Taxonomie für Methoden der Erklärbaren Künstlichen Intelligenz

Projekt: Offenes Innovationslabor KI zur Förderung gemeinwohlorientierter (Go-KI)

Autoren: Nils Ole Breuer, Yonatan Demissie, Aray Karjauv, Tobias Küster

Datum: 03.02.2025

Inhaltsübersicht

Gefördert durch:

Bundesministerium für Arbeit und Soziales

aufgrund eines Beschlusses des Deutschen Bundestages

Einführung

Das Forschungsgebiet der erklärbaren KI (Explainable AI / XAI) erhält derzeit viel Aufmerksamkeit. Diese Aufmerksamkeit führt zu einer Flut neuer Erklärungsmethoden und -ansätze, die Licht ins Dunkle von intransparenten KI Modellen bringen sollen. Gleichzeitig führt dies aber zu einer unüberschaubaren und nicht verwaltbaren Menge an Methoden. Um den Überblick nicht zu verlieren, können Taxonomien für erklärbare KI hilfreich sein.

In diesem Dokument stellen wir unsere XAI-Taxonomie vor, die im Rahmen des Go-KI-Projekts entwickelt wurde. Im Gegensatz zu anderen XAI-Taxonomie-Ansätzen, die XAI-Methoden entweder auf der Basis von Eigenschaften der generierten Erklärungen kategorisieren und gruppieren, z.B. Feature-Zuweisungsmethoden, oder darauf, wie die Erklärungen generiert werden, folgt unsere Taxonomie einem menschenzentrierten Ansatz (Breuer et al., 2024). Wir kategorisieren und gruppieren XAI-Methoden danach, wie sie ausgewählt werden müssen, um ein bestimmtes Erklärungsziel zu erreichen, welches durch das zu erklärende System und die Zielgruppe der Erklärung definiert ist.

Das Dokument beginnt mit einer Literaturübersicht, welche zwei Arbeiten beschreibt, die ebenfalls mithilfe einer eigenen Taxonomie mehrere XAI Methoden kategorisieren. Die neuere Version des *XAI Method Finders* – eine Anwendung, die im Rahmen des Go-KI-Projekts entwickelt wurde, um passende XAI Methoden zu empfehlen – verwendet einen LLM-Assistenten, der diese Arbeiten als Wissensbasis nutzt, um Entscheidungen zu treffen (Breuer et al., 2024). In den nächsten beiden Kapiteln beschreiben wir unsere XAI Taxonomie und stellen die darauf basierende Kategorisierung bestimmter XAI Methoden vor.

Literaturübersicht

Schwalbe und Finzel (2024) erstellen eine systematische Übersicht über Umfragen zu XAI-Methoden und -Konzepten, und Ding et al. (2022) führten eine Auswertung der Anwendung und Herausforderungen von XAI-Methoden durch. Beide Arbeiten kategorisieren XAI-Methoden auf der Grundlage der von ihnen vorgeschlagenen Taxonomien.

Die wesentlichen Ergebnisse dieser beiden Arbeiten sind in den Abbildungen und Tabellen auf den folgenden Seiten wiedergegeben. Abbildung 1 und 2 zeigen die vorgeschlagene Taxonomie der beiden Arbeiten. Diese Taxonomien werden dann verwendet, um XAI-Methoden zu kategorisieren.

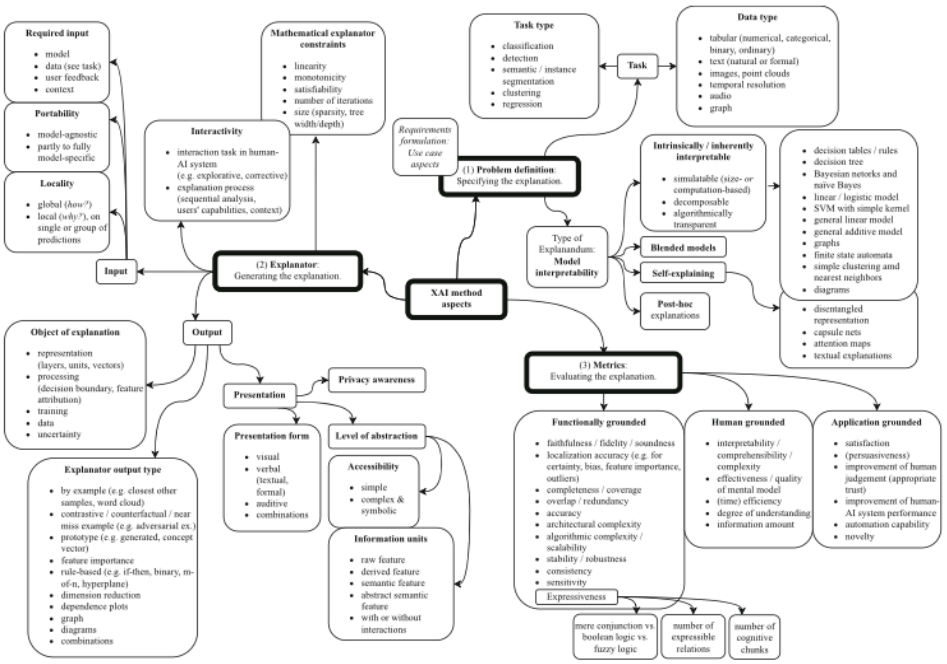


Abbildung 1: Vorgeschlagene Taxonomie (Schwalbe et al., 2024).

