Aus der Klinik für Innere Medizin II – Gastroenterologie/Endoskopie,

UKS - Universitätsklinikum des Saarlandes, Homburg/Saar

Direktor: Prof. Dr. med. Jörn Schattenberg

Prospektive Evaluation von Anwendungen der künstlichen Intelligenz in der gastroenterologischen Endoskopie

Dissertation zur Erlangung des Grades eines Doktors der Medizin

der Medizinischen Fakultät

der UNIVERSITÄT DES SAARLANDES

2025

vorgelegt von: Mandy Thérèse Marthe Simon

geb. am: 26.12.1996 in Luxemburg-Stadt, Luxemburg

Tag der Promotion: 11.09.2025

Dekan: Univ.-Prof. Dr. med. dent. Matthias Hannig

Berichterstatter: Priv.-Doz. Dr. med. M. Casper

Prof. Dr. A. Keller

Gewidmet meinen Eltern, Daniel und Nicole,

sowie meinem Ehemann Dominic.

Inhaltsverzeichnis

1 Zusammenfassung

1.1 Zusammenfassung

Hintergrund: Die publizierten Studien von Guimarães P. et al. von 2020 [29] und 2022 [30] testeten in einem vorselektierten Validierungsdatensatz zwei selbst etablierte Algorithmen zur Erkennung präkanzeröser Bedingungen des proximalen Magens und der eosinophilen Ösophagitis und erzielten eine diagnostische Genauigkeit von 93% bzw. 91%. Die vorliegende Studie testet die entsprechenden Algorithmen nun prospektiv unter reellen Bedingungen in einem nicht vorselektierten Kollektiv.

Patienten und Methoden: Insgesamt wurden 650 Patienten mittels Ösophagogastroduodenoskopie untersucht. Dabei wurden vom Untersuchenden bis zu 6 Bilder in vorher definierten endoskopischen Positionen (Loc 1-4 proximaler Magen, davon 2 in Inversion; Loc 5 und 6 Ösophagus) aufgenommen und für die Auswertung ausgewählt. Es erfolgte eine Einschätzung hinsichtlich des Vorliegens einer proximalen Atrophie bzw. eosinophilen Ösophagitis durch den Untersucher. Nach Analyse der ausgewählten Bilder durch die künstlichen Intelligenzen wurden die Ergebnisse statistisch ausgewertet und mit denen der Untersucher verglichen. Das tatsächliche Vorliegen der Erkrankung (Grundwahrheit) wurde für die Atrophie histopathologisch und für die eosinophile Ösophagitis histopathologisch bzw. kombiniert klinisch/histopathologisch definiert. Letztendlich konnten für die Atrophie 1.681 Bilder (Loc 1 392, Loc 2 339, Loc 3 446, Loc 4 504 Bilder) von 583 Patienten mit vorliegender Histologie ausgewertet werden. Für die eosinophile Ösophagitis wurden 351 Bilder (Loc 5 169, Loc 6 182 Bilder) von 204 Patienten mit Biopsie bzw. 936 Bilder (Loc 5 438, Loc 6 498 Bilder) von 550 Patienten mit klinisch/histopathologisch definierter Erkrankungssituation ausgewertet.

Ergebnisse: In der Gesamtkohorte wiesen 22 von 583 auswertbaren Patienten eine atrophische Gastritis und 7 von 550 auswertbaren Patienten eine eosinophile Ösophagitis auf. Für die präkanzerösen Bedingungen des proximalen Magens finden sich für die Loc 1-4 33, 146, 90 und 37 falsch-positive bzw. 3, 1, 6 und 10 falsch-negative Beurteilungen durch den Algorithmus. Es zeigt sich eine balanced Accuracy für Loc 1-4 von 86%, 74%, 72% und 67%, eine Sensitivität von 81%, 94%, 65% und 41% und eine Spezifität von 91%, 55%, 79% und 92%. Durch Kombination der einzelnen Lokalisationen konnte keine relevante Verbesserung der Ergebnisse erzielt werden. Die Untersucher geben auf Patientenebene 16 falsch-positive und 13 falsch-negative Beurteilungen ab und erzielen eine balanced Accuracy von 69%, eine Sensitivität von 41% und eine Spezifität von 97%.

Der Algorithmus für die eosinophile Ösophagitis weist in der Histologie-Kohorte für Loc 5 und Loc 6 10 bzw. 16 falsch-positive und 0 bzw. 3 falsch-negative Befunde auf. Dadurch zeigt sich

für Loc 5 und Loc 6 eine balanced Accuracy von 97% bzw. 74%, eine Sensitivität von 100% bzw. 57% sowie eine Spezifität von 94% bzw. 91%. In der klinisch/histopathologisch definierten Kohorte ergeben sich für Loc 5 und Loc 6 21 bzw. 30 falsch-positive und 0 bzw. 3 falsch-negative Befunde mit einer balanced Accuracy von 95% bzw. 98%, eine Sensitivität von 100% für beide und eine Spezifität von 91% bzw. 95%. Durch Kombination der Lokalisationen 5 und 6 konnte keine relevante Verbesserung der Ergebnisse erreicht werden. Die Untersucher geben auf Patientenebene bei beiden Kohorten 7 falsch-positive und 1 falsch-negative Beurteilung ab. Für die Histologie-Kohorte ergeben sich eine balanced Accuracy von 91%, eine Sensitivität von 86% und eine Spezifität von 96%. In der klinisch/histopathologischen Kohorte liegen die entsprechenden Werte bei 92%, 86% und 99%.

In der Fehleranalyse fand sich eine relevante Anzahl an Untersuchungen mit in der Qualität des Bildmaterials begründeten erklärbaren Ursachen für eine Fehleinschätzung (Loc 1-6: 58%, 9,5%, 52%, 55%, 95%, 100%). Weiterhin sind Normalbefunde aus Loc 2 im ursprünglichen Trainingsdatensatz stark unterrepräsentiert (4%). Somit dürfte die hohe Rate an sonst nicht erklärbaren falsch-positiven Befunden durch eine Unterrepräsentation vergleichbarer Normalbefunde im ursprünglichen Trainingsdatensatz begründet sein.

Schlussfolgerung: Diese Arbeit zeigt klar, dass zur Beurteilung der klinischen Nutzbarkeit von KI-Algorithmen prospektive Validierungen in unselektierten Kohorten unabdingbar sind. Der Algorithmus zur Erkennung von präkanzerösen Bedingungen des proximalen Magens offenbarte deutliche Schwächen mit zahlreichen falsch-positiven und falsch-negativen Auswertungen. Dies ist durch den limitierten Trainingsdatensatz bei einem Krankheitsbild mit großer Bandbreite an endoskopischen Erscheinungsbildern und großem Organ mit variablem Bildwinkeln erklärt. Bei Verwendung von Bildern aus einer endoskopischen Standardposition (Loc 1) zeigt der Algorithmus unter Realbedingungen durchaus eine zufriedenstellende Performance, während der Algorithmus zur Beurteilung von Bildern aus anderen Positionen (insbesondere Loc 2) aufgrund unzureichender Abbildung im Trainingsdatensatz nicht nutzbar ist.

Für die eosinophile Ösophagitis ergibt sich eine deutlich stabilere Performance des Algorithmus unter Realbedingungen. Dies ist durch die deutlich geringere Variabilität der Bilddarstellung erklärt, sodass das Spektrum der Erscheinungsbilder in dem größeren initialen Trainingsdatensatz besser abgebildet wurde. Die Unterschiede zwischen Loc 5 und 6 sind am ehesten durch weniger repräsentatives Bildmaterial in Loc 6 bedingt, weswegen die Benutzung von Bildern der unteren Speiseröhre (Loc 5) zu empfehlen ist.

1.2 Abstract

Background: The published studies by Guimarães P. et al. from 2020 [29] and 2022 [30] tested two self-established algorithms for the detection of precancerous conditions of the proximal stomach and eosinophilic esophagitis in a preselected validation dataset and achieved a diagnostic accuracy of 93% and 91%, respectively. The present study now prospectively tests the corresponding algorithms under real-life conditions in a non-pre-selected population.

Patients and methods: A total of 650 patients were examined using esophagogastroduodenoscopy. The examiner took up to 6 images in predefined endoscopic positions (Loc 1-4 proximal stomach, 2 of them in inversion; Loc 5 and 6 esophagus) and selected them for evaluation. The examiner assessed the presence of proximal atrophy or eosinophilic esophagitis. After analysis of the selected images by artificial intelligence, the results were statistically evaluated and compared with those of the examiners. The actual presence of the disease (ground truth) was defined as histopathologically for atrophy and histopathologically or combined clinically/histopathologically for eosinophilic esophagitis. Ultimately, 1,681 images (Loc 1 392, Loc 2 339, Loc 3 446, Loc 4 504 images) of 583 patients with available histology were analyzed for atrophy. For eosinophilic esophagitis, 351 images (Loc 5 169, Loc 6 182 images) of 204 patients with biopsy and 936 images (Loc 5 438, Loc 6 498 images) of 550 patients with clinically/histopathologically defined disease were analyzed.

Results: In the total cohort, 22 of 583 evaluable patients had atrophic gastritis and 7 of 550 evaluable patients had eosinophilic esophagitis. For the precancerous conditions of the proximal stomach, there were 33, 146, 90, 37 false-positive and 3, 1, 6, 10 false-negative evaluations by the algorithm for Loc 1-4. There was a balanced accuracy for Loc 1-4 of 86%, 74%, 72% and 67%, a sensitivity of 81%, 94%, 65% and 41% and a specificity of 91%, 55%, 79% and 92%. No relevant improvement in the results could be achieved by combining the individual localizations. The examiners made 16 false-positive and 13 false-negative assessments at patient level and achieved a balanced accuracy of 69%, a sensitivity of 41% and a specificity of 97%. The algorithm for eosinophilic esophagitis shows 10 and 16 false-positive and 0 and 3 false-negative findings in the histology cohort for Loc 5 and Loc 6, respectively. This results in a balanced accuracy of 97% and 74% for Loc 5 and Loc 6, a sensitivity of 100% and 57% and a specificity of 94% and 91%. In the clinically/histopathologically defined cohort, there were 21 and 30 false-positive and 0 and 3 false-negative findings for Loc 5 and Loc 6, respectively, with a balanced accuracy of 95% and 98%, a sensitivity of 100% for both and a specificity of 91% and 95%. No relevant improvement in the results could be achieved by combining localizations 5 and 6. The investigators gave 7 false-positive and 1 false-negative assessment at patient level for both cohorts. For the histology cohort, there was a balanced accuracy of

91%, a sensitivity of 86% and a specificity of 96%. In the clinical/histopathological cohort, the corresponding values are 92%, 86% and 99%.

In the error analysis, a relevant number of examinations were found with explainable causes for a misjudgment due to the quality of the image material (Loc 1-6: 58%, 9,5%, 52%, 55%, 95%, 100%). Furthermore, normal findings from Loc 2 are strongly underrepresented in the original training data set (4%). Thus, the high rate of otherwise unexplained false-positive findings may be due to an underrepresentation of comparable normal findings in the original training dataset.

Conclusion: This work clearly shows that prospective validations in unselected cohorts are essential to assess the clinical utility of AI algorithms. The algorithm for the detection of precancerous conditions of the proximal stomach revealed significant weaknesses with numerous false-positive and false-negative evaluations. This is explained by the limited training data set for a disease with a wide range of endoscopic appearances and a large organ with variable image angles. When using images from a standard endoscopic position (Loc 1), the algorithm performs satisfactorily under real conditions, while the algorithm cannot be used to evaluate images from other positions (especially Loc 2) due to insufficient mapping in the training data set.

For eosinophilic oesophagitis, the performance of the algorithm is significantly more stable under real conditions. This is due to the significantly lower variability of the image representation, so that the spectrum of symptoms was better represented in the larger initial training data set. The differences between Loc 5 and 6 are most likely due to less representative image material in Loc 6, which is why the use of images of the lower esophagus (Loc 5) is recommended.

2 Einleitung

2.1 Eosinophile Ösophagitis

2.1.1 Anatomie des Ösophagus

Der Ösophagus, auch Speiseröhre genannt, befindet sich im hinteren unteren Mediastinum und erstreckt sich über eine Länge von ca. 25cm, beginnend auf der Höhe des 6. bis 7. Halswirbelkörpers und endend auf der Höhe des 11. Brustwirbelkörpers. Seine Hauptaufgabe ist der Weitertransport der Speisen vom Rachen zum Magen [23].

MAKROSKOPIE

Der Ösophagus wird in seinem Verlauf in drei Abschnitte unterteilt und besitzt drei physiologische Engstellen (siehe Tabelle 1 und 2) [6,23,66].

Tabelle 1 - Abschnitte des Ösophagus

Pars cervicalis	- Halsteil, ca. 8 cm lang - liegt der Halswirbelsäule direkt an
Pars thoracica	- Brustteil, ca. 16 cm lang und somit der längste Teil - tritt durch das Zwerchfell in die Bauchhöhle
Pars abdominalis	- Bauchteil, ca. 1-3cm lang und somit der kürzeste Teil - mündet in die Kardia des Magens - in Ruhe geschlossen, öffnet sich nur beim Schluckakt

Tabelle 2 - Engstellen des Ösophagus

Obere Enge	- Constrictio pharyngooesophagealis bzw. cricoidea auf Höhe des Ringknorpels - Ösophagusmund, engste Stelle, ca. 15 cm von der vorderen Zahnreihe entfernt - **Bestandteile**: zirkulärer Sphinktermuskel, submuköser Venenplexus
Mittlere Enge	- Constrictio partis thoracicae bzw. bronchoaortica - Aortenenge, ca. 25 cm von der vorderen Zahnreihe entfernt - **Bestandteile**: Aortenbogen, linker Hauptbronchus
Untere Enge	- Constrictio diaphragmatica bzw. phrenica - Zwerchfellenge, Durchtritt durch Hiatus oesophageus, ca. 40cm von der vorderen Zahnreihe entfernt - **Bestandteile**: Zwerchfell

HISTOLOGIE

Sein Wandaufbau (siehe Abb. 1) besteht von innen nach außen aus der Tunica mucosa, Tunica submucosa, Tunica muscularis aus Ring- und Längsmuskulatur und der Tunica serosa. Die Tunica mucosa besteht aus den drei Schichten Lamina epithelialis mucosae (aus mehrschichtig unverhorntem Plattenepithel), Lamina propria mucosae (enthält den Venenplexus und Schleimdrüsen) und der Lamina muscularis mucosae (aus glatter Muskulatur). Die Muskulatur der Tunica muscularis variiert je nach Höhe des Ösophagus, so befindet sich im oberen Drittel nur quergestreifte Muskulatur und im unteren Drittel nur glatte Muskulatur. In dem Drittel dazwischen sind beide Muskulaturarten vorzufinden [66].

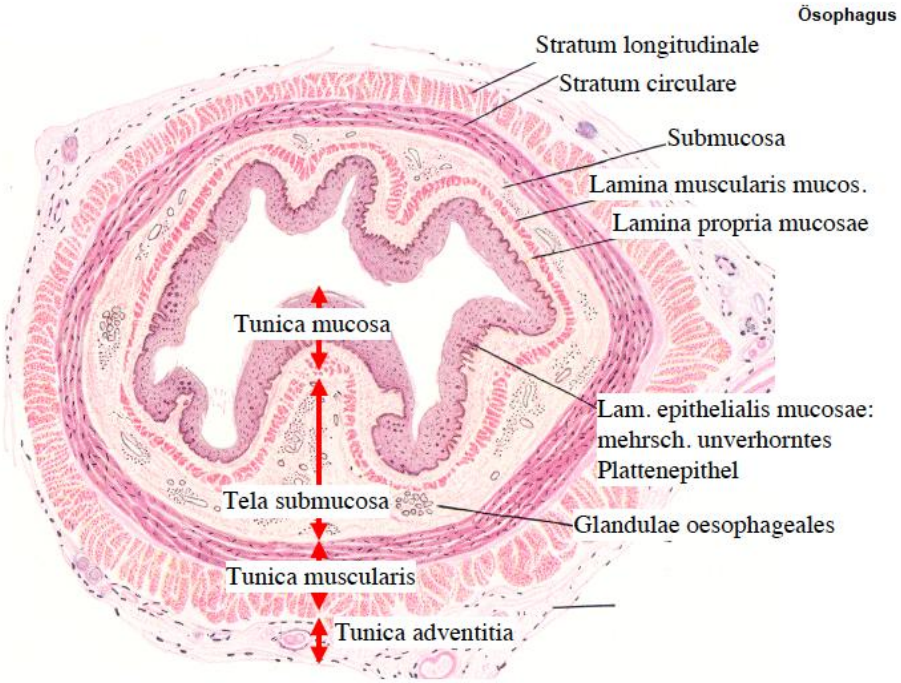


Abbildung 1 - Wandaufbau des Ösophagus im Querschnitt

Quelle: https://histohelp.files.wordpress.com/2011/09/screen-capture7.png

FUNKTIONSWEISE

Weitere wichtige Bestandteile des Ösophagus sind der obere Ösophagussphinkter (OÖS), der Luft- und Nahrungsweg voneinander trennt, sowie der untere Ösophagussphinkter (UÖS) [66]. Dieser besteht jedoch nicht aus einem einzelnen Muskel, sondern ist vielmehr ein Zusammenspiel mehrerer Mechanismen, um einen Aufstieg der Magensäure aus dem Magen, somit einen Reflux, physiologischerweise zu verhindern:

- die spiralig angeordnete Muskulatur der Tunica muscularis strahlt schraubenförmig am Übergang vom Ösophagus in den Magen nach innen und engt somit das Lumen ein [6]. Durch eine Peristaltikwelle beim Schluckakt wird die Spannung der Muskulatur gelöst und das Lumen geöffnet,
- in der Bauchhöhle herrscht ein höherer Druck als in der Brusthöhle [6]. Somit wird ein Druckgradient erzeugt, der den UÖS zusammendrückt und dadurch dessen Verschluss verstärkt,
- der His-Winkel entsteht durch die spitzwinklige Einmündung des Ösophagus in den Magen. Er beträgt physiologisch beim Erwachsenen ca. 60-65°,

- ein Venenplexus, welcher in Ruhe mit Blut gefüllt ist. Erst beim Schlucken entleert sich das Blut,
- der enge Durchtritt durch den Zwerchfellhiatus verstärkt zusätzlich die Einengung des Lumens,
- das Lig. phrenicooesophageale fixiert das untere Ösophagusende am Zwerchfell, wodurch das Bestehen des His-Winkels gewährleistet wird [6].

GEFÄßVERSORGUNG

Die Pars cervicalis wird arteriell über die Rr. oesophageales der A. thyroidea inferior aus dem Truncus thyrocervicalis versorgt mit venösem Abfluss über die entsprechenden Vv. oesophageales der V. thyroidea inferior sowie dem lymphalen Abfluss über die Nll. cervicales profundii [66].

Die Pars thoracica hingegen erhält ihre arterielle Versorgung über die Rr. oesophageales direkt aus der Aorta. Die Vv. oesophageales mit Mündung in die V. azygos bzw. hemiazygos sorgen für den venösen Rückfluss. Die Lymphe wird u.a. über die tiefen zervikalen Lymphknoten sowie die des Mediastinums, also die Nll. paratracheales, Nll. tracheales superior und inferior sowie Nll. juxtaoesophageales abtransportiert [66].

Der abdominelle Anteil wird arteriell über die Rr. oesophageales der A. gastrica sinistra aus dem Truncus coeliacus versorgt. Der Rückfluss wird über die gleichen Venen wie bei der Pars thoracica gewährleistet. Der Lymphabfluss gelingt über die Nll. gastrici sinistri und Nll. coeliaci. [66].

INNERVATION

Der paarige N. vagus ist für die sensible sowie parasympathische Funktion des Ösophagus zuständig, dies betrifft also die Wandmotilität, die Sphinkterfunktion sowie die Drüsensekretion. Die sympathische Versorgung erfolgt v.a. über das thorakale Ganglion stellatum und den Grenzstrang und beeinflusst das Schmerzempfinden und die Engstellung der Gefäße. Zusätzlich befindet sich in der Tunica submucosa der Plexus submucosus (Meissner-Plexus) bzw. in der Tunica muscularis der Plexus myentericus (Auerbach-Plexus) des autonomen enterischen Nervensystems [66].

2.1.2 **Definition der eosinophilen Ösophagitis**

Bei der eosinophilen Ösophagitis (EoE) handelt es sich um eine immunologische, chronisch verlaufende Erkrankung des Ösophagus, bei welcher sich symptomatisch v.a. eine Dysphagie zeigt [53]. Hierbei ist eine eosinophile Inflammation das typische histologische Bild [50]. Ein wichtiges Kriterium zur Diagnosestellung der EoE ist das Ausschließen anderer systemischer und/oder lokaler Ursachen einer Eosinophilie des Ösophagus, wie z.B. eine gastroösophageale Refluxkrankheit, eine Achalasie oder andere primäre Motilitätsstörungen des Ösophagus, ein Hypereosinophilensyndrom, ein Morbus Crohn sowie Hauterkrankungen mit Ösophagusbeteiligung (z.B. Pemphigus oder Lichen), Autoimmunerkrankungen, Vaskulitiden, eine Graft-versus-Host-Erkrankung, eine Pseudodivertikulose oder fungale, virale oder parasitäre Infektionen [53].

In ehemaligen Leitlinien der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- & Stoffwechselkrankheiten (DGVS) wurde zusätzlich die Differenzierung einer PPI-responsiven Eosinophilie (PPI-REE) erläutert [53]. Es hat sich jedoch gezeigt, dass sich die PPI-REE und EoE auf keiner diagnostischen Ebene unterscheiden lassen, wodurch diese Einteilung zunehmend verlassen wurde [52].

2.1.3 **Epidemiologie und Risikofaktoren**

Die erstmalige Beschreibung der EoE als eigenständiges Krankheitsbild erfolgte im Verlauf der 1990er Jahre und stellte zum damaligen Zeitpunkt eine absolute Rarität dar [5]. Mit der Industrialisierung der Länder entwickelte sich diese jedoch zum jetzigen Stand zu einer der häufigsten Inflammationen des Ösophagus. Sowohl Prävalenz, also die Anzahl der Erkrankten in einem bestimmten Zeitraum in einer bestimmten Population, als auch die Inzidenz, also die Anzahl der Neuerkrankungen in einem bestimmten Zeitraum in einer bestimmten Population, sind über die letzten 20 Jahre um ca. das 20-fache angestiegen, obwohl die Anzahl der durchgeführten Endoskopien und Biopsien kaum zugenommen hat [20,26]. Zum aktuellen Zeitpunkt wird die durchschnittliche Prävalenz auf ca. 35 pro 100.000 Einwohner geschätzt, basierend auf diesbezüglichen veröffentlichten Publikationen. Dabei zeigt sich bei nicht selektionierter Population eine Prävalenzrate von 4,8 – 7,3%, wohingegen bei Endoskopien, die primär durch Dysphagien indiziert sind, sogar 10-25% eine EoE aufzeigen [20,74].

Die EoE manifestiert sich im Durchschnitt zwischen dem 30. und 50. Lebensjahr [68].

Als mögliche Risikofaktoren werden aktuell v.a. atopische Vorerkrankungen wie z.B. Asthma bronchiale, atopische Dermatitis oder allergische Rhinitis diskutiert, da entsprechende Patienten häufiger eine EoE aufweisen als die Normalbevölkerung. Vermutlicher Auslöser sind hierbei Aero- und Nahrungsallergene, die zu einer nicht-IgE-vermittelten Immunreaktion führen [27,88].

Ebenfalls gibt es Hinweise für eine familiäre Häufung mit genetischer Prädisposition für die EoE, wovon v.a. Männer ein bis zu 64-faches erhöhtes Risiko für die Erkrankung zeigen [2].

2.1.4 **Klinik und Verlauf**

Die Hauptsymptome der EoE sind die Dysphagie (70-80%) und retrosternales Brennen (30-55%), welche von unspezifischen Beschwerden wie Übelkeit und Erbrechen, Lumenverlegung des Ösophagus durch Nahrung und daraus resultierender Nahrungsverweigerung bzw. -vermeidung begleitet werden können [19].

Seit kurzem wird der Symptomkomplex FIRE beschrieben. FIRE ist die Abkürzung für „food-induced immediate response of the esophagus“ und beschreibt ein von der eigentlichen Dysphagie unabhängiges unangenehmes bis schmerzhaftes Empfinden nach unmittelbarem Kontakt der Ösophagusschleimhaut mit bestimmten Lebensmitteln, v.a. Wein, Gemüse und Früchte [7].

Die Folge dieser Symptome ist häufig, dass Patienten (bewusst) ihr Essverhalten verändern, indem sie z.B. besser kauen, mehr trinken beim Essen, mehr Zeit zum Essen benötigen oder sogar bestimmte Nahrungsmittel meiden [3]. Die Patienten geben bei gezielten Nachfragen infolgedessen eine reduzierte sogenannte gesundheitsbezogene Lebensqualität an (HRQOL, health reduced quality of life) [92].

Außerdem leiden EoE-Patienten häufiger unter sozialen, körperlichen und psychologischen Belastungen, die sich u.a. in lebensmittelspezifischen Ängsten, Angstzuständen und Depressionen äußern [92].

Studien mit Placebo-behandelten Patienten zeigen, dass eine unbehandelte EoE im Verlauf zu einer chronischen und rezidivierenden Inflammation führt, die zwar einerseits eine Abnahme der Eosinophilie zeigt, jedoch gleichzeitig auch eine zunehmende Fibrosierung des Ösophagus. Spezifische Komplikationen hiervon sind v.a. Motilitätsstörungen, Strikturen und Stenosen der Speiseröhre mit höherem Risiko für das Auftreten von Bolusobstruktionen [53,91].

2.1.5 **Diagnostik**

Zur Diagnosestellung und Verlaufskontrolle der therapierten EoE bilden die Endoskopie sowie die Biopsie und die daraus resultierende Histologie die wichtigsten Säulen [53]. Die Diagnose wird bei Vorliegen einer ösophagealen Dysfunktion und Gewebseosinophilie unter Abwesenheit alternativer Ursachen für eine Eosinophilie gestellt [6].

In der Endoskopie zeigen sich auffällige strukturelle Befunde des Ösophagus. Merkmale der akuten Entzündung sind weiße Exsudate, welche Mikroabszessen aus Eosinophilen entsprechen, Längsfurchen und ein Schleimhautödem. Der chronischen Entzündung lassen sich die sogenannte Trachealisierung, also eine Ringbildung der Schleimhaut, ein verkleinertes Lumen und Strikturen des Ösophagus zuordnen [21]. Abbildung 2 zeigt die unterschiedlichen Erscheinungsbilder der EoE beim Endoskopieren.

Insgesamt ist in beiden Stadien eine „Crepe paper mucosa“ (Krepppapier-Mukosa), also eine vermehrte Verletzbarkeit der Ösophagusschleimhaut [1], sowie ein „tug sign“, entsprechend einem harten Widerstand bei Biopsieentnahme, möglich [1,58]. Retrospektive Studien haben gezeigt, dass sich bei Erwachsenen zu 80% Längsfurchen, 64% Ringe, 28% ein verkleinertes Lumen, 16% Mikroabszesse und 12% Strikturen vorfinden lassen [84]. Diese genannten Auffälligkeiten können vereinzelt oder kombiniert auftreten, wobei sie bei 90% der EoE-Patienten gefunden werden können. Weiterhin hat sich zeigen lassen, dass die Auffindungsrate der auffälligen Befunde mit der Erfahrung des Endoskopikers korreliert [45].

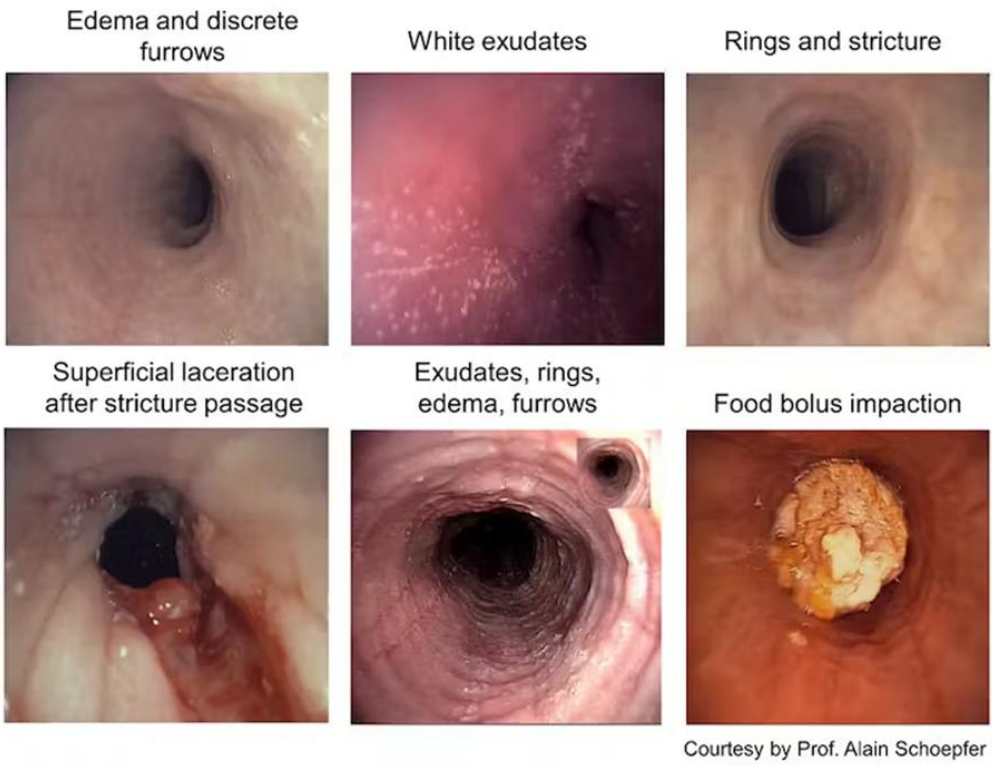


Abbildung 2 - Auffällige Befunde bei EoE
Quelle: https://www.paediatrieschweiz.ch/eosinophile-osophagitis-update-2022/

Die endoskopische Befundung soll anhand der EREFS-Klassifikation (Akronym für Exsudate, Ringe, Ödem, Furchen, Strikturen, siehe Abb. 3) standardisiert werden [82]. Diese wurde erstmalig 2013 von Hirano et al. [40] veröffentlicht und kürzlich in modifizierter, vereinfachter Form von Schoepfer et al. vorgeschlagen [82]. In dieser Klassifikation werden die einzelnen endoskopischen Befunde als Major- oder Minor-Kriterium zugeordnet und anhand der Ausprägung bewertet um somit einen Zusammenhang zwischen diesen, den histologischen und therapeutischen Ergebnissen herzustellen [82].

Majorbefunde	
	Grad
	Ringe
0	Keine
1	Gering (dezent erkennbar)
2	Moderat (deutliche Ringe, Passage mit Standardgastroskop möglich)
3	Schwer (deutliche Ringe, Passage mit Standardgastroskop nicht möglich)
	Exsudat
0	Kein
1	Mild (</= 10 % der ösophagealen Oberfläche)
2	Schwer (> 10 % der ösophagealen Oberfläche)
	Furchen
0	Keine
1	Vorhanden
	Ödem
0	Kein (mukosale Gefäße sichtbar)
1	Vorhanden (mukosale Gefäße nicht oder vermindert sichtbar)
	Striktur
0	Keine
1	Vorhanden
Minorbefunde	
	Krepppapierzeichen (mukosale Lazeration bei Endoskoppassage)
0	Kein
1	Vorhanden

Abbildung 3 - Modifizierter EREFS-Score nach Schoepfer et al.

Quelle: S2k-Leitlinie Gastroösophageale Refluxkrankheit und eosinophile Ösophagitis der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (DGVS), 2023, S. 902 [53]

Während der Endoskopie sollte eine sogenannte Stufenbiopsie insbesondere aus auffälligen Arealen, also v.a. aus weißlichen Exsudaten [51] und Längsfurchen [35], erfolgen, da hier aufgrund einer hohen Entzündungsaktivität eine hohe Anzahl an eosinophilen Granulozyten zu erwarten ist [53]. Ein unauffälliger histologischer Befund kann durch das oft unregelmäßige Befallsmuster im Sinne eines falsch-negativen Ergebnisses eine EoE nicht ausschließen [45]. Eine ausreichend hohe Sensitivität, also die Wahrscheinlichkeit eines richtig-positiven Ergebnisses bei erkranktem Patienten, wird durch die Entnahme von mindestens 6 Biopsien aus mindestens zwei Abschnitten, also z.B. unterer und oberer Ösophagus, gewährleistet [14,53]. Bei Erstdiagnose werden zusätzliche Probeentnahmen aus Magen und Duodenum empfohlen [53].

Die anschließende histologische Auswertung erfolgt in Hämatoxylin-Eosin-Färbung (HE) und fokussiert sich v.a. auf die Anzahl der Eosinophilen [14]. Für die Diagnosestellung einer EoE werden > 15 Eosinophile pro hochauflösendes Gesichtsfeld (HPF, Standardgröße 0,3mm²) als ausreichend gewertet [14,53]. Andere auffällige Befunde sind eosinophile Mikroabszesse

(siehe Abb. 4), eine Hyperplasie - also eine Vermehrung der Zellen – der Basalzone, vergrößerte interzelluläre Abstände, eine Verhornungsstörung sowie Fibrose der Epithelzellen der Tunica mucosa [14].

Aktuelle Studien untersuchen den Nutzen des EoE-spezifischen-Histologie-Score (EoEHSS) zur zukünftigen standardisierten Befundung [53].

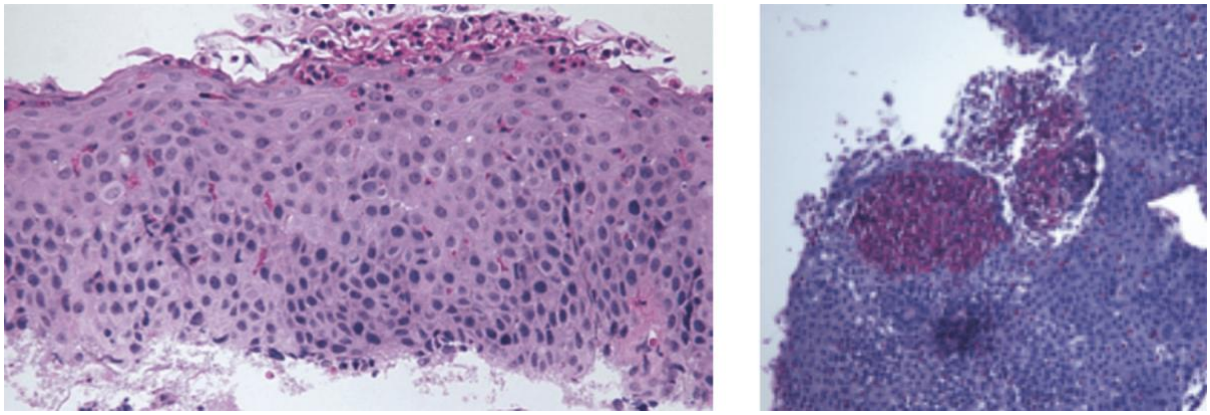


Abbildung 4 - Links: eosinophile Infiltration. Rechts: eosinophiler Mikroabszess
Quelle: https://www.eoe.de/diagnose/

Weitere nicht-invasive Methoden, v.a. zur Verlaufskontrolle zum Verzicht weiterer Biopsien, werden derzeit untersucht, z.B. die absolute Eosinophilenzahl im Blut oder weitere Biomarker im Atem, Speichel, Stuhl oder sogar Urin. Diese zeigen jedoch bisher eine unzureichende Sensitivität und Spezifität, also die Wahrscheinlichkeit für richtig-negative Ergebnisse bei gesunden Patienten, und werden somit aktuell nicht von den Leitlinien empfohlen [39,53]. Es wird ebenfalls von einer routinemäßigen Testung auf Allergien mittels Pricktest oder Serum-IgE abgeraten, da der Auslöser der EoE nicht ausreichend sichergestellt werden kann [53].

2.1.6 **Therapie**

Bei Nachweis einer aktiven EoE soll eine Therapie erfolgen [53]. Diese besteht grundsätzlich zunächst aus einer Induktionstherapie, um eine Remission zu erhalten, welche anschließend bei Erfolg durch eine Erhaltungstherapie ergänzt wird. Die Remission sollte nach 8-12 Wochen klinisch, endoskopisch und histologisch überprüft werden [52]. Abbildung 5 gibt einen Überblick über die Therapiemöglichkeiten bei Erwachsenen und Kindern.

Das effektivste und somit bevorzugte Präparat der Induktionstherapie ist Budesonid in Form einer Tablette (topisches Glucocorticoid), welche täglich 2x 1 mg während 8 bis 12 Wochen eingenommen wird. Die wichtigste, aber dennoch selten auftretende Nebenwirkung ist eine lokale Candidiasis. Systemische Glucocorticoide sollten hingegen nicht angewendet werden. [52,53]

Eine Alternative stellt eine hochdosierte PPI-Therapie (Protonenpumpeninhibitor) dar, beispielsweise Omeprazol täglich 2x 40mg. Hierzu gibt es laut den aktuellen Leitlinien der DGVS jedoch nur wenige Studien, die eine PPI-Therapie direkt mit einer topischen Glucocorticoid-Therapie vergleichen und keine Placebo-kontrollierten Studien. Außerdem wird das frühere Konzept der PPI-responsiven (PPI-REE) EoE nicht mehr erörtert. Es gibt lediglich Risikofaktoren, wie junges Alter, ein BMI ≤ 25,2 kg/m2 sowie der Nachweis einer Eosinophilie > 460 pro mm^3 im Blut für eine höhere Wahrscheinlichkeit des Nichtansprechens auf eine alternative PPI-Therapie. Im Falle eines Ansprechens besteht aber somit die Möglichkeit einer langfristigen steroidfreien Remission der Erkrankung und rechtfertigt entsprechend den Versuch einer PPI-Therapie [53].

Eine weitere Möglichkeit, v.a. bei Versagen oder Unverträglichkeit vorangegangener Behandlungsversuchen, ist die 6-Food-Eliminationsdiät, bei welcher die häufigsten Nahrungsmittelallergien auslösenden Nahrungsmittel vermieden werden. Die Patienten sollten dementsprechend auf Kuhmilchproteine, Weizen, Soja, Nüsse, Eier und Fisch sowie Meeresfrüchte verzichten [42,53]. Dies stellt eine große Herausforderung für die Patienten dar, weswegen eine begleitende Ernährungsberatung empfohlen wird, um die Compliance zu gewährleisten sowie Fehl- und Mangelernährung zu verhindern [53]. Im Falle einer Remission oder eines Ansprechens der Erkrankung auf die Eliminationsdiät werden dann schrittweise Nahrungsmittel wieder eingeführt mit dem Ziel, das auslösende Nahrungsallergen zu identifizieren, um dieses dann dauerhaft zu meiden [42,53]. Sollte es dennoch anschließend zu einer Reexposition kommen kann dies zu einem erneuten Ausbruch der EoE führen [52,53].

Anschließend wird die Erhaltungstherapie durchgeführt, um den chronischen Progress der Erkrankung zu verhindern [20]. Die bevorzugte Methode ist das Präparat oder die Vorgehensweise, welche erfolgreich eine Remission induziert hat und somit häufige Rezidive verhindern soll. Bei Weitergabe der topischen Glucocorticoiden oder der PPI sollte individuell eine mögliche Dosisreduzierung erfolgen, welche dennoch eine Rezidivfreiheit gewährleistet. Es liegen jedoch weniger Daten zur Erhaltungstherapie mit PPI und Eliminationsdiät vor als mit topischen Glucocorticoiden [52,53].

Der Erfolg der remissionserhaltenden Therapie sollte alle 1-2 Jahre durch klinische, endoskopische und histologische Verlaufskontrollen dokumentiert werden [53].

Zur Therapie der entstehenden Komplikationen können verschiedene interventionelle Vorgehensweisen genutzt werden [53]. Die wichtigste stellt die endoskopische Dilatation des Ösophagus bei Schluckbeschwerden mittels Ballons oder Savary-Bougies dar, sie kann jedoch nicht die eigentliche Entzündung beeinflussen und muss häufig wiederholt durchgeführt werden [81]. Wichtige, aber seltene Komplikationen hiervon sind die Ösophagusperforation und postinterventionelle Blutungen [53,81].

