A Experimental Setup

A.1 Black Box Models: Training and Performance

For tabular data, we train four models: a logistic regression model, a gradient-boosted tree model (50 estimators), a random forest model (50 estimators), and a densely-connected feed-forward neural network (with 3 hidden layers with relu activation consisting of 50, 100, and 50 neurons, respectively). For the COMPAS dataset, we train the four models based on a 80%-20% train-test split of the dataset, using features to predict COMPAS risk score group. The test accuracies of the four models are 0.75, 0.79, 0.79, and 0.73, respectively. For the German credit dataset, we train the same four models based on a 80%-20% train-test split of the dataset, using features to predict credit risk group. The test accuracies of the four models are 0.65, 0.70, 0.73, and 0.64, respectively.

For text data, we trained a widely-used LSTM-based text classifier, based on 120,000 training samples and 7,600 test samples, to predict the news category of the article from which a sentence was obtained. The model performs with 90.67% accuracy. The architecture comprises of an embedding layer of dimension 300, followed by an LSTM layer of hidden size 256 connected to a four-dimensional output layer.

For image data, we use the pre-trained ResNet-18 model [58] and analyze explanations generated for predictions made to classify images to one of the 1000 classes. This model performs 69.758 % and 89.078 % on Accuracy@1 and Accuracy@5 metrics*, respectively.

A.2 Explanation Methods

For tabular data, the perturbation-based explanation methods (LIME and KernelSHAP) were applied to explain all four models while the gradient-based explanation methods (Vanilla Gradients, Integrated Gradients, Gradient*Input, and SmoothGRAD) were applied to explain the logistic regression and neural network models to explain samples from the test set (1,198 samples for the COMPAS dataset and 200 samples for the German Credit dataset). Because gradients are not computed for tree-based models, the gradient-based explanation methods were not applied to the random forest and gradient-boosted tree models. When applying explanation methods with a sample size hyperparameter (LIME, KernelSHAP, Integrated Gradients, SmoothGRAD), we performed a convergence check and selected the sample size at which an increase in the number of samples does not significantly change the explanations. Change in explanations at the current versus previous sample size is measured by L2 distance of feature attributions, L2 distance of the ranks of all features, and the six metrics. For both COMPAS and German Credit datasets, we used 2,000 samples for relevant methods (i.e. LIME, KernelSHAP, SmoothGRAD, and Integrated Gradients).

For text data, we applied all six explanation methods on the LSTM-based classifier to explain predictions for 7,600 samples in the test set. For LIME and KernelSHAP, we follow the convergence analysis described above and find that attributions do not change significantly beyond 500 perturbations; hence, we use 500 perturbations for LIME and KernelSHAP. Integrated Gradients explanations were generated using 500 steps which is higher than the recommended number of steps mentioned in [5]. SmoothGRAD explanations were generated using 500 samples to get the most confident attribution which is significantly higher the recommended number of 50 samples [7].

For image data, we applied all six explanation methods on the ResNet-18 model [58] to explain predictions for the PASCAL VOC 2012 test set of 1,449 samples. Integrated Gradients explanations were generated using 400 steps, significantly higher than the recommendation of 300 [5], to obtain a stable and confident attribution map. Similarly, SmoothGRAD explanations were generated using a sample size of 200 which is also higher than the recommended sample size of 50 [7]. For LIME and KernelSHAP, we chose 100 perturbations to train the surrogate model as we did not notice any significant changes in attributions beyond 50 perturbations. KernelSHAP and LIME were used to compute attributions of super-pixels annotated in PASCAL VOC 2012 segmentation maps. Due to a larger feature space in images compared to the previous tabular and text datasets, disagreement metrics based on top-k features may not provide a clear picture. Hence, we use Rank Correlation and cosine distance between attribution maps generated by a pair of explanation methods as the

*`https://pytorch.org/vision/stable/models.html`

disagreement metric. Higher cosine distance between attribution maps indicate larger disagreement between explanation methods.

