Can We Trust a Neural Network Prediction?

Methods and Pitfalls for Explaining Black Boxes

Niklas Koenen[1,2] & Marvin N. Wright[1,2,3]

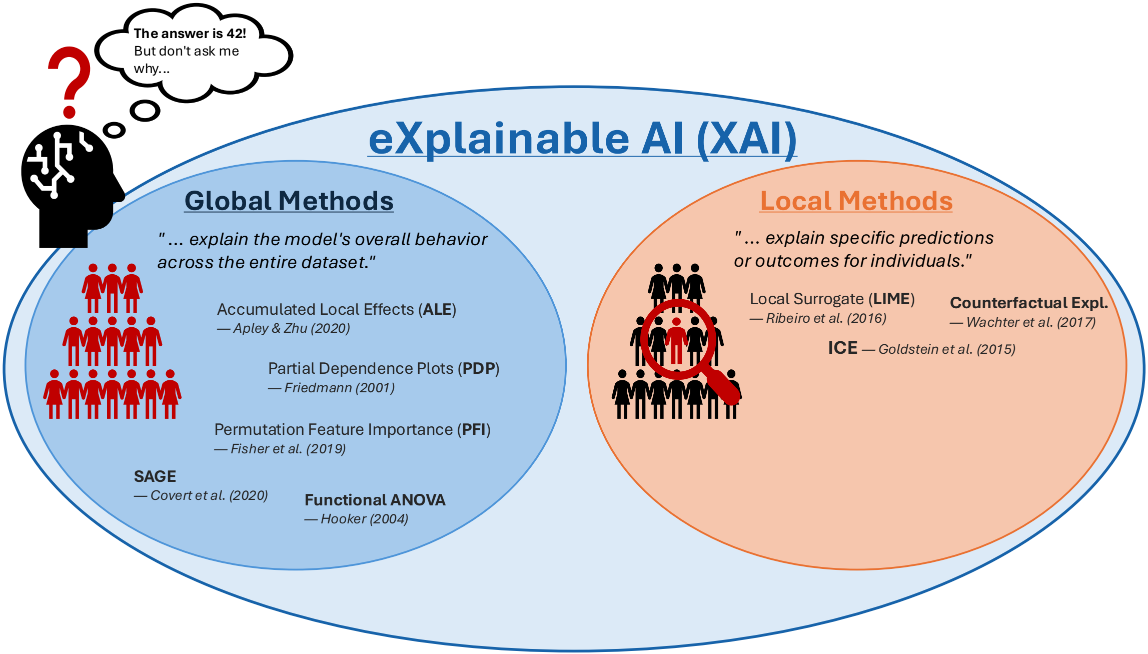
[1]Leibniz Institute for Prevention Research & Epidemiology – BIPS

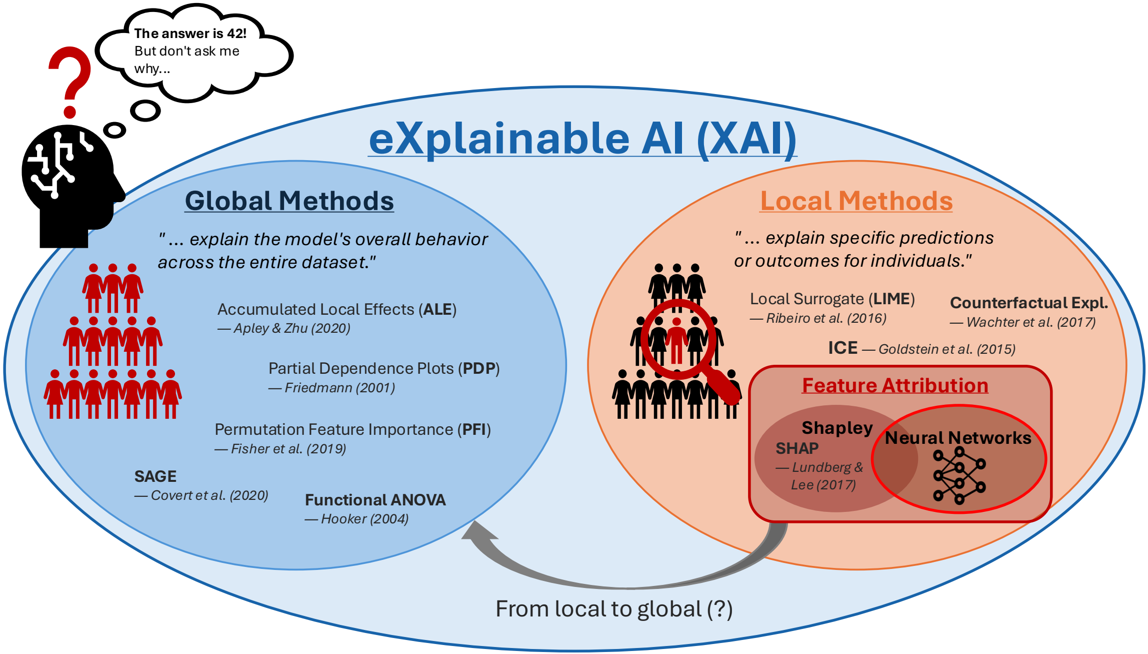
[2]Faculty of Mathematics and Computer Science, University of Bremen