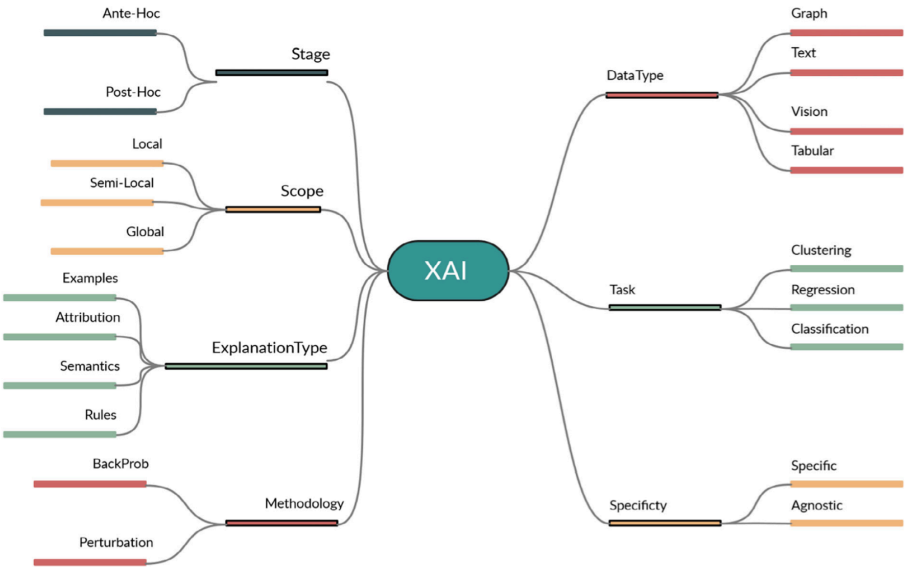


Abbildung 2: Vorgeschlagene Taxonomie (Ding et al., 2022).

Tabelle 1 zeigt, wie XAI-Methoden von Schwalbe (2024) kategorisiert wurden, und Tabellen 2 bis 5 zeigen die Kategorisierung von XAI-Methoden nach Ding et al. (2022), die für KI-Lösungen auf der Grundlage visueller Daten, KI-Lösungen auf der Grundlage tabellarischer Daten, textbasierte KI-Lösungen bzw. Graph-basierte KI-Lösungen verwendet werden.

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Self-explaining and blended models								
–	Hendricks et al. (2016)	cls		s		p	sym/vis	rules/fi
–	Kim et al. (2018b)	any		s		p	sym/vis	rules/fi
ProtoPNet	Chen et al. (2019a)	cls,img		s		p/r	vis	proto/fi
Capsule Nets	Sabour et al. (2017)	cls		s		r	sym	fi
Semantic Bottlenecks, ReNN, Concept Whitening	Losch et al. (2019), Wang (2018), Chen et al. (2020)	any		s		r	sym	fi
Logic Tensor Nets	Donadello et al. (2017)	any		b	✓	p/r	sym	rule
FoldingNet	Yang et al. (2017)	any,pcl		b		p	vis	fi/red
Neuralized clustering	Kauffmann et al. (2019)	any		b		p	vis	fi
Black-box heatmapping								
LIME, SHAP	Ribeiro et al. (2016), Lundberg and Lee (2017)	cls	✓	p		p	vis	fi/con
RISE	Petsiuk et al. (2018)	cls,img	✓	p		p	vis	fi
D-RISE	Petsiuk et al. (2021)	det,img	✓	p		p	vis	fi
CEM	Dhurandhar et al. (2018)	cls,img	✓	p		p	vis	fi/con
White-box heatmapping								
Sensitivity analysis	Baehrens et al. (2010)	cls		p		p	vis	fi
Deconvnet, (Guided) Backprop.	Zeiler and Fergus (2014), Simonyan et al. (2014), Springenberg et al. (2015)	img		p		p	vis	fi
CAM, Grad-CAM	Zhou et al. (2016), Selvaraju et al. (2017)	cls,img		p		p	vis	fi
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	fi
Concept-wise Grad-CAM	Zhou et al. (2018)	cls,img		p		p/r	vis	fi
SIDU	Muddamsetty et al. (2021)	cls,img		p		p	vis	fi
LRP	Bach et al. (2015)	cls		p		p	vis	fi

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Pattern attribution	Kindermans et al. (2018)	cls		p		p	vis	fi
–	Fong and Vedaldi (2017)	cls		p		p	vis	fi
SmoothGrad, Integrated Gradients	Smilkov et al. (2017), Sundararajan et al. (2017)	cls		p		p	vis	fi
Integrated Hessians	Janizek et al. (2020)	cls		p		p	vis	fi
Global representation analysis								
Feature Visualization	Olah et al. (2017)	img		p	✓	r	vis	proto
NetDissect	Bau et al. (2017)	img		p	✓	r	vis	proto/fi
Net2Vec	Fong and Vedaldi (2018)	img		p	(✓)	r	vis	fi
TCAV	Kim et al. (2018a)	any		p	✓	r	vis	fi
ACE	Ghorbani et al. (2019)	any		p	✓	r	vis	fi
–	Yeh et al. (2020)	any		p	✓	r	vis	proto
IIN	Esser et al. (2020)	any		p	(✓)	r	vis/sym	fi
Explanatory Graph	Zhang et al. (2018)	img		p	(✓)	p/r	vis	graph
Dependency plots								
PDP	Friedman (2001)	any	✓	p		p	vis	plt
ICE	Goldstein et al. (2015)	any	✓	p	✓	p	vis	plt
Rule extraction								
TREPAN, C4.5, Concept Tree	Craven and Shavlik (1995), Quinlan (1993), Renard et al. (2019)	cls	✓	p	✓	p	sym	tree
VIA	Thrun (1995)	cls	✓	p	✓	p	sym	rules
DeepRED	Zilke et al. (2016)	cls		p	✓	p	sym	rules
LIME-Aleph	Rabold et al. (2018)	cls	✓	p		p	sym	rules
CA-ILP	Rabold et al. (2020)	cls		p	✓	p	sym	rules
NBDT	Wan et al. (2020)	cls		p	✓	p	sym	tree

Name	Cite	Task	Model-agnostic?	Transp.	Global?	Obj. Expl.	Form	Type
Interactivity								
CAIPI	Teso and Kersting (2019)	cls,img	✓	p		r	vis	fi/con
EluciDebug	Kulesza et al. (2010)	cls	✓	p		r	vis	fi,plt
Crayons	Fails and Olsen Jr (2003)	cls,img	✓	t		p	vis	plt
LearnWithME	Schmid and Finzel (2020)	cls	✓	t	✓	p, r	sym	rules
Multi-modal phrase-critic model	Hendricks et al. (2018)	cls,img		p	✓	p	vis,sym	plt,rules
Inspection of the training								
–	Shwartz-Ziv and Tishby (2017)	any		p	✓	t	vis	dist
Influence functions	Koh and Liang (2017)	cls		p	✓	t	vis	fi/dist
Data analysis methods								
t-SNE, PCA	van der Maaten and Hinton (2008), Jolliffe (2002)	any	✓	p	✓	d	vis	red
k-means, spectral clustering	Hartigan and Wong (1979), von Luxburg (2007)	any	✓	p	✓	d	vis	proto

Abbreviations by column: *image data*=img, *point cloud data*=pcl; *Trans.*=transparency, *post-hoc*=p, *transparent*=t, *self-explaining*=s, *blended*=b; *processing*=p, *representation*=r, *development during training*=t *data*=d; *visual*=vis, *symbolic*=sym, *plot*=plt; *feature importance*=fi, *contrastive*=con, *prototypical*=proto, *decision tree*=tree, *distribution*=dist

Tabelle 1: Kategorisierung von XAI Methoden (Schwalbe et al., 2024).

