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

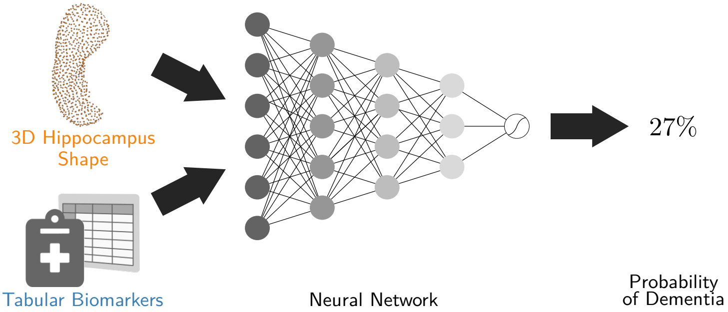
Sebastian Pölsterl, Christina Aigner and Christian Wachinger

Artificial Intelligence in Medical Imaging, Ludwig-Maximilians-Universität, Munich

International Conference on Medical Image Computing and Computer Assisted Intervention

Sept. 27th – Oct. 1st 2021

Alzheimer's Disease Diagnosis

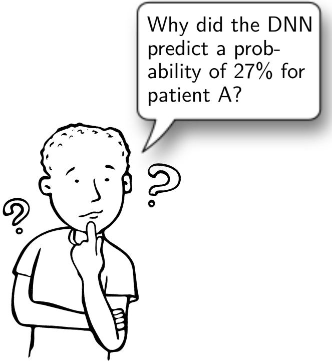


- Assume we have successfully trained a DNN f to *accurately* predict AD diagnosis from the **hippocampus shape** and **tabular biomarkers** of an individual:

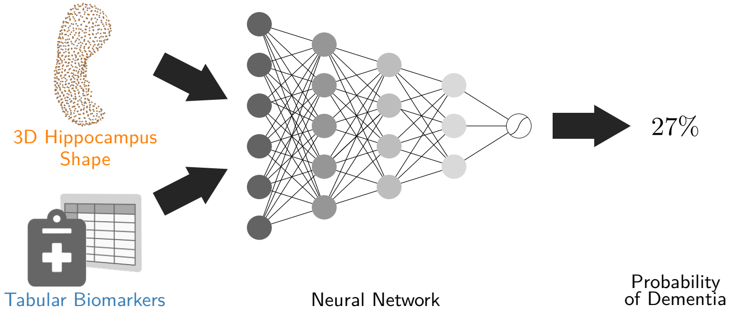
$$f : \mathbb{R}^{K \times 3} \times \mathbb{R}^{D} \rightarrow [0; 1].$$

Explainable Artificial Intelligence (XAI)

- Predictions by a DNN are opaque, therefore we require **post-hoc explainability** techniques.
- Our objective:
  inform the user about the decision making process.



XAI in Alzheimer's Disease



- The input data are **heterogeneous**.
- Point clouds are **non-Euclidean**.
- Requires networks that **differ substantially from standard CNNs**.

Axioms of an Explanation

	Completeness	Null Player	Symmetry	Scale Invariance	Linearity	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, Taly, et al., 2017)	✓	✓	✓	✓	✓	✗	✓
Shapley Value (Shapley, 1953)	✓	✓	✓	✓	✓	✓	✓

See Ancona et al. (2019), Montavon (2019), and Sundararajan, Taly, et al. (2017) for proofs.

Shapley Value (Shapley, 1953)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Definition (Shapley Value)

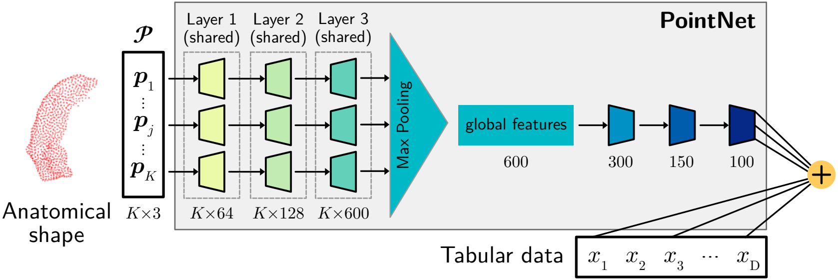
$$s_i(\mathbf{z} \mid f) = \frac{1}{|\mathcal{F}|!} \sum_{\mathcal{S} \subseteq \mathcal{F} \backslash \{i\}} |\mathcal{S}|! \cdot (|\mathcal{F}| - |\mathcal{S}| - 1)! \underbrace{[g(\mathcal{S} \cup \{i\}) - g(\mathcal{S})]}_{=\Delta_i}.$$

