Hochschule Offenburg
offenburg.university

Seminar: Erklärbare KI
Einführung

Dr.-Ing. Xiao Zhao

06.10.2022

EMI Elektrotechnik, Medizintechnik und Informatik

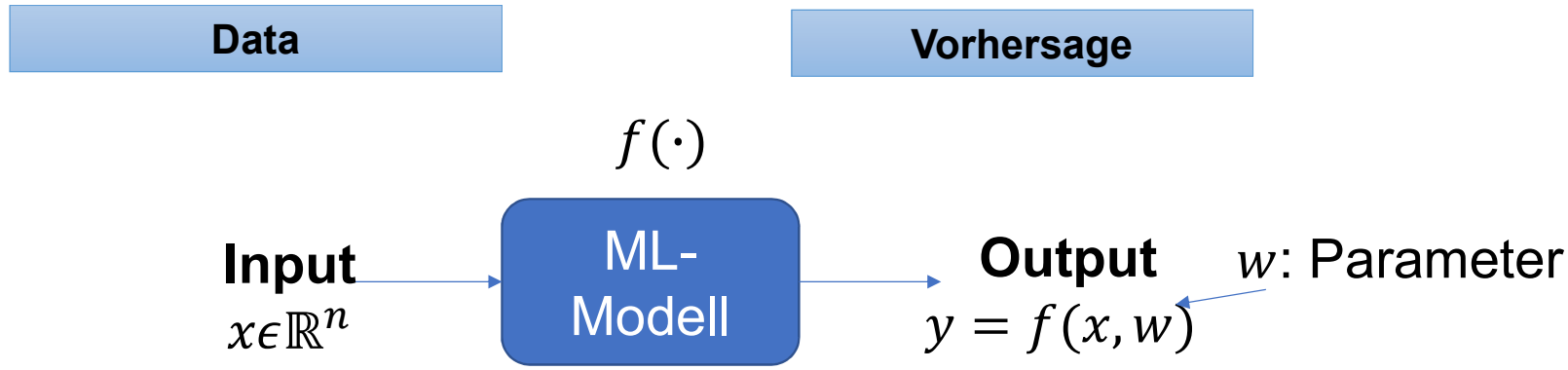
Inhalt

Hochschule Offenburg
offenburg.university

- XAI Definition & Motivation
- XAI Terminologien
- XAI Beispiele

XAI = Explainable AI

Was ist ein ML-Modell?



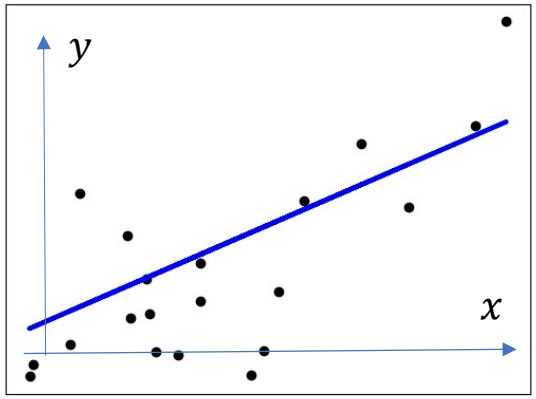
- Viele ML-Modelle können durch eine nicht-lineare mathematische Funktion $f(\cdot)$ dargestellt werden

Beispiele: ML-Modell

Hochschule Offenburg
offenburg.university

- **Linear Regression**:

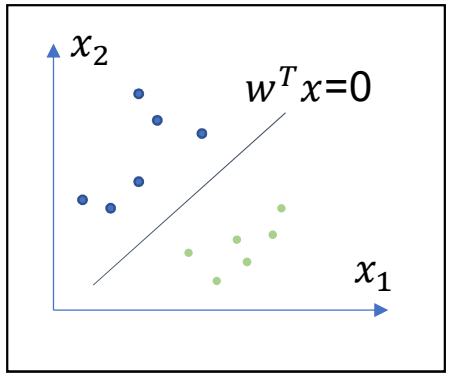
$$y = a_0 + a_1 x_1 + \cdots + a_n x_n$$



Linear Regression

- **Support Vector Machine:**

$$y = \begin{cases} 1, & if\ w^T x \geq 0 \\ 0, & else \end{cases}$$



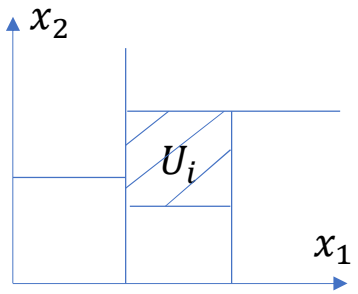
SVM

Beispiele: ML-Modell

Hochschule Offenburg
offenburg.university

- **Entscheidungsbaum:**

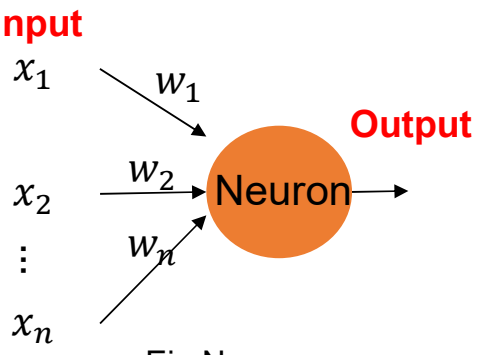
$$y = \sum_i I_i(x)\,,\ wo\ I_i(x) = \begin{cases} 1, \text{if } x \epsilon U_i \\ 0,\ else \end{cases}$$



