

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

Authors: Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari,
Zhiwei Steven Wu, and Himabindu Lakkaraju
(Harvard, MIT, Drexel, CMU)
Feb, 2022 (170+ citations!)

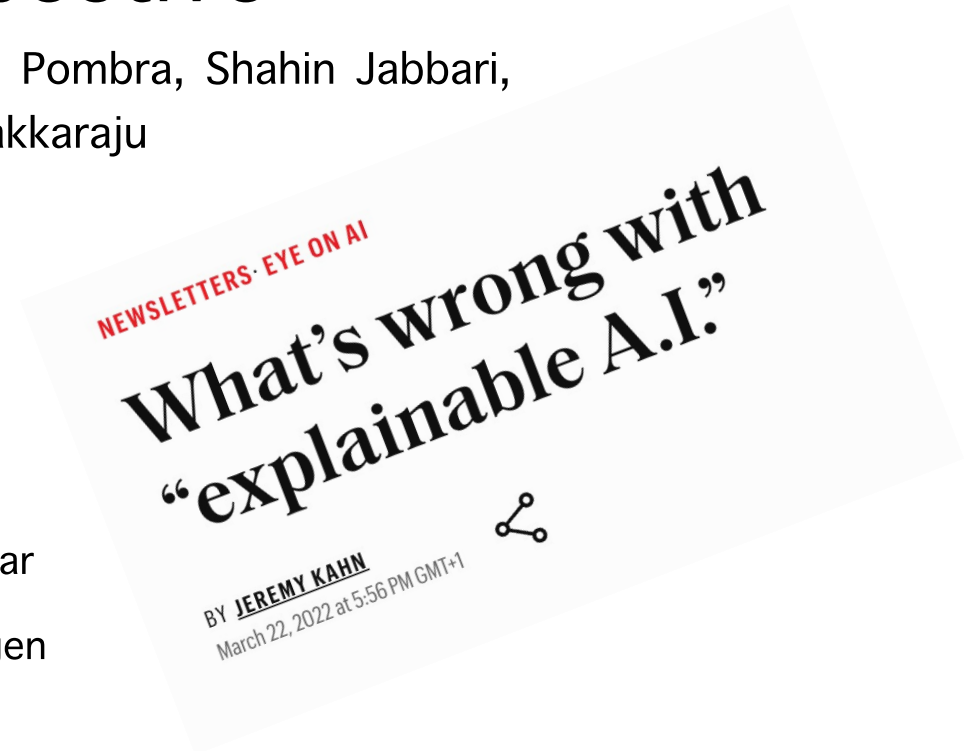
Presenter: Swagatam Haldar
Explainable Machine Learning Seminar
Summer 2024, University of Tübingen



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

Authors: Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari,
Zhiwei Steven Wu, and Himabindu Lakkaraju
(Harvard, MIT, Drexel, CMU)
Feb, 2022 (170+ citations!)

Presenter: Swagatam Haldar
Explainable Machine Learning Seminar
Summer 2024, University of Tübingen



Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

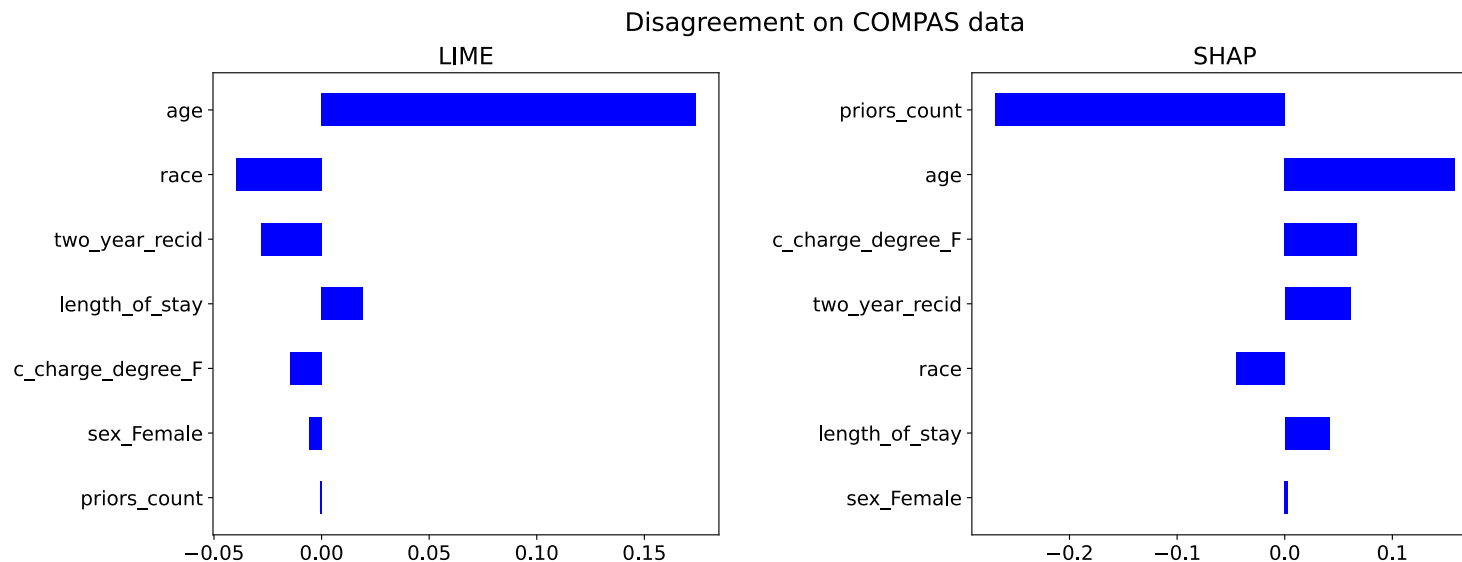
Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

