Fraunhofer

Benjamin Fresz | Quynh Anh Tran

XAI für das Debugging von ML-Modellen

Hrsg.: Thomas Bauernhansl, Marco Huber, Werner Kraus

Im Rahmen des

KI Fortschrittszentrum
LERNENDE SYSTEME UND KOGNITIVE ROBOTIK

Vorwort

Künstliche Intelligenz (KI) ist eine der zentralen Technologien für die Zukunft. Ihre Einführung und der Einsatz fordern Unternehmen im besonderen Maß heraus.

Mit dem KI-Fortschrittszentrum des Fraunhofer IAO und Fraunhofer IPA unterstützen wir Unternehmen dabei, das Potenzial von KI zu erkennen und dieses wirtschaftlich nutzbar zu machen. 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.

Diese Studie diskutiert den Einsatz von Methoden der erklärbaren KI (XAI) für das Debugging von KI-Modellen. Während XAI-Methoden im Idealfall das Debugging von KI-Modellen vereinfachen, stellt im Moment die Auswahl der besten XAI-Methode für einen gegebenen Anwendungsfall noch ein Problem dar. Um hier zu unterstützen, gibt diese Arbeit eine Übersicht über Kategorisierungsansätze für XAI-Verfahren, geht auf die Evaluierungsprobleme jetziger Erklärungsmethoden ein und stellt anhand sieben aktueller Studien den Stand der Forschung zu XAI für das Debugging dar. Abschließend werden allgemeine Empfehlungen zum Einsatz von XAI für das Debugging formuliert.

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

Management Summary

Die rasante Entwicklung Künstlicher Intelligenz (KI) konfrontiert Unternehmen mit neuen Herausforderungen: Während KI-Systeme zunehmend leistungsfähiger werden, wird ihre Entscheidungsfindung auch undurchsichtiger. Die klassische empirische Bewertung von KI-Systemen anhand von Testdaten erweist sich für sicherheitskritische Anwendungen als unzureichend, da verschiedene Problemfelder die Zuverlässigkeit dieser Methodik untergraben. Leaking, bei dem KI-Systeme unbeabsichtigt Informationen über Testdaten erhalten, unausgewogene Datensätze mit unterrepräsentierten Teilgruppen und sogenannte »Scheinkorrelationen«, das Lernen von irrelevanten Mustern in den Daten, können zu Fehlentscheidungen führen.

Erklärbare KI (engl. Explainable AI, XAI) wird als mögliche Lösung für diese Herausforderungen beschrieben und soll beim Debugging von KI-Modellen unterstützen. Die wachsende Zahl von XAI-Publikationen und eine Vielzahl unterschiedlicher Ansätze stellen Entwicklungsverantwortliche jedoch vor eine zentrale Herausforderung: Welche XAI-Methode eignet sich für welchen Anwendungsfall?

Die Analyse der XAI-Landschaft offenbart eine beeindruckende, aber auch überwältigende Methodenvarianz. Viele XAI-Verfahren sind verfügbar, aber allgemein akzeptierte und eindeutige Kriterien, um diese zu kategorisieren, existieren nicht. Für die Einordnung von XAI-Verfahren bekannt sind konzeptionelle, ergebnis- und funktionsbasierte Ansätze. Diese bieten zwar Orientierungshilfen für die Vorauswahl, allerdings lassen sich die Verfahren teils nicht eindeutig einordnen und evaluieren.

Grundsätzlich sind XAI-Methoden mithilfe zweier Möglichkeiten zu bewerten: Mittels automatisierter Metriken und mittels Nutzerstudien – entweder auf einfacheren Aufgaben oder in realistischen Anwendungsszenarien. Allerdings sind automatische Bewertungsmetriken teils unzuverlässig und inkonsistent, mit starken Abhängigkeiten von Implementierungsdetails, die zu unterschiedlichen Rankings der Methoden führen können. Durch unklare Definitionen ist die Vergleichbarkeit von verschiedenen Metriken oft nicht gegeben. Menschbasierte Evaluierungen hingegen, obwohl unverzichtbar für die Praxistauglichkeit, leiden unter einem unvermeidbar hohen Aufwand und methodischen Schwächen in vielen Studien. Widersprüchliche Ergebnisse zwischen verschiedenen Untersuchungen sind gängig, und als besonders problematisch ist die Diskrepanz zwischen subjektivem und objektivem Verständnis zu sehen: Nutzer überschätzen ihr Verständnis der KI-Systeme durch XAI-Erklärungen, ohne dass dies mit einer tatsächlichen Verbesserung ihrer Leistung zum Beispiel im Debugging einhergeht.

Die Analyse sieben relevanter Studien zum XAI-Einsatz beim Debugging zeigt, wie stark die Methoden vom Kontext abhängen: Verschiedene Fehlertypen und Datensätze erfordern verschiedene XAI-Ansätze. Zusätzlich beeinflusst die Expertise der Nutzer die Effektivität erheblich. Während erfahrene Entwickler und Entwicklerinnen von komplexeren Erklärungen profitieren können, können dieselben Methoden bei weniger erfahrenen kontraproduktiv wirken. Insgesamt ist die Formulierung allgemeiner Hinweise zur erfolgreichen Nutzung von XAI für das Debugging schwierig, aber die Interaktivität von XAI-Verfahren, die parallele Nutzung mehrerer Verfahren und ausführliche Interpretationshilfen (wie Schulungen) scheinen Vorteile zu bringen.

Insgesamt ergibt sich, dass Unternehmen XAI-Methoden für das Debugging verwenden können. Für einen möglichst erfolgreichen Einsatz, zum Beispiel in sicherheitskritischen Anwendungen, ist aber noch ein hoher manueller Aufwand – insbesondere für das Testen in der realen Anwendung – notwendig, da nur eine Nutzerstudie validieren kann, wie effektiv die gewählte XAI-Methode in der realen Anwendung ist. Zudem sollte XAI nicht als alleiniges Validierungstool, sondern als Teil des Qualitätssicherungsprozesses für KI-Systeme verwendet werden.

1. Einleitung und Motivation

In den letzten Jahren können zunehmend mehr Anwendungsfälle mit Künstlicher Intelligenz (KI) gelöst werden. Dabei basieren die heutigen KI-Verfahren auf der Fähigkeit, Muster aus sogenannten Trainingsdaten zu extrahieren und diese für die Bearbeitung von neuen Datenpunkten zu verwenden. Mittels dieser Verfahren ist es inzwischen möglich, KI in Bereichen anzuwenden, die traditionell menschlicher Intelligenz vorbehalten waren.

Dabei unterscheiden sich die verwendeten Verfahren in ihrem Aufbau und ihrer Funktionsweise, aber das grundsätzliche Konzept der Mustererkennung bleibt bestehen. Eine Ausnahme bilden hier sogenannten Expertensysteme, bei denen, wie schon im Namen impliziert, Experten oder Expertinnen explizite Regeln einprogrammieren. Heutzutage wird unter dem Begriff KI allerdings meist von Verfahren des Maschinellen Lernens (ML) ausgegangen, bei denen das Verhalten durch aus den Trainingsdaten extrahierte Muster bestimmt wird. Daraus resultieren – je nach verwendetem Verfahren – meist »Black Box-Modelle«, deren Entscheidungsfindung nicht einsehbar oder zu komplex für menschliches Verständnis ist.

Deshalb wird die Fähigkeit eines ML-Modells, ein bestimmtes Problem zu lösen, üblicherweise empirisch definiert. Für ein ML-Modell, das eingesetzt werden soll, um im Automotive-Bereich Kabelbäume in die dafür vorgesehenen Buchsen einzustecken, könnte so eine Bewertung lauten: »Für 98,7 Prozent der gegebenen Datenpunkte ist der Einsteckvorgang bei der Kabelbaummontage erfolgreich.«

Allerdings reicht in bestimmten Fällen die empirische Auswertung von ML-Modellen für einen bestimmten Anwendungsfall nicht aus. Hier könnte die Nutzung von Methoden der sogenannten erklärbaren KI (engl. Explainable AI bzw. XAI) unterstützen. Im Folgenden wird kurz beschrieben, welche Probleme bei der empirischen Bewertung von ML-Modellen auftreten können und wie sich die ideale Rolle von XAI hier gestaltet, bevor in den folgenden Kapiteln darauf eingegangen wird, was es aktuell beim Einsatz von XAI zu beachten gilt und wie die Umsetzung für den Debugging-Fall in der Praxis aussehen kann.

1.1. Schwierigkeiten bei der Bewertung von KI-Verfahren

Für die Bewertung von KI-Systemen bzw. ML-Modellen wird meist – im Gegensatz zu den für das Training verwendeten Trainingsdaten – ein spezifisch hierfür vorgehaltener Datensatz, die sogenannten Testdaten, verwendet. Von diesem Datensatz wird angenommen, dass er den in der realen Anwendung entstehenden Daten ähnlich ist bzw. der gleichen statistischen Verteilung entstammt. Auf diesem Datensatz wird dann der Restfehler des KI-Systems nach dem Training erhoben.

Dabei kann »Leaking« auftreten, zu Deutsch etwa »auslaufen« oder auch »durchstechen (von Informationen)«. Leaking bedeutet in diesem Kontext, dass das ML-Modell dank der Trainingsdaten Informationen über die Testdaten hat, die in der realen Anwendung nicht zur Verfügung stehen. Ein Beispiel für ein ML-Modell mit Leaking findet sich in der Vorhersage von Lungenentzündungen basierend auf Röntgenscans [20]. Wenn sich hier Bilder derselben Erkrankten sowohl im Trainings- als auch im Testdatensatz befinden, kann das ML-Modell bereits über die Assoziation einer Person mit der zugehörigen Diagnose das zugehörige Ergebnis prädizieren, anstatt die relevanten Informationen aus dem Scan zu verwenden. Dies führt somit dazu, dass das ML-Modell auf den Trainingsdaten eine gute Performance erzielt. In der tatsächlichen Anwendung für neue Erkrankte würde ein solches System eine deutlich schlechtere Performance bieten, da für diese Personen das Diagnoseergebnis nicht in den Trainingsdaten vorgegeben ist.

Ebenfalls kann es sein, dass die Performance der Modelle je nach betrachteter Teilmenge unterschiedlich ist. Wenn in den Trainingsdaten bestimmte Eigenschaften unterrepräsentiert oder meist mit bestimmten Ergebnissen verknüpft sind, kann die Performance eines ML-Modells auf den meisten Teilmengen gut sein, aber für andere Teilmengen (fast) nie erfolgreich. Bei der oben genannten Kabelbaummontage könnte das KI-System zum Beispiel für einen bestimmten Steckertyp nicht ausreichend trainiert sein. Insbesondere, wenn die Performance von KI-Systemen von sensitiven Merkmalen wie Alter, Geschlecht oder Hautfarbe abhängen könnte, ist Vorsicht geboten.

Zudem sind seltene Fälle in Datensätzen meist unterrepräsentiert. Für selten auftretende Szenarien, zum Beispiel Beinahe-Unfälle im Straßenverkehr, stehen üblicherweise sowohl in den Trainings- als auch in den Testdaten nur wenige Datenpunkte zur Verfügung. Zusätzlich ist es schwierig, weitere Datenpunkte aufzunehmen. Dennoch sollten – insbesondere, wenn es sich um sicherheitskritische Szenarien handelt – für solche Situation verlässliche Schlüsse über das Verhalten des ML-Modells vorhanden sein.

Wie oben beschrieben, wenden ML-Modelle üblicherweise bekannte Muster auf neue Daten an. Allerdings kann es auch passieren, dass die erkannten Muster nicht denen entsprechen, die ein Mensch verwenden würde, und somit in der Anwendung unerwartete Probleme verursachen oder nicht über den gegebenen Datensatz hinaus generalisieren. Ein bekanntes, wenn auch akademisches, Beispiel bezieht sich auf ein ML-Modell, das für ein Eingabebild ausgeben sollte, ob jeweils ein Wolf oder ein Husky abgebildet war. Hierbei zeigte sich allerdings, dass das ML-Modell wesentlich einen sogenannten Shortcut (dt. Abkürzung) für die Entscheidungsfindung gelernt hatte: Sobald sich im Hintergrund des gezeigten Tiers Schnee befand, wurde das Tier als Wolf klassifiziert, während andere Hintergründe zur Klassifikationsentscheidung Husky führten [21] (siehe Abbildung 1). ML-Modelle, die auf Basis eines solchen Shortcuts entscheiden, werden auch »Clever-Hans-Prädiktoren« genannt [13].

Diese verschiedenen Fehlerfälle in der Bewertung von ML-Modellen zeigen, dass die rein empirische Prüfung der Performance nicht ausreicht, um ML-Modelle so zu prüfen, dass ihr Einsatz in sicherheitsrelevanten Anwendungen angemessen verargumentiert werden kann.

Abbildung 1: Links: Bild, das als »Wolf« klassifiziert wurde. Rechts: LIME-Erklärung für die Klassifikation. Basierend auf der Erklärung zeigt sich, dass das Modell einen Shortcut für die Unterscheidung von Wölfen und Huskys – Schnee im Hintergrund des Bilds – gelernt hat. Grafik aus [21].