- Average over all subsets $\mathcal{S} \subseteq \mathcal{F} \backslash \{i\}$ ($\mathcal{F}$ comprises all features of the input $\mathbf{z}$).
- $g(\mathcal{S})$ measures the impact of feature set $\mathcal{S}$ (Sundararajan and Najmi, 2020):

$$g(\mathcal{S}) = f(\mathbf{z}_{\mathcal{S}}; \mathbf{z}^{\text{bl}}_{\mathcal{F} \backslash \mathcal{S}}) - f(\mathbf{z}^{\text{bl}}), \qquad \mathbf{z}^{\text{bl}}_{\mathcal{F} \backslash \mathcal{S}} : \text{Replace features} \notin \mathcal{S} \text{ with a baseline value.}$$

- Shapley value scales **exponentially in the number of features**.
  $\Rightarrow$ Need to approximate it.

Estimation of Shapley Value



Wide and Deep Network proposed in Pölsterl et al. (2020).

☺ Tabular feature: only depends on the i-th weight of the last linear layer.

☹ Point of the hippocampus: depends on the *entire* PointNet.
$\Rightarrow$ Need to approximate the Shapley value.

Approximate Shapley Value (Fatima et al., 2008)

- Explicitly sum over all sets $\mathcal{S}$ of equal size to obtain **linear** runtime:

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

$$\approx \frac{1}{|\mathcal{F}|} \sum_{k=0}^{|\mathcal{F}|-1} \mathbb{E}_k(\Delta_i)$$

- Only need to estimate $\mathbb{E}_k(\Delta_i)$:

$$\mathbb{E}_k(\Delta_i) = \mathbb{E}_k[f(\mathbf{z}_{\mathcal{S} \cup \{i\}}; \mathbf{z}^{\mathsf{bl}}_{\mathcal{F} \setminus \mathcal{S} \cup \{i\}})] - \mathbb{E}_k[f(\mathbf{z}_{\mathcal{S}}; \mathbf{z}^{\mathsf{bl}}_{\mathcal{F} \setminus \mathcal{S}})].$$

Shapley Values of Anatomical Shape

Objective:

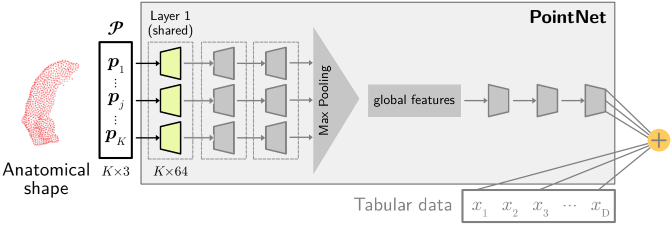
- Estimate $\mathbb{E}_k[f(\mathbf{z}_\mathcal{S}; \mathbf{z}^{\text{bl}}_{\mathcal{F} \setminus \mathcal{S}})]$.

Problem:

☹ $f(\mathbf{z}_\mathcal{S}; \mathbf{z}^{\text{bl}}_{\mathcal{F} \setminus \mathcal{S}})$ depends on the *entire* PointNet.

Solution:

- Represent output of first layer as a **normal distribution**.
- The objective becomes **propagating aleatoric uncertainty**.
- Transform remaining layers into a **Lightweight Probabilistic Deep Network** (Gast and Roth, 2018).



- **Objective**: Estimate $\mathbb{E}_k[f(\mathbf{z}_\mathcal{S}; \mathbf{z}^{\text{bl}}_{\mathcal{F}\setminus\mathcal{S}})]$.
- First PointNet layer yields $\mathbf{h}_j = \left(\sum_{l=1}^3 p_{jl} W_{l\,1}, \ \ldots \ , \sum_{l=1}^3 p_{jl} W_{l\,64}\right)^\top$.
- Whether $j \in \mathcal{S}$ is random, we only know $|\mathcal{S}| = k$.

Normal Approximation (II)

Objective:

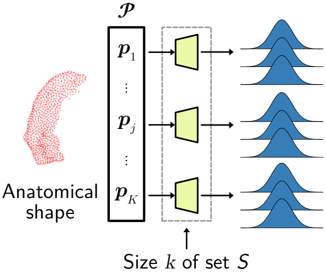
- Approximate output of first layer with a **normal distribution**.

