



# Evaluating the Fidelity of Explanations for Convolutional Neural Networks in Alzheimer's Disease Detection

---

Bjarne C. Hiller

2025-03-09

University of Rostock

# Motivation

# Data and Preprocessing

T1-weighted MRIs

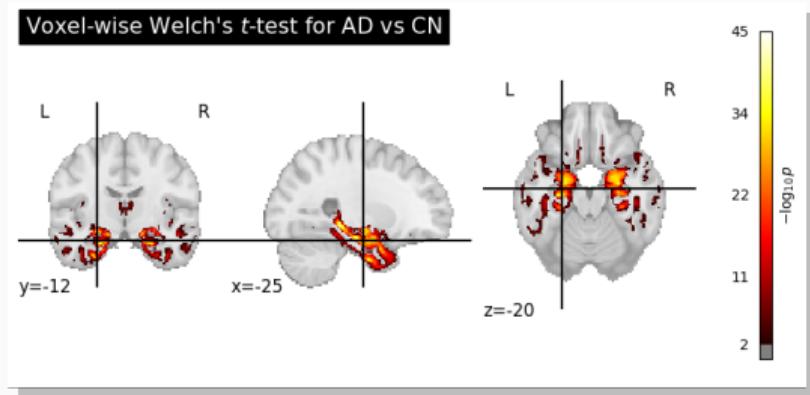
$$AD = 189V = 254$$

skull stripping

grey matter segmentation

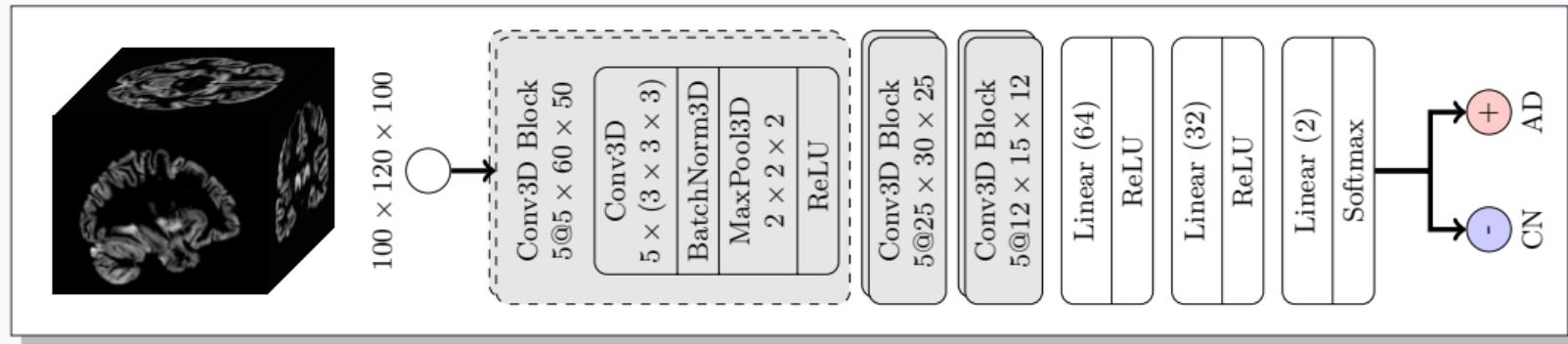
non-linear registration

Voxel-wise Welch's  $t$ -test for AD vs CN

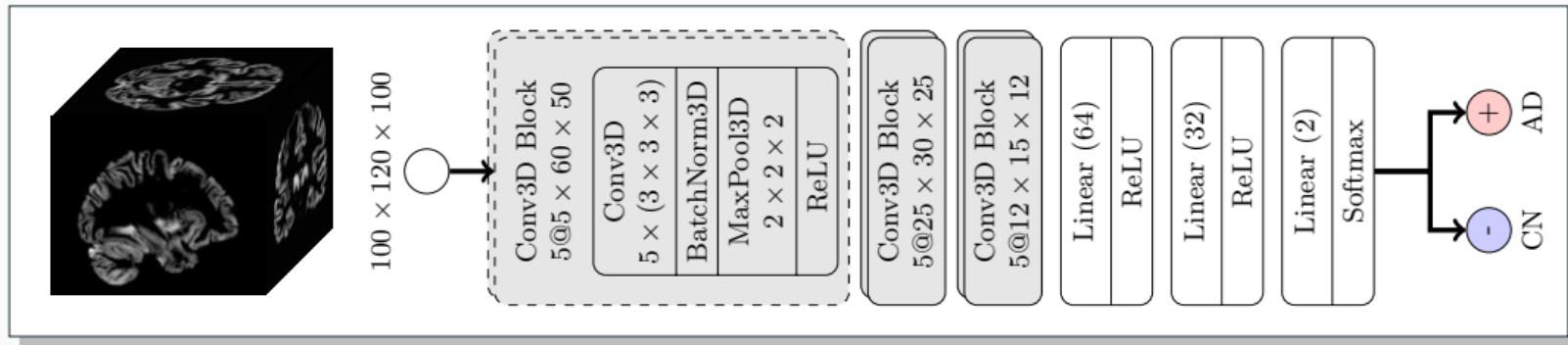


Voxel-wise Welch's  $t$ -test ( $\alpha = 0.01$ )

# A CNN Model for AD vs CN Classification



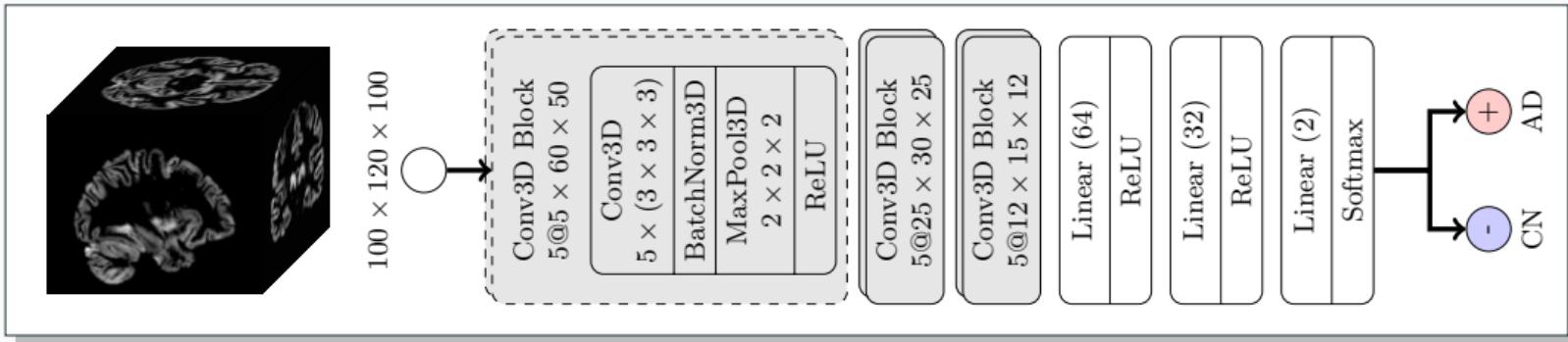
# A CNN Model for AD vs CN Classification