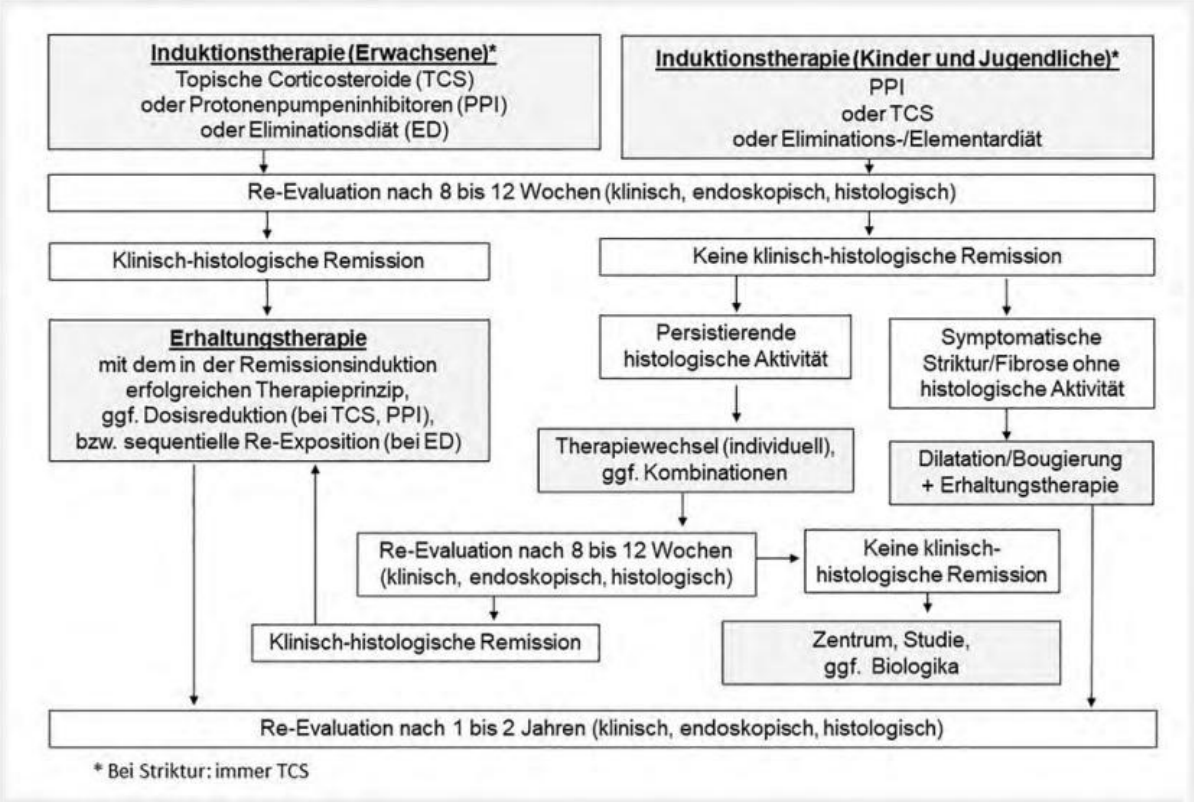


Abbildung 5 - Therapiealgorithmus nach der Leitlinie der DGVS

Quelle: S2k-Leitlinie Gastroösophageale Refluxkrankheit und eosinophile Ösophagitis der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (DGVS), 2023, S. 905 [53]

2.2 Atrophische Gastritis

2.2.1 Anatomie des Magens

MAKROSKOPIE

Der Magen lässt sich makroskopisch von kranial nach kaudal in fünf Abschnitte einteilen (siehe Abb. 6):

- die Kardia ist nur wenige Millimeter breit und mit muzinproduzierenden Drüsenzellen, sogenannten Nebenzellen, ausgekleidet. Sie wird vom Ösophagus durch die Z-Linie getrennt und stellt eine Verbindung zu diesem her,
- der Fundus ist der höchste Teil des Magens und besitzt Pepsinogen-bildende Haupt sowie Intrinsic-Faktor- und Magensäure-bildende Parietal-/Belegzellen,
- der Korpus bildet den größten Anteil mit der kleinen und großen Kurvatur und besitzt die gleichen Zellen wie der Fundus. Die Incisura angularis ist ein Knick in der kleinen Kurvatur und bildet somit die Grenze zwischen Korpus und Antrum bzw. Pylorus,
- das Antrum fungiert als Übergangskanal zum Pylorus und enthält Neben- und Gastrin-produzierende G-Zellen,
- der Pylorus stellt den Schließmuskel des Magens am Übergang zum Duodenum dar und hat als Aufgabe den Speisebrei portioniert weiterzuleiten [11].

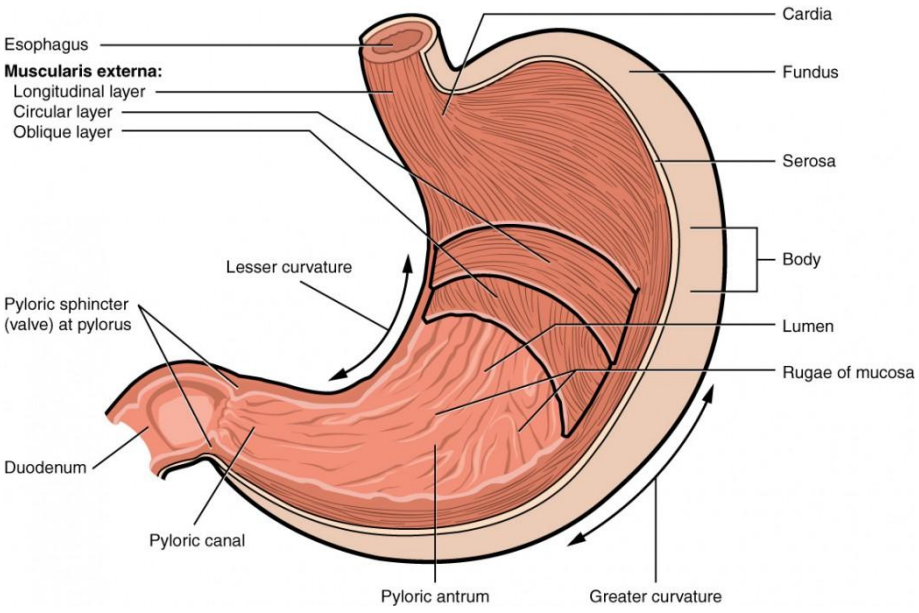


Abbildung 6 - Aufbau des Magens

Quelle: https://courses.lumen-learning.com/suny-ap2/chapter/the-stomach/

HISTOLOGIE

Histologisch ist der Aufbau der Magenwand ähnlich der des Ösophagus, also von innen nach außen aus Tunica mucosa, Tunica submucosa, Tunica muscularis mit Ring- und Längsmuskulatur, und Tunica serosa. Anders als beim Ösophagus besteht die Tunica mucosa des Magens aus einem muzinbedeckten einschichtigem Zylinderepithel. Weiterhin kennzeichnend sind die Areae gastricae, ein pflastersteinartiges Schleimhautrelief, in denen sich die Foveolae gastricae vertiefen, die Einmündungsstellen der Magendrüsen. Die Tunica mucose und Tunica submucosa bilden zusammen die makroskopisch erkennbaren Magenfalten (Plicae gastricae), welche mit zunehmender Füllung des Magens verstreichen. In der Tunica muscularis findet sich eine zusätzliche dritte Muskelschicht, die schräge Muskulatur (Fibrae obliquae), und bildet somit die innerste Muskelschicht. Die eigentliche Ringmuskulatur verdickt sich zusätzlich am Pylorus zum Schließmuskel [11]. Die Abbildung 7 stellt den histologischen Aufbau dar.

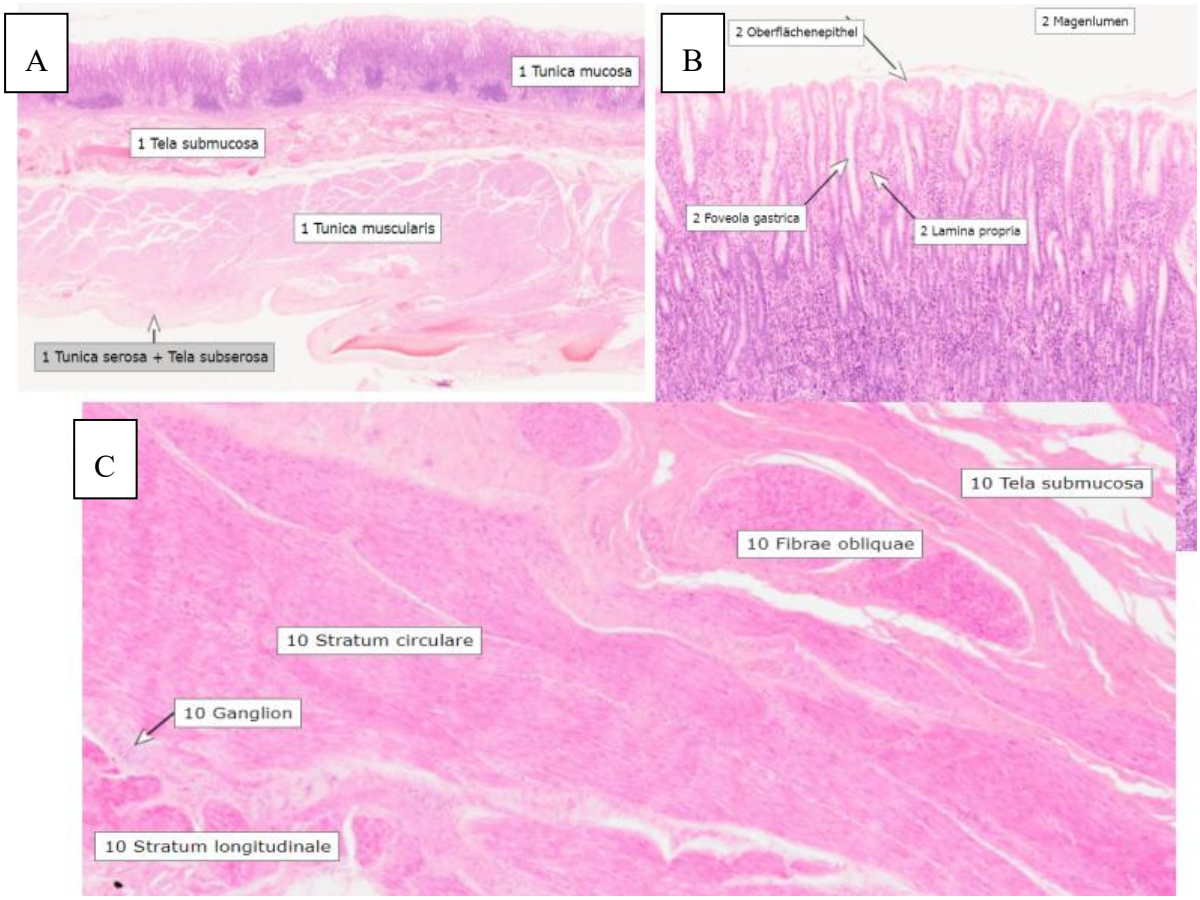


Abbildung 7 - Histologischer Aufbau des Magens

A: Übersichtsaufnahme. **B:** Tunica mucosa. **C:** Tunica muscularis

Quelle: Virtuelles Mikroskop der mikroskopischen Anatomie des Universitätsklinikum des Saarlandes

FUNKTIONSWEISE

Die grundsätzliche Funktion des Magens besteht aus der Verdauung der Nahrung zu einem Speisebrei. Das Muzin der Nebenzellen bildet auf der Magenschleimhaut einen schützenden Schleimfilm und verhindert somit u.a. Verletzungen durch die Magensäure sowie eine Selbstverdauung. Das von den Hauptzellen sezernierte Pepsinogen ist die inaktive Vorstufe des Pepsins, welches von der Magensäure erst aktiviert wird und somit die Proteine in der Nahrung spaltet [11,70]. Das Gastrin der G-Zellen stimuliert zum einen die Pepsinogen- und zum anderen die Salzsäure-Produktion. Auch ECL-Zellen (enterochromaffin-like cell) in der Nähe der Parietalzellen regen diese zur Magensäureproduktion durch Ausschüttung von Histamin an [11,69,70].

Die Parietalzellen sind für die Bildung der Magensäure zuständig, indem sie Salzsäure (HCl) sezernieren. Für diese Funktion besitzen sie eine H+/K+-ATPase, welche Protonen im Austausch gegen Kalium in die Zelle pumpt. Dabei entsteht Bicarbonat (HCO_3-), welches durch einen HCO_3-/Cl- -Antiporter im Austausch aus der Zelle befördert wird. Am apikalen Pol der Parietalzelle befinden sich Cl- -Kanäle, wodurch Cl- ins Magenlumen sezerniert wird und zusammen mit dem Proton des Bicarbonats somit die Salzsäure entsteht [11,69,70].

Zusätzlich bilden die Parietalzellen Intrinsic-Faktor. Dieses Protein ist notwendig um im Ileum die Aufnahme von Vitamin B12 (Cobalamin) zu ermöglichen, welches hingegen u.a. für die Erythropoese essenziell ist [11,70].

GEFÄßVERSORGUNG

Die Gefäßversorgung ist für die verschiedenen Anteile des Magens unterschiedlich. Die große Kurvatur wird von der A. gastrica sinistra aus dem Truncus coeliacus und der A. gastrica dextra aus der A. hepatica communis versorgt mit venösem Abfluss über die Vv. gastricae sinistra und dextra in die V. portae [11].

Die kleine Kurvatur erhält ihre arterielle Versorgung aus der A. gastroomentalis sinistra aus der A. splenica des Truncus coeliacus und der A. gastroomentalis dextra aus der A. gastroduodenalis, wobei das venöse Blut über die gleichnamigen Venen wieder abfließt und in der V. portae mündet [11].

Der Fundus wird über die Aa. gastricae breves aus der A. splenica versorgt und wird über die entsprechenden Venen wieder in die V. portae abtransportiert, wohingegen der Blutfluss für

die Magenhinterwand über die A. und V. gastrica posterior der A. und V. splenica reguliert wird [11].

Der Lymphabfluss ist für alle Magenanteile identisch über verschiedene abdominale Lymphknoten, unter anderem über die Nll. gastrici dextri und sinistri, Nll. splenici, Nll. hepatici, Nll. pylorici und Nl. cardiacus [11].

INNERVATION

Der Magen wird, wie der Ösophagus, parasympathisch über den paarigen N. vagus versorgt. Dabei stellt durch die Magendrehung bedingt der linke N. vagus den Truncus vagalis anterior und der rechte N. vagus den Truncus vagalis posterior dar. Die sympathische Versorgung, v.a. für das Schmerzempfinden, erfolgt über Fasern aus dem sympathischen Grenzstrang von Th6 bis TH9 [11].

2.2.2 **Definition der atrophischen Gastritis**

Die atrophische Gastritis (AG) beschreibt eine mit Verlust der mukosalen Drüsenzellen einhergehende Magenschleimhautentzündung und betrifft meist Antrum, Corpus und/oder Fundus. Dieser Verlust entwickelt sich im Verlauf zur intestinalen Metaplasie (IM), also die Umwandlung in eine andere epitheliale Zellart des Gastrointestinaltraktes (GIT), und birgt das Risiko der Entstehung einer intraepithelialen Dysplasie, die eine direkte Tumorvorstufe darstellt. Während die atrophische Gastritis und intestinale Metaplasien präkanzeröse Bedingungen darstellen, sind die geringgradigen und hochgradigen Dysplasien echte präkanzeröse Läsionen, die sich zu einem Magenkarzinom entwickeln können [29,89,99].

Ein Modell zur Erklärung dieser Tumorentstehung ist die Correa-Sequenz (siehe Abb. 8). Diese beschreibt die stufenweise Entwicklung eines Magenkarzinoms ausgehend von einer ursprünglichen physiologischen Magenschleimhaut [16]. Erstmalig wurde der entsprechende Verlauf 1988 von Correa P. [15] beschrieben, wobei aus einer unauffälligen Magenschleimhaut v.a. durch eine Infektion mit Helicobacter pylori eine nicht-atrophische Gastritis entsteht. Diese kann im unbehandelten Zustand zu einer atrophischen Gastritis führen, welche anschließend zu intestinalen Metaplasien führt. Aus diesen wiederrum entstehen schlussendlich Dysplasien und somit möglicherweise ein Magenkarzinom [16].

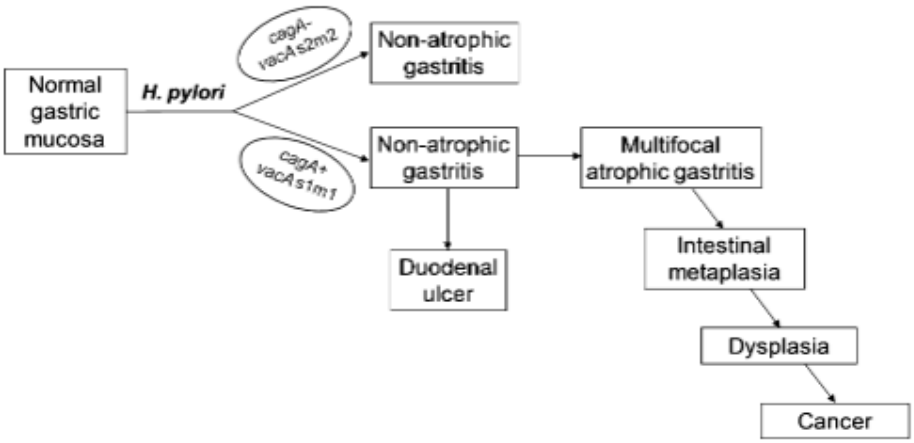


Abbildung 8 - Correa-Sequenz zur Entstehung von Magenkarzinomen

Quelle: The gastric precancerous cascade. Correa P., Piazuelo M. Journal of Digestive Diseases 2012, page 3 [16]

2.2.3 **Epidemiologie und Risikofaktoren**

Die Hauptauslöser für eine atrophische Gastritis sind vor allem eine Helicobacter-pylori (HP)-Infektion sowie die Autoimmungastritis, welche häufig zunächst unentdeckt bleiben und somit progressiv verlaufen. In Abhängigkeit der untersuchten Population ändert sich die Häufigkeit der ursächlichen Auslöser [90].

In Asien ist am häufigsten die HP-assoziierte Gastritis mit Prävalenzen zwischen 54-76% [24], wohingegen die Prävalenz in Westeuropa bei lediglich 34% liegt [98]. Die langjährige Infektion breitet sich im Verlauf vom Antrum über die kleine Kurvatur bis nach proximal in den Fundus aus [85]. Die Besiedlung mit dem gramnegativen Stäbchen HP führt zu einer reduzierten Muzin- und gleichzeitig erhöhten Magensäureproduktion. Durch das Enzym Urease kann es sich selbst vor der Zerstörung durch die Magensäure schützen, indem Harnstoff zu Ammoniak umgewandelt wird [8]. Die Erstinfektion tritt v.a. bei Kindern durch eine fäkal-orale Übertragung auf und ist abhängig von sozioökonomischen Bedingen, wie beispielsweise den häuslichen Hygienebedingungen oder der Kochgewohnheiten [90]. Eine Infektion erhöht das Risiko für die Entstehung eines Magenkarzinoms, v.a. im distalen Bereich, um den Faktor 2-3 [103].

In Europa wiederum spielt die Hauptrolle für die Entstehung einer atrophischen Gastritis die Autoimmungastritis, auch als Typ-A-Gastritis bekannt. Die Prävalenz in Deutschland ist durch oftmals asymptomatische Patienten unklar, jedoch zeigte die Studie von Zhang Y. et al. von 2013, dass bei fast 20% der untersuchten Patienten entsprechende Autoantikörper vorliegen

[101]. Durch die Bildung von Parietalzellantikörper werden die körpereigenen Parietalzellen im proximalen Magenanteil angegriffen. Zusätzlich sind möglicherweise auch Intrinsic-Faktor-Antikörper nachweisbar. Somit greifen diese vor allem die Produktion der Magensäure und des Intrinsic-Faktors an, was wiederum zu einer vermehrten Stimulation der ECL-Zellen führt [85,90].

Insgesamt tritt die Autoimmungastritis häufiger bei Patienten auf, die bereits andere Autoimmunerkrankungen haben. Am häufigsten sind dies Diabetes mellitus Typ 1, Vitiligo oder autoimmunvermittelte Schilddrüsenerkrankungen [72]. Ihr Verlauf ist im Vergleich zur HP-Gastritis rascher progredient [90]. Dennoch ist das Risiko der Entstehung eines Magenkarzinoms bei Patienten mit Autoimmungastritis wesentlich geringer als bei Patienten mit HP-Infektion [76].

Es gibt zwei klinisch etablierte Klassifikationssysteme zur Risikostratifizierung bezüglich der Karzinogenese bei atrophischer Gastritis und/oder intestinalen Metaplasien. Für die AG erfolgt dieses Staging mittels OLGA-Klassifikation (Operative Link on Gastritis Assessment, siehe Abb. 9) bzw. bei intestinalen Metaplasien mittels OLGIM-Klassifikation (Operative Link on Gastric Intestinal Metaplasia Assessment, siehe Abb. 10). Beide teilen die präkanzerösen Bedingungen anhand ihres Schweregrades in entsprechende Stadien ein. Obschon OLGIM weniger untersucherabhängig ist werden oftmals beide Schemata gemeinsam verwendet und zeigen somit die bestmögliche Risikoprädiktion. Dabei zeigt sich, dass Patienten mit Stadium OLGA III/IV bzw. OLGIM III/IV ein erhöhtes Risiko für ein Magenkarzinom aufweisen [9,54,77,103].

Insgesamt wird bei Vorliegen von präkanzerösen Bedingungen im Sinne einer fokalen Atrophie oder intestinalen Metaplasien unabhängig ihres Ursprunges ein Untersuchungsintervall mittels Endoskopie und Biopsieentnahme nach Sydney-Protokoll [46] (siehe Abb. 12) von 3 Jahren empfohlen [77].

		Corpus			
Atrophy Score		No	Mild	Moderate	Severe
Antrum	No	Stage 0	Stage I	Stage II	Stage II
	Mild	Stage 1	Stage I	Stage II	Stage III
	Moderate	Stage II	Stage II	Stage III	Stage IV
	Severe	Sage III	Stage III	Stage IV	Stage IV

Abbildung 9 – OLGA

Quelle: OLGA- and OLGIM-Based Staging in the Patients with Gastritis and Endoscopy Indications. Mansour-Ghanaei F., 2022, S. 96 [54]

		Corpus			
Intestinal Metaplasia Score		No	Mild	Moderate	Severe
Antrum	No	Stage 0	Stage I	Stage II	Stage II
	Mild	Stage 1	Stage I	Stage II	Stage III
	Moderate	Stage II	Stage II	Stage III	Stage IV
	Severe	Sage III	Stage III	Stage IV	Stage IV

Abbildung 10 - OLGIM

Quelle: OLGA- and OLGIM-Based Staging in the Patients with Gastritis and Endoscopy Indications. Mansour-Ghanaei F., 2022, S. 97 [54]

2.2.4 **Klinik und Verlauf**

Symptomatisch zeigt sich die AG selbst unauffällig. Die meisten Patienten zeigen unspezifische Symptome entsprechend der zugrunde liegenden chronischen Gastritis. Die Hauptsymptome sind hierbei v.a. Oberbauchschmerzen und Übelkeit, gefolgt von einem vorzeitig einsetzendem Sättigungs- und Völlegefühl. Andere Symptome können Dyspepsie, Reflux, ein Nüchternschmerz oder postprandialer Schmerz, sowie selten Sodbrennen oder Regurgitationen sein [72].

Im Verlauf führt die AG bei Befall des proximalen Magens durch den Untergang der diversen Zellen in der Magenschleimhaut zu einer Hypo- oder sogar Achlorhydrie und führt durch diese mangelhafte Produktion an Magensäure zu eingeschränkten Funktionen des Magens sowie zu einer Fehlbesiedlung mit oropharyngealen Mikroorganismen. Es kommt zu einer gestörten Verdauung auf der Ebene der gastralen Phase und somit möglicherweise auch zu einem Mangel an Vitaminen und Spurenelementen [72,90].

So kommt es durch die mangelhafte Bildung von Intrinsic-Faktor zu einer verminderten Resorption von Vitamin B12 (Cobalamin) im Dünndarm. Der Mangel an Vitamin B12 kann zu einer megaloblastären, perniziösen Anämie mit den Symptomen Tachykardie, Blässe, Schwindel und Belastungsdyspnoe führen. Typisch ist außerdem eine sogenannte Hunter-Glossitis, also eine rötliche Schwellung der Zunge mit brennenden Missempfindungen [72]. Bei langjährigem Verlauf mündet der Mangel an B12 in eine funikuläre Myelose, also einer Demyelinisierung und Atrophie des Rückenmarks und somit in einer Neuropathie mit u.a. spastischen Paresen und sensorischer Ataxie [72,96]. Auch unspezifische neurologische Symptome im Sinne von kognitiven Defiziten wie Depressionen, Gedächtnisverlust und Apathie sind möglich [72].

Oftmals besteht bei chronischer Erkrankung zusätzlich ein Mangel an Eisen und führt somit zu einer mikrozytären, hypochromen Anämie, welche sich ebenfalls mit Blässe, Schwindel und Belastungsdyspnoe präsentiert. Dieser Mangel entsteht hauptsächlich durch die Hypo- bzw. Achlorhydrie, wodurch eine verminderte Eisenresorption stattfindet. Der Eisenmangel wird durch blutende Mikroläsionen der Magenschleimhaut, eine entzündungsbedingte erhöhte Bildung des Hepcidins und somit eine verminderte Eisenaufnahme im Darm sowie das Konkurrieren mit HP bei entsprechender Infektion um das in der Nahrung enthaltene Eisen zusätzlich begünstigt [72].

Durch die bereits erläuterte Correa-Sequenz ist die größte Gefahr einer AG die Entstehung eines Magenkarzinoms, welches ohne zugrundeliegende AG seltener in der Allgemeinbevölkerung auftritt [90]. Auch die Entstehungsrate von neuroendokrinen Tumoren (NET) ist erhöht, da diese mit den höheren Gastrin-Spiegeln in Folge der Hypo- bzw. Achlorhydrie assoziiert sind [72].

2.2.5 **Diagnostik**

Der Goldstandard zur Diagnosestellung der AG ist die ÖGD (Ösophagogastroduodenoskopie) mit Biopsieentnahme und entsprechender histologischer Untersuchung [67].

Die Endoskopie wird mit HD-WLE (High-Definition-Weißlicht-Endoskopie), ggf. ergänzt um virtuelle Chromoendoskopie-Verfahren wie z.B. NBI (Narrow Band Imaging) [67], durchgeführt. Die NBI-Technik führt durch Nutzung von grünem und blauen Licht dazu, dass der rote Farbanteil aus dem Bild rausgefiltert und der blaue Farbanteil im Lichtspektrum betont wird. Als Bildeffekt werden somit z.B. Blutgefäße dunkel und somit stärker kontrastiert dargestellt. Mittels Einstellung der Wellenlänge des Lichtes kann außerdem die Eindringtiefe des Lichtes in das Gewebe beeinflusst werden [78].

Die ÖGD sollte unter ausreichender Luftinsufflation und bei möglichst sauberer Magenschleimhaut (z.B. durch Nüchternheit, Mukolytika) erfolgen, um eine gute Sicht für den Untersucher zu ermöglichen. Es wird das gesamte Magenlumen auf Gesamteindruck, Farbe und Textur der Mukosa, sowie Architektur der Magenfalten und Sichtbarkeit submukosaler Blutgefäße untersucht und fotodokumentiert. Hinweise auf eine mögliche AG bei der ÖGD sind z.B. multifo-

kale Blässe der Magenwand, durchschimmernde submukosale Blutgefäße, Verlust der Magenfältelung und/oder sichtbare Abgrenzbarkeit der atrophischen Läsionen durch auffällige lokale Magenschleimhautdefekte. Bereits daraus entstandene IM zeigen sich als weißliche, noduläre Areale mit teils gyriertem Muster und eventuellen hyperämischen Läsionen [85].

Ein Hinweis auf eine mögliche IM ist unter NBI-Technik das „Light Blue Crest“-Zeichen (LBC), welches feine, blau-weiße Linien auf dem Oberflächenepithel des Magens bezeichnet und entspricht histologisch dem Rand des Bürstensaums [95]. Weiterhin steht auch das unter NBI-Technik sichtbare „white opaque field“ (WOF)-Zeichen in Zusammenhang mit IM und anderen gastralen Tumoren, welches durch die Lichtstreuung an mikroskopisch kleinen Lipidtröpfchen entsteht und in der Mukosa akkumuliert. Beide Auffälligkeiten haben eine hohe Spezifität (>90%) für eine IM. Das LBC-Zeichen hat außerdem eine hohe Sensitivität von >90%, wohingegen diese für das WOF-Zeichen nur bei 50% für eine IM liegt [9,85].

Abbildung 11 zeigt die verschiedenen Auffälligkeiten bei AG und IM beim Endoskopieren.

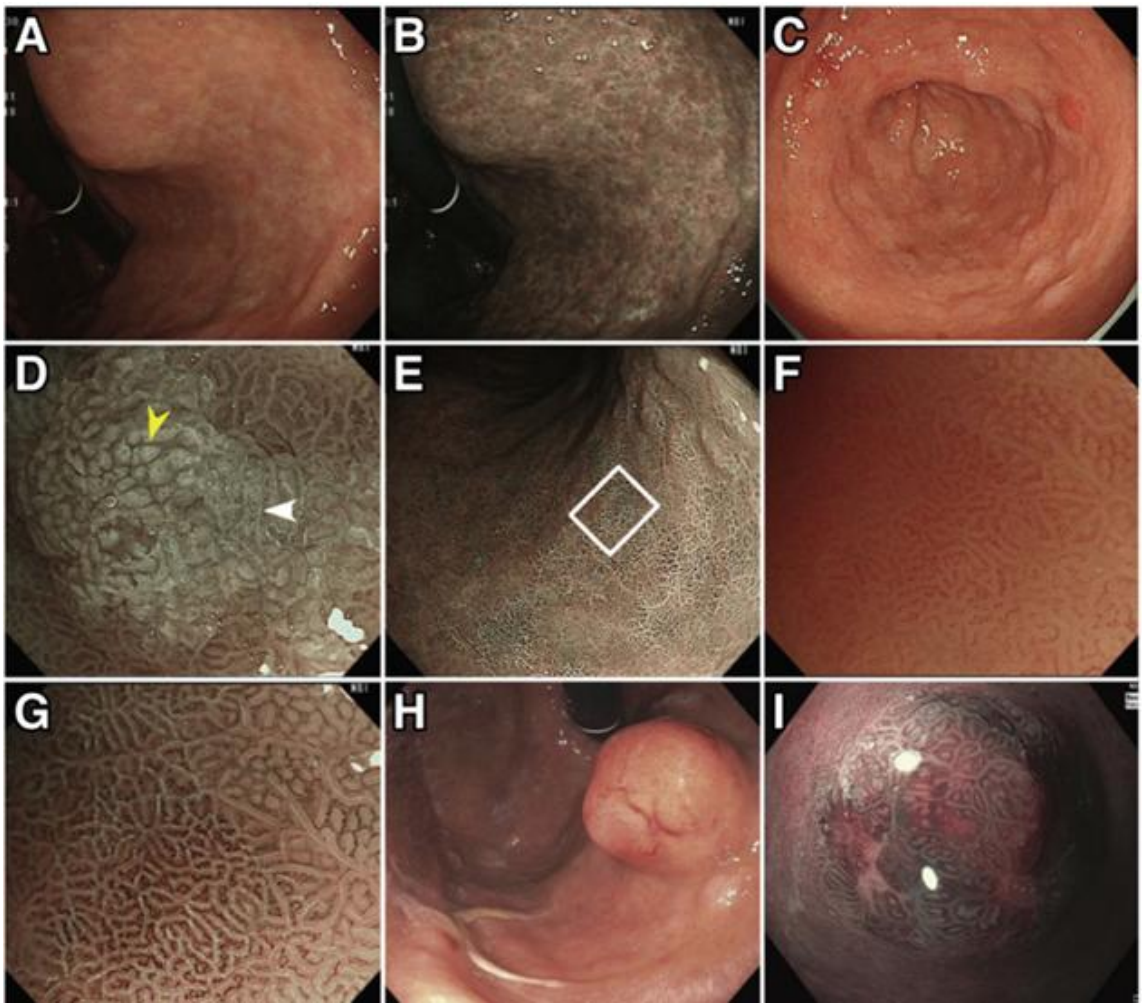


Abbildung 11 - Endoskopische Erscheinungsbilder der atrophischen Gastritis (AG), Intestinalen Metaplasie (IM) und neuroendokrinen Tumoren (NET)

A: AG - HD-WLE., Blässe, Verlust der Magenfältelung

B: AG – NBI.

C: IM – HD-WLE, noduläres Erscheinungsbild.

D: IM – NBI vergrößert, LBC-Zeichen (weißer Pfeil), WOF (gelber Pfeil).

E: IM – NBI ohne Vergrößerung.

F-G: IM – vergrößert.

H: NET – HD-WLE. **I:** NET – NBI.

Quelle: AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis: Expert Review. Shailja Shah, 2021, S. 19 [85]

Bei der Endoskopie wird empfohlen Biopsien nach dem aktualisierten Sydney-Protokoll [46] zu entnehmen (siehe Abb. 12). Dies entspricht also fünf Biopsien, welche zur histologischen Untersuchung in unterschiedlichen, beschrifteten Gefäßen eingesendet werden. Dies ermöglicht somit zusätzlich zur breiten topographischen Abdeckung eine hohe Zuverlässigkeit für einen HP-Nachweis [103]. Zwei Biopsien stammen aus dem Antrum jeweils entlang der kleinen und großen Kurvatur mit einem Abstand von 2-3 cm zum Pylorus. Weitere zwei werden aus dem Korpus entnommen, davon eine aus der kleinen Kurvatur mit proximal 4 cm Abstand zur Incisura angularis und eine aus dem mittleren Bereich der großen Kurvatur mit 8 cm Abstand zur Kardia. Die letzte Biopsie wird aus der Incisura angularis selbst entnommen. Zusätzlich sollen weitere Biopsien aus makroskopisch sichtbar auffälligen Arealen entnommen werden [46,85,103].

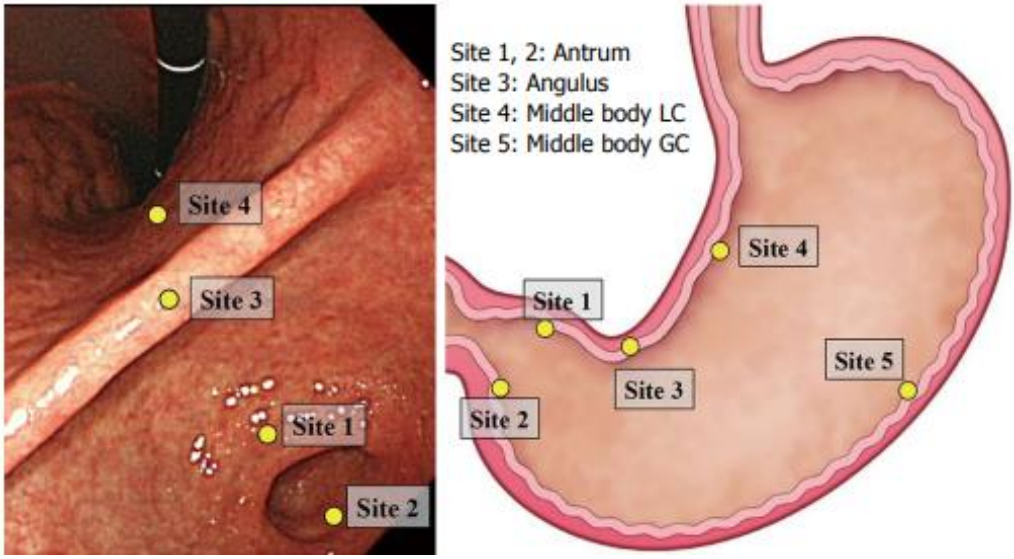


Abbildung 12 - Modifiziertes Sydney-Protokoll

LC: lesser curvature (kleine Kurvatur)

GC: greater curvature (große Kurvatur)

Quelle: Can endoscopic atrophy predict histological atrophy? Historical study in United Kingdom and Japan Kono, 2015, S. 3 [46]

Histologisch zeigt sich grundsätzlich eine chronische Inflammation der Mukosa. Definitionsgemäß kommt es zur Verkleinerung der Mukosadrüsen und Parietalzellen mit anschließendem vollständigem Verlust. Dieser Verlust wird entweder durch risikoarmes Bindegewebe oder durch Epithelien mit metaplastischen Charakteristika, sprich IM mit dem Risiko der Ausbildung von Dysplasien, ausgeglichen. Bei ursächlicher HP-Infektion tritt dieser Vorgang zunächst im Antrum auf und breitet sich nach oral über die kleine Kurvatur, seltener Korpus und Fundus, weiter aus. Dabei bilden sich anfangs kleine Areale der Atrophie, welche sich im Verlauf zu größeren Flächen ausdehnen. Die Parietalzellen atrophieren eher unvollständig. Die Autoimmungastritis-bedingte AG hingegen breitet sich vom Korpus ausgehend aus und zeigt einen vollständigen Verlust der Parietalzellen. Zusätzlich werden die Hauptzellen durch IM ersetzt und es kommt zur ECL-Zell-Hyperplasie, wobei diese sich meistens in der Nähe der Parietalzellen befinden [85]. Abbildung 13 zeigt entsprechende mögliche histologische Befunde.

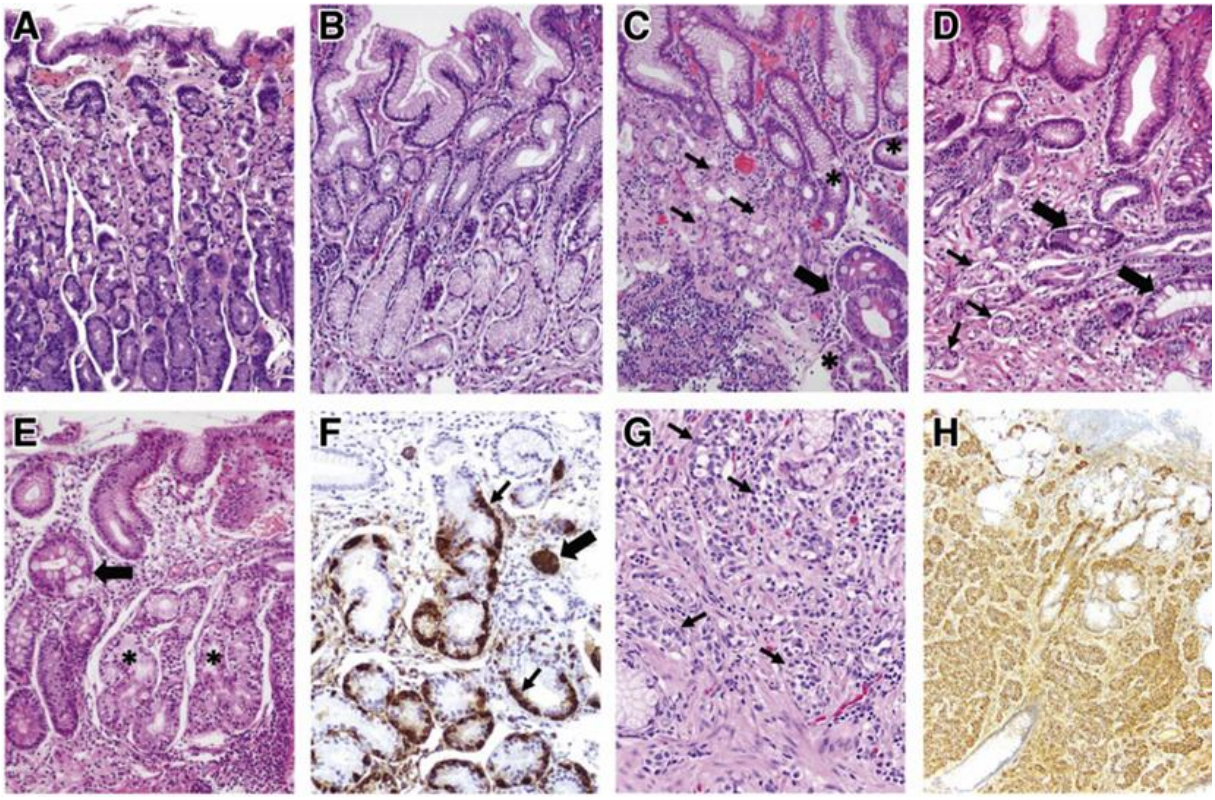


Abbildung 13 - Histologische Befunde bei Normalbefund, atrophischer Gastritis (AG) und neuroendokrinen Tumoren (NET)

A: Normalbefund der gastralen Mukosa mit kurzen Foveolae und dicht gepackten, geraden Drüsen aus überwiegend Parietalzellen.

B: Normalbefund der antralen Mukosa mit breiteren Foveolae und weniger dichten Drüsen aus überwiegend Nebenzellen.

C: Korpusmukosa mit chronischer Inflammation, teilweise Verlust der Parietalzellen, Metaplasien (Sterne) und IM (dicker Pfeil). Verbleibende Parietalzellen (schmale Pfeile).

D: Antrummukosa mit Schrumpfung, Drüsenverlust (schmale Pfeile) und IM (dicke Pfeile) umgeben von fibromuskulärem Gewebe der Lamina propria.

C-D: HP-bedingte AG.

E: Vollständiger Verlust der Parietalzellen, Hauptzellen durch intestinale Metaplasien ersetzt (Pfeil und Sterne).