Entscheidungsbaum

- **Ein Neuron**:

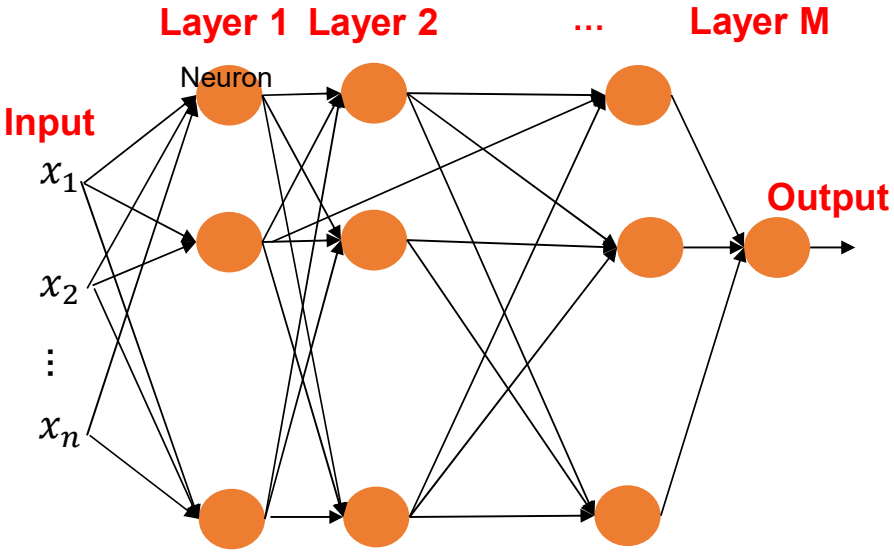
$$y = \sigma(\sum w_i x_i)$$



Ein Neuron

- **Neuronennetz**:

$$y = \sigma(\ldots\sigma(\ldots, \sigma(\sum w_i x_i), \ldots))$$



Neuronennetz

Was ist das Problem dabei?

- ML-Modelle sind **Blackboxes**
  - nur Eingangs- und Ausgangsvariablen sind beobachtbar
  - kein Zugang zu der inneren Parameter und Struktur eines trainierten Modell
- ML-Modelle sind **kompliziert**
  - viele Parameter
  - viele mathematische Operationen
  - Tiefe Struktur, z.B. Neuronennetz
- **Nur Entscheidung, keine Erklärung**

Definition: Erklärbare KI

- *Auf English*: Explainable AI (**XAI**) ≈ interpretable AI

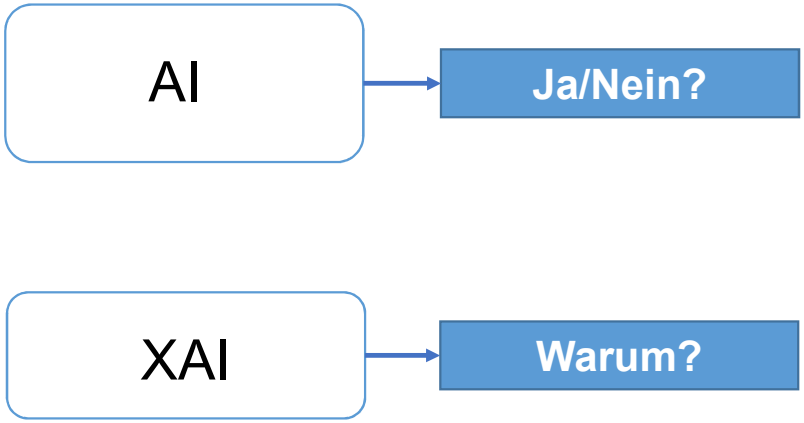
- „XAI ist eine Reihe von Prozessen und Methoden, die es **menschlichen Anwendern** ermöglichen, die von maschinellen Lernalgorithmen erzeugten Ergebnisse und Ausgaben zu **verstehen** und ihnen zu **vertrauen**.“

  — **IBM**

- „XAI umfasst eine Reihe von Tools und Frameworks, mit denen Sie die **Vorhersagen** Ihrer Modelle für maschinelles Lernen **verstehen** und **interpretieren** können.“

