig alle 39 unterscheidbaren Erkrankungen erkennen kann. Die Fundusfotos wurden aus verschiedenen Krankenhäusern in China, einem Datensatz aus USA und vier öffentlich verfügbaren Datensätzen zusammengestellt. Aus diesem Datensatz werden in dieser Arbeit 1000 Fundusfotos zum Transfer Learning des CNN verwendet.

4. Methoden und Versuchsaufbau

In diesem Kapitel werden die gewählten Methoden und Machine Learning Modelle vorgestellt, die für die Experimente verwendet wurden. Dazu wird zunächst ein Überblick über den gesamten Versuchsaufbau gegeben. Im zweiten Unterkapitel werden die Datenvorverarbeitungsschritte vorgestellt und der Aufbau der Datensätze für die Modelle beschrieben. Darauf folgt die Vorstellung der gewählten Machine Learning Architekturen und des Modelltrainings. Im letzten Abschnitt werden die Entwicklungs- und Ausführungsumgebung zur Durchführung der Experimente beschrieben.

4.1. Vorgehen und Überblick über Methoden

Um einen klaren Überblick über die Datensätze zu erhalten, wie sich diese zusammensetzen und welches Modell auf welchem Datensatz trainiert wird, folgt in diesem Abschnitt eine grobe Beschreibung der Datensätze und der Ablauf der Experimente. In den folgenden Unterkapiteln wird dann zusätzlich auf die Datensätze und die Modelle eingegangen.

In Abbildung 4.1 werden in einem Schaubild alle verwendeten Datensätze und Modelle zusammengefasst. Die einzelnen Artefakte und Schritte werden im folgenden Abschnitt erläutert.

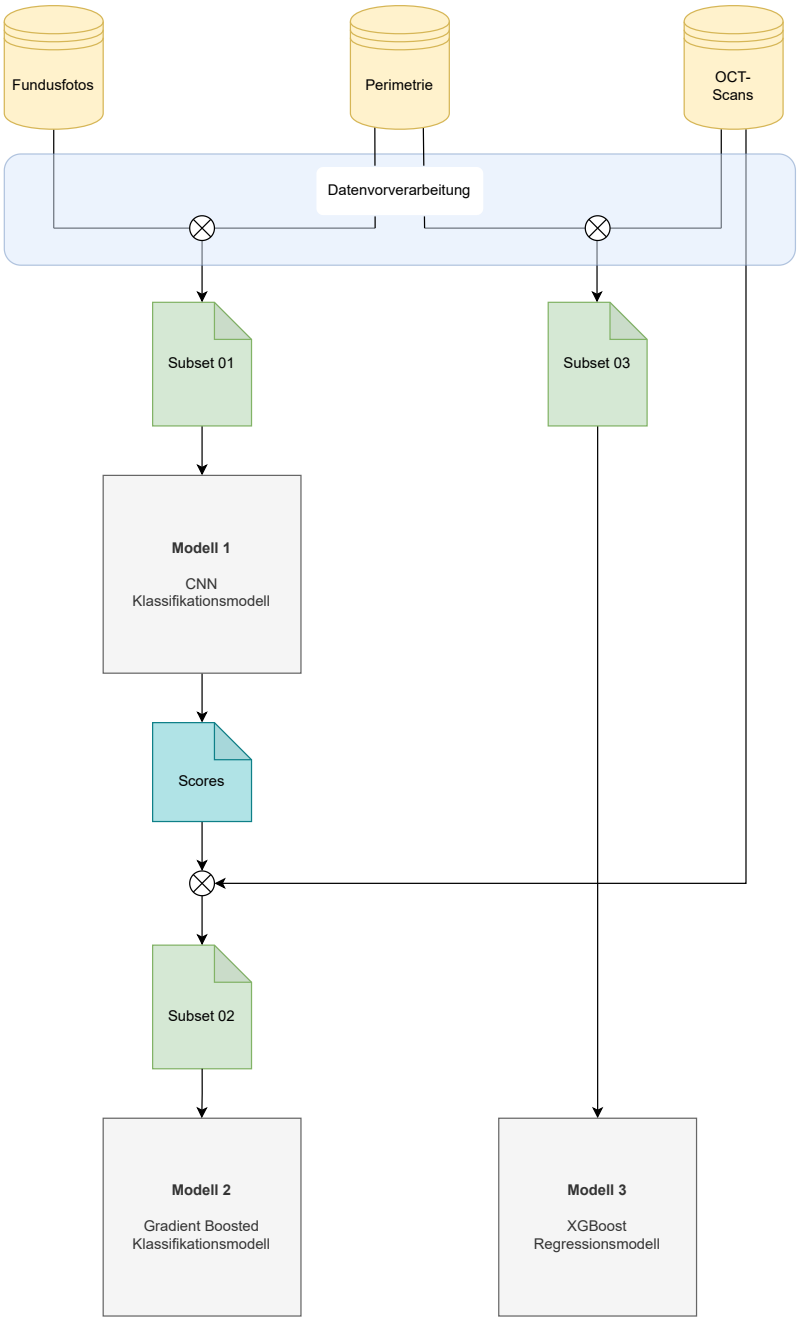


Abbildung 4.1.: Schaubild zu Verknüpfung der Datensätze und Ablauf der Experimente.

Im ersten Schritt wurden alle Datensätze der drei Datenquellen Fundusfotos, Perimetrie und OCT-Scans vorverarbeitet und auf der bereinigten Datenbasis die Datensätze, die für das Modelltraining benötigt werden, erstellt. Im folgenden Abschnitt 4.2 wird genauer darauf eingegangen, welche Schritte hierbei angewendet wurden.

Der erste Datensatz setzt sich aus den Fundusfotos und den Gesichtsfeldmessung zusammen. Zur leichteren Lesbarkeit und um Verwechslung zu vermeiden, wird dieser Datensatz nachfolgend als *Subset 01* bezeichnet. Das *Subset 01* zeichnet sich im wesentlichen durch die Verknüpfung von Fundusfoto und der Einteilung in eine Klasse aus, sodass jedes Fundusfoto des Datensatzes ein Label erhält.
Mit dem *Subset 01* wurde das erste Modell der Versuchsreihe trainiert, dem CNN Klassifikationsmodell. Ziel des Modells ist die Vorhersage der Hodapp-Stadien 1 bis 3. Also eine Einteilung von milder bis schwerer Form von Gesichtsfelddefekten.
Neben der Klassifikation in drei Klassen, wurde das Modell für weitere Untersuchungen auch auf zwei Klassen trainiert. Dabei fokussierte sich das Modelltraining darauf, die Unterscheidung zwischen *gesunden* Augen und Augen mit Defekten zu unterscheiden. Der Input des Modells war in diesem Fall zum einen, Fundusfotos mit der Zuteilung des Hodapp-Stadiums 0 als eine Klasse und die Hodapp-Stadien 1 bis 3 vereint zu der zweiten Klasse. Es geht hier also nicht, wie im ersten Modelltraining, um die Quantifizierung der Defekte, sondern um die Erkennung von Augen, die als *gesund* gelten und Augen denen bereits ein Defekt, unabhängig von der Schwere, zugeordnet werden.

Das zweite Modell der Versuchsreihe stellt eine Erweiterung des ersten Modells dar. Bei dem Gradient Boosted Klassifikationsmodell, wurden die Ergebnisse des CNN Klassifikationsmodell mit Metadaten verknüpft. Dadurch wurde untersucht, inwieweit die RNFL, MRW und Metadaten, wie Alter und Geschlecht, die Vorhersagen der Hodapp-Stadien verbessern können.
Für dieses Modell muss ein neuer Datensatz erstellt werden, in dem die Ergebnisse des CNN Klassifikationsmodells mit den Werten der RNFL und der MRW sowie den Metadaten kombiniert werden. Dieser Datensatz wird nachfolgend als *Subset 02* bezeichnet und kann als eine Erweiterung des *Subset 01* betrachtet werden. Er beinhaltet die Scores des CNN Klassifikationsmodells, den Hodapp-Stadien und den Werten der RNFL und der MRW.

Der dritte Datensatz setzt sich aus der Gesichtsfeldmessung und den

OCT-Scans zusammen. Dieser Datensatz wird nachfolgend als *Subset 03* bezeichnet und verknüpft die Messungen aus der Perimetrie mit den Werten der RNFL und MRW. Auf diesem Datensatz wurde das dritte Modell der Versuchsreihe trainiert, dem XGBoost Regressionsmodell. Ziel des Modells ist die Vorhersage der ortsabhängigen retinalen Sensitivität des Auges. Da es sich hierbei um ein Regressionsmodell handelt, erfolgt die Vorhersage des Modells auf kontinuierlichen reellen Werten. In diesem Fall auf der retinalen Sensitivität, die aus der numerischen Sensitivitätsgrafik der Gesichtsfeldmessung extrahiert wurde (vgl. Kapitel 2.1.2). Durch dieses Modell wird untersucht, inwieweit anhand der RNFL und der MRW Werte, Defekte im Gesichtsfeld vorhergesagt werden können.

4.2. Datenvorverarbeitung

4.2.1. Datenbereinigung

Bei großen Datenmengen ist es ganz natürlich, dass einige Datensätze für den Einsatz von Machine Learning Methoden nicht die notwendige Qualität aufweisen. Daher müssen zunächst einige Daten aussortiert werden, da sie fehlerhaft sind oder ein nicht verwendbares Format aufweisen. Zum Beispiel sind in den Fundusfotografien misslungene Aufnahmen vorhanden, bei denen der Sehnervenkopf nicht zu sehen ist oder eine zu starke Reflexion entstanden ist und somit nicht verwendet werden können.
Aus der Perimetrie werden nur die Untersuchungsergebnisse verwendet, die auch als zuverlässig gelten. Durch die Bereinigung anhand der Zuverlässigkeitsindizes entfallen mehr als 50% der Untersuchungen.
Auch in den Ergebnissen der OCTs, sind einige fehlerhafte Untersuchungen enthalten, in der keine genauen Werte zur RNFL oder MRW gemessen wurden und daher aus dem Datensatz entfernt wurden.
In der folgenden Tabelle 4.1 wird die Datenmenge nach Bereinigung und Aussortierung aller nicht verwendbaren Daten gezeigt.

Nach der Bereinigung von fehlerhaften Daten folgt die Aufbereitung der verbliebenen Daten.
Dafür wird zu Beginn auf allen vorhandenen Datensätzen das Datumsformat angepasst, da die Datensätze unterschiedliche Formate aufweisen und nicht dem Standardformat entsprechen. Dies ist ein für weitere Analysen notwendiger Schritt, da die Untersuchungen später unter anderem anhand des Datums zugeordnet werden. Schließlich werden die Ergebnisse und Da-

Datensatz	Anzahl	Dateityp
Fundusfotos	15374	JPEG
Perimetrie	6826	PDF
OCT-Scans	10447	PDF

Tabelle 4.1.: Übersicht über bereinigten Datensatz

ten gleicher Untersuchungsarten zu einem Datensatz zusammengeführt und dabei eventuell noch enthaltene Duplikate entfernt. Die nun vereinheitlichten und anonymisierten Daten werden dann je Untersuchung als eine Tabelle in eine Datenbank geschrieben.

Neben der Vorverarbeitung der Metadaten müssen auch die Bilddaten vorverarbeitet werden. Um die Fundusfotografien in das CNN geben zu können, wird von allen Bildern ein Ausschnitt um den Sehnervenkopf angefertigt. Der Ausschnitt ist nötig, da auf allen Fundusfotos nur ein kreisrunder Ausschnitt des Augenhintergrunds abgebildet wird. Hinter dem kreisrunden Ausschnitt ist der Hintergrund auf dem Bild schwarz und enthält somit keinerlei Information. Daher wird ein quadratischer Ausschnitt um den Sehnervenkopf angefertigt, sodass nur der Augenhintergrund mit dem Sehnervenkopf auf dem Bild zu sehen ist (vgl. Abbildung 2.5).
Die Bilder werden über eine statische Regel zugeschnitten. Hierfür wird ein Rahmen anhand von Koordinaten im Bild definiert, der den maximal größten Ausschnitt aus dem Fundusfoto anfertigt.
Da die Lage des Sehnervenkopfes teilweise auch am Rand oder sogar über dem Rand des Bildes liegt, werden alle Fundusfotografien manuell geprüft und gegebenenfalls aussortiert. So wird sichergestellt, dass sich nur Bilder im Datensatz befinden, die auch die erforderlichen Biomarker aufweisen.

4.2.2. Feature Engineering

Für die Erkennung von Erkrankungen anhand von medizinischen bildgebenden Verfahren ist es in der Regel notwendig, dass diese Daten zunächst von medizinisch geschultem Personal gesichtet und in Klassen eingeteilt werden, da das Labeling nicht durch einen Laien vorgenommen werden kann. Da der vorliegende Datensatz jedoch noch keine Label enthält, muss für das Training des CNN zunächst die Information geschaffen werden, anhand der das Modell

die Zuteilungsregeln erlernen soll. Die Daten liegen also sozusagen in ihrer Rohform vor. Für das Training des CNN wurden daher im nächsten Schritt die relevanten Informationen aus den Daten extrahiert, die als Inputvariablen verwendet werden.
Diesen Prozess nennt man Feature Engineering. Hierbei wird Fach- oder Domänenwissen angewendet, um Features aus den Rohdaten zu extrahieren. Diese Features enthalten in der Regel relevante Informationen, die für das Training eines Machine Learning Modells verwendet werden können. Dieser Prozess wird in Machine Learning Projekten als eine Schlüsselrolle bezeichnet und hat einen starken Einfluss auf die Qualität und die Vorhersagegüte eines Modells. [43, Kap. 4.8]

Mit Hilfe von Feature Engineering werden folgende Werte aus den Daten extrahiert bzw. generiert, die im Datensatz noch nicht enthalten sind:

- Wert der Mean Deviation aus der Gesichtsfeldmessung
- Zuverlässigkeitsindizes aus der Gesichtsfeldmessung
- Numerische Sensitivitätsgrafik aus der Gesichtsfeldmessung
- Klasseneinteilung in Hodapp-Stadien
- RNFL und MRW Werte aus den OCT-Scans

Extraktion der Mean Deviation
Die gemessene Mean Deviation bei einer Gesichtsfeldmessung ist bisher nur aus PDF Dokumenten zu entnehmen. Theoretisch müssten die Rohdaten dem System vorliegen, da es schließlich das PDF mit dem Wert erstellt. Auf diese Daten hat der Endnutzer des Geräts, in diesem Fall das UKM, jedoch keinen Zugriff. Um die Werte zu extrahieren, müssen die Daten also aus den PDF Dokumenten gelesen werden.
Das gewählte Vorgehen hierbei ist es also, durch ein Python-Skript mit einem PDF Parser die Dokumente maschinell auszulesen. Anschließend werden auf den ausgelesenen Text mit *Regular Expression* die Einträge zur Mean Deviation gefiltert und extrahiert.
Hierbei hat sich jedoch die Herausforderung ergeben, dass nicht alle Dokumente der Gesichtsfeldmessung auch als ein maschinenlesbares PDF Dokument vorliegen. Nur etwa 30% der Dokumente liegen in einem PDF Format vor, das durch den PDF Parser ausgelesen werden konnte. Die restlichen Dateien konnten durch das Python-Skript nicht ausgelesen werden, da diese PDF Dateien aus einer Bilddatei bestehen. Da es sich hierbei allerdings um

einen sehr großen und wichtigen Anteil der Daten handelt, wird eine weitere Methode angewendet die es ermöglicht, auch die nicht maschinenlesbaren Dokumente zu extrahieren.
Hierfür wird *Optical Character Recognition* (kurz OCR) verwendet. Mit OCR können Bilder von Texten in ein maschinenlesbares Textformat umgewandelt werden. Für die Erkennung der MD, in dem nicht maschinenlesbaren PDF Dokument, wird die entsprechende Stelle aus dem PDF bzw. Bild ausgeschnitten. Anschließend wird der Ausschnitt von der OCR Software analysiert und die Ergebnisse der Analyse werden gespeichert.

