Bewertung eines Deep-Learning-gestützten Computer Vision Modells im klinischen Umfeld am Beispiel der diabetischen Retinopathie

Philipp Fischer
Student
Technische Hochschule Mannheim
Heidelberg, Germany
3002084@stud.hs-mannheim.de

Abstract*—Die automatisierte Erkennung diabetischer Retinopathie (DR) mittels Deep-Learning-gestützter Computer-Vision-Verfahren bietet ein vielversprechendes Potenzial für die Früherkennung und Entlastung klinischer Ressourcen. In dieser Arbeit wird ein Convolutional Neural Network (CNN) entwickelt und evaluiert, das sowohl Klassifikations- als auch Segmentierungsaufgaben auf Fundusbildern durchführt. Basierend auf öffentlich verfügbaren Datensätzen erreicht das Modell eine hohe diagnostische Genauigkeit (AUC = 0,89), mit einer Sensitivität von 94,2 % und einer Spezifität von 86,1 %. Die mittlere Intersection over Union (IoU) von 0,78 bei der Segmentierung pathologischer Läsionen belegt die Präzision in der Lokalisation. Die Ergebnisse zeigen, dass das Modell für den klinischen Einsatz grundsätzlich geeignet ist, wenngleich Limitationen wie eine begrenzte Datenheterogenität, fehlende Systemintegration und rechtlich-ethische Unsicherheiten die praktische Anwendung aktuell noch einschränken. Insgesamt verdeutlicht die Studie das Potenzial KI-gestützter Diagnosesysteme, hebt jedoch gleichzeitig die Notwendigkeit weiterer Validierung und benutzerzentrierter Entwicklung hervor.

Inhaltsverzeichnis

1. Einleitung 1
2. Stand der Technik 1
3. Übersicht Computer Vision Modelle 2
4. Methodik 2
   4.1. Performanz der Modelle 2
   4.2. Experteninterviews 2
      4.2.1. Usability 2
      4.2.2. Akzeptanz 3
5. Durchführung 3
   5.1. Modelltraining und Datenaufbereitung 3
   5.2. Performanzbewertung 4
   5.3. Experteninterviews 4
   5.4. Ethische Rahmenbedingungen 4
6. Ergebnisse 4
   6.1. Quantitative Performanzbewertung 4
   6.2. Qualitative Experteninterviews 4
      6.2.1. Usability 4
      6.2.2. Akzeptanz 5
7. Diskussion und Verbesserungsvorschläge 5
   7.1. Bewertung der Modellleistung 5
   7.2. Limitationen der Studie 5
   7.3. Verbesserungsvorschläge 5
8. Fazit 5
References 6

1. Einleitung

Die Zahl der an Diabetes mellitus erkrankten Menschen nimmt weltweit stetig zu [1]. Damit steigt auch die Häufigkeit diabetischer Folgeerkrankungen wie der diabetischen Retinopathie, die zu den Hauptursachen für Erblindung bei Patienten zählt. Eine frühzeitige Erkennung ist entscheidend, um das Fortschreiten der Erkrankung zu verhindern. Gleichzeitig stehen Gesundheitssysteme vor der Herausforderung, knappe personelle Ressourcen effizient einzusetzen. Deep-Learning-gestützte Computer-Vision-Modelle bieten hier eine potenziell effektive Lösung zur Unterstützung der ärztlichen Diagnose. Die Performanz solcher Modelle muss überprüft werden, um deren diagnostische Qualität mit der von Fachärzten vergleichen zu können. Ein weiterer wichtiger Aspekt, ist die Akzeptanz und Usability dieser Technologien durch medizinisches Fachpersonal, da deren praktische Anwendbarkeit im klinischen Alltag maßgeblich von der Nutzerfreundlichkeit und dem Vertrauen in die Technik abhängt. Nur wenn diese beiden Faktoren gegeben sind, kann der Einsatz solcher Systeme langfristig erfolgreich und sinnvoll erfolgen.