  — **Google**

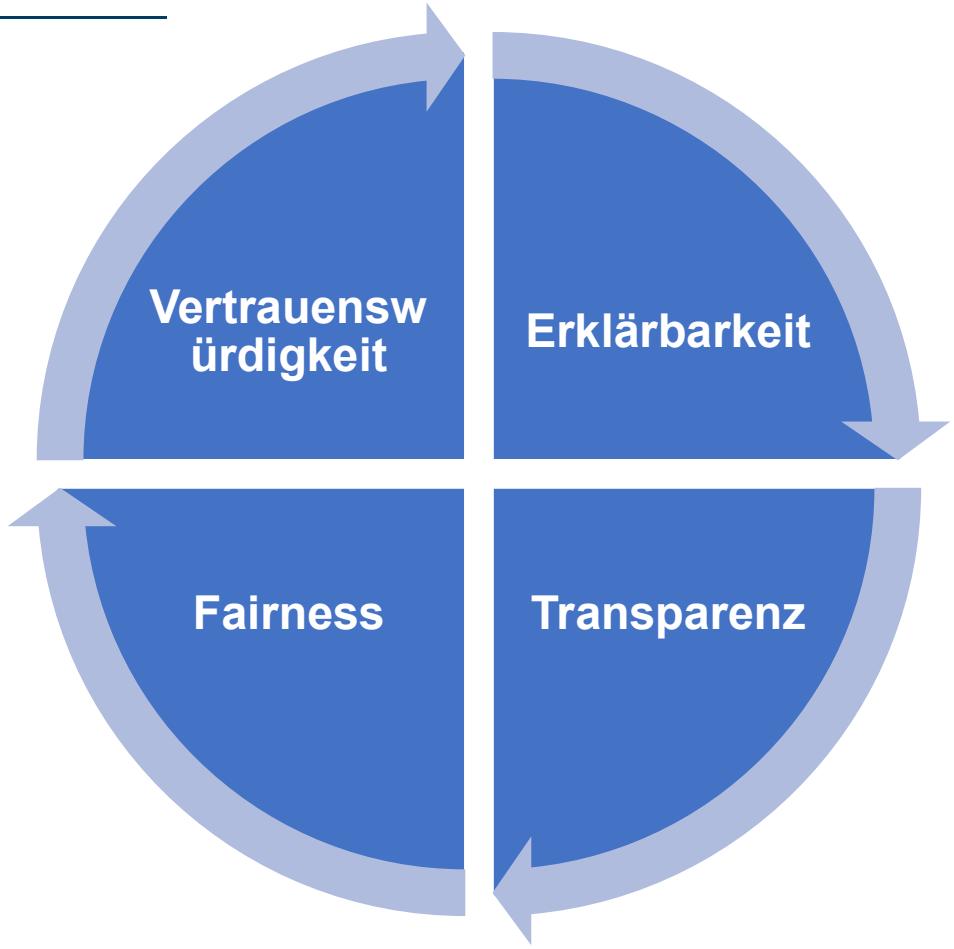
Motivation: Erklärbare KI



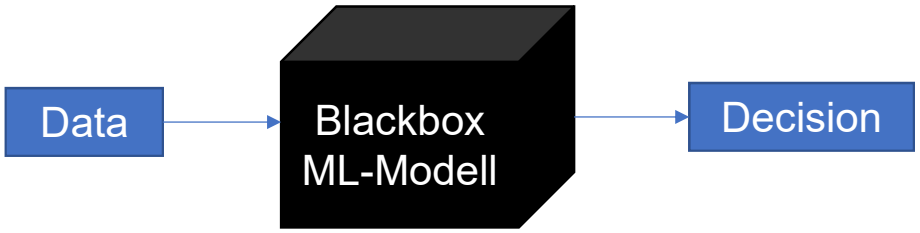
Typische Fragen:

- **Wie** wird eine Entscheidung getroffen?
- Welche **Features** sind wichtig für die Entscheidung?
- Wird die Entscheidung aus **sinnvollen Gründen** getroffen?
- Wird die Entscheidung zu **fairen Bedingungen** getroffen?

Motivation: Erklärbare KI



Interpretierbarkeit von ML-Modellen



Business Manager — Wie kann ich KI-Entscheidungen vertrauen?

Kundensupport — Wie soll ich auf Kundenbeschwerden reagieren?

Data Scientist — Wie bekomme ich bessere Modelle?

Regulierungsbehörde — Sind KI-Entscheidungen fair?