B Results from Empirical Analysis of Disagreement Problem

B.1 Tabular Data: COMPAS and German Credit Datasets

The tabular datasets consist of the COMPAS and German Credit datasets. The full set of figures showing explanation disagreement for the two datasets, over four models, measured by six metrics, displayed in two formats (metric mean in heatmap and metric distribution in boxplot) for varying values of top-k features can be found in the code repository accompanying this paper. Here, we elaborate on trends of explanation disagreement at the level of metrics, models, and datasets.

At the level of metrics, explanations show stronger disagreement when measured by stricter metrics (such as signed rank agreement) than by less strict metrics (such as feature agreement and sign agreement). Generally, as the number of top-k features increases, metrics show stronger disagreement between explanations. The exception is the feature agreement metric which, by definition equals one (indicating perfect agreement) when the k is the total number of features. For both datasets, the majority of models and explanation pairs have several points for which the rank correlation metric is near zero or even negative, indicating disagreeing, and even opposing, explanations.

At the level of models, explanations for more complex models (such as the neural network) tend to show similar or stronger disagreement than explanations for less complex models (such as the logistic regression model). As discussed in the main section of the paper, as the complexity of the black box model increases, it may be more difficult to explain the model's decision-making process and more difficult to disentangle the contribution of each feature to the model's prediction. Thus, the higher the model complexity, the more difficult it may be for different explanation methods to generate the true explanation and the more likely it may be for different explanation methods to generate differently false explanations, leading to stronger disagreement among explanation methods.

At the level of datasets, explanations for the German Credit dataset showed similar or stronger levels of disagreement than those for the COMPAS dataset. As discussed in the main paper, one possible reason is that the German Credit dataset has more features than the COMPAS dataset, resulting in a larger number of possible ranking and sign combinations assigned by a given explanation method and making it less likely for two explanation methods to produce consistent explanations.

B.2 Text Data: AG_News Dataset

In this section, we analyse the trends in pair-wise disagreement w.r.t the feature agreement metric with increasing number of features considered for measuring disagreement($k = [3, 7, 11]$). We observe that agreement between gradient based methods is significantly more compared to that with perturbation-based explanation method, which is more prominant for larger values of k.

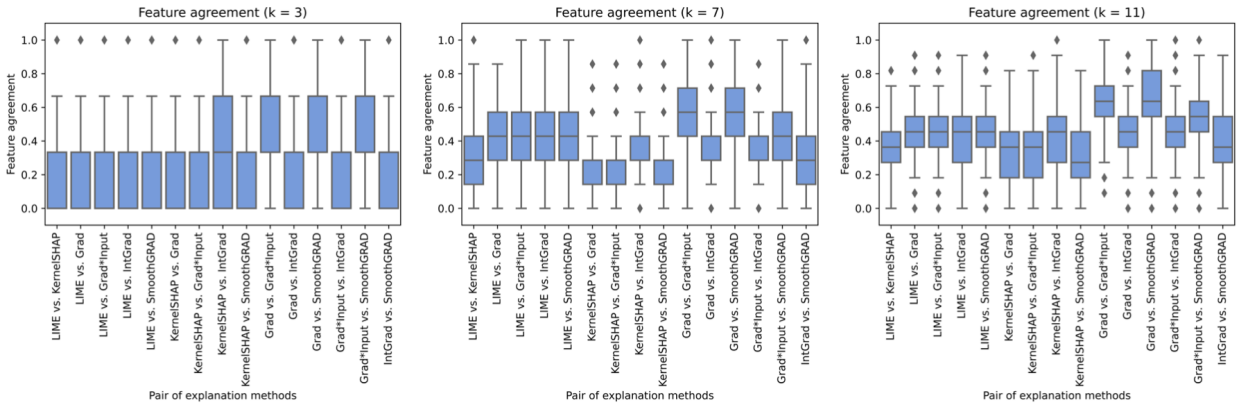


Figure 7: Box plot for feature agreement on AG_News dataset for k = [3,7,11]

B.3 Image Data: ImageNet Dataset

In Table 2, we analyse disagreement between perturbation-based explanation methods when feature attributions are computed at super-pixel level, and observe a significantly high agreement in terms of all the proposed disagreement metrics. However, there is little to no agreement when its computed at pixel level in terms of the rank correlation between explanations, shown in Figure 8.

