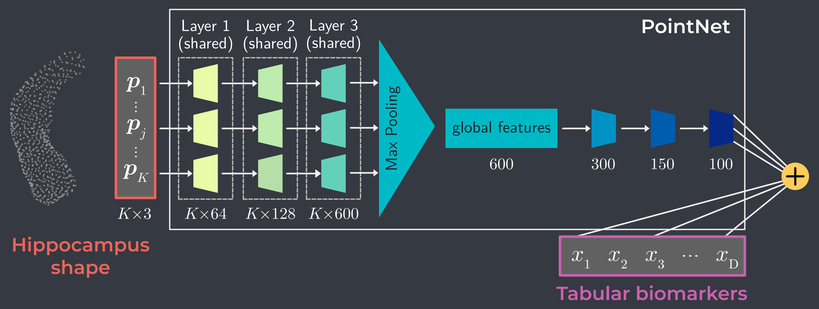
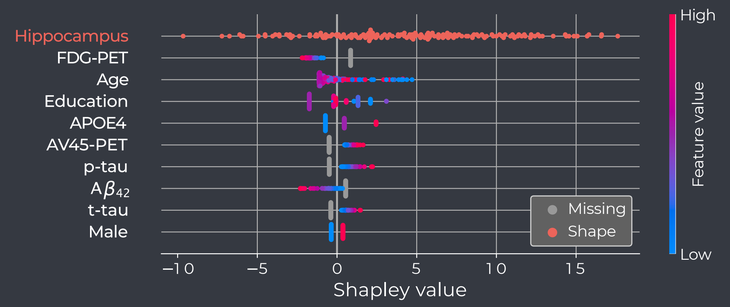
Faithful and scalable **explanation of decisions** of a deep neural network from **anatomical shape** and **tabular biomarkers**.





Scalable, Axiomatic Explanations of Deep Alzheimer's Diagnosis from Heterogeneous Data

Sebastian Pölsterl, Christine Aigner, and Christian Wachinger

LMU KLINIKUM DER UNIVERSITÄT MÜNCHEN AI med

Explainable Artificial Intelligence (XAI)

- *Goal*: **Explain Alzheimer's diagnosis** made by a deep neural network f from **anatomical shape** and **tabular biomarkers**. Most existing work on XAI focus on CNN, but
  1. Inputs are *heterogeneous*,
  2. Shapes are *non-Euclidean*,
  3. Network *differs substantially from a standard CNN*.
- Axioms of explanations:

	Comp-leteness	Null Player	Symmetry	Scale Invariance	Linear	Continuity	Implement. Invariance
Occlusion (Zeiler and Fergus, 2014)	✗	✓	✓	✓	✓	✗	✓
Guided Grad-CAM (Selvaraju et al., 2017)	✗	✓	✓	✓	✓	✗	✓
Layer-wise relevance prop. (Bach et al., 2015)	✓	✓	✓	✓	✓	✓	✗
DeepLift (Shrikumar et al., 2017)	✓	✓	✓	✓	✓	✓	✗
Integrated Gradients (Sundararajan et al., 2017)	✓	✓	✓	✓	✓	✗	✓
Shapley Value (Shapley, 1953)	✓	✓	✓	✓	✓	✓	✓

Methods

- Shapley Value of feature i of input $\mathbf{z}$:

$$\Delta_i = f(\mathbf{z}_{\mathcal{S}\cup\{i\}}; \mathbf{z}^{\mathrm{bl}}_{\mathcal{F}\setminus\mathcal{S}\cup\{i\}} - f(\mathbf{z}_{\mathcal{S}}; \mathbf{z}^{\mathrm{bl}}_{\mathcal{F}\setminus\mathcal{S}})$$

$$s_i(\mathbf{z} \mid f) = \frac{1}{|\mathcal{F}|!} \sum_{\mathcal{S}\subseteq\mathcal{F}\setminus\{i\}} |\mathcal{S}|! \cdot (|\mathcal{F}| - |\mathcal{S}| - 1)! \cdot \Delta_i$$

$$= \frac{1}{|\mathcal{F}|!} \sum_{k=0}^{|\mathcal{F}|-1} \sum_{\substack{\mathcal{S}\subseteq\mathcal{F}\setminus\{i\} \\ |\mathcal{S}|=k}} k!(|\mathcal{F}| - k - 1)! \cdot \Delta_i$$