AUC ROC	Accuracy
$0.95 \pm 0.02$	87.64%

5-fold Cross Validation Results

# A CNN Model for AD vs CN Classification

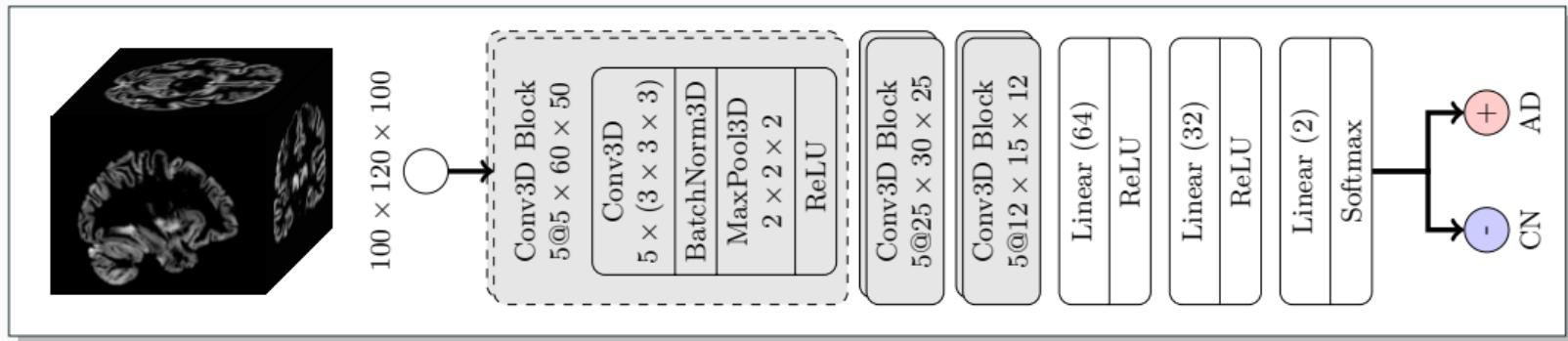


AUC ROC	Accuracy
$0.95 \pm 0.02$	87.64%

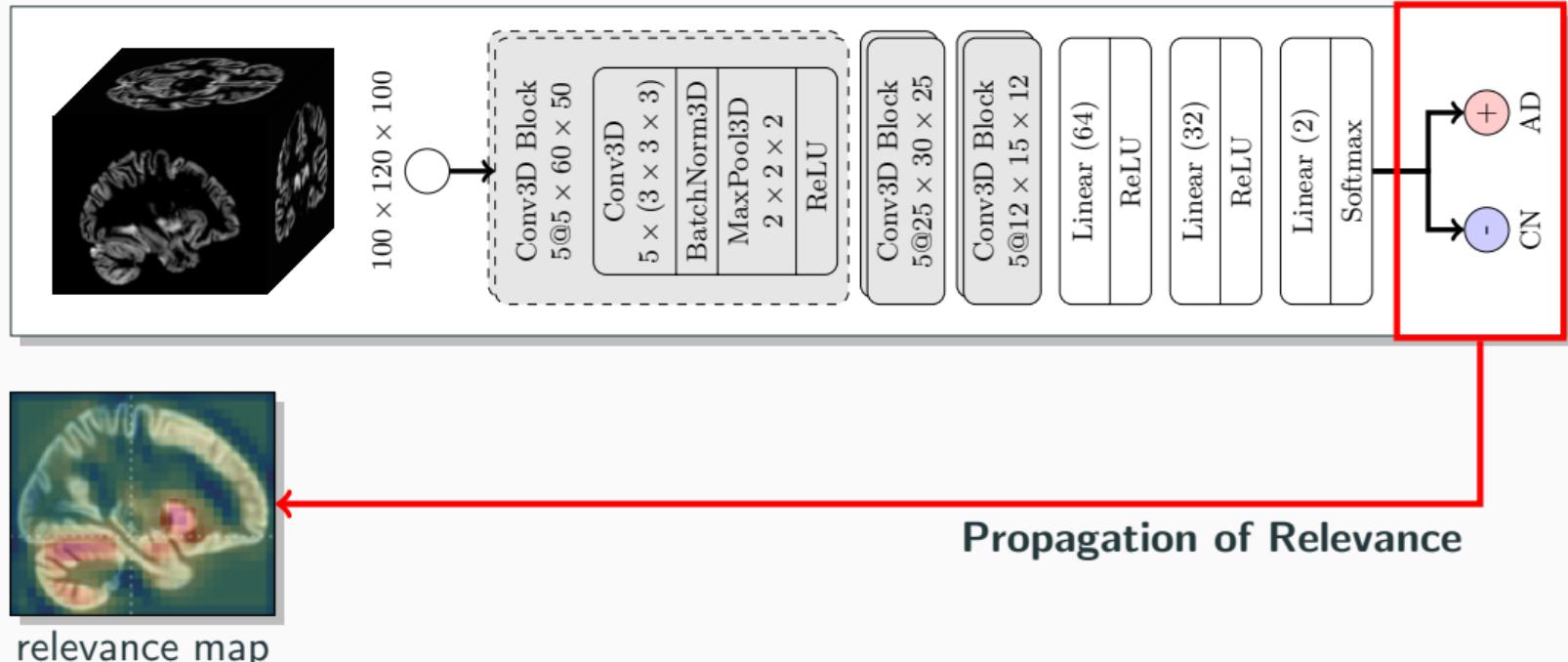
Acceptable Performance...  
...but can we **trust** the model?