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

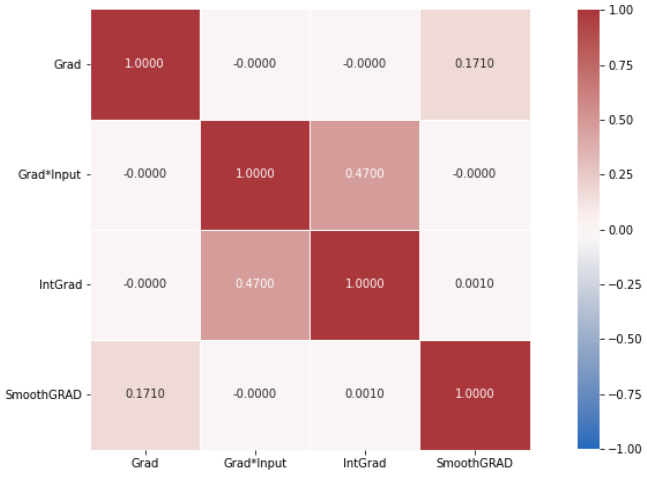
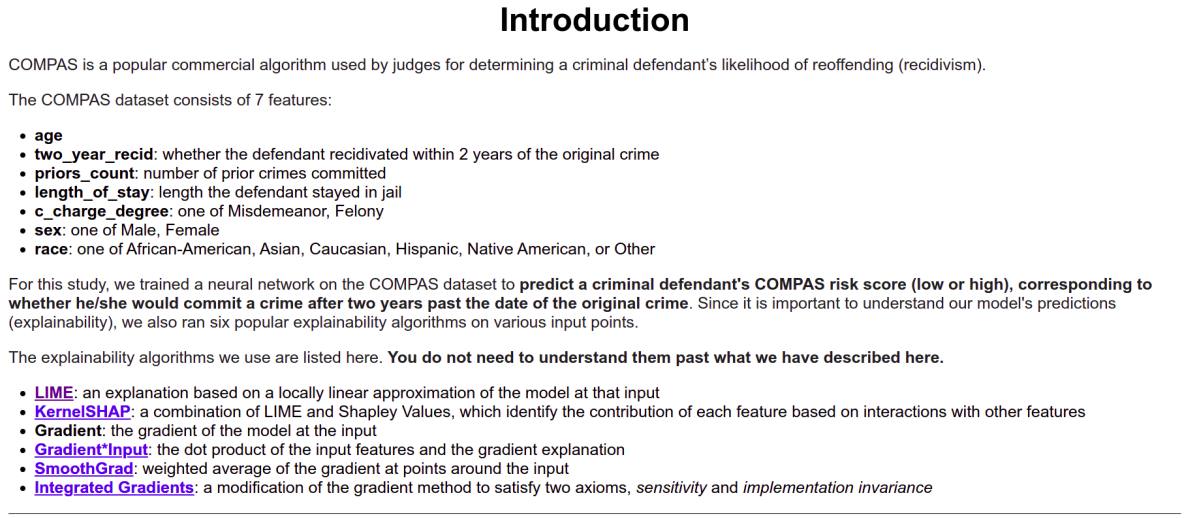


Figure 8: Rank correlation for explanations computed at pixel level by gradient-based explanation methods

C Omitted Details from Section 5

C.1 Screenshots of UI

In Figures 9 and 10, we present screenshots of the UI that participants are presented with before beginning the study. The purpose of this introduction page is to familiarize the participants with the COMPAS prediction setting, the six explainability methods we use, and the explainability plots we show in each of the prompts.



Introduction

COMPAS is a popular commercial algorithm used by judges for determining a criminal defendant's likelihood of reoffending (recidivism).

The COMPAS dataset consists of 7 features:

- **age**
- **two_year_recid**: whether the defendant recidivated within 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**: one of Misdemeanor, Felony
- **sex**: one of Male, Female
- **race**: one of African-American, Asian, Caucasian, Hispanic, Native American, or Other

For this study, we trained a neural network on the COMPAS dataset to **predict a criminal defendant's COMPAS risk score (low or high), corresponding to whether he/she would commit a crime after two years past the date of the original crime**. Since it is important to understand our model's predictions (explainability), we also ran six popular explainability algorithms on various input points.