1.2. XAI zur Evaluation von ML-Modellen

Eine weitere Möglichkeit, um die Funktionsweise von ML-Modellen zu prüfen, sind Methoden aus dem Bereich der erklärbaren KI. Diese versuchen, grundsätzlich die Entscheidungsfindung von ML-Modellen zu »erklären«, d.h. zusätzliche Informationen zur Verfügung zu stellen, die es Menschen ermöglichen, die Entscheidungsfindung eines ML-Modells nachvollziehen zu können. Dies könnte grundsätzlich ermöglichen, Fehlerfälle wie die oben genannten Shortcuts zu entdecken und somit im Debugging von ML-Modellen zu unterstützen oder bei seltenen Datenpunkten ein tiefgreifenderes Verständnis der Funktionsweise des Modells zu erhalten. Dabei beschreibt der Begriff Debugging grundsätzlich alle Tätigkeiten, die darauf ausgelegt sind, Fehlerfälle in KI-Systemen zu erkennen und zu beheben.

Das Feld der erklärbaren KI ist jedoch selbst zu einer Herausforderung geworden. Mit der exponentiell wachsenden Zahl von XAI-Publikationen – mehrere hundert pro Jahr – und einer Vielzahl unterschiedlicher Ansätze stehen Entwicklerinnen und Entwickler vor der Frage: Welche XAI-Methode eignet sich für welche Anwendung?

Das folgende Kapitel gibt einen Überblick über die bestehenden Kategorisierungsansätze für XAI, bevor in Kapitel 3 die Schwierigkeiten bei der Evaluation von XAI beschrieben werden. Kapitel 4 stellt mehrere aktuelle Studien bezüglich XAI für das Debugging dar und führt diese in Empfehlungen für die praktische Nutzung von XAI zusammen.

2. Kategorisierungen von XAI-Methoden

Es gibt einige Publikationen, die Übersichten über die bestehenden XAI-Methoden bieten – so viele, dass im Jahr 2022 sogar versucht wurde, eine Kategorisierung der verschiedenen XAI-Kategorisierungen zu erstellen [25]. Diese Veröffentlichung ist die wesentliche Grundlage für das folgende Kapitel. Dabei lassen sich die bestehenden Kategorisierungen grob in drei Ansätze einteilen:

- Konzeptioneller Ansatz: »In welcher Phase der Entwicklung kann die Methode verwendet werden?« und »Was wird erklärt: Modell oder Einzelentscheidung?«
- Ergebnisbasierter Ansatz: »Wie sieht das Ergebnis dieser XAI-Methode aus?«
- Funktionsbasierter Ansatz: »Was macht diese XAI-Methode?«

Diese unterschiedlichen Ansätze sollen helfen, XAI-Methoden für spezifische Anwendungsfälle auszuwählen, bedürfen selbst allerdings wiederum einer Einführung. Deshalb werden im Folgenden alle Ansätze kurz vorgestellt. Zusätzlich wird noch auf einen gemischten Ansatz sowie die möglichen Stakeholder und Zielsetzungen beim Einsatz von XAI eingegangen.

2.1. Konzeptionelle Einordnung

Um einen generellen Überblick über die Landschaft von XAI-Methoden zu bekommen, bietet sich der konzeptionelle Ansatz an. In diesem werden – bestenfalls voneinander unabhängige – Konzepte verwendet, um XAI-Methoden einordnen zu können. Diese Dimensionen sind in der Regel hierarchisch organisiert und über verschiedene Publikationen hinweg unterschiedlich, wobei mehrere wesentliche Begrifflichkeiten gängig sind. Das sind insbesondere:

- **Phase der XAI-Anwendung**, aufgeteilt in ante-hoc (»vorher«) und post-hoc (»nachher«).
  Bei Ante-hoc-Ansätzen werden Modelle von Grund auf als interpretierbar konzipiert (z.B. Entscheidungsbäume oder lineare Regression). Diese Modelle sind intrinsisch verständlich, weisen jedoch oft eine geringere Leistungsfähigkeit auf als komplexere Modelle. Diese kann jedoch je nach Anwendungsfall ausreichend sein.
  Post-hoc-Erklärungen werden nach dem Training eines bereits existierenden Modells generiert. Diese Methoden sind besonders relevant für komplexe Modelle wie tiefe neuronale Netze oder Random Forests.
- **Anwendbarkeit der Methode**, insbesondere modell-agnostisch und modell-spezifisch.
  Modell-agnostische Methoden funktionieren unabhängig vom zugrundeliegenden Modelltyp und können auf beliebige ML-Modelle angewendet werden (z.B. Local Interpretable Model-agnostic Explanations (LIME) [21], SHapley Additive ExPlanations (SHAP) [15]).

Modell-spezifische Methoden sind auf bestimmte Modelltypen zugeschnitten und nutzen deren interne Struktur zur Erklärung (z.B. gradientenbasierte Methoden für neuronale Netze).

- **Umfang der Erklärung**, insbesondere lokale und globale Erklärungen.
  Lokale Erklärungen fokussieren sich auf einzelne Vorhersagen oder Datenpunkte und erklären, warum ein bestimmtes Ergebnis generiert wurde. Globale Erklärungen versuchen, das gesamte Verhalten eines Modells zu erfassen und zu erklären.

Die beschriebenen Konzepte sind ebenfalls in Abbildung 2 dargestellt.

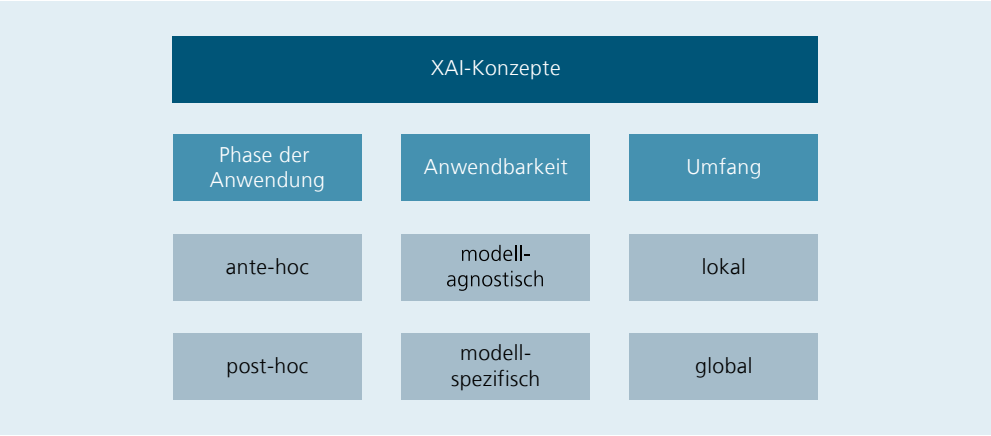


Abbildung 2: Wesentliche Konzepte für XAI-Systeme.

Die zuvor vorgestellten Konzepte eignen sich, um einen generellen Überblick über die Landschaft an XAI-Methoden zu erhalten. Allerdings gibt es auch Ansätze, deutlich spezifischere Konzepte zu definieren, die für die Kategorisierung von XAI-Methoden verwendet werden können (siehe Sektion 2.4).

2.2. Ergebnisbasierte Einordnung

Eine praxisnahe Perspektive auf die Einordnung von XAI-Methoden bietet der ergebnisbasierte Ansatz. Im Gegensatz zu funktionsbasierten oder konzeptionellen Taxonomien ist dieser Ansatz darauf fokussiert, welche Art von Ergebnis eine Erklärbarkeitsmethode produziert. Dies ist insbesondere dann hilfreich, wenn Anwenderinnen und Anwender eine klare Präferenz für einen bestimmten Erklärungstypus haben. Dabei gibt es drei zentrale Ergebniskategorien:

- **Feature Importance (Merkmalswichtigkeit)**
  Diese Methoden identifizieren und quantifizieren, welche Eingabemerkmale am stärksten zur Entscheidung des ML-Modells beigetragen haben. Beispiele hierfür sind LIME [21], SHAP [15] und DeepLIFT (Deep Learning Important FeaTures) [23]. Diese Techniken liefern oft numerische Gewichtungen oder visuelle Heatmaps bzw. Saliency Maps, die insbesondere für Fachleute verständlich sind, da sie direkt auf die ursprünglichen Eingabedaten Bezug nehmen. Diese Art von Erklärung wird auch gerne für Bilder verwendet, wie in Abbildung 1 und Abbildung 3 dargestellt.
  Eine Erweiterung der Merkmalswichtigkeit stellen Konzept-Erklärungen dar. Hier wird versucht, eine Erklärung nicht mehr die Wichtigkeit einzelner Dimensionen der Eingabedaten (wie Pixel) anzeigen zu lassen, sondern die Eingabedaten in Konzepte zu zerlegen (wie Hundegesichter oder Fellmuster für den Fall in Abbildung 3) und deren Relevanz für die Ausgabe des ML-Modells darzustellen.

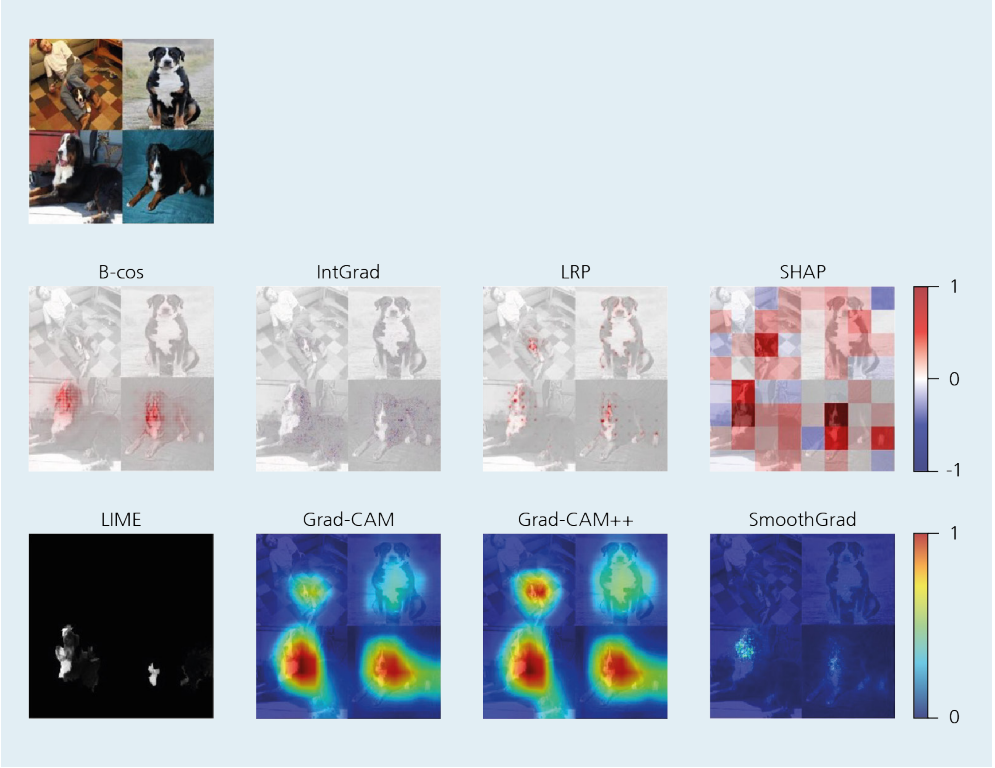


Abbildung 3: Saliency Maps (Merkmalswichtigkeit auf Bilddaten) auf dem ImageNet-Datensatz. In der oberen Reihe ist das Originalbild dargestellt, nachfolgend die Erklärungen der XAI-Methoden. Alle Methoden zeigen eine Erklärung im Sinne der wichtigsten Bildelemente für die Klassifikation »Berner-Sennenhund«. Bild aus [12].

- **Surrogate Models (Ersatzmodelle)**
  Diese Methoden approximieren das zu erklärende ML-Modell durch ein einfacheres, intrinsisch erklärbares Modell (z.B. Entscheidungsbäume oder lineare Regression). LIME verwendet beispielsweise lokale Surrogatmodelle, um einzelne Vorhersagen zu erklären, verdeutlicht damit aber auch, dass die Kategorien der XAI-Methoden teils nicht klar trennbar sind. Abgesehen von LIME werden Surrogatmodelle vor allem für globale Erklärungen eingesetzt, wodurch allerdings meist auch Informationen über das Verhalten des komplexen Modells verloren gehen.
- **Examples (Beispiele)**
  Methoden in dieser Kategorie erklären durch das Präsentieren bestimmter Beispiele aus den Trainingsdaten. Dazu zählen Prototypen-Erklärungen, die repräsentative Datenbeispiele für bestimmte Klassen zeigen, oder kontrafaktische Erklärungen, die darstellen, wie möglichst geringfügige Änderungen an den Eingabedaten zu anderen Vorhersagen führen würden. Ein Beispiel für eine kontrafaktische Erklärung könnte zum Beispiel im Falle einer automatisierten Kreditentscheidung die Aussage »Wenn Ihr Gehalt um 10.000 € höher wäre, hätten Sie einen Kredit erhalten« sein.

Mittels der vorgestellten Kategorien lassen sich XAI-Methoden nach ihrem Ziel auswählen, wenn ein bestimmter Erklärungstyp bevorzugt wird. Allerdings zeigt sich, wie auch in Abbildung 3 zu sehen, dass verschiedene XAI-Methoden zwar Erklärungen des gleichen Typs liefern, sich aber in den Details der erstellten Erklärung unterscheiden. Um diese Unterschiede zu beachten, bietet es sich an, einen anderen Kategorisierungsansatz ergänzend hinzuzunehmen.

2.3. Funktionsbasierte Einordnung

Abbildung 4: Kategorisierung der Funktionsweise von XAI-Methoden basierend auf [14, 25].