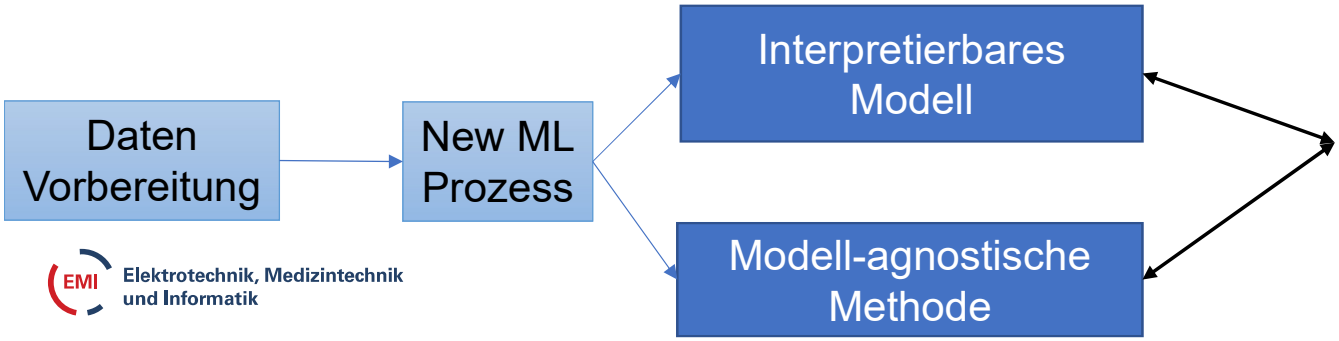
Klassische KI vs. Erklärbare KI

Hochschule Offenburg
offenburg.university

- **Klassische KI**

- User: Warum diese Entscheidung?

- **Erklärbare KI**



- User: Entscheidung verstehen

EMI Elektrotechnik, Medizintechnik und Informatik

Wann wird XAI benötigt?

- **Kritische Systeme**
  - Zugplanungssystem
  - Kraftwerk
  - Militärisches System

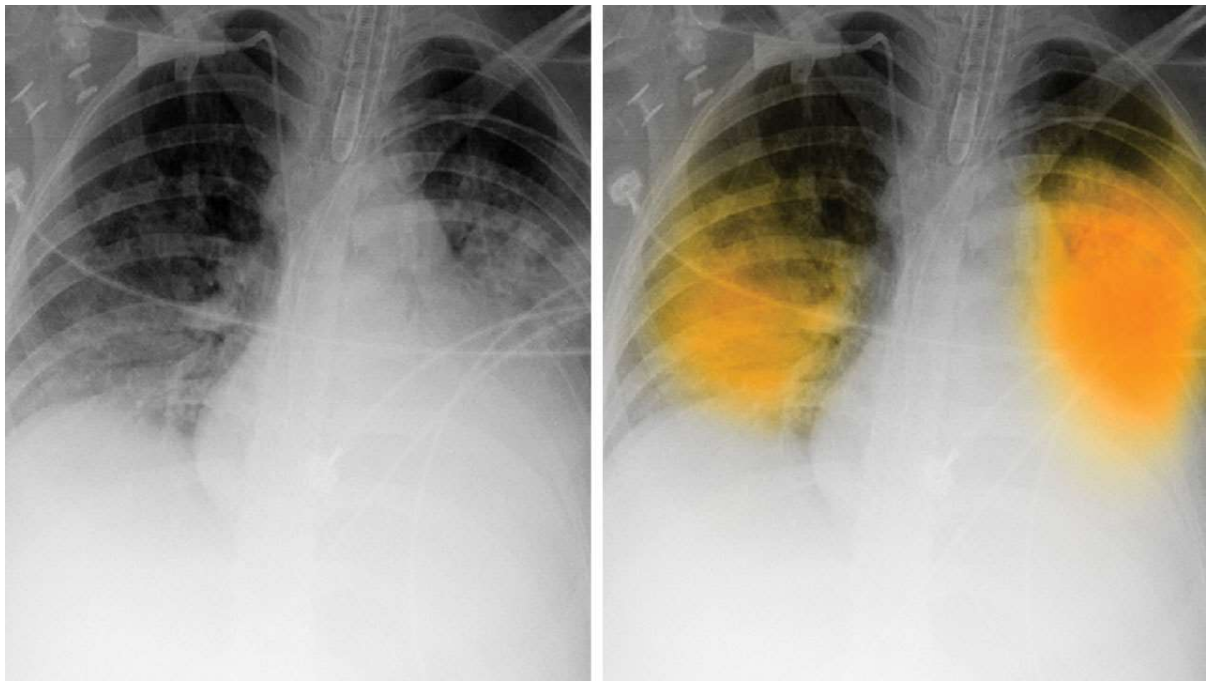
Wichtige Entscheidung ⟹ XAI

[Quelle: en.wikipedia.org]

Wann wird XAI benötigt?