The explainability algorithms we use are listed here. **You do not need to understand them past what we have described here.**

- **LIME**: an explanation based on a locally linear approximation of the model at that input
- **KernelSHAP**: a combination of LIME and Shapley Values, which identify the contribution of each feature based on interactions with other features
- **Gradient**: the gradient of the model at the input
- **Gradient*Input**: the dot product of the input features and the gradient explanation
- **SmoothGrad**: weighted average of the gradient at points around the input
- **Integrated Gradients**: a modification of the gradient method to satisfy two axioms, *sensitivity* and *implementation invariance*

Figure 9: This is a screenshot of the first half of the introductory page, describing our COMPAS risk score prediction setting and briefly summarizing the six explainability algorithms used (with links to their corresponding papers for the interested participant).

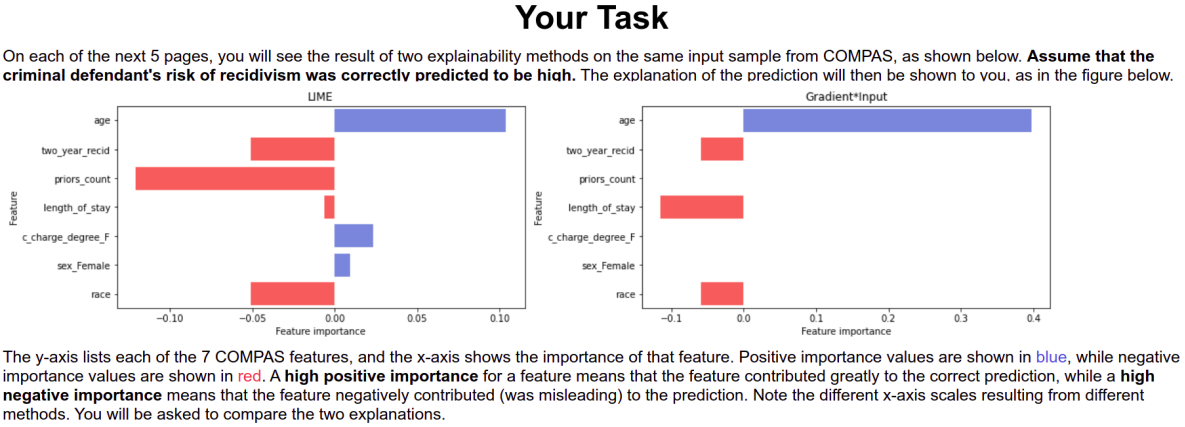


Figure 10: This is a screenshot of the second half of the introductory page, describing the concrete task and an explanation of what is shown in the explainability plots.

C.2 Prompts Used

In this section, we share the 15 prompts that we showed users. Each prompt highlights a pair of different explainability algorithms on a COMPAS data point. For each pair, we chose the data point from the entire

COMPAS set that maximized the rank correlation between the explanations.

C.3 User Study Questions

In each of the five prompts, we asked participants the following questions, which we refer to as *Set 1.* Questions 3-4 were only shown if the user selected *Mostly agree*, *Mostly disagree*, or *Completely disagree* to Question (1).

1. To what extent do you think the two explanations shown above agree or disagree with each other? (choice between *Completely agree, Mostly agree, Mostly disagree, Completely disagree*)
2. Please explain why you chose the above answer.
3. Since you believe that the above explanations disagree (to some extent), which explanation would you rely on? (choice between *Algorithm 1 explanation, Algorithm 2 explanation, It depends*)
4. Please explain why you chose the above answer.

After answering all five prompts, the user was then asked the following set of questions, which we refer to as *Set 2.* Questions 4-9 were only shown if the user selected *Yes* to Question 3.

