Analyzing Explainer Robustness via Probabilistic Lipschitzness of Prediction Functions

Zulqarnain Khan*
Northeastern University
khanzu@ece.neu.edu

Davin Hill*
Northeastern University
dhill@ece.neu.edu

Aria Masoomi
Northeastern University
masoomi.a@northeastern.edu

Josh Bone
Northeastern University
bone.j@northeastern.edu

Jennifer Dy
Northeastern University
jdy@ece.neu.edu

Abstract

Machine learning methods have significantly improved in their predictive capabilities, but at the same time they are becoming more complex and less transparent. As a result, explainers are often relied on to provide interpretability to these *black-box* prediction models. As crucial diagnostics tools, it is important that these explainers themselves are robust. In this paper we focus on one particular aspect of robustness, namely that an explainer should give similar explanations for similar data inputs. We formalize this notion by introducing and defining *explainer astuteness*, analogous to astuteness of prediction functions. Our formalism allows us to connect explainer robustness to the predictor's *probabilistic Lipschitzness*, which captures the probability of local smoothness of a function. We provide lower bound guarantees on the astuteness of a variety of explainers (e.g., SHAP, RISE, CXPlain) given the Lipschitzness of the prediction function. These theoretical results imply that locally smooth prediction functions lend themselves to locally robust explanations. We evaluate these results empirically on simulated as well as real datasets.

1 INTRODUCTION

Machine learning models have made significant improvements in their ability to predict and classify data. However, these gains in predictive power have come at the cost of increasingly opaque models, which has resulted in a proliferation of explainers that seek to provide transparency for these black-box models. Given the importance of these explainers, it is essential to understand the factors that contribute to their robustness and effectiveness.

In this paper we focus on explainer robustness. A robust explainer is one where *similar inputs results in similar explanations*. As an example, consider two patients given the same diagnosis in a medical setting. These two patients share identical symptoms and are demographically very similar, therefore a clinician would expect that factors influencing the model decision should be similar as well. Prior work in explainer robustness suggests that this expectation does not always hold true (Alvarez-Melis and Jaakkola, 2018; Ghorbani et al., 2019); small changes to the input samples can result in large shifts in explanation.

For this reason we investigate the theoretical underpinning of explainer robustness. Specifically, we focus on investigating the connection between explainer robustness and smoothness of the black-box model being explained. In this work, we propose and formally define *Explainer Astuteness*, which characterizes the ability of a given explainer to provide robust explanations. Explainer Astuteness enables the evaluation of different explainers for prediction tasks where having robust explanations is critical. We then establish a theoretical connection between explainer astuteness and the *probabilistic Lipschitzness* of the black-box model that is being explained. Since probabilistic Lipschitzness is

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

*equal contribution

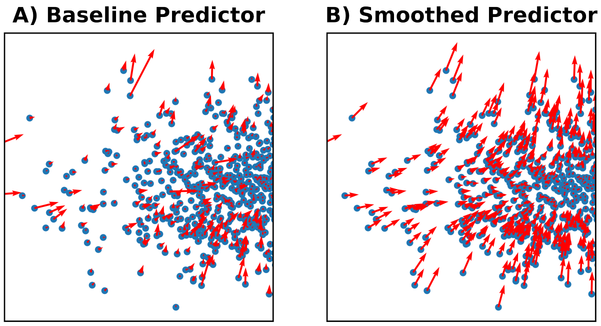

Figure 1: **Smoother black-box predictors lead to more astute explainers and robust explanations.** Samples from a simulated dataset are plotted in blue, with red arrows representing respective post-hoc SHAP explanations plotted as a vector. **A)** When a neural network (NN) is trained with no Lipschitz constraints, explanations of nearby points can vary significantly, as evidenced by the arrows varying in length and direction. **B)** When the NN is retrained with Lipschitz regularization suggested by Gouk et al. (2021), explanations are observed to be more aligned in length and direction, indicating higher robustness.

a measure of the probability that a function is smooth in a local neighborhood, our results demonstrate how the smoothness of the black-box model itself impacts the astuteness of the explainer (Fig. 1).

Contributions:

- We formalize *explainer astuteness*, which captures how well a given explainer provides similar explanations to similar points. This formalism allows us to evaluate different explainers based on the robustness of their explanations.