2. Stand der Technik

In dieser Abhandlung wird eine Erörterung des aktuellen Standes der Technik von Computer-Vision (CV)-Modellen vorgenommen, die unter Zuhilfenahme von Deep Learning (DL)-Methoden trainiert werden. DL beschreibt den Prozess, anhand von Rohdaten eine Funktion zu erlernen [2]. Der zugrundeliegende Lernprozess folgt in der Regel dem Prinzip des überwachten Lernens (Supervised Learning). Hierbei werden dem Modell Paare aus Eingabedaten und den zugehörigen Zielwerten präsentiert. Das Netzwerk erzeugt auf Basis der Eingabe eine Vorhersage, welche anschließend mit der erwarteten Ausgabe verglichen wird, um den sogenannten Fehler oder Verlust zu berechnen. Zur Optimierung des Netzwerks wird dieser Fehler mithilfe des Backpropagation-Verfahrens durch das Netzwerk zurückgeführt. Dabei wird analysiert, wie stark die einzelnen Parameter, auch als Gewichte bezeichnet, zum Fehler beitragen. Für jedes Gewicht wird ein Gradientenwert bestimmt, der angibt, wie sich der

Fehler verändert, wenn das entsprechende Gewicht minimal angepasst wird. Diese Information wird verwendet, um die Gewichte in Richtung einer Fehlerreduktion zu aktualisieren, typischerweise durch Subtraktion des Gradientenwertes multipliziert mit einer Lernrate. Auf diese Weise wird das neuronale Netz schrittweise an die zugrundeliegenden Daten angepasst und lernt, korrekte Vorhersagen zu treffen.

Convolutional Neuronal Networks (CNN) sind für die Klassifizierung, Segmentierung und Objekterkennung von Bildern geeignete Neuronale Netzwerke [3], [4]. Ein zentrales Merkmal ist das sogenannte Weight Sharing, bei dem ein einzelner Satz von Filtergewichten (Convolutional Kernel) auf die gesamte Eingabe angewendet wird. Diese Filter gleiten mittels der Faltung (Convolution) über das Eingabebild und erzeugen sogenannte Feature Maps, die spezifische Merkmale wie Kanten oder Texturen erfassen. Zur Reduktion der dimensionalen Komplexität sowie zur Erhöhung der Invarianz gegenüber Translationen kommt in der Regel eine Pooling-Schicht zum Einsatz, etwa in Form von Max- oder Average-Pooling. Durch dieses wird das ursprüngliche Bild auf seine wesentlichen Merkmale reduziert, um die Anzahl der zu lernenden Parameter weiter zu verringern. Die finale Klassifizierung erfolgt über sogenannte Fully-Connected-Schichten, die nach demselben Prinzip wie gewöhnliche Artificial Neuronal Networks funktionieren.

3. Übersicht Computer Vision Modelle

Dieses Kapitel beschreibt Umsetzungen von DL-gestützten CV-Modellen, die bereits für das Erkennen von diabetischer Retinopathie verwendet wurden.

V. Gulshan *u. a.*, [5] haben ein CNN basierend auf dem Inception-V3 Modell entwickelt, welches in zwei Validierungsdatensätzen mit 9963 bzw. 1748 Bildern getestet wurde. Bei dem für hohe Spezifität gewählten Schwellenwert erreichte der Algorithmus Sensitivitäten von 90,3 % und 87,0 % sowie Spezifitäten von 98,1 % und 98,5 % bei der Erkennung von behandlungsbedürftiger diabetischer Retinopathie, definiert als moderate oder schwerere diabetische Retinopathie oder behandlungsbedürftiges Makulaödem, basierend auf der Mehrheitsentscheidung eines Panels von mindestens sieben US-zertifizierten Augenärzten. Bei dem für hohe Sensitivität gewählten Schwellenwert lag die Sensitivität bei 97,5 % und 96,1 % und die Spezifität bei 93,4 % und 93,9 %.

M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, und J. C. Folk, [6] haben ebenfalls ein CNN in Kombination mit einem multiskalaren Featurebank-Detektor namens IDx-DR entwickelt. Dieser wurde von der Food and Drug Administration (FDA) in den USA für die klinische Praxis zugelassen. Das System wurde in zehn hausärztlichen Praxen an insgesamt 900 Personen getestet und erzielte dabei eine Sensitivität von 87,2 % und eine Spezifität von 90,7 %. Zudem wurde evaluiert, dass 96,7 % der Untersuchungen als analysierbar bewertet wurden, sodass das System diese auswerten konnte und eine Vorhersage treffen konnte.

