Aus der Fakultät für Medizin
der Universität Regensburg
Prof. Dr. med. Oliver Pech
Gastroenterologie und interventionelle Endoskopie

Investigating the Generalizability of a Computer-Aided Detection System for Barrett's Neoplasia by Application on Untrained Domains

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Medizin

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Leona Brenner

2026

Aus der Fakultät für Medizin
der Universität Regensburg
Prof. Dr. med. Oliver Pech
Gastroenterologie und interventionelle Endoskopie

Investigating the Generalizability of a Computer-Aided Detection System for Barrett's Neoplasia by Application on Untrained Domains

Inaugural – Dissertation
zur Erlangung des Doktorgrades
der Medizin

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Leona Brenner

2026

Dekan:	Prof. Dr. Dirk Hellwig
1. Berichterstatter:	Prof. Dr. Oliver Pech
2. Berichterstatter:	Prof. Dr. Hans-Stefan Hofmann
Tag der mündlichen Prüfung:	30. Oktober 2026

Table of Contents

0 Zusammenfassung (Deutsch) .....
Abbreviations .....
1 Introduction **19**
1.1 Objectives of the Dissertation 20
2 Theoretical Background **21**
2.1 Barrett's Esophagus 21
2.1.1 Diagnostics 21
2.1.2 Esophageal Adenocarcinoma 23
2.1.3 Treatment 23
2.1.4 Surveillance Guidelines 24
2.2 Introduction to Neural Networks 25
2.2.1 Convolutional Neural Networks 26
2.3 Domain Shift 27
2.4 Use of Deep Learning in Endoscopy 27
3 Methodology **30**
3.1 Study Setting and Design 30
3.2 Design of the Computer Aided Detection System 32
3.3 Data Acquisition 33
3.3.1 *Test set 1* 36
3.3.2 *Test set 2* 36
3.3.3 *Test set 3* and *Test set 4* 36
3.4 Performance Metrics 37
4 Results **40**
4.1 Performance on *Test set 1* 40
4.2 Performance on *Test set 2* 41
4.3 Performance on *Test set 3* 41
4.4 Performance on *Test set 4* 42
4.5 Comparison between *Test set 1* and *2* 43
4.6 Comparison between *Test set 2* and *3* 44
4.7 Comparison between *Test set 3* and *4* 44
5 Discussion **46**
5.1 Performance on Images from New Endoscopy System 46
5.2 Performance on Images from New Medical Center 52

5.3 Performance on Posttreatment Images 54
5.4 Evaluation of Applicability in Non-Expert Settings 55
5.5 The Domain Shift Problem 56
5.6 Need for External Validation 58
5.7 Explainable Artificial Intelligence 60
5.8 Collaboration for Common Datasets 61
5.9 A General Frame for Artificial Intelligence 62
5.10 Limitations 64
6 Conclusion 65
7 Appendix 66
7.1 List of Tables 66
7.2 List of Figures 66
8 References .....
Danksagung .....
Selbstständigkeitserklärung .....

0 Zusammenfassung (Deutsch)

Die folgende Zusammenfassung gibt einen Überblick über die englischsprachige Dissertation mit dem Titel *Untersuchung der Verallgemeinbarkeit eines computergestützten Erkennungssystems für Barrett-Neoplasien durch Anwendung auf untrainierte Domänen.*

0.1 Einleitung und Zielsetzung

Ösophaguskarzinome sind weit verbreitet und gehören zu den Krebsarten mit hoher Mortalität (1). Während Plattenepithelkarzinome weltweit noch überwiegen, nimmt die Prävalenz von Adenokarzinomen stetig zu, insbesondere in Ländern mit hohem Einkommen (1). Eine frühe endoskopische Entdeckung ist daher prognostisch entscheidend. Trotz neuer Entwicklungen und Verbesserungen der modernen endoskopischen Bildgebung bleibt die Interpretation diskreter Veränderungen der Mukosa schwierig (2). Vor allem nicht-spezialisierten Endoskopiker:innen fehlt häufig die Routine und Erfahrung in der Erkennung von frühen neoplastischen Läsionen im Barrett-Ösophagus (3,4). Künstliche Intelligenz kann hier unterstützen, da sie einen schnellen Lernprozess hat und unabhängig von der Erfahrung des Untersuchers, Müdigkeit, Stress oder Konzentrationsverlust arbeitet (5,6). So könnte ein verlässlicher und objektiver Qualitätsstandard in der Routine-Endoskopie flächendeckend gesichert werden (7). Dies würde vor allem in der Überwachung des Barrett-Ösophagus eine Rolle spielen, was häufig in nicht-spezialisierten Zentren stattfindet. Gerade dort ist eine Entdeckung früher Dys- und Neoplasien essentiell, um die Patient:innen für eine korrekte Behandlung an Expertenzentren weiterzuleiten (8). Allerdings sind viele auf künstlicher Intelligenz basierende Systeme in nur einem Zentrum trainiert und entwickelt, was ihre Übertragbarkeit einschränkt. Für einen breiten klinischen Einsatz müssen sie auch bei wechselnden Bedingungen wie unterschiedlichem Endoskopie-Equipment, Patientenkollektiven und Behandler:innen zuverlässig funktionieren (9).

Diese Arbeit untersucht, wie robust ein bereits bestehendes computergestütztes Erkennungssystem für die Entdeckung von frühen Barett-Neoplasien unter solchen Veränderungen bleibt.

Dazu ändern wir schrittweise das Endoskopiesystem zur Bildaufnahme (von Fujifilm zu Olympus), die Herkunft der Bilddaten (bekanntes vs. unbekanntes Zentrum) und den Behandlungsstatus des Ösophagus (prä- vs. posttherapeutisch). Ziel ist es, die

Generalisierbarkeit des Algorithmus in neuen klinischen Settings, in denen posttherapeutische Überwachungsendoskopien und neue Bildquellen zu erwarten sind, zu bewerten.

0.2 Theoretischer Hintergrund

0.2.1 Barrett-Ösophagus

Barrett-Ösophagus bezeichnet die Umwandlung von Plattenepithel zu Zylinderepithel im Ösophagus, meist als Folge chronischer epithelialer Schädigung durch Magensäure (10,11). Die Prävalenz beträgt etwa 5,57 %, variiert jedoch regional (12–14). Der Barrett-Ösophagus gilt als prämaligne Veränderung und stellt den einzig bekannten Vorläufer für ösophageale Adenokarzinome dar (15,16). Die maligne Transformation durchläuft typischerweise zunächst die niedriggradige und hochgradiger Dysplasie, bevor sich ein mukosales und in Folge submukosales Adenokarzinom entwickelt (2,10). Angesichts eines zehnfachen Anstiegs der Inzidenz ösophagealer Adenokarzinome in den letzten 25 Jahren bei gleichzeitig schlechter Prognose ist eine frühzeitige Erkennung von entscheidender Bedeutung (16–18).

Diese erfolgt endoskopisch und basiert auf visueller Identifikation von Zylinderepithel im distalen Ösophagus mit mukosalen Veränderungen und einer Biopsie zur histopathologischen Bestätigung von dys- oder neoplastischen Zellen (11).

Die Therapie richtet sich nach dem jeweiligen Stadium der Neoplasie. Bei geringer Tumortiefe erfolgt bevorzugt eine endoskopische Resektion (19), fortgeschrittenere Stadien erfordern häufig eine chirurgische Intervention in Kombination mit weiteren Verfahren (20). Endoskopische Verfahren sind weniger invasiv, erfordern jedoch engmaschige Nachsorge wegen möglicher Rezidive (21,22). Chirurgische Eingriffe zeigen geringere Rezidivraten, sind jedoch mit einem deutlich erhöhten Komplikationsrisiko verbunden (22,23).

Die European Society for Gastrointestinal Endoscopy empfiehlt beim Barrett-Ösophagus regelmäßige Überwachung mittels Endoskopie, angepasst an die Länge des Barrett-Segments. Bei Segmenten ab 10 cm Länge wird die Überwachung und Behandlung in einem spezialisierten Zentrum empfohlen (19).

0.2.2 Neuronale Netzwerke

Machine Learning ist ein Teilgebiet der Künstlichen Intelligenz (KI) und beschreibt die Fähigkeit von Computern, aus Erfahrungen zu lernen, ohne explizit programmiert zu werden (24,25). Eine besonders leistungsfähige Methode ist das Deep Learning, das auf tiefen

neuronalen Netzwerken basiert. Diese Netzwerke bestehen aus künstlichen Neuronen, die, ähnlich wie die Neuronen im Gehirn, Informationen von vorgeschalteten Neuronen erhalten, diese verarbeiten und an folgende Neurone weitergeben (26). Mittels dieser Methode können neuronale Netzwerke unabhängig von menschlichem Einfluss aus Rohdaten relevante Merkmale definieren (25,27).

Konvolutionale neuronale Netzwerke sind auf Bilddaten spezialisiert. Sie bestehen aus mehreren Netzwerkschichten (Convolution- und Pooling-Schichten) zur Merkmalsextraktion und nachgeschalteten Schichten für die Klassifikation (28). Wenn zusätzliche Schichten nicht zu einer weiteren Verbesserung des Modells führen, können spezialisierte Architekturen wie z.B. Residuale Neuronale Netzwerke oder U-Nets zu verbesserten Ergebnissen führen (29).

0.2.3 Domain Shift

Ein zentrales Ziel im Machine Learning ist die Entwicklung universell einsetzbarer Algorithmen (30). Eine große Herausforderung ist dabei die sogenannte Generalisierbarkeit, also die Fähigkeit eines Modells, auch bei bisher unbekannten Daten gute Ergebnisse zu liefern (31). Deep-Learning-Modelle zeigen häufig sehr gute Leistungen in der sogenannten Quell-Domäne, also dem Datensatz, auf dem sie trainiert wurden. Werden sie jedoch auf andere Datensätze (Ziel-Domäne) angewendet, sinkt die Leistung oft deutlich, da die Modelle zu stark auf die Trainingsdaten spezialisiert sind (32).

Dieses Problem existiert auch in der Endoskopie: Bilder unterscheiden sich je nach Hersteller des Endoskops, Software oder Aufnahmebedingungen. Während diese Unterschiede für das menschliche Auge minimal sind, können sie für eine künstliche Intelligenz zu einer sehr viel schlechteren Leistung führen. Dieses Phänomen wird als *Domain Shift Problem* bezeichnet (30).

0.2.4 Deep Learning in der Endoskopie

In der Endoskopie konzentriert sich die Forschung zu Computerassistierten Detektionssystemen (CAD-Systemen) auf die Erkennung und Klassifikation von Polypen, Blutungen sowie pathologischen Oberflächenveränderungen, auch in Bezug auf Barrett-Ösophagus (33). Hier wird Künstliche Intelligenz vor allem zur Detektion und Bewertung neoplastischer Läsionen eingesetzt, basierend auf verschiedenen Bild- und Farbgebungstechniken (34–45).

0.3 Material und Methoden

Diese Studie entstand in Zusammenarbeit zwischen den Amsterdam University Medical Centers, dem Krankenhaus Barmherzige Brüder Regensburg sowie der Technischen Universität Eindhoven. Ziel war es, die Übertragbarkeit eines bestehenden CAD-Systems zur Erkennung von Neoplasien bei Barrett-Ösophagus auf neue Bedingungen zu untersuchen, darunter ein unbekanntes Endoskopiesystem, ein neues Krankenhaus und neuer Behandlungsstatus (posttherapeutischer Ösophagus (nach Behandlung)).

Das CAD-System, entwickelt von Groof et al., nutzt Konvolutionale Neuronale Netzwerke (CNN) um während einer Gastroskopie dys- und neoplastische Läsionen zu markieren und die am besten geeignete Stelle für eine Biopsie vorzuschlagen (41).

Für die Vergleiche, auf die sich unsere Studie stützt, wurden vier Testdatensätze mit unterschiedlichen Charakteristika zusammengestellt:

a) Bekanntes Zentrum und Endoskopiesystem (Fujifilm), bekannter Behandlungsstatus (prätherapeutisch) *(Test Set 1)*
b) Bekanntes Zentrum, **neues Endoskopiesystem (Olympus)**, bekannter Behandlungsstatus (prätherapeutisch) *(Test Set 2)*
c) **Neues Zentrum, neues Endoskopiesystem (Olympus)**, bekannter Behandlungsstatus (prätherapeutisch) *(Test Set 3)*
d) **Neues Zentrum, neues Endokopiesysytem (Olympus), neuer Behandlungsstatus (posttherapeutisch)** *(Test Set 4)*

Test Set 1 enthält Bilder aus der externen Validierung des Algorithmus und entspricht sowohl hinsichtlich Zentrum (Amsterdam University Medical Centers) als auch Endoskopiesystem (Fujifilm) den Trainingsbedingungen. *Test Set 2* stammt aus dem ebenfalls bekannten Zentrum in Amsterdam, wurde jedoch mit einem neuen Endoskopiesystem (Olympus) aufgenommen. *Test Set 3* kommt aus einem neuen Zentrum in Regensburg mit ebenfalls neuem Endoskopiesystem (Olympus). *Test Set 4* stammt ebenfalls aus Regensburg, enthält jedoch Bilder nach endoskopischer Behandlung (posttherapeutisch), was das Erscheinungsbild des Ösophagus deutlich verändern kann.

Alle Datensätze umfassen Bilder mit nicht-dysplastischem sowie neoplastischem Barrett-Ösophagus (beinhaltend Adenokarzinom, hochgradige Dysplasie, ohne niedriggradige Dysplasie). Ziel war es, die Performance des CAD-Systems unter drei Bedingungen zu

vergleichen: bei unterschiedlichem Endoskopiesystem (*Test Set 1* vs. *2*), bei Anwendung in einem neuen Zentrum (*Test Set 2* vs. *3*) und bei posttherapeutischen Bildern (*Test Set 3* vs. *4*), um die Anwendbarkeit und Generalisierbarkeit des Algorithmus in realen, klinisch heterogenen Settings zu beurteilen.

Es folgt eine Übersicht über die Zusammensetzung der Testdatensätze mit Anzahl an Bildern und Patient:innen, Anteil an neoplastischen und nicht-dysplastischen (NDBE) Bildern und beteiligten Zentren:

a) *Test Set 1*: 160 Bilder (80 NDBE, 80 neoplastisch; 160 Patient:innen)
   Amsterdam University Medical Centers, Universitair Ziekenhuis Leuven, Centre Hospitalier Universitaire de Nantes, St. Antonius Ziekenhuis, Catharina Ziekenhuis Eindhoven
b) *Test Set 2*: 524 Bilder (391 NDBE, 133 neoplastisch; 97 Patient:innen)
   Amsterdam University Medical Centers
c) *Test Set 3*: 213 Bilder (8 NDBE, 205 neoplastisch; 72 Patient:innen)
   Krankenhaus Barmherzige Brüder, Regensburg
d) *Test Set 4*: 422 Bilder (385 NDBE, 37 neoplastisch; 84 Patient:innen)
   Krankenhaus Barmherzige Brüder, Regensburg

Die genutzten Ein- und Ausschlusskriterien der Bilddaten können der Originalarbeit entnommen werden.

0.4 Ergebnisse

In diesem Abschnitt werden die Resultate der Anwendung des CAD-Systems auf die jeweiligen Test Sets präsentiert.

Die folgende Tabelle gibt eine Übersicht über die wichtigsten Metriken.

Tabelle 1. Performance des CAD-Systems auf unsere Testdatensätze.

	Genau-igkeit	**Sensiti-vität**	**Spezifi-tät**	**F1**	**AUC**	**NPV**	**PPV**	**FNR**	**FPR**
Test Set 1	0,8875	0,9125	0,8625	0,8868	0,8875	0,9079	0,8690	0,0875	0,1375
Test Set 2	0,7462	0,3759	0,8721	0,5254	0,6240	0,8042	0,5	0,6241	0,1279
Test Set 3	0,6901	0,6878	0,75	0,7176	0,7189	0,0857	0,9860	0,3122	0,25
Test Set 4	0,6588	0,6486	0,6597	0,6541	0,6542	0,9513	0,1548	0,3514	0,3403

Abkürzungen: F1 = F1 Score; AUC = Area under the receiver operating characteristic curve; NPV = Negativer prädiktiver Wert; PPV = Positiver prädiktiver Wert; FNR = Falsch-Negativ-Rate; FPR = Falsch-Positiv-Rate; Genauigkeit entspricht *Accuracy*

ROC-Kurven zeigen graphisch den Kompromiss des Modells zwischen richtig-positiven und falsch-positiven Werten bei verschiedenen Schwellenwerten. Dies sagt aus, wie gut das Modell unabhängig des konkret gewählten Schwellenwerts klassifiziert.

AUC (Area Under the ROC Curve) ist die Fläche unter der ROC-Kurve und beschreibt, mit welcher Wahrscheinlichkeit ein Klassifikationsmodell einen zufälligen positiven Fall höher bewertet als einen zufälligen negativen Fall. Damit kann unabhängig vom gewählten Schwellenwert eine Aussage über die Leistung des Modells getroffen werden.

F1 ist in unserem Fall das harmonische Mittel zwischen Sensitivität und Spezifität.

0.4.1 Vergleiche zwischen den Test Sets

Der Algorithmus zeigt auf *Test Set 1* (keine neuen Bedingungen) generell eine bessere Performance als auf *Test Set 2* (neues Endoskopiesystem: Olympus). Die Genauigkeit sinkt um 14%, die Sensitivität um fast 60%, negativer und positiver prädiktiver Wert weisen Unterschiede von etwa 10 % bzw. 37 % auf. In *Test Set 1* wurden 8,7% der neoplastischen Bilder übersehen, in *Test Set 2* war dies bei 62,4% der Fall. Der F1-Score ist bei *Test Set 2* 36 %

niedriger, die AUC fällt um 26 %. Nur die Spezifität bleibt mit ca. 87 % nahezu gleich. Die entsprechenden ROC-Kurven sind in der untenstehenden Grafik dargestellt.

Der Vergleich der Performance auf *Test Set 2* und *Test Set 3* beschreibt zusätzlich zum neuen Endoskopiesystem die Datenherkunft aus einem neuen Zentrum als unbekannten Faktor. Genauigkeit und Spezifität sind bei *Test Set 2* höher, dafür ist die Sensitivität für *Test Set 3* fast doppelt so hoch. Auch F1-Score und AUC fallen bei *Test Set 3* besser aus (19% bzw. 9%). Der negativ prädiktive Wert ist bei *Test Set 2* deutlich höher (80 % vs. 8,6 %), während der positive prädiktive Wert für *Test Set 3* fast 50 % besser ist (98,6 %). Insgesamt wurden in *Test Set 2* rund 62 % der Neoplasien übersehen, in *Test Set 3* etwa 31 %.

Beim Vergleich von *Test Set 3* (prätherapeutisch) und *Test Set 4* (posttherapeutisch) zeigt die Performance des CAD-Systems auf *Test Set 3* insgesamt etwas bessere Werte in fast allen Bereichen. Die Genauigkeit liegt bei 69 % in *Test Set 3* gegenüber 66 % in *Test Set 4*, die Sensitivität ist um 4 % und die Spezifität um 9 % höher. Der F1-Score beträgt 75 % vs. 65 %, die AUC zeigt einen Unterschied von 6 %. Der negative prädiktive Wert ist für *Test Set 4* deutlich höher (95 % vs. 8,6 %), der positive prädiktive Wert dagegen bei *Test Set 3* (98,6 % vs. 15,5 %). Die Rate übersehener Neoplasien ist in beiden Sets ähnlich (31 % vs. 35 %).

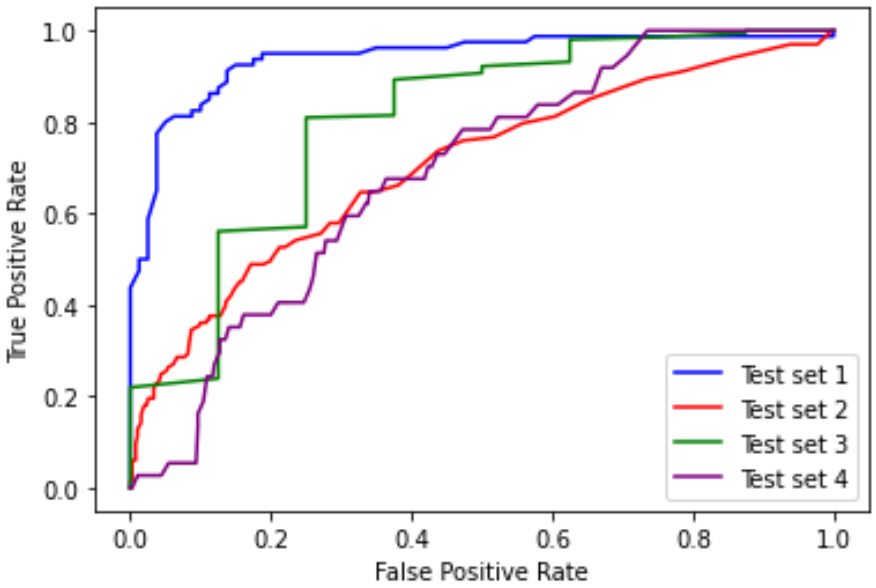


Abbildung 1. ROC-Kurven von *Test Set 1*, *2*, *3* und *4*.

