Erklärbare Künstliche Intelligenz

Impuls zur Digitalkonferenz 2024

Nils Gumpfer, M. Sc. [1,2]

[1]Kompetenzzentrum für Informationstechnologie, Technische Hochschule Mittelhessen, Friedberg
[2]Promotionszentrum für Ingenieurswissenschaften, Forschungscampus Mittelhessen, Gießen

27. November 2024

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

KITE
Kompetenzzentrum für
Informationstechnologie

Können Sie diesen Satz beenden?

Frankfurt am Main ist eine Stadt in ...

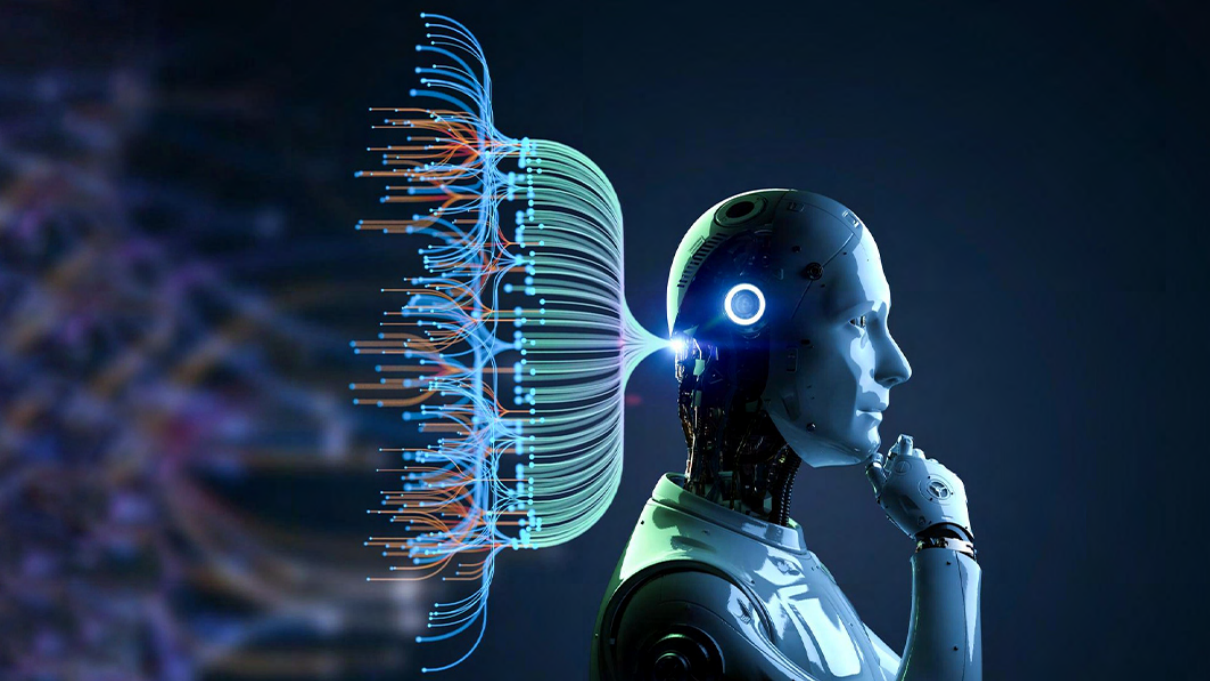
Frankfurt am Main ist eine Stadt in **Hessen**