Class	Method	YEAR	Ref.	Data Type	AH / PH	G / L	A / S
SM	SHAP	2007	[181]	ANY	PH	L	A
	LIME	2016	[237]	ANY	PH	L	A
	ϵ-LRP	2015	[16]	ANY	PH	L	S
	INTGRAD	2017	[276]	ANY	PH	L	S
	DEEPLIFT	2017	[265]	ANY	PH	L	S
	SMOOTHGRAD	2017	[269]	IMG	PH	L	S
	XRAI	2019	[138]	ANY	PH	L	S
	GRADCAM	2017	[255]	IMG	PH	L	S
	GRADCAM++	2018	[53]	IMG	PH	L	S
	CAM	2016	[328]	IMG	PH	L	S
	IS-CAM	2020	[203]	IMG	PH	L	S
	Score-CAM	2020	[288]	IMG	PH	L	S
	SSCAM	2020	[287]	IMG	PH	L	S
	LayerCAM	2021	[132]	IMG	PH	L	S
	XGrad-CAM [87]	2020	[88]	IMG	PH	L	S
	Smooth Grad-CAM pp	2019	[210]	IMG	PH	L	S
	RISE	2018	[219]	IMG	PH	L	S
CA	TCAV	2018	[146]	IMG	PH	L	A
	ACE	2019	[94]	IMG	PH	G	A
	CONCEPTSHAP	2020	[310]	IMG	PH	G	A
	CACE	2019	[96]	IMG	AH	G	A
CF	CEM	2018	[73]	IMG	PH	L	A
	ABELE	2020	[98]	IMG	PH	L	A
	L2X	2018	[56]	ANY	PH	L	A
	GUIDED PROTO	2019	[180]	IMG	PH	L	A
PR	MMD-CRITIC	2016	[145]	ANY	IN	G	A
	–	2017	[153]	ANY	PH	L	A
	PROTONET	2019	[55]	IMG	AH	G	S

L = Local; G = Global; A = Agnsotic; S = specific, AH = ant-hoc; PH = post-hoc, IMG = IMAGE.

Tabelle 2: XAI-Methoden für KI-Lösungen für visuelle Daten (Ding et al., 2022).

Type	Name	Ref.	YEAR	Data Type	AH / PH	G / L	A / S
FI	SHAP	[181]	2007	ANY	PH	G / L	A
	LIME	[237]	2016	ANY	PH	L	A
	LRP	[16]	2015	ANY	PH	L	A
	DALEX	[27]	2020	ANY	PH	L / G	A
	NAM	[4]	2020	TAB	PH	L	S
	CIU	[11]	2020	TAB	PH	L	A
	MAPLE	[222]	2018	TAB	PH / IN	L	A
RB	ANCHOR	[238]	2018	TAB	PH	L / G	A
	LORE	[99]	2018	TAB	PH	L	A
	LRI	[295]	2000	TAB	AH	L	S
	MLRULE	[72]	2008	TAB	AH	G / L	S
	RULEFIT	[87]	2008	TAB	AH	G / L	S
	SCALABLE-PRL	[307]	2017	TAB	AH	G / L	A
	RULEMATRIX	[193]	2018	TAB	PH	G / L	A
	IDS	[164]	2016	TAB	AH	G / L	S
	DECTEXT	[33]	2002	TAB	PH	G	S
	STA	[329]	2016	TAB	PH	G	S
	SKOPERULE	[87]	2020	TAB	PH	L / G	A
PR	GLOCALX	[257]	2019	TAB	PH	L / G	A
	MMD-CRITIC	[145]	2016	TAB	AH	G	S
	PROTODASH	[104]	2019	TAB	AH	G	A
	TSP	[277]	2020	TAB	PH	L	S
CF	PS	[28]	2011	TAB	IN	G / L	S
	CEM	[73]	2018	ANY	PH	L	S
	DICE	[201]	2020	ANY	PH	L	A
	FACE	[225]	2020	ANY	PH	L	A
	CFX	[7]	2020	TAB	PH	L	S

L = Local; G = Global; A = Agnostic; S = specific, AH = ant-hoc; PH = post-hoc, TAB = Tabular.

Tabelle 3: XAI-Methoden für KI-Lösungen für tabellarische Daten (Ding et al., 2022).

Type	Name	Ref.	YEAR	Data Type	AH / PH	G / L	A / S
SM	LIME	[237]	2016	ANY	PH	L	A
	INTGRAD	[276]	2017	ANY	PH	L	S
	L2X	[56]	2018	ANY	PH	L	A
	DEEPLIFT	[265]	2017	ANY	PH	L	S
	LIONETS	[199]	2019	ANY	PH	L	S
CA	–	[296]	2014	TXT	PH	L	S
	EXBERT	[124]	2019	TXT	PH	L	S
	–	[279]	2017	TXT	PH	L	S
PR	ANCHOR	[238]	2018	TXT	PH	L	A
	QUINT	[1]	2017	TXT	PH	L	S
	CRIAGE	[220]	2019	TXT	PH	L	S
	LASTS	[101]	2020	TXT	PH	L	S
	XSPELLS	[165]	2020	TXT	PH	L	S
	–	[232]	2019	TXT	PH	L	S
	DOCTORXAI	[213]	2020	ANY	PH	L	S

L = Local; G = Global; A = Agnsotic; S = specific, AH = ant-hoc; PH = post-hoc, TXT = TEXT.

Tabelle 4: XAI-Methoden für textbasierte KI-Lösungen (Ding et al., 2022).