Extraktion der Zuverlässigkeitsindizes
Neben der Mean Deviation werden auch die Zuverlässigkeitsindizes aus dem Dokument der Gesichtsfeldmessung extrahiert. In Kapitel 2.1.2 wird die Bedeutung und Relevanz der Indizes aus der Gesichtsfeldmessung erklärt.
Es wird davon ausgegangen, dass die Werte eine bedeutsame Aussagekraft über die Qualität der Messung und somit auch auf das Ergebnis haben. Daher werden die Indizes extrahiert und genutzt, um einen weiteren Bereinigungsschritt auf den Daten auszuführen, indem unzuverlässige Messergebnisse entfernt werden.

Extraktion der numerischen Sensitivitätsgrafik
Zusätzlich ist es gelungen die numerische Sensitivitätsgrafik aus der Gesichtsfeldmessung zu extrahieren. Anhand der Werte soll die ortsaufgelöste Vorhersage von Gesichtsfelddefekten ermöglicht werden. Dies trifft allerdings nur auf die maschinenlesbaren PDF Dokumente zu. Auf allen anderen Dokumenten war es leider nicht möglich, die Sensitivitätsgrafik mittel OCR auszulesen.

Bilden der Hodapp-Stadien
In Kapitel 2.1.3 wurde der medizinische Hintergrund zu Hodapp-Stadien bereits erklärt. Im vorherigen Abschnitt wurde aufgezeigt, wie die Mean Deviation aus der Gesichtsfeldmessung extrahiert wird.
Nun wird auf Grundlage der MD für alle Gesichtsfeldmessungen eine Spalte mit der Einteilung in das entsprechende Hodapp-Stadium hinzugefügt. Die Einteilung in Klassen sieht wie folgt aus:

- Klasse 0: MD zwischen 0 dB und 2 dB
- Klasse 1: MD zwischen 0 dB und -6 dB
- Klasse 2: MD zwischen -6 dB und -12 dB

- Klasse 3: MD von weniger als -12

Durch die Einteilung nach Hodapp-Stadium sind nun Klassen bzw. Label definiert, mit denen das CNN zur Klassifikation trainiert wird.

Extraktion der RNFL und MRW Werte
Die RNFL Werte aus den OCT-Scans werden, wie auch die MD Werte, durch ein Python-Skript ausgelesen. Bei den Ergebnissen der Messung der Nervenfaserschichtdicke handelt es sich ebenfalls um maschinenlesbare PDF Dokumente. Die benötigten Werte werden auch hier mittels *Regular Expression* aus dem Text des PDF gefiltert und extrahiert. Entsprechend erhält man je Untersuchung, neben den Stammdaten, die Werte der sechs Sektoren sowie des globalen Durchschnittswerts (Global (G), Temporal (T), Temporal-Superior (TS), Temporal-Inferior (TI), Nasal (N), Nasal-Superior(NS) und Nasal-Inferior (NI)). Je Sektor werden zwei Werte extrahiert: Zum einen der absolute Wert der RNFL in Mikrometer und zum anderen der prozentuale Wert der die Abweichung zur Norm beschreibt.
Analog zu den RNFL Werten wird im gleichen Schritt auch die MRW Werte extrahiert. Die Werte können auf die gleich Art ausgelesen werden und liefern je Sektor den absoluten und den prozentualen Wert.

4.2.3. Aufbau der Datensätze

Nachdem die einzelnen Datensätze bereinigt und gegebenenfalls Feature Engineering auf ihnen angewendet wurde, werden die Trainingsdatensätze aufgestellt. Dafür werden die einzelnen Datensätze zu einer Schnittmenge zusammengeführt. (Für eine veranschaulichte Übersicht siehe auch Abbildung 4.1.)
Der erste Datensatz, das *Subset 01*, setzt sich aus den Fundusfotografien und der Perimetrie zusammen. Dazu werden alle Fundusfotografien mit den Ergebnissen der Gesichtsfeldmessung eines Patienten zusammengeführt. Um den Datensatz nicht zu sehr einzuschränken, wird dabei ein zeitlicher Abstand von sieben Monaten zwischen der Aufnahme des Fundusfotos und der Gesichtsfeldmessung toleriert. Somit umfasst der Datensatz *Subset 01* insgesamt 6126 Paare von bereinigten Fundusfotografien mit dem eindeutig zuordenbaren Hodapp-Stadium aus der Perimetrie.
Für das Training von Machine Learning Modellen wird der Datensatz typischerweise in drei Teile aufgeteilt. Der Datensatz *Subset 01* wird für das Training des CNN in einem Verhältnis von 70% Trainings-, 20% Validierungs- und 10% Testdaten aufgeteilt. Da es in der Literatur unterschiedliche No-

menklaturen für die Datensatz-Splits gibt, werden die Bezeichnungen der Datensätze, wie sie in dieser Arbeit verwendet werden, hier definiert.
Der Trainingsdatensatz ist der Datensatz, auf dem der Lernalgorithmus die Mustererkennung und die Anpassung der Gewichte vornimmt. Die Bilder dienen dem Lernalgorithmus als Beispiele, auf denen gelernt wird.
Der Validierungsdatensatz ist ebenfalls Teil des Trainingsprozesses des CNN. Er wird verwendet, um die Ergebnisse auf den Trainingsdaten zu evaluieren und die Hyperparameter des Modells anzupassen.
Der Testdatensatz umfasst nur Daten, die nicht in das Training des CNN gegeben wurden. Diese Daten sind dem Modell also unbekannt und ermöglichen eine unabhängige Bewertung der Modellgüte.
Für die Evaluierung des CNN wird für den Testdatensatz eine Patientengruppe gebildet, die bewusst nicht in dem Trainings- und Validierungsdatensatz enthalten ist. Somit kann mit dieser Gruppe eine unabhängige Evaluierung durchgeführt werden, bei der die Vorhersage nicht durch das Training beeinflusst wird.

In Tabelle 4.2 sind Aufbau und Umfang des bereinigten Datensatzes *Subset 01* aufgeführt.

Trainingsdatensatz				
Klasse	0	1	2	3
Anzahl Bilder	525	2226	571	966
Anzahl Patienten	318	1224	330	431
Anzahl Augen	412	1627	387	555
Alter (Jahren)	49.2 ± 16.5	47.1 ± 20.9	56.7 ± 19.9	58.4 ± 20.9
Anteil Frauen (%)	51.6	57.4	56.4	48.4
Mean Deviation (dB)	0.71 ± 0.56	-2.42 ± 1.59	-8.67 ± 1.76	-21.96 ± 6.41
Validierungsdatensatz				
Klasse	0	1	2	3
Anzahl Bilder	150	636	164	276
Anzahl Patienten	94	498	114	152
Anzahl Augen	121	574	127	197
Alter (Jahren)	43.8 ± 14.3	49.1 ± 20.9	54.1 ± 17.19	59.1 ± 18.7
Anteil Frauen (%)	56	62.3	50	47.8
Mean Deviation (dB)	0.68 ± 0.51	-2.38 ± 1.58	-8.43 ± 1.70	-22.24 ± 6.19
Testdatensatz				
Klasse	0	1	2	3
Anzahl Bilder	75	318	81	138
Anzahl Patienten	64	246	49	66
Anzahl Augen	69	266	55	82
Alter (Jahren)	53.5 ± 15.6	50.8 ± 20.0	57.6 ± 16.15	59.21 ± 20.56
Anteil Frauen (%)	50.6	64.5	60.5	52.2
Mean Deviation (dB)	0.81 ± 0.61	-2.44 ± 1.62	-8.96 ± 1.80	-22.43 ± 6.97

Tabelle 4.2.: Übersicht über demographische und klinische Eigenschaften der Augen und Patienten, die in Trainings-, Test- und Validierungsdatensatz des *Subset 01* enthalten sind.

Der zweite Datensatz, *Subset 02*, setzt sich aus den Scores des CNN Klassifikationsmodells, den Hodapp-Stadien und den Werten der RNFL und der MRW zusammen.
Die Scores drücken die Wahrscheinlichkeit für die Zuordnung in einer der drei Klassen aus. Je Klasse wird dem Datensatz also ein Score hinzugefügt.

Anhand der Patienten ID und des Datums werden je Patient nur OCT-Scans, die einen maximalen zeitlichen Abstand von sieben Monaten haben, verknüpft. Das *Subset 02* umfasst 1652 Dateneinträge. Auf diesem Datensatz wird das Gradient Boosted Klassifikationsmodell trainiert.
Da dieser Datensatz von *Subset 01* abstammt und Überschneidungen zum *Subset 03* aufweist, wird dieser Datensatz nicht gesondert in einer Tabelle mit den demografischen und klinischen Eigenschaften dargestellt.

Der dritte Datensatz, *Subset 03*, umfasst die Schnittmenge von Gesichtsfeldmessung und der OCT-Scans eines Patienten. Auch hier werden die Untersuchungen je Patient mit einem maximalen zeitlichen Abstand von sieben Monaten zugeordnet. Der Datensatz *Subset 03* umfasst 1372 Paare von Gesichtsfeldmessungen und OCT-Scans. Auf diesem Datensatz wird das Modell 3 - XGBoost Regressionsmodell - trainiert. Der Datensatz wird für das Modelltraining in einem Verhältnis von 70% Trainings- und 30% Testdaten aufgeteilt. In Tabelle 4.3 sind Aufbau und Umfang des Datensatzes *Subset 03* aufgeführt.

Trainingsdatensatz				
Hodapp-Stadium	0	1	2	3
Anzahl Datensätze	146	442	125	247
Anzahl Patienten	93	253	72	117
Anzahl Augen	115	328	79	144
Alter (Jahren)	46.5 ± 16.5	48.1 ± 21.4	62.5 ± 18.5	57.2 ± 20.6
Anteil Frauen (%)	46.6	61.4	56.8	54.7
Mean Deviation (dB)	0.68 ± 0.56	-2.20 ± 1.54	-8.98 ± 1.63	-21.81 ± 6.7
RNFLT_G (μm)	95.96 ± 12.11	95.57 ± 30.38	74.37 ± 36.50	59.81 ± 25.56
MRW_G (μm)	298.76 ± 64.27	308.61 ± 115.95	220.30 ± 109.29	190.32 ± 107.93
Testdatensatz				
Hodapp-Stadium	0	1	2	3
Anzahl Datensätze	48	207	59	98
Anzahl Patienten	42	160	47	65
Anzahl Augen	43	180	49	73
Alter (Jahren)	50.0 ± 17.6	45.1 ± 21.9	62.3 ± 18.2	63.7 ± 23.7
Anteil Frauen (%)	33.3	62.8	49.2	52.0
Mean Deviation (dB)	0.67 ± 0.68	-2.12 ± 1.51	-8.61 ± 1.83	-22.39 ± 6.68
RNFLT_G (μm)	93.52 ± 10.83	94.83 ± 24.00	72.56 ± 21.56	69.28 ± 60.19
MRW_G (μm)	294.96 ± 72.47	308.07 ± 104.56	216.76 ± 98.25	208.43 ± 122.09

Tabelle 4.3.: Übersicht über demographische und klinische Eigenschaften der Augen und Patienten, die in Trainings- und Testdatensatz des *Subset 03* enthalten sind.

4.3. Machine Learning Modelle

Modell 1 - EfficientNet Klassifikationsmodell

Nach der Datenvorverarbeitung und dem Aufstellen der Datensätze, können die Daten an die Machine Learning Modelle übergeben und das Training begonnen werden. Im folgenden Abschnitt wird auf die Wahl und den Aufbau der gewählten Modelle eingegangen sowie der Ablauf des Modelltrainings.