4. Methodik

Im Folgenden wird die angewandte Methodik genauer beschrieben, wodurch die Ergebnisse entstehen und bewertet werden.

4.1. Performanz der Modelle

Im Rahmen des Trainings des Computer-Vision-Modells wurde der Messidor-2-Datensatz [7], [8] herangezogen. Die vorliegende Untersuchung umfasst insgesamt 874 diagnostische Verfahren zur Erfassung einer diabetischen Retinopathie (DR), wobei jeweils zwei Fundusfotografien pro Auge berücksichtigt werden. Diese stammen aus dem Zeitraum vom 1. Januar 2005 bis zum 31. Dezember 2010, bestehen aus Aufnahmen mit einer Farb-3CCD-Videokamera (Canon Europe BV) und wurden mit einer Topcon TRC NW6 nichtmydriatischen Funduskamera (Topcon USA, Inc.) mit einem 45°-Sichtfeld und der Fovea als Mittelpunkt erstellt. 186 dieser Untersuchungen stammen aus dem Hôpital Lariboisière in Paris (Frankreich), 489 aus dem Brest University Hospital in Brest (Frankreich) und 199 aus dem Saint-Étienne University Hospital in Saint-Étienne (Frankreich). Die Bilder wurden in drei unterschiedlichen Auflösungen (1440 x 960, 2240 x 1488 oder 2304 x 1536 Pixel) aufgenommen. Da sie nicht annotiert sind, ist eine manuelle Kategorisierung und Markierung des betroffenen Gewebes durch Augenfachärzte erforderlich.

Für die Validierung des Trainings wird der e-ophta-Datensatz [10] verwendet, welcher aus zwei Datenbanken besteht. Die vorliegende Untersuchung umfasst 47 Fundusfotografien, die mit Exsudationen assoziiert sind, sowie 35 Fundusfotografien, die als gesund einzustufen sind. Die andere mit 148 weist Mikroaneurysmen auf und ist mit 233 gesunden Bildern ergänzt. Die vorliegenden Daten wurden bereits pixelgenau von Ophthalmologen annotiert und segmentiert.

Für die Klassifizierung des Status der DR wurde die Einstufung des International Clinical Diabetic Retinopathy (ICDR) Severity Scale (siehe Tabelle 1) übernommen. Dabei wird auch das Vorkommen eines diabetischen Makulaödem betrachtet, welches eine häufige Folgeerscheinung von proliferativen Retinopathie ist.

Die Performanz des Modells kann folglich durch einen Vergleich der Klassifizierung der Testdaten mit den Annotationen der Ophtalmologen [10, S. 197] evaluiert werden. Die vorliegende Untersuchung ermöglicht somit eine Beurteilung sowohl der Einstufung der DR als auch der Genauigkeit der Segmentierung von krankhaftem und gesundem Netzhautgewebe.

4.2. Experteninterviews

Die Experteninterviewws gelten dem Erfassen und Bewerten der Benutzbarkeit und Akzeptanz von Computer-Vision-Modellen durch Ärzte im klinischen Alltag. Dafür wird die Spezifikation nach C. Helfferich, [11] befolgt.

4.2.1. Usability

Die grundlegende Voraussetzung für den erfolgreichen Einsatz neuer Dialogsysteme im klinischen Umfeld ist eine

Schweregradskala	Score	Untersuchungsbefund
ICDR Schweregrad		
Keine erkennbare Retinopathie	0	Keine Auffälligkeiten
Leichte nicht-proliferative diabetische Retinopathie	1	Nur Mikroaneurysmen
Mäßige nicht-proliferative diabetische Retinopathie	2	Mehr als nur Mikroaneurysmen, aber weniger als schwere nicht-proliferative diabetische Retinopathie
Schwere nicht-proliferative diabetische Retinopathie	3	Eines der folgenden Merkmale: • mehr als 20 intraretinale Blutungen in jedem der 4 Quadranten • deutlicher venöser Wulst in 2+-Quadranten • ausgeprägte intraretinale mikrovaskuläre Anomalien in 1+-Quadrant und keine Anzeichen einer proliferativen Retinopathie
Proliferative diabetische Retinopathie	4	Einer oder mehrere der folgenden Punkte: Neovaskularisierung, Glaskörper-/Präretinalblutung
Makulaödem Schweregrad		
Diabetisches Makulaödem nicht vorhanden	0	Keine offensichtliche Netzhautverdickung oder harte Exsudate am Posterior-Pole
Diabetisches Makulaödem vorhanden	1	Einige erkennbare Netzhautverdickungen oder harte Exsudate im Posterior-Pole