Example:

	age	two_year_recid	priors_count	length_of_stay	c_charge_degree_F	sex_Female	race
1979	45	0	20	46	0	0	1

COMPAS, RF



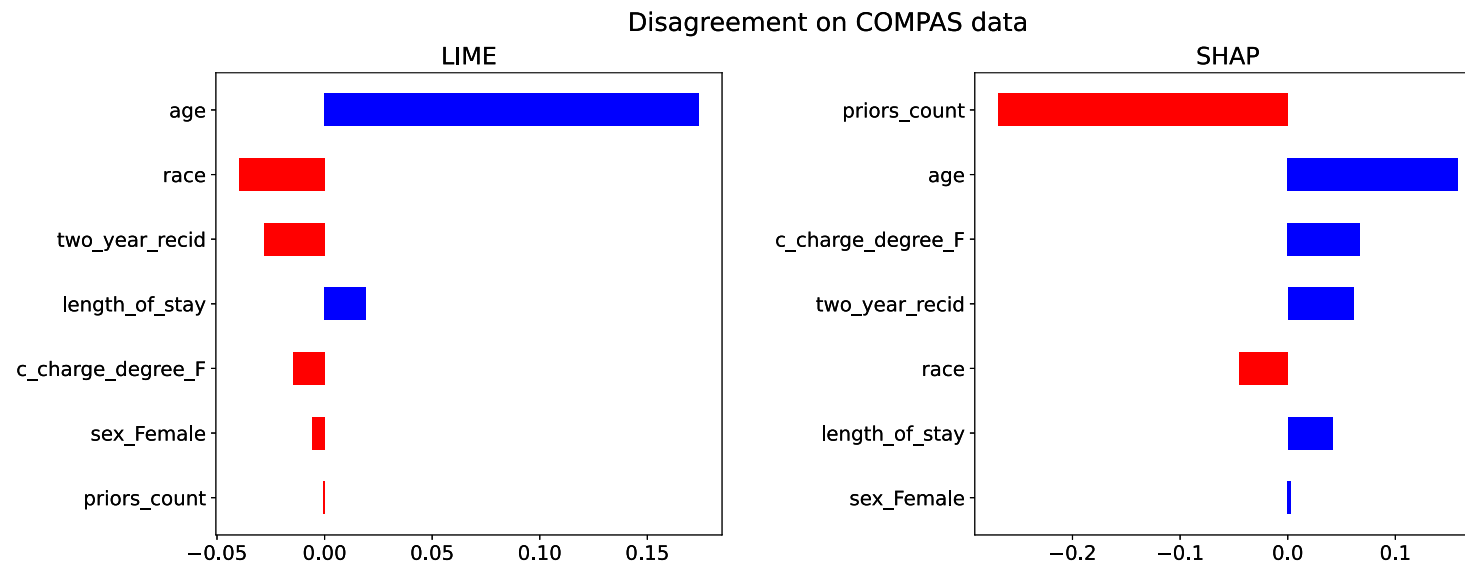
Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

Example:

	age	two_year_recid	priors_count	length_of_stay	c_charge_degree_F	sex_Female	race
1979	45	0	20	46	0	0	1