F: Chromogranin-A-Färbung, ECL-Zell-Hyperplasie (Pfeile).

E-F: Autoimmungastritis-bedingte AG.

G-H: NET aus ECL-Zellen in HE- bzw. Chromogranin-A-Färbung.

Quelle: AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis: Expert Review Shailja Shah, 2021, S. 19 [85]

Zusätzlich kann die endoskopische und histologische Diagnosesicherung mittels serologischer Diagnostik durch Kombination verschiedener Marker Rückschlüsse auf die Lokalisation der AG (Antrum oder Korpus) sowie die noch vorhandene Magenfunktion erlauben. PG-I ist ein Proenzym und wird ausschließlich von den Drüsenzellen in Fundus und Korpus produziert, wohingegen PG-II, ebenfalls ein Proenzym, von allen Drüsenzellen des Magens sezerniert wird. G-17 wiederum wird im Antrum bzw. Pylorus von G-Zellen ausgeschüttet. Bei einer AG im Korpus kommt es zum Absinken der PG-I-Werte (< 70 ng/ml) und der PG-Ratio (< 3) und korreliert gleichzeitig mit der Progression und Schwere der AG. G-17 wäre in diesem Falle erhöht durch die entstehende Hypo- bzw. Achlorhydrie. Bei einer AG im Antrum hingegen ist ein erniedrigter G-17-Wert zu erwarten, da die entsprechenden G-Zellen atrophieren. Insgesamt weist ein erhöhter PG-II-Wert auf eine mögliche Inflammation der Mukosa hin, kann aber im Verlauf bei multifokaler AG erniedrigt sein [56].

2.2.6 **Therapie**

Die Therapie bei der AG besteht v.a. in der Behandlung von Mangelzuständen und das Erkennen von weiteren Komorbiditäten sowie Komplikationen wie die maligne Entartung der IM. Eine Heilung ist in der Regel nicht möglich [103].

Eine wichtige Säule stellt die Behandlung von HP-positiven Patienten dar. Eine medikamentöse Eradikation des HP kann in frühen Stadien der AG eine Rückbildung ermöglichen und soll dementsprechend oberste Priorität haben [85]. Die Basis dieser Therapie stellt zunächst ein hochdosierter PPI 2x täglich für mindestens 10 Tage in Kombination mit verschiedenen antibiotischen Schemata dar [57].

Die zweite wichtige Säule stellen Kontrollendoskopien dar. Bei fortgeschrittener AG oder IM OLGA/OLGIM III/IV werden diese im Durchschnitt alle 3 Jahre empfohlen, wobei der Zeitraum auch kürzer oder sogar länger je nach individuellen Risikofaktoren ausfallen kann [103]. Darunter fallen u.a. familiäres Auftreten sowie Migration aus Ländern mit höheren Inzidenzraten des Magenkarzinoms, eine langjährige Raucheranamnese, Ernährungsweisen und eine persistierende HP-Infektion. Bei einer AG im Rahmen einer Autoimmungastritis gibt es keine klaren Empfehlungen zum Intervall, die Tendenz liegt hierbei jedoch bei alle 3-5 Jahre [85,103].

Bei einem Teil der Patienten mit AG entsteht im Verlauf ein gastraler NET, somit ist die Früherkennung einer AG auch in Hinblick auf die Entstehung eines NET essenziell. Die Therapie ist abhängig von der Größe sowie Infiltrationstiefe bzw. Stadium des Tumors. Ziel ist eine endoskopische En-bloc- oder chirurgische Resektion mit anschließenden Kontroll-ÖGDs im Verlauf [85,103].

Zusätzlich soll bei begründetem Verdacht jeder Patient auf einen Mangel, v.a. an Vitamin B12 oder Eisen, untersucht und dieser entsprechend behandelt werden. Es wird zusätzlich empfohlen bei Patienten mit Autoimmungastritis eine Untersuchung auf autoimmunbedingte Schilddrüsenerkrankungen durchzuführen, da hierbei die höchste Assoziation besteht [85].

2.3 Künstliche Intelligenz

Künstliche Intelligenzen (KI) sind Computerprogramme, welche sich wiederholende Muster in großen Datensätzen erkennen und analysieren können. Speziell das „Lernen an Beispielen“ zur Bildverarbeitung, auch *Deep Learning* (DL) genannt, ähnelt in seiner Arbeitsweise dem Nervensystem des Menschen, da die Bilddaten über mehrere verknüpfte Schichten verarbeitet werden [10,80]. Die modernste Variante des „*Deep Learnings*“ sind dabei die *Convolutional Neuronal Networks* (*CNN*, gefaltetes künstliches neuronales Netz), deren Fokus auf der Bildverarbeitung und der eigenständigen Interpretation und Klassifizierung der Daten liegt [10]. Diese werden mittlerweile vielfältig im medizinischen Bereich, beispielsweise zur endoskopischen Bildanalyse von Polypen oder Adenomen, eingesetzt [36,43].

2.3.1 Historische Entwicklung

Den Ursprung der heutigen neuronalen Netze bildet der Entwurf des „Perceptrons“ in den 1950er Jahren von Frank Rosenblatt. „Perceptron“ lässt sich von dem englischen Begriff „perceive“ bzw. dem lateinischen „percipere“ ableiten und bedeutet so viel wie „Wahrnehmung“. Es war das erste einfache künstliche neuronale Netzwerk, welches auf der Basis von „trial and error“ lernfähig war und somit mit dem Gehirn des Menschen verglichen wurde [75,80]. So bildete sich in den späten 1960er Jahren durch weitere Forschung die neurowissenschaftliche Hypothese, dass sich komplexe Objekte von Ebene zu Ebene weiter in Einzelheiten zerlegen lassen, bis eine bestimmte Nervenzelle das spezifische Objekt erkennt und darauf durch Aktionspotentiale reagiert und den Vorgang abschließt [80].

In den 1970er Jahren war der Nutzen der KIs begrenzt, sie dienten v.a. als Suchmaschinen in der Medizin [34]. Mit der Weiterentwicklung konnten diese dann in den 2000er Jahren bereits einfache Datensätze analysieren, jedoch noch keine Sprache oder Bilder [80].

Der eigentliche Durchbruch der heutigen CNN-Technik gelang in der „ImageNet-Challenge“. Dies ist ein jährlicher Wettbewerb seit 2010 bei dem weltweit jeder seine kreierten automatisierten Algorithmen zur Bilderkennung von alltäglichen Gegenständen einreichen kann, z.B. Erkennen von Katzenrassen und Flugzeugmodellen. Trainiert werden diese privaten Algorithmen mit einem zur Verfügung gestellten Datensatz aus tausenden Kategorien an Objekten. Abschließend werden diese Algorithmen mit einem speziellen, bisher unbekannten Test-Da-

tensatz auf ihre Genauigkeit geprüft und somit der Sieger verkündet. Bis zum Jahre 2012 wurden diese Algorithmen hauptsächlich auf Basis von „machine learning“, also automatisierte Informationen aus Formen, Strukturen, Farben und Farbübergängen kombinieren, erstellt und ergaben deswegen lange Fehlerraten über 25% [80]. Erst ab dem Durchbruch mittels CNN war es möglich komplexere Bilder aufgrund einer abstrakteren Bildanalyse zu kategorisieren und somit die Fehlerrate schnell unter 3% zu senken [47,80].

Mittlerweile werden sie v.a. zur Bildauswertung, Sprachverarbeitung und Automatisierung einfacher Prozesse verwendet, so dass validierte KIs in der Medizin seit 2010 zur Polypenerkennung in der Endoskopie ihre Rolle eingenommen haben [34,80]. Es gelingt in der Zwischenzeit auch einigen KIs Menschen in Strategiespielen zu besiegen und bessere Ergebnisse als diese zu erzielen [43].

2.3.2 **Relevante Definitionen zur künstlichen Intelligenz**

Machine Learning

„*Machine Learning*“ (maschinelles Lernen) beschreibt die Lernfähigkeit von Algorithmen, welche es ihnen ermöglicht aus Datensätzen Vorhersagen oder Entscheidungen zu treffen. Dadurch können sie ohne gezielte Programmierung anhand von neuen Daten Modelle erstellen [10]. Dies wird im Alltag beispielsweise bei den heutigen Wettervorhersagen, insbesondere um Wetterveränderungen der nächsten Stunden präzise zu erkennen, genutzt [12]. In der Medizin hingegen können solche Systeme atypische Muster von normalen unterscheiden [10].

Deep Learning

Eine spezifische und heutzutage häufig angewandte Methode des „*Machine Learnings*“ ist das bekannte „*Deep Learning*“ (tiefes Lernen) [22]. Es beschreibt die Fähigkeit von künstlichen Intelligenzen über mehrere verknüpfte Schichten aus neuronalen Netzwerken verschiedene Muster und Charakteristiken von Daten zu erkennen. Somit stellt es eine wesentlich komplexere Verarbeitungsmethode dar und wird aufgrund seiner höheren Leistungsfähigkeit für Sprach- und Bildverarbeitung genutzt [10].

Convolutional Neuronal Network

Ein komplexerer Ansatz des „*Deep Learnings*“ sind *CNN* (*Convolutional Neuronal Network*), also tiefe neuronale Netzwerke, deren Fokus auf der Verarbeitung von Bildern und deren Daten liegt. Dies erreichen sie, indem sie mittels „*Deep Learning*“ über mehrere zusammenarbeitende Schichten die spezifischen Charakteristiken der Bilder automatisiert erlernen und extrahieren. Anschließend werden die gewonnenen Daten durch die KI eigenständig interpretiert und letztendlich in entsprechende Kategorien klassifiziert. Die wichtigsten Schichten der CNNs setzen sich aus *Convolutional Layers*, *Pooling Layers* und *Fully Connected Layers* zusammen [10].

2.3.3 **Funktionsweise der CNN-basierten künstlichen Intelligenz**

Der erste Schritt bevor eine KI Bilder zur Analyse verarbeiten kann ist die eigentliche Bildaufnahme und die Optimierung der Bildqualität und -größe. Anschließend werden diese der KI präsentiert, welche die Bilder in Form von zweidimensionalen Pixelwerten wahrnimmt. Dabei werden die Anordnung dieser zweidimensionalen Werte als Matrix bezeichnet und beschreibt zusätzlich die unterschiedlichen Farbkomponenten [49].

Somit hat ein Farbbild der Pixelgröße 224x224 automatisch drei Matrizen durch die Farben Rot, Grün und Blau (RGB), wobei jedem Unterschied in der Farbintensität ein eigener Wert zugeteilt wird. Es entsteht als Eingabe in der KI also eine 224x224x3-Matrix, wodurch für jedes Bild komplexe Datenstrukturen entstehen, welche wiederum für die Analyse und Mustererkennung der Bildinhalte durch Algorithmen notwendig sind [47,49].

Nachdem der Algorithmus das Bild erhalten hat, wird dieses mehrere Schichten durchlaufen bevor ein Ergebnis zustande kommt (siehe Abb. 14). Dabei besitzt jede CNN *Convolutional Layers, Pooling Layers*, und *Fully Connected Layers*, wobei optional je nach Fragestellung auch noch so genannte Aktivierungs- und Softmax-Schichten vorhanden sein können [47,49].

Convolutional Layers

Das Kernstück der künstlichen Intelligenz besteht aus mehreren, nacheinander geschalteten *Convolutional Layers*, also Faltungsschichten. Ihre Hauptaufgabe ist die Erkennung der verschiedenen Bildmerkmale, indem ein Filter unterschiedlicher Größe, beispielsweise ein 3x3-

Bereich, über die unterschiedlichen Abschnitte des Bildes geschoben wird. Anschließend wendet der Filter mathematische Berechnungen an um sogenannte Merkmalskarten (*Feature Maps*) zu erstellen [49].

Zum besseren Verständnis der Funktionsweise der Faltungsschichten kann ein solcher Filter z.B. Ecken im Bild erkennen. Dabei wird beim Verschieben des Filters über das Bild in jedem Bereich ein neuer Wert errechnet, um an der entsprechenden Stelle eine mögliche Ecke zu beschreiben. Dabei entsteht beim Erkennen einer neuen Ecke jedes Mal eine Merkmalskarte, um dieses Merkmal erneut im Bild zu kennzeichnen. Dabei spielt die entscheidende Rolle nicht die genaue Positionierung der Ecke, sondern ihre Lage zu anderen Merkmalen [49].

Die Komplexität dieser Faltungsschichten wird bei tieferen Schichten größer. Dadurch können die ersten Schichten z.B. Ecken, Kanten, (Gewebe-)Texturen oder einfache Muster erkennen, wobei die tieferen Schichten durch die Kombination der bereits bestehenden Merkmalskarten eigenständige Anomalien oder sogar Tumoren erkennen können [47,79,80].

Pooling Layers

Anschließend werden die entstandenen Werte, welche durch die große Anzahl an Merkmalskarten erstellt wurden, über mehrere Pooling-Schichten reduziert [79]. Dadurch wird insgesamt die räumliche Größe unter Beibehaltung der wichtigsten Informationen und die Rechenlast des Algorithmus verringert [49,79].

Die häufigste Methode ist das Max-Pooling. Dabei wird der größte Wert in einem bestimmten Bereich ausgewählt und beibehalten, um eine neue Merkmalskarte daraus zu generieren, welche die Werte der bisherigen Merkmalskarte zusammenfasst [79]. Veranschaulicht bedeutet dies, dass bei Betrachtung eines 3x3-Bereiches des Bildes mit den Werten (1, 2, 3, 4, 5) der Wert 5 erhalten bleibt und eine neue Merkmalskarte ergibt.

Eine weitere Methode ist das Sub-Sampling. Dabei wird aus dem betrachteten Bereich der Durchschnittswert aus den erstellten Werten ermittelt und für die weitere Bearbeitung beibehalten [49].

Der Vorteil dieser Schichten ist, dass sie die künstliche Intelligenz beständig gegenüber Veränderungen der Bildqualität und -größe machen, jedoch auch weniger anfällig für Fehler bei Verzerrungen und Änderungen der Merkmalspositionen wird [49]. Es hat sich jedoch gezeigt, dass Max-Pooling bei großen Änderungen bei der Darstellung der Merkmale konstantere Ergebnisse zeigt [47].

Fully Connected Layers

Die letzten Schichten einer CNN-basierten künstlichen Intelligenz stellen eine oder mehrere *Fully Connected Layers* dar [47,49]. Sie sind mit den vorherigen Schichten verbunden und erhalten somit sämtliche Informationen über jegliche Merkmale des analysierten Bildes [47].

In diesen Schichten erfolgt somit die Interpretation der miteinander verknüpften Bildinformationen und kann diese entsprechend klassifizieren. Die Ausgabe kann also Diagnosen zustimmen bzw. ausschließen (Ja oder Nein) oder Läsionen konkret anzeigen (z.B. Polypen) [80].

Optionale Aktivierungs- und Softmax-Schichten

Die *Aktivierungsschichten* finden sich meist nach den *Convolutional Layers* (selten nach den *Fully Connected Layers*) und wenden eine *ReLu* (*Rectified Linear Unit*) -Funktion an. Es werden alle negativen Werte einer Merkmalskarte auf null gesetzt, wobei positive Werte unverändert bleiben [47]. Dies bedeutet beispielsweise, dass durch die Durchführung der ReLu-Funktion innerhalb einer Merkmalskarte ein Wert von -3 auf 0 gesetzt wird, wohingegen ein Wert von 1 unverändert bleiben würde.

Dadurch wird die Merkmalskarte also umgewandelt indem sie zusätzlich nicht-lineare Eigenschaften erlernt und beschleunigt somit das weitere Erlernen von komplexeren Mustern [47].

Die *Softmax-Schichten* ermöglichen die zusätzliche Angabe einer Wahrscheinlichkeit in Prozent für die Aussage der Fully Connected-Schicht [47,80]. Somit kann der Algorithmus selbst anzeigen, wie sicher er sich bei der jeweiligen Aussage ist [80]. Beispielsweise kann er beim Nachweis bzw. Ausschluss einer Erkrankung somit 80% für die Klassifizierungsmöglichkeit „Ja“ und 20% für die Möglichkeit „Nein“ eingeben.

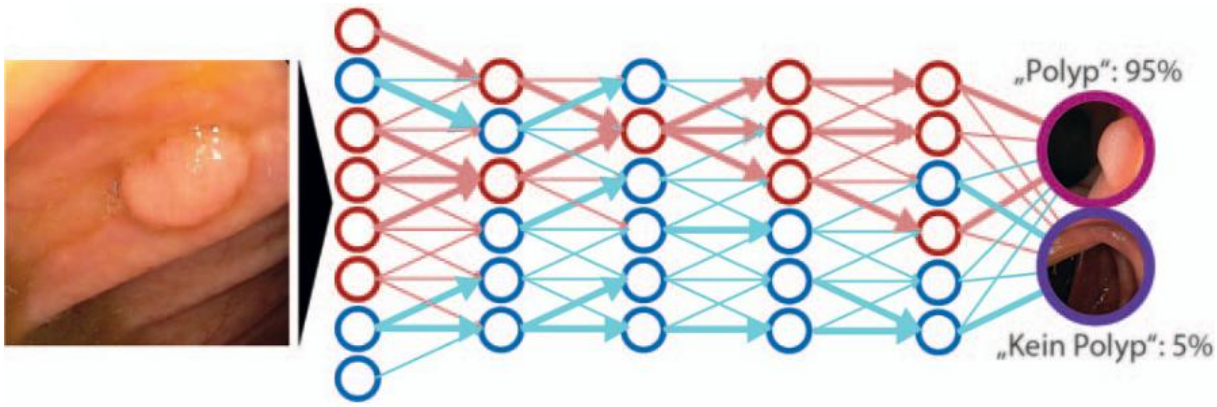


Abbildung 14 – Vereinfachte Darstellung des Aufbaus einer CNN-basierten künstlichen Intelligenz anhand einer Polypenerkennung

Links wird das zu analysierende Bild des Polyps für den Algorithmus und rechts dessen Ergebnis dargestellt. Durch Eingabe des Bildes beginnt der Algorithmus über *Convolutional Layer* die verschiedenen Merkmale zu erkennen, wobei die roten und blauen Pfeile stark vereinfacht die Entstehung unterschiedlicher Merkmalskarten darstellen. Diese werden über weitere *Convolutional Layer* und *Pooling Layer* weiterverarbeitet (keine genaue Trennung in der Darstellung, Schichten zwischen Beginn und Ende). Die letzten Schichten (große Kreise rechts) stellen die *Fully Connected Layers* dar, welche eine Interpretation durchführen (Polyp vs. kein Polyp). Zusätzlich sind diese hier mit einer *Softmax-Schicht* gekoppelt, um die prozentuale Wahrscheinlichkeit des Ergebnisses anzuzeigen (95% vs. 5%).

Quelle: Angelehnt an „Artificial Intelligence in Endoscopy: Deep Neural Nets for Endoscopic Computer Vision - Methods & Perspectives“ von Schmitz R. et al, 2019 [80]

2.3.4 **Bisherige Einsatzbereiche in der Medizin**

Allgemeiner Nutzen in der Medizin

Mittlerweile wurden vielfältige KI-Systeme für verschiedene Bereiche der Medizin, u.a. Anästhesiologie, Kardiologie, Neurologie, Ophthalmologie und die Onkologie etabliert. Der Fokus liegt meist auf der Verarbeitung von Patientendaten, Fragebögen, Blutwerten oder Beatmungs- und Vitalparametern [48]. Zahlreiche Arbeiten zur Nutzung von bildanalysierenden KIs stammen auch aus dem Bereich der Radiologie, Dermatologie und Gastroenterologie, wobei hier v.a. Endoskopie-Bilder verwendet werden [44,48,100]. Bei dem Großteil der KIs unabhängig des Systems und der Disziplin werden hierbei statische Daten, also einzelne Bilder, Fotodokumentationen oder Werte zu einem bestimmten Zeitpunkt, statt dynamische Daten wie Vi-

deoaufnahmen von mehreren Untersuchungen und somit der Verlauf über einen längeren Zeitraum, verwendet [48]. Allgemein wird zur Beurteilung der Bildergebnisse bei bilderverarbeitenden KIs das Ergebnis der Histopathologie als Grundwahrheit angenommen [43].

Eine große Hürde bei der Entwicklung von KIs ist die klinische Validierung, um diese somit im Alltag zu benutzen. Häufig funktionieren KIs an einem erprobten und vorselektierten Datensatz, versagen mutmaßlich jedoch dann in Studien bei der Testung an einem randomisierten Patientenklientel. Eine Ursache hierfür sind z.B. zu kleine Test-Datensätze, wodurch die KI nicht alle Aspekte und Erscheinungsbilder einer Erkrankung häufig genug analysiert. Andererseits zeigen unselektierte Datensätze oft auch nicht optimale Bilder im Gegensatz zu den qualitativ hochwertigen aus dem Training [57].

Ein großer Nutzbereich der bildanalysierenden KIs ist die Radiologie, wobei hier zur klinischen Validierung besonders retrospektive Kohortenstudien eine Rolle spielen [44]. Die meiste Anwendung findet sich in der Auswertung von MRT-Bildern in der Neuroradiologie [65]. Dabei werden diese genutzt, um Tumoren zu erkennen und zu klassifizieren [44], jedoch auch um mögliche Merkmale im Sinne einer möglichen Alzheimer-Demenz oder um mögliche Herde passend zu einer multiplen Sklerose zu erkennen. Andere Applikationen können einheitliche Krankheitsbilder wie Schlaganfälle, intrazerebrale Blutungen oder intrakranielle Aneurysmen erkennen [65]. Klinisch validierte Programme sind z.B. „e-ASPECTS“© (Alberta Stroke Program Early CT Score) zur Auswertung nativer CT-Bilder hinsichtlich ischämischer Schlaganfälle [64] und z.B. „e-CTA“© (Computed Tomography Angiography) zur Detektion von entsprechenden intrakraniellen Gefäßverschlüssen [28] der Firma Brainomix.

Ein weiterer großer Anwendungsbereich ist das Einschätzen von pulmonalen Noduli in CT-Bildern bezüglich ihrer möglichen bildmorphologischen Malignität und somit eine mögliche Früherkennung von Lungenkarzinomen [25,44], beispielsweise mittels „CAD“© (computer assisted diagnosis system) als klinisch validiertes Tool [60]. Mithilfe der Auswertung mittels KI wurden weniger Knoten in den untersuchten Lungen übersehen und somit potenzielle Malignitäten in früheren Stadien erkannt. Dabei entstehen wenige falsch-positive Ergebnisse, welche zusätzliche Diagnostik zur weiteren Einschätzung benötigen. Gleichzeitig wurde durch ein paralleles Auswerten durch die KI und die Radiologen die Bearbeitungszeit reduziert [25].

Auch zur Einschätzung von Leberpathologien werden Studien zur Anwendung von KIs durchgeführt. Die meisten zielen dabei auf die Erkennung von hepatozellulären Karzinomen ab, wobei auch wenige zur Erkennung von hepatischen Metastasen getestet werden [4]. Andere

Studien wie die von Cheng et al. untersuchen KIs zur Klassifizierung von unterschiedlichen hepatischen nodulären Läsionen anhand histopathologischer Bilder, u.a. hepatozelluläres Adenom, Karzinom, dysplastischer Nodulus oder fokale noduläre Hyperplasie [13]. Andere klinisch validierte KIs wie „QP-Liver“© von Quibim können in MRT-Bildern die Leber segmentieren zur vereinfachten Beurteilung des Leberverfettungsgrades und zur Eisenquantifizierung bei chronischen Lebererkrankungen [55].

Nutzen in der gastroenterologischen Endoskopie

Die Nutzung von KI-Unterstützungssystemen in der gastrointestinalen Endoskopie ist mittlerweile gut etabliert. Eine Studie von Repici A. et al von 2020 zeigt, dass durch Nutzung einer KI-Unterstützung die Polypendetektionsrate von 40% auf 54% angehoben werden kann. Besonders kleinere Läsionen und Polypen von unter 10 mm werden durch die KIs besser erkannt, ohne dabei Zeitverzögerungen bei der Untersuchung zu verursachen [71].

Eine klinisch validierte KI für die Erkennung von intestinalen Läsionen wie Polypen oder Adenomen ist „GI-Genius“© von Medtronic [36]. Sie benötigt keinen separaten Bildschirm zur Analyse und kann somit in Echtzeit auf dem gleichen Monitor bereits auffällige Läsionen für den Untersucher mittels grünem Kasten markieren [71]. Dabei zeigt sie bei mehr als drei Viertel der Untersuchungen eine schnellere Echtzeiterkennung der Läsionen von bis zu 3,8 Sekunden und weist dabei bei weniger als 1% der Bildframes ein falsch-positives Ergebnis auf. Insgesamt werden somit weniger malignitätsverdächtige Läsionen übersehen und gleichzeitig wird die Untersuchung objektiver und die Auswertung beschleunigt [36].

Eine weitere KI zur Erkennung von Polypen ist „ENDO-AID“© von Olympus, welche in einer retrospektiven Studie von Wong YT et al. 2022 klinisch validiert wurde. Dabei zeigte sich eine verbesserte Polypendetektionsrate mit KI von fast 65% im Gegensatz zu 46% bei den Untersuchungen ohne. Die KI kann die verdächtigen Polypen direkt am gleichen Monitor ebenfalls mit einem grünen Kasten markieren oder in einem seitlichen kleinen zusätzlichen Bild die eigentliche Auffälligkeit markieren [97].

Ein weiterer Ansatz ist die endoskopische Erkennung eines Barrett-Ösophagus, eine Umwandlung des ösophagealen Plattenepithels in intestinales Zylinderepithel durch den Kontakt mit der Magensäure im Rahmen einer Refluxerkrankung, welcher je nach Ausprägung eine

mögliche Präkanzerose für ein Ösophaguskarzinom darstellt [94]. In einer retrospektiven Studie von de Groof A. et al von 2020 wurden die Ergebnisse einer KI, welche einzelne Bilder in neoplastische oder nicht-dysplastische Barrett-Ösophagus einteilt, mit denen der endoskopierenden Ärzten verglichen. Dabei erreichte die KI eine Accuracy von 88%, eine Sensitivität von 72% und eine Spezifität von 83%, wohingegen die Untersucher jeweils 73%, 72% und 74% erreichten und somit insgesamt etwas schlechter abschnitten [17]. Eine weitere Methode zeigt die Studie von Hussein M. et al von 2022, in welcher eine KI die entsprechenden Veränderungen des Ösophagus bereits während der Untersuchung am Bildschirm erkennt und anzeigt, um gezielte Biopsien durchzuführen. Auch hier ist die Performance des Algorithmus besser als die der Untersucher [41].

Doch auch für die Erkennung eines manifesten Ösophaguskarzinoms gibt es bereits Ansätze. So zeigt die Arbeit von Guo L. et al aus dem Jahr 2020 zeigt eine KI, die während der Untersuchung frühe Stadien eines Plattenepithelkarzinoms erkennen kann. Dabei werden einzelne Lokalisationen gelb angezeigt, wenn eine hohe Wahrscheinlichkeit für ein Plattenepithelkarzinom vorliegt bzw. blau, wenn diese niedrig ist. Sie wurde anhand von einzelnen Bildern als auch mit Videos trainiert und optimiert [33]. Weiterhin ist auch die Erkennung der Invasionstiefe von ösophagealen Plattenepithelkarzinomen ein wichtiger Faktor. Damit hat sich die Publikation von Shimamoto Y. et al von 2020 beschäftigt, in der ein Algorithmus getestet wurde, der während der Untersuchung bereits Berechnungen zur Invasionstiefe macht und in den meisten relevanten diagnostischen Parametern bessere Ergebnisse als die Untersucher erzielte [87]. Auch die Studie von Nagakawa K. et al aus dem Jahr 2019 untersuchte die Ergebnisse einer KI zur Berechnung der Invasionstiefe, welche auch hier besser ausfielen als bei den Untersuchern. Dabei beschränkten sie sich auf eine Ausbreitung innerhalb der Mukosa (oberflächlichste Schicht) und submukosale Anteile [62].

Weitere Möglichkeiten der KIs in der gastroenterologischen Endoskopie ist die Erkennung eines gastroösophagealen Refluxkrankheit (GERD, gastroesophageal reflux disease) oder auch Refluxösophagitis genannt, bei der es durch aufsteigende Magensäure zu Verletzungen der ösophagealen Schleimhaut kommt [94]. Die Studie von Gulati et al. von 2019 untersuchte eine KI zur Erkennung einer Refluxkrankheit während der Untersuchung und erreichte eine Sensitivität von 67% und eine Spezifität von 92% [31].

Auch wird versucht mittels Algorithmen manifeste Magenkarzinome zu erkennen. Die Arbeit von Ueyama H. et al von 2021 präsentiert einen solchen Algorithmus, der in der Lage ist, anhand der fotodokumentierten Untersuchung viele Bilder in kurzer Zeit zu analysieren und

zeigt dabei eine diagnostische Genauigkeit von fast 99%. Dabei zeigt sich, dass die nicht erkannten frühen Magenkarzinome schwer von Gastritiden zu unterscheiden waren [93]. Weiterhin untersucht die Studie von Nagao S. et al aus 2020 eine KI zur Erkennung der Invasionstiefe beim Magenkarzinom. Dafür wurden über 16.000 Bilder von 1.084 Patienten von 2013 – 2019 retrospektiv durch die KI ausgewertet und zeigte dabei eine Accuracy von 94%, sie kann jedoch zum aktuellen Zeitpunkt noch nicht während einer laufenden ÖGD angewendet werden [61].

Eigene KI-Algorithmen zur Erkennung der proximalen Atrophie und der eosinophilen Ösophagitis

Beide KI-Algorithmen zur Detektion von präkanzerösen Bedingungen des proximalen Magens und der EoE basieren auf dem CNN-Prinzip des „*Deep Learnings*“. Vereinfacht dargestellt bedeutet dies, dass die importierten Bilder anhand mehrerer Merkmale Schritt für Schritt über mehrere Ebenen anhand vorab erlernter Kriterien analysiert werden [29,80].

Der Algorithmus für die proximale Atrophie wurde mit einem Datensatz von 200 Bildern von 101 Patienten trainiert, wovon 100 Bilder von 37 Patienten histologisch den Nachweis einer atrophischen Gastritis hatten und 100 Bilder ohne von 64 Patienten stammten. Die gemachten Bilder entstanden durch unterschiedliche Endoskope mittels HD-WLE-Technik und wurden anonymisiert extrahiert. Die hierfür verwendeten Bilder stammten aus Fundus und Corpus und wurden nicht-standardisiert durchgeführt, bedeutet also z.B. unterschiedliche Winkel des Endoskops, Kontamination der Mukosa durch Essen und Mukus oder schlechte Belichtung sowie unzureichende Insufflation des Magens. In mehreren Durchläufen wurden dem Algorithmus diese 200 Bilder gezeigt, um diesen zu trainieren und optimieren [29].

In einem unabhängigen Test wurden dem Algorithmus zur proximalen Atrophie anschließend in der Studie von Guimarães P. et al. 2020 insgesamt 70 neue Bilder von 35 Patienten gezeigt, darunter 30 Bilder mit Atrophie von 13 Patienten und 40 Bilder ohne von 22 Patienten, welche ebenfalls durch sechs Endoskopiker beurteilt wurden. Ziel hierbei war es, die KI klinisch zu validieren ohne weitere Optimierungsarbeiten durch diesen Datensatz durchzuführen. Die KI zeigte eine klinische Genauigkeit (ACC, Accuracy) von 93% bei einer Sensitivität von 100% und einer Spezifität von 87,5%. Die Endoskopiker hingegen hatten bei allen drei genannten Werten lediglich 80%. Somit schnitt die KI besser ab und zeigte eine falsch-positiv Rate von 12,5% [29].

Der Algorithmus für die eosinophile Ösophagitis kann eine EoE, eine ösophageale Candidiasis oder einen gesunden Ösophagus unterscheiden. Dafür wurde dieser mit einem Datensatz von insgesamt 406 Bildern von 103 Patienten trainiert und durch mehrere Durchläufe optimiert, davon stammten 164 Bilder von 25 Patienten mit histologisch nachgewiesener EoE, 107 von 46 Patienten mit ösophagealer Candidiasis und 135 Bilder von 32 Patienten mit Normalbefunden. Die Bilder sind unter ähnlichen Bedingungen entstanden wie für den Algorithmus der proximalen Atrophie, also unter HD-WLE-Technik und ähnliche nicht-standardisierten Aufnahmetechniken [30].

Auch dieser Algorithmus wurde anschließend in einem unabhängigen Test in der Studie von Guimarães P. et al. 2022 mit neuen Bildern bezüglich seiner Leistungsfähigkeit geprüft. Dabei wurden insgesamt 78 Bilder von 31 Patienten genutzt. Darunter waren 26 Bilder von 7 Patienten mit nachgewiesener EoE, 25 Bilder von 16 Patienten mit ösophagealer Candidiasis und 27 Bilder von 8 Patienten mit Normalbefund. Dabei erreichte die KI eine Accuracy von 91,5%, wohingegen die Untersucher eine Accuracy von 83% erreichen. Die Sensitivität der KI lag bei 87% und die Spezifität bei 94% [30].

2.4 Ziele und Fragestellung

Die Arbeit zielt darauf ab, zwei selbst entwickelte KI-Algorithmen zur Bildauswertung in der gastroenterologischen Endoskopie hinsichtlich ihrer Ergebnisse zur Erkennung präkanzeröser Bedingungen des proximalen Magens bzw. der EoE in einem unselektierten, prospektiv analysierten Patientenkollektiv zu evaluieren. Dabei wird das Ergebnis der KI mit der Einschätzung der die Endoskopie durchführenden Ärzte und dem tatsächlichen Vorliegen einer Erkrankung als Grundwahrheit verglichen und statistisch ausgewertet. Hiervon kann dann abgeleitet werden, ob die Systeme im klinisch-endoskopischen Alltag verlässliche Ergebnisse liefern.

3 Material und Methodik

3.1 Patientenkollektiv

Bei der vorliegenden Arbeit handelt es sich um eine prospektive Studie, in die insgesamt 650 evaluierbare von insgesamt 830 aufgeklärten Patienten eingeschlossen wurden, welche im Zeitraum von Februar 2021 bis April 2022 in der Klinik für Innere Medizin II, Schwerpunkt Gastroenterologie/Endoskopie, des Universitätsklinikum des Saarlandes endoskopiert wurden. In diesem Zeitraum wurde allen ambulant, sowie teilweise stationär, mittels Ösophagogastroduodenoskopie untersuchten Patienten die Teilnahme an der Studie angeboten und die Patienten für die Studie aufgeklärt.

Insgesamt konnten 180 untersuchte Patienten wegen unterschiedlicher Gründe wie unvollständige Einverständniserklärung, Mehrfachvorstellung oder Nutzung von Bildmaterial dieser Patienten bei der Etablierung der Algorithmen nicht ausgewertet werden.

In die Studie wurden alle Patienten aufgenommen, die im angegebenen Zeitraum untersucht wurden, und von denen eine Zustimmung zur Verwendung des Bildmaterials vorlag, unabhängig von möglichen Vorerkrankungen oder der Indikation zur Untersuchung.

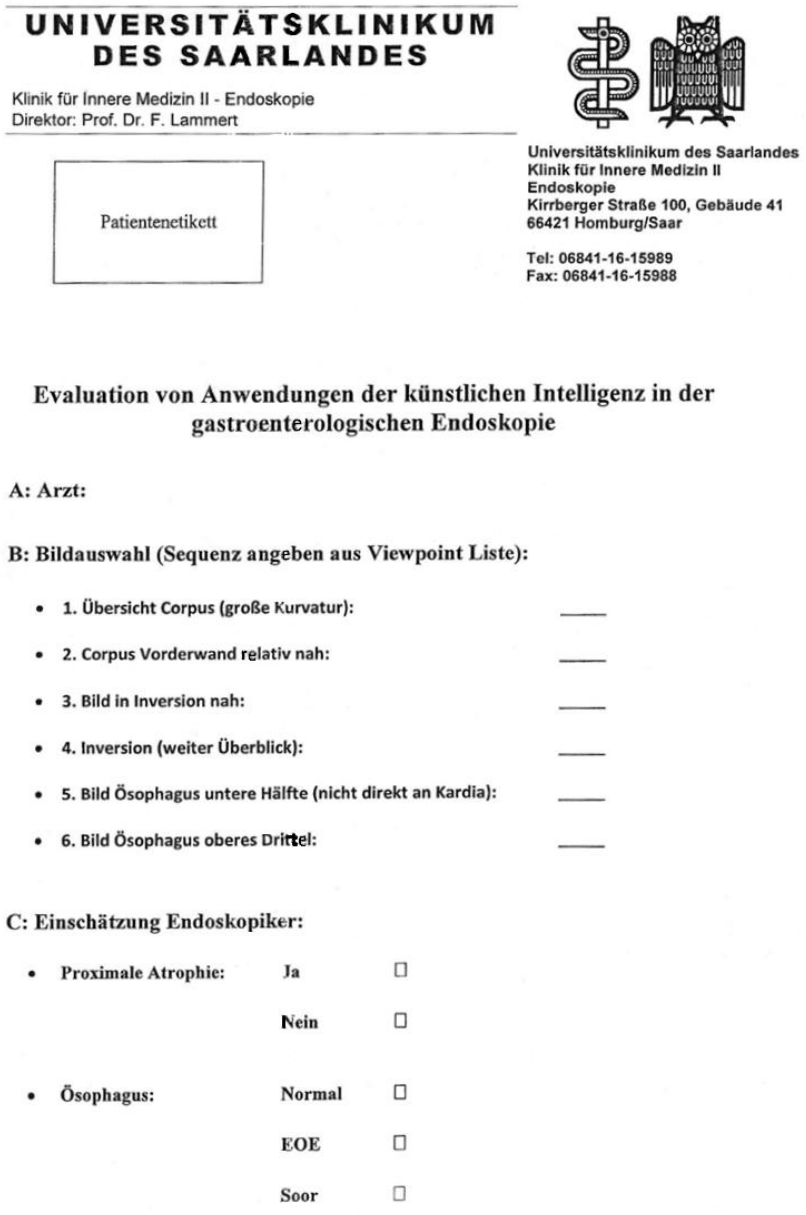
Die Ausschlusskriterien waren:

1. eine fehlende oder unvollständige Einwilligung
2. fehlende Möglichkeit zur Durchführung einer Gastroskopie
3. bereits früherer Einschluss und Auswertung, um somit eine Dopplung zu verhindern

Vor der Durchführung der Untersuchung wurden den Patienten zusätzlich der wissenschaftliche Hintergrund sowie die Konsequenzen einer Einwilligung erläutert. Bei schriftlichem Einverständnis des Patienten wurde das Bildmaterial zu den erklärten Studienzwecken verwendet.

Die durchführenden Untersucher wurden angewiesen, eine Bilddokumentation in 6 vorgegebenen Endoskoppositionen während der Untersuchung vorzunehmen. Nach der Untersuchung wurde von den durchführenden Endoskopikern ein Evaluationsbogen (siehe Abb. 15) ausgefüllt, in dem die sechs für die Auswertung ausgewählten Bilder angegeben wurden. Zudem erfolgte eine Einschätzung durch die Untersucher, ob aus ihrer Sicht eine atrophische Gastritis/intestinale Metaplasie oder EoE bzw. ösophageale Candidiasis vorlagen.

Ferner wurden für die Auswertung Basisdaten wie Alter und Geschlecht, die Indikation der Endoskopie, sowie das Setting der Untersuchung erfasst.



UNIVERSITÄTSKLINIKUM DES SAARLANDES

Klinik für Innere Medizin II - Endoskopie
Direktor: Prof. Dr. F. Lammert

Patientenetikett

Universitätsklinikum des Saarlandes
Klinik für Innere Medizin II
Endoskopie
Kirrberger Straße 100, Gebäude 41
66421 Homburg/Saar

Tel: 06841-16-15989
Fax: 06841-16-15988

Evaluation von Anwendungen der künstlichen Intelligenz in der gastroenterologischen Endoskopie

A: Arzt:

B: Bildauswahl (Sequenz angeben aus Viewpoint Liste):

- 1. Übersicht Corpus (große Kurvatur): ____
- 2. Corpus Vorderwand relativ nah: ____
- 3. Bild in Inversion nah: ____
- 4. Inversion (weiter Überblick): ____
- 5. Bild Ösophagus untere Hälfte (nicht direkt an Kardia): ____
- 6. Bild Ösophagus oberes Drittel: ____

C: Einschätzung Endoskopiker:

- Proximale Atrophie: Ja ☐
  Nein ☐
- Ösophagus: Normal ☐
  EOE ☐
  Soor ☐

Abbildung 15 - Evaluationsbogen

Bei dem selbsterstellten Evaluationsbogen für die Endoskopie wird oben ein Patientenetikett aufgeklebt, der bzw. die untersuchenden Ärzte notiert sowie die entsprechenden Bilder ausgewählt und zugeordnet. In dem unteren Abschnitt gibt der Arzt seine Einschätzung zum Magen und dem Ösophagus ab.