Im ersten Teil der Experimente wurde ein CNN Klassifikationsmodell

trainiert, das die Gesichtsfelddefekte anhand der Fundusfotos quantifiziert und über die Hodapp-Stadien in Klassen einteilt.
Als CNN Architektur wurde die EfficientNet-B4 Architektur verwendet. Für die Entwicklung des CNN wurden Teile des Programmcodes aus der vorangegangenen Masterarbeit von Robin Baudisch verwendet.[10] Die allgemeinen technischen Grundlagen zu CNNs werden bereits in Kapitel 2.5.2 gegeben. An dieser Stelle wird nun spezifisch auf die Besonderheiten des EfficientNet eingegangen.
Die EfficientNet Architektur nutzt eine besondere Skalierungsmethode, bei der alle Dimensionen des Faltungsnetzes einheitlich skaliert werden. Die Wahl der Modellkomplexität ist bei der Entwicklung von CNNs eine Herausforderung, die in der Regel mit viel manueller Arbeit verbunden ist. Das Modelltuning und die Verbesserung der Accuracy ist zudem auch eine Frage der Rechenleistung. In den vergangenen Jahren wurden zahlreiche neuartige CNN Architekturen vorgestellt, die bessere Performance bei gleichzeitig geringerer Modellkomplexität versprechen.
Im Gegensatz zur herkömmlichen CNN Architekturen skaliert die EfficientNet Architektur die Netzwerkbreite, -tiefe und -auflösung einheitlich über feste Skalierungkoeffizienten. Auf anderen Architekturen war es üblich, nur einen der drei Dimensionen zu skalieren.[52]
Diese Koeffizienten werden definiert als Tiefe $d = \alpha^{\phi}$, Breite $w = \beta^{\phi}$, Auflösung $r = \gamma^{\phi}$ und werden einheitlich nach dem Parameter ϕ skaliert. Dadurch erreicht die EfficientNet Architektur eine besser Performance, bei gleichzeitig geringerer Rechenkapazität. [52]

Das EfficientNet Modell wird im ersten Schritt vortrainiert. Dieses Vorgehen wird auch als *Transfer Learning* bezeichnet.
Transfer Learning ist in Machine Learning eine Methode, bei der Gelerntes anhand eines Problems auf ein verwandtes Problem übertragen wird. Das Ziel dieser Methode ist es, die Performance eines Modells allein über den Transfer von domänenspezifischem Wissen zu verbessern.[58]
Dieses Verfahren wird häufig auf Klassifikationsproblemen in Verbindung mit dem CIFAR Datensatz angewendet, um die Erkennung von Objekten zu verbessern. Der CIFAR Datensatz ist ein frei verfügbarer Datensatz, der 60000 32x32 Farbbilder mit 10 verschiedenen Klassen von Objekten umfasst. Die Objekte zeigen verschiedene Fahrzeug- und Tierarten. [33]
Durch das Transfer Learning sind die Parameter und Gewichte des Modells vorjustiert und das CNN muss nicht von grundauf neu trainiert werden.
In einem zweiten Durchlauf wird das EfficientNet erneut auf einem freien Datensatz von Fundusfotos trainiert. Hierfür wird ein frei verfügbarer Da-

tensatz von klassifizierten Fundusfotos verwendet. (vgl. 3.2).

Ein Schwäche des Trainingsdatensatzes ist die unausgewogene Datenmenge je Klasse. Unausgewogene Datensätze haben in der Regel einen negativen Einfluss auf die Genauigkeit eines Modells. Durch die ungleiche Verteilung der Klassen kann es passieren, dass es bei der Vorhersage zu Verzerrung kommt, indem die stärkste repräsentierte Klasse des Datensatzes am häufigsten vorhergesagt wird.[12] Um der ungleichen Verteilung entgegenzuwirken, wurde eine Gewichtung der Klassen vorgenommen.

Eine weitere gängige Methode bei unausgeglichenen Klassen und einer generell geringen Datenmenge, ist das Augmentieren der Bilder. *Augmentierung* in Machine Learning ist ein Verfahren zur Verbesserung des Datensatzes, indem die vorliegenden Trainingsdaten durch leichte Veränderungen vervielfacht werden.[56] Klassische Methoden der Augmentierung sind beispielsweise eine Rotation oder Spiegelung des Bildes. Auch können verschiedene Bildeffekte oder Filter auf die Bilder angewendet werden, die zum Beispiel einen künstliche Unschärfe über das Bild legen. Je mehr Daten ein Machine Learning Algorithmus zur Verfügung hat, desto effektiver kann ein Modell im Training lernen. [48] Daher ist es bei Datensätzen mit einem geringen Umfang oder bei stark unausgeglichenen Klassen gängige Praxis, die Menge der Bilder mit Hilfe von Augmentierung zu erhöhen. Auch kann durch die Augmentierung ein Overfitting des Modells, aufgrund eines geringen Datensatzes, verhindert werden. [49]
Zur Augmentierung wurde im Training zum einen eine Rotation von bis zu 10 Grad verwendet. Auch wurde eine horizontale Spiegelung der Achse angewendet, was die Unterscheidung zwischen rechten und linken Auge nachahmt. Zudem wurde ein Ausschnitt um das Zentrum des Bildes vorgenommen.

In Abbildung 4.2 wir exemplarisch gezeigt, wie Augmentierung auf die Fundusfotos angewendet wird.

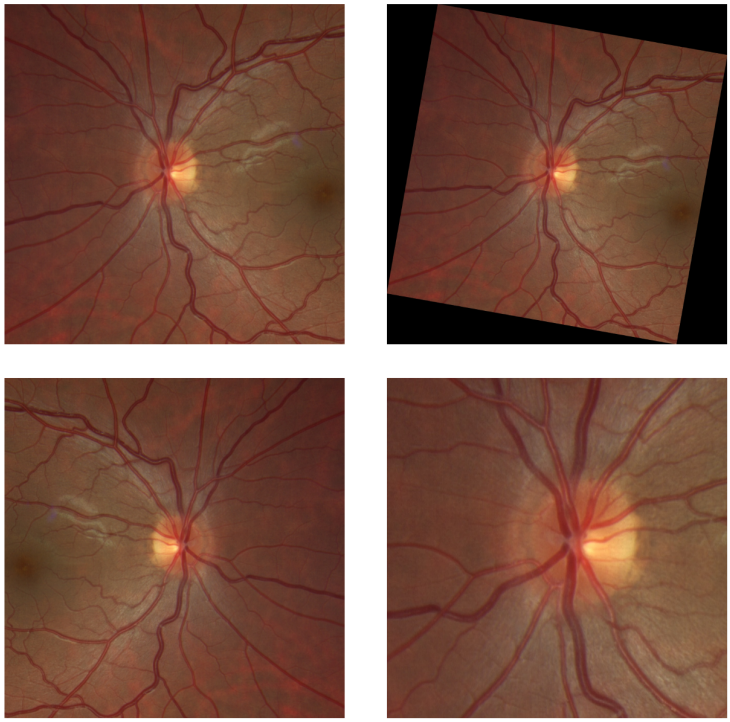


Abbildung 4.2.: Anwendung der Augmentierung auf Fundusfotos.
(Oben links: Fundusfoto vor Augmentierung. Oben rechts: Rotation des Bildes. Unten links: Horizontale Spiegelung. Unten rechts: Zentraler Ausschnitt des Bildes.)

Zur weiteren Prävention gegen Overfitting des Modells, verfügt das EfficientNet über Dropout Layer. Den Hidden-Layern des EfficientNet wurde eine Droput-Rate von 40% hinzugefügt. Es konnte gezeigt werden, dass durch das Hinzufügen von zufälligem Dropout in einem Modell es zu einer Verbesserungen bei der Objekterkennung unterstützen kann. [29]

Nach Abschluss des Trainings wird das Modell gespeichert. Anhand des gespeicherten Modells können die Vorhersagewerte, die das Modell anhand der Trainings-, Validierungs- und Testdaten trifft, jederzeit abgerufen werden oder auch neue Daten zur Erkennung übergeben werden.
Das EfficientNet-B4 Modell wurde mit den in Tabelle 4.4 gezeigten Hyperparametern trainiert.

Parameter	Parameterwert
Batch Size	16
learning_rate	0.001
Epochen	60
Weight Decay	0.001
Momentum	0.9
Nesterov	True

Tabelle 4.4.: Übersicht über die Wahl der Hyperparameter für das Training des EfficientNet.

Modell 2 - Gradient Boosted Klassifikationsmodell

Mit dem ersten Modell, dem CNN Klassifikationsmodell, wird die Erkennung und Quantifizierung von Gesichtsfelddefekten anhand von Fundusfotos prozessiert. Neben den Fundusfotos wurden im nächsten Schritt die Untersuchungsergebnisse aus den OCT-Scans der Nervenfaserschicht einbezogen. Da die RNFL Werte aus den OCT-Scans eine hohe Aussagefähigkeit über das Sehvermögen haben, verspricht das Einbeziehen dieser Werte einen positiven Effekt auf die Vorhersagegenauigkeit des Modells. Daher werden die RNFL Werte mit den Ergebnissen des Deep Learning Modells zur Klassifizierung auf Fundusfotos kombiniert.

Neben den RNFL Werten wurde in den Experimenten auch die MRW Werte einbezogen. Anhand der Analyse soll festgestellt werden, welchen Einfluss die Information der MRW auf die Vorhersage von Gesichtsfelddefekten hat. Für die Kombination des CNN Klassifikationsmodells und den Werten der RNFL bzw. MRW wird ein Gradient Boosted Klassifikationsmodell verwendet. Die theoretischen Grundlagen des Gradient Boosting werden in 2.5.3 gegeben.

Die Inputdaten für das Gradient Boosted Klassifikationsmodell setzten sich aus den Ergebnissen des CNN Klassifikationsmodell und den je Fundusbild zuordenbaren Messwerten der RNFL und MRW zusammen. Hierfür werden zunächst die Scores aus der Klassifikation des CNN Modells für jeden Datenpunkt des Trainings-, Validierungs- und Testdatensatzes generiert. Dieser Score drückt die prozentuale Sicherheit des Modells aus, zu welcher Klasse der Datenpunkt gehört. Die Inputdaten setzten sich also aus rein numerischen Werten zusammen.

Für die Experimente wurden die prozentualen und absoluten Messwerte der

RNFL und MRW einbezogen. Es wird Teil der Untersuchung sein, ob einer der Werten einen positiven Effekt auf die Vorhersagegenauigkeit haben wird. Der Output des Modells ist ebenfalls die Zuteilung nach Hodapp-Stadium je Datenpunkt. Also eine Quantifizierung der Defekte im Auge, wie schon bei dem ersten CNN Modell.
Daher erfolgt die Evaluierung auf den bekannten Evaluierungsmetriken für Klassifikationsproblemen und erlaubt den direkten Performance-Vergleich zum CNN Modell.

Das Gradient Boosted Modell wurde nach den in Tabelle 4.5 angegebenen Parametern trainiert.

Parameter	**Parameterwert**
loss	log_loss
learning_rate	0.1
max_depth	10
min_samples_leaf	1
min_samples_split	20
n_estimators	100
subsample	0.5

Tabelle 4.5.: Übersicht über die Wahl der Parameterwerte für das Training des Gradient Boosted Klassifikationsmodells.

Modell 3 - XGBoost Regressionsmodell
Im nächsten Schritt soll nach der Quantifizierung nun auch untersucht werden, inwiefern die Daten eine Verortung von Gesichtsfelddefekten ermöglichen. Hierfür wurde ein Regressionsmodell verwendet, das anhand der RNFL und MRW Werte vorhersagen soll, wie stark und in welchen Arealen des Auges das periphere Sehen geschwächt ist.
Der XGBoost Algorithmus ist eine Sonderform des Gradient Boosting. Vollständig steht XGBoost für *eXtreme Gradient Boosting* und soll verdeutlichen, dass es sich hierbei um eine Erweiterung des Gradient Boosting handelt. Der entscheidende Erfolgsfaktor hinter dem XGBoost Algorithmus ist seine Skalierbarkeit. Die Ausführung des Modells ist mehr als zehnmal schneller als andere bekannte Baum Algorithmen. Zudem hat der Algorithmus, im Vergleich zu anderen, eine bessere Performance auch auf geringen Datenmengen. [15]

Für die Vorhersage dient als Zielvariable diesmal nicht die Mean Deviation aus der Perimetrie, sondern die retinale Sensitivität. Die einzelnen Werte der retinalen Sensitivität erlauben eine noch genauere und vor allem ortsaufgelöste Aussage über die Sehleistung des Patienten, im Vergleich zu der Mean Deviation.
Für die Vorhersage der retinalen Sensitivität wurden die gemessenen Werte aus der numerischen Sensitivitätsgrafik je Quadrant zu einem Durchschnittswert zusammengefasst. In Abbildung 4.3 wird die Einteilung der Quadranten auf Grundlage der numerischen Sensitivitätsgrafik aus der Perimetrie veranschaulicht. Dabei repräsentieren die beiden oberen Quadranten die peripheren Sektoren Superior,die beiden unteren Inferior. [5]

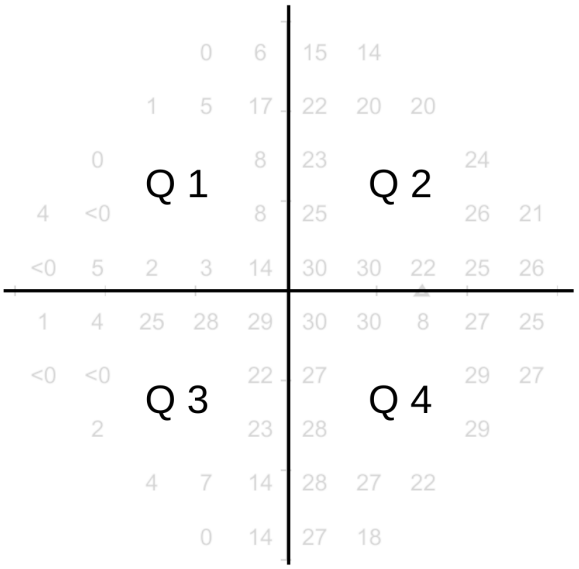


Abbildung 4.3.: Einteilung der Quadranten der retinalen Sensitivität.

Die Vorhersage des Modells wurde je Quadrant getroffen. Dies erlaubt eine Verortung von Defekten auf den linken bzw. rechten Quadranten der Sektoren Superior und Inferior. Diese Art der Vorhersage wäre allein anhand der Fundusfotos nicht ohne weiteres möglich, da hier die Ausrichtung des Sehnervenkopfes im Bild nicht eindeutig ist. Ohne Expertenwissen kann bei Betrachtung eines Fundusfotos also nicht die Lage der peripapillären Sektoren bestimmt werden. Diese Zuordnung ist zwischen den Werten der RNFL bzw. MRW und den Quadranten der retinalen Sensitivität deutlich einfacher, da die Lage der Sektoren in jeder Untersuchung gleich und eindeutig ist.

Das XGBoost Modell wurde nach den in Tabelle 4.6 angegebenen Parametern trainiert.

Parameter	**Parameterwert**
learning_rate	0.1
max_depth	10
n_estimators	500
subsample	0.5
booster	gbtree

Tabelle 4.6.: Übersicht über die Wahl der Parameterwerte für das Training des XGBoost Regressionsmodells.

4.4. Ausführungsumgebung

Die Ausführung der Experimente erfolgte auf eigener GPU Hardware. Für die Umsetzung und Entwicklung der Modelle, wurden Date Science übliche Python Libraries verwendet.
Für die Umsetzung des Convolutional Neural Network Modells zur Klassifikation wurde das Python Framework *PyTorch* verwendet.[45]
Für die Umsetzung des Gradient Boosted Klassifikationsmodells und des XGBoost Regressionsmodells wurde die Python Library *Scikit-learn* verwendet [47].
In Tabelle 4.7 sind die Ausführungsumgebung und Hardware aufgeführt, die zum Training der Machine Learning Modelle verwendet werden.