Hochschule Offenburg
offenburg.university

- Medizinische Diagnose:
  - Image-basierte Diagnose
  - Symptom-basierte Diagnose
  - Daten-basierte Diagnose



[Quelle: www.science.org]

Inhalt

- XAI Definition & Motivation
- XAI Terminologien
- XAI Beispiele

Definition: Interpretierbare ML-Modelle

- Interpretierbare Modelle sind **einfache ML-Modelle**, deren Parameter oder Struktur zur Interpretation des Modells und der Nachvollziehbarkeit der Modellvorhersagen direkt verwendet werden können.
- **Beispiele**:
  - Lineare Regression:
    - Großer Koeffizient ⇒ wichtiges Feature
  - Entscheidungsbaum:
    - $x_1 > 10, x_2 < 5 \Rightarrow$ y=1
  - Decision Rule:
    - *Wenn* „Lage=gut“, *dann* „Hauspreis > 1 Mio.“

XAI: Intrinsische Methode & Post-hoc Methode

Hochschule Offenburg
offenburg.university

- **Intrinsische Methode**: Methode für *interpretierbare Modelle*, die wenige Parameter beinhalten oder deren Strukturen einfach sind.
  - Nur für einfaches ML-Modell
  - Unterschiedliche interpretierbare Modelle haben unterschiedliche intrinsische Methoden

- **Post-hoc Methode:** Methode, die das Modell nach dem Training (post hoc) analysiert.
  - Geeignet für komplizierte ML-Modelle
  - Gleiche Methode für unterschiedliche Modelltypen

Modell-spezifische vs. Model-agnostische Methoden

Hochschule Offenburg
offenburg.university

- **Modell-spezifische Methoden**:
  - Nur für bestimmte Modelle

- **Modell-agnostische Methoden**:
  - Für unterschiedliche Modelle

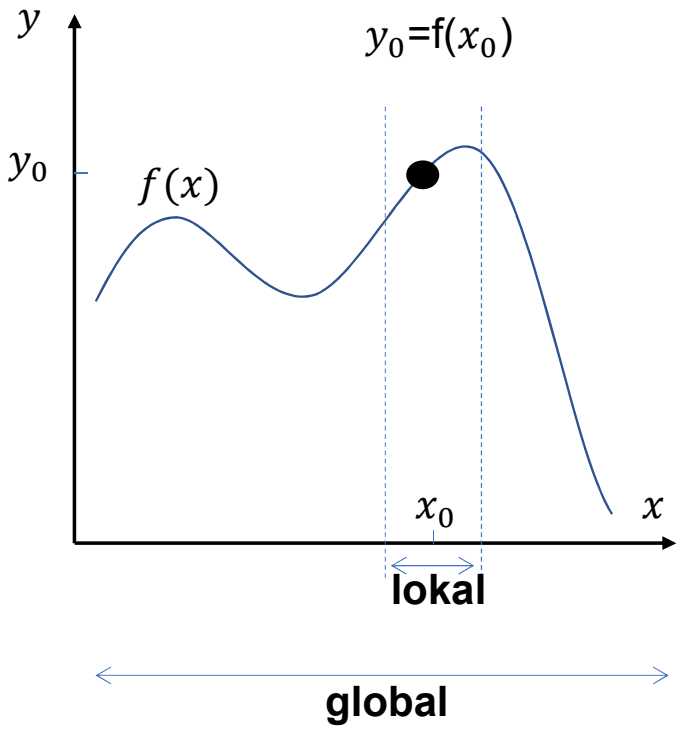
XAI: Lokale Methode vs. globale Methode

Hochschule Offenburg
offenburg.university

Lokal & Global in Mathematik:

- Betrachten $y = f(x)$, $x \epsilon R^n$, $y \epsilon R^m$

Lokal	Global
in einer kleinen **Nachbarschaft U** von Punkt x_0	für **alle** $x \epsilon R^n$
Frage: Wie sieht y in U aus?	Frage: Wie sieht y in R^n aus?



XAI: Lokale Methode vs. globale Methode

Hochschule Offenburg
offenburg.university

Lokale Methode	Globale Methode
in einer kleinen **Nachbarschaft** $\boldsymbol{U}$ von Bespiel $\boldsymbol{x_0}$	für **alle** Beispiele $x \epsilon R^n$
Frage: Wie sieht die Inferenz y **in** $\boldsymbol{U}$ aus?	Frage: Wie sieht die Inferenz y **in** $\boldsymbol{R^n}$ aus?

XAI: Lokale Methode vs. globale Methode

Hochschule Offenburg
offenburg.university

- **Lokale Methode**:
  - interessiert sich für das Verhalten des Modells, wenn die Eingabedaten nicht sehr von einem Datenpunkt x0 abweichen.