3.2 Diagnostische Verfahren

3.2.1 Endoskopische Untersuchung

Bei der Ösophagogastroduodenoskopie (ÖGD) wird die Schleimhaut des Gastrointestinaltraktes mithilfe eines Videoendoskopes beurteilt. Dabei werden Bilder zur Dokumentation gemacht und eventuell Biopsien zur histologischen Untersuchung entnommen.

Die Untersuchungen im Rahmen dieser Studie wurden von elf unterschiedlichen Untersuchern mit verschiedenen Generationen von Olympus-Endoskopen (HD-Endoskope; high definition) durchgeführt (GIF-H180, GIF-H190 und GIF-1TH190). Für die Auswertung wurde nur Weißlicht-Bildmaterial ohne Verwendung virtueller Chromoendoskopie genutzt [29,30]. Anschließend wurden die vom Endoskopiker ausgewählten Bilder aus dem Endoskopie-Dokumentationssystem (Viewpoint) exportiert und anonymisiert als JP(E)G-Dateien (Joint Photographic (Experts) Group) zum Hochladen in die beiden Apps verwendet [29,30]. Somit war das Vorgehen insgesamt identisch zu der bereits von Guimarães P. et al. 2020 und 2022 publizierten Validierung der Algorithmen in einem selektierten Kollektiv [29,30].

3.2.2 Histologische Untersuchung

Die eingesendeten Biopsien, aus Magen und Speiseröhre wurden in HE-Färbung untersucht. Zusätzlich wurde teilweise ein Urease-Schnelltest zur Untersuchung auf HP durchgeführt.

3.2.3 Ethik

Die Studie wurde von der Ethikkommission der Ärztekammer des Saarlandes (Saarbrücken, Deutschland; #45/20) genehmigt.

3.3 Datengewinnung und statistische Methoden

3.3.1 Evaluationsbogen

Um vergleichbares Bildmaterial durch die unterschiedlichen Untersucher zu erhalten, wurden diese vor Beginn der Studie schriftlich instruiert, in welchen Gerätepositionen Bilder zu dokumentieren sind. Um die Bildauswahl der Endoskopiker und deren Einschätzung zum Vorliegen einer präkanzerösen Bedingung des Magens oder einer EoE für die spätere Auswertung nachvollziehbar dokumentieren zu können, wurde ein Dokumentationsbogen entworfen und unmittelbar nach der Untersuchung vom Untersucher ausgefüllt (siehe Abb. 15). Auf diesem sollte optimalerweise ein Patientenetikett oder wahlweise dessen Name und Geburtsdatum zur Zuordnung, sowie der untersuchende Arzt, hinterlegt werden. Gefordert wurden sechs verschiedene Lokalisationen für Bildaufnahmen. Die unterschiedlichen Lokalisationen (Loc) sind:

1. Übersichtsaufnahme des Corpus mit großer Kurvatur
2. Nahaufnahme der Vorderwand des Corpus
3. Nahaufnahme des Fundus in Inversion
4. Übersichtsaufnahme des Fundus in Inversion
5. Untere Hälfte des Ösophagus, mit Abstand zur Kardia
6. Oberes Drittel des Ösophagus

Dies bedeutet also, dass für die Diagnostik der proximalen Atrophie vier bzw. für die EoE zwei verschiedene Bilder für die jeweilige App als Ziel angesetzt wurden.

Im letzten Abschnitt des Evaluationsbogens (siehe Abb.15) sollte der Endoskopiker noch seine persönliche Einschätzung abgeben, ob eine proximale Atrophie vorliegt und wie das Erscheinungsbild des Ösophagus wirkte (Normal vs. EoE vs. Soor).

Die Abbildung 16 zeigt gesunde bzw. Abbildung 17 zeigt auffällige Beispielbilder (atrophische Gastritis in A-D; EoE in E und F) für die jeweiligen Lokalisationen.

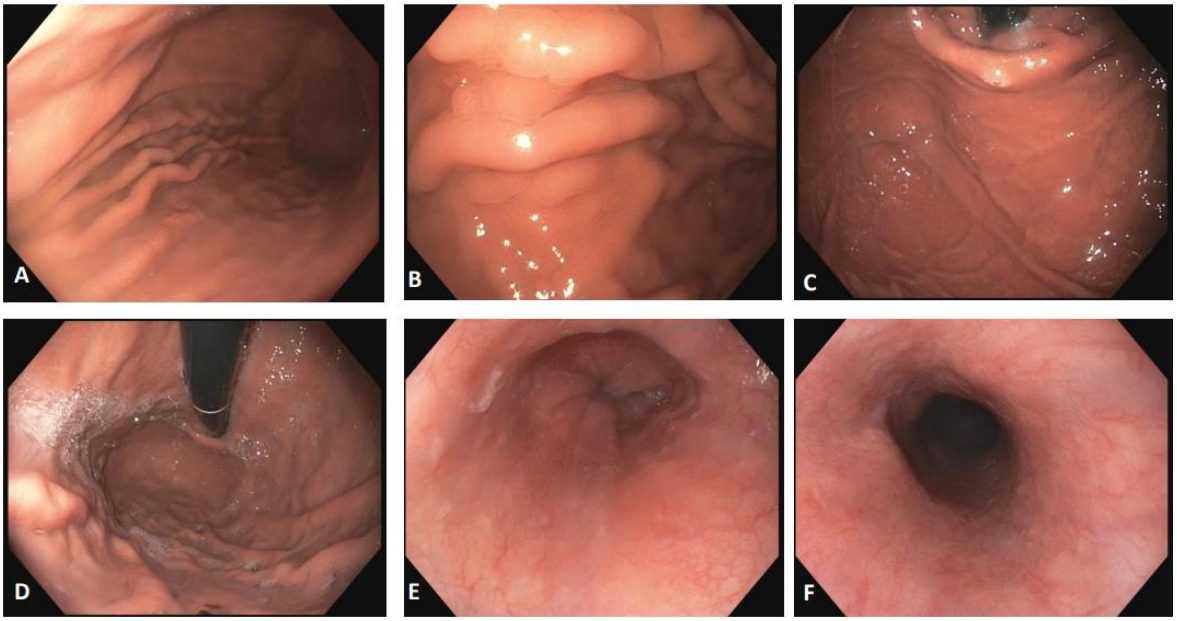


Abbildung 16 - Gesunde Beispielbilder

A: Lokalisation 1 (Übersicht Corpus). **B:** Lokalisation 2 (Nahaufnahme Corpus). **C:** Lokalisation 3 (Nahaufnahme Inversion). **D:** Lokalisation 4 (Übersicht Inversion). **E:** Lokalisation 5 (untere Hälfte Ösophagus). **F:** Lokalisation 6 (oberes Drittel Ösophagus).

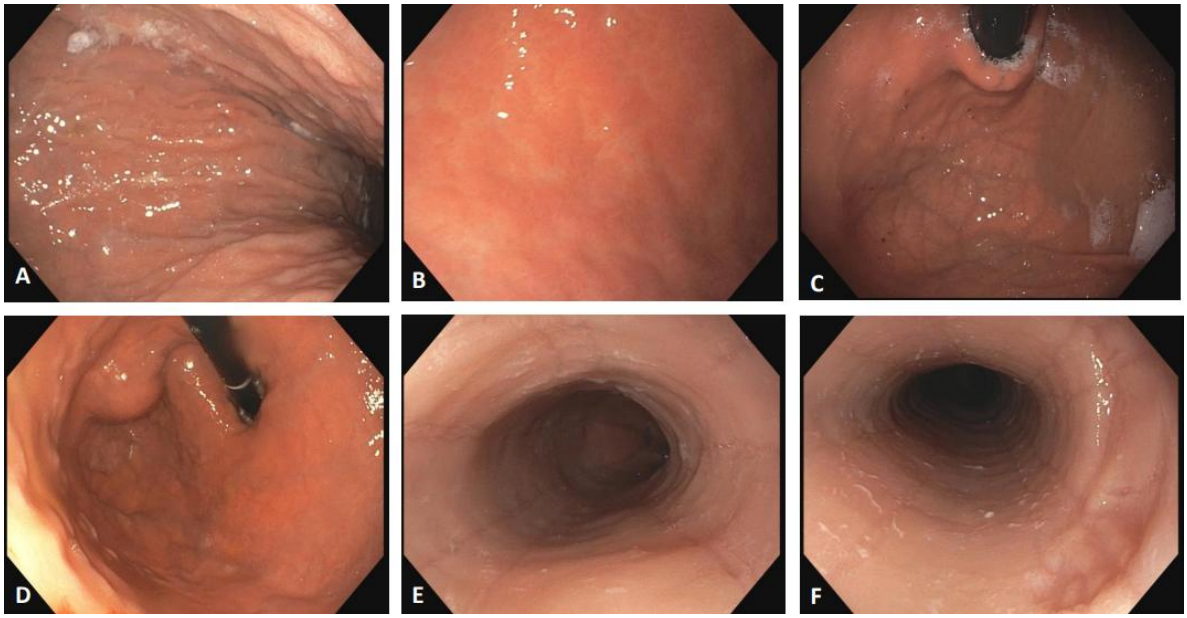


Abbildung 17 - Auffällige Beispielbilder

A: Lokalisation 1 (Übersicht Corpus). **B:** Lokalisation 2 (Nahaufnahme Corpus). **C:** Lokalisation 3 (Nahaufnahme Inversion). **D:** Lokalisation 4 (Übersicht Inversion). **E:** Lokalisation 5 (untere Hälfte Ösophagus). **F:** Lokalisation 6 (oberes Drittel Ösophagus).

3.3.2 **Eingesetzte KI-Algorithmen zur Bildanalyse**

Die beiden im Rahmen der Arbeit zur Bildanalyse eingesetzten KI-Algorithmen basieren auf dem CNN-Prinzip des *Deep Learnings*.

Die Atrophie-App wurde ursprünglich mit insgesamt 200 Bildern trainiert, jeweils 100 ohne Erkrankung und 100 mit proximaler AG [29].

Die EoE-App hingegen wurde mit insgesamt 406 Bildern trainiert, wovon 135 ohne Erkrankung, 107 mit Candidiasis und 164 mit EoE vorlagen [30].

Die Validierung der Algorithmen erfolgte in unabhängigen Validierungskohorten aus erkrankten und gesunden Personen. Für den Atrophie-Algorithmus lagen hierfür 70 neue Bilder von 35 Patienten vor, wovon 30 Bilder von 13 Patienten eine Atrophie zeigten und 40 Bilder von 22 Patienten ohne Atrophie vorlagen. Dabei erreichte der Algorithmus eine Accuracy von 93% bei einer Sensitivität von 100% und einer Spezifität von 87,5% [29]. Für den EoE-Algorithmus hingegen wurden 78 neue Bilder von 31 Patienten ausgesucht, worunter 26 Bilder von 7 Patienten mit nachgewiesener EoE, 25 Bilder von 16 Patienten mit ösophagealer Candidiasis und 27 Bilder von 8 Patienten mit Normalbefund stammten. Dabei erreichte dieser Algorithmus eine Accuracy von 91,5% bei einer Sensitivität von 87% und einer Spezifität von 94% [30].

3.3.3 **Statistik**

Die Erfassung der Daten sowie die statistische Analyse (Alter, Geschlechterverteilung, Setting, Anzahl der Bilder pro Lokalisation und je App) erfolgte mit Hilfe von Microsoft Excel. Die anschließende statistische Auswertung und Berechnung erfolgte mittels IBM ® SPSS ® Statistics Version 24.0.0.0 (© Copyright IBM Corporation and its licencors 1989, 2016) für Windows. Tabellen und Grafiken wurden mit Microsoft Word 2010 und SPSS erstellt.

Bei der deskriptiven Statistik wurden absolute und relative Häufigkeiten angegeben und entsprechende Diagramme wie Kreis-, Balken und Venn-Diagramme eingefügt.

4 Ergebnisse

4.1 Patientencharakteristika

Im Zeitraum Februar 2021 bis April 2022 wurden 650 Patienten in der Klinik der Inneren Medizin II am Universitätsklinikum des Saarlandes endoskopiert, deren Einverständnis zur Nutzung der Endoskopie-Bilder vorlag.

Von diesen Patienten wurden 578 (88,9%) ambulant und 72 (11,1%) stationär untersucht (siehe Abb. 18).

Tabelle 3 und Abb. 20 zeigen zusätzlich die Verteilung der ambulanten und stationären Patienten in Altersgruppen aufgeteilt.

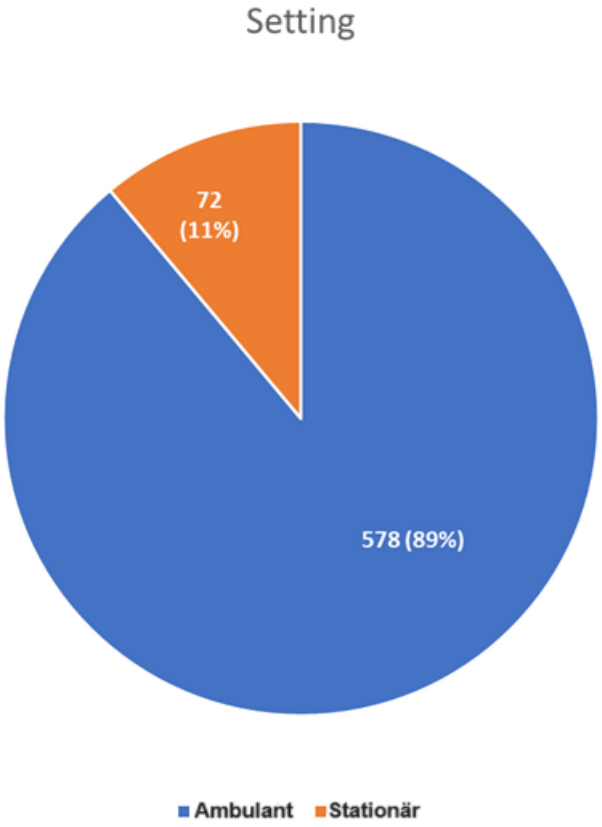


Abbildung 18 – Verteilung der Patienten anhand ihrer Art des Aufenthaltes

Es zeigt sich in orange der Anteil der stationären und in blau der ambulanten Patienten von insgesamt 650 eingeschlossenen Patienten.

GESCHLECHTS- UND ALTERSVERTEILUNG

Das Patientenkollektiv bestand aus 353 (54,3%) Frauen und 297 (45,7%) Männern (siehe Abb. 19). Die Altersspanne zum Zeitpunkt der Untersuchung erstreckt sich von 18 bis 97 Jahren mit einem mittleren Alter von 54,4 Jahren.

Tabelle 3 sowie Abb. 20 zeigen die Altersverteilung in Altersgruppen von 10 Jahren. Dabei beziehen sich die Prozentangaben auf die Anzahl der Patienten im entsprechenden Setting, also auf 578 ambulante sowie 72 stationäre Patienten.

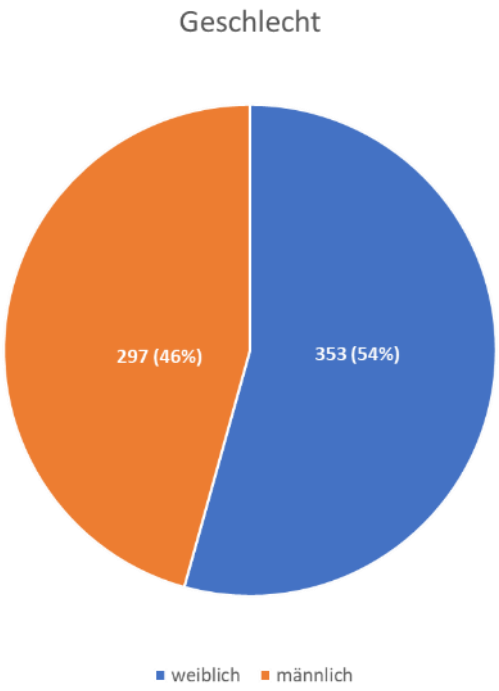


Abbildung 19 - Geschlechterverteilung

Es zeigt sich in blau der Anteil an weiblichen und in orange der Anteil der männlichen Patienten von insgesamt 650 eingeschlossenen Patienten.

Tabelle 3 - Verteilung der Patienten anhand ihrer Art des Aufenthaltes in Altersgruppen

Die Prozentangaben beziehen sich auf die jeweils untersuchte Kohorte, also auf die jeweils stationäre und ambulante Anzahl an Patienten.

	Ambulant		Stationär	
Altersgruppe	n	%	n	%
18-20 Jahre	11	1,9%	0	0%
21-30 Jahre	42	7,3%	1	1,4%
31-40 Jahre	75	13%	5	6,9%
41-50 Jahre	96	16,6%	8	11,1%
51-60 Jahre	158	27,3%	13	18,1%
61-70 Jahre	114	19,7%	18	25%
71-80 Jahre	65	11,2%	19	26,4%
81-90 Jahre	16	2,8%	6	8,3%
91-100 Jahre	1	0,2%	2	2,8%
Gesamt	578	100%	72	100%

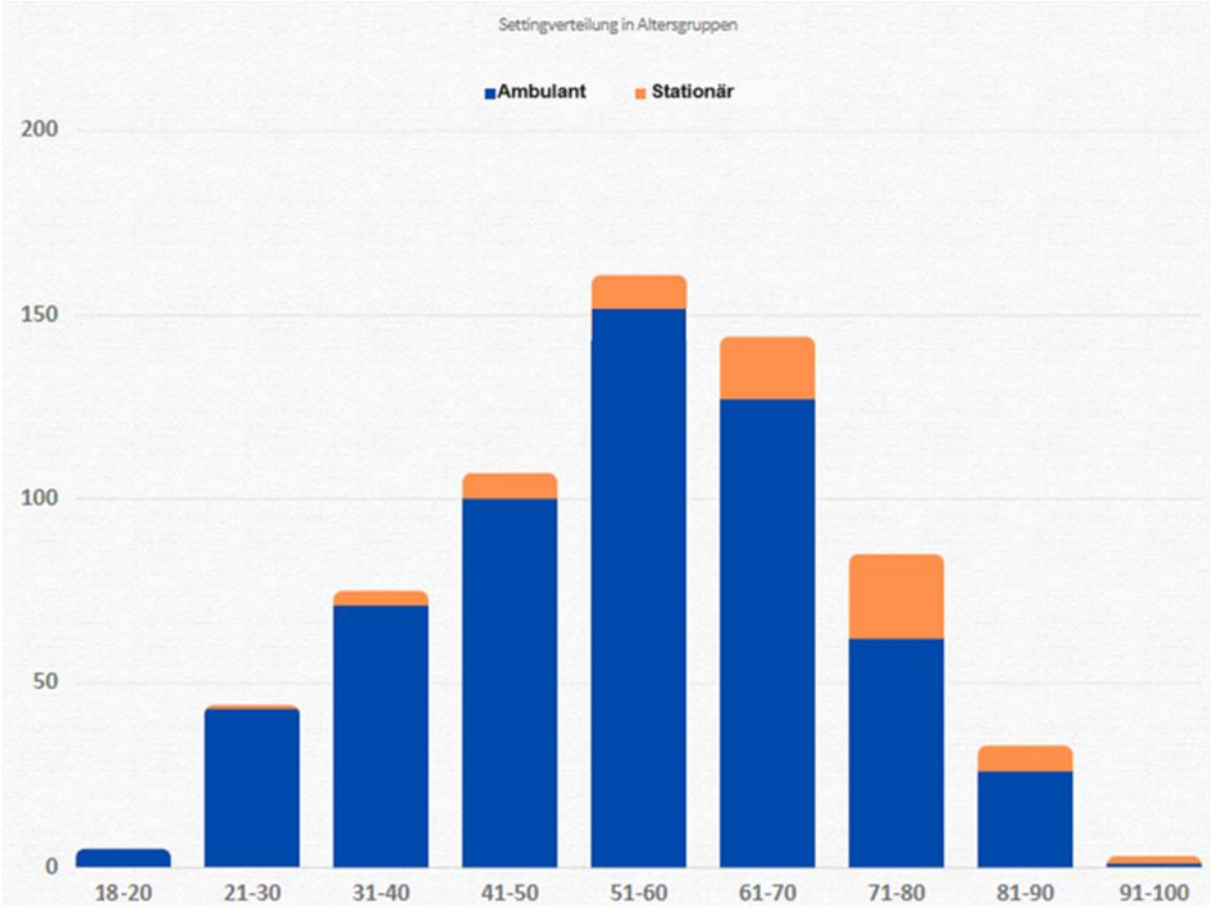


Abbildung 20 - Verteilung der Patienten anhand ihrer Art des Aufenthaltes in Altersgruppen

Es zeigt sich in verschiedenen Altersgruppen in blau der Anteil der ambulanten Patienten und in orange der Anteil an stationären Patienten von insgesamt 650 eingeschlossenen Patienten.

Indikation zur Durchführung einer ÖGD

Bei den 650 untersuchten Patienten lagen unterschiedliche Beweggründe für die Durchführung einer ÖGD vor (siehe Tabelle 4). So lagen beim Großteil der Patienten (182, 28%) abdominelle Beschwerden im Sinne von Oberbauchschmerzen vor. Ein weiterer Teil (123, 19%) wurde endoskopiert um beispielweise andere Interventionen wie eine PEG-Anlage (perkutane endoskopische Gastrostomie), Blutstillung oder eine HP-Diagnostik durchzuführen. Weiterhin gaben 88 Patienten (13,5%) Refluxbeschwerden und 24 Patienten (3,7%) Dysphagie als Symptomatik an. Bei 67 Patienten (10,3%) war eine anstehende Operation und bei 58 Patienten (8,9%) eine Tumorsuche bzw. Umfelddiagnostik die Indikation. Weitere Gründe zur Endoskopie waren bei 34 Patienten (5,2%) eine bekannte Leberzirrhose oder Hepatopathie bzw. bereits vorhandene Varizen, sowie bei 21 Patienten (3,2%) V.a. Blutungen/Anämien und bei

18 Patienten (2,8%) eine chronisch-entzündliche Darmerkrankung. Bei 35 Patienten (5,4%) wurde keine Angabe über die Indikation der Untersuchung gemacht.

Tabelle 4 - Indikationen zur Durchführung einer Endoskopie

Unter „Sonstiges“ wurden Indikationen für eine ÖGD wie andere Interventionen, beispielsweise eine PEG-Anlage, Blutstillung oder eine HP-Diagnostik zusammengefasst.

Endoskopie-Indikation	**Anzahl %** n = 650 (100%)
Abdominelle Beschwerden (Schmerzen)	182 (28%)
Sonstiges	123 (19%)
Refluxbeschwerden	88 (13,5%)
Status vor OP	67 (10,3%)
Tumorsuche und Umfelddiagnostik	58 (8,9%)
Keine Angabe	35 (5,4%)
Leberzirrhose, Hepatopathie, Varizen	34 (5,2%)
Dysphagie	24 (3,7%)
V.a. Blutungen, Anämie	21 (3,2%)
Chronisch-entzündliche Darmerkrankung	18 (2,8%)

4.2 Charakteristika der Endoskopie-Bilder

Gesamtzahl der vorhandenen Bilder

Im Verlauf der Datensammlung ergab sich, dass nicht bei jedem Patienten vier Bilder für die Atrophie-App bzw. zwei Bilder für die EoE-App zur Verfügung standen. Entweder wurde kein Bild der geforderten Lokalisation dokumentiert, oder es erfolgte eine falsche Bildzuordnung. Alle falsch zugeordneten Bilder (Bild entspricht nicht der geforderten Position z.B. Bild aus dem Duodenum oder Antrum) wurden nicht in die Auswertung aufgenommen. In Tabelle 5 wurde aufgeschlüsselt, wie viele Bilder pro Lokalisation zur Verfügung standen. Dabei definieren Loc 1 bis 4 die Bilder für die Atrophie-App und Loc 5 bis 6 für die EoE-App.

Tabelle 5 - Anzahl der Bilder nach Lokalisation (vor Entfernung von Bildern)

Lokalisation	Anzahl der Bilder (%)
Loc 1	415 (15,1%)
Loc 2	358 (13%)
Loc 3	487 (17,8%)
Loc 4	546 (19,9%)
Loc 5	438 (16%)
Loc 6	498 (18,2%)
Anzahl der potenziellen Bilder insgesamt	2.742 (100%)

Daraus resultiert, dass insgesamt 2.742 Bilder ausgewertet hätten werden können, davon 1.806 aus dem Magen und 936 Bilder aus dem Ösophagus. Die meisten Bilder standen für die Loc 4 (546; 19,9%) der Atrophie-App zur Verfügung. Am wenigsten Bilder sind für die Loc 2 (358; 13%) der Atrophie-App vorhanden gewesen. Für die EoE-App waren für beide Lokalisationen eine ähnlich hohe Anzahl an Bildern verfügbar. Schlussendlich wurden jedoch weniger Bilder ausgewertet, dies wird in einem späteren Kapitel dieser Arbeit näher erläutert.

Verteilung der Bilder pro Patient, App und Lokalisation

Interessant war anschließend die Fragestellung, wie viele Bilder pro Patienten für die endgültige Auswertung zur Verfügung standen. Dabei wurden nur die korrekt zugeordneten und somit auswertbaren Bilder berücksichtigt. Das nachfolgende Unterkapitel verdeutlicht, wie viele Bilder insgesamt gefehlt bzw. nicht auswertbar waren und somit nicht in den folgenden Tabellen enthalten sind. Tabelle 6 verdeutlicht, dass nur für circa ein Drittel der Patienten alle sechs geforderten Bilder für die Auswertung mithilfe der beiden Algorithmen zur Verfügung stand (225; 34,6%). Für fast ein Viertel der Patienten waren hingegen nur ein oder zwei der geforderten Bilder verfügbar (insgesamt 140; 21,5%).

Tabelle 6 - Anzahl der Bilder pro Patient

Anzahl der Bilder	**Anzahl der Patienten (%)**
1 von 6 Bildern	61 (9,4%)
2 von 6 Bildern	79 (12,2%)
3 von 6 Bildern	67 (10,3%)
4 von 6 Bildern	101 (15,5%)
5 von 6 Bildern	117 (18%)
6 von 6 Bildern	225 (34,6%)
Insgesamte Anzahl	650 (100%)

Diese Angaben wurden dann in Tabelle 7 und 8 nochmals explizit für die jeweilige App aufgeschlüsselt, um darzustellen, wie viele Bilder eines Patienten für die beiden Algorithmen vorhanden waren. Dabei zeigt sich, dass über ein Drittel der Patienten alle benötigten Bilder für die Atrophie-App (247; 38%) bzw. sogar fast zwei Drittel für die EoE-App (386; 59,4%) hatten.

Tabelle 7 - Anzahl der auswertbaren Bilder pro Patient für die Atrophie-App

Vorhandene Bilder	**Anzahl (%)**
Keine Bilder	14 (2,2%)
1 von 4 Bildern	109 (16,8%)
2 von 4 Bildern	131 (20,1%)
3 von 4 Bildern	149 (22,9%)
4 von 4 Bildern	247 (38%)
Insgesamte Anzahl	650 (100%)

Es zeigt sich also, dass für den Atrophie-Algorithmus nur Bilder von 636 von 650 Patienten (97,9%) zur Verfügung standen (mindestens 1 Bild), da bei 14 von 650 Patienten (2,2%) keine Bilder zu entsprechenden Lokalisationen vorlagen.

Tabelle 8 - Anzahl der auswertbaren Bilder pro Patient für die EoE-App

Vorhandene Bilder	**Anzahl (%)**
Keine Bilder	100 (15,4%)
1 von 2 Bildern	164 (25,2%)
2 von 2 Bildern	386 (59,4%)
Insgesamte Anzahl	650 (100%)

Auch für die EoE-App zeigt sich somit, dass bei 100 von 650 Patienten (15,4%) keine Bilder zu entsprechenden Lokalisationen vorlagen und somit nur 550 Patienten (84,6%) zur Auswertung (mindestens 1 Lokalisation) verfügbar waren.

Fehlende und nicht auswertbare Bilder

Durch die Tabellen 7 und 8 wird also deutlich, dass einige Bilder für die einzelnen Lokalisationen gefehlt haben. Insgesamt fehlten 992 Bilder, davon 672 (67,7%) für die Atrophie und 320 (32,3%) für die EoE. Die meisten Bilder haben für Loc 2 der Atrophie-App (236; 23,8%) bzw. Loc 5 der EoE-App (182; 18,4%) gefehlt (siehe Abb. 21)

Weiterhin konnten von den insgesamt 2.742 Bildern 126 Bilder für die Atrophie-App und 28 Bilder für die EoE-App aufgrund falscher Lokalisationszuordnung (z.B. Bild aus dem Duodenum oder dem Magenantrum oder doppelte Bilder aus gleicher Magenlokalisation) bzw. nicht repräsentativer Areale (z.B. Barrett-Schleimhaut im distalen Ösophagus; Loc 5) nicht zur Auswertung genutzt werden (siehe Abb. 21). Sie wurden dementsprechend nicht mithilfe der Algorithmen ausgewertet, um eine Verfälschung der Ergebnisse durch unpassende Bilder zu verhindern. Tabelle 9 zeigt die Verteilung der aus der Auswertung entfernten Bilder. Dabei beziehen sich die Prozentangaben in Tabelle 9 auf die Anzahl der vorhandenen Bilder je Lokalisation aus Tabelle 5.

Tabelle 9 - Anzahl der entfernten Bilder bei falscher Lokalisation

Die Prozentangabe der entfernten Bilder bezieht sich auf die insgesamt vorhandene Anzahl je Lokalisation aus Tabelle 5. Die entfernten Bilder bei vorliegender Candidiasis im Ösophagus (Loc 5 und 6) sind hier nicht mit einbezogen worden.

Lokalisation	Anzahl der entfernten Bilder (%)
Loc 1	47 von 415 (11,3%)
Loc 2	57 von 358 (15,9%)
Loc 3	6 von 487 (1,2%)
Loc 4	16 von 546 (2,9%)
Loc 5	24 von 438 (5,5%)
Loc 6	4 von 498 (0,8%)
Anzahl der entfernten Bilder insgesamt	154 von 2.742 (5,6%)

Es wurden also insgesamt 154 von 2.742 Bildern (5,6%) entfernt und nicht ausgewertet. Dabei zeigt sich, dass für Loc 2 (57 von 358; 15,9%) für die Atrophie-App am häufigsten ein falsch lokalisiertes Bild von den Untersuchern im Evaluationsbogen angegeben war. Am wenigsten falsch lokalisierte Bilder gab es für Loc 6 (4 von 498; 0,8%) für die EoE-App.

Weiterhin wurden für die EoE-App für Loc 5 7 von 438 Bildern (1,6%) und für Loc 6 11 von 498 Bildern (2,2%) aus der Auswertung genommen, da diese eine Candidiasis zeigen (siehe Abb. 21). Zwar hat der Algorithmus die Möglichkeit die Pilzinfektion zu erkennen, jedoch liegt durch die geringen Fallzahlen der Fokus der Arbeit auf der Erkennung der EoE.

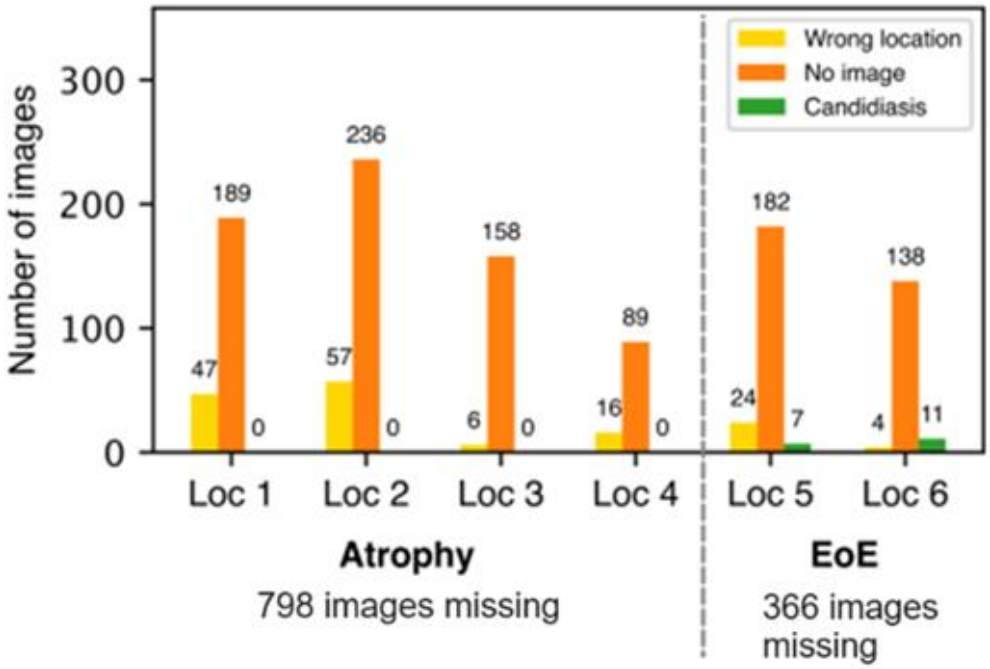


Abbildung 21 - Anzahl und Verteilung der fehlenden Bilder

Es zeigen sich für die einzelnen Lokalisationen die Anzahl an falsch lokalisierten und dadurch entfernten Bildern (gelb), die nicht vorhandenen Bilder (orange) sowie durch Candidiasis-Befall entfernte Bilder (grün). Unter dem Bild zeigt sich die Gesamtzahl an fehlenden Bildern für die Atrophie-App (Loc 1 bis Loc 4) sowie für die EoE-App (Loc 5 und Loc 6)

4.3 Biopsien

Nicht bei allen 650 Patienten wurden Biopsien des Magens und des Ösophagus durchgeführt.

Es wurden von 583 (91,7%) von 636 Patienten mit verfügbarem Bildmaterial aus dem Magen bzw. 204 (37,1%) von 550 Patienten mit Bildmaterial aus dem Ösophagus Biopsien entnommen. Da das histopathologische Ergebnis die Grundwahrheit für die Algorithmen bildet, wurde ein positives Biopsie-Ergebnis als eine positive Fallzahl und umgekehrt gewertet. Somit zeigen sich 22 (3,8%) positive und 561 (96,2%) negative Ergebnisse für die präkanzerösen Bedingungen des Magens, sowie 7 (3,4%) positive und 197 (96,6%) negative Ergebnisse für die EoE. Die wesentlich niedrigere Anzahl der Biopsien für die EoE lässt sich v.a. dadurch erklären, dass im Ösophagus meist nur welche entnommen werden, sofern Auffälligkeiten sichtbar sind oder eine ösophageale Dysfunktion (typische Beschwerden) vorlag. Abbildung 22 zeigt die Biopsien und deren Ergebnisse auf die Lokalisationen der beiden Erkrankungen verteilt.

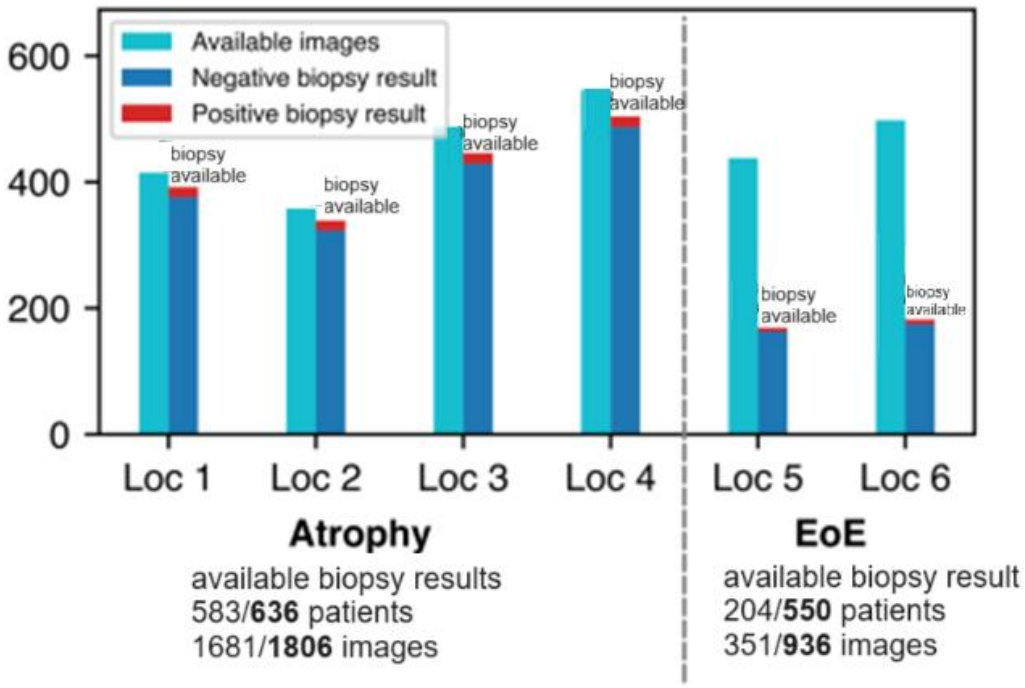


Abbildung 22 - Anzahl der Biopsien pro Lokalisation

In hellblau ist pro Lokalisation gekennzeichnet, wie viele Bilder zur Verfügung stehen. Dunkelblau kennzeichnet die Anzahl an negativen und rot die Anzahl an positiven Biopsien pro Lokalisation. Unterhalb des Balkendiagramms ist somit bei Berücksichtigung nur der Patienten mit vorhandenen Biopsien die Anzahl an Patienten und Bilder für die Atrophie und die EoE angegeben.

4.4 Ergebnisse zu präkanzerösen Bedingungen des proximalen Magens

Um aussagekräftige Ergebnisse zur AG zu erhalten, wurden nur Patienten berücksichtigt, welchen als Goldstandard zur Diagnosesicherung eine Biopsie entnommen wurde. Somit beziehen sich alle Werte auf 583 (91,7%) von 636 Patienten mit verbleibenden 1.681 (93,1%) von 1.806 Bildern. Dies bedeutet gleichzeitig, dass die Bildanzahl pro Lokalisation sich für die endgültige Auswertung der KI verringert und somit für Loc 1 392, für Loc 2 339, für Loc 3 446 und für Loc 4 504 Bilder berücksichtigt wurden. Außerdem muss beachtet werden, dass die Untersucher ihr Ergebnis pro Patient insgesamt angeben im Gegensatz zur KI, welche einzelne Bilder bewertet hat.

4.4.1 Ergebnisse der Untersucher auf Patientenebene

Insgesamt wurden von den 583 Patienten während der Untersuchung durch die Endoskopiker 25 (4,3%) als positiv und 558 (95,7%) als negativ eingeschätzt. Bei genauerer Betrachtung wird deutlich, dass von den laut Biopsie 22 (3,8%) positiven Patienten 9 (40,9%) richtig-positiv und 13 (59,1%) falsch-negativ eingeschätzt wurden. Insgesamt wurden somit also bei 561 (96,2%) negativen Patienten 16 (2,9%) falsch-positive und 545 (97,1%) richtig-negative Einschätzungen abgegeben (siehe Tabelle 10).

Tabelle 10 - Einschätzungen der Untersucher zur proximalen Atrophie

Die Prozentangaben der positiven und negativen Patienten sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl von 583 Patienten. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der 22 positiven Patienten. Die richtig-negativen und falsch-positiven Prozentangaben beziehen sich auf die Anzahl der 561 negativen Patienten.

	Einschätzung der Untersucher
Positive Patienten	22 (3,8%)
Negative Patienten	561 (96,2%)
Positive Einschätzung (PP)	25 (4,3%)
Negative Einschätzung (PN)	558 (95,7%)
Richtig-positiv (RP)	9 (40,9%)
Richtig-negativ (RN)	545 (97,1%)
Falsch-positiv (FP)	16 (2,9%)
Falsch-negativ (FN)	13 (59,1%)

Statistische Metriken

Bei 583 Untersuchungen mit 22 (3,8%) positiven und 561 (96,2%) negativen Biopsien ergibt sich eine *Prävalenzrate* von 4% (0,04). Daraus ergibt sich also für eine statistische Auswertung das Problem einer unbalancierten Verteilung.

Die *Sensitivität*, also die Wahrscheinlichkeit mit welcher ein positiver Patient als positiv eingeschätzt wurde, ergab 41% für die Untersucher. Die *Spezifität* hingegen, also die Wahrscheinlichkeit für eine negative Einschätzung bei tatsächlich auch gesundem Patienten, ergibt in diesem Fall 97%. Daraus ergibt sich die *Precision* (auch *positiver prädiktiver Wert*, *PPV* genannt), also die Wahrscheinlichkeit, dass bei einem als erkrankt eingeschätztem Patienten tatsächlich die Erkrankung vorliegt. Sie liegt bei 36%.