Hardware/Software Komponente	**Spezifikation**
GPU	AMD Radeon RX 6800 XT
CPU	AMD Ryzen 7 3700X 8-Core Processor
RAM	32GB
OS	Fedora Linux 37 (Kernel Version: 6.3, 64-bit)
Python Version	3.11
PyTorch Version	2.0.1 ROCm 5.4.2

Tabelle 4.7.: Übersicht über Ausführungsumgebung

5. Ergebnisse

In diesem Kapitel werden die Ergebnisse der durchgeführten Experimente vorgestellt.
Der Versuchsaufbau und die Architektur der Machine Learning Modelle werden bereits in Kapitel 4.1 beschrieben.
Im ersten Unterkapitel werden die CNN Klassifikationsmodelle evaluiert. Zunächst werden die Ergebnisse des Modells zur Quantifizierung der Gesichtsfeldschäden gezeigt. Darauf folgen die Ergebnisse des Modells auf zwei Klassen, zur Erkennung und Unterscheidung von *gesunden* Augen und Augen mit einem Gesichtsfelddefekt.
Im zweiten Unterkapitel wird das Gradient Boosted Klassifikationsmodell vorgestellt, das eine Erweiterung des CNN Klassifikationsmodell zur Quantifizierung der Gesichtsfelddefekte darstellt.
Darauf folgen die Ergebnisse zum XGBoost Regressionsmodell, mit dem die Verortung von Defekten im Auge vorhergesagt wurde.
Am Ende des Kapitels werden die wichtigsten Ergebnisse und die Erkenntnisse zusammengefasst, bevor im letzten Kapitel das Fazit der Arbeit gezogen wird.

5.1. Evaluation - CNN Klassifikationsmodell

5.1.1. CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien

Zunächst werden der Verlauf der Accuracy und des Loss für die Trainings- und Validierungsdaten aus dem Modelltraining vorgestellt.
In Abbildung 5.1 wird der Verlauf der Accuracy, in Abbildung 5.2 der des Loss je Epoche, grafisch abgebildet.
Auf den Trainingsdaten lässt sich ein typischer Verlauf beobachten. Im Verlauf des Trainings verbessert sich die Accuracy, während sich der Loss verringert.
Auf den Validierungsdaten allerdings ist dieser Verlauf nicht zu beobachten. Die Accuracy stagniert schon früh im Training und erreicht nicht das gleiche Niveau, wie auf den Trainingsdaten. Auch der Loss verringert sich im Verlauf der Trainings nicht weiter in Richtung eines Minimus, sondern steigt leicht an.
Obwohl einige Methoden zur Prävention von Overfitting getroffen wurden, weist der Trainingsverlauf den typischen Verlauf von Overfitting auf. Trotz der sehr kleinen Lernrate im Trainingsverlauf, sind auch die starken Schwankungen im Lernprozess auf den Validierungsdaten auffällig.
Auch nach Erhöhung der Epochen konnte keine Konvergenz des Modells festgestellt werden. Folglich schafft es das Modell nicht, ein globales Minimum zu finden, um den Verlust zu verringern und lässt auf eine bleibende Unsicherheit des Modells schließen.

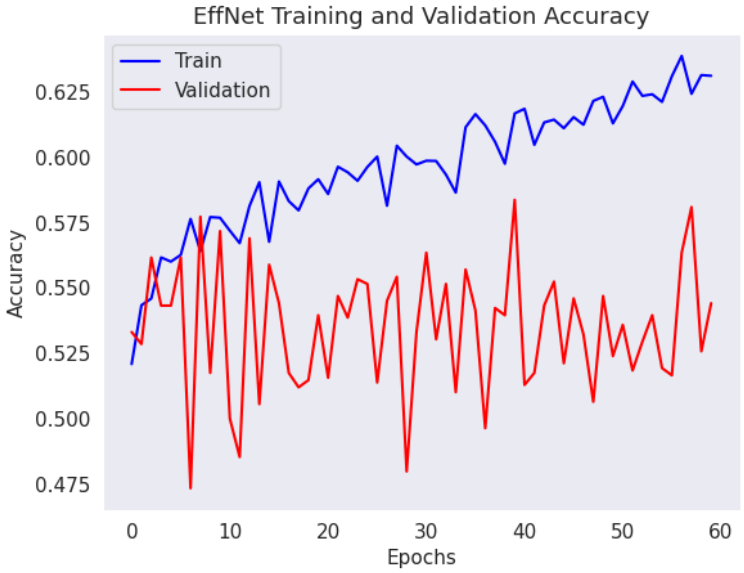


Abbildung 5.1.: Verlauf der Accuracy im Modelltraining des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

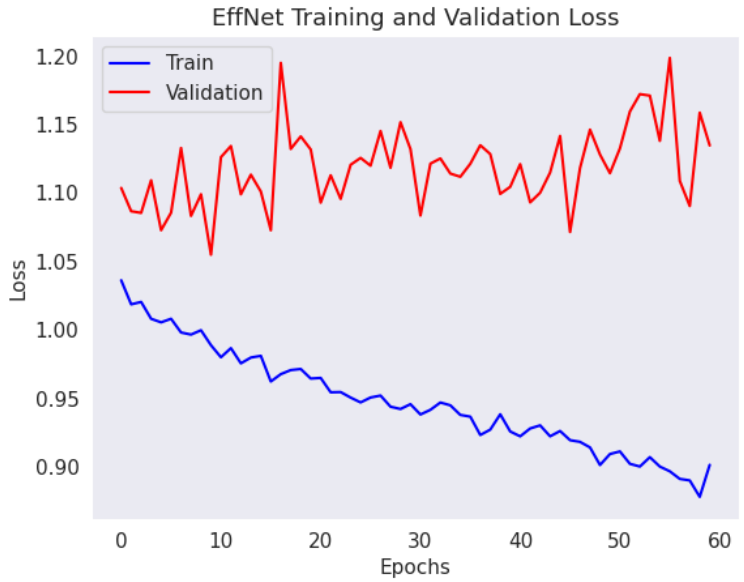


Abbildung 5.2.: Verlauf des Loss im Modelltraining des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Das Modell erzielt bei der Erkennung der drei Hodapp-Stadien einen

Micro-Average Recall von 0.54. Das Modell weist somit erhebliche Schwächen in der Unterscheidung der Klassen auf und ist in der Genauigkeit kaum besser als raten.
Bei Betrachtung des F_1-Score je Klasse wird deutlich, dass sich die Genauigkeit der Vorhersage zwischen den Klassen stark unterscheidet.
Bei Fundusfotos, die dem Stadium 1 angehören, ist sich das Modell mit einem F_1-Score von 0,71 deutlich sicherer, als bei der Einteilung in das Stadium 3, wo dieser bei nur 0,42 liegt. Besonders schlecht erkennt das Modell Defekte des Stadiums 2, wo der F_1-Score bei nur 0,15 liegt. Hier ist also festzustellen, dass die Vorhersage der Klassen 2 und 3 selbst durch zufälliges Vorhersagen besser wäre.
Durch die ungleiche Klassenverteilung der Datenmenge ist vorzugsweise der Micro-Average Wert in der Bewertung und dem Vergleich der Modelle heranzuziehen. Der Micro-Average F_1-Score liegt bei 0.56. Daran wird deutlich, dass die Vorhersagegenauigkeit des Modells insgesamt verzerrt wird durch die ungleichen Klassengrößen.

Tabelle 5.1 zeigt die Metriken im Detail.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,74	0,69	0,71
Stadium 2	0,15	0,22	0,18
Stadium 3	0,45	0,4	0,42
Micro-Average	0,58	0,54	0,56
Macro-Average	0,45	0,44	0,44

Tabelle 5.1.: Übersicht über Evaluationsmetriken auf den Testdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Die Confusion Matrix (vgl. Tabelle 5.2) stellt noch einmal die vom Modell getroffenen Klassenteilungen gegenüber. Die Zahl im Feld ist der absolute Wert der richtig bzw. falsch klassifizierten Bilder. In Klammern ist der Anteil der True Positive im Verhältnis der False Positive je Klasse dargestellt. Dies ist gleichzeitig auch der Recall.
Hier wird die besonders hohe Unsicherheit der Klassen 2 deutlich, bei der mehr Bilder falsch als richtig eingeteilt wurden. Selbst die Anzahl der True Positive klassifizierten Bilder des Stadiums 1, erreicht für den medizinischen Kontext keine besonders gute Quote.

Confusion Matrix - Testdaten			
	Stadium 1	Stadium 2	Stadium 3
Stadium 1	219 (0.69)	67 (0.21)	32 (0.1)
Stadium 2	28 (0.35)	18 (0.22)	35 (0.43)
Stadium 3	49 (0.36)	34 (0.25)	55 (0.4)

Tabelle 5.2.: Confusion Matrix auf den Testdaten des CNN Klassifikationsmodells zur Erkennung der Hodapp-Stadien.

Die ROC-Kurve in Abbildung 5.3 verdeutlicht die Performance des Modells grafisch. Hier wird besonders die schlechte Erkennung des Hodapp-Stadiums 2 erneut deutlich. Ein Wert nahe der Winkelhalbierenden in der ROC-Kurve bedeutet eine gleiche Quote in der Zuteilung von True Positive und False Negative. Der Verlauf der Kurve für die Klasse 2 verläuft erkennbar unterhalb der Winkelhalbierenden. Die Einteilung ist somit nicht besser als zufälliges Raten und die Klasse ist in ihren Merkmalen nicht unterscheidbar.

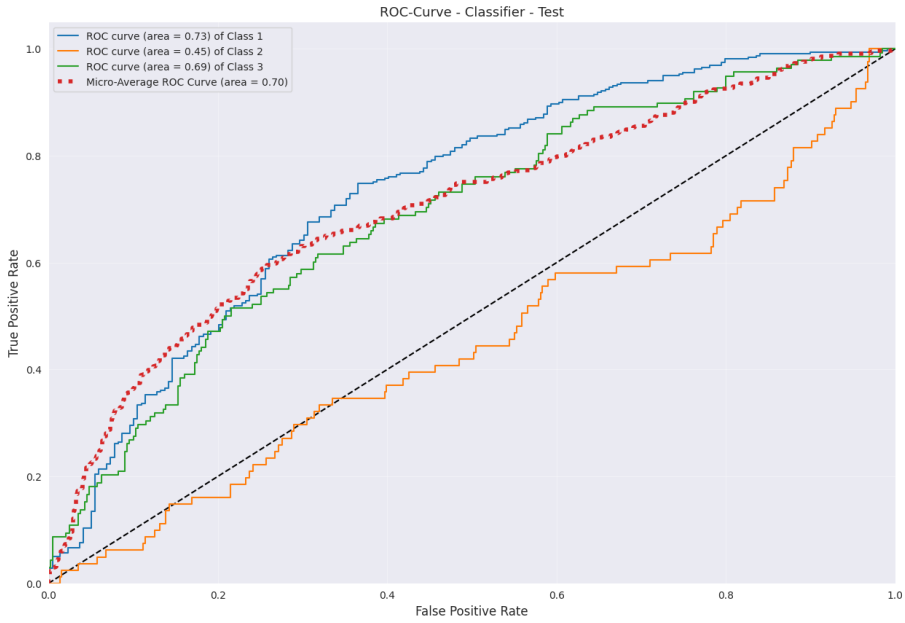


Abbildung 5.3.: ROC-Kurve und AUC-Scores auf Testdaten des CNN Klassifikationsmodells zur Erkennung der Hodapp-Stadien.

Im Anhang sind die Evaluationsmetriken des Modells auch für die Trainings- und Valdierungsdaten abgebildet. Dabei ist festzustellen das auch im Training

bereits die Unsicherheiten in der Klasseneinteilung bestehen.

Vergleich der Modellgüte vor Bereinigung des Fixationsverlusts

Um besser nachvollziehen zu können welchen Einfluss die Daten auf die Vorhersage des Modells haben, werden anschließend noch die Ergebnisse des CNN vorgestellt, das auf Trainingsdaten trainiert wurde, bevor diese um Fixationsverluste bereinigt wurden. Die Fixationsverluste betreffen die Messungen aus der Perimetrie und geben Aufschluss über die Genauigkeit der Untersuchungsergebnisse. (vgl. Kapitel 2.1.2)
In Tabelle 5.3 werden die Evaluationsmetriken und in 5.4 die Confusion Matrix des Modells gezeigt.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,76	0,84	0,8
Stadium 2	0,36	0,06	0,11
Stadium 3	0,55	0,67	0,6
Micro-Average	0,65	0,68	0,65
Macro-Average	0,55	0,52	0,5

Tabelle 5.3.: Übersicht über Evaluationsmetriken des CNN Klassifikationsmodells auf Testdaten vor Bereinigung um Fixationsverluste.

Confusion Matrix - Testdaten			
	Stadium 1	Stadium 2	Stadium 3
Stadium 1	276 (0.84)	6 (0.02)	48 (0.14)
Stadium 2	44 (0.54)	5 (0.06)	32 (0.40)
Stadium 3	44 (0.31)	2 (0.02)	97 (0.67)

Tabelle 5.4.: Confusion Matrix des CNN Klassifikationsmodells auf den Testdaten vor Bereinigung um Fixationsverluste.

Insgesamt hat das Modell mit einem Micro-Average F_1-Score von 0.65 eine bessere Vorhersagegüte als vor der Bereinigung des Fixationsverlusts. Bei Betrachtung der einzelnen Klassen fallen aber wieder starke Schwankungen auf. Die Genauigkeit der Klasse 1 schneidet mit einem F_1-Score von 0.8

deutlich besser ab. Ebenso Klasse 3, mit einem F_1-Score von 0.6. Anders aber die Klasse 2, bei der im Vergleich ein noch schlechterer F_1-Score von nur 0.1 erreicht wird.
Obwohl das Modell nach den Metriken insgesamt eine bessere Vorhersagegüte aufweist, wird bei genauerer Betrachtung deutlich, dass sich hier nur eine Verschiebung zwischen den Klassen ergeben hat. Die Verbesserung in der Vorhersage der Klasse 1 und 3 führten zu einer Verschlechterung der Klasse 2. Die Modellgüte wurde also nur auf Kosten der Genauigkeit des Modells erhöht. Daraus kann abgeleitet werden, dass die Bereinigung des Fixationsverlusts die Trennschärfe zwischen den Klassen verbessert hat.