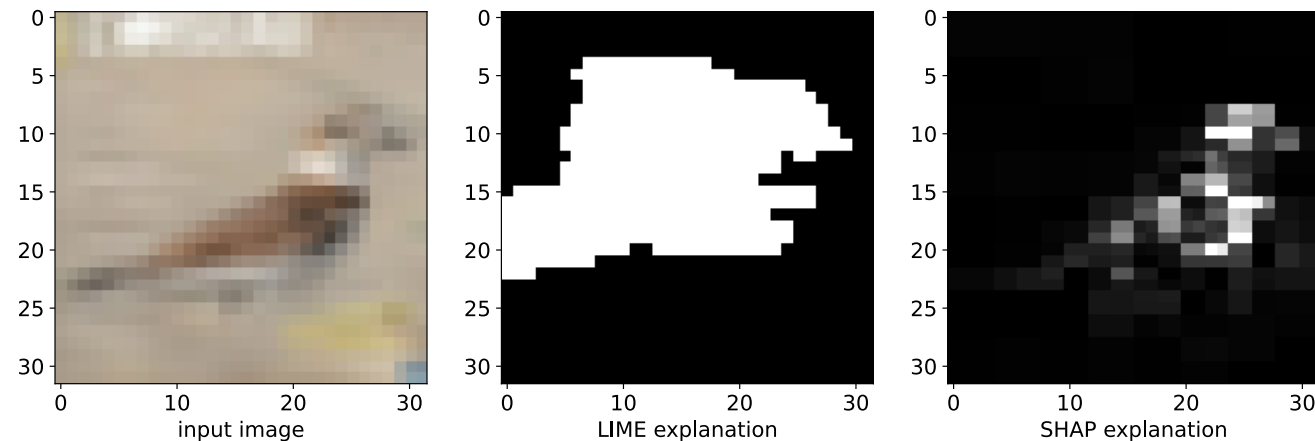
COMPAS, RF



Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

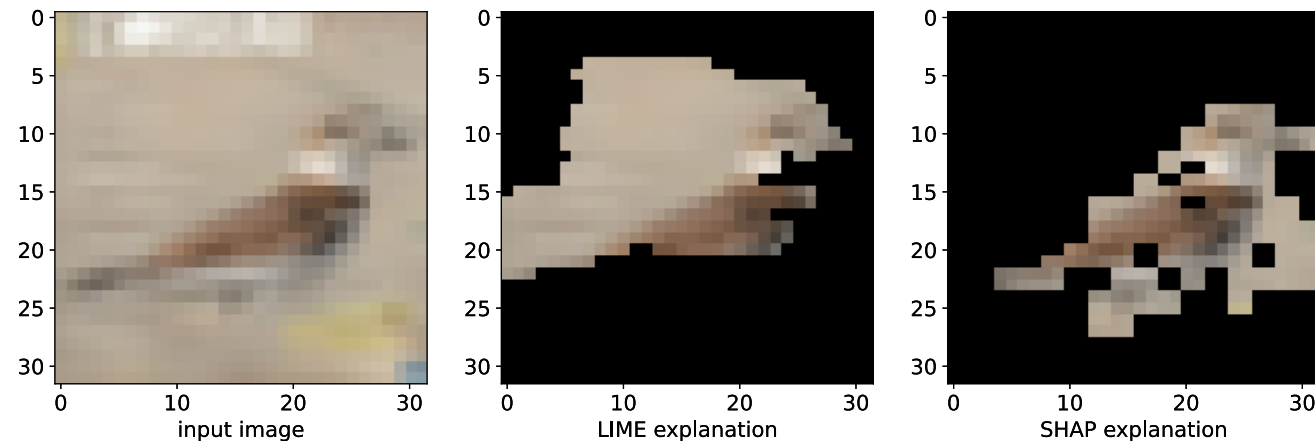
Example: CIFAR10, VGG19



Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

Example: CIFAR10, VGG19



Problem addressed

Motivation: Often explanations for the same prediction disagree across ML algorithms, what can we do about this, if anything?

> *Practitioner's perspective:*

- interviews with 25 data scientists/practitioners to understand what constitutes disagreement between explanations
- Then online study with 25 participants to observe how they resolve that in practice

> Contributions:

- Identify/point out that this problem exists through interviews and online study
- Develop *principled evaluation metrics* to compare explanations quantitatively

But what is an *explanation*?

Most overloaded term currently in use in this seminar! Local vs. global, feature attribution, counterfactuals, training data attributions etc.

Scope of this paper: **local, feature attribution based**

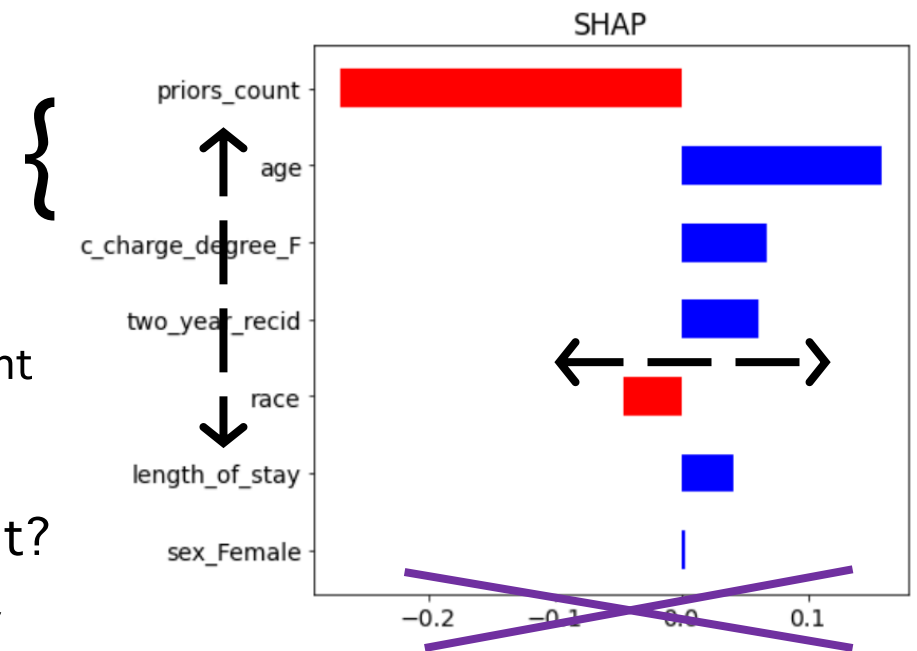
Explanation: an ordered/ranked list of <feature, importance> pairs