Der funktionsbasierte Ansatz zur Kategorisierung von XAI-Methoden klassifiziert Erklärbarkeitsverfahren nach ihrem zugrundeliegenden Funktionsprinzip – also danach, wie eine Methode Informationen aus einem ML-Modell extrahiert, um Erklärungen zu generieren. Hierbei sind fünf wesentliche Funktionskategorien zu unterscheiden:

- **Lokale Perturbationen** verändern die Eingaben eines Modells leicht, um die Wichtigkeit bestimmter Merkmale für die Vorhersage zu ermitteln. Durch systematische Variation der Eingabewerte kann analysiert werden, wie sich diese auf die Ausgabe auswirken.
- **Strukturelle Ansätze** nutzen spezifische Eigenschaften der ML-Modelle, z. B. deren Gradienten, aus, um Erklärungen zu erzeugen. Sie verwenden die interne Struktur des Modells, um Erklärungen abzuleiten.
- **Meta-Erklärungen** arbeiten nicht direkt auf einem ML-Modell, sondern auf Erklärungen, die von anderen Methoden generiert wurden. Sie aggregieren und vergleichen mehrere Erklärungen, um eine robustere und konsistentere Erklärung zu liefern.
- Bei **Anpassungen der Modellstruktur** wird versucht, komplexe Modelle durch Veränderungen von deren Architektur zu vereinfachen. Mittels Ersetzen komplexer Komponenten durch einfachere wird das Modellverhalten transparenter gemacht.
- Ebenfalls enthalten ist die **Extraktion von Beispielen**, die bereits in der ergebnisbasierten Einordnung in Sektion 2.2 beschrieben wurde.

Die funktionsbasierte Einordnung ist insbesondere für die tiefere inhaltliche Auseinandersetzung mit XAI-Methoden nützlich, da sich aus der Funktionsweise bestimmte weitere Eigenschaften (siehe Kapitel 3) ergeben können.

2.4. Mischungen der vorherigen Ansätze

Die zuvor vorgestellten Ansätze werden in der Praxis oft in Kombination verwendet, um eine umfängliche Beschreibung von XAI-Methoden zu erhalten. Die bisher wohl umfänglichste Kategorisierung bieten die sogenannten Co-12 Eigenschaften von Nauta et al. [19] (siehe Tabelle 1). Zwei der dort beschriebenen Eigenschaften sind Compactness, die Größe der Erklärung, und Compositionality, deren Format und Organisation. Diese beschreiben wesentlich die ergebnisbasierte Einordnung aus Sektion 2.2, während die Funktionsweise der XAI-Methode in diesem Ansatz nur indirekt über die Auswirkungen auf andere Eigenschaften Beachtung findet. Weitere Eigenschaften wie Correctness – die Übereinstimmung der Erklärung mit der inneren Logik des Modells – wurden in den vorherigen Kategorisierungsansätzen höchstens indirekt, z. B. mittels der funktionsbasierten Einordnung, berücksichtigt.

Tabelle 1: Co-12 Eigenschaften von Nauta et al. [19] zur Kategorisierung von XAI-Methoden.

Bezug	Kurzform	Beschreibung
Inhalt	Correctness	Übereinstimmung der Erklärung mit dem (Black-Box-)Modell
	Completeness	Grad der Abdeckung des Black-Box-Modells durch Erklärungen
	Consistency	Determinismus und Implementierungsunabhängigkeit (Gedanke: identische Eingaben sollten die gleiche Erklärung haben)
	Continuity	Kontinuität und Generalisierbarkeit der Erklärung (ähnliche Eingaben sollten ähnliche Erklärungen haben)
	Contrastivity	Diskriminativität von Erklärungen in Bezug auf andere Ereignisse/Ziele (Antworten auf »Warum nicht B stattdessen?«)
	Covariate Complexity	Komplexität der Merkmalsinteraktionen zwischen verschiedenen Eingabedimensionen
Darstellung	Compactness	Größe der Erklärung
	Compositionality	Format und Organisation der Erklärung
	Confidence	Vorhandensein und Genauigkeit probabilistischer Informationen, z. B. Modell- und Erklärungssicherheit
Nutzende	Context	Relevanz der Erklärung für Nutzende und deren Bedürfnisse
	Coherence	Übereinstimmung mit Vorwissen und Überzeugungen
	Controllability	Interaktion und Kontrollierbarkeit der Erklärung

In der Praxis stellt es sich allerdings als schwierig heraus, spezifische Bewertungen der einzelnen Eigenschaften vorzunehmen, mittels derer die XAI-Methoden dann verglichen werden können. Dieses Problem wird in Kapitel 3 ausführlicher beschrieben.

2.5. Anwendungsfälle und Stakeholder bei XAI-Verfahren

Während sich diese Studie nur der Auswahl einer XAI-Methode für das Debugging von ML-Modellen widmet, ist generell darauf hinzuweisen, dass der Auswahl eines geeigneten XAI-Verfahrens stets zugrunde liegen sollte, wem, wann, was und wie erklärt wird [30]. Dies spiegelt sich auch in den zuvor vorgestellten Kategorisierungen wider – je nach gewählter Kategorisierung in unterschiedlicher Form. Dabei ist das Feld der möglichen Stakeholder für XAI und deren Zielsetzungen recht umfangreich. Je nach Publikation wird eine unterschiedliche Anzahl an möglichen Stakeholdern für XAI angegeben. Wang et al. [30] benennen z. B. die in Tabelle 2 aufgeführten neun unterschiedlichen mögliche Stakeholder mit verschiedenen Zielen.

Tabelle 2: Zusammenfassung der Stakeholder und zugehöriger Ziele für Erklärbarkeit von Wang et al. [30].

Kurzform	Zielsetzung
KI-Fachleute	Einblick in interne Logik komplexer ML-Modelle
KI-Entwickler	Unterstützung bei Feature Auswahl und Modelloptimierung
Validierungsteam	Überprüfung der Entscheidungslogik des Modells
Modellbetreiber	Sicherstellung korrekter und effizienter Interaktion mit dem Modell
Entscheider	Verständnis spezifischer Modellentscheidungen
Regulierungsbehörden	Unterstützung bei der Zuordnung von Verantwortung
Datensubjekte	Schutz persönlicher Dateninformationen
Von Entscheidung betroffene Personen	Verständnis von Entscheidungen, die sie betreffen
Allgemeine Endnutzer	Aufbau angemessenen Vertrauens in das System

Der gewählte Use Case des Debuggings mittels XAI bestimmt die Auswahl der relevanten Stakeholder-Gruppe, und zwar des Validierungsteams. Während auch das Entwicklungsteam die gleiche Zielsetzung verfolgen könnte, stellt sich hierdurch keine wesentliche Änderung ein: Sowohl eine hohe Expertise bezüglich KI als auch die Zielsetzung der Fehlersuche mittels XAI sind bei beiden Stakeholdern gegeben. Hier zeigt sich auch, wieso der vorliegende Anwendungsfall für diese Studie gewählt wurde: Es existiert eine klare und unmissverständliche Zielsetzung und für die Zielgruppe der KI-Fachleute, -Entwickler und Validierungsteams liegen bisher die meisten Studien dazu vor, welche Bedingungen Erklärungen erfüllen müssen. Für andere Anwendungsfälle, wie zum Beispiel der Aufbau von »angemessenem« Vertrauen sind sowohl die Zielgruppe als auch ein messbares und eindeutiges Ziel deutlich schwieriger zu definieren.

Für das Debugging von ML-Modellen mittels XAI ergeben sich auch noch weitere Einordnungen: Es sollte eine Ergebnisdarstellung gewählt werden, die für KI-Entwicklerinnen und -Entwickler eingängig und leicht anzuwenden ist. Eingängig für diese Zielgruppe sind z.B. Feature-Importance-Methoden (siehe Sektion 2.2), da diese direkt die Verbindung zwischen Modell-Eingabe und -Ausgabe darstellen. Leicht anzuwenden sind generell Post-hoc- und Modell-agnostische Methoden (siehe Sektion 2.1). Während für das Debugging globale Erklärungen zum Finden systematischer Fehler in ML-Modellen interessant wären, stehen diese noch nicht ausreichend und gut nutzbar zur Verfügung. Deshalb kommen eher lokale Methoden zum Einsatz.
Für neuronale Netzwerke wird Merkmalswichtigkeit üblicherweise mittels der Gradienten oder lokaler Perturbationen generiert. Damit ist – nach den oben beschriebenen Kategorien – eine Klasse an Verfahren für das Debugging ausgewählt. Wieso die spezifische Methodenauswahl im konkreten Fall dennoch ein Problem darstellen kann, wird im folgenden Kapitel beschrieben.

3. Evaluation von XAI-Methoden

Neben der generellen Unübersichtlichkeit des Forschungsfelds XAI bedingen auch weitere Probleme die Schwierigkeiten bei der Auswahl geeigneter Methoden für einen Anwendungsfall. Im vorhergehenden Kapitel wurden verschiedene Ansätze zur Kategorisierung von XAI-Methoden vorgestellt. Um eine Kategorisierung für die Auswahl einer konkreten Methode zu nutzen, sollte die Kategorisierung möglichst genau alle Eigenschaften einer XAI-Methode erfassen (z.B. nach dem Ansatz aus Sektion 2.4). Dafür ist es nötig, die jeweiligen spezifischen Eigenschaften zu bewerten. Die Schwierigkeiten bei der Bewertung von XAI-Methoden werden in diesem Kapitel – nach einer kurzen Einführung der generellen Kritikpunkte am Forschungsfeld XAI – näher beschrieben.

3.1. Kritikpunkte am Forschungsfeld XAI

Um die Herausforderungen im Feld XAI systematisch anzugehen, benennen Longo et al. [14] 28 offene Probleme, während Weber et al. [31] mit dem provokativen Titel „XAI is in Trouble" auf die offenen Fragen um XAI hinweisen. Dabei lassen sich die Kritikpunkte hauptsächlich in drei zentrale Problemfelder einteilen:

3.1.1. Konzeptionelle Unschärfe und mangelnde terminologische Klarheit

Wie Longo et al. betonen, herrscht im XAI-Bereich erhebliche begriffliche Unschärfe, insbesondere bezüglich der Unterscheidung zwischen »Explainability«, »Interpretability« und »Transparency« (dt. Erklärbarkeit, Interpretierbarkeit und Transparenz). Insbesondere der Begriff Transparenz umfasst oft mehr als nur Erklärbarkeit und Interpretierbarkeit. So wird in Transparenz auch oft der Zugang zu Modellen oder Daten oder auch Informationen über die grundsätzliche Funktionsweise eines ML-Modells gesehen. Erklärbarkeit und Interpretierbarkeit sind hingegen deutlich spezifischer, teils bezieht sich Erklärbarkeit wesentlich auf Post-hoc-Modelle und Interpretierbarkeit auf Ante-hoc-Verständlichkeit von ML-Modellen.

Diese begriffliche Unschärfe führt dazu, dass Forschungsergebnisse schwer vergleichbar sind und praktische Anwendungen oft auf unsicheren theoretischen Fundamenten aufbauen. Wie Weber et al. [31] kritisch anmerken, wird XAI oft als universelles Heilmittel für alle Probleme von Black-Box-Modellen präsentiert, ohne dass der konkrete Nutzen für spezifische Szenarien systematisch untersucht wurde.

3.1.2. Praktische Einschränkungen

Eine zweite Kritikrichtung betrifft die praktische Anwendbarkeit von XAI-Methoden. Viele XAI-Methoden weisen folgende Einschränkungen auf [31]:

- **Robustheitsprobleme**: Viele XAI-Methoden sind empfindlich gegenüber kleinen Störungen in den Eingabedaten oder Modellparametern, was zu inkonsistenten Erklärungen führt.
- **Kontextunabhängigkeit**: Die meisten XAI-Methoden liefern Erklärungen, die nicht auf den spezifischen Kontext abgestimmt sind. Beispielsweise sind Erklärungen, die für das Debugging von Shortcuts entwickelt wurden, teils ungeeignet, um Nicht-Fachleuten eine Entscheidung näher zu erklären.
- **Fehlende Kausalität**: Die meisten XAI-Methoden identifizieren Korrelationen statt kausaler Zusammenhänge. Dennoch werden die generierten Erklärungen oft kausal interpretiert.

3.1.3. Evaluierung von XAI

Der wohl gravierendste Kritikpunkt betrifft die mangelnde methodische Fundierung der Evaluierung von XAI-Methoden. Dabei gibt es drei zentrale Probleme bei der Evaluierung [14]:

- **Fehlende Evaluierung mit Menschen**: Nur ein geringer Anteil von XAI-Veröffentlichungen enthält Nutzerstudien, um die tatsächliche Verständlichkeit der Erklärungen für die Zielgruppe zu prüfen. Dies ist besonders problematisch für das Debugging, da Entwicklerinnen und Entwickler oft »inmates running the asylum« [16] sind – sie entwickeln Erklärungen, die für sie selbst verständlich sind, aber nicht notwendigerweise für andere Stakeholder.
- **Mangelnder Evaluierungsrahmen**: Es existiert kein standardisierter Rahmen zur Evaluierung von XAI-Methoden. Die wenigen existierenden Evaluierungsmetriken sind oft auf spezifische Anwendungsfälle oder Datentypen beschränkt und lassen sich nicht auf andere Szenarien übertragen.
- **Begrenzte Reproduzierbarkeit**: Viele Nutzerstudien mit Menschen weisen methodische Schwächen auf, was zu einer schlechten Reproduzierbarkeit und nicht generalisierbaren Erkenntnissen führt.