1. (Optional) What is your name?
2. What is your occupation? (eg: PhD student, software engineer, etc.)
3. Have you used explainability methods in your work before? (*Yes/No*)
4. What do you use explainability methods for?
5. Which data modalities do you run explainability algorithms on in your day to day workflow? (eg: tabular data, images, language, audio, etc.)
6. Which explainability methods do you use in your day to day workflow? (eg: LIME, KernelSHAP, SmoothGrad, etc.)
7. Which methods do you prefer, and why?
8. Do you observe disagreements between explanations output by state of the art methods in your day to day workflow?
9. How do you resolve such disagreements in your day to day workflow?

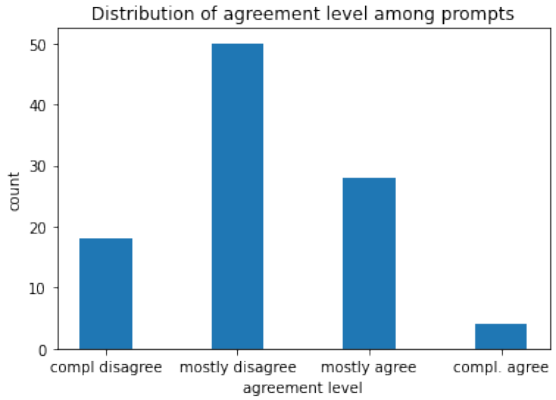
C.4 Further analysis of overall agreement levels

In this section, we present further plots analyzing responses to questions (1) in Set 1. As shown in Figure 12a, only 32% of responses were *Mostly Agree/Completely Agree* and 68% were *Mostly Disagree/Completely Disagree*, indicating that participants experienced the disagreement problem. We also grouped the responses by prompt, shown in Figure 12b, highlighting that different pairs of algorithms can have different levels of disagreement. We removed prompts with less than 4 total responses. We see that there are varying levels of disagreements among prompts. For example, all participants who were shown the Gradient vs. SmoothGrad prompt believed they agreed to some extent, while all participants who were shown the Gradient vs. Integrated Gradients prompt believed they disagreed to some extent.

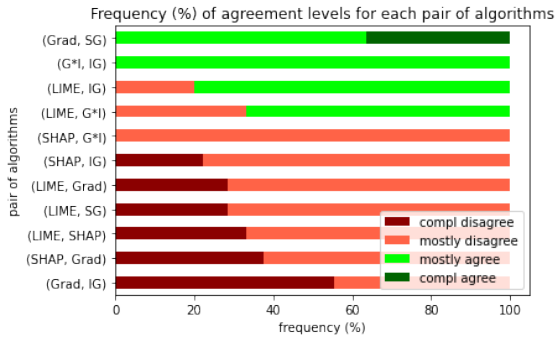
C.5 Further analysis of reasons participants chose specific algorithms

In this section, we analyze the responses to Set 1, Question (3) in Section C.3. We saw, in 5.2.2, that algorithms such as KernelSHAP were favored over other algorithms. In Table 3, we list the top reasons the four most frequently chosen algorithms were preferred, showcasing direct quotes from participants.

Figure 11: Images showing the 15 prompts we used. Each prompt shows the explanation of the same input point with two different interpretability algorithms.



(a) This figure shows the distribution of responses in aggregation over all prompts. The x-axis shows the four possible responses, and the y-axis shows the number of times that response was chosen. Observe that in 68% of cases, participants indicated that the prompts mostly or completely disagreed.



(b) This figure shows the distribution of responses, sorted by prompt. The y-axis shows the pair of explainability algorithms shown in the prompt, and the x-axis shows the frequency that each response was chosen.

Figure 12: These figures show the distribution of answers to Question (1) in Set (1) from Section C.4 in aggregation over all participants.

C.6 Analysis of reasons participants chose neither algorithm

In this section, we analyze the responses to Set 1, Question (4) in Section C.3, focusing on when participants selected *"It depends"* in Question (3), which was chosen in 38% of cases. Again, we present an overarching summary of the reasons participants made this decision in Table 4.

C.7 Further analysis of concluding questionnaire

In this section, we extend the analysis presented in 5.2.3, analyzing the responses to questions in Set 2 of Section C.3. As stated in 5.2.3, we received a total of 20 positive responses to Question (3), but one declined to answer Questions (4) through (9). Therefore, we analyze the remaining 19 responses.