- <feature, importance> or <pixel, importance> or <superpixel, importance> or <word/token, importance> pairs
- *Complications might arise in defining “what is a feature” with discretization of continuous features, segmentation, tokenization etc., I will not get into that.*

Insights from interview with practitioners

The paper uses these insights to formalize the notion of explanation disagreement:

- > What people **looked at** to determine disagreement?
 - Top features are different
 - Ordering among top features is different
 - Direction/influence of contribution of top feature is different
 - Relative ordering of meaningful features are different
- > What they **did not look at** to determine disagreement?
 - The actual values of feature importance – they have wildly different scales across explanations



How to compare explanations quantitatively?

The authors propose to use six metrics to measure how similar (or dissimilar) are **feature attribution**-based explanations:

A. (Dis)agreement w.r.t top-k features

- (FA) how many features (fraction) within the topk are same?
- (RA) Is the order of top-k features same?
- (SA) Are the signs of top-k features identical in both explanations?
- (SRA) another metric combining both signs and ranks of top-k features (**strictest**)

B. (Dis)agreement w.r.t user-provided features of interest

Here we can take any set of features

- (RC) rank correlation (viewing explanations as ordered rank lists)
- (PRA) pairwise rank agreement: checks relative ordering of every pair of features



higher
means
more
agreement
 \forall 6 metrics

Metric definition: Signed rank agreement (SRA)

For any two explanations E_a and E_b ,

$$\text{SRA} = \frac{|\bigcup_{s \in S} \{s \mid s \in \text{top_features}(E_a, k) \wedge s \in \text{top_features}(E_b, k) \wedge \text{sign}(E_a, s) = \text{sign}(E_b, s) \wedge \text{rank}(E_a, s) = \text{rank}(E_b, s)\}|}{k}$$

For a given set of features $F = \{f_1, f_2, \dots, f_n\}$, and r_s is the Spearman rank correlation coefficient

$$\text{Rank correlation (RC)} = r_s(\text{Ranking}(E_a, F), \text{Ranking}(E_b, F))$$

How to compare explanations quantitatively?

The authors propose to use six metrics to measure how similar (or dissimilar) are **feature attribution**-based explanations:


A. (Dis)agreement w.r.t top-k features

- (FA) how many features (fraction) within the topk are same?
- (RA) Is the order of top-k features same?
- (SA) Are the signs of top-k features identical in both explanations?
- (SRA) another metric combining both signs and ranks of top-k features (**strictest**)

B. (Dis)agreement w.r.t user-provided features of interest

Here we can take any set of features

- (RC) rank correlation (viewing explanations as ordered rank lists)
- (PRA) pairwise rank agreement: checks relative ordering of every pair of features



lower
means
more
disagreement
 \forall 6 metrics

Empirical verification of disagreement: Ingredients

Datasets:

- > Tabular (**COMPAS**, German credit)
- > Image (Imagenet-1000)
- > Text (news category prediction)

Models:

- > Tabular (LR, MLP, RF, GBDT)
- > Image (ResNet18)
- > Text (LSTM based classifier)

- All models were trained to a reasonable accuracy
- For RF & GBDT, they only used LIME and KernelSHAP.
- **Metrics** as defined in last slide.