0.5 Diskussion

0.5.1 Veränderte Bedingung: Neues Endoskopiesystem

Die Anwendung unseres CAD-Systems auf Bilder aufgenommen mit einem neuen Endoskopiesystem (Olympus) führte zu einem Leistungsabfall in wichtigen Bereichen wie Sensitivität und positivem prädiktiven Wert.

Sowohl Fujifilm als auch Olympus sind Hersteller qualitativ hochwertiger Endoskopiesysteme. Unterschiede bestehen jedoch in Auflösung, Lichtverhältnissen und Kontrast, was ein unterschiedliches Abschneiden KI-basierter Modelle auf Systemen verschiedener Hersteller erklären könnte (46).

Generell werden die meisten CAD-Systeme für Barrett-Klassifizierung mit Endoskopiebildern von nur einem System trainiert. Dabei nutzen die meisten Studien Olympus- oder Fujifilm-Systeme (37–39,41,42,44), es gibt eine Studie zu einem Pentaxsytem (43). Lediglich Abdelrahim et al. entwickelten ein System basierend auf Fujifilm-, Olympus- und Pentax-Bildgebung (34). Für die endoskopische Erkennung von Barrett-Neoplasien gibt es derzeit zwei für den klinischen Einsatz zugelassene KI-Systeme (WISE VISION und CADU), die beide mit allen großen Endoskopherstellern kompatibel sein sollen. Beide erfüllen EU-Sicherheitsanforderungen (CE-Zertifizierung), aber transparente Informationen zu Entwicklung und Testung fehlen (47,48).

Kritisch ist, dass in *Test Set 2* 62% aller neoplastischen Bilder fälschlicherweise als unauffällig klassifiziert wurden. Ähnliche Fehlerquoten zeigen sich in *Test Set 3* (31,2%) und *4* (35,1%). Falsch-positive Befunde können im klinischen Alltag leicht durch eine risikoarme Biopsie ausgeschlossen werden, während sich verpasste Läsionen zu höheren Tumorstadien mit schlechterer Prognose entwickeln können.

Qualitativ schlechtere Datenbeschriftung in *Test Set 2* als Grund der schwächeren Performance erscheint unwahrscheinlich, da alle Testdaten unabhängig und fachgerecht beschriftet wurden. Auch sind nach menschlicher Wahrnehmung die neoplastischen Läsionen in *Test Set 2* nicht subtiler und damit schwieriger zu erkennen als in *Test Set 1*.

Deshalb gehen wir davon aus, dass der Wechsel zu einem anderen Endoskopie-System (hier: Olympus) für den Leistungsabfall verantwortlich ist.

0.5.2 Veränderte Bedingung: Neues Zentrum

Die Anwendung des CAD-Systems auf Bilder aus einem neuen Zentrum (*Test Set 3*) zeigte überraschenderweise teils bessere Leistungswerte als bei *Test Set 2*, obwohl sowohl das Endoskopiesystem als auch das Zentrum neu waren. Im Vergleich zu *Test Set 1* war jedoch ein Leistungsabfall sichtbar, was auf beide Faktoren zurückzuführen sein könnte. Die Ergebnisse deuten darauf hin, dass der Wechsel des Endoskopiesystems die größere Rolle spielt. Durch die Schiefe von 0,038 von *Test Set 3* mit nur 8 nicht-dysplastischen Bildern ist der informative Wert der Ergebnisse jedoch begrenzt.

Ein möglicher Grund für den Leistungsabfall ist *Overfitting*, eine Überanpassung an das Trainingsset. Dadurch kann das System neue Daten schlechter verarbeiten (49). Um Overfitting zu vermeiden, kommen technische Maßnahmen wie Batch Normalization, Datenaugmentation oder eine vereinfachte Architektur zum Einsatz (28), wie auch bei unserem CAD-System geschehen (41). Vor allem aber größere und heterogenere Datensätze für das Training sind hilfreich (50).

Studien wie von Abdelrahim et al. und Hussein et al. nutzten Trainingsdaten aus mehreren Zentren und mit verschiedenen Endoskopiesystemen, um die Generalisierbarkeit zu verbessern (34,43). Fockens et al. erstellten das bisher größte und diverseste Datenset mit über 5 Millionen allgemeinen Endoskopiebildern zum *Pretraining* und fast 14 000 Barrett-spezifischen Bildern aus 15 Zentren in 7 Ländern (37).

Alternativ könnte ein CAD-System gezielt auf das jeweilige Zentrum angepasst werden, was jedoch zusätzliche Ressourcen erfordert.

Die extreme Werte des positiven (98,6%) und negativen (8,6%) prädiktiven Wertes in *Test Set 3* variieren mit Krankheitsprävalenz (bei uns die Schiefe des Test Sets) und werden deshalb nicht weiter bewertet.

0.5.3 Geänderte Bedingung: Neuer Behandlungsstatus (posttherapeutisch)

Unser CAD-System stuft posttherapeutische nicht-dysplastische Bilder häufiger fälschlich als neoplastisch ein (34% in *Test Set 4*, verglichen mit 25% in *Test Set 3* und 12,8% in *Test Set 2*). Ein direkter Vergleich zwischen *Test Set 3* und *4* ist aufgrund der sehr geringen Anzahl an NDBE-Bildern in *Test Set 3* jedoch nur eingeschränkt aussagekräftig.

Der positive (15,5 %) und negative (95,1%) prädiktive Wert sind aufgrund der starken Ungleichverteilung der Daten (mehr nicht-dysplastische als neoplastische Bilder) schwer zu interpretieren, spiegeln aber auch die Realität wider: Auch im klinischen Alltag ist die Inzidenz von hochgradiger Dysplasie und Adenokarzinom bei Barrett-Patient:innen sehr niedrig (~1 % pro Jahr) (51).

Fockens et al. nutzen als neuen Ansatz gezielt ein unausgewogenes Test Set, um die klinische Realität besser abzubilden. Ihr positiver prädiktiver Wert lag bei 45% (38). Wegen unterschiedlicher Charakteristika der Test Sets sind Werte jedoch nicht direkt vergleichbar.

0.5.4 Das Domain Shift Problem

Leistungsabfälle durch sogenannte *Domain Shifts* sind in der KI-Forschung ein bekanntes Problem und wurden auch in dieser Studie festgestellt. Die neuen Domänen sind in unserem Fall Veränderungen im Bildmaterial durch ein neues Endoskopiesystem, ein neues medizinisches Zentrum sowie posttherapeutische Veränderungen des Ösophagus.

Verschiedene Arten von Domain Shift liegen vor.
Covariate Shift beschreibt Veränderungen im Input, und besteht bei unserem System in Unterschieden in der Bildaufnahme (z.B. neues Kamerasystem, klinikeigene Untersuchungsabläufe) und einer anderen Patientenkohorte. *Label Shift* bezeichnet die Veränderung der Verteilung der Outcomes, und liegt ebenfalls vor, da sich das Verhältnis von neoplastischen zu nicht-neoplastischen Bildern zwischen den Test Sets ändert. *Concept Shift* tritt auf, wenn sich die Beziehung von Input zu Output ändert, was vor allem in *Test Set 4* vorliegt. Vorliegende Schleimhautänderungen sind durch vorherige invasive Therapie nicht mehr zwingend neoplastische Läsionen.

Diese Shifts sind Hauptfaktoren für die nachlassende CAD-Performance. Lösungen sind größere, heterogene Datensätze, idealerweise prospektiv aus verschiedenen Zentren erhoben und in klar getrennte Trainings-, Validierungs- und Testsätze aufgeteilt (52).

Neue Ansätze wie die Metrik *Representation Shift* quantifizieren Domain Shifts und können die Generalisierbarkeit eines Modells besser bewerten (31). *Unsupervised Domain Adaption*, also das Training mit nicht gelabelten Ziel-Daten oder KI-generierten Bildern, zeigt vielversprechende Ergebnisse (53), und wurde auch in der Segmentierung bei Barrett-Ösophagus bereits erfolgreich angewendet (54).

0.5.5 Externe Validierung

Ein zentrales Problem für die breite Anwendung von KI-Systemen in der Medizin sind fehlende geeigneten Studienformate für objektive Vergleiche und Bewertungen. Randomisierte, verblindete Studien sind schwer mit KI umsetzbar (9). Auch externe Validierung bleibt selten: Nur 55 % der KI-Studien zur Diagnostik oberer gastrointestinaler Neoplasien wurden external validiert, und nur 8% in realer klinischer Umgebung getestet (55).

Die Vergleichbarkeit verschiedener CNN-Modelle ist eingeschränkt, da jede Arbeitsgruppe normalerweise ihren eigenen Bild- und Testdatensatz mit eigenen und oft nicht beschriebenen Charakteristika hat und Ergebnisse nicht einheitlich präsentiert werden.

Auch im Bereich KI-gestützter Detektion von Barrett-Neoplasien ist externale Validierung rar, folgende Studien stechen durch ihre Verfahren hervor: das hier genutzte System von Groof et al. führte doppelte externe Validierung und Benchmarking mit nicht-spezialisierten Endoskopiker:inenn durch (41). Abdelrahim et al. validierten ihr Modell external auf Videos aus drei externen europäischen Zentren, auch zeitlich unabhängig von Trainingsdaten mit zusätzlichem Benchmarking (34). Fockens et al. erstellten Test Sets mit verschiedenen Charakteristika und benchmarkten sowohl Expert:innen als auch nicht-spezialisierte Endoskopiker:innen mit und ohne CAD-Unterstützung (37,38).

Unsere Studie zeigt, dass selbst ein gut funktionierendes, external validiertes System bei neuen Daten deutlich schlechter abschneiden kann. Dennoch ist eine weitreichendere externe Validierung essenziell für Aussagen zur Generalisierbarkeit und sicheren Anwendung.

0.5.6 Gemeinsame Datensätze

Diese Studie zeigt, dass es für robuste KI-Modelle größere, heterogenere Datensätze aus Zusammenarbeit möglichst diverser Zentren braucht. Datenschutz und begrenzte Kapazitäten von Expert:innen erschweren jedoch die Datengewinnung im medizinischen Bereich.

Erste Konsortien wie ARGOS (56) und BONS-AI (37,38) bündeln Daten und Wissen mehrerer Zentren. Dieser Ansatz der internationalen Zusammenarbeit sollte in großem Stil weltweit weiterverfolgt werden. Öffentliche, auf Endoskopie spezialisierte Bilddatenbanken wie GastroNet (37) und HyperKvasir (57) fördern die Entwicklung guter Systeme, benötigen aber breitere Beteiligung.

Ein Anreiz könnte sein, wissenschaftliche Credits für Veröffentlichung von Datensätzen zu bekommen (58). Alternativ ermöglichen technische Methoden wie Föderales Lernen oder Swarm Learning ein Training auf geteilten Daten zwischen Zentren, aber ohne zentrale Datenspeicherung (58,59).

0.5.7 Rahmenbedingungen für Künstliche Intelligenz

Während dieser Studie ist die eingeschränkte Vergleichbarkeit mit anderen Arbeiten aufgefallen. Unterschiedliche Testdatensätze, unklare Metriken und fehlende Angaben zur Datengewinnung und -auswahl erschweren objektive Vergleiche. Studien sollten deshalb transparent und möglichst detailliert dokumentieren, woher Daten stammen, wie sie aufbereitet wurden und welche Patientengruppen vertreten sind.

Die FUTURE-AI-Guidelines bieten bereits ein Rahmenkonzept mit Leitlinien für eine verlässlichere und sichere Entwicklung und Implementierung von KI in der Gesundheitsversorgung (60). Auch Fachgesellschaften wie die American Society for Gastrointestinal Endoscopy geben klare Leistungskriterien für KI-Systeme vor. Solche Standards sollten helfen, Vergleichbarkeit und Sicherheit zu schaffen, ohne Innovationsfreiheit einzuschränken.

Außerdem sollte überlegt werden, ob nicht eine Testung von Endoskopiker:innen mit und ohne CAD-Unterstützung aussagekräftiger für den klinischen Nutzen ist, als die alleinige Angabe von CAD-Performance. Dies wurde bisher alleinig von Fockens et al. getestet (37). Auch ein unabhängiger Benchmarking-Datensatz, auf dem jedes System final getestet wird, könnte für objektivere Vergleiche sorgen (52).

0.5.8 Limitationen

Unsere Studie weist mehrere Einschränkungen auf. Bilder mit niedriggradiger Dysplasie wurden ausgeschlossen trotz klinischer Relevanz. *Test Set 3* besteht aus nur acht NDBE-Bildern, was die Aussagekraft der Ergebnisse limitiert. Unsere Testdatensätze sind unterschiedlich ausgewogen (NDBE vs. neoplastisch), was die Vergleichbarkeit der positiven und negativen prädiktiven Werte zwischen den Sets erschwert. Der Algorithmus wurde nicht mit posttherapeutischen Bildern trainiert, obwohl diese im *Test Set 4* enthalten waren, sodass Zufallsergebnisse nicht ausschließbar sind. Die gemessenen Werte wurden nicht statistisch getestet und es wurde keine Power-Analyse durchgeführt, weshalb wir die Ergebnisse nicht

sicher als signifikant werten können. Es wurde nur ein einzelner Algorithmus untersucht, weshalb sich die Ergebnisse nicht ohne Weiteres auf andere Modelle übertragen lassen.

0.6 Fazit

Diese Studie untersuchte die Generalisierbarkeit eines KI-Systems zur Detektion von Barrett-Neoplasien auf unbekannten Domänen. Die Leistung sank bei unbekannten Faktoren, wie bei Bildern eines neuen Endoskopiesystems, aus anderen Zentren und bei dem Modell unbekanntem posttherapeutischem Behandlungsstatus. Wahrscheinliche Ursachen sind Overfitting, unzureichendes spezifisches Training und Domain Shifts. Für die weitere Entwicklung sollten vielfältigere Daten (z. B. verschiedene Endoskopiesysteme und Behandlungsstadien, mehrere Zentren) einbezogen werden. Erklärbare KI, lokale Modellanpassung und mehr Zusammenarbeit zwischen Forschungsgruppen könnten Prozesse und Ergebnisse verbessern.

Diversere, größere Trainings- und Testdatensätze, standardisierte Rahmendbedingungen und weitreichendere externe Validierung sollten angestrebt werden und sind essentiell für eine verantwortungsbewusste und sichere klinische Anwendung.

Abbreviations

AI	artificial intelligence
AUC	area under the curve
CAD	computer-aided detection
CNN	convolutional neural network
EAC	esophageal adenocarcinoma
ESGE	European Society for Gastrointestinal Endoscopy
HGD	high-grade dysplasia
LGD	low-grade dysplasia
NBI	narrow band imaging
NDBE	non-dysplastic Barrett's esophagus
ROC	receiver operating characteristic
WLE	white light endoscopy

1 Introduction

Esophageal cancer is the eleventh most common cancer worldwide and ranges seventh place regarding mortality, causing around 445 000 deaths each year (1). Although esophageal squamous cell carcinoma is still more common, prevalence of esophageal adenocarcinoma is rising. Among high-income countries, esophageal adenocarcinoma account already for two-thirds of esophageal cancer (1). Due to alterations in lifestyle, we can expect a further increase throughout the world. Therefore, research in esophageal adenocarcinoma is essential to prevent future generations from this trend in rising incidence and mortality (1,61).

Over the past years, various improvements in endoscopic detection of neoplasia have already been developed. Different colour enhancements like chromoendoscopy intensify the esophageal mucosa's surface structure and vascularization - either dye-based or virtually (2,62,63). Volumetric laser endomicroscopy creates high-resolution cross-sectional images of the esophagus using infrared light (64) and confocal laser endomicroscopy allows real-time optical biopsy (65,66).

These advanced imaging technologies work on making everything possible visible. But now, that imaging has significantly been improved, we still need to recognize what we see. The technologies have to be used, and for that, we need people with knowhow. Currently, this seems to be a problem. "Technology has evolved quicker than our ability as a society and community to create robust, high-quality educational programs for endoscopists of all levels.", state Kolb and Wani (2). Due to the low prevalence of early neoplasia in the esophagus, general endoscopists are not familiar with the detection. Neoplastic lesions are difficult to recognize, so despite all the technological improvements, non-specialized endoscopists might easily miss them (3,4).

There are two solutions conceivable: Either, we offer better training for non-expert endoscopists, together with improvement at the expert level (2). Or else, we further implement artificial intelligence (AI) in the field. An AI system can be trained quickly on specific tasks like different new imaging modalities (5).

Artificial intelligence has some obvious advantages compared to human endoscopists: factors like tiredness or loss in concentration after long working days, mental wellbeing, time pressure and challenges in intra- and interobserver agreement do not play a role (6). AI standardizes diagnostic procedures independent of of the endoscopist's experience, making endoscopy more reliable and faster (7).

This could be particularly important for surveillance, which often takes place in non-expert centers or community hospitals. Although Barrett's esophagus treatment is recommended to do in expert centers, the first points of care are often non-specialized institutions with less experienced staff. To keep a potential invasive treatment as minimal as possible, it is essential to detect the subtle lesions at an early stage during surveillance (8). This can be guaranteed by referral to expert centers – or potentially by future support of inexperienced endoscopists by AI systems.

Still, limitations exist. AI systems are often trained in a single center. This may bias the performance of the system - possibly even subconsciously - due to institutional habits. Additionally, each center treats an individual patient population which may differ from centers in other locations. All this may not make a big difference to educated endoscopists, but it does to a specifically trained algorithm. Its performance can become unpredictable when used in new settings. To adopt a computer-aided detection (CAD) system in general clinical use, it has to show consistency and generalizability. It has to be able to deal with different endoscopists, endoscopic equipment, clinical software and patient selections (9).

Only if this can be guaranteed, a universal use in non-expert centers or community hospitals to support general endoscopists would be possible. This shall be evaluated in this thesis.

1.1 Objectives of the Dissertation

With this thesis, we aim to make a statement regarding the following:

We will test the performance of an existing CAD system on different untrained domains. In this process, we will stepwise add for the algorithm new elements. First, we change the endoscopy system used for image acquisition from Fujifilm to Olympus, then the centers of data recruitment from known to unknown to the algorithm. In the final step, we will change the treatment status of the esophagus in the imagery, moving from pretreatment to posttreatment cases.

Objective: We want to evaluate the generalizability of the CAD system when applied to images that differ to the training and validation data in (I) endoscopy system, (II) data acquisition center, and (III) treatment status of the esophagus.

General objective: We want to get an idea, to which extent the algorithm is applicable in a new setting, where posttreatment surveillance and new image sources are usual.

2 Theoretical Background

2.1 Barrett's Esophagus

Barrett's esophagus describes the transformation from physiological esophageal squamous epithelium to columnar epithelium. This metaplastic process often takes place during the process of healing after an acute injury to the esophageal mucosa. These injuries usually arise from gastric acid damaging the esophageal tissue. The tissue then heals to columnar epithelium, which better protects from the acidic environment (10,11,15).

Barrett's esophagus shows a prevalence of about 5,57% in the general population - but with a wide range in study results from 1,6% in Sweden up to 6,8% in the US. The numbers could be even higher in the older generation (12–14).

The development of this metaplastic process is favoured by chronic gastroesophageal reflux disease, male gender, central obesity, positive family history of Barrett's esophagus or esophageal adenocarcinoma and smoking (14,67). For the first four risk factors, Qumseya et al. were able to show an increase about 1,2% per additional factor (68). In general, prevalence for gastroesophageal reflux disease and Barrett's esophagus is higher in Western populations (69).

The risk with Barrett's esophagus is the progression to esophageal or gastroesophageal junction adenocarcinoma, which are currently difficult to treat and should therefore be prevented (15,16).

2.1.1 Diagnostics

Barrett's esophagus is the only known precursor for esophageal adenocarcinoma (11,15,16). Thus, early diagnosis is essential. The European Society of Gastrointestinal Endoscopy's parameters for the diagnosis of Barrett's esophagus are columnar epithelium in the distal esophagus, a minimum metaplasia length of 1 cm and the occurrence of specialized intestinal metaplasia in the histopathologic sample (70).

That means, endoscopy is necessary for diagnosis - firstly, to optically detect columnar epithelium in the distal esophagus and secondly, to take a biopsy for histopathological confirmation of specialized intestinal metaplasia (11).

The endoscopist has to identify the histological border between squamous epithelium (esophagus) and columnar epithelium (stomach). Additional columnar epithelium caused by

Barrett's syndrome shifts this separation further proximal. The normally clear, circumferential line then transitions to an irregular, frayed line (11,15).