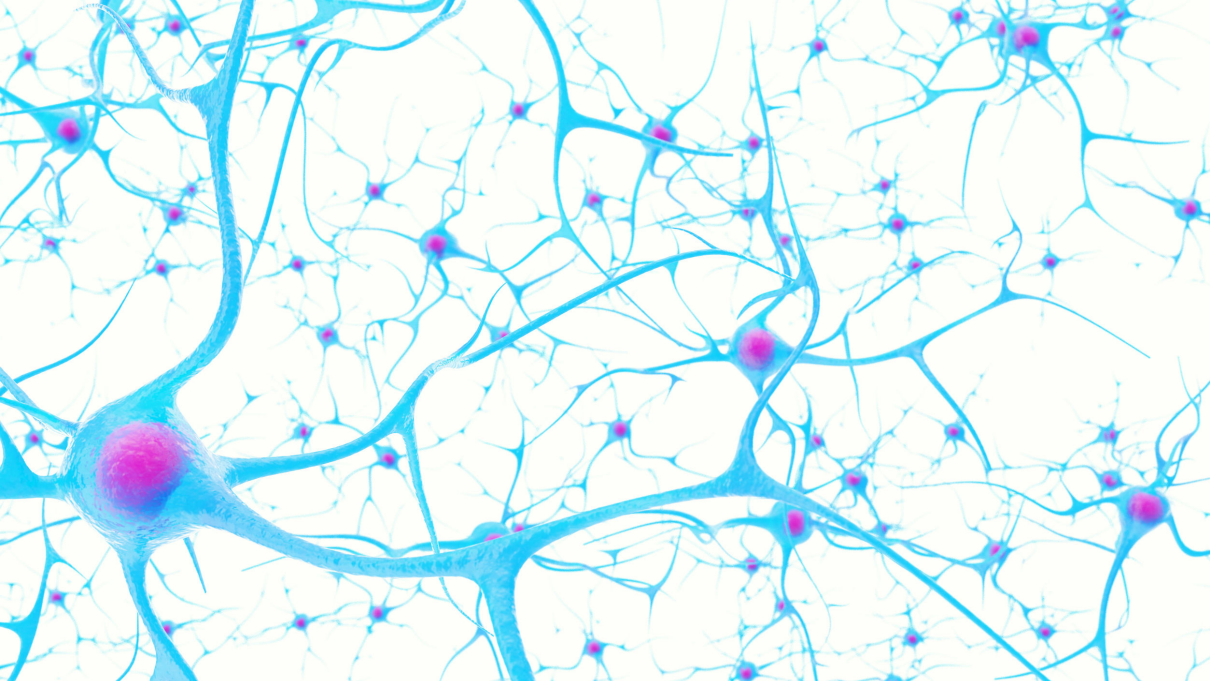
Frankfurt am Main ist eine Stadt in **Deutschland**

Frankfurt am Main ist eine Stadt in **Europa**

Warum können wir das?