Der *F1-Score* bildet einen gewichteten Mittelwert von *Precision* und *Sensitivität*. Dadurch erhalten beide Werte eine ähnlich starke Gewichtung und somit eignet er sich bei solch unausgewogenen Gruppen besser und ergibt für die Untersucher 0,38.
Der F1-Score berechnet sich wie folgt:

$$\frac{2 \times RP}{2 \times RP + FP + FN}$$

Die diagnostische Genauigkeit zur Veranschaulichung des Anteils der korrekten Vorhersagen, auch *Accuracy* (ACC) genannt, berechnet sich auf 95% für die Untersucher. Durch die unbalancierten Kohorten mit wenigen positiven und vielen negativen Fällen ist dieser Wert jedoch nicht aussagekräftig. Um das Ergebnis der *ACC* realistischer auszudrücken, nutzen wir die *balanced Accuracy* (*bACC*), bei der die Summe von Sensitivität und Spezifität durch zwei geteilt wird und somit künstlich eine 50/50-Kohorte erzeugt wird. Sie eignet sich bei ungleich großen Kohortengruppen und ergibt im Fall der Untersucher 69%. Somit ist diese deutlich niedriger als die *ACC* ohne Angleichung.

Zusätzlich lässt sich als Alternative zu Sensitivität und Spezifität die *Likelihood-Quotienten/Ratio (LR)* berechnen, welche es ermöglicht, abzuschätzen, wie hoch die Wahrscheinlichkeit anhand des Testergebnisses ist, ob die Erkrankung vorliegt oder nicht. Sie beschreibt mittels Faktor wie viel häufiger das Ergebnis bei Kranken im Gegensatz zu Gesunden vorkommt und kann somit als positive (*LR+*) oder negative (*LR-*) Ratio angegeben werden. Der Vorteil ist, dass dieser Wert sich unabhängig von der Krankheitsprävalenz berechnen lässt, wodurch sie bei unbalancierten Gruppengrößen besser funktionieren als andere Metriken.

Tabelle 11 zeigt die Reichweite und Bedeutung der Werte. Dabei würde ein Wert von genau 1 bedeuten, dass die Wahrscheinlichkeit ein positives Testergebnis unabhängig des Gesundheitszustandes zu erhalten gleich groß ist und somit der Test kaum diagnostische Bedeutung besitzt. Umso höher der Wert für *LR+* bzw. umso niedriger für *LR-*, desto besser ist der Test für die entsprechende Interpretation geeignet. Dies bedeutet also umso höher *LR+* ist, desto besser erkennt der Test erkrankte Patienten. Umso höher *LR-* ist, desto besser erkennt der Test gesunde Patienten.

Tabelle 11 - Interpretation der Wahrscheinlichkeit mittels Likelihood-Ratio

Quelle: Wahrscheinlichkeitsverhältnis (Likelihood Ratio) – Alternative zu Sensitivität und Spezifität von Schwarzer G. et al, 2002 [83]

LR+	**LR-**	**Bedeutung**
> 10	< 0,1	Überzeugende diagnostische Evidenz
5 – 10	0,1 – 0,2	Hohe diagnostische Evidenz
2 – 5	0,2 – 0,5	Schwache diagnostische Evidenz
1 - 2	0,5 – 1	Kaum relevante diagnostische Evidenz

LR+ gibt den Faktor der Wahrscheinlichkeit an, mit dem der Patient bei positivem Test krank ist. Er berechnet sich aus dem Quotienten von Sensitivität durch 1 - Spezifität (oder anders formuliert aus dem Quotienten der Richtig-positiven durch die Falsch-positiven) und ergibt für die Untersucher 14,34. Somit zeigen die Aussagen der Untersucher mit hoher diagnostischer Evidenz eine hohe Wahrscheinlichkeit, dass der Patient bei positivem Test tatsächlich erkrankt ist.

LR- hingegen gibt den Faktor der Wahrscheinlichkeit an, mit dem ein negativ getesteter Patient gesund ist. Er berechnet sich aus dem Quotienten von 1 - Sensitivität durch Spezifität und ergibt hier 0,61. Dies bedeutet, dass die Aussage der Untersucher zum Erkennen von gesunden Patienten eine wesentlich geringere diagnostische Bedeutung hat. Sie ist also weniger geeignet, um tatsächlich gesunde Patienten zu erkennen.

Das *Cohens Kappa* (*Kappa*, K) gibt vereinfacht erklärt an, wie oft das Ergebnis der Untersucher mit dem Ergebnis der Biopsien übereinstimmt und setzt dies in ein Verhältnis. Es kann angeben, wie zuverlässig beide Methoden das gleiche Ergebnis messen. Es beträgt in diesem Fall 0,36. Dabei gibt ein Wert näher der 1 eine hohe Übereinstimmung an bzw. bedeutet näher 0, dass es eher eine zufällige Übereinstimmung ist. Bei unbalancierten Gruppengrößen ist dieser Wert etwas zuverlässiger als andere Metriken.

Der letzte berechnete Wert ist die *Area under the curve* (*AUC*) der ROC-Kurve und beträgt für die Untersucher 0,69. Diese kann berechnet oder, wie in diesem Fall, abgeschätzt werden. Die Abschätzung erfolgt, wenn kein Zugriff auf die ROC-Kurven zur genauen Berechnung vorliegt. Dies liegt daran, dass die Untersucher ähnlich wie die Algorithmen lediglich „Ja“ oder „Nein“-Aussagen erstellen und somit keine Prozentangaben zur Wahrscheinlichkeit zur Verfügung stehen. Sie nimmt Werte zwischen 0,5 und 1 an und gibt die Leistung des Tests an, wobei ein Wert von 0,5 eher eine zufälliges Ergebnis angibt und somit eine schlechte Testleistung. Tabelle 12 gibt einen Überblick über die einzelnen besprochenen Metriken.

Tabelle 12 - Statistische Metriken der Einschätzungen der Untersucher zu den präkanzerösen Bedingungen des proximalen Magens

	Einschätzung Endoskopiker
Gesamtzahl	583
Prävalenzrate	0,04
Accuracy (ACC)	0,95
Balanced Accuracy (bACC)	0,69
Precision (PPV)	0,36
Sensitivität	0,41
Spezifität	0,97
F1-Score	0,38
Area under the curve (AUC)	0,69
Kappa	0,36
LR +	14,34
LR -	0,61

4.4.2 **Ergebnisse des Atrophie-Algorithmus auf Bildebene**

Die Ergebnisse der KI ergeben sich aus den verschiedenen Lokalisationen mit unterschiedlicher Anzahl an vorhandenen Bildern, so dass eine individuelle Anzahl an vorhandenen positiven Biopsien jeweils angegeben wird. Es handelt sich also um Auswertungen auf Ebene einer Einzellokalisation und nicht um eine Auswertung auf Patientenebene. Tabelle 13 zeigt die einzelnen Werte.

Für Loc 1 sind von der Atrophie-App von insgesamt 392 Bildern 46 (11,7%) als positiv und 346 (88,3%) als negativ ausgewertet worden. Somit ergeben sich unter Annahme des histopathologischen Ergebnisses als Grundwahrheit bei 16 (4,1%) positiven Bildern 13 (81,3%) richtig-positive (RP) und 3 (18,7%) falsch negative (FN) bzw. bei 376 (95,9%) negativen Bildern 33 (8,8%) falsch-positive (FP) und 343 (91,2%) richtig-negative (RN) Ergebnisse.

In Loc 2 wurden von 339 Bildern 161 (47,5%) positiv und 178 (52,5%) negativ bewertet. Bei tatsächlich 16 (4,7%) positiven Bildern ergeben sich somit 15 (93,7%) RP und 1 (6,3%) FN bzw. bei 323 (95,3%) negativen Bildern 146 (45,2%) FP und 177 (54,8%) RN Auswertungen.

Bei 446 Bildern für Loc 3 sind 101 (22,7%) positiv und 345 (77,3%) negativ vorausgesagt worden. Es lagen 17 (3,8%) positive Bilder vor woraus sich entsprechend 11 (64,7%) RP und 6 (35,3%) FN Fälle ergeben. Bei 429 (96,2%) negativen Bildern ergeben sich 90 (21%) FP und 339 (79%) RN.

Loc 4 hat 504 Bilder zur Verfügung, wovon 44 (8,7%) positiv und 460 (91,3%) negativ bewertet wurden. Bei 17 (3,4%) positiven Bildern ergeben sich somit 7 (41,2%) RP und 10 (58,8%) FN bzw. bei 487 (96,6%) negativen Bildern 37 (7,6%) FP und 450 (92,4%) RN Ergebnisse.

Insgesamt wird daraus ersichtlich, dass mit weitem Abstand Loc 2 mit 45,2% (146 von 323) die höchste Rate an FP Bildern hat, dahinter folgend Loc 3 mit 21% (90 von 429). Bei Loc 1 (33; 8,8%) und Loc 4 (37; 7,6%) liegt diese Fehlerrate unter 10%. Dem gegenübergestellt zeigt jedoch Loc 4 mit 58,8% (10 von 17) die höchste Anfälligkeit für FN Befunde, gefolgt von Loc 3 mit 35,3% (6 von 17).

Tabelle 13 - Einschätzungen der Atrophie-App zu den präkanzerösen Bedingungen des proximalen Magens

Die Prozentangaben der positiven und negativen Biopsien sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl der vorhandenen Bilder pro Lokalisation. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der positiven bzw. die richtig-negativen und falsch-positiven Prozentangaben auf die Anzahl der negativen Biopsien pro Lokalisation.

	Loc 1	**Loc 2**	**Loc 3**	**Loc 4**
Bilder insgesamt	392	339	446	504
Positive Biopsien	16 (4,1%)	16 (4,7%)	17 (3,8%)	17 (3,4%)
Negative Biopsien	376 (95,9%)	323 (95,3%)	429 (96,2%)	487 (96,6%)
Positive Einschätzung (PP)	46 (11,7%)	161 (47,5%)	101 (22,7%)	44 (8,7%)
Negative Einschätzung (PN)	346 (88,3%)	178 (52,5%)	345 (77,3%)	460 (91,3%)
Richtig-positiv (RP)	13 (81,3%)	15 (93,7%)	11 (64,7%)	7 (41,2%)
Richtig-negativ (RN)	343 (91,2%)	177 (54,8%)	339 (79%)	450 (92,4%)
Falsch-positiv (FP)	33 (8,8%)	146 (45,2%)	90 (21%)	37 (7,6%)
Falsch-negativ (FN)	3 (18,7%)	1 (6,3%)	6 (35,3%)	10 (58,8%)

Statistische Metriken

Die *Prävalenzrate* beträgt sowohl für Loc 1 mit 392 Patienten, wovon 16 (4,1%) positiv und 376 (95,9%) negativ biopsiert wurden, als auch für Loc 3 mit 446 Patienten, wovon wiederum 17 (3,8%) positive und 429 (96,2%) negative Biopsien vorliegen, jeweils 4%. Loc 2 mit 339 Patienten und 16 (4,7%) positiven und 323 (95,3%) negativen Biopsien weist mit 5% eine etwas höhere Rate auf. Loc 4 hingegen mit 504 Bildern und 17 (3,4%) positiven und 487 (96,6%) negativen Fällen hat mit 3% die niedrigste Prävalenzrate.

Die *Sensitivität* ist für Loc 2 mit 94% am höchsten. Loc 4 hat die schlechteste Sensitivität mit 41%. Loc 1 mit 81% und Loc 3 mit 65% liegen dazwischen.

Die *Spezifität* hingegen zeigt sich für Loc 4 mit 92% und Loc 1 mit 91% am höchsten. Danach folgt Loc 3 mit 79% sowie Loc 2 mit 55% und ist dementsprechend am schlechtesten.

Die *Precision* (*PPV*) also die Wahrscheinlichkeit, dass bei einem als erkrankt eingeschätztem Patienten tatsächlich die Erkrankung vorliegt, ist bei allen vier Lokalisationen schlecht ausgefallen. So ist sie mit 28% für Loc 1 am höchsten und für Loc 2 mit 9% am niedrigsten. Loc 4 (16%) und Loc 3 (11%) liegen ebenfalls im eher unteren Bereich. Dies lässt sich durch die unbalancierte Kohorte erklären.

Insgesamt spiegeln sich diese bisherigen Werte auch im *F1-Score* wider. So beträgt dieser für Loc 1 0,42. Die anderen drei Lokalisationen haben einen größeren Abstand zu diesem Wert und siedeln sich noch weiter im unteren Bereich an und liegen für Loc 4 bei 0,23, für Loc 3 bei 0,19 und für Loc 2 bei 0,17.

Die *Accuracy* für Loc 1 und Loc 4 beträgt jeweils 91%. Interessant ist hierbei dann die *bACC* zur mathematischen Angleichung an eine 50/50-Kohorte, welche für Loc 1 dann zwar immer noch hoch bei 86% liegt, für Loc 4 hingegen ist diese mit 67% wesentlich schlechter. Loc 3 hingegen hat eine *ACC* von 78% und eine *bACC* von 72%, so dass die balancierte Kohorte zwar etwas schlechtere Ergebnisse erzielt, aber immer noch in einem ähnlichen Bereich liegt. Loc 2 hingegen hat in der reellen Auswertung eine *ACC* von lediglich 57%, wohingegen sie für die *bACC* mit 74% wesentlich besser abschneidet.

LR+, der Faktor für die Wahrscheinlichkeit krank zu sein bei positivem Ergebnis, ist für Loc 1 mit 9,26 am höchsten, gefolgt von Loc 4 mit 5,42 und zeigen somit eine hohe diagnostische

Aussagekraft. Loc 3 mit 3,08 und Loc 2 mit 2,07 sind niedriger im Vergleich zu den beiden anderen Lokalisationen, wodurch auch ihre Aussagekraft weniger bedeutsam ist.

LR-, der Faktor für die Wahrscheinlichkeit gesund zu sein bei negativem Ergebnis, ist für Loc 4 mit 0,64 am höchsten und hat somit eine geringe diagnostische Aussagekraft. Anschließend zeigt Loc 3 mit 0,45 das zweithöchste Ergebnis. Loc 1 mit 0,21 und Loc 2 mit 0,11 sind vergleichsweise niedrig und haben somit eine höhere Wahrscheinlichkeit, dass der Patient wirklich gesund ist bei negativem Test.

Auch *Cohens Kappa* zeigt sich für Loc 1 am höchsten mit 0,38. Für Loc 4 beträgt dieser 0,19, für Loc 3 0,13 und für Loc 2 0,09 und ist somit niedriger im Vergleich zu Loc 1.

Der letzte verbleibende Wert, die *AUC,* errechnet sich für Loc 1 auf 0,86 und ist somit am höchsten. Loc 2 mit 0,74 und Loc 3 mit 0,72 liegen sehr nahe beieinander. Der niedrigste Wert ergibt sich somit für Loc 4 mit 0,67. Auch im Falle der KI wird dieser geschätzt, da die KI keine Prozentangaben zur Wahrscheinlichkeit angibt und somit nur die Ergebnisse „Ja“ oder „Nein“ zur Verfügung stehen. Tabelle 14 gibt einen Überblick über die berechneten Werte für die vier verschiedenen Lokalisationen der proximalen Atrophie.

Tabelle 14 - Statistische Metriken der Einschätzung der Atrophie-App zur proximalen Atrophie

	Loc 1	Loc 2	Loc 3	Loc 4
Gesamtzahl Bilder	392	339	446	504
Prävalenzrate	0,04	0,05	0,04	0,03
Accuracy (ACC)	0,91	0,57	0,78	0,91
Balanced Accuracy (bACC)	0,86	0,74	0,72	0,67
Precision (PPV)	0,28	0,09	0,11	0,16
Sensitivität	0,81	0,94	0,65	0,41
Spezifität	0,91	0,55	0,79	0,92
F1-Score	0,42	0,17	0,19	0,23
Area under the curve (AUC)	0,86	0,74	0,72	0,67
Kappa	0,38	0,09	0,13	0,19
LR +	9,26	2,07	3,08	5,42
LR -	0,21	0,11	0,45	0,64

4.4.3 Ergebnisse zu kombinierten Lokalisationen des Atrophie-Algorithmus

Mit der Fragestellung, ob durch geschickte Kombination der Ergebnisse verschiedener Lokalisationen die Real-life Performance des Algorithmus verbessert werden kann, wurden unterschiedliche Kombinationen in allen Varianten ausprobiert. Dabei war das Ziel am Ende eine möglichst niedrige falsch-negative (FN) Anzahl zu haben und dass somit keine, bis wenig erkrankte Patienten von der KI übersehen werden bei gleichzeitig akzeptabler Rate an falsch-positiven (FP) Befunden. Bei der Kombination wird ein Patient mit positivem Befund in mindestens einer Lokalisation als positiv gewertet. Abbildung 23 zeigt dabei die drei besten Ergebnisse in Hinblick auf möglichst wenig falsch negative Befunde (Betrachtung aller vs. drei Lokalisationen und Wertung als erkrankt, wenn in einer Lokalisation positiv vs. isolierte Betrachtung von Lokalisation 2). So ist ersichtlich, dass bei Kombination von Lokalisation 1, 2, 3 und 4 die größte Menge an Bildern (583) vorhanden war und lediglich 1 FN (0,2%), aber gleichzeitig eine hohe Anzahl an FP Ergebnissen (232; 39,8%) resultierte. Bei Entfernung von Loc 4 zeigt sich dann, dass sich bei leicht reduzierter Bildanzahl (524) die Anzahl an FN (1; 0,2%) und FP Ergebnissen (211; 40,3%) kaum verändert. Ebenfalls die alleinige Betrachtung von Lokalisation 2 zeigt ein vergleichbares Resultat bei 1 FN (0,3%) und 146 FP (43,1%) Auswertungen bei insgesamt 339 vorhandenen Bildern.

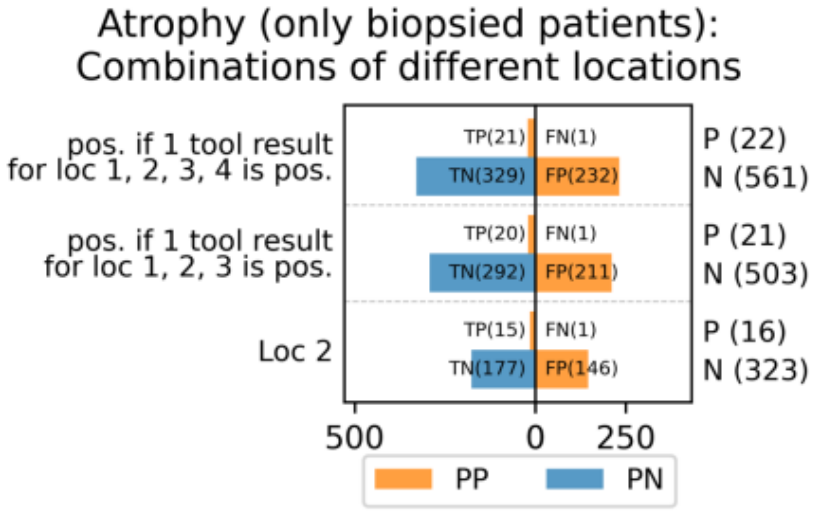


Abbildung 23 - Kombinierte Lokalisationen für die Atrophie-App

Orange: positive Einschätzung des Algorithmus (PP)

Blau: negative Einschätzung des Algorithmus (PN)

TP = richtig-positiv, TN = richtig-negativ, FP = falsch-positiv FN = falsch-negativ

P = Anzahl positiv ausgewerteter Ergebnisse insgesamt (richtig- und falsch-positiv)

N = Anzahl negativ ausgewerteter Ergebnisse insgesamt (richtig- und falsch-negativ)

Auf der linken Seite sind die richtig-positiven und richtig-negativen Ergebnisse des Algorithmus anhand unterschiedlicher Kombinationen aufgelistet. Auf der rechten Seite hingegen sind die entsprechenden Falschaussagen aufgelistet.

Statistische Metriken der kombinierten Lokalisationen

Tabelle 15 zeigt, dass für die Kombination aus allen vier Lokalisationen 583, für die Kombination aus Loc 1, 2 und 3 524 und für Loc 2 allein 339 Patienten beachtet wurden. Somit ergeben sich *Prävalenzraten* von jeweils 3%, 4% und 5%.

Die *Sensitivität* ergibt für die verglichenen Kombinationen 95% und zeigt ein ähnliches Ergebnis wie Loc 2 allein mit 94%. Die *Spezifität* hingegen ist durch die hohe Anzahl an falsch-negativen Ergebnissen für die Kombinationen mit 15% vergleichsweise schlecht. Loc 2 allein zeigt hingegen 55%. Insgesamt resultiert hieraus sowohl für die Kombinationen als auch für Loc 2 eine *Precision* von 8% bzw. 9%.

Der *F1-Score* ist für Loc 2 allein mit 0,17 ähnlich niedrig wie für die Kombinationen mit jeweils 0,15. Auch die *Accuracy* mit 57-60% sowie die *balanced Accuracy* mit 74-77% zeigen für alle drei verglichenen Auswertungsmethoden ähnliche Ergebnisse.

Auch die Werte für *LR*+ mit einem Faktor von 2,07-2,30 liegen für alle drei Messbereiche in einem ähnlichen Bereich. *LR*- hingegen zeigt sich mit 0,77 für die Kombination aus allen vier Lokalisationen schlechter als für Loc 2 allein mit 0,11. Die Kombination aus Loc 1, 2 und 3 weist mit 0,08 die beste negative Vorhersage auf.

Tabelle 15 – Statistische Metriken zu den kombinierten Lokalisationen der Atrophie-App

	Loc 1, 2, 3 und 4	**Loc 1, 2 und 3**	**Loc 2**
Gesamtzahl Patienten	583	524	339
Prävalenzrate	0,03	0,04	0,05
Accuracy (ACC)	0,60	0,59	0,57
Balanced Accuracy (bACC)	0,77	0,76	0,74
Precision (PPV)	0,08	0,08	0,09
Sensitivität	0,95	0,95	0,94
Spezifität	0,15	0,15	0,55
F1-Score	0,15	0,15	0,17
LR +	2,30	2,27	2,07
LR -	0,77	0,08	0,11

4.5 Ergebnisse zur eosinophilen Ösophagitis

Da bei Endoskopien Biopsien des Ösophagus nur bei sichtbaren Auffälligkeiten oder klinisch vorliegender ösophagealer Dysfunktion (z.B. Dysphagie) entnommen werden, ist der Anteil der Patienten mit durchgeführter Biopsie relativ niedrig (204 von 550). Neben einer Auswertung, die nur Patienten mit vorliegendem Biopsieergebnis berücksichtigt, präsentieren wir auch eine Auswertung, bei der alle Patienten ohne ösophageale Dysfunktion und fehlender Biopsie als negativ berücksichtigt werden, da die Definition der Erkrankung sowohl das Vorliegen einer bioptisch gesicherten Eosinophilie wie auch eine ösophageale Dysfunktion fordert.

4.5.1 Ergebnisse der Untersucher auf Patientenebene bei Betrachtung aller Patienten

Es liegen von 550 Patienten 13 (2,4%) positive und 537 (97,6%) negative Einschätzungen von den Untersuchern vor. Bei genauerer Betrachtung wird deutlich, dass von den 7 (1,3%) Patienten mit EoE 6 (85,7%) richtig-positiv (RP) und 1 (14,3%) falsch-negativ (FN) eingeschätzt wurden. Insgesamt wurden somit also bei 543 (98,7%) Patienten ohne EoE 7 (1,3%) falsch-positive (FP) und 536 (98,7%) richtig-negative (RN) Einschätzungen abgegeben (siehe Tabelle 16).

Tabelle 16 - Einschätzungen der Untersucher zur eosinophilen Ösophagitis bei Betrachtung aller Patienten

Die Prozentangaben der positiven und negativen Patienten sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl von 550 Patienten. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der 7 positiven Patienten. Die richtig-negativen und falsch-positiven Prozentangaben beziehen sich auf die Anzahl der 543 negativen Patienten.
EoE = eosinophile Ösophagitis

	Einschätzung Untersucher
Patienten mit EoE	7 (1,3%)
Patienten ohne EoE	543 (98,7%)
Positive Einschätzung (PP)	13 (2,4%)
Negative Einschätzung (PN)	537 (97,6%)
Richtig-positiv (RP)	6 (85,7%)
Richtig-negativ (RN)	536 (98,7%)
Falsch-positiv (FP)	7 (1,3%)
Falsch-negativ (FN)	1 (14,3%)

Statistische Metriken

Von 550 untersuchten Patienten mit 7 (1,3%) Erkrankten und 543 (98,7%) Patienten ohne EoE ergibt sich eine *Prävalenzrate* von 1% für die EoE. Auch hier besteht das Problem einer unbalancierten Verteilung durch die geringe Anzahl an positiven Fällen.

Für die Untersucher ergibt sich eine *Sensitivität* von 86% und eine *Spezifität* von 99% mit einer resultierenden *Precision* von 46% durch die starke Beeinflussung durch die geringe Prävalenz und gleichzeitig großen Kohorte an negativen Fällen.

Der *F1-Score* beträgt in diesem Falle 0,6. Die *Accuracy* zeigt sich mit 99% besonders hoch und bleibt auch in diesem hohen Bereich bei dem Versuch die Gruppen anzugleichen, also eine *bACC* von 92%.

Auch *LR+* zeigt sich besonders hoch mit einem Faktor von 66,49 für die Erkrankung bei positivem Testergebnis. *LR-* hingegen zeigt sich niedrig mit einem Faktor von 0,14, was bedeutet,

dass die Patienten bei negativer Einschätzung durch den Endoskopiker mit hoher Wahrscheinlichkeit tatsächlich gesund sind.

Die *AUC*, beträgt 0,92 und zeigt somit eine gute Leistung des Tests durch die Untersucher. Tabelle 17 zeigt die einzelnen Werte.

Tabelle 17 - Statistische Metriken der Einschätzung der Untersucher zur eosinophilen Ösophagitis bei allen Patienten

	Einschätzung Untersucher
Gesamtzahl Patienten	550
Prävalenzrate	0,01
Accuracy (ACC)	0,99
Balanced Accuracy (bACC)	0,92
Precision (PPV)	0,46
Sensitivität	0,86
Spezifität	0,99
F1-Score	0,6
Area under the curve (AUC)	0,92
LR +	66,49
LR -	0,14

4.5.2 Ergebnisse der Untersucher auf Patientenebene bei Betrachtung der biopsierten Patienten

Von 204 durchgeführten Biopsien sind 7 (3,4%) positiv und 197 (96,6%) negativ ausgefallen. Die Angaben der Untersucher sind 13 (6,4%) positive und 191 (93,6%) negative Befunde. Daraus resultieren 6 (85,7%) richtig-positive (RP) und 1 (14,3%) falsch-negative (FN) bzw. 7 (3,5%) falsch-positive (FP) und 190 (96,5%) richtig-negative (RN) Auswertungen (siehe Tabelle 18) und deckt sich somit mit den bisherigen Ergebnissen unter *4.5.1 Ergebnisse der Untersucher auf Patientenebene bei Betrachtung aller Patienten*.

Tabelle 18- Einschätzungen der Untersucher zur eosinophilen Ösophagitis bei Betrachtung biopsierter Patienten

Die Prozentangaben der positiven und negativen Patienten sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl von 204 biopsierten Patienten. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der 7 positiven Biopsien. Die richtig-negativen und falsch-positiven Prozentangaben beziehen sich auf die Anzahl der 197 negativen Biopsien.

	Einschätzung Untersucher
Positive Biopsien	7 (3,4%)
Negative Biopsien	197 (96,6%)
Positive Einschätzung (PP)	13 (6,4%)
Negative Einschätzung (PN)	191 (93,6%)
Richtig-positiv (RP)	6 (85,7%)
Richtig-negativ (RN)	190 (96,5%)
Falsch-positiv (FP)	7 (3,5%)
Falsch-negativ (FN)	1 (14,3%)

STATISTISCHE METRIKEN

Bei 204 biopsierten Patienten waren 7 (3,4%) positiv und 197 (96,6%) negativ wodurch sich eine leicht höhere *Prävalenzrate* von 3% im Gegensatz zur Berechnung in der gesamten Kohorte ergibt.

Die *Sensitivität* beträgt 86% die *Spezifität* 96%. Die *Precision* liegt weiterhin bei 46%.

Der *F1-Score* beträgt nach wie vor 0,6. Die *Accuracy* ist leicht gesunken auf 96%, wobei die *bACC* sich mit 91% kaum verändert hat zur Auswertung der gesamten Patientenkohorte.

LR+ ist niedriger und berechnet sich auf einen Faktor von 24,12, wobei *LR-* hingegen mit einem Faktor von 0,15 in einem ähnlichen Bereich bleibt.

Die geschätzte *AUC* ist mit 0,91 im Vergleich zu 0,92 bei Auswertung aller untersuchten Patienten auch kaum verändert. Die Tabelle 19 fasst die einzelnen Werte nochmals zusammen.

Tabelle 19 - Statistische Metriken der Einschätzung der Untersucher zur eosinophilen Ösophagitis bei biopsierten Patienten

	Einschätzung Untersucher
Gesamtzahl Patienten	204
Prävalenzrate	0,03
Accuracy (ACC)	0,96
Balanced Accuracy (bACC)	0,91
Precision (PPV)	0,46
Sensitivität	0,86
Spezifität	0,96
F1-Score	0,6
Area under the curve (AUC)	0,91
LR +	24,12
LR -	0,15

4.5.3 Ergebnisse der eosinophilen Ösophagitis-App auf Bildebene bei Betrachtung aller Patienten

Auch für die EoE ergeben sich für jede Lokalisation unterschiedliche Ergebnisse, da unterschiedliche Anzahlen an Bildern zur Verfügung stehen.

Loc 5 hat 438 Bilder zur Verfügung, davon sind 27 (6,2%) positiv und 411 (93,8%) negativ vom Algorithmus eingeschätzt worden. Bei 6 (1,4%) Patienten mit EoE ergeben sich somit 6 (100%) richtig-positive (RP) und 0 (0%) falsch-negative (FN) bzw. bei 432 (98,6%) gesunden Patienten 21 (4,9%) falsch-positive (FP) und 411 (95,1%) richtig-negative (RN) Auswertungen.

Für Loc 6 gibt es 498 Bilder mit 34 (6,8%) positiv und 464 (93,2%) negativ durch den Algorithmus vorausgesagten Ergebnissen. Es ergaben sich bei 7 (1,4%) Patienten mit EoE somit 4 (57,1%) RP und 3 (42,9%) FN bzw. bei 491 (98,6%) gesunden Patienten 30 (6,1%) FP und 461 (93,9%) RN Resultate.

Daraus ergibt sich, dass Loc 6 mit 6,1% (30 von 491) eine leicht höhere Rate an FP Befunden hat als Loc 5 mit 4,9% (21 von 432). Außerdem ist ebenfalls das Ergebnis der FN Befunde bei Loc 6 mit 42,9% (3 von 7) höher als bei Loc 5 mit 0% (0 von 6).

Tabelle 20 zeigt die einzelnen Werte.

Tabelle 20 - Einschätzungen der eosinophilen Ösophagitis-App bei Betrachtung aller Patienten

Die Prozentangaben der positiven und negativen Patienten sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl der vorhandenen Bilder pro Lokalisation. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der 6 bzw. 7 positiven Patienten. Die richtig-negativen und falsch-positiven Prozentangaben beziehen sich auf die Anzahl der 432 bzw. 491 negativen Patienten.
EoE = eosinophile Ösophagitis

	Loc 5	**Loc 6**
Gesamtzahl Bilder	438	498
Patienten mit EoE	6 (1,4%)	7 (1,4%)
Patienten ohne EoE	432 (98,6%)	491 (98,6%)
Positive Einschätzung (PP)	27 (6,2%)	34 (6,8%)
Negative Einschätzung (PN)	411 (93,8%)	464 (93,2%)
Richtig-positiv (RP)	6 (100%)	4 (57,1%)
Richtig-negativ (RN)	411 (95,1%)	461 (93,9%)
Falsch-positiv (FP)	21 (4,9%)	30 (6,1%)
Falsch-negativ (FN)	0 (0%)	3 (42,9%)

STATISTISCHE METRIKEN

Trotz der unterschiedlichen Anzahl an Patienten für Loc 5 (438, davon 6 positiv und 432 negativ) und Loc 6 (498, davon 7 positiv und 491 negativ) ergibt sich für beide Lokalisationen eine *Prävalenzrate* von 1%.

Die *Sensitivität* ist für Loc 5 mit 100% am höchsten. Für Loc 6 beträgt diese lediglich 57%. Die *Spezifität* hingegen ist für beide Lokalisationen ähnlich hoch mit 95% für Loc 5 bzw. 94% für Loc 6. Die daraus resultierende *Precision* ist für beide Lokalisationen niedrig, wobei Loc 5 mit 22% etwas höher berechnet ist als Loc 6 mit 12%. Auch in diesem Falle lässt sich dies nach wie vor durch die unbalancierte Kohorte erklären.

Der *F1-Score* ist für Loc 5 etwas besser mit 0,36. Loc 6 beläuft sich auf einen Wert von 0,2. Die *Accuracy* befindet sich für beide Lokalisationen in einem ähnlichen Bereich, wobei sie für die Loc 5 mit 95% leicht höher ist als für Loc 6 mit 93%. Interessant ist die *bACC*, welche für Loc 5 sogar etwas höher ist mit 98%. Für Loc 6 hingegen ist diese auf 76% gesunken.

LR+ ist für Loc 5 höher mit einem Faktor von 20,57 zu einem Faktor von 9,35 für Loc 6. *LR-* hingegen beträgt für Loc 5 0 und für Loc 6 0,46. Somit liegt bei negativer Vorhersage mit hoher Wahrscheinlichkeit ein negativer Befund vor.

Die *AUC* liegt für Loc 5 bei 0,98 und für Loc 6 bei 0,76. Dies bedeutet, dass Loc 5 eine bessere Testleistung zeigt als Loc 6. Tabelle 21 gibt einen Überblick über die einzelnen Werte.

Tabelle 21 - Statistische Metriken der Einschätzung der eosinophilen Ösophagitis-App bei Betrachtung aller Patienten

	Loc 5	**Loc 6**
Gesamtzahl Bilder	438	498
Prävalenzrate	0,01	0,01
Accuracy (ACC)	0,95	0,93
Balanced Accuracy (bACC)	0,98	0,76
Precision (PPV)	0,22	0,12
Sensitivität	1	0,57
Spezifität	0,95	0,94
F1-Score	0,36	0,2
Area under the curve (AUC)	0,98	0,76
LR +	20,57	9,35
LR -	0	0,46

4.5.4 Ergebnisse zu kombinierten Lokalisationen der eosinophilen Ösophagitis-App bei Betrachtung aller Patienten

Auch für die EoE-App wurde ausprobiert ob durch Kombination der Lokalisationen (Wertung als positiv, wenn in einer der beiden Lokalisation positiv) eine möglichst niedrige Anzahl an falsch-negativen (FN) Ergebnissen möglich ist. Abbildung 24 zeigt die zwei besten Ergebnisse mit dem Ziel, kleinere Balken auf der rechten Seite der falsch zugeordneten Ergebnisse im Vergleich zur linken Seite mit den richtig zugeordneten Ergebnissen zu haben. So ergab sich für die Kombination beider Lokalisationen eine Anzahl von 0 FN Ergebnissen bei einer Anzahl von insgesamt 550 Patienten. Jedoch lag die Anzahl der falsch-positiven (FP) Auswertungen bei 50 (9,1%).

Das beste Ergebnis zeigt die alleinige Verwendung von Lokalisation 5. Diese zeigt bei 438 Auswertungen ebenfalls 0 FN Ergebnisse, aber lediglich eine Anzahl von 21 (4,9%) FP Ergebnissen und somit weniger als bei Verwendung beider Lokalisationen.

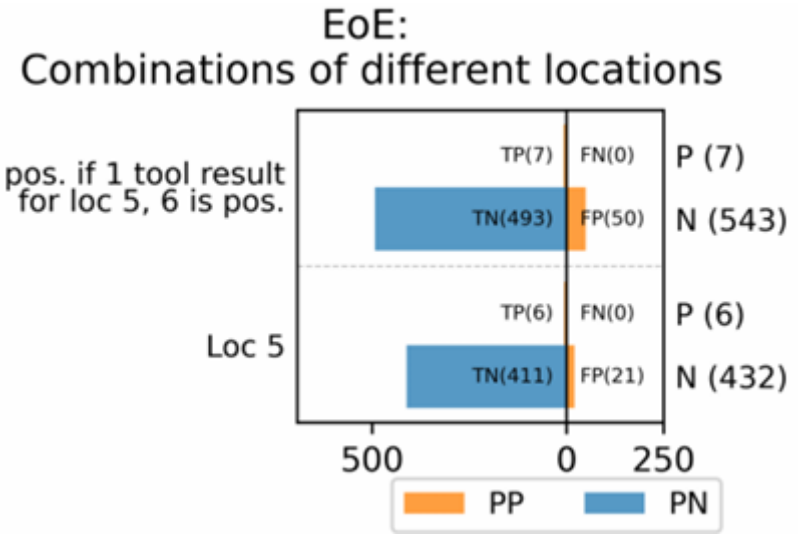


Abbildung 24 - Kombinierte Lokalisationen für die eosinophile Ösophagitis-App

Orange: positive Einschätzung des Algorithmus (PP)

Blau: negative Einschätzung des Algorithmus (PN)

TP = richtig-positiv, TN = richtig-negativ, FP = falsch-positiv FN = falsch-negativ

P = Anzahl positiv ausgewerteter Ergebnisse insgesamt (richtig- und falsch-positiv)

N = Anzahl negativ ausgewerteter Ergebnisse insgesamt (richtig- und falsch-negativ)

Auf der linken Seite sind die richtig-positiven und richtig-negativen Ergebnisse des Algorithmus anhand unterschiedlicher Kombinationen aufgelistet. Auf der rechten Seite hingegen sind die entsprechenden Falschaussagen aufgelistet.

Statistische Metriken der kombinierten Lokalisationen

Bei der Kombination von Loc 5 und 6 beträgt die *Prävalenzrate* weiterhin 1%.

Die *Sensitivität* bleibt weiterhin bei 100%, wohingegen die *Spezifität* für beide Lokalisationen zusammen leicht absinkt (91%) im Vergleich zur einzelnen Loc 5 (95%). Die daraus resultierende *Precision* ist die Kombination weiter gesunken auf 13%. Dies lässt sich weiterhin durch die unbalancierte Kohorte erklären.

Der *F1-Score* ist für Loc 5 etwas besser mit 0,36 als für die Kombination mit 0,22. Auch die *Accuracy* (91%) sowie die balanced Accuracy (95%) sind etwas niedriger als für die Loc 5 allein (ACC 95%, bACC 98%).

LR+ ist für Loc 5 höher mit einem Faktor von 20,57 zu einem Faktor von 10,86 für beide Lokalisationen zusammen. *LR-* hingegen beträgt für beide Methoden 0 und zeigt, dass somit bei negativer Vorhersage mit hoher Wahrscheinlichkeit ein negativer Befund vorliegt.

Tabelle 22 gibt einen Überblick über die einzelnen Werte.

Tabelle 22 - Statistische Metriken zu den kombinierten Lokalisationen der eosinophilen Ösophagitis-App

	Loc 5 und 6	**Loc 5**
Gesamtzahl Patienten	550	438
Prävalenzrate	0,01	0,01
Accuracy (ACC)	0,91	0,95
Balanced Accuracy (bACC)	0,95	0,98
Precision (PPV)	0,13	0,22
Sensitivität	1	1
Spezifität	0,91	0,95
F1-Score	0,22	0,36
LR +	10,86	20,57
LR -	0	0

4.5.5 Ergebnisse der eosinophilen Ösophagitis-App auf Bildebene bei Betrachtung biopsierter Patienten

Durch die unterschiedliche Bildanzahl pro Lokalisation variieren auch die darin enthaltenen Biopsien in der Auswertung. Tabelle 23 zeigt die einzelnen Werte.

Für Loc 5 stehen 169 Bilder mit 6 (3,5%) positiven und 163 (96,5%) negativen Biopsien zur Verfügung. Die KI gab dabei 16 (9,5%) positive und somit 6 (100%) richtig-positive (RP) und 0 (0%) falsch-negative (FN) bzw. 153 (90,5%) negative und dementsprechend 153 (93,9%) richtig-negative (RN) und 10 (6,1%) falsch-positive (FP) Ergebnisse an.

Loc 6 hingegen hatte 182 Bilder zur Auswertung und somit 7 (3,8%) positive und 175 (96,2%) negative Biopsien. Insgesamt wurden 20 (11%) positive und 162 (89%) negative Befunde von der KI angegeben. Daraus resultieren 4 RP (57,1%) und 3 (42,9%) FN bzw. 159 (90,9%) RN und 16 (9,1%) FP Ergebnisse. Hierbei zeigt sich also insgesamt wie bereits unter *4.5.3 Ergebnisse der eosinophilen Ösophagitis-App auf Bildebene bei Betrachtung aller Patienten*, dass Loc 5 (10 FP, 0 FN) etwas bessere Ergebnisse als Loc 6 (16 FP, 3 FN) zeigt.

Tabelle 23 - Einschätzung der eosinophilen Ösophagitis-App bei biopsierten Patienten

Die Prozentangaben der positiven und negativen Patienten sowie die positive und negative Einschätzung beziehen sich auf die Gesamtzahl der vorhandenen Bildern von biopsierten Patienten pro einzelne Lokalisation. Die richtig-positiven und falsch-negativen Prozentangaben beziehen sich auf die Anzahl der 6 bzw. 7 positiven Biopsien. Die richtig-negativen und falsch-positiven Prozentangaben beziehen sich auf die Anzahl der 163 bzw. 175 negativen Biopsien.