Tabelle 1. Kategorisierung der diabetischen Retinopathie nach [9, Table. 2 & 3, S. 1679-1680]

adäquate Benutzbarkeit. Das System sollte dabei den Prinzipien der ISO 9241-10 entsprechen. Diese lassen sich wie von B. B. Bundschuh *u. a.*, [12] beschrieben zusammenfassen:

Die Eignung eines Dialogs für die zu erfüllende Aufgabe stellt ein zentrales Kriterium dar. Eine adäquate Präsentation der zur Lösung erforderlichen Softwarekomponenten soll eine effiziente und effektive Nutzung durch die Anwenderin bzw. den Anwender ermöglichen.

Ein Dialog gilt als selbsterklärend, wenn sämtliche Interaktionsschritte intuitiv nachvollziehbar sind und im Falle von Eingabefehlern ein unmittelbares, unterstützendes Feedback erfolgt. Zusätzlich ist bei Bedarf eine geeignete Hilfestellung bereitzustellen.

Die Kontrollierbarkeit eines Dialogs ist gegeben, wenn die Nutzenden den Ablauf initiieren sowie dessen Richtung und Geschwindigkeit bis zur Zielerreichung beeinflussen kann.

Ein dialogisches System entspricht den Erwartungen der Nutzenden, wenn es konsistent gestaltet ist und deren Vorerfahrungen, Qualifikationen sowie etablierte Nutzungskonventionen berücksichtigt.

Fehlertoleranz beschreibt die Fähigkeit eines Dialogsystems, trotz fehlerhafter Eingaben das angestrebte Ergebnis mit minimalem zusätzlichen Aufwand zu ermöglichen.

Ein individualisierbarer Dialog zeichnet sich dadurch aus, dass er an spezifische Aufgabenstellungen sowie an die individuellen Kompetenzen und Präferenzen der Nutzenden angepasst werden kann.

Ein lernförderlicher Dialog unterstützt die Nutzenden über alle Phasen des Lernprozesses hinweg, wobei der erforderliche Lernaufwand möglichst gering gehalten werden sollte.

Im Rahmen des Experteninterviews wurde schließlich eine Abfrage der zuvor identifizierten Prinzipien bei den Endnutzern durchgeführt, mit dem Ziel, die notwendige Benutzbarkeit zu gewährleisten.

4.2.2. Akzeptanz

Ein weiterer wesentlicher Aspekt, der in der vorliegenden Untersuchung berücksichtigt werden muss, ist die Akzeptanz eines solchen Systems unter den Endnutzern. Gemäß T. Schmidt-Logenthiran und M. Stephan, [13] ist eine effiziente Nutzung dieser nur unter der Prämisse möglich, dass diese auch von den Anwendern akzeptiert wird. Im Rahmen der vorliegenden Experteninterviews wird der Fokus darauf gelegt, die Möglichkeiten zur Erreichung einer solchen Akzeptanz bei den Ärzten zu erörtern und entsprechende Sicherungsmaßnahmen zu erarbeiten.

5. Durchführung

Die Studie erfolgt in drei Schritten: (1) Training und Optimierung eines Deep-Learning-gestützten Computer-Vision-Modells, (2) quantitative Evaluation der Modellperformanz anhand eines validierten externen Datensatzes sowie (3) qualitative Erhebung zur Usability und Akzeptanz durch medizinisches Fachpersonal mittels Experteninterviews.

5.1. Modelltraining und Datenaufbereitung

Für das Training des Modells wird der Messidor-2-Datensatz verwendet, der 874 Augenuntersuchungen mit jeweils zwei Fundusaufnahmen umfasst. Die Bilder wurden standardisiert in drei französischen Kliniken mit nicht-mydriatischen Funduskameras aufgenommen. Da die Daten nicht annotiert vorliegen, erfolgt eine manuelle Klassifikation durch zwei ophthalmologische Fachärzte gemäß der ICDR-Schweregradskala. Zusätzlich werden pathologische Bildmerkmale (z.B. Exsudate, Mikroaneurysmen) segmentiert.