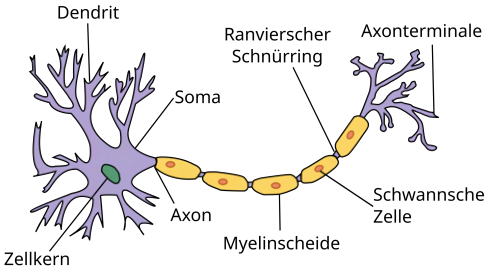

Google



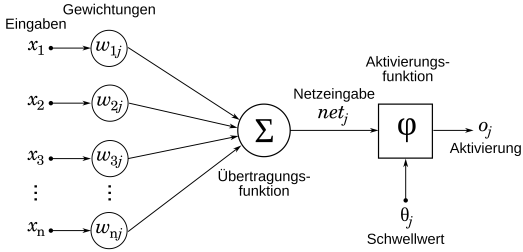


Künstliche Neuronen

Biologisches Neuron

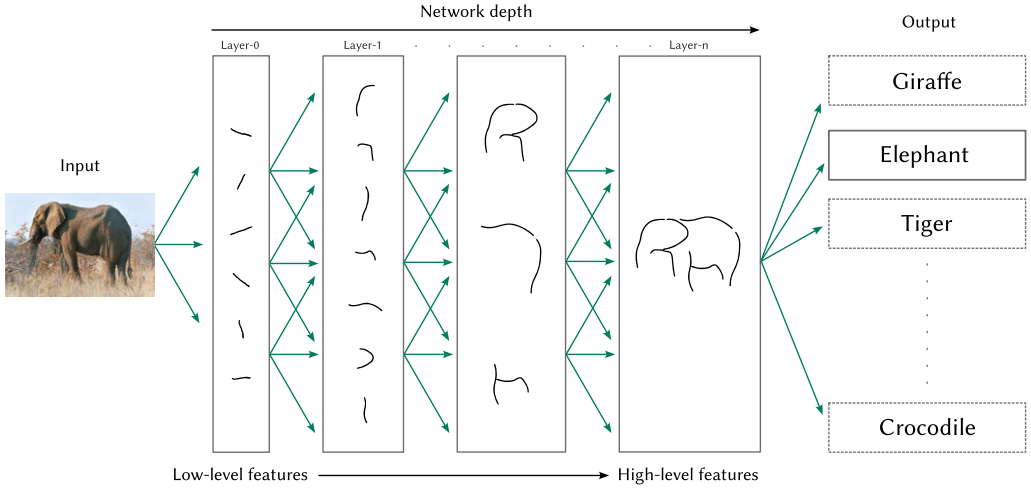


Künstliches Neuron

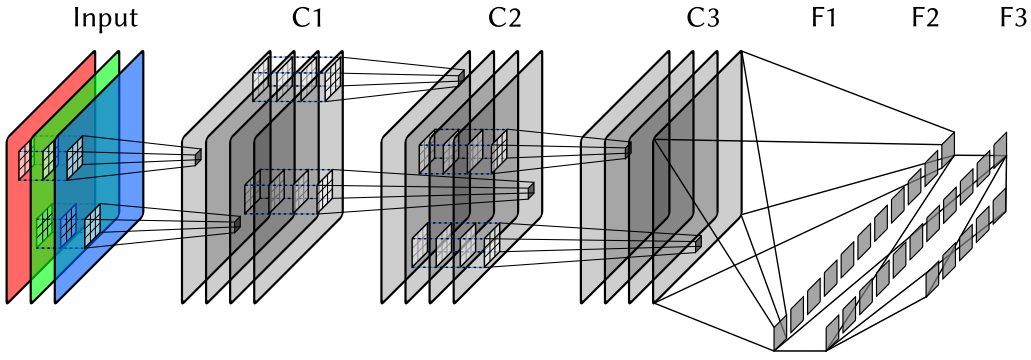


Copyrights: Felix Andrews, CC BY-SA 3.0, via Wikimedia Commons

Künstliche Neuronale Netze

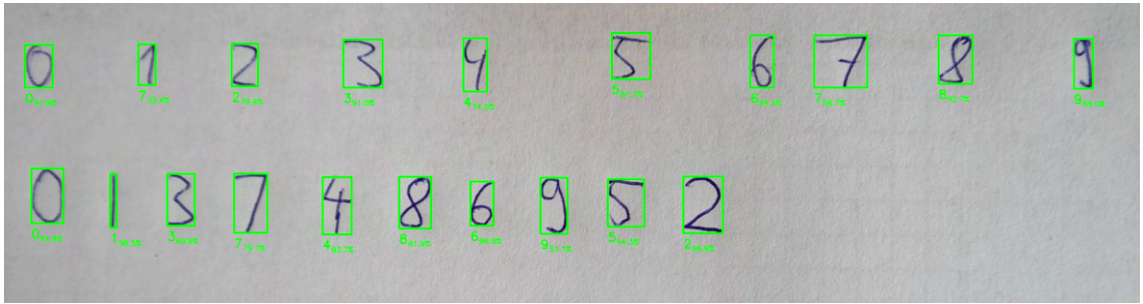


Künstliche Neuronale Netze



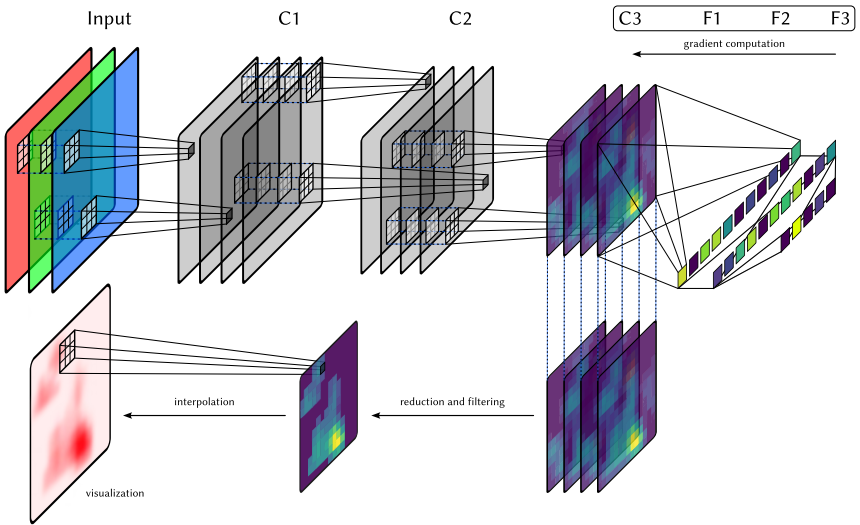
Convolutional Neural Network (CNN)

Künstliche Neuronale Netze



Anwendungsbeispiel für CNNs: Schrifterkennung

Erklärbarkeit



Erklärbarkeit

Erklärbarkeit

Grad-CAM Erklärung für "Fahrrad"

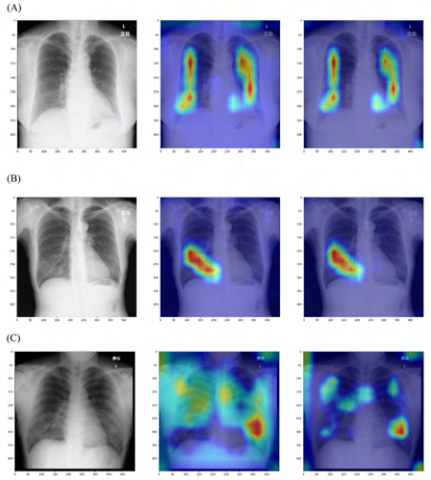
Grad-CAM: Selvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, ICCV (2017) 618–626