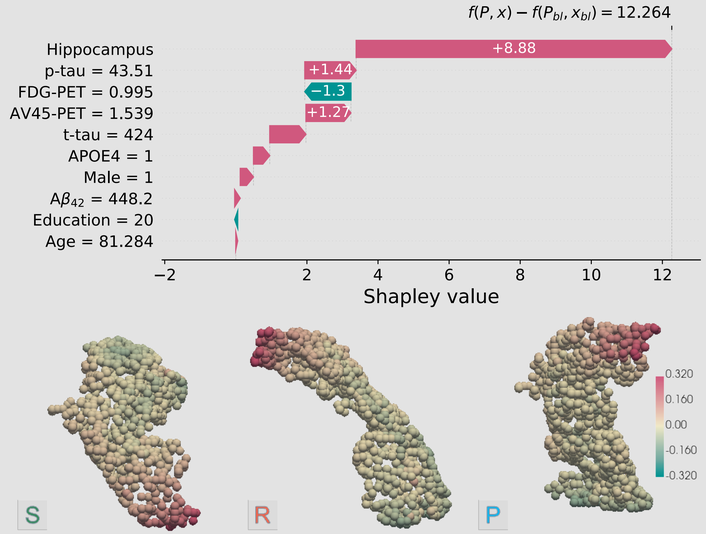
- ☹ **Scales exponentially** in the number of features $|\mathcal{F}|$.
- $\mathbf{z}^{\mathrm{bl}}_{\mathcal{F}\setminus\mathcal{S}}$: Simulate absence of point by replacing it with projection onto the convex hull.
- Approximate Shapley value:
  1. Represent output of first layer as *normal distribution* using sampling theory (Ancona et al., 2019; Cochran, 1977).
  2. Propagate distributions by converting layers into a *Lightweight Probabilistic Deep Network* (LDPN, Gast and Roth, 2018).

$$s_i(\mathbf{z} \mid f) \approx \underbrace{\mathbb{E}_k[f(\mathbf{z}_{\mathcal{S}\cup\{i\}}; \mathbf{z}^{\mathrm{bl}}_{\mathcal{F}\setminus\mathcal{S}\cup\{i\}})]}_{\text{Output of LPDN}} - \underbrace{\mathbb{E}_k[f(\mathbf{z}_{\mathcal{S}}; \mathbf{z}^{\mathrm{bl}}_{\mathcal{F}\setminus\mathcal{S}})]}_{\text{Output of LPDN}}$$

- ☺ Approximate Shapley value **scales linearly**.

Experiment: Real Data from ADNI

- Use Wide and Deep Network with left hippocampus shape ($K = 1024$ points) and 9 tabular biomarkers.
- 1308 visits for training, 169 for hyper- parameter tuning.
- Balanced accuracy: 0.942 on test set (176 patients).
- Explanation of diagnosis for *individual* patient:



Experiment: Synthetic Data

- Use PointNet to classify 100 point clouds of "X" and "I".
- Compare against *exact* Shapley value.

Method	MSE	SRC	NDCG	NE
SV Exact	0	1	1	65,538
Occlusion	16.1311	0.3180	0.8659	17
SV Sampling	0.0008	0.9505	0.9986	32,000
SV Sampling	0.0340	0.5440	0.9481	512
Proposed	0.0443	0.6918	0.9641	512

MSE: mean squared error. SRC: Spearman's rank correlation.
NDCG: normalized discounted cumulative gain. NE: forward passes.