In Question (4), we found that study participants use explainability methods for a variety of reasons such as understanding models, debugging models, help explain models to clients, research. In Question (5), we found that 16 of 19 participants employed explanations for tabular data, 6 of 19 participants for text and language data, 11 of 19 participants for image data, and 1 of 19 for audio data. In Question (6), we found that 14 of 19 participants used LIME, 14 of 19 participants used SHAP, and 13 of 19 participants used some sort of gradient-based methods. Participants also indicated using methods like GradCAM, dimensionality reduction, MAPLE, and rule-based methods. In Question (7), 9 of 19 participants stated that they preferred both LIME and SHAP, with another 3 of 19 participants stating LIME only. We showcase some intriguing answers from Question (7) below:

- "LIME and SHAP seem to be the most universally applicable and I can understand."
- "Methods with underlying theoretical justifications such as KernelSHAP and Integrated Gradients"
- "lime and shap ... [easy to implement] and can work with black box"
- "shap and lime because ... [they are] easy to understand and have standard implementations"
- "LIME, because everything else isn't necessarily capturing what I actually want to know about the local behavior"

Algorithm	Reasons that algorithm was chosen in disagreement
KernelSHAP	• [36%] SHAP is better for tabular data (*"SHAP is more commonly used [than Gradient] for tabular data"*) • [25%] SHAP is more familiar (*"More information present + more familiarity"*) • [14%] SHAP is a better algorithm overall (*"SHAP seems more methodical than LIME"*, *"SHAP is a more rigorous approach [than LIME] in theory"*)
SmoothGrad	• [33%] SmoothGrad paper is newer or better (*"SmoothGrad is apparently more robust"*, *"SmoothGrad is often considered improved verison of grad"*) • [58%] Reasons based on the explainability map shown (*"directionality of the attributions ... [agree] with intuition"*, *"gradient has unstability problems [, so] smoothgrad"*)
LIME	• [54%] LIME is better for tabular data (*"I use LIME for structured data."*) • [15%] LIME is more familiar/easier to interpret (*"I am more familiar with LIME"*, *"LIME is easy to interpret"*)
Integrated Gradients	• [86%] Integrated Gradients paper is better (*"IG came after gradients and paper shows improvements"*, *"integrated gradients paper showed improvements [over Gradient × Input]"*

Table 3: Reasons participants chose the top four most favored explainability algorithms (KernelSHAP, SmoothGrad, LIME, and Integrated Gradients) over others when explanations disagreed.

Rationale	Representative Quote
1. Need more information	• *"need to see the final prediction of the model and the feature values"*
2. Pick neither explanation	• *"No compelling reason to choose one over the other. Both don't align with intuition."*
3. Unsure/Don't know	• *"I'm not sure which of the two methods is more trustworthy"*
4. Would consult an expert	• *"I would ask a domain expert for his/her opinion"*
5. Combine explanations	• *"I would combine both – note that age might be doing weird things, but that length of stay and race both contribute to a negative prediction"*
6. Depends on use case	• *"The two methods have different interpretations - it depends on if I'm more interested in comparing my explanation to some baseline individual state versus just interested in understanding the immediate local behavior"*

Table 4: Reasons people answered *"It depends"* after being asked to choose between disagreements

Finally, we provide additional quotes highlighting the responses to Questions (8) and (9), which were briefly analyzed in Section 5.2.3. These are shown in Table 5.

Category of Response	Samples Quotes
1. Make arbitrary decisions (50%).	• *"Such disagreements are resolved by data scientists picking their favorite algorithm"* • *"I try to use rules of thumb based on results in research papers and/or easy to understand outputs."* • *"I favor lime and shap because there is well documented packages on github"*
2. Unsure/Don't know/Don't resolve (36%)	• *"there is no clear answer to me. I hope research community can provide some guidance"* • *"unfortunately there is no good answer at my end ... I hope you can help me with finding an answer"*
3. Use other metrics (fidelity) (14%).	• *"By quantitative assessment of feature importance methods that assess specific properties like faithfulness"* • *"I might try and use some metric to measure fidelity."*

Table 5: Representative quotes highlighting themes of how participants address the disagreement problem in their day to day work