Diese Kritikpunkte sowie die generelle Unübersichtlichkeit des Forschungsfelds verdeutlichen die Schwierigkeiten bei der Auswahl geeigneter Methoden für einen Anwendungsfall. Insbesondere die Evaluierung von XAI ist eine erhebliche Hürde, XAI im Debugging-Prozess effektiv zu nutzen. Um hier ein tieferes Verständnis zu bieten, werden die Probleme bei der Evaluierung von XAI-Methoden – sowohl automatisiert als auch basierend auf Nutzerstudien – im Folgenden detaillierter beschrieben.

3.2. Grundkonzepte der Bewertung von XAI

Das zentrale Problem bei der Auswahl von XAI-Methoden wurde bereits 2017 von Doshi-Velez und Kim [11] beschrieben: Durch unklare Definitionen, was Erklärbarkeit und Interpretierbarkeit wirklich sind und wie diese messbar gemacht werden können, kommen Ergebnisse zustande, die sich untereinander nicht vergleichen lassen. Insgesamt zeigt sich, dass Interpretierbarkeit nicht als universelles Merkmal, sondern als kontextabhängiges Konzept verstanden werden muss, was sich in den neueren Kategorisierungen in Kapitel 2 ebenfalls zeigt.

Weiterhin haben Doshi-Velez und Kim eine dreigliedrige Taxonomie zur Bewertung von XAI-Methoden entwickelt. Diese ist aufgeteilt in technische bzw. automatisierte Bewertungen, Studien mit Menschen mit vereinfachten Aufgaben und Studien mit realen Anwendern in realistischen Anwendungsszenarien. Diese Kategorien werden bezeichnet als funktionsbasierte, menschbasierte und anwendungsbasierte Evaluierung. Dabei steigen Anwendungsbezug, aber auch Aufwand und die zugehörigen Kosten mit jedem der Fälle an (Abbildung 5). Es zeigt sich, dass durch die Kontextabhängigkeit von XAI rein automatische Metriken nicht ausreichend sind, um den praktischen Nutzen einer Methode in der realen Anwendung vorherzusagen. Dennoch können automatische Bewertungsmetriken nützlich sein, um so z.B. XAI-Methoden, die automatisch prüfbare Eigenschaften nicht erfüllen, bereits ohne Nutzerstudie auszuschließen. Eine solche Eigenschaft könnte Correctness sein, also dass Erklärung und Modellverhalten übereinstimmen. Wenn diese nicht gegeben ist, ist auch in einer späteren Studie kein großer Nutzen einer XAI-Methode zu erwarten, womit der Aufwand für die Studie vergebens wäre.

Abbildung 5: Evaluierungsmöglichkeiten für XAI basierend auf [11].

3.3. Automatische Bewertung von XAI

Wie zuvor beschrieben, kann die automatische Bewertung gewisser Eigenschaften von XAI-Methoden den Aufwand verringern, der bis zu einem tatsächlichen Einsatz dieser führt. Aber auch bei der automatischen Bewertung von XAI sind gewisse Herausforderungen festzustellen, die im Folgenden näher erklärt werden.

3.3.1. Mangelnde Zuverlässigkeit von Bewertungsmetriken

Eine der zentralen Kritiken an automatisierten XAI-Bewertungen betrifft deren statistische Zuverlässigkeit [28]. Dabei weisen gängige Metriken wie AOPC (Area Over the Perturbation Curve) erhebliche Unzuverlässigkeiten auf.

Besonders problematisch ist die geringe »inter-rater reliability«, gemessen durch Krippendorff's α-Statistik. Bei der Anwendung auf verschiedene Bilder produzieren die meisten Metriken inkonsistente Rankings der XAI-Methoden. So erreicht selbst die beste Konfiguration in den Experimenten von Tomsett et al. [28] einen α-Wert von lediglich 0,48 – deutlich unter dem

Schwellenwert von 0,65, der für zuverlässige Metriken typischerweise erwartet wird. Dies bedeutet, dass Rankings von XAI-Methoden, die mit diesen Metriken für einen Datensatz erstellt werden, kaum verlässlich auf neue Bilder übertragbar sind.

Darüber hinaus zeigen dieselben Studien, dass verschiedene Metriken oftmals nicht das gleiche zugrundeliegende Konzept messen. So erreichen bestimmte Metrik-Paare nur schwache Korrelationen, während für andere die Korrelationen je nach verwendeter XAI-Methode stark schwanken. Dies deutet auf eine mangelnde interne Konsistenz der Bewertungsmetriken hin. Um diesen Problemen zu begegnen, wurden in den letzten Jahren weitere Bewertungsmetriken entwickelt [12], deren Zusammenhang zu bestehenden allerdings noch klargestellt werden muss.

3.3.2. Abhängigkeit von Implementierungsdetails

Ein weiterer Kritikpunkt betrifft die hohe Empfindlichkeit der Bewertungsmetriken gegenüber kleinen Änderungen in der Implementierung [28, 29]. So kann bei perturbationsbasierten Methoden die Wahl der spezifischen Perturbation (z. B. Ersetzen durch Mittelwert vs. zufällige Werte) zu völlig unterschiedlichen Rankings führen.

Interessanterweise wirken sich diese Implementierungsdetails unterschiedlich auf verschiedene XAI-Methoden aus. So sind gradientenbasierte Methoden wie Integrated Gradients [27] oder SmoothGrad [24] stärker von der Wahl der Perturbationsmethode betroffen als Methoden basierend auf Class Activation Mapping (CAM) wie Grad-CAM. Dies erschwert einen fairen Vergleich verschiedener XAI-Ansätze, da die Wahl der Bewertungskonfiguration das Ergebnis maßgeblich beeinflussen kann.

3.3.3. Einfluss der Dichte einer visuellen Erklärung

Die genauen visuellen Details einer Erklärung, wie in Abbildung 3 dargestellt, spielen für deren Bewertungsergebnis ebenfalls eine wichtige Rolle [29]. Gradientenbasierte Methoden wie Integrated Gradients (IntGrad) erzeugen typischerweise sehr spärliche Erklärungen mit wenigen stark hervorgehobenen Pixeln, während CAM-basierte Methoden wie Grad-CAM deutlich dichtere Erklärungen liefern.

Diese Unterschiede haben erhebliche Auswirkungen auf die Bewertung. Gängige Metriken sind die Betrachtung der Fläche oberhalb/unterhalb der Perturbationskurve beim Hinzufügen (Insertion) oder Entfernen (Deletion) von Pixeln basierend auf deren Wichtigkeit innerhalb einer Feature-Importance-Erklärung. Hier zeigt sich, dass beim Hinzufügen von Pixeln die dichteren Erklärungen bevorzugt werden, während beim Entfernen die weniger dichten Erklärungen eine bessere Bewertung erhalten. Diese Ergebnisse können manipuliert werden. So führt zum Beispiel das Glätten einer weniger dichten Erklärung zu deutlich besseren Ergebnissen bei den Insertion-Metriken. Dies zeigt, dass die Bewertungsergebnisse stark von der Struktur der Erklärung abhängen und nicht notwendigerweise die tatsächliche Qualität der XAI-Methode widerspiegeln.

3.3.4. Inkonsistenz zwischen verschiedenen Metriken

Eine weitere Herausforderung ist die mangelnde Kohärenz zwischen verschiedenen Bewertungsmetriken. So zeigen Insertion- und Deletion-Metriken typischerweise eine starke negative Korrelation – was gut bei der einen Metrik abschneidet, schneidet bei der anderen schlecht ab. Dabei sollen beide Metriken die Correctness bewerten, d. h., ob Modellverhalten und Erklärung übereinstimmen. Auch weitere Metriken, wie das sogenannte Pointing Game (eine vereinfachte Aufgabe für die Lokalisierung von Objekten in Bildern) weisen nahezu keine Korrelation zu anderen Bewertungsmetriken für die eigentliche Aufgabe auf.

3.3.5. Aufgaben-spezifische Artefakte

Wie bereits zuvor beschrieben, sind XAI-Methoden grundsätzlich kontextabhängig. Das gilt auch für Bewertungsmetriken, deren Ergebnisse stark von den verwendeten Aufgaben und Datensätzen abhängen können [12, 32]. So zeigen einige XAI-Methoden wie Guided Backpropagation auf Standardaufgaben wie MNIST oder ImageNet schlechte Ergebnisse bei sogenannten »Sanity Checks«, während sie auf künstlich konstruierten Aufgaben, bei denen die korrekte Erklärung bekannt ist, deutlich besser abschneiden.

Dies deutet darauf hin, dass einige beobachtete »Fehlschläge« von XAI-Methoden weniger auf inhärente Schwächen der Methoden zurückzuführen sind als vielmehr auf Artefakte der verwendeten Datensätze. Bei MNIST und ImageNet beispielsweise enthält jedes Bild typischerweise nur ein zentriertes Objekt, was dazu führen kann, dass auch zufällige oder edge-basierte Erklärungen gut scheinen, da sie zufällig das relevante Objekt hervorheben.

3.4. Menschbasierte Bewertung von XAI

Die menschbasierte Evaluierung von XAI-Methoden ist, wie in Abschnitt 3.2 beschrieben, unverzichtbar, um die tatsächliche Wirkung von Erklärungen auf reale Anwenderinnen und Anwender zu prüfen. Gleichzeitig sind aber auch hier eine Reihe von Herausforderungen bei der korrekten Anwendung und Validität der Ergebnisse festzustellen [17, 22].

3.4.1. Methodische Herausforderungen bei Nutzerstudien

Ein zentraler Kritikpunkt an menschbasierter Evaluation betrifft die methodischen Schwächen vieler Nutzerstudien. So werden zum Beispiel oft nur menschbasierte, aber nicht anwendungsbasierte Evaluierungen durchgeführt. Dabei zeigt sich allerdings, dass Ergebnisse auf vereinfachten Aufgaben deutlich von Ergebnissen in realen Anwendungen abweichen können [7]. Dies ist besonders problematisch, wenn die gewählten Aufgaben nicht die eigentlichen kognitiven Anforderungen der Zielanwendung widerspiegeln.

Weiterhin sind inadäquate Messgrößen ein Problem. Die Definition von Messgrößen wie »Vertrauen«, »Verständnis« oder »Usability« ist oft unklar und inkonsistent. Dadurch können die Ergebnisse verschiedener Studien nur unzureichend verglichen werden.

Außerdem weisen viele Studien eine unzureichende Stichprobengröße auf oder rekrutieren Teilnehmer, die nicht repräsentativ für die eigentliche Nutzergruppe sind. Dabei variieren die Teilnehmerzahlen stark – von wenigen Fachleuten bis zu über 1000 Nicht-Fachleuten – ohne dass eine klare Begründung für die gewählte Stichprobengröße gegeben wird [22].

Ebenfalls entspricht das experimentelle Design teils nicht den Ansprüchen für gute Nutzerstudien. Die Wahl zwischen »between-subjects«, »within-subjects« oder »mixed Designs« ist oft nicht ausreichend begründet. Dabei beschreibt between-subjects, dass Studienteilnehmer jeweils nur eine Parameterkonfiguration der Studie sehen, was die notwendige Stichprobengröße erhöht, während beim within-subjects-Design die Teilnehmer mehrere Parameterkonfigurationen sehen. Dies kann allerdings dazu führen, dass im Verlauf der Studie die Teilnehmer in der ursprünglichen Aufgabe besser abschneiden und die Ergebnisse somit verfälscht werden.

3.4.2. Inkonsistente Ergebnisse zwischen Studien

Ein weiterer wesentlicher Kritikpunkt betrifft die mangelnde Reproduzierbarkeit der Ergebnisse zwischen verschiedenen Nutzerstudien. So werden zwischen den Studien heterogene Effekte auf die menschliche Leistung in einer Aufgabe gemessen [17]. Insbesondere die Effekte von Saliency Maps sind äußerst heterogen. Während einige Studien Leistungsvorteile feststellen, zeigen andere Null-Effekte oder sogar Leistungseinbußen. Dies macht es schwierig, allgemeine Aussagen über die Wirksamkeit von XAI-Methoden zu treffen.

Oft besteht zudem eine Diskrepanz zwischen subjektivem und objektivem Verständnis [8]. Prägend hierfür ist die sogenannte »Illusion of Explanatory Depth« (»Illusion der Erklärungstiefe«), bei der Nutzerinnen und Nutzer glauben, das Modell durch XAI zu verstehen, tatsächlich aber nicht in der Lage sind, ihre vermeintlichen Erkenntnisse praktisch anzuwenden [9].

Die Effektivität von XAI-Methoden hängt stark von der Art der Aufgabe ab [17]. So sind Saliency Maps bei Aufgaben bezüglich des ML-Modells selbst (z. B. Fehleridentifikation im Modell) tendenziell hilfreicher als bei Aufgaben bezüglich der verwendeten Daten wie Bilder (z. B. menschliche Klassifikation eines Bilds mithilfe von ML und XAI).

3.4.3. Bias und Verzerrung durch menschliche Faktoren

Ein weiterer kritischer Aspekt menschbasierter Evaluation ist die Anfälligkeit für menschliche Bias: Nutzerinnen und Nutzer neigen dazu, ML-Prädiktionen zu übernehmen, selbst wenn diese falsch sind. Saliency Maps können dieses Verhalten verstärken, indem sie sie glauben lassen, das ML-Modell sei korrekt, was zu einer höheren Übereinstimmung mit ML-Prädiktionen führt, ohne dass dies mit einer höheren Leistung in der eigentlichen Aufgabe einhergeht [17].

Durch die Menge an dargestellten Informationen können XAI-Erklärungen, insbesondere Merkmalswichtigkeit, auch fehlinterpretiert werden. Saliency Maps können dazu führen, dass fälschlicherweise angenommen wird, dass die Merkmale, die bei korrekten Prädiktionen verwendet wurden, relevant sind. Die gleichen Merkmale sind aber bei falschen Prädiktionen irrelevant [17].