Explanation methods:

- > Perturbation based:
LIME, KernelSHAP
- > Gradient based:
Vanilla gradient (VG),
Gradient * Input (G*I),
Integrated Gradients (IG),
SmoothGrad (SG)

Result computation

1. Take a test data point
2. Compute explanation E_a and E_b for it
3. Calculate $\text{metric}(E_a, E_b)$
4. Average over all test points

		Rank correlation				
LIME	1.000	0.800	0.022	0.624	0.630	0.016
Kernel SHAP	0.800	1.000	0.189	0.440	0.450	0.184
Grad	0.022	0.189	1.000	-0.001	-0.018	0.980
Grad* Input	0.624	0.440	-0.001	1.000	0.988	0.005
IntGrad	0.630	0.450	-0.018	0.988	1.000	-0.012
Smooth GRAD	0.016	0.184	0.980	0.005	-0.012	1.000
	LIME	Kernel SHAP	Grad	Grad* Input	IntGrad	Smooth GRAD

Main results: tabular modality

COMPAS,
NN

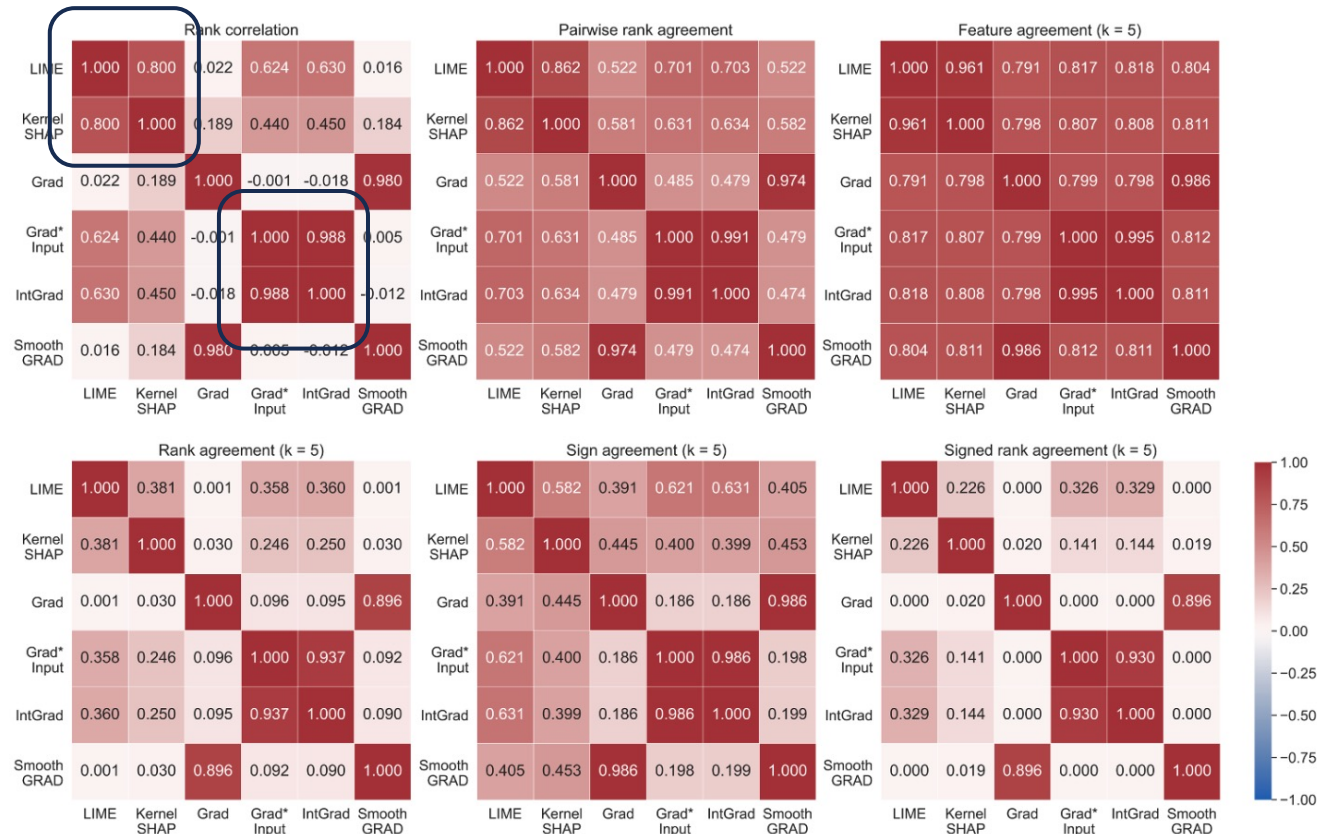


Figure 1: Disagreement between explanation methods for neural network model trained on COMPAS dataset measured by six metrics: rank correlation and pairwise rank agreement across **all features**, and feature, rank, sign, and signed rank agreement across top $k = 5$ features. **Heatmaps show the average metric value over test set data points for each pair of explanation methods, with lighter colors indicating stronger disagreement.** Across all six heatmaps, the standard error ranges between 0 and 0.009.