5.1.2. CNN Klassifikationsmodell zur Erkennung von Gesichtsfelddefekten

Im Vergleich zum vorherigen Modell, werden nun die Ergebnisse des CNN Klassifikationsmodell vorgestellt, das nur auf zwei Klassen trainiert wurde. Bei diesem Klassifikationsmodell wird untersucht, wie gut *gesunde* Augen von Augen mit Gesichtsfelddefekten unterschieden werden können. Die Auswertung erfolgt also auf der Klasse 0, *gesund*, und Klasse 1, mit Gesichtsfelddefekt, die alle Bilder der Hodapp-Stadien 1 bis 3 zugeordnet werden können.

Die folgenden zwei Abbildungen 5.4 und 5.5 zeigen den Verlauf der Accuracy und des Loss während des Modelltrainings.

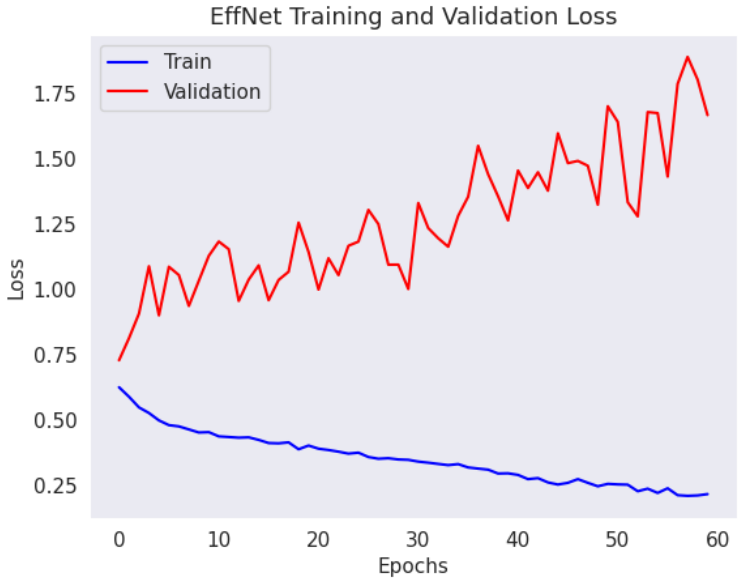


Abbildung 5.4.: Verlauf der Accuracy im Modelltraining des CNN Klassifikationsmodells mit zwei Klassen.

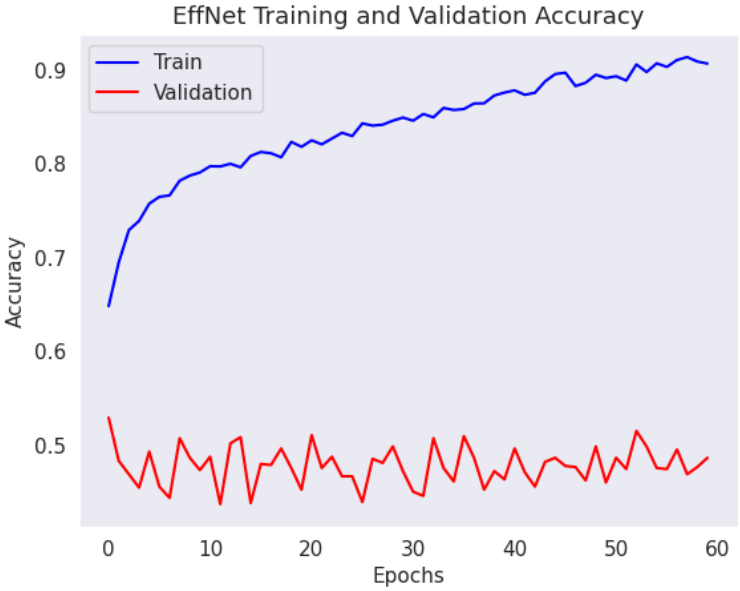


Abbildung 5.5.: Verlauf des Loss im Modelltraining des CNN Klassifikationsmodells mit zwei Klassen.

Wie auch im Modell zuvor, wird in diesem Modelltraining keine Konvergenz

erreicht. Auf den Trainingsdaten verbessert sich die Accuracy mit Fortschritt des Trainings während der Loss abnimmt. Auf dem Validierungsdatensatz aber stagniert die Accuracy auf einem gleichen Niveau über das gesamte Training. Der Loss dagegen steigt sogar, anstatt sich zu verringern. Hier sind jedoch keine so deutlichen Ausreißer in den Kurven, wie beim ersten Modell, zu erkennen.

Die Evaluationsmetriken (Tabelle 5.6) und die Confusion Matrix (Tabelle 5.5) zeigen die Ergebnisse der Auswertung auf den unabhängigen Testdaten. Für die Erkennung des Stadiums 0 erreicht das Modell einen F_1-Score von 0.73. Für Klasse 1 liegt der F_1-Score bei 0.79. Mit einer Accuracy von 0,77, ist das Modell in der Vorhersage insgesamt besser und genauer als das vorherige Modell, das auf drei Klassen trainiert wurde.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 0	0,72	0,73	0,73
Stadium 1	0,8	0,79	0,79
Accuracy			0,77

Tabelle 5.5.: Übersicht über Evaluationsmetriken auf den Testdaten des CNN Klassifikationsmodells mit zwei Klassen.

Confusion Matrix - Testdaten		
	Stadium 0	Stadium 1
Stadium 0	55 (0.73)	20 (0.27)
Stadium 1	21 (0.21)	79 (0.79)

Tabelle 5.6.: Confusion Matrix auf den Testdaten des CNN Klassifikationsmodell mit zwei Klassen.

Anhand der ROC-Kurve (Abbildung 5.7) ist zu sehen, dass beide Klassen einen hohen AUC-Score von 0.86 erreichen. Vergleicht man den Micro-Average AUC-Score, so schneidet das Modell in der Vorhersagegüte besser ab als das vorherige Modell (AUC-Score=0.70).

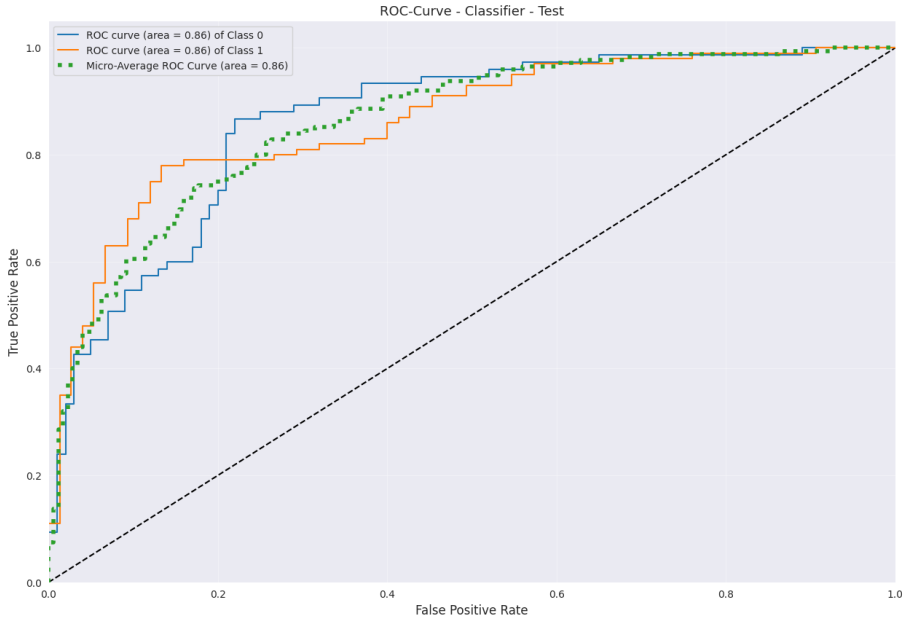


Abbildung 5.6.: ROC-Kurve und AUC-Scores auf den Testdaten für das CNN Klassifikationsmodells mit zwei Klassen.

5.2. Evaluation - Gradient Boosted Klassifikationsmodell

An dieser Stelle werden die Ergebnisse des Gradient Boosted Klassifikationsmodell gezeigt, die durch die Kombination aus den Ergebnissen des CNN Klassifikationsmodell und den RNFL und MRW Werten erzielt werden.

Die Evaluierung erfolgt anhand der Evaluationsmetriken und der Confusion Matrix (vgl. Tabelle 5.7 und 5.8).
Auf den Testdaten erreicht das Modell einen Micro-Average F_1-Score von 0.85, was eine deutliche Verbesserung zum CNN Klassifikationsmodell darstellt (0.56). Bei Betrachtung der einzelnen Klassen, lässt sich für jede der Klassen eine Verbesserung in der Vorhersage erkennen. Hier erreicht das Modell einen sehr guten F_1-Score für die Klasse 1 (0.93) und gute Werte für die Klassen 2 (0.59) und 3 (0.73).
Diese Werte liegen deutlich über den Ergebnissen des ersten CNN Modells (Klasse 1: 0.71, Klasse 2: 0.18 und Klasse 3: 0.42).
Weiterhin auffällig ist jedoch, dass die Vorhersage auf Klasse 1 am stärksten

und auf Klasse 2 am schwächsten bleibt.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,91	0,95	0,93
Stadium 2	0,58	0,61	0,59
Stadium 3	0,81	0,67	0,73
Micro-Average	0,85	0,86	0,85
Macro-Average	0,76	0,75	0,75

Tabelle 5.7.: Übersicht über Evaluationsmetriken des Gradient Boosted Klassifikationsmodells auf allen Variablen.

Confusion Matrix - Testdaten			
	Stadium 1	Stadium 2	Stadium 3
Stadium 1	121 (0.95)	4 (0.03)	2 (0.02)
Stadium 2	2 (0.11)	11 (0.61)	5 (0.28)
Stadium 3	10 (0.23)	4 (0.10)	29 (0.67)

Tabelle 5.8.: Confusion Matrix auf den Testdaten des Gradient Boosted Evaluationsmetriken auf allen Variablen.

Auch anhand der ROC-Kurve (Abb.: 5.7) wird deutlich, dass sich die Vorhersagegenauigkeit für alle Klassen verbessert hat. Mit einem Micro-Average AUC-Score von 0.96 auf den Testdaten, wird die deutliche Verbesserung in der Trennschärfe des Modells deutlich. Es ist demnach festzustellen, dass sich durch die Kombination der Scores und der RNFL und MRW Werte die Modellgüte deutlich verbessert hat.

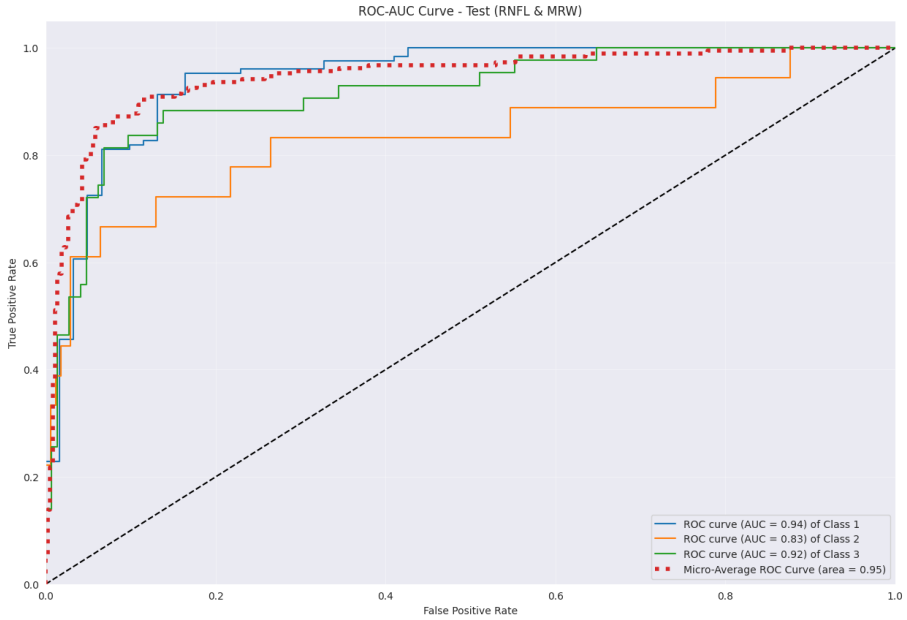


Abbildung 5.7.: ROC-Kurve auf den Testdaten des Gradient Boosted Klassifikationsmodells auf allen Variablen.

Da das Modell bisher alle Variablen in die Vorhersage einbezieht hat, ist es möglich, dass die Vorhersage durch sich gegenseitig beeinflussende Variablen verzerrt wird. Daher wird im nächsten Schritt die Feature Importance und anschließend die Korrelation zwischen den Variablen berücksichtigt. Das Entfernen von überflüssigen Variablen kann einen positiven Einfluss auf die Erklärbarkeit und Güte des Modells haben.
Die Tabelle 5.9 zeigt die Feature Importance der 15 Variablen mit der höchsten Wichtigkeit, gemessen an der *Gini Importance*, für das Modell. Hieran wird deutlich das die Scores, die aus dem CNN Modell stammen, eine hohe Wichtigkeit für das Modell haben.
Zudem weisen nicht nur die RNFL Werte, sondern auch die Werte der MRW eine hohe Wichtigkeit auf. Die MRW Werte wirken sich also ebenfalls positiv auf die Performance des Modells aus.
Eine weitere Erkenntnis ist, dass die absoluten Werte der RNFL bzw. der MRW eine höhere Wichtigkeit für das Modell haben als die prozentualen Werte.
Das Alter, als eines der zehn wichtigsten Feature, weist ebenfalls einen positiven Beitrag zur Modellgüte auf. An letzter Stelle der Feature Importance mit nur 0.000622 steht das Geschlecht. Hier ist also davon auszugehen, dass das Geschlecht keinen Einfluss auf die Vorhersage hat.

Variablen	**Feature Importance**
RNFLT_TS	0.190221
Score 3	0.092605
RNFLT_G	0.086739
Score 1	0.079772
Score 2	0.063514
RNFLT_TI	0.059851
MRW_NI	0.037525
MRW_N	0.029400
MRW_TI	0.028321
Alter	0.026868
MRW_T	0.025042
RNFLT_N	0.024325
RNFLT_TS_Prc	0.023703
RNFLT_NS	0.023196
RNFLT_NI	0.022826

Tabelle 5.9.: Übersicht über die Feature Importance des Gradient Boosted Klassifikationsmodells auf 15 Variablen.