Erklärbarkeit

Grad-CAM Erklärung für "Burg"

Grad-CAM: Selvaraju et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, ICCV (2017) 618–626

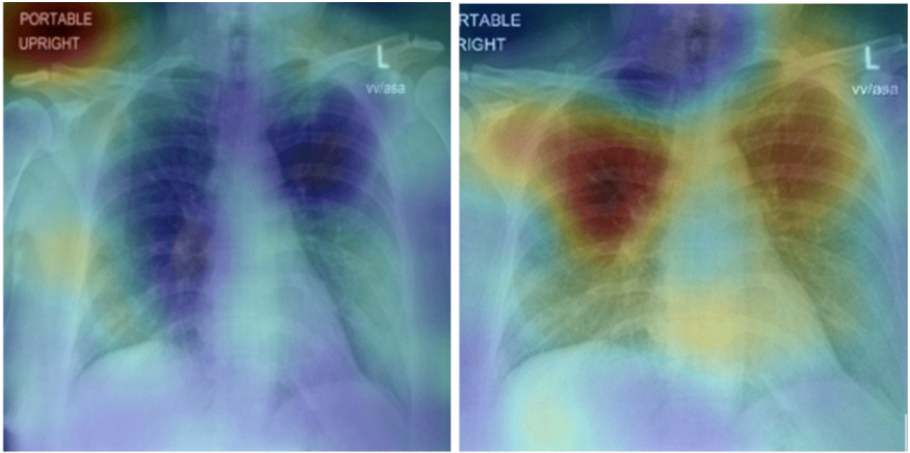
Richtig aus den richtigen Gründen



Röntgen-Aufnahmen mit Grad-CAM Erklärungen

Miyazaki et al., Computer-aided diagnosis of chest X-ray for COVID-19 diagnosis [...], Sci Rep 13, 17533 (2023)

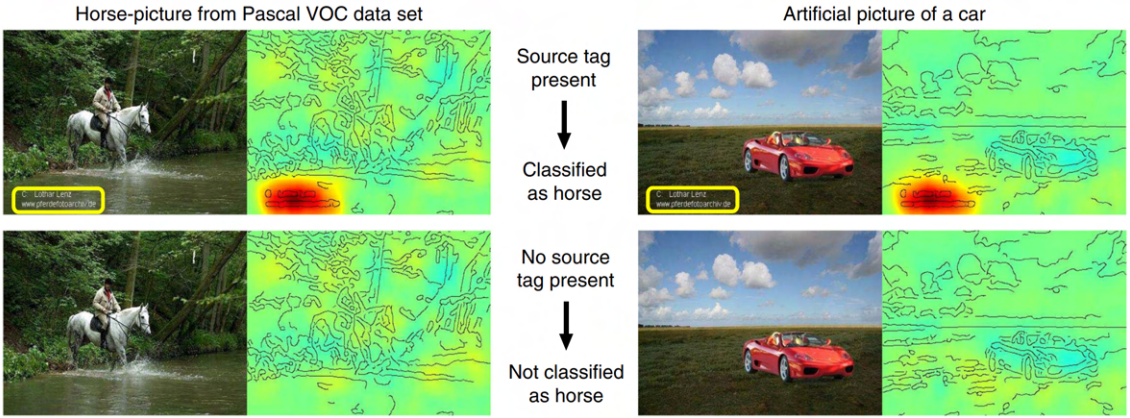
Richtig aus den richtigen Gründen?



Weitere Röntgen-Aufnahmen mit Grad-CAM Erklärungen

Mumuni et al., Improving deep learning with prior knowledge and cognitive models [...], Cogn Sys Res 84, 101188 (2024)

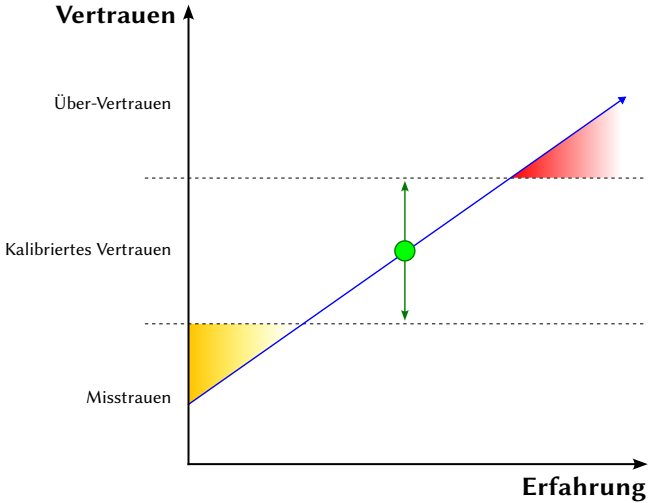
Richtig aus den richtigen Gründen?