- > high FA (k=5) for all explanations
(caveat: only 7 features!)
- > RC: higher for (VG, SG), and (G*I, IG) pairs
- > More features lead to more disagreement
- > More complex models (hard to approximate) lead to less agreement

Main results: text modality

Text clf,
LSTM

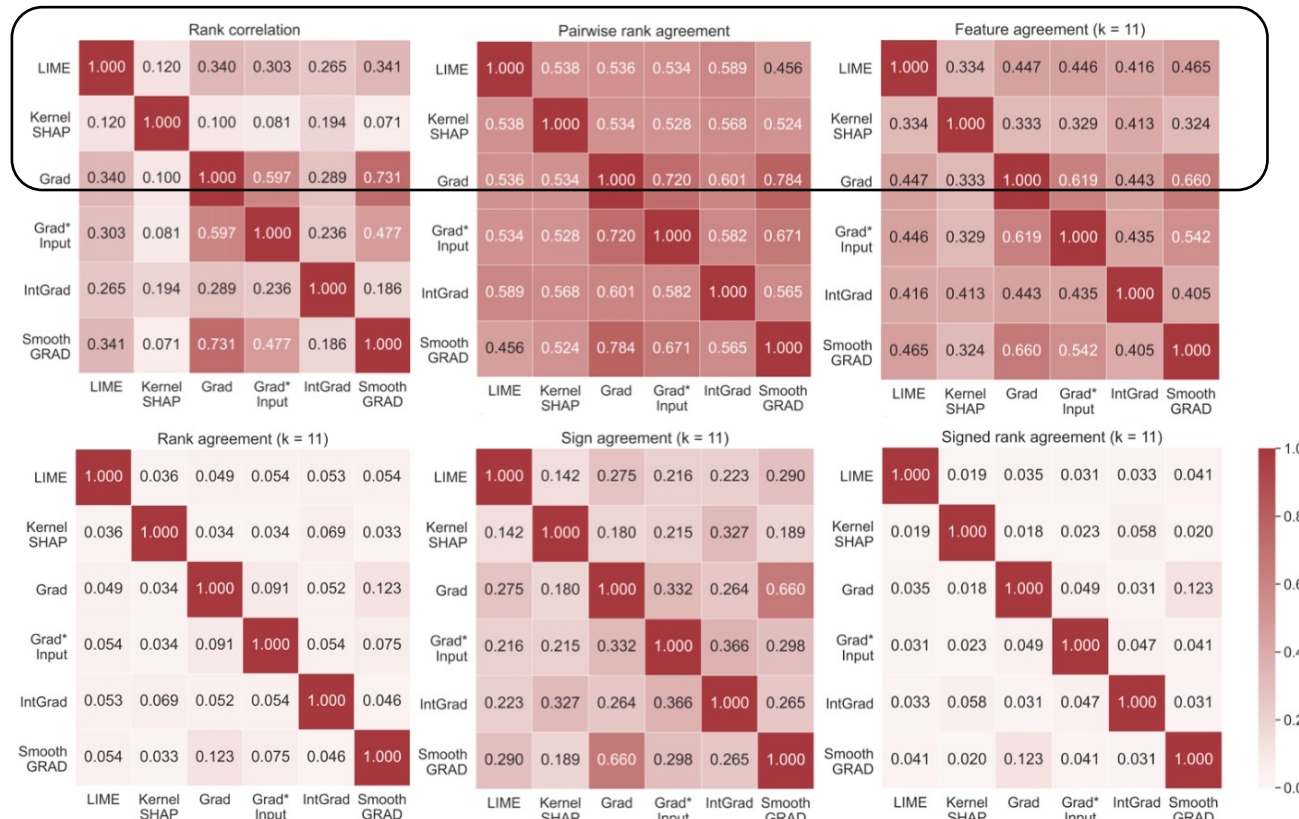


Figure 4: Disagreement between explanation methods for the LSTM model trained on the AG_News dataset using $k = 11$ features for metrics operating on top- k features, and all features for other metrics. Each heatmap shows the metric value averaged over test data for each pair of explanation methods. Lighter colors indicate more disagreement. Standard error ranges from 0.0 to 0.0025 for all six metrics.

> more features so all plots look lighter!

> very low agreement between methods here

> LIME agrees more with other explanations than KernelSHAP (look at first 2 rows top)

Main results: image modality

> unlike tabular & text, here **more agreement** between LIME and SHAP (probably because they used the same superpixels, too coarse)

superpixel level

Metrics	ResNet-18
Rank correlation	0.8977
Pairwise rank agreement	0.9302
Feature agreement	0.9535
Rank agreement	0.8478
Sign agreement	0.9218
Signed rank agreement	0.8193