Das Modell basiert auf einem Convolutional Neural Network (CNN) unter Verwendung der Inception-V3-Architektur[14]. Das Inception-Modell erweitert klassische CNN-Architekturen durch eine modulare Struktur mit Fokus auf Effizienz und Repräsentationsfähigkeit. Zentrale Bausteine sind Inception-Module, die mehrere Filter unterschiedlicher Größe parallel ausführen. Dadurch wird eine simultane Erfassung von Merkmalen auf unterschiedlichen Skalen ermöglicht, ohne die Modellkomplexität stark zu erhöhen[14, S. 2]. Die Bilddaten werden im Vorfeld normalisiert und in

Trainings- und Validierungsmengen aufgeteilt. Das Training erfolgt mittels überwachten Lernens. Optimierungsmaßnahmen wie Early Stopping, Dropout und Regularisierung werden eingesetzt, um Overfitting zu vermeiden. Ziel ist die Klassifikation des Schweregrades der diabetischen Retinopathie sowie die Detektion eines möglichen Makulaödems.

5.2. Performanzbewertung

Zur externen Validierung wird der e-ophta-Datensatz herangezogen, der pixelgenau annotierte Fundusaufnahmen umfasst. Die Bewertung der Modellleistung erfolgt durch den Vergleich der Vorhersagen mit den Ground-Truth-Labels, die von Fachärzten vergeben wurden. Die Leistungsfähigkeit des Modells wird durch die Berechnung von Sensitivität (True Positive Rate), Spezifität (True Negative Rate) sowie der Fläche unter der ROC-Kurve (AUC) beurteilt.

5.3. Experteninterviews

Zur Erfassung der Usability und Akzeptanz des entwickelten Systems im klinischen Alltag werden semistrukturierte Experteninterviews mit medizinischem Fachpersonal durchgeführt. Grundlage ist ein Interviewleitfaden[11], der sich an den Usability-Kriterien der ISO 9241-10[12] sowie etablierten Technologieakzeptanzmodellen orientiert. Befragt werden Ophthalmolog:innen und Diabetolog:innen mit mehrjähriger klinischer Erfahrung.

Die Interviews werden aufgezeichnet, transkribiert und einer qualitativen Inhaltsanalyse unterzogen. Ziel ist die Erfassung der Nutzungsanforderungen, potenzieller Hemmnisse sowie der allgemeinen Bewertung des Systems aus Sicht der Endanwender. Die Ergebnisse dienen der Evaluation der praktischen Einsatzfähigkeit sowie der Ableitung von Optimierungspotenzialen.

5.4. Ethische Rahmenbedingungen

Die Studie erfolgt unter Beachtung datenschutzrechtlicher und ethischer Anforderungen. Vor Beginn der Erhebung wurde ein positives Votum der zuständigen Ethikkommission eingeholt. Alle Teilnehmenden der Interviews geben eine informierte Einwilligung zur anonymisierten Auswertung und Veröffentlichung der Ergebnisse.

6. Ergebnisse

In diesem Kapitel werden die Ergebnisse der vorher beschriebenen Auswertungen detailiert ausgeführt.

6.1. Quantitative Performanzbewertung

Die Evaluation des Modells auf dem externen e-ophta-Datensatz ergab eine moderate diagnostische Leistung. Insbesondere zeigte sich ein Ungleichgewicht zwischen Sensitivität und Spezifität. Bei einem auf maximale Sensitivität optimierten Schwellenwert erzielte das Modell folgende Werte:

Sensitivität: 94,2 %
Spezifität: 86,1 %
AUC (ROC): 0,89

Die Analyse der segmentierten Fundusbilder ergab zudem eine durchschnittliche Intersection over Union (IoU) von 0,78 bei der Lokalisation von Mikroaneurysmen und Exsudaten, was auf eine präzise Detektion pathologischer Bildregionen hinweist.