Beispiel für den "Clever Hans Effect"

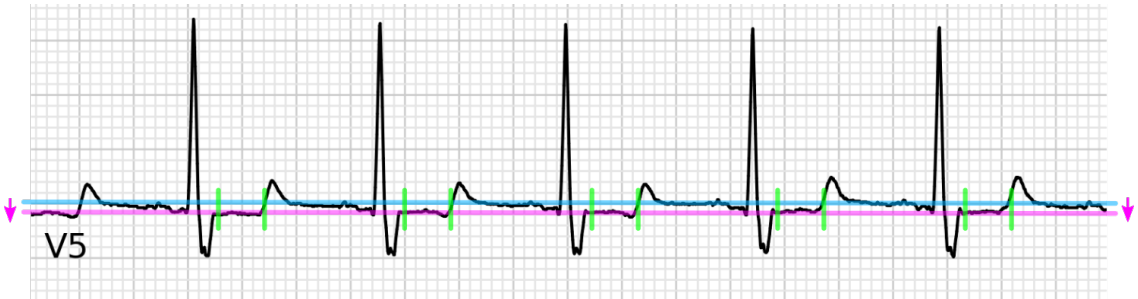
Lapuschkin, S., Wäldchen, S., Binder, A. et al. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 10, 1096 (2019)

Kalibrierung von Vertrauen



nach B. M. Muir (1994)

Anwendungsbeispiel aus der Kardiologie



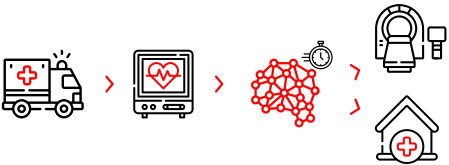
LRP-SIGN Erklärung für Myokardischämie

LRP-SIGN: Gumpfer et al., SIGNed explanations: Unveiling relevant features by reducing bias, Information Fusion 99 (2023)

Forschungstransfer

RISKO

HERMIQS



HESSEN Hessisches Ministerium für Digitalisierung und Innovation

digitales.hessen DISTR@L

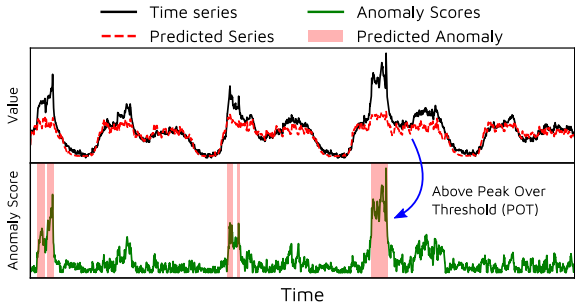
CRS medical

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

THM TECHNISCHE HOCHSCHULE MITTELHESSEN

Forschungstransfer

Forschungsgruppe TimeXAI



Quelle: Tuli et al., TranAD: deep transformer networks for anomaly detection in multivariate time series data. Proc. VLDB Endow. 15, 6, 2022), 1201–1214.

Bundesministerium für Bildung und Forschung

THM TECHNISCHE HOCHSCHULE MITTELHESSEN

KITE Kompetenzzentrum für Informationstechnologie

Résumé

Positiv

Richtig eingesetzt, kann KI viele Aufgaben erleichtern

Vorsicht

Verhalten von KI kann "intelligent" wirken

Empfehlung

Chancen nutzen, aber Grenzen von KI kennen(-lernen)

Vielen Dank.

Acknowledgement:

Prof. Dr. Michael Guckert, Prof. Dr. Bernhard Seeger, Dr. Jennifer Hannig

Funding:

German Ministry for Education and Research via de.NBI Cloud*
(031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A)

Hessian Ministry of Digital Strategy and Development via Distr@l*
(2100664A, 2100932A)

Hessian Center for Artificial Intelligence (hessian.ai)*

* Sponsors had no influence on study designs, statistical analyses, or drafts of publications.

Bundesministerium für Bildung und Forschung

HESSEN Hessisches Ministerium für Digitalisierung und Innovation

digitales.hessen DISTR@L

hessian.AI

THM TECHNISCHE HOCHSCHULE MITTELHESSEN

KITE Kompetenzzentrum für Informationstechnologie