5-fold Cross Validation Results

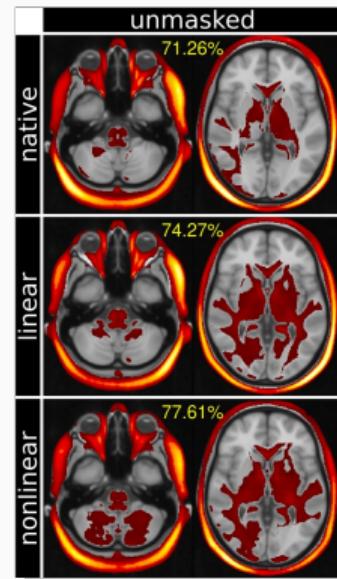
# Attribution Maps: What did the network look at?



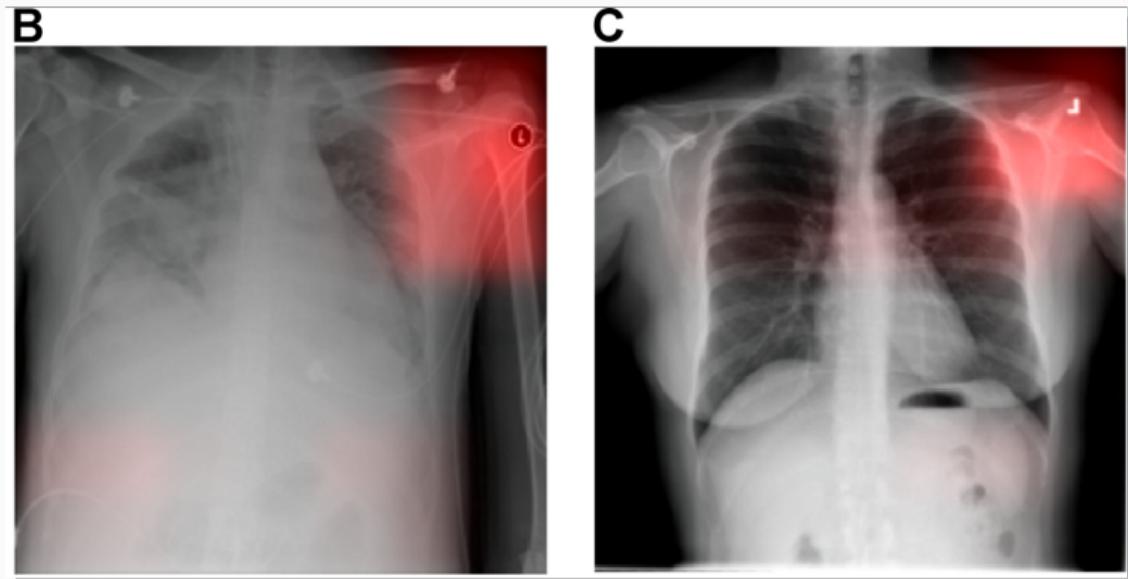
# Attribution Maps: What did the network look at?



# Why should I care?



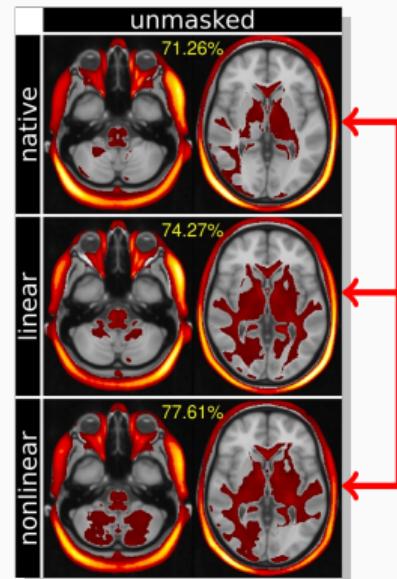
(a) From: Tinauer et al. [1]



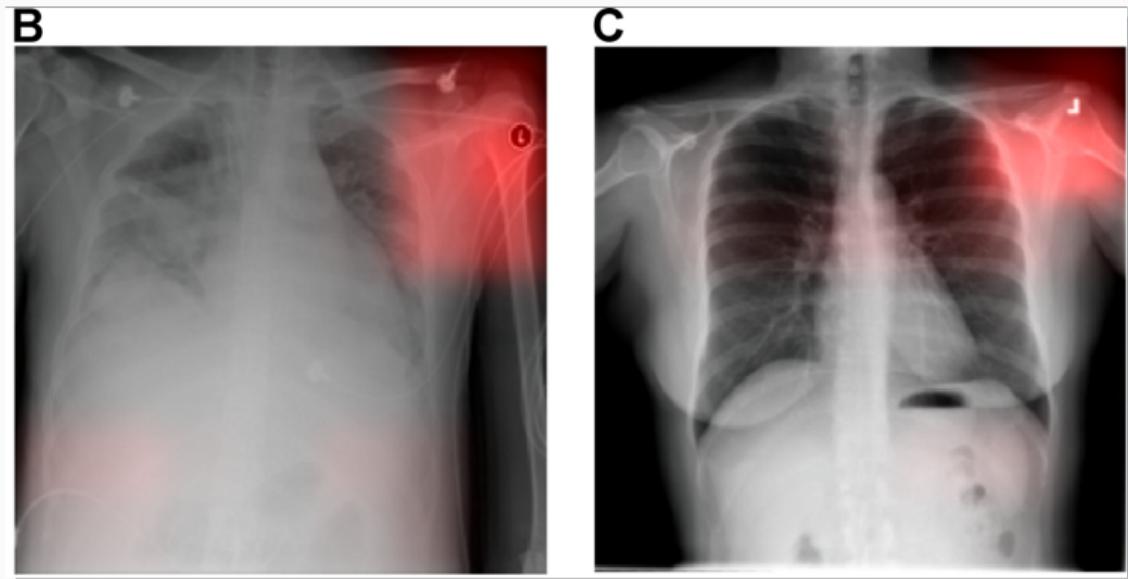
(b) From: Zech et al. [2]

Attribution maps can reveal **shortcut learning**: Neural Networks can use features outside of the brain parenchyma (a) or X-ray side marker tokens (b) for classification.