- **Globale Methode:**
  - daran interessiert sind, die Eigenschaft des Modells für alle möglichen Beispiele zu verstehen

Beispiele von lokalen Methoden

- Local Interpretable Model-agnostic Explainations (LIME)
- Anchor Methode
- Layer-wise relevance propagation (LRP)

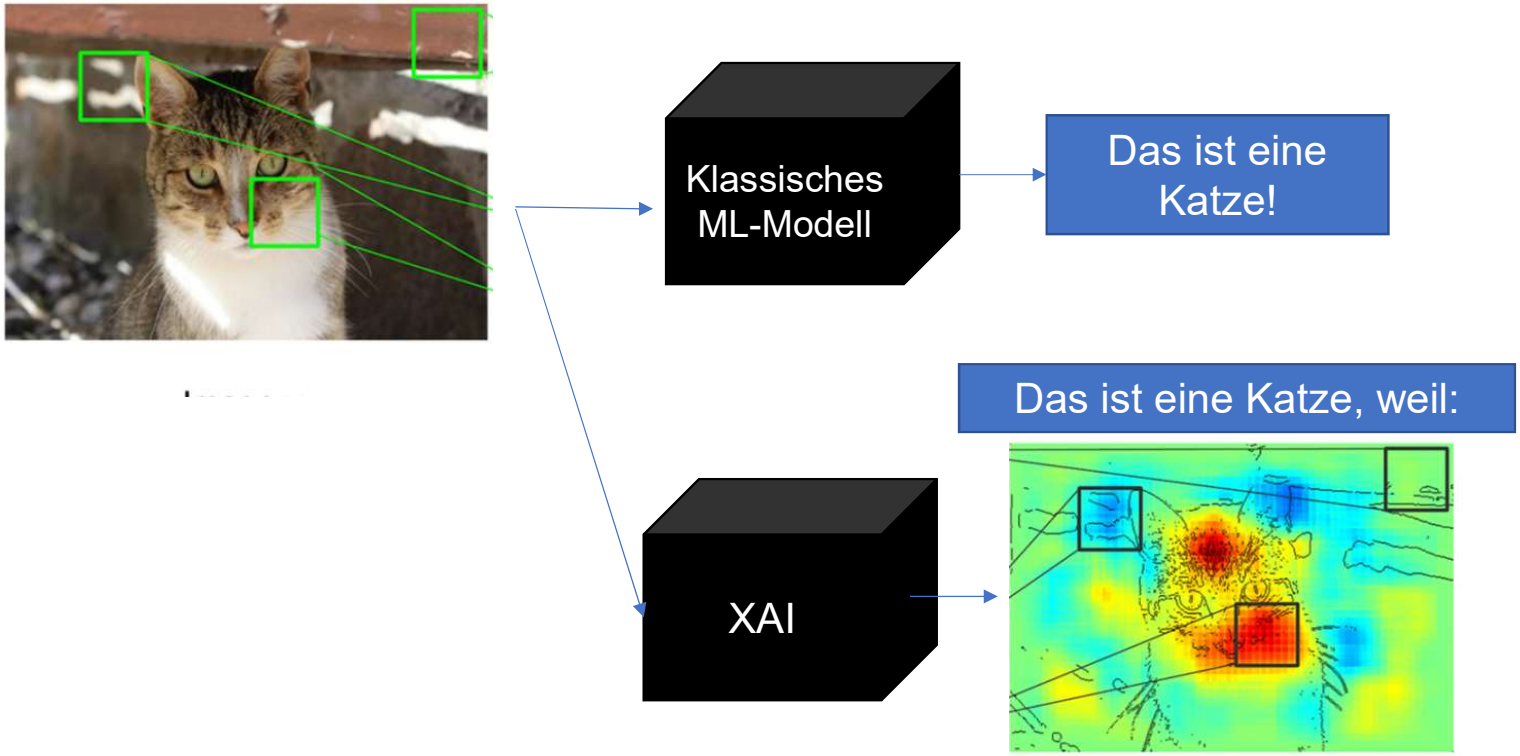
Beispiele von globalen Methoden

- Permutation Feature Importance
- Partial Dependence Plot (PDP)
- Accumulated Local Effect (ALE)
- Adversarial Examples

Inhalt

- XAI Definition & Motivation
- XAI Terminologien
- XAI Beispiele

XAI Beispiel: LRP Methode



Bach, et al., "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLoS ONE, 10, 7, 2015

XAI Beispiel: Anchor Methode

(a) Original image

(b) Anchor for "beagle"

Ribeiro, Singh, and Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2018.