Technique	Category	Learning	Task	Target	Black box	Direction	Design
Sensitivity analysis [18223]	Instance-based	No	GC/NC	N/E/NF	No	Backward	✗
GBP [18]		No		N/E/NF	No		✗
CAM [223]		No	GC	N	No		✗
Grad-CAM [223]		No		N	No		✗
GNNExplainer [312]		Yes	GC/NC	E/NF	Yes	Forward	Yes
PGExplainer [182]		Yes	G	E	No		Yes
GraphMask [250]		Yes		E	No		Yes
ZORRO [15]		No		N/NF	Yes		Yes
Causal Screening [95]		No		E	Yes		Yes
SubgraphX [317]		Yes		Subgraph	Yes		Yes
LRP [18253]		No		N	No	Backward	No
Excitation BP [54]		No		N	No	Backward	No
GNN-LRP [223]		No		Walk	No	Backward	Yes
GraphLime [128]		Yes	NC	NF	Yes	Forward	No
RelEx [324]		Yes		N / E	Yes	Forward	Yes
PGM-Explainer [282]		Yes	GC/NC	N	Yes	Forward	Yes
XGNN [315]	Model-level	Yes	GC	Subgraph	Yes	Forward	Yes

Tabelle 5: XAI-Methoden für Graph-basierte KI-Lösungen (Ding et al., 2022).

Klassifikation für Menschenzentrierte XAI

Basierend auf diesen früheren Ansätzen, aber auch auf unserer Forschung und unseren Erfahrungen im Go-KI Projekt, wurde eine Klassifikation für Menschenzentrierte Erklärbare KI ausgearbeitet. In dieser werden XAI-Methoden basierend auf Eigenschaften in drei Kategorien klassifiziert:

- Die Zielgruppe, für die die Erklärung bestimmt ist, deren Ziele und Bedürfnisse,
- die Domäne, also insb. Die Art der Daten und das Problem das gelöst werden soll,
- und die Art der Erklärung.

Im Folgenden werden diese drei Kategorien näher beschrieben.

Klassifikation: Zielgruppe

Rolle

Aus der relevanten Zielgruppe und ihren Aufgaben im Umgang mit der KI-Anwendung eröffnen sich unterschiedliche Möglichkeiten, aber auch Notwendigkeiten für die Erklärung. Ein Einfühlen in die Zielgruppe ist entscheidend für eine gute Erklärung. Die Zielgruppe einer Erklärung muss nicht gleichbedeutend sein mit der direkten Nutzergruppe der KI-Anwendung.

- **Betroffene**: Die Zielgruppe ist von der Entscheidung der KI-Anwendung betroffen, beispielsweise als Patient oder Antragsteller.
- **Benutzer**: Die Zielgruppe ist in direktem Kontakt mit der KI-Anwendung, beispielsweise ein Pathologe an einer digitalen Workstation, oder ein Bürger im Kontakt mit einem Dialogsystem.
- **Entwickler**: Die Zielgruppe ist selbst Entwickler der KI-Anwendung und könnte diese verändern oder verbessern.
- **Eigentümer**: Die Zielgruppe ist Eigentümer der KI-Anwendung, ist rechtlich oder finanziell für diese zuständig.
- **Validator**: Die Zielgruppe ist Validator der KI-Anwendung, d.h. sie haben genügend Fach- und Hintergrundwissen, um die Entscheidungen der KI zu bewerten, einschließlich hinsichtlich geltender Richtlinien, Gesetze oder ethischer Grundsätze.

Diese Unterscheidung hebt hervor, dass die Zielgruppe bspw. die KI-Anwendung benutzen kann, ohne selbst von deren Entscheidungen betroffen zu sein. Zugleich kann die Zielgruppe aber auch mehrere Rollen vereinen, beispielsweise als Benutzer und Betroffene, oder Eigentümer und Validator.

Fachwissen

Das Wissensniveau der Zielgruppe ist entscheidend für die Komplexität der Erklärung. Beispiel: Wenn eine KI-Anwendung erklärt werden soll, die Werkzeuge an der Form des Griffes erkennt, kann die Erklärung für einen Handwerker komplexer sein als für ein Grundschulkind. Je nach Zielgruppe und Erklärziel kann hierbei sowohl Fachwissen in der Problem-Domäne (bspw. Pathologie) als auch Fachwissen im Bereich Machine Learning relevant sein.

Erklärziel

Nach dem Bestimmen der Rolle und des Fachwissens der Zielgruppe ist es wichtig, die Zielsetzung der Erklärung zu identifizieren. Das heißt, welches Ziel soll die Erklärung für die spezifische Zielgruppe erfüllen? In der Literatur wird zwischen sechs Zielkategorien differenziert:

- **Vertrauen**: Die Erklärung soll das Vertrauen und die Akzeptanz der Zielgruppe in die KI-Anwendung erhöhen.
- **Legitimieren**: Die Erklärung soll die Entscheidung der KI-Anwendung legitimieren, also auf gesetzlicher, rechtlicher und ethischer Ebene absichern.
- **Evaluieren**: Die Erklärung soll dabei helfen, die Qualität der KI-Anwendung zu evaluieren, bspw. systematische Fehler und Biases aufdecken.
- **Verbessern**: Die Erklärung soll dabei helfen die KI-Anwendung zu verbessern, bspw. deren Genauigkeit oder Geschwindigkeit.
- **Lernen**: Die Erklärung soll die Zielgruppe etwas über die KI-Anwendung lernen lassen, zum generellen Verhalten des Algorithmus oder Charakteristiken der Daten.

Das Erklärziel korreliert oft stark mit der Rolle/den Rollen der Zielgruppe, kann aber auch von dieser unabhängig sein, und wie bei den Rollen kann eine Erklärung auch mehrere Erklärziele bedienen.

Klassifikation: Domäne

Datentyp

Neben der Zielgruppe ist auch die Art der Daten, die von der KI-Anwendung genutzt werden, von Bedeutung. Je nach Datentyp können unterschiedliche Erklärbarkeitswerkzeuge von Nutzen sein.

- **Bilder**: Die KI-Anwendung verarbeitet Bilddaten, bspw. Im Kontext von Computer-Vision, Bild- und Gesichtserkennung, OCR, Bildgeneratoren, etc.
- **Text**: Die KI-Anwendung verarbeitet Text, etwa in Form von Benutzereingaben (Chatbots), Sprache, Freitext-Dokumenten, Übersetzungen, etc.
- **Tabellarische Daten**: Die KI-Anwendung verarbeitet tabellarische Daten, d.h. in der Regel numerische oder kategorische Daten, die über einen längeren Zeitraum oder für verschiedene Fälle/Begebenheiten erhoben wurden.