## Why should I care?



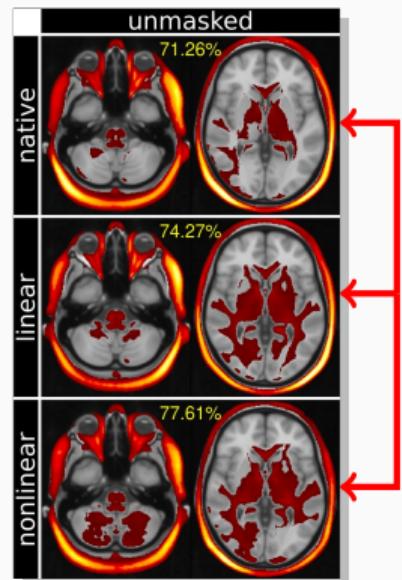
(a) From: Tinauer et al. [1]



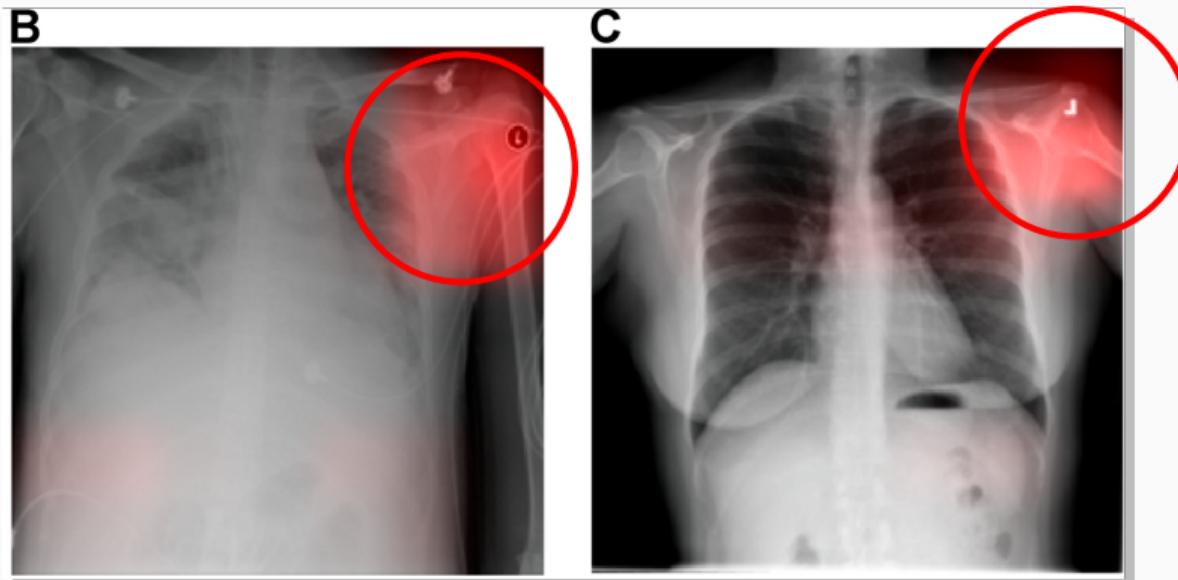
(b) From: Zech et al. [2]

Attribution maps can reveal **shortcut learning**: Neural Networks can use features outside of the brain parenchyma (a) or X-ray side marker tokens (b) for classification.

## Why should I care?

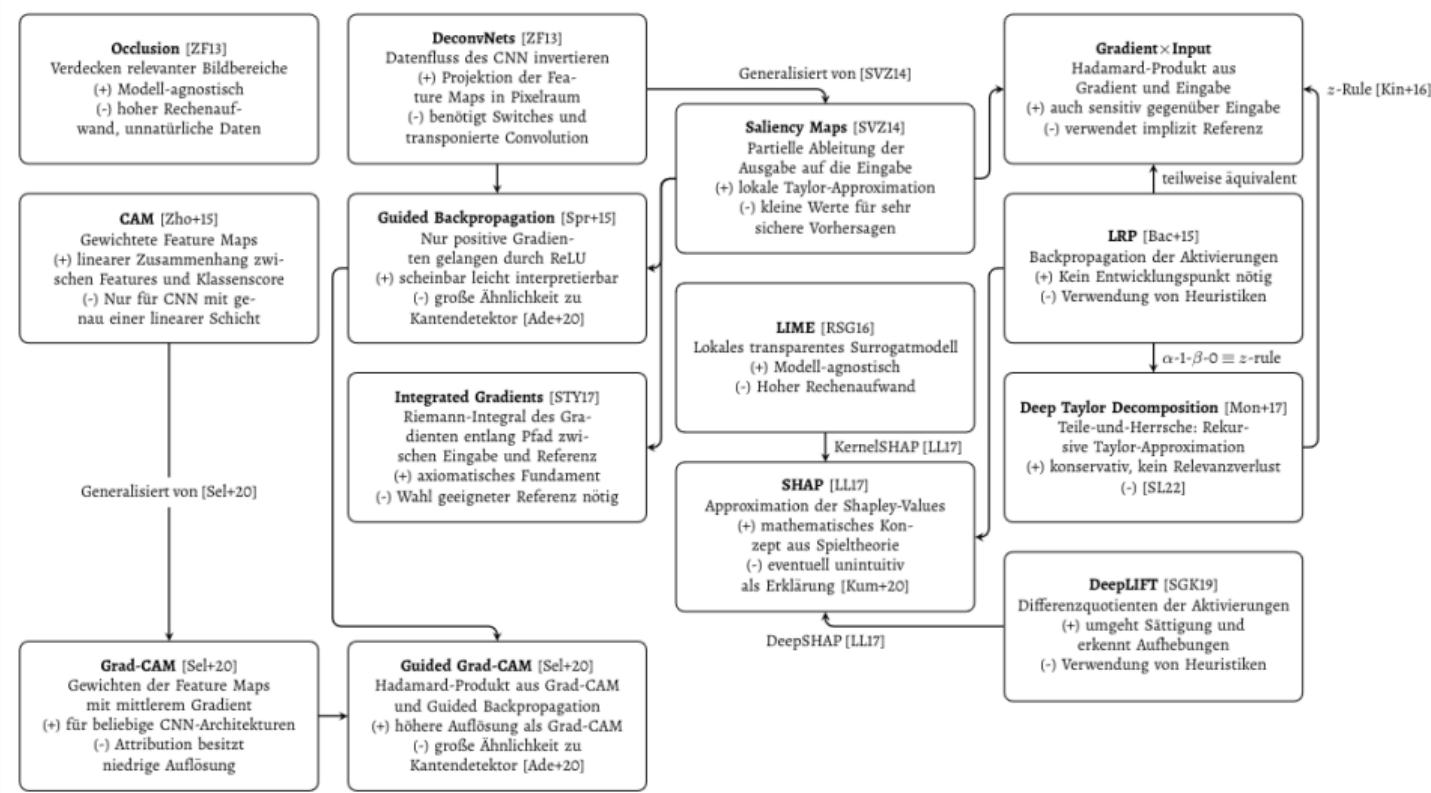


(a) From: Tinauer et al. [1]