	Loc 5	**Loc 6**
Gesamtzahl Bilder	169	182
Positive Biopsien	6 (3,5%)	7 (3,8%)
Negative Biopsien	163 (96,5%)	175 (96,2%)
Positive Einschätzung (PP)	16 (9,5%)	20 (11%)
Negative Einschätzung (PN)	153 (90,5%)	162 (89%)
Richtig-positiv (RP)	6 (100%)	4 (57,1%)
Richtig-negativ (RN)	153 (93,9%)	159 (90,9%)
Falsch-positiv (FP)	10 (6,1%)	16 (9,1%)
Falsch-negativ (FN)	0 (0%)	3 (42,9%)

STATISTISCHE METRIKEN

Bei Reduktion der Auswertung auf lediglich biopsierte Patienten ergibt sich für beide Lokalisationen trotz unterschiedlicher Patientenanzahlen eine *Prävalenzrate* von 4% und ist somit höher als bei der Auswertung aller Patienten.

Die *Sensitivität* ist für beide Lokalisationen unverändert (Loc 5 100%, Loc 6 57%), wobei die *Spezifität* in beiden Fällen leicht gesunken ist (Loc 5 94%, Loc 6 91%). Die sich daraus ergebende *Precision* ist für Loc 5 auf 38% und Loc 6 auf 20% berechnet und somit etwas besser.

Auch der *F1-Score* liegt etwas höher und beträgt für Loc 5 0,55 und Loc 6 0,3. Die *Accuracy* hingegen mit 94% für Loc 5 und 90% für Loc 6 hat sich leicht verschlechtert. Die *bACC* ist fast unverändert geblieben (Loc 5 97%, Loc 6 74%) und stimmt somit bereits mit der errechneten für alle Patienten überein.

LR+ ist in diesem Falle für beide Lokalisationen etwas gesunken und beträgt für Loc 5 16,3 und für Loc 6 6,25. *LR-* hingegen ist für Loc 5 mit 0 und für Loc 6 mit 0,47 unverändert geblieben.

Zuletzt wurde wieder der *AUC-Wert* abgeschätzt. Dieser errechnet sich für Loc 5 auf 0,97 und Loc 6 auf 0,74 und weicht somit kaum von dem Wert des Ergebnisses der Auswertung aller Patienten durch den Algorithmus ab. Somit zeigt Loc 5 nach wie vor eine bessere Testleistung als Loc 6.

Tabelle 24 zeigt die einzelnen Werte im Überblick.

Tabelle 24 - Statistische Metriken der Einschätzung der eosinophilen Ösophagitis-App bei biopsierten Patienten

	Loc 5	**Loc 6**
Gesamtzahl Bilder	169	182
Prävalenzrate	0,04	0,04
Accuracy (ACC)	0,94	0,9
Balanced Accuracy (bACC)	0,97	0,74
Precision (PPV)	0,38	0,2
Sensitivität	1	0,57
Spezifität	0,94	0,91
F1-Score	0,55	0,3
Area under the curve (AUC)	0,97	0,74
LR +	16,3	6,25
LR -	0	0,47

4.6 Fehleranalyse der Auswertung durch die Algorithmen

Es gibt verschiedene Fehlermöglichkeiten, die eine Fehleinordnung durch die KI in den sechs Lokalisationen begründen können. Um die Fehler besser verstehen zu können, wurden die falsch klassifizierten Bilder nachträglich durch einen erfahrenen Endoskopiker beurteilt. Die Fehleranzahl errechnet sich aus der Summe der falsch-positiv (FP) und falsch-negativ (FN) eingeordneten Bilder und beträgt insgesamt 380 Bilder. Sie ist am höchsten für Loc 2 der Atrophie-App mit insgesamt 147 (38,7%) falsch bewerteten Bildern und am niedrigsten für Loc 5 der EoE-App mit insgesamt 21 (5,5%).

Die möglichen Fehler wurden unterschiedlichen Klassen zugeordnet. Folgende Ursachen für falsch zugeordnete Bilder wurden identifiziert: Bildkontamination mit Blut oder Sekreten, Untersuchungsfehler wie z.B. zu wenig Luftinsufflation oder übermäßige Luftgabe, schlechte Bildqualität wie ein verschwommenes Bild, andere sichtbare konkurrierende Erkrankung, durchschimmernde Gefäße ohne Atrophie. In einem großen Teil der fehlklassifizierten Bilder konnte aber keine klare Ursache für die Fehlklassifizierung identifiziert werden. Am häufigsten ist insgesamt eine solche unbekannte Ursache (216; 56,8%), wobei dies v.a. Loc 2 der Atrophie-App mit 133 (90,5%) von 147 falsch bewerteten Bildern betrifft. Dahinter folgen Untersuchungsfehler (50; 13,2%), insbesondere für Loc 3 der Atrophie-App mit 17 (17,7%) von 96 potenziellen Fehlern.

In einem Teil der Fälle muss davon ausgegangen werden, dass bei tatsächlich vorliegender Erkrankung Biopsien in nicht repräsentativen Arealen entnommen wurden, sodass in diesen Fällen das Biopsieergebnis als Grundwahrheit anzuzweifeln ist und tatsächlich vom Vorliegen einer Atrophie auszugehen ist.

Insgesamt sind somit von den fehlerhaft ausgewerteten Bildern für Loc 1 21 von 36 Bildern (58,3%), Loc 2 14 von 147 Bildern (9,5%), Loc 3 50 von 96 Bildern (52%), Loc 4 26 von 47 Bildern (55,3%), Loc 5 20 von 21 Bildern (95,2%), und Loc 6 alle 33 Bilder (100%) auf mögliche erklärbare Ursachen im Zusammenhang mit der Bild- und Biopsiequalität zurückzuführen.

Tabelle 25 und Abbildung 26 geben eine Übersicht über die unterschiedliche Anzahl der möglichen Fehler mit Verteilung auf die einzelnen Lokalisationen.

Tabelle 25 - Anzahl der potenziellen Fehlerarten nach Lokalisation

Die Prozentangaben der Ursachen in den Spalten der einzelnen Lokalisationen beziehen sich auf den Anteil an der Gesamtanzahl an Fehlern pro Lokalisation. Die rechte „Gesamt"-Spalte zählt die Anzahl des Fehlers insgesamt für alle sechs Lokalisationen zusammen. Diese Prozentangabe bezeichnet den Anteil des Fehlers an der Gesamtzahl von 380 Fehlern. Die Prozentangabe der letzten „Gesamt"-Zeile unterhalb der Lokalisationen gibt den Anteil der Fehler der jeweiligen Lokalisation als Anteil an der Gesamtzahl von 380 Fehlern an.

Fehler	**Loc 1**	**Loc 2**	**Loc 3**	**Loc 4**	**Loc 5**	**Loc 6**	**Gesamt**
Kontamination	7 (19,4%)	2 (1,3%)	8 (8,3%)	7 (14,9%)	1 (4,8%)	3 (9,1%)	**28** (7,4%)
Untersuchungsfehler	8 (22,2%)	1 (0,7%)	17 (17,7%)	10 (21,2%)	4 (19%)	10 (30,3%)	**50** (13,2%)
Schlechte Bildqualität	0	6 (4,1%)	2 (2,1%)	4 (8,5%)	9 (42,8%)	11 (33,3%)	**32** (8,4%)
Zeichen anderer Erkrankungen	2 (5,6%)	1 (0,7%)	2 (2,1%)	2 (4,3%)	5 (23,8%)	7 (21,2%)	**19** (5%)
Durchscheinende Gefäße	0	0	14 (14,6%)	1 (2,1%)	0	0	**15** (3,9%)
Biopsie nicht repräsentativ	4 (11,1%)	4 (2,7%)	7 (7,3%)	2 (4,3%)	1 (4,8%)	2 (6,1%)	**20** (5,3%)
Unbekannt	15 (41,7%)	133 (90,5%)	46 (47,9%)	21 (44,7%)	1 (4,8%)	0	**216** (56,8%)
Gesamt	**36** (9,5%)	**147** (38,7%)	**96** (25,3%)	**47** (12,3%)	**21** (5,5%)	**33** (8,7%)	**380** (100%)

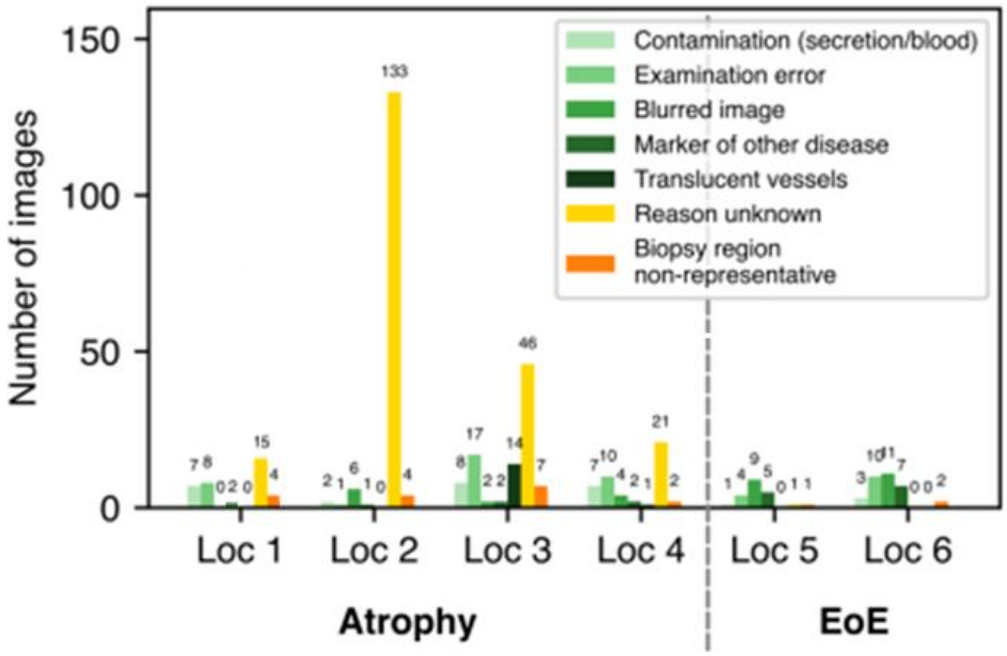


Abbildung 25 - Anzahl und Art der potenziellen Fehler pro Lokalisation

Loc 1 bis 4 zeigen die Fehler für die Atrophie-App bzw. Loc 5 und 6 für die EoE-App. Die unterschiedlichen Balken beschreiben die verschiedenen Ursachen für eine mögliche Fehlklassifizierung.

4.7 Vergleich der Ergebnisse

4.7.1 Vergleich für die präkanzerösen Bedingungen des proximalen Magens

Absolute Zahlen

Insgesamt wurden sowohl durch die Untersucher als auch durch die KI die meisten Patienten richtig zugeordnet (TP und TN), wodurch die falschen Ergebnisse (FP und FN) wesentlich geringer ausfallen (siehe Abb. 26). Durch die jeweils unterschiedliche Anzahl an untersuchten Patienten werden die absoluten Zahlen betrachtet, welche sich jedoch nur auf die biopsierten Patienten (583) beziehen. Dabei zeigt sich, dass die KI durchschnittlich mehr falsch-positive (FP) Ergebnisse für die einzelnen Lokalisationen hat, wobei die niedrigste Anzahl Loc 1 mit 33 (8,8%) FP auf 376 negative Patienten darstellt, als die Untersucher mit 16 (2,9%) FP Auswertungen auf 561 negative Patienten. Der Vorteil der KI ist jedoch die niedrigere Anzahl an falsch-negativen (FN) Ergebnissen im Vergleich zu den Untersuchern. Die KI gibt als schlechtestes Ergebnis Loc 4 mit 10 (58,8%) FN auf 17 positive Patienten bzw. am besten Loc 2 mit 1 (6,3%) FN auf 16 positive Patienten an, wobei die Untersucher 13 (59,1%) FN Auswertungen auf 22 positive Biopsien haben. Somit ist Loc 2 hinsichtlich des Erkennens der erkrankten Patienten am zuverlässigsten, zeigt aber gleichzeitig eine sehr hohe FP-Rate mit 146 (45,2%) von 323 negativen Biopsien.

Es lässt sich aus diesen Zahlen nicht klar ableiten, ob die KI oder die Untersucher bessere Ergebnisse liefern. So zeigen die Untersucher eine niedrige FP-Rate und die KI eine niedrige FN-Rate.

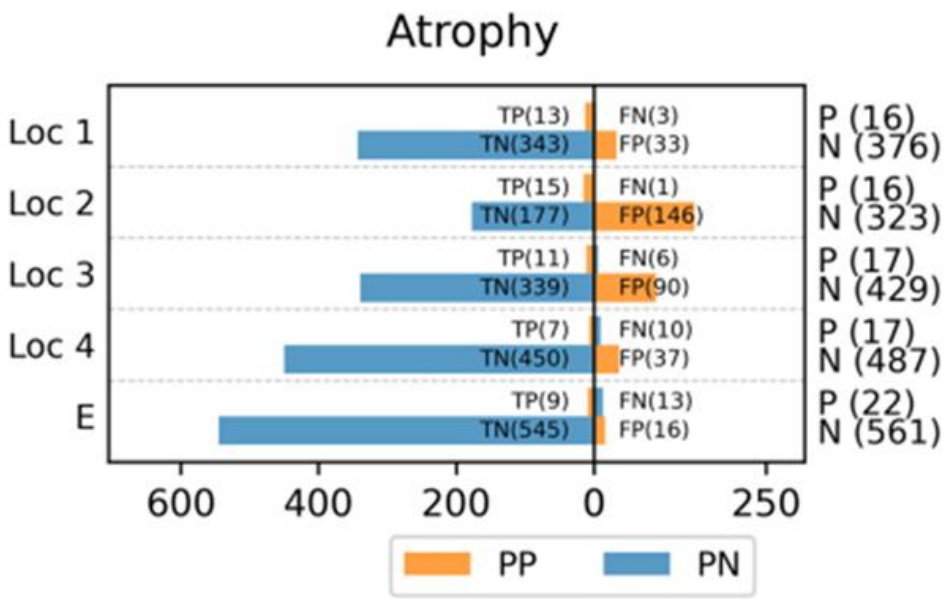


Abbildung 26 - Einschätzung der Endoskopiker und Atrophie-App

Loc 1 bis 4 = Atrophie-Algorithmus. E = Endoskopiker

Orange: positive Einschätzung des Algorithmus und der Endoskopiker (PP)

Blau: negative Einschätzung des Algorithmus und der Endoskopiker(PN)

TP = richtig-positiv, TN = richtig-negativ, FP = falsch-positiv FN = falsch-negativ

P = Anzahl positiv ausgewerteter Ergebnisse insgesamt (richtig- und falsch-positiv)

N = Anzahl negativ ausgewerteter Ergebnisse insgesamt (richtig- und falsch-negativ)

Auf der linken Seite sind die richtig-positiven und richtig-negativen Ergebnisse des Algorithmus und der Untersucher aufgelistet. Auf der rechten Seite hingegen sind die entsprechenden Falschaussagen aufgelistet.

Statistische Metriken

Die *Prävalenzrate* ist für die Untersucher und für die einzelnen Lokalisationen der KI in einem ähnlich niedrigen Bereich zwischen 3% und maximal 5%.

Die Untersucher sowie Loc 4 der KI haben mit 41% die schlechteste *Sensitivität* im Vergleich zu 65% für Loc 3 sowie 81% für Loc 1. Loc 2 hat mit 94% die höchste Sensitivität, was sich durch die hohe Anzahl an richtig-positiven (RP, 15 von 16) bzw. niedrige Anzahl an falsch-negativen (FN, 1 von 16) Ergebnissen erklären lässt.

Die *Spezifität* hingegen ist für die Untersucher mit 97% am größten durch den großen Anteil an richtig-negativen (RN) bzw. geringen Anteil an falsch-positiven (FP) Ergebnissen. Für die KI ist die beste Lokalisation Loc 4 mit 92% bzw. die schlechteste ist Loc 2 mit 55%. Insgesamt lässt sich dies durch die durchgehend höheren FP Ergebnisse erklären. Insgesamt fällt in dieser Studie die *Spezifität* besser aus als die *Sensitivität* durch den überwiegenden Anteil an tatsächlich negativen Patienten.

Die *Precision* ist sowohl für die KI als auch für die Untersucher niedrig ausgefallen, wobei letztere mit 36% besser abschneiden als die KI mit maximal 28% für Loc 1 bzw. am wenigsten für Loc 2 mit 9%. Dies ist v.a. durch die hohe Anzahl an FP Patienten bei der KI zu erklären trotz der besseren Einschätzung der RP Patienten. Der aus *Precision* und *Sensitivität* resultierende *F1-Score* ist wiederum für Loc 1 mit 0,42 am besten ausgefallen. Loc 2 ist auch hier mit 0,17 am schlechtesten. Die Untersucher haben einen Wert von 0,38 und sind somit sehr nah am Score von Loc 1.

Bei der *Accuracy* hingegen haben die Untersucher das beste Ergebnis mit 95%, gefolgt von Loc 1 und 4 mit jeweils 91%. Am schlechtesten hat Loc 2 der KI mit 57% abgeschnitten. Bei Angleichung der negativen und positiven Fällen zur Berechnung der *bACC* hingegen zeigt Loc 1 mit 86% das beste Ergebnis. Loc 4 sinkt ab auf 67% und Loc 2 verbessert sich sogar auf 74%. Loc 3 ist in beiden Fällen bei 78% bzw. 72% und verändert sich somit kaum. Die Untersucher verschlechtern sich jedoch auf 69% und würden im Falle der *bACC* also ein schlechteres Ergebnis als die KI zeigen.

LR+ zeigt bei den Untersuchern mit einem Faktor von 14,34 das höchste Ergebnis. Für die KI ist dies für Loc 1 mit 9,26 am höchsten bzw. für Loc 2 mit 2,07 am niedrigsten. Somit ist die Wahrscheinlichkeit für einen Patienten bei positivem Testergebnis tatsächlich krank zu sein bei den Untersuchern höher.

LR- zeigt den besten Wert für Loc 2 der KI mit 0,11, wodurch es sehr wahrscheinlich ist, dass Patienten mit einem negativem Ergebnis tatsächlich gesund sind. Loc 1 mit 0,22 und Loc 3 mit 0,45 befinden sich in einem ähnlich niedrigen Bereich. Die Untersucher mit 0,61 sowie Loc 4 mit 0,64 hingegen schneiden ähnlich schlecht ab. Da die Werte näher an der Eins sind ist es wahrscheinlicher, dass Patienten trotz negativem Ergebnis krank sein könnten.

Die geschätzte *AUC* ist für die Loc 1 der KI mit 0,86 am besten ausgefallen. Loc 2 und 3 befinden sich mit 0,74 bzw. 0,72 im mittleren Bereich. Loc 4 mit 0,67 und das Ergebnis der Untersucher mit 0,69 stellen die schlechtesten Werte dar. Somit zeigen Loc 1, Loc 2 und Loc 3 eine bessere Testleistung als Loc 4 und die Untersucher, welche sogar näher an der 0,5 sind und somit einer schlechten Testleistung entspricht.

Insgesamt zeigen sowohl die Untersucher als auch die KI Bereiche in denen sie besser oder schlechter abschneiden. Es lässt sich nicht eindeutig sagen, wer allgemein betrachtet ein besseres Auswertungsergebnis erzielt, da dieses auch abhängig von der Zielsetzung (z.B. wenig FP Ergebnisse, wenig FN Ergebnisse) ist. So hat die KI je nach Lokalisation die besten Ergebnisse für *Sensitivität*, *F1-Score*, *bACC*, *LR-* und *AUC*, wohingegen die Untersucher bessere Ergebnisse in *Spezifität*, *Precision*, *Accuracy* und *LR+* erreicht haben. Jedoch muss bei all diesen Werten bedacht werden, dass manche sich besser eignen für unbalancierte Kohorten wie in dieser Studie und andere wiederum für ausgeglichene Fallzahlen. Aus ärztlicher Sicht ist eine möglichst geringe Rate falsch negativer Patienten bei akzeptabler Anzahl falsch positiver Patienten anzustreben. Unter diesen Bedingungen würde Loc 2 mit nur einem FN Ergebnis am besten abschneiden, jedoch zeigt diese gleichzeitig eine sehr hohe FP-Rate (146). Entsprechend ist zum aktuellen Zeitpunkt die KI am zuverlässigsten bei Benutzung von Loc 1, da hier lediglich 3 FN und 33 FP Ergebnisse aufgetreten sind.

VENN-DIAGRAMME

Venn-Diagramme sind eine Art der Darstellung von Mengendiagrammen, um alle möglichen Beziehungen zwischen den vorhandenen Mengen zu visualisieren. So ermöglichen sie sowohl das Erkennen von möglichen Zusammenhängen als auch das Fehlen von solchen. Abbildung 27 stellt dies für die proximale AG dar. Aus Gründen der Übersicht ist ein Diagramm für die positiven und ein Diagramm für die negativen Ergebnisse getrennt dargestellt worden und beinhalten die Auswertungen der vier Lokalisationen der App, der Untersucher und der Biopsie. Ziel ist es zu erkennen, ob und wie häufig diese drei Untersuchungsmethoden im Ergebnis übereinstimmen.

Bei den positiven Ergebnissen (linkes Diagramm) zeigt sich, dass die größte Überlappung (25) zwischen Loc 2 und Loc 3 stattfindet, jedoch ohne mit den positiven Biopsien oder den positiven Auswertungen der Untersucher übereinzustimmen. Die restlichen vorzufindenden Überlappungen fallen sehr gering aus, wobei sich nur bei 2 positiven Ergebnissen in der Mitte sowohl die KI als auch die Untersucher und die Biopsieergebnisse einig sind. Loc 1 zeigt für 13 Ergebnisse, Loc 2 für 94, Loc 3 für 42, Loc 4 für 20 und die Untersucher für 6 positive Aussagen keine einzige Überlappung untereinander. Eine einzige positive Biopsie von 22 wurde weder von den Untersuchern noch von der KI als positiv richtig ausgewertet.

Bei den negativen Ergebnissen (rechtes Diagramm) findet sich die größte Überlappung (105) zwischen Loc 1, Loc 3, Loc 4, den Untersuchern und den Biopsien. Danach folgen 81 Überlappungen zwischen Loc 4, den Untersuchern und den Biopsien, sowie 53 weitere noch zusätzlich mit Loc 3. Weitere 49 negative Ergebnisse überschneiden sich zwischen Loc 1, Loc

4, den Untersuchern und den Biopsien. Mittig finden sich 100 überlappende Ergebnisse, in denen sich alle Untersuchungsmethoden einig sind. Die restlichen Überlappungen fallen zahlenmäßig wesentlich geringer aus. In diesem Diagramm lässt sich zusätzlich feststellen, dass bei keiner Auswertung für Loc 1, Loc 2 und Loc 3 keine Überschneidung mit anderen aufgetreten ist. Für Loc 4 zeigen sich lediglich 4 und für die Untersucher 5 nicht überlappende Ergebnisse. Von den insgesamt 561 möglichen negativen Ergebnissen sind 2 Biopsien weder von der KI noch von den Untersuchern richtig ausgewertet worden.

Gemeinsam zeigen diese beiden Diagramme bei wie vielen einzelnen Patienten sich schlussendlich Untersucher und KI im Ergebnis einig waren. Dabei fällt zusätzlich für die KI auf, dass sich manche Lokalisationen ähnlich sind, so z.B. Loc 1 und 2 sowie Loc 3 und 4.

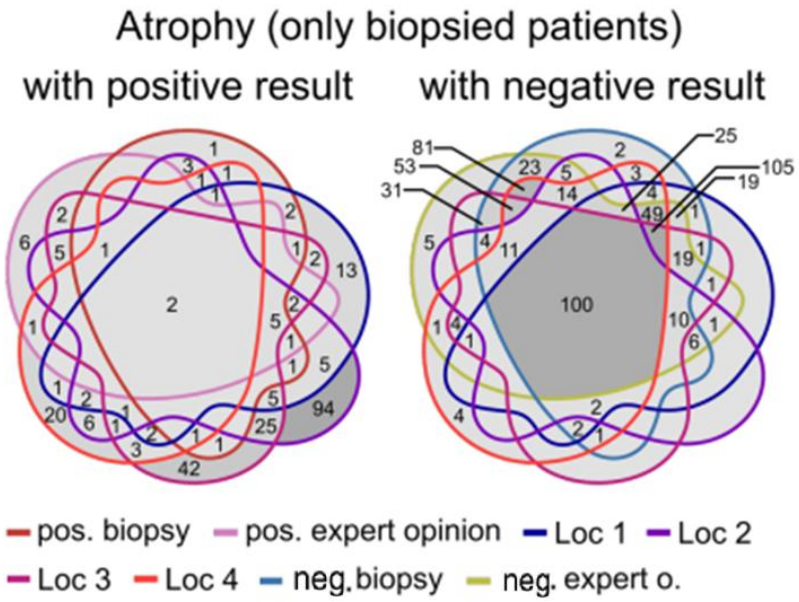


Abbildung 27 - Venn-Diagramme zu präkanzerösen Bedingungen des proximalen Magens

Das linke Diagramm zeigt die Übereinstimmungen der positiven Ergebnisse zwischen den Biopsien, Untersuchern und dem Atrophie-Algorithmus. Das rechte Diagramm zeigt die entsprechenden Übereinstimmungen für die negativen Ergebnisse.

4.7.2 **Vergleich für die eosinophile Ösophagitis**

ABSOLUTE ZAHLEN

Ähnlich wie für die proximale AG wurden die meisten Patienten richtig als positiv und negativ identifiziert und somit wurden nur wenige falsch zugeordnet. Im Falle der EoE werden die absoluten Zahlen mit Einbezug aller Patienten (550) betrachtet (siehe Abb. 28). Die KI gab für Loc 5 21 (4,9%) falsch-positive (FP) auf 432 negative und für Loc 6 30 (6,1%) FP Ergebnisse auf 491 negative Untersuchungen an und somit mehr als die Untersucher mit 7 (1,3%) FP Auswertungen auf 543 negative Patienten. Loc 6 hat 3 (42,9%) falsch-negative (FN) auf 7 positive Patienten erkannt. Im Gegensatz dazu haben die Untersucher nur 1 (14,3%) FN Aussage auf 7 positive Biopsien gehabt. Insgesamt zeigt jedoch Loc 5 mit 0 FN Ergebnissen auf 6 positive Patienten das beste Ergebnis und hat somit keinen betroffenen Patienten übersehen.

Letztendlich gilt für die Auswertung der EoE dieselbe Schlussfolgerung wie für die proximale Atrophie, nämlich dass die Beurteilung stark durch die geringe Anzahl an positiven Fällen, was jedoch der klinischen Realität entspricht, beeinflusst wird und somit nicht ganz einfach abgeleitet werden kann, welche Ergebnisse wirklich zuverlässiger sind. Es lässt sich jedoch zumindest erkennen, dass die Untersucher weniger FP Aussagen getätigt haben als die KI.

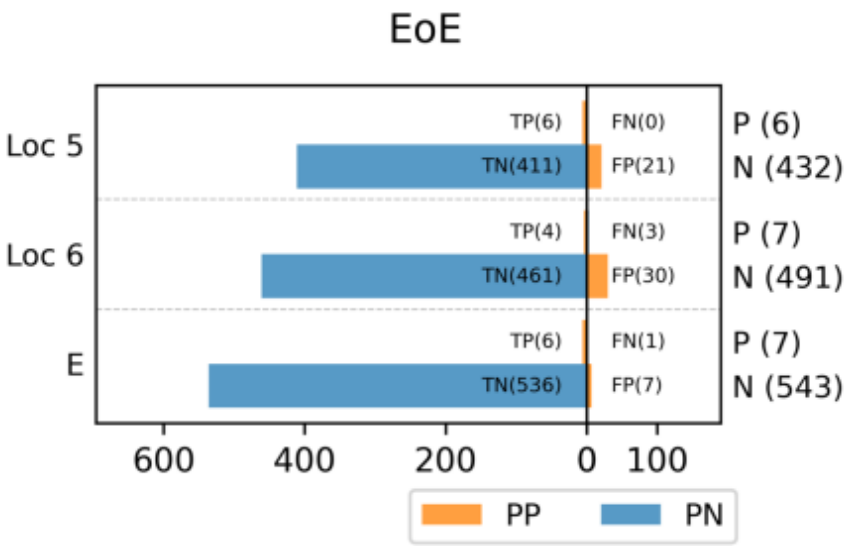


Abbildung 28 - Einschätzung der Endoskopiker und der eosinophilen Ösophagitis-App

Loc 5 und 6 = eosinophiler Ösophagitis-Algorithmus. E = Endoskopiker

Orange: positive Einschätzung des Algorithmus und der Endoskopiker (PP)

Blau: negative Einschätzung des Algorithmus und der Endoskopiker (PN)

TP = richtig-positiv, TN = richtig-negativ, FP = falsch-positiv FN = falsch-negativ

P = Anzahl positiv ausgewerteter Ergebnisse insgesamt (richtig- und falsch-positiv)

N = Anzahl negativ ausgewerteter Ergebnisse insgesamt (richtig- und falsch-negativ)

Auf der linken Seite sind die richtig-positiven und richtig-negativen Ergebnisse des Algorithmus und der Untersucher aufgelistet. Auf der rechten Seite hingegen sind die entsprechenden Falschaussagen aufgelistet.

Statistische Metriken

Die beste *Sensitivität* weist Loc 5 mit 100% auf, gefolgt von den Untersuchern mit 86% und letztlich Loc 6 mit lediglich 57%.

Dahingegen weisen die Untersucher bei Einbezug aller Patienten das bessere Ergebnis für die *Spezifität* mit 99% auf, wohingegen die KI mit Loc 5 nur 95% bzw. mit Loc 6 nur 94% erreicht. Dies lässt sich dadurch erklären, dass die Untersucher weniger FP Ergebnisse (7) als die KI hatten (21 bei Loc 5 bzw. 30 FP bei Loc 6).

Die *Precision* für alle Patienten ist wie im Fall der proximalen AG für beide Lokalisationen des Algorithmus niedrig ausgefallen, dabei zeigt Loc 6 mit 12% das schlechteste Ergebnis, wohingegen Loc 5 22% erreicht. Die Untersucher hingegen haben mit 46% den höchsten Wert. Der

F1-Score fiel für die Untersucher mit 0,6 am besten aus. Die KI erreicht für Loc 5 für alle Patienten 0,36 und für Loc 6 0,2.

Die *Accuracy* ist für beide Auswertungsmethoden bei Einbezug aller Patienten sehr hoch, wobei auch hier die Untersucher mit 99% ein besseres Ergebnis als die KI erzielen. Diese weist für Loc 5 eine Genauigkeit von 95% bzw. für Loc 6 von 93% auf. Bei Angleichung der Kohorten und somit Berechnung der *bACC* hingegen zeigt sich, dass die Untersucher sich auf 92% verschlechtern und Loc 5 der KI sich auf 98% verbessert. Loc 6 hingegen zeigt eine hohe Verschlechterung auf 76%.

LR+ zeigt bei allen Patienten mit einem Faktor von 66,49 für die Untersucher das beste Ergebnis. Für die KI fallen wesentlich niedrigere Werte von 20,57 für Loc 5 bzw. 9,35 für Loc 6 auf.

Für *LR-* hingegen zeigt Loc 5 mit einem Faktor von 0 das beste Ergebnis. Die Untersucher sind mit 0,14 sehr nahe an diesem Ergebnis dran, was in beiden Fällen für den Patienten bedeutet, mit hoher Wahrscheinlichkeit gesund zu sein bei negativem Testergebnis. Lediglich Loc 6 schneidet mit 0,46 für alle Patienten schlecht ab.

Zuletzt ist die geschätzte *AUC* für alle Patienten für Loc 5 mit 0,98 am besten ausgefallen, wobei die Untersucher mit 0,92 einen ähnlich hohen Wert erzielt haben. Lediglich Loc 6 ist etwas niedriger mit 0,76 und weist somit eine schlechtere Testleistung als die anderen beiden vor.

Insgesamt zeigen sich auch bei der EoE unterschiedliche Werte, welche für die Untersucher oder KI besser ausfallen, wodurch sich nicht eindeutig sagen lässt, welche Methode allgemein bessere Ergebnisse erzielt. So hat die KI bei Betrachtung aller Patienten je nach Lokalisation die besten Ergebnisse für *Sensitivität*, *bACC*, *LR-* und *AUC*, wohingegen die Untersucher bessere Ergebnisse in *Spezifität*, *Precision*, *F1-Score*, *Accuracy* und *LR+* erreicht haben. Jedoch muss nach wie vor bedacht werden, dass diese Metriken sich unterschiedlich verhalten bei unbalancierten Kohorten. Insgesamt weist aber Loc 5 für die KI ein deutlich besseres Ergebnis auf als Loc 6, das vergleichbar mit dem Ergebnis der Untersucher ist.

VENN-DIAGRAMME

Auch für die EoE wurden zwei Venn-Diagramme erstellt (siehe Abb. 29), wobei für diese alle 550 Patienten einbezogen worden sind. Ähnlich wie für die proximale AG wurden zur Übersicht zwei getrennte Diagramme für die positiven und negativen Ergebnisse mit den Auswertungen der zwei Lokalisationen der KI, der Untersucher und der Grundwahrheit (Vorliegen einer Biopsie und/oder ösophagealer Dysfunktion) zur Darstellung der überlappenden Ergebnisse erstellt.

Das linke Diagramm mit den positiven Einschätzungen zeigt die zwei größten Überschneidungen mit jeweils 3 Ergebnissen zwischen Grundwahrheit, Untersuchern, Loc 5 und Loc 6 sowie zwischen Grundwahrheit, Untersuchern und Loc 5. Die restlichen vorzufindenden Überlappungen fallen geringer aus, wobei davon zwei Überschneidungen zwischen positiver Untersucheraussage und Loc 6 sowie eine zwischen der Grundwahrheit und Loc 6 stattfinden. Loc 5 zeigt für 19 positive Einschätzungen, Loc 6 für 27 und die Untersucher für 3 positive Einschätzungen keine einzige Überlappung. Jedoch liegt im Gegensatz zu den Ergebnissen zur proximalen AG für kein Verfahren eine unerkannte positive Grundwahrheit vor.

Das rechte Diagramm mit den negativen Ergebnissen zeigt die größte Überschneidung (340) zwischen den beiden Lokalisationen der KI, den Untersuchern und der Grundwahrheit. Außerdem überschneiden sich 119 Ergebnisse zwischen Loc 6, den Untersuchern und der Grundwahrheit bzw. weitere 112 zwischen den Untersuchern und der Grundwahrheit sowie 66 zwischen Loc 5, den Untersuchern und der Grundwahrheit. Die restlichen Überlappungen fallen zahlenmäßig wesentlich geringer aus. Ebenfalls zeigt sich, dass bei einem Ergebnis der Untersucher sowie 3 Einschätzungen der KI von Loc 6 keine einzige Überschneidung mit den anderen stattfindet. Außerdem wurde eine positive Grundwahrheit weder von der KI noch von den Untersuchern richtig vorhergesagt, was auf nicht repräsentatives Bildmaterial hinweisen könnte.

Insgesamt zeigt sich durch die geringe Fallzahl an positiven Patienten im linken Diagramm, dass die meisten Patienten von Loc 5 und den Untersuchern (jeweils 6) richtig erkannt wurden, wohingegen Loc 6 nur bei drei Ergebnissen mit den Untersuchern und der Grundwahrheit übereinstimmte. Zusätzlich zeigt Loc 5 eine wesentlich geringere Anzahl (19) an falsch-positiven (FP) Ergebnissen als Loc 6 (27), wobei die Untersucher insgesamt am wenigsten (3) falsch eingeordnet haben. Somit sind Loc 5 und die Untersucher für die positiven Fälle besser geeignet als Loc 6.

Bei den negativen Ergebnissen zeigt sich, dass alle Auswertungsmethoden sich beim Großteil der Patienten einig sind. Loc 6 hat jedoch fast doppelt so viele Fälle übereinstimmend eingeordnet wie Loc 5, wohingegen diese jedoch keine Fälle ohne Übereinstimmungen vorzeigt. Zusätzlich zeigt sich ein negatives Biopsie-Ergebnis ohne Übereinstimmung von KI und Untersucher, was für eine falsch-negative Histologie sprechen könnte. Die Untersucher zeigen ähnliche Ergebnisse wie Loc 6. Insgesamt lässt sich aus diesem Diagramm schließen, dass in den meisten Fällen vermutlich eine der beiden Lokalisationen ausreichend wäre für ein richtig-negatives Ergebnis, wobei Loc 5 etwas überlegen ist, da diese keine falsch-negativen und weniger falsch-positive Zuordnungen vorweist.

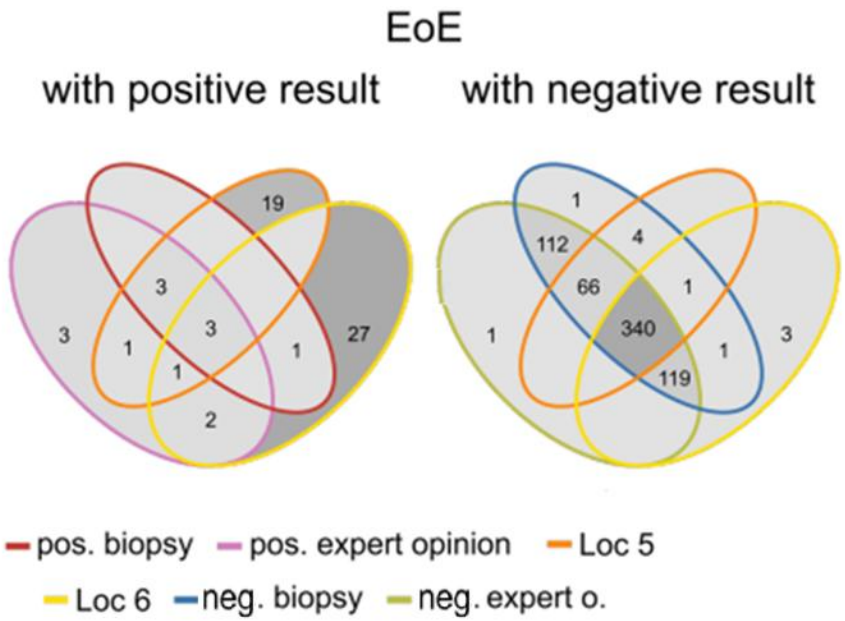


Abbildung 29 - Venn-Diagramme zur eosinophilen Ösophagitis

Das linke Diagramm zeigt die Übereinstimmungen der positiven Ergebnisse zwischen der Grundwahrheit, Untersuchern und dem Atrophie-Algorithmus. Das rechte Diagramm zeigt die entsprechenden Übereinstimmungen für die negativen Ergebnisse.

5 Diskussion

5.1 Kritische Betrachtung der eigenen Untersuchung

Verteilung der negativen und positiven Fälle

Die unterschiedliche Anzahl an positiven sowie negativen Fällen erschwert die statistische Auswertung. So sind die Gruppen der gesunden Patienten für beide Erkrankungen merklich größer als die der Erkrankten, wodurch verschiedene statistische Werte an Aussagekraft verlieren. Dennoch repräsentieren sie gleichzeitig die Realität, da die Wahrscheinlichkeit einen erkrankten Patienten zu untersuchen durch die niedrige Prävalenz wesentlich geringer ist. Dementsprechend sind die absoluten Anzahlen v.a. an falsch-positiven (FP) und falsch-negativen (FN) Ergebnissen aussagekräftiger zur Einschätzung der Auswertung als die relativen Werte der Metriken, da es aus medizinischer Sicht wichtig ist wenig FN bei gleichzeitig akzeptabler Anzahl FP Einschätzungen zu haben, um somit keine positiven Patienten zu übersehen und um zu viel Überdiagnostik bei eigentlich gesunden Patienten zu vermeiden.

Durch die geringe *Prävalenzrate* der Erkrankungen werden sowohl die *Sensitivität*, also ein positiver Test bei erkrankten Patienten, und die *Precision*, die Anzahl der tatsächlich erkrankten Patienten bei positivem Test, stark beeinflusst, da beide die Anzahl der erkrankten Patienten zur Berechnung berücksichtigen und von einer hohen Anzahl an positiven Fällen profitieren würden. Somit erklärt sich durch die geringe Anzahl an nachweislich erkrankten Patienten wieso beide Werte für beide Erkrankungen in sämtlichen Auswertungen dieser Arbeit eher schlecht ausfallen. Im Gegensatz dazu zeigt in allen Berechnungen die *Spezifität*, also ein negativer Test bei gesunden Patienten, hohe Werte. Da dieser statistische Wert zur Berechnung die Anzahl an gesunden Patienten berücksichtigt, wird dieser von einer großen gesunden Kohorte positiv beeinflusst. Weiterhin sinkt auch die *Accuracy* durch diese ungleiche Verteilung der Patientengruppen.