Problemtyp

Ein zentraler Aspekt ist die zugrunde liegende mathematische Problemstellung, die eine KI-Anwendung löst. Oftmals lassen sich diese aber auf drei Hauptprobleme, die man mit maschinellem Lernen lösen kann, zusammenfassen. XAI Methoden sind häufig auf einzelne Probleme spezialisiert und können nicht für jedes der drei Hauptprobleme genutzt werden.

- **Regression**: Wird verwendet, um die Auswirkung einer unabhängigen Variablen auf eine andere Variable zu bestimmen, insbesondere wenn die abhängige Variable ein kontinuierlicher Wert ist.

- **Klassifikation**: Ordnet einen Eintrag bestimmten Kategorien oder Klassen zu. Es geht um die Klassifizierung von Eingabedaten in vorgegebene Klassen.
- **Generierung**: Wird verwendet, um neue Datenmuster zu erzeugen. Durch die Modellierung der Datenverteilung können Datenmuster erzeugt werden, die dem ursprünglichen Datensatz ähnlich sind, sich aber von ihm unterscheiden.

Modelltyp

Zur Lösung einer bestimmten mathematischen Problemstellung können unterschiedliche Typen von Modellen des maschinellen Lernens zum Einsatz kommen. Welches Modell man für die Anwendung benutzt, ist ausschlaggebend für die Auswahl der XAI Methode, da diese oftmals nur für bestimmte Modelltypen entwickelt wurden. XAI Methoden, die unabhängig vom Modell angewendet werden können, nennt man **Modell-agnostisch**.

- **Convolutional Neural Network (CNN)**: Es handelt sich um eine Art künstliches neuronales Netz, das zur Verarbeitung visueller Daten verwendet wird. Es funktioniert, indem es kleine Teile von Bildern analysiert und diese Teile zu einem größeren Bild kombiniert.
- **Multilayer Perceptron**: Es handelt sich um eine Art künstliches neuronales Netz, das zur Verarbeitung von Textdaten verwendet wird. Es arbeitet, indem es den Kontext des Textes versteht und diesen Kontext nutzt, um den Text aussagekräftiger zu machen.
- **Large Language Model**: Es handelt sich um ein fortschrittliches künstliches neuronales Netz, das auf das Verstehen und Erzeugen natürlicher Sprache spezialisiert ist. Es ist in der Lage, komplexe sprachliche Muster zu erkennen, den Kontext zu verstehen und menschenähnlichen Text zu erzeugen.

Klassifikation: Erklärung

Welche Erklärungsmethode zu einem Problem passt, hängt davon ab, welche Fragen im Umgang mit der KI die Erklärung beantworten soll. Wie ein Arzt, der seine Patienten befragt, um ihren Gesundheitszustand zu erfahren, kann man die KI-Anwendung befragen, um über den “Zustand”, also die Güte des KI-Ergebnisses, zu lernen. Das kann zum Beispiel sein: “Wieso habe ich dieses Ergebnis bekommen?”, oder auch “Wann hätte ich ein anderes Ergebnis bekommen?”. In der Literatur werden drei Kategorien von Fragen unterschieden.

Wie funktioniert die KI-Anwendung im Allgemeinen?

Allgemeine Erklärungen von KI-Anwendungen, auch **globale Erklärungen** genannt, geben Einblick, welche allgemeinen Konzepte der Algorithmus lernt, um Vorhersagen zu treffen. Unter Konzepten versteht man Ideen, Muster und Merkmale.

Beispiel

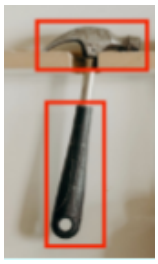
Damit ein Objekt zur Kategorie “Hammer” gehört, muss es einen Griff haben.

Abbildung 3:
https://gt-arc.github.io/go-ki-demo/xai_method_finder/

Wie funktioniert die KI-Anwendung im konkreten Fall?

Erklärungen für spezifische Anwendungsfälle, auch **lokale Erklärungen** genannt, erklären die Vorhersage des KI-Algorithmus in einzelnen Fällen.

Beispiel



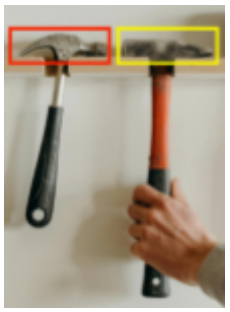
Damit ein Objekt zur Klasse “Latthammer” gehört, muss es einen Griff und einen gebogenen Hammerkopf haben.

Abbildung 4
https://gt-arc.github.io/go-ki-demo/xai_method_finder/

Ab wann ändert sich die Vorhersage?

Eine mögliche Vorgehensweise zum Verstehen der KI-Anwendung ist das Betrachten von Grenzfällen. Das heißt, was man verändern muss, damit die aktuelle Vorhersage eine andere wird. Diese Vorgehensweise wird **kontrafaktische Erklärung** genannt.

Beispiel



Ein Objekt gehört zur Klasse "Latthammer", wenn es einen gebogenen Hammerkopf hat. Wenn der Hammerkopf gerade ist, gehört das Objekt zur Klasse “Schlosserhammer”.

Abbildung 5
https://gt-arc.github.io/go-ki-demo/xai_method_finder/

XAI Steckbriefe

Die im vorigen Abschnitt beschriebenen Kategorien wurden exemplarisch auf verschiedene bekannte XAI-Methoden (siehe Abschnitt Literaturübersicht) angewendet. Daraus entstanden sind die "XAI-Steckbriefe" (siehe https://go-ki.org/ergebnisse), die die wesentlichen Eigenschaften dieser Methoden (nach unserer Klassifikation für Menschenzentrierte XAI) kompakt zusammenfassen. Im Folgenden sind die XAI-Steckbriefe auch in diesem Dokument wiedergegeben.