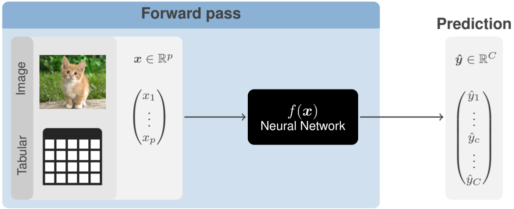
[3]Department of Public Health, University of Copenhagen

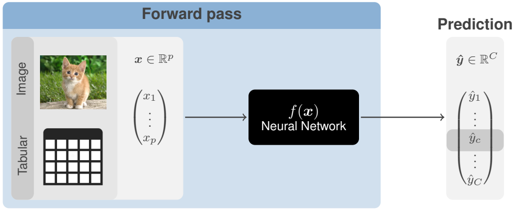
December 9th, 2024

32nd International Biometric Conference – Atlanta


The answer is 42!
But don't ask me why...
eXplainable AI (XAI)
Global Methods
"... explain the model's overall behavior across the entire dataset."
Accumulated Local Effects (ALE)
— Apley & Zhu (2020)
Partial Dependence Plots (PDP)
— Friedmann (2001)
Permutation Feature Importance (PFI)
— Fisher et al. (2019)
SAGE
— Covert et al. (2020)
Functional ANOVA
— Hooker (2004)
Local Methods
"... explain specific predictions or outcomes for individuals."
Local Surrogate (LIME)
— Ribeiro et al. (2016)
Counterfactual Expl.
— Wachter et al. (2017)
ICE — Goldstein et al. (2015)

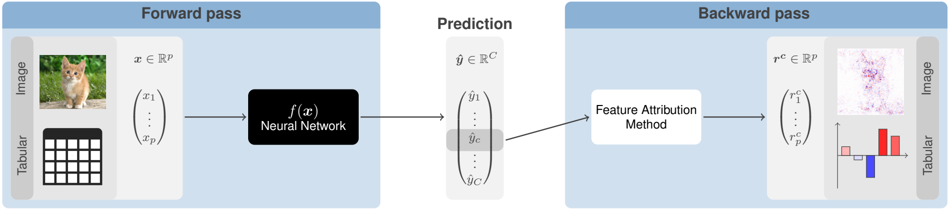

The answer is 42!
But don't ask me why...
eXplainable AI (XAI)
Global Methods
"... explain the model's overall behavior across the entire dataset."
Accumulated Local Effects (ALE)
— Apley & Zhu (2020)
Partial Dependence Plots (PDP)
— Friedmann (2001)
Permutation Feature Importance (PFI)
— Fisher et al. (2019)
SAGE
— Covert et al. (2020)
Functional ANOVA
— Hooker (2004)
Local Methods
"... explain specific predictions or outcomes for individuals."
Local Surrogate (LIME)
— Ribeiro et al. (2016)
Counterfactual Expl.
— Wachter et al. (2017)
ICE — Goldstein et al. (2015)
Feature Attribution
Shapley
SHAP
— Lundberg & Lee (2017)
Neural Networks
From local to global (?)