In Abbildung 5.8 wird die Korrelations-Heatmap mit den zuvor als wichtigsten identifizierten Variablen gezeigt. Eine besonders starke positive Korrelation wird zwischen den Variablen des zentralen Sektor (RNFLT_G) der RNFL Werte und den sechs peripapillären Sektoren deutlich. Gleiches gilt für die MRW Werte. Dies ist nicht überraschend, da der zentrale Wert einen Durchschnittswert über alle Sektoren darstellt. Entsprechend weisen die sechs peripapillären Sektoren und der zentrale Sektor eine Überschneidung in der Erklärbarkeit des Modells auf.

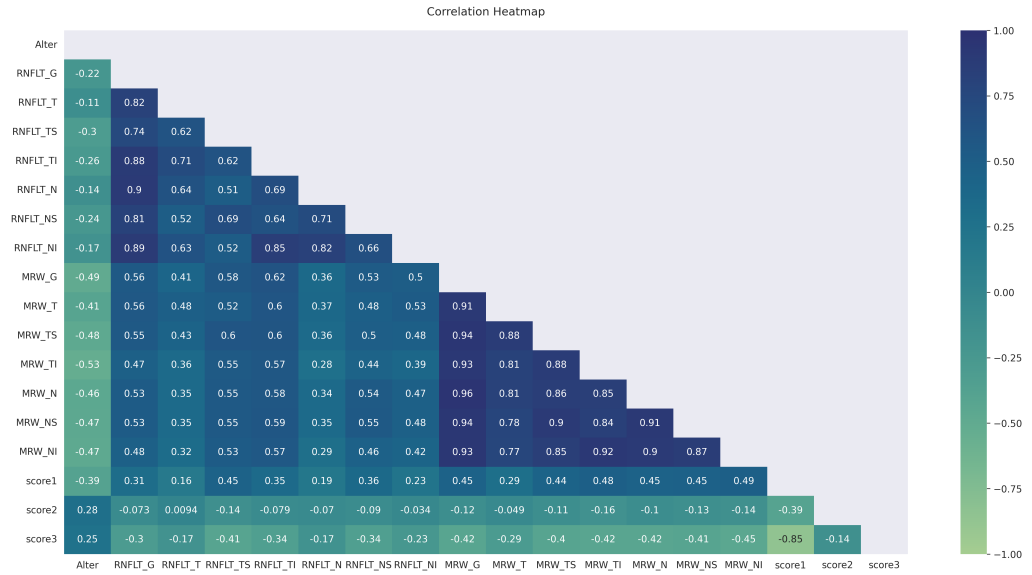


Abbildung 5.8.: Korrelations-Plot über alle Variablen.

Anhand der Erkenntnisse aus der Feature Importance sowie der Korrelation zwischen den Variablen, wurde das Modell erneut trainiert. In einem ersten Durchgang werden die Variablen RNFLT_G und MRW_G aus dem Modell entfernt, da der zentrale Wert nahezu vollständig aus den anderen sechs Sektoren erklärt werden kann.
Zur Evaluierung wurden auf den Testdaten des neu trainierten Modell erneut die Metriken berechnet. Es zeigt sich, dass die Metriken nahezu identisch zu dem Modell vor Bereinigung der zentralen Sektoren RNFLT_G und MRW_G sind.
In einem zweiten Durchlauf werden nun nur die Variablen der zentralen Sektoren beibehalten. Alle anderen sechs Sektoren werden entfernt. Das Szenario wird also quasi umgedreht. Dies sollte zu einer Verschlechterung des Modells führen, da deutlich weniger erklärende Variablen in das Modell einfließen. Es wäre allerdings von Interesse festzustellen, inwieweit sich die Modellgüte verändert, wenn nur ein zentraler Wert für die Vorhersage verwendet wird.
Die Evaluationsmetriken beider neu trainierten Modelle werden in Tabelle 5.10 und 5.11 gezeigt.
Die Ergebnisse decken sich mit den getroffenen Annahmen. Das Modell, das um die Variablen der zentralen Sektoren RNFLT_G und MRW_G bereinigt wurde, erreicht die nahezu identische Modellgüte wie das Modell, das alle Variablen umfasst.

Dagegen hat sich die Genauigkeit des zweiten Modells verschlechtert. Der Recall der Klasse 2 fällt jedoch auf weniger als die Hälfte des vorherigen Wertes. Demnach tragen die zentralen Sektoren RNFLT_G und MRW_G alleine nicht so gut zur Erklärbarkeit des Modells bei.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,9	0,95	0,92
Stadium 2	0,61	0,61	0,61
Stadium 3	0,8	0,65	0,72
Micro-Average	0,85	0,85	0,85
Macro-Average	0,77	0,74	0,75

Tabelle 5.10.: Übersicht über Evaluationsmetriken des Gradient Boosted Klassifikationsmodells, bei dem die Variablen um den zentralen Sektor der RNFL und MRW Messung bereinigt wurden.

Evaluationsmetriken - Testdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,84	0,93	0,88
Stadium 2	0,36	0,28	0,31
Stadium 3	0,76	0,6	0,68
Micro-Average	0,78	0,79	0,78
Macro-Average	0,65	0,6	0,62

Tabelle 5.11.: Übersicht über Evaluationsmetriken des Gradient Boosted Klassifikationsmodells, bei dem die Variablen um alle Sektoren bis auf den zentralen bereinigt wurden.

5.3. Evaluation - XGBoost Regressionsmodell

An dieser Stelle werden die Ergebnisse des XGBoost Regressionsmodells vorgestellt.
In Abbildung 5.9 wird die Feature Importance anhand des F-Scores veranschaulicht. Überraschend ist hier, dass das Alter eine hohe Wichtigkeit für das Modell aufweist. In ihrer Wichtigkeit stehen die RNFL Werte vor denen der MRW. Das Geschlecht eines Patienten hat, anders als das Alter, keine

bedeutende Wichtigkeit für das Modell.
Wie auch schon beim Gradient Boosted Klassifikationsmodell, werden für die Auswertung des Modells die zentralen Sektoren der RNFL und MRW entfernt. Diese Werte haben in ihrer Aussage eine Überschneidung zu den sechs Sektoren. Daher werden die redundanten Informationen entfernt.

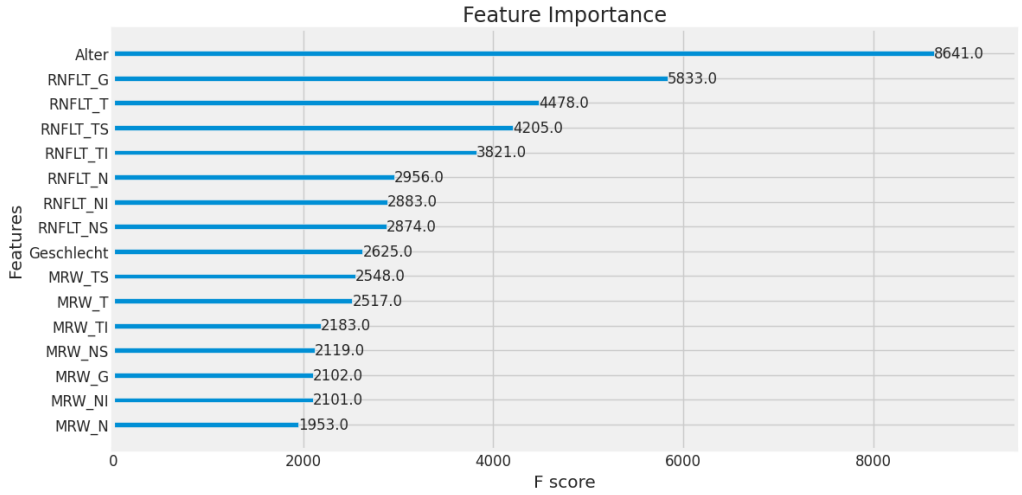


Abbildung 5.9.: Barplot mit der Feature Importance des XGBoost Regressionsmodell auf allen Variablen.

Das finale Modell wird also auf den sechs sektoralen Werten der RNFL und MRW sowie dem Alter und Geschlecht im Input trainiert. Je Quadrant wird ein eigenes Modell trainiert.

Die Ergebnisse des Regressionsmodells auf den Testdaten werden in Tabelle 5.12 gezeigt. Zur Evaluierung des Modells wird der wahre durchschnittliche Wert je Quadrant dem durchschnittlichen Wert der Modellvorhersage gegenübergestellt. Der durchschnittliche Wert je Quadrant wird anhand aller Dateneinträge des Testdatensatzes ermittelt. Zusätzlich wird je Quadrant der R^2 und der MAE ermittelt.
Für alle Quadranten liegt der MAE für die Vorhersage der retinalen Sensitivität bei einem Wert von etwa 4 dB. Trotz eines R^2 Wert von nur knapp 0.6, liegen die Vorhersagen durch das Modell sehr nahe an den wahren Werten. Bei der Vorhersage der retinalen Sensitivität liegt die Standardabweichung des Modells für alle Quadranten bei etwa 7 dB. Die Standardabweichung der retinalen Sensitivität der wahren Werte liegt um einen Wert von 9 dB. Zwischen den wahren Werten der retinalen Sensitivität und der Vorhersage

durch das Modell ist eine Abweichung von etwa 2 dB zu Beobachtungen.

XGBoost Regressions Evaluation - Testdaten				
	y_pred (dB)	y_true (dB)	R^2	MAE (dB)
Quadrant 1	21.75 (7.25)	21.39 (9.95)	0.538	4.711
Quadrant 2	21.68 (7.53)	21.78 (9.57)	0.561	4.497
Quadrant 3	22.88 (7.37)	22.55 (9.65)	0.563	4.333
Quadrant 4	22.86 (7.24)	22.98 (9.18)	0.574	4.100

Tabelle 5.12.: Gegenüberstellung der durchschnittlichen retinale Sensitivität je Quadrant aus der Perimetrie und der durchschnittlichen Werte aus der Vorhersage des Modells auf den Testdaten. Zusätzlich werden R^2 und der MAE je Modell gezeigt.

In der Abbildung 5.10 wird ein Scatterplot gezeigt, in dem die geschätzten Werte des Modells und die wahren gemessenen Werte aus der Gesichtsfeldmessung gegenübergestellt werden. Anhand des Plots wird deutlich, dass sich das Modell bei der Vorhersage von hohen Werten bereits sehr sicher ist. Hier besteht eine hohe Korrelation zwischen den wahren und vorhergesagten Werten. Die Schwäche des Modells wird aber erneut bei Betrachtung der Werte unterhalb eines Werts von etwa 25 deutlich. Das Modell ist sich bei *gesunden* Augen oder Augen mit nur sehr schwachen Defekten sehr sicher. Liegen aber mittlere bis starke Gesichtsfelddefekte vor, so weicht die Vorhersage stärker von den wahren Werte ab. Diese schwächeren Bereiche in der Vorhersage entsprechen in der Klassifikation den Hodapp-Stadien 2 und 3.

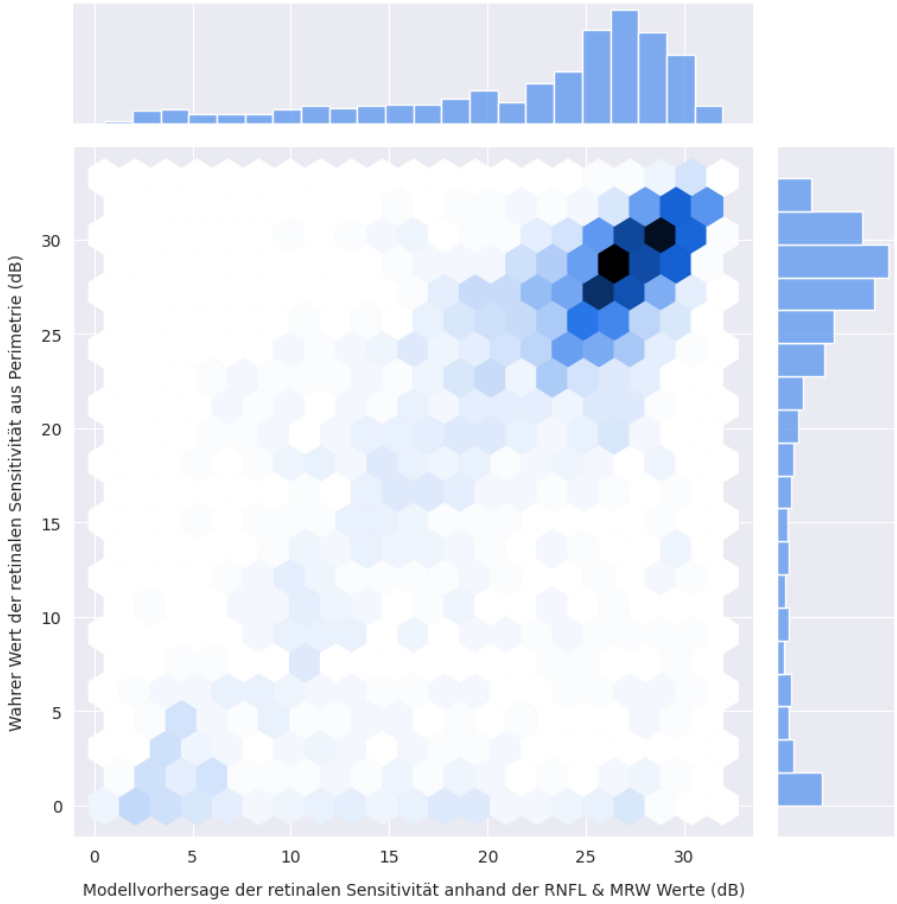


Abbildung 5.10.: Scatterplot und Histogramme zur Illustration der Bezeichnung zwischen den Vorhersagen aus dem Deep Learning Modell anhand der RNFL und MRW Werten und den wahren retinale Sensitivität (dB) aus der Perimetrie. Der Plot wurde auf den Testdaten erstellt.

Die Beobachtungen des Regressionsmodells decken sich also mit denen des Klassifikationsmodells. Es ist anzunehmen, dass auch hier die Datengrundlage zu einer Verzerrung der Genauigkeit zwischen den Klassen führt. Da im Trainingsdatensatz die meisten Datenpunkte von Augen mit keinen oder nur geringen Defekten sind, ist sich das Modell hier in der Vorhersage am sichersten. Auf die anderen Datenpunkten kann das Modell die Werte nur deutlich schlechter vorhersagen.

5.4. Zusammenfassung der Ergebnisse

In diesem Abschnitt werden die wichtigsten Erkenntnisse und Ergebnisse der Experimente zusammengefasst.