LRP (Layer-wise Relevance Propagation)

Methode

Layer-Wise-Relevance Propagation (LRP) ist eine Methode, um die Relevanz konkreter Eingabewerte eines neuronalen Netzes zu einer spezifischen Ausgabe zu berechnen. Die Relevanz kann anschließend z.B. für bildbasierte Netze durch eine sog. Heatmap dargestellt werden (siehe unten). Für die Berechnung der Relevanz wird der zu untersuchende Ergebniswert des Netzes durch Anwendung eigens dafür entwickelter Übertragungsregeln "zurück" gespielt, wobei aber die Summe der Relevanz jeder Netzwerkschicht gleich bleibt.

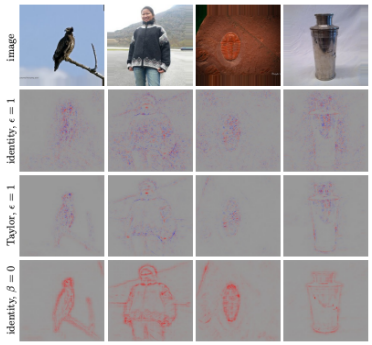


Abbildung 6: Darstellung der relevanten Bereiche als Heatmap (Binder et al., 2016)

Anwendungsbereich

Ursprünglich wurde die Methode für Convolutional Neural Networks (CNNs) und Support Vector Machines (SVMs) entwickelt, inzwischen aber auch auf andere Arten von Modellen wie Recurrent Neural Networks (RNNs), One-Class Support Vector Machines und K-Means-Clustering erweitert. Sie bietet eine sehr schnelle Berechnung bis hin zur Echtzeitfähigkeit.

Einschränkungen

Die Methode benötigt Zugriff auf das neuronale Netzwerk und sollte an die Architektur angepasst werden. Es kann im Vergleich zu anderen Methoden mehr Expertenwissen erfordern, um die verschiedenen LRP-Varianten korrekt anzuwenden. Einstellungen einzelner Parameter können großen Einfluss auf die Qualität der Erklärung haben.

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-spezifisch
- Erklärungsansatz: Post-hoc Erklärung
- Erklärziel: Evaluieren, Lernen

Referenzen

- Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In Artificial Neural Networks and Machine Learning – ICANN 2016: 25th International Conference on Artificial Neural Net- works, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25, pages 63–71. Springer, 2016.

SHAP (SHapley Additive exPlanations)

Methode

Auf Basis der kooperativen Spieltheorie berechnet SHAP für jedes Eingabemerkmal einen Wert, der ausdrückt, wie stark der Einfluss des individuellen Merkmals auf das Vorhersageergebnis ist. SHAP soll dadurch eine relativ einfache Interpretation der Eingabemerkmale für eine breite Zielgruppe ermöglichen. Diese Methode findet aufgrund ihrer theoretischen Fundierung, der vielfältigen Verwendungsmöglichkeiten und hoher Genauigkeit eine breite Anwendung in Wirtschaft und Forschung.

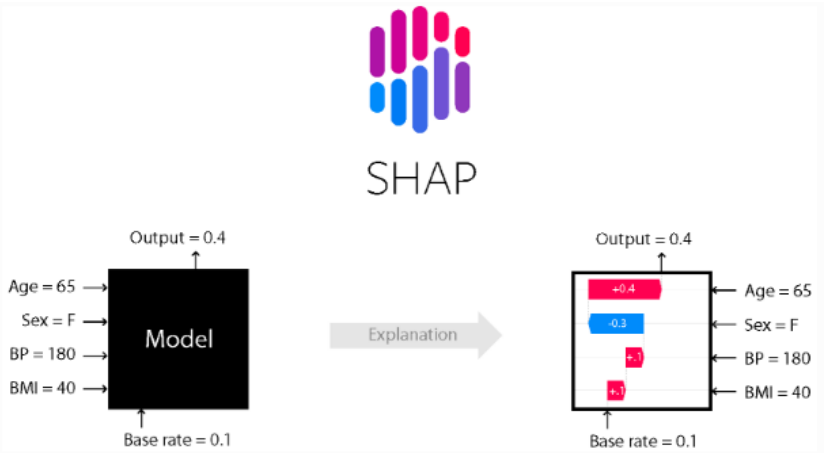


Abbildung 7: Gewichtung der Eingabemerkmale nach SHAP (https://shap.readthedocs.io)

Anwendungsbereich

SHAP wird für eine Vielzahl von Modellen empfohlen, einschließlich lineare Modelle, baumbasierte Modelle (wie Entscheidungsbäume und Random Forests), Support Vector Machines (SVMs), neuronale Netze und Ensemblemodelle.

Einschränkungen

SHAP ist rechenintensiv, und die Interpretation der SHAP Werte kann bei korrelierten Merkmalen oder anderen Wechselwirkungen sehr komplex sein. Ferner gibt es mögliche Einschränkungen bei der Erfassung aller Aspekte des Modellverhaltens, z.B. bei ungleicher Verteilung im Datensatz.

Kategorisierung

- Erklärungsart: Lokal/Global
- Modell: Modell-agnostisch
- Erklärungsansatz: Post-hoc Erklärung
- Erklärziel: Evaluieren, Lernen

Referenzen

- Scott Lundberg. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.

LIME (Local Interpretable Model-Agnostic Explanations)

Methode

LIME generiert lokale Erklärungen. Es vereinfacht komplexes Modellverhalten durch simple und dadurch interpretierbare Regeln. Durch einen Prozess, der als Perturbation Sampling bezeichnet wird, stört LIME Merkmale von Eingabeinstanzen und bewertet die daraus resultierende Auswirkung auf die Ausgabe des Modells. LIME bewertet die Merkmale von Eingabeinstanzen am wichtigsten, deren Störung den größten Einfluss auf das Endergebnis hat.

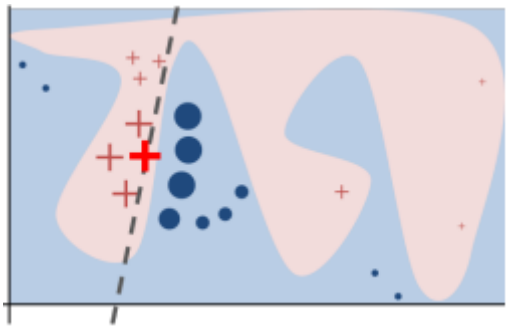


Abbildung 8: Beispiel für Sampling in LIME: Das fette rote Kreuz ist die Prediction, andere Kreuze/Kreise die Samples, und der Hintergrund die Ground-Truth (Ribeiro et al., 2016)

Anwendungsbereich

LIME ist in der Lage, Vorhersagen beliebiger Modellarchitekturen zu erklären, wie z.B. Random Forests, neuronale Netze und mehr. Im Vergleich zu SHAP wird LIME oft für Datensätze mit vielen Merkmalen bevorzugt, da der Sampling-Ansatz von LIME hilft, die Komplexität zu reduzieren.

