Fraunhofer
IPA

FRAUNHOFER-INSTITUT FÜR PRODUKTIONSTECHNIK UND AUTOMATISIERUNG IPA

NINA SCHAAF | SASKIA JOHANNA WIEDENROTH | PHILIPP WAGNER

ERKLÄRBARE KI IN DER PRAXIS

ANWENDUNGSORIENTIERTE EVALUATION VON XAI-VERFAHREN

HRSG.: THOMAS BAUERNHANSL | MARCO HUBER | WERNER KRAUS

KI Fortschrittszentrum
LERNENDE SYSTEME

Nina Schaaf, Saskia Johanna Wiedenroth, Philipp Wagner
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

ERKLÄRBARE KI IN DER PRAXIS

Anwendungsorientierte Evaluation von xAI-Verfahren

Herausgeber
Thomas Bauernhansl, Marco Huber, Werner Kraus

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S·TEC

VORWORT

Künstliche Intelligenz (KI) ist eine der zentralen Technologien für die Zukunft. Ihre Einführung und der Einsatz fordern Unternehmen im besonderen Maß heraus. Es gilt, das Potenzial zu erkennen und dieses wirtschaftlich nutzbar zu machen. Lassen Sie sich dabei durch Europas größte Forschungskooperation auf dem Gebiet der KI, Cyber Valley, begleiten.

Mit dem KI-Fortschrittszentrum von Fraunhofer IAO und Fraunhofer IPA unterstützen wir Unternehmen dabei, das Potenzial von KI nutzbringend einzusetzen. An der Schnittstelle zwischen anwendungsorientierter Wirtschaft und exzellenter Forschung des Cyber-Valley-Konsortiums entwickeln wir innovative KI-Anwendungen für die Praxis und treiben damit die Kommerzialisierung von KI voran. Erklärtes Ziel ist dabei, menschzentrierte KI-Lösungen zu entwickeln.Denn nur wenn Menschen mit einer neuen Technologie intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann ihr Potenzial optimal ausgeschöpft werden.

Die Studienreihe »Lernende Systeme« des KI-Fortschrittszentrums gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Dabei werden übergreifende Themen wie Zuverlässigkeit, Erklärbarkeit (xAI), cloudbasierte Plattformen, Technologien und Einführungsstrategien diskutiert. Zudem werden einzelne Anwendungsbereiche in der Wissensarbeit, Bauwirtschaft, Produktion und dem Kundenservice im Detail beleuchtet.

Die vorliegende Studie vergleicht neun verschiedene Methoden zur Erklärung von KI-Modellen hinsichtlich unterschiedlicher quantitativer Qualitätskriterien. Darüber hinaus werden Möglichkeiten vorgestellt, um die Erklärungsverständlichkeit zu untersuchen. Die Evaluationsergebnisse sowie ein zusätzlicher Vergleich von Open-Source-Bibliotheken für Erklärungsmethoden unterstützen Leser*innen bei der Auswahl von geeigneten Erklärungsmethoden für den eigenen Anwendungsfall.

Wir wünschen Ihnen eine spannende Lektüre, und freuen uns, wenn wir in Zukunft auch Sie mit unserer Expertise auf Ihrem Weg zur menschzentrierten KI unterstützen dürfen.

Thomas Bauernhansl, Marco Huber, Werner Kraus
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA

INHALT

ABBILDUNGEN

TABELLEN

1 MANAGEMENT SUMMARY

In den letzten Jahren kommen Methoden der künstlichen Intelligenz (KI) in unterschiedlichsten Anwendungen immer mehr zum Einsatz. Diese zunehmende Beliebtheit sowie die steigende Komplexität und die dadurch bedingte Intransparenz der eingesetzten Algorithmen haben zu einem starken Anstieg an Arbeiten im Forschungsfeld der erklärbaren künstlichen Intelligenz (xAI) geführt. Die in diesem Kontext entwickelten Methoden zur Erklärung komplexer »Black-Box«-Modelle verfolgen unterschiedlichste Ansätze und unterscheiden sich sowohl im Vorgehen als auch im Ergebnis. Die vorliegende Studie beschäftigt sich deshalb mit der Frage, nach welchen Kriterien Anwender*innen für ihren Anwendungsfall geeignete xAI-Methoden bewerten und auswählen können.

Hierfür stellt die Studie Möglichkeiten zur quantitativen wie auch zur qualitativen Evaluation von xAI-Verfahren und deren Ergebnissen – den Erklärungen – vor. Die Resultate eines im Rahmen der Studie durchgeführten Benchmarkings liefern einen objektiven Qualitätsvergleich ausgewählter xAI-Methoden für unterschiedliche Anwendungsfälle. Weiterhin werden diverse Möglichkeiten diskutiert, um die Verständlichkeit von Erklärungen zu evaluieren, sowie Richtlinien zur Gestaltung von Erklärungen besprochen.

Um Anwender*innen den praktischen Einstieg für den Einsatz von xAI-Verfahren für eigene Anwendungsfälle zu erleichtern, bietet die Studie zusätzlich einen Überblick über existierende Open-Source-Softwarebibliotheken, die eines oder mehrere xAI-Verfahren bereitstellen. Diese Bibliotheken werden anhand diverser Kriterien bewertet. So können sich Anwender*innen einfach einen Überblick über die technischen Rahmenbedingungen der einzelnen Implementierungen verschaffen und damit auch deren Eignung für den eigenen Anwendungsfall einschätzen.

2 EINLEITUNG

Weil die Datenmengen immer größer werden und günstige Rechenleistung gut verfügbar ist, steigt das Interesse an Methoden der künstlichen Intelligenz (KI) und speziell am maschinellen Lernen (ML). Das gilt für verschiedenste Anwendungsdomänen wie die Produktion, das Finanzwesen oder die Medizin. Allen voran Deep-Learning-Ansätze, das heißt die Erstellung tiefer künstlicher neuronaler Netze mittels großer Datensätze, hat stark an Bedeutung gewonnen. Häufig übertreffen Modelle, die mit Deep Learning erstellt wurden, sogar den Menschen [1]. Allerdings haben viele der eingesetzten ML-Modelle »Black-Box«-Charakter. Das bedeutet, dass die gelernten Zusammenhänge so komplex sind, dass Menschen – und selbst KI-Experten – sie nicht mehr gänzlich nachvollziehen können. Für manche Anwendungen, beispielsweise die maschinelle Übersetzung oder Empfehlungssysteme, ist die mangelnde Nachvollziehbarkeit unkritisch. Für andere Anwendungsgebiete hingegen ist ein berechtigtes Interesse vorhanden, eingesetzte ML-Modelle auch erklären und verstehen zu können. Beispielsweise hat eine von der EU-Kommission eingesetzte Expertengruppe die Transparenz als ein zentrales Element zur Gestaltung einer vertrauenswürdigen KI definiert [2]. Vor allem dann, wenn maschinelle Fehlentscheidungen gravierende Auswirkungen auf den Menschen haben können, ist es unerlässlich, dass die Entscheidungen der Modelle nachprüfbar sind. Beispiele für solche Fehlentscheidungen finden sich bereits heute in unterschiedlichsten Anwendungsfeldern, etwa dem Justizsystem [3], der Medizin [4] oder dem autonomen Fahren [5].

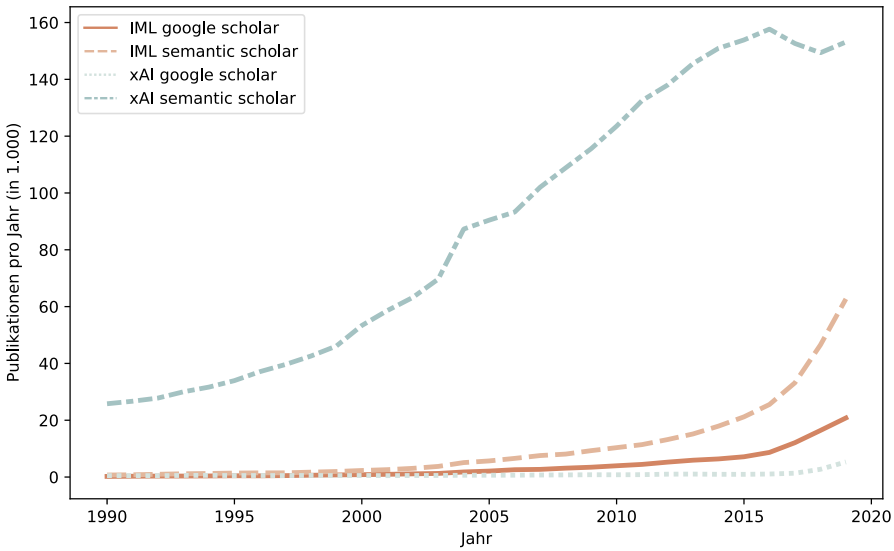

Abbildung 1: Anzahl der wissenschaftlichen Publikationen zu den Begriffen »Explainable Artificial Intelligence« (xAI) und »Interpretable Machine Learning« (IML) in den Jahren 1990 bis 2019. Bild: Fraunhofer IPA

Bereits seit den 1990er Jahren existieren Forschungsbestrebungen, um interpretierbare KI-Algorithmen zu entwickeln sowie existierende Black-Box-Modelle zu erklären. Diese Anstrengungen sind dem Forschungsfeld »Explainable Artificial Intelligence« (xAI, dt. erklärbare künstliche Intelligenz) zuzuordnen, auch bekannt unter dem Begriff »Interpretable Machine Learning«. In den letzten Jahren sind zahllose neue Erklärungsmethoden für Black-Box-Modelle entwickelt worden (s. Abbildung 1). Diese verfolgen unterschiedlichste Ansätze, um KI-Modelle erklärbar zu machen.

Das breite und diverse Methodenangebot sowie die fehlende Vereinheitlichung einer Schnittstelle zur praktischen Umsetzung erschweren allerdings die Auswahl für den eigenen Anwendungsfall deutlich. Es fehlen eine Übersicht und Bewertungskriterien, die es ermöglichen, die angebotenen Methoden angemessen einordnen zu können.

Genau hier setzt diese Studie an. Sie bietet zunächst eine Übersicht über gängige Unterscheidungsmerkmale von xAI-Methoden. Ergänzend werden mögliche Evaluationskriterien vorgestellt, um unterschiedliche Methoden und Frameworks zu bewerten. Die quantitative Evaluation ausgewählter Methoden hinsichtlich diverser Bewertungskriterien sowie eine Übersicht über bestehende xAI-Softwarebibliotheken sollen es Anwender*innen erleichtern, das passende Verfahren für ihren Anwendungsfall auszuwählen. Ein weiterer Aspekt der Studie liegt darauf, Möglichkeiten zu betrachten, Erklärungen und ihre Verständlichkeit und Tauglichkeit in der Praxis evaluieren zu können. Hierzu werden unterschiedliche Anforderungen an Verständlichkeit sowie darauf aufbauend Möglichkeiten zur Evaluation von Erklärungsverständlichkeit diskutiert.

Die übergeordnete Fragestellung der Studie lautet:

»Welche Evaluations- und Auswahlmöglichkeiten haben Anwender*innen beim Einsatz von xAI-Methoden für den eigenen Anwendungsfall und was ist bei der praktischen Anwendung zu beachten?«

Die Beantwortung dieser Fragen wird anhand von drei Aspekten untersucht:

1. Benchmarking und qualitative Bewertung von xAI-Methoden
2. Analyse existierender Softwarebibliotheken für xAI-Methoden
3. Betrachtung von Möglichkeiten, um die Verständlichkeit und Praxistauglichkeit von Erklärungen zu evaluieren

Jeder dieser Teilbereiche befasst sich mit unterschiedlichen Fragestellungen in Bezug auf die Bewertung und Auswahl von xAI-Methoden.

Die Studie ist folgendermaßen aufgebaut. Kapitel 3 bietet einen Überblick über das Forschungsfeld xAI und stellt eine Taxonomie zur Kategorisierung unterschiedlicher Erklärungsmethoden vor. Kapitel 4 bietet zunächst eine Übersicht über die in der Studie betrachteten Erklärungsverfahren, Benchmark-Datensätze und Evaluationsmetriken. Des Weiteren werden die in der Studie untersuchten xAI-Softwarebibliotheken sowie Kriterien zur Auswahl eines Frameworks für die Nutzung im eigenen Anwendungsfall vorgestellt. Anschließend folgt mit Kapitel 5 eine Präsentation der Ergebnisse. Dies umfasst zum einen das Benchmarking und eine Bewertung der verschiedenen xAI-Softwarebibliotheken. Weiterhin werden unterschiedlichen Möglichkeiten und Kriterien zur Evaluation der Verständlichkeit von Erklärungen diskutiert. Zuletzt folgt die Diskussion der Ergebnisse inklusive eines Ausblicks.

.

3 ERKLÄRBARE KÜNSTLICHE INTELLIGENZ (XAI)

In vielen Anwendungsbereichen bieten KI-Methoden ein enormes Potenzial, bestehende Prozesse zu vereinfachen, zu optimieren oder neue Möglichkeiten auszuschöpfen. So wurden 2019 in Deutschland bereits schätzungsweise knapp 220 Milliarden Euro Umsatz durch KI-Anwendungen beeinflusst – ein Großteil davon im produzierende Gewerbe [6]. Dem Wunsch, diese Potenziale zu nutzen, steht jedoch oftmals der Black-Box-Charakter vieler ML-Modelle im Weg. Häufig ist eine hohe Vorhersagegenauigkeit alleine nicht ausreichend: In einer 2020 vom Bitkom durchgeführten Umfrage sprachen sich 85 Prozent der Teilnehmenden dafür aus, dass KI-Software in Deutschland besonders gründlich geprüft werden soll [7]. Die Nutzung von Methoden zur Erklärung von Black-Box-Modellen kann die Hürden für einen erfolgreichen Einsatz von KI also stark senken. Daher wird im Folgenden das Forschungsfeld »erklärbare künstliche Intelligenz«, welches die Erklärung von KI-Systemen zum Ziel hat, kurz vorgestellt. Einen detaillierten Überblick über das Themenfeld xAI bietet ein von Burkart und Huber veröffentlichtes Übersichtspapier [8].

3.1 Überblick

Erklärbare künstliche Intelligenz verfolgt das Ziel, dem Menschen die Ergebnisse von KI-Systemen verständlich zu machen [9]. Obwohl bereits seit den 1990er Jahren an dem Thema geforscht wird (s. Abbildung 1), existiert bis heute keine standardisierte Definition von xAI. Oft wird ebenfalls die Bezeichnung »interpretable Machine Learning« verwendet, um Forschungsbestrebungen zu beschreiben, die KI-Systeme besser interpretierbar machen sollen. Auch die Begriffe *Interpretierbarkeit* und *Erklärbarkeit* sind nicht eindeutig definiert. Während manche Arbeiten die Begriffe bewusst – oftmals auch unterschiedlich – voneinander abgrenzen [10, 11], gebrauchen andere Arbeiten die Begriffe synonym. Auch die vorliegende Studie unterscheidet nicht explizit zwischen Erklärbarkeit und Interpretierbarkeit – beide Begriffe werden dem menschlichen Verstehen von KI-Systemen zugeordnet.

Drei Gesetze der Erklärbarkeit

Angelehnt an die drei von Isaac Asimov definierten Robotergesetze, die das Zusammenleben zwischen Menschen und Robotern definieren sollen, haben Biecek und Burzykowski [12] drei Gesetze der Erklärbarkeit für KI-Modelle abgeleitet. Diese legen Anforderungen fest, die jedes KI-Modell erfüllen sollte [12]:

1. **Vorhersagevalidierung:** für jede Vorhersage eines KI-Modells sollte überprüft werden können, wie stark die Evidenz ist, die die Vorhersage stützt.
2. **Vorhersagerechtfertigung:** für jede Vorhersage sollte man verstehen können, welche Variablen die Vorhersage in welchem Umfang beeinflussen.
3. **Vorhersageerwartung**: für jede Vorhersage sollte man in der Lage sein zu verstehen, wie sich die Vorhersage bei einer Veränderung der Werte der im Modell enthaltenen Variablen ändern würde.

Es gibt zwei Möglichkeiten, diese Anforderungen zu erfüllen. Die erste ist, White-Box-Modelle zu verwenden. Alternativ können xAI-Verfahren zur Erklärung von Black-Box-Modellen genutzt werden (s. nächster Abschnitt).

Black-Box vs. White-Box

Grundsätzlich gibt es zweierlei Ansätze, um KI-Systeme erklärbar zu machen. Ersterer hat das Ziel, inhärent interpretierbare KI-Modelle zu entwickeln, sodass deren Entscheidungen direkt für den Menschen nachvollziehbar sind. Die zweite Herangehensweise beschäftigt sich mit der Entwicklung von Methoden und Verfahren, die versuchen, Black-Box-Modelle im Nachhinein zu erklären. Dies wirft jedoch die Frage auf, welche Eigenschaften ein KI-Modell grundsätzlich aufweisen muss, um als interpretierbar zu gelten. Lipton [13] definiert hierfür drei Kriterien:

1. **Simulierbarkeit:** Nutzende sollten in der Lage sein, mithilfe der Eingangsdaten und der Modellparameter in angemessener Zeit jeden für eine Vorhersage benötigten Berechnungsschritt nachvollziehen zu können.
2. **Zerlegbarkeit:** Jeder Bestandteil des KI-Modells (d. h. Eingabedaten, Parameter und Berechnung) ist intuitiv erklärbar. Dies bedeutet jedoch auch, dass ein besonders umfangreiches und komplexes Aufbereiten der Rohdaten für das KI-Modell (Feature-Engineering und Feature-Extraction) die Interpretierbarkeit negativ beeinflusst.
3. **Algorithmische Transparenz:** Der eingesetzte Lernalgorithmus selbst ist nachvollziehbar.

Erfüllt ein KI-Modell also diese Eigenschaften, kann es als interpretierbar gelten. Diese Art von Modelle werden auch als »White-Box«-Modelle bezeichnet.

Datenerklärbarkeit

Zusätzlich zur Erklärung von KI-Modellen existieren Arbeiten, die sich mit der Erklärung bzw. Analyse der zugrundeliegenden Daten selbst befassen. Ein häufig genutzter Ansatz ist die Identifikation von sogenannten Prototypen [14], d. h. Dateninstanzen, die repräsentativ für alle oder eine Gruppe von Daten sind. Eine andere Möglichkeit ist, Techniken zur Dimensionalitätsreduktion wie t-Distributed Stochastik Neighbor Embedding (t-SNE) oder Hauptkomponentenanalysen einzusetzen. Hierbei werden die zumeist hochdimensionalen Eingabedaten in einen niedrigdimensionalen Raum transformiert, sodass nur besonders aussagekräftige Merkmale erhalten bleiben – oder anders ausgedrückt: man erhält die »Essenz« der Eingabedaten [15, S. 11].

Bedarfe und Zielgruppen

KI-Modelle sind mittlerweile in einer Vielzahl von Anwendungsfällen in Gebrauch. Diese reichen vom Dienstleistungssektor über die Finanzwirtschaft bis hin zum produzierenden Gewerbe. Über alle Branchen hinweg ist ein gestiegenes Bedürfnis nach Erklärungen der eingesetzten Methode zu beobachten. Laut einer vom IIT Berlin durchgeführten Studie ist Erklärbarkeit unter anderem für den Gesundheitssektor und die Produktion stark von Bedeutung [16].

Über die Entwicklung und den Einsatz von KI-Modellen hinweg kommen unterschiedliche Nutzergruppen mit den Modellen in Berührung. Tomsett et al. [10] plädieren daher dafür, auch die Erklärbarkeit von KI-Modellen nutzergruppenbezogen zu betrachten. Unterschiedliche Anwender*innen haben unterschiedliche Ansprüche und Erwartungen an KI-Modelle, weshalb auch die Interpretierbarkeit der Modelle in diesen Kontext gesetzt werden muss. Dabei definieren die Autoren sechs verschiedene Nutzergruppen:

- *KI-Entwickler*innen:* Eigentümer*innen oder Entwickler*innen des KI-Systems
- *Anwender*innen:* Personen, die direkt mit dem KI-System interagieren
- *Ausführende*: Personen, die basierend auf KI-Systemen Entscheidungen treffen
- *Entscheidungsbetroffene:* Personen, die unmittelbar von einer KI-Entscheidung betroffen sind, z. B. Endkund*innen
- *Daten-Subjekte:* Personen, deren Daten genutzt wurden, um das KI-System zu trainieren
- *Prüfende:* Personen, die das KI-System begutachten bzw. auditieren

Diese Kategorisierung von Nutzergruppen kann dabei helfen, zielgerichtet Anforderungen an die Interpretierbarkeit der eingesetzten KI-Modelle abzuleiten. Eine vom IIT Berlin durchgeführte Befragung hat ergeben, dass Erklärbarkeit derzeit noch hauptsächlich für die KI-Entwickler*innen selbst relevant ist [16]. Jedoch gehen die Befragten davon aus, dass Erklärbarkeit in den nächsten fünf bis zehn Jahren für viele Zielgruppen relevant werden wird, die heute noch relativ wenig mit Erklärungen konfrontiert sind, z. B. Endnutzende sowie Prüfende (intern und extern).

3.2 Eigenschaften von Erklärungsmethoden

Die enormen Forschungsbestrebungen zur »Explainable Artificial Intelligence« haben in den letzten Jahren zahllose neue Methoden zur Erklärung von KI-Modellen hervorgebracht. Allerdings fehlt es dem Forschungsbereich teilweise noch an allgemeingültigen Definitionen und Taxonomien. So ist etwa bisher keine standardisierte Taxonomie zur Kategorisierung unterschiedlicher Erklärungsmethoden vorhanden. Es gibt jedoch einige Eigenschaften, anhand derer sich diverse Herangehensweisen zur Erklärung von KI-Modellen unterscheiden lassen. Obwohl keine einheitliche Definition dieser Taxonomien vorliegt, werden im Folgenden häufig genannte Eigenschaften kurz erläutert.

Intrinsische Motivation: Grundsätzlich werden zwei Ansätze bei der Herstellung von Erklärbarkeit für KI-Systeme verfolgt (s. Abbildung 2). Ist die Nachvollziehbarkeit von ML-Modellen von Beginn an ein zentrales Kriterium, kann auf Modelle zurückgegriffen werden, die inhärent interpretierbar sind ***(ante-hoc)***. Solche von Natur aus interpretierbaren Modelle werden auch als White-Box-Modelle bezeichnet. Soll Erklärbarkeit hingegen für ein bereits existierendes Black-Box-Modell hergestellt werden, sind ***post-hoc*** Erklärungsmethoden zu nutzen.

Modellabhängigkeit: Des Weiteren können Methoden zur Herstellung von Erklärbarkeit ***modellspezifisch*** sein, also nur für eine Art von KI-Modell (z. B. neuronale Netze) funktionieren, oder ***modellagnostisch*** und somit für verschiedene Modellarten anwendbar sein (s. Abbildung 2).

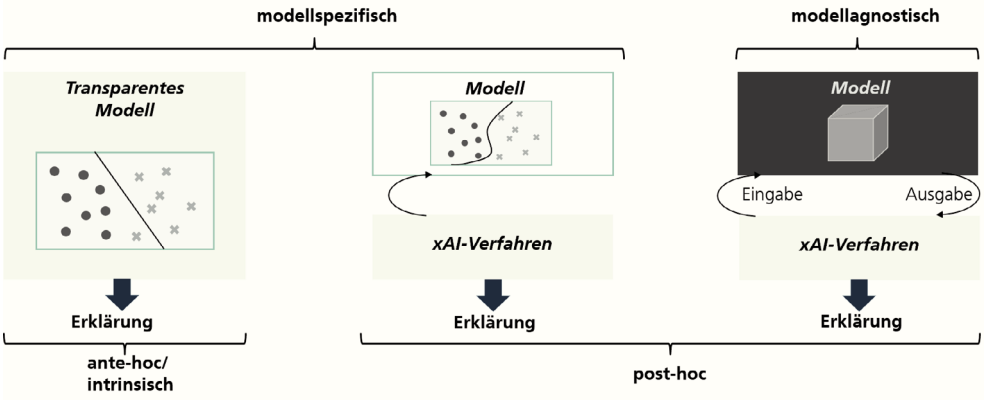

Abbildung 2: Demonstration des Unterschieds zwischen ante-hoc und post-hoc Erklärungsansätzen sowie modellagnostischen und modellspezifischen xAI-Verfahren. Bild: Fraunhofer IPA

Geltungsbereich (Scope): ***Globale*** Erklärbarkeit – manchmal auch Modellerklärbarkeit genannt – setzt voraus, dass die erzeugten Erklärungen für das Modell als Ganzes gelten. Globale Erklärbarkeit ermöglicht etwa, Einblicke in Nichtlinearitäten oder Wechselwirkungen in den Eingabedaten zu erlangen. Im Gegensatz dazu zielt die ***lokale*** oder Ausgabeerklärbarkeit darauf ab, die Gründe für die Entstehung einer einzelnen Prognose (oder einer Gruppe von Prognosen) des untersuchten ML-Modells zu verstehen. Eine Gegenüberstellung der beiden Ansätze ist in Abbildung 3 dargestellt.

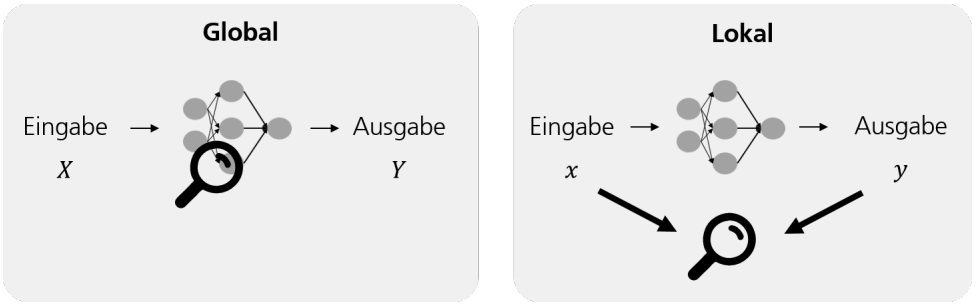

Abbildung 3: Unterschied zwischen globalen und lokalen xAI-Verfahren. Bild: Fraunhofer IPA

Lernaufgabe: Beim überwachten maschinellen Lernen werden die beiden Lernaufgaben ***Klassifikation*** und ***Regression*** unterschieden. Während erstere das Ziel hat, einer Dateninstanz eine von mehreren diskreten Klassen zuzuweisen, hat letztere kontinuierliche Ausgabewerte als Ergebnis. Da nicht jede Erklärungsmethode gleichermaßen auf beide Lernaufgaben angewandt werden kann, bietet sich hierdurch ein weiteres Merkmal zur Kategorisierung an.

Daten: Erklärungsansätze können zudem hinsichtlich der Art der verwendeten Daten kategorisiert werden. Grundsätzlich können Daten in Form von Text, Bildern (IMG) oder in tabellarischer Form (TAB) – sowohl numerisch als auch kategorisch – vorliegen. Einige Erklärungsverfahren sind für alle Datentypen universell verwendbar, während andere in ihrem Anwendungsbereich auf einen oder zwei Datentypen beschränkt sind.

Ergebnis der Erklärungsmethode: Verschiedene Erklärungsmethoden bringen Erklärungen in unterschiedlicher Form hervor und liefern somit kein einheitliches Ergebnis. Grundsätzlich lassen sich fünf verschiedene Ergebnistypen unterscheiden (s. auch Abbildung 4):

- ***Merkmalsrelevanz:*** Einem oder mehreren Merkmalen wird ein numerischer Wert zugewiesen. Dieser informiert darüber, wie relevant das Merkmal für eine bestimmte Entscheidung oder eine Gruppe von Entscheidungen ist.
- ***Datenpunkt(e):*** Neu erzeugte Datenpunkte oder Datenbeispiele aus den Trainingsdaten.
- ***Text:*** Erklärungen in Form von natürlicher Sprache oder textuellen Regeln (Wenn-dann Regeln).
- ***Modellinterna:*** Komponenten oder Parameter des zu erklärenden Modells.
- ***Neues ML-Modell:*** Erklärung eines existierenden Modells mithilfe eines neu generierten Modells. Dieses neu generierte Modell wird oftmals auch als Stellvertretermodell oder Surrogatmodell bezeichnet.

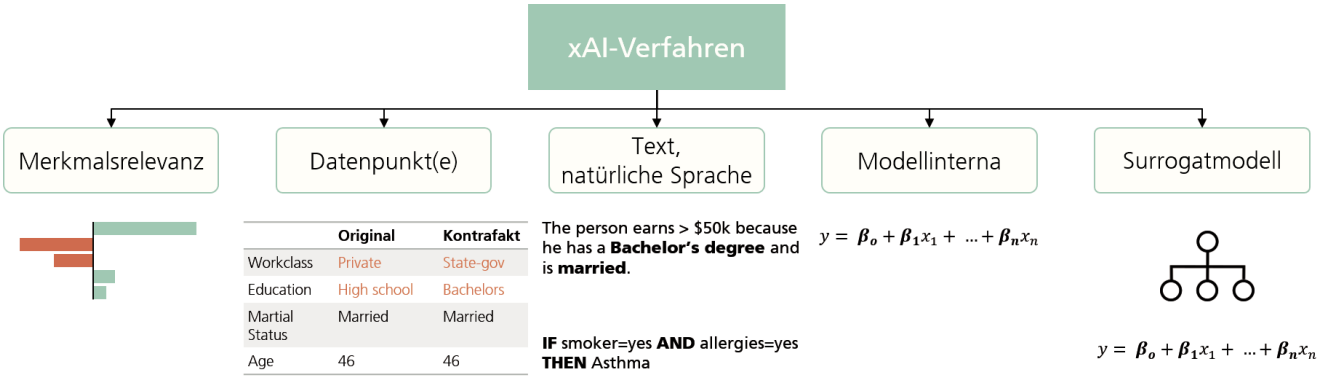

Abbildung 4: xAI-Verfahren können unterschiedliche Ergebnisse liefern: Merkmalsrelevanzwerte, Datenpunkte, Text, Modellinterna oder Surrogatmodelle. Bild: Fraunhofer IPA, Visualisierung Merkmalsrelevanz links: LIME [17]

Eigenschaft der Erklärung: Zuletzt weisen natürlich auch die Erklärungen selbst unterschiedliche Eigenschaften auf. Diese sind meist relativ abstrakt definiert – bis heute fehlt eine formale Definition von Erklärungseigenschaften. Einige Beispiele für Erklärungseigenschaften, entnommen aus [18, S. 17], sind:

- ***Genauigkeit:*** Vorhersagegenauigkeit für unbekannte Daten.
- ***Wiedergabetreue*** gibt an, inwieweit eine post-hoc Erklärung das Verhalten des Black-Box-Modells widerspiegelt. Die Wiedergabetreue ist eine besonders erstrebenswerte Eigenschaft, da nur bei korrekt wiedergegebenem Modellverhalten durch Analyse der Erklärungen sinnvolle Rückschlüsse gezogen werden können.
- ***Konsistenz*** gibt an, ob für ein definiertes Modell ähnliche Datenpunkte ähnliche Erklärungen erhalten.
- ***Verständlichkeit und Praxistauglichkeit*** gibt an, wie gut Nutzende die Erklärung verstehen und für ihre individuelle Aufgabe nutzen können. Die Verständlichkeit ist eine der wichtigsten und zugleich am schwersten zu formalisierenden Erklärungseigenschaften.

3.3 Evaluation von xAI-Methoden

Für einen erfolgreichen praktischen Einsatz von xAI-Methoden ist es von größter Bedeutung, die Verfahren bzw. die Erklärungen hinsichtlich diverser Kriterien (z. B. der Verständlichkeit) evaluieren zu können. Je nach verfolgter Zielsetzung sind hierbei unterschiedliche Ansätze zu präferieren. Doshi-Velez und Kim [19] definieren drei Herangehensweisen, um Erklärbarkeit zu evaluieren:

1. Anwendungsbezogene Evaluation
2. Nutzerbezogene Evaluation
3. Funktional-basierte Evaluation

Die anwendungsbezogene Evaluation beschreibt die Durchführung von Nutzerstudien im Kontext realer Anwendungsfälle. Hierbei werden die Experimente mit Angehörigen der Anwendungszielgruppe (z. B. Domänenexpert*innen) durchgeführt. Bei der nutzerbezogenen Evaluation wird ebenfalls mit Nutzerstudien gearbeitet – allerdings in vereinfachter Form. Die Experimente können hier auch mit Laien, außerhalb der realen Anwendungsumgebung, durchgeführt werden. Geeignet ist diese Herangehensweise, wenn hauptsächlich allgemeine Fragestellungen bzw. Eigenschaften von Erklärungen im Vordergrund stehen, beispielsweise welche Art von Erklärungen unter Zeitdruck besonders gut verständlich sind [19].

Die dritte Evaluationsmöglichkeit, die funktional-basierte Evaluation, verfolgt eine andere Methodik – hierfür sind keine Nutzerstudien durchzuführen. Stattdessen werden formale Definitionen von Erklärbarkeit verwendet, um die Erklärungsgüte zu messen. Eine Schwierigkeit hierbei ist, geeignete Metriken zu definieren, welche die Erklärbarkeit auf quantifizierbare Werte abbilden.

3.4 Beispielhafte Anwendungsfälle

Im Folgenden werden zwei Beispiele für praktische Anwendungsfälle beschrieben, bei welchen der Einsatz von xAI-Methoden für diverse Zielgruppen interessant ist.

3.4.1 Vorhersage des Zustandes von Fahrzeugkomponenten

In modernen Fahrzeugen fallen immer mehr Daten an. Diese Daten können wiederum dafür genutzt werden, potenzielle Fehler im Fahrzeug frühzeitig zu erkennen oder die Lebensdauer bestimmter Fahrzeugkomponenten präzise vorherzusagen. Beispielsweise lassen sich basierend auf historischen Daten der Zustand und die Alterung der Fahrzeugbatterie vorhersagen. Die zwischen den aufgezeichneten Daten und der zu ermittelnden Zielgröße herrschenden Zusammenhänge sind oftmals so komplex, dass sie nicht mehr manuell modellierbar sind. KI-Modelle, allen voran tiefe neuronale Netze, haben hingegen den Vorteil, dass sie auch komplexe Korrelationen abbilden können. Jedoch sind die vom Modell gelernten Zusammenhänge dann für den Menschen nicht mehr nachvollziehbar.

Genau diese Nachvollziehbarkeit kann jedoch für zweierlei Zielgruppen durchaus relevant sein. Bei der Wartung des Fahrzeugs könnte, basierend auf einem durch ein KI-Modell berechneten hohen Verschleißgrad, etwa ein Batteriewechsel vorgeschlagen werden. Für verantwortliche Mechaniker*innen können Erklärungen der Modellvorhersage dann z. B. darauf hinweisen, was den Verschleiß verursacht hat (lokale Erklärbarkeit). Für KI-Expert*innen dienen Erklärungen hingegen als Unterstützungswerkzeug bei der Modellentwicklung. Globale und lokale Erklärungen werden dazu genutzt, das Modellverhalten besser zu verstehen. Beispielsweise kann geprüft werden, welche Eingangsmerkmale einen besonders hohen Einfluss auf die Modellentscheidung haben. Diese Ergebnisse können mit Expertenwissen abgeglichen und so überprüft werden, ob das Modell ggf. irrelevante Merkmale zur Entscheidungsfindung nutzt. Unter Zuhilfenahme von Erklärungen ist es den Entwicker*innen also möglich, das Modell zu debuggen, potenzielle Fehler aufzudecken und ggf. zu beheben. Um Erklärungsmethoden zu diesem Zweck nutzen zu können, müssen Entwickler*innen sich jedoch sicher sein können, dass die Erklärungen tatsächlich die Entscheidungsweise des zugrundeliegenden KI-Modells abbilden. Es sind also Metriken nötig, mit denen die Güte der generierten Erklärungen bewertet werden kann.

Handelt es sich bei dem Anwendungsfall um eine besonders sicherheitskritische Applikation, können Erklärungen von KI-Modellen alleine nur einen begrenzen Beitrag leisten. In diesem Fall sollten zusätzlich noch weitere Verfahren eingesetzt werden, um die Modelle abzusichern. Eine Möglichkeit hierfür bieten Verifikationsverfahren. Mithilfe solcher Techniken kann beispielsweise die Robustheit eines KI-Modells gegenüber Schwankungen in den Eingabedaten sichergestellt werden. Eine aktuelle Studie des KI-Fortschrittszentrums »Lernende Systeme und Kognitive Robotik« des Fraunhofer IPA befasst sich mit diesem Thema und beleuchtet hierzu aktuelle Methoden und Möglichkeiten zur Entwicklung zuverlässiger KI-Modelle [20].

Abbildung 5: Bei der sensorgestützten Schüttgutsortierung wird ein Materialstrom sensorisch erfasst, bewertet und Einzelpartikel pneumatisch abgetrennt. Mehr Informationen zur sensorgestützten Sortierung: Fraunhofer IOSB (http://spr.iosb.fraunhofer.de). Bild: © Fraunhofer IOSB / M. Zentsch

3.4.2 Optische Qualitätsinspektion

In vielen Branchen ist die Schüttgutinspektion ein wesentlicher Bestandteil des Qualitätskontrollprozesses. In der Agrar- und Lebensmittelbranche können so beispielsweise Fremdkörper in Tee und Kräutern erkannt und entfernt werden (s. Abbildung 5). Zum Einsatz kommen hierfür verschiedene bildgebende Sensortechnologien wie Farb-, Ultraviolett- oder Hyperspektralkameras. Dadurch erkennbare Materialeigenschaften wie Form, Farbe oder Fluoreszenz können als Sortierkriterien herangezogen werden. Für schwierige Sortieraufgaben werden Multi-Sensor-Systeme benötigt, die eine komplexe Parametrierung von Sortierkriterien erfordern. Diese Aufgabe können Menschen jedoch kaum noch handhaben. KI-Methoden können helfen, da die Sortierkriterien direkt anhand von Beispielen erlernt werden und nicht in Form von manuell erstellten Regeln durch Experten abgebildet werden müssen.

Allerdings sind die von einem KI-Modell erlernten Sortierkriterien nicht mehr direkt ersichtlich: Die eingebüßte Nachvollziehbarkeit kann einen Vertrauens- und Kontrollverlust zur Folge haben. Auch hier interessieren sich diverse Zielgruppen für die Erklärbarkeit der eingesetzten KI-Modelle. Beispielsweise könnten sich Anlagenbetreibende fragen, ob ein KI-Modell, dessen Entscheidungen nicht nachvollziehbar sind, ohne Weiteres in eine Anlage integriert werden sollte. Erklärbarkeit kann dafür eingesetzt werden, die generelle Entscheidungsweise des Modells aufzuzeigen (globale Erklärbarkeit) und somit das Vertrauen in die Technologie zu stärken. Systemintegrator*innen hingegen werden vielmehr direkt einzelne Entscheidungen hinterfragen, z. B. wie eine bestimmte Sortierentscheidung zustande kommt (lokale Erklärbarkeit). Zuletzt können Entwickler*innen die Erklärbarkeit der KI-Modelle als Debugging-Werkzeug nutzen. In diesem Fall sind sowohl lokale als auch globale Erklärbarkeit von Interesse. Entwickler*innen können hierdurch beispielsweise verstehen, welche Sortierkriterien das Modell erlernt hat und wie sie die Sortierung eventuell noch optimieren können.

4 STUDIENINHALTE

Dieses Kapitel führt in die in der Studie untersuchten Inhalte ein. Speziell werden die Teilbereiche »Benchmarking« und »Analyse von xAI-Softwarebibliotheken« vorgestellt.

4.1 Umfang und Eingrenzung

Die im Rahmen der Studie beschriebenen Arbeiten und Analysen werden für einen eingeschränkten Anwendungsbereich, zusammengefasst in Abbildung 6, durchgeführt. Konkret werden Erklärungstechniken für Anwendungen des *überwachten maschinellen Lernens*, einem Teilbereich der künstlichen Intelligenz, betrachtet, da sich ein Großteil der xAI-Forschungsbestrebungen auf diesen Bereich fokussiert. Des Weiteren werden post-hoc Erklärungsverfahren untersucht. Aufgrund der Tatsache, dass in der Produktion mehrheitlich *tabellarische* Daten (Sensordaten) und *Bilddaten* anfallen, wurden die in Kapitel 5.1 vorgestellten Ergebnisse des Benchmark-Tests mit ebensolchen Daten durchgeführt. Grundlage hierfür sind überwacht gelernte ML-Modelle. Die Erklärung der Daten selbst (siehe Kapitel 3.1) wird in der vorliegenden Studie nicht betrachtet. Wie aus [16] hervorgeht, ist Erklärbarkeit derzeit noch hauptsächlich für *KI-Expert*innen* relevant. Daher sind die Studieninhalte größtenteils auch auf ebendiese Zielgruppe ausgelegt. Zuletzt werden die in Kapitel 5.3 diskutierten Empfehlungen, um die Verständlichkeit von Erklärungen zu evaluieren, auf das Erklärungsformat der *Visualisierung* beschränkt. Dies ist darin begründet, dass derzeitige Erklärungen fast ausschließlich als Visualisierungen bereitgestellt

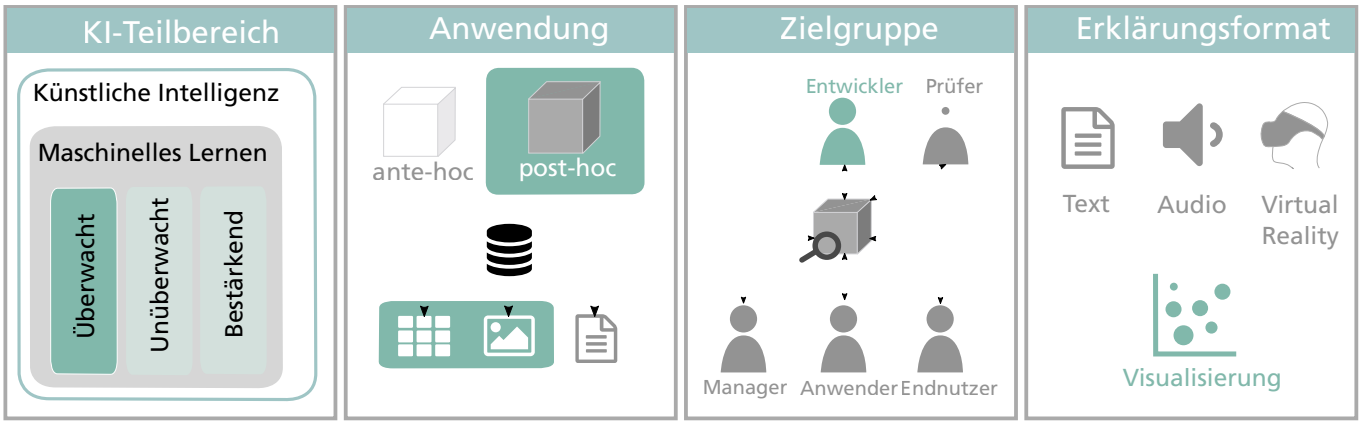

Abbildung 6: Eingrenzung der in der Studie betrachteten KI-Teilbereiche, Geltungsbereiche und Zielgruppen. Bild: Fraunhofer IPA

(vgl. auch die Methoden der in Kapitel 4.3.1 vorgestellten Softwarebibliotheken) und gewünscht werden (vgl. [16]). Visualisierungen sind hier definiert als die Darstellung einer Erklärung in Form einer (interaktiven) Grafik oder eines Bildes.

4.2 Benchmarking von xAI-Methoden

Das Benchmarking verfolgt das Ziel der funktional-basierten Evaluation (vgl. Kapitel 3.3) verschiedener xAI-Methoden. In der Literatur sind bisher wenige umfassende Analysen zur quantitativen Bewertung unterschiedlicher xAI-Verfahren und Erklärungen beschrieben. Einzelne Arbeiten beschäftigen sich gezielt mit der Evaluation bestimmter Erklärungstypen (z. B. Heatmaps) [21, 22]. Aus diesem Grund existiert auch kein standardisierter »Pool« an Evaluationsmetriken, die universell auf verschiedene Erklärungen anwendbar wären. Ziel des Benchmarkings von xAI-Methoden ist es daher, eine erste Annäherung an eine anwendungsübergreifende Evaluation mit bisher verfügbaren Evaluationsmethoden durchzuführen. Zu diesem Zweck wird für mehrere beispielhafte Anwendungsfälle eine Auswahl derzeit populärer xAI-Methoden betrachtet.

4.2.1 Anwendungsfälle

Das Benchmarking verschiedener Erklärungsmethoden wird auf Basis von fünf Anwendungen durchgeführt. Die hierfür genutzten Open-Source-Datensätze werden in zahlreichen wissenschaftlichen Untersuchungen verwendet. Die ausgewählten Datensätze decken sowohl Klassifikations- als auch Regressionsprobleme ab. Zudem ist die binäre Klassifikation genauso wie die Unterscheidung in mehr als zwei Klassen berücksichtigt. Des Weiteren inkludieren die tabellarischen Daten sowohl kategorische als auch numerische Daten. Die verwendeten Datensätze werden im Folgenden kurz beschrieben. Nähere Details sind in Tabelle 6 im Anhang aufgeführt.

Adult: Der öffentlich verfügbare UCI Datensatz »Adult« [23] umfasst Zensusdaten von knapp 49.000 Personen. Die Klassifikationsaufgabe besteht darin, auf Basis der Personenmerkmale wie Alter, Bildungsgrad oder Geschlecht zu bestimmen, ob eine Person mehr oder weniger als 50.000 US-Dollar pro Jahr verdient (binäre Klassifikation).

Sensorless Drive Diagnosis: Ebenfalls in der UCI Datenbank [23] enthalten ist ein Datensatz zur sensorlosen Überwachung eines Synchronmotors [24]. Der Datensatz umfasst ca. 58.000 Instanzen. Die 49 Merkmale wurden aus den elektrischen Strom-Antriebssignalen des Motors extrahiert. Zu unterscheiden sind elf Klassen, die verschiedene Fehlerzustände darstellen.

Bike Sharing: Dieser UCI Datensatz [23] umfasst über 17.000 Instanzen und 16 Merkmale wie Jahreszeiten, Wochentage oder Temperatur. Die Daten können dazu für die Regressionsaufgabe genutzt werden, die Anzahl der ausgeliehenen Fahrräder pro Stunde zu prognostizieren.

MNIST: Die MNIST-Datenbank [25] umfasst 70.000 Bilder von handgeschriebenen Ziffern der Zahlen 0 bis 9 und kann für Bildklassifikationsaufgaben eingesetzt werden. Die MNIST-Datenbank wird in vielen wissenschaftlichen Publikationen als Benchmarkdatensatz herangezogen.

ImageNet ist eine visuelle Datenbank mit über 14 Millionen handannotierten Bildern, die in mehr als 20.000 Kategorien eingeordnet sind. Die ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [26] ist ein Wettbewerb, in dem Algorithmen für Objekterkennung und Bildklassifikation hinsichtlich ihrer Leistung auf dem ImageNet-Datensatz bewertet werden. Hierfür wird eine Auswahl von 1.000 nicht-überlappenden Klassen betrachtet, die Objekte und Tiere umfassen.

4.2.2 Untersuchte xAI-Methoden

Im Folgenden werden diejenigen xAI-Methoden vorgestellt, die im Rahmen des Benchmarkings näher betrachtet werden. Grundlage für die Methodenauswahl sind die in Tabelle 9 und Tabelle 10 gelisteten Verfahren. Die dort enthaltenen Methoden sind in mindestens einer der in Kapitel 4.3.1 beschriebenen Softwarebibliotheken integriert. Ein zentrales Auswahlkriterium für die in der Studie betrachteten xAI-Methoden ist deren Verbreitung bzw. Popularität (s. Abbildung 7 sowie Tabelle 9 und Tabelle 10). Für die Erklärung von Bilddaten werden sowohl modellspezifische als auch modellagnostische Verfahren geprüft und zudem auf eine gewisse Diversität des Detailgrades der Erklärungen geachtet (s. Abbildung 10). Für tabellarische Daten sollen zumindest manche Verfahren auch für Regressionsaufgaben anwendbar sein. Zusätzlich wird hier auf eine möglichst hohe Diversität der Ergebnistypen wertgelegt (s. Kapitel 3.2).

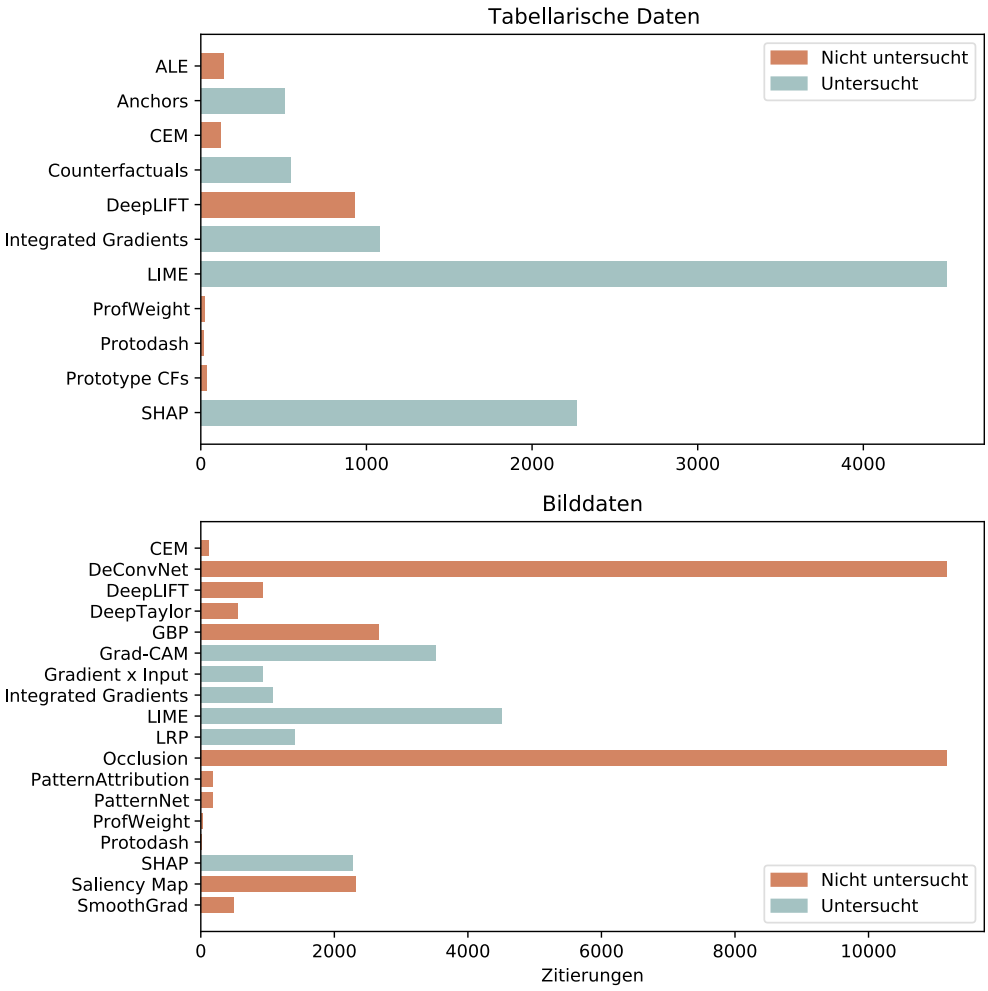

Abbildung 7: Verbreitung (Zitierungen laut Google Scholar, Stand: 14.01.2020) ausgewählter xAI-Verfahren. Die in der Studie betrachteten Verfahren sind grün markiert. Bild: Fraunhofer IPA

Da es im Rahmen dieser Studie nicht möglich ist, die Funktionsweise der Methoden noch individuell zu erklären, werden die Verfahren ergebnisorientiert kategorisiert und vorgestellt. Tabelle 1 fasst die untersuchten Methoden und ihre individuellen Eigenschaften zusammen.

Merkmalsrelevanz

Viele Erklärungsmethoden liefern sogenannte Merkmalsrelevanzen. Dies bedeutet, dass mehreren oder allen Eingabemerkmalen ein numerischer Wert zugewiesen wird, der anzeigt, wie relevant das jeweilige Merkmal für eine bestimmte Entscheidung oder alle Entscheidungen eines Modells ist. Mit *KernelSHAP* [27] und *LIME* [28] lassen sich Merkmalsrelevanzen für einzelne Entscheidungen erstellen (s. Abbildung 8 und Abbildung 9). Auch Integrated Gradients (IG) [29] ist nutzbar, um Merkmalsrelevanzen zu berechnen.

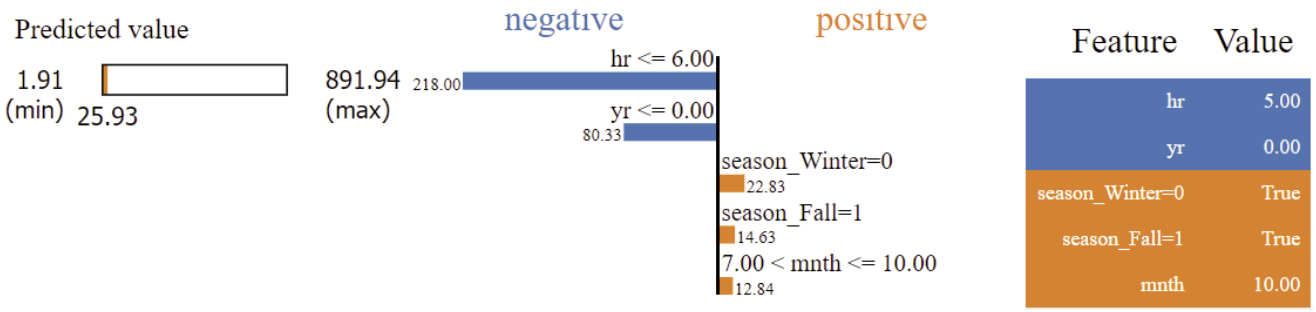

Abbildung 8: Erklärung tabellarischer Daten des Bike Sharing Datensatzes mit LIME. Links im Bild ist die Modellvorhersage für den zu erklärenden Datenpunkt dargestellt (25.93 ausgeliehene Fahrräder pro Stunde), rechts sind die Werte für fünf Merkmale des zu erklärenden Datenpunkts zusammengefasst. In der Mitte sind die Relevanzen für die rechts gelisteten Merkmale (hr, yr, season_Winter, season_Fall, mnth) dargestellt. Bild: Fraunhofer IPA, Visualisierungsfunktion: LIME [17]

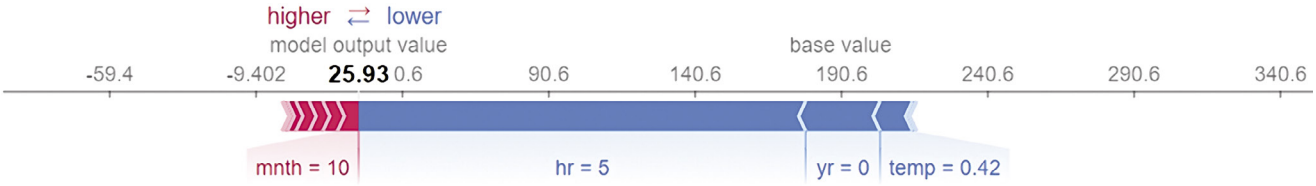

Abbildung 9: Erklärung tabellarischer Daten des Bike Sharing Datensatzes mit KernelSHAP. Fett markiert ist die Modellvorhersage für den zu erklärenden Datenpunkt (25.93 ausgeliehene Fahrräder pro Stunde). »base value« ist die durchschnittliche Vorhersage über alle Datenpunkte. Die farbigen Balken stellen die Relevanzwerte für die aufgeführten Merkmale dar. Eine blaue Hervorhebung bedeutet, dass der spezifische Merkmalswert weniger ausgeliehene Fahrräder im Vergleich zum »base value« zur Folge hat. Bild: Fraunhofer IPA, Visualisierungsfunktion: SHAP force_plot [30]

Sonderfall Merkmalsrelevanz: Heatmap

Eine Heatmap teilt jedem Pixel basierend auf einer Bewertungsfunktion einen Wert zu und kann demnach als Bild visualisiert werden. Die durch verschiedene Erklärungsmethoden generierten Heatmaps können sich in ihrer Auflösung (Detaillierungsgrad) unterscheiden. So liefern Methoden wie *IG*, *Layer-wise Relevance Propagation* (LRP) [31] und *Gradient* × *Input [*32] pixelgenaue Visualisierungen, wohingegen *Grad-CAM* [33] lediglich eine grobe Lokalisierungskarte generiert. *LIME* und *KernelSHAP* unterteilen das Bild in Superpixel und weisen jedem Superpixel einen Relevanzwert zu (s. Abbildung 10).

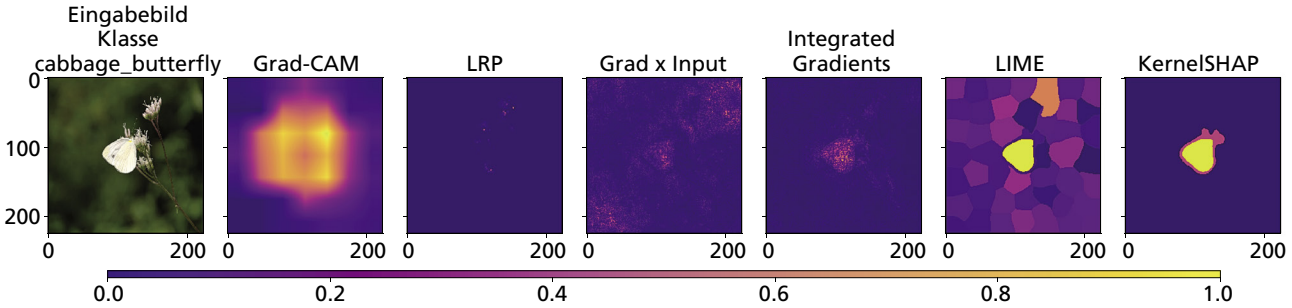

Abbildung 10: Beispiele für Heatmaps, die mithilfe unterschiedlicher Verfahren für ein Bild eines Kohlweißlings generiert wurden. Bild: Fraunhofer IPA; Schmetterlingsbild: Yongseok Lee (Pixabay)

Textuelle Erklärung: Wenn-Dann-Regeln

Wenn-Dann-Regeln sind sowohl direkt aus den Daten erlernbar als auch dazu geeignet, Black-Box-Modelle post-hoc zu erklären. Wenn-Dann-Regeln haben die Form:

WENN Temperatur > 10 °C und Windstärke = gering **DANN** spiele_Tennis = Ja.

Die Methode Anchors [34] liefert Erklärungen in Form solch einfacher Regeln für einzelne Datenpunkte.

Datenpunkt

Kontrafaktische Erklärungen [35] vermitteln, welches die kleinstmöglichen Veränderungen an den Eingangsdaten sind, die durchgeführt werden müssten, um eine andere Entscheidung des ML-Modells zu erhalten. Diese Erklärungsform liefert als Ergebnis einen oder mehrere neue, kontrafaktische Datenpunkte zurück (s. Abbildung 11).

Query instance (original outcome : 0)

	Age	Workclass	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
0	28.0	Private	10.0	Married-civ-spouse	Craft-repair	Husband	White	Male	0.0	0.0	40.0	United-States	0.321724

Diverse Counterfactual set (new outcome : 1)

	Age	Workclass	Education-Num	Marital Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
0	28.0	Private	10.0	Never-married	Craft-repair	Husband	White	Male	18709.8	1376.1	40.0	United-States	0.808
1	28.0	Private	10.0	Married-civ-spouse	Craft-repair	Husband	White	Male	78571.1	0.0	40.0	United-States	0.786

Abbildung 11: Kontrafaktische Erklärungen für den »Adult«-Datensatz. Die Unterschiede zum Originaldatenpunkt sind in rot hervorgehoben. Bild: Fraunhofer IPA, Visualisierungsfunktion: DiCE [36]

Stellvertretermodell/Surrogat

Zuletzt ist es möglich, ein Black-Box-Modell mithilfe eines White-Box-Modells zu erklären. Hierfür wird ein von Natur aus interpretierbares Modell wie z. B. ein *Entscheidungsbaum* (s. Abbildung 12) erstellt, das die Entscheidungen des komplexeren Black-Box-Modells approximieren soll. Diese Stellvertretermodelle werden auch als »Surrogate« bezeichnet. Ein Black-Box-Modell lässt sich mithilfe des Surrogat-Ansatzes durch ein White-Box-Modell abbilden. Im Rahmen der Studie wird ein Surrogat-Ansatz mithilfe eines Entscheidungsbaums (EB) realisiert. Ein solcher Entscheidungsbaum lässt sich zudem in eine Menge von Wenn-Dann Regeln übersetzen.

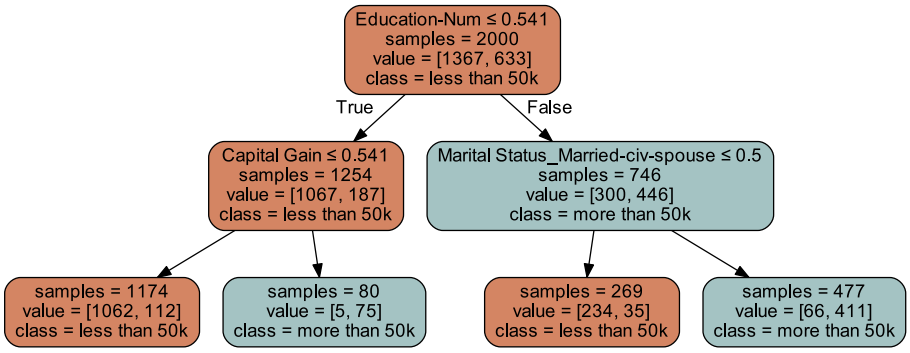

Abbildung 12: Kleiner Entscheidungsbaum mit zwei Ebenen. Bild: Fraunhofer IPA, Visualisierungsfunktion: scikit-learn [37]

Methode	Ergebnis	Scope		Modell agnostisch	Lernaufga.		Daten	
		Lokal	Global		Klass.	Regr.	IMG	TAB
LIME	Merkmals-relevanz	●		●	●	●	●	●
KernelSHAP	Merkmals-relevanz	●	▲	●	●	●	●	●
LRP	Merkmals-relevanz	●			●	▲	●	▲
Integrated Gradients	Merkmals-relevanz	●			●	▲	●	●
Grad-CAM	Merkmals-relevanz	●			●	▲	●	
Gradient × Input	Merkmals-relevanz	●			●	▲	●	▲
Anchors	Regeln	●		●	●		▲	●
EB-Surrogat	Surrogat, Regeln	●	●	●	●	●		●
Kontrafaktische Erklärungen	Datenpunkt	●		O	●		O	●

Tabelle 1: Übersicht und Kategorisierung der untersuchten xAI-Methoden hinsichtlich der Attribute Ergebnis, Scope, Modellabhängigkeit, Lernaufgabe und Daten.

Legende: ● erfüllt O mit Einschränkungen ▲ erfüllt, jedoch nicht Gegenstand der Studie

4.2.3 Evaluationsmetriken

Um die Güte eines Black-Box-Modells zu testen, gibt es viele Evaluations-Metriken, z. B. Accuracy, F1-Score, Precision und Recall. Im Gegensatz dazu fehlen einheitliche Metriken, um die Qualität der Erklärungen zu bewerten. Dennoch werden in der Fachliteratur einige Ansätze diskutiert, mit denen sich Verfahren zur Erklärbarkeit objektiv bewerten lassen. Die hier berücksichtigten Metriken werden im Folgenden kurz vorgestellt. Wie aus Tabelle 2 ersichtlich wird, sind einige der Metriken auf alle betrachteten Erklärungsverfahren gleichermaßen anwendbar, andere hingegen sind in ihrer Anwendbarkeit auf einen bestimmten Ergebnistyp oder eine Lernaufgabe beschränkt. Details zur praktischen Umsetzung der Metriken sind im Anhang aufgeführt.

Erklärungsinvarianz

Stabilität: Eine allgemein gültige Metrik für die Erklärbarkeit ist die Stabilität [38]. Erfüllt eine Erklärungsmethode diese Eigenschaft, bedeutet dies, dass man für den gleichen Datenpunkt stets die gleiche Erklärung erhält. Wird die Eigenschaft hingegen nicht erfüllt, kann dies aufgrund inkonsistenter Ergebnisse Verwirrungen und schlimmstenfalls einen Vertrauensverlust in die Erklärungen zur Folge haben. Zudem erschweren instabile Erklärungen das Ableiten von Handlungsoptionen.

Trennbarkeit (engl. separability) stellt sicher, dass unterschiedliche Datenpunkte unterschiedliche Erklärungen aufweisen [39]. So wird ausgeschlossen, dass beispielsweise zwei komplett gegensätzliche Datenpunkte die gleiche Erklärung enthalten. Da Entscheidungsbäume und Wenn-Dann-Regelsätze mehrere Datenpunkte über einen Baumpfad bzw. eine Regel einordnen, ist Trennbarkeit für diese Modelle nicht erstrebenswert.

Konsistenz ist dann für eine Erklärungsmethode gegeben, wenn sie ähnliche Erklärungen für nur geringfügig unterschiedliche Instanzen generiert [38]. Die Konsistenz bezieht sich dabei immer auf ein festgelegtes Modell. Um die Konsistenz eines Datenpunktes zu berechnen, werden die Eingangsmerkmale geringfügig verändert und im Anschluss geprüft, ob sich die Erklärung dadurch signifikant verändert.

Erklärungsähnlichkeit: Im besten Fall sollten sich die Erklärungen, die mithilfe unterschiedlicher xAI-Verfahren für denselben Datenpunkt generiert wurden, ähnlich sein. Das Phänomen, wenn dasselbe Ereignis durch verschiedene, widersprüchliche Erklärungen durch unterschiedliche Akteure erklärt wird, ist auch als »Rashomon Effekt« bekannt [40]. Tritt dieses Phänomen auf, d. h. wird ein Datenpunkt durch verschiedene Erklärungen unterschiedlich erklärt, stehen Nutzende vor der Herausforderung, zu entscheiden, welche Erklärung besser ist. Bestenfalls sollte daher eine starke Diskrepanz zwischen verschiedenen Erklärungen vermieden werden

oder Strategien zur Auflösung möglicher Erklärungskonflikte vorhanden sein. Um die Erklärungsähnlichkeit zu ermitteln, wird die Ähnlichkeit der am wichtigsten bewerteten Merkmale für jeden Datenpunkt betrachtet.

Wiedergabetreue

Die ***Wiedergabetreue*** (engl. fidelity) gibt an, inwieweit eine Erklärung der Vorhersage des Modells ähnlich ist. Diese Metrik kann in direkter Form lediglich für ML-Modelle berechnet werden (z. B. Entscheidungsbäume). Für andere Erklärungsergebnisse werden daher Approximationen (s. AOPC im übernächsten Absatz) benötigt, was einen Vergleich zwischen unterschiedlichen Erklärungsergebnissen schwierig bis unmöglich macht. Um die Wiedergabetreue zu berechnen, wird untersucht, in wie vielen Fällen das Erklärungsmodell und das Black-Box-Modell die gleiche Entscheidung getroffen haben [41]. Je besser die Erklärung die Vorhersage imitiert, desto höher ist der Wiedergabetreue-Wert.

Regel-Wiedergabetreue: Die Wiedergabetreue von Wenn-Dann-Regeln hängt eng zusammen mit der Regelabdeckung (engl.: coverage). Diese gibt an, für wie viele Datenpunkte des Datensatzes die Regel gilt. Die Wiedergabetreue informiert dann darüber, für wie viele von der Regel abgedeckten Datenpunkte das ML-Modell die gleiche Entscheidung trifft wie die betreffende Regel. Die Regel-Wiedergabetreue wird also für jede einzelne Regel berechnet.

AOPC: Diese speziell für Heatmaps entwickelte Metrik misst in gewisser Weise die Wiedergabetreue der Heatmap hinsichtlich der Modellentscheidung. Es wird angenommen, dass sich die Vorhersagegenauigkeit eines Modells verschlechtert, sobald relevante Bildbereiche durch zufällige Pixelwerte ersetzt werden. Folglich werden systematisch die relevanten Bildbereiche in absteigender Reihenfolge verändert und die Auswirkung auf die Klassifikationswahrscheinlichkeit beobachtet. Je höher der AOPC-Wert, desto besser erfasst die Heatmap die relevanten Bildregionen [21]. Da die AOPC-Metrik ursprünglich für die Bildklassifikation entwickelt wurde, ist sie nicht ohne Weiteres auf Regressionsanwendungen übertragbar. Ein Vorschlag [42] sieht jedoch vor, anstatt der Klassifikationswahrscheinlichkeit den Vorhersagefehler zu betrachten. Auch eine direkte Anwendung der AOPC-Metrik für tabellarische Daten ist nicht möglich. Allerdings kann auch hier eine Anpassung vorgenommen werden, sodass die einzelnen Merkmale nacheinander durch Zufallswerte ersetzt werden.

Funktionale Anforderungen

Laufzeit: Ein ebenfalls wichtiges Merkmal eines Erklärungsverfahrens ist die Zeit, die nötig ist, um eine Erklärung zu erzeugen. Diese Eigenschaft ist insbesondere hinsichtlich der praktischen Anwendung von Erklärungsmethoden relevant. Wird ein ML-System in einem Kontext angewandt, in dem eine schnelle Reaktionszeit gefordert ist, muss die Erklärung folglich ebenso schnell verfügbar sein.

Metrik	Lernaufgabe		Ergebnis				
	Klass.	Regr.	Heat map	Merkmals-relevanz	Daten-punkt	Regeln	Surrogat (EB)
Stabilität	●	●	●	●	●	●	●
Trennbarkeit	●	●	●	●	●	●	●
Konsistenz	●	●	●	●	●	●	●
Erklärungs-ähnlichkeit	●	●	●	●	●		
AOPC	●	O	●	O			
Wiedergabe-treue	●	●				●	●
Regel-Wieder-gabetreue	●					●	●*
Laufzeit	●	●	●	●	●	●	●

Tabelle 2: Anwendbarkeit der betrachteten Evaluationsmetriken auf unterschiedliche Erklärungsergebnisse und Lernaufgaben.

** wenn der EB in Wenn-Dann-Regeln übersetzt wird*

Legende: ● anwendbar O mit Modifikationen anwendbar

Wie Tabelle 2 zu entnehmen ist, können Stabilität, Trennbarkeit und Laufzeit für alle Erklärungen berechnet werden. Die Konsistenz ist lediglich für Heatmaps nicht anwendbar. Dies ist darin begründet, dass eine geringfügige Änderung der Eingangsmerkmale bei Bilddaten ein verrauschtes Bild zur Folge hätte, das dem Originalbild semantisch nicht ähnlich wäre. Um die Wiedergabetreue zu evaluieren, müssen drei verschiedene Metriken angewendet werden, was eine eingeschränkte Vergleichbarkeit der Ergebnisse zur Folge hat. Eine Berechnung der Wiedergabetreue für kontrafaktische Erklärungen (Ergebnis: Datenpunkt) ist nicht nötig, da das ML-Modell direkt dafür genutzt wird, den neu erzeugten Datenpunkt zu klassifizieren und somit eine Wiedergabetreue zum Modell unmittelbar gegeben ist.

4.2.4 Eingesetzte ML-Modelle

Random Forest

Random Forests sind ML-Verfahren, die für Klassifikations- und Regressionsaufgaben eingesetzt werden können. Ein Random Forest besteht aus vielen einzelnen Entscheidungsbäumen, die ein sogenanntes *Ensemble* bilden. Von einem Ensemble spricht man, wenn eine Menge von ML-Modellen kombiniert wird, um ein neues ML-Modell zu bilden, das eine bessere Performanz

erreicht als die individuellen Modelle alleine. Die einzelnen Entscheidungsbäume eines Random Forest werden hierbei nach dem *Bagging*-Prinzip kombiniert. Das bedeutet, dass die Entscheidungsbäume unabhängig voneinander erstellt werden und zwar so, dass jeder Baum auf einer anderen Untermenge der Trainingsdaten, basierend auf einer zufällig gezogenen Strichprobe (mit Zurücklegen), trainiert wird. Um schließlich mithilfe des Random Forest für Regressionsaufgaben Vorhersagen zu erstellen, werden die Entscheidungen der einzelnen Entscheidungsbäume gemittelt. Bei der Verwendung zur Klassifizierung erhält ein Random Forest von jedem Baum eine Klassenvorhersage und klassifiziert dann mithilfe des Mehrheitsvotums [43, S. 592].

Künstliche neuronale Netze (KNN)

Künstliche neuronale Netze (KNN) sind ein weiteres ML-Verfahren, das für Klassifikations- und Regressionsaufgaben nutzbar ist. Inspiriert sind KNN von den Lernprozessen im menschlichen Gehirn. Sie haben ihren Ursprung bereits in den 1940er Jahren [44, S. 12]. Die Grundeinheit eines neuronalen Netzes ist ein künstliches Neuron – ein Element, das gewichtete Eingaben entgegennimmt, verarbeitet und eine Ausgabe erzeugt. Im Grunde ist ein KNN nichts anderes als eine Kombination mehrerer Neuronen. Neuronale Netze sind typischerweise in Schichten angeordnet und gerichtet miteinander verbunden, d. h. jede Schicht besteht aus einem oder mehreren Neuronen und erhält ihre Eingabe von der vorhergehenden Schicht. Die Anzahl der Eingabemerkmale bestimmt die Menge der Neuronen in der Eingabeschicht, die Menge der Neuronen in der Ausgabeschicht hängt von der gewünschten Anzahl an Ergebnissen ab. Dazwischen können beliebig viele verdeckte Schichten liegen. Je mehr Schichten ein Netz hat, desto »tiefer« ist es, d. h. der Begriff »Deep Learning« bezieht sich auf die Verwendung vieler verdeckter Schichten.

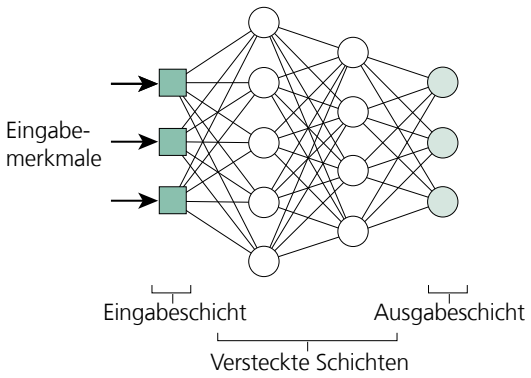

Abbildung 13: Schematische Darstellung eines vollvernetzten Feedforward-Netzes mit einer Eingabeschicht, zwei verdeckten Schichten und einer Ausgabeschicht. Bild: Fraunhofer IPA

Es existieren verschiedene Möglichkeiten, neuronale Netze aufzubauen, woraus verschiedene Typen neuronaler Netze resultieren. Sogenannte *Feedforward*-Netze bilden einen gerichteten Informationsfluss von der Eingabe- in die Ausgabeschicht ab. Konkret bedeutet dies, dass jede Schicht ihre Eingabe von der vorherigen Schicht erhält. Rekurrente Netze hingegen speisen ihre Ausgaben wiederholt als Eingaben ein, wodurch ein »Kurzzeitgedächtnis« implementiert werden kann. Diese Art von Netzwerkstruktur eignet sich besonders für die Arbeit mit Zeitreihendaten. In *vollvernetzten* KNN ist jedes Neuron einer Schicht mit jedem Neuron der nachfolgenden Schicht verbunden. Bei faltenden (engl. convolutional) neuronalen Netzen haben die verdeckten Schichten spezielle Strukturen. In den faltenden Schichten sind z. B. im Gegensatz zu vollvernetzten Schichten nicht mehr alle Neuronen miteinander vernetzt. Anwendung finden faltende neuronale Netze (CNN) vor allem bei der Verarbeitung von Bilddaten.

4.3 Analyse von xAI-Softwarebibliotheken

Aufgrund der großen Forschungsbestrebungen der letzten Jahre existieren mittlerweile einige ***Open-Source-Softwarebibliotheken***, die Implementierungen von xAI-Verfahren anbieten. Die durchgeführte Analyse bietet einen Überblick über die technischen Anforderungen und Features dieser Bibliotheken. Dies soll es Anwender*innen erleichtern, diejenigen Bibliotheken eingrenzen zu können, die für den eigenen Anwendungsfall und die damit verbundenen Anforderungen am besten geeignet sind.

4.3.1 Vorstellung Softwarebibliotheken

Die meisten der im Rahmen der Studie betrachteten Bibliotheken bündeln mehrere xAI-Methoden. Lediglich die Standalone-Bibliotheken der Methoden *SHAP* und *LIME* werden ebenfalls hinzugezogen, weil die Verfahren sehr verbreitet sind. Im Folgenden werden die unterschiedlichen Bibliotheken kurz beschrieben. Eine detaillierte Auflistung der von den Frameworks angebotenen xAI-Methoden ist dem Anhang (Tabelle 9, Tabelle 10) zu entnehmen.

- **InterpretML** [45]: Das von Microsoft entwickelte Projekt bietet verschiedene White-Box-Modelle (z. B. Entscheidungsbäume, Regellisten) sowie modell-agnostische Methoden, um Black-Box-Modelle zu erklären. Die angebotenen Methoden sind ausgelegt für tabellarische Daten.
- **AIX360** [46]: AI Explainability 360, entwickelt von IBM, bündelt verschiedene modellagnostische und modellspezifische Verfahren zur lokalen und globalen Erklärbarkeit sowie zur Erklärung von Daten.

- **Skater** [47]: Das von Oracle angebotene Framework ist die älteste der untersuchten Bibliotheken (besteht seit 2017). Skater bietet hauptsächlich modellagnostische Verfahren für lokale und globale Erklärbarkeit. Zusätzlich sind zwei lokale Erklärungsverfahren für tiefe neuronale Netze integriert.
- **alibi** [48]: Hinter der Entwicklung von alibi steht die auf KI und DevOps spezialisierte Firma Seldon Technologies Ltd. Ebenso wie AIX360 bietet auch alibi verschiedene modellagnostische und modellspezifische Verfahren zur lokalen und globalen Erklärbarkeit.
- **iNNvestigate** [49]: DeepExplain [50]: Beide Bibliotheken bündeln verschiedene Verfahren zur Erklärung tiefer neuronaler Netze und werden von Wissenschaftlern entwickelt.
- **Tf-explain** [51]: Das französische DeepTech Startup Sicara bietet mit tf-explain ein weiteres Framework zur Erklärung tiefer neuronaler Netze.
- **Captum** [52]: Ebenso wie die vorigen Bibliotheken ist auch Captum auf die Erklärung neuronaler Netze spezialisiert. Captum ist speziell für den Einsatz für PyTorch ausgelegt; Facebook treibt die Entwicklung voran.
- **SHAP** [27]: SHAP ist die Implementierung des gleichnamigen Erklärungsverfahrens. Die Autoren stellen dies selbst bereit.
- **LIME** [28]: Auch die Implementierung von *LIME* stellen die Autoren direkt zur Verfügung.

4.3.2 Bewertungskriterien

Im Folgenden werden Kriterien vorgestellt, die für die im vorigen Kapitel beschriebenen xAI-Softwarebibliotheken geprüft wurden. Diese adressieren unterschiedliche Anforderungen, die im Rahmen eines individuellen Anwendungsfalls gegeben sein können und somit die Auswahl einer geeigneten Bibliothek beeinflussen.

Unterstützte ML-Modellformate und -versionen
Dieses Kriterium informiert darüber, welche der gängigsten ML-Modellformate die jeweilige Implementierung unterstützt. Konkret überprüft wurde die Unterstützung für:

- Scikit-Learn
- Keras (mit Tensorflow Backend)
- PyTorch

Trainingsdaten
Sind zur Erklärungsgenerierung die Trainingsdaten nötig oder ist eine Erklärung eines spezifischen Datenpunktes ohne die Zugabe der Daten möglich? In manchen Anwendungsfällen kann es vorkommen, dass zum Zeitpunkt der Erklärungsgenerierung die Originaldaten, die zum Training des ML-Modells verwendet wurden, nicht mehr vorliegen oder schwer zu beschaffen sind. Eventuell handelt es sich dabei auch um sensible Daten, die besonders sorgfältig behandelt werden müssen. In diesem Fall ist es vorteilhaft, wenn die xAI-Methode(n) keinen Zugriff auf die Trainingsdaten benötigen, um eine Erklärung zu generieren.

Visualisierung
Bietet die Bibliothek eine eingebaute Funktion, um die Ergebnisse darzustellen?

Dokumentation
Existiert eine detaillierte Dokumentation? Dies beinhaltet sowohl eine Übersicht über alle angebotenen Funktionen und Methoden der Bibliothek als auch eine detaillierte Dokumentation der einzelnen Methoden selbst. Diese ist nötig, damit Entwickler die Methoden problemlos verwenden und parametrieren können.

Beispiele
Gibt es Beispiele, anhand derer der Prozess zur Erklärungserstellung sowie die Ergebnisse nachvollziehbar werden? Für welche Datentypen werden Beispiele bereitgestellt?

Metriken
Bietet die Softwarebibliothek Metriken, um die Erklärungen zu evaluieren?

Softwareaktualität (GitHub)
Wie regelmäßig wird die Bibliothek aktualisiert? Generell sollte, wenn möglich, auf regelmäßig gepflegte Repositories zurückgegriffen werden, da es so wahrscheinlicher ist, zukünftig von Fehlerbehebungen oder Updates zu profitieren. Die Aktualität bzw. Pflege der Git Repositories werden anhand von zwei Indikatoren erfasst:

- Letztes Release: Wann wurde zuletzt eine neue Version der Software veröffentlicht?
- Letzter Commit: Wann wurden zuletzt Änderungen am Code (z. B. Fehlerbehebungen) durchgeführt?

5 ERGEBNISSE

Dieses Kapitel beschreibt die Studienergebnisse. Zunächst werden die Ergebnisse des Benchmarkings vorgestellt, bei dem ausgewählte xAI-Methoden für unterschiedliche Beispielanwendungen verglichen und quantitativ bewertet werden. Die Ergebnisse des Vergleichs von xAI-Softwarebibliotheken ermöglichen, die softwareseitigen Rahmenbedingungen einzuschätzen, wenn xAI für die eigene Anwendung genutzt werden soll. Zuletzt empfiehlt die Studie, wie die Verständlichkeit von Erklärungen evaluiert werden kann.

5.1 Benchmarking

Für vier der fünf Anwendungsfälle des Benchmarkings wurden eigene ML-Modelle generiert. Da zur Erstellung der kontrafaktischen Erklärungen ein differenzierbares Modell benötigt wird, wurden die Anwendungsfälle »Adult« und »Sensorless Drive Diagnosis« mittels neuronaler Netze umgesetzt. Für das Benchmarking des ImageNet-Anwendungsfalls wurde auf zwei vortrainierte, im Rahmen der ILSVRC entwickelte Convolutional Neural Networks zurückgegriffen: ResNet50 [53] und MobileNetV2 [54]. Die Nutzung zweier Netze für denselben Datensatz ist hauptsächlich hinsichtlich der Laufzeitmessung begründet. Ziel ist es, die Laufzeiten zur Erstellung einzelner Erklärungen sowohl für relativ leichtgewichtige (MobileNetV2), als auch für komplexere Modelle (ResNet50) zu messen. Die Laufzeitmessungen selbst wurden auf einer Dual-Core CPU mit 8GB Arbeitsspeicher durchgeführt, da nicht davon auszugehen ist, dass in jedem Anwendungskontext eine leistungsstarke Grafikkarte (GPU) vorhanden ist, welche die Erklärungserzeugung ggf. beschleunigen könnte.

Tabelle 3 bietet eine Übersicht über die im Benchmarking verwendeten ML-Modelle. Detailinformationen zu Performance-Werten und Komplexität der Modelle sind Tabelle 7 im Anhang zu entnehmen. Zusätzlich ist dort die Anzahl der Modellparameter angegeben, um die ermittelten Laufzeiten zur Erklärungsgenerierung angemessen einordnen zu können.

	Datensatz	ML-Modell
TAB	Bike Sharing	Random Forest
	Adult	Neuronales Netz (Feedforward)
	Drive Diagnostics	Neuronales Netz (Feedforward)
IMG	MNIST	Neuronales Netz (CNN)
	ImageNet I	Neuronales Netz (CNN) (MobileNetV2)
	ImageNet II	Neuronales Netz (CNN) (ResNet50)

Tabelle 3: Übersicht über die für das Benchmarking eingesetzten ML-Modelle.

Tabelle 4 zeigt die Ergebnisse für die Auswertung der in Kapitel 4.2.3 beschriebenen Evaluationsmetriken (außer der Erklärungsähnlichkeit). Für die Kriterien Stabilität, Trennbarkeit und Konsistenz ist jeweils angegeben, ob die jeweilige Methode das Kriterium erfüllt oder nicht.

Es ist ersichtlich, dass *LIME*, *KernelSHAP* und *Anchors* das Stabilitäts-Kriterium verletzen. Das bedeutet, dass bei mehrmaliger Erklärungserstellung für den selben Datenpunkt verschiedene Erklärungen erzeugt werden können. Dies liegt daran, wie die Verfahren ihre Erklärungen generieren: Alle drei Methoden greifen hierfür auf Samplingansätze zurück. Die Nicht-Erfüllung des Trennbarkeits-Kriteriums für *Anchors* und *Entscheidungsbaum-Surrogate* ist durch das Konzept begründet, das hinter den Methoden steht. Ein Entscheidungsbaum ist so aufgebaut, dass ein einzelner Baumpfad zur Vorhersage für mehrere Datenpunkte herangezogen werden kann. *Anchors* versucht ebenfalls, Regeln zu finden, die auf mehr als nur einen Datenpunkt angewandt werden können. *IG* erfüllt das Kriterium Trennbarkeit für den Sensorless Drive Diagnosis Datensatz, nicht jedoch für den Adult Datensatz. Die Konsistenz konnte, wie bereits in Kapitel 4.2.3 erwähnt, lediglich für die tabellarischen Daten berechnet werden. Hier ist Anchors die einzige Methode, die das Kriterium nicht erfüllt.

Um die Wiedergabetreue zu evaluieren, kommen die drei Metriken der (Regel-) Wiedergabetreue und AOPC zum Einsatz. Weil die Wiedergabetreue ungleich berechnet wird, ist es nicht möglich, die Methoden über die verschiedenen Metriken hinweg zu vergleichen. Jedoch können für die tabellarischen Daten *LIME*, *KernelSHAP* und *IG* mittels AOPC verglichen werden. Hier fällt auf, dass *IG* beim Sensorless Drive Diagnosis Datensatz gleiche AOPC-Werte wie *KernelSHAP* erreicht. Beim Adult Datensatz liegt der AOPC-Wert für *IG* hingegen deutlich unter denen von *KernelSHAP*. Dies könnte an der Art der jeweiligen Merkmale der Datensätze liegen. Während der Sensorless Drive Diagnosis nur numerische Daten umfasst, besteht der Adult-Datensatz zu einem Großteil aus kategorischen Daten. Hierfür scheint *IG* (ohne Modifikation) nicht gut geeignet zu sein. Bestärkt wird diese Annahme zudem dadurch, dass *IG* für den

Adult-Datensatz das Trennbarkeits-Kriterium verletzt. Mittels der Regel-Wiedergabetreue lassen sich *Anchors* und *Entscheidungsbaum-Surrogate* miteinander vergleichen. Hier erreicht *Anchors* deutlich bessere Ergebnisse. Ein Nachteil der aus den *Entscheidungsbaum-Surrogaten* extrahierten Regeln ist, dass diese meist eine sehr geringe Abdeckung aufweisen, d. h. nur wenige Datenpunkte klassifizieren. Genau diese Regeln erreichen oftmals auch geringe Werte für die Wiedergabetreue. Eine Möglichkeit, diesem Problem zu begegnen, ist, den Entscheidungsbaum zu stutzen (engl. pruning), d. h. überspezialisierte oder redundante Teile des Baumes zu entfernen. Das reduziert die Regeln und erhöht die Gesamt-Wiedergabetreue.

Für die Bilddaten kann für alle Methoden ein AOPC-Wert berechnet und die einzelnen Verfahren dadurch direkt miteinander verglichen werden. Aus den für beide Bilddatensätze und drei Netze errechneten AOPC-Werten ergibt sich die in Tabelle 4 aufgeführte Rangfolge. Demzufolge erreichen *LIME* und *IG* über alle Datensätze und Netze hinweg die besten, *KernelSHAP* hingegen die schlechtesten Werte. Im Allgemeinen erzielen alle Methoden, mit Ausnahme von *KernelSHAP*, relativ ähnliche AOPC-Werte.

	Methode	Stabilität	Trennbarkeit	Konsistenz	Wiedergabetreue			Laufzeit
					Wiedergabetreue	AOPC	Regel-Wiedergabetreue	
TAB	LIME	x	●	●	–	2	–	schnell
	KernelSHAP	x	●	●	–	1	–	schnell
	IG	●	O	●	–	2	–	sofort
	Anchors	x	x	x	–	–	85/100	langsam
	Kontrafakte	●	x	●	●	●	●	langsam
	Surrogat (EB)	●*	x	●	90/100	–	68/100	sofort
IMG	Grad-CAM	●	●	–	–	2	–	schnell
	LRP	●	●	–	–	2	–	sofort
	Grad x Input	●	●	–	–	2	–	sofort
	IG	●	●	–	–	1	–	langsam
	LIME	x	●	–	–	1	–	langsam
	KernelSHAP	x	●	–	–	3	–	langsam

** sofern der Baum nicht neu trainiert wird*

Legende:

● erfüllt O teilweise erfüllt x nicht erfüllt – nicht prüfbar

Tabelle 4: Ergebnisse des Benchmarkings für die Kriterien Stabilität, Trennbarkeit, Konsistenz, Wiedergabetreue und Laufzeit.

Die Erklärungsähnlichkeiten können für tabellarische Daten lediglich für vier der sechs Methoden berechnet werden, da hierfür Merkmalsrelevanzwerte benötigt werden. Bei *LIME*, *KernelSHAP* und *IG* sind die Relevanzwerte direkt ersichtlich. Für die kontrafaktischen Erklärungen wird für jedes Merkmal der Unterschied zwischen den Originaldaten und dem kontrafaktischen Datenpunkt betrachtet: je größer der Unterschied, desto relevanter das Merkmal. Abbildung 14 zeigt die Ähnlichkeiten der mittels *LIME*, *KernelSHAP* und *IG* generierten sowie der *kontrafaktischen* Erklärungen für die Datensätze Adult und Sensorless Drive Diagnosis. Für den Bike-Datensatz konnten nur die mittels *LIME* und *KernelSHAP* generierten Erklärungen verglichen werden. Dargestellt ist die durchschnittliche Übereinstimmung der Top-10-Merkmale, gemittelt über alle generierten Erklärungen. Es ist erkennbar, dass sich sowohl für den Adult, als auch für den Sensorless Drive Diagnosis Datensatz die Erklärungen von *KernelSHAP* und *IG* relativ stark ähneln. Beim Adult Datensatz ähneln sich darüber hinaus die mittels *IG* generierten und die *kontrafaktischen* Erklärungen. Über alle Datensätze hinweg scheint die Übereinstimmung zwischen *LIME* und den restlichen xAI-Methoden eher gering auszufallen.

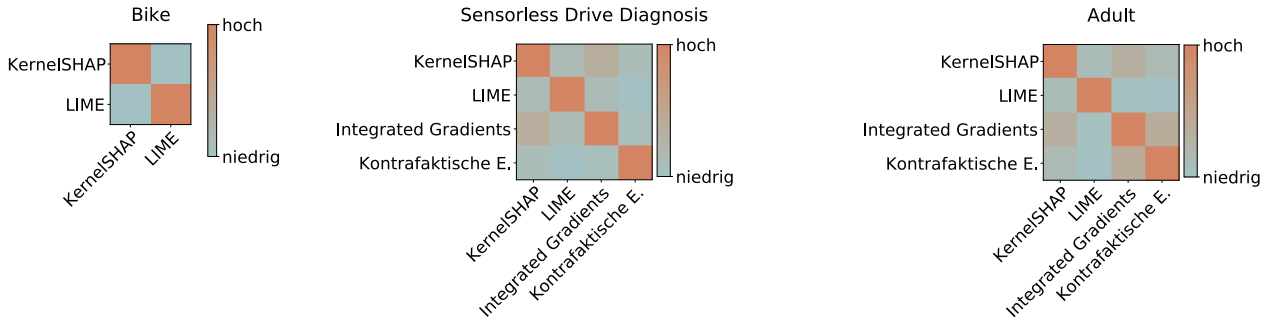

Abbildung 14:
Erklärungsähnlichkeiten für tabellarische Daten anhand der Datensätze Bike (links), Adult (Mitte) und Sensorless Drive Diagnosis (rechts). Rote Farben bedeuten höhere Übereinstimmungen. Bild: Fraunhofer IPA

Abbildung 15 stellt die Erklärungsähnlichkeiten für die Datensätze MNIST und ImageNet dar. Es ist zu erkennen, dass für den MNIST-Datensatz *LRP*, *Gradient* × *Input* und *IG* beinahe identische Erklärungen generieren. Dieses Ergebnis ist wenig überraschend, da alle drei Verfahren eine ähnliche Arbeitsweise basierend auf Rückpropagierung aufweisen. *Gradient* × *Input* und *IG*

propagieren den Gradienten von der Ausgabeschicht durch das Netzwerk zurück bis zu Eingabeschicht, während *LRP* für die Rückpropagierung spezielle Regeln nutzt. Auch beim ImageNet-Datensatz ähneln sich die von *LRP* und *Gradient × Input* erzeugten Erklärungen. Für die ResNet-Architektur sind zudem Ähnlichkeiten zwischen *LIME* und *KernelSHAP* zu beobachten, die bei MobileNet weniger stark ausgeprägt sind. Zuletzt ist erkennbar, dass sich die mit *LIME* und *Grad-CAM* erstellten Erklärungen bei allen drei getesteten Netzwerkarchitekturen ähneln.

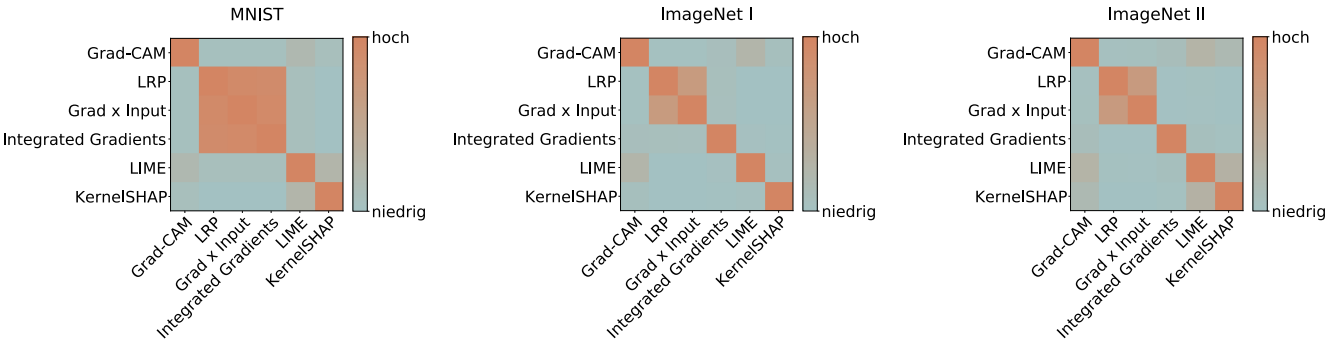

Abbildung 15: Erklärungsähnlichkeiten für Bilddaten anhand der Datensätze MNIST (links) und ImageNet mit den Netzwerkarchitekturen MobileNet (Mitte) und ResNet (rechts). Rote Farben bedeuten höhere Übereinstimmungen. Bild: Fraunhofer IPA

Die durchschnittlichen Laufzeiten, die zur Erklärungserstellung für die unterschiedlichen Bilddaten-Anwendungsfälle benötigt werden, sind in Abbildung 16 dargestellt. Es ist zu erkennen, dass über alle Netze und Datensätze hinweg die Erklärungserstellung mit *LRP* und *Gradient × Input* am schnellsten ist. Dicht dahinter folgt *Grad-CAM*. Die Ausprägungen der Laufzeiten für *IG*, *LIME* und *KernelSHAP* variieren hingegen. Für kleine Eingabebilder (MNIST, ResNet50 (112x112), MobileNetV2 (112x112)) sind die Laufzeiten von IG vergleichbar mit den Zeiten von *LIME* und *KernelSHAP*. Mit der Größe des Eingabebildes und Netzes steigt die Laufzeit jedoch enorm: für den Fall ResNet50 (224x224) liegt die Laufzeit für *IG* durchschnittlich bei über 70 Sekunden. Gerade für Anwendungsfälle, in denen eine schnelle Reaktionszeit erforderlich ist, sollte daher eher auf eine der schnelleren Methoden zurückgegriffen werden. Ist eine besonders schnelle Laufzeit erforderlich, kann, wenn möglich, zur Erklärungsgenerierung auf GPUs zurückgegriffen werden, da diese die Ausführung nochmals deutlich beschleunigen können.

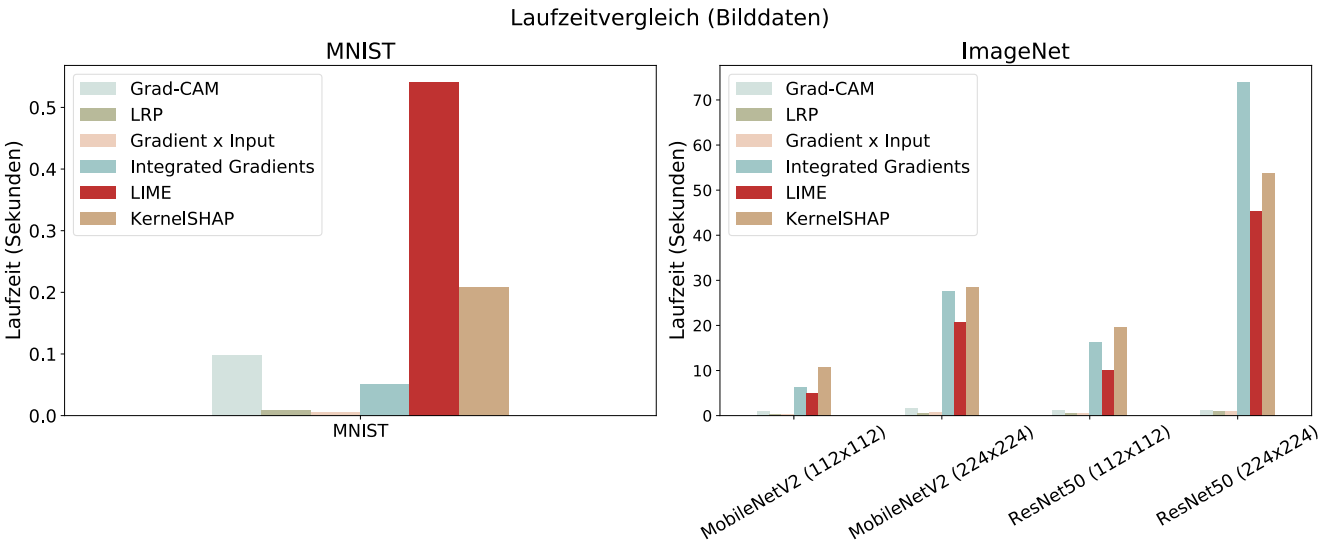

Abbildung 16: Laufzeiten der untersuchten Erklärungsmethoden für Bilddaten. Bild: Fraunhofer IPA

Abbildung 17 zeigt die durchschnittlichen Laufzeiten der Erklärungsgenerierung für die Anwendungen in Bezug auf tabellarische Daten. Die Laufzeit zur Erklärungserstellung mit *Entscheidungsbaum-Surrogaten* (hier gemessen: Training des Entscheidungsbaums) und *IG* ist innerhalb weniger Millisekunden abgeschlossen. Für *KernelSHAP* und *LIME* bewegt sich die Laufzeit im Rahmen von maximal einer Sekunde. Besonders zeitintensiv ist die Erklärungsgenerierung für *Anchors* und *kontrafaktische Erklärungen*.

Bei der Betrachtung der Laufzeiten ist zudem die starke Abhängigkeit von der jeweils genutzten Software zu berücksichtigen. Es ist davon auszugehen, dass nicht jede Implementierung grundlegend im Hinblick auf Effizienzkriterien optimiert ist. Daher ist es ratsam, hinsichtlich der Laufzeit zur Erklärungsgenerierung unterschiedliche Implementierungen miteinander zu vergleichen.

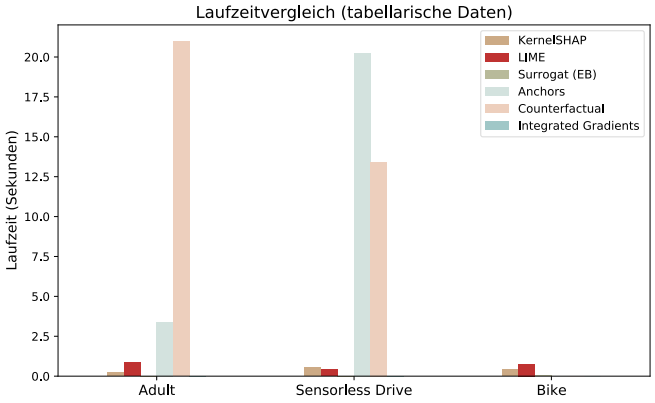

Abbildung 17:
Laufzeiten der untersuchten Erklärungsverfahren für tabellarische Daten. Bild: Fraunhofer IPA

Aus den Ergebnissen des Benchmarkings lässt sich kein klarer »Sieger« ableiten. Im Hinblick auf die AOPC-Werte scheint für die Erklärung von Anwendungen zur Bildklassifikation lediglich *KernelSHAP* nicht optimal zu sein. *IG* und *LIME* überzeugen zwar mit besonders guten AOPC-Werten, benötigen jedoch vergleichsweise relativ viel Zeit, um eine Erklärung zu generieren. Zudem verletzt *LIME* das Stabilitäts-Kriterium. Berücksichtigt man alle Kriterien, sind *LRP* und *Gradient × Input* für die Anwendung zur Erklärung von Bilddaten gut einsetzbar. Diese Methoden erfüllen die Anforderungen an Stabilität und Trennbarkeit und können mit einer schnellen Laufzeit (im Bereich von max. 1 Sekunde) überzeugen. Spielt die Laufzeit hingegen eine untergeordnete Rolle, ist zudem *IG* zu empfehlen.

Für die Anwendung auf tabellarische Daten ist *Anchors* nach Betrachtung der durchgeführten Experimente nicht gut geeignet. Die Methode verstößt nicht nur gegen die Kriterien Stabilität und Konsistenz, sondern benötigt auch vergleichsweise viel Zeit, um eine Erklärung zu generieren. Weil drei verschiedene Kriterien zur Bestimmung der Wiedergabetreue verwendet werden, sind die unterschiedlichen Methoden hierzu kaum vergleichbar. *LIME* und *KernelSHAP* liefern sowohl bei Stabilität, Trennbarkeit und Konsistenz als auch bei der Laufzeit eine ähnliche Performance. Lediglich bei der Betrachtung des AOPC-Kriteriums liegt *KernelSHAP* vor *LIME*. *Kontrafaktische Erklärungen* benötigen zwar vergleichsweise viel Zeit zur Erklärungsgenerierung, dafür sind diese die einzige Erklärungsform, für die eine Wiedergabetreue zum Modell stets gegeben ist. Zuletzt überzeugen *Entscheidungsbaum-Surrogate*, indem sie die Kriterien Stabilität und Konsistenz erfüllen, gute Wiedergabetreue-Werte und besonders kurze Laufzeiten haben.

Zusammenfassend lässt sich festhalten, dass speziell für tabellarische Daten aufgrund der hohen Diversität der möglichen Erklärungsergebnisse keine xAI-Methode pauschal zu empfehlen ist. Vielmehr sollte Anwender*innen eine geeignete Erklärungsmethode für den eigenen Anwendungsfall sorgfältig und mit besonderem Augenmerk auf die jeweiligen Anforderungen auswählen. Hierfür sind die im Benchmarking erlangten Ergebnisse eine erste Hilfestellung.

5.2 xAI-Softwarebibliotheken

Im Folgenden wird die Auswertung der analysierten Softwarebibliotheken anhand der in Kapitel 4.3.2 vorgestellten Auswahlkriterien vorgestellt.

Aus Tabelle 5 wird ersichtlich, dass, bis auf eine Ausnahme, alle Softwarebibliotheken Tensorflow (Keras) Modelle unterstützen. Jedoch werden zum Teil unterschiedliche Anforderungen an die Tensorflow-Version gestellt. Dies kann dann zu Kompatibilitätsproblemen führen, wenn verschiedene Bibliotheken genutzt werden sollen, um ein Modell mit einer bestimmten Tensorflow Version zu erklären. Die meisten Bibliotheken unterstützen auch PyTorch-Modelle zumindest mit Einschränkungen (Nutzung in modell-agnostischer Form als Black-Box-Modell). Eine deutlich geringere Unterstützung weisen hingegen scikit-learn-Modelle auf. Diese werden nur von *InterpretML* und *LIME* vollumfänglich unterstützt. Die restlichen Bibliotheken bieten entweder nur eine teilweise Unterstützung (nur ein Teil der angebotenen Methoden kann mit scikit-learn-Modellen arbeiten) oder aber keinerlei Möglichkeit zur Nutzung von scikit-learn-Modellen. Wird also beim eigenen Anwendungsfall ein nicht-differenzierbares Modell eingesetzt, etwa ein baumbasiertes Modell wie Random Forest, ist die Auswahl an verfügbaren xAI-Softwarebibliotheken deutlich limitiert.

InterpretML benötigt als einzige Bibliothek für alle angebotenen Methoden Zugriff auf zumindest einen Auszug der Trainingsdaten. Einige andere Verfahren kommen gänzlich ohne Bereitstellung der Trainingsdaten aus. Eine (teilweise) Bereitstellung von Visualisierungsfunktionen ist bis auf eine Ausnahme *(DeepExplain)* für alle untersuchten Bibliotheken gegeben. Dies kann es Anwender*innen erleichtern, die Erklärungen auszuwerten.

Besonders wichtig für einen schnellen und möglichst reibungslosen Einsatz der xAI-Bibliotheken ist eine umfangreiche Dokumentation der angebotenen Methoden. Die meisten der untersuchten Implementierungen verfügen über eine detaillierte Dokumentation. Allerdings gibt es auch einige Ausnahmen, in denen gar keine *(InterpretML, DeepExplain)* oder lediglich eine sporadische *(tf-explain)* Dokumentation bereitgestellt wird. In diesem Fall müssen sich Entwickler*innen soweit möglich an den bereitgestellten Beispielen orientieren. Dies erschwert es, den vollen Funktionsumfang und die Parametrierung der Methoden auszunutzen, und kann somit zu unnötigen Verzögerungen und möglichen Fehleinstellungen bei der Entwicklung führen.

Zuletzt wird deutlich, dass der Großteil der untersuchten xAI-Bibliotheken regelmäßig weiterentwickelt und aktualisiert wird. Lediglich *skater*, *iNNvestigate* und *tf-explain* haben seit mindestens einem Jahr kein neues Release-Update erhalten.

Tabelle 5: Vergleich verschiedener xAI-Softwarebibliotheken. Stand: 29.01.2021.

Bibliothek	ML-Modell			Trainingsdaten	Visualisierung	Dokumentation	Beispiele	Metriken	Softwareaktualität (GitHub)	
	Scitkit-Learn	Tensorflow (Keras)	PyTorch						Letztes Release	Letzter Commit
InterpretML [45]	●	●	●	●	●		TAB		01/2021	01/2021
InterpretML DiCE [365]		● ≥ 1.13	●		●	*	TAB		09/2020	01/2021
AIX360 [46]	o ≥ 0.21.2	● 1.14	o	o	o	●	TAB, IMG, TEXT	●	10/2020	12/2020
Skater [47]	o ≥ 0.18	● ≥ 1.4.0	o	o	●	●	TAB, IMG, TEXT		09/2018	06/2020
Alibi [48] **I**	o	● ≥ 2.0	o	o	o	●	TAB, IMG, TEXT	●	10/2020	01/2021
iNNvestigate [49]		Nur Keras 1.12			●	●	IMG, TEXT	●	06/2019	10/2020
DeepExplain [50]		● > 1.0					IMG		–	08/2020
Tf-explain [51]		● ≥ 2.0			●	● Unvollst.	IMG		02/2020	01/2021
Captum [52]			●		●	●	TAB, IMG, TEXT		01/2021	01/2021
SHAP [27]	o	● ≥ 2.0	●	o	●	●	TAB, IMG, TEXT		01/2021	01/2021
LIME [28]	● ≥ 0.18	● ≥ 2.0	●	o	o	●	TAB, IMG, TEXT		04/2020	01/2021

** Webseite vorhanden, ohne detaillierte Beschreibung*

Legende: ● erfüllt o teilweise / nicht für alle enthaltenen Methoden

5.3 Empfehlungen zur Evaluation der Verständlichkeit und Praxistauglichkeit visueller Erklärungen

Dieses Kapitel diskutiert verschiedene Möglichkeiten, um die Verständlichkeit von Erklärungen mithilfe von Nutzerstudien zu evaluieren. Diese sind sowohl auf die anwendungsbezogene als auch auf die nutzerbezogene Herangehensweise zur xAI-Evaluation (vgl. Kapitel 3.3) anwendbar. Wie bereits erwähnt, liegt der Fokus hier auf der Evaluation visueller Erklärungen, d. h. der Darstellung als (interaktive) Grafik oder Bild.

Hierfür wird zunächst auf potenzielle Herausforderungen eingegangen, die beim Gebrauch und der Evaluation von Erklärungen auftreten können. Anschließend folgt eine Beschreibung von Möglichkeiten für die Nutzerstudienkonzeption und -durchführung sowie die Vorstellung von Gestaltungsempfehlungen für die Erklärungsvisualisierung.

5.3.1 Erklärungen im Nutzungskontext

Die Visualisierungen und insbesondere interaktive Visualisierungen von Erklärungen hängen, wie andere interaktive Systeme auch, von verschiedenen, verketteten Kriterien ab, um gebrauchstauglich zu sein: von den Nutzenden an sich, deren Vorkenntnissen und Fähigkeiten, den spezifischen Zielen und den zielführenden Aufgaben – im jeweiligen Nutzungskontext. Wenn diese Kriterien berücksichtigt werden, können nutzerzentrierte Visualisierungen erstellt und nach den gleichen Kriterien auch evaluiert werden. Damit wird auch klar, dass es ***keine optimale Visualisierungsart kontextunabhängig für alle Zielgruppen*** geben kann. Beispielsweise benötigen Nutzende, die KI-Anwendungen zu Instandsetzungs- oder Wartungszwecken in unregelmäßigen Abständen prüfen, andere Unterstützung als solche, die die ML-Modelle regelmäßig mit der gleichen Erklärungsmethode prüfen. Des Weiteren können unterschiedliche Nutzergruppen auch unterschiedliche Erwartungen an die Erklärungen stellen [10].

Bevor Erklärungen mithilfe von Visualisierungstechniken dargestellt und anschließend ausgewertet werden können, müssen Nutzende die Limitierungen und Vorteile der Erklärungen verstehen können. Erst wenn potenzielle Einschränkungen und der Erklärungsnutzen verstanden sind, können die Erklärung verstanden, falsch priorisierte Merkmale identifiziert oder substanzielles Vertrauen in die KI gewonnen werden. Bevor die Ergebnisse mit Visualisierungen analysiert werden, müssen daher die nachfolgend aufgelisteten Punkte verständlich sein bzw. bewusstgemacht werden. In zwei Fällen werden zudem Umsetzungsmöglichkeiten vorgestellt, mithilfe derer den Herausforderungen begegnet werden kann.

Herausforderung 1: Post-hoc Erklärungen werden nachträglich generiert. Sie entsprechen daher nicht zwingend der tatsächlichen Abbildung des Entscheidungswegs.

> **Umsetzungsmöglichkeit:** Darstellung eines Falls, in dem es vorkommt, dass verschiedene Erklärungsmethoden aus dem Ergebnis des gleichen Modells widersprüchliche Erklärungen generieren. Dies kann das Vertrauen in die Erklärung und damit einhergehend auch in die KI zunächst schwächen. Langfristig kann aber genau diese Transparenz Vertrauen schaffen und dazu führen, dass Nutzende die Fähigkeiten und Grenzen der eingesetzten Erklärungen verstehen.

Herausforderung 2: Wird lokale Erklärbarkeit betrachtet, kann die Priorität einzelner Merkmale in einer Visualisierung nicht zwingend auf mehrere Dateninstanzen verallgemeinert werden. Es muss deutlich werden, dass die Ausprägung der Merkmalsrelevanz/die Veränderung des Merkmals/die Auswahl des Merkmals von dem spezifischen Merkmalswert selbst (z. B. sehr niedriger Wert), jedoch auch von Merkmalswerten korrelierender Merkmale abhängen kann.

> **Umsetzungsmöglichkeiten:**
> - Abstrakte Darstellung der Veränderungen von relevanten Merkmalen und Korrelationen, wobei es keine konkreten Merkmalsbezeichnungen gibt.
> - Greifbare, konkrete Beispiele mit wenigen Merkmalen, welche die einzelnen Korrelationsmöglichkeiten im relevanten Anwendungskontext beschreiben.

Herausforderung 3: Es ist nicht immer einfach, den Nutzen der Erklärung klarzumachen. Wozu befähigen die Erklärungen die Nutzenden? Die Antwort darauf muss auch wieder abhängig vom Nutzungskontext sein, aber hilft den Nutzenden, sie zu motivieren, sich mit den Daten in angemessenem Umfang auseinanderzusetzen.

5.3.2 Evaluationsmöglichkeiten auf Basis von Nutzerstudien

Wenn die Evaluation nicht nur die Verständlichkeit an sich bewerten, sondern auch Verbesserungsmöglichkeiten hierzu aufzeigen soll, bietet es sich an, auch die Kausalitäten während der Interaktion mit der Erklärungsmethode zu untersuchen. Beobachtet man Nutzerinteraktionen, wird nachvollziehbar, *warum* Nutzende bestimmte Erklärungen nicht richtig interpretieren können. Denn während die kognitive Belastung durch komplexe Visualisierungen Interpretationsfehler bedingen kann, sind möglicherweise auch schon Teile der Erklärung fehlerfördernd, wenn die Nutzenden sie nicht korrekt wahrnehmen oder erkennen können. Die bei einer Nutzerstudie erzielten qualitativen Ergebnisse klären also, woher Verständnisprobleme möglicherweise kommen und geben Hinweise, wie einzelne Erklärungsmethoden optimierbar sind.

Studienaufbau
Um eine solche Nutzerstudie durchzuführen, sollte pro Proband*in ein Set von mindestens fünf Erklärungen getestet werden. Diese Erklärungen sollten jeweils einem der folgenden Zustände entsprechen:

- Richtige Entscheidung des KI-Modells mit
  - einem oder mehreren zu hoch gewichteten Merkmal(en)
  - einem oder mehreren zu niedrig gewichteten Merkmal(en)
  - gut gewichteten, eindeutigen Merkmalen
- Falsche Entscheidung des KI-Modells mit
  - einem oder mehreren zu hoch gewichteten Merkmal(en)
  - einem oder mehreren zu niedrig gewichteten Merkmal(en)

Falls mehrere Varianten der Erklärungsmethode verglichen werden sollen, können A/B-Tests mit zwei Probandengruppen helfen, die Vor- und Nachteile zu analysieren.

Studiendurchführung: Beobachtung
Eine Möglichkeit ist, die Beobachtung durch die *Think-Aloud-Methode* zusammen mit *Eye-Tracking-Analysen* zu begleiten. Bei der Think-Aloud-Methode wird der/die Proband*in gebeten, alle Gedanken und Intentionen laut auszusprechen. Eye-Tracking-Analysen helfen, die Wahrnehmung der Proband*innen nachzuvollziehen. *Eye-Tracking* ermöglicht es, Situationen zu erfassen, in denen Proband*innen spezifische Elemente besonders lange betrachten (Fixationen) oder unerwartete Betrachtungspfade einschlagen (insbesondere schnelle Augenbewegungen (Saccaden)). Gemeinsam mit den Think-Aloud-Ergebnissen können sich Studienleitende so den kognitiven Prozessen nähern. Fügt man die Fixationspunkte aller Proband*innen zusammen, können anschließend Wahrnehmungs-Heatmaps erzeugt werden, die z. B. zeigen, welche möglicherweise relevanten Bereiche die Proband*innen übersehen haben.

Studiendurchführung: Abfrage
Nachdem die Studienleitenden die Reaktionen auf einzelne Erklärungen beobachtet haben, können sie Verständnisfragen stellen. Um diese vergleichen zu können, empfiehlt sich ein strukturierter Fragebogen. Idealerweise gibt es aufgabenspezifische Fragen, die die folgenden Vorschläge erweitern können:

- Warum hat sich die KI für dieses Ergebnis entschieden? Interpretieren Sie die Erklärung in Ihren eigenen Worten.
  - Welche Merkmale hatten großen Einfluss auf die Entscheidung der KI?
  - Welche Merkmale hatten geringen Einfluss auf die Entscheidung der KI?

- Gab es Merkmale, die zu großen Einfluss auf die Entscheidung der KI hatten? Wenn ja, welche?
- Gab es Merkmale, die zu geringen Einfluss auf die Entscheidung der KI hatten? Wenn ja, welche?
- Welche Merkmale müssten wie verändert werden, damit anstelle Ergebnis A Ergebnis B erscheint? (bei kontrafaktischen Erklärungsmethoden)

Abschließend zu allen Betrachtungen kann die Erklärung durch folgende Elemente im Likert-Stil [55] evaluiert werden, um Indizien zur Praxistauglichkeit abzutasten.

1. Ich habe das Gefühl, die Erklärung verstanden zu haben.
2. Ich habe lange gebraucht, um die Erklärung zu verstehen.
3. Ich finde es anstrengend, mich mit dieser Erklärung zu befassen.
4. Ich finde den Detailgrad der Erklärung angemessen.
5. Ich finde die Erklärung intuitiv.
6. Ich finde die Interaktion mit der Erklärung intuitiv (Dieser Bereich ist von der Interaktivität in den jeweiligen Erklärungen abhängig. Falls interaktive Elemente vorhanden sind, z. B. um den Detailgrad der Merkmale zu verändern, können diese auch nochmal genauer und aufgabenspezifisch abgefragt werden.)
7. Ich kann mir vorstellen, welche Merkmale für die betrachten Erklärungen besonders relevant sind.
8. Ich verstehe, wie die Merkmale in Beziehung mit anderen Merkmalen stehen.

Bei der Nutzung der Items muss beachtet werden, dass eine hohe Zustimmung nicht in allen Fällen ein gutes Nutzungsergebnis bedeutet (vgl. Item zwei und drei).

5.3.3 Gestaltung von xAI-Visualisierungen

Im Rahmen der Studie diskutierten die Autor*innen in mehreren User-Experience- und Usability-Fachkreisen über nutzerzentrierte xAI-Methoden und entsprechende Gütekriterien und führten gemeinsame Workshops durch, z. B. auf dem World Usability Day Stuttgart, einem Barcamp auf der Mensch und Computer Konferenz und bei UX Lean Coffees der German UPA [56]. Daraus resultierten mögliche Facilitatoren sowie die folgenden Überlegungen, um xAI-Visualisierungen sinnvoll zu gestalten.

Nutzerführung/Interaktivität
Wird eine explorative oder geleitete Darstellung von Erklärungsmethoden bevorzugt? Ob die Erklärungsmethode eher explorativ oder verstärkt systemgeleitet dargestellt werden sollte, ist stark vom Nutzungskontext abhängig. Gestaltprinzipien [57] helfen, Wahrnehmungen zu len-

ken und mentale Modelle zu nutzen. Zudem kann eine schrittweise Darstellung der Erklärungsbereiche in Betracht gezogen werden. Dies reduziert eine visuelle Überreizung, indem schrittweise der jeweilige Bereich erklärt, eindeutige Hinweise in Legenden gegeben und vor allem interaktive Möglichkeiten genutzt werden. Zuletzt ist der Abstraktionsgrad der Darstellung zu hinterfragen. Welche bzw. wie viele Merkmale sollten Nutzende sehen? Sollten konkrete Bezeichnungen interaktiv oder ausführlich bereitstehen? Ob diese unterstützend oder eher reizüberflutend wirken, muss im Nutzungskontext getestet werden.

Bereitstellung von Beispielen
Es kann sinnvoll sein, beispielhaft Ergebnisse zu erklären, bei denen die Merkmale eindeutig und greifbar sind.

Mentale Modelle/Kultur
Bei der Gestaltung von Erklärungen sollten mentale Modelle von Nutzenden berücksichtig werden, unter anderem hinsichtlich der verwendeten Diagrammart. So kann beispielsweise Bias vermieden werden, wenn man grundsätzliche Heuristiken für Datenvisualisierung beachtet, z. B. beim Umgang von nicht linearen Achsenskalierungen. Des Weiteren ist auch auf interkulturelle Unterschiede in der Datenvisualisierung zu achten. Eine Hilfestellung hierfür bietet z. B. die Checkliste zur Berücksichtigung interkultureller Aspekte in der menschzentrierten Gestaltung [58].

Barrierefreiheit
Ein weiteres wichtiges Element bei der Darstellung von Erklärungen ist die Barrierefreiheit. Beispielsweise könnten Nutzende Einschränkungen hinsichtlich der Wahrnehmung von Farben und Kontrasten haben. Aus diesem Grund sollten grundsätzliche Prinzipien zur barrierefreien Gestaltung eingehalten werden. Anwendbar ist beispielsweise die gemeinsame Verwendung von additiven und subtraktiven Farbräumen und entsprechenden Einstellungsmöglichkeiten. Des Weiteren sollten Grafiken und Tabellen mit textuellen Beschreibungen in leichter Sprache versehen werden. Mit der alternativen Erklärungsmethode können Nutzende ihre eigene Interpretation prüfen und Barrierefreiheit wird gefördert.

Übergeordnetes Ziel (Effektivität/Effizienz)
Grundsätzlich sollte stets die Frage nach dem übergeordneten Ziel der Anwendung beantwortet werden. Steht Effizienz oder Effektivität im Vordergrund? Hierbei gilt es, etwa zu prüfen, ob hinsichtlich Effizienz Zeit wirklich der wesentliche Faktor in der Interaktion mit der Visualisierung ist oder ob andere Kriterien wie der menschliche Aufwand oder die verbrauchten Ressourcen ebenfalls betrachtet werden. Anschließend kann untersucht werden, welche zusätzlichen Hilfestellungen hierbei möglich sind. Um die Effizienz zu testen, bietet es sich beispielsweise an, mit

einem Zeit-MVP (Minimum Viable Product) in der Visualisierung zu arbeiten. Dieses adressiert die Frage: Was wäre, wenn Nutzende nur wenige Sekunden Zeit hätten, um zu entscheiden, ob die KI sinnvoll entschieden hat oder nicht?

Ästhetik/Design
Zuletzt spielt die Ästhetik der dargestellten Erklärung eine zentrale Rolle. Ein ansprechendes Design hilft Nutzenden, sich komplexen Datenkonstrukten zu öffnen und Zusammenhänge zu erschließen. Schlechtes Design blockiert. Zum Testen der subjektiven Wahrnehmung kann beispielsweise VisAWI [59] genutzt werden.

5.4 Tipps und Tricks

Wie aus der Studie hervorgeht, ist die Auswahl geeigneter xAI-Methoden für einen bestimmten Anwendungsfall nicht trivial. Besonders wichtig sind hierbei die spezifischen Gegebenheiten der zu erklärenden Anwendung. Dies umfasst sowohl die Art der verwendeten Daten und Modelle als auch die Anforderungen an die xAI-Methoden und Erklärungen selbst. Das in Kapitel 5.1 vorgestellte Benchmarking adressiert vor allem die Bewertung von xAI-Verfahren und Erklärungen. Mithilfe der Ergebnisse des Benchmarkings lassen sich unterschiedliche xAI-Methoden hinsichtlich diverser Gütekriterien und Eigenschaften einordnen und bewerten.

Um das Methodenangebot bereits im Voraus eingrenzen können, können gezielte Fragen den Auswahlprozess hinsichtlich der zu erklärenden Anwendung unterstützen. Die im Folgenden aufgelisteten Fragen – abgeleitet von den in Kapitel 3.2 beschriebenen Eigenschaften von Erklärungsmethoden – helfen, Anforderungen zu charakterisieren:

- Soll ein Klassifikations- oder ein Regressionsmodell erklärt werden?
- Sind einzelne Entscheidungen oder allgemeine Modellzusammenhänge zu erklären? [lokal, global]
- Ist Modellunabhängigkeit [modellagnostisch] oder ist ein speziell für das vorliegende Modell ausgelegtes Verfahren gewünscht? [modellspezifisch]
  - Ist das Modell differenzierbar?
- Welcher Datentyp soll erklärt werden? [Bilddaten, tabellarische Daten, beides]
  - Tabellarische Daten: Sind kategorische Daten vorhanden? [kategorisch, gemischt, numerisch]
- Welcher Art sollen die Erklärungen sein? [Merkmalsrelevanz, Regeln, Datenpunkt, Surrogat]

Auf Basis dieser Fragen kann ein strukturierter Auswahlprozess in Form eines Graphen erstellt werden. In Abbildung 18 ist ein Beispiel für einen Entscheidungsbaum zur Methodeneingrenzung dargestellt. Die dort aufgeführten Fragen und xAI-Methoden sind zugeschnitten auf den in Kapitel 4.1 definierten Studienumfang sowie die in Kapitel 4.2.2 vorgestellten xAI-Methoden. Für Anwendungsfälle mit anderen Anforderungen, z. B. die Betrachtung von Zeitreihendaten, sind die oben aufgelisteten Fragen dementsprechend zu modifizieren oder zu erweitern sowie weitere xAI-Verfahren zu berücksichtigen. Somit deckt der in Abbildung 18 gezeigte Auswahlprozess lediglich einen Teilbereich möglicher xAI-Methoden und KI-Anwendungsfälle ab.

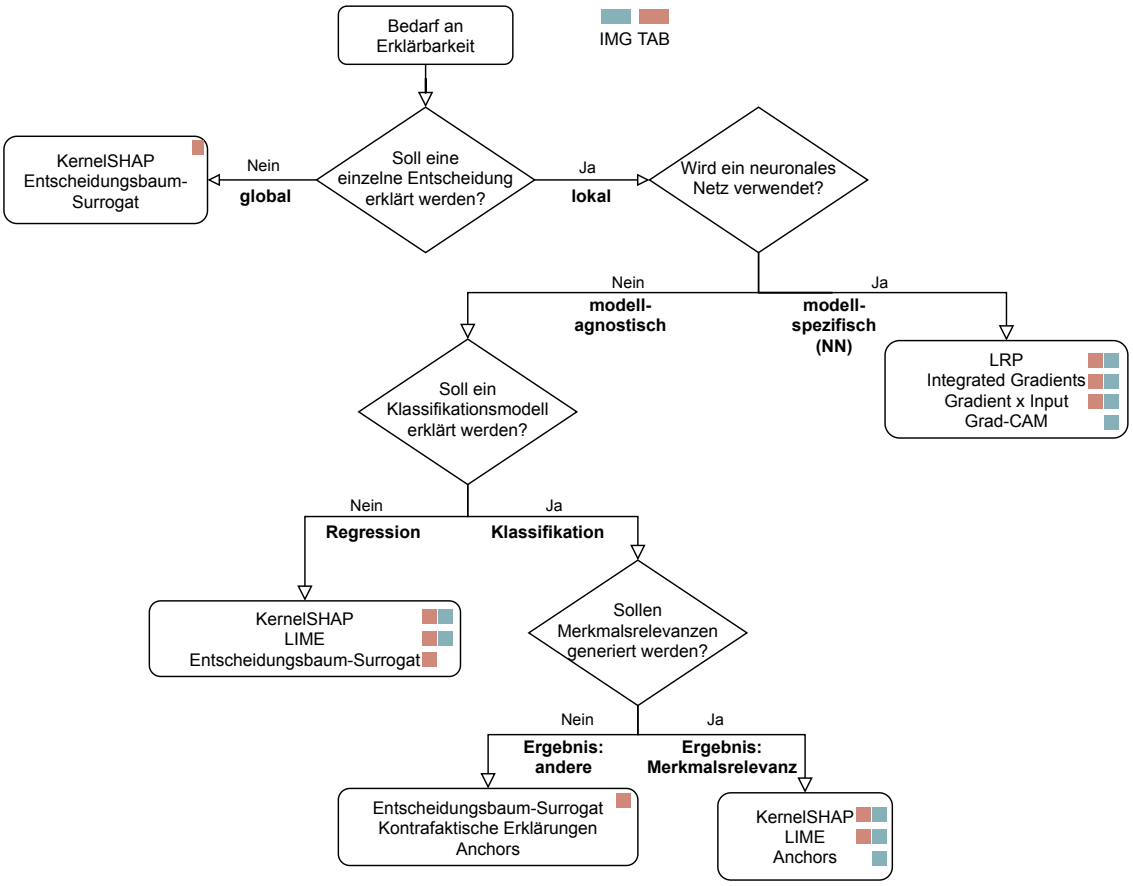

Abbildung 18: Entscheidungsbaum als Hilfestellung zur Auswahl von xAI-Verfahren. Bild: Fraunhofer IPA

6 FAZIT

In der vorliegenden Studie wurde untersucht, anhand welcher Kriterien xAI-Methoden für den eigenen Anwendungsfall eingegrenzt und ausgewählt werden können. Hierzu analysierte die Studie drei Teilfragestellungen: die quantitative Evaluation von xAI-Methoden, die Beleuchtung der softwareseitigen Rahmenbedingungen existierender Softwarebibliotheken sowie die Frage, wie verständlich visualisierte Erklärungen sind.

Allen drei Fragestellungen ist gemein, dass die Wahl geeigneter xAI-Methoden stets vor allem vom betrachteten Anwendungsfall und dessen Anforderungen abhängt. Es ist daher unerlässlich, bei der Auswahl von xAI-Methoden die spezifischen Rahmenbedingungen zu berücksichtigen.

Die im Rahmen des Benchmarkings durchgeführte quantitative Evaluation ist ein erster Schritt, um die Güte verschiedener Erklärungsverfahren für unterschiedliche Anwendungen vergleichbar zu machen. Obwohl nicht alle Gütekriterien anwendungsübergreifend vergleichbar sind, können die Ergebnisse dennoch bei der Einordnung der Methoden helfen. Auch die in Kapitel 5.2 dargestellten technischen Rahmenbedingungen verschiedener Open-Source xAI-Softwarebibliotheken können die Methodenauswahl für die eigene Anwendung strukturieren und vereinfachen. Die Studie zeigt jedoch auch auf, dass gerade bei der quantitativen Evaluation von xAI-Methoden die Entwicklung weiterer – bestenfalls anwendungsübergreifender – Gütekriterien erforderlich ist.

Lässt sich die Erklärungsgüte verlässlich bewerten, ist dies ein essentieller Faktor für ein ausgeprägtes Vertrauen in die Methoden. Des Weiteren hat die Studie gezeigt, dass das Fehlen einer einheitlichen Programmierschnittstelle die Anwendung von xAI-Methoden unter Umständen erschweren kann. Um diesem Problem zu begegnen, entwickelt das Fraunhofer IPA derzeit eine Toolbox, die mehrere xAI-Methoden unter einer einheitlichen Schnittstelle bündelt.

Die Ergebnisse des Benchmarkings (vgl. Tabelle 4) und Vergleichs von Softwarebibliotheken können bei der Auswahl von xAI-Methoden für den eigenen Anwendungsfall helfen. Die mithilfe dieser Methoden generierten Erklärungen müssen jedoch auch für Anwender*innen verständlich sein. Es ist daher wichtig, ebenfalls ein besonderes Augenmerk darauf zu legen, wie die generierten Erklärungen dargestellt werden. Wenn es um die Gestaltung von Erklärungsvisualisierungen geht, muss immer die gesamte User Experience (UX) rund um die Erklärungsmethoden evaluiert werden, um langfristig positive Nutzungsergebnisse zu erzielen. Eine gute UX führt dazu, dass Nutzende eher den Willen haben, die Entscheidungen von KI-Anwendungen verstehen zu wollen. Zudem ermöglicht eine gute UX, KI-Modelle langfristig zu verbessern.

Zusätzlich ist zu beachten, dass Effizienz, Effektivität und Zufriedenheit die Praxistauglichkeit der xAI-Methoden beschreiben und daher technische oder nutzerzentrierte Anforderungen nicht trennbar sind. Die Wiedergabetreue einer Erklärung ist bei der Auswahl aus beiden Betrachtungswinkeln wichtig, da die Ziele der Nutzende in diesem Zusammenhang auch mit den Businesszielen korrelieren. Um in Zukunft also mit nutzerzentrierten Erklärungsmethoden arbeiten zu können, müssen Entwickler*innen und Designer*innen noch stärker zusammenarbeiten.

7 LITERATURVERZEICHNIS

[1] OpenAI, *OpenAI Five Defeats Dota 2 World Champions.* [Online]. Verfügbar unter: https://openai.com/blog/openai-five-defeats-dota-2-world-champions/ (Zugriff am: 16. November 2020).

[2] High-Level Expert Group on AI, »Ethics guidelines for trustworthy AI«. Report, European Commission, Brussels, 2019. [Online]. Verfügbar unter: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Zugriff am: 23. November 2020.

[3] J. Angwin, J. Larson, S. Mattu und L. Kirchner, *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.* [Online]. Verfügbar unter: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Zugriff am: 16. November 2020).

[4] J. R. Zech, *What are radiological deep learning models actually learning?* [Online]. Verfügbar unter: https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98 (Zugriff am: 23. November 2020).

[5] S. Levin und J. Carrie, *Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian*. [Online]. Verfügbar unter: https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe (Zugriff am: 16. November 2020).

[6] M. Brandt, *Künstliche Intelligenz rechnet sich.* [Online]. Verfügbar unter: https://de.statista.com/infografik/16992/umsatz-der-in-deutschland-durch-ki-anwendungen-beeinflusst-wird/ (Zugriff am: 25. Februar 2021).

[7] A. Berg und S. Dehmel, »Künstliche Intelligenz: Von der Strategie zum Handeln«. Berlin, 28. Sep. 2020.

[8] N. Burkart und M. F. Huber, »A Survey on the Explainability of Supervised Machine Learning«, *jair,* Jg. 70, S. 245–317, 2021, doi: 10.1613/jair.1.12228.

[9] A. Adadi und M. Berrada, »Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)«, *IEEE* Access, Jg. 6, S. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[10] R. Tomsett, D. Braines, D. Harborne, A. Preece und S. Chakraborty, »Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems« in *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018),* Stockholm, Sweden, 2018, S. 8–14.

[11] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter und L. Kagal, »Explaining Explanations: An Overview of Interpretability of Machine Learning«, 2019. arXiv:1806.00069.

[12] P. Biecek und T. Burzykowski, *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. With examples in R and Python.* Chapman and Hall/CRC, New York, 2021. [Online]. Verfügbar unter: https://pbiecek.github.io/ema/

[13] Z. C. Lipton, »The Mythos of Model Interpretability«, 2016. arXiv:1606.03490v3.

[14] B. Kim, R. Khanna und O. Koyejo, »Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability« in *Proceedings of the 30th International Conference on Neural Information Processing Systems,* 2016, S. 2288–2296.

[15] K. P. Murphy, *Machine learning: A probabilistic perspective.* Cambridge, Mass.: MIT Press, 2012.

[16] T. Kraus, L. Ganschow, M. Eisenträger und S. Wischmann, »Erklärbare KI – Anforderungen, Anwendungsfälle und Lösungen«, Institut für Innovation und Technik, Berlin, 2021.

[17] *lime.* 0.2.0.0, 2021. [Online]. Verfügbar unter: https://github.com/marcotcr/lime

[18] D. V. Carvalho, E. M. Pereira und J. S. Cardoso, »Machine Learning Interpretability: A Survey on Methods and Metrics«, *Electronics*, Jg. 8, Nr. 8, 2019, Art. no. 832, doi: 10.3390/electronics8080832.

[19] F. Doshi-Velez und B. Kim, »Towards A Rigorous Science of Interpretable Machine Learning«, 2017. arXiv:1702.08608.

[20] X. Wu, M. El-Shamouty und P. Wagner, »Zuverlässige KI – KI in sicherheitskritischen industriellen Anwendungen einsetzen«, Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Stuttgart, 2021.

[21] W. Samek, A. Binder, G. Montavon, S. Lapuschkin und K.-R. Müller, »Evaluating the Visualization of What a Deep Neural Network Has Learned« (eng), *IEEE transactions on neural networks and learning systems,* Jg. 28, Nr. 11, S. 2660–2673, 2017, doi: 10.1109/TNNLS.2016.2599820.

[22] A. Osman, L. Arras und W. Samek, »Towards Ground Truth Evaluation of Visual Explanations«, 2020. arXiv:2003.07258v1.

[23] D. Dua und C. Graff, *UCI Machine Learning Repository*. [Online]. Verfügbar unter: http://archive.ics.uci.edu/ml.

[24] V. Lohweg *et al.*, »Sensorlose Zustandsüberwachung an Synchronmotoren« in *Schriftenreihe des Instituts für Angewandte Informatik, Automatisierungstechnik am Karlsruher Institut für Technologie,* Bd. 46, *Proceedings of the 23rd Workshop Computational Intelligence: Dortmund, 5. – 6. Dezember 2013*, F. Hoffmann und E. Hüllermeier, Hg., Karlsruhe: KIT Scientific Publ, 2013, S. 211–225.

[25] Y. LeCun, L. Bottou, Y. Bengio und P. Haffner, »Gradient-based learning applied to document recognition«, *Proc. IEEE,* Jg. 86, Nr. 11, S. 2278–2324, 1998, doi: 10.1109/5.726791.

[26] O. Russakovsky et al., »ImageNet Large Scale Visual Recognition Challenge«, *International Journal of Computer Vision (IJCV),* Jg. 115, Nr. 3, S. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[27] S. M. Lundberg und S.-I. Lee, »A Unified Approach to Interpreting Model Predictions« in *Advances in Neural Information Processing Systems 30,* I. Guyon et al., Hg., Curran Associates, Inc., 2017, S. 4765–4774.

[28] M. T. Ribeiro, S. Singh und C. Guestrin, »"Why Should I Trust You?": Explaining the Predictions of Any Classifier«, 2016. arXiv:1602.04938.

[29] M. Sundararajan, A. Taly und Q. Yan, »Axiomatic Attribution for Deep Networks« in *Proceedings of the 34th International Conference on Machine Learning – Volume 70,* 2017, S. 3319–3328.

[30] S. M. Lundberg *et al.*, »Explainable machine-learning predictions for the prevention of hypoxaemia during surgery«, *Nature Biomedical Engineering,* Jg. 2, Nr. 10, S. 749, 2018.

[31] A. Binder, S. Bach, G. Montavon, K.-R. Müller und W. Samek, »Layer-wise Relevance Propagation for Deep Neural Network Architectures« in *Lecture Notes in Electrical Engineering, Information Science and Applications (ICISA) 2016,* Kuinam J. Kim und Nikolai Joukov, Hg., Singapore: Springer Singapore, 2016, S. 913–922, doi: 10.1007/978-981-10-0557-2_87.

[32] A. Shrikumar, P. Greenside und A. Kundaje, »Learning Important Features Through Propagating Activation Differences«, 2017. arXiv:1704.02685v2.

[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh und D. Batra, »Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization«, 2016. arXiv:1610.02391v3.

[34] M. T. Ribeiro, S. Singh und C. Guestrin, »Anchors: High-Precision Model-Agostic Explanations« in AAAI, New Orleans, Louisiana, USA, 2018, S. 1527–1535.

[35] S. Wachter, B. Mittelstadt und C. Russell, »Counterfactual explanations without opening the black box: automated decisions and the GDPR«, *Harvard Journal of Law and Technology,* Jg. 31, Nr. 2, S. 841–887, 2018.

[36] R. K. Mothilal, A. Sharma und C. Tan, »Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations« in *Conference on Fairness, Accountability, and Transparency,* 2020, S. 607–617.

[37] F. Pedregosa et al., »Scikit-learn: Machine Learning in Python«, *Journal of Machine Learning Research,* Jg. 12, S. 2825–2830, 2011.

[38] K. Sokol und P. Flach, »Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches« in FAT* '20: *Conference on Fairness, Accountability, and Transparency,* Barcelona Spain, 2020, S. 56–67, doi: 10.1145/3351095.3372870.

[39] M. R. Honegger, »Shedding Light on Black Box Machine Learning Algorithms«. Master Thesis, Institute of Information Systems and Marketing (IISM), Karlsruhe Institute of Technology, Karlsruhe, 2018.

[40] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2019. [Online]. Verfügbar unter: https://christophm.github.io/interpretable-ml-book/

[41] Z.-H. Zhou, »Rule extraction: Using neural networks or for neural networks?«, *J. Comput. Sci. & Technol.,* Jg. 19, Nr. 2, S. 249–253, 2004, doi: 10.1007/BF02944803.

[42] J. Zhang, C. Petitjean, F. Yger und S. Ainouz, »Explainability for Regression CNN in Fetal Head Circumference Estimation from Ultrasound Images« in *Lecture Notes in Computer Science, Interpretable and Annotation-Efficient Learning for Medical Image Computing,* J. S. Cardoso et al., Hg., Cham: Springer International Publishing, 2020, S. 73–82, doi: 10.1007/978-3-030-61166-8_8.

[43] T. Hastie, R. Tibshirani und J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* 12. Aufl. New York, NY: Springer, 2017.

[44] I. Goodfellow, Y. Bengio und A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Verfügbar unter: http://www.deeplearningbook.org

[45] H. Nori, S. Jenkins, P. Koch und R. Caruana, »InterpretML: A Unified Framework for Machine Learning Interpretability«, 2019. arXiv:1909.09223.

[46] V. Arya et al., »One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques«, 2019. arXiv:1909.03012v2.

[47] *skater.* 1.1.2. Oracle, 2018. [Online]. Verfügbar unter: https://github.com/oracle/Skater

[48] *Alibi: Algorithms for monitoring and explaining machine learning models,* 2019. [Online]. Verfügbar unter: https://github.com/SeldonIO/alibi

[49] Maximilian Alber et al., »iNNvestigate Neural Networks!«, *Journal of Machine Learning Research,* Jg. 20, Nr. 93, S. 1–8, 2019. [Online]. Verfügbar unter: http://jmlr.org/papers/v20/18-540.html

[50] M. Ancona, E. Ceolini, C. Öztireli und M. Gross, »Towards better understanding of gradient-based attribution methods for Deep Neural Networks«, 2017. arXiv:1711.06104.

[51] *tf-explain.* 0.2.1. Sicara SAS, 2020. [Online]. Verfügbar unter: https://github.com/sicara/tf-explain

[52] *Captum.* 0.3.0. Facebook Inc., 2020. [Online]. Verfügbar unter: https://captum.ai/

[53] K. He, X. Zhang, S. Ren und J. Sun, »Deep Residual Learning for Image Recognition« in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2016, S. 770–778, doi: 10.1109/CVPR.2016.90.

[54] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov und L.-C. Chen, »MobileNetV2: Inverted Residuals and Linear Bottlenecks« in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 2018, S. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[55] R. Likert, »A technique for the measurement of attitudes« in Bd. 22, *Archives of Psychology,* R. S. Woodworth, Hg., New York, USA, 1932, S. 5–55.

[56] S. J. Wiedenroth und N. Schaaf, »Workshop: Usability von erklärbarer KI (XAI)« in *world usability day,* Stuttgart, 2020.

[57] M. Wertheimer, »Untersuchungen zur Lehre von der Gestalt« in *Psychologische Forschung: Zeitschrift für Psychologie und ihre Grenzwissenschaften,* K. Koffka, W. Köhler, M. Wertheimer, K. Goldstein und H. Gruhle, Hg., Berlin: Justus Springer, 1923, S. 301–350.

[58] R. Heimgärtner, A. Beck, K. Proschek, A. Solanki, O. Lange und M. Kostrubov, »Berücksichtigung interkultureller Aspekte in der menschzentrierten Gestaltung: Erläuterungen zur "Checkliste 2019" des AK Interkulturalität der German UPA«, Gesellschaft für Informatik e.V. und die German UPA, Hamburg, 2019.

[59] M. Moshagen und M. Thielsch, »A short version of the visual aesthetics of websites inventory«, *Behaviour & Information Technology,* Jg. 32, Nr. 12, S. 1305–1311, 2013, doi: 10.1080/0144929X.2012.694910.

[60] *keras-gradcam,* 2019. [Online]. Verfügbar unter: https://github.com/eclique/keras-gradcam

[61] H. Zhang, J. Chen, H. Xue und Q. Zhang, »Towards a Unified Evaluation of Explanation Methods without Ground Truth«, 2019. arXiv:1911.09017v1.

[62] V. Lai, J. Z. Cai und C. Tan, »Many Faces of Feature Importance: Comparing Built-in and Post-hoc Feature Importance in Text Classification«, 2019. arXiv:1910.08534v1.

KI-FORTSCHRITTSZENTRUM

Das KI-Fortschrittszentrum »Lernende Systeme« unterstützt Firmen dabei, die wirtschaftlichen Chancen der Künstlichen Intelligenz und insbesondere des Maschinellen Lernens für sich zu nutzen. In anwendungsnahen Forschungsprojekten und in direkter Kooperation mit Industrieunternehmen arbeiten die Stuttgarter Fraunhofer-Institute für Arbeitswirtschaft und Organisation IAO sowie für Produktionstechnik und Automatisierung IPA daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg.

Europas größte Forschungskooperation auf dem Gebiet der KI

Das KI-Forschungszentrum ist Forschungspartner des Cyber Valley, einem Konsortium aus den renommierten Universitäten Tübingen und Stuttgart, dem Max-Planck-Institut für intelligente Systeme und einigen führenden Industrieunternehmen. In gemeinsamen Forschungslabors werden Grundlagenforschung und anwendungsorientierte Entwicklung zu aktuellen wie auch zukünftigen Bedarfen behandelt und vorangetrieben.

Menschzentrierte KI

Alle Aktivitäten des Zentrums verfolgen das Ziel, eine menschzentrierte KI zu entwickeln, der die Menschen vertrauen und die sie akzeptieren. Nur wenn Menschen mit neuen Technologien intuitiv interagieren und vertrauensvoll zusammenarbeiten, kann deren Potenzial optimal ausgeschöpft werden. Daher konzentrieren sich die Forschungsaktivitäten unter anderem auf die Themen Erklärbarkeit, Datenschutz, Sicherheit und Robustheit von KI-Technologien.

Studienreihe »Lernende Systeme«

Die Studienreihe »Lernende Systeme« gibt Einblick in die Potenziale und die praktischen Einsatzmöglichkeiten von KI. Nähere Informationen und die aktuellen Versionen der Studien finden Sie unter: www.ki-fortschrittszentrum.de/studien

FRAUNHOFER-GESELLSCHAFT

Die Fraunhofer-Gesellschaft mit Sitz in Deutschland ist die weltweit führende Organisation für anwendungsorientierte Forschung. Mit ihrer Fokussierung auf zukunftsrelevante Schlüsseltechnologien sowie auf die Verwertung der Ergebnisse in Wirtschaft und Industrie spielt sie eine zentrale Rolle im Innovationsprozess. Sie ist Wegweiser und Impulsgeber für innovative Entwicklungen und wissenschaftliche Exzellenz. Mit inspirierenden Ideen und nachhaltigen wissenschaftlich-technologischen Lösungen fördert die Fraunhofer-Gesellschaft Wissenschaft und Wirtschaft und wirkt mit an der Gestaltung unserer Gesellschaft und unserer Zukunft.

Interdisziplinäre Forschungsteams der Fraunhofer-Gesellschaft setzen gemeinsam mit Vertragspartnern aus Wirtschaft und öffentlicher Hand originäre Ideen in Innovationen um, koordinieren und realisieren systemrelevante, forschungspolitische Schlüsselprojekte und stärken mit werteorientierter Wertschöpfung die deutsche und europäische Wirtschaft. Internationale Kooperationen mit exzellenten Forschungspartnern und Unternehmen weltweit sorgen für einen direkten Austausch mit den einflussreichsten Wissenschafts- und Wirtschaftsräumen.

Die 1949 gegründete Organisation betreibt in Deutschland derzeit 75 Institute und Forschungseinrichtungen. Rund 29 000 Mitarbeiterinnen und Mitarbeiter, überwiegend mit natur- oder ingenieur-wissenschaftlicher Ausbildung, erarbeiten das jährliche Forschungsvolumen von 2,8 Milliarden Euro. Davon fallen 2,4 Milliarden Euro auf den Leistungsbereich Vertragsforschung. Rund zwei Drittel davon erwirtschaftet Fraunhofer mit Aufträgen aus der Industrie und mit öffentlich finanzierten Forschungsprojekten. Rund ein Drittel steuern Bund und Länder als Grundfinanzierung bei, damit die Institute schon heute Problemlösungen entwickeln können, die in einigen Jahren für Wirtschaft und Gesellschaft entscheidend wichtig werden.

Die Wirkung der angewandten Forschung geht weit über den direkten Nutzen für die Auftraggeber hinaus: Fraunhofer-Institute stärken die Leistungsfähigkeit der Unternehmen, verbessern die Akzeptanz moderner Technik in der Gesellschaft und sorgen für die Aus- und Weiterbildung des dringend benötigten wissenschaftlich-technischen Nachwuchses.

Hochmotivierte Mitarbeiterinnen und Mitarbeiter auf dem Stand der aktuellen Spitzenforschung stellen für uns als Wissenschaftsorganisation den wichtigsten Erfolgsfaktor dar. Fraunhofer bietet daher die Möglichkeit zum selbstständigen, gestaltenden und zugleich zielorientierten Arbeiten und somit zur fachlichen und persönlichen Entwicklung, die zu anspruchsvollen Positio-

nen in den Instituten, an Hochschulen, in Wirtschaft und Gesellschaft befähigt. Studierenden eröffnen sich aufgrund der praxisnahen Ausbildung und des frühzeitigen Kontakts mit Auftraggebern hervorragende Einstiegs- und Entwicklungschancen in Unternehmen.

Namensgeber der als gemeinnützig anerkannten Fraunhofer-Gesellschaft ist der Münchner Gelehrte Joseph von Fraunhofer (1787–1826). Er war als Forscher, Erfinder und Unternehmer gleichermaßen erfolgreich.

Fraunhofer IPA

Das **Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA**, kurz Fraunhofer IPA, ist mit annähernd 1000 Mitarbeiterinnen und Mitarbeitern eines der größten Institute der Fraunhofer-Gesellschaft. Der gesamte Haushalt beträgt 76 Millionen Euro. Forschungsschwerpunkte des Instituts sind organisatorische und technologische Aufgaben aus der Produktion. Methoden, Komponenten und Geräte bis hin zu kompletten Maschinen und Anlagen werden entwickelt, erprobt und umgesetzt. 15 Fachabteilungen arbeiten interdisziplinär, koordiniert durch 6 Geschäftsfelder, vor allem mit den Branchen Automotive, Maschinen- und Anlagenbau, Elektronik und Mikrosystemtechnik, Energie, Medizin- und Biotechnik sowie Prozessindustrie zusammen. Das Fraunhofer IPA orientiert seine Forschung an der wirtschaftlichen Produktion nachhaltiger und personalisierter Produkte.

»Zentrum für Cyber Cognitive Intelligence«

Das **Zentrum für Cyber Cognitive Intelligence CCI** des Fraunhofer IPA ist ein industrienaher Forschungs- und Entwicklungspartner für die Umsetzung von Applikationen im Bereich Künstliche Intelligenz (KI) und insbesondere Maschinelles Lernen (ML) in der produzierenden Industrie. Ziel des Zentrums ist es, sowohl die KI-Forschung als auch den Technologietransfer von KI und ML in die Anwendung voranzutreiben.

Durch die Vernetzung von Produktionsanlagen und die fortschreitende Digitalisierung werden Daten in großen Mengen verfügbar. Diese Daten werden zunehmend mit ML- bzw. KI-basierten Verfahren ausgewertet und nutzbar gemacht. Dies bietet beachtliche Vorteile für die Industrie: Zum einen sind Leistungssprünge in der Nutzung von Maschinen und Anlagen in Bezug auf Qualität, Flexibilität und Effizienz zu erwarten. Zum anderen entstehen neue Automatisierungslösungen. Hierbei werden nicht zuletzt mit ML ausgestattete Roboter vermehrt Einzug in alle Arbeits- und Alltagsumgebungen halten.

Team »Zuverlässige KI-Systeme«

Der vollumfängliche Einsatz von KI-Funktionen in immer mehr Anwendungsbereichen wie z. B. kollaborierenden Robotersystemen, autonomem Fahren, Medizintechnik oder digitalisierter Produktion stellt vermehrt hohe Anforderungen an die funktionale Sicherheit, Nachvollzieh-

barkeit und Akzeptanz. Die Gruppe »Zuverlässige KI-Systeme« des Fraunhofer IPA forscht und entwickelt Methoden zur Bewerkstelligung erklärbarer, verifizierbarer und robuster maschineller Lernverfahren mit dem Ziel, Vertrauen in KI-Lösungen zu stärken und Unternehmen bei der Umsetzung, Implementierung und Absicherung von KI-Funktionalitäten zu helfen.

ANHANG

Benchmarking

Datensätze

Tabelle 6 bietet eine Übersicht über die Eigenschaften der im Benchmarking verwendeten Datensätze.

Datensatz	Datentyp	Anz. Daten	Anz. verwendete Daten	Anz. Features bzw. Bildgröße	Lernaufgabe	Anz. Klassen
Adult	Tabellarisch (numerisch, kategorisch)	48.842	2000	14	Klassifikation	2
Sensorless Drive Diagnosis	Tabellarisch (numerisch)	58.509	2000	48	Klassifikation	11
Bike Sharing	Tabellarisch (numerisch, kategorisch)	17.389	2000	16	Regression	–
MNIST	Bild	70.000	10.000	28 x 28 x 1	Klassifikation	10
ImageNet	Bild	1.25 Mio	1000	224 x 224 x 3	Klassifikation	1000

Tabelle 6: Übersicht über die für das Benchmarking verwendeten Datensätze.

Modelle

Tabelle 7 fasst Art und Performance der im Benchmarking genutzten ML-Modelle zusammen. Zusätzlich ist die Anzahl der Parameter angegeben, um die ermittelten Laufzeiten (s. Kapitel 4.1) zur Erklärungsgenerierung angemessen einordnen zu können.

	Datensatz	MI-Modell	Accuracy (Test)	MAE (Test)	Anzahl Parameter
TAB	Bike Sharing	Random Forest	–	24.76	300 Bäume
	Adult	Neuronales Netz	85.14 %	–	ca. 3.500 Gewichte
	Sensorless Drive Diagnosis	Neuronales Netz	99.4 %	–	ca. 20.000 Gewichte
IMG	MNIST	Neuronales Netz	98.8 %	–	ca. 130.000 Gewichte
	ImageNet I	Neuronales Netz (MobileNetV2)	74.7%	–	ca. 3.5 Mio. Gewichte

Tabelle 7: Spezifikationen der für das Benchmarking eingesetzten ML-Modelle.

xAI-Methoden

Zur Durchführung des Benchmarkings von xAI-Methoden kamen folgende Implementierungen zum Einsatz:

- **Integrated Gradients, ε-LRP, Gradient × Input:** iNNvestigate [49].
- **LIME** [28], **KernelSHAP** [27], **Anchors** [34]: Repositories der Autoren.
- **Kontrafaktische Erklärungen**: alibi [48]. Speziell wurden die Module »CounterFactual« (Datensatz »Sensorless Drive Diagnosis«) und »CounterfactualProto« (Datensatz »Adult«) verwendet.
- **Grad-CAM**: Git-Repository [60].
- **Entscheidungsbäume**: Softwarebibliothek scikit-learn [37].

Details zur Parametrierung der einzelnen Methoden sind in Tabelle 8 aufgeführt. Für diejenigen Verfahren, die nicht in der Tabelle gelistet sind (*Grad-CAM, Gradient × Input, Entscheidungsbäume*), wurden die Standardeinstellungen der jeweiligen Implementierungen verwendet.

Methode	ImageNet	MNIST	Adult	Drive	Bike
LIME	n_steps: 100, segmentation: slic (skimage)	n_steps: 1000, segmentation: slic (skimage)	n_steps: 1000	n_steps: 1000	n_steps: 1000
KernelSHAP	n_steps: 100, segmentation: slic (skimage)	n_steps: 1000, segmentation: slic (skimage)	n_steps: 1000	n_steps: 1000	n_steps: _1000
ε-LRP	ε: 0.01	ε: 0.01	–	–	–
IG	n_steps: 64, background: black	n_steps: 64, background: black	–	–	–
Anchors	–	–	threshold: 0.95	threshold: 0.95	–
Kontrafaktische Erklärungen	–	–	–	max_iter: 500 lam_init: 0.1 c_init: 1 c_steps: 5	max_iter: 500 lam_init: 0.1 target_proba: 0.9 tol: 0.1 max_lam_steps: 10 learning_rate_init: 0.1

Tabelle 8: Parametrierung der untersuchten xAI-Methoden.

Metriken

Die in Kapitel 4.2.3 beschriebenen Metriken zur Evaluation von xAI-Verfahren wurden wie folgt umgesetzt:

Stabilität: Zweimalige Generierung von Erklärungen für die gleiche Auswahl von Daten. Im Anschluss wird geprüft, ob die Erklärungen beider Durchläufe sich unterscheiden. Falls ja, ist das Stabilitätskriterium verletzt.

Trennbarkeit

- **Tabellarische Daten**: Durchsuchen des Datensatzes nach den beiden am dichtesten beieinanderliegenden Datenpunkten (gemessen an der Vektornorm zwischen beiden Datenpunkten). Für diese Datenpunkte wird jeweils eine Erklärung erstellt. Sind diese Erklärungen nicht identisch, ist das Trennbarkeits-Kriterium erfüllt.
- **Bilddaten**: Generierung von Erklärungen für jedes Bild im Datensatz. Im Anschluss wird geprüft, ob die generierten Erklärungen Duplikate enthalten. Falls ja, so ist das Trennbarkeits-Kriterium verletzt.

Konsistenz: Zweimalige Generierung von Erklärungen für die gleiche Auswahl von Daten. Im ersten Durchlauf werden Erklärungen für die Originaldaten generiert. Im zweiten Durchlauf werden die Daten minimal verändert: für alle kontinuierlichen Merkmale wird 0.1 * Merkmalsstandardabweichung auf den jeweiligen Merkmalswert addiert. Im Anschluss wird der Unterschied zwischen den Erklärungen beider Durchläufe betrachtet. Liegt die durchschnittliche Abweichung nahe Null, ist das Konsistenz-Kriterium erfüllt.

Erklärungsähnlichkeit: Generierung von Erklärungen (Relevanzwerten) für alle Daten des Datensatzes. Für Bilddaten folgt anschließend die Anwendung der in [61] eingeführten »Mutual Verfication«-Metrik. Für tabellarische Daten wird die in [62] beschriebene Jaccard-Ähnlichkeit der Top-10 Merkmale berechnet. Um Relevanzwerte für kontrafaktische Erklärungen zu berechnen, werden die Merkmalsdifferenzen zwischen den Originaldaten und den kontrafaktischen Erklärungen betrachtet.

AOPC

- **Tabellarische Daten:** Generierung von Erklärungen (Relevanzwerten) für alle Daten des Datensatzes. Nun werden für jeden Datenpunkt die Merkmalswerte schrittweise (in absteigender Reihenfolge ihrer Relevanz) verändert. Dabei werden kontinuierliche Werte durch einen Zufallswert und kategorische Daten durch ihren gegenteiligen Wert ersetzt. Nach jedem Veränderungsschritt wird für den neu entstandenen Datenpunkt eine Modellvorhersage erstellt und der Unterschied zur Originalvorhersage (Vorhersage für den ursprünglichen Datenpunkt) berechnet. So können die durchschnittlichen Vorhersageunterschiede für jeden Veränderungsschritt berechnet und aufsummiert werden (AOPC-Wert). Je höher der AOPC-Wert, desto besser bildet die Erklärung die Modellentscheidung ab.
- **Bilddaten:** Analog zum Vorgehen bei tabellarischen Daten. Der einzige Unterschied liegt in der Veränderung des Originalbildes. Bei jedem Veränderungsschritt wird eine Bildregion der Größe 9x9 Pixel (ImageNet) bzw. 1x1 Pixel (MNIST) durch Zufallswerte einer Gleichverteilung ersetzt. Insgesamt werden 100 (ImageNet) bzw. 125 (MNIST) Veränderungsschritte durchgeführt, sodass ca. 15 Prozent der Pixelwerte des Bildes ausgetauscht werden.

Regel-Wiedergabetreue: Generierung von Regeln für alle Daten des Datensatzes. Für jede Regel wird im Datensatz nach Datenpunkten gesucht, die die Regel abdeckt. Nun wird für diese Auswahl an Daten eine Modellvorhersage erstellt und mit der Vorhersage für die Regel selbst verglichen. Die Wiedergabetreue ist der Anteil der übereinstimmenden Vorhersagen zwischen der Regel und der von ihr abgedeckten Datenpunkte.

Laufzeit: Messung der Laufzeit zur Erklärungsgenerierung für 100 Datenpunkte. Im Anschluss wird über diese 100 Messwerte gemittelt.

xAI-Softwarebibliotheken

Tabelle 9 und Tabelle 10 geben Auskunft über die angebotenen xAI-Methoden der zehn untersuchten Softwarebibliotheken.

xAI-Methode	InterpretML	AIX360	Skater	Alibi	iNNvestigate
Explainable Boosting	●				
Entscheidungsbaum	●		●		
Lineare Regression	●				
Logistische Regression	●				
SHAP Kernel	●	●		●	
SHAP Tree		●		●	
SHAP Deep		●			
SHAP Gradient		●			
SHAP Linear		●			
LIME Tabular	●	●	●		
LIME Image		●			
LIME Text		●			
Morris Sensitivity Analysis	●				
Partial Dependence Plot	●		●		
LRP			●		●
Integrated Gradients			●	●	●
Scalable Bayesian Rulelist			●		
Occlusion			●		
Skater Feature Importance			●		
Accumulated Local Effects				●	
Anchors				●	
CEM		●		●	
Counterfactuals				●	
Prototype Counterfactuals				●	
Boolean Decision Rules		●			
Generalized Linear Rule Models		●			
ProfWeight		●			
Protodash		●			
DIPVAE		●			
Gradient					●
Gradient × Input					●
Deep Taylor					●

xAI-Methode	Inter-pretML	AIX360	Skater	Alibi	iNNvesti-gate
Pattern Attribution					●
DeepLIFT					●
Smoothgrad					●
DeConvNet					●
Guided Backpropagation					●
PatternNet					●
Shapely Value Sampling					
Saliency Maps					
Grad-CAM					
Guided Grad-CAM					
Feature Ablation					
Feature Permutation					

Tabelle 9: Verfügbarkeit verschiedener xAI-Methoden in den untersuchten xAI-Softwarebibliotheken InterpretML, AIX360, Skater, Alibi und iNNvestigate.

Tabelle 10: Verfügbarkeit verschiedener xAI-Methoden in den untersuchten xAI-Softwarebibliotheken DeepExplain, Tf-explain, Captum, SHAP und LIME.

xAI-Methode	Deep-Explain	TF-Explain	Captum	SHAP	LIME
Explainable Boosting					
Entscheidungsbaum					
Lineare Regression					
Logistische Regression					
SHAP Kernel				●	
SHAP Tree				●	
SHAP Deep			●	●	
SHAP Gradient			●	●	
SHAP Linear				●	
LIME Tabular					●
LIME Image					●
LIME Text					●
Morris Sensitivity Analysis					
Partial Dependence Plot					
LRP	●				
Integrated Gradients	●	●	●		
Scalable Bayesian Rulelist					

xAI-Methode	Deep-Explain	TF-Explain	Captum	SHAP	LIME
Occlusion	●	●	●		
Skater Feature Importance					
Accumulated Local Effects					
Anchors					
CEM					
Counterfactuals					
Prototype Counterfactuals					
Boolean Decision Rules					
Generalized Linear Rule Models					
ProfWeight					
Protodash					
DIPVAE					
Gradient		●			
Gradient × Input	●	●	●		
Deep Taylor					
Pattern Attribution					
DeepLIFT	●		●		
Smoothgrad		●			
DeConvNet			●		
Guided Backpropagation			●		
PatternNet					
Shapely Value Sampling	●		●		
Saliency Maps	●		●		
Grad-CAM		●			
Guided Grad-CAM			●		
Feature Ablation			●		
Feature Permutation			●		

IMPRESSUM

Kontaktadresse
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA
Nobelstraße 12, 70569 Stuttgart

Nina Schaaf
Telefon +49 711 970-1971
nina.schaaf@ipa.fraunhofer.de

Herausgeber
Thomas Bauernhansl, Marco Huber, Werner Kraus

Titelbild
© Molnia – stock.adobe.com / Fraunhofer IPA

Satz und Gestaltung
Armin Zebrowski, komwerb Agentur

URN-Nummer
urn:nbn:de:0011-n-6306675

Online verfügbar als Fraunhofer-ePrint
http://publica.fraunhofer.de/dokumente/N-630667.html

Gefördert durch das Ministerium für Wirtschaft, Arbeit und Wohnungsbau Baden-Württemberg

Gefördert durch

Baden-Württemberg
MINISTERIUM FÜR WIRTSCHAFT, ARBEIT UND WOHNUNGSBAU

CyberValley

S-TEC