Forward pass
Image
Tabular
$x \in \mathbb{R}^p$
x_1
x_p
$f(x)$
Neural Network
Prediction
$\hat{y} \in \mathbb{R}^C$
$\hat{y}_1$
$\hat{y}_c$
$\hat{y}_C$



Forward pass
Image
Tabular
$x \in \mathbb{R}^p$
x_1
x_p
$f(x)$
Neural Network
Prediction
$\hat{y} \in \mathbb{R}^C$
$\hat{y}_1$
$\hat{y}_c$
$\hat{y}_C$

What is feature attribution?



- Assigns an attribution value of a selected output to each input variable
  - $\longrightarrow$ also known as contribution or relevance
  - $\longrightarrow$ can be positive (in red) or negative (blue)
- Utilizes the
  - $\longrightarrow$ layer-wise architecture of a neural network
  - $\longrightarrow$ automatic differentiation engine of the training
- Extremely efficient and fast
  - $\longrightarrow$ thanks to deep learning libraries like PyTorch, Keras/TensorFlow etc.
  - $\longrightarrow$ generally only one forward and one backward pass is needed (no optimization or estimation)

Which Method Should I Use?

Saliency/Gradient!!
Keep it simple

DeepLIFT!!
Need a reasonable reference

SmoothGrad!!
Smooth for clarity

IntegratedGradient!!
Need axiomatic consideration

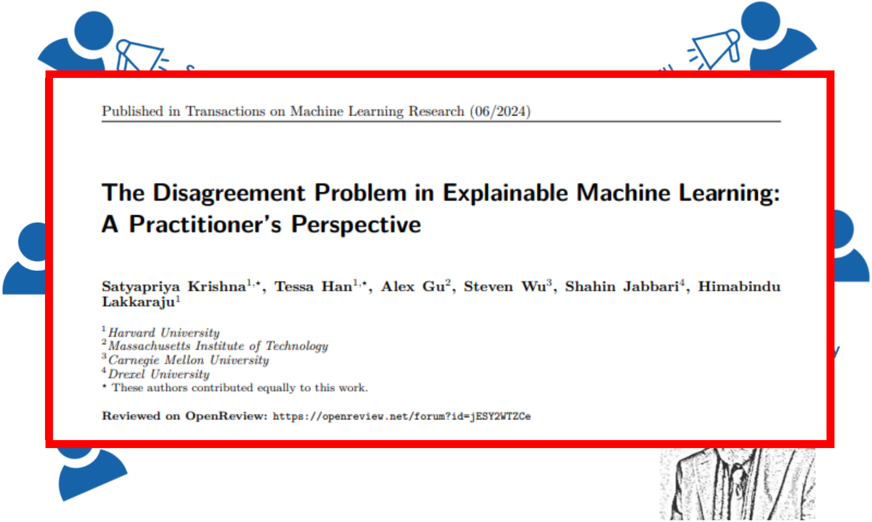
User

LRP!!
Class-sensitivity

Come on, guys!
There is only one
FAIR option.

Lloyd S. Shapley

Which Method Should I Use?



Published in Transactions on Machine Learning Research (06/2024)

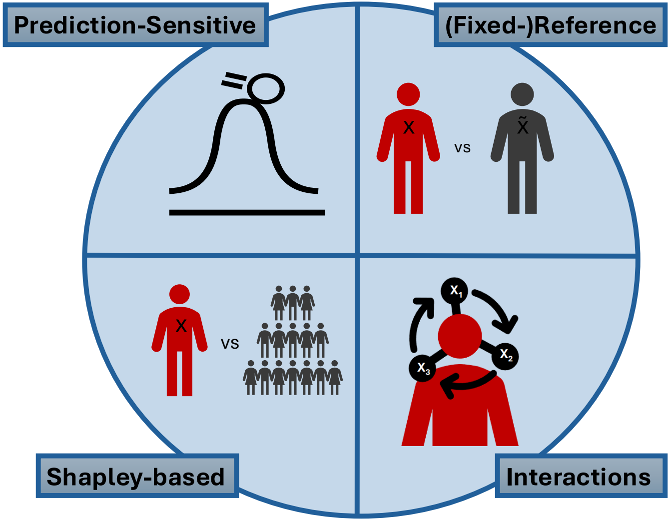
The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective

Satyapriya Krishna[1,*], Tessa Han[1,*], Alex Gu[2], Steven Wu[3], Shahin Jabbari[4], Himabindu Lakkaraju[1]

[1] *Harvard University*
[2] *Massachusetts Institute of Technology*
[3] *Carnegie Mellon University*
[4] *Drexel University*
[*] These authors contributed equally to this work.