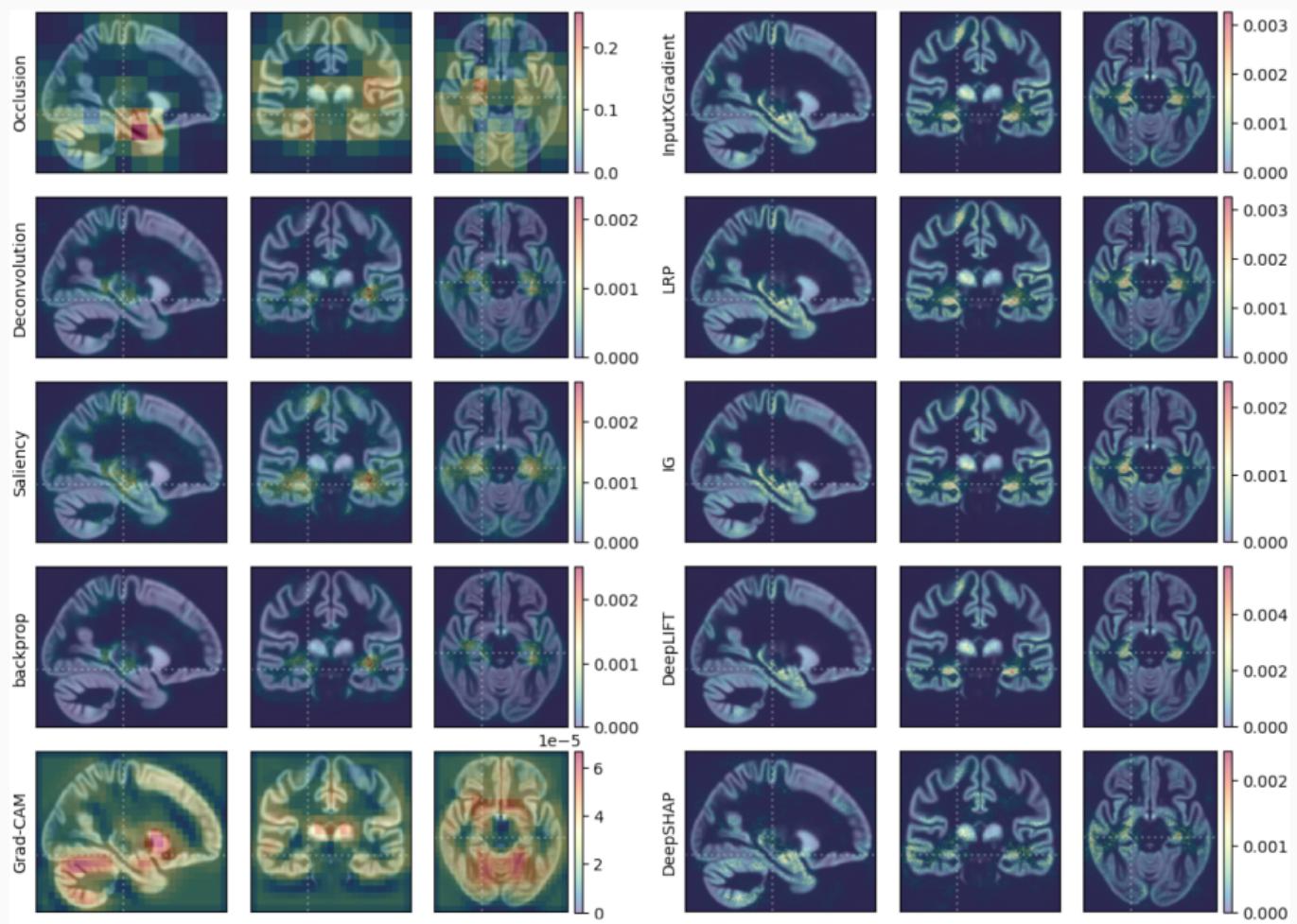


(b) From: Zech et al. [2]

Attribution maps can reveal **shortcut learning**: Neural Networks can use features outside of the brain parenchyma (a) or X-ray side marker tokens (b) for classification.



Popular feature attribution methods for Deep Neural Networks and their Relationships



## Total Relevance per ROI

	Occlusion	IG	DeepLIFT	DeepSHAP
1.	Precuneus_L	Temporal_Mid_L	Temporal_Mid_L	Calcarine_L
2.	Precuneus_R	Temporal_Mid_R	Temporal_Mid_R	Precentral_R
3.	Postcentral_L	Temporal_Inf_L	Temporal_Inf_L	Calcarine_R
4.	Supp_Motor_Area_L	Precentral_R	Precuneus_R	Cerebellum_6_R
5.	Supp_Motor_Area_R	Postcentral_L	Precuneus_L	Precentral_L
6.	Postcentral_R	Frontal_Mid_L	Temporal_Inf_R	Lingual_L
7.	Precentral_L	Postcentral_R	Parietal_Inf_L	Postcentral_R
8.	Cingulum_Mid_R	<b>Hippocampus_L</b>	Frontal_Mid_L	Postcentral_L
9.	Frontal_Sup_Medial_L	Temporal_Inf_R	<b>Hippocampus_L</b>	Lingual_R
10.	Cingulum_Mid_L	Parietal_Inf_L	Supp_Motor_Area_R	Cuneus_L