Zudem können Erklärungen durch die Menge an dargestellten Informationen kognitiv überlasten, was wiederum die Leistung bei Aufgaben beeinträchtigen kann. Dies kann dazu führen, dass Nutzerinnen und Nutzer trotz besserer Erklärungen schlechtere Entscheidungen treffen [1].

3.4.4. Kontextabhängigkeit der Effekte

Ein weiterer zentraler Kritikpunkt betrifft die starke Kontextabhängigkeit der Effekte von XAI-Methoden. Wie in Abschnitt 3.4.2 bereits angedeutet, hängen die Effekte von Erklärungen stark von der Genauigkeit des ML-Modells ab, und das wiederum unterschiedlich je nach Anwendungsfall [17]. Bei Aufgaben, bei denen Erklärungen bezüglich des ML-Modells interpretiert werden sollen, wie dem Debugging von Modellen, sind Saliency Maps tendenziell hilfreicher, wenn das Modell falsch liegt. Bei bildfokussierten Aufgaben, also solchen, bei denen aus dem ML-Modell Neues gelernt werden soll, sind sie eher dann nützlich, wenn das Modell richtig liegt.

Zusätzlich sind die Ergebnisse der Studien oft vom spezifischen Datensatz abhängig. Gerade bei Bilddaten hängt die Effektivität von Saliency Maps davon ab, welche Eigenschaften die verwendeten Bilder haben. Sie sind eher hilfreich, wenn die Klassenzuordnung durch lokalisierte, eindeutige Objekte bestimmt wird, weniger hilfreich hingegen bei komplexen Szenen oder bei Klassendifferenzierung durch Objekt-Objekt-Beziehungen.

Die Effektivität von XAI-Methoden variiert auch je nach persönlichen Merkmalen wie Erfahrung und Expertise. Dabei ist es wahrscheinlich, dass Fachleute andere Anforderungen an Erklärungen haben als Nicht-Fachleute, was in vielen Studien zu wenig berücksichtigt wird [17].

3.5. Praktische Implikationen

Die vorangehenden Kritikpunkte an der Evaluierung von XAI-Methoden scheinen theoretisch, haben allerdings für die Anwendung in spezifischen Szenarien konkrete Auswirkungen: Um sicherzugehen, dass eine XAI-Methode für einen Anwendungsfall wie gewünscht funktioniert, ist eine Kombination aus verschiedenen Evaluierungsmethoden entscheidend. Dabei können automatisierte Metriken und menschbasierte Evaluationen durchaus Informationen über die Performance von XAI-Methoden liefern. Allerdings kann nur eine Studie mit echten Anwendungsszenarien und echten Personen auf realistischen Ergebnissen ein aussagekräftiges Ergebnis über die Performance im realen Einsatz liefern. Dabei ist insbesondere auf ein korrektes und aussagekräftiges Studiendesign zu achten.

Um nicht für jeden Einsatz von XAI-Methoden eine spezifische Nutzerstudie durchführen zu müssen, ist es allerdings auch möglich, sich an bestehenden Studien zum gleichen Thema zu orientieren. Für das Debugging von ML-Modellen mittels XAI stellt das nächste Kapitel eine Übersicht von Studien vor.

4. Studien zu XAI für Debugging-Aufgaben

Die zuvor dargelegten Herausforderungen bei der Evaluation von XAI-Methoden verdeutlichen ein zentrales Dilemma: Während es theoretisch zahlreiche XAI-Methoden gibt, ist es aufgrund der inkonsistenten und oft widersprüchlichen Evaluationsergebnisse schwierig zu bestimmen, welche Methoden für spezifische Anwendungsfälle tatsächlich geeignet sind. Diese Problematik ist besonders ausgeprägt beim Debugging von ML-Modellen, da hier die Korrektheit und Zuverlässigkeit der Erklärungen sehr wichtig sind.

Während Kapitel 2 verschiedene Kategorisierungsansätze für XAI-Methoden vorstellte und deren theoretische Einordnung ermöglichte, zeigten die in Kapitel 3 beschriebenen Evaluationsprobleme, dass sowohl automatisierte Metriken als auch menschbasierte Bewertungen erhebliche methodische Schwächen aufweisen. Diese Erkenntnisse werfen die praktische Frage auf: Wie können Entwicklungs- und Validierungsteams trotz dieser Unsicherheiten fundierte Entscheidungen über den Einsatz von XAI-Methoden für das Debugging treffen?

Im Gegensatz zu anderen Anwendungsfällen für XAI – wie der Schaffung angemessenen Vertrauens in ML-Modelle – hat die Anwendung für das Debugging einen praktischen Vorteil: Ob und wie viele Fehler Nutzerinnen und Nutzer in einem Datensatz erkennen, ist klar und eindeutig messbar. Eine XAI-Methode ist dann nützlich, wenn sie dabei hilft, Fehler in Daten bzw. Modellen zu erkennen.

Dennoch zeigen die generell existierenden und die in diesem Kapitel vorgestellten Studien zu XAI für das Debugging eine heterogene Landschaft verschiedener Ansätze und teils widersprüchlicher Ergebnisse. Diese Heterogenität spiegelt auch die allgemeinen Herausforderungen des XAI-Feldes wider, die zuvor beschrieben wurden.

Im Folgenden werden einige empirischen Studien dargestellt und ihre praktischen Implikationen für Debugging-orientierte XAI-Anwendungen aufgezeigt. Ein besonderer Fokus liegt dabei auf der praktischen Relevanz der Ergebnisse. Im Gegensatz zu rein theoretischen Ansätzen betonen alle vorgestellten Studien die Notwendigkeit menschzentrierter Evaluation – denn was technisch als »gute Erklärung« gilt, muss auch für die Zielgruppe verständlich und nützlich sein.

Die folgenden Studien liefern konkrete Erkenntnisse darüber, wie der Einsatz von XAI in der Praxis gestaltet sein kann. Durch die Zusammenfassung dieser Forschungsergebnisse soll dieses Kapitel bei der Nutzung von XAI in realen Anwendungsfällen unterstützen.

4.1. Übersicht relevanter Studien

4.1.1. Debugging Tests for Model Explanations

Diese Studie von Adebayo et al. [5] ist eine der ersten systematischen Untersuchungen, die die praktische Nützlichkeit von Post-hoc-Erklärungen für das Debugging von ML-Modellen evaluiert. Die Autoren greifen ein zentrales Problem auf: Während XAI-Methoden oft als universelles Werkzeug zur Modellvalidierung beworben werden, fehlen bislang klare Richtlinien, wann und wie sie effektiv eingesetzt werden können. Die Studie entstand im Kontext der wachsenden Kritik an XAI, insbesondere bezüglich der »Sanity Checks« [3], die Adebayo bereits 2018 entwickelt hatte.

Die Studie testet, ob Methoden zur Merkmalsattribution (z. B. Grad-CAM, LIME) dabei helfen können, verschiedene Arten von Modellfehlern zu identifizieren. Dazu wurden künstliche Fehler in Trainingsdaten und Modellarchitekturen eingeführt. Diese enthielten Parameter-Bugs (Modellgewichte werden vor der Vorhersage neu initialisiert), Label-Bugs (10 Prozent der Trainingsdaten sind falsch gelabelt), Scheinkorrelationen (Tiere werden immer vor spezifischen Hintergründen präsentiert) und Testzeit-Bugs (Daten außerhalb der Trainingsdatenverteilung).

In einer Nutzerstudie sollten 54 Fach- und Nicht-Fachleute anhand von Modellvorhersagen und zugehörigen Erklärungen beurteilen, ob das Modell für den kommerziellen Einsatz bereit ist.

Tabelle 3: Zusammenfassung der Ergebnisse für »Debugging Tests for Model Explanations« [5].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Als Qualitätssicherungstester für ein Modell zur Tierklassifikation für Hundebilder sollte basierend auf Ausgabe und Erklärung bewertet werden, ob das trainierte Modell bereit zum Verkauf sei.
Kernergebnisse	• Scheinkorrelationen (z. B. Hintergrundabhängigkeit) werden durch Heatmaps erkennbar. • Label-Bugs bleiben oft unentdeckt. • Nutzer verlassen sich stärker auf Modellvorhersagen als auf Erklärungen. • XAI-Methoden, die Gradienten nutzen, sind ungeeignet für die Diagnose von Parameter-Bugs.
Debugging-Empfehlungen	• Erklärungsmethoden sollten gezielt für spezifische Fehlertypen validiert werden. • Automatische Checks und Nutzerstudien sollten kombiniert verwendet werden. • Je nach Szenario sollte eine spezifische XAI-Methode ausgewählt werden, z. B. Grad-CAM für Scheinkorrelationen.

Diese Studie zeigt, dass nicht alle Fehlerarten mit den gleichen XAI-Methoden identifiziert werden können. Dies ist in der Praxis besonders problematisch, da die jeweils vorliegende Fehlerart nicht bekannt ist. Somit sollten verschiedene XAI-Methoden genutzt werden. Zusätzlich kommt die Studie zu der Erkenntnis, dass Anwender mehr auf die Modellvorhersagen als auf die Erklärungen achten. Hier könnten Informationen oder Schulungen zur korrekten Nutzung von XAI-Ergebnissen den praktischen Nutzen von XAI verbessern.

4.1.2. Post-hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlations

Diese Folgestudie baut auf den Erkenntnissen der zuvor vorgestellten Studie auf und fokussiert sich speziell auf die Erkennung von unbekannten Scheinkorrelationen, hier Spurious Signals (Scheinsignale) genannt [4] – ein besonders praxisnahes Problem, da in der Realität die Fehlerfälle in Datensätzen meist nicht im Vorhinein bekannt sind. Für diese Studie wurden drei Anwendungen genutzt, darunter zwei medizinische (die Schätzung des Alters mittels Hand-Röntgenbilder und die Erkennung von Arthritis in Knie-Röntgenbildern) sowie die Klassifikation von Hunderassen.

Die Studie testet drei Darstellungsweisen von XAI-Erklärungen (Feature Attribution, konzeptbasierte Erklärungen und Beispiele aus den Trainingsdaten) auf ihre Fähigkeit, unbekannte Scheinsignale zu erkennen. Dazu wurden künstliche Artefakte in medizinische Bilddatensätze eingeführt. Die verwendeten Artefakte waren dabei sichtbare Tags (z. B. Krankenhausetiketten), ein weißer Streifen im Bild oder kaum sichtbare gaußsche Unschärfe im Hintergrund.

In einer Nutzerstudie mit 200 Personen (jeweils die Hälfte davon Fachleute bzw. Nicht-Fachleute) wurden zwei Gruppen gebildet: Eine Gruppe wusste über die Scheinsignale Bescheid, die andere nicht.

Tabelle 4: Zusammenfassung der Ergebnisse für »Post-hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation« [4].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Als Qualitätssicherungstester für Modelle für den jeweiligen Einsatz sollte bewertet werden, ob das trainierte Modell bereit zum Einsatz sei.
Kernergebnisse	• Post-hoc-Erklärungen sind nur effektiv, wenn das Scheinsignal vorab bekannt ist. Dabei sind konzeptbasierte Erklärungen am effektivsten. Insbesondere schlecht sichtbare Verzerrungen werden am wenigsten erkannt. • Selbst wenn Modelle nicht von Scheinsignalen abhängen, werden diese von Methoden zur Merkmalsattribution als relevant angezeigt.
Debugging-Empfehlungen	• Merkmalsattribution sollte nur bei Verdacht auf sichtbare Shortcuts eingesetzt werden. • Um konzeptbasierte Erklärungen effektiv einsetzen zu können, sollten Scheinsignale bereits im Voraus bekannt und definiert sein, um spezifisch auf Abhängigkeit von diesen Signalen testen zu können. • Die Darstellung von Beispielen ist dann nützlich, wenn diese das Scheinsignal enthalten.

Diese Studie zeigt, dass zusätzliche Validierungsschritte notwendig sind, die über XAI-Methoden hinausgehen. Dies gilt insbesondere für unbekannte Scheinsignale, die in der Praxis das größte Problem darstellen. Die Empfehlung, konzeptbasierte Erklärungen für vordefinierte Konzepte zu nutzen, bietet einen praktischen Ansatzpunkt für Domänenexpertinnen und -experten, die spezifische Fehlerquellen vermuten.

4.1.3. How Can Explainability Methods Support Bug Identification in Computer Vision Models?

Diese Studie von Balayn et al. [6] fokussiert sich ebenfalls mit dem Ziel, das Debugging von ML-Modellen mittels XAI zu untersuchen, nur auf ML-Fachleute bzw. Anwender. Durch Einflüsse aus dem Bereich der Human-Computer-Interaction liefert diese Arbeit zusätzlich Informationen zu möglichen Darstellungsformen von XAI. Hierfür wurde mit einer Anwendergruppe ein interaktives Werkzeug (»Explainability Probe«) entwickelt, das verschiedene Erklärungstypen kombiniert, insbesondere lokale Erklärungen (Saliency Maps), globale Erklärungen im Sinne von statistischen Zusammenfassungen der Zusammenhänge zwischen Klassen, textuelle Erklärungen bezüglich der Konzepte je Klasse und Kontrafakte. Diese Erklärungen konnten über eine interaktive Konfusionsmatrix erkundet werden.