Reviewed on OpenReview: https://openreview.net/forum?id=jESY2WTZCe

Which Method Should I Use? – It depends...



Simulation

Features (real):

BMI, Age, Gender and HbA1c

Outcome (binary)

Diabetes (synth.)

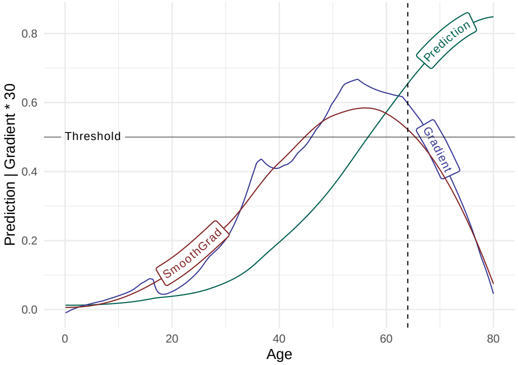
Prediction-Sensitive Methods

"How sensitive is the prediction w.r.t. this feature?"

Gradient[1]: $\frac{\partial f(\boldsymbol{x})}{\partial x_i}$ **SmoothGrad**[2]: $\sum_{k=1}^{K} \frac{\partial f(\boldsymbol{x} + \boldsymbol{\varepsilon^{(k)}})}{\partial x_i + \varepsilon_i^{(k)}} \quad \left(\varepsilon^{(k)} \sim \mathcal{N}(0, I\sigma)\right)$

Interpretation:

- Increasing patient's age increases model's prediction for diabetes



[1] Simonyan et al. (2014) ● [2] Smilkov et al. (2017)

Prediction-Sensitive Methods

"How sensitive is the prediction w.r.t. this feature?"

Gradient[1]: $\frac{\partial f(\boldsymbol{x})}{\partial x_i}$ **SmoothGrad**[2]: $\sum_{k=1}^{K} \frac{\partial f(\boldsymbol{x} + \boldsymbol{\varepsilon}^{(\boldsymbol{k})})}{\partial x_i + \varepsilon_i^{(k)}}$ $\left(\varepsilon^{(k)} \sim \mathcal{N}(0, I\sigma)\right)$

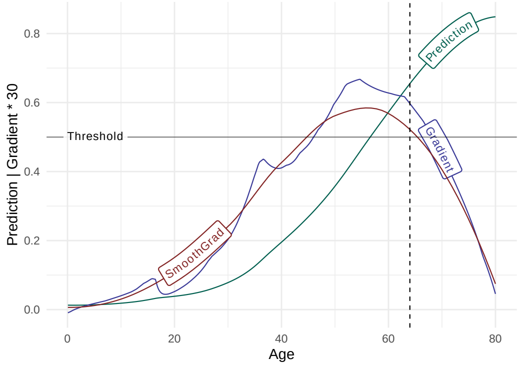
Interpretation:

- Increasing patient's age increases model's prediction for diabetes

Pitfalls:

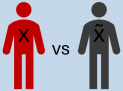
- Show output sensitivity, not attributions!
- Depend on feature scaling

  $\rightarrow$ complicating comparisons for tabular data
- Provide point-specific insights
- **SmoothGrad**
  - Reduces noise ("area-gradient") but depends on smoothing parameters
  - Is Gaussian noise always the best choice?



[1] Simonyan et al. (2014) • [2] Smilkov et al. (2017)

(Fixed-)Reference Methods



"What are the features' contributions for predicting $f(x)$ instead of $f(\tilde{x})$?"