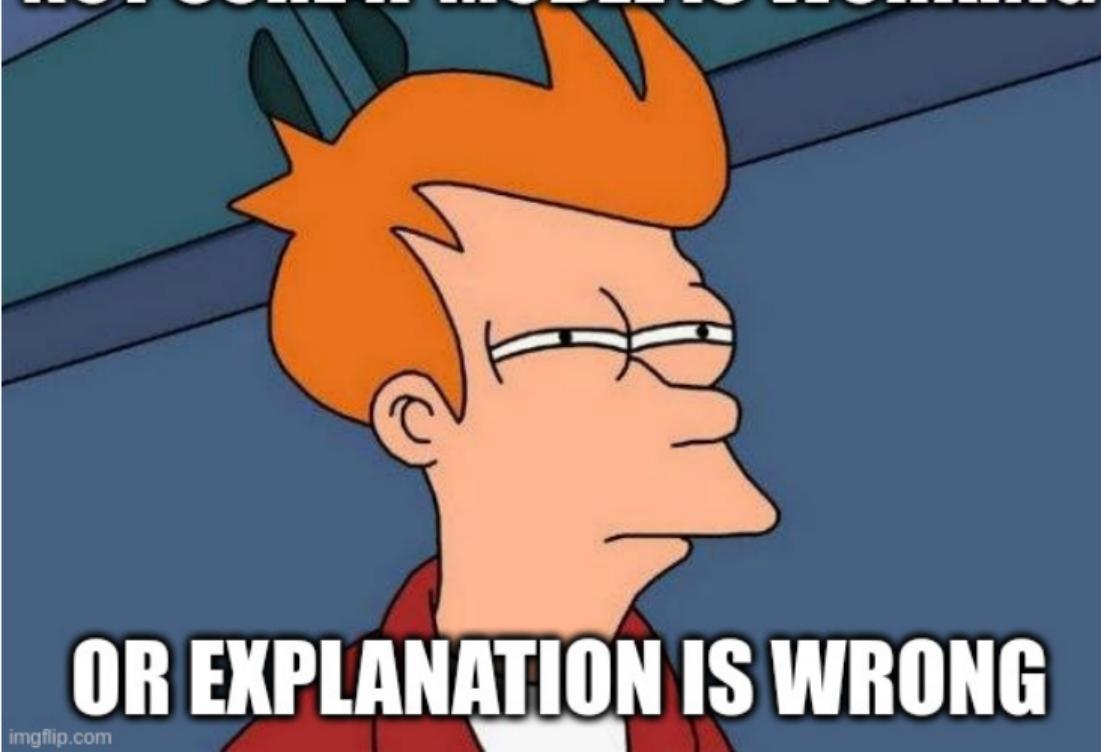
Top 10 AAL ROIs by total relevance for class AD

## Mean Relevance per ROI

	Occlusion	IG	DeepLIFT	DeepSHAP
1.	Supp_Motor_Area_L	<b>Hippocampus_L</b>	<b>Hippocampus_L</b>	Calcarine_L
2.	Supp_Motor_Area_R	<b>Hippocampus_R</b>	<b>Hippocampus_R</b>	Calcarine_R
3.	Rolandic_Oper_L	ParaHippocampal_R	Parietal_Inf_R	Vermis_10
4.	Cingulum_Mid_L	Heschl_L	Amygdala_L	Vermis_7
5.	Cingulum_Mid_R	Parietal_Inf_L	ParaHippocampal_R	Vermis_6
6.	Paracentral_Lobule_R	Thalamus_R	Parietal_Inf_L	Vermis_9
7.	Precuneus_L	Rolandic_Oper_L	Calcarine_R	Vermis_8
8.	Precuneus_R	Temporal_Inf_L	Supp_Motor_Area_R	Cuneus_R
9.	Heschl_L	Temporal_Mid_L	Temporal_Inf_L	Cerebellum_6_R
10.	Frontal_Med_Orb_R	Supp_Motor_Area_R	SupraMarginal_L	Precentral_R

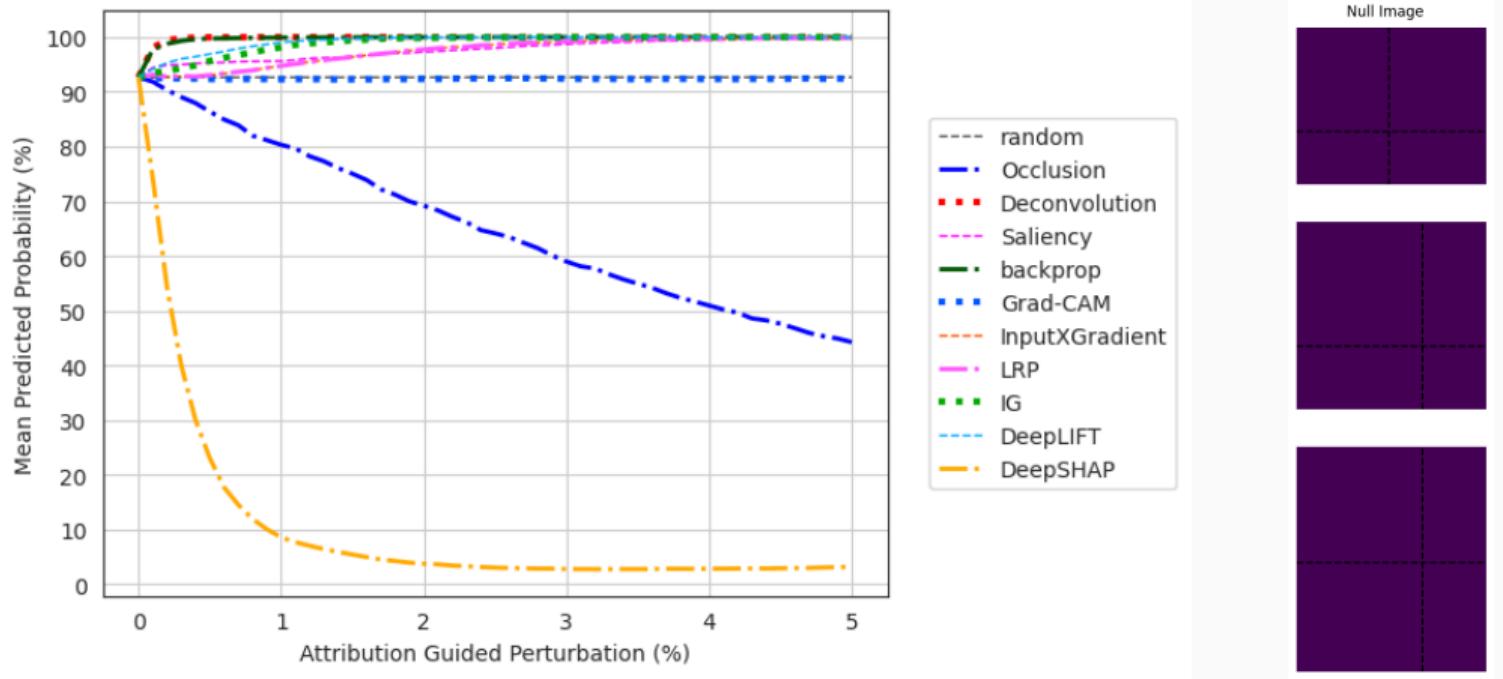
Top 10 AAL ROIs by mean relevance per voxel for class AD

**NOT SURE IF MODEL IS WORKING**



imgflip.com

But can the **explanation** be trusted?