Im Gegensatz zu den oben genannten Werten hat die *Prävalenzrate* auf andere kaum bis keinen Einfluss. Dies trifft auf die *balanced Accuracy* zu, wodurch die ungleiche Verteilung der Fälle auf 50/50 angeglichen wird. Dadurch zeigen sich je nach betrachteter Auswertung oder Lokalisation bessere Werte. Auch die *Likelihood-Quotienten*, welche die Wahrscheinlichkeit für das Vorliegen der Erkrankung anhand des Testergebnisses angeben, werden nicht von der Prävalenz beeinflusst. Somit funktionieren diese zwei statistischen Werte besser bei unbalancierten Gruppengrößen.

Ursprüngliches Training zur Etablierung der Algorithmen

Ein weiteres Problem ist das ursprünglich durchgeführte Training der beiden KI-Algorithmen selbst. So wurden diese jeweils mit einem vorselektierten Datensatz trainiert. Dabei wurde für den Algorithmus der proximalen Atrophie kein Wert auf ausgeglichene Repräsentation der vier hier getesteten Lokalisationen gelegt, sondern allgemein repräsentatives Bildmaterial für Atrophie oder Normalbefunde aus dem proximalen Magenanteil ausgewählt. Bei nachträglicher Betrachtung des Trainingsdatensatzes für die Atrophie-App zeigte sich bei 100 Bildern mit Atrophie folgende Bildverteilung: Loc 1 25 Bilder (25%), Loc 2 25 Bilder (25%), Loc 3 32 Bilder (32%), Loc 4 18 Bilder (18%). Für die Normalbefunde lag bei 100 Bildern folgende Bildverteilung vor: Loc 1 30 Bilder (30%), Loc 2 4 Bilder (4%), Loc 3 40 Bilder (40%), Loc 4 26 Bilder (26%). Die hohe Rate an falsch-positiven (FP) Aussagen für Loc 2 könnte der geringen Anzahl an Bildern mit Normalbefunden im Trainingsdatensatz geschuldet sein. Dies führt wahrscheinlich dazu, dass der Algorithmus Bilder in Nahaufnahme fälschlicherweise als Atrophie wertet, da Bilder in Nahaufnahme aus Loc 2 überwiegend in der Atrophie-Kohorte repräsentiert waren. Entsprechend zeigt Loc 2 aufgrund dieses Umstandes die niedrigste falsch-negativ (FN) Rate.

Aufgrund dieser Ergebnisse muss angenommen werden, dass der Atrophie-Algorithmus für Loc 2 nicht ausreichend trainiert wurde. Somit für Bilder aus dieser Lokalisation keine verlässliche Aussage mithilfe des Algorithmus getroffen werden kann. Dies trifft somit auch für alle Kombinationen zu, die diese Lokalisation miteinbeziehen.

Für den Algorithmus der eosinophilen Ösophagitis spielt das Problem der verschiedenen Lokalisationen kaum eine Rolle, da Bilder aus den Loc 5 und 6 sehr ähnlich sind. In der Auswertung zeigte Loc 5 weniger FP als auch falsch-negative (FN) Ergebnisse als Loc 6. Die kombinierte Auswertung der beiden Lokalisationen bringt keine Vorteile. Die Unterschiede in den Ergebnissen zwischen Loc 5 und Loc 6 dürften in der Qualität der eingesetzten Bilder mit hier nicht so gut erkennbaren Zeichen der EoE liegen, sodass allgemein geschlussfolgert werden kann, dass ein einzelnes Bild aus der unteren Ösophagushälfte ausreichend sein sollte, um ein verlässliches Ergebnis der EoE-App zu erhalten.

Datengewinnung

Weiterhin stellt auch die Datengewinnung selbst ein Problem dar. So haben viele Patienten nicht alle Bilder für die jeweilige App, sondern z.B. nur eines von zwei für die EoE- bzw. weniger als drei von vier für die Atrophie-App. Dies erschwert die definitive Beurteilung. Der wichtigste Faktor scheint aber die Qualität der Bilder selbst im Sinne von schlechter Qualität zu sein, sprich unscharfe oder verschmutzte Bilder, Untersuchungsfehler (z.B. zu viel oder zu wenig Luftinsufflation bei der Endoskopie) sowie Bildgewinnung aus nicht repräsentativen Arealen des Magens oder Ösophagus. Diese Umstände erschweren die richtige Auswertung sowohl für die menschlichen Untersucher als auch für die künstlichen Intelligenzen (siehe Kapitel 4.6).

Beurteilung auf Patienten- und Bildebene

Außerdem stellen für den Vergleich der Ergebnisse zwischen Untersucher und jeweiliger KI die Beurteilungsmethoden ein weiteres Hindernis dar. So beurteilen die Untersucher für die jeweilige Erkrankung auf Patientenebene (ein positiver Gesamteindruck führt nach Gewichtung der Eindrücke zur Bewertung), im Gegensatz zur KI, welche nur einzelne mitunter ggf. auch nicht repräsentative Bilder beurteilen kann, ohne die Fähigkeit aus mehreren Bildern einen Zusammenhang bilden zu können. Dies zeigt sich bei der KI in den Venn-Diagrammen anhand der Anzahl fehlender Übereinstimmungen zwischen den einzelnen Lokalisationen sowie die unterschiedlichen positiven und negativen Aussagen zwischen den unterschiedlichen Lokalisationen. Insgesamt zeigt sich jedoch für die EoE eine größere Übereinstimmung zwischen den Aussagen der KI, was auf die reduzierte Bildanzahl (zwei statt vier wie für die Atrophie-App) und die geringeren anatomischen Unterschiede zwischen Loc 5 und 6 zurückzuführen ist.

Biopsien als Grundwahrheit

Schlussendlich stellen auch die Biopsien ein Problem dar, da das histopathologische Ergebnis als Grundwahrheit für den Vergleich der Ergebnisse der KI und der Untersucher angenommen wird. Einerseits ist die mangelhafte Anzahl der Biopsien für die jeweiligen Algorithmen und den daraus erhaltenen Patientenpool eine Herausforderung. So liegt für den Atrophie-Algorithmus

für verhältnismäßig viele Patienten (583 von 636) eine Biopsie vor, wodurch nur wenige aus der statistischen Auswertung ausgeschlossen werden mussten. Für den EoE-Algorithmus hingegen haben lediglich 204 von 550 Patienten eine Biopsie. Dies lässt sich dadurch erklären, dass im Gegensatz zu Magenbiopsien solche aus dem Ösophagus nur bei begründetem Verdacht (klinische Symptomatik oder auffälliges endoskopisches Erscheinungsbild) entnommen werden. Da die Definition der EoE die ösophageale Dysfunktion mitbeinhaltet können zur Vergrößerung der Kohorte Patienten ohne solche als negativ gewertet werden woraus für die Auswertung des EoE Algorithmus zwei Kohorten resultieren (alle Patienten vs. nur biopsierte Patienten). Interessanterweise fallen die Ergebnisse für beide Auswertungsgruppen ähnlich aus.

Andererseits stellen jedoch auch die vorhandenen Biopsien selbst ein Problem dar, da diese eventuell aus nicht repräsentativen Arealen entnommen wurden und somit auch bei Vorliegen der Erkrankung negativ ausfallen können. Dies zeigt sich besonders bei Betrachtung der Venn-Diagramme der negativen Ergebnisse, bei welcher für die Atrophie zwei und für die EoE eine negative Biopsie sowohl von der KI als auch von den Untersuchern als positiv beurteilt worden sind. Dies lässt also darauf schließen, dass auch die Biopsie als Grundwahrheit für die Beurteilung nicht fehlerfrei ist und es somit auch vorstellbar ist, dass in diesen Fällen die angenommene Grundwahrheit falsch ist.

5.2 Vergleich der eigenen Ergebnisse des Atrophie-Algorithmus mit den bisher veröffentlichten und anderen aus der Literatur

Vergleich mit der Studie zur proximalen Atrophie von Guimarães et al.

Der gleiche Algorithmus zur Erkennung der präkanzerösen Bedingungen des proximalen Magens wurde 2020 in der Studie von Guimarães P. et al. mit insgesamt 70 Bildern von 35 Patienten getestet, wovon 30 Bilder von 13 Patienten mit verdächtigen Läsionen und 40 Bilder von 22 Patienten ohne stammten [29]. Beim Vergleich der unterschiedlichen Kohorten wird also deutlich, dass die ausgewertete Kohorte zur Atrophie aus dieser Arbeit mit 583 biopsierten Patienten und insgesamt 1.681 Bildern, davon 22 Patienten mit positiver Biopsie und 561 negativen Patienten, weitaus größer ist. Außerdem ist die Verteilung der positiven und negativen Bilder in der Vergleichsstudie gleichmäßiger aufgeteilt und unterliegt nicht so einem starken Ungleichgewicht wie in dieser Arbeit. Zusätzlich wurde in der Studie von Guimarães P. et al. keine Rücksicht auf die Lokalisation der Bilder genommen, wohingegen in dieser Arbeit die Gesamtzahl der analysierten Bilder sich auf vier verschiedene Lokalisationen aufteilt. Somit liegen 392 Bilder für Loc 1, 339 für Loc 2, 446 für Loc 3 und 504 Bilder für Loc 4 vor. Ein weiteres Unterscheidungsmerkmal bei den Kohorten der beiden Studien ist, dass die Bilder und Patienten für die Studie von Guimarães P. et al. anders als für die vorliegende Arbeit gezielt ausgesucht wurden [29].

In der kleineren Studie von Guimarães P. et al. erreichte der Atrophie-Algorithmus eine *Accuracy* von 93%, eine *Sensitivität* von 100% und eine *Spezifität* von 87,5% [29]. In dieser Arbeit erreicht die gleiche KI je nach betrachteter Lokalisation eine *Accuracy* zwischen 57% (Loc 2) und 91% (Loc 1 und 4) und zeigt somit eine etwas schlechtere diagnostische Genauigkeit. Die *Sensitivität* liegt hier zwischen 41% (Loc 4) und 94% (Loc 2) und die *Spezifität* zwischen 55% (Loc 2) und 92% (Loc 4). Somit ist die *Sensitivität* bei allen vier Lokalisationen niedriger als in der vorangegangenen Studie, wohingegen die *Spezifität* für Loc 4 (und Loc 1 mit 91%) etwas höher ist.

Diese Differenz der Werte der KI in dieser Arbeit im Vergleich zur Studie von Guimarães P. et al. zeigt also deutlich den Einfluss von unterschiedlich großen Kohorten und die geringe Anzahl an positiven Fälle auf die Aussagekraft der statistischen Berechnungen. Außerdem führen die unterschiedlich hohen falsch-positiven (FP) und falsch-negativen (FN) Aussagen des Algorithmus je nach Lokalisation zu starken Schwankungen der untersuchten Werte.

Zusätzlich wurde in der Studie von Guimarães P. et al. das Ergebnis der KI mit denen der Untersucher verglichen, welche sowohl in der *Accuracy*, sowie *Sensitivität* als auch in der *Spezifität* 80% erreichten und entsprechend insgesamt niedrigere Werte als der Algorithmus aufwiesen [29]. In der vorliegenden Studie erreichen die Untersucher eine *Accuracy* von 95%, welche somit höher ist als die von der KI und die der Untersucher in der Vorstudie. Die *Sensitivität* hingegen liegt bei 41% und ist somit schlechter als in der Vorstudie, jedoch ähnlich niedrig wie für Loc 4 der KI bzw. schlechter als für Loc 2 der KI. Die *Spezifität* hingegen weist bei den Untersuchern 97% auf und ist somit besser als die KI und die Untersucher in der Vorstudie. Die besseren Ergebnisse der Untersucher sind dadurch begründet, dass in der Voruntersuchung die Endoskopiker wie der Algorithmus Einzelbilder ausgewertet haben. Im Gegensatz dazu erfolgte in der hier vorliegenden Arbeit die Beurteilung nach eigener Durchführung einer kompletten ÖGD in Kenntnis von Vorbefunden und Laborwerten.

Vergleich mit anderen KI-basierten Studien zu präkanzerösen Bedingungen des proximalen Magens

Eine vergleichbare Arbeit ist die 2020 veröffentlichte Studie von Zhang Y. et al, in welcher retrospektiv 5.470 Bilder von 1.699 unterschiedlichen Patienten, davon 803 Frauen, 892 Männer und 4 divers, aus dem Magenantrum mittels einer künstlichen Intelligenz ausgewertet wurden, wobei das histopathologische Ergebnis als Grundwahrheit angenommen wurde. Von diesen 5.470 Bildern wurden 2.428 Bilder gesunden Patienten zugeordnet. Die restlichen 3.042 Bilder zeigten unterschiedliche Schweregrade einer atrophischen Gastritis, wovon 1.458 ein mildes, 1.348 ein mäßiges, 38 ein schweres Stadium und 198 Bilder ein nicht einzuordnenden Schweregrad zeigten [102]. Somit ist die verwendete Kohorte in der Studie größer als in der vorliegenden Arbeit, jedoch wurde in der Studie immer die gleiche Lokalisation des Magens genutzt statt vier verschiedene.

Von den 5.470 Bildern wurden 70%, also 3.829 Bilder, zufällig für das Training des Algorithmus genutzt, wobei eine fünffache Kreuzvalidierung (Five-fold Cross-validation) durchgeführt wurde [102]. Dies bedeutet, dass der entsprechende Datensatz in fünf gleich große Anteile aufgeteilt wurde und der Algorithmus anschließend insgesamt fünfmal trainiert wird. Dabei sind bei jedem Training vier Datensätze als Trainingsmaterial und einer als Validierungsmaterial vorhanden und wird insgesamt fünfmal mit unterschiedlichem Validierungsdatensatz durchgeführt. Die genutzte KI in unserer Arbeit wurde nicht durch eine entsprechend große Anzahl an Bildern und nicht durch diese Methode trainiert [29].

Die restlichen 30% (1.641 Bilder) wurden als reiner Testdatensatz aufgespart, diese Bilder wurden also zu keinem Moment vom Algorithmus zum Trainieren genutzt. Als Ergebnis zeigen sich anschließend eine *Accuracy*, eine *Sensitivität* und eine *Spezifität* von durchschnittlich 94%. Bei Betrachtung der Schweregrade der Atrophie zeigt sich für ein mildes Stadium eine *Accuracy* von 93%, für ein mäßiges Stadium von 95% und für ein schweres Stadium von 99% [102]. In der vorliegenden Arbeit wurde also eine ähnliche Anzahl an Bildern (1.681 Bilder) zur Validierung genutzt, wobei die *Accuracy* ihr bestes Ergebnis für Loc 1 und 4 mit 91% zeigt und somit ähnlich hoch ist. Die *Sensitivität* zeigt sich mit bis zu 94% für Loc 4 und die *Spezifität* mit 92% ebenfalls für Loc 4 und sind ebenfalls im ähnlichen Bereich wie in der Vergleichsstudie.

Zusätzlich wurden in der Studie von Zhang Y. et al ebenfalls die Ergebnisse des Algorithmus mit denen der Untersucher verglichen. Dabei wurde bei den Untersuchern anhand des Erfahrungsgrades zwischen Anfängern (zwei) und Fortgeschrittenen (drei) unterschieden. Die Anfänger zeigten in allen drei Werten (*Accuracy*, *Sensitivität* und *Spezifität*) ca. 60%, wohingegen die Fortgeschrittenen in den drei genannten Metriken Werte von ca. 90% zeigten. Somit zeigte sich ein deutlicher Unterschied zwischen den Untersuchern, jedoch auch zwischen den Anfängern und dem Algorithmus [102]. In der vorliegenden Arbeit wurde keine Differenzierung zwischen den Untersuchern durchgeführt. Es zeigt sich eine ähnlich hohe *Accuracy* mit 95%, eine bessere *Spezifität* mit 97%, jedoch eine schlechtere *Sensitivität* mit 41%.

Die großen Unterschiede zwischen der vorliegenden Arbeit und der Studie von Zhang Y. et al [102] liegen also vor allem in der Kohortengröße zum Trainieren sowie in der betrachteten Lokalisation (hier vier verschiedene im proximalen Magen vs. nur Antrum). Auch der Datengewinn unterscheidet sich dadurch, dass die Vergleichsstudie retrospektiv durchgeführt wurde, wobei in der vorliegenden Arbeit die Patienten und Bilder prospektiv generiert wurden. Ebenfalls wurde in dieser Arbeit keine Rücksicht auf das Stadium der atrophischen Gastritis und die Expertise der Untersucher genommen. Die Vergleichsstudie macht keine Angaben zu den absoluten Zahlen bezüglich falsch-positiver (FP) und falsch-negativer (FN) Ergebnisse, wodurch diese nicht verglichen werden können.

Mit unserer Studie vergleichbare prospektive Ansätze für Algorithmen zur Erkennung der proximal betonten Atrophie existieren nicht. Die große Mehrzahl der Arbeiten zur Atrophie stammen aus Fernost und beziehen sich auf die distale, zumeist Helicobacter-assoziierte Atrophie. Ein Beispiel ist die Studie von Nakashima H. et al. aus dem Jahr 2018, in welcher eine KI anhand von Bildern aus der kleinen Magenkurvatur des Korpus in drei unterschiedlichen Aufnahmetechniken eine HP-Infektion aufgrund von Merkmalen wie z.B. atrophischer Gastritis und intestinaler Metaplasie erkennen und farblich markieren soll. Es standen 222 Patienten, davon 105 HP-positiv getestet anhand des spezifischen IgG-Antikörpers, mit insgesamt 2.214

Bildern zur Verfügung. Die KI wurde im Voraus mit insgesamt 1.944 Bildern trainiert (je Aufnahmetechnik insgesamt 648 Bilder durch Drehungen um 90°). Anschließend wurde die KI mit 180 ausgewählten Bildern (60 Bilder je Aufnahmetechnik, davon jeweils 30 mit positivem und 30 mit negativem Befund) getestet. Je nach Aufnahmetechnik variiert die Sensitivität zwischen 66,7% und 96,7%, die Spezifität liegt zwischen 60% und 86,7%. Zusätzlich wurde die AUC angegeben, welche je nach Aufnahmetechnik zwischen 0,66 und 0,96 lag [63]. Somit unterscheidet sich diese Arbeit von der vorliegenden darin, dass auf ausgeglichene Fallzahlen geachtet wurden. Zudem wurde nur eine Magenregion von der KI evaluiert, statt mehrere Bilder aus unterschiedlichen Magenanteilen. Weiterhin wurden verschiedene Aufnahmetechniken untersucht statt nur eine. Es werden keine Prozentangaben zur Accuracy gemacht und keine absoluten Zahlen bezüglich FP und FN angegeben. Auch wird kein direkter Vergleich mit der Performance der Untersucher angestellt.

Eine weitere relevante Arbeit ist die von Shichijo S. et al. von 2019. Auch hier soll eine KI eine HP-Infektion anhand von Bildmerkmalen wie Atrophie, diffuse Rötung oder mukosale Schwellung erkennen. Insgesamt wurden von 735 positiven und 1.015 negativen Patienten 32.208 Bilder sowohl aus dem proximalen wie auch distalen Magen ausgewählt, um diesen Algorithmus zu trainieren. Anschließend wurden weitere 11.481 Bilder von 397 Patienten, davon 72 positive und 325 negative Patienten, zum Testen ausgesucht. Je nach Bewertungsmethode der KI (alle Bilder auf einmal oder eingeteilt nach Lokalisation im Magen) variiert die Accuracy zwischen 83,1% und 87,7%, die Sensitivität von 81,9% und 88,9% sowie die Spezifität von 83,4% und 87,4%. Im Vergleich lag die durchschnittliche Accuracy der Untersucher bei 82,4 %, die durchschnittliche Sensitivität bei 79% und die durchschnittliche Spezifität bei 83,2% [86]. Hier wurde also ähnlich wie in der vorliegenden Arbeit kein großer Wert auf balancierte Fallzahlen gelegt. Weiterhin wurden in der Studie sowohl die distalen wie auch die proximalen Magenanteile untersucht und wie in den bereits vorherigen vorgestellten Arbeiten keine Angaben zu absoluten FP und FN Zahlen gemacht.

Meist stellt in Arbeiten zur KI in der Endoskopie die genutzte Validierungskohorte einen Teil der Gesamtkohorte dar. Dadurch zeigt die genutzte künstliche Intelligenz oft keine unbeeinflusste Auswertung und entspricht keiner Performance unter reellen Bedingungen. Entsprechend müssen die KIs durch prospektive Daten geprüft werden, um ihre Performance und Zuverlässigkeit realistisch einschätzen zu können.

5.3 Vergleich der eigenen Ergebnisse des eosinophilen Ösophagitis-Algorithmus mit den bisher veröffentlichten und anderen aus der Literatur

Vergleich mit der Studie zur eosinophilen Ösophagitis von Guimarães et al

Guimarães P. et al. testete in der Studie von 2022 den auch hier verwendeten Algorithmus zur Detektion der eosinophilen Ösophagitis mit 78 neuen Bildern von 31 Patienten. 7 Patienten mit 26 Bilder hatten eine nachgewiesene EoE, 16 Patienten mit 25 Bildern eine Candidiasis und 8 Patienten mit 27 Bildern einen Normalbefund [30]. Bei Vernachlässigung der Patienten mit Candidiasis wird deutlich, dass ähnlich wie in der Studie zur Atrophie die Anzahl an gesunden zu an einer EoE erkrankten Patienten und deren Bilder ähnlich verteilt ist. Auch in diesem Fall ist die Kohorte der vorliegenden Studie bei Betrachtung der Patienten mit vorliegender Biopsie (204 von möglichen 550) mit insgesamt 351 Bildern größer als in der Vorstudie. Gleichzeitig besteht bei 7 Patienten mit positiver Biopsie und 197 gesunden Patienten auch hier eine ungleichmäßigere Verteilung. Ähnlich wie in den Studien zur Atrophie wurden in der Studie von Guimarães P. et al. die untersuchten Bilder gezielt ausgesucht und es wurde keine Rücksicht auf die Lokalisation genommen, wohingegen in der vorliegenden Arbeit 169 Bilder für Loc 5 und 182 für Loc 6 vorliegen.

In der Vergleichsstudie mit balancierter kleinerer Kohorte erreichte die KI eine *Accuracy* von 91,5%, eine *Sensitivität* von 87% und eine *Spezifität* von 94% [30]. In der vorliegenden Studie erreicht die identische KI für Loc 5 eine *Accuracy* von 94% und ist somit sogar etwas besser, wohingegen Loc 6 mit 90% etwas schlechter ist. Die *Sensitivität* liegt für Loc 5 bei 100% und ist entsprechend besser, wohingegen Loc 6 lediglich 57% hat. Die *Spezifität* beträgt für Loc 5 94% und für Loc 6 91%.

Auch hier zeigt sich bei den Unterschieden zwischen den Werten der Einfluss durch die unbalancierte größere Kohorte und die sehr geringe Fallzahl positiver Fälle. Zusätzlich ist die Differenz der *Sensitivität* zwischen den beiden Lokalisationen dadurch zu erklären, dass Loc 5 6 von 6 erkrankte und Loc 6 nur 4 von 7 erkrankte Patienten richtig erkannt hat.

In der vorangegangenen Studie von Guimarães P. et al. wurde die *Accuracy* der Untersucher im Vergleich zur KI mit 83% angegeben [30]. In dieser Arbeit erreichen die Untersucher eine höhere diagnostische Genauigkeit von 96%. Die besseren Ergebnisse der Untersucher sind

wiederum dadurch begründet, dass in der hier vorliegenden Arbeit die Untersucher nach vollständiger ÖGD auf Patientenebene entschieden haben, und nicht losgelöstes Bildmaterial beurteilen mussten.

Vergleich mit anderen KI-basierten Studien zur eosinophilen Ösophagitis

Die Studie von Römmele C. et al aus dem Jahr 2022 untersucht ebenfalls mittels alternativem Algorithmus (AI-EoE) die Erkennung von eosinophiler Ösophagitis (EoE). Die Daten wurden retrospektiv durch zwei Fachärzte der Inneren Medizin und Gastroenterologie anhand der histopathologischen Ergebnisse als Grundwahrheit ausgesucht und stammten von Patienten, die zwischen Juli 2010 und Mai 2020 entsprechende Untersuchungen und Ergebnisse vorweisen konnten. Dabei wurden sowohl gesunde als auch nachweislich erkrankte Bilder unabhängig der Lokalisation im Ösophagus ausgesucht. Insgesamt wurden 1.272 Bilder, davon 401 Bilder von 61 Patienten mit nachweislicher EoE und 871 Bilder von 393 Patienten mit Normalbefund genutzt, um die KI mittels fünffacher Kreuzvalidierung zu trainieren und zu testen. Zusätzlich wurde überprüft, ob die KI durch die Erweiterung der eigenständigen Analyse der EREFS-Klassifizierung optimiert werden kann (AI-EoE-EREFS) [73]. Wichtigste Unterschiede unserer Arbeit sind somit einerseits die prospektive Vorgehensweise sowie die Unterscheidung zwischen Bildern aus dem proximalen und distalen Ösophagus. Der ursprüngliche Trainingsdatensatz für unsere KI mit 406 Bildern von 103 Patienten, davon 164 Bilder von 25 Patienten mit histologisch nachgewiesener EoE, 107 von 46 Patienten mit ösophagealer Candidiasis und 135 Bilder von 32 Patienten mit Normalbefunden ist kleiner gewesen als in der Vergleichsstudie [30].

Anschließend wurde die KI in der Studie von Römmele C. et al mit einem externen Datensatz von insgesamt 200 ausgesuchten Bildern, davon 100 Bilder mit Erkrankung und 100 Normalbefunde, geprüft. Zunächst wurde die KI ohne Erweiterung durch die EREFS-Klassifikation (AI-EoE) und anschließend nochmals mit (AI-EoE-EREFS) geprüft. Dabei zeigte die AI-EoE sowohl für *Accuracy* als auch *Sensitivität* und *Spezifität* Werte von 93% [73]. In der vorliegenden Arbeit wurden 204 von möglichen 550 Patienten mit Biopsie als Grundwahrheit mit 351 Bildern berücksichtigt. Dabei zeigt unsere KI je nach Lokalisation im prospektiven Datensatz ähnliche Werte mit einer *Accuracy* bis zu 94%, eine *Sensitivität* bis zu 100% und eine *Spezifität* bis zu 5 94%.

Auch in der Arbeit von Römmele C. et al. wurden die Ergebnisse der KI mit denen von menschlichen Untersuchern verglichen. Dafür entschied man sich für zwei Anfänger, zwei Fortgeschrittene und zwei fertig ausgebildete Fachärzte, welche in zwei Gruppen aufgeteilt wurden,

wodurch sich in jeder Gruppe jeweils einer der entsprechenden Erfahrungskategorie befand. Sie werteten den externen Datensatz von 200 Bildern aus. Dabei sollte eine Gruppe ähnlich wie die KI AI-EoE die 200 Bilder ohne Beachtung der EREFS-Klassifizierung beurteilen. Dabei erreichte der Anfänger eine Accuracy von 77%, der Fortgeschrittene von 92% und der Facharzt von 97%. Die Sensitivität lag beim Anfänger bei 56%, beim Fortgeschrittenen bei 87% und beim Facharzt bei 96%. Beim Vergleich der Spezifität untereinander zeigte sich beim Anfänger und beim Fortgeschrittenen ein Wert von 97%, wohingegen der Facharzt bei 99% lag [73]. In der vorliegenden Studie wurde keine Differenzierung der Auswertung anhand der Erfahrung der Untersucher gemacht. Die Untersucher erreichen eine *Accuracy* von 96%, eine *Sensitivität* von 86% und eine *Spezifität* von 96% und somit durchschnittliche Werte, die vergleichbar mit denen der fortgeschrittenen Untersucher waren. Es muss aber wiederum betont werden, dass im Vergleich zur Arbeit von Römmele C et al. in unserer Arbeit eine Beurteilung auf Patientenebene erfolgte und keine Auswertung von Einzelbildern.

Insgesamt unterscheidet sich die Herangehensweise beider Studien darin, dass die vorliegende Arbeit prospektiv durchgeführt wurde und sich damit der Anteil der Fälle auf einem deutlich niedrigeren Niveau bewegt, das aber der klinischen Realität in einem Zuweiserzentrum entspricht. Bei den Untersuchern wurde hier keine Berücksichtigung der bisherigen Erfahrungen durchgeführt und die Auswertung erfolgte nach Durchführung einer kompletten ÖGD. Außerdem hat die Vergleichsstudie von Römmele C. et al [73] keine Angaben über die absoluten Zahlen der falsch-positiven (FP) und falsch-negativen (FN) Auswertungen gemacht, wodurch ein solcher Vergleich nicht möglich ist.

Auch bei dieser Erkrankung lässt sich keine weitere Studie zur Detektion einer eosinophilen Ösophagitis mittels endoskopischen Bilder durch eine künstliche Intelligenz finden. Es gibt jedoch anders als zu den Studien des Magens für ösophageale Erkrankungen viele prospektive Studien. Die Studie von de Groof AJ et al aus dem Jahr 2020 untersuchte prospektiv die Früherkennung einer Barrett-Neoplasie in der Endoskopie mittels KI [18]. Eine andere prospektive Studie ist beispielsweise die Studie von Gulati et al. aus dem Jahr 2019, welche die Diagnose einer GERD (gastroösophageale Refluxkrankheit) mittels KI durch endoskopische Bilder überprüfte [32].

5.4 Vorteile durch Nutzung künstlicher Intelligenz in der Medizin

Effizienz und Zeitersparnis

Messmann H. postulierte bereits 2022, dass sich durch die Nutzung von KI in der Endoskopie die Untersuchungszeit gegebenenfalls verlängern könnte [56]. Erklären würde sich dies dadurch, dass die Untersucher bei der Diagnostik zu sehr auf optimale Gegebenheiten für den Algorithmus wie beispielsweise Bildqualität und Strukturdarstellung achten würden. Ein weiterer Grund wäre die Einschätzung positiv detektierter Befunde durch die künstliche Intelligenz auf ihre Richtigkeit [56]. Die Studie von Hassan C. et al von 2020 zeigt, dass sich die durchschnittliche Untersuchungszeit pro Koloskopie durch 30-60 falsch-positive Befunde durch den Algorithmus um ca. eine Minute verlängert [37]. Zusätzlich zeigt sich, dass die meisten falsch-positiven Befunde auf Unregelmäßigkeiten oder Verschmutzungen der Kolonschleimhaut zurückzuführen sind [37]. Durch regelmäßige Verwendung der Algorithmen durch die Untersucher und die daraus entstehende Vertrautheit mit der Arbeitsweise der künstlichen Intelligenz und die entsprechend bessere Einschätzbarkeit ihrer Aussagen, würde sich diese geringe Untersuchungsverlängerung reduzieren oder gänzlich verhindern lassen [56].

Einige bereits kommerziell erhältliche künstliche Intelligenzen können nachweislich die Arbeitsschritte bei der Endoskopie beschleunigen und somit v.a die Befunderstellung vereinfachen. So kann der Algorithmus von Odin Vision© in Echtzeit den Sauberkeitsgrad der Schleimhaut erkennen und gibt dem Untersucher an, wie viel Prozent der Schleimhaut sichtbar ist. Weiterhin können anatomische Strukturen, welche standardmäßig fotodokumentiert werden müssen oder zur Orientierung auf den Bildern sichtbar sein müssen, automatisch markiert werden [56]. Andere Algorithmen wie der von Olympus© kann während der Gastroskopie erkennen, an welcher Lokalisation das Bild gemacht wurde und erkennt gleichzeitig die Qualität davon. Diese Informationen werden dann automatisch in den Untersuchungsbefund eingefügt, so dass ein manuelles Eintragen nicht mehr notwendig ist [56].

Insgesamt wird also deutlich, dass sich ein effektives Zeitersparnis erst nach Erlernen der Arbeitsweise durch die Algorithmen etabliert, da die Untersucher lernen müssen die Ergebnisse richtig einzuschätzen. Weiterhin kann aber die Nachbearbeitungszeit durch die Durchsicht der Bilder und die Befunderstellung reduziert und effizienter gestaltet werden. Um Algorithmen in der Praxis nutzen zu können, müssen diese in ihrer Funktionsweise so optimiert

werden, dass sie unter Echtzeit während der Untersuchung auswerten können. Dies ist sicherlich ein großer Nachteil und eine Limitation der von uns etablierten und in dieser Arbeit prospektiv evaluierten Algorithmen.

Fehlerreduktion und gezielte Probeentnahmen

Eine weitere Überlegung von Messmann H. war, dass durch die Nutzung der Algorithmen die Untersucher abgelenkt werden könnten. Dies würde zu vermehrten Übersehen möglicher Auffälligkeiten und zu vermehrten endoskopiebedingten Komplikationen führen [56].

Eine 2021 publizierte Metaanalasye von Hassan C. kann diese Überlegung widerlegen. Dabei wurden die Adenomdetektionsrate (ADR) und die Anzahl an Adenomen pro Koloskopie (APC) in Abhängigkeit ihrer Größe bei Koloskopien zwischen Kontrollgruppen und Gruppen mit Verwendung von Algorithmen verglichen. Es zeigt sich, dass in der KI-Gruppe die ADR mit 36,6% höher ist, wohingegen die Kontrollgruppe 25,2% erreicht. Insgesamt erreicht die Gruppe mit den Algorithmen auch bessere Ergebnisse in der APC. Sie detektieren sowohl mehr Adenome < 10mm, als auch bei Größen > 10mm, sowohl in proximalen als auch distalen Darmabschnitten. Bezüglich der Morphologie wurden in der Algorithmus-Gruppe mehr flache Adenome erkannt [38]. Somit können mehr klinisch relevante Läsionen erkannt werden, wodurch sich anfänglich die Untersuchungszeit leicht erhöhen kann. Daraus entsteht jedoch der Vorteil, dass mehr gezielte Probeentnahmen durchgeführt werden und somit sich insgesamt die Anzahl an Biopsien reduziert [57].

Objektivität und Unterstützung bei Entscheidungen

Bei der Verwendung von künstlichen Intelligenzen muss bedacht werden, dass sie lediglich Assistenzsysteme darstellen [22,56]. Die eingegebenen Informationen werden rein rechnerisch anhand bisher erlernter Musteranalyse verarbeitet, um somit eine objektive und unbeeinflusste Auswertung auszugeben. Sie fungieren also als eine Art Zweitbeurteiler für die Untersuchung [56]. Dies kann vor allem für junge Ärzte in Weiterbildung mit wenig Erfahrung eine Hilfestellung sein, jedoch immer unter der Beachtung, dass eine kritische Hinterfragung der Ergebnisse notwendig ist. Es muss auch bedacht werden, dass der Algorithmus gerade aufgrund seiner Objektivität keine Rücksicht auf mögliche Vorerkrankungen oder sonstige patientenindividuelle Faktoren und deren Einfluss auf die analysierten Daten nimmt [10].

5.5 Grenzen bei der Verwendung von künstlicher Intelligenz in der Medizin

Abhängigkeit von Training und Datensätzen

Die potenzielle Leistungsfähigkeit einer KI ist von zwei Hauptkriterien abhängig, nämlich einerseits von der Menge und andererseits von der qualitativen Diversität der Trainingsdaten [73].

Ein Algorithmus benötigt eine große Menge an unterschiedlichen Daten, um von einem bestimmten Krankheitsbild alle möglichen Darstellungsmöglichkeiten in angemessenem Ausmaß zu erlernen [73]. Er benötigt also unterschiedliche Bildwinkel des gleichen Abschnittes, um diesen weiterhin als solchen erkennen zu können. Weiterhin benötigt er aus dem gleichen Organ verschiedene Lokalisationen, damit die Auffälligkeit nicht einer bestimmten Struktur zugeordnet wird. Optimalerweise sieht er das Trainingsmaterial sowohl von weiblichen wie auch männlichen, von jungen und alten, sowie Menschen unterschiedlicher ethnischer Herkunft.

Weiterhin spielt die qualitative Diversität der Trainingsdaten eine Rolle [73]. So wird der Algorithmus zum Erlernen von krankheitsspezifischen Merkmalen meist nur mit qualitativ hochwertigen Bildern, sprich scharf eingestellte Bilder ohne Verschmutzungen und ohne Verzerrungen, trainiert. Es werden wichtige anatomische Schnittpunkte vollständig abgebildet und störende Fremdkörper entfernt. Dies kann dann in der klinischen Realität zu einer schlechteren Performance führen, wie unsere Daten klar zeigen. Auch die Art der genutzten Daten hat Einfluss auf die Entwicklung der künstlichen Intelligenz. So beschreibt auch bereits Messmann H., dass ein Algorithmus, welcher endoskopische Bilder in einer bestimmten Aufnahmetechnik zu interpretieren gelernt hat, nicht ohne weiteres Training Bilder in einer anderen Technik zuverlässig beurteilen könnte, obschon die untersuchten Merkmale die gleichen sind [56]. Gleiches gilt für die Nutzung bei Live-Untersuchungen, ein Algorithmus wird keine zuverlässigen Ergebnisse liefern, sofern er nicht in genau dieser Untersuchungstechnik trainiert wurde.

Gleichzeitig entstehen aus diesen Kriterien jedoch auch die Probleme. Eine künstliche Intelligenz, die unter perfekten Bedingungen entwickelt wurde, kann auch nur perfekte Daten zuverlässig beurteilen und entspricht nicht den reellen Untersuchungsbedingungen. Dies zeigt sich auch in dieser Arbeit am deutlichsten bei dem Algorithmus für die atrophische Gastritis im proximalen Magenabschnitt. Es wird deutlich, dass beim Training nur qualitativ hochwertige Bilder genutzt wurden und nicht auf ausgeglichene Repräsentation der Magenanteile geachtet

wurde. Somit ergeben sich die hohen Zahlen an falsch-positiven und falsch-negativen Befunde bei wenig trainierten Lokalisationen (siehe hohe Anzahl falsch-positiver Bewertungen aus Loc 2 aufgrund unterrepräsentierter Normalbefunde in dieser Lokalisation im Trainingsdatensatz) und bei verschmutzten oder gar unscharfen Bildern.

Klinische Validierung

Ein weiteres Problem, welches aufgrund des Vorgehens bei den Trainingsdatensätzen und auch durch zu geringe Mengen an Daten entsteht, ist die Art der klinischen Validierung. Die meisten klinischen Validierungen werden an retrospektiv zusammengestellten Kollektiven durchgeführt, wodurch also die untersuchte Kohorte und deren Material ausgesucht wurde und somit nicht den reellen Bedingungen entsprechen [22]. Zwar erzielen die Algorithmen in diesen Fällen gute Ergebnisse, würden jedoch bei einer prospektiven klinischen Validierungen automatisch schlechtere Ergebnisse erzielen, da dann auch beispielsweise wie im medizinischen Alltag unscharfe oder verschmutzten Bilder aus unterschiedlichen Winkeln beurteilt werden müssten, welche nicht vom Algorithmus erlernt wurden. So versucht dieser zwar anhand bereits erlernter Merkmale das Bild zu analysieren, kann jedoch manche Eigenschaften wie z.B. bisher nicht gesehene anatomische Strukturen oder Sekrete und Blut nicht richtig einordnen. Das daraus entstehende Problem sind falsch-positive und falsch-negative Aussagen durch eine erlernte verallgemeinerte Beurteilung und somit eine gewisse Unzuverlässigkeit für den alltäglichen Einsatz. Diese Schwächen werden in unserer Arbeit vor allem für die Magen-KI aufgezeigt.

Insgesamt mangelt es dem Algorithmus also an Kenntnissen über die Varianz der Darstellung, sowohl der Bilder als auch der Erkrankung, unter realistischen unvoreingenommenen Bedingungen [56,73]. Um die eigentliche Zuverlässigkeit im Alltag also besser beurteilen zu können, müssen wie in unserer Arbeit mehr prospektive klinische Validierungen, die den endoskopischen Alltag imitieren, durchgeführt werden [22].

Transparenz

Wie bereits in dieser Arbeit anhand der Fehleranalyse festgestellt, kann es schwierig sein, die Ergebnisse der künstlichen Intelligenz nachzuvollziehen. Durch ihr eigenständiges Lernen und Entscheiden ist es oft nicht klar ersichtlich, nach welchen Kriterien der Algorithmus seine Entscheidungen trifft. Dies kann zwar durch die selektive Auswahl an zur Verfügung gestelltem

Material zum Teil beeinflusst werden, jedoch funktioniert die künstliche Intelligenz nach dem „Black-Box“-Prinzip, es sind also nur die Eingabe- und Ausgabedaten bekannt, ohne zu wissen, was in den Zwischenebenen passiert [30,94]. So kann man oft nur anhand Ähnlichkeiten der Bilder mutmaßen wieso die falsch-positiven oder falsch-negativen Ergebnisse entstehen und so die Diagnose des Algorithmus entsprechend ausfällt.

Im medizinischen Alltag ist diese Intransparenz ein Problem, da somit weder die Fähigkeiten des Untersuchers verbessert werden können noch durch die fehlende Nachvollziehbarkeit für die Anwender die Akzeptanz für die algorithmische Aussage gefördert wird [94].