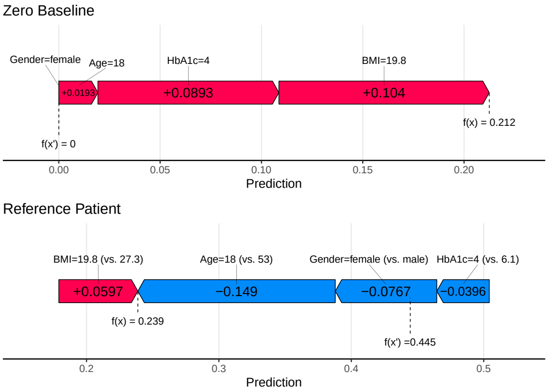
*Grad×Input[3]: $\frac{\partial f(\boldsymbol{x})}{\partial x_i} \cdot x_i$ **IntGrad**[5]: $(x_i - \tilde{x}_i) \int_{\alpha=0}^{1} \frac{\partial f(\tilde{\boldsymbol{x}} + \alpha(\boldsymbol{x} - \tilde{\boldsymbol{x}}))}{\partial x_i} d\alpha$

***LRP**[4] **DeepLIFT**[3]

*Approximation and fixed reference ($\tilde{x} = 0$)

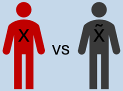
Interpretation:

- Decomposition of $f(x) - f(\tilde{x})$ in feature-wise effects
- Did this feature change (from x to $\tilde{x}$) argue against or for an increase?



[3] Shrikumar et al. (2017) ● [4] Montavon et al. (2019) ● [5] Sundararajan et al. (2017) ● [6] Koenen & Wright (2024)

(Fixed-)Reference Methods



"What are the features' contributions for predicting $f(x)$ instead of $f(\tilde{x})$?"

*Grad×Input[3]: $\frac{\partial f(x)}{\partial x_i} \cdot x_i$ **IntGrad**[5]: $(x_i - \tilde{x}_i) \int_{\alpha=0}^{1} \frac{\partial f(\tilde{x} + \alpha(x - \tilde{x}))}{\partial x_i} d\alpha$

***LRP**[4] **DeepLIFT**[3]

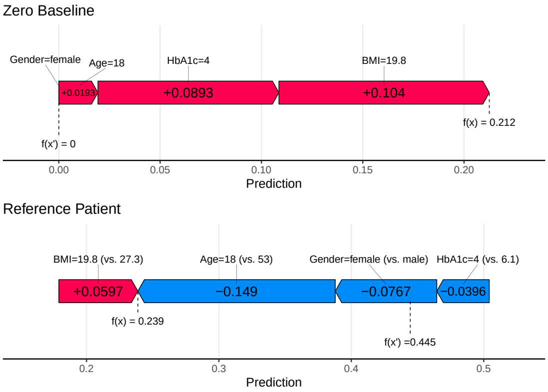
*Approximation and fixed reference ($\tilde{x} = 0$)

Interpretation:

- Decomposition of $f(x) - f(\tilde{x})$ in feature-wise effects
- Did this feature change (from x to $\tilde{x}$) argue against or for an increase?

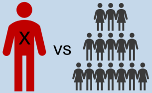
Pitfalls:

- Explanations are highly sensitive to the choice of reference
- Other reference means other question
- Grad$\times$Input and LRP struggle with non-linearity[6]
- Typical references (e.g., 0) often fall outside data distribution



[3] Shrikumar et al. (2017) • [4] Montavon et al. (2019) • [5] Sundararajan et al. (2017) • [6] Koenen & Wright (2024)

Shapley-based Methods



"What are the features' contributions for predicting $f(x)$ compared to $\mathbb{E}_X\left[f(X)\right]$?"