Die Auswertung des CNN Klassifikationsmodells hat gezeigt, dass alleine anhand der Fundusfotos keine zuverlässige Vorhersage der Hodapp-Stadien möglich ist. Die Accuracy des Modells liegt nur knapp über 50% und ist somit nur ein wenig besser als eine zufällige Entscheidung. Für den Einsatz in der Praxis, besonders im medizinischen Kontext, ist dies keine ausreichende Performance. Bei Betrachtung der Vorhersagegenauigkeit je Klasse kann abgeleitet werden, dass die Größe des Datensatzes einen deutlichen Einfluss auf die Genauigkeit des Modells hat. Deutlich unterrepräsentierte Klassen wie Bilder des Stadium 2, können mit einer nur geringen Sicherheit erkannt werden, während Gesichtsfelddefekte des Stadium 1 auf Fundusfotos bereits mit hoher Sicherheit erkannt und zugeordnet werden.
Zudem ist davon auszugehen, dass die Klasseneinteilung anhand der Hodapp-Stadien eine schwierige Aufgabe für das CNN darstellt. Denn die Klassen sind an der Mean Deviation orientiert, die im Grunde eine numerische Skala darstellt. Die Klassengrenzen sind daher für das Modell sehr unscharf und könnten somit schnell zu einer Fehlerklassifizierung führen. Diese Überlegung deckt sich mit der Beobachtung aus dem Modelltraining des CNN Klassifikationsmodell mit nur zwei Klassen, bei dem die Unterscheidung zwischen *gesunden* Augen und Augen, in denen Gesichtsfelddefekte vorliegen, trainiert wurde. Hier ist sich das Modell deutlicher sicher in der Einteilung beider Klassen. Dennoch lässt sich anhand des Modelltrainings beider Modelle eine Unsicherheit der Schätzung beobachten. Eine weitere Ursache für die Ungenauigkeit des Modells könnte sein, dass die Informationen über die Ausprägung der Defekte schlicht nicht in Fundusfotos enthalten ist.
Die Bereinigung des subset 01 anhand der Fixationsverluste hat sich als sinnvoll erwiesen. Die Vorhersagegenauigkeit des Modells hat sich verbessert, nachdem die Datensätze über dem Schwellwert entfernt wurden.
Durch die Verknüpfung der Vorhersage anhand von Fundusfotos und dem Datensatz der RNFL, MRW sowie weiteren Metadaten konnte eine Verbesserung des Modells erreicht werden. Die RNFL Werte eignen sich somit für die Vorhersage des Gesichtsfeldes eines Patienten. Auch zeigte das Einbeziehen der MRW Werte eine Verbesserung der Vorhersage. Somit gelten auch die Werte der MRW als geeigneter Prädiktor für Gesichtsfelddefekte.
Dabei bewirken die sechs sektoralen Werte der RNFL bzw. MRW eine höhere Vorhersagegüte, als nur der zentrale Wert.

Das Einbeziehen der Metadaten, Alter und Geschlecht eines Patienten haben nur teilweise einen positiven Effekt auf die Modellgüte gehabt. Das Geschlecht hat in beiden Modellen kaum zur Erklärbarkeit oder Verbesserung des Modells beigetragen. Folglich hat das Geschlecht keine Bedeutung für die Vorhersage von Gesichtsfelddefekten. Das Alter dagegen trug in beiden Modellen zur Modellgüte bei. Daher ist daraus zu schließen, dass das Alter einen Einfluss auf die Vorhersage von Gesichtsfelddefekten hat.
Durch das Regressionsmodell konnte gezeigt werden, dass anhand der Ergebnisse des OCT-Scans, eine gute Vorhersage der retinalen Sensitivität möglich ist. Die Abweichung zwischen wahren und vorhergesagten Werten von durchschnittlich 4,41 dB weist allerdings auch auf noch vorhandene Ungenauigkeiten des Modells hin. Für die Praxis ist hier zu prüfen, welche Abweichung medizinisch noch vertretbar ist.

6. Fazit

Die Untersuchungen in dieser Arbeit haben gezeigt, dass sich ungelabelte Fundusfotos nicht zweifelsfrei zur Erkennung und Vorhersage von Gesichtsfelddefekten eignen. Die Vorhersagegenauigkeit des CNN Klassifikationsmodells, zur Vorhersage von milden bis schweren Gesichtsfelddefekten, beträgt auf den Testdaten eine Accuracy von 54%.
Dabei ist zu beobachten, dass die Vorhersagegenauigkeit von der Klasse abhängt. Eine Schwachstelle des Modells ist daher auf die Datengrundlage zurückzuführen, die sehr wahrscheinlich zu einer Verzerrung in der Vorhersage führt. Die Klasse mit den meisten Dateneinträgen erzielt die höchste Genauigkeit, während die am wenigsten repräsentierte Klasse die geringste Vorhersagegenauigkeit erreicht. In allen Modellen, die in dieser Arbeit trainiert wurden, konnte diese Verzerrung der Vorhersage in Richtung der am meisten vertretenen Klasse beobachtet werden.
Es konnte aber bestätigt werden, dass die Bereinigung des Datensatzes anhand der Fixationsverlusten sich positiv auf die Qualität der Vorhersage ausgewirkt hat. Für weitere Untersuchungen ist es daher sinnvoll, einen Datensatz zu verwenden, in dem von Beginn an nur Gesichtsfeldmessungen einbezogen werden, die als zuverlässig eingestuft werden können.
Aufgrund dieser Ergebnisse können ungelabelte Fundusfotos aktuell nicht als ernsthafte Alternative zur Gesichtsfeldmessung in der Praxis empfohlen werden. Für den klinischen Einsatz sind daher noch weitere Untersuchungen notwendig.
Weitere Untersuchungen konnten zeigen, dass die Hinzunahme der Messungen aus den OCT-Scans einen positiven Effekt auf die Modellgüte hat. Die Vorhersage der Hodapp-Stadien mit dem Gradient Boosted Klassifikationsmodell konnte auf eine Accuracy von 86% auf den Testdaten verbessert werden. Jedoch sind die Ergebnisse, die aus dem CNN Modell kommen zu hinterfragen, da die Performance des CNNs noch deutliche Schwächen aufweist. Dies wird auch daran deutlich, dass die Varianz in der Genauigkeit zwischen den Klassen bestehen bleibt.
Erste Untersuchungen zur Verortung der Gesichtsfelddefekte konnten zeigen, dass die Werte der RNFL und MRW ein guter Prädiktor für die Vorhersage der retinalen Sensitivität sind. Dennoch weist die Modellgüte mit einem

R^2 von unter 60% noch deutliche Unsicherheiten auf. Für weiterführende Arbeiten zu diesem Thema, wäre es von Interesse, die retinale Sensitivität genauer vorherzusagen, und nicht nur anhand eines Durchschnittswertes je Quadrant.

Anhand der Beobachtungen und Erkenntnisse in dieser Arbeit können einige Empfehlungen und Ideen formuliert werden, die für zukünftige Arbeiten mit einer verwandten Fragestellung hilfreich sein könnten. Für weitere Untersuchungen ist es empfehlenswert, auf einen ausgeglichenen Datensatz zu achten, der alle Klassen gleichermaßen repräsentiert.
Weiterhin würde es sich auf die Untersuchungen positiv auswirken, wenn der Zugriff auf Untersuchungsergebnisse wie die der Perimetrie vereinfacht möglich wäre. Um die Experimente in dieser Arbeit durchführen zu können, war es zunächst mit hohem Aufwand verbunden, die Daten aufzubereiten, sodass sie für das Modelltraining verwendet werden konnten.
Zusätzlich könnte die Qualität der Vorhersage möglicherweise verbessert werden, indem ein gelabelter Datensatz verwendet wird, in dem Vorerkrankungen bereits eindeutig zugeordnet werden können. Ein möglicher Grund für die Ungenauigkeiten in der Quantifizierung und Erkennung der Gesichtsfelddefekte könnte die zu diverse Repräsentation von Augenerkrankungen in den Inputdaten sein. Aufgrund der fehlenden Label ist nicht bekannt, welche Erkrankungen tatsächlich auf den Fundusfotos vorhanden sind. Im Lernprozess kann dieser Umstand es erschweren, eindeutige und markante Features in den Daten zu erkennen. Durch die Vorfilterung von Vorerkrankungen fällt im zweiten Schritt die Einschätzung von Gesichtsfelddefekten möglicherweise genauer aus.
Um das periphere Sehen eines Patienten gänzlich ohne die Perimetrie zu messen, wäre es nach den Erkenntnissen dieser Arbeit denkbar, ein dreistufiges Verfahren anzuwenden. Im ersten Schritt könnte anhand der Fundusfotos eine Vorfilterung vorgenommen werden, bei der eventuell vorliegende Vorerkrankungen erkannt werden. Im zweiten Schritt könnten dann die Ergebnisse der Vorklassifizierung genutzt werden, um in einem zweiten Modell Gesichtsfelddefekte zu erkennen. Dabei hätten Fundusfotos, auf denen eine Glaukomerkrankung erkannt wird, eine deutlich höhere Wahrscheinlichkeit auch Gesichtsfelddefekte zu beinhalten. Im dritten Schritt könnten, wie auch in dieser Arbeit, die Messungen der Nervenfaserschicht einbezogen werden, um die Vorhersage zu verbessern.

Für die Bestimmung des peripheren Sehens und der exakten Messung von Gesichtsfelddefekten bleibt die Perimetrie, nach den Erkenntnissen dieser Arbeit, zunächst unersetzlich. Auch wenn anhand der Ergebnisse dieser Arbeit kein gänzlich positives Fazit gezogen werden konnte, so bleiben die Fragestellungen dieser Arbeit von hohem Interesse. Es ist davon auszugehen, dass sich auch die Forschung in der Medizintechnik weiterentwickeln wird und in wenigen Jahren eventuell bereits verbesserte bildgebende Verfahren möglich werden, anhand derer auch Gesichtsfelddefekte zuverlässiger erkannt werden können.

Appendix

A. Zusätzliche Metriken und Abbildungen der Ergebnisse

A.1. CNN Klassifikationsmodell mit drei Klassen

Evaluationsmetriken - Trainingsdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,82	0,77	0,79
Stadium 2	0,39	0,44	0,41
Stadium 3	0,62	0,66	0,64
Micro-Average	0,7	0,69	0,69
Macro-Average	0,61	0,62	0,61

Tabelle A.1.: Übersicht über Evaluationsmetriken auf den Trainingsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Confusion Matrix - Trainingsdaten			
	Stadium 1	Stadium 2	Stadium 3
Stadium 1	1704 (0.77)	250 (0.11)	272 (0.12)
Stadium 2	197 (0.35)	250 (0.44)	124 (0.22)
Stadium 3	186 (0.19)	140 (0.14)	640 (0.66)

Tabelle A.2.: Confusion Matrix auf den Trainingsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

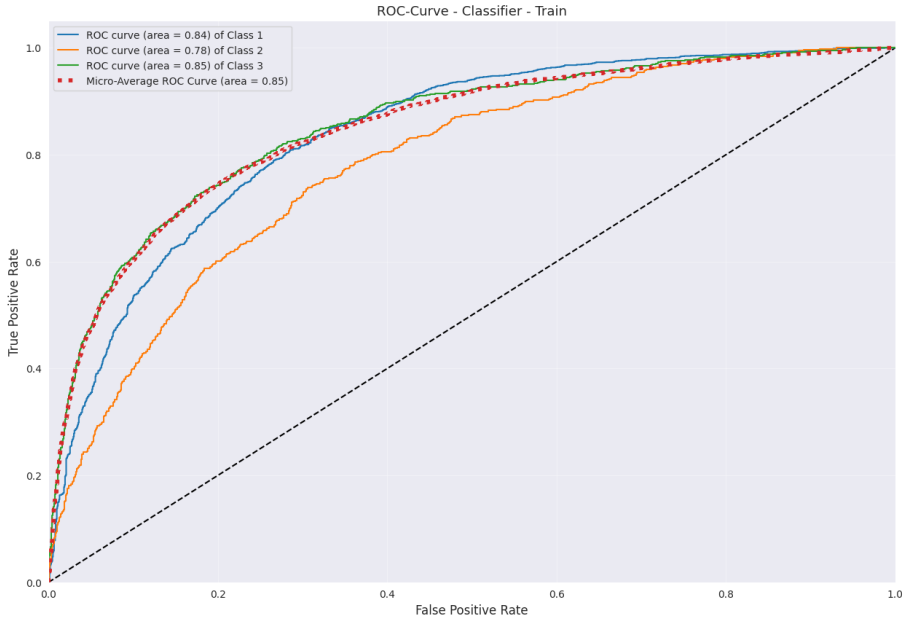


Abbildung A.1.: ROC-Kurve und AUC-Scores auf den Trainingsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Evaluationsmetriken - Validierungsdaten			
	Precision	Recall	F_1-Score
Stadium 1	0,75	0,71	0,73
Stadium 2	0,1	0,15	0,12
Stadium 3	0,46	0,38	0,42
Micro-Average	0,58	0,54	0,56
Macro-Average	0,44	0,41	0,42

Tabelle A.3.: Übersicht über Evaluationsmetriken auf den Validierungsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Confusion Matrix - Validierungsdaten			
	Stadium 1	Stadium 2	Stadium 3
Stadium 1	454 (0.71)	122 (0.19)	60 (0.09)
Stadium 2	78 (0.48)	24 (0.15)	62 (0.38)
Stadium 3	71 (0.26)	100 (0.36)	105 (0.38)

Tabelle A.4.: Confusion Matrix auf den Validierungsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

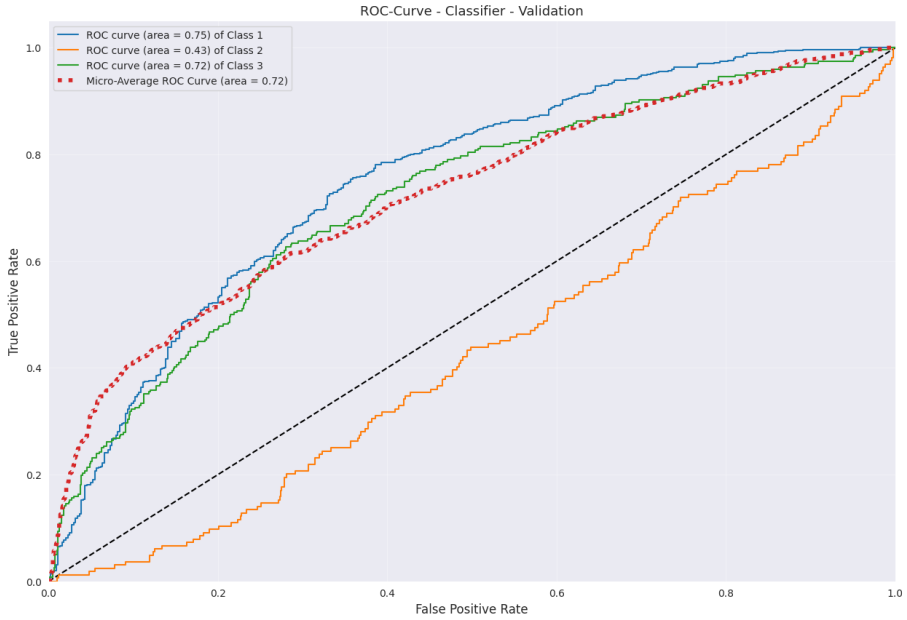


Abbildung A.2.: ROC-Kurve und AUC-Scores auf den Validierungsdaten des CNN Klassifikationsmodell zur Erkennung der Hodapp-Stadien.