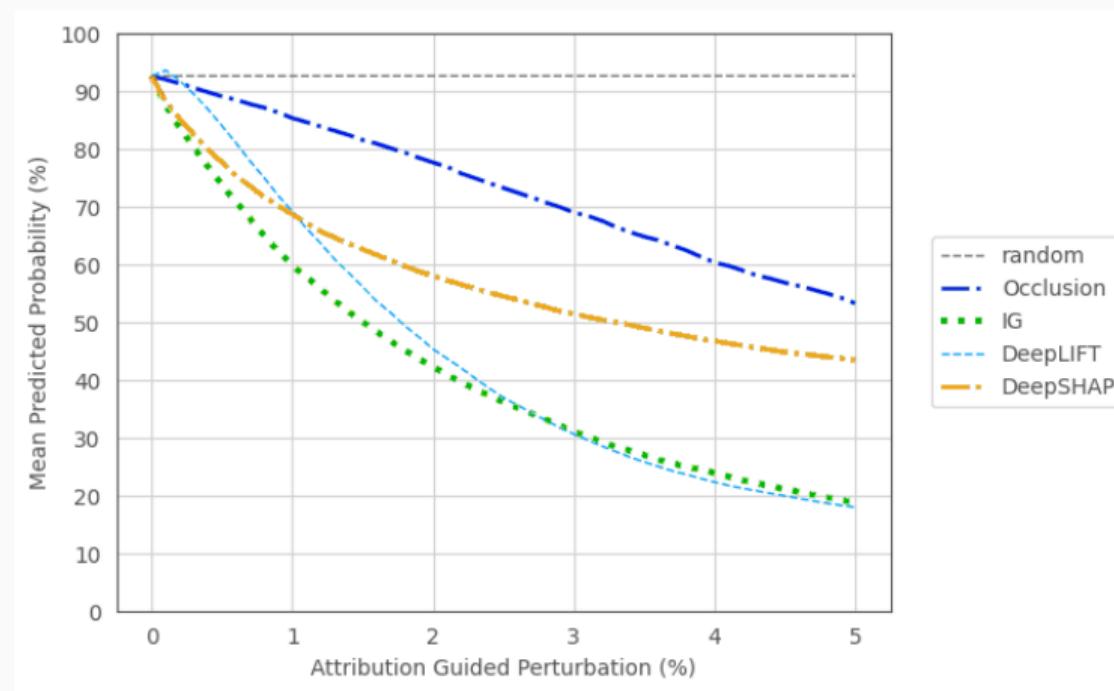


Mean Predicted AD Probability when replacing voxels by the null image baseline

## Hypothesis

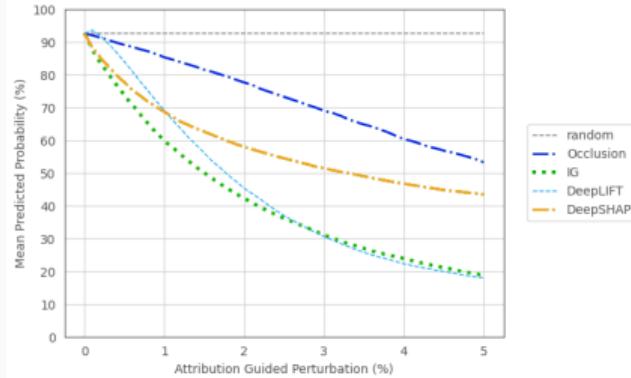
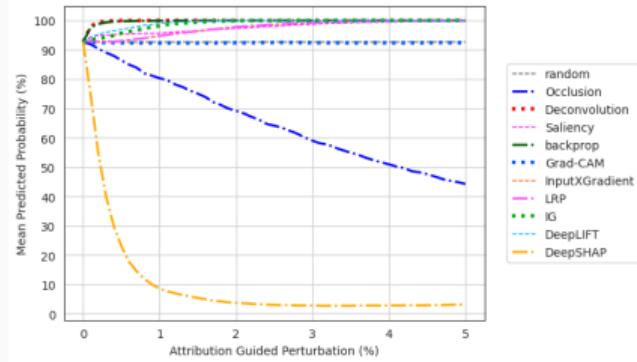
Null image corresponds to "maximum atrophy".

# AD to CN Perturbation: Using the CN mean as Attribution Baseline



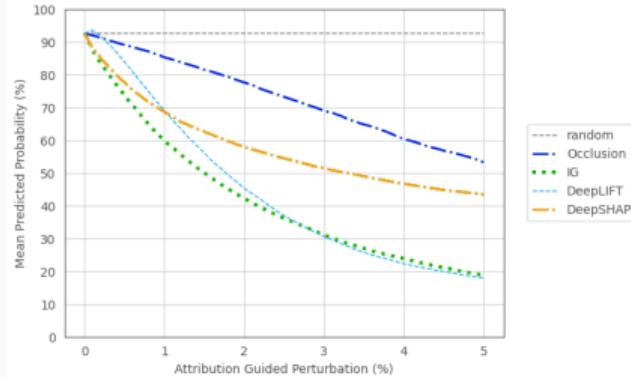
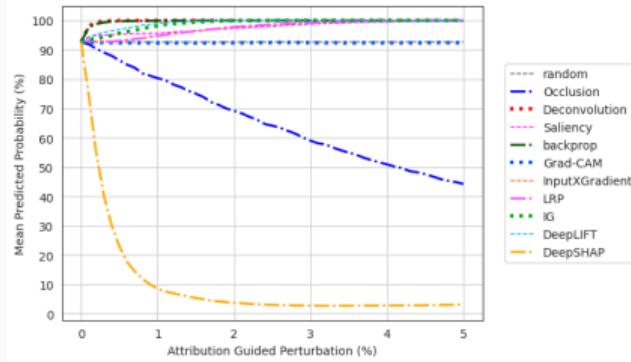
Mean Predicted AD Probability when replacing voxels by the CN mean