Einschränkungen

Wegen der benötigten Stichproben ist LIME rechenintensiv. Die Zuverlässigkeit und Vollständigkeit der Erklärungen hängt vom Umfang des Samplings ab. Es können systemische Muster oder Bias unerkannt bleiben, sowie irreführende Erklärungen auftreten. Näherungen erfassen nicht die volle Komplexität bestimmter Modelle oder ihrer Wechselwirkungen.

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-agnostisch
- Erklärungsansatz: Post-hoc Erklärung
- Erklärziel: Vertrauen, Verbessern

Referenzen

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

ProtoSeg (Interpretable Semantic Segmentation with Prototypical Parts)

Methode

ProtoSeg ist ein Modell für interpretierbare Segmentierung, das Prototypen als Erklärmethode verwendet. ProtoSeg lernt Prototypen, d.h. Referenzbeispiele, für jede Klasse und verwendet diese, um Segmentierungen mit Patches (Fällen) aus der Trainingsmenge zu erzeugen und so zu erklären. Es besteht aus einem Backbone-Netz, einem Prototypen Layer und einem fully-connected Layer. Während das Backbone-Netz das Bild segmentiert, verbinden die Prototypen- und die fully-connected Layer die Segmente mit ähnlichen Bildausschnitten .

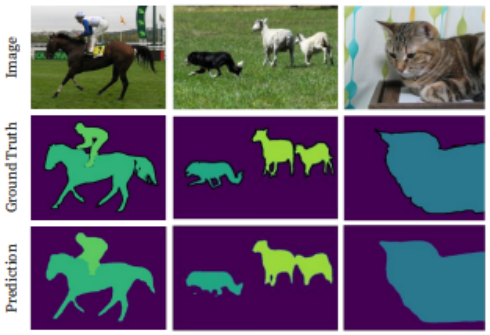


Abbildung 9: Hervorhebung der erkannten Entitäten (Sacha et al., 2023)

Anwendungsbereich

ProtoSeg ist ein interpretierbares Modell, d.h. es erzeugt erklärbare Ergebnisse per Design. Es ist daher sowohl ein Modell als auch eine Erklärbarkeitsmethode. Vereinfacht gesagt handelt es sich bei ProtoSeg um eine XAI-Methode für CNN, da ProtoSeg selbst ein Convolutional Neural Network (CNN) ist.

Einschränkungen

Es handelt sich um eine spezifische Methode, die nur für Segmentierungsaufgaben geeignet und daher weder Modell- noch Aufgaben-agnostisch ist. Die Präzision kann im Vergleich zu anderen Segmentierungsmodellen ohne integrierte Erklärbarkeit beeinträchtigt sein (Accuracy-Interpretability Trade-off).

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-spezifisch
- Erklärungsansatz: Interpretable Model
- Erklärziel: Vertrauen, Lernen

Referenzen

- Mikolaj Sacha, Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protoseg: Interpretable semantic segmentation with prototypical parts. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1481–1492, 2023.

CEM (Contrastive Explanation Method)

Methode

CEM ist eine Methode, die mit kontrastiven Erklärungen arbeitet. Diese erläutern, wieso ein bestimmtes Ergebnis und nicht ein anderes erreicht wurde. Sie unterteilen erklärende Elemente in Pertinent Positive (PP) und Pertinent Negative (PN) Faktoren. Die PP sind eine minimale Menge an Faktoren, die für ein bestimmtes Ergebnis sprechen. Die PN sind die minimale Menge an zusätzlichen Faktoren, die zu einem anderen Ergebnis führen würden. Ein Beispiel für PN wären die Bilder in der rechten Spalte der folgenden Abbildung, wo pinke Bereiche kennzeichnen, was erforderlich wäre, um statt der ursprünglichen Zahl (linke Spalte) eine andere Zahl zu erkennen.

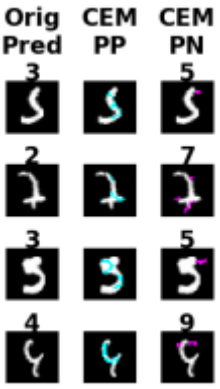


Abbildung 10: PP und PN am Beispiel einer Zeichenerkennung (Dhurandhar et al., 2018)

Anwendungsbereich

CEM ist eine Erklärbarkeitsmethode für Klassifizierungsaufgaben, kann aber auch für Regression oder Clustering verwendet werden. Sie bietet sich daher für verschiedene Methoden wie neuronale Netze, Random Forests, logistische Regression oder Support Vector Machines (SVM) an.

Einschränkungen

Die Interpretierbarkeit kann bei komplexen Modellen eingeschränkt sein. Die Methode hat Schwierigkeiten bei der Verarbeitung kategorischer oder diskreter Merkmale und ist empfindlich gegenüber dem Ausmaß der Störung der Eingabemerkmale.

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-spezifisch
- Erklärungsansatz: Post-hoc Erklärung
- Erklärziel: Vertrauen, Lernen

Referenzen

- Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Advances in neural information processing systems, 31, 2018.

Anchors (High-Precision Model Agnostic Explanations)

Methode

Die Anchor-Methode sucht nach interpretierbaren Bedingungen (Anchors), die die Entscheidungsgrenze um einen bestimmten Datenpunkt herum ausreichend beschreiben. Die generierten Regeln haben die Form "WENN Bedingung DANN Vorhersage" und zielen darauf ab, die Bedingungen zu erfassen, unter denen die Vorhersage des Modells konsistent ist. Die sich daraus

ergebenden Anchors liefern verständliche und von Menschen überprüfbare Erklärungen für die Modellvorhersagen.

Abbildung 11: Originalbild und Anchors für Visual-Question-Answering (Ribeiro et al., 2018)

Anwendungsbereich

Anchors können aufgrund ihrer Modellunabhängigkeit auf eine Vielzahl von linearen und nichtlinearen Modellen angewendet werden. Solange das Modell Vorhersagewahr- scheinlichkeiten oder Konfidenzwerte für jede Klasse liefert, kann es verwendet werden.