GradSHAP[7]: $\mathbb{E}_{\substack{\tilde{x}\sim X \\ \alpha\sim\mathcal{U}[0,1]}}\left[(x_i-\tilde{x}_i)\frac{\partial f(\tilde{x}+\alpha(x-\tilde{x}))}{\partial x_i}\right]$

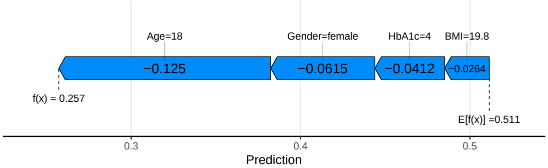
DeepSHAP[8] (rescale or reveal-cancel)

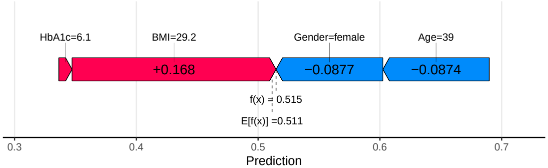
Interpretation:

- Decomposition of $f(x) - \mathbb{E}\left[f(X)\right]$ in feature-wise effects
- What is the marginal contribution of this feature?

  $\rightarrow$ answers a more natural question

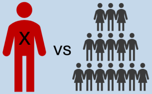
  $\rightarrow$ axiomatic and theoretical foundation





[7] Erion et al. (2021) • [8] Lundberg & Lee (2017)

Shapley-based Methods



"What are the features' contributions for predicting $f(x)$ compared to $\mathbb{E}_X\left[f(X)\right]$?"

GradSHAP[7]: $\mathbb{E}_{\substack{\tilde{x}\sim X \\ \alpha\sim\mathcal{U}[0,1]}}\left[(x_i-\tilde{x}_i)\frac{\partial f(\tilde{x}+\alpha(x-\tilde{x}))}{\partial x_i}\right]$

DeepSHAP[8] (rescale or reveal-cancel)

Interpretation:

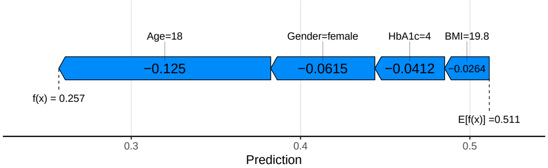
- Decomposition of $f(x) - \mathbb{E}\left[f(X)\right]$ in feature-wise effects
- What is the marginal contribution of this feature?

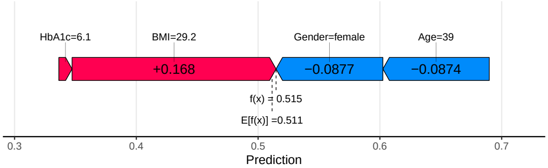
  $\rightarrow$ answers a more natural question

  $\rightarrow$ axiomatic and theoretical foundation

Pitfalls:

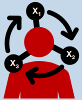
- Still marginal (no conditional Shapley values)
- Higher computational costs
- Requires suitable (reference) dataset





[7] Erion et al. (2021) • [8] Lundberg & Lee (2017)

Interaction-based Methods



"Is there a combined effect of features on $f(x)$?"

Hessian: $\frac{\partial^2 f(x)}{\partial x_i \, \partial x_j}$

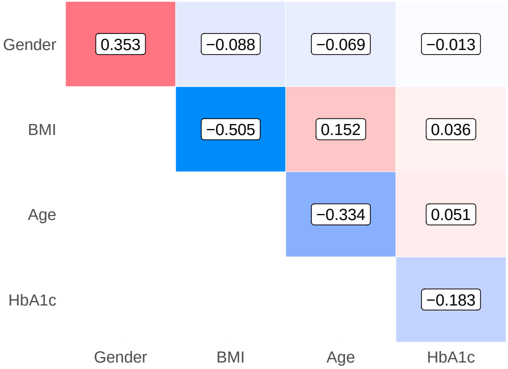
ExpectedHessian[9]

IntHessian[9]: $(x_i - \tilde{x}_i)\,(x_j - \tilde{x}_j) \int_0^1 \int_0^1 \alpha\beta \frac{\partial^2 f(\tilde{x} + \alpha\beta(x - \tilde{x}))}{\partial x_i \, \partial x_j} \, d\alpha \, d\beta$

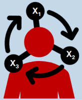
Interpretation:

- Decomposition of $f(x) - f(\tilde{x})$ / $f(x) - \mathbb{E}_X\left[f(X)\right]$ in feature-wise main (diagonal) and two-way interaction effects (**IntHessian**/ **ExpHessian**)
- Reveals local interaction effects and strength (**Hessian**)

f(x) = 0.039 vs. f(x') = 0.569



Interaction-based Methods



"Is there a combined effect of features on $f(x)$?"