Kosten

Die Studie von Mori Y. et al aus dem Jahr 2020 zeigt, dass sich die Kosten für die eigentliche Untersuchung durch Reduktion der Biopsien um bis zu 19% senken lassen [59]. Es muss jedoch bedacht werden, dass die Anschaffung sowie Wartung des benötigten technischen Equipments bereits Kosten mit sich bringt. Auch die Lizenz zur Nutzung des eigentlichen Algorithmus sowie die ständige Verbesserung davon und der menschliche Kundensupport mit entsprechendem Fachpersonal bei aufkommenden Problemen oder Fehlermeldungen bringen dauerhafte und höhere Kosten mit sich.

Verantwortung

Die genutzten Algorithmen können keine Entscheidungen übernehmen, welche über ihre erlernte Bildanalyse hinaus gehen und somit auch keine Verantwortung übernehmen. Entsprechend liegt die Verantwortung beim Anwender, welcher selbst kritisch die Auswertung beurteilen und die daraus weiteren (Therapie-)Schritte entscheiden muss [22,56].

Eine Idealvorstellung wäre ein automatisiertes objektives System, um schnelle und effiziente Entscheidungen in Echtzeit zu treffen [10]. Realistisch wird dies jedoch nicht sein, da jeder Patient und jeder Krankheitsverlauf individuell betrachtet werden muss, um die für ihn optimale Diagnostik und Therapie durchzuführen. Die wenigsten Patienten haben einen Lehrbuch-Verlauf, da oft auch die Vorerkrankungen Einfluss haben, welche durch einen Algorithmus nicht beachtet werden können, wodurch ein gewisser erfahrener Menschenverstand von Fachärzten weiterhin notwendig sein wird [10].

5.6 Schlussfolgerungen / Konklusionen

Künstliche Intelligenzen sind bereits heute in der Endoskopie zur Polypen- und Adenomdetektion häufig genutzte Assistenzsysteme. Es muss grundsätzlich aber beachtet werden, dass diese Systeme nur so gut sind wie ihr ursprüngliches Training und sie somit mit falsch-positiven oder falsch-negativen Einschätzungen im Alltag einhergehen. Auch können die Algorithmen nicht den Einfluss von Vorerkrankungen auf die verschiedenen Krankheitsbilder beurteilen. Somit liegt die endgültige Entscheidung und Verantwortung bezüglich der therapeutischen Konsequenz immer beim Anwender und können nicht abgegeben werden.

Die beiden vorgestellten Algorithmen in dieser Arbeit beziehen sich auf zwei Krankheitsbilder, die bisher wenig Aufmerksamkeit im Hinblick auf KI erhalten haben. Es zeigt sich sowohl beim Algorithmus für die proximale Atrophie als auch für die eosinophile Ösophagitis, dass solche Algorithmen aufgrund von limitierten Daten häufig nicht vielfältig genug trainiert werden können. Wie auch andere künstliche Intelligenzen erreichten beide Algorithmen in vorherigen kleineren klinischen Tests in selektierten Datensätzen bessere Ergebnisse (für die Atrophie eine *Accuracy* von 93%, *Sensitivität* von 100% und eine *Spezifität* von 87,5%; für die EoE eine *Accuracy* von 91,5%, eine *Sensitivität* von 87% und eine *Spezifität* von 94%). Dies liegt daran, dass die in den Publikationen genutzten Validierungskohorten häufig aus retrospektiven Datensätzen zusammengestellt werden. Dies bedingt, dass in den selektierten Datensätzen oft die Verteilung von positiven und negativen Fällen nicht der wahren Krankheitsprävalenz entspricht und suboptimales oder nicht ganz eindeutiges und dadurch erschwert auszuwertendes Bildmaterial nicht berücksichtigt wird. Dies bedingt dann beschönigte Ergebnisse zur Leistungsfähigkeit von Algorithmen. Im Gegensatz dazu wurde diese Arbeit prospektiv durchgeführt und sollte die reellen alltäglichen Bedingungen in der Endoskopie imitieren und damit die wahre Performance ungeschönt abbilden.

Insbesondere der Algorithmus zur Erkennung von präkanzerösen Bedingungen des proximalen Magens offenbarte deutliche Schwächen mit Auftreten zahlreicher falsch-positiver und falsch-negativer Auswertungen (falsch-positiv: Loc 1 33, Loc 2 146, Loc 3 90, Loc 4 37; falsch-negativ: Loc 1 3, Loc 2 1, Loc 3 6, Loc 4 10). Dies ist durch den limitierten Trainingsdatensatz bei einem Krankheitsbild mit großer Bandbreite an endoskopischen Erscheinungsbildern und großem Organ mit variablem Bildwinkeln erklärt. Eine limitierte Performance bei schlechtem Bildmaterial mit offensichtlichen Bildfehlern betrifft menschliche Untersucher wie die KI gleichermaßen. Die schlechte Qualität der ausgewählten Einzelbilder kommt bei der Performance

der menschlichen Untersucher in dieser Arbeit jedoch nicht zum Tragen, da die Beurteilung anhand des Eindruckes bei der Untersuchung erfolgte. Insofern dürfte die wahre Leistungsfähigkeit des Algorithmus bei doch in der Zweitbeurteilung hohem Anteil an Bildmaterial mit signifikanten Limitationen (siehe Kapitel *4. 6 Fehleranalyse*; für Loc 1 36, Loc 2 147, Loc 3 96, Loc 4 47 Bilder), im Falle der Elimination von solchem Bildmaterial besser ausfallen. Bei Verwendung von Bildern aus einer endoskopischen Standardposition (Überblick Corpus; Loc 1) zeigt der Algorithmus unter Realbedingungen durchaus eine zufriedenstellende Performance, die eine Nutzung als unabhängiger Zweitbegutachter ermöglichen könnte (*Accuracy* 91%, *balanced Accuracy* 86%, *Sensitivität* 81%, *Spezifität* 91%). Bei anderen Einstellungen (insbesondere Loc 2) ist die Beurteilung des Algorithmus aufgrund sehr vieler Fehlbeurteilungen jedoch unbrauchbar (*Accuracy* 57%, *balanced Accuracy* 74%, *Sensitivität* 94%, *Spezifität* 55%). Grundsätzlich kann aus ethisch medizinischer Sicht eine etwas erhöhte Rate an falsch-positiven Befunden besser toleriert werden als falsch-negative Beurteilungen (Loc 1 33 falsch-positive und 3 falsch-negative Auswertungen, Loc 2 hingegen 146 bzw. 1).

Für die EoE ergibt sich eine stabilere Performance des Algorithmus unter Realbedingungen (*Accuracy* Loc 5 95%, Loc 6 93%; *balanced Accuracy* Loc 5 98%, Loc 6 76%; *Sensitivität* Loc 5 100%, Loc 6 57%; *Spezifität* Loc 5 95%, Loc 6 94%). Dies ist in erster Linie dadurch erklärbar, dass im Ösophagus anatomisch bedingt eine deutlich geringere Variabilität der Bilddarstellung vorliegt, sodass das Spektrum der Erscheinungsbilder in dem auch repräsentativeren Trainingsdatensatz besser abgebildet wurde. Dass der Algorithmus die einzelnen Auffälligkeiten, die bei der EoE erkennbar sind, stabil wahrnehmen kann, wurde bereits in der initialen Validierung nachgewiesen. Die Unterschiede zwischen Loc 5 und Loc 6 sind marginal, sodass die unterschiedlichen Ergebnisse am ehesten daher rühren, dass die Kriterien der EoE in Loc 6 weniger klar oder nicht abgebildet waren, sodass hier mehr falsch-negative Befunde entstanden sind (Loc 5 21 falsch-positive und keine falsch-negative, Loc 6 30 falsch-positive und 3 falsch-negative Auswertungen). Die Ergebnisse aus Loc 5 belegen, dass die Limitationen durch teils schlechtes Bildmaterial wie im Vorkapitel für den Atrophie-Algorithmus aufgeführt auch hier gelten, sodass bei Elimination klar insuffizienter Bilder eine nochmals bessere Leistungsfähigkeit zu erwarten ist (siehe Kapitel *4.6 Fehleranalyse*; für Loc 5 21 und Loc 6 33 Bilder).

Schlussfolgernd zeigt diese Arbeit, dass bisher veröffentlichte retrospektive Studien zwar Anhaltspunkte zur Zuverlässigkeit der künstlichen Intelligenzen geben, jedoch zur besseren Beurteilung mehr prospektive klinische Validierungen unter reellen Bedingungen benötigt wer-

den. Außerdem müssen KI-Systeme optimalerweise unter Echtzeit-Untersuchungen funktionieren, damit eine tatsächliche Zeitersparnis entsteht und die Systeme praktikabel werden. Weiterhin wären deutlich größere und variablere Trainingsdatensätze erforderlich, damit diese Systeme verlässliche Zweitbeurteiler bei Untersuchungen werden könnten um dann vor allem junge Ärzte in Weiterbildung, aber auch erfahrene Fachärzte effizient unterstützen zu können.

6 Literaturverzeichnis

1. Abe Y, Sasaki Y, Yagi M, Mizumoto N, Onozato Y, Umehara M, Ueno Y (2022) Endoscopic Diagnosis of Eosinophilic Esophagitis: Basics and Recent Advances. Diagnostics 12:

2. Alexander ES, Martin LJ, Collins MH, Kottyan L, Sucharew H, He H, Mukkada VA, Succop PA, Pablo Abonia J, Foote H, Eby MD, Grotjan TM, Greenler AJ, Dellon ES, Demain JG, Furuta GT, Gurian LE, Harley JB, Hopp RJ, Kaul A, Nadeau KC, Noel RJ, Putnam PE, von Tiehl KF, Rothenberg ME (2014) Twin and family studies reveal strong environmental and weaker genetic cues explaining heritability of eosinophilic esophagitis. J Allergy Clin Immunol 134:1084–1092

3. Alexander R, Alexander JA, Ravi K, Geno D, Tholen C, Mara K, Katzka DA (2019) RESEARCH CORRESPONDENCE Measurement of Observed Eating Behaviors in Patients With Active and Inactive Eosinophilic Esophagitis. Clinical Gastroenterology and Hepatology 17:2371–2373

4. Allaume P, Rabilloud N, Turlin B, Bardou-Jacquet E, Loréal O, Calderaro J, Khene ZE, Acosta O, De Crevoisier R, Rioux-Leclercq N, Pecot T, Kammerer-Jacquet SF (2023) Artificial Intelligence-Based Opportunities in Liver Pathology—A Systematic Review. Diagnostics 13:

5. Attwood SE STDT et al. (1993) Esophageal eosinophilia with dysphagia. A distinct clinicopathologic syndrome. Dig Dis Sci 109–116

6. Aumüller. G, Aust G, Conrad A, Engele J, Kirsch J, Maio G, Mayerhofer A, Mense S, Reißig D (2020) Duale Reihe Anatomie. Thieme, Stuttgart

7. Biedermann L, Holbreich M, Atkins D, Chehade M, Dellon ES, Furuta GT, Hirano I, Gonsalves N, Greuter T, Gupta S, Katzka DA, De Rooij W, Safroneeva E, Schoepfer A, Schreiner P, Simon D, Simon HU, Warners M, Bredenoord AJ, Straumann A (2021) Food-induced immediate response of the esophagus—A newly identified syndrome in patients with eosinophilic esophagitis. Allergy 76:339–347

8. Bittencourt De Brito B, Antônio França Da Silva F, Soares AS, Pereira A, Luísa M, Santos C, Sampaio MM, Henrique P, Neves M, Freire De Melo F (2019) Pathogenesis and clinical management of Helicobacter pylori gastric infection. World J Gastroenterol 25:5578–5589

9. Botezatu A, Bodrug N (2021) Chronic atrophic gastritis: an update on diagnosis. Med Pharm Rep 94:7–14

10. Chan HP, Samala RK, Hadjiiski LM, Zhou C (2020) Deep Learning in Medical Image Analysis. Springer

11. Chaudhry SR, Liman MNP, Peterson DC (2022) Anatomy, Abdomen and Pelvis: Stomach. StatPearls

12. Chen L, Han B, Wang X, Zhao J, Yang W, Yang Z (2023) Machine Learning Methods in Weather and Climate Applications: A Survey. Applied Sciences (Switzerland) 13:

13. Cheng N, Ren Y, Zhou J, Zhang Y, Wang D, Zhang X, Chen B, Liu F, Lv J, Cao Q, Chen S, Du H, Hui D, Weng Z, Liang Q, Su B, Tang L, Han L, Chen J, Shao C (2022) Deep Learning-Based Classification of Hepatocellular Nodular Lesions on Whole-Slide Histopathologic Images. Gastroenterology 162:1948-1961.e7

14. Collins MH (2014) Histopathologic Features of Eosinophilic Esophagitis and Eosinophilic Gastrointestinal Diseases. Gastroenterol Clin North Am 43:257–268

15. Correa P (1988) A human model of gastric carcinogenesis. Cancer Res

16. Correa P, Piazuelo MB (2012) The gastric precancerous cascade. J Dig Dis 13:2–9

17. de Groof AJ, Struyvenberg MR, van der Putten J, van der Sommen F, Fockens KN, Curvers WL, Zinger S, Pouw RE, Coron E, Baldaque-Silva F, Pech O, Weusten B, Meining A, Neuhaus H, Bisschops R, Dent J, Schoon EJ, de With PH, Bergman JJ (2020) Deep-Learning System Detects Neoplasia in Patients With Barrett's Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking. Gastroenterology 158:915-929.e4

18. de Groof AJ, Struyvenberg MR, Fockens KN, van der Putten J, van der Sommen F, Boers TG, Zinger S, Bisschops R, de With PH, Pouw RE, Curvers WL, Schoon EJ, Bergman JJGHM (2020) Deep learning algorithm detection of Barrett's neoplasia with high accuracy during live endoscopic procedures: a pilot study (with video). Gastrointest Endosc 91:1242–1250

19. Dellon ES, Gibbs WB, Fritchie KJ, Rubinas TC, Wilson LA, Woosley JT, Shaheen NJ (2009) Clinical, endoscopic, and histologic findings distinguish eosinophilic esophagitis from gastroesophageal reflux disease.

20. Dellon ES, Hirano I Epidemiology and Natural History of Eosinophilic Esophagitis.

21. Dellon ES, Kim HP, Sperry SLW, Rybnicek DA, Woosley JT, Shaheen NJ A phenotypic analysis shows eosinophilic esophagitis is a progressive fibrostenotic disease.

22. Ebigbo A, Palm C, Probst A, Mendel R, Manzeneder J, Prinz F, Souza LA de, Papa JP, Siersema P, Messmann H (2019) A technical review of artificial intelligence as applied to gastrointestinal endoscopy: clarifying the terminology. Endosc Int Open 7:E1616

23. E.D.D. D, F.R.C.S., L.R.C.P., L.D.S., R.C.S. (1948) THE APPLIED ANATOMY AND PHYSIOLOGY OF THE PHARYNX AND OESOPHAGUS.

24. Eusebi LH, Zagari RM, Bazzoli F (2014) Epidemiology of Helicobacter pylori Infection. Helicobacter 19:1–5

25. Ewals LJS, van der Wulp K, van den Borne BEEM, Pluyter JR, Jacobs I, Mavroeidis D, van der Sommen F, Nederend J (2023) The Effects of Artificial Intelligence Assistance on the Radiologists' Assessment of Lung Nodules on CT Scans: A Systematic Review. J Clin Med 12:

26. Giriens B, Yan P, Safroneeva E, Zwahlen M, Reinhard A, Nydegger A, Vavricka S, Sempoux C, Straumann A, Schoepfer AM (2015) Escalating incidence of eosinophilic esophagitis in Canton of Vaud, Switzerland, 1993–2013: a population-based study. Allergy 70:1633–1639

27. González-Cervera J, Arias Á, Redondo-González O, Cano-Mollinedo MM, Terreehorst I, Lucendo AJ (2017) Association between atopic manifestations and eosinophilic esophagitis: A systematic review and meta-analysis. Annals of Allergy, Asthma and Immunology 118:582-590.e2

28. Grunwald IQ, Kulikovski J, Reith W, Gerry S, Namias R, Politi M, Papanagiotou P, Essig M, Mathur S, Joly O, Hussain K, Wagner V, Shah S, Harston G, Vlahovic J, Walter S, Podlasek A, Fassbender K (2019) Collateral Automation for Triage in Stroke: Evaluating Automated Scoring of Collaterals in Acute Stroke on Computed Tomography Scans. Cerebrovasc Dis 47:217

29. Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M (2020) Deep-learning based detection of gastric precancerous conditions. Gut 69:4–6

30. Guimarães P, Keller A, Fehlmann T, Lammert F, Casper M (2022) Deep learning-based detection of eosinophilic esophagitis. Endoscopy 54:299–304

31. Gulati S, Bernth J, Liao J, Poliyivets D, Chatu S, Emmanuel A, Haji A, Liu H, Hayee B (2019) OTU-07 Near focus narrow and imaging driven artificial intelligence for the diagnosis of gastro-oesophageal reflux disease. A4.1-A4

32. Gulati S, Bernth J, Liao J, Poliyivets D, Chatu S, Emmanuel A, Haji A, Liu H, Hayee B (2019) NEAR FOCUS NARROW BAND IMAGING DRIVEN ARTIFICIAL INTELLIGENCE FOR THE DIAGNOSIS OF GASTROESOPHAGEAL REFLUX DISEASE. Gastrointest Endosc 89:AB633

33. Guo LJ, Xiao X, Wu CC, Zeng X, Zhang Y, Du J, Bai S, Xie J, Zhang Z, Li Y, Wang X, Cheung O, Sharma M, Liu J, Hu B (2020) Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc 91:41–51

34. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25:1315–1360

35. Gupta SK, Fitzgerald JF, Chong SKF, Croffie JM, Collins MH (1997) Vertical lines in distal esophageal mucosa (VLEM): a true endoscopic manifestation of esophagitis in children? Gastrointest Endosc 45:485–489

36. Hassan C, Wallace MB, Sharma P, Maselli R, Craviotto V, Spadaccini M, Repici A (2020) New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut 69:799–800

37. Hassan C, Badalamenti M, Maselli R, Correale L, Iannone A, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Craviotto V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rösch T, Repici A (2020) Computer-aided detection-assisted colonoscopy: classification and relevance of false positives. Gastrointest Endosc 92:900-904.e4

38. Hassan C, Spadaccini M, Iannone A, Maselli R, Jovani M, Chandrasekar VT, Antonelli G, Yu H, Areia M, Dinis-Ribeiro M, Bhandari P, Sharma P, Rex DK, Rösch

T, Wallace M, Repici A (2021) Performance of artificial intelligence in colonoscopy for adenoma and polyp detection: a systematic review and meta-analysis. Gastrointest Endosc 93:77-85.e6

39. Hines BT, Rank MA, Wright BL, Marks LA, Hagan JB, Straumann A, Greenhawt M, Dellon ES (2018) Minimally-Invasive Biomarker Studies in Eosinophilic Esophagitis: A Systematic Review. Ann Allergy Asthma Immunol 121:218

40. Hirano I, Moy N, Heckman MG, Thomas CS, Gonsalves N, Achem SR (2013) Endoscopic assessment of the oesophageal features of eosinophilic oesophagitis: validation of a novel classification and grading system. Gut 62:489–495

41. Hussein M, González-Bueno Puyal J, Lines D, Sehgal V, Toth D, Ahmad OF, Kader R, Everson M, Lipman G, Fernandez-Sordo JO, Ragunath K, Esteban JM, Bisschops R, Banks M, Haefner M, Mountney P, Stoyanov D, Lovat LB, Haidry R (2022) A new artificial intelligence system successfully detects and localises early neoplasia in Barrett's esophagus by using convolutional neural networks. United European Gastroenterol J 10:528–537

42. Kagalwalla AF, Shah A, Li BUK, Sentongo TA, Ritz S, Manuel-Rubio M, Jacques K, Wang D, Melin-Aldana H, Nelson SP (2011) Identification of specific foods responsible for inflammation in children with eosinophilic esophagitis successfully treated with empiric elimination diet. J Pediatr Gastroenterol Nutr 53:145–149

43. Kather JN, Krause J, Luedde T (2020) Künstliche Intelligenz in der Gastroenterologie. DMW - Deutsche Medizinische Wochenschrift 145:1450–1454

44. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, Mathur P, Islam S, Yeom KW, Lawlor A, Killeen RP (2022) Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). Eur Radiol 32:7998–8007

45. Kim HP, Vance RB, Shaheen NJ, Dellon ES (2012) The prevalence and diagnostic utility of endoscopic features of eosinophilic esophagitis: a meta-analysis. Clin Gastroenterol Hepatol 10:988-996.e5

46. Kono S, Gotoda T, Yoshida S, Oda I, Kondo H, Gatta L, Naylor G, Dixon M, Moriyasu F, Axon A (2015) Can endoscopic atrophy predict histological atrophy? Historical study in United Kingdom and Japan. World J Gastroenterol 21:13113–13123

47. Krizhevsky A, Sutskever I, Hinton GE ImageNet Classification with Deep Convolutional Neural Networks.

48. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY (2022) Randomized Controlled Trials of Artificial Intelligence in Clinical Practice: Systematic Review. J Med Internet Res 24:

49. Lecun Y, Eon Bottou L, Bengio Y, Abstract| PH Gradient-Based Learning Applied to Document Recognition.

50. Liacouras CA, Furuta GT, Hirano I, Atkins D, Attwood SE, Bonis PA, Burks AW, Chehade M, Collins MH, Dellon ES, Dohil R, Falk GW, Gonsalves N, Gupta SK, Katzka DA, Lucendo AJ, Markowitz JE, Noel RJ, Odze RD, Putnam PE, Richter JE, Romero Y, Ruchelli E, Sampson HA, Schoepfer A, Shaheen NJ, Sicherer SH, Spechler S, Spergel JM, Straumann A, Wershil BK, Rothenberg ME, Aceves SS (2011) Eosinophilic esophagitis: Updated consensus recommendations for children and adults. Journal of Allergy and Clinical Immunology 128:3-20.e6

51. Lim JR, Gupta SK, Croffie JM, Pfefferkorn MD, Molleston JP, Corkins MR, Davis MM, Faught PP, Steiner SJ, Fitzgerald JF (2004) White specks in the esophageal mucosa: An endoscopic manifestation of non-reflux eosinophilic esophagitis in children. Gastrointest Endosc 59:835–838

52. Lucendo AJ, Molina-Infante J, Arias Á, von Arnim U, Bredenoord AJ, Bussmann C, Amil Dias J, Bove M, González-Cervera J, Larsson H, Miehlke S, Papadopoulou A, Rodríguez-Sánchez J, Ravelli A, Ronkainen J, Santander C, Schoepfer AM, Storr MA, Terreehorst I, Straumann A, Attwood SE (2017) Guidelines on eosinophilic esophagitis: evidence-based statements and recommendations for diagnosis and management in children and adults. United European Gastroenterol J 5:335

53. Madisch A, Koop H, Miehlke S, Leers J, Lorenz P, Jansen PL, Pech O, Schilling D, Labenz J (2023) S2k-Leitlinie Gastroösophageale Refluxkrankheit und eosinophile ösophagitis der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (DGVS) - März 2023 - AWMF-Registernummer: 021-013. Z Gastroenterol 61:862–933

54. Mansour-Ghanaei F, Joukar F, Yeganeh S, Sadeghi M, Daryakar A, Sepehrimanesh M (2022) OLGA- and OLGIM-Based Staging in the Patients with Gastritis and Endoscopy Indications. Turkish Journal of Gastroenterology 33:95–102

55. Martí-Aguado D, Jiménez-Pastor A, Alberich-Bayarri Á, Rodríguez-Ortega A, Alfaro-Cervello C, Mestre-Alagarda C, Bauza M, Gallén-Peris A, Valero-Pérez E,

Ballester MP, Gimeno-Torres M, Pérez-Girbés A, Benlloch S, Pérez-Rojas J, Puglia V, Ferrández A, Aguilera V, Escudero-García D, Serra MA, Martí-Bonmatí L (2022) Automated Whole-Liver MRI Segmentation to Assess Steatosis and Iron Quantification in Chronic Liver Disease. Radiology 302:345–354

56. Messmann H, Ebigbo A, Hassan C, Repici A, Mori Y (2022) How to Integrate Artificial Intelligence in Gastrointestinal Practice. Gastroenterology 162:1583–1586

57. Miotto R, Wang F, Wang S, Jiang X, Dudley JT (2017) Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform 19:1236–1246

58. Moawad FJ, Robinson CL, Veerappan GR, Summers TA, Maydonovitch CL, Wong RK (2013) The tug sign: An endoscopic feature of eosinophilic esophagitis. American Journal of Gastroenterology 108:1938–1939

59. Mori Y, Kudo S ei, East JE, Rastogi A, Bretthauer M, Misawa M, Sekiguchi M, Matsuda T, Saito Y, Ikematsu H, Hotta K, Ohtsuka K, Kudo T, Mori K (2020) Cost savings in colonoscopy with artificial intelligence-aided polyp diagnosis: an add-on analysis of a clinical trial (with video). Gastrointest Endosc 92:905-911.e1

60. Murchison JT, Ritchie G, Senyszak D, Nijwening JH, van Veenendaal G, Wakkie J, van Beek EJR (2022) Validation of a deep learning computer aided system for CT based lung nodule detection, classification, and growth rate estimation in a routine clinical population. PLoS One 17:

61. Nagao S, Tsuji Y, Sakaguchi Y, Takahashi Y, Minatsuki C, Niimi K, Yamashita H, Yamamichi N, Seto Y, Tada T, Koike K (2020) Highly accurate artificial intelligence systems to predict the invasion depth of gastric cancer: efficacy of conventional white-light imaging, nonmagnifying narrow-band imaging, and indigo-carmine dye contrast imaging. Gastrointest Endosc 92:866-873.e1

62. Nakagawa K, Ishihara R, Aoyama K, Ohmori M, Nakahira H, Matsuura N, Shichijo S, Nishida T, Yamada T, Yamaguchi S, Ogiyama H, Egawa S, Kishida O, Tada T (2019) Classification for invasion depth of esophageal squamous cell carcinoma using a deep neural network compared with experienced endoscopists. Gastrointest Endosc 90:407–414

63. Nakashima H, Kawahira H, Kawachi H, Sakaki N (2018) Artificial intelligence diagnosis of helicobacter pylori infection using blue laser imaging-bright and linked color imaging: A single-center prospective study. Ann Gastroenterol 31:462–468

64. Neuhaus A, Seyedsaadat SM, Mihal D, Benson J, Mark I, Kallmes DF, Brinjikji W (2020) Region-specific agreement in ASPECTS estimation between neuroradiologists and e-ASPECTS software. J Neurointerv Surg 12:720–724

65. Olthof AW, Van Ooijen PMA, Mehrizi MHR Promises of artificial intelligence in neuroradiology: a systematic technographic review.

66. Petrov R V., Su S, Bakhos CT, Abbas AES (2019) Surgical Anatomy of Paraesophageal Hernias. Thorac Surg Clin 29:359–368

67. Pimentel-Nunes P, Libânio D, Marcos-Pinto R, Areia M, Leja M, Esposito G, Garrido M, Kikuste I, Megraud F, Matysiak-Budnik T, Annibale B, Dumonceau JM, Barros R, Fléjou JF, Carneiro F, Van Hooft JE, Kuipers EJ, Dinis-Ribeiro M (2019) Management of epithelial precancerous conditions and lesions in the stomach (MAPS II): European Society of Gastrointestinal Endoscopy (ESGE), European Helicobacter and Microbiota Study Group (EHMSG), European Society of Pathology (ESP), and Sociedade Portuguesa de Endoscopia Digestiva (SPED) guideline update 2019. Endoscopy 51:365–388

68. Prasad GA, Alexander JA, Schleck CD, Zinsmeister AR, Smyrk TC, Elias RM, Richard G, Iii L, Talley NJ Epidemiology of Eosinophilic Esophagitis over 3 Decades in Olmsted County, Minnesota.

69. Prinz C, Kajimura M, Scoyt D, Helander H, Shin J, Besancon M, Bamberg K, Hersey S, Sachs G (1992) Acid Secretion and the H,K ATPase of Stomach.

70. Ramsay PT, Carr A (2011) Gastric Acid and Digestive Physiology. Surgical Clinics of North America 91:977–982

71. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, Ferrara E, Spadaccini M, Alkandari A, Fugazza A, Anderloni A, Galtieri PA, Pellegatta G, Carrara S, Di Leo M, Craviotto V, Lamonaca L, Lorenzetti R, Andrealli A, Antonelli G, Wallace M, Sharma P, Rosch T, Hassan C (2020) Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. Gastroenterology 159:512-520.e7

72. Rodriguez-Castro KI, Franceschi M, Noto A, Miraglia C, Nouvenne A, Leandro G, Meschi T, De' Angelis GL, Mario F Di (2018) Clinical manifestations of chronic atrophic gastritis. Acta Biomedica 89:88–92

73. Römmele C, Mendel R, Barrett C, Kiesl H, Rauber D, Rückert T, Kraus L, Heinkele J, Dhillon C, Grosser B, Prinz F, Wanzl J, Fleischmann C, Nagl S, Schnoy E, Schlottmann J, Dellon ES, Messmann H, Palm C, Ebigbo A (2022) An artificial intelligence algorithm is highly accurate for detecting endoscopic features of eosinophilic esophagitis. Sci Rep 12:

74. Ronkainen J, Talley NJ, Aro P, Storskrubb T, Johansson S-E, Lind T, Bolling-Sternevald E, Vieth M, Stolte M, Walker MM, Agréus L, CeFAM R Prevalence of oesophageal eosinophils and eosinophilic oesophagitis in adults: the population-based Kalixanda study.

75. Rosenblatt F. (1957) The Perceptron. A Perceiving and Recocgnizing Automaton.

76. Rugge M, Bricca L, Guzzinati S, Sacchi D, Pizzi M, Savarino E, Farinati F, Zorzi M, Fassan M, Dei Tos AP, Malfertheiner P, Genta RM, Graham DY (2023) Autoimmune gastritis: long-term natural history in naïve Helicobacter pylori-negative patients. Gut 72:30–38

77. S3-Leitlinie Magenkarzinom - Diagnostik und Therapie der Adenokarzinome des Magens und ösophagogastralen Übergangs Leitlinienprogramm Onkologie - August 2019 - AWMF-Registernummer: 032/009OL.

78. Salvador ASSIRATI F, Lyoiti HASHIMOTO C, Anuar DIB R, Henrique Souza FONTES L, Navarro-rodriguez T HIGH DEFINITION ENDOSCOPY AND „NARROW BAND IMAGING“ IN THE DIAGNOSIS OF GASTROESOPHAGEAL REFLUX DISEASE.

79. Scherer D, Müller A, Behnke S Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition.

80. Schmitz R, Werner R, Rösch T (2019) Artificial Intelligence in Endoscopy: Deep Neural Nets for Endoscopic Computer Vision - Methods & Perspectives. Z Gastroenterol 57:767–780

81. Schoepfer AM, Gonsalves N, Bussmann C, Conus S, Simon HU, Straumann A, Hirano I (2010) Esophageal dilation in eosinophilic esophagitis: Effectiveness, safety, and impact on the underlying inflammation. American Journal of Gastroenterology 105:1062–1070

82. Schoepfer AM, Hirano I, Coslovsky M, Roumet MC, Zwahlen M, Kuehni CE, Hafner D, Alexander JA, Dellon ES, Gonsalves N, Leung J, Bussmann C, Collins MH,

Newbury RO, Smyrk TC, Woosley JT, Yang G-Y, Romero Y, Katzka DA, Furuta GT, Gupta SK, Aceves SS, Chehade M, Spergel JM, Falk GW, Meltzer BA, Comer GM, Straumann A, Safroneeva E (2018) Variation in Endoscopic Activity Assessment and Endoscopy Score Validation in Adults with Eosinophilic Esophagitis Running head: Endoscopy score for adults with EoE. Clinical Gastroenterology and Hepatology

83. Schwarzer G, Türp JC, Antes G (2002) Wahrscheinlichkeitsverhältnis (Likelihood Ratio) - Alternative zu Sensitivität und Spezifität.

84. Sgouros SN, Bergele C, Mantides A (2006) Eosinophilic esophagitis in adults: a systematic review. 211–217

85. Shah SC, Piazuelo MB, Kuipers EJ, Li D (2021) AGA Clinical Practice Update on the Diagnosis and Management of Atrophic Gastritis: Expert Review. Gastroenterology 161:1325-1332.e7

86. Shichijo S, Endo Y, Aoyama K, Takeuchi Y, Ozawa T, Takiyama H, Matsuo K, Fujishiro M, Ishihara S, Ishihara R, Tada T (2019) Application of convolutional neural networks for evaluating Helicobacter pylori infection status on the basis of endoscopic images. Scand J Gastroenterol 54:158–163

87. Shimamoto Y, Ishihara R, Kato Y, Shoji A, Inoue T, Matsueda K, Miyake M, Waki K, Kono M, Fukuda H, Matsuura N, Nagaike K, Aoi K, Yamamoto K, Inoue T, Nakahara M, Nishihara A, Tada T (2020) Real-time assessment of video images for esophageal squamous cell carcinoma invasion depth using artificial intelligence. J Gastroenterol 55:1037–1045

88. Simon D, Cianferoni A, Spergel JM, Aceves S, Holbreich M, Venter C, Rothenberg ME, Terreehorst I, Muraro A, Lucendo AJ, Schoepfer A, Straumann A, Simon HU (2016) Eosinophilic esophagitis is characterized by a non-IgE-mediated food hypersensitivity. Allergy 71:611–620

89. Sipponen P, Maaroos HI (2015) Chronic gastritis. Scand J Gastroenterol 50:657–667

90. Sipponen P, Maaroos HI (2015) Chronic gastritis. Scand J Gastroenterol 50:657–667

91. Straumann A, Spichtin H-P, Grize L, Bucher KA, Beglinger C, Simon H-U (2003) Natural History of Primary Eosinophilic Esophagitis: A Follow-up of 30 Adult Patients for Up to 11.5 Years.

92. Taft TH, Guadagnoli L, Edlynn E (2019) Anxiety and Depression in Eosinophilic Esophagitis: A Scoping Review and Recommendations for Future Research.

93. Ueyama H, Kato Y, Akazawa Y, Yatagai N, Komori H, Takeda T, Matsumoto K, Ueda K, Matsumoto K, Hojo M, Yao T, Nagahara A, Tada T (2021) Application of artificial intelligence using a convolutional neural network for diagnosis of early gastric cancer based on magnifying endoscopy with narrow-band imaging. J Gastroenterol Hepatol 36:482–489

94. Visaggi P, De Bortoli N, Barberio B, Savarino V, Oleas R, Rosi EM, Marchi S, Ribolsi M, Savarino E (2022) Artificial Intelligence in the Diagnosis of Upper Gastrointestinal Diseases. J Clin Gastroenterol 56:23

95. Wang L, Huang W, Du J, Chen Y, Yang J (2014) Diagnostic yield of the light blue crest sign in gastric intestinal metaplasia: A meta-analysis. PLoS One 9:

96. Wolpert F, Baráth K, Brakowski J, Renzel R, Linnebank M, Gantenbein AR (2015) Funicular myelosis in a butcher: it was the cream cans. Case Rep Neurol Med 2015:1–3

97. Wong YT, Tai TF, Wong KF, Leung SK, Lam SM, Wong SY, Lo YY, Yan KM, Tam SK, Wong MF, Chan HL (2022) The study on artificial intelligence (AI) colonoscopy in affecting the rate of polyp detection in colonoscopy: A single centre retrospective study. Surg Pract 26:115–119

98. Y Hooi JK, Ying Lai W, Khoon Ng W, Y Suen MM, Underwood FE, Tanyingoh D, Malfertheiner P, Graham DY, S Wong VW, Y Wu JC, L Chan FK, Y Sung JJ, Kaplan GG, Ng SC (2017) Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. Gastroenterology 153:420–429

99. Yin Y, Liang H, Wei N, Zheng Z (2022) Prevalence of chronic atrophic gastritis worldwide from 2010 to 2020: an updated systematic review and meta-analysis. Ann Palliat Med 11:3697–3703

100. Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML (2020) Artificial Intelligence in Dermatology: A Primer. Journal of Investigative Dermatology 140:1504–1512

101. Zhang Y, Weck MN, Schöttker B, Rothenbacher D, Brenner H (2013) Gastric parietal cell antibodies, Helicobacter pylori infection, and chronic atrophic gastritis: evidence from a large population-based study in Germany. Cancer Epidemiol Biomarkers Prev 22:821–826

102. Zhang Y, Li F, Yuan F, Zhang K, Huo L, Dong Z, Lang Y, Zhang Y, Wang M, Gao Z, Qin Z, Shen L (2020) Diagnosing chronic atrophic gastritis by gastroscopy using artificial intelligence. Digestive and Liver Disease 52:566–572

103. (2023) Aktualisierte S2k-Leitlinie Helicobacter pylori und gastroduodenale Ulkuskrankheit der Deutschen Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten (DGVS) – Juli 2022 – AWMF-Registernummer: 021–001. Z Gastroenterol 61:544–606

7 Abbildungsverzeichnis

8 Tabellenverzeichnis

9 Abkürzungsverzeichnis

ADR	Adenomdetektionsrate
AG	Atrophische Gastritis
AWMF	Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V.
CNN	Convolutional Neuronal Network
CT	Computertomographie
DGVS	Deutsche Gesellschaft für Gastroenterologie, Verdauungs- und Stoffwechselkrankheiten
DL	Deep Learning
ECL	enterochromaffin-like cells
EoE	Eosinophile Ösophagitis
EoE-HSS	EoE-spezifischer-Histologie-Score
EREFS	Exsudate, Ringe, Ödem, Furchen, Strikturen
FN	falsch-negativ
FP	falsch-positiv
GERD	Gastroösophageale Reflux-Krankheit
GIT	Gastrointestinaltrakt
HD-WLE	High-Definition-Weißlicht-Endoskopie
HE	Hämatoxylin-Eosin-Färbung
HP	Helicobacter pylori
HPF	High-power field (hochauflösendes Gesichtsfeld)
HRQOL	Health related quality of life
IM	intestinale Metaplasie
JPG	Joint Photographic (Experts) Group
KI	Künstliche Intelligenz
LBC	Light Blue Crest
Loc	Lokation
MRT	Magnetresonanztomographie
NBI	Narrow Band Imaging
NET	neuroendokrine Tumoren
ÖGD	Ösophago-Gastro-Duodenoskopie
OLGA	Operative Link on Gastritis Assessment
OLGIM	Operative Link on Gastric Intestinal Metaplasia Assessment
OÖS	oberer Ösophagussphinkter
PG	Pepsinogen
PN	predicted negative

PP	predicted positive
PPI	Protonenpumpeninhibitor
PPI-REE	Protonenpumpeninhibitor-responsive Eosinophilie
PPV	positive predicted value
ReLu	Rectified Linear Unit
RN	richtig-negativ
RP	richtig-positiv
SPSS	Statistical Package for the Social Sciences
UKS	Universitätsklinikum des Saarlandes
UÖS	unterer Ösophagussphinkter
WOF	white opaque field

10 Publikation

Publikation in Planung und Arbeit in Zusammenarbeit u.a. mit Dr. M. Casper und Frau A. Engel.

11 Danksagung

Mein Dank gilt zunächst einmal meinem Doktorvater, Dr. M. Casper. Sie haben mir immer und zu jeder Zeit schnell mit Rat zur Seite gestanden und mir viele offene Fragen beantwortet. Sie waren immer geduldig und haben mir, wenn nötig auch 5x erklärt wieso, weshalb, warum wir etwas so tun sollten, wie wir es taten. Außerdem haben Sie viel zur Organisation beigetragen und viele Evaluationsbögen überprüft wodurch wir eine größere Anzahl an Patienten und Bilder generieren konnten.

Außerdem bedanke ich mich herzlich bei Frau A. Engel. Sie haben mir viele statistische Aspekte in langen Meetings erklärt und mir auch immer bei offenen Fragen weitergeholfen. Außerdem haben Sie mir viele Berechnungen erklärt und gezeigt, und mir bei vielen Abbildungen und Diagrammen geholfen.

Ebenfalls bedanke ich mich beim gesamten Team der Inneren II der IMED des UKS. Ohne diese motivierte Zusammenarbeit wären nicht so viele ausgefüllte Evaluationsbögen entstanden. Auch bedanke ich mich speziell bei Dr. O. Linn, indem Sie mir bei der Datensammlung viele weitere interessante Aspekte erläuterten, die im Verlauf der Arbeit benötigt wurden.

Zuletzt gilt mein Dank meinem Ehemann, Caroline, Emilie und Kevin, die mich alle immer wieder motiviert haben, sich mein Leid angehört haben und mich versucht haben zum Schreiben zu motivieren. Ihr durftet euch viele offene Fragen anhören und lange Monologe, aber ihr habt sie alle tapfer durchgestanden und mir euer offenes Ohr geboten.

12 Lebenslauf

Aus datenschutzrechtlichen Gründen wird der Lebenslauf in der elektronischen Fassung der Dissertation nicht veröffentlicht.