Solution:

- *Sampling theory* suggests approximation with a normal distribution (Ancona et al., 2019; Cochran, 1977):

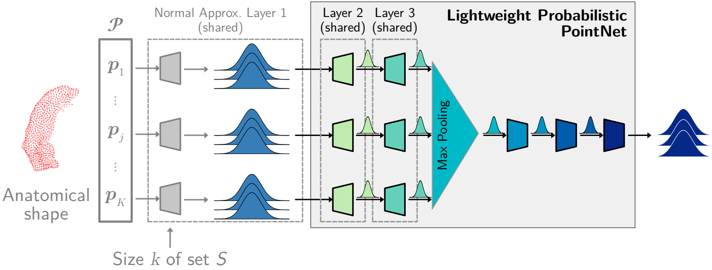
$$\mathbb{E}_k[h_{jm}] = \frac{k}{|\mathcal{F}|} h_{jm},$$

$$\mathbb{V}_k(h_{jm}) = k \frac{|\mathcal{F}| - k}{|\mathcal{F}| - 1} \left[\frac{1}{|\mathcal{F}|} \sum_{l=1}^{3} (p_{jl} W_{lm})^2 - \left(\frac{1}{|\mathcal{F}|} h_{jm} \right)^2 \right].$$

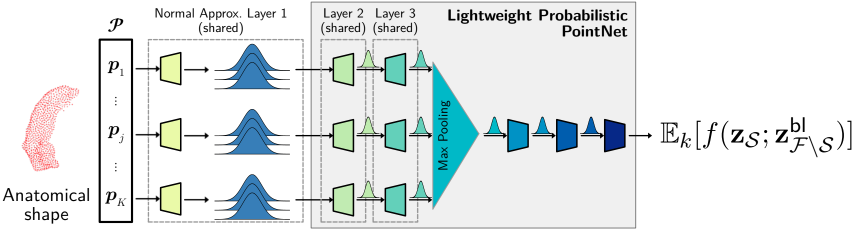


Propagating Aleatoric Uncertainty

- Outputs of first layer are approximated by independent normal distributions.
- **Propagate distributions** using a Lightweight Probabilistic Deep Network (Gast and Roth, 2018).
- Replace layers with their probabilistic counterpart:
  ReLU, batch-norm, and max-pooling, fully-connected.



Efficient Shapley Value Estimation



- Require $2|\mathcal{F}|$ forward passes:

$$s_i(\mathbf{z} \mid f) \approx \frac{1}{|\mathcal{F}|} \sum_{k=0}^{|\mathcal{F}|-1} \underbrace{\mathbb{E}_k[f(\mathbf{z}_{\mathcal{S}\cup\{i\}}; \mathbf{z}^{\text{bl}}_{\mathcal{F}\setminus\mathcal{S}\cup\{i\}})]}_{\text{Output of LPDN}} - \underbrace{\mathbb{E}_k[f(\mathbf{z}_{\mathcal{S}}; \mathbf{z}^{\text{bl}}_{\mathcal{F}\setminus\mathcal{S}})]}_{\text{Output of LPDN}} .$$

- Runtime: $\mathcal{O}(|\mathcal{F}|)$.

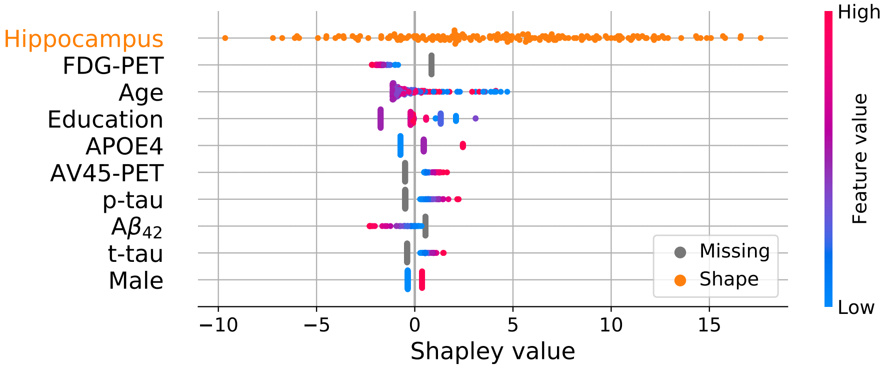
Experiments

1. Quantitative evaluation on synthetic data.
2. Qualitative evaluation on data from the Alzheimer's Disease Neuroimaging Initiative.