# Conclusion



## Take-Aways

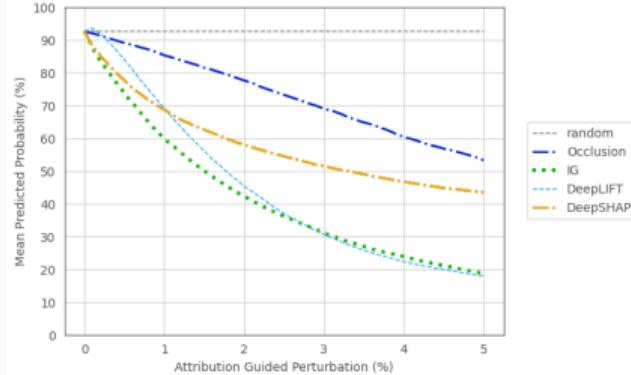
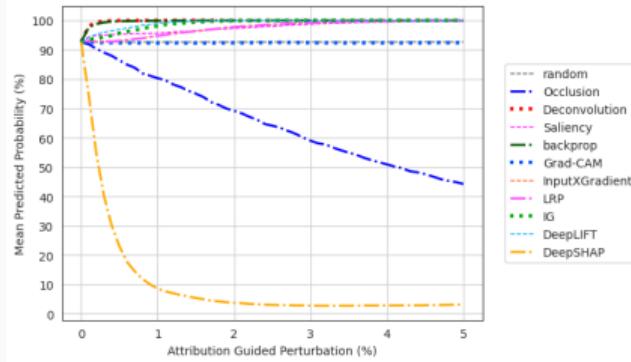
# Conclusion



## Take-Aways

1. Perturbation tests offer a **model-agnostic fidelity metric**.

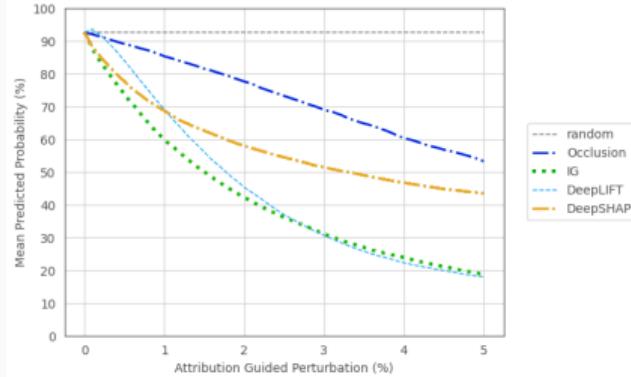
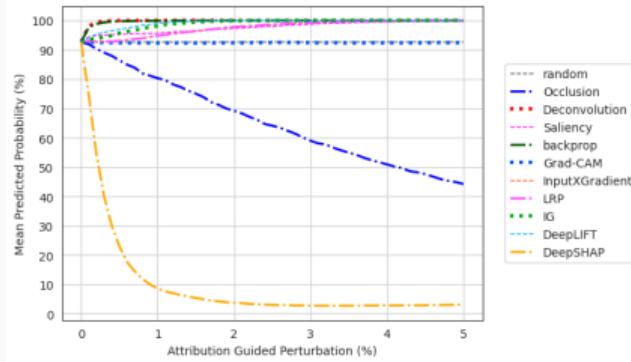
# Conclusion



## Take-Aways

1. Perturbation tests offer a **model-agnostic fidelity metric**.
2. The **attribution baseline** should be chosen carefully.

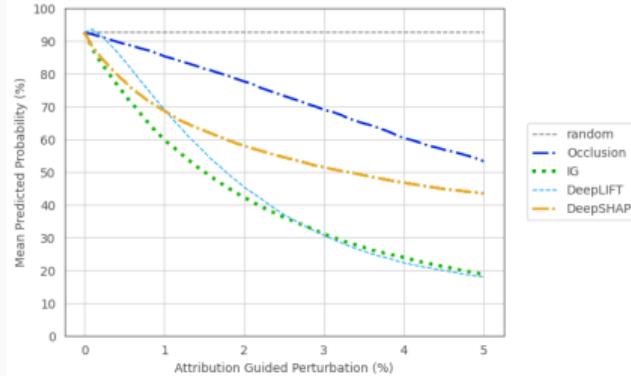
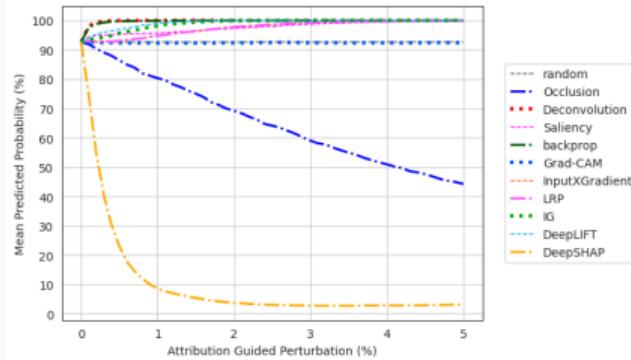
# Conclusion



## Take-Aways

1. Perturbation tests offer a **model-agnostic fidelity metric**.
2. The **attribution baseline** should be chosen carefully.
3. Attribution Maps **need interpretation** to actually explain anything.

# Conclusion



## Take-Aways

1. Perturbation tests offer a **model-agnostic fidelity metric**.
2. The **attribution baseline** should be chosen carefully.
3. Attribution Maps **need interpretation** to actually explain anything.

# Meet the Team



University of Rostock



Thomas Kirste



Martin Becker



Sebastian Bader



Bjarne Hiller

DZNE



Martin Dyrba



Devesh Singh



Thanks for your Attention!

See you on GitHub!  
bckrlab/ad-fidelity



# References i

- [1] Christian Tinauer et al. "Interpretable brain disease classification and relevance-guided deep learning". en. In: **Scientific Reports** 12.1 (Nov. 2022). Publisher: Nature Publishing Group, p. 20254. ISSN: 2045-2322. DOI: 10.1038/s41598-022-24541-7. URL: <https://www.nature.com/articles/s41598-022-24541-7> (visited on 10/12/2024).
- [2] John R. Zech et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". en. In: **PLOS Medicine** 15.11 (June 2018). Publisher: Public Library of Science, e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683. URL: <https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002683> (visited on 10/12/2024).