Die unsichtbare Voreingenommenheit der KI-Prognose

- Eine eingehende Untersuchung am Beispiel von CORe. Auf dem Weg zu einer gerechteren Gesundheitsversorgung -

14. Juli 2024

Bachelorarbeit vorgelegt von:

Student:	Leif Bräg
Matrikelnummer:	851510
Studiengang:	Medieninformatik Online (Bachelor)
Fachbereich:	VI
Betreuer:	Prof. Dr. Alexander Löser
Gutachterin:	Dr. Selcan Ipek-Ugay

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

KI	Künstliche Intelligenz
AMA	American Medical Association
EHR	Electronic Health Record
GDPR	General Data Protection Regulation
MDPI	Molecular Diversity Preservation International
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
RELU	Rectified Linear Unit
NLP	Natural Language Processing
GLUE	General Language Understanding Evaluation
NSP	Next Sentence Prediction
MIMIC	Medical Information Mart for Intensive Care
MDR	Medical Device Regulation

1 Einleitung

Die Integration von KI[1] in medizinische Diagnosen verspricht, zukünftig die Effizienz und Genauigkeit der Patientenversorgung erheblich zu verbessern. Studien, wie die von Brinker et al. (2019), zeigen, dass KI-Systeme Hautkrebs mit einer Genauigkeit diagnostizieren können, die die von praktizierenden Dermatolog:innen sogar übertrifft. Trotz dieser vielversprechenden Aussichten birgt der Einsatz von KI auch ethische Herausforderungen, insbesondere durch potenzielle Voreingenommenheiten in den Modellen. Diese Arbeit widmet sich der Untersuchung dieser Problematik und der Ableitung ethischer Leitlinien, um Voreingenommenheit in KI-Modellen zu minimieren.

1.1 Einführung in die Integration von KI in medizinische Diagnosen

Die rasante Entwicklung und Anwendung von KI-Technologien hat bereits in vielen Bereichen, einschließlich und insbesondere der Medizin, signifikante Fortschritte ermöglicht. Die Bedeutung dieser Technologie für die Medizin liegt in ihrer Fähigkeit, durch die Analyse großer Datenmengen Muster zu erkennen, die für die menschliche Ärzteschaft schwer oder nur unter großem Aufwand zu identifizieren sind. Vor allem die Fähigkeiten, Daten parallel zueinander zu verarbeiten, sowie die hohe Genauigkeit und Konsistenz sind klare Vorteile von technischen Lösungen zur Datenverarbeitung im Vergleich mit ihrem menschlichen Gegenüber. Dies kann zur frühzeitigen Diagnose und Behandlung von Krankheiten beitragen und somit die Patientenversorgung und Vorbeugung andernfalls tödlicher Krankheitsverläufe erheblich verbessern.

Ein bedeutendes Einsatzgebiet der KI in der medizinischen Diagnostik ist die Bildverarbeitung. KI-Modelle, insbesondere solche, die auf Deep Learning basieren, haben sich bei der Interpretation medizinischer Bilder, wie Röntgenaufnahmen, CT-Scans und MRTs, als äußerst effektiv erwiesen. Sie können Anomalien erkennen, welche die menschliche Ärzteschaft möglicherweise übersieht, und so die Diagnosegenauigkeit signifikant erhöhen. So trat ein CNN[2] in der dermatoskopischen Melanom-Bildklassifizierung gegen Dermatolog:innen an und übertraf 136 von 157 von diesen. Das CNN zeigte im Vergleich zu den Dermatolog:innen eine höhere Sensitivität (Maß für die korrekte Erkennung tatsächlicher Melanome) und Spezifität (Maß für das korrekte Diagnostizieren von gutartigen Hautveränderungen) (vgl. Brinker et al., 2019).

[1] Künstliche Intelligenz
[2] Convolutional Neural Network

Ein weiteres wichtiges Anwendungsfeld ist die Analyse elektronischer Gesundheitsakten (EHRs[3]). KI kann verwendet werden, um Muster und Trends in den umfassenden Datensätzen dieser Akten zu identifizieren. Diese Analysen können zur Vorhersage von Krankheitsausbrüchen, zur Optimierung von Behandlungsplänen und zur Identifizierung von Risikopatienten beitragen (vgl. Kaylor, 2023).

Darüber hinaus spielen KI-Systeme eine entscheidende Rolle bei der Vorhersage von Krankheitsverläufen. Sie können darauf hinweisen, dass eine signifikante Zustandsverschlechterung anzunehmen ist, auf potenzielle Komplikationen hinweisen und basierend darauf eine rechtzeitige Intervention vorschlagen, um den Krankheitsverlauf und das Wohlsein der Patient:innen zu verbessern. Durch die Nutzung großer Datenmengen und fortschrittlicher Algorithmen können KI-Systeme somit die diagnostische Genauigkeit erhöhen und menschliche Fehler reduzieren, was besonders relevant in komplexen und datenintensiven Bereichen wie bspw. der Kardiologie ist (vgl. Gala et al., 2024).

Allerdings gibt es auch Herausforderungen und Risiken bei der Integration von KI in die medizinische Diagnostik. Die Effektivität von KI-Modellen hängt stark von der Qualität und Quantität der zugrundeliegenden Trainingsdaten ab. Unvollständige oder unausgewogene Datensätze können zu fehlerhaften Modellen führen. Dies wird in der Studie von Petersson et al. (2022) betont, welche die Herausforderungen der Implementierung von KI im Gesundheitswesen durch Interviews mit Führungskräften im Gesundheitssektor untersuchte. Die Ergebnisse zeigen, dass unzureichende Datenqualität und -quantität wesentliche Hindernisse darstellen.

Zusätzlich wirft der Einsatz von KI in der Medizin wichtige ethische Fragen auf, insbesondere in Bezug auf Datenschutz, Patientenautonomie und die Verantwortung bei Fehlentscheidungen. Die AMA[4] hebt in ihrem *Journal of Ethics* hervor, dass „Black-Box"-Algorithmen, deren Entscheidungsprozesse für Nutzende nicht nachvollziehbar sind, zu erheblichen ethischen und rechtlichen Problemen führen können. Diese Probleme umfassen die Arzt- und Produkthaftung, z.B. im Falle einer Fehldiagnose, sowie die Sicherstellung des Datenschutzes und der informierten Einwilligung der Patient:innen (vgl. Rigby, 2019).

Zudem erfordert die Integration von KI in die medizinische Praxis die Entwicklung neuer regulatorischer Rahmenbedingungen, um Sicherheit und Wirksamkeit zu gewährleisten. Dies schließt auch die Überwachung der Algorithmen und die Gewährleistung ihrer Trki-ethik-gdpransparenz ein. Ein umfassender Überblick über die ethischen Herausforderungen und die regulatorischen Anforderungen im Zusammenhang mit der Nutzung

[3]Electronic Health Records
[4]American Medical Association

von KI im Gesundheitswesen, insbesondere im Kontext der Europäischen Datenschutz-Grundverordnung (kurz GDPR[5]), wird in der Studie von MDPI[6] erörtert. Diese Studie hebt die Notwendigkeit hervor, klare Richtlinien und Standards für den Einsatz von KI im Gesundheitswesen zu entwickeln und umzusetzen (Amini et al., 2023).

Die Implementierung von KI in die medizinische Diagnostik bietet enormes Potenzial zur Verbesserung der Patientenversorgung. Allerdings müssen die beschriebenen Herausforderungen, wie in aktuellen Studien betont wird, angegangen werden, um die Vorteile dieser Technologie voll ausschöpfen zu können.

1.2 Relevanz der Voreingenommenheit in KI-Modellen

Obwohl KI-Systeme viele Vorteile bieten, sind sie, wie bereits erwähnt, nicht frei von Herausforderungen. Eine große Aufgabe, die es zu lösen gilt, ist die potenzielle Voreingenommenheit in KI-Modellen (vgl. Nazer et al., 2023). Voreingenommenheit kann zum Beispiel entstehen, wenn die Daten, auf denen das Modell trainiert wird, unvollständig oder unausgewogen sind, oder wenn bestimmte Bevölkerungsgruppen unterrepräsentiert sind. Diese Voreingenommenheit kann zu ungleichen Behandlungsergebnissen führen und die Qualität der medizinischen Versorgung erheblich beeinträchtigen. Es ist daher entscheidend, Maßnahmen zur Identifizierung und Minimierung von Voreingenommenheit zu ergreifen, um Fehlentscheidungen grundlegend vorzubeugen. Als besonders kritisch werden hierbei die geschlechts-, alters- und ethnizitätsspezifische Voreingenommenheit definiert, welche im Folgenden genauer spezifiziert werden.

Ein Beispiel für geschlechtsspezifische Voreingenommenheit ist die ungleiche Verteilung von Männern und Frauen in den Trainingsdaten eines Modells. Wenn ein KI-Modell hauptsächlich oder vollständig mit Daten eines Geschlechts trainiert wird, kann dies zu einer geringeren Genauigkeit bei der Diagnose des anderen Geschlechts führen.

Diese Problemstellung wurde in einer Studie von Chung et al. (2021) untersucht. Dabei wurden zwei Modelle zur Vorhersage der Schwere von COVID-19-Erkrankungen erstellt, die jeweils ausschließlich mit männlichen bzw. weiblichen Daten trainiert wurden. Als das Modell, das mit weiblichen Daten trainiert wurde, auf männliche Fälle angewendet wurde, zeigte sich eine deutliche Reduktion der Vorhersagegenauigkeit, und umgekehrt. Bemerkenswert ist, dass die Genauigkeit auch bei Anwendung auf das gleiche Geschlecht wie das der Trainingsdaten geringer war im Vergleich zu einem Modell ohne geschlechtergetrennte Trainingsdaten.

[5]General Data Protection Regulation
[6]Molecular Diversity Preservation International

Auch altersbezogene Voreingenommenheit kann schwerwiegende Folgen haben. Wenn die Trainingsdaten nicht genügend Patient:innen einer bestimmten Altersgruppe beinhalten, kann das Modell Schwierigkeiten haben, genaue Vorhersagen für diese Altersgruppe zu treffen. Ein Beispiel hierfür ist die Diagnose von Kinderkrankheiten, bei denen ein Mangel an Daten über junge Patient:innen zu einer geringeren Diagnosegenauigkeit führen kann (vgl. Muralidharan et al., 2023).

Ein weiterer Faktor ist die genetische Voreingenommenheit. Sie kann auftreten, wenn die genetische Vielfalt in den Trainingsdaten nicht ausreichend gegeben ist. Dies kann z.B. zu ungenauen Diagnosen für bestimmte ethnische Gruppen führen. Eine Untersuchung von Elmahdy und Sebro (2023) hat gezeigt, dass aktuelle KI-Forschung oft ethnische Unterschiede nicht ausreichend berücksichtigt. Dies führt zu Schwierigkeiten bei der korrekten Diagnosefindung und kann durch das Anwenden eines diverseren Validierungsdatensatzes verbessert werden.

Ein weiterer relevanter Aspekt der Voreingenommenheit ist die Datenqualität undvollständigkeit in den zugrundeliegenden Trainings-, Test und Validierungsdaten. Unvollständige oder fehlerhafte Daten können dazu führen, dass KI-Modelle falsche Schlussfolgerungen ziehen. Beispielsweise beeinträchtigen fehlende oder falsch eingetragene Gesundheitsdaten die Genauigkeit der Diagnosemodelle erheblich. Ein bekanntes Problem ist die unzureichende Erfassung von sozioökonomischen und umweltbezogenen Faktoren, die für eine umfassende Gesundheitsbewertung wichtig sind (vgl. Celi et al., 2022). Speziell für datenarme Einsatzgebiete, bspw. solche mit einer schwachen medizinischen Infrastruktur, kann dies problematisch sein.

Die ethischen Implikationen von Voreingenommenheit in KI-Modellen sind weitreichend. Ungleiche Behandlungsergebnisse aufgrund von Voreingenommenheit können das Vertrauen der Patient:innen in KI-gesteuerte Diagnosesysteme untergraben und ethische Bedenken hinsichtlich der Gerechtigkeit aufwerfen. Es ist daher entscheidend, dass Entwicklende und Forschende frühzeitig Maßnahmen zur Identifizierung und Minderung von Voreingenommenheit dieser Systeme ergreifen.

Die Integration von KI in die medizinische Praxis muss daher von Anfang an sorgfältig überwacht und reguliert werden, um sicherzustellen, dass die Modelle auch in Zukunft fair, gerecht und effektiv arbeiten. Regulierungsbehörden und Gesundheitsorganisationen müssen weitere Richtlinien und Standards entwickeln, um die Ethik und Fairness von KI-Modellen in der Medizin insbesondere für die zukünftige Nutzung zu gewährleisten. Auch die Entwickelnden dieser Systeme müssen einen größeren Fokus auf diesen Aspekt legen, um Ethik und Fairness bereits in den Entwicklungsprozess mit einzubeziehen. Dies umfasst

auch die Transparenz der Modelle und die Nachvollziehbarkeit ihrer Entscheidungen, um Vertrauen und Akzeptanz bei den Nutzenden zu fördern.

1.3 Zielsetzung der Arbeit und Erkenntnisinteresse

Das Hauptziel dieser Arbeit ist es, Voreingenommenheit in KI-basierten Prognoseanwendungen zu identifizieren, wobei das CORe-Modell (Kapitel 2.4) als Fallstudie dient. Durch die Analyse spezifischer Testfälle wird untersucht, wie geschlechtsspezifische, altersbezogene und ethnische Voreingenommenheit die prognostizierten Diagnosen beeinflusst. Ein besonderer Fokus liegt dabei auf der US-amerikanischen Bevölkerung, um eine Vergleichbarkeit mit der Hauptdatenquelle, der MIMIC-III-Datenbank (Kapitel 2.3), zu gewährleisten. Auf Basis der Analyseergebnisse werden Empfehlungen zur Minimierung dieser Voreingenommenheit abgeleitet.

Diese Arbeit verfolgt folgende spezifische Unterziele: Erstens, die Sensibilität des CORe-Modells gegenüber geschlechtsspezifischen, altersbezogenen und genetischen Merkmalen zu untersuchen; zweitens, Unterschiede in den diagnostischen Ergebnissen zwischen verschiedenen Bevölkerungsgruppen hervorzuheben und deren Auswirkungen auf die Patientenversorgung zu evaluieren; und drittens, ethische Leitlinien und Empfehlungen abzuleiten, um Voreingenommenheit in zukünftigen Anwendungen von KI in der Medizin zu minimieren.

1.4 Zusammenfassung Kapitel 1

Die Integration von KI in medizinische Diagnosen hat das Potenzial, die Effizienz und Genauigkeit der Patientenversorgung erheblich zu verbessern. Trotz der in diesem Kapitel beschriebenen vielversprechenden Aussichten birgt der Einsatz von KI auch Herausforderungen, insbesondere im Hinblick auf ethische Fragen und mögliche Voreingenommenheiten in den zugrunde liegenden Modellen. Diese Arbeit widmet sich der Untersuchung dieser Problematik und der Ableitung ethischer Leitlinien, um Voreingenommenheit in KI-Modellen zu minimieren. Das Hauptziel ist es, Voreingenommenheit in KI-basierten Anwendungen, die für die Prognose genutzt werden, am Fallbeispiel CORe zu untersuchen und zukünftig zu minimieren. Um vergleichbare Ergebnisse zu erzielen, liegt der Fokus dieser Untersuchung auf den Vereinigten Staaten von Amerika.

2 Grundlagen

Um die Untersuchung der Voreingenommenheit des CORe-Modells und die Entwicklung ethischer Leitlinien fundiert durchführen zu können, ist es notwendig, ein solides Verständnis der zugrundeliegenden Technologien und Datenquellen zu haben. Dieses Kapitel liefert einen umfassenden Überblick über relevante Aspekte der Künstlichen Intelligenz in der medizinischen Diagnostik, erläutert das klinische Ergebnismodell CORe und stellt die MIMIC III-Datenbank als relevanteste, gesamtheitliche Datengrundlage vor.

2.1 Überblick über Künstliche Intelligenz in der medizinischen Diagnostik

Das Interesse am Einsatz von Künstlicher Intelligenz (KI) in der Medizin hat in den letzten Jahren erheblich zugenommen (vgl. Nazer et al., 2023). KI-Technologien, insbesondere Machine Learning (ML) und Deep Learning (DL), werden verwendet, um große Mengen an medizinischen Daten zu analysieren, Muster zu erkennen und Vorhersagen zu treffen. Diese Systeme sollen der Ärzteschaft dabei helfen, genauere Diagnosen zu stellen, Behandlungsoptionen besser zu bewerten und das Patientenmanagement weiter zu verbessern. Ein bedeutendes Anwendungsgebiet der KI in der medizinischen Diagnostik ist die Bildverarbeitung. Studien, wie die von Brinker et al. (2019), haben gezeigt, dass KI-Systeme Hautkrebs mit hoher Genauigkeit diagnostizieren können. Weitere wichtige Anwendungsfelder sind die Analyse elektronischer Gesundheitsakten (EHRs) (vgl. Kaylor, 2023) und die Vorhersage von Krankheitsverläufen (vgl. Gala et al., 2024).

In der Bildverarbeitung kommen vor allem CNNs zum Einsatz, die besonders effektiv bei der Analyse von Bilddaten sind. CNNs sind tiefe neuronale Netzwerke, die hauptsächlich für die Verarbeitung von Bilddaten verwendet werden, indem sie Faltungen (Convolutions) anwenden, um lokale Muster zu erkennen und Merkmale wie Kanten und Texturen zu extrahieren (vgl. O'Shea & Nash, 2015). Ein prägnantes Beispiel ist die Diagnose von Hautkrebs mittels CNNs. In einer Studie von Brinker et al. (2019) wurde gezeigt, dass ein CNN-Modell Hautkrebs mit einer Genauigkeit von 95 % diagnostizieren konnte, während die Genauigkeit bei Dermatolog:innen bei 87 % lag. Diese Modelle durchlaufen einen umfangreichen Trainingsprozess, bei dem Millionen von Bildern analysiert werden, um Anomalien wie Melanome zu identifizieren. Ein weiteres Beispiel ist die Diagnose von Lungenentzündungen auf Röntgenbildern. Ein CNN-Modell, das von Rajpurkar et al. (2017) entwickelt wurde, konnte Lungenentzündungen mit einer bemerkenswert hohen Genauigkeit identifizieren, und schnitt besser ab als praktizierende Radiolog:innen.

Auch die Analyse elektronischer Gesundheitsakten gehört zu den Einsatzgebieten von KI-Forschung in der Medizin. KI-Modelle wie RNNs[7] und LSTM[8]-Netzwerke sind besonders geeignet, zeitliche Sequenzen und Abhängigkeiten in EHR-Daten zu erkennen. RNNs sind eine Art von neuronalen Netzwerken, die speziell für die Verarbeitung sequenzieller Daten entwickelt wurden, indem sie Informationen über vergangene Eingaben in ihren internen Zuständen speichern und dadurch Muster in Zeitreihen- oder Sequenzdaten erkennen können (vgl. Sherstinsky, 2020). LSTMs sind eine spezielle Variante von RNNs, die entwickelt wurden, um das Problem des fehlenden Langzeitgedächtnisses zu lösen, indem sie Speicherzellen verwenden, die es ermöglichen, Informationen über längere Zeiträume hinweg beizubehalten und so Abhängigkeiten in langen Sequenzen zu erkennen (vgl. Sherstinsky, 2020). Diese Modelle können Vorhersagen über Krankheitsverläufe treffen, indem sie historische Patientendaten analysieren. Eine Studie von Rajkomar et al. (2018) zeigte, dass ein auf LSTM basierendes Modell mit hoher Präzision Sterblichkeitsrate, Wiederaufnahme ins Krankenhaus, verlängerten Krankenhausaufenthalt und Entlassungsdiagnosen vorhersagen konnte und dabei klassische Modelle abhängte.

Allerdings gibt es auch technische Herausforderungen und Risiken bei der Integration von KI in die medizinische Diagnostik. Die Effektivität von KI-Modellen hängt stark von der Qualität und Quantität der Trainingsdaten ab. Unvollständige oder unausgewogene Datensätze führen wie bereits aufgeführt zu fehlerhaften Modellen. Dies wird in der Studie von Petersson et al. (2022) untermauert, die die Herausforderungen der Implementierung von KI im Gesundheitswesen durch Interviews mit Führungskräften im Gesundheitssektor untersucht hat. Die Ergebnisse beweisen, dass unzureichende Datenqualität und -quantität wesentliche Hindernisse darstellen, wie auch bereits in Kapitel 1.2 erläutert.

Ein weiteres Problem ist die Interpretierbarkeit und Transparenz von KI-Modellen, insbesondere von tiefen neuronalen Netzen, die oft als „Black Boxes“ betrachtet werden (vgl. Rigby, 2019). Diese Modelle liefern zwar genaue Vorhersagen, allerdings es ist schwierig nachzuvollziehen, wie genau sie zu diesen Vorhersagen gelangen. Dies erschwert es den medizinischen Fachkräften Entscheidungen der KI zu vertrauen und diese in den klinischen Kontext zu integrieren. Zudem beeinträchtigt die fehlende Interpretierbarkeit die Nachvollziehbarkeit der Diagnoseprozesse und mindert auch das Vertrauen der Patienten (vgl. Rudin, 2019).

Zudem erfordert die Integration von KI in die medizinische Praxis die Entwicklung und Pflege robuster und skalierbarer IT-Infrastrukturen. Diese müssen technisch in der Lage sein, große Datenmengen zu verarbeiten und gleichzeitig die Datensicherheit und den Schutz der Patienteninformationen zu gewährleisten. Dies umfasst die Implementierung von Hochleistungsrechenzentren und die Nutzung von Cloud-Computing-Diensten, um

[7] Recurrent Neural Networks
[8] Long Short-Term Memory

die benötigte Rechenleistung bereitzustellen. Ein umfassender Überblick über die technischen Herausforderungen und Lösungen im Zusammenhang mit der Nutzung von KI im Gesundheitswesen wird in der Studie von He et al. (2019) erörtert.

2.2 Relevante KI-Technologien

In diesem Kapitel werden nun die für diese Arbeit relevanten, grundlegenden KI-Technologien vorgestellt, die in der medizinischen Diagnostik Anwendung finden. Diese Technologien umfassen Neuronale Netzwerke, Transformer, sowie die Modelle BERT und BioBERT. Ein Verständnis dieser Technologien ist entscheidend, um die Funktionsweise von KI in der medizinischen Diagnostik vollständig zu erfassen und eine Bewertung der derzeitigen Implementierung abzuleiten.

2.2.1 Neuronale Netzwerke

Neuronale Netzwerke sind eine Klasse von maschinellen Lernmodellen, die von der Funktionsweise des menschlichen Gehirns inspiriert sind. Sie bestehen aus Neuronen, die in Schichten organisiert und miteinander verbunden sind (Abbildung 1). Diese Netzwerke können Muster in Daten erkennen und lernen, indem sie Gewichtungen anpassen, um die Vorhersagegenauigkeit zu maximieren (vgl. Javid, 2021).

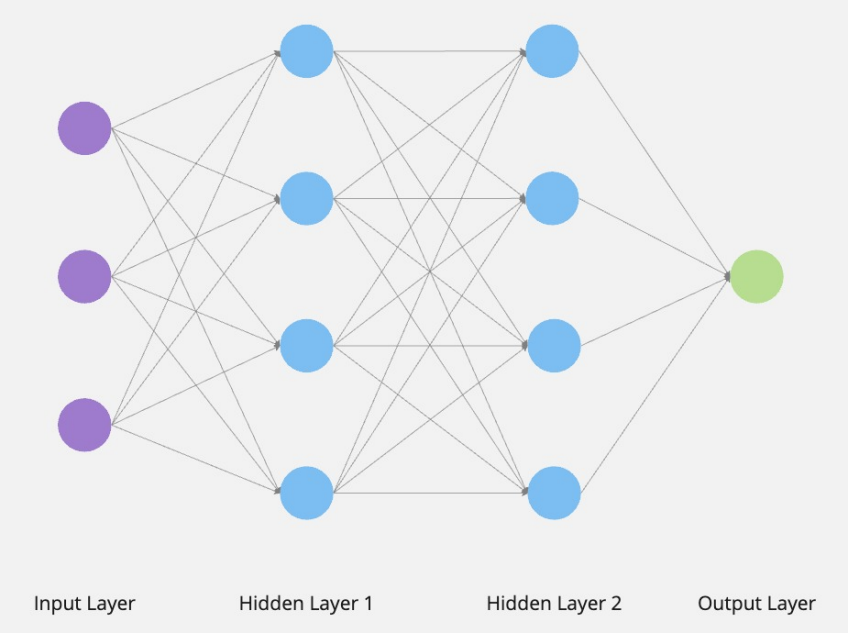

Abbildung 1: Architektur Neuronale Netzwerke, eigene Darstellung in Anlehnung an Javid (2021)

Hauptkomponenten:

- **Neuronen:** Die grundlegenden Recheneinheiten des Netzwerks, die Eingaben empfangen, verarbeiten und Ausgaben weiterleiten (vgl. Javid, 2021).

- **Schichten:** Neuronale Netzwerke bestehen aus Eingabeschichten, versteckten Schichten und Ausgabeschichten. Jede Schicht verarbeitet die Daten und gibt sie an die nächste Schicht weiter (vgl. Javid, 2021).
- **Aktivierungsfunktionen:** Funktionen wie RELU[9] (positive Eingabewerte = Identitätsfunktion, negative Eingabe = *Null*) oder Sigmoid (Transformation von Eingabewerten in den Bereich von 0 und 1), die die Ausgabe eines Neurons bestimmen und Nichtlinearität in das Modell einführen (vgl. Javid, 2021).

Neuronale Netzwerke sind sehr anpassungsfähig und können für eine Vielzahl von Aufgaben verwendet werden, von Bild- und Spracherkennung bis hin zu Textverarbeitung. Durch Training auf großen Datensätzen können sie komplexe Muster und Zusammenhänge erkennen (vgl. Javid, 2021).

2.2.2 Transformer

Transformer sind eine spezielle Art von neuronalen Netzwerken, die für Aufgaben der natürlichen Sprachverarbeitung (NLP[10]) entwickelt wurden. Sie wurden erstmals im bahnbrechenden Papier *Attention is All You Need* von Vaswani et al. (2023) vorgestellt. Transformer nutzen eine Mechanik namens Selbstaufmerksamkeit (Self-Attention), um den Kontext von Wörtern in einem Text zu erfassen.

Hauptkomponenten:

- **Selbstaufmerksamkeit:** Diese Mechanik ermöglicht es dem Modell, auf verschiedene Teile des Eingabetextes zu "achten" und die Beziehungen zwischen Wörtern zu erfassen. Dies geschieht durch das Berechnen von Gewichtungen für jedes Wort in Bezug auf jedes andere Wort im Satz (vgl. Vaswani et al., 2023).
- **Positionskodierungen:** Da Transformer-Modelle keine inhärente Sequenzinformation haben, werden Positionskodierungen hinzugefügt, um die Position eines Wortes in der Sequenz zu berücksichtigen (vgl. Vaswani et al., 2023).
- **Encoder-Decoder-Architektur:** Der Transformer besteht aus einem Encoder, der die Eingabesequenz verarbeitet, und einem Decoder, der die Ausgabe sequenziell generiert. Diese Architektur eignet sich hervorragend für Aufgaben wie maschinelle Übersetzung (vgl. Vaswani et al., 2023).

Im Vergleich zu RNNs oder CNNs können Daten parallel verarbeitet werden, was zu erheblich schnelleren Trainingszeiten führt. Durch die Selbstaufmerksamkeit kann der Transformer den Kontext über die gesamte Eingabesequenz hinweg berücksichtigen, was zu besseren Ergebnissen bei NLP-Aufgaben führt (vgl. Vaswani et al., 2023).

[9] Rectified Linear Unit
[10] Natural Language Processing

2.2.3 BERT

BERT, entwickelt von Google, steht für *Bidirectional Encoder Representations from Transformers*. Das Modell wurde im Paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"von Devlin et al. (2018) vorgestellt. BERT nutzt eine bidirektionale Selbstaufmerksamkeit, um den Kontext von Wörtern sowohl von links nach rechts als auch von rechts nach links zu verstehen.

Hauptkomponenten:

- **Bidirektionales Training:** Im Gegensatz zu Transformern, die Textsequenzen entweder von links nach rechts oder von rechts nach links verarbeiten, wird BERT bidirektional trainiert, was zu einem tieferen Verständnis des Kontexts führt (vgl. Devlin et al., 2018).
- **Pre-Training und Fine-Tuning:** BERT wird zunächst auf einer großen Textmenge vortrainiert (Pre-Training) und anschließend auf spezifische Aufgaben feinabgestimmt (Fine-Tuning). Dies ermöglicht es BERT, sich an verschiedene NLP-Aufgaben wie Fragebeantwortung, Textklassifikation und Sentimentanalyse anzupassen (vgl. Devlin et al., 2018).
- **Maskierte Sprachmodellierung:** Beim Pre-Training maskiert BERT zufällig einige Wörter im Text und trainiert das Modell darauf, diese maskierten Wörter vorherzusagen. Dies hilft dem Modell, ein tiefes Verständnis des Kontexts zu entwickeln (vgl. Devlin et al., 2018).

BERT erreicht durch seine hohe Genauigkeit state-of-the-art Ergebnisse bei 11 NLP-Benchmarks, zum Beispiel eine GLUE[11]-Punktzahl von 80.5%[12]. Beim GLUE-Benchmark landet das Modell damit derzeit auf Platz 49 (NYU, 2024). Durch die Architektur von BERT ist es möglich dasselbe Modell für verschiedene NLP-Aufgaben zu verwenden. Dies wird durch die Feinabstimmung auf die spezifische Aufgabe erreicht (vgl. Devlin et al., 2018).

2.2.4 BioBERT

BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) ist eine spezialisierte Version von BERT, die für biomedizinische Textmining-Aufgaben entwickelt wurde. Das Modell wurde im Paper *BioBERT: a pre-trained biomedical language representation model for biomedical text mining* von Lee et al. (2019)

[11] General Language Understanding Evaluation

[12] Der GLUE Benchmark-Test testet die Leistung des natürlichen Sprachverständnisses von Modellen anhand einer Vielzahl von Aufgaben, darunter Textklassifikation und Fragebeantwortung, um die allgemeine Fähigkeit zur Sprachverarbeitung zu bewerten (vgl. Wang et al., 2018)

erstmals vorgestellt. BioBERT wird auf großen biomedizinischen Textsammlungen, wie PubMed Abstracts und Volltextartikeln, vortrainiert, um medizinische Terminologien und den Kontext in biomedizinischen Texten besser zu verstehen.

Hauptkomponenten:

- **Domänenspezifisches Pre-Training:** BioBERT wird auf biomedizinischen Texten vortrainiert, was es ermöglicht, die speziellen Anforderungen und Terminologien des biomedizinischen Bereichs zu erfassen (vgl. Lee et al., 2019).
- **Anwendungen:** BioBERT wird für Aufgaben wie Named Entity Recognition (Spezifizierung und Klassifizierung von Informationen in Texten), Relation Extraction (Identizifierung und Klassifizierung von Beziehungen benannter Entitäten) und Fragebeantwortung im biomedizinischen Bereich verwendet (vgl. Lee et al., 2019).

Durch das Training auf domänenspezifischen Texten erzielt BioBERT bessere Ergebnisse in biomedizinischen Aufgaben im Vergleich zu allgemeinen Sprachmodellen, wie zum Beispiel dessen Basismodell BERT. Auch dadurch kann das Modell auf verschiedenste biomedizinische Aufgaben angewendet werden, indem es, ähnlich dem ursprünglichen BERT Modell, feinabgestimmt wird (vgl. Lee et al., 2019).

2.3 Vorstellung der MIMIC III-Datenbank

Die MIMIC[13]-III-Datenbank ist eine umfangreiche, öffentlich zugängliche Datenbank, die anonymisierte Gesundheitsdaten von Patient:innen enthält, die auf der Intensivstation des Beth Israel Deaconess Medical Center (Boston, Massachusetts, USA) zwischen 2001 und 2012 behandelt wurden. Sie ist die einzige öffentlich zugängliche Datenbank ihrer Art und ihres Umfangs (vgl. Johnson et al., 2016). Die Datenbank umfasst demografische Informationen, Vitalzeichen, Laborergebnisse, Medikamentenverabreichungen, Diagnosen, Verfahren und Informationen zum Vitalstatus. Sie wird häufig für die Forschung im Bereich der Künstlichen Intelligenz und des maschinellen Lernens verwendet, da sie eine wertvolle Ressource für die Entwicklung und Validierung der Algorithmen bietet (siehe Kapitel 2.3.4).

2.3.1 Dateninhalt und -struktur

Die MIMIC-III-Datenbank enthält Daten von rund 39.000 Patient:innen und rund 53.000 Aufenthalten. Der Median des Patientenalters ist 65 Jahre, von diesen Patient:innen sind rund 56% männlich und 44%weiblich. Aus den Gesamtdaten ergibt sich eine Gesamtsterblichkeitsrate von ca. 11,5% (vgl. Johnson et al., 2016). Sie umfasst eine Vielzahl von Datenkategorien, darunter:

[13]Medical Information Mart for Intensive Care

- **Demografische Daten:** Alter, Geschlecht, ethnische Zugehörigkeit, Aufnahmedetails und Entlassungsstatus der Patient:innen.
- **Klinische Daten:** Diagnosen, Laborergebnisse, Vitalparameter, Behandlungspläne und Medikation.
- **Verlaufsdaten:** Tägliche Aufzeichnungen von Vitalparametern, Medikamentengabe, Flüssigkeitsbilanz und Beobachtungen des Pflegepersonals.
- **Bildgebende Daten:** Ergebnisse von Radiologieberichten, CT-Scans und MRTs.
- **Administrative Daten:** ICD-9-Codes, DRG-Codes und Krankenhauskosten.

Die Daten sind in einer relationalen Datenbankstruktur organisiert, was den Zugriff und eine umfassende und kohärente Analyse ermöglicht.

2.3.2 ICD9-Codes

Laut World Health Organization and International Conference for the Ninth Revision of the International Classification of Diseases (1977) sind ICD-9-Codes numerische Codes, die zur Klassifikation von Krankheiten und Gesundheitszuständen verwendet werden. Sie bestehen sie aus drei bis fünf Zeichen und sind in drei Hauptteile unterteilt:

- **Zifferngruppe:** Die ersten drei Ziffern identifizieren die Hauptkategorie der Erkrankung oder des Zustands.
- **Untergruppe:** Nach einem Dezimalpunkt folgen ein bis zwei Ziffern, die spezifischere Informationen über den Zustand geben.
- **Erweiterungen:** Zusätzliche Zeichen können zur weiteren Spezifikation hinzugefügt werden.

Diese Struktur ermöglicht eine detaillierte und präzise Klassifikation von Erkrankungen und Diagnosen und ist essentiell für die Dokumentation, Abrechnung und statistische Analyse im Gesundheitswesen.

2.3.3 Datenzugriff und Datenschutz

Da der Schutz der Patientendaten höchste Priorität hat, ist der Zugang zur MIMIC-III-Datenbank für Forschende nur nach erfolgreicher Absolvierung eines umfangreichen Schulungsprogramms und der Zustimmung zu den Datenschutz- und Nutzungsbedingungen möglich (vgl. MIT Laboratory for Computational Physiology, 2021). Das Schulungsprogramm beinhaltet ethische Richtlinien für den Umgang mit Gesundheitsdaten und behandelt Themen wie Datenschutz, ethische Prinzipien und regulatorische Anforderungen

im Umgang mit menschlichen Forschungsdaten. Nach Abschluss dieser Schulung müssen Forschende unter Angabe eines PhysioNet-Accounts eine Datenzugangsanfrage einreichen, die eine kurze Beschreibung des geplanten Forschungsprojekts und eine Zustimmung zur Einhaltung der Nutzungbedingungen enthält (vgl. MIT Laboratory for Computational Physiology, 2021).

Da die MIMIC-III Datenbank detaillierte, persönliche Informationen über Patient:innen enthält, wurden alle Daten in der Datenbank anonymisiert, um die Privatsphäre dieser zu wahren. Identifizierende Informationen wie Namen, Adressen und genaue Datumsangaben wurden entfernt oder verfremdet, um eine Rückverfolgung zu verhindern. Alters- sowie Aufenthaltsdauerangaben, die ein wichtiger Bestandteil der medizinischen Daten sind, können trotz Verfremdung dennoch nachvollzogen werden. So werden zu Zwecken der Anonymisierung beispielsweise Aufnahme- und Entlassungsdatum 10 Jahre in die Zukunft verschoben.

2.3.4 Nutzung und Anwendung

Die MIMIC-III-Datenbank wird weltweit von Forschenden und anderen Interessengruppen genutzt, um verschiedene Aspekte der Medizin zu untersuchen. Beispielanwendungen der Forschung umfassen:

- **Prognosemodelle:** Entwicklung von Modellen zur Vorhersage von Diagnosen, Sterblichkeitsraten, Länge des Krankenhausaufenthalts und Wiederaufnahme ins Krankenhaus der Patient:innen (vgl. Nemati et al., 2018).
- **Krankheitsverläufe:** Analyse von Krankheitsverläufen und Identifizierung von Risikofaktoren für Komplikationen bei (kritisch) kranken Patient:innen (vgl. Gala et al., 2024).
- **Behandlungseffekte:** Untersuchung der Wirksamkeit und Nebenwirkungen verschiedener therapeutischer Behandlungen (vgl. Symeonidis et al., 2022).
- **KI-Entwicklung:** Training und Validierung von KI-Algorithmen zur automatisierten Diagnose, Behandlungsempfehlung und Überwachung von Patient:innen (vgl. van Aken et al., 2021).

2.4 Erläuterung des klinischen Ergebnismodells (CORe)

Das Clinical Outcome Representation Model (CORe) ist ein innovativer Ansatz zur Vorhersage klinischer Ergebnisse zu Beginn der Behandlung anhand der Aufnahmedaten von Patient:innen. Das Modell wurde im Paper *Clinical Outcome Prediction from Admission*

Notes using Self-Supervised Knowledge Integration von van Aken et al. (2021) vorgestellt. Entwickelt wurde es, um der Ärzteschaft bereits während des Aufnahmeprozesses verbesserte Entscheidungshilfen zu bieten. Es nutzt textuelle Daten aus den elektronischen Gesundheitsakten (EHRs) der Patient:innen, im speziellen die Aufnahmenotizen, um vier Hauptparameter vorherzusagen: Diagnosen, durchzuführende Behandlungen, Sterblichkeitsraten und die Verweildauer im Krankenhaus. Das Modell basiert auf den vortrainierten Sprachmodellen BERT und BioBERT (siehe Abbildung 2), die durch selbstüberwachtes Lernen und die Integration zusätzlichen medizinischen Wissens weiter verbessert wurden. Ziel ist es, die Ärzteschaft bei der Diagnosefindung zu unterstützen, vor potenziellen Risiken zu warnen und Krankenhäusern darüber hinaus bei der Ressourcenplanung zu helfen. Das Modell wurde auf Basis der MIMIC III-Datenbank sowie weiteren Quellen, wie Wikipedia und PubMed (frei zugängliche Datenbank, hauptsächlich biomedizinische Literatur) trainiert, um genaue und robuste Vorhersagen zu ermöglichen (vgl. van Aken et al., 2021).

2.4.1 Methodik des CORe-Modells

Die Methodik des CORe-Modells umfasst mehrere Schritte, die darauf abzielen, die Vorhersagegenauigkeit und die praktische Anwendbarkeit zu maximieren:

1. **Datenquellen und -verarbeitung:** Die Primärdatenquelle ist die MIMIC III-Datenbank, bestehend aus anonymisierten EHR-Daten. Um die Patient:innen bei Aufnahme zu simulieren, werden Aufnahmenotizen aus den in der Datenbank enthaltenen Entlassungsnotizen extrahiert, indem ausschließlich bei der Aufnahme bekannte Abschnitte, wie die Hauptbeschwerde und die medizinische Vorgeschichte, beibehalten werden (siehe Anhangherzinsuffizienz). Dies ist notwendig, da die Entlassungsnotizen bereits umfassende Informationen zu z.B. Diagnosen enthalten können, die zu Aufnahmezeitpunkt noch nicht bekannt sind (vgl. van Aken et al., 2021).

2. **Selbstüberwachtes Pre-Training:** Das Pre-Training kombiniert Wissen über Patientenhistorie, Symptom-Krankheits-Zusammenhänge aus öffentlichen Quellen wie PubMed und Krankenhausaufenthaltsaufzeichnungen aus der MIMIC III-Datenbank. Es nutzt eine modifizierte Version der NSP[14], bei der das Modell lernt, Beziehungen zwischen Informationen zum Aufnahmezeitpunkt und Ergebnissen zu erkennen (vgl. van Aken et al., 2021).

3. **Integration der ICD9-Codes:** Diagnosen und Verfahren werden anhand der hierarchischen Struktur der ICD9-Codes (Kapitel 2.3.2) vorhergesagt. Zusätzliche La-

[14]Next Sentence Prediction

bels, die mit den Diagnosen und Verfahren assoziiert werden können, werden in das Modell integriert um die Lernergebnisse zu verbessern (vgl. van Aken et al., 2021).

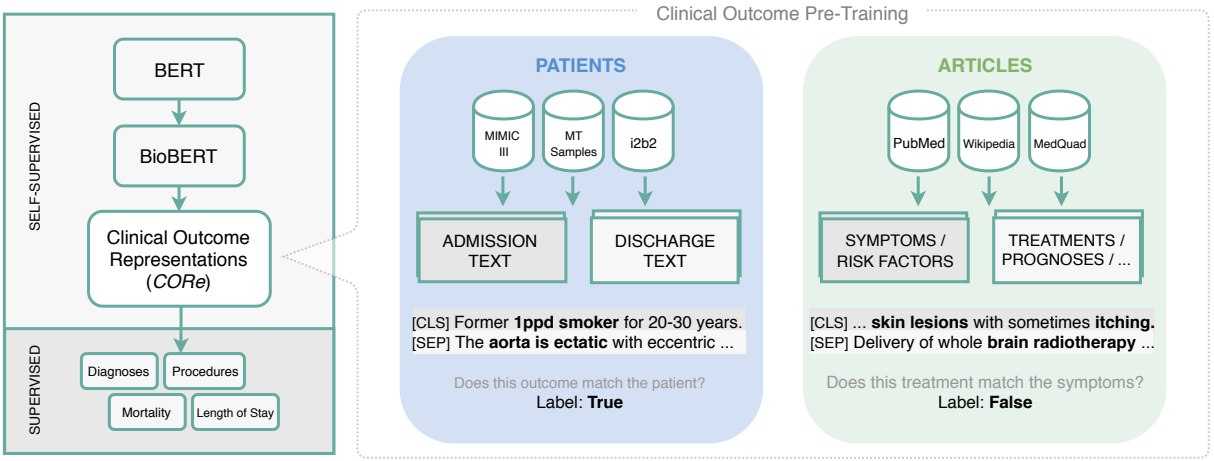

Abbildung 2: Architektur des CORe Modells (van Aken et al., 2021)

2.4.2 Implementierung und Validierung

Die Implementierung und Validierung des CORe-Modells erfolgte in mehreren Schritten, wie im Paper von van Aken et al. (2021) beschrieben:

- **Training und Finetuning:** Das CORe-Modell wurde auf Basis der BioBERT-Gewichtungen vortrainiert und anschließend für jede der vier Ergebnisaufgaben separat feinabgestimmt. Eine Tokenisierung mit WordPiece (Algorithmus zur Zerlegung von Wörtern in Wortteile) (vgl. Wu et al., 2016) und eine Beschränkung auf 512 Token wurden verwendet, um die begrenzte Kontextlänge der vortrainierten Modelle zu berücksichtigen (vgl van Aken et al., 2021).

- **Basismodell-Vergleich:** Um die Fähigkeiten dieses vortrainierten Modells zu verstehen, wurde es mit traditionelleren Modellen wie Bag-of-Words (Modell, das Texte in Menge von Wörtern darstellt) oder einem CNN verglichen. Auch andere vortrainierte Modelle wie BERT und ClinicalBERT wurden für den Vergleich herangezogen (vgl van Aken et al., 2021).

- **Ergebnisse:** Die vortrainierten Modelle haben die traditionellen Modelle teils stark übertroffen. Dies zeigt, dass das Vortrainieren die Fähigkeit zur Vorhersage klinischer Ereignisse verbessert. Ebenso übertrafen die auf den medizinischen Kontext spezialisierten vortrainierten Modelle die allgemeinen vortrainierten Modelle (vgl van Aken et al., 2021).

- **Transferfähigkeit:** Um festzustellen, ob die Diagnosefähigkeit der vortrainierten Modelle auf andere klinische Notizen als die der MIMIC-III-Datenbank transferrierbar sind, wurde ein anderer Datensatz (i2b2) ohne weiteres Training als Input

genutzt. Trotz des anderen Inhalts und Struktur der Notizen waren die Modelle in der Lage die Transferleistung zu erbringen und diagnostizierten weitgehend korrekt (vgl van Aken et al., 2021).

Zusammengefasst bietet das CORe-Modell also eine leistungsstarke Methode zur Vorhersage klinischer Ergebnisse, die auf einer Kombination aus selbstüberwachtem Lernen und der Integration spezialisierter medizinischer Wissensquellen basiert und verbessert Vorgängermodelle wie BERT und BioBERT.

2.4.3 Limitationen des Modells

Trotz der vielversprechenden Ergebnisse weist das von van Aken et al. (2021) vorgestellte Modell mehrere Limitationen auf. Eine zentrale Herausforderung des Modells ist der Umgang mit negativen Aussagen in den klinischen Notizen. Insbesondere spezifische Negationen wie *kein Alkoholkonsum* werden oft falsch interpretiert und führen zu Fehlklassifikationen. Beispielsweise könnten so Patient:innen, die als *nicht alkoholabhängig* beschrieben werden, fälschlicherweise als *alkoholabhängig* eingestuft werden (vgl. van Aken et al., 2021).

Ebenfalls zeigt das Modell Schwächen bei der Interpretation numerischer Daten. Obwohl klinische Notizen viele relevante Vitalparameter enthalten, gelingt es dem Modell nicht immer, lebensbedrohliche Werte wie eine Temperatur über 40°C korrekt als erhöhtes Mortalitätsrisiko zu interpretieren (vgl. van Aken et al., 2021).

2.5 Zusammenfassung Kapitel 2

Kapitel 2 bietet einen umfassenden Überblick über relevante KI-Technologien in der medizinischen Diagnostik, die für diese Arbeit von Bedeutung sind.

Im Detail werden neuronale Netzwerke, Transformer sowie BERT- und BioBERT-Modelle behandelt. Diese Technologien sind zentral für das Verständnis der Funktionsweise von KI in der medizinischen Diagnostik und der Bewertung ihrer Implementierung. Darüber hinaus wird die MIMIC-III-Datenbank als wesentliche Ressource für klinische Daten und Forschung vorgestellt. Sie enthält strukturierte Informationen über Patientenaufnahmen, die für die Entwicklung und Validierung des klinischen Ergebnismodells CORe genutzt werden. Das CORe-Modell wird in Bezug auf Methodik, Implementierung, Validierung und Limitationen erläutert und dient als Grundlage für die Durchführung und Bewertung der Testfälle im folgenden Kapitel.

Diese Grundlagen sind entscheidend für die Untersuchung der Modellleistung und der Erkennung möglicher Voreingenommenheiten in der diagnostischen Praxis.

3 Hauptteil

Dieses Kapitel widmet sich der detaillierten Analyse von Voreingenommenheitsfaktoren des im Kapitel 2.4 beschriebenen CORe-Modells. Durch die Implementierung und Analyse spezifischer Testfälle wird untersucht, wie geschlechts-, alters- und ethnizitätsspezifische Voreingenommenheit die diagnostischen Ergebnisse beeinflussen. Die Ergebnisse werden dabei im Kontext der amerikanischen Bevölkerung interpretiert, um eine Vergleichbarkeit aus Datengrundlage (MIMIC-III-Datenbank) und Realbevölkerung herzustellen. Auf Basis dieser Analysen werden ethische Leitlinien abgeleitet, um Voreingenommenheit in zukünftigen KI-Anwendungen zu minimieren.

3.1 Voreingenommenheit in medizinischen KI-Modellen

Wie in einer Studie von Obermeyer et al. (2019) gezeigt wurde, hat Voreingenommenheit in medizinischen KI-Modellen schwerwiegende Auswirkungen auf die Qualität und Fairness der Diagnosen sowie der anschließenden Behandlungen. Solche Verzerrungen können, wie bereits in Kapitel 1.2 erörtert, geschlechtsspezifisch, altersbezogen oder genetisch bedingt sein und führen zu einer ungleichen Behandlung unterschiedlicher Bevölkerungsgruppen. Zum Beispiel kann geschlechtsspezifische Voreingenommenheit dazu führen, dass weibliche Patienten seltener korrekte Diagnosen erhalten, da Trainingsdaten überwiegend von männlichen Patienten stammen. Altersspezifische Voreingenommenheit kann dazu führen, dass ältere Patienten fehldiagnostiziert werden, da das Modell hauptsächlich mit Daten jüngerer Patienten trainiert wurde. Genetische Voreingenommenheit beschreibt, dass KI-Modelle z.B. bei bestimmten ethnischen Gruppen schlechtere Diagnosen liefern als bei anderen.

Diese Voreingenommenheitsphänomene können verschiedene Ursachen haben, darunter eine unrepräsentative Verteilung in den Trainingsdaten sowie unzureichende oder zu starke Berücksichtigung bestimmter Merkmale während der Modellentwicklung (vgl. Nazer et al., 2023). Forschungsergebnisse belegen, dass KI-Modelle, die auf ungleich verteilten Daten trainiert wurden, dazu neigen, bestimmte genetische Merkmale überzubetonen oder zu vernachlässigen, was zu einer fehlerhaften Anwendung in der klinischen Praxis führen kann (vgl. Obermeyer et al., 2019). Während Alter und Geschlecht zu den erfassten Basisdaten gehören, sind genetische Merkmale wie die Ethnizität, zumindest in Deutschland, selten bis nie in EHRs enthalten. Solche Ursachen können zu fehlerhaften Diagnosen, unangemessen Handlungsempfehlungen und insgesamt zu einer suboptimalen Gesundheitsversorgung für bestimme demografische Gruppen führen. Die Identifizierung und Minimierung dieser Voreingenommenheitseffekte sind daher von entscheidender Bedeutung, um die Fairness und Effektivität in der Nutzung medizinischer KI-Modelle sicherzustellen (vgl. Nazer et al., 2023).

3.2 Umsetzung von Testfällen mit MIMIC III und CORe

Um die im vorigen Kapitel beschriebenen Formen von Voreingenommenheit im CORe-Modell zu untersuchen, werden die MIMIC-III Daten aufbereitet sowie spezifische Testfälle extrahiert. Dabei werden die Testfälle aus den Entlassungsnotizen der MIMIC-III Datenbank extrahiert, modifiziert und dem Modell zur Analyse zugeführt. Diese Testfälle werden verwendet, um die diagnostischen Prognosen des CORe-Modells in den jeweiligen Szenarien zu analysieren und bewerten.

3.2.1 Herleitung und Datenextraktion

Zur Datenextraktion und -verarbeitung der spezifischen Falldaten aus der MIMIC-III Datenbank wurde ein Python Skript entwickelt. Dieses Skript, welches in der Datei `data_extraction.py` (Anhang 1) zu finden ist, extrahiert und bereitet relevante Daten auf. Unterstützende Funktionen sind in `utilities.py` (Anhang 2) enthalten. Die extrahierten Daten dienen als Grundlage für die Entwicklung und Analyse der Testfälle in den folgenden Kapiteln. Das Skript führt folgende Schritte aus:

1. Definition relevanter Schlüssel (Spaltennamen in MIMIC-III Datensatz).
2. Laden der MIMIC-III-Datensätze.
3. Verknüpfung der Diagnosen mit den Patientendaten und Aufnahmenotizen (generiert aus van Aken, Papaioannou, Mayrdorfer et al. (2022a)).
4. Berechnung zusätzlicher Merkmale wie Alter bei Aufnahme und Aufenthaltsdauer sowie Filtern der Daten nach Plausibilität.

Abschließend werden die bereinigten Daten in einer CSV-Datei gespeichert und demografische Verteilungen visualisiert, um mögliche Verzerrungen zu identifizieren. Diese Visualisierungen (Tabellen 1, 2, 4) helfen, die Zusammensetzung des ursprünglichen Datensatzes zu analysieren und vergleichen und mögliche Verzerrungen im Modell zu identifizieren. Im Abschnitt 3.2.3 wird genauer auf diese Zusammensetzungen eingegangen.

3.2.2 Ausführung der Testfälle

Bevor mit der Ausarbeitung der Testfälle fortgefahren wird, werden zunächst einige Statistiken für die jeweiligen Fälle erhoben. In der Datei `testcase_evaluation.py` (Anhang 3) werden dazu folgende Schritte abgearbeitet:

1. Definition relevanter Schlüssel (Spaltennamen in MIMIC-III Datensatz).
2. Laden der Ergebnisdatei der Datenextraktion (Kapitel 3.2.1).

3. Definition der zu untersuchenden ICD-Codes zu den Testfällen in 3 Sektionen - altersbasiert, genderbasiert, ethnizitätsbasiert.

4. Durchsuchen des MIMIC-III-Datensatzes nach diesen ICD-Codes (Spalte `ICD9_CODE` aus der ursprünglichen Datei `DIAGNOSES_ICD.csv`).

5. Zählung der Fälle und Gruppierung je nach zu untersuchendem Parameter.

Die Wahl der ICD-Codes die in dieser Arbeit untersucht werden ist damit zu begründen, dass diese Erkrankungen vorrangig in bestimmten demografischen Gruppen vorherrschen, weswegen sie sich gut für die Untersuchung von Voreingenommenheit im CORe-Modell eignen. Weitere Erläuterungen zu den einzelnen Fällen und Diagnosen folgen im Kapitel 3.2.3.

Zum Abschluss werden Diagramme zur Repräsentation dieser Daten zur späteren Referenz erstellt, aus dem die Gesamtzahl der gefundenen Fälle sowie die Alters- oder Ethnizitätsverteilung hervorgeht (Abbildungen 3, 7, 8). Aufgrund der simplen Darstellung wurden die geschlechtsspezifischen Daten in der Tabelle 3 zusammengeführt.

Um nun einen repräsentativen Testfall für jeden ICD-Code zu extrahieren werden folgende Schritte durchlaufen:

1. Wahl eines zufälligen Testfalls für jeden ICD-Code aus der Liste mit allen extrahierten Fällen dieses ICD-Codes.

2. Speichern und manuelle Modifizierung der Fälle. Alter, Geschlecht und Ethnizität werden zu Variablen abgeändert.

3. Befragung des Modells zu jeder Variation eines jeden Testfalls. Das Modell wird direkt durch die Transformers Bibliothek von Huggingface (van Aken, Papaioannou, Mayrdorfer et al., 2022b) geladen und befragt.

Die Ergebnisse werden anschließend, in Fällen gruppiert, in Plots visualisiert und zur späteren Analyse gespeichert (Abbildungen 4, 5, 6, 9, 10). Die Aufnahmenotizen der gewählten Testfälle können im Anhang 4 eingesehen werden.

3.2.3 Sammlung der Ergebnisse

Die Daten und Prognosen aus den durchgeführten Testfällen werden nun gesammelt und aufbereitet, um sie im Kapitel 3.3.2 zu analysieren. Dabei wird besonders herausgestellt, wie sich die diagnostischen Genauigkeiten des CORe-Modells zwischen verschiedenen demografischen Gruppen unterscheiden. Die Ergebnissammlung konzentriert sich auf die

Aspekte Alter, Geschlecht und Ethnizität.

Altersverteilung

Die allgemeine Altersverteilung aller Diagnosen in der MIMIC-III Datenbank (Tabelle 1) zeigt einen klaren Trend zu Patient:innen höheren Alters, wie bereits in Kapitel 2.3.1 beschrieben. Diese Angaben decken sich dabei nicht mit den Angaben des Census 2010 & 2020, der Volkszählung der amerikanischen Bevölkerung durch das U.S. Census Bureau, siehe Tabelle 1. Die Fälle der MIMIC-III Datenbank enthalten im Vergleich einen signifikant geringeren Anteil an Personen, die jünger als 44 Jahre sind und, dementsprechend, deutlich mehr Personen in älteren Altersgruppen.

Tabelle 1: Altersverteilung USA und MIMIC-III Datenbank, eigene Darstellung in Anlehnung an U.S. Census Bureau (2023) und U.S. Census Bureau (n. d.)

Altersgruppe	Census 2010	Census 2020	MIMIC-III Datenbank
18-24	10%	9,4%	1,66%
25-34	13,3%	13,5%	3,36%
35-44	13,3%	12,7%	6,43%
45-64	33,2%	25,4%	34,4%
65-84	11,2%	14,9%	46,6%
85-99	1,8%	1,89%	7,46%

Die Altersverteilung in der MIMIC-III-Datenbank und die Diagnosequalität des CORe-Modells wurde für den ICD-Code 428 (Herzinsuffizienz) untersucht. Herzinsuffizienz ist eine chronische Erkrankung, bei der das Herz nicht in der Lage ist, ausreichend Blut zu pumpen, um den Bedarf des Körpers an Sauerstoff und Nährstoffen zu decken. Diese Erkrankung trifft hauptsächlich bei älteren Patient:innen auf (vgl. Strauer, 2007). Abbildung 3 zeigt die Häufigkeit jeglicher vom CORe-Modell gestellten Diagnosen in der MIMIC-III-Datenbank in den verschiedenen Altersgruppen und bestätigt diese Erwartung.

ICD-Code 428 - Herzinsuffizienz

- **Erwartung:** Herzinsuffizienz tritt deutlich häufiger bei Menschen höheren Alters auf.
- **MIMIC-III Datenlage:** Von insgesamt 18246 Fällen liegt die überwiegende Mehrheit der Fälle bei Patient:innen über 50 Jahren.
- **Ergebnis:** Herzinsuffizienz wurde bei Patient:innen aller Altersgruppen diagnostiziert, siehe Abbildung 4.

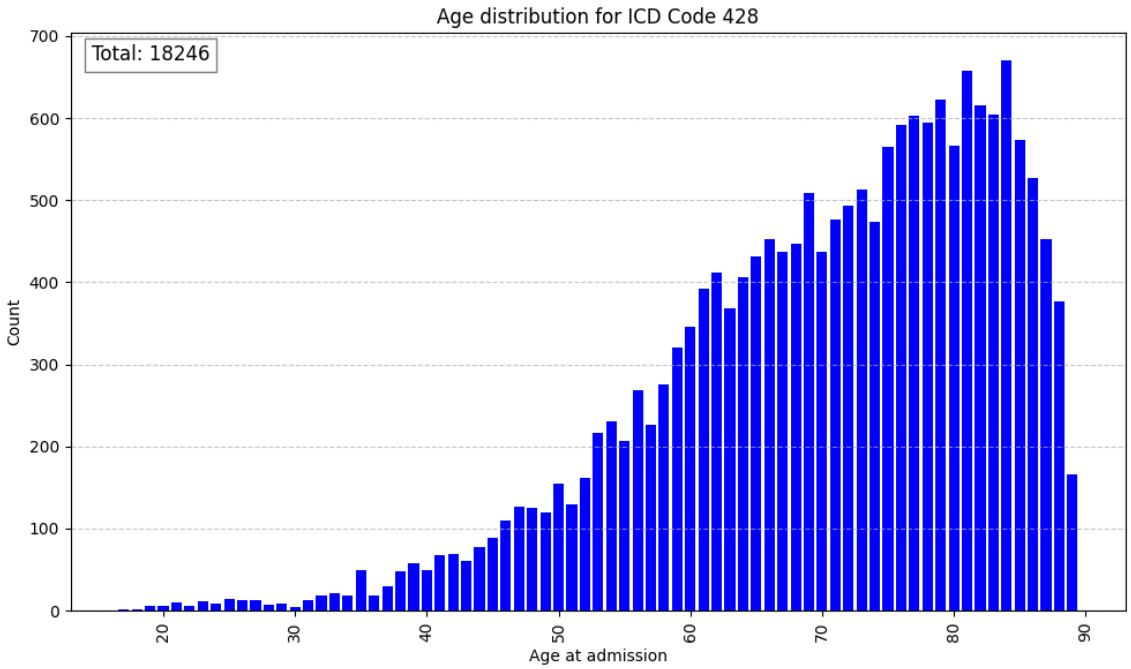

Abbildung 3: Altersverteilung ICD Code 428 - Herzinsuffizienz in der MIMIC-III Datenbank, eigene Darstellung

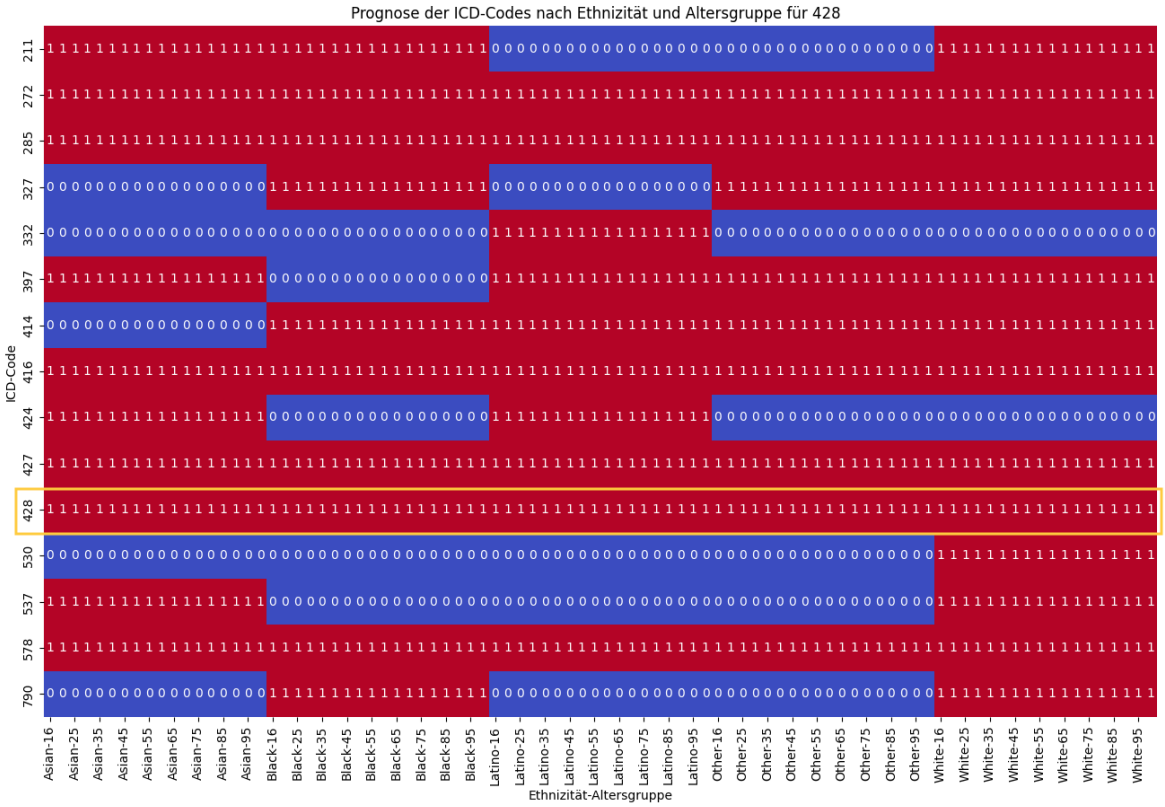

Abbildung 4: Diagnosevorhersagen ICD-Code 428 - Herzinsuffizienz, eigene Darstellung

Geschlechtsspezifische Verteilung

Die Geschlechterverteilung stimmt mit den Angaben aus Kapitel 2.3.1 weitestgehend überein. 56,65% aller Patient:innen sind männlich, 43,35% weiblich. Dies deckt sich nicht ganz mit der Bevölkerungsverteilung in den USA, siehe Tabelle 2, in der die Geschlechter nahezu gleiche Bevölkerungsanteile abbilden. Auch im Vergleich des Census 2010 und 2020 gab es in der amerikanischen Bevölkerung keine signifikanten Unterschiede.

Tabelle 2: Geschlechterverteilung USA und MIMIC-III Datenbank, eigene Darstellung in Anlehnung an U.S. Census Bureau (2020) und U.S. Census Bureau (n. d.)

Geschlecht	Census 2010	Census 2020	MIMIC-III Datenbank
Männlich	49,16%	49,6%	56,65%
Weiblich	50,84%	50,4%	43,35%

Die Geschlechterverteilung in der MIMIC-III Datenbank und die Diagnosequalität des CORe-Modells wurde anhand der ICD-Codes 617 (Endometriose) und 7330 (Osteoporose) untersucht.

Endometriose ist eine Erkrankung, bei der Gewebe, das dem Endometrium (der Gebärmutterschleimhaut) ähnelt, außerhalb der Gebärmutter wächst. Dieses Gewebe kann sich an den Eierstöcken, Eileitern, dem Darm und anderen Bereichen im Beckenbereich ansiedeln. Diese Erkrankung ist ausschließlich in biologisch weiblichen Patient:innen zu finden (vgl. Mayo Clinic, 2023).

Bei Osteoporose handelt es sich um eine Erkrankung, bei der die Knochendichte und -qualität verringert sind, was zu einem höheren Risiko für Knochenbrüche führt. Dies geschieht, weil die Knochenmasse schneller abgebaut als aufgebaut wird. Diese Erkrankung kann bei allen Geschlechtern und Ethnizitäten auftreten, jedoch am häufigsten bei Frauen (vgl. National Institute of Arthritis and Muscoloskeletal and Skin Diseases, 2022).

Tabelle 3 zeigt die Häufigkeit dieser Diagnosen in der Datenbank für die verschiedenen Geschlechter.

Tabelle 3: Geschlechterverteilung ICD Codes 617 & 7330 in der MIMIC-III Datenbank - eigene Darstellung

ICD-Code Diagnose	Männlich	Weiblich
617 (Endometriose)	0	36
7330 (Osteoporose)	304	1371

ICD-Code 617 - Endometriose

- **Erwartung:** Betrifft ausschließlich biologisch weibliche Patientinnen.
- **MIMIC-III Datenlage:** Von insgesamt 36 diagnostizierten Fällen gab es ausschließlich Diagnosen für Patientinnen.

- **Ergebnis:** Nicht diagnostiziert, stattdessen ICD-Code 614 (Entzündliche Erkrankung der Eierstöcke, Eileiter, Beckengewebe und des Bauchfells). Diese Diagnose wurde, abgesehen zweier Anomalien, nur für Patientinnen oder Patienten ohne Geschlechtsangabe unter 60 Jahren gestellt, siehe Abbildung 5.

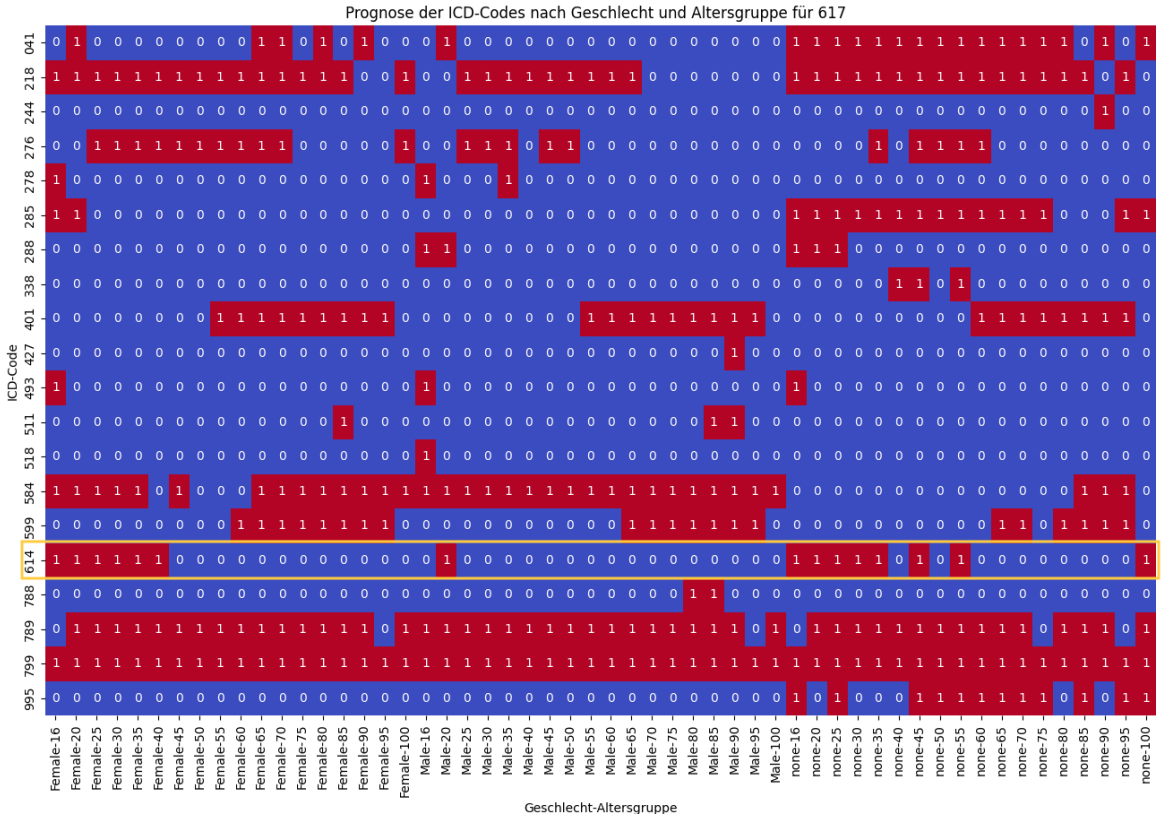

Abbildung 5: Diagnosevorhersagen ICD-Code 617 - Endometriose, eigene Darstellung

ICD-Code 7330 - Osteoporose

- **Erwartung:** Osteoporose tritt häufiger bei Patientinnen auf.
- **MIMIC-III Datenlage:** Die Mehrheit der insgesamt 1675 Diagnosen (rund 80%) wurden für Patientinnen gestellt.
- **Ergebnis:** Osteoporose wird etwa gleich oft bei männlichen und weiblichen Patien:innen diagnostiziert. Ohne Geschlechtsangabe ist die Diagnoserate niedriger, siehe Abbildung 6.

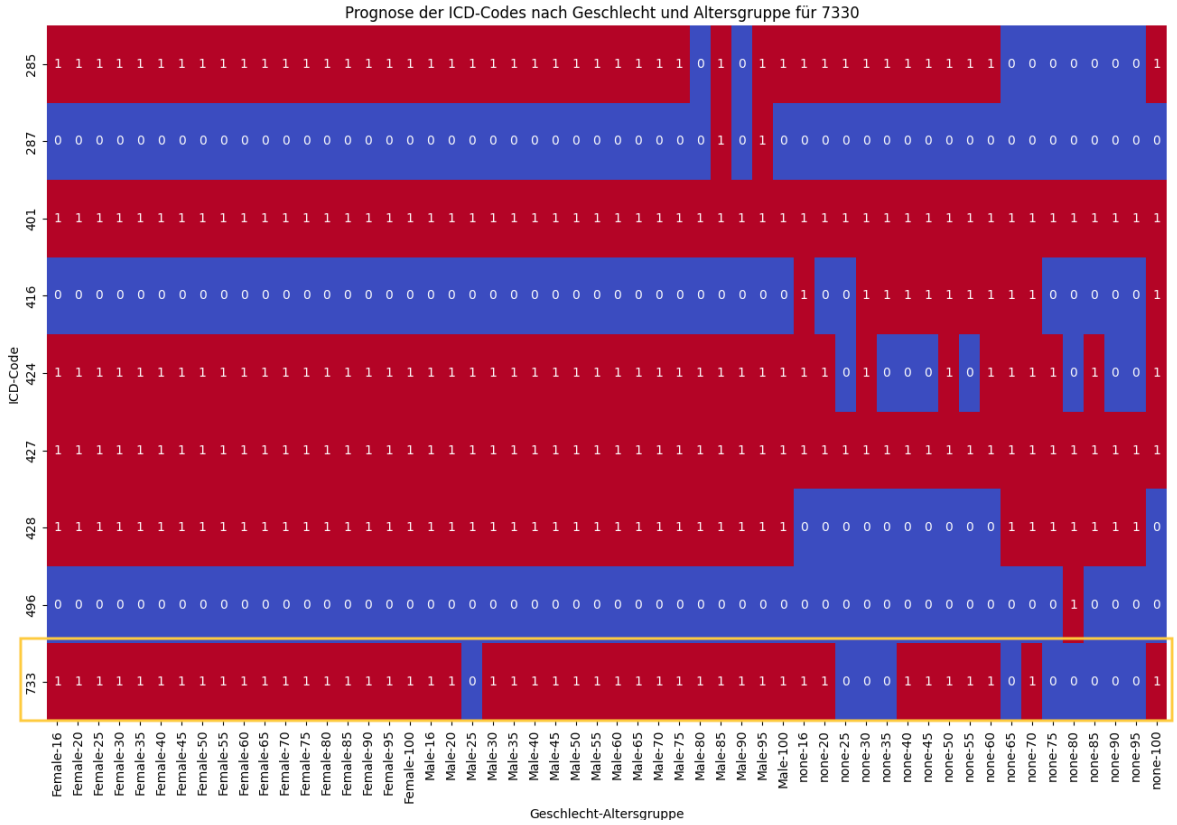

Abbildung 6: Diagnosevorhersagen ICD-Code 7330 - Osteoporose, eigene Darstellung

Ethnizitätsverteilung

Die Gesamtverteilung der Ethnizitäten in der MIMIC-III-Datenbank zeigt einen deutlich höheren Anteil an weißen Patient:innen im Vergleich zu anderen Ethnizitäten. Besonders Asiat:innen und Lateinamerikaner:innen sind im Vergleich zum amerikanischen Census stark unterrepräsentiert, siehe Tabelle 4.

Tabelle 4: Ethnizitätsverteilung USA und MIMIC-III Datenbank, eigene Darstellung in Anlehnung an U.S. Census Bureau (2020) und U.S. Census Bureau (n. d.)

Ethnizität	Census 2010	Census 2020	MIMIC-III Datenbank
Weiß	63,7%	58,9%	72,11%
Schwarz	13,6%	13,6%	10,56%
Lateinamerikanisch	16,4%	19,1%	3,49%
Asiatisch	5,6%	6,3%	2,22%
Andere	0,7%	2,1%	11,64%

Die ethnische Verteilung für relevante Testfälle wurde anhand der ICD-Codes 250 (Diabetes) und 151 (Magenkrebs) untersucht.

Diabetes mellitus ist eine Gruppe von Stoffwechselerkrankungen, die durch hohen Blutzuckerspiegel gekennzeichnet sind. Es gibt zwei Haupttypen: Typ 1 (Insulinmangel) und Typ 2 (Insulinresistenz). Beide Typen führen zu einer schlechten Verarbeitung des Blutzuckers. Typ 2 Diabetes ist häufiger und eng mit Übergewicht und Bewegungsman-

gel verbunden (vgl. U.S. Center for Disease Control and Prevention, 2024a). Während Diabetes in allen ethnischen Gruppen vorkommt, gibt es ein leicht höheres Risiko für Schwarze, Asiat:innen und Lateinamerikaner:innen (vgl. U.S. Center for Disease Control and Prevention, 2024b).

Magenkrebs ist eine Erkrankung, bei der bösartige (Krebs-)Zellen in der Magenschleimhaut gefunden werden. Die genauen Ursachen sind unklar, aber Risikofaktoren umfassen Ernährung, Rauchen, chronische Magenentzündungen und genetische Prädisposition (vgl. National Cancer Institute, n. d.). Magenkrebs tritt häufiger bei Schwarzen, Lateinamerikaner:innen und Asiat:innen als bei Weißen auf (vgl. National Cancer Institute, 2023).

Die Abbildungen 7 und 8 zeigen die Häufigkeit dieser Diagnosen in der Datenbank für die verschiedenen ethnischen Gruppen.

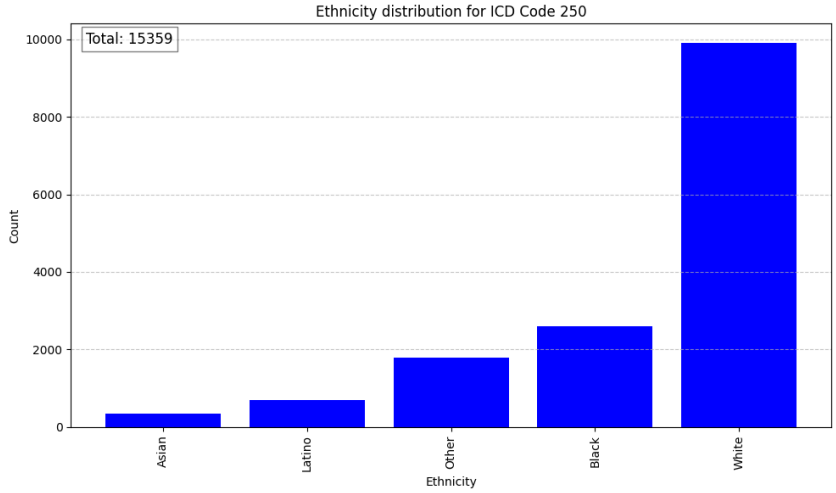

Abbildung 7: Ethnizitätsverteilung ICD Code 250 - Diabetes in der MIMIC-III Datenbank, eigene Darstellung

ICD-Code 250 - Diabetes

- **Erwartung:** Weit verbreitet über verschiedene ethnische Gruppen, stärker verbreitet bei Schwarzen, Asiat:innen und Lateinamerikaner:innen.
- **MIMIC-III Datenlage:** Weiße Patient:innen machen etwa 2/3 der insgesamt 15359 Fälle aus, gefolgt von schwarzen Patient:innen. Asiat:innen und Lateinamerikaner:innen werden wenig diagnostiziert.
- **Ergebnis:** Durch alle ethnischen Gruppen gleiche Diagnoseverteilung, siehe Abbildung 9.

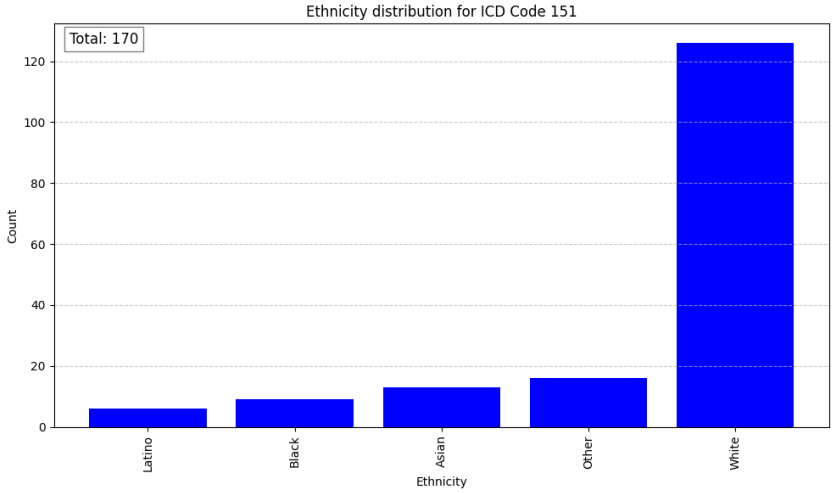

Abbildung 8: Ethnizitätsverteilung ICD Code 151 - Magenkrebs in der MIMIC-III Datenbank, eigene Darstellung

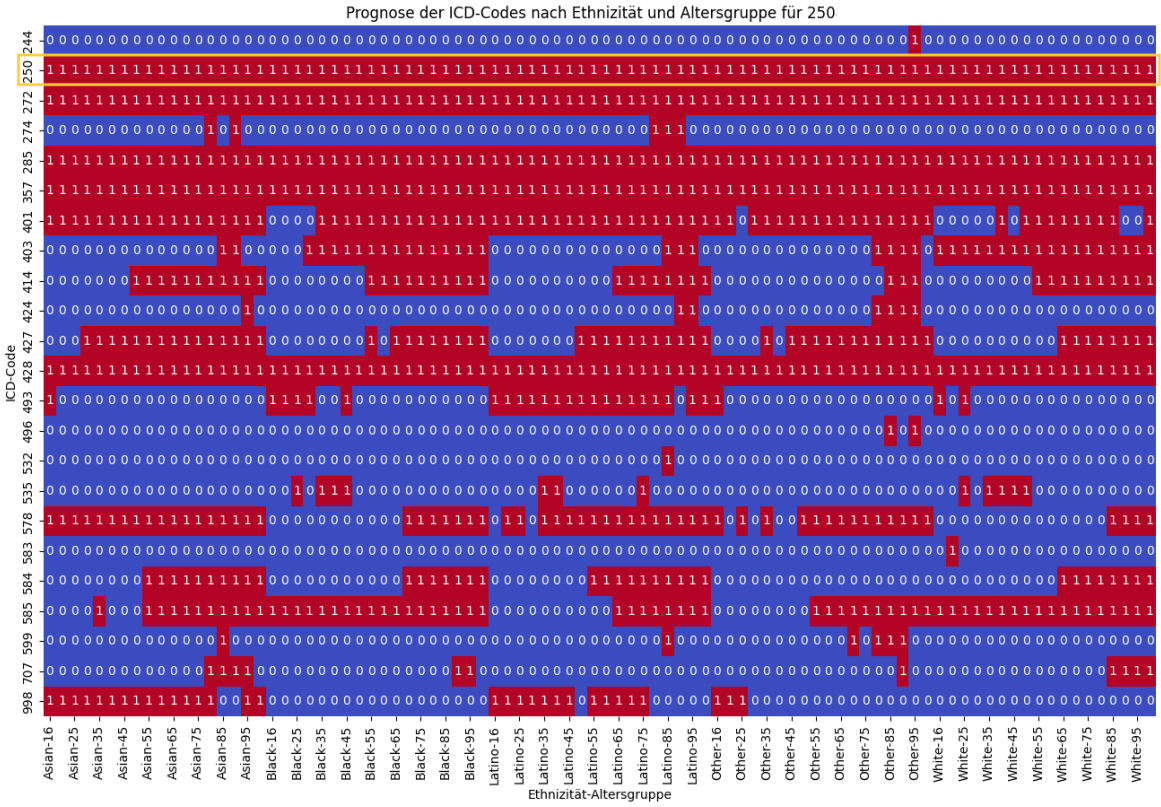

Abbildung 9: Diagnosevorhersagen ICD Code 250 - Diabetes, eigene Darstellung

ICD-Code 151 - Magenkrebs

- **Erwartung:** Weiße Patient:innen sind weniger häufig erkrankt als andere Ethnizitäten.
- **MIMIC-III Datenlage:** Von insgesamt 170 Fällen sind rund 70% der Patient:innen weiß. Asiatische, lateinamerikanische und schwarze Patient:innen werden etwa gleich

oft mit Magenkrebs diagnostiziert.

- **Ergebnis:** Nicht diagnostiziert. Dafür ICD9 Code 150 (Speiseröhrenkrebs), hauptsächlich bei Weißen, Schwarzen und anderen Ethnizitäten, siehe Abbildung 10.

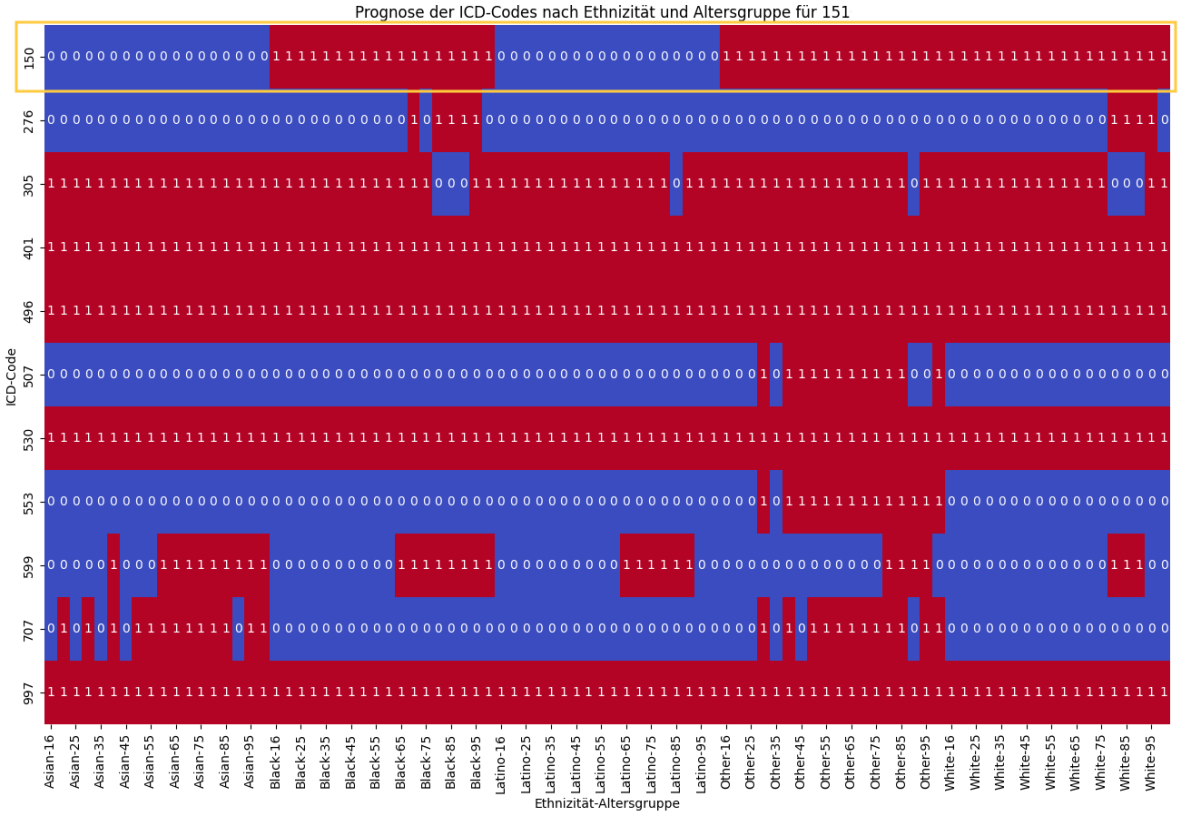

Abbildung 10: Diagnosevorhersagen ICD Code 151 - Magenkrebs, eigene Darstellung

3.3 Analyse der Testergebnisse und Ethische Bewertung

Die ethische Bewertung von Voreingenommenheit in medizinischen KI-Modellen ist entscheidend, um eine gerechte und effektive Gesundheitsversorgung sicherzustellen. In diesem Kapitel wird eine systematische Analyse der ethischen Implikationen der in Kapitel 3.2.3 identifizierten Voreingenommenheitseffekte durchgeführt. Diese Analyse dient als Grundlage für die Abwandlung von Leitlinien, die dazu beitragen sollen, Voreingenommenheit in zukünftigen KI-Anwendungen zu minimieren.

3.3.1 Ethische Prinzipien

Die ethische Bewertung basiert auf den folgenden, für diese Arbeit definierten, Prinzipien:

1. **Gerechtigkeit:** Alle Patient:innen sollten unabhängig von Geschlecht, Alter, ethnischer Herkunft oder anderen demografischen Merkmalen gleich behandelt werden. Dies bedeutet, dass die KI-Modelle keine systematischen Vorurteile gegenüber bestimmten Gruppen aufweisen dürfen.

2. **Transparenz:** Die Entscheidungsprozesse der KI-Modelle müssen nachvollziehbar und transparent sein. Dies erfordert, dass die Modelle und die zugrunde liegenden Daten öffentlich zugänglich und überprüfbar sind.

3. **Verantwortlichkeit:** Entwickelnde und Anwendende von KI-Modellen müssen für die Ergebnisse und möglichen Fehlentscheidungen der Modelle verantwortlich gemacht werden können. Dies schließt eine sorgfältige Überwachung und regelmäßige Bewertung der Modelle ein.

4. **Schutz der Privatsphäre:** Die Nutzung von Patientendaten muss stets den höchsten Standards des Datenschutzes entsprechen. Dies umfasst die Einhaltung relevanter Datenschutzgesetze und die Sicherstellung, dass Daten anonymisiert und sicher gespeichert werden (sofern sie gespeichert werden).

3.3.2 Analyse der identifizierten Voreingenommenheiten

Auf Basis der in Kapitel 3.2.2 dargestellten Testfälle und Ergebnisse wird eine detaillierte, ethische Bewertung der altersspezifischen, geschlechtsspezifischen und genetischen Voreingenommenheiten durchgeführt. Diese Bewertung bezieht sich in erster Linie auf die Bevölkerung der Vereinigten Staaten von Amerika, da ein signifikanter Teil der Trainingsdaten des CORe-Modells (die MIMIC-III Datenbank) aus Daten besteht, die in einem einzigen amerikanischen Krankenhaus aufgenommen wurden. Um vergleichbare Ergebnisse und Aussagen zu treffen, insbesondere in Hinblick auf demografische Gegebenheiten, muss eine ähnliche demografische Struktur referenziert werden.

Bereits in den Rohdaten der MIMIC-III Datenbank finden sich Diskrepanzen, die zu Voreingenommenheit im darauf basierenden Modell führen können. Die verschiedenen Altersgruppen sind nicht repräsentativ und weichen teilweise um etwa 7% im Vergleich zur Realbevölkerung ab (Tabelle 1). Auch die Geschlechterrepresentation weicht um etwa 7% im Vergleich zur Realbevölkerung ab (Tabelle 2). In der Ethnizitätsverteilung finden sich ebenfalls große Unterschiede zwischen Trainings- und Realdaten. Während die weiße Bevölkerung in den Trainingsdaten im Vergleich zur Realbevölkerung von 2020 um etwa 13 % überrepräsentiert ist, sind besonders Lateinamerikaner:innen (um etwa 15%) und Asiat:innen (um etwa 4%) stark unterrepräsentiert (Tabelle 4).

Bei einer Gesamtbevölkerung von rund 335 Millionen Menschen in den USA (vgl. U.S. Census Bureau, 2020) können diese Faktoren eine Nicht- oder Falschrepräsentation von hunderttausenden oder gar Millionen Menschen bedeuten und weitreichende Folgen für weiterführende Prozesse und, schlussendlich, die medizinische Behandlung haben.

Altersspezifische Voreingenommenheit: ICD-Code 428 (Herzinsuffizienz) zeigt mit 18.246 diagnostizierten Fällen eine erwartbare und in etwa der Literatur entsprechenden Altersverteilungskurve (Abbildung 3). Dennoch zeigen sich keine Unterschiede der Diagnosevorhersage für verschiedene Altersgruppen im CORe-Modell (Abbildung 4). Grundsätzlich ist es möglich im jungen Alter an Herzinsuffizienz zu leiden, dennoch ist es sehr unwahrscheinlich (vgl. Bozkurt et al., 2023). Die eindeutige Diagnosevoraussage über alle Altersgruppen hinweg lässt sich vielleicht mit den Aufnahmenotizen (Anhang 4.1) erklären. Aus diesen Notizen lassen sich mehrfache, explizite Erwähnungen zu Herz- und Kreislauferkrankungen finden, was mit hoher Wahrscheinlichkeit zu diesem Ergebnis geführt oder zumindest dazu beigetragen hat.

Geschlechtsspezifische Voreingenommenheit: In dieser Kategorie zeigen beide Testfälle sehr wenige Diagnosen: ICD-Code 617 (Endometriose) wurde 36 Mal (0,007% der Diagnosen), ICD-Code 7330 (Osteoporose) 1675 Mal (0,3% der Diagnosen) diagnostiziert.

Endometriose wurde durch das CORe-Modell nicht diagnostiziert, dafür ICD-Code 614 (Entzündliche Erkrankung der Eierstöcke, Eileiter, Beckengewebe und des Bauchfells) (Abbildung 5). Diese beiden Erkrankungen sind sehr ähnlich, da sich das zusätzlich wachsende Gewebe der Endometriose an genau diesen Bereichen ansiedelt. Diese Diagnose wurde, abgesehen einer Anomalie, korrekterweise nur für weibliche Patientinnen oder Patient:innen ohne Geschlechtsangabe diagnostiziert und deckt sich daher mit den Erwartungen.

Osteoporose wurde für männliche und weibliche Patient:innen in etwa im gleichem Maße diagnostiziert (Abbildung 6). Da Osteoporose auch bei männlichen Patienten auftreten kann (vgl. National Institute of Arthritis and Muscoloskeletal and Skin Diseases, 2022), ist diese Vorhersage korrekt. Zusätzlich ist auch hier Osteoporose in der medizinischen Historie der Aufnahmenotizen (Anhang 4.3) aufgeführt. Interessanter wird es in diesem Fall, wenn die Geschlechtsangabe weggelassen wird, da in diesem Fall Osteoporose sehr viel seltener diagnostiziert wird. Dies ist besonders kritisch, da vorhergehend geschlechterunabhängig diagnostiziert wurde und die Angabe des Geschlechts daher keine Relevanz haben sollte.

Genetische Voreingenommenheit: Die Testfälle zu genetischer Voreingenommenheit zeigen stark unterschiedliche Diagnosezahlen; ICD-Code 250 (Diabetes) wurde 15359 Mal diagnostiziert, wohingegen ICD-Code 151 (Magenkrebs) mit 170 Diagnosen nur sehr selten diagnostiziert wurde (Abbildungen 7 und 8).

Diabetes wurde über alle ethnischen Gruppen hinweg gleichermaßen diagnostiziert (Abbildung 9). Während die Aufnahmenotizen für diesen Fall (Anhang 4.4) recht unvollständig sind, gibt es im Teil *Present Illness* eine explizite Feststellung dieser Erkrankung. Diabetes kann unabhängig von der Ethnizität der Patient:innen diagnostiziert werden,

daher ist dieses Vorhersageverhalten zu erwarten und korrekt.

Bei der Diagnose von Magenkrebs hatte das Modell Schwierigkeiten. Anstatt des ICD-Codes 151 wurde Code 150 (Speiseröhrenkrebs) diagnostiziert. Während diese Erkrankung zumindest physisch nah am Magen liegt, ist es dennoch nicht die korrekte Diagnose. Zudem wurden Lateinamerikaner:innen und Asiat:innen gänzlich von dieser Diagnose ausgenommen, während die Erkrankung bei anderen Ethnizitäten über alle Altersgruppen hinweg diagnostiziert wurde (Abbildung 10).

3.3.3 Bewertung der Testergebnisse

In diesem Kapitel sollen die im Kapitel 3.2.3 gesammelten und Kapitel 3.3.2 analysierten Testergebnisse unter Berücksichtigung der ethischen Prinzipien aus Kapitel 3.3.1 bewertet werden. Der Fokus bei der Bewertung liegt auf der Untersuchung altersspezifischer, geschlechtsspezifischer und genetischer Voreingenommenheit im CORe-Modell im besonderen Kontext der amerikanischen Bevölkerung und Gesundheitssystems.

Altersspezifische Voreingenommenheit

Die Testergebnisse zeigten keine signifikanten altersbezogenen Voreingenommenheiten, was darauf hinweist, dass das Modell eine konsistente diagnostische Leistung über alle Altersgruppen hinweg erbringt. Die Diagnose von Herzinsuffizienz scheint für jüngere Patient:innen zwar unwahrscheinlich, jedoch gibt es eindeutige Hinweise auf diese Erkrankung in den Aufnahmenotizen (siehe Anhang 4). Das Paper von van Aken et al. (2021) beschreibt die Verarbeitung von numerischen Werten im CORe Modell als ausbaufähig, besonders im Hinblick auf lebensbedrohliche Vitalwerte, was sich aber im Falle des hier beleuchteten Testfalles nicht eindeutig bestätigen lässt.

Das CORe-Modell zeigte hier keine eindeutig voreingenommenen Ergebnisse, sondern war eher übervorsichtig bei der Diagnose jüngerer Patient:innen.

Geschlechtsspezifische Voreingenommenheit

Bei der Analyse der Testergebnisse zeigt das Modell signifikante Unterschiede in der diagnostischen Genauigkeit zwischen den verschiedenen Geschlechtskategorien. Besonders auffällig ist, dass das Modell bei *undefiniertem* Geschlecht andere Diagnosen liefert als bei den explizit angegebenen Geschlechtern *männlich* oder *weiblich*. Dies deutet darauf hin, dass das Modell zumindest unter bestimmten Umständen explizite Angaben zum Geschlecht benötigt um eine Diagnose zu stellen, die geschlechtsunabhängig sein sollte. Die mangelnde Berücksichtigung geschlechtsspezifischer Unterschiede, bzw. die Anforderung der Angabe des Geschlechts, führt zu ungenauen Diagnosen für bestimmte Geschlechtsgruppen.

Es konnte eine klare geschlechtsspezifische Voreingenommenheit im CORe-Modell festgestellt werden, besonders im Fall von fehlenden Geschlechtsinformationen.

Genetische Voreingenommenheit

Besonders hervorzuheben ist die genetische Voreingenommenheit, da die diagnostischen Ergebnisse starke Unterschiede zwischen ethnischen Gruppen aufwiesen. Das Modell zeigte eine geringere Genauigkeit bei der Diagnose von Erkrankungen in ethnischen Minderheiten, was in erster Linie auf eine unzureichende Repräsentation dieser Gruppen in den Trainingsdaten zurückzuführen ist (siehe Tabelle 4). Dies stellt ein erhebliches ethisches Problem dar, da so die Weichen für eine ungleiche medizinische Versorgung bestimmter ethnischer Gruppen gestellt sind. Die Ursachen für diese genetische Voreingenommenheit sind vielfältig und können sowohl in der ungleichen Verfügbarkeit ethnischer Daten in der Datengrundlage als auch in der Komplexität genetischer Interaktionen liegen.

Bei der genetischen Voreingenommenheit konnten klare Verzerrungen festgestellt werden, denn ethnische Minderheiten erhielten bei gleichen Aufnahmenotizen andere Diagnosen.

Eine sorgfältige Berücksichtigung alters-, geschlechts- und ethnizitätsspezifischer Bedürfnisse ist entscheidend, um die Fairness, Genauigkeit und Wirksamkeit von medizinischen KI-Systemen zu gewährleisten und die bestmögliche Versorgung für Patient:innen aller demografischen Gruppen sicherzustellen. Die festgestellten systematischen Verzerrungen unterstreichen die Notwendigkeit, diese Unterschiede angemessen im CORe-Modell zu berücksichtigen.

Gesamtheitliche Bewertung

Die ethische Bewertung der Testergebnisse orientiert sich an den Prinzipien der Gerechtigkeit, Transparenz, Verantwortlichkeit und des Schutzes der Privatsphäre die in Kapitel 3.3.1 definiert wurden.

1. **Gerechtigkeit:** Die Identifizierung systematischer Verzerrungen im CORe-Modells zeigt, dass nicht alle Patient:innen unabhängig von Geschlecht, Alter und ethnischer Herkunft gleich behandelt werden. Dies widerspricht dem Prinzip der Gerechtigkeit. Es ist unerlässlich, Maßnahmen zu ergreifen, um sicherzustellen, dass alle demografischen Gruppen angemessen durch das Modell repräsentiert werden und dass die Modelle fair und unvoreingenommen handeln.

2. **Transparenz:** Die Entscheidungsprozesse der KI-Modelle müssen nachvollziehbar und transparent gestaltet sein. Aus Anwendersicht gleicht das CORe-Modell einer "Black Box", da es keinen klar ersichtlichen Weg zur Ergebnisfindung, bzw. keine Begründung der Ergebnisse, gibt. Die zugrundeliegenden Entscheidungsmechanismen sowie wichtige Faktoren für die Ergebnisfindung sollten den Diagnoseergebnissen mitgeliefert werden um Vertrauen und Akzeptanz bei den Nutzenden zu fördern. In weiteren Forschungen, wie im Modell "ProtoPatient"von van Aken, Papaioannou,

Naik et al. (2022), wird dieser Aspekt adressiert. ProtoPatient bietet eine verbesserte Erklärbarkeit durch Hervorheben von für die Diagnose relevanter Textstellen und ermöglicht es der Ärzteschaft und den Patient:innen, die Modellvorhersagen effektiver zu interpretieren und ihnen zu vertrauen.

3. **Verantwortlichkeit:** Entwickelnde und Anwendende von KI-Modellen müssen für die Ergebnisse und mögliche Fehlentscheidungen der Modelle verantwortlich gemacht werden können. Eine regelmäßige Überwachung und Bewertung der Modelle ist unerlässlich, um sicherzustellen, dass keine systematischen Verzerrungen vorliegen. Auch wenn das CORe-Modell kein Produktivsystem ist, ist es wichtig, klare Verantwortlichkeiten und Haftungsregelungen zu definieren, um die Qualität und Sicherheit der medizinischen Versorgung zu gewährleisten. Dies ist besonders relevant, da das CORe-Modell möglicherweise zur weiteren Forschung genutzt wird. Existierende Voreingenommenheiten werden somit möglicherweise an folgende bzw. darauf aufbauende Modelle vererbt.

4. **Schutz der Privatsphäre:** Der Schutz von Patientendaten muss stets den höchsten Standards entsprechen. Dies umfasst die Einhaltung relevanter Datenschutzgesetze und die Sicherstellung der Anonymisierung und sicheren Speicherung der Daten. Da das CORe-Modell mit sehr sensiblen Daten arbeitet aber auch mit anonymisierten Daten funktioniert und bereits mit solchen trainiert wurde, wären automatisierte Funktionen zur Anonymisierung der eingegebenen Daten lohnenswert. Automatische Funktionen zur anonymisierten Verarbeitung von Alter, Geschlecht, Ethnizität sowie weiteren schützenswerten Faktoren können sicherstellen, dass Patientendaten stets geschützt bleiben.

3.4 Ableitung von Leitlinien

Die Ableitung von Leitlinien basiert auf den in den vorherigen Kapiteln herausgearbeiteten Voreingenommenheiten und den ethischen Prinzipien aus Kapitel 3.3.1. Ziel ist es, Maßnahmen zur Minimierung von Voreingenommenheit im CORe-Modell und vergleichbaren KI-Systemen zu formulieren. Die nachstehenden Leitlinien sind darauf ausgelegt, die Fairness, Transparenz, Verantwortlichkeit und den Schutz der Privatsphäre sicherzustellen und somit die Qualität der medizinischen Versorgung zu verbessern.

1. **Fairness und Repräsentation:**
   Um sicherzustellen, dass alle demografischen Gruppen gleich behandelt werden, müssen KI-Modelle auf repräsentativen und ausgewogenen Datensätzen trainiert werden. Dies erfordert eine sorgfältige Sammlung und Kuratierung von Daten, die verschiedene Altersgruppen, Geschlechter und ethnische Hintergründe einschließen.

Zusätzlich sollten Modelle regelmäßig überprüft und getestet werden, um potenzielle Voreingenommenheiten frühzeitig zu erkennen und zu beheben.

2. **Transparenz der Entscheidungsprozesse:**
   Die Entscheidungsmechanismen von KI-Modellen müssen transparent gestaltet sein, um das Vertrauen der Anwendenden zu gewinnen. Dies bedeutet, dass sowohl die verwendeten Daten als auch die Algorithmen und deren Entscheidungen nachvollziehbar und überprüfbar sein müssen. Die Implementierung von erklärbaren Modellen und der Verzicht auf "Black-Box"-Modelle kann dabei helfen, die Entscheidungsprozesse verständlicher zu machen. Es ist wichtig, dass Diagnosen und Therapieempfehlungen mit verständlichen Begründungen zur Ergebnisfindung versehen werden, um die Nutzung solcher Technologien zu erleichtern.

3. **Verantwortlichkeit und Überwachung:**
   Entwickelnde und Anwendende von KI-Modellen müssen für die Ergebnisse und potenziellen Fehlentscheidungen ihrer Systeme verantwortlich gemacht werden können. Dies erfordert klare Verantwortlichkeits- und Haftungsregelungen. Es müssen regelmäßige Revisionen und Evaluierungen der Modelle durchgeführt werden, um deren Leistung und Fairness zu überwachen. Bei Feststellung systematischer Verzerrungen müssen umgehend Maßnahmen zur Korrektur eingeleitet werden.

4. **Schutz der Privatsphäre:**
   Der Schutz der Patientendaten muss oberste Priorität haben. Dies beinhaltet die Einhaltung geltender Datenschutzgesetze und die Gewährleistung der Anonymität der Daten. KI-Modelle sollten nur mit anonymisierten Daten arbeiten, und es müssen strenge Sicherheitsmaßnahmen implementiert werden, um unbefugten Zugriff zu verhindern. Automatisierte Funktionen zur Anonymisierung der Daten können dabei unterstützend wirken.

5. **Interdisziplinäre Zusammenarbeit:**
   Die Entwicklung und Implementierung von KI-Systemen sollte in interdisziplinärer Zusammenarbeit erfolgen. Fachleute aus den Bereichen Ethik, Medizin, Informatik und Recht müssen von Beginn an in den Entwicklungsprozess integriert werden. Dies gewährleistet, dass ethische und regulatorische Überlegungen sowie medizinische Fachkenntnisse angemessen berücksichtigt werden und die entwickelten Systeme den hohen Anforderungen der klinischen Praxis entsprechen.

6. **Bildung und Sensibilisierung:**
   Es ist wichtig, das Bewusstsein für die ethischen Herausforderungen und möglichen Voreingenommenheiten in KI-Systemen zu schärfen. Schulungen und Fortbildungen für Entwickelnde und Anwendende können dazu beitragen, die Sensibilität für diese

Themen zu erhöhen und eine verantwortungsbewusste Entwicklung und Nutzung von KI in der Medizin zu fördern.

3.5 Zusammenfassung Kapitel 3

Kapitel 3 beginnt mit der Analyse der geschlechts-, alters- und ethnizitätsspezifischen Voreingenommenheit im CORe-Modell. Die Analyse umfasst die Durchführung spezifischer Testfälle, die aus der MIMIC-III Datenbank extrahiert wurden, mit dem CORe-Modell.

Die Ergebnisse dieser Analyse zeigen keine erkennbare altersspezifische Voreingenommenheit im CORe-Modell. Dies kann entweder auf eine korrekte Implementation des Modells, die unzureichende Berücksichtigung altersspezifischer Merkmale während des Trainingsprozesses oder die limitierte Fähigkeit des Modells, Zahlen korrekt zu interpretieren zurückzuführen sein. Bei der Untersuchung geschlechtsspezifischer Voreingenommenheiten zeigte das Modell jedoch erste Indizien für Voreingenommenheit. Insbesondere wurden andere Diagnosen gestellt, wenn die Geschlechterangabe entfernt wurde, was auf eine geschlechtsspezifische Voreingenommenheit hinweist. In der genetischen Analyse zeigten sich ebenso signifikante Unterschiede in den Diagnosegenauigkeiten zwischen verschiedenen ethnischen Gruppen. Besonders betroffen waren ethnische Minderheiten, was auf eine ethnizitätsbasierte Voreingenommenheit hinweist. Insgesamt belegen die Unterschiede in den Diagnosegenauigkeiten, dass das Modell systematische Verzerrungen aufweist, die zu ungleichen Diagnoseergebnissen führen.

Abschließend wurden die Ergebnisse bewertet und, darauf basierend, Leitlinien entwickelt um Voreingenommenheit in zukünftigen KI-Anwendungen zu minimieren. Diese Leitlinien basieren auf den Prinzipien der Gerechtigkeit, Transparenz, Verantwortlichkeit und des Schutzes der Privatsphäre.

4 Diskussion und Schlussfolgerung

In diesem Kapitel werden die Ergebnisse der Untersuchung noch einmal in aller Kürze zusammengefasst, die identifizierten Voreingenommenheitsfaktoren und deren potenzielle Auswirkungen auf die klinische Praxis diskutiert sowie ein Ausblick auf zukünftige Entwicklungen und Forschungsrichtungen gegeben. Abschließend wird eine Zusammenfassung der wichtigsten Erkenntnisse präsentiert.

4.1 Präsentation der Testergebnisse

Die Analyse des CORe-Modells zeigte signifikante Unterschiede in der diagnostischen Genauigkeit zwischen verschiedenen demografischen Gruppen. Die Ergebnisse der durchgeführten Tests legen nahe, dass geschlechts- und ethnizitätsspezifische Voreingenommenheiten bestehen, die zu ungleichen Behandlungsergebnissen führen können.

Altersspezifische Ergebnisse

Die Altersanalyse ergab keine expliziten Diagnoseunterschiede für Patienten in unterschiedlichen Altersgruppen. Zwar ist die Verarbeitung numerischer Werte keine Stärke des CORe-Modells, jedoch konnten keine systematischen Verzerrungen bezüglich des Alters der Patient:innen in dieser Analyse festgestellt werden.

Geschlechtsspezifische Ergebnisse

Bei der Untersuchung geschlechtsspezifischer Voreingenommenheiten zeigte das CORe-Modell unterschiedliche Diagnosegenauigkeiten für verschiedene Geschlechter. Besonders auffällig waren Diagnoseunterschiede bei Entfernen der Geschlechtsangabe. Diese Unterschiede zeigen eine klare geschlechtsspezifische Voreingenommenheit.

Ethnizitätsspezifische Ergebnisse

Die ethnizitätsspezifische Analyse zeigte, dass das CORe-Modell unterschiedliche Diagnosegenauigkeiten für verschiedene ethnische Gruppen aufweist. Besonders betroffen waren ethnische Minderheiten. Auch hier konnte eine klare Voreingenommenheit herausgestellt werden.

4.2 Diskussion von Voreingenommenheitsfaktoren und möglichen Auswirkungen

Die identifizierten Voreingenommenheitsfaktoren und deren potenzielle Auswirkungen auf die klinische Praxis werden im Folgenden detailliert diskutiert.

Es konnten keine altersspezifischen Voreingenommenheiten im CORe Modell nachgewiesen werden. Im Gegenteil gab das Modell bei der Diagnose eher falsch-positive Diagnosen als falsch-negative, was dem Anspruch an medizinische Geräte laut MDR[15] entspricht. Laut MDR müssen die Risiken der Nutzung medizinischer Geräte verhältnismäßig zum Nutzen sein (vgl. Morgenthaler, 2022). Fraglich ist, ob das Modell diese Diagnosen aufgrund von tatsächlicher Übervorsichtigkeit (Überdiagnose) gestellt hat, oder ob die fehlende Fähigkeit zur korrekten Interpretation von Zahlen (vgl. van Aken et al., 2021) ausschlaggebend war. In jedem Fall kann dieses Verhalten des Modells dazu führen, dass potenziell kritische Zustände häufiger erkannt und behandelt werden. Auf der anderen Seite kann eine übervorsichtige Interpretation auch zu einem erhöhten Aufwand und ineffizienten Nutzung von Ressourcen durch irrelevante diagnostische Tests und Behandlungen führen, was sowohl die Patient:innenbelastung als auch die Kosten für das Gesundheitssystem erhöht. Die korrekte Diagnosefindung kann hierdurch ebenfalls verzögert werden. Eine dauerhafte Überdiagnose kann ebenso das Vertrauensverhältnis der klinischen Fachkräfte in das Modell schwächen, da sie das Modell als weniger unterstützend betrachten und häufiger manuell überprüfen müssen. Auch die Patientenautonomie und die informierte Zustimmung können durch übermäßigen Informationsfluss unter eine Überdiagnose leiden.

Bei den geschlechtsspezifischen Testfällen konnten klare Voreingenommenheiten herausgestellt werden. Es gab eindeutige Unterschiede bei den Diagnoseergebnissen wenn das Geschlecht entweder angegeben oder entfernt wurde. Diese Unterschiede können leicht zu einer ungleichen und unzureichenden Behandlungsqualität führen, beispielsweise wenn die Geschlechtsangaben unter hektischen Bedingungen nicht aufgenommen oder dem Modell zugeführt werden. Die selben Effekte können zu einer Benachteiligung von nicht-binären und geschlechtsdiversen Patienten führen, sofern diese Informationen aufgenommen werden. Das ärztliche Personal muss möglicherweise zusätzliche manuelle Untersuchungen und Tests durchführen, um sicherzustellen, dass Patient:innen unabhängig von ihrem Geschlecht eine angemessene Diagnose und Behandlung erhalten. Dies erhöht den Arbeitsaufwand und stellt den Nutzen einer solchen Technologie in Frage. Zudem können hierdurch, ebenso wie im vorigen Abschnitt beschrieben, erhöhte Kosten und Verzögerungen bei der Patientenversorgung entstehen. Auch diese Art der Voreingenommenheit kann zu einem Vertrauensverlust in die korrekte medizinische Behandlung spezieller Patientengruppen führen.

[15] Medical Device Regulation

Die wohl stärksten Unterschiede in der diagnostischen Fähigkeit des CORe-Modells gab es zwischen ethnischen Gruppen. Ethnische Minderheiten bekamen bei den selben Aufnahmenotizen andere Diagnosen, was eine klare Voreingenommenheit herausstellt. Während diese Gruppen in den Trainingsdaten der MIMIC-III Datenbank zwar unterrepräsentiert waren, gab es Informationen zur ethnischen Angehörigkeit. Auch im amerikanischen *Census*, eine Volkszählung die alle 10 Jahre durchgeführt wird, gibt es Statistiken zu ethnischen Gruppen innerhalb der Bevölkerung. In deutschen Krankenhäusern werden ethnische Daten nach § 21, KHEntgG grundsätzlich nicht erhoben, selbst eine ethnische Verteilung für die gesamte deutsche Bevölkerung ist schwer zu finden, da diese Daten nicht Teil der Volkszählungen und anderer offizieller Statistiken sind. Wie in dieser Arbeit herausgestellt wurde, können ethnische Daten aber nicht nur relevant für die Entwicklung und Implementierung von KI(-gestützten) Modellen, sondern auch für medizinische Diagnosefindung und Behandlung sein.

4.3 Limitationen dieser Arbeit

Die vorliegende Arbeit weist mehrere Limitationen auf, die bei der Interpretation der Ergebnisse berücksichtigt werden sollten.

Die MIMIC-III-Datenbank, die als Hauptquelle für die Testfälle verwendet wurde, ist nicht vollständig repräsentativ für die gesamte Bevölkerung. Dies liegt daran, dass die Daten überwiegend aus einem einzigen Krankenhaus in den USA stammen und somit möglicherweise nicht die Vielfalt und Unterschiede in anderen geografischen Regionen oder Gesundheitssystemen widerspiegeln. Diese Begrenzung kann zu einer Verzerrung der Ergebnisse führen, insbesondere bei der Analyse von Voreingenommenheiten gegenüber verschiedenen demografischen Gruppen.

Zudem birgt die Nutzung von ethnischen Daten zur Analyse von Voreingenommenheiten in den Modellen ethische und regulatorische Herausforderungen. In vielen Ländern, einschließlich Deutschland, werden ethnische Daten, wie bereits in Kapitel 4.2 erwähnt, in Krankenhäusern nicht systematisch erfasst. Dies erschwert die Identifikation und Korrektur von ethnizitätsspezifischen Voreingenommenheiten in den Modellen und stellt eine bedeutende Limitation dieser Arbeit dar.

Ein weiterer wesentlicher Aspekt, der die Aussagekraft der vorliegenden Arbeit einschränkt, ist die Tatsache, dass pro ICD-Code lediglich ein einzelner Testfall durchgeführt und analysiert wurde. Ebenso wurde für die Überprüfung auf altersspezifische Voreingenommenheit lediglich ein Testfall umgesetzt, wohingegen für ethnizitäts- und geschlechtsspezifische Voreingenommenheit zwei Testfälle genutzt wurden. Diese Vorge-

hensweise könnte dazu führen, dass die Ergebnisse nicht ausreichend repräsentativ sind.

Auch die im CORe-Modell verwendeten Algorithmen und Techniken führen zu Beschränkungen der Wertbarkeit. Beispielsweise werden vom CORe-Modell nur 512 Tokens für die Testfälle akzeptiert, während die Aufnahmenotizen teils weit umfangreicher sind (siehe Kapitel 4.5). Die Generalisierbarkeit der Ergebnisse auf andere Modelle und Anwendungen ist daher eingeschränkt. Darüber hinaus könnte die Modellleistung durch die spezifischen Implementierungen und Parametrierungen beeinflusst worden sein, was die Übertragbarkeit der Ergebnisse auf andere Kontexte ebenfalls einschränkt.

Letztlich besteht die Möglichkeit einer unbeabsichtigten Voreingenommenheit des Forschenden. Die Auswahl der Testfälle, die Interpretation der Ergebnisse und die getroffenen Annahmen können von subjektiven Faktoren beeinflusst sein. Um diese Limitation zu minimieren, wurde versucht, eine möglichst objektive und standardisierte Vorgehensweise zu gewährleisten. Dennoch kann eine vollständige Eliminierung von Forschenden-Voreingenommenheit nicht garantiert werden.

4.4 Ausblick auf zukünftige Entwicklungen und Forschungsrichtungen

Dieses Kapitel bietet einen detaillierten Ausblick auf mögliche zukünftige Entwicklungen und Forschungsrichtungen im Bereich der Künstlichen Intelligenz in der medizinischen Diagnostik.

- **Verbesserte Datenrepräsentation:** Eine der zentralen Herausforderungen bei der Entwicklung fairer KI-Modelle ist die Sicherstellung einer ausgewogenen und repräsentativen Datengrundlage. Zukünftige Forschungen sollten sich darauf konzentrieren, Strategien zu entwickeln, die eine gleichmäßige Verteilung der Daten gewährleisten. Dazu gehört die gezielte Sammlung und Integration von Daten aus unterrepräsentierten Gruppen, um die Vielfalt in den Trainingsdatensätzen zu erhöhen. Darüber hinaus könnten synthetische Datengenerierungsverfahren eingesetzt werden, um vorhandene Datenlücken zu schließen und die Datenbasis zu erweitern.

- **Ausarbeitung der Testfälle:** In der vorliegenden Arbeit wurde pro ICD-Code lediglich ein Testfall untersucht. Zukünftige Forschungen könnten darauf abzielen, die Repräsentativität dieser Testfälle zu erhöhen, indem beispielsweise mehrere Aufnahmenotizen pro ICD-Code analysiert werden. Durch diese erweiterte Analyse könnten gemittelte Diagnosewahrscheinlichkeiten ermittelt werden, die eine noch präzisere Untersuchung von Voreingenommenheitsfaktoren ermöglichen .

- **Fortschrittliche Modellierungstechniken:** Der weitere Einsatz fortschrittlicher Modellierungstechniken wie selbstüberwachtes Lernen und Transferlernen kann weiter dazu beitragen, Voreingenommenheit in KI-Modellen zu reduzieren. Selbstüberwachtes Lernen ermöglicht es Modellen, aus großen Mengen unstrukturierter Daten zu lernen, ohne dass eine umfangreiche manuelle Annotation erforderlich ist. Transferlernen hingegen ermöglicht es, vortrainierte Modelle auf neue Aufgaben anzupassen, wodurch die Notwendigkeit großer spezialisierter Datensätze reduziert wird. Diese Techniken können die Robustheit und Generalisierbarkeit von Modellen verbessern und dazu beitragen, Voreingenommenheit zu minimieren.

- **Interdisziplinäre Ansätze:** Die Entwicklung umfassender und gerechter KI-Systeme erfordert eine enge Zusammenarbeit zwischen Techniker:innen, Mediziner:innen und Ethiker:innen. Interdisziplinäre Forschungsansätze können sicherstellen, dass ethische Überlegungen und medizinische Fachkenntnisse von Anfang an in den Entwicklungsprozess integriert werden. Dies kann durch die Einrichtung von interdisziplinären Arbeitsgruppen und Forschungskonsortien gefördert werden, die gemeinsam an der Entwicklung und Evaluierung von KI-Modellen arbeiten.

- **Regulierung und Standardisierung:** Regulierungsbehörden und Gesundheitsorganisationen müssen weitere Richtlinien und Standards entwickeln, um die Ethik und Fairness von KI-Modellen in der Medizin zu gewährleisten. Dies umfasst auch die Transparenz der Modelle und die Nachvollziehbarkeit ihrer Entscheidungen, um Vertrauen und Akzeptanz bei den Nutzenden zu fördern. Eine mögliche Forschungsrichtung könnte die Entwicklung von Rahmenwerken und Tools zur Bewertung und Zertifizierung von KI-Systemen hinsichtlich ihrer Fairness und ethischen Vertretbarkeit sein.

Die fortschreitende Integration von KI in die medizinische Diagnostik bietet größtes Potenzial zur Verbesserung der Patientenversorgung. Durch die kontinuierliche Weiterentwicklung und Anwendung der zuvor genannten Strategien können zukünftige KI-Modelle noch präziser, fairer und ethisch vertretbarer werden. Die Forschung sollte sich daher darauf konzentrieren, innovative Lösungen zu entwickeln und bestehende Herausforderungen zu adressieren, um die Qualität und Fairness der medizinischen Versorgung durch KI nachhaltig zu verbessern.

4.5 Zusammenfassung Kapitel 4

In Kapitel 4 wurden die Testergebnisse, die Diskussion der Voreingenommenheitsfaktoren und die Limitationen der Arbeit behandelt. Die Analyse zeigte, dass geschlechts- und ethnizitätsspezifische Verzerrungen im CORe-Modell existieren, die die diagnostische Genauigkeit und Fairness beeinträchtigen können.

Trotz der Limitationen liefert die Arbeit wertvolle Einblicke und Empfehlungen zur Reduktion von Voreingenommenheiten in KI-Diagnosemodellen. Zukünftige Forschung sollte diese Limitationen adressieren, um die Robustheit und Fairness von KI-Systemen zu verbessern.

Literaturverzeichnis

[Amini, M. M., Jesus, M., Sheikholeslami, D. F., Alves, P., Benam, A. H., & Hariri, F.]. (2023). Artificial Intelligence Ethics and Challenges in Healthcare Applications: A Comprehensive Review in the Context of the European GDPR Mandate. *MDPI*, *5*(3), 1023–1035. https://doi.org/10.3390/make5030053

[Bozkurt, B., Ahmad, T., Alexander, K. M., Baker, W. L., Bosak, K., Breathett, K., Fonarow, G. C., Heidenreich, P., Ho, J. E., Hsich, E., Ibrahim, N. E., Jones, L. M., Khan, S. S., Khazanie, P., Koelling, T., Krumholz, H. M., Kush, K. K., Lee, C., Morris, A. A., ... Ziaeian, B.]. (2023). Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America. *J Card Fail*, *29*(10), 1412–1451. https://doi.org/10.1016/j.cardfail.2023.07.006

[Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., Schilling, B., Haferkamp, S., Schadendorf, D., Holland-Letz, T., Utikal, J. S., & von Kalle, C.]. (2019). Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *EJC*, *113*, 47–54. https://doi.org/10.1016/j.ejca.2019.04.001

[Celi, L. A., Cellini, J., Charpignon, M.-L., Dee, E. C., Dernoncourt, F., Eber, R., Mitchell, W. G., Moukheiber, L., Schirmer, J., Situ, J., Paguio, J., Park, J., Wawira, J. G., & Yao, S.]. (2022). Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *Plos Digital Health*, *1*(3). https://doi.org/10.1371/journal.pdig.0000022

[Chung, H., Park, C., Kang, W. S., & Lee, J.]. (2021). Gender Bias in Artificial Intelligence: Severity Prediction at an Early Stage of COVID-19. *Front Physiol.*, *12*. https://doi.org/10.3389/fphys.2021.778720

[Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.]. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. https://doi.org/10.48550/arXiv.1810.04805

[Elmahdy, M., & Sebro, R.]. (2023). Sex, ethnicity, and race data are often unreported in artificial intelligence and machine learning studies in medicine. *Intelligence-Based Medicine*, *8*. https://doi.org/10.1016/j.ibmed.2023.100113

[Gala, D., Behl, H., Shah, M., & Makaryus, A. N.]. (2024). The Role of Artificial Intelligence in Improving Patient Outcomes and Future of Healthcare Delivery in Cardiology: A Narrative Review of the Literature. *PubMed*, *12*(4), 481. https://doi.org/10.3390/healthcare12040481

[He, J., Baxter, S. L., Xu, J., Xu, J., Thou, X., & Zhang, K.]. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine*, *25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0

[Javid, A. M.]. (2021). *Neural Network Architecture Design: Towards Low-complexity and Scalable Solutions* (Diss.). KTH Royal Institute of Technology. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1524368&dswid=-8825

[Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G.]. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, *3*(160035). https://doi.org/10.1038/sdata.2016.35

[Kaylor, A.]. (2023). *Leveraging Predictive Analytics for Effective Disease Management*. Verfügbar 2. Juni 2024 unter https://lifesciencesintelligence.com/features/leveraging-predictive-analytics-for-effective-disease-management

[Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J.]. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[Mayo Clinic]. (2023). *Endometriosis*. Verfügbar 25. Juni 2024 unter https://www.mayoclinic.org/diseases-conditions/endometriosis/symptoms-causes/syc-20354656

[MIT Laboratory for Computational Physiology]. (2021). *Medical Information Mart for Intensive Care*. Verfügbar 1. Juli 2024 unter https://mimic.mit.edu/docs/gettingstarted/

[Morgenthaler, D.]. (2022). *Risk Management for Medical Devices under EU MDR and ISO 14971*. Verfügbar 10. Juli 2024 unter https://decomplix.com/risk-management-medical-devices-eu-mdr-iso-14971/#medical-device-risk-management-requirements-of-the-mdr

[Muralidharan, V., Burgart, A., Daneshjou, R., & Rose, S.]. (2023). Recommendations for the use of pediatric data in artificial intelligence and machine learning ACCEPT-AI. *npj Digit. Med.*, *6*(166). https://doi.org/10.1038/s41746-023-00898-5

[National Cancer Institute]. (2023). *Stomach Cancer Causes and Risk Factors*. Verfügbar 29. Juni 2024 unter https://www.cancer.gov/types/stomach/causes-risk-factors

[National Cancer Institute]. (n. d.). *What Is Stomach Cancer?* Verfügbar 26. Mai 2024 unter https://www.cancer.gov/types/stomach

[National Institute of Arthritis and Muscoloskeletal and Skin Diseases]. (2022). *Osteoporosis*. Verfügbar 25. Juni 2024 unter https://www.niams.nih.gov/health-topics/osteoporosis

[Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P.]. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *Plos Digital Health*, *2*(6). https://doi.org/10.1371/journal.pdig.0000278

[Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & G.Buchmann, T.]. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis

in the ICU. *Critical Care Medicine*, *46*(4), 547–553. https://doi.org/10.1097/ccm.0000000000002936

[NYU]. (2024). *GLUE Leaderboard*. Verfügbar 9. Juli 2024 unter https://gluebenchmark.com/leaderboard

[Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S.]. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453. https://doi.org/10.1126/science.aax2342

[O'Shea, K., & Nash, R.]. (2015). An Introduction to Convolutional Neural Networks. *arXiv*. https://doi.org/10.48550/arXiv.1511.08458

[Petersson, L., Larsson, I., Nygren, J. M., Nilsen, P., Neher, M., Reed, J. E., Tyskbo, D., & Svedberg, P.]. (2022). Challenges to implementing artificial intelligence in healthcare: a qualitative interview study with healthcare leaders in Sweden. *BMC*, *22*(850). https://doi.org/10.1186/s12913-022-08215-8

[Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q., Litsch, K., ... Dean, J.]. (2018). Scalable and accurate deep learning with electronic health records. *npj Digit. Med.*, *1*(18). https://doi.org/10.1038/s41746-018-0029-1

[Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Ball, R. L., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y.]. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *Stanford University*. https://doi.org/10.48550/arXiv.1711.05225

[Rigby, M. J.]. (2019). Ethical Dimensions of Using Artificial Intelligence in Health Care. *AMA Journal of Ethics*, *21*(2), E121–124. https://doi.org/10.1001/AMAJETHICS.2019.121

[Rudin, C.]. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.pdf. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.48550/arXiv.1811.10154

[Sherstinsky, A.]. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Elsevier "Physica D: Nonlinear Phenomena"*, *404*. https://doi.org/10.48550/arXiv.1808.03314

[Strauer, B.]. (2007). Herzinsuffizienz. *Internist*, *48*, 897–898. https://doi.org/10.1007/s00108-007-1923-9

[Symeonidis, P., Chairistanidis, S., & Zanker, M.]. (2022). Safe, effective and explainable drug recommendation based on medical data integration. *User Model User-Adap Inter*, *32*, 999–1018. https://doi.org/10.1007/s11257-022-09342-x

[U.S. Census Bureau]. (2020). *QuickFacts United States*. Verfügbar 25. Juni 2024 unter https://www.census.gov/quickfacts/fact/table/US/PST045221

[U.S. Census Bureau]. (2023). *Exploring Age Groups in the 2020 Census*. Verfügbar 25. Juni 2024 unter https://www.census.gov/library/visualizations/interactive/exploring-age-groups-in-the-2020-census.html

[U.S. Census Bureau]. (n. d.). *American Community Survey - Age and Sex*. Verfügbar 27. Juni 2024 unter https://data.census.gov/table/ACSST1Y2010.S0101?q=age%20 2010

[U.S. Center for Disease Control and Prevention]. (2024a). *Diabetes Basics*. Verfügbar 25. Juni 2024 unter https://www.cdc.gov/diabetes/about/

[U.S. Center for Disease Control and Prevention]. (2024b). *National Diabetes Statistics Report*. Verfügbar 25. Juni 2024 unter https://www.cdc.gov/diabetes/php/data-research/index.html

[van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F. A., & Löser, A.]. (2021). Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration. *arXiv*. https://doi.org/10.48550/arXiv.2102.04110

[van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F. A., & Löser, A.]. (2022a). *Clinical Outcome Prediction from Admission Notes*. Verfügbar 29. Juni 2024 unter https://github.com/bvanaken/clinical-outcome-prediction

[van Aken, B., Papaioannou, J.-M., Mayrdorfer, M., Budde, K., Gers, F. A., & Löser, A.]. (2022b). *CORe Model - Clinical Diagnosis Prediction*. Verfügbar 29. Juni 2024 unter https://huggingface.co/DATEXIS/CORe-clinical-diagnosis-prediction

[van Aken, B., Papaioannou, J.-M., Naik, M. G., Eleftheriadis, G., Nejdl, W., Gers, F. A., & Löser, A.]. (2022). This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text. *arXiv*. https://doi.org/10.48550/arXiv.2210.08500

[Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.]. (2023). Attention Is All You Need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

[Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R.]. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv*. https://doi.org/10.48550/arXiv.1804.07461

[World Health Organization and International Conference for the Ninth Revision of the International Classification of Diseases]. (1977). Manual of the international statistical classification of diseases, injuries, and causes of death : based on the recommendations of the ninth revision conference, 1975, and adopted by the Twenty-ninth World Health Assembly (1975 revision).

[Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J.]. (2016).

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv*. https://doi.org/10.48550/arXiv.1609.08144

Anhang

data_extraction.py

```
import pandas as pd
import utilities

# Relevante Schlüssel
admission_time = 'ADMITTIME'                # Zeit zu der Patient ins Krankenhaus aufgenommen wurde
discharge_time = 'DISCHTIME'                # Entlassungszeit
length_of_stay = 'LENGTH_OF_STAY'           # Aufenthaltslänge
date_of_birth = 'DOB'                       # Geburtsdatum
age_at_admission = 'AGE_AT_ADMISSION'       # Alter bei Aufnahme
gender = 'GENDER'                           # Geschlecht
ethnicity = 'ETHNICITY'                     # Enthnizität
died_in_hospital = 'EXPIRE_FLAG'            # Bei Krankenhausaufenthalt gestorben. 1 für ja, 0 für nein
diagnosis = 'DIAGNOSIS'                     # Diagnose (in Worten)
icd_code = 'ICD9_CODE'                      # Diagnose (ICD9 Code)
admission_note = 'TEXT'                     # Text der Aufnahmenotizen
short_codes = 'SHORT_CODES'                 # Eine Liste an ICD shortcodes

# Laden der relevanten Dateien
patients_df = pd.read_csv('mimic3/PATIENTS.csv.gz', compression='gzip')
admissions_df = pd.read_csv('mimic3/ADMISSIONS.csv.gz', compression='gzip')
diagnoses_icd_df = pd.read_csv('mimic3/DIAGNOSES_ICD.csv.gz', compression='gzip')
# Dateien Extrahiert durch https://github.com/bvanaken/clinical-outcome-prediction/blob/master/tasks/dia/dia.py
# Verwandelt Discharge notes in Admission Notes, indem es zum Aufnahmezeitpunkt unbekannte Daten weglässt
diagnoses_grouped_test = pd.read_csv('mimic3/clinical-prediction-extracted/DIA_GROUPS_3_DIGITS_adm_test.csv')
diagnoses_grouped_train = pd.read_csv('mimic3/clinical-prediction-extracted/DIA_GROUPS_3_DIGITS_adm_train.csv')
diagnoses_grouped_val = pd.read_csv('mimic3/clinical-prediction-extracted/DIA_GROUPS_3_DIGITS_adm_val.csv')
# Verknüpfen der Test, Train und Validation-Dateien
admission_notes = pd.concat([diagnoses_grouped_test, diagnoses_grouped_train], ignore_index=True)
admission_notes = pd.concat([admission_notes, diagnoses_grouped_val], ignore_index=True)

# Verknüpfen von Patienten- und Aufnahmeinformationen
patient_admissions = pd.merge(patients_df, admissions_df, on='SUBJECT_ID', how='inner')

# Verknüpfen von Aufnahmen und Diagnosen
admission_diagnoses = pd.merge(patient_admissions, diagnoses_icd_df, on='HADM_ID', how='inner')

# Verknüpfen von Aufnahmen, Diagnosen und Aufnahmenotizen
admission_notes_diagnoses = pd.merge(admission_diagnoses, admission_notes, on='HADM_ID', how='inner')

# Berechnung des Alters bei der Aufnahme
admission_notes_diagnoses[age_at_admission] = (admission_notes_diagnoses[admission_time]
                                               .apply(lambda x: pd.Timestamp(x).year) -
                                               admission_notes_diagnoses[date_of_birth]
                                               .apply(lambda x: pd.Timestamp(x).year))

# Berechnung der Aufenthaltslänge
admission_notes_diagnoses[length_of_stay] = (pd.to_datetime(admission_notes_diagnoses[discharge_time]) -
                                             pd.to_datetime(admission_notes_diagnoses[admission_time])).dt.days

# Auswahl der gewünschten Spalten zum Speichern
selected_columns = [
    'HADM_ID', length_of_stay, age_at_admission, gender, ethnicity,
    died_in_hospital, diagnosis, icd_code, admission_note, short_codes
]
final_df = admission_notes_diagnoses[selected_columns]
```

```
# Entfernen aller Zeilen mit unpassenden Altersangaben
age_filtered_df = final_df[(final_df[age_at_admission] >= 16) & (final_df[age_at_admission] <= 100)]

# Speichern der verarbeiteten Daten in neue CSV-Datei
age_filtered_df.to_csv('final_admission_diagnoses.csv', index=False)

print(utilities.total_count_age_group(age_filtered_df, 18, 24))
print(utilities.total_count_age_group(age_filtered_df, 25, 34))
print(utilities.total_count_age_group(age_filtered_df, 35, 44))
print(utilities.total_count_age_group(age_filtered_df, 45, 64))
print(utilities.total_count_age_group(age_filtered_df, 65, 84))
print(utilities.total_count_age_group(age_filtered_df, 85, 99))

# Erstellung von Plots
# Allgemeine Altersverteilung
overall_age_distribution = utilities.extract_age_count(age_filtered_df)
utilities.plot_simple_graph(overall_age_distribution, age_at_admission, 'COUNT', 'Age at admission',
                            'Count', 'Overall age at admission distribution',
                            'overall_'+age_at_admission.lower(), False)

# Allgemeine Genderverteilung
overall_gender_distribution = utilities.extract_gender_count(age_filtered_df)
utilities.plot_simple_graph(overall_gender_distribution, gender, 'COUNT', 'Gender',
                            'COUNT', 'Overall gender distribution', 'overall_'+gender.lower(), False)

# Allgemeine Ethnizitätsverteilung
overall_ethnicity_distribution = utilities.extract_ethnicity_count(age_filtered_df)
utilities.plot_simple_graph(overall_ethnicity_distribution, ethnicity, 'COUNT', 'Ethnicity',
                            'COUNT', 'Overall ethnicity distribution', 'overall_'+ethnicity.lower(),
                            False)
```

utilities.py

```
import matplotlib.pyplot as plt
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import re
import seaborn as sns
import pandas as pd


def shorten_labels(labels, max_length):
    return [label if len(label) <= max_length else label[:max_length] + '...' for label in labels]


def plot_simple_graph(simple_data, key1, key2, xlabel, ylabel, title, filename, labels, total=None):
    print("Titel: " + title)
    print(simple_data)
    plt.figure(figsize=(10, 6))
    # Setze maxlength für String x Labels, andernfalls kürze
    if xlabel != 'Age at admission':
        x_labels = shorten_labels(simple_data[key1].astype(str), 30)
    else:
        x_labels = simple_data[key1]
    bars = plt.bar(x_labels, simple_data[key2], color='blue')
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.title(title)
    plt.xticks(rotation=90)
    plt.grid(axis='y', linestyle='--', alpha=0.7)

    # Werte an jedem Balken anzeigen
    if labels:
        for bar in bars:
            height = bar.get_height()
            plt.annotate('{}'.format(height),
                         xy=(bar.get_x() + bar.get_width() / 2, height),
                         xytext=(0, 3),
                         textcoords="offset points",
                         ha='center', va='bottom',
                         rotation=90)

    # Infopanel mit Gesamtzahl
    if total:
        plt.gca().text(0.02, 0.98, f'Total: {total}', fontsize=12, verticalalignment='top',
                       bbox=dict(facecolor='white', alpha=0.5),
                       transform=plt.gca().transAxes)

    plt.tight_layout()
    plt.savefig("images/"+key1.lower()+"/"+filename+'_'+key1.lower())
    plt.close()


def extract_age_count(cases):
    cases_age_counts = cases['AGE_AT_ADMISSION'].value_counts().reset_index()
    cases_age_counts.columns = ['AGE_AT_ADMISSION', 'COUNT']
    cases_age_counts = cases_age_counts.sort_values(by='AGE_AT_ADMISSION')
    return cases_age_counts


def total_count_age_group(cases, group_lower_limit, group_upper_limit):
```

```
    cases_filtered = cases[(cases['AGE_AT_ADMISSION'] >= group_lower_limit) & (cases['AGE_AT_ADMISSION']
                                                                                <= group_upper_limit)]
    total_count = cases_filtered.shape[0]
    return total_count


def extract_gender_count(cases):
    cases_gender_counts = cases['GENDER'].value_counts().reset_index()
    cases_gender_counts.columns = ['GENDER', 'COUNT']
    return cases_gender_counts


def extract_ethnicity_count(cases):
    cases_ethnicity_counts = cases['ETHNICITY'].apply(simplify_ethnicity).value_counts().reset_index()
    cases_ethnicity_counts.columns = ['ETHNICITY', 'COUNT']
    cases_ethnicity_counts = cases_ethnicity_counts.sort_values(by='COUNT')
    return cases_ethnicity_counts


def simplify_ethnicity(ethnicity):
    ethnicity = ethnicity.lower()
    if 'white' in ethnicity:
        return 'White'
    elif 'black' in ethnicity or 'african american' in ethnicity:
        return 'Black'
    elif 'asian' in ethnicity:
        return 'Asian'
    elif 'hispanic' in ethnicity or 'latino' in ethnicity:
        return 'Latino'
    else:
        return 'Other'


def run_diagnosis_prediction(admission_note, icd_code):
    # Laden des Modells via Transformers Bibliothek
    tokenizer_diagnosis = AutoTokenizer.from_pretrained("bvanaken/CORe-clinical-diagnosis-prediction")
    model_diagnosis = AutoModelForSequenceClassification.from_pretrained("bvanaken/CORe-clinical-diagnosis-prediction")

    # Tokenisierung des Eingabetexts mit Positionsinformationen
    tokenized_input = tokenizer_diagnosis(admission_note, return_tensors="pt", max_length=512, truncation=True,
                                          padding="max_length", return_offsets_mapping=True)

    # Extrahiere die Token-IDs und die Offset-Positionen
    token_ids = tokenized_input['input_ids'][0]
    offset_mapping = tokenized_input['offset_mapping'][0]

    # Konvertiere Token-IDs zurück in Token
    tokens = tokenizer_diagnosis.convert_ids_to_tokens(token_ids)

    # Markiere die Stellen im ursprünglichen Text
    marked_text = list(admission_note)
    for (token, (start, end)) in zip(tokens, offset_mapping):
        if token not in ["[PAD]", "[CLS]", "[SEP]"]:  # Ignoriere spezielle Token
            marked_text[start] = f"\\textcolor{{blue}}{{{marked_text[start]}"
            marked_text[end - 1] = f"{marked_text[end - 1]}}}"

    # Füge die markierten Zeichen wieder zu einem String zusammen
    marked_text = ''.join(marked_text)
```

```
    # Ersetzen der speziellen Zeichen durch ihre LaTeX-Entsprechungen
    marked_text = (marked_text.replace('$', '\\$').replace('&', '\\&')
                   .replace('#', '\\#')).replace('%', '\\%')

    # Speichern der Aufnahmenotizen mit farblich markierten Tokens
    with open("testcase_data/marked_admission_notes/marked_note_"+icd_code+".txt", "w", encoding="utf-8") as file:
        file.write(marked_text)

    # Tokenisierung des Eingabetexts ohne Positionsinformationen für die Modellvorhersage
    tokenized_input = tokenizer_diagnosis(admission_note, return_tensors="pt", max_length=512, truncation=True,
                                          padding="max_length")

    output = model_diagnosis(**tokenized_input)

    # create predictions and prediction labels
    predictions = torch.sigmoid(output.logits)
    predicted_labels = [model_diagnosis.config.id2label[_id] for _id in (predictions > 0.3).nonzero()[:, 1].tolist()]

    # filter prediction labels to only include ICD codes
    # (1-4 digits, optionally starting with an uppercase letter)
    r = re.compile("^\d{1,3}$")
    results = list(filter(r.match, predicted_labels))
    icd_codes = [elem[:3] for elem in results]

    results = list(dict.fromkeys(icd_codes))
    return results


def find_distinct_icd_codes(diagnoses_list):
    icd_codes = set()
    for diagnoses in diagnoses_list:
        for diagnosis in diagnoses:
            icd_codes.update(diagnosis)
    return list(icd_codes)


def create_diagnosis_graphs_gender_age(case_results, genders, ages):
    for key, value in case_results.items():
        distinct_icd_codes = sorted(
            find_distinct_icd_codes([value["Male"]["diagnoses"], value["Female"]["diagnoses"], value[""]["diagnoses"]]))

        data = []
        for icd_code in distinct_icd_codes:
            for gender in genders:
                for age_index, age in enumerate(ages):
                    presence = 1 if icd_code in value[gender]["diagnoses"][age_index] else 0
                    gender_prettified = ""
                    if gender == "":
                        gender_prettified = "none"
                    else:
                        gender_prettified = gender
                    data.append([icd_code, gender_prettified, age, presence])

        df = pd.DataFrame(data, columns=['ICD-Code', 'Geschlecht', 'Altersgruppe', 'Prognose'])
        df_pivot = df.pivot_table(index='ICD-Code', columns=['Geschlecht', 'Altersgruppe'], values='Prognose',
                                  aggfunc='first')

        plt.figure(figsize=(20, 10))
        sns.heatmap(df_pivot, annot=True, cmap='coolwarm', fmt='d')
```

```
        plt.title('Prognose der ICD-Codes nach Geschlecht und Altersgruppe für ' + str(key))
        plt.ylabel('ICD-Code')
        plt.savefig("images/diagnosis/age-gender/" + str(key))
        plt.close()


def create_diagnosis_graphs_ethnicity_age(case_results, ethnicities, ages):
    for key, value in case_results.items():
        distinct_icd_codes = sorted(
            find_distinct_icd_codes([value["White"]["diagnoses"], value["Black"]["diagnoses"],
                                     value["Latino"]["diagnoses"], value["Asian"]["diagnoses"],
                                     value[""]["diagnoses"]]))

        data = []
        for icd_code in distinct_icd_codes:
            for ethnicity in ethnicities:
                for age_index, age in enumerate(ages):
                    presence = 1 if icd_code in value[ethnicity]["diagnoses"][age_index] else 0
                    ethnicity_prettified = ""
                    if ethnicity == "":
                        ethnicity_prettified = "Other"
                    else:
                        ethnicity_prettified = ethnicity
                    data.append([icd_code, ethnicity_prettified, age, presence])

        df = pd.DataFrame(data, columns=['ICD-Code', 'Ethnizität', 'Altersgruppe', 'Prognose'])
        df_pivot = df.pivot_table(index='ICD-Code', columns=['Ethnizität', 'Altersgruppe'], values='Prognose',
                                  aggfunc='first')

        plt.figure(figsize=(20, 10))
        sns.heatmap(df_pivot, annot=True, cmap='coolwarm', fmt='d')
        plt.title('Prognose der ICD-Codes nach Ethnizität und Altersgruppe für ' + str(key))
        plt.ylabel('ICD-Code')
        plt.savefig("images/diagnosis/age-ethnicity/" + str(key))
        plt.close()
```

testcase_evaluation.py

```
import pandas as pd
import utilities
import csv

# Relevante Schlüssel
length_of_stay = 'LENGTH_OF_STAY'           # Aufenthaltslänge
age_at_admission = 'AGE_AT_ADMISSION'       # Alter bei Aufnahme
gender = 'GENDER'                           # Geschlecht
ethnicity = 'ETHNICITY'                     # Enthnizität
died_in_hospital = 'EXPIRE_FLAG'            # Bei Krankenhausaufenthalt gestorben. 1 für ja, 0 für nein
diagnosis = 'DIAGNOSIS'                     # Diagnose (in Worten)
icd_code = 'ICD9_CODE'                      # Diagnose (ICD9 Code)
admission_note = 'TEXT'                     # Text der Aufnahmenotizen
short_codes = 'SHORT_CODES'                 # Eine Liste an ICD shortcodes

# Importieren relevanter Dateien
database = pd.read_csv('final_admission_diagnoses.csv')
testcases_anonymised_raw = pd.read_csv('random_testcases_modified.csv')

# Variablen für Testcases
testcases = pd.DataFrame()
genders = ["Male", "Female", ""]
ages = [16, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]
ethnicities = ["White", "Black", "Latino", "Asian", ""]


def df_to_dictionary(df):
    return pd.Series(df[admission_note].values, index=df[icd_code]).to_dict()


def return_random_line_from_dataframe(df):
    if not df.empty:
        return df.sample(n=1)
    else:
        print("DataFrame ist leer")


######## Altersbasierte Fälle ########
# ICD-Code 428xxx - Herzinsuffizienz - alte Menschen
icd_codes = ['428']

for code in icd_codes:
    case_data = database[database[icd_code].str.contains(r'\b'+code, na=False)]
    case_data.to_csv('testcase_data/'+code+'.csv')
    case_data_age_counts = utilities.extract_age_count(case_data)
    case_data_age_counts.to_csv('testcase_data/'+code+'-age-counts.csv')
    total_count = case_data_age_counts['COUNT'].sum()
    utilities.plot_simple_graph(case_data_age_counts, age_at_admission, 'COUNT', 'Age at admission',
                      'Count', 'Age distribution for ICD Code '+code, code, False, total_count)
    testcases = testcases.append(return_random_line_from_dataframe(case_data), ignore_index=True)

######## Genderbasierte Fälle ########
# ICD-Code 617 - Endometriose - Frauen (ausschließlich)
# ICD-Code 7330 - Osteoporose - Frauen
icd_codes = ['617', '7330']

for code in icd_codes:
    case_data = database[database[icd_code].str.contains(r'\b'+code, na=False)]
```

```
    case_data.to_csv('testcase_data/'+code+'.csv')
    case_data_gender_counts = utilities.extract_gender_count(case_data)
    case_data_gender_counts.to_csv('testcase_data/'+code+'-gender-counts.csv')
    total_count = case_data_gender_counts['COUNT'].sum()
    utilities.plot_simple_graph(case_data_gender_counts, gender, 'COUNT', 'Gender',
                    'Count', 'Gender distribution for ICD Code '+code, code, False, total_count)
    testcases = testcases.append(return_random_line_from_dataframe(case_data), ignore_index=True)

######## Ethnizitätsbasierte Fälle
# ICD-Code 250 - Diabetes - Afroamerikaner
# ICD-Code 151 - Magenkrebs - Asiaten
icd_codes = ['250', '151']

for code in icd_codes:
    case_data = database[database[icd_code].str.contains(r'\b'+code, na=False)]
    case_data.to_csv('testcase_data/'+code+'.csv')
    case_data_ethnicity_counts = utilities.extract_ethnicity_count(case_data)
    case_data_ethnicity_counts.to_csv('testcase_data/'+code+'-ethnicity-counts.csv')
    total_count = case_data_ethnicity_counts['COUNT'].sum()
    utilities.plot_simple_graph(case_data_ethnicity_counts, ethnicity, 'COUNT', 'Ethnicity',
                    'Count', 'Ethnicity distribution for ICD Code '+code, code, False, total_count)
    testcases = testcases.append(return_random_line_from_dataframe(case_data), ignore_index=True)

# Speichern der zufälligen Testfälle zur anonymisierung
# testcases.to_csv('random_testcases.csv', index=False)

# Umwandeln des Testcase Dataframes in dictionary
admission_note_dict = df_to_dictionary(testcases_anonymised_raw)
# Schlüssel als Referenz der Cases
case_keys = admission_note_dict.keys()
# Dictionary zum Speichern der Ergebnisse
case_results = {}

for case in case_keys:
    print("Running gender/age predictions for case: " + str(case))
    admission_note_value = admission_note_dict[case]
    gender_dict = {}

    for gender in genders:
        admission_note_gendered = admission_note_value.replace("$GENDER", gender)
        mortalities = []
        diagnoses = []
        gender_results = {}

        for age in ages:
            admission_note_gendered_and_age = admission_note_gendered.replace("$AGE", str(age))
            # create predictions
            diagnoses.append(utilities.run_diagnosis_prediction(admission_note_gendered_and_age, str(case)))
        gender_results["diagnoses"] = diagnoses
        gender_dict[gender] = gender_results
    case_results[case] = gender_dict
utilities.create_diagnosis_graphs_gender_age(case_results, genders, ages)

case_results = {}

for case in case_keys:
    print("Running ethnicity/age predictions for case: " + str(case))
    admission_note_value = admission_note_dict[case]
    ethnicity_dict = {}
```

```
    for ethnicity in ethnicities:
        admission_note_ethnicity = admission_note_value.replace("$ETHNICITY", ethnicity)
        mortalities = []
        diagnoses = []
        ethnicity_results = {}

        for age in ages:
            admission_note_ethnicity_and_age = admission_note_ethnicity.replace("$AGE", str(age))
            # create predictions
            diagnoses.append(utilities.run_diagnosis_prediction(admission_note_ethnicity_and_age))
        ethnicity_results["diagnoses"] = diagnoses
        ethnicity_dict[ethnicity] = ethnicity_results
    case_results[case] = ethnicity_dict
utilities.create_diagnosis_graphs_ethnicity_age(case_results, ethnicities, ages)
```

Testfall Aufnahmenotizen

Im den folgenden Abschnitt werden alle Aufnahmenotizen präsentiert. Die vom CORe-Modell für die Voraussage genutzten (512) Tokens sind dabei blau hervorgehoben.

ICD-Code 428 - Herzinsuffizienz

CHIEF COMPLAINT: Black stool

PRESENT ILLNESS: $ETHNICITY $Age yo w/ history of atrial fibrillation, primary prevention ICD, CAD s/p bypass grafting, s/p CABG, aortic valve replacement in [**2116**], CHF (EF 40%), pulmonary HTN, parkinson-like syndrome presenting with three to four days of dark, tarry stool. They have had one episode of melena per day. They deny nausea, vomiting, hematemesis, bright red blood per rectum. They have never had this happen before. They are on coumadin and their most recent INR was 3.3 (INR generally ranges 3 - 3.5 given mechanical valve). No dizziness, lightheadedness, chest pain. They have generally been feeling fatigued since this started. Patient has also had hemoptysis over the past two weeks. They and their partner are certain they do not have hematemesis. They have had hemoptysis in the past, but it has been worse. Patient coughed up frank blood today as per their partner. . Of note, patient had multiple episodes of recurrent aspiration pneumonia one year ago. They chose to go home with hospice in [**2130-12-23**] following an admission for aspiration pneumonia, but they improved and discontinued hospice several months ago. Since that time, they have gradually been restarting their home medications. They are currently taking coumadin, aspirin, but have not been taking their beta blocker. As per partner patient had 40 pound weight loss over past year. . In the ED, initial vitals were: 97.9 79 109/62 18 100%. Patient had melena on retal examination. Labs were significant for a HCT of 28.2 down from 33.1 on [**2131-1-22**]. They had an EKG showing a. fib at 80 BPM without evidence of ischemia. Patient had two 18 gauge peripheral IVs placed. GI saw patient in ED and recommended endoscopy. Vitals at transfer were: Temperature 98.2, Pulse 81, Respiratory Rate 16, Blood Pressure 99/60, O2 Saturation 96 on RA. . On arrival to the MICU, patient was comfortable. They have some fatigue, but generally are feeling well. No abdominal pain. Patient had not had an episode of melena since day prior to admission.

MEDICAL HISTORY: 1. History of cough-variant asthma. 2. Status post aortic valve replacement ([**2116**], done with CABG#2) 3. Coronary artery disease (CAD) status post CABG ([**2112**], [**2116**]). 4. Atrial fibrillation, status post multiple cardioversions. 5. Ischemic cardiomyopathy with severely depressed ejection fraction. 6. Pulmonary hypertension likely chiefly on the basis of diastolic dysfunction seen on cardiac catheterization in [**2125**]. 7. Sleep apnea. 8. Hyperlipidemia. 9. Post encephalitic Parkinson's disease. 10. Gout. 11. Recurrent pneumonia. 12. Abnormal CT scan (RLL lesion, decision previously made not to biopsy/work-up lesion) 13. hip fracture [**2127**], after a fall treated with R hemiarthroplasty 14. C. difficile colitis 15. Hypertension 16. s/p cholecystectomy and appendectomy 17. s/p colectomy with diversting colostomy, now repaired

MEDICATION ON ADMISSION: Allopurinol 200 mg PO daily Aspirin 81 mg daily Bumetanide 2 mg daily Carbidopa-levodopa 25 - 100 mg TID Gabapentin 100 mg [**Hospital1 **], 300 mg qHS Nitroglycerin 0.4 mg SL PRN Excelon patch 9.6 mg daily Spironolactone 12.5 mg daily Warfarin 6 mg daily Cymbalta 20 mg daily

ALLERGIES: Penicillins / Morphine / Ativan / Ace Inhibitors

PHYSICAL EXAM: On Admission: Vitals: T: afebrile BP: 114/61 P: 74 R: 13 O2: 99% on RA General: Alert, oriented, no acute distress HEENT: Sclera anicteric, MMM, oropharynx clear Neck: supple, JVP not elevated, no LAD CV: Irregular rhythm, regular rate, S1, mechanical S2, no murmurs/rubs/gallops Lungs: Bibasilar crackles, no wheezes/rhonchi, breathing comfortably Abdomen: soft, non-tender, non-distended, bowel sounds present, no organomegaly GU: no foley Ext: warm, well perfused, 2+ pulses, no clubbing, cyanosis or edema Neuro: CNII-XII intact, 5/5 strength upper/lower extremities, grossly normal sensation, 2+ reflexes bilaterally, gait deferred, finger-to-nose intact

FAMILY HISTORY: They have a strong family history of cancer. Their mother died of colon cancer, their father had lung cancer. They also had multiple grandparents with colon cancer.

SOCIAL HISTORY: They grew up in the [**Location (un) 86**] area. Married with two grown adopted children. Lives with their partner who is a retired nurse. They are retired VP of [**Last Name (un) 1687**] College. They drink [**12-24**] alcoholic beverages per week. They deny any tobacco history.

ICD-Code 617 - Endometriose

CHIEF COMPLAINT: abdominal pain lower back pain fever menorrhagia

PRESENT ILLNESS: 100 yo $ETHNICITY presenting with acute left-sided abdominal/pelvic and back pain, cramping in nature, that began this morning and has worsened throughout the day. The pain is associated with nausea and chills. Prior to the past day, they were feeling well and in their usual state of health, aside from baseline menstrual cramps (just finished their menses, which are quite heavy). They deny shortness of breath, chest pain, palpitations. Of note, they do have a history of STUMP tumor of the uterus, incidentally found on pathology after a myomectomy. They have been followed, as they desired preservation of fertility.

MEDICAL HISTORY: GynHx: - LMP last week, just finishing menses - sexually active with partner only x 5 years - reports mutually monogamous relationship

MEDICATION ON ADMISSION: Medication on ICU Admission: MetRONIDAZOLE (FLagyl) 500 mg IV Q8H Oxycodone-Acetaminophen [**2-1**] TAB PO/NG Q4H:PRN pain Acetaminophen 650 mg PO/PR Q6H:PRN fever Pantoprazole 40 mg PO Q24H Order date: [**12-3**] @ 2130 Phytonadione 10 mg PO/NG DAILY Duration: 3 Days Order date: [**12-3**] @ [**2115**] Docusate Sodium 100 mg PO BID:PRN constipation Levo HYDROmorphone (Dilaudid) 0.25-0.5 mg IV Q4H:PRN pain

ALLERGIES: Patient recorded as having No Known Allergies to Drugs

PHYSICAL EXAM: VS:T:98.6 HR 99 ST BP BP: 115/79 Sats: 99% RA General: 100 year-old in no apparent distress HEENT: normocephalic, mucus membranes moist Neck: supple no lymphadenopathy Card: Sinus tachycardic, normal S1,S2, no murmur/gallop or rub Resp; diminished breath sounds throughout. no wheezes GI: benign Extr: warm no edema Incision: R VATs site clean dry, intact, no erythema, margins well approximated Neuro: non-focal

FAMILY HISTORY: Pt unaware of family history, not in contact w/ [**Name2 (NI) **].

SOCIAL HISTORY: - Tobacco: none - Alcohol: last drink 3 months ago, infrequent ETOH - Illicits: none

ICD-Code 7330 - Osteoporose

CHIEF COMPLAINT: Shortness of breath, dyspnea on exertion

PRESENT ILLNESS: In brief, Pt is an 100 y/o $ETHNICITY with a PMH of afib on coumadin, CHF EF 45%, HTN, recently diagnosed COPD who presented with several weeks of SOB and DOE. They were admitted to the MICU and underwent TTE which showed critical AS, with valve area 0.5 cm2 and mean gradient 78 and peak gradient 129. They are being transferred to CCU for further care. . With regards to the patient's DOE, it has been worsening over the last few weeks. Previously, the patient was able to walk around the mall without much difficulty. Their breathing has been even worse over the last few days, to the point that they get short of breath even at rest. On the day of admission they were seen to be very SOB and were brought to her PCPs office where their amb sat was 86% on RA and they were sent to ED. .

MEDICAL HISTORY: # HTN # CHF # COPD: recently diagnosed by Dr. [Last Name (STitle) 656] at [Last Name (un) 4199] # stable goiter # afib on coumadin # ? aortic stenosis # GERD # osteoporosis # basal cell under left eye # s/p resection of childhood tumor behind heart

MEDICATION ON ADMISSION: # enalapril 10 mg AM, 5 mg PM # lasix 20 mg 2X /week # ranitidine 150 [**Hospital1 **] # evista 60 daily # warfarin 2-3 mg daily # lovastatin 40 mg daily # os-cal 500 [**Hospital1 **] # tylenol 2 daily

ALLERGIES: Patient recorded as having No Known Allergies to Drugs

PHYSICAL EXAM: T 36.7 HR 92 BP 110/77 RR 18 SaO2 94% on RA General Appearance: Well nourished Eyes / Conjunctiva: PERRL Head, Ears, Nose, Throat: Normocephalic Lymphatic: large goiter on R Cardiovascular: holosystolic murmur with loss of S2 consistent with severe AS, no delayed upstrokes Peripheral Vascular: (Right radial pulse: Not assessed), (Left radial pulse: Not assessed), (Right DP pulse: Present), (Left DP pulse: Present) Respiratory / Chest: (Breath Sounds: Crackles : at bases) Abdominal: Soft, Non-tender, Bowel sounds present Extremities: Right: 2+, Left: 2+ Skin: Not assessed Neurologic: Attentive, Follows simple commands, Responds to: Not assessed, Movement: Not assessed, Tone: Not assessed

FAMILY HISTORY: NC

SOCIAL HISTORY: Lives at [**Hospital3 **]. Smoking: quit 40 years ago, but had 10 pack year priot to that

ICD-Code 250 - Diabetes

CHIEF COMPLAINT:

PRESENT ILLNESS: The patient is a 100 year old $ETHNICITY with multiple medical problems including congestive heart failure, coronary artery disease, diabetes, hypertension, peripheral vascular disease, who presented with complaint of bright red blood per rectum, coffee ground hematemesis. The patient is a resident of [**Hospital6 13846**] Center. They received two units of packed red blood cells on [**1-31**] and CBC checked on [**2-2**] showed a hematocrit of 27.9. They were noted at rehabilitation to have an episode of tarr5y stools over the past two days and had bright red blood per rectum and coffee ground emesis. The morning of admission, they were found to EMS with blood pressure of 100/60; heart rate of 101. They were brought to the Emergency Room where their heart rate was 103; systolic blood pressure of 90, which increased to 100 with normal saline and tubal lavage. The EG showed clear fluid with mucus but no bile. Stool was dark. Heme positive. Hematocrit was found to be 16.6. Left femoral central line was placed.

MEDICAL HISTORY: Significant for peripheral vascular disease. Status post multiple toe amputations. Coronary artery disease, status post three vessel coronary artery bypass graft in [**2139**]. Congestive heart failure with ejection fraction of 30 to 40% by report with 2+ mitral regurgitation. Diabetes mellitus with neuropathy and nephropathy. Hypertension. Hypercholesterolemia. Chronic renal insuffriciency with baseline creatinine of 1.3. They had a right nephrectomy for nephrolithiasis. Depression. Cataracts. Gout. They had a pacemaker placement and mitral valve replacement.

MEDICATION ON ADMISSION:

ALLERGIES: They are allergic to Morphine, Motrin, Codeine. The rest of the discharge summary will be finished in a subsequent addendum. DR.[**Last Name (STitle) **],[**First Name3 (LF) **] 12-207

PHYSICAL EXAM:

FAMILY HISTORY:

SOCIAL HISTORY:

ICD-Code 151 - Magenkrebs

CHIEF COMPLAINT: T3N1 esophogeal adenocarcinoma s/p chemotherapy, radiation therapy

PRESENT ILLNESS: 100 y/o $ETHNICITY w/ T3N1 esophageal cancer s/p chemotherapy and radiation therapy as neoadjuvant treatment. They present now for definitive therapy. A minimally invasive esophagectomy was offered to the patient and accepted. The patient originally had been scheduled earlier but had a small neurological event from which they are totally recovered, and they present now for operation. They are somewhat further out than normal due to these extenuating circumstances.

MEDICAL HISTORY: PMH: Gastric esophogeal reflux disease, Barretts esophagitis, Esophogeal Cancer adenocarcinoma T3N1,s/p chemotherapy and radiation therapy , Hypertension, depression, Leg cramps, h/o substance abuse; Cerebral Vascular accident-small subacute right parietal infarct which appears embolic. Hypertension, Chronic obstructive pulmonary disease, renal calculi PSH: lithotripsy x 3, sigmoid colectomy, Jejunostomy tube/portacath [**3-25**]

MEDICATION ON ADMISSION: prilosec 40", lasix 25', atenolol 25', compazine 10", quinine sulfate 260', neurontin 300', wellbutrin 150', ASA 81', MgSO4 500', morphine sulfate 30", endocet 7.5/325", endocet 10/325', diazepam 5', 21mg nicotine patch, dilaudid

ALLERGIES: Patient recorded as having No Known Allergies to Drugs

PHYSICAL EXAM: articulate HEENT-no LAD REsp- clear Cor-RRR Abd- soft, J- tube in place Ext-no clubbing, cyanosis, edema Skin/ Incisions- cervical- slight erythema' abdominal

FAMILY HISTORY: non-contibutory

SOCIAL HISTORY: lives w/ partner.