Table 2: Disagreement on ImageNet between LIME and KernelSHAP

> Gradient based explanations show high disagreement (likely because too granular)

pixel level explanations

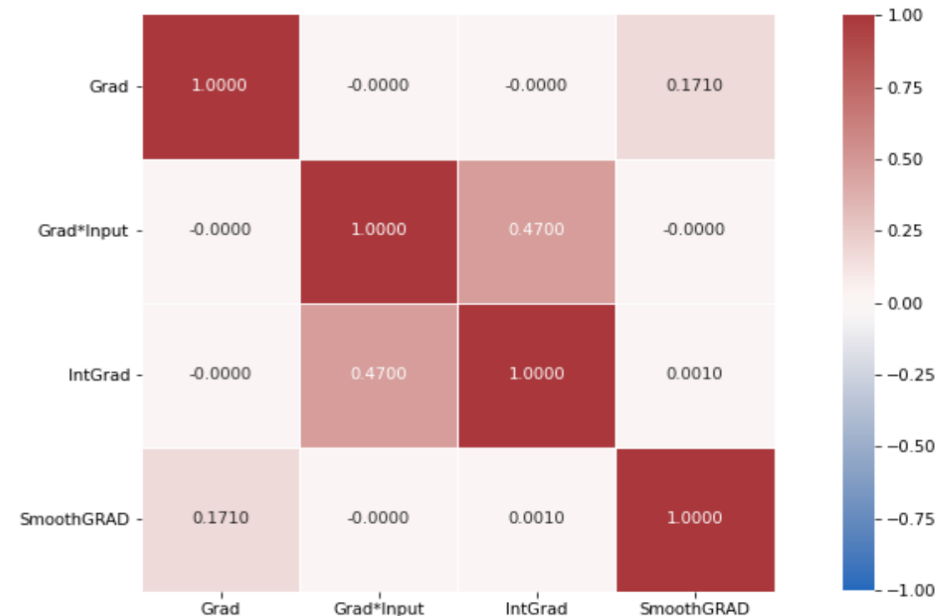
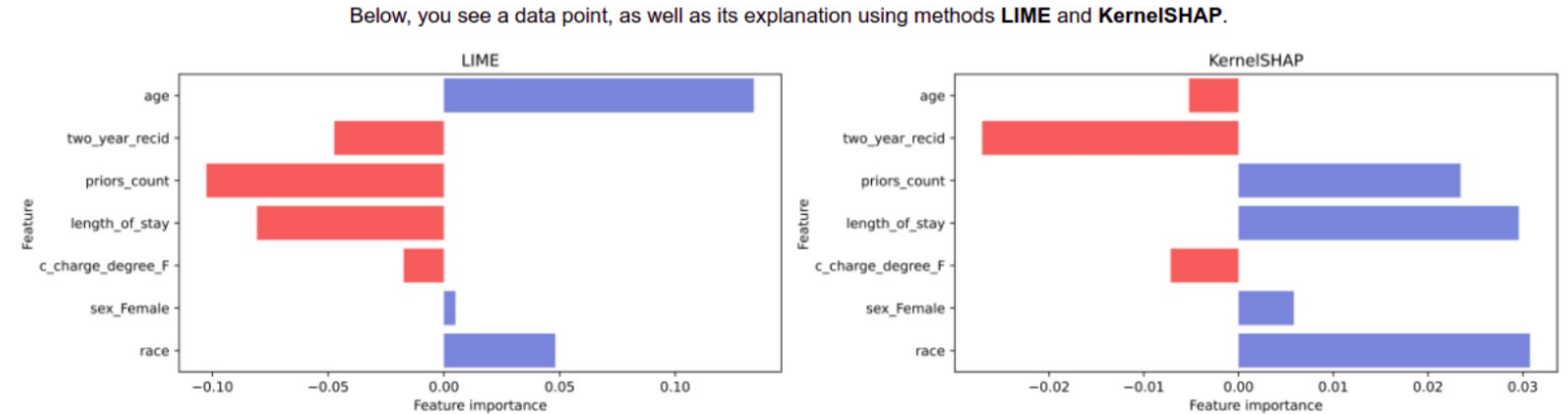


Figure 25: Rank correlation for explanations computed at pixel level by gradient-based explanation methods

Qualitative online study

Goal of the study:
To understand
how people
resolve or choose
between two
disagreeing
explanations

One example
prompt from the
study →



As a reminder, the 7 features of the COMPAS dataset are **age**, **two_year_recid** (whether the defendant recidivated after 2 years of the original crime), **priors_count** (number of prior crimes committed), **length_of_stay** (length the defendant stayed in jail), **c_charge_degree** (whether the previous charge was a Misdemeanor or Felony), **sex**, and **race**

To what extent do you think the two explanations shown above agree or disagree with each other?

☐ Completely agree ☒ Mostly agree ☐ Mostly disagree ☐ Completely disagree

Please explain why you chose the above answer.

Since you believe that the above explanations disagree (to some extent), which explanation would you rely on?