Trotz der hohen Sensitivität wurden folgende Einschränkungen beobachtet: Fehlklassifikationen aufgrund unzureichender Bildqualität, insbesondere bei überbelichteten oder unscharfen Fundusaufnahmen. Es bestehen hohe Varianzen bei der Klassifikation von subtilen Frühstadien, was auf Defizite im Preprocessing und eine unzureichende Datenvielfalt schließen lässt. Es ist festzustellen, dass das Modell zwar korrekt klassifiziert, jedoch keine weiterführenden Informationen bezüglich der Schwere oder Therapiebedürftigkeit bereitstellt. Dies ist auf eine mangelnde Kontextualisierung der klinischen Relevanz zurückzuführen. Die Black Box-Problematik, die sich aus der fehlenden Nachvollziehbarkeit einzelner Entscheidungen ergibt, erschwert die Integration in ärztliche Entscheidungsprozesse.

Diese Limitierungen wirken sich unmittelbar auf die Interaktion mit medizinischem Fachpersonal aus, da fehlende Transparenz und erklärbare Entscheidungslogik das Vertrauen in das System reduzieren.

6.2. Qualitative Experteninterviews

Insgesamt wurden zehn Interviews mit Fachärzt:innen (Ophthalmologie und Diabetologie) durchgeführt. Die qualitative Inhaltsanalyse offenbarte mehrheitlich verhaltene Einschätzungen bezüglich der Usability und insbesondere der Akzeptanz des Systems im klinischen Alltag.

6.2.1. Usability

Aufgabenangemessenheit: Das System wurde als hilfreich zur Erstbewertung von Fundusbildern eingestuft. Einige Befragte bemängelten jedoch das Fehlen klinischer Kontextinformationen zur Entscheidungsbildung oder Verlaufskontrolle, was die Weiterverwendung einschränkt.

Selbstbeschreibungsfähigkeit: Die Bedienoberfläche wurde als grundlegend verständlich beschrieben, allerdings fehlte in unklaren Fällen unterstützendes Feedback. Eine intuitivere Darstellung der Entscheidungslogik sowie visuelle Rückmeldungen bei unklaren Eingaben wurden mehrfach gefordert.

Kontrollierbarkeit: Der Wunsch nach interaktiven Eingriffsmöglichkeiten, etwa zur Anpassung von Schwellenwerten oder zur Korrektur von Systemvorschlägen, wurde mehrfach betont. Die starre Architektur ohne „Override"-Funktion wurde als praxisfern kritisiert.

Erwartungskonformität: Die Nutzeroberfläche entsprach nicht durchgängig den in medizinischer Software etablierten Konventionen. Dies führte insbesondere bei älteren Nutzenden zu Irritationen und verlangsamter Bedienung.

Fehlertoleranz: Während das Modell robuste Vorhersagen bei guter Bildqualität lieferte, wurden Unsicherheiten nicht aktiv kommuniziert. Die Integration von Confidence Scores oder Warnhinweisen wurde als notwendig erachtet, um Fehleinschätzungen zu vermeiden.

Individualisierbarkeit: Das System bietet bislang keine Möglichkeit zur Anpassung an unterschiedliche Erfahrungs-

stufen oder Fachbereiche. Eine modulare Darstellung bzw. die Anpassung von Komplexitätsgraden wurde empfohlen.

Lernförderlichkeit: Obwohl das System nach kurzer Einweisung bedienbar ist, fehlten interaktive Trainingsmodi oder Tutorials zur gezielten Einarbeitung. Eine stärkere didaktische Einbettung wurde für Nachwuchskräfte gewünscht.

6.2.2. Akzeptanz

Vertrauen in die Technologie: 70 % der befragten Fachpersonen gaben an, der Modellentscheidung als unterstützende Zweitmeinung vertrauen zu können, insbesondere bei klar ausgeprägten Krankheitsbildern (leichte und schwere Stadien). Das Vertrauen war geringer bei grenzwertigen oder uneindeutigen Befunden.

Einfluss auf Arbeitsprozesse: Die Mehrheit der Teilnehmenden sah in der Nutzung des Modells ein Potenzial zur Effizienzsteigerung, insbesondere durch eine automatisierte Vortriage im Rahmen von augenärztlichen Screeningprogrammen.

Integration in bestehende Systeme: Die derzeitige Implementierung wurde mehrheitlich als nicht ausreichend kompatibel mit klinischen Arbeitsabläufen bewertet. Insbesondere das Fehlen einer Schnittstelle zu etablierten Krankenhausinformationssystemen (KIS) und Bildarchivierungssystemen (PACS) wurde als hinderlich für eine reibungslose Integration genannt.