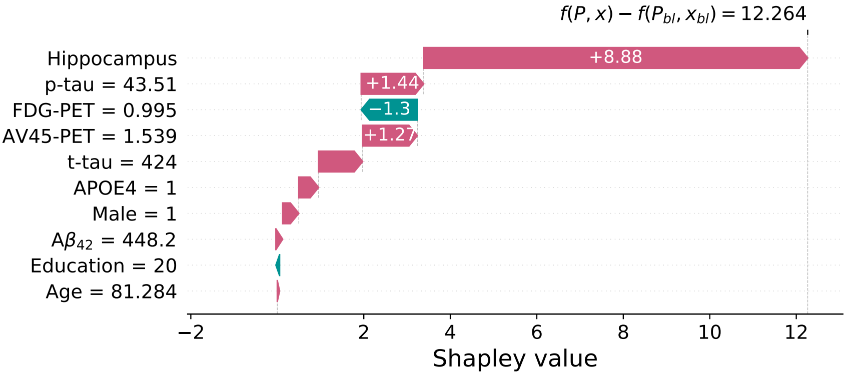
Experiments – Alzheimer's Disease Diagnosis

- **Data**: T1 MRI from the Alzheimer's Disease Neuroimaging Initiative (Jack et al., 2008).
- **Network**: Wide and Deep PointNet (Pölsterl et al., 2020).
- **Anatomical shape**: Left hippocampus point cloud (1024 points).
- **Tabular data**:
  - 9 features (demographics, APOE4, CSF, AV45-PET, FDG-PET).
  - Explicitly encode missing values via indicator variables.
- **Balanced accuracy**: 0.942 on the test data.

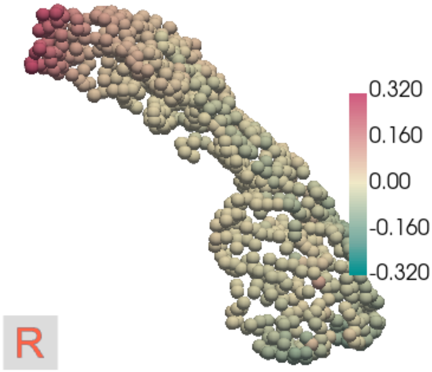
Shapley Values of 167 Correctly Classified Patients



Shapley Values of Individual Patient



Shapley Values of Hippocampus



Conclusion

- An axiomatic approach based on the Shapley value to explain predictions of a DNN.
- Approximation of the Shapley value requires a quadratic (instead of exponential) number of network evaluations.
- Explain Alzheimer's diagnosis of a DNN from anatomical shape and tabular biomarkers.

Thanks For Your Attention!

sebastian.poelsterl@med.uni-muenchen.de

www.ai-med.de

github.com/ai-med

AI_Medic

Lab for AI in Medical Imaging

Founding sources: Bavarian State Ministry of Science and the Arts, Federal Ministry of Education and Research.

References I

Ancona, M., C. Oztireli, and M. Gross (2019). “Explaining Deep Neural Networks with a Polynomial Time Algorithm for Shapley Value Approximation”. In: *Proc. of the 36th International Conference on Machine Learning*. Vol. 97, pp. 272–281.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek (July 2015). “On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation”. In: *PLOS ONE* 10.7, e0130140.

Cochran (1977). *Sampling Techniques*. 3rd. John Wiley & Sons.

Fatima, S. S., M. Wooldridge, and N. R. Jennings (Sept. 2008). “A linear approximation method for the Shapley value”. In: *Artificial Intelligence* 172.14, pp. 1673–1699.

Gast, J. and S. Roth (2018). “Lightweight Probabilistic Deep Networks”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3369–3378.

Jack, C. R., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, et al. (2008). “The Alzheimer’s disease neuroimaging initiative (ADNI): MRI methods”. In: *Journal of Magnetic Resonance Imaging* 27.4, pp. 685–691.

References II

Montavon, G. (2019). "Gradient-Based Vs. Propagation-Based Explanations: An Axiomatic Comparison". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, pp. 253–265.

Pölsterl, S., I. Sarasua, B. Gutiérrez-Becker, and C. Wachinger (2020). "A Wide and Deep Neural Network for Survival Analysis from Anatomical Shape and Tabular Clinical Data". In: *Machine Learning and Knowledge Discovery in Databases*, pp. 453–464.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra (2017). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Shapley, L. S. (1953). "A value for n-person games". In: *Contributions to the Theory of Games* 2.28, pp. 307–317.

Shrikumar, A., P. Greenside, and A. Kundaje (2017). "Learning Important Features Through Propagating Activation Differences". In: *Proc. of the 34th International Conference on Machine Learning*. Vol. 70, pp. 3145–3153.

Sundararajan, M. and A. Najmi (2020). "The many Shapley values for model explanation". In: *Proc. of the 37th International Conference on Machine Learning*. Vol. 119, pp. 9269–9278.

References III

Sundararajan, M., A. Taly, and Q. Yan (2017). "Axiomatic Attribution for Deep Networks". In: *Proc. of the 34th International Conference on Machine Learning*. Vol. 70, pp. 3319–3328.

Zeiler, M. D. and R. Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *European Conference on Computer Vision (ECCV)*, pp. 818–833.