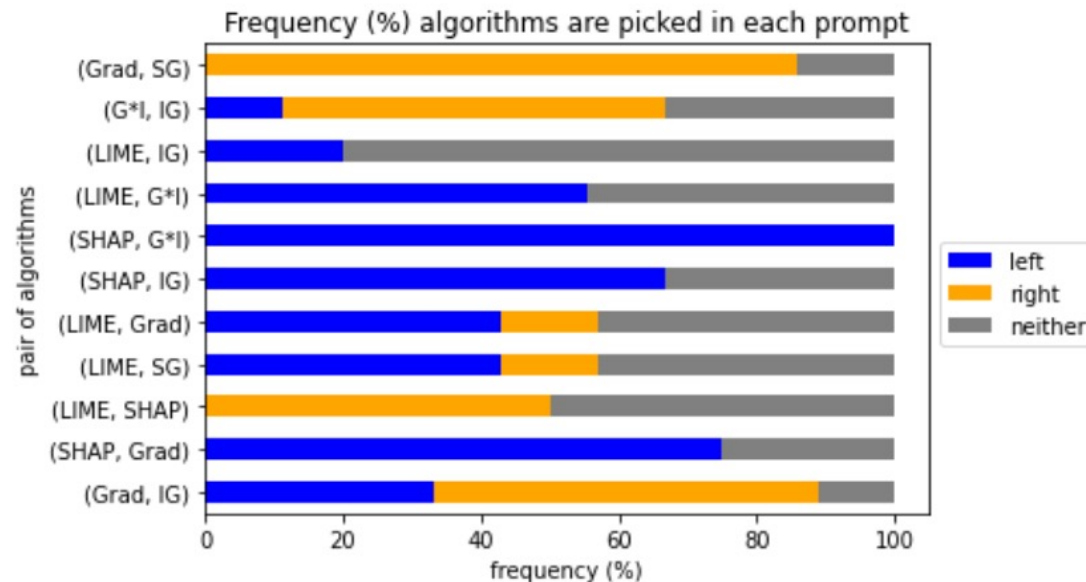
☐ LIME explanation ☒ KernelSHAP explanation ☐ It depends

Please explain why you chose the above answer.

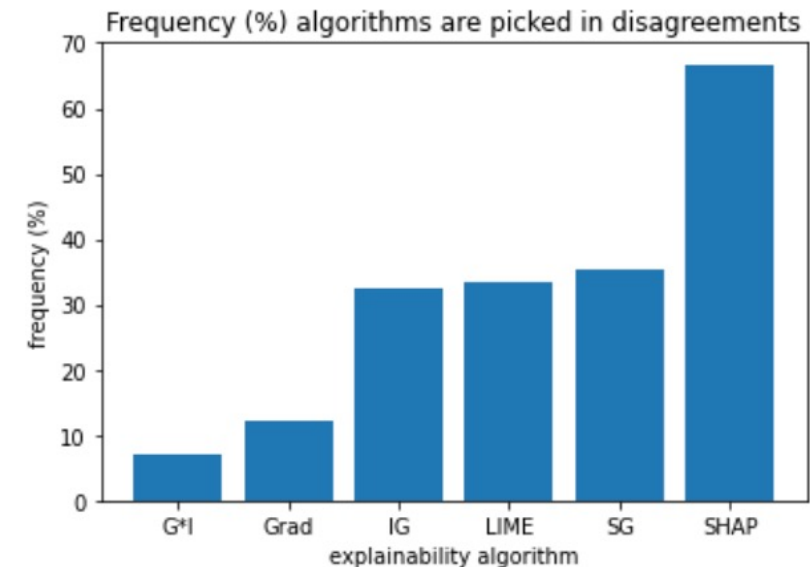
Insights from the online study

- > Majority of respondents observed and acknowledged disagreement
- > Often certain explanations are favoured over others:
 - SHAP is the most popular while (Gradient*Input) is least chosen
 - SmoothGrad is often favoured over VanillaGrad (probably because of less noise)
- > How do they resolve disagreements in practice?
 - Rely on heuristics such as “lime/shap is better for tabular data”
 - Choosing the explanation that matches their *intuition* better
 - Using other metrics (fidelity, stability) that quantify the “goodness of fit” of the explanation to the model
 - Other things also matter like ease of implementation, groundedness of theory, recency of publication, documentation of packages

Insights from the online study



(a) The frequency with which each of the explanations in a pair is selected upon disagreement. The blue, gold, and grey bars show the percentage of participants (X axis) that picked the left, right, and neither algorithm when presented with the pair of algorithms shown on the Y axis.



(b) The frequency with which each of the explanations was chosen when there is a disagreement. X axis indicates the explainability algorithms and Y axis indicates the frequency.

Thoughts

- > This paper is more of a **practitioners'** perspective on explanation disagreement which motivates the metrics
 - Interview to understand what exactly is disagreement
 - Online study to find out how people resolve such disagreements
- > They do not investigate further on the reasons behind why explanations disagree
- > Their studies also find out that people mostly rely on heuristics, intuitions when they use explainability methods
- > No mention of other explanation methods
- > This paper raises more questions than answers it provides

Thank you for your attention!
Questions?