XAI Beispiel: Medizinische Diagnose

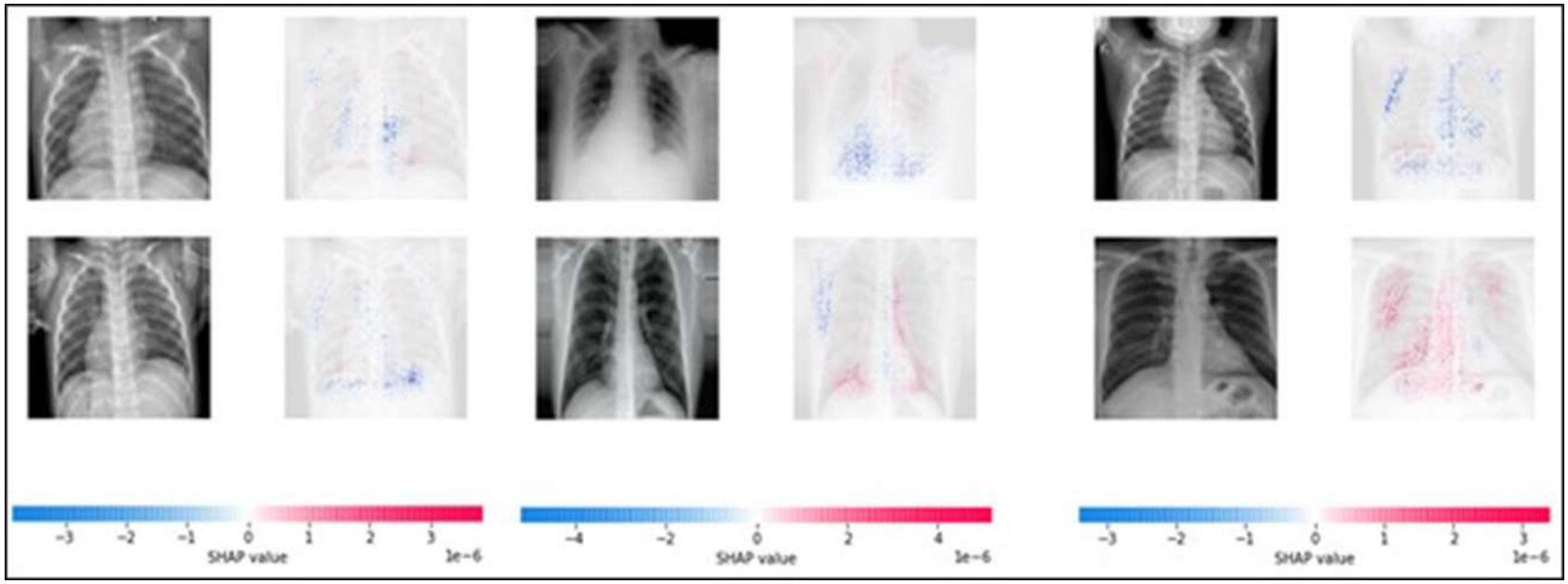
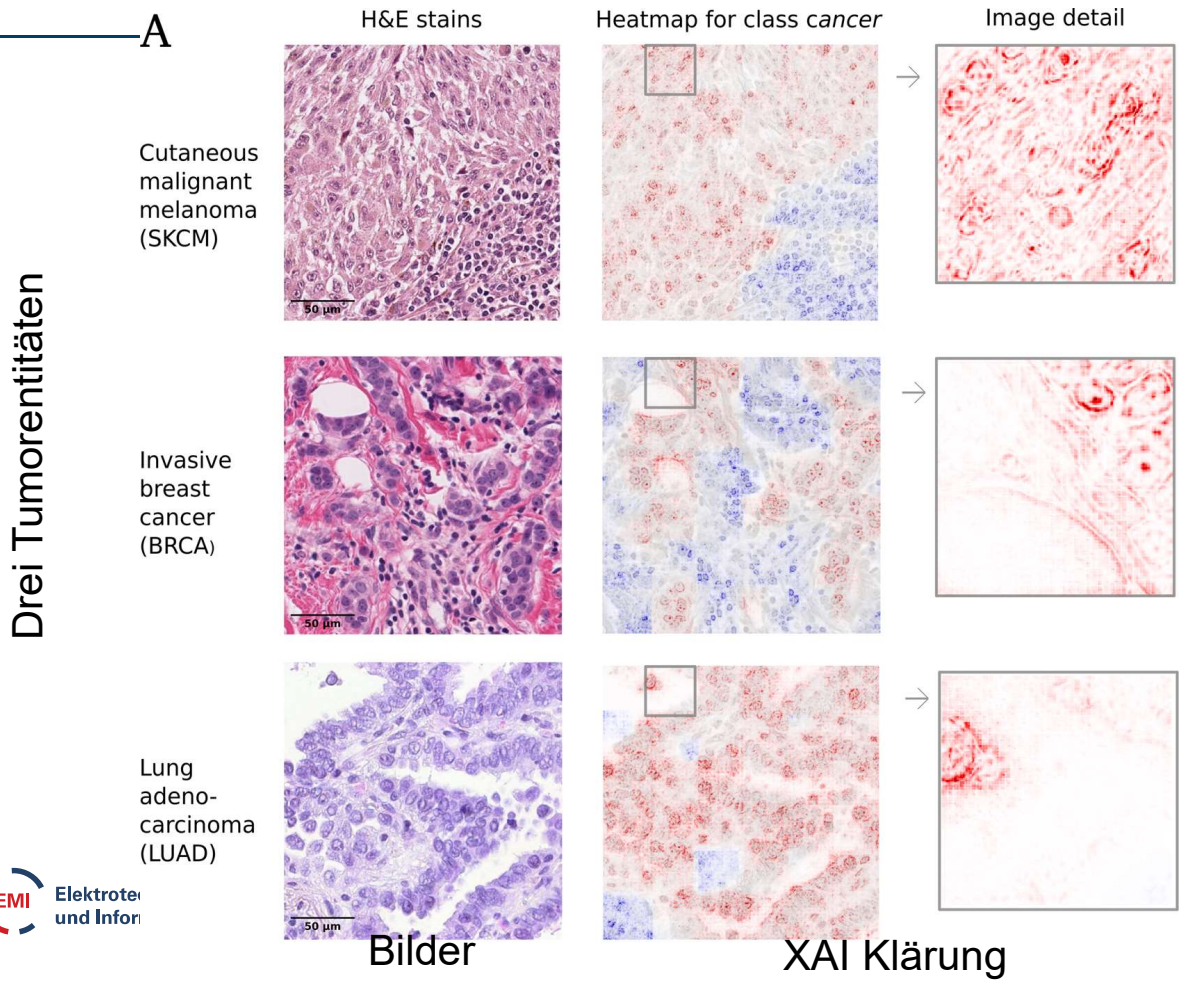
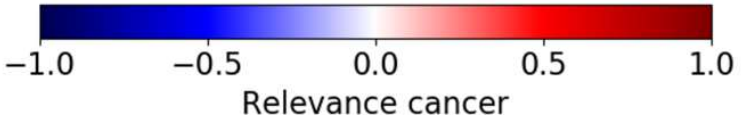