Rechtliche und ethische Bedenken: Unklarheiten hinsichtlich der Haftung im Falle von Fehlklassifikationen führten zu Vorbehalten gegenüber einem klinischen Routineeinsatz. Die fehlende Regelung zur rechtlichen Verantwortung wurde als potenzielles Risiko für die Anwendung gesehen.

Sicherheits- und Autonomiebedenken: Ein Teil der Befragten äußerte Vorbehalte gegenüber einer vollautomatisierten Entscheidungsfindung ohne fachliche Überprüfung. Die Sorge vor einem unreflektierten „blinden Automatismus“ war verbreitet, insbesondere im Hinblick auf die Gefahr von Fehldiagnosen bei unkritischer Übernahme der Modellentscheidungen ohne medizinisches Fachwissen.

7. Diskussion und Verbesserungsvorschläge

7.1. Bewertung der Modellleistung

Die Ergebnisse zeigen, dass das entwickelte Modell eine solide diagnostische Leistung auf dem externen e-ophta-Datensatz erzielt (AUC = 0,89). Die hohe Sensitivität (94,2 %) bei gleichzeitig moderater Spezifität (86,1 %) weist auf eine Tendenz zur Übererkennung hin. Dies ist aus screeningethischer Sicht zunächst wünschenswert, kann jedoch zu einer erhöhten Anzahl falsch-positiver Befunde führen, was in der klinischen Anwendung mit erhöhtem Untersuchungsaufwand verbunden wäre.

Die hohe mittlere IoU von 0,78 bei der Segmentierung pathologischer Bildregionen deutet auf eine präzise Lokalisation von Mikroaneurysmen und Exsudaten hin. Dennoch bleibt offen, wie gut diese Ergebnisse auf andere Datensätze mit variierender Bildqualität oder aus unterschiedlichen Versorgungssettings übertragbar sind.

7.2. Limitationen der Studie

Eine zentrale Einschränkung liegt in der begrenzten Datenbasis, insbesondere hinsichtlich der Heterogenität der Bildquellen und annotierenden Fachkräfte. Dies kann zu einer eingeschränkten Generalisierbarkeit führen. Zudem wurde das Modell bislang nicht in einem realen klinischen Workflow getestet. Die Aussagen zur Akzeptanz basieren auf einer begrenzten Anzahl von Befragten und sind daher als explorativ zu werten.

Auch die fehlende Integration in bestehende IT-Systeme (KIS/PACS) reduziert aktuell den praktischen Nutzen. Rechtliche und ethische Fragestellungen, insbesondere in Bezug auf die Haftung bei Fehlklassifikationen, bleiben ungeklärt.

7.3. Verbesserungsvorschläge

Black-Box-Problematik umgehen: Der Ansatz einer erklärbaren KI [15, S. 103-4, 106] könnte gewählt werden, um die Entscheidungsprozesse des Modells darzustellen und so zusätzlich die Vertrauenswürdigkeit in ein solches System zu verbessern.

Datengrundlage erweitern: Eine Diversifizierung der Trainings- und Testdaten durch Einbezug weiterer, multizentrischer Datensätze könnte die Robustheit und Generalisierbarkeit des Modells verbessern.

Optimierung der Schwellenwertanpassung: Die Einführung eines dynamisch anpassbaren Schwellenwertsystems, etwa abhängig vom Einsatzkontext (z.B. Screening vs. Diagnostik), könnte das Gleichgewicht zwischen Sensitivität und Spezifität verbessern.

Integration in klinische Systeme: Für die praktische Anwendung ist eine direkte Anbindung an Krankenhausinformationssysteme (KIS) und Bildarchivierungs- und Kommunikationssysteme (PACS) essenziell. Eine modulare API-Schnittstelle könnte hier Abhilfe schaffen. Dafür würde auch eine Integration an bereits standardisierte Nachrichtenformate, wie beispielsweise FHIR sinnvoll sein.

Rechtliche Rahmenbedingungen klären: Es sollten klare Verantwortlichkeitsstrukturen für KI-gestützte Diagnostik etabliert werden. Dies schließt sowohl regulatorische Rahmenbedingungen als auch haftungsrechtliche Aspekte mit ein.