Hessian: $\dfrac{\partial^2 f(x)}{\partial x_i\, \partial x_j}$

ExpectedHessian[9]

IntHessian[9]: $(x_i - \tilde{x}_i)\,(x_j - \tilde{x}_j) \displaystyle\int_0^1 \int_0^1 \alpha\beta \frac{\partial^2 f(\tilde{x} + \alpha\beta(x - \tilde{x}))}{\partial x_i\, \partial x_j}\, d\alpha\, d\beta$

Interpretation:

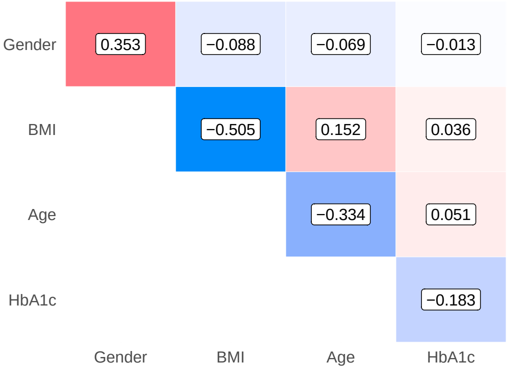
- Decomposition of $f(x) - f(\tilde{x})$ / $f(x) - \mathbb{E}_X\left[f(X)\right]$ in feature-wise main (diagonal) and two-way interaction effects (**IntHessian**/ **ExpHessian**)
- Reveals local interaction effects and strength (**Hessian**)

Pitfalls:

- Higher computational costs
- Not possible for ReLU networks

  $\longrightarrow$ vanishing 2nd derivative
- (similar to the other three groups)

f(x) = 0.039 vs. f(x') = 0.569

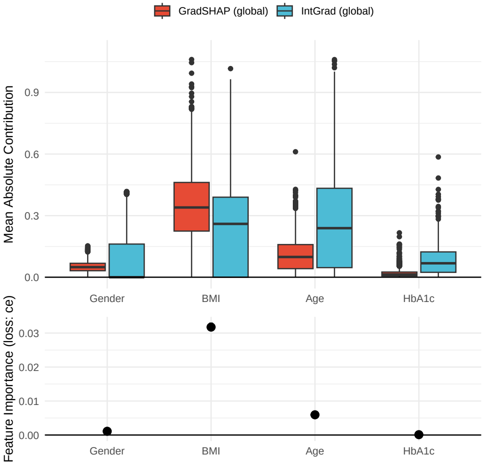


[9] Janizek et al. (2021)

From Local to Global Explanations

- **So far:** Only explanations of individuals
- Aggregating local explanations for global insights (Lundberg et al. (2020))
- Gives relative importance among features
- Can we do so for other feature attribution methods?
  - Prediction-sensitive: Global feature's sensitivity (but depends on scaling)
  - Reference-based: Global effect against single reference
  - Shapley-based: Marginal global effect

$\Rightarrow$ Only explaining the model's prediction (not necessarily aligning with data-based or loss-based importance measures)



Key Takeaways & Future Work

Key Takeaways:

- Feature Attribution Method $\neq$ *feature attribution* $\rightarrow$ answer different questions!
- Choice of method depends on the question: sensitivity, (baseline/marginal) attribution, or interaction
- Each reference value answers another question
- No recommendation for vague approximations (like LRP, Grad$\times$Input)
- Can be aggregated to global importance measures $\rightarrow$ feature selection
- Prediction-based! $\rightarrow$ *"What does the model see for the prediction" – not true to the data*

Future Work:

- Adopting methods for loss-based insights $\rightarrow$ performance attribution
- Extending methods for conditional (not marginal) values

Thank you for your attention!

Slides, references and reproduction material

R Package `innsight`

Contact
Niklas Koenen
Leibniz Institute for Prevention Research
and Epidemiology – BIPS
Achterstraße 30
D-28359 Bremen
koenen@leibniz-bips.de

I have no current or past relationships with commercial entities.