Fig. : Shapley values acquired for classification of several example images. Note that this technique can identify both positively and negatively influential pixels.

Fuhrman, et al., “A review of explainable and interpretable AI with applications in COVID-19 imaging," Medical Physics, 49, 1, 2022.

XAI Beispiel: Medizinische Diagnose

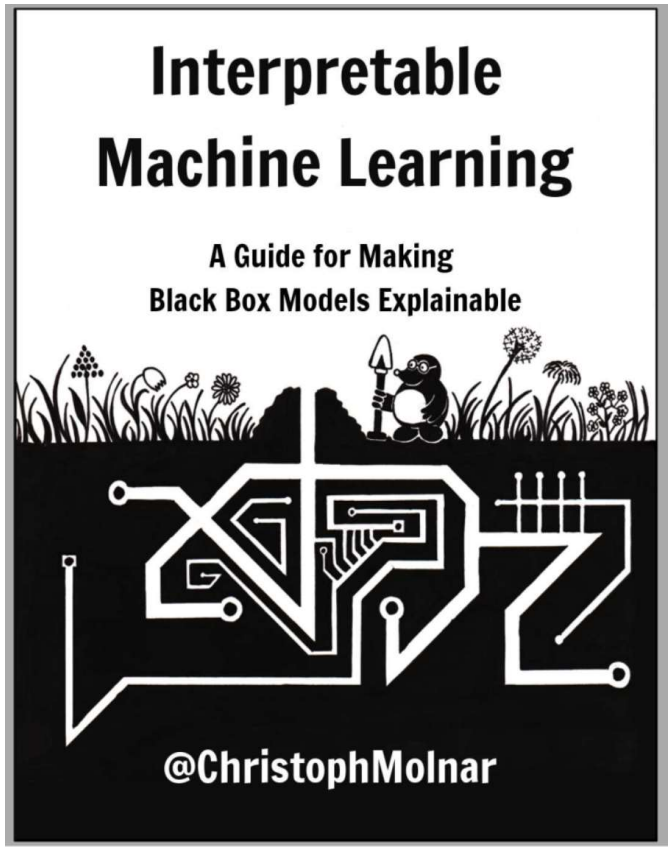




Hägele, *et al.,* "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Sci Rep* **10**, 6423, 2020.

Literatur Empfehlungen

- https://christophm.github.io/interpretable-ml-book



10 Themen in Moodle

Hochschule Offenburg
offenburg.university

- Generalized Linear Model (GLM), Generalized Additive models (GAM) and their application (Zhao)
- Sequential covering and its application to a case study (Zhao)
- Layer-wise relevance propagation (LRP) and its application to a case study (Zhao)
- Permutation feature importance and its application to a case study (Zhao)
- Partial Dependence Plot (PDP) and its application to a case study (Zhao)
- Accumulated Local Effect (ALE) plot and its application to a case study (Zhao)
- Apply linear regression-based surrogate model in Local Interpretable Model-agnostic Explanations (LIME) (Zhao)
- SHAP (SHapley Additive exPlanations) and its application to a case study (Zhao)
- Counterfactual explanations and its application to a case study (Zhao)
- Adversarial examples and its application to a case study (Zhao)

Kontakt

Hochschule Offenburg
offenburg.university

- Email: xiao.zhao@hs-offenburg.de
- Tel: 0781 205 1167
- Büro: STB 0.16 (EG, IMLA)
- Arbeitszeit bei HSO: Donnerstag + Freitag
- Sprechstunde:
  - Donnerstag: 13:30-17:30
  - Termin