Einschränkungen

Die Methode bietet möglicherweise kein umfassendes, globales Verständnis des Verhaltens des Modells und ist empfindlich gegenüber Störungen in den Daten. Anchor Regeln sind von Natur aus vereinfacht und können die Feinheiten des zugrundeliegenden Modells nicht vollständig erfassen.

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-agnostisch
- Erklärungsansatz: Post-hoc Erklärung
- Erklärziel: Vertrauen, Evaluieren, Legitimieren, Lernen

Referenzen

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

t-SNE (t-Stochastic Neighbor Embedding)

Methode

Das Hauptziel von Methoden wie PCA, t-SNE oder UMAP besteht darin, komplexe, hochdimensionale Datensätze in einem Raum mit niedriger Dimensionalität abzubilden. Dies sind klassischerweise zwei oder drei Dimensionen, da sich diese gut graphisch darstellen lassen. In möglichen Darstellungen werden ähnliche Datenpunkte näher beieinander und sich unterscheidende Punkte weiter

voneinander entfernt gruppiert. Dies trägt dazu bei, zugrunde liegende Muster und Cluster in den Daten aufzudecken, die im ursprünglichen, hochdimensionalen Raum möglicherweise nicht sofort ersichtlich gewesen wären.

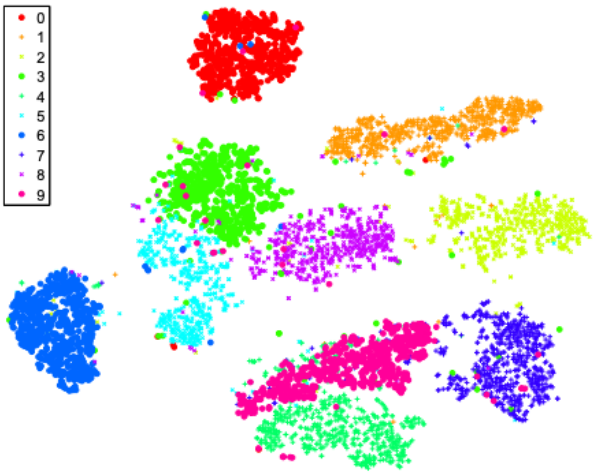


Abbildung 12: Cluster nach Dimensionsreduktion (Van der Maaten and Hinton, 2008)

Anwendungsbereich

Solche Methoden können auf eine Breite von Modellen und Datensätzen angewendet werden, da es sich in erster Linie um eine non-lineare Technik zur Reduktion von Dimensionalität und Datenvisualisierung handelt, die über klassische ML-Modelle hinausgeht. Es kann für Supervised und Unsupervised Learning, NLP oder Feature Engineering verwendet werden.

Einschränkungen

Es handelt sich um keine eigenständige XAI-Methode, die daher nur in Kombination mit anderen XAI-Methoden verwendet werden sollte. Sie liefert nur statische, zwei- oder dreidimensionale Darstellungen der Daten, mit einer starken Abhängigkeit der Ergebnisse von den Eingabeparametern.

Kategorisierung

- Erklärungsart: Lokal
- Modell: Modell-agnostisch
- Erklärungsansatz: Datensatzerklärung
- Erklärziel: Evaluieren, Lernen

Referenzen

- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(11), 2008.

Zusammenfassung

In diesem Dokument stellen wir unsere XAI-Taxonomie und die darauf basierende Kategorisierung bestimmter XAI Methoden vor, die im Rahmen des Go-KI-Projekts entwickelt wurden. Mit unserer XAI-Taxonomie ist es möglich, die große Menge an XAI-Methoden, die es aktuell gibt, strukturiert zu kategorisieren, und hierbei neben technischen insbesondere auch menschliche Faktoren zu berücksichtigen. Dies hilft dabei, für einen bestimmten Use-Case passende XAI-Methoden zu finden. Zudem ist unser Ansatz im Gegensatz zu anderen Ansätzen menschenzentriert, wodurch auch auf die Eigenschaften der Zielgruppe eingegangen wird. Dadurch wird nicht nur eine passende XAI-Methode, sondern auch eine zur Zielgruppe passende Darstellung der Erklärung gewährleistet. Der Ansatz wurde exemplarisch in Form von “XAI-Steckbriefen” auf verschiedene etablierte XAI-Methoden angewendet, kann aber auch dazu genutzt werden, weitere Methoden zu beschreiben.

Der hier vorgestellte Ansatz könnte weiter ausgebaut werden, indem man die in der Taxonomie vorgestellten Kategorien detaillierter unterteilt, um sicherzustellen, dass die Darstellung der Erklärung so gut wie möglich zu der Zielgruppe passt: In der neueren Version des Method Finders wurde zum Beispiel in der Kategorie ‘Fachwissen’ zwischen dem Fachwissen in der Anwendungsdomäne und dem Fachwissen in der KI-Domäne der Zielgruppe unterschieden. Außerdem wurde zwischen theoretischem und praktischem Fachwissen der Zielgruppe unterschieden. Dies hilft dabei zu wissen, wie komplex bestimmte Teile der Erklärung dargestellt werden sollten.

Referenzen

- Nils Ole Breuer and Sahin Albayrak: Towards Automated Human-Centered Recommendation of Explainable AI Solutions. Multimodal, Affective and Interactive eXplainable AI Workshop, 27th European Conference on Artificial Intelligence, 2024. https://ceur-ws.org/Vol-3803/paper6.pdf

- Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38(5):3043–3101, 2024.

- Weiping Ding, Mohamed Abdel-Basset, Hossam Hawash, and Ahmed M Ali. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Information Sciences, 615:238–292, 2022.

- Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. Layer-wise relevance propagation for neural networks with local renormalization layers. In Artificial Neural Networks and Machine Learning – ICANN 2016: 25th International Conference on Artificial Neural Net- works, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25, pages 63–71. Springer, 2016.

- Scott Lundberg. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ”Why should I trust you?” explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144, 2016.

- Mikolaj Sacha, Dawid Rymarczyk, Lukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protoseg: Interpretable semantic segmentation with prototypical parts. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1481–1492, 2023.

- Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. Advances in neural information processing systems, 31, 2018.

- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of machine learning research, 9(11), 2008.