- We establish a lower bound on explainer astuteness which is dependent on the smoothness of the black-box model. We characterize model smoothness with *probabilistic Lipschitzness*, which allows for a broader class of black-box models than with standard Lipschitzness.

- We investigate the astuteness of three classes of explainers: (1) Shapley value-based (e.g., SHAP), (2) explainers that simulate mean effect of features (e.g., RISE), and (3) explainers that simulate individual feature removal (e.g., CXPlain).

- Our theoretical results are empirically validated on a mix of simulated and real datasets. Experiments show that increasing the smoothness of black-box models also improves the astuteness of the three classes of investigated explainers.

2 RELATED WORKS

Explainers. A wide variety of explainers have been proposed in the literature (Guidotti et al., 2018; Arrieta et al., 2020). We focus on *post-hoc* methods, specifically feature attribution and feature selection explainers. Feature attribution explainers provide continuous-valued importance scores to each of the input features. Some models such as CXPlain (Schwab and Karlen, 2019), PredDiff (Zintgraf et al., 2017) and feature ablation explainers (Lei et al., 2018) calculate feature attributions by simulating individual feature removal, while other methods such as RISE (Petsiuk et al., 2018) calculate the mean effect of a feature's presence to attribute importance to it. Lundberg and Lee (2017) unify six different feature attribution explainers under the SHAP framework. Other works have extended feature attributions to higher-order explanations (Lundberg et al., 2018; Masoomi et al., 2021; Torop et al., 2023). In contrast, feature selection explainers provide binary explanations for each feature and include individual selector approaches such as L2X (Chen et al., 2018) and INVASE (Yoon et al., 2018), and group-wise selection approaches such as gI (Masoomi et al., 2020).

In this work, we focus on *removal-based explainers*. Removal based feature explainers are a class of 25 methods defined by Covert et al. (2020) that define a feature's influence through the impact of removing it from a model. This includes popular approaches including KernelSHAP, LIME, DeepLIFT (Lundberg and Lee, 2017), mean effect based methods such as RISE (Petsiuk et al., 2018), and individual effects based methods such as CXPlain (Schwab and Karlen, 2019), PredDiff (Zintgraf et al., 2017), permutation tests (Strobl et al., 2008), and feature ablation explainers (Lei et al., 2018). All of these methods simulate feature removal either explicitly or implicitly. For example, SHAP explicitly considers effect of using subsets that include a feature as compared to the effect of removing that feature from the subset.

Explainer Robustness. Similarly, there has been a recent increase in research focused on analyzing different aspects of explainer behavior and reliability. Yin et al. (2021) propose stability and sensitivity as measures of faithfulness of explainers to the classifier decision-making process. Yeh et al. (2019) investigate infidelity and sensitivity of explanations under perturbations. Li et al. (2020) explore connections between explainers and model generalization. Agarwal et al. (2022c) investigate theoretical guarantees for stability of Graph Neural Network explainers. The term "robustness" for explainers has also been used in different contexts, such as in relation to distribution shifts (Lakkaraju et al., 2020; Upadhyay et al., 2021), model geometry (Dombrowski et al., 2019; Wang et al., 2020;

Hill et al., 2022), or adversarial attacks (Ghorbani et al., 2019; Rieger and Hansen, 2020). We follow the definition by Alvarez-Melis and Jaakkola (2018), who empirically show that robustness, in the sense that explainers should provide similar explanations for similar inputs, is a desirable property and how forcing this property yields better explanations. Recently, Agarwal et al. (2021) explore the robustness of LIME (Ribeiro et al., 2016) and SmoothGrad (Smilkov et al., 2017), and prove that, for these two methods, their robustness is related to the maximum value of the gradient of the predictor function.

Our work is related to the work by Alvarez-Melis and Jaakkola (2018) and Agarwal et al. (2021) on explainer robustness. However, instead of enforcing explainers to be robust themselves (Alvarez-Melis and Jaakkola, 2018), our theoretical results suggest that ensuring robustness of explanations also depends on the smoothness of the black-box model that is being explained. Our results are complementary to the results obtained by Agarwal et al. (2021) in that our theorems