Literatur

[1] Deutschen Ophthalmologischen Gesellschaft (DOG). *Stellungnahme der Deutschen Ophthalmologischen Gesellschaft zur Glaukomvorsorge*. Aug. 2012. URL: https://www.dog.org/wp-content/uploads/2009/08/Glaukomvorsorge-Stand-August-20121.pdf (besucht am 15. 04. 2023).

[2] Deutschen Ophthalmologischen Gesellschaft (DOG). *Stellungnahme der Deutschen Ophthalmologischen Gesellschaft zur Glaukomfrüherkennung*. Aug. 2015. URL: https://www.dog.org/wp-content/uploads/2023/03/SN-Glaukom-August-2015.pdf (besucht am 15. 04. 2023).

[3] Kassenärztliche Bundesvereinigung (KBV). URL: https://gesundheitsdaten.kbv.de/cms/html/16393.php (besucht am 16. 05. 2023).

[4] Kassenärztliche Bundesvereinigung (KBV). URL: https://gesundheitsdaten.kbv.de/cms/html/17023.php (besucht am 15. 08. 2023).

[5] Sayaka Adachi u. a. „Factors associated with developing a fear of falling in subjects with primary open-angle glaucoma“. In: *BMC Ophthalmology* 18 (2018). DOI: 10.1186/s12886-018-0706-5.

[6] *Anatomie - Bau des Augapfels*. 2019. URL: https://www.dog.org/wp-content/uploads/2023/03/glaukom-patienteninfo_20230303.pdf (besucht am 16. 04. 2023).

[7] Berufsverband der Augenärzte (BVA). URL: https://www.augeninfo.de/offen/index.php?themenseite=Augenaerzte (besucht am 15. 08. 2023).

[8] Berufsverband der Augenärzte (BVA) und Deutschen Ophthalmologischen Gesellschaft (DOG). *Patienteninfo zu Glaukom*. URL: https://www.dog.org/wp-content/uploads/2023/03/glaukom-patienteninfo_20230303.pdf (besucht am 16. 04. 2023).

[9] Franziska Barbara Baudisch. „Die Optische Kohärenztomographie (OCT) - Analyse der retinalen Nervenfaserschichtdicke und Retinafundusdicke nach Alter, Papillengröße und Refraktionsfehler sowie bei Patienten mit Optikopathien und arterieller Hypertonie“. Diss. Ludwig-Maximilians-Universität zu München, 2018.

[10] Robin Baudisch. „Vergleich verschiedener Ansätze des Supervised- und Semi-Supervised Learning zur robusten Erkennung von Glaukomerkrankungen mittels medizinischer Bild- und Metadaten“. Diss. Hochschule Darmstadt, 2021.

[11] Christopher M. Bishop. *Pattern recognition and machine learning*. New York, 2009.

[12] Mateusz Buda, Atsuto Maki und Maciej A. Mazurowski. „A systematic study of the class imbalance problem in convolutional neural networks“. In: *Neural Networks* 106 (2018), S. 249–259. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2018.07.011.

[13] Inc. Carl Zeiss Meditec. *Humphrey Field Analyzer - A Guide to Interpretation*. URL: https://www.zeiss.com/content/dam/Meditec/us/download/Glaucoma%20Landing%20Page/hfasinglefieldguidehfa5268.pdf (besucht am 10. 05. 2023).

[14] Ling-Ping Cen u. a. „Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks“. In: *Nature Communications* 12.1 (Aug. 2021). DOI: 10.1038/s41467-021-25138-w. URL: https://doi.org/10.1038/s41467-021-25138-w.

[15] Tianqi Chen und Carlos Guestrin. „XGBoost“. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145/2939672.2939785.

[16] Mark Christopher u. a. „Deep Learning Approaches Predict Glaucomatous Visual Field Damage from OCT Optic Nerve Head En Face Images and Retinal Nerve Fiber Layer Thickness Maps“. In: *Ophthalmology* 127.3 (März 2020), S. 346–356. DOI: 10.1016/j.ophtha.2019.09.036. URL: https://doi.org/10.1016/j.ophtha.2019.09.036.

[17] *Decision Referral: Combining strengths of radiologists and AI*. URL: https://www.vara.ai/decision-referral (besucht am 25. 06. 2023).

[18] Holger Dietze. *Die optometrische Untersuchung*. Thieme, 2008.

[19] R.K. Parrish und D.R. Anderson E. Hodapp. *Clinical Decisions in Glaucoma*. Mosby, 1993.

[20] Philip Enders u. a. „Evaluation of two-dimensional Bruch's membrane opening minimum rim area for glaucoma diagnostics in a large patient cohort“. In: *Acta Ophthalmologica* 97.1 (März 2018), S. 60–67. DOI: 10.1111/aos.13698. URL: https://doi.org/10.1111/aos.13698.

[21] Jerome H. Friedman. „Greedy function approximation: A gradient boosting machine.“ In: *The Annals of Statistics* 29.5 (Okt. 2001). DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[22] *Fundusfotografie: Fotoverlaufskontrolle des Augenhintergrunds*. 11. Okt. 2019. URL: https://www.aumedo.de/fundusfotografie/ (besucht am 27. 05. 2023).

[23] *GBE-Themenheft: Blindheit und Sehbehinderung*. 2017. URL: https://www.rki.de/DE/Content/Gesundheitsmonitoring/Gesundheitsberichterstattung/GBEDownloadsT/blindheit.pdf?__blob=publicationFile (besucht am 18. 04. 2023).

[24] „Glaukom im Fokus“. In: *Kompass Ophthalmologie* 6.4 (2020), S. 196–201. DOI: 10.1159/000512220. URL: https://doi.org/10.1159/000512220.

[25] Heidelberg Engineering GmbH. *Gesichtsfelduntersuchungen*. URL: https://www.augenwissen.de/untersuchungen/gesichtsfelduntersuchungen/ (besucht am 12. 04. 2023).

[26] Ian Goodfellow. *Deep Learning – Grundlagen, aktuelle Verfahren und Algorithmen, neue Forschungsansätze*. 2018.

[27] Chen Guo, Minzhong Yu und Jing Li. „Prediction of Different Eye Diseases Based on Fundus Photography via Deep Transfer Learning“. In: *Journal of Clinical Medicine* 10.23 (Nov. 2021), S. 5481. DOI: 10.3390/jcm10235481. URL: https://doi.org/10.3390/jcm10235481.

[28] Trevor Hastie. *The elements of statistical learning : data mining, inference, and prediction*. New York [u.a.], 2009.

[29] Geoffrey E. Hinton u. a. *Improving neural networks by preventing co-adaptation of feature detectors*. 2012. DOI: 10.48550/ARXIV.1207.0580. URL: https://arxiv.org/abs/1207.0580.

[30] Gareth James u. a. *An Introduction to Statistical Learning: with Applications in R*. 2017. DOI: 10.1007/978-1-4614-7138-7.

[31] Ph.D. Joy N. Carroll und Chris A. Johnson. *The University of Iowa - Department of Ophthalmology and Visual Sciences*. URL: https://eyerounds.org/tutorials/VF-testing/ (besucht am 16. 04. 2023).

[32] Edward Korot u. a. „Predicting sex from retinal fundus photographs using automated deep learning“. In: *Scientific Reports* 11.1 (Mai 2021). DOI: 10.1038/s41598-021-89743-x. URL: https://doi.org/10.1038/s41598-021-89743-x.

[33] Alex Krizhevsky. „Learning Multiple Layers of Features from Tiny Images“. In: *University of Toronto* (Mai 2012).

[34] C. Lamirel. „Optical Coherence Tomography“. In: *Encyclopedia of the Neurological Sciences (Second Edition)*. Hrsg. von Michael J. Aminoff und Robert B. Daroff. Second Edition. Oxford: Academic Press, 2014, S. 660–668. ISBN: 978-0-12-385158-1. DOI: https://doi.org/10.1016/B978-0-12-385157-4.00171-8. URL: https://www.sciencedirect.com/science/article/pii/B9780123851574001718.

[35] Y. LeCun u. a. „Backpropagation Applied to Handwritten Zip Code Recognition“. In: *Neural Computation* 1.4 (1989), S. 541–551. DOI: 10.1162/neco.1989.1.4.541.

[36] Christian Leibig u. a. „Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis“. In: *The Lancet Digital Health* 4.7 (Juli 2022), e507–e519. DOI: 10.1016/s2589-7500(22)00070-x. URL: https://doi.org/10.1016/s2589-7500(22)00070-x.

[37] *Limitations of Mammograms*. 14. Jan. 2023. URL: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html (besucht am 25. 06. 2023).

[38] J. Luebke u. a. „Abhängigkeit der Größe der OCT-Bruch-Membran-Öffnung von Hornhautkorrekturfaktoren – eine Pilotstudie“. In: *Klinische Monatsblätter für Augenheilkunde* 234.07 (Aug. 2016), S. 918–923. DOI: 10.1055/s-0042-109702. URL: https://doi.org/10.1055/s-0042-109702.

[39] MBA) Malik Y. Kahook(MD) und Robert J. Noecker(MD. *How Do You Interpret a 24-2 Humphrey Visual Field Printout?* URL: https://assets.bmctoday.net/glaucomatoday/pdfs/GT1107_10.pdf (besucht am 09. 05. 2023).

[40] Pamela McCorduck. „Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence“. In: 1979. URL: https://api.semanticscholar.org/CorpusID:111014295.

[41] Felipe A. Medeiros, Alessandro A. Jammal und Atalie C. Thompson. „From Machine to Machine: An OCT-Trained Deep Learning Algorithm for Objective Quantification of Glaucomatous Damage in Fundus Photographs“. In: *Ophthalmology* 126.4 (Apr. 2019), S. 513–521. DOI: 10.1016/j.ophtha.2018.12.033. URL: https://doi.org/10.1016/j.ophtha.2018.12.033.

[42] Mayank Mishra. *Convolutional Neural Networks, Explained*. 2020. URL: https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939 (besucht am 27. 07. 2023).

[43] Andreas C. Muller. *Introduction to machine learning with Python : a guide for data scientists*. 2017.

[44] Lucas Pascal u. a. „Multi-task deep learning for glaucoma detection from color fundus images“. In: *Scientific Reports* 12.1 (Juli 2022). DOI: 10.1038/s41598-022-16262-8. URL: https://doi.org/10.1038/s41598-022-16262-8.

[45] Adam Paszke u. a. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. 2019. DOI: 10.48550/ARXIV.1912.01703. URL: https://arxiv.org/abs/1912.01703.

[46] Josh Patterson. *Deep Learning*. 2017.

[47] F. Pedregosa u. a. „Scikit-learn: Machine Learning in Python“. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830.

[48] Luis Perez und Jason Wang. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning*. 2017. DOI: 10.48550/ARXIV.1712.04621. URL: https://arxiv.org/abs/1712.04621.

[49] Connor Shorten und Taghi M. Khoshgoftaar. „A survey on Image Data Augmentation for Deep Learning“. In: *Journal of Big Data* 6.1 (Juli 2019). DOI: 10.1186/s40537-019-0197-0. URL: https://doi.org/10.1186/s40537-019-0197-0.

[50] Edward H Shortliffe. „Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases.“ In: *Proceedings of the Annual Symposium on Computer Application in Medical Care* (1977), S. 66–69.

[51] Kanae Takahashi u. a. „Confidence interval for micro-averaged F1 and macro-averaged F1 scores“. In: *Applied Intelligence* 52 (März 2022). DOI: 10.1007/s10489-021-02635-5.

[52] Mingxing Tan und Quoc V. Le. „EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks“. In: (2019). DOI: 10.48550/ARXIV.1905.11946. URL: https://arxiv.org/abs/1905.11946.

[53] Atalie C. Thompson, Alessandro A. Jammal und Felipe A. Medeiros. „A Deep Learning Algorithm to Quantify Neuroretinal Rim Loss From Optic Disc Photographs“. In: *American Journal of Ophthalmology* 201 (Mai 2019), S. 9–18. DOI: 10.1016/j.ajo.2019.01.011. URL: https://doi.org/10.1016/j.ajo.2019.01.011.

[54] Vincent Weber. „Vergleich morphometrischer Vermessungen von Makropapillen mittels optischer Kohärenztomographie und konfokaler LaserScanning Tomographie zur Glaukomdiagnostik“. Diss. Universität zu Köln, 2022.

[55] *Weißbuch zur Situation der ophthalmologischen Versorgung in Deutschland*. Sep. 2012. URL: https://www.dog.org/wp-content/uploads/2013/03/DOG_Weissbuch_2012_fin.pdf (besucht am 15. 04. 2023).

[56] Suorong Yang u. a. *Image Data Augmentation for Deep Learning: A Survey*. 2022. DOI: 10.48550/ARXIV.2204.08610. URL: https://arxiv.org/abs/2204.08610.

[57] Camila S. Zangalli u. a. „Minimum Rim Width and Peripapillary Retinal Nerve Fiber Layer Thickness for Diagnosing Early to Moderate Glaucoma“. In: *Journal of Glaucoma* 32.6 (Dez. 2022), S. 526–532. DOI: 10.1097/ijg.0000000000002156. URL: https://doi.org/10.1097/ijg.0000000000002156.

[58] Fuzhen Zhuang u. a. *A Comprehensive Survey on Transfer Learning*. 2019. DOI: 10.48550/ARXIV.1911.02685. URL: https://arxiv.org/abs/1911.02685.

[59] Universitätsspital Zürich. URL: https://www.usz.ch/krankheit/glaukom/ (besucht am 18. 04. 2023).