Wie in den vorhergehenden Studien wurden Fehlerquellen ebenfalls bewusst in den Datensatz eingebracht. Gewählt wurde ein Datensatz zur Klassifikation verschiedener Vogelarten. In diesem Fall wurden Scheinsignale (z. B. Kakteen im Hintergrund bei Gilaspechten), unvollständige Merkmale (z.B. Farbe Rot nur bei männlichen Hakengimpeln), Trainings- und Testdaten aus unterschiedlichen Verteilungen und irrelevante Merkmale (z. B. Wasser im Hintergrund bei Kappensägern) eingefügt.

Diese Daten wurden von 18 ML-Praktikern mit unterschiedlichen Erfahrungsgraden mittels des interaktiven Werkzeugs untersucht.

Tabelle 5: Zusammenfassung der Ergebnisse für »How Can Explainability Methods Support Bug Identification in Computer Vision Models?« [6].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Die Teilnehmer sollten ein Vogelklassifikationsmodell untersuchen und entscheiden, ob es bereit für den Einsatz ist oder ob und wie nachgebessert werden muss.
Kernergebnisse	• Zwei Drittel identifizierten Bugs durch XAI-Methoden – durchschnittlich 3,5 Bugs pro Person, bis zu sieben Bugs bei erfahrenen Entwicklern. • Erfahrene Entwickler nutzten interaktive Werkzeuge effektiver als weniger erfahrene. • Saliency Maps allein sind oft unzureichend, eine Kombination mit weiteren Informationen oder Erklärungen ist notwendig. • Teilnehmern fielen in »Explainability Traps« (z. B. Confirmation Bias bei Heatmaps, mangelnde Nutzung ungewohnter Informationen und übermäßiges Vertrauen in Erklärungen).
Debugging-Empfehlungen	• Interaktive Werkzeuge mit mehreren Erklärungstypen (z. B. die Konfusionsmatrix-Visualisierung) helfen beim Debugging. • Die Anleitung sollte an den Erfahrungsgrad der Zielgruppe angepasst werden und Informationen zu »Explainability Traps« enthalten.

Diese Studie liefert konkrete Designprinzipien für XAI-Werkzeuge und zeigt, dass die Effektivität von XAI stark vom Erfahrungsgrad der Nutzer abhängt. Erklärungen interaktiv zu gestalten, bietet grundsätzlich einen Vorteil für deren Nutzung, wobei die Größe des Vorteils von der jeweiligen fachlichen Expertise abhängt. Die Identifikation von »Explainability Traps« hilft zudem dabei, typischen Fehlern bei der Interpretation von Erklärungen vorzubeugen.

4.1.4. From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation

Eines der Probleme von Saliency Maps ist, dass diese zwar zeigen, wo in den Eingabedaten relevante Informationen zu finden sind, aber nicht, was das Modell eigentlich erkannt hat. Die gleiche Einschränkung gilt nicht notwendigerweise für konzeptbasierte Erklärungen. Deshalb wurde von Achtibat et al. [2] ein konzeptbasierter Erklärungsansatz namens Concept Relevance Propagation vorgestellt. Dieser soll insbesondere für das Debugging von ML-Modellen hilfreich sein und wurde deshalb ebenfalls mittels einer Nutzerstudie evaluiert. Hierfür wurde der Ansatz mit bestehenden Saliency-Map-Methoden (Integrated Gradients, SHAP, Grad-CAM und Layerwise Relevance Propagation) verglichen.

Als Shortcut für die Studie wurde ein schwarzer Bildrand in den ImageNet-Datensatz eingebaut. Die Studie wurde mit 125 Personen, 25 pro Erklärungsmethode und rekrutiert über Amazon Mechanical Turk, durchgeführt. Weiterhin wurden von den Autoren selbst Wasserzeichen im ImageNet-Datensatz, als in einem real vorhandenen Datensatz existierende Shortcuts, untersucht.

Tabelle 6: Zusammenfassung der Ergebnisse für »From Attribution Maps to Human-Understandable Explanations through Concept Relevance Propagation« [2].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Hat der schwarze Rand im Bild die Vorhersage beeinflusst? (Antwort: Ja/Nein, mit zusätzlicher Einschätzung der Sicherheit der Entscheidung und Bewertung der visuellen Deutlichkeit der Erklärung)
Kernergebnisse	• Mittels konzeptbasierter Erklärungen lassen sich viele Shortcuts erkennen, während Grad-CAM schlechter abschneidet und Integrated Gradients keinen nennenswerten Mehrwert liefert. • Nutzer der konzeptbasierten Erklärungen empfanden diese, vermutlich wegen der höheren Komplexität, als am undeutlichsten, trotz bester Fehlererkennung. • Integrated Gradients erhielt die beste Bewertung bei der Deutlichkeit der Erklärung, obwohl die Erkennung von Shortcuts am schlechtesten funktionierte.
Debugging-Empfehlungen	Konzeptbasierte Erklärungen und Saliency Maps sind hilfreich für das Debugging. Zusätzlich können Beispielerklärungen einen Mehrwert liefern. Für komplexe Erklärungen sollte eine Schulung bereitstehen, um diese angemessen nutzen zu können.

Diese Studie zeigt, dass komplexere Erklärungen zwar deutlich schwieriger zu verstehen sind, aber dennoch effektiver bei der Identifikation von versteckten Shortcuts sein können. Allerdings wurden in dieser Studie vor allem optisch auffällige Shortcuts verwendet. Zusätzlich zeigt sich, dass Erklärungen, die als gut wahrgenommen werden, nicht zwangsweise die Fehlererkennung verbessern. So wurden die Erklärungen von Integrated Gradients mit hohen Werten in der Sicherheit der Befragten und der visuellen Deutlichkeit der Erklärung bewertet, während sie gleichzeitig für die Fehlererkennung nicht hilfreich waren.

4.1.5. Designing a Direct Feedback Loop between Humans and Convolutional Neural Networks through Local Explanations

Nützlicher als die reine Diagnose von Fehlern mittels XAI wäre es, die erzeugten Erklärungen direkt zur Korrektur der bestehenden Fehler zu nutzen. Ein solcher Ansatz findet sich bei Sun et al. [26]. Sie stellen DeepFuse vor, ein interaktives System, das es ermöglichen soll, mittels XAI Convolutional Neural Networks (CNNs) zu debuggen, die in der Bildverarbeitung gängig sind. Als Grundlage für die Erklärungen in DeepFuse dient die Erzeugung von Erklärungen mittels Grad-CAM.

Inhalt der Studie ist dabei ein Modell zur Klassifikation des Geschlechts von Personen anhand von Bildern. Das Modell lernt einen grünen Stern als Shortcut für »männlich«, wofür bei einem Drittel der männlichen Trainingsbilder ein Stern oben links platziert wurde. In einer zweistufigen Nutzerstudie mit zwölf erfahrenen CNN-Entwicklerinnen und -Entwicklern (sechs aus der Industrie, sechs aus der Wissenschaft) sollten diese zunächst die »Reasonability« (Plausibilität) der Grad-CAM-Heatmaps beurteilen und bei nicht plausibler Merkmalswichtigkeit neue Zielobjekt-Grenzen annotieren. Diese Daten wurden zum Finetuning des Modells verwendet. Am nächsten Tag bewerteten sie auf gleiche Weise das feingetunte Modell, wobei DeepFuse eine Reasonability-Matrix als Dashboard bereitstellte – eine Erweiterung der Konfusionsmatrix, in der die Korrektheit einer Vorhersage und die Plausibilität der Merkmalsattribution manuell bewertet werden. Dabei gelten Shortcuts als »unreasonable accurate«, da zwar die Vorhersage des ML-Modells korrekt ist, aber die verwendeten Merkmale nicht plausibel sind.

Tabelle 7: Zusammenfassung der Ergebnisse für »Designing a Direct Feedback Loop between Humans and Convolutional Neural Networks through Local Explanations« [26].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Beurteilung der Attribution als plausibel (auf relevanten Bildbereichen) oder als unplausibel und Annotation neuer Zielobjekt-Grenzen. Im zweiten Teil erneute Bewertung des verbesserten Modells.
Kernergebnisse	Grad-CAM und damit DeepFuse führten zu einer höheren Genauigkeit des Modells und zu plausibleren Erklärungen bei allen Teilnehmern bei hoher Nutzerfreundlichkeit. Mit Grad-CAM und DeepFuse konnten alle den verwendeten Shortcut entdecken.
Debugging-Empfehlungen	Grad-CAM kann zur Shortcut-Erkennung und -Korrektur eingesetzt werden. Bei der Nutzung von XAI hilft Interaktivität für bessere Performance.

In der Praxis ist es hilfreich, Shortcuts in Modellen zu erkennen, allerdings entstehen Vorteile wesentlich durch die Korrektur dieser. Die Studie skizziert einen möglichen Ansatzpunkt, wie dies gelingen kann, obschon dieser Ansatz durch die manuelle Annotation der Ergebnisse und Erklärungen mit hohem Aufwand verbunden ist. Leider bietet die Studie nur Ergebnisse für recht offensichtliche Shortcuts wie den verwendeten grünen Stern in den Bildern.

4.1.6. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods

Im Gegensatz zu den vorherigen Studien verfolgt die von Colin et al. [10] eine andere Zielsetzung: Können Menschen mithilfe von Erklärungen tatsächlich das Verhalten von ML-Modellen vorhersagen und damit echtes Verständnis entwickeln? Die Annahme und Verbindung zum Debugging dabei ist, dass dieses nur dann gelingen kann, wenn XAI-Erklärungen dazu befähigen, das tatsächliche Modellverhalten vorherzusagen.

Aus der Fragestellung ergibt sich bereits das Studiendesign: Das Forschungsteam testete, ob sechs gängige XAI-Methoden Menschen dabei helfen können, die Entscheidungen von ML-Modellen auf neuen, ungesehenen Daten vorherzusagen. Als Methoden kamen dabei Saliency, Gradient × Input, Integrated Gradients, SmoothGrad, Occlusion und Grad-CAM zum Einsatz. Getestet wurde in drei Szenarien: die Klassifikation von Husky vs. Wolf mit Schnee-Shortcut im Hintergrund (siehe Abschnitt 1.1), die Klassifikation von Kitfuchs vs. Rotfuchs mit ähnlichen Erscheinungsbildern und die Unterscheidung von zwei ähnlichen Pflanzenblättern basierend auf feinen Unterschieden. Dafür wurden 1.150 mehrheitlich Nicht-Fachleute ohne KI-Kenntnisse über Amazon Mechanical Turk rekrutiert. Nachdem überprüft worden war, ob die Teilnehmer aufmerksam genug waren, standen je Datensatz 240 Personen (30 pro Bedingung in 8 Bedingungen) zur Verfügung. Diesen wurden entweder keine XAI-Erklärungen gezeigt, Erklärungen basierend auf den oben genannten Methoden oder eine zufällige Merkmalsattribution. Nach einer Trainingsphase sollten sie auf neuen Daten die Prädiktion des ML-Modells vorhersagen. Zusätzlich wurden noch technische Metriken für XAI, insbesondere eine Deletion-Metrik und Sparsity, getestet.

Tabelle 8: Zusammenfassung der Ergebnisse für »What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods« [10].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Nach Training mit/ohne/mit zufälliger Merkmalsattribution: Welche Vorhersage wird das ML-Modell für diese Bilder treffen?
Kernergebnisse	Je nach Szenario sind Methoden unterschiedlich gut oder teils sogar schädlich. Insbesondere gut sind Grad-CAM, Occlusion und SmoothGrad für die Shortcut-Erkennung. Saliency, SmoothGrad und Integrated Gradients sind hingegen hilfreich, um neue Strategien (Unterscheidung der Blätter) aus Modellen zu lernen. Nur wenige der Methoden führen tatsächlich zu einem besseren Modellverständnis der Befragten, während die automatische Evaluierung von XAI keinen Schluss auf den Nutzen der Erklärungen für sie zulässt.
Debugging-Empfehlungen	Die Methodenauswahl sollte stets anwendungsbasiert erfolgen. Bessere XAI-Methoden sollten beschreiben, was in einem Bild erkannt wird, nicht nur wo. Je nach Szenario sollten mehrere XAI-Methoden kombiniert werden.

Diese Studie verdeutlicht, dass eine rein technische Evaluierung von XAI zwar wegen des entstehenden Aufwands wünschenswert wäre, im Moment technisch allerdings noch nicht möglich ist: Eine Erklärung kann laut Metrik »faithful«, d. h. korrekt sein (siehe Abschnitt 2.4 für die Definition), aber Menschen trotzdem in die Irre führen und sogar bei der Erkennung von Shortcuts schaden. Deshalb sollten Methoden im spezifischen Anwendungsszenario getestet werden.

4.1.7. Interpretability is in the Eye of the Beholder: Human Versus Artificial Classification of Image Segments Generated by Humans Versus XAI

Wie auch die Studie von Colin et al. zuvor zeigte, muss eine technisch »gute« Erklärung nicht unbedingt für Menschen verständlich und nützlich sein. So geht auch die Studie von Müller et al. [18] davon aus, dass hohe Modell-Fidelity nicht automatisch zu hoher menschlicher Interpretierbarkeit führt. Diese Annahme wird anhand der Segmentierung von Bildern, d. h. der Zuteilung verschiedener Bildteile zu bestimmten Klassen, getestet.