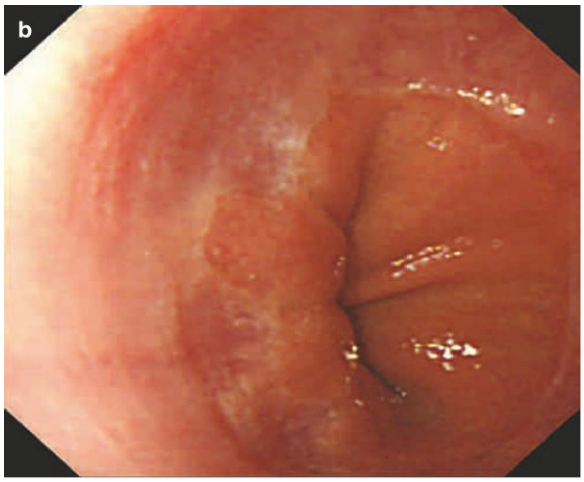


Figure 1. Physiological gastroesophageal junction. Landmark for the gastroesophageal junction is the proximal end of the gastric folds. Here, it is similar to the histological border; visible as separating line between reddish columnar epithelium which belongs to the stomach and whitish esophageal squamous epithelium (128).

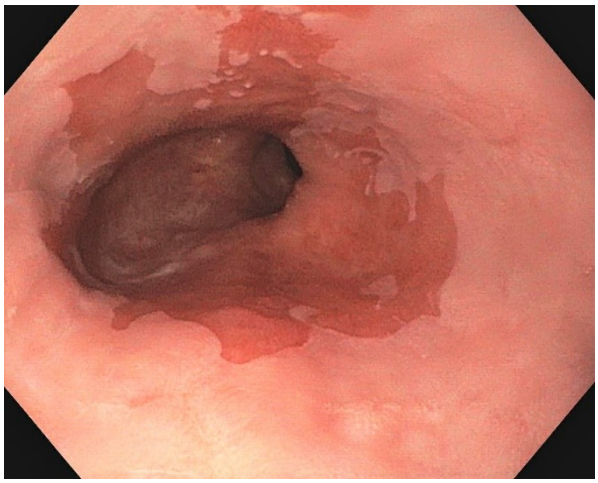


Figure 2. Barrett's esophagus. The salmon-coloured squamous epithelium is pierced by a reddish tongue of Barrett's epithelium. The separating line is clearly irregular and frayed.

Endoscopy can be a demanding process. Challenges in the endoscopic evaluation of esophagus and stomach are motility, hiatal hernia, and the handling of insufflation. Over-insufflation can cause errors like flat gastric folds (15).

For an examination as objective as possible, maximal vertical extent and circumferential extent of metaplasia are measured and documented, called the Prague Classification. Also, metaplasia gets classified into short segment Barrett's esophagus with vertical extension below 3 cm or long segment Barrett's esophagus above 3 cm (11,71).

2.1.2 Esophageal Adenocarcinoma

Several steps lead from Barrett's metaplasia to adenocarcinoma. In a metaplastic process, a stimulus brings a differentiated tissue to transform into another differentiated tissue (72). This can progress to dysplasia, defined as "neoplastic epithelium that is confined to the basement membrane" (15). Neoplasia describes the autonomous new formation of tissue without regulation (72).

Esophageal adenocarcinoma develop after passing through various stages of dysplasia. The metaplasia in Barrett's esophagus is followed by low-grade dysplasia, high-grade dysplasia and then potentially progresses to intramucosal or invasive carcinoma (2,10).

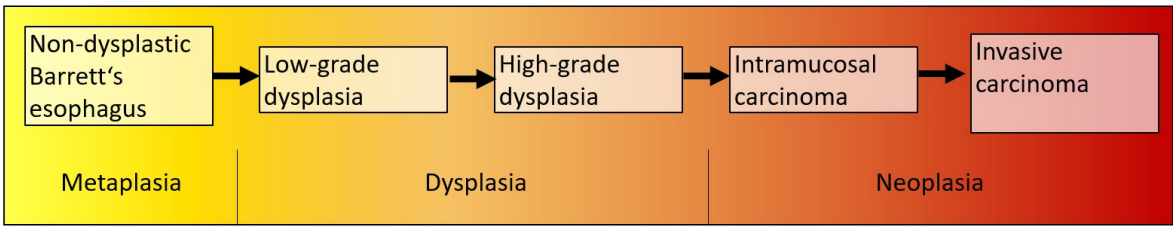


Figure 3. Progression of Barrett's esophagus.

In Western populations, the incidence of adenocarcinoma increased 10-fold in the past 25 years (16,17,73). Due to alterations in lifestyle, numbers of adenocarcinoma are rising and occur in high-income countries nowadays more often than squamous cell carcinoma (61).

In 1000 Barrett's esophagus patients, about 6 will develop an adenocarcinoma (51,74). Low-grade dysplasia progresses to high-grade dysplasia with a probability of 26,5%, of whom then 19-59% develop to cancer (75,76).

Esophageal adenocarcinoma have a poor prognosis, showing an average survival time after diagnosis of 0,8 years and a 5-year all-cause mortality rate of 89% (18).

2.1.3 Treatment

Barrett's esophagus can be treated through medication, endoscopic resection, ablative therapies, or surgery, depending on the presence and stage of lesions.

Endoscopic resection is the preferred treatment for mucosal adenocarcinoma (T1a carcinoma), followed by ablation of the remaining Barrett's mucosa (19), achieving a 93,4% neoplasia

eradication rate (77). For submucosal tumors (T1b), treatment choice should be discussed individually in an interdisciplinary setting, depending on infiltration depth, differentiation grade, and vascular involvement. Additional treatment can involve chemotherapy, radiotherapy or surgery (19). For further advanced stages, surgery is option of choice for resection (20).

Endoscopic resection is minimally invasive, allows for histopathological assessment, and offers high long-term remission rates (78). However, repeated surveillance endoscopies and possibly further treatments are needed as short-term recurrence is possible (22).

Surgical treatment (esophagectomy) ensures the lowest recurrence (23) but comes with high complication risks: over a third of patients face serious complications (22), nearly 10% needed reoperation in an English study, and the 90-day mortality rate remains at 4.1% (79,80).

Despite the effectiveness of surgery, early cancer detection is key to avoiding more invasive and riskier options.

2.1.4 Surveillance Guidelines

The European Society of Gastrointestinal Endoscopy (ESGE) recommends endoscopic surveillance of Barrett's esophagus (70). ESGE's proposed surveillance intervals depend on the length of the Barrett's segment and are shown for non-dysplastic Barrett's esophagus in the following table:

Table 1. Surveillance intervals for Barrett's esophagus recommended by the European Society of Gastrointestinal Endoscopy (19,70).

Length Barrett's segment	Surveillance
<1 cm	no endoscopic surveillance, no routine biopsies
1 cm - 3 cm	surveillance interval of 5 years
3 cm - 10 cm	surveillance interval of 3 years
>10 cm	referral to expert center

Additionally, biopsies should be taken first from visible mucosal lesions and abnormalities and second as part of a 4-quadrant biopsy according to the Seattle protocol. In this procedure, random samples from the Barrett mucosa are taken in an interval of 2 cm. An expert pathologist then has to confirm a diagnosis of dysplasia (70,81).

In case of detected or successfully eradicated dysplasia, surveillance intervals are kept shorter for the first 5-10 years (19).

Routine surveillance procedures can be carried out in community hospitals. In case of a Barrett's segment >10 cm, presence of dysplasia (low-grade dysplasia, high-grade dysplasia, early cancer) and for the first endoscopic follow-up after endoscopic eradication of neoplasia, ESGE recommends treatment in an expert center (19).

2.2 Introduction to Neural Networks

Machine learning is a branch of artificial intelligence and is defined by the capability of a computer to learn from experience, without being explicitly programmed (24,25).

A powerful type of machine learning is deep learning with neural networks as its core technology. Neural networks are inspired by the human brain. Just like biological neurons receive inputs via dendrites from multiple other neurons, categorize it into a single signal and then transmit it, artificial neurons receive inputs, process them, and pass the output to other neurons in the network (26). When a neural network has more than one hidden layer between the input and output, it is called a deep neural network - the foundation of deep learning (82).

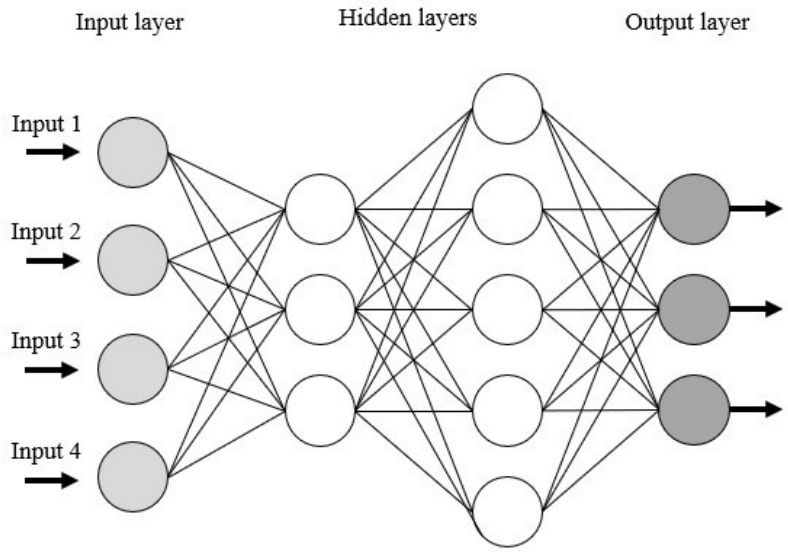


Figure 4. Simplified working mechanism of a neural network.

What makes these deep neural networks special is their ability to independently learn features from raw data without human input. In traditional machine learning, human experts had to

define relevant features manually. In contrast, neural networks identify features on their own through a process of training (25,27). This offers a more detailed analysis and more abstract and complex representations of data (82).

2.2.1 Convolutional Neural Networks

Convolutional neural networks (CNN) are a special type of neural network designed to process data with grid pattern, such as images. They are especially used in image recognition tasks, due to their ability to process data presented in multiple arrays. In imaging, data consists of three arrays for the three colour channels with pixel values presented as numbers in a two-dimensional grid (28).

A typical CNN is composed of three types of layers: Convolutional and pooling layers are repeated multiple times and extract features from the input. These are followed by one or more fully connected layers, which summarize extracted features into the final output (27).

At some point, additional layers do not lead to further performance improvement anymore. Instead, it has been shown that different deep architectures are more effective (29).
Two such approaches used in this thesis are briefly introduced here:

- Residual neural networks (ResNets) use short cut connections to let the network learn only the difference (residual) between layers, leading to a more effective training of the neural network (29,83).
- U-Nets are designed for segmentation tasks. Segmentation means object recognition; pixels are divided into belonging or not belonging to the object of interest. A U-Net architecture allows the network to combine detailed and global image information and thereby to keep a better overview over the whole image (83,84).

2.3 Domain Shift

In machine learning, it is usually the goal to develop a universal algorithm that can be used in different settings worldwide (30). One key challenge in achieving this is ensuring generalization, meaning the ability of the model to perform well on unseen data, which was not part of the training set. A model trained on a specific type or domain of data typically struggles to perform well when applied to an entirely different domain (31). Especially in deep learning, models can be very specialized to the dataset or domain on which they have been trained, the so-called *source domain*. There, they show strong performances by being able to distinguish the smallest details - but sometimes these algorithms are too specialized, having then difficulties and performance drops when applied on another dataset or domain, the *target domain*. Thus, when a model works well in the source domain, it is not given that it shows the same performance in the target domain (32).

In endoscopy, images can differ considerably depending on the endoscopy company, the software used or the type of filming. To the human eye, the differences may be minimal. However, artificial intelligence is trained and tested on a limited, often homogeneous amount of data. If the AI is now used in a new environment that does not correspond exactly to the training and test conditions (which is usually the case in real life application), this can significantly limit the performance. The AI is used on a new domain, that is why to this situation is referred as *domain shift problem*. Additionally, different distributions in the source and target domain can cause problems (30).

2.4 Use of Deep Learning in Endoscopy

Fields of research for CAD systems in endoscopy are mainly detection and classification of polyps, gastrointestinal bleeding and differentiation between pathologic and physiologic regions, but also regarding Barrett's esophagus, different approaches are being developed (33).

In the field of Barrett's esophagus, AI is mainly used for support in detecting and evaluating neoplastic lesions. Algorithms have been developed for endoscopy with different colouring and imaging techniques like white light endoscopy (WLE) (41,45,85), narrow band imaging (NBI) (86–88) or volumetric laser endomicroscopy (89,90), T staging of neoplastic lesions (91,92), real-time pathology (93,94), histology (95) and segmentation in endoscopic ultrasound (96).

In early endoscopic detection of dysplasia and adenocarcinoma, there are different promising approaches.

The research group around Mendel and Ebigbo was among the first to demonstrate the high potential of AI in detecting neoplasia in Barrett's esophagus. They conducted multiple studies on still images with small, single-center datasets using a combined CNN-ResNet architecture (35,45). A follow-up publication based on the same deep learning system presented one of the first uses in real-time in the endoscopic suite. The model was used on a live video working with image freezes, and showing results of a sensitivity of 83,7% and specificity of 100%. The algorithm was trained on a small, internal dataset without videos and not validated externally (36).

Gathwary et al. compared different deep learning methods used for CNN-based development of CAD models. Their most successful model performed with a sensitivity of 96% and specificity of 92% on 100 still images (40).

Hashimoto et al. conducted a pilot study for detection of early neoplasia using CNN architectures. Their system runs at high speed, which could facilitate use in real time, but lacks external validation. Sensitivity is at 96,4%, specificity at 94,2% and accuracy comes to 95,4% for binary classification, including WLE, NBI, standard focus and near focus (42).

The Dutch-led ARGOS consortium and the subsequently emerged more international BONS-AI (Barrett's Oesophagus Imaging for Artificial Intelligence) consortium have published several approaches.

The CAD system developed by Groof et al., which is also used for this thesis, was designed as hybrid CNN-ResNet-U-Net architecture. It categorizes WLE imagery and shows results of 88% accuracy, 88% sensitivity and 89% specificity for classification. The system was externally validated and outperformed non-expert endoscopists (41).

The same research group used a CAD system with again ResNet-U-Net architecture for diagnosing neoplastic or non-dysplastic lesions in Barrett's esophagus in NBI zoom imagery. The algorithm got tested on images and on videos, which makes it more realistic for a real-life setting. The performance on NBI-zoom images shows an accuracy of 84%, sensitivity of 88% and specificity 78%, while the values for NBI videos come to 83%, 85% and 83% respectively (44).

In a later work, the consortium presented another CNN-based model which was trained on one of the most heterogenous and largest Barrett's esophagus image datasets to date. It was then tested on different test sets representing different clinical settings, for example imitation of application in surveillance or more general clinical conditions, and also on prospectively collected images. The model was also benchmarked with general endoscopists. It showed for all test sets a high sensitivity (between 84% and 100%) and for every test set a specificity of 65% (38).

In a following publication, the group expanded the dataset and additionally trained the developed CAD model on videos. Also, they included a new benchmarking process testing the performance of endoscopists with and without additional help of the AI. Through their CAD model, general endoscopists improved their detection of neoplastic lesions from 74% to 88% (37). Subsequently, they tested the model during live endoscopic procedures in the endoscopic suite on a small patient number, resulting in a sensitivity of 100% and a specificity of 65% (39).

The first classification and segmentation CNN-model based on a Pentax endoscopy system has been developed by Hussein et al. in an international multi-center study. The model, being the first one based on a Pentax endoscopy system and including virtual chromoendoscopy technique, was trained on prospectively collected images and video frames. The sensitivity and specificity were 91% and 79% for virtual chromoendoscopy, and 92% and 73% for WLE (43).

Abdelrahim et al. developed a CNN-based CAD system particularly for real-time video analysis. The model was developed using a large image and video dataset from different international centers and including multiple different endoscopic systems. The group performed external image and video validation with results of 95,3% sensitivity and 94,5% specificity for images and 93,8% sensitivity and 90,7% specificity for videos (34).

3 Methodology

3.1 Study Setting and Design

This study was performed as collaboration between the Department of Gastroenterology and Hepatology of the Amsterdam University Medical Centers, location VUmc, the Netherlands, the Department of Gastroenterology and Interventional Endoscopy of the Hospital Barmherzige Brüder in Regensburg, Germany and the Department of Electrical Engineering of the Eindhoven University of Technology, the Netherlands.

The objective was to evaluate the generalizability of an existing CAD system using imagery from an endoscopic system previously unknown to the algorithm, a new medical center and an unfamiliar treatment status. This status is referred to as *posttreatment* esophagus, describing a Barrett's esophagus after treatment for neoplasia.

The mentioned CAD system was designed to support endoscopists in the early detection of neoplasia as a consequence of Barrett's esophagus. Using convolutional neural networks, it is able to mark areas containing neoplastic lesions in endoscopic images and to propose the best suitable spot for a biopsy. The CAD system was already fully developed and validated by Groof et al. (41).

For our study, we collected four different test sets with different characteristics. To carry out the comparisons our study is based on, we needed image sets from

a) the same medical center and the same endoscopic system the algorithm was trained on, same treatment status (pretreatment) *(Test set 1*)
b) the same medical center but different endoscopic system than the algorithm was trained on, same treatment status (pretreatment) (*Test set 2*)
c) a different medical center and different endoscopic system, same treatment status (pretreatment) (*Test set 3*)
d) a different medical center and different endoscopic system, new treatment status (posttreatment) (*Test set 4*)

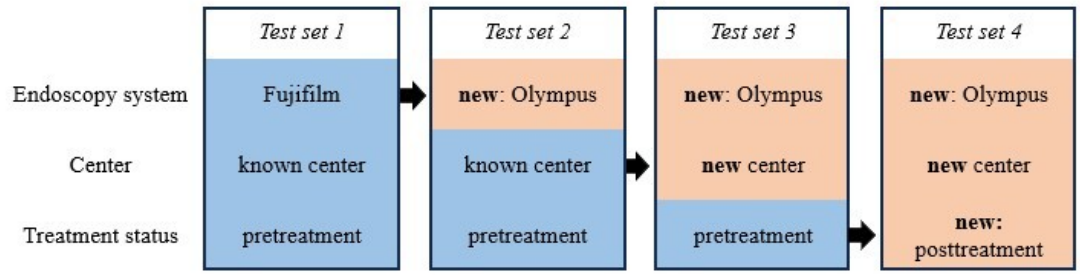


Figure 5. Overview of the characteristics of the test sets.

The first test set consists of images from the external validation sets used in the development of the algorithm. The CAD system was partly trained with images from these centers and this endoscopy system, Fujifilm. Both factors were stable compared to the training of the algorithm. The second test set was collected in the Amsterdam University Medical Centers, the Netherlands, which also contributed images to the training, internal and external validation sets and was therefore known to the algorithm. However, the images from *Test set 2* were from an Olympus endoscopic system (instead of Fujifilm like before) - a new factor to the algorithm. *Test set 3* comes from the hospital Barmherzige Brüder in Regensburg, Germany. It combines a new medical center and a to the CAD system new endoscopic system as also an Olympus system was used in the hospital. *Test set 4*, which was also collected in Regensburg, shows the same characteristics as *Test set 3* and consists additionally only of posttreatment images. This means, any kind of endoscopic treatment (ablation, endoscopic resection, esophagectomy) has been carried out in the esophagus, resulting in an altered appearance.

Each test set contains non-dysplastic and neoplastic images of the Barrett's esophagus. The CAD system classified the images of each test set into either non-dysplastic Barrett's esophagus (NDBE) or neoplastic Barrett's esophagus. Images with esophageal adenocarcinoma (EAC) and high-grade dysplasia (HGD) were considered neoplastic, low-grade dysplastic lesions (LGD) were excluded. From these predictions, we calculated several performance metrics. These metrics were used to compare

1. *Test set 1* and *Test set 2* to evaluate if a different endoscopic system impacts the algorithm's performance (Fujifilm vs. Olympus endoscopy system)
2. *Test set 2* and *Test set 3* to evaluate to which extent imagery obtained in unknown medical centers influences the algorithm's performance

3. *Test set 3* and *Test set 4* to investigate how the algorithm deals with posttreatment images instead of pretreatment images, which it was trained on

Finally, based on the results of the mentioned comparisons, we wish to make a statement on the applicability and generalizability of the CAD system in a non-expert setting.

3.2 Design of the Computer Aided Detection System

The CAD system, often simply referred to as *algorithm* or *model* in this manuscript, was designed by the research group of Groof et al. It was created to support endoscopists by highlighting areas containing early neoplasia. For this serves an individually designed fully-residual, hybrid ResNet-U-Net model. This combination supports both classification and segmentation. The CAD system is able to outline the region of the neoplastic lesion and to mark the best spot for a biopsy.

The network was developed and trained on five different datasets, which are summarized in the following table.

Table 2. Overview of datasets for training and validation of the used CAD system.