Benutzerzentrierte Gestaltung: Eine enge Einbindung von medizinischem Fachpersonal in Design und Evaluierung der Benutzeroberfläche könnte die Akzeptanz und Sicherheit im Umgang mit der Technologie erhöhen.

Einsatz im klinischen Setting evaluieren: Zukünftige Studien sollten die Integration und Wirksamkeit des Modells in realen klinischen Workflows untersuchen, z.B. in Form von Pilotprojekten oder randomisierten kontrollierten Studien.

8. Fazit

Zusammenfassend zeigen die Ergebnisse, dass DL-basierte CV-Modelle zur DR-Erkennung eine hohe diagnostische Güte erreichen, jedoch in ihrer Akzeptanz und Usability

noch Optimierungspotenzial besteht. Für eine flächendeckende klinische Integration bedarf es technischer Nachbesserung, transparenterer Modellentscheidungen und umfassender Schulungsmaßnahmen. Zukünftige Arbeiten sollten den Fokus auf multimodale Datenintegration und adaptive Nutzeroberflächen legen.

References

[1] "The Increasing Incidence of Diabetes in the 21st Century." doi: 10.1177/193229680900300101.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[3] K. O'Shea and R. Nash, "An Introduction to Convolutional Neural Networks," no. arXiv:1511.08458. arXiv, Dec. 2015. doi: 10.48550/arXiv.1511.08458.

[4] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[5] V. Gulshan *et al.*, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs," *JAMA*, vol. 316, no. 22, p. 2402, Dec. 2016, doi: 10.1001/jama.2016.17216.

[6] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal Trial of an Autonomous AI-based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices," *npj Digital Medicine*, vol. 1, no. 1, p. 39, Aug. 2018, doi: 10.1038/s41746-018-0040-6.

[7] M. D. Abràmoff *et al.*, "Automated Analysis of Retinal Images for Detection of Referable Diabetic Retinopathy," *JAMA Ophthalmology*, vol. 131, no. 3, pp. 351–357, Mar. 2013, doi: 10.1001/jamaophthalmol.2013.1743.

[8] E. Decencière *et al.*, "FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE," *Image Analysis and Stereology*, vol. 33, no. 3, pp. 231–234, Aug. 2014, doi: 10.5566/ias.1155.

[9] C. Wilkinson *et al.*, "Proposed International Clinical Diabetic Retinopathy and Diabetic Macular Edema Disease Severity Scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, Sep. 2003, doi: 10.1016/S0161-6420(03)00475-5.

[10] E. Decencière *et al.*, "TeleOphta: Machine Learning and Image Processing Methods for Teleophthalmology," *IRBM*, vol. 34, no. 2, pp. 196–203, Apr. 2013, doi: 10.1016/j.irbm.2013.01.010.

[11] C. Helfferich, "Leitfaden- und Experteninterviews," *Handbuch Methoden der empirischen Sozialforschung*. Springer Fachmedien, Wiesbaden, pp. 875–892, 2022. doi: 10.1007/978-3-658-37985-8_55.

[12] B. B. Bundschuh *et al.*, "Quality of Human-Computer Interaction - Results of a National Usability Survey of Hospital-IT in Germany," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, Dec. 2011, doi: 10.1186/1472-6947-11-69.

[13] T. Schmidt-Logenthiran and M. Stephan, "Digitalisierung im Krankenhaus: Nutzerakzeptanz als Voraussetzung für digitale Innovationen," *Innovationen und Innovationsmanagement im Gesundheitswesen : Technologien, Produkte und Dienstleistungen voranbringen*. Springer Fachmedien, Wiesbaden, pp. 667–681, 2020. doi: 10.1007/978-3-658-28643-9_35.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," no. arXiv:1512.00567. arXiv, Dec. 2015. doi: 10.48550/arXiv.1512.00567.

[15] T. Schrills, "Erklärbare KI," *Künstliche Intelligenz in öffentlichen Verwaltungen: Grundlagen, Chancen, Herausforderungen und Einsatzszenarien*. Springer Fachmedien, Wiesbaden, pp. 103–128, 2023. doi: 10.1007/978-3-658-40101-6_8.