In der Studie selbst wurde überprüft, wie gut Menschen verschiedene Segmenttypen interpretieren können. Die Segmente wurden dabei auf vier Arten erzeugt: Durch menschliche Blickdaten (Gaze), menschliche Zeichnungen (Drawing), Grad-CAM-Merkmalsattribution und XRAI-Segmentierungen. Als Daten wurden die drei Bildtypen Objekte, Indoor-Szenen und Landschaften verwendet. Mit den Segmenten auf diesen Bildern wurden drei Experimente durchgeführt: Bei der groben Unterscheidungsaufgabe sollten die Befragten entscheiden, ob ein Segment zur vorgegebenen Klasse passt (z.B. »Wüste«). Bei der feinen Unterscheidungsaufgabe ging es um schwierigere Klassendifferenzierungen (z. B. zwei Landschaften wie »Wüste« vs. »Weizenfeld«). Im dritten Experiment bewerteten die Befragten subjektiv, wie gut ein Segment eine Kategorie repräsentiert.

Die Evaluierung erfolgte mit insgesamt 48 Personen aus dem Teilnehmerpool der TU Dresden, dabei 24 pro Klassifikationsexperiment und 32 der 48 Personen für Experiment drei.

Tabelle 9: Zusammenfassung der Ergebnisse für »Interpretability is in the Eye of the Beholder: Human Versus Artificial Classification of Image Segments Generated by Humans Versus XAI« [18].

Kriterium	Zusammenfassung
Aufgabe für Teilnehmer	Teilnehmer sollten entscheiden, ob ein Bildsegment zur vorgegebenen Klasse gehört. Zusätzlich sollte im letzten Experiment bewertet werden, wie gut die generierten Segmente zu einer Klasse passen.
Kernergebnisse	Menschliche Segmente, sowohl per Blickdaten oder manuell annotiert, sind für Menschen am besten verständlich. Grad-CAM ist moderat hilfreich, insbesondere für Indoor- und Landschaftsbilder, während XRAI am schwersten interpretierbar ist. XRAI versagt besonders bei Indoor-Szenen. Die Interpretierbarkeit der Segmente hängt stark von der Segmentgenerierung, dem Bildtyp und der Aufgabe ab.
Debugging-Empfehlungen	XAI-Methoden sollten immer im konkreten Anwendungskontext getestet werden, z. B. mit dem konkreten Bildtyp und den Endanwenderinnen und -anwendern.

Für die praktische Anwendung bedeutet diese Studie, dass XAI-Tools nicht pauschal als »gut« oder »schlecht« bewertet werden können – ihre Effektivität hängt stark vom spezifischen Anwendungskontext ab. Zusätzlich zeigt die Studie, dass von Menschen annotierte Segmente am besten für Menschen verständlich sind, während die durch XAI generierten Erklärungen diesen im Moment (noch) nicht ebenbürtig sind.

4.2. Schlussfolgerungen und Empfehlungen

Die vorliegenden Studien zu XAI-Methoden für das Debugging von ML-Modellen offenbaren ein komplexes Bild, das sowohl Potenziale als auch erhebliche Grenzen der derzeitigen Technologien aufzeigt. Die Studien machen deutlich, dass XAI kein universelles Allheilmittel für die Validierung und Fehlerbehebung von KI-Systemen darstellt, sondern vielmehr ein kontextsensitives Instrument, dessen Einsatz auf den spezifischen Anwendungsfall abgestimmt werden muss.

Wie in den Kapiteln 2 und 3 erwähnt, zeigt sich in den Studien eindeutig, dass die Effektivität von XAI-Methoden hochgradig kontextabhängig ist. Der praktische Nutzen der XAI-Methoden hängt vom Fehlertyp, den betrachteten Daten und der Expertise der Nutzerinnen und Nutzer ab. So sind bestimmte XAI-Methoden zum Beispiel bei der Identifikation von bekannten Shortcuts effektiv, erweisen sich aber bei Label-Bugs oder unbekannten Shortcuts als unzureichend. Somit sollte in der praktischen Anwendung bereits eine Vermutung existieren, welche Fehlerquellen wahrscheinlich sind.

Besonders problematisch ist die Feststellung von mehreren Studien, dass technisch »gute« Erklärungen – gemessen an automatisierten Metriken – nicht zwangsläufig für Menschen verständlich und nützlich sind. Teilweise ist gegenteiliges der Fall: So können XAI-Erklärungen zu schlechteren Ergebnissen führen als die Nutzung keiner Erklärungen. Dies ist sehr herausfordernd für die praktische Anwendung und verdeutlicht, dass rein automatisierte Evaluierungen nicht ausreichen, um den tatsächlichen praktischen Nutzen von XAI-Methoden zu bewerten.

Die Forschungslage zeigt, dass die Integration von Interaktivität in XAI-Tools einen Mehrwert für das Debugging schaffen kann. Hier scheint es insbesondere effektiv zu sein, aktiv falsche Erklärungen korrigieren zu können und diese Informationen zum Nachtraining der ML-Modelle zu verwenden. Allerdings ist hierbei hoher manueller Aufwand zu erwarten.

Ein weiterer kritischer Aspekt, der aus den Studien hervorgeht, ist die Notwendigkeit von Schulungen und Informationen für die Nutzerinnen und Nutzer von XAI-Tools. Die Identifikation von typischen »Explainability Traps« wie Confirmation Bias bei Heatmaps oder die falsche Interpretation von Erklärungen bei Unsicherheit zeigt, dass der alleinige Einsatz von XAI-Methoden ohne angemessene Schulung sogar kontraproduktiv sein kann.

Eine Schwierigkeit der vorgestellten Studien ist ebenfalls die Auswahl der XAI-Methoden: Es stehen so viele zur Auswahl, dass in jeder Studie andere verwendet wurden. Dabei scheint Grad-CAM insgesamt recht gute Ergebnisse zu liefern, was aber auch dadurch bedingt sein kann, dass diese Methode bereits früh als Open-Source-Implementierung zur Verfügung stand und dies somit den Einsatz in vielen Studien erleichterte, während andere Methoden aufgrund späterer oder schlechterer Verfügbarkeit teils nur einmal in den Studien genutzt wurden.

Zusammenfassend bleibt das Bild, dass XAI-Methoden zwar einen praktischen Vorteil für das Debugging von ML-Modellen bieten können, für ihren effektiven Einsatz aber noch diverse Fragen offenbleiben. So scheint insbesondere die Frage, welche XAI-Methode im spezifischen Einsatzszenario besonders effektiv ist, nur mit einer Nutzerstudie und in der Folge mit hohem Aufwand beantwortbar zu sein.

5. Fazit

Dieses Whitepaper stellte den Einsatz von XAI für das Debugging von ML-Modellen vor und beschrieb die zugehörigen Herausforderungen. Dabei zeigt sich ein komplexes Bild der aktuellen Möglichkeiten und Grenzen jetziger Erklärungen für ML-Modelle.

In Kapitel 1 wurde kurz eingeführt, was der mögliche Vorteil von XAI für das Debugging von ML-Modellen sein könnte, und es wurde beschrieben, dass insbesondere die Auswahl der besten XAI-Methode für einen gegebenen Anwendungsfall ein Problem darstellt.

Um einen Überblick über bestehende Möglichkeiten zur Kategorisierung – und damit erste Kriterien zur Auswahl von XAI-Methoden – zu geben, wurden in Kapitel 2 ebensolche Kategorisierungen vorgestellt, insbesondere konzeptionelle, ergebnisbasierte, funktionsbasierte und gemischte Einordnungen. Um diese Kategorisierungen in der Praxis nutzen zu können, muss es möglich sein, XAI-Methoden eindeutige Eigenschaften zuzuordnen. Hier sind manche der XAI-Konzepte problematisch, da für sie bisher noch keine klaren Bewertungsverfahren existieren.

Die bestehenden Probleme bei der Bewertung von XAI wurden sowohl für die automatische Bewertung als auch für Nutzerstudien in Kapitel 3 vorgestellt. Dabei zeigt sich, dass es noch keinen Goldstandard für die Bewertung von XAI gibt und deshalb eine Kombination von automatischen und menschzentrierten Bewertungen stattfinden sollte. Grundsätzlich gilt, dass ein XAI-Verfahren insbesondere dann gut zu bewerten ist, wenn man die Performance in der echten Anwendung direkt messen kann, was im Debugging-Kontext zum Beispiel für die Erkennung von Shortcuts möglich ist.

Kapitel 4 bot einen Überblick über bestehende Studien bezüglich der Performance von XAI für das Debugging von ML-Modellen. Hier ergibt sich ein eher heterogenes Bild. Insgesamt zeigt sich zwar, dass XAI-Methoden grundsätzlich beim Debugging von ML-Modellen helfen können. Es gibt aber keine allgemein am besten funktionierende Methode, sondern sie müssen immer für den spezifischen Einzelfall evaluiert werden.

Diese Untersuchung macht insgesamt deutlich, dass XAI kein Allheilmittel für die Validierung und Fehlerbehebung von KI-Systemen darstellt. Die Effektivität von XAI-Methoden hängt maßgeblich vom spezifischen Kontext ab – vom Typ der zu identifizierenden Fehler, den verwendeten Daten, der gewählten Modellarchitektur und nicht zuletzt von der Expertise der Anwenderinnen und Anwender. Während beispielsweise Methoden zur Merkmalsattribution wie Grad-CAM bei der Identifikation visuell offensichtlicher und im Vorhinein bekannter Shortcuts erfolgreich sein können, gilt dies nicht gleichermaßen über alle Datensätze oder Anwendungsfälle hinweg.

Insgesamt ergibt sich das Bild, dass XAI nicht als isolierte Technologie für das Debugging von ML-Modellen verwendet werden sollte, sondern als Bestandteil eines umfassenden Validierungs- und Debugging-Prozesses. Idealerweise sollte dieser Prozess eine methodische Vorgehensweise

beinhalten, die mit der Identifikation wahrscheinlicher Fehlerquellen beginnt, gezielt XAI-Methoden auswählt, menschzentrierte Evaluation durchführt, interaktives Tooling implementiert und dazu passende Schulungen bereitstellt.

Die Analyse der vorliegenden Studien unterstreicht die Unerlässlichkeit menschzentrierter Evaluierungen für XAI-Methoden. Nur durch Nutzerstudien in realistischen Anwendungsszenarien lässt sich der tatsächliche Nutzen einer XAI-Methode – sowohl generell als auch für das Debugging – bestimmen. Dies erfordert jedoch erhebliche Ressourcen und Expertise im Bereich der empirischen Forschung, was den praktischen Einsatz erschwert.

Generell vielversprechend erweisen sich interaktive XAI-Tools, die ermöglichen, aktiv mit den Erklärungen zu interagieren und Feedback zu geben. Solche Systeme können nicht nur bei der Fehleridentifikation helfen, sondern auch zur direkten Modellverbesserung beitragen. Allerdings ist auch hier der Entwicklungs- und Pflegeaufwand hoch.

Ebenfalls sollte bei der Nutzung von XAI darauf geachtet werden, dass Nutzerinnen und Nutzer angemessen informiert und geschult sind, weil der bloße Einsatz von XAI ansonsten kontraproduktiv sein kann. Typische »Explainability Traps« wie Confirmation Bias oder die Fehlinterpretation von Saliency Maps können zu falschen Schlussfolgerungen führen. Eine erfolgreiche Implementierung von XAI für das Debugging erfordert daher Schulungsprogramme und klare Richtlinien für die Interpretation von Erklärungen.

Zusammenfassend lässt sich festhalten, dass XAI für das Debugging von ML-Modellen ein hilfreiches Tool darstellen kann, der jetzige Stand der Forschung allerdings noch diverse Fragen zu einem praktikablen Einsatz offenlässt. Im Moment erfordert der praktische Einsatz noch ein tiefes Verständnis sowohl der technischen als auch der menschlichen Dimensionen des Problems.

Firmen, die bereits jetzt XAI zum Debugging von ML-Modellen einsetzen wollen, können sich an folgenden Hinweisen orientieren:

- Methoden sollten basierend auf spezifischen Fehlertypen ausgewählt werden: Identifizieren Sie zunächst die wahrscheinlichsten Fehlerquellen in Ihrem Anwendungskontext und wählen Sie XAI-Methoden entsprechend aus.
- XAI ist keine alleinige Lösung für die Validierung von KI-Systemen: Setzen Sie XAI als Teil eines umfassenden Validierungsprozesses ein, nicht als isolierte Lösung.
- Um bezüglich der Performance sichergehen zu können, sind im Moment Nutzerstudien notwendig: Wenn Ihr Anwendungsfall hohe Anforderungen an die KI-Validierung stellt, zum Beispiel in sicherheitskritischen Bereichen, planen Sie Ressourcen für die empirische Evaluierung der gewählten XAI-Methoden in Ihrem spezifischen Kontext ein.
- Nutzerinnen und Nutzer sollten für den XAI-Einsatz geschult sein: Entwickeln Sie Schulungsprogramme und etablieren Sie klare Prozesse für die Interpretation von XAI-Ergebnissen.
- Interaktive Tools sind zu bevorzugen: Erwägen Sie den Einsatz interaktiver XAI-Systeme, die direktes Feedback und iterative Verbesserungen ermöglichen.

Diese Tipps können heute schon helfen, XAI-Methoden möglichst gewinnbringend einzusetzen. Für das Forschungsfeld XAI, insbesondere im Debugging-Kontext, gilt, dass die Zukunft von XAI vermutlich nicht in der Entwicklung einer universellen »besten« Methode liegt, sondern in der Schaffung adaptiver, kontextsensitiver Systeme, die sich an die spezifischen Bedürfnisse und Arbeitsweisen anpassen. Weiterhin ist zu hoffen, dass über die nächsten Jahre weitere, untereinander gut vergleichbare Studien durchgeführt werden, die eindeutige Rückschlüsse über die bestehenden Evaluierungsmethoden für XAI und die Performance von XAI selbst zulassen. Nur durch diese menschzentrierte Herangehensweise kann das Potenzial erklärbarer KI für die Verbesserung der Qualität und Zuverlässigkeit von KI-Systemen vollständig ausgeschöpft werden.

6. Transparenzerklärung

Für die Verfeinerung des Entwurfes dieser Studie wurde FhGenie (Version Qwen3) verwendet. Es wurden jedoch keine Textausschnitte kopiert, ohne dass diese zuvor von den Autoren gründlich Korrektur gelesen und überarbeitet worden wären. Alle zitationsfähigen Passagen sind als solche gekennzeichnet.

7. Literaturverzeichnis

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM Digital Library. Association for Computing Machinery, New York,NY,United States, 1–14. DOI: https://doi.org/10.1145/3313831.3376615.

[2] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through Concept Relevance Propagation. Nat Mach Intell 5, 9, 1006–1019. DOI: https://doi.org/10.1038/s42256-023-00711-8.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps.

[4] Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. 2022. Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation http://arxiv.org/pdf/2212.04629.

[5] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. DOI: https://doi.org/10.48550/ARXIV.2011.05429.

[6] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. 2022. How can Explainability Methods be Used to Support Bug Identification in Computer Vision Models? In CHI Conference on Human Factors in Computing Systems. ACM Digital Library. Association for Computing Machinery, New York,NY,United States, 1–16. DOI: https://doi.org/10.1145/3491102.3517474.

[7] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, 454–464. DOI: https://doi.org/10.1145/3377325.3377498.

[8] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In 26th International Conference on Intelligent User Interfaces. ACM, New York, NY, USA, 307–317. DOI: https://doi.org/10.1145/3397481.3450644.

[9] Julien Colin, Thomas Fel, Remi Cadene, and Thomas Serre. 2021. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. arXiv https://arxiv.org/pdf/2112.04417.

[10] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. 2022. What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods. Advances in neural information processing systems 35, 2832–2845.

[11] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv https://arxiv.org/pdf/1702.08608.

[12] Benjamin Fresz, Lena Lörcher, and Marco Huber. 2024. Classification Metrics for Image Explanations: Towards Building Reliable XAI-Evaluations. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. ACM Digital Library. Association for Computing Machinery, Erscheinungsort nicht ermittelbar, 1–19. DOI: https://doi.org/10.1145/3630106.3658537.

[13] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications, 10, 1, 1096 https://arxiv.org/pdf/1902.10178.

[14] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2023. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions http://arxiv.org/pdf/2310.19775.pdf.

[15] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions.

[16] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences.

[17] Romy Müller. 2024. How explainable AI affects human performance: A systematic review of the behavioural consequences of saliency maps http://arxiv.org/pdf/2404.16042.

[18] Romy Müller, Marius Thoß, Julian Ullrich, Steffen Seitz, and Carsten Knoll. 2025. Interpretability is in the Eye of the Beholder: Human Versus Artificial Classification of Image Segments Generated by Humans Versus XAI. International Journal of Human–Computer Interaction 41, 4, 2371–2393. DOI: https://doi.org/10.1080/10447318.2024.2323263.

[19] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2022. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI.

[20] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.

[21] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. »Why Should I Trust You?«. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778.

[22] Yao Rong, Tobias Leemann, Thai-trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2022. Towards Human-centered Explainable AI: User Studies for Model Explanations https://arxiv.org/pdf/2210.11584.

[23] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. PMLR 70:3145-3153.

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise.

[25] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA, 2239–2250. DOI: https://doi.org/10.1145/3531146.3534639.

[26] Tong S. Sun, Yuyang Gao, Shubham Khaladkar, Sijia Liu, Liang Zhao, Young-Ho Kim, and Sungsoo R. Hong. 2023. Designing a Direct Feedback Loop between Humans and Convolutional Neural Networks through Local Explanations. Proc. ACM Hum.-Comput. Interact. 7, CSCW2, 1–32. DOI: https://doi.org/10.1145/3610187.

[27] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks.

[28] Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun Preece. 2019. Sanity Checks for Saliency Metrics https://arxiv.org/pdf/1912.01451.

[29] Benjamin Vandersmissen and Jose Oramas. 2023. On The Coherence of Quantitative Evaluation of Visual Explanations https://arxiv.org/pdf/2302.10764.pdf.

[30] Ziming Wang, Changwu Huang, and Xin Yao. 2024. A Roadmap of Explainable Artificial Intelligence: Explain to Whom, When, What and How? ACM Trans. Auton. Adapt. Syst. 19, 4, 1–40. DOI: https://doi.org/10.1145/3702004.

[31] Rosina O. Weber, Adam J. Johs, Prateek Goel, and João M. Silva. 2024. XAI is in trouble. AI Magazine 45, 3, 300–316. DOI: https://doi.org/10.1002/aaai.12184.

[32] Gal Yona and Daniel Greenfeld. 2021. Revisiting Sanity Checks for Saliency Maps https://arxiv.org/pdf/2110.14297.pdf

8. Abbildungsverzeichnis

9. Tabellenverzeichnis

KI-Fortschrittszentrum

Das KI-Fortschrittszentrum »Lernende Systeme und Kognitive Robotik« 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 Produktionstechnik und Automatisierung IPA sowie für Arbeitswirtschaft und Organisation IAO daran, Technologien aus der KI-Spitzenforschung in die breite Anwendung der produzierenden Industrie und der Dienstleistungswirtschaft zu bringen. Unterstützt werden sie dabei vom Institut für Arbeitswissenschaft und Technologiemanagement IAT der Universität Stuttgart. Finanzielle Förderung erhält das Zentrum vom Ministerium für Wirtschaft, Arbeit und Tourismus Baden-Württemberg.

Mission
Das KI-Fortschrittszentrum ist der anwendungsorientierte Zweig von Cyber Valley, Europas größter Forschungskooperation im Bereich der Künstlichen Intelligenz. Darüber hinaus ist das KI-Fortschrittszentrum Bestandteil von S-TEC, dem Stuttgarter Technologie- und Innovationscampus: www.s-tec.de

Das KI-Fortschrittszentrum schlägt die Brücke von der KI-Spitzenforschung in den Mittelstand und macht KI-Technologien für die Wirtschaft in Baden-Württemberg und darüber hinaus nutzbar. Als führender Innovationspartner für den Mittelstand arbeitet das Zentrum an Themen, die für den Einsatz von KI und Robotik branchenübergreifend von zentraler Bedeutung sind, beispielsweise Autonomie, Effizienz und Nachhaltigkeit, Mensch-Maschine-Interaktion sowie Vertrauen. Das KI-Fortschrittszentrum informiert Unternehmen über Technologietrends und deren Einsatzpotenziale und unterstützt sie bedarfsgerecht und niedrigschwellig bei der Entwicklung und Umsetzung von ambitionierten KI-Innovationen, damit sie die wirtschaftlichen Chancen der KI künftig noch besser nutzen können.

Vision
Das KI-Fortschrittszentrum ist ein Leuchtturm für erfolgreichen Technologietransfer in den Mittelstand und ermöglicht Unternehmen einen wirtschaftlichen und verantwortungsvollen Einsatz von Künstlicher Intelligenz und Robotik für unternehmerischen Erfolg sowie individuellen und gesellschaftlichen Nutzen.

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

Fraunhofer

Die Fraunhofer-Gesellschaft mit Sitz in Deutschland ist eine der führenden Organisationen für anwendungsorientierte Forschung. Im Innovationsprozess spielt sie eine zentrale Rolle – mit Forschungsschwerpunkten in zukunftsrelevanten Schlüsseltechnologien und dem Transfer von Forschungsergebnissen in die Industrie zur Stärkung unseres Wirtschaftsstandorts und zum Wohle unserer Gesellschaft. Seit ihrer Gründung als gemeinnütziger Verein im Jahr 1949 nimmt sie eine einzigartige Position im Wissenschafts- und Innovationssystem ein.

Knapp 32 000 Mitarbeitende an 75 Instituten und selbstständigen Forschungseinrichtungen in Deutschland erarbeiten das jährliche Finanzvolumen von 3,6 Mrd. €. Davon entfallen 3,1 Mrd. € auf das zentrale Geschäftsmodell von Fraunhofer, die Vertragsforschung. Im Vergleich zu anderen öffentlichen Forschungseinrichtungen bildet die Grundfinanzierung durch Bund und Länder lediglich das Fundament des jährlichen Forschungshaushalts. Sie ist die Basis für wegweisende Vorlaufforschung, die in den kommenden Jahren für Wirtschaft und Gesellschaft bedeutend wird. Das entscheidende Alleinstellungsmerkmal ist der hohe Anteil an Wirtschaftserträgen, der Garant ist für die enge Zusammenarbeit mit Wirtschaft und Industrie und die stetige Marktorientierung der Fraunhofer-Forschung: 2024 beliefen sich die Wirtschaftserträge auf 867 Mio. € des laufenden Haushalts. Ergänzt wird das Forschungsportfolio durch im Wettbewerb eingeworbene öffentliche Projektmittel, wobei eine ausgewogene Balance zwischen öffentlichen und wirtschaftlichen Erträgen angestrebt wird.

Wir produzieren Zukunft
Das Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, kurz Fraunhofer IPA, realisiert hoch innovative und nachhaltige Lösungen in der Produktionstechnik und Automatisierung für verschiedenste Zukunftsbranchen. Dies können Methoden, Komponenten und Geräte bis hin zu kompletten Maschinen und Anlagen sein. Die Lösungen stehen stets in Verbindung mit den strategischen Eckpfeilern des Instituts »Mass Sustainibility« und »Mass Personalization«. Seine Hauptaufgabe sieht das Institut im Wissens-, Innovations- und Technologietransfer von Forschungsergebnissen in Applikationen, um die Wettbewerbsfähigkeit der Unternehmen zu stärken. Dabei versteht es sich als unabhängiger Ansprechpartner, der neutral berät und Unternehmen mit genau auf deren Bedürfnisse zugeschnittenen Projektteams unterstützt.

Autorinnen und Autoren

Benjamin Fresz, M. Sc.
Fraunhofer-Institut für Produktionstechnik und Automatisierung IPA, Stuttgart

Benjamin Fresz studierte an der Universität Stuttgart Technische Kybernetik (B. Sc.) und Autonome Systeme (M. Sc.). Am Fraunhofer IPA ist er auf interdisziplinäre Forschung am Grenzbereich von Informatik, Recht und Sozialwissenschaften spezialisiert. Sein Interesse gilt insbesondere dem rechtskonformen Einsatz von KI-Systemen im Sinne der KI-Verordnung und der rechtskonformen und effektiven Nutzung von XAI. Im Rahmen des KI-Fortschrittszentrums war er an einigen AI-Explorer-Projekten und Quick Checks beteiligt.

Quynh Anh Tran

Quynh Anh Tran studiert Digital Business Management (B. Sc.) an der Universität Hohenheim. Im Rahmen ihrer studentischen Tätigkeit beschäftigte sie sich gemeinsam mit Benjamin Fresz mit Fragestellungen zur erklärbaren KI im Debugging. Ihre Interessen liegen zudem in den Bereichen generative KI und digitale Geschäftsprozesse.

Impressum

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

Benjamin Fresz
Telefon +49 711 970-1404
benjamin.fresz@ipa.fraunhofer.de

Quynh Anh Tran
quynhanh.tran@uni-hohenheim.de

Fraunhofer-Publica
http://dx.doi.org/10.24406/publica-7088

Titelbild
© LAONG – Adobe Stock

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

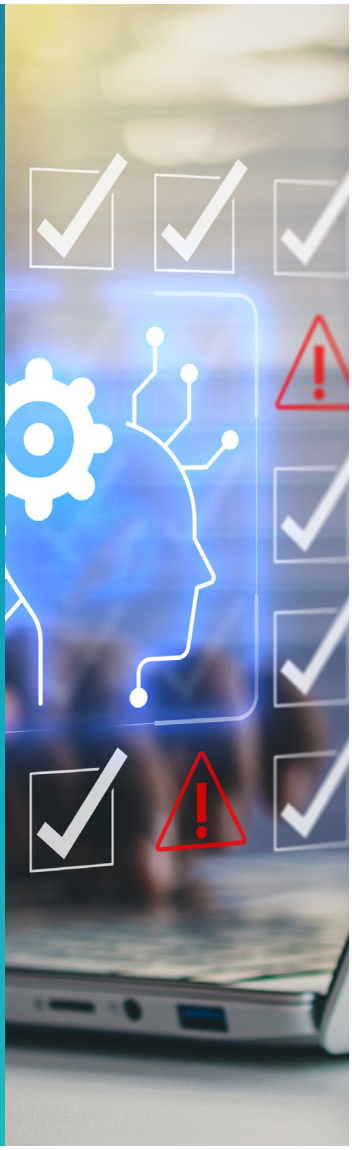
Kontakt

KI-Fortschrittszentrum
Fraunhofer IAO und Fraunhofer IPA
Nobelstraße 12
70569 Stuttgart
www.ki-fortschrittszentrum.de

Benjamin Fresz
Telefon +49 711 970-1404
benjamin.fresz@ipa.fraunhofer.de

Quynh Anh Tran
quynhanh.tran@uni-hohenheim.de

Fraunhofer IPA
Nobelstraße 12
70569 Stuttgart
www.ipa.fraunhofer.de



Gefördert durch

Baden-Württemberg
Ministerium für Wirtschaft,
Arbeit und Tourismus

Partner

Cyber Valley

S-TEC

Universität Stuttgart
Institut für Arbeitswissenschaft und
Technologiemanagement IAT