Dataset and purpose	**Number of images**	**Endoscopy system**	**Centers of recruitment**
Dataset 1: Pretraining	494,364	Olympus HQ190, HQ180, HQ260, Fujifilm 700 series	Amsterdam University Medical Centers
Dataset 2: Refinement training	1247 (690 NDBE, 557 neoplastic)	Olympus HQ190, HQ180, HQ260	Amsterdam University Medical Centers
Dataset 3: Refinement training and internal validation	297 (168 NDBE, 129 neoplastic)	Fujifilm 700 series	Amsterdam University Medical Centers, Catharina Hospital Eindhoven
Dataset 4: External validation	80 (40 NDBE, 40 neoplastic)	Fujifilm 700 series	Amsterdam University Medical Centers, Catharina Hospital Eindhoven,

			University Hospital Leuven, University Hospital Nantes
Dataset 5: External validation	80 (40 NDBE, 40 neoplastic)	Fujifilm 700 series	Amsterdam University Medical Centers, Catharina Hospital Eindhoven, University Hospital Leuven

Abbreviations: NDBE = non-dysplastic Barrett's esophagus, CAD system = computer aided detection system, HQ = high quality

The pretraining was carried out with a specific gastrointestinal dataset. In this step, the CAD system was supposed to gain knowledge about general image features like edges and colours. This part has been conducted via semi-supervised learning.
In the following refinement training, the algorithm got trained with Barrett's esophagus specific images for the first time. With *Dataset 3*, refinement and tuning of parameters were continued. For internal validation, 4-fold-cross-validation was applied, meaning the validation data got split into 4 parts, of which one was used for validation and the other 3 for further training. This was repeated with 4 different validation sets.
The algorithm then got tested on two individual external validation sets. This step was used to evaluate the performance of the CAD system on unseen data.

From *Dataset 2* on, data was hand-labelled by 2 experts and the neoplastic lesions within the images were delineated by 2 to 6 experts.

This should serve as a brief overview on the CAD system which our study is based on. Further detailed information can be found in the original paper by Groof et al. (41).

3.3 Data Acquisition

Four test sets were used in this study. As described in the chapter *Study Setting and Design*, the test sets need to fulfil individual characteristics. This resulted in the composition of four different test sets from centers in Amsterdam and Eindhoven in the Netherlands, Leuven in Belgium, Nantes in France, and Regensburg in Germany. All participating hospitals are specialized centers for Barrett's esophagus.

Table 3. Overview of included test sets. Changed parameters to the previous test set are written in bold letters.

	Characteristics of test set	**Number of images and patients**	**Centers of recruitment**
Test set 1	Same center Same endoscopic system (Fujifilm) Pretreatment	160 (80 NDBE, 80 neoplasia; 160 patients)	Amsterdam University Medical Centers, Universitair Ziekenhuis Leuven, Centre Hospitalier Universitaire de Nantes, St. Antonius Ziekenhuis, Catharina Ziekenhuis Eindhoven
Test set 2	Same center **Different endoscopic system (Olympus)** Pretreatment	524 (391 NDBE, 133 neoplasia; 97 patients)	Amsterdam University Medical Centers
Test set 3	**Different center** Different endoscopic system (Olympus) Pretreatment	213 (8 NDBE, 205 neoplasia; 72 patients)	Hospital Barmherzige Brüder, Regensburg
Test set 4	Different center Different endoscopic system (Olympus) **Posttreatment**	422 (385 NDBE, 37 neoplasia; 84 patients)	Hospital Barmherzige Brüder, Regensburg

Abbreviations: NDBE = non-dysplastic Barrett's esophagus

The image collection and labelling happened independently by different teams in all centers of recruitment. Nevertheless, inclusion and exclusion criteria were similar and read as follows:

Table 4. Inclusion and exclusion criteria for the test sets.

Inclusion criteria	**Exclusion criteria**
- Images of non-dysplastic Barrett's esophagus (NDBE) without any visible abnormalities, or: - Images of high-grade dysplasia (HGD) with visible lesions, or: - Images of early adenocarcinoma (EAC) with visible lesions - Attached histopathological report for current or prior endoscopy - Imaging technique: white light endoscopy (WLE) - Images taken in overview without focus on a lesion or non-dysplastic mucosa	- Images of low-grade dysplasia (LGD) - Neoplastic images (HGD, EAC) without visible lesions - NDBE images with visible lesions (except *Test set 4*) - Images in near focus mode - Images focussed on the lesion - Application of acetic acid - Tools in the image - Blood in the image - For *Test set 1*: non-Fujifilm imagery - For *Test set 2, 3, 4*: non-Olympus imagery

Images could be taken in the antegrade position, in retrograde position, which we call *retroflexion image* in this thesis, or with a distal cap attached to the endoscope, here called *cap image*. For better understanding, some examples will follow.

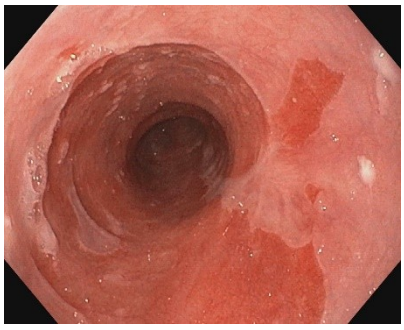


Figure 6. Example for image.

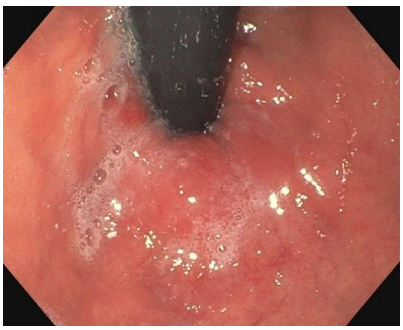


Figure 7. Example for retroflexion image.

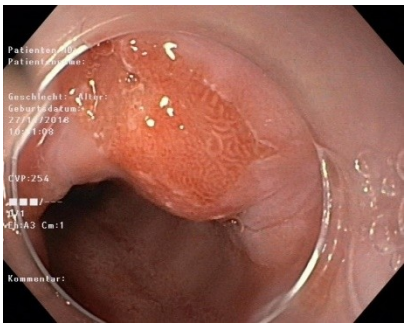


Figure 8. Example for cap image.

A detailed description of the recruitment process is described in the following subchapters.

3.3.1 *Test set 1*

The images for *Test set 1* were originally collected for the development of the CAD system we use for this thesis. The used images were selected from two external validation sets. Centers of collection are the Amsterdam University Medical Centers, Netherlands, Universitair Ziekenhuis Leuven, Belgium, Center Hospitalier Universitaire de Nantes, France, St. Antonius Ziekenhuis, Netherlands, and Catharina Ziekenhuis Eindhoven, Netherlands.

Images were collected retrospectively and prospectively. The utilized endoscopy system was Fujifilm 700 series. Two experts hand-labelled the images and the correlated pathology report verified the state of dysplasia. The neoplastic area was delineated by 3 to 6 experts.
All of the 160 images from the external validation datasets were used for our study. These 160 images show an even distribution of 80 NDBE and 80 neoplastic images derived from 160 different patients. All images are from a pretreatment period. The dataset is described as heterogenous regarding lesion subtlety, with partly very subtle lesions (41).

3.3.2 *Test set 2*

This test set consists of images recruited in the Amsterdam University Medical Centers, Netherlands. Images were collected retrospectively for the inclusion in another study and also reviewed for quality for inclusion in that study. An Olympus system was used for endoscopy.

We randomly selected 524 images, divided into 391 non-dysplastic images and 133 neoplastic images. The 391 non-dysplastic images include 25 cap images and 88 retro images, representing 68 patients. The neoplastic images consist of 83 HGD and 50 EAC images with 11 cap and 33 retro images, coming from 29 different patients. A correlated pathology report confirmed the classification into NDBE or neoplastic. All images are pretreatment.

3.3.3 *Test set 3* and *Test set 4*

For *Test set 3* and *4*, images were acquired retrospectively from the hospital Barmherzige Brüder Regensburg, Department for Gastroenterology and Interventional Endoscopy. The hospital uses an endoscopic system by Olympus, mainly the 190 series was used.

In total, 750 WLE images of 103 patients were collected, taken in 357 endoscopies. Of these 750 WLE images, 259 were classified as pretreatment and 491 as posttreatment. Of the 259 pretreatment images, 21 were non-dysplastic (including 0 cap images and 0 retro images) and

238 neoplastic (including 36 cap images and 10 retro images).
The 491 posttreatment images are divided into 450 non-dysplastic images (including 4 cap images and 14 retro images) and 41 neoplastic images (including 1 cap image and 4 retro images).

These images were reviewed by an expert of the Amsterdam University Medical Centers to guarantee high quality.

Eventually included into our study were 213 pretreatment images, of which 8 were non-dysplastic (0 cap images, 0 retro images) and 205 neoplastic (41 cap images, 8 retro images; 31 HGD, 174 EAC). Non-dysplastic pretreatment images were taken in 5 different patients, neoplastic pretreatment images in 67 patients.

We included in total 422 posttreatment images, containing 385 NDBE images (with 4 cap images and 10 retro images) and 37 neoplastic images (with 1 cap image and 4 retro images; 6 HGD and 31 EAC). Non-dysplastic posttreatment images came from 70 different patients, neoplastic posttreatment images from 14 patients.

The endoscopies took place between 31.05.2011 and 14.04.2021 and the age of patients ranges between 41 and 88 years.

3.4 Performance Metrics

As some metrics from the field of machine learning might be unknown to the reader, we will give a brief overview of the metrics playing a role in the results section.

Firstly, a confusion matrix is a basic way to present prediction results for binary classification. Several metrics like sensitivity or specificity can be calculated from the information in the matrix.

Table 5. Confusion matrix.

	True class 1 (positive)	True class 2 (negative)
Predicted class 1 (positive)	True Positive (TP)	False Positive (FP)
Predicted class 2 (negative)	False Negative (FN)	True Negative (TN)

Secondly, metrics we use in this thesis are presented with a short explanation and the corresponding equation (97–99).

True Positive Rate (TPR)/Sensitivity/ Recall	Out of all images with neoplastic ground truth, how many were correctly classified as neoplastic?	$TPR = \frac{TP}{TP + FN}$
True Negative Rate (TNR)/Specificity	Out of all images with non-neoplastic ground truth, how many were correctly classified as non-neoplastic?	$TNR = \frac{TN}{TN + FP}$
Positive Predictive Value (PPV)/ Precision	Out of all as neoplastic classified images, how many were truly neoplastic?	$PPV = \frac{TP}{TP + FP}$
Negative Predictive Value (NPV)	Out of all as non-neoplastic classified images, how many were truly non-neoplastic?	$NPV = \frac{TN}{TN + FN}$
False Negative Rate (FNR)/Miss Rate	Out of all images with neoplastic ground truth, how many were misclassified as false non-neoplastic? How many neoplastic images did we miss?	$FNR = \frac{FN}{FN + TP}$
False Positive Rate (FPR)	Out of all images with non-neoplastic ground truth, how many were misclassified as false neoplastic?	$FPR = \frac{FP}{FP + TN}$
Accuracy	Which proportion of predictions in total was correct?	$Acc. = \frac{TP + TN}{TP + FP + TN + FN}$
Receiver operating characteristic (ROC) curve	shows graphically the performance of a classification algorithm at all possible thresholds	curve with FPR on x-axis and TPR on y-axis
Area under the ROC curve (AUC)	With which probability is a neoplastic result ranked above a non-neoplastic result?	$AUC = \int_0^1 t \times df$

Skew Ratio	How imbalanced is the dataset?	$Skew = \frac{TN + FP}{TP + FN}$

Special attention has to be given to the F1-Score, whose formula was slightly amended in this thesis. Instead of the positive predictive value, an alternative formula with the specificity was applied.

F1-Score	harmonic mean between specificity and sensitivity	$F1 = 2 \times \frac{Sens. \times Spec.}{Sens. + Spec.}$

In machine learning, the term *recall* is often used synonymously to the true positive rate or sensitivity and the term *precision* to positive predictive value. To avoid confusion, we will stick to the terms sensitivity and positive predictive value.

Metrics can be divided into threshold metrics, which are sensitive to the threshold but do not consider how narrow the decision was, and rank metrics, which evaluate the ordering of positive predictions above negative predictions. Threshold metrics include among others accuracy and F1-Score. A rank metric is AUC (99).

4 Results

This chapter presents the results obtained from applying the CAD system to the specified test sets. The table below provides an overview of important metrics, which will be discussed in detail for each test set in the following sections.

Table 6. Performance results of the CAD system on our test sets.

	Accuracy	**Sensitivity**	**Specificity**	**F1**	**AUC**	**NPV**	**PPV**	**FNR**	**FPR**
Test set 1	0,8875	0,9125	0,8625	0,8868	0,8875	0,9079	0,8690	0,0875	0,1375
Test set 2	0,7462	0,3759	0,8721	0,5254	0,6240	0,8042	0,5	0,6241	0,1279
Test set 3	0,6901	0,6878	0,75	0,7176	0,7189	0,0857	0,9860	0,3122	0,25
Test set 4	0,6588	0,6486	0,6597	0,6541	0,6542	0,9513	0,1548	0,3514	0,3403

Abbreviations: CAD = computer aided detection; F1 = F1 score; AUC = Area under the receiver operating characteristic curve; NPV = negative predictive value; PPV = positive predictive value; FNR = false negative rate; FPR = false positive rate

4.1 Performance on *Test set 1*

Test set 1 shows the initial performance of the developed CAD system on Fujifilm imagery from the center it was trained in.

Table 7. Confusion matrix for *Test set 1*.

	NDBE	Neoplasia	
Predicted NDBE	69	7	76
Predicted neoplasia	11	73	84
	80	80	160

Abbreviations: NDBE = non-dysplastic Barrett's esophagus

Test set 1 consists of 160 images, with 80 images being non-dysplastic and 80 images being neoplastic. The test set is balanced with a skew ratio of 1. The CAD system detects 91,3% of neoplastic images as neoplastic and correctly classifies 86,3% of non-dysplastic images. This

results in an accuracy of 88,8%. Of all as non-dysplastic classified images, 90,8% are actually non-dysplastic (negative predictive value), while 86,9% of all as neoplastic classified images are truly neoplastic. The F1 score presents with 88,7% and AUC with 88,8%.

4.2 Performance on *Test set 2*

Test set 2 shows the performance of the developed CAD system on Olympus imagery with no altered participating center.

Table 8. Confusion matrix for *Test set 2*.

	NDBE	Neoplasia	
Predicted NDBE	341	83	424
Predicted neoplasia	50	50	100
	391	133	524

Abbreviations: NDBE = non-dysplastic Barrett's esophagus

Test set 2 contains 524 images, of which 391 images are non-dysplastic and 133 images are neoplastic. This results in a skew of 2,94, which means that the test set contains nearly three times as many NDBE images as neoplastic images. The algorithm shows on *Test set 2* a value of 87,2% for specificity, while sensitivity, F1 score and AUC are lower. The sensitivity of 37,6% means that of all 133 images with neoplasia nearly two third were misclassified as non-dysplastic. In comparison, of all non-dysplastic images, 87,2% were correctly identified and only 12,8% showed a wrong classification. The F1 score is just above 50% (exact: 52,5%) and the threshold-independent AUC comes to 62,4%. Of all images, 74,6% were correctly classified (accuracy). 80,4% of all as non-dysplastic classified images were truly non-dysplastic, while of the 100 images with predicted neoplasia, only 50 were actually neoplastic.

4.3 Performance on *Test set 3*

Test set 3 shows the performance of the developed CAD system on Olympus imagery in a new participating medical center.

Table 9. Confusion matrix for *Test set 3*.

	NDBE	Neoplasia	
Predicted NDBE	6	64	70
Predicted neoplasia	2	141	143
	8	205	213

Abbreviations: NDBE = non-dysplastic Barrett's esophagus

In *Test set 3*, of in total 213 images, 8 images were non-dysplastic and 205 images were neoplastic. In this case, the skew is 0,038 - representing an imbalance with more neoplastic images than NDBE images. It shows high results for the metric positive predictive value with 98,6%. In absolute numbers, this means, that of all 143 as neoplastic classified images, 141 were actually neoplastic and only 2 images were misclassified as non-dysplastic. In comparison, the negative predictive value is at 8,6% - of all as non-dysplastic classified images, only 8,6% were truly non-dysplastic. The accuracy is at 69%. Sensitivity shows that 68,8% of all neoplastic images are correctly classified as neoplastic, whereas according to specificity, 75% of non-dysplastic images are correctly detected as non-dysplastic. The F1 score presents with 71,8% and AUC with 71,9%.

4.4 Performance on *Test set 4*

Test set 4 shows the performance of the developed CAD system on Olympus imagery in a new participating center with the new treatment status *posttreatment*.

Table 10. Confusion matrix for *Test set 4*.

	NDBE	Neoplasia	
Predicted NDBE	254	13	267
Predicted neoplasia	131	24	155
	385	37	422

Abbreviations: NDBE = non-dysplastic Barrett's esophagus

Test set 4 consist of 422 images, including 385 non-dysplastic and 37 neoplastic images. The skew presents an imbalance with 10,41 times more non-dysplastic than neoplastic images. The accuracy is at 65,9%, with a sensitivity of 64,9% and a specificity of 66%. The F1 score comes to 65,4%. The AUC shows that a neoplastic image gets classified above a non-dysplastic image

with a probability of 65,4%. According to the negative predictive value, 95,1% of as non-dysplastic classified images are actually non-dysplastic, whereas the positive predictive value of 15,5% means that of all as neoplastic classified images, only 15,5% were truly neoplastic.

4.5 Comparison between *Test set 1* and *2*

This comparison describes the change in endoscopy system.

The algorithm shows on *Test set 2* generally a weaker performance than on *Test set 1*. Accuracy is about 14% lower, negative predictive value and positive predictive value present differences of about 10% and 37%, respectively. Sensitivity is about three fifth decreased (from 91% to 38%), only specificity is with a difference of 1% nearly similar (86% and 87%). The F1 score, which summarizes sensitivity and specificity, differs about 36% with a lot better performance on *Test set 1* than on *Test set 2*. In *Test set 1,* 8,7% of neoplastic images were missed, in *Test set* 2 this was the case for 62,4%. For the threshold-independent performance presented by AUC shows the result on *Test set 1* a 26% higher value than on *Test set 1*. The corresponding ROC curves are displayed in the following plot.

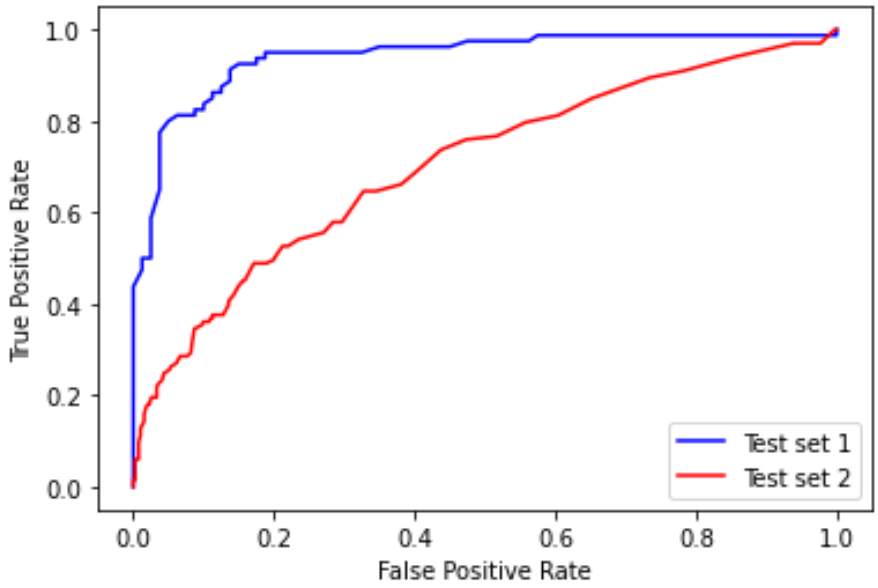


Figure 9. Receiver operating characteristic curves for *Test set 1* and *2*.

4.6 Comparison between *Test set 2* and *3*

The new factor in this comparison is the collection of images from a new medical center. In total, more images were classified correctly in *Test set 2*, being presented by a 5,6% difference in accuracy. Also, specificity is in *Test set 2* about 12% higher. On the other hand, values for sensitivity are in *Test set 3* nearly twice as high as in *Test set 2*. The F1 score is in *Test set 3* about 19% and AUC about 9% increased. Clear differences can be seen in negative predictive value and positive predictive value. Negative predictive value for *Test set 2* presents with 80,4%, compared to 8,6% for *Test set 3*. The positive predictive value is the other way round with 50% for *Test set 2* and 98,6% for *Test set 3* - a nearly 50% difference. In *Test set 2*, about 62% and in *Test set 3* about 31% of neoplastic images were missed.

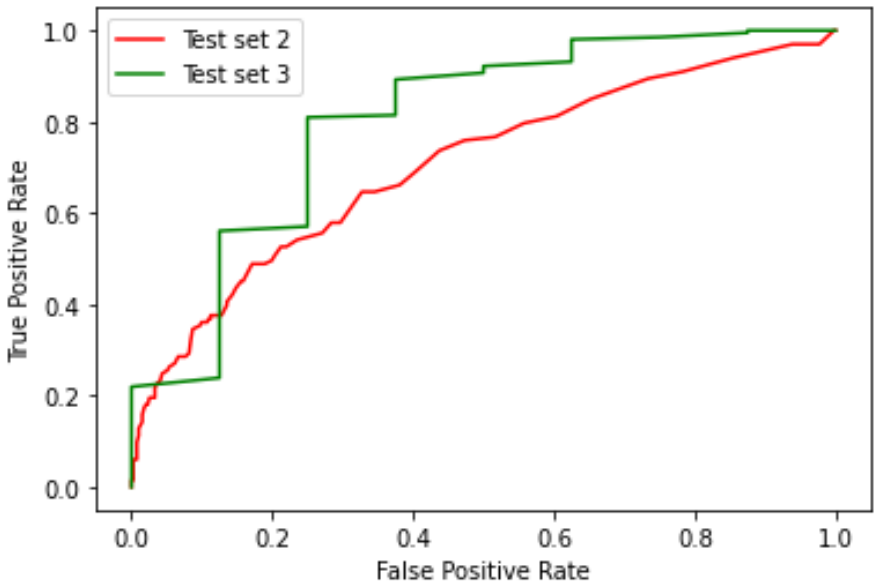


Figure 10. Receiver operating characteristic curves for *Test set 2* and *3*.

4.7 Comparison between *Test set 3* and *4*

In this comparison, the new treatment status *posttreatment* is introduced.

The performance on *Test set 3* is superior in all metrics except negative predictive value and false positive rate. Still, differences are mostly below 10%. Accuracy is about 3% higher in *Test set 3* with 69%. Sensitivity and specificity differ about 4% and 9%, resulting in a F1 score of 75% for *Test set 3* and 65% for *Test set 4*. The AUC has a difference of 6%, which can also be observed in the graphic. The low negative predictive value of 8,6% for *Test set 3* stands compared to the eleven times higher value of 95,1% for *Test set 4*. Also, for the positive

predictive value there is a big difference of 83% - as in *Test set 3* 98,6% and in *Test set 4* 15,5% of as neoplastic classified images were truly neoplastic. A similar number of neoplastic images has been missed - about 31% in *Test set 3* and 35% in *Test set 4*.

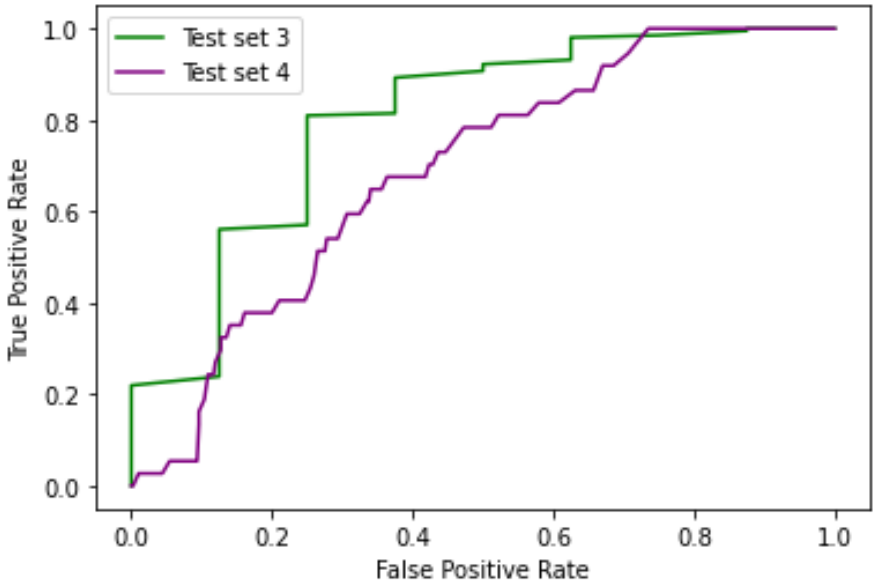


Figure 11. Receiver operating characteristic curves for *Test set 3* and *4*.

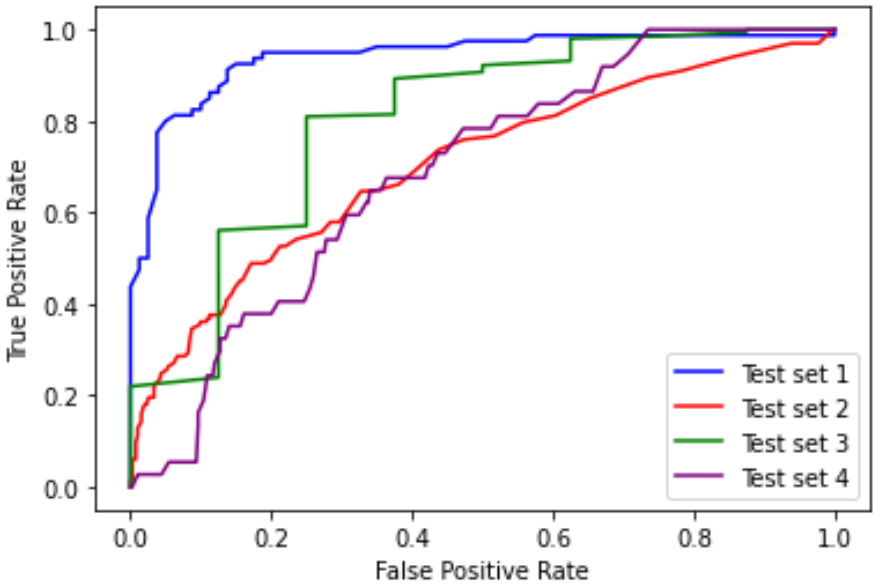


Figure 12. Receiver operating characteristic curves for *Test set 1*, *2*, *3* and *4*.

5 Discussion

With esophageal adenocarcinoma showing a poor survival prognosis, early detection is essential. With a lower tumor stage, riskier surgical treatment can be avoided and overall survival rate will be improved (18,100). As non-expert endoscopists were shown to have difficulties detecting the rare and often subtle neoplastic lesions, AI systems can support endoscopists in the detection process (8). An AI system should be applicable in different medical centers, which implies individual patient populations including different treatment status and technical equipment like endoscopy systems. To which extent these factors influence the performance of a CAD system is investigated in this manuscript. In the following paragraphs the results of this study and clinical implications will be discussed.

5.1 Performance on Images from New Endoscopy System

In this study, the application of our CAD system on images by a new endoscopy system led to a drop in performance. The decrease in for clinical settings important metrics like sensitivity (from 91,3% in *Test set 1* to 37,6% in *Test set 2*) and the positive predictive value (from 86,9% to 50%) indicates that this trained CAD system is not generalizable on endoscopy systems by different companies. This was here shown for the case of a Fujifilm trained system on Olympus imagery.

Ko et al. found out that both Fujifilm and Olympus have excellent endoscopy systems which do not show any weaknesses in comparison of quality - meaning, the drop did most likely not happen due to decreased quality of Olympus imagery. Still, differences in definition, light and shade, boundary and brightness exist (46). It is therefore not surprising that a CNN-based CAD system performs slightly differently on endoscopy systems from different manufacturers.

However, most CAD systems for Barrett's classification are trained on endoscopic imagery from just a single manufacturer (see *Table 11*) (35,37,38,40,42–44,86,87,101–106). Groof et al. mention additional Olympus imagery in pretraining and refinement phases, but all specific training was done on Fujifilm imagery (41). As this is the CAD system used in this study, even the early involvement with Olympus imagery apparently does not lead to an equal performance as with Fujifilm imagery.

To date, Abdelrahim et al. developed the only CAD model for Barrett's neoplasia using imagery of more than one endoscopy system. They used images made with Fujifilm, Olympus and

Pentax endoscopes for training, validation and testing, therefore representing three major manufacturers. The individual proportion of these camera platforms in the datasets used remains unclear (34).

In endoscopic detection of Barrett's neoplasia, there are currently two AI systems approved for clinical use. WISE VISION, developed by NEC Corporation, can be connected to endoscopes from Olympus, Fujifilm and Pentax. Their special knowledge in facial recognition technology was transferred to neoplasia detection, but no further technical details are found publicly. According to the company, the system was "trained with more than 1 million Barrett's esophagus endoscopy images" (47). This cannot be independently verified as no information on the training, validation and testing process is publicly available. The only information on performance is a sensitivity of over 90%, but the source for this is a non-public presentation held in 2021 (47). However, a clinical trial on a real-life evaluation of WISE VISION is currently underway (107).

The second certified system is CADU by Odin Vision, now belonging to Olympus. According to the company, CADU is compatible with all major endoscopy manufacturers (48). No further information on development of the system and performance evaluation were found publicly available. A study on the in-vivo performance of CADU is currently running (108).

Both systems fulfilled the requirements of CE mark labelling, meaning essential European safety and performance standards. WISE VISION has this certification since 2021, Odin Vision since 2024. Both companies did not answer a personal request for more transparent information on their development, testing processes and results.

Several studies mention the need for further validation on different endoscopy systems and our findings support this demand (42,44,109).
Some studies are funded by one of the manufacturers, which could result in a conflict of interest (37,38).

The literature research revealed that most research in AI in Barrett's esophagus happens with Olympus or Fujifilm endoscopy systems. Only one study used solely Pentax imagery to develop their neural network (43), another one included Pentax amongst other manufacturers (34). Indeed, no study was found for any other system, even though other high-quality systems like Karl Storz exist. The lack of external validation across different endoscopy systems might reinforce the dominance of specific manufacturers, as AI systems may only perform reliably on devices used during training. As different systems have different strengths and weaknesses, it

should be of doctor's and patient's interest to keep products of different endoscopy companies competitive. This means that CAD systems should be usable independent of the endoscopy system (46).

Another thing that stands out is that 62% of all neoplastic images in *Test set 2* were misclassified as non-dysplastic and would have been missed. For our cause it is important to have a low false negative rate (which corresponds to a high sensitivity). False positives can easily be excluded via low-risk biopsy and subsequent pathological staging, while false negatives may evolve to higher tumor stages with poorer prognosis and riskier treatment options (2,75). Because of the importance of the false negative rate for the clinical setting, we will have a closer look at possible reasons for the high miss rate. Also, this is not only the case for *Test set 2*, but also for *Test set 3* with 31,2% of missed neoplastic images and *Test set 4* with 35,1% of missed tumors.

Possibly, the images in the training data were simpler compared to maybe more subtle lesions in our test sets. We classify this reason as rather unlikely as to our human perception the external validation sets of the algorithm contain subtle cases which we rate as more subtle than our *Test set 2*.

It might be that the data labelling for *Test set 2*, *3* and *4* was of worse quality, resulting in poor performance due to label noise. Label noise describes the appearance of mistakes in labelling which can happen due to interexpert variability, subjectivity in labelling, or high expenditure leading to cheaper non-expert labelling (110). We rate this as an improbable reason for the drop in performance between *Test set 1* and the other test sets. *Test set 2* was labelled completely independently and by a different team than *Test set 3* and *4* - and still showing a similar drop in sensitivity. Also, all test sets were labelled by medical trained staff and reviewed by an expert.

As the high miss rate is present in all Olympus test sets, we suspect that the new endoscopy system is the reason for this. To avoid missed neoplasia following a domain switch, training with images of different endoscopy systems and matching external validation should be sought.

Table 11. Overview of studies on CAD systems for neoplasia detection in Barrett's esophagus (adenocarcinoma).

Study	Design	Endoscopy company	Data collection method	Training set (images, (patients))	Valildation set (images, (patients))	Test set (images, (patients))	Centers	Characterization of Test sets
Abdelrahim, 2023 (34)	ex-vivo, video-based	**Fujifilm, Olympus, Pentax**	retro. + prosp.	816 original images of neo (96) and NDBE (65), after augmentation: 75 198 frames neo, 1 014 973 NDBE	image-based validation: 107 neo (20); 364 NDBE (14)	Video-based **external** validation (real-time): 32 videos neo (32), 43 NDBE (43) benchmarking with 6 non-exp. endoscopists	4 (European)	Test set: prospectively collected from external centers HD imagery
Groof, 2020 (41)	ex-vivo, image-based	Fujifilm	retro. + prosp.	1. ImageNet 2. GastroNet: 494 364 images (15 286) 3. 690 neo images (198); 557 NDBE (216)	129 neo (54); 128 NDBE (41)	**External** validation: 1. 40 neo (40), 40 NDBE (40) 2. 40 neo (40), 40 NDBE (40) **benchmarking** with 53 non-exp. endoscopists	4 (European)	Test set 1: heterogenous regarding lesions, 4 centers Test set 2: prospective collection, subtle lesions, 3 centers HD imagery
Ebigbo, 2019 (35)	ex-vivo, image-based	n.a.	n.a.	248 images of 50 neo lesions and 63 NDBE (WLE and NBI) no specification on exact data set split		n.a.	1 (European)	HD imagery
Ebigbo, 2020 (36)	**in-vivo**, video-based	n.a.	prosp.	129 images, no further specification	n.a.	36 images neo, 26 NDBE (total 14), as image freezes from real-time endoscopies **benchmarking** with expert endoscopists	1 (European)	n.a.

Fockens, 2023a (38)	ex-vivo, image-based	Olympus	retro.	1. ImageNet 2. GastroNet: 494 364 (15 286) 3. 1480 neo (435), 2333 NDBE (592)	233 neo (129), 374 NDBE (73)	**External** validation: 1. 50 neo (50), 150 NDBE (150) 2. 50 neo (50), 50 NDBE (50) 3. 50 neo (39), 150 NDBE (74) **benchmarking** on test set 1 with 52 non-exp. endoscopists	**15** (European, Australian)	Test set 1: retrospective acquisition, more subtle lesions Test set 2: retrospective acquisition, heterogenous lesions Test set 3: prospective acquisition high quality imagery
Fockens, 2023b (37)	ex-vivo, video-based	Olympus	retro. + prosp.	1. ImageNet 2. GastroNet: 5 084 494 images 3. 6251 neo (1296), 7595 NDBE (1095)	Image validation: 100 neo (58), 100 NDBE (36) Video validation: 77 neo (58), 103 NDBE (33)	**External** validation: Image test set: 82 neo (46), 327 NDBE (73) Video test set: 71 neo (45), 180 NDBE (66) **benchmarking** with 28 exp. and group of >40 non-exp. endoscopists: image test set: 100 neo (50), 300 NDBE (125) video test set: 47 neo (47), 141 NDBE (82)	**15** (European, Australian)	prospective acquisition without focus on neo lesion Image test set: representing daily clinical practice benchmarking test set with more subtle lesions high quality imagery
Fockens, 2023c (39)	**in-vivo**, video-based	Olympus	retro. + prosp.	6237 neo (1304), 7595 NDBE (1103) further information see (source Fockens 2023a)	see (source Fockens 2023a)	video test set in real-time: 15 neo (15), 15 NDBE (15)	**15** (European, Australian)	standardized WLE pullback video and overview videos

Ghatwary, 2019 (40)	ex-vivo, image-based	n.a.	retro.	50 neo, 50 NDBE (total 39), data augmentation	n.a.	leave-one-patient-out cross validation	1 (Euro-pean)	HD imagery
Hashimoto, 2020 (42)	ex-vivo, image-based	Olympus	retro.	1. ImageNet 2. 691 neo (70), 683 NDBE (30)	n.a.	225 neo; 233 NDBE	1 (North-Ameri-can)	high quality imagery WLE, NBI, standard focus, near focus heterogenous lesions, many subtle lesions
Hussein, 2020 (43)	ex-vivo, image-based	**Pentax**	prosp.	classification model: 31 neo videos, 31 NDBE, 2 normal esophagus (total 118) 148 936 frames segmentation model: 1. external data 2. 94 neo images (30)	classification: 25 161 frames (6 neo, 5 NDBE) segmentation: 12 neo images (6)	classification: 168 iscan neo frames (28), 96 iscan NDBE (16); 108 WLE neo frames (18), 84 WLE NDBE (14) segmentation: 86 iscan neo (28)	4 (Euro-pean)	pullback video HD imagery i-scan, WLE
Mendel, 2017 (45)	ex-vivo, image-based	n.a.	retro.	50 neo (22), 50 NDBE (17), data augmentation	n.a.	leave-one-patient-out cross validation	1 (Euro-pean)	HD imagery
Struyvenberg, 2021 (44)	ex-vivo, image-based	Olympus	retro.	1. GastroNet: 494 364 (15 286) 2. 690 WLE neo (198), 557 WLE NDBE (216) 3. 112 NBI neo (50), 71 NBI NDBE (50) 4. 59 neo NBI videos (17), 98 NDBE NBI videos (33)	validated on training set 3	testing on training set 3 and 4	3 (Euro-pean)	high quality imagery

Abbreviations: NDBE = non-dysplastic Barrett's esophagus, neo = neoplastic, retro. = retrospective, prosp. = prospective, HD = high definition, WLE = white light endoscopy, NBI = narrow band imaging, exp. = expert, non-exp. = non-expert

5.2 Performance on Images from New Medical Center

The application of a CAD system on images from an unknown center does probably affect the system's performance. We used *Test set 2* as control dataset to review whether a performance change to *Test set 3* was due to a different endoscopy system or a new center. But still, we do not know to which extent each of these two factors impact the performance. Also, *Test set 3* is highly imbalanced with a skew of 0,038 and only 8 non-dysplastic images, compared to 205 neoplastic images. Thus, its informative value is limited.

We cannot draw a clear conclusion from the comparison of *Test set 2* and *Test set 3*. Unexpectedly, metrics like sensitivity, F1 score and AUC did even improve in *Test set 3* - although the algorithm has to deal with two new factors here. But if compared with *Test set 1*, a drop in performance for *Test set 3* is visible, which could be caused by both new endoscopy system and new center. However, our results suggest that a shift in endoscopy system has a bigger influence than switching from one expert center to another.

A cause for the decreased performance compared to *Test set 1* could be overfitting, a common challenge in deep learning. A machine learning model can get over-specialized on a specific dataset, meaning it rather memorizes irrelevant parameters than learning a generalizable structure. It will then have a high performance on this dataset, but perform poorly on unknown datasets (49). This is also why external validation on external datasets is necessary for a significant performance evaluation.
Solutions against overfitting can be more training data, or more technical solutions like regularization with dropout, weight decay, batch normalization, data augmentation and a less complex deep learning architecture (28).

For the design of the used CAD system, Groof et al. tried to minimize overfitting via batch normalization, data augmentation and the special, custom-made architecture of the algorithm, making it less complex (41).

Overfitting is prone to happen due to small datasets (50). Some of the more current studies confront this issue with a lot larger and heterogenous datasets than previously (see *Table 11*).

Abdelrahim et al. used following measures against overfitting: First, they collected training and testing data (images and videos) from four centers, geographically and temporarily independent. This is important as data from different centers means data from different patient cohorts, from endoscopists with different endoscopy styles and endoscopy systems. Second, imagery by three different endoscopy systems was included (34). This augments heterogeneity

as every endoscopy system has slightly different parameters regarding lightning or colouring (46). Third, all training, validation and test sets were independent, thus ensuring that the performance measures were not influenced. If data from the training set appears again in the test set, the CAD system might remember the image and show a better performance than on unknown data. This is how they received a comparatively large Barrett's specific dataset with 816 images for training and after augmentation techniques even 1 090 171 training frames. Externally and prospectively collected data for testing avoided getting a false-good performance due to possible overfitting (34).

Hussein et al. worked with independent, not overlapping training, validation and test sets as well. From 116 videos, they created 148 936 frames to train their classification model. All these videos were collected prospectively from four different European centers (43). Prospective collection is helpful because retrospectively collected images or videos might be biased in focussing on a lesion for documentation for example. It could be easier for a CAD system to detect areas of interest in this imagery than in real-life, then resulting in a performance drop. For their segmentation model, external data was used for general training, then specified with 94 more neoplastic images (43).

An impressive dataset has been created by Fockens et al.. The group built a heterogenous dataset with data from 15 centers in 7 countries. These centers do still represent mainly European and high-income countries (Australia as non-European country), so there is still room for even more heterogenous datasets. But in comparison with most other studies in the field, this is still one of the most diverse datasets so far and thus more robust against overfitting. In their current studies, first a huge dataset with general images from endoscopy in the upper gastrointestinal tract is used, consisting by now of 5 084 494 images. Then, more specific training was performed with the largest Barrett specific image set to date with 13 846 images from 2391 patients. For external validation, the group used prospectively collected data from external centers and divided the data into different test sets representing different types of heterogeneity (37).

Other CAD systems in esophageal endoscopy use different approaches against overfitting.

Mendel et al. mention their batch size and weight decay (45) and Hashimoto et al. use dropout, weight decay and data augmentation (42). Some research groups use cross-validation as tool against overfitting (35,44,45,86,106).

Especially between different centers slight differences in endoscopic procedure, in quality of taking images, in implementation of documentation guidelines or handling of the endoscope will always be found. Together with spectrum bias due to single-center's individual patient populations and selection bias due to institutional habits, generalizability will be limited (9). Thus, to establish secure applicability in various centers, mentioned problems have to be solved.

Larger, more heterogenous test sets which better include these differences would be an improvement. A more individualized training concept with the algorithm being fine-tuned on an image test set from the new hospital could promise even better results, but at the same time resulting in more costs, time and effort.

For the extreme values of the positive predictive value (98,6%) and negative predictive value (8,6%) in *Test set 3* it has to be kept in mind that these metrics vary with disease prevalence, in this case represented by our skewed test set. They are likely to differ with a more balanced or reversely skewed dataset and are thus not further discussed despite their extremity.

5.3 Performance on Posttreatment Images

The pretreatment CAD system classifies a posttreatment NDBE image more likely as neoplastic than a pretreatment NDBE image. 34% of NDBE images in our posttreatment *Test set 4* were falsely classified as neoplastic, compared to 25% in *Test set 3* and 12,8% in *Test set 2*.

We planned to compare *Test set 4* with *Test set 3* as these images were taken in the same center and labelled by the same person, so that these biases can be neglected. But as *Test set 3* only consists of 8 NDBE images, we have to treat all metrics except sensitivity critically. With such a small NDBE image set, the possibility of chance cannot be ruled out. That is why we also take *Test set 2* for comparison.

The values for positive predictive value (15,5%) and negative predictive value (95,1%) in *Test set 4* have to be considered with caution because - as already discussed in the previous chapter - these metrics vary with disease prevalence or the balance of datasets. With a lot more NDBE than neoplastic images, there are in absolute numbers more false positives than true positives, which leads to a poor positive predictive value. At the same time we have to keep in mind that the real-life distribution is similarly imbalanced with an incidence of 1% of high grade dysplasia and adenocarcinoma per year among Barrett's patients (51).

Horie et al., who developed a model for detection of squamous cell carcinoma and adenocarcinoma in the esophagus, also deal with low positive predictive value (39%) regarding their CAD system but state that in this context false positives are more acceptable than false negatives. Additionally, the positive predictive value of the CAD system has to be compared with the human positive predictive value in order to truly being able to make a judgement. Their test set shows a skew of 5,9 (103).

However, Fockens et al. take a new approach in their publication. They intentionally use an unbalanced dataset in one of their test sets with a distribution of 1:3 of neoplastic to NDBE images, as this would be more representative of the real distribution in everyday clinical practice. This results in a positive predictive value of 45%. Their other test sets show positive predictive values of 75% and 45%, but as every test set has unique features, these values are not directly comparable (38).

5.4 Evaluation of Applicability in Non-Expert Settings

The used algorithm works very well for Fujifilm images from centers it was trained in. Our results suggest that the CAD system cannot easily be used in new centers with possibly an endoscopy system by another company. For an appropriate, responsible use in new centers or with new endoscopy systems, the performance has to be improved. This is important as especially inexperienced or non-specialized endoscopists, who this algorithm is supposed to support, might not recognize possible mistakes made by the AI.

For our cause, sensitivity has a higher priority than the positive predictive value. False positives can be ruled out via biopsy, while early detection protects from the high mortality (18,63). According to our results, we would miss 62,4% of high-grade dysplasia or neoplasia when applied to a new endoscopy system and 31,2% with the combination of new endoscopy system and new center. Endoscopists on their own would most likely have a better detection rate than this (41). Even if the CAD system is only thought to support endoscopists, such a performance might make endoscopists feel insecure and eventually bias them in also missing lesions.

A posttreatment algorithm would be of great use for surveillance endoscopies after eradication therapies - often taking place in centers which are not specialized in Barrett's esophagus. After treatment for neoplasia, patients are prone to develop new tissue, with the esophagus looking differently to prior to treatment. This makes it harder for non-expert endoscopists to differentiate between a posttreatment or a neoplastic esophagus.

However, our algorithm is not suitable for use in posttreatment surveillance endoscopies. Too many neoplastic images are being missed, while at the same time an unacceptable amount of NDBE images is falsely classified as neoplastic.

Another limitation might be that community hospitals or non-specialized surveillance centers often do not have the same high-quality equipment as expert centers. Endoscopic image material might thus be of lower quality (8).

This is a problem, as AI models are usually trained on high-quality images with strict inclusion criteria from expert centers but are intended to be used rather in community hospitals, then potentially leading to a drop in performance due to domain change (52,111). A solution could be more heterogenous datasets also regarding data quality, reflecting more realistically the clinical everyday life and equipment in community hospitals. Additionally, the implementation of additional quality assurance algorithms could help with a better and more standardized endoscopic performance in general (111).

All in all, the CAD system is not generalizable enough for applicability in a new and non-expert setting, meaning in our case with new endoscopy systems and new image origins.

5.5 The Domain Shift Problem

Performance drops due to domain shifts are a well-known issue in the world of artificial intelligence. In our study, we show that our model is similarly affected by the domain shift problem. The new domains are in our case data obtained from an unknown endoscopy system, a new medical center and images of an altered esophagus after previous treatment.

We can identify a mix between different types of domain shift affecting our model.

- Covariate shift: This occurs when the distribution of *input* features changes between source and target domains (112). In our case, it can be pinned down on a population drift between the original and the new medical centers, as well as a shift in data acquisition. This shift is caused on one hand by the change of endoscopy system (and therefore changes in colouring, contrast, lighting) and on the other hand by variations in endoscopy procedures between centers.
- Label shift: The distribution of positive and negative *outcomes* changes from source to target domain, while input features stay the same (112). Label shift is present as the ratio of neoplastic to NDBE imagery changes from source to target domain.

- Concept shift: Concept shift occurs when the relationship between input features and output labels changes (113). This can be observed in *Test set 4*, containing posttreatment imagery. In this case, not every mucosal change indicates a possible neoplastic lesion, but can instead be consequence of a previous treatment.

These domain shifts are the main factors contributing to the performance decline of our CAD system.

Different solutions are being discussed by the scientific community.
In general, larger and more heterogeneous datasets lead to more robust models. Especially prospective data from multiple centers and independent training, validation and test sets are essential (52). This shall be discussed in further depth in the chapter *Collaboration for Common Datasets*.
Schmitz highlights that the common approach of splitting videos frame-by-frame to obtain larger image datasets leads to overestimation of a model's performance as the frames show quite similar images (114).

Stacke et. al propose a new metric, the representation shift. This metric quantifies the domain shift and can thus predict the risk of a decrease in the performance. The representation shift could be used for monitoring the progress of a model in development or for an initial control of a model in a new setting for its generalizability (31).

Unsupervised domain adaption has been proposed to address the challenge of limited training data, particularly in the medical field with its strict data protection regulations. Labelled data is especially scarce, as annotation is time-consuming, expensive, and requires trained experts. In unsupervised domain adaption, a model is getting additional training with unlabelled data from the target domain to prepare for domain change, or even without any target domain data, for example by using AI-generated training data (53). Celik et al. applied an unsupervised domain adaption model for segmentation in Barrett's esophagus. They trained the model unsupervised (without manually labelled images) on WLE imagery as source domain, and then applied it on NBI and post-acetic-acid WLE as target domain. It shows a stable performance with accurate results in all different imaging conditions, thus showing the effectiveness and promising future of unsupervised domain adaption techniques (54).

5.6 Need for External Validation

A general challenge for the widespread application of AI systems is the lack of appropriate trials which would make an objective comparison and safety evaluation possible. The medical "gold standard" of blinded randomized-controlled trials is difficult to carry out with AI (9). A 2019 meta-analysis on AI-based diagnostic analysis of medical images conducted by Kim et al. found that fewer than 6% of the included studies did an external validation of any kind (115). In AI detection of upper gastrointestinal neoplasia, 55% went through some kind of external validation after all, but only 8% were tested in a live, clinical, in-vivo setting (55).

Comparability between current CNN-based approaches is complicated because of several factors, which we discuss in more depth elsewhere : Working groups usually have their own image databases, so that due to different image origins models can hardly be compared directly. Often, different metrics are given and also testing on different image datasets and the balance of datasets produce bias and limit comparability. We experienced this even in our study, where positive predictive value, negative predictive value and accuracy could not be used for comparison due to differently balanced test sets.

In general, our study shows the necessity for more and stricter external validation in machine learning. A well-performing single-center CAD system does not necessarily perform the same way on slightly modified images, as shown in our case.

Research on prediction models shows that they generally perform poorer on new populations (116), which we also proved for this classification model. While internal validation uses data from the population or database which was used to construct the model (117), external validation is based on structurally altered data. Only external validation investigates generalizability and hence needs data from e.g. different centers or geographical locations (118).

In CNN-based systems for the esophagus, internal validation is quite common (see *Table 11*) (36,42,43). Several research groups make use of cross validation, an internal validation approach for smaller datasets (35,40,44,45). However, it should be clarified that the meaning of the term *validation* differs slightly between the field of machine learning and medicine. In machine learning, it describes a part of the model development - it is the fine tuning of the algorithm on a new internal dataset (119).

In contrast, external validation - giving a statement about robustness and generalizability - remains rare. Studies that stand out due to their advanced testing processes are the following:

The CAD system used in this study by Groof et al. was one of the first models to be externally validated. After internal cross-validation, they performed a double external validation with 80 images each. Some images originated from hospitals that had previously been involved in training, while one and two additional centers were included in the two external validation sets. Additionally, performance was benchmarked against non-expert endoscopists (41).

Abdelrahim et al. first validated their model on 471 images from the same center but from a different time period than training. Aligned with their aim to develop a video-based model, an external validation based on 75 prospectively collected videos was carried out. These videos come from three external centers across Europe and are therefore temporally and geographically distinct from the training data. This set was also benchmarked against six non-expert endoscopists (34).

Fockens et al. used a subset of their training data (excluded prior to training) for internal validation and adjustment of parameters (607 images), carefully selected based on lesion subtlety. For testing, they created three independent test sets with different characteristics. The test sets are independent from the datasets from training or validation, but it remains unclear, which of the 14 external centers provided data for which datasets. The first test set, including 200 retrospectively collected images, consists of more subtle lesions, representing images expected in surveillance endoscopies. They made the interesting choice of deliberately using an unbalanced dataset with a ratio of 1:3 of neoplastic to NDBE images, stating that this would better represent the distribution in clinical practice. The second test set consists of 100 retrospectively collected images with a 1:1 ratio and greater variation in lesion subtlety, representing endoscopy in clinical everyday life. The third test set contains 250 prospectively collected images. The first test set was also benchmarked against 52 non-expert endoscopists (38).

A following publication by Fockens et al. continued with extensive validation and testing. The CAD system was internally validated on 200 prospectively collected images (manually enriched with subtle lesions) and 180 videos. The group designed several prospectively collected test sets comprising both images and videos. The first test set reflects routine clinical practice and includes 409 images and 188 videos, each with a neoplastic to NDBE ratio of 1:4. In addition, 28 expert endoscopists benchmarked a second test set, a special benchmarking set of images and videos (400 images (1:3), 188 videos (1:3)). Furthermore, over 64 non-expert endoscopists were assessed on their performance both with and without the CAD system (37).

Both publications by Fockens et al. stand out for their detailed and transparent reporting of dataset collection and composition, with only the exact external center contributions remaining unclear. They are also the first ones to benchmark the performance of non-expert endoscopists with and without help of CAD, which is an important step towards clinical interpretability.

Regarding the external validation process of Groof et al. for the CAD system also used for this study, it has to be noted that - even with this previously externally validated algorithm - performance declined on our test sets. This shows that a lot more independent external validation is needed before considering clinical implementation.

In colonoscopy, one of the more advanced fields in medical AI, the need to establish better clinical trial designs to demonstrate efficacy has been identified as one of the top ten research priorities in working with AI (120). This highlights the importance of aligning AI development with fitting and reliable clinical evaluations to ensure patient safety.

5.7 Explainable Artificial Intelligence

In the previous discussion of our results, we had difficulties explaining performance drops and rises. We expected to see a drop in performance between *Test set 1* and *2*, but were surprised that the application on *Test set 3* with two unknown factors (endoscopy system and new medical center) led to a partly better performance than on *Test set 2* with only one new factor (endoscopy system). Neural networks are in the literature often compared to "black boxes" as it is opaque to us what happens in the hidden layers between input and output. This is problematic for quality control and interpretation of the results - which is particularly important in the medical sector where decisions have huge influence on people's wellbeing (121).

In our case, too, it would be helpful for our comprehension and further research to better understand what led to the differences in performance of the CAD system. The field of explainable artificial intelligence (XAI) aims for better interpretability and explainability of AI systems. For CNN-based systems, different techniques can highlight the image pixels which had high impact on the classification. Besides guided backpropagation (122), integrated gradients (123) and Deep Learning Important FeaTures (DeepLIFT) (124), especially the method of saliency (125) showed promising results in a Barrett's esophagus specific comparing study (126). All these methods differ in technical details, like the weighting of input features, input values or gradient (126).

5.8 Collaboration for Common Datasets

In this study, we found out that larger and more heterogenous datasets for developing AI are needed. We discussed that multi-center studies tend to produce more robust models than single-center studies. Especially in the medical field, data scarcity due to data protection laws and limited capacity of experts to label imagery are present. In addition to technical solutions (see chapter *The Domain Shift Problem*), we believe that the scientific community should tackle this issue in collaboration.

First collaborations between different centers began to appear during the past years. The ARGOS consortium (56) and the BONS-AI (Barrett's Oesophagus Imaging for Artificial Intelligence) consortium (37,38) take a step in the right direction and combine their knowledge and databases with 3 and 15 international centers, respectively. This should be continued on a bigger scale and if possible, also with wider geographical extent. To date, only European and one Australian center are part of these consortiums (37–39,56).

Public data bases with specified imagery of gastrointestinal endoscopy would support better development processes. Usually, CAD systems are first trained on general huge databases like ImageNet to learn general image characteristics. Afterwards, domain specific training is needed.
The BONS-AI consortium created GastroNet. This database consists of 5 084 494 unlabelled endoscopic images from general upper gastrointestinal endoscopy (37). HyperKvasir is another existing large and open database for endoscopic imagery with 110 079 images and 374 videos (but only 94 available images for Barrett's esophagus) (57). Other teams with large datasets should contribute to these public databases.

Schmitz and Kather propose new regulations on getting scientific credit for the publication of datasets. This could be a first step easy to implement by scientific journals. It is already common practise in other research areas to reward dataset publications as full-rank scientific contributions (58). Furthermore, they think it would be best to combine present datasets as they already represent different centers, endoscopists, geographical regions and endoscopy systems. However, very likely this will not happen due to legal and practical obstacles. Instead, techniques like swarm learning or federated learning can be used. This allows centers to train their models on shared data but without having the data stored on their own servers (58,59).

5.9 A General Frame for Artificial Intelligence

Another challenge we encountered during our study is the difficulty of comparing different studies' findings. Each study utilizes its own test and benchmarking sets, and while they try their best for transparency by describing their composition, performance comparison remains challenging.

To us, the following question arose: How can we present results in a way that makes them easier to compare?
As previously discussed, some metrics cannot be compared in isolation, as they depend on the balance of the test sets used. In general, performance is influenced not only by the CAD system's capabilities itself but also by the specific test set used.

First of all, this is a call for transparency. Researchers must publish as much information as possible on the data used. Where and when was the data collected? By whom? Using which systems? What were inclusion and exclusion criteria? What was the exact distribution of positive and negative cases? Was data edited or augmented? How many different patients does the data represent? What are characteristics of test sets?
To some researchers, this might seem obvious, but many studies in the field miss crucial information. In their meta-analysis on AI in upper gastrointestinal neoplasia, Arribas et al. found that of 36 included studies, only 13 gave information on the number of included patients for image extraction. Ten studies did not report on the number of images used for the training set (55).

Also, we should consider which metrics truly provide meaningful information on the performance of a model. Typically, the stand-alone performance of a CAD model is reported. However, this does not accurately reflect its performance in a real-world setting - the CAD model is designed to assist the endoscopist rather than operate independently. Therefore, the most relevant metric for us is the model's ability to improve the endoscopist's performance. The most meaningful evaluation would involve assessing the performance of non-expert endoscopists - who are the primary users of the system - both with and without the model's assistance (111). While some studies have benchmarked expert and/or non-expert endoscopists versus the model (36,38,41), direct comparisons of endoscopists' performances with and without the CAD system have to date only been conducted by Fockens et al. (37).

Van der Sommen et al. also propose the creation of independent benchmarking datasets with an adequate sample size, provided by the endoscopy community. These datasets should be

strictly reserved for testing, with no access for training or validation (52). This approach would enhance the comparability of performance metrics.

Further thought, standardized methods for development and evaluation of new systems in this relatively new field of AI in endoscopy are needed. This is also a focus of the group developing the FUTURE-AI guidelines (60). Based on an international, multi-domain expert consensus, they have established guidelines for the trustworthy and safe implementation of AI in healthcare. Their framework consists of the key pillars fairness, universality, traceability, usability, robustness, and explainability, providing a clear and comprehensive guide. Additionally, they have introduced the *FUTURE-AI medical imaging checklist* to support those involved in the development of CAD models (60). Researchers should align their studies more closely with these frameworks. It might also be helpful to have more specified guidelines tailored for more defined research fields, such as AI in endoscopic imaging in our case.

The American Society for Gastrointestinal Endoscopy established even more specific regulations. In their *Preservation and Incorporation of Valuable Endoscopic Innovations* guidelines for integrating new imaging technologies in the context of Barrett's esophagus, they define precise values for performance thresholds that technologies have to meet to replace random biopsies with targeted biopsies. In the case of high-grade dysplasia or esophageal adenocarcinoma, they set for example a minimum performance of 90% sensitivity, 98% negative predictive value, and 80% specificity (127).

We see a clear need for more standards in in the development and reporting on AI models. However, it is important to ensure that the implementation of new general standards does not block innovation and creativity in research. These standards should not restrict research groups from developing new models or exploring new approaches. Overall, a more transparent and standardized framework for developing and presenting CAD models would benefit the entire community.

5.10 Limitations

Our study has several limitations.

First, images of low-grade dysplasia were excluded. Still, low-grade dysplasia is important to detect as it is a preliminary stage of high-grade dysplasia and adenocarcinoma (10). Future CAD systems should consider including low grade dysplastic imagery in training and testing to make it more practicable for real-life settings.

Second, *Test set 3* only consisted of 8 NDBE images. With such a small sample, a classification by chance cannot be ruled out. We tried to take this limitation into consideration and restrained from putting too much weight on results based on *Test set 3*.

Third, our test sets show very different levels of balance, meaning they have different proportions of NDBE to neoplastic images. This affects the metrics positive predictive value and negative predictive value. Thus, the comparability between the test sets using these two metrics is limited.

Fourth, the algorithm was not trained on posttreatment images. Posttreatment data differs from pretreatment data by appearance, for example the mucosa looks different. Still, we used the algorithm to classify posttreatment images in *Test set 4*. Because of the missing specific training, we cannot say for sure if the performance was just a coincidence or if it is attributable to the skills of the CAD system.

Fifth, we did not statistically test the observed differences between the groups, nor did we perform a power analysis beforehand. This means, we cannot formally assess whether the observed variations are statistically significant or might have occurred by chance. However, it should be noted that this approach is relatively common in the field of AI research, particularly in exploratory studies. Also, our main objective was to generate insights rather than providing definitive conclusions regarding the clinical applicability of the system.

Sixth, we investigated only a single algorithm. Therefore, our conclusions only apply to this specific model. It would have been better to evaluate multiple algorithms in order to be able to make a more robust and generally valid statement. However, this would have exceeded the scope of this thesis.

6 Conclusion

This research investigated the generalizability of a computer-aided diagnosis (CAD) system for Barrett's neoplasia by evaluating its performance on previously unseen domains. We assessed the algorithm's performance on different test sets: one with images from a different endoscopy system, one with images from a new medical center, and one with post-treatment images. Although our test sets were partly too small and imbalanced, we can conclude that the CAD system, in its current form, should not yet be used in new, especially non-expert clinical settings.

The observed decline in performance likely results from several factors: overfitting to the original training data, substantial domain shifts, and insufficient domain-specific training.

For the future development of our CAD system, several adjustments should be made. First, Olympus imagery should be added to training and validation sets to stabilize performance. Second, training on data from multiple medical centers would create more heterogeneity in terms of institutional habits and patient population. Third, the algorithm should be trained on posttreatment imagery. Alternatively, separate algorithms based on the structure of the existing CAD system for Olympus imagery or posttreatment imagery could be developed.

We recommend the introduction of explainable AI methods into medical deep learning systems to enhance transparency and clinical trustworthiness. Further, we advocate for stronger collaboration within the endoscopic AI community to develop larger and more diverse datasets, regarding data quality, geographic and temporal diversity and variety of used endoscopic systems. Alternatively, the option of individualizing CAD models by fine-tuning with local hospital data for an improved performance should be explored. More standardized and transparent frameworks for training, testing and reporting results, possibly including independent benchmarking sets, could improve comparability and reproducibility between models.

Above all, transparent and qualitative external validation should be carried out as standard for new models. Feasible trials for AI algorithms in endoscopy should be developed in order to obtain reliable scientific results. This is essential for the responsible and safe implementation of AI in the treatment of patients.

7 Appendix

7.1 List of Tables

Table 1. Surveillance intervals for Barrett's esophagus recommended by the European Society of Gastrointestinal Endoscopy. 24
Table 2. Overview of datasets for training and validation of the used CAD system. 32
Table 3. Overview of included test sets 34
Table 4. Inclusion and exclusion criteria for the test sets 35
Table 5. Confusion matrix. 37
Table 6. Performance results of the CAD system on our test sets. 40
Table 7. Confusion matrix for *Test set 1*. 40
Table 8. Confusion matrix for *Test set 2*. 41
Table 9. Confusion matrix for *Test set 3*. 42
Table 10. Confusion matrix for *Test set 4*. 42
Table 11. Overview of studies on CAD systems for neoplasia detection in Barrett's esophagus (adenocarcinoma). 49

7.2 List of Figures

Figure 1. Physiological gastroesophageal junction. 22
Figure 2. Barrett’s esophagus. 22
Figure 3. Progression of Barrett’s esophagus. 23
Figure 4. Simplified working mechanism of a neural network. 25
Figure 5. Overview of the characteristics of the test sets. 31
Figure 6. Example for image. 35
Figure 7. Example for retroflexion image. 35
Figure 8. Example for cap image. 35
Figure 9. Receiver operating characteristic curves for *Test set 1* and *2*. 43
Figure 10. Receiver operating characteristic curves for *Test set 2* and *3*. 44
Figure 11. Receiver operating characteristic curves for *Test set 3* and *4*. 45
Figure 12. Receiver operating characteristic curves for *Test set 1*, *2*, *3* and *4*. 45

8 References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63. eng. doi:10.3322/caac.21834 Cited in: PubMed; PMID 38572751.
2. Kolb JM, Wani S. Barrett's esophagus: current standards in advanced imaging. Transl Gastroenterol Hepatol. 2021;614. eng. doi:10.21037/tgh.2020.02.10 Cited in: PubMed; PMID 33409408.
3. Boerwinkel DF, Swager A-F, Curvers WL, Bergman JJGHM. The clinical consequences of advanced imaging techniques in Barrett's esophagus. Gastroenterology. 2014;146(3):622-629.e4. eng. doi:10.1053/j.gastro.2014.01.007 Cited in: PubMed; PMID 24412487.
4. Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. N Engl J Med. 2011;365(15):1375–83. eng. doi:10.1056/NEJMoa1103042 Cited in: PubMed; PMID 21995385.
5. Chowdhury M, Sadek AW. Advantages and Limitations of Artificial Intelligence [Internet]. 2012;(E-C168). Available from: https://trid.trb.org/view/1225151
6. Struyvenberg MR, Groof AJ de, Bergman JJ, van der Sommen F, With PHN de, Konda VJA, Curvers WL. Advanced Imaging and Sampling in Barrett's Esophagus: Artificial Intelligence to the Rescue? Gastrointestinal Endoscopy Clinics of North America. 2021;31(1):91–103. eng. doi:10.1016/j.giec.2020.08.006 Cited in: PubMed; PMID 33213802.
7. Parasher G, Wong M, Rawat M. Evolving role of artificial intelligence in gastrointestinal endoscopy. World J Gastroenterol. 2020;26(46):7287–98. eng. doi:10.3748/wjg.v26.i46.7287 Cited in: PubMed; PMID 33362384.
8. Schölvinck DW, van der Meulen K, Bergman JJGHM, Weusten BLAM. Detection of lesions in dysplastic Barrett's esophagus by community and expert endoscopists. Endoscopy. 2017;49(2):113–20. eng. doi:10.1055/s-0042-118312 Cited in: PubMed; PMID 27855466.
9. Abadir AP, Ali MF, Karnes W, Samarasena JB. Artificial Intelligence in Gastrointestinal Endoscopy. Clin Endosc. 2020;53(2):132–41. eng. doi:10.5946/ce.2020.038 Cited in: PubMed; PMID 32252506.
10. Stein HJ, Siewert JR. Barrett's esophagus: pathogenesis, epidemiology, functional abnormalities, malignant degeneration, and surgical management. Dysphagia. 1993;8(3):276–88. eng. doi:10.1007/BF01354551 Cited in: PubMed; PMID 8359051.
11. Spechler SJ. Clinical practice. Barrett's Esophagus. N Engl J Med. 2002;346(11):836–42. eng. doi:10.1056/NEJMcp012118 Cited in: PubMed; PMID 11893796.
12. Hayeck TJ, Kong CY, Spechler SJ, Gazelle GS, Hur C. The prevalence of Barrett's esophagus in the US: estimates from a simulation model confirmed by SEER data. Dis Esophagus. 2010;23(6):451–7. eng. doi:10.1111/j.1442-2050.2010.01054.x Cited in: PubMed; PMID 20353441.

13. Rex DK, Cummings OW, Shaw M, Cumings MD, Wong RKH, Vasudeva RS, Dunne D, Rahmani EY, Helper DJ. Screening for Barrett's esophagus in colonoscopy patients with and without heartburn. Gastroenterology. 2003;125(6):1670–7. eng. doi:10.1053/j.gastro.2003.09.030 Cited in: PubMed; PMID 14724819.
14. Ronkainen J, Aro P, Storskrubb T, Johansson S-E, Lind T, Bolling-Sternevald E, Vieth M, Stolte M, Talley NJ, Agréus L. Prevalence of Barrett's esophagus in the general population: an endoscopic study. Gastroenterology. 2005;129(6):1825–31. eng. doi:10.1053/j.gastro.2005.08.053 Cited in: PubMed; PMID 16344051.
15. Naini BV, Chak A, Ali MA, Odze RD. Barrett's oesophagus diagnostic criteria: endoscopy and histology. Best Pract Res Clin Gastroenterol. 2015;29(1):77–96. eng. doi:10.1016/j.bpg.2014.11.004 Cited in: PubMed; PMID 25743458.
16. Thrift AP. The epidemic of oesophageal carcinoma: Where are we now? Cancer Epidemiol. 2016;4188–95. eng. doi:10.1016/j.canep.2016.01.013 Cited in: PubMed; PMID 26851752.
17. Thrift AP, Whiteman DC. The incidence of esophageal adenocarcinoma continues to rise: analysis of period and birth cohort effects on recent trends. Ann Oncol. 2012;23(12):3155–62. eng. doi:10.1093/annonc/mds181 Cited in: PubMed; PMID 22847812.
18. Verbeek RE, Leenders M, Kate FJW ten, van Hillegersberg R, Vleggaar FP, van Baal JWPM, van Oijen MGH, Siersema PD. Surveillance of Barrett's esophagus and mortality from esophageal adenocarcinoma: a population-based cohort study. Am J Gastroenterol. 2014;109(8):1215–22. eng. doi:10.1038/ajg.2014.156 Cited in: PubMed; PMID 24980881.
19. Weusten BLAM, Bisschops R, Dinis-Ribeiro M, Di Pietro M, Pech O, Spaander MCW, Baldaque-Silva F, Barret M, Coron E, Fernández-Esparrach G, Fitzgerald RC, Jansen M, Jovani M, Marques-de-Sa I, Rattan A, Tan WK, Verheij EPD, Zellenrath PA, Triantafyllou K, Pouw RE. Diagnosis and management of Barrett esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Guideline. Endoscopy. 2023;55(12):1124–46. eng. doi:10.1055/a-2176-2440 Cited in: PubMed; PMID 37813356.
20. Obermannová R, Alsina M, Cervantes A, Leong T, Lordick F, Nilsson M, van Grieken NCT, Vogel A, Smyth EC. Oesophageal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2022;33(10):992–1004. eng. doi:10.1016/j.annonc.2022.07.003 Cited in: PubMed; PMID 35914638.
21. Bennett C, Green S, DeCaestecker J, Almond M, Barr H, Bhandari P, Ragunath K, Singh R, Jankowski J. Surgery versus radical endotherapies for early cancer and high-grade dysplasia in Barrett's oesophagus. Cochrane Database Syst Rev. 2020;5CD007334. eng. doi:10.1002/14651858.CD007334.pub5 Cited in: PubMed; PMID 32442322.
22. Pech O, Bollschweiler E, Manner H, Leers J, Ell C, Hölscher AH. Comparison between endoscopic and surgical resection of mucosal esophageal adenocarcinoma in Barrett's esophagus at two high-volume centers. Ann Surg. 2011;254(1):67–72. eng. doi:10.1097/SLA.0b013e31821d4bf6 Cited in: PubMed; PMID 21532466.
23. Zehetner J, DeMeester SR, Hagen JA, Ayazi S, Augustin F, Lipham JC, DeMeester TR. Endoscopic resection and ablation versus esophagectomy for high-grade dysplasia and

intramucosal adenocarcinoma. J Thorac Cardiovasc Surg. 2011;141(1):39–47. eng. doi:10.1016/j.jtcvs.2010.08.058 Cited in: PubMed; PMID 21055772.

24. Mintz Y, Brodie R. Introduction to artificial intelligence in medicine. Minim Invasive Ther Allied Technol. 2019;28(2):73–81. eng. doi:10.1080/13645706.2019.1575882 Cited in: PubMed; PMID 30810430.
25. Bini SA. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty. 2018;33(8):2358–61. eng. doi:10.1016/j.arth.2018.02.067 Cited in: PubMed; PMID 29656964.
26. Kriegeskorte N, Golan T. Neural network models and deep learning. Curr Biol. 2019;29(7):R231-R236. eng. doi:10.1016/j.cub.2019.02.034 Cited in: PubMed; PMID 30939301.
27. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44. eng. doi:10.1038/nature14539 Cited in: PubMed; PMID 26017442.
28. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018;9(4):611–29. eng. doi:10.1007/s13244-018-0639-9 Cited in: PubMed; PMID 29934920.
29. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR): IEEE; 6/27/2016 - 6/30/2016. p. 770–8.
30. Kim MJ, Kim SH, Kim SM, Nam JH, Hwang YB, Lim YJ. The Advent of Domain Adaptation into Artificial Intelligence for Gastrointestinal Endoscopy and Medical Imaging. Diagnostics (Basel). 2023;13(19). eng. doi:10.3390/diagnostics13193023 Cited in: PubMed; PMID 37835766.
31. Stacke K, Eilertsen G, Unger J, Lundstrom C. Measuring Domain Shift for Deep Learning in Histopathology. IEEE J Biomed Health Inform. 2021;25(2):325–36. eng. doi:10.1109/JBHI.2020.3032060 Cited in: PubMed; PMID 33085623.
32. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn. 2010;79(1-2):151–75. doi:10.1007/s10994-009-5152-4
33. Liedlgruber M, Uhl A. Computer-aided decision support systems for endoscopy in the gastrointestinal tract: a review. IEEE Rev Biomed Eng. 2011;473–88. eng. doi:10.1109/RBME.2011.2175445 Cited in: PubMed; PMID 22273792.
34. Abdelrahim M, Saiko M, Maeda N, Hossain E, Alkandari A, Subramaniam S, Parra-Blanco A, Sanchez-Yague A, Coron E, Repici A, Bhandari P. Development and validation of artificial neural networks model for detection of Barrett's neoplasia: a multicenter pragmatic nonrandomized trial (with video). Gastrointest Endosc. 2023;97(3):422–34. eng. doi:10.1016/j.gie.2022.10.031 Cited in: PubMed; PMID 36283443.
35. Ebigbo A, Mendel R, Probst A, Manzeneder J, Souza LA de, Papa JP, Palm C, Messmann H. Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma. Gut. 2019;68(7):1143–5. eng. doi:10.1136/gutjnl-2018-317573 Cited in: PubMed; PMID 30510110.

36. Ebigbo A, Mendel R, Probst A, Manzeneder J, Prinz F, Souza LA de, Papa J, Palm C, Messmann H. Real-time use of artificial intelligence in the evaluation of cancer in Barrett's oesophagus. Gut. 2020;69(4):615–6. eng. doi:10.1136/gutjnl-2019-319460 Cited in: PubMed; PMID 31541004.

37. Fockens KN, Jong MR, Jukema JB, Boers TGW, Kusters CHJ, van der Putten JA, Pouw RE, Duits LC, Montazeri NSM, van Munster SN, Weusten BLAM, Alvarez Herrero L, Houben MHMG, Nagengast WB, Westerhof J, Alkhalaf A, Mallant-Hent RC, Scholten P, Ragunath K, Seewald S, Elbe P, Baldaque-Silva F, Barret M, Ortiz Fernández-Sordo J, Villarejo GM, Pech O, Beyna T, van der Sommen F, With PH de, Groof AJ de, Bergman JJ. A deep learning system for detection of early Barrett's neoplasia: a model development and validation study. Lancet Digit Health. 2023;5(12):e905-e916. eng. doi:10.1016/S2589-7500(23)00199-1 Cited in: PubMed; PMID 38000874.

38. Fockens KN, Jukema JB, Boers T, Jong MR, van der Putten JA, Pouw RE, Weusten BLAM, Alvarez Herrero L, Houben MHMG, Nagengast WB, Westerhof J, Alkhalaf A, Mallant R, Ragunath K, Seewald S, Elbe P, Barret M, Ortiz Fernández-Sordo J, Pech O, Beyna T, van der Sommen F, With PH de, Groof AJ de, Bergman JJ. Towards a robust and compact deep learning system for primary detection of early Barrett's neoplasia: Initial image-based results of training on a multi-center retrospectively collected data set. United European Gastroenterol J. 2023;11(4):324–36. eng. doi:10.1002/ueg2.12363 Cited in: PubMed; PMID 37095718.

39. Fockens K, Jukema J, Jong M, Boers T, Kusters K, van der Putten J, Pouw R, Duits L, Weusten B, Herrero LA, Houben M, Nagengast W, Westerhof J, Alkhalaf A, Mallant-Hent R, Scholten P, Ragunath K, Seewald S, Elbe P, Baldaque-Silva F, Barret M, Fernández-Sordo JO, Villarejo GM, Pech O, Beyna T, van der Sommen F, With P de, Groof J de, Bergman J. VIDEO-BASED COMPUTER AIDED DETECTION SYSTEM DETECTS BARRETT'S NEOPLASIA WITH HIGH ACCURACY DURING LIVE ENDOSCOPIC PROCEDURES: A MULTI-CENTER PILOT AND FEASIBILITY STUDY. Gastrointest Endosc. 2023;97(6):AB1001-AB1002. doi:10.1016/j.gie.2023.04.1547

40. Ghatwary N, Zolgharni M, Ye X. Early esophageal adenocarcinoma detection using deep learning methods. Int J Comput Assist Radiol Surg. 2019;14(4):611–21. eng. doi:10.1007/s11548-019-01914-4 Cited in: PubMed; PMID 30666547.

41. Groof AJ de, 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, With PH de, Bergman JJ. 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. 2020;158(4):915-929.e4. eng. doi:10.1053/j.gastro.2019.11.030 Cited in: PubMed; PMID 31759929.

42. Hashimoto R, Requa J, Dao T, Ninh A, Tran E, Mai D, Lugo M, El-Hage Chehade N, Chang KJ, Karnes WE, Samarasena JB. Artificial intelligence using convolutional neural networks for real-time detection of early esophageal neoplasia in Barrett's esophagus (with video). Gastrointest Endosc. 2020;91(6):1264-1271.e1. eng. doi:10.1016/j.gie.2019.12.049 Cited in: PubMed; PMID 31930967.

43. 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. A new artificial intelligence system successfully detects and localises early neoplasia in Barrett's esophagus by using convolutional neural networks. United European Gastroenterol J. 2022;10(6):528–37. eng. doi:10.1002/ueg2.12233 Cited in: PubMed; PMID 35521666.

44. Struyvenberg MR, Groof AJ de, van der Putten J, van der Sommen F, Baldaque-Silva F, Omae M, Pouw R, Bisschops R, Vieth M, Schoon EJ, Curvers WL, With PH de, Bergman JJ. A computer-assisted algorithm for narrow-band imaging-based tissue characterization in Barrett's esophagus. Gastrointest Endosc. 2021;93(1):89–98. eng. doi:10.1016/j.gie.2020.05.050 Cited in: PubMed; PMID 32504696.

45. Mendel R, Ebigbo A, Probst A, Messmann H, Palm C. Barrett's Esophagus Analysis Using Convolutional Neural Networks. In: Maier-Hein gFH, Deserno gLM, Handels H, Tolxdorff T, editors. Bildverarbeitung für die Medizin 2017. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 80–5. (Informatik aktuell).

46. Ko WJ, An P, Ko KH, Hahm KB, Hong SP, Cho JY. Image Quality Analysis of Various Gastrointestinal Endoscopes: Why Image Quality Is a Prerequisite for Proper Diagnostic and Therapeutic Endoscopy. Clin Endosc. 2015;48(5):374–9. eng. doi:10.5946/ce.2015.48.5.374 Cited in: PubMed; PMID 26473119.

47. NEC's AI supports doctors to detect neoplasia in Barrett's esophagus during endoscopic procedures [Internet]. 2021 [cited 2025 Mar 27]. Available from: https://www.nec.com/en/press/202105/global_20210528_01.html

48. Odin Vision. CADU A new era of Cloud AI enabled Gastroscopy [Internet] [cited 2025 Mar 27]. Available from: https://odin-vision.com/wp-content/uploads/2022/10/CADU-brochure.pdf

49. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019;19(1):64. eng. doi:10.1186/s12874-019-0681-4 Cited in: PubMed; PMID 30890124.

50. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. eng. doi:10.1038/s41591-018-0300-7 Cited in: PubMed; PMID 30617339.

51. Yousef F, Cardwell C, Cantwell MM, Galway K, Johnston BT, Murray L. The incidence of esophageal cancer and high-grade dysplasia in Barrett's esophagus: a systematic review and meta-analysis. Am J Epidemiol. 2008;168(3):237–49. eng. doi:10.1093/aje/kwn121 Cited in: PubMed; PMID 18550563.

52. van der Sommen F, Groof J de, Struyvenberg M, van der Putten J, Boers T, Fockens K, Schoon EJ, Curvers W, With P de, Mori Y, Byrne M, Bergman JJGHM. Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. Gut. 2020;69(11):2035–45. eng. doi:10.1136/gutjnl-2019-320466 Cited in: PubMed; PMID 32393540.

53. Guan H, Liu M. Domain Adaptation for Medical Image Analysis: A Survey. IEEE Trans Biomed Eng. 2022;69(3):1173–85. eng. doi:10.1109/TBME.2021.3117407 Cited in: PubMed; PMID 34606445.

54. Celik N, Gupta S, Ali S, Rittscher J. Unsupervised Adversarial Domain Adaptation For Barrett's Segmentation; 09/12/2020.

55. Arribas J, Antonelli G, Frazzoni L, Fuccio L, Ebigbo A, van der Sommen F, Ghatwary N, Palm C, Coimbra M, Renna F, Bergman JJGHM, Sharma P, Messmann H, Hassan C, Dinis-Ribeiro MJ. Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. Gut. 2020;70(8):1458–68. eng. doi:10.1136/gutjnl-2020-321922 Cited in: PubMed; PMID 33127833.

56. Groof J de, van der Sommen F, van der Putten J, Struyvenberg MR, Zinger S, Curvers WL, Pech O, Meining A, Neuhaus H, Bisschops R, Schoon EJ, With PH de, Bergman JJ. The Argos project: The development of a computer-aided detection system to improve detection of Barrett's neoplasia on white light endoscopy. United European Gastroenterol J. 2019;7(4):538–47. eng. doi:10.1177/2050640619837443 Cited in: PubMed; PMID 31065371.

57. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD, Johansen D, Griwodz C, Stensland HK, Garcia-Ceja E, Schmidt PT, Hammer HL, Riegler MA, Halvorsen P, Lange T de. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific Data. 2020;7(1):283. eng. doi:10.1038/s41597-020-00622-y Cited in: PubMed; PMID 32859981.

58. Schmitz R, Kather JN. Artificial intelligence in Barrett's oesophagus and the need for shared and combined data. United European Gastroenterol J. 2022;10(6):525–7. eng. doi:10.1002/ueg2.12260 Cited in: PubMed; PMID 35704382.

59. Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, Salto-Tellez M, Alwers E, Cifci D, Ghaffari Laleh N, Seibel T, Gray R, Hutchins GGA, Brenner H, van Treeck M, Yuan T, Brinker TJ, Chang-Claude J, Khader F, Schuppert A, Luedde T, Trautwein C, Muti HS, Foersch S, Hoffmeister M, Truhn D, Kather JN. Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat Med. 2022;28(6):1232–9. eng. doi:10.1038/s41591-022-01768-5 Cited in: PubMed; PMID 35469069.

60. Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, Aussó S, Alberich LC, Marias K, Tsiknakis M, Colantonio S, Papanikolaou N, Salahuddin Z, Woodruff HC, Lambin P, Martí-Bonmatí L. FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging; 20/09/2021.

61. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–49. eng. doi:10.3322/caac.21660 Cited in: PubMed; PMID 33538338.

62. Mizuno H, Gono K, Takehana S, Nonami T, Nakamura K. Narrow band imaging technique. Techniques in Gastrointestinal Endoscopy. 2003;5(2):78–81. doi:10.1053/tgie.2003.50001

63. Qumseya B, Sultan S, Bain P, Jamil L, Jacobson B, Anandasabapathy S, Agrawal D, Buxbaum JL, Fishman DS, Gurudu SR, Jue TL, Kripalani S, Lee JK, Khashab MA, Naveed M, Thosani NC, Yang J, DeWitt J, Wani S. ASGE guideline on screening and

surveillance of Barrett's esophagus. Gastrointest Endosc. 2019;90(3):335-359.e2. eng. doi:10.1016/j.gie.2019.05.012 Cited in: PubMed; PMID 31439127.

64. Trindade AJ, Smith MS, Pleskow DK. The new kid on the block for advanced imaging in Barrett's esophagus: a review of volumetric laser endomicroscopy. Therap Adv Gastroenterol. 2016;9(3):408–16. eng. doi:10.1177/1756283X16639003 Cited in: PubMed; PMID 27134668.
65. ASGE Technology Committee. Confocal laser endomicroscopy. Gastrointest Endosc. 2014;80(6):928–38. eng. doi:10.1016/j.gie.2014.06.021 Cited in: PubMed; PMID 25442092.
66. Newton RC, Kemp SV, Shah PL, Elson D, Darzi A, Shibuya K, Mulgrew S, Yang G-Z. Progress toward optical biopsy: bringing the microscope to the patient. Lung. 2011;189(2):111–9. eng. doi:10.1007/s00408-011-9282-7 Cited in: PubMed; PMID 21336942.
67. Shaheen NJ, Falk GW, Iyer PG, Gerson LB. ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. Am J Gastroenterol. 2016;111(1):30-50; quiz 51. eng. doi:10.1038/ajg.2015.322 Cited in: PubMed; PMID 26526079.
68. Qumseya BJ, Bukannan A, Gendy S, Ahemd Y, Sultan S, Bain P, Gross SA, Iyer P, Wani S. Systematic review and meta-analysis of prevalence and risk factors for Barrett's esophagus. Gastrointest Endosc. 2019;90(5):707-717.e1. eng. doi:10.1016/j.gie.2019.05.030 Cited in: PubMed; PMID 31152737.
69. Kang JY. Systematic review: geographical and ethnic differences in gastro-oesophageal reflux disease. Aliment Pharmacol Ther. 2004;20(7):705–17. eng. doi:10.1111/j.1365-2036.2004.02165.x Cited in: PubMed; PMID 15379831.
70. Weusten B, Bisschops R, Coron E, Dinis-Ribeiro M, Dumonceau J-M, Esteban J-M, Hassan C, Pech O, Repici A, Bergman J, Di Pietro M. Endoscopic management of Barrett's esophagus: European Society of Gastrointestinal Endoscopy (ESGE) Position Statement. Endoscopy. 2017;49(2):191–8. eng. doi:10.1055/s-0042-122140 Cited in: PubMed; PMID 28122386.
71. Cohen J, Safdi MA, Deal SE, Baron TH, Chak A, Hoffman B, Jacobson BC, Mergener K, Petersen BT, Petrini JL, Rex DK, Faigel DO, Pike IM. Quality indicators for esophagogastroduodenoscopy. Gastrointest Endosc. 2006;63(4 Suppl):S10-5. eng. doi:10.1016/j.gie.2006.02.018 Cited in: PubMed; PMID 16564907.
72. Pschyrembel W. Pschyrembel Klinisches Wörterbuch. 266th ed. Des Verlages dP-R, editor. Berlin, Boston: De Gruyter; 2017. 2377 p. Available from: https://www.degruyter.com/doi/book/10.1515/9783110539233 ger.
73. Conio M, Cameron AJ, Romero Y, Branch CD, Schleck CD, Burgart LJ, Zinsmeister AR, Melton LJ, Locke GR. Secular trends in the epidemiology and outcome of Barrett's oesophagus in Olmsted County, Minnesota. Gut. 2001;48(3):304–9. eng. doi:10.1136/gut.48.3.304 Cited in: PubMed; PMID 11171817.
74. Sikkema M, Jonge PJF de, Steyerberg EW, Kuipers EJ. Risk of esophageal adenocarcinoma and mortality in patients with Barrett's esophagus: a systematic review and meta-analysis. Clin Gastroenterol Hepatol. 2010;8(3):235-44; quiz e32. eng. doi:10.1016/j.cgh.2009.10.010 Cited in: PubMed; PMID 19850156.

75. Al-Kasspooles MF, Hill HC, Nava HR, Smith JL, Douglass HO, Gibbs JF. High-grade dysplasia within Barrett's esophagus: controversies regarding clinical opinions and approaches. Ann Surg Oncol. 2002;9(3):222–7. eng. doi:10.1007/BF02573058 Cited in: PubMed; PMID 11923127.

76. Phoa KN, van Vilsteren FGI, Weusten BLAM, Bisschops R, Schoon EJ, Ragunath K, Fullarton G, Di Pietro M, Ravi N, Visser M, Offerhaus GJ, Seldenrijk CA, Meijer SL, Kate FJW ten, Tijssen JGP, Bergman JJGHM. Radiofrequency ablation vs endoscopic surveillance for patients with Barrett esophagus and low-grade dysplasia: a randomized clinical trial. JAMA. 2014;311(12):1209–17. eng. doi:10.1001/jama.2014.2511 Cited in: PubMed; PMID 24668102.

77. Desai M, Saligram S, Gupta N, Vennalaganti P, Bansal A, Choudhary A, Vennelaganti S, He J, Titi M, Maselli R, Qumseya B, Olyaee M, Waxman I, Repici A, Hassan C, Sharma P. Efficacy and safety outcomes of multimodal endoscopic eradication therapy in Barrett's esophagus-related neoplasia: a systematic review and pooled analysis. Gastrointest Endosc. 2017;85(3):482-495.e4. eng. doi:10.1016/j.gie.2016.09.022 Cited in: PubMed; PMID 27670227.

78. Bennett C, Green S, DeCaestecker J, Almond M, Barr H, Bhandari P, Ragunath K, Singh R, Jankowski J. Surgery versus radical endotherapies for early cancer and high-grade dysplasia in Barrett's oesophagus. Cochrane Database Syst Rev. 2012;11CD007334. eng. doi:10.1002/14651858.CD007334.pub4 Cited in: PubMed; PMID 23152243.

79. National Oesophago-gastric Audit 2016. The NHS Information Centre. Annual report of the national oesophago-gastric audit 2016. [Internet]. 2016 [cited 2021 Jun 22]. Available from: https://files.digital.nhs.uk/publicationimport/pub21xxx/pub21561/clin-audi-supp-prog-oeso-gast-2016-rep.pdf

80. National Oesophago-gastric Audit 2015. The NHS Information Centre. Annual report of the national oesophago-gastric audit 2015. [Internet] [cited 2021 Jun 22]. Available from: https://files.digital.nhs.uk/publicationimport/pub19xxx/pub19627/clin-audi-supp-prog-oeso-gast-2015-rep.pdf

81. Phoa KN, Pouw RE, Bisschops R, Pech O, Ragunath K, Weusten BLAM, Schumacher B, Rembacken B, Meining A, Messmann H, Schoon EJ, Gossner L, Mannath J, Seldenrijk CA, Visser M, Lerut T, Seewald S, Kate FJ ten, Ell C, Neuhaus H, Bergman JJGHM. Multimodality endoscopic eradication for neoplastic Barrett oesophagus: results of an European multicentre study (EURO-II). Gut. 2016;65(4):555–62. eng. doi:10.1136/gutjnl-2015-309298 Cited in: PubMed; PMID 25731874.

82. Currie G, Hawk KE, Rohren E, Vial A, Klein R. Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. J Med Imaging Radiat Sci. 2019;50(4):477–87. eng. doi:10.1016/j.jmir.2019.09.005 Cited in: PubMed; PMID 31601480.

83. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, van der Laak JAWM, van Ginneken B, Sánchez CI. A survey on deep learning in medical image analysis. Med Image Anal. 2017;4260–88. eng. doi:10.1016/j.media.2017.07.005 Cited in: PubMed; PMID 28778026.

84. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer

International Publishing; 2015. p. 234–41. (Lecture Notes in Computer Science; vol. 9351).

85. van der Sommen F, Zinger S, Schoon EJ, With PHN de. Computer-aided detection of early cancer in the esophagus using HD endoscopy images. In: Novak CL, Aylward S, editors. Medical Imaging 2013: Computer-Aided Diagnosis: SPIE; 2013. 86700V.
86. Boschetto D, Gambaretto G, Grisan E. Automatic classification of endoscopic images for premalignant conditions of the esophagus. In: Gimi B, Krol A, editors. Medical Imaging 2016: Biomedical Applications in Molecular, Structural, and Functional Imaging: SPIE; 2016. p. 978808.
87. Kage A, Raithel M, Zopf S, Wittenberg T, Münzenmayer C. Narrow-band imaging for the computer assisted diagnosis in patients with Barrett's esophagus. In: Karssemeijer N, Giger ML, editors. Medical Imaging 2009: Computer-Aided Diagnosis: SPIE; 2009. 72603S.
88. Sharma P, Bergman JJGHM, Goda K, Kato M, Messmann H, Alsop BR, Gupta N, Vennalaganti P, Hall M, Konda V, Koons A, Penner O, Goldblum JR, Waxman I. Development and Validation of a Classification System to Identify High-Grade Dysplasia and Esophageal Adenocarcinoma in Barrett's Esophagus Using Narrow-Band Imaging. Gastroenterology. 2016;150(3):591–8. eng. doi:10.1053/j.gastro.2015.11.037 Cited in: PubMed; PMID 26627609.
89. Klomp S, van der Sommen F, Swager A-F, Zinger S, Schoon EJ, Curvers WL, Bergman JJ, With PHN de. Evaluation of image features and classification methods for Barrett's cancer detection using VLE imaging. In: Armato SG, Petrick NA, editors. Medical Imaging 2017: Computer-Aided Diagnosis: SPIE; 2017. 101340D.
90. Swager A-F, van der Sommen F, Klomp SR, Zinger S, Meijer SL, Schoon EJ, Bergman JJGHM, With PH de, Curvers WL. Computer-aided detection of early Barrett's neoplasia using volumetric laser endomicroscopy. Gastrointest Endosc. 2017;86(5):839–46. eng. doi:10.1016/j.gie.2017.03.011 Cited in: PubMed; PMID 28322771.
91. Knabe M, Welsch L, Blasberg T, Müller E, Heilani M, Bergen C, Herrmann E, May A. Artificial intelligence-assisted staging in Barrett's carcinoma. Endoscopy. 2022;54(12):1191–7. eng. doi:10.1055/a-1811-9407 Cited in: PubMed; PMID 35354219.
92. Ebigbo A, Mendel R, Rückert T, Schuster L, Probst A, Manzeneder J, Prinz F, Mende M, Steinbrück I, Faiss S, Rauber D, Souza LA de, Papa JP, Deprez PH, Oyama T, Takahashi A, Seewald S, Sharma P, Byrne MF, Palm C, Messmann H. Endoscopic prediction of submucosal invasion in Barrett's cancer with the use of artificial intelligence: a pilot study. Endoscopy. 2021;53(9):878–83. eng. doi:10.1055/a-1311-8570 Cited in: PubMed; PMID 33197942.
93. Veronese E, Grisan E, Diamantis G, Crosta C, Trovato C. Hybrid patch-based and image-wide classification of confocal laser endomicroscopy images in Barrett's esophagus surveillance. In: ; 2013. p. 362–5.
94. Hong J, Park B, Park H. Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC): IEEE; 7/11/2017 - 7/15/2017. p. 2892–5.

95. Faghani S, Codipilly DC, David V, Moassefi M, Rouzrokh P, Khosravi B, Agarwal S, Dhaliwal L, Katzka DA, Hagen C, Lewis J, Leggett CL, Erickson BJ, Iyer PG. Development of a deep learning model for the histologic diagnosis of dysplasia in Barrett's esophagus. Gastrointest Endosc. 2022;96(6):918-925.e3. eng. doi:10.1016/j.gie.2022.06.013 Cited in: PubMed; PMID 35718071.
96. Zhang Z, Bai L, Ren P, Hancock ER. High-order graph matching kernel for early carcinoma EUS image classification. Multimed Tools Appl. 2016;75(7):3993–4012. doi:10.1007/s11042-015-3108-1
97. Liu Y, Zhou Y, Wen S, Tang C. A Strategy on Selecting Performance Metrics for Classifier Evaluation. International Journal of Mobile Computing and Multimedia Communications. 2014;6(4):20–35. doi:10.4018/IJMCMC.2014100102
98. Naser MZ, Alavi A. Insights into Performance Fitness and Error Metrics for Machine Learning. 2020.
99. Jeni LA, Cohn JF, La Torre F de. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. Int Conf Affect Comput Intell Interact Workshops. 2013;2013245–51. eng. doi:10.1109/ACII.2013.47 Cited in: PubMed; PMID 25574450.
100. Codipilly DC, Chandar AK, Singh S, Wani S, Shaheen NJ, Inadomi JM, Chak A, Iyer PG. The Effect of Endoscopic Surveillance in Patients With Barrett's Esophagus: A Systematic Review and Meta-analysis. Gastroenterology. 2018;154(8):2068-2086.e5. eng. doi:10.1053/j.gastro.2018.02.022 Cited in: PubMed; PMID 29458154.
101. Cai S-L, Li B, Tan W-M, Niu X-J, Yu H-H, Yao L-Q, Zhou P-H, Yan B, Zhong Y-S. Using a deep learning system in endoscopy for screening of early esophageal squamous cell carcinoma (with video). Gastrointest Endosc. 2019;90(5):745-753.e2. eng. doi:10.1016/j.gie.2019.06.044 Cited in: PubMed; PMID 31302091.
102. Guo L, Xiao X, Wu C, Zeng X, Zhang Y, Du J, Bai S, Xie J, Zhang Z, Li Y, Wang X, Cheung O, Sharma M, Liu J, Hu B. Real-time automated diagnosis of precancerous lesions and early esophageal squamous cell carcinoma using a deep learning model (with videos). Gastrointest Endosc. 2020;91(1):41–51. eng. doi:10.1016/j.gie.2019.08.018 Cited in: PubMed; PMID 31445040.
103. Horie Y, Yoshio T, Aoyama K, Yoshimizu S, Horiuchi Y, Ishiyama A, Hirasawa T, Tsuchida T, Ozawa T, Ishihara S, Kumagai Y, Fujishiro M, Maetani I, Fujisaki J, Tada T. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. Gastrointest Endosc. 2019;89(1):25–32. eng. doi:10.1016/j.gie.2018.07.037 Cited in: PubMed; PMID 30120958.
104. Kanesaka T, Lee T-C, Uedo N, Lin K-P, Chen H-Z, Lee J-Y, Wang H-P, Chang H-T. Computer-aided diagnosis for identifying and delineating early gastric cancers in magnifying narrow-band imaging. Gastrointest Endosc. 2018;87(5):1339–44. eng. doi:10.1016/j.gie.2017.11.029 Cited in: PubMed; PMID 29225083.
105. Souza LA de, Afonso LCS, Palm C, Papa JP. Barrett's Esophagus Identification Using Optimum-Path Forest. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI): IEEE; 10/17/2017 - 10/20/2017. p. 308–14.
106. van der Sommen F, Zinger S, Curvers WL, Bisschops R, Pech O, Weusten BLAM, Bergman JJGHM, With PHN de, Schoon EJ. Computer-aided detection of early

neoplastic lesions in Barrett's esophagus. Endoscopy. 2016;48(7):617–24. eng. doi:10.1055/s-0042-105284 Cited in: PubMed; PMID 27100718.

107. Portsmouth Hospitals NHS Trust. Real-life Prospective Evaluation of Computer-aided Detection (CAD) of Barrett's Neoplasia (MetaVision) [Internet]. 2023 [cited 2025 Mar 27]. Available from: https://clinicaltrials.gov/study/NCT05965921?locStr=Portsmouth,%20UK&country=United%20Kingdom&state=England&city=Portsmouth&cond=Real-life%20Prospective%20Evaluation%20of%20Computer-aided%20Detection%20(CAD)%20of%20Barrett%27s%20Neoplasia&rank=1
108. NHS Health Research Authority. Feasibility of CADU system and bio-impedence in Barrett's Oesophagus [Internet]. 2024 [updated 2025 Mar 27; cited 2025 Mar 27]. Available from: https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/feasibility-of-cadu-system-and-bio-impedence-in-barretts-oesophagus/
109. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, Liu P, Li L, Song Y, Di Zhang, Li Y, Xu G, Tu M, Liu X. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813–9. eng. doi:10.1136/gutjnl-2018-317500 Cited in: PubMed; PMID 30814121.
110. Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst. 2014;25(5):845–69. eng. doi:10.1109/TNNLS.2013.2292894 Cited in: PubMed; PMID 24808033.
111. Jong MR, Groof AJ de. Advancement of artificial intelligence systems for surveillance endoscopy of Barrett's esophagus. Dig Liver Dis. 2024;56(7):1126–30. eng. doi:10.1016/j.dld.2023.11.038 Cited in: PubMed; PMID 38071181.
112. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset Shift in Machine Learning: The MIT Press; 2008.
113. Kouw WM, Loog M. An introduction to domain adaptation and transfer learning; 31/12/2018.
114. Schmitz R, Werner R, Repici A, Bisschops R, Meining A, Zornow M, Messmann H, Hassan C, Sharma P, Rösch T. Artificial intelligence in GI endoscopy: stumbling blocks, gold standards and the role of endoscopy societies. Gut. 2022;71(3):451–4. eng. doi:10.1136/gutjnl-2020-323115 Cited in: PubMed; PMID 33479051.
115. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol. 2019;20(3):405–10. eng. doi:10.3348/kjr.2019.0025 Cited in: PubMed; PMID 30799571.
116. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, Vickers AJ, Ransohoff DF, Collins GS. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73. eng. doi:10.7326/M14-0698 Cited in: PubMed; PMID 25560730.

117. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130(6):515–24. eng. doi:10.7326/0003-4819-130-6-199903160-00016 Cited in: PubMed; PMID 10075620.

118. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338(may28 1):b605. eng. doi:10.1136/bmj.b605 Cited in: PubMed; PMID 19477892.

119. Park SH, Kressel HY. Connecting Technological Innovation in Artificial Intelligence to Real-world Medical Practice through Rigorous Clinical Validation: What Peer-reviewed Medical Journals Could Do. J Korean Med Sci. 2018;33(22):e152. eng. doi:10.3346/jkms.2018.33.e152 Cited in: PubMed; PMID 29805337.

120. Mori Y, Kudo S-E, Misawa M, Saito Y, Ikematsu H, Hotta K, Ohtsuka K, Urushibara F, Kataoka S, Ogawa Y, Maeda Y, Takeda K, Nakamura H, Ichimasa K, Kudo T, Hayashi T, Wakamura K, Ishida F, Inoue H, Itoh H, Oda M, Mori K. Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study. Ann Intern Med. 2018;169(6):357–66. eng. doi:10.7326/M18-0249 Cited in: PubMed; PMID 30105375.

121. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharm. 2016;13(5):1445–54. eng. doi:10.1021/acs.molpharmaceut.5b00982 Cited in: PubMed; PMID 27007977.

122. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net; 21.12.2014.

123. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks; 04.03.2017.

124. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. PMLR 70:3145-3153 [Internet]. Available from: https://arxiv.org/pdf/1704.02685

125. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps; 20.12.2013.

126. Souza LA de, Mendel R, Strasser S, Ebigbo A, Probst A, Messmann H, Papa JP, Palm C. Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box. Comput Biol Med. 2021;135104578. eng. doi:10.1016/j.compbiomed.2021.104578 Cited in: PubMed; PMID 34171639.

127. Sharma P, Savides TJ, Canto MI, Corley DA, Falk GW, Goldblum JR, Wang KK, Wallace MB, Wolfsen HC. The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on imaging in Barrett's Esophagus. Gastrointest Endosc. 2012;76(2):252–4. eng. doi:10.1016/j.gie.2012.05.007 Cited in: PubMed; PMID 22817781.

128. Chun HJ, Yang S-K, Choi M-G, editors. Clinical gastrointestinal endoscopy: A comprehensive atlas. 2nd ed. Singapore: Springer Singapore; 2018.

Danksagung

Zunächst gilt mein Dank Prof. Pech für die kontinuierliche Betreuung, Erreichbarkeit und Begleitung während der gesamten Arbeit. Ebenso danke ich dem Team der Endoskopie im Krankenhaus Barmherzige Brüder Regensburg, sowie Frau Bleicher für ihre hilfreiche organisatorische Unterstützung.

Kiki und Jelmer danke ich herzlichst für ihre großartige Hilfe, die ausführlichen Antworten auf meine Fragen und dafür, dass sie – trotz geografischer Distanz – stets erreichbar waren und mir den Einstieg in die Forschung maßgeblich erleichtert haben. Außerdem danke ich den Partnergruppen aus Amsterdam und Eindhoven für die Ermöglichung der Nutzung ihres CAD-Systems für diese Arbeit.

Für gemeinsames Co-Working, mentale Unterstützung und motivierenden Zuspruch möchte ich mich bei Moritz und Anneke bedanken. Johannes danke ich für sein offenes Ohr bei technischen Fragen.

Zu guter Letzt danke ich meiner Familie für ihren uneingeschränkten Rückhalt und das wertvolle Feedback.

Selbstständigkeitserklärung

Hiermit erkläre ich, Leona Brenner, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Person) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt. Ich erkläre hiermit ausdrücklich, dass KI-Unterstützung ausschließlich zur stilistischen Überarbeitung, Verbesserung der sprachlichen Lesbarkeit sowie zum Korrekturlesen (Rechtschreibung und Grammatik) eingesetzt wurden. Alle wissenschaftlichen Inhalte, Studiendesigns, Datenerhebungen, Auswertungen, Interpretationen und Schlussfolgerungen sowie die Literaturauswahl stammen uneingeschränkt aus meiner wissenschaftlichen Leistung.

_____ _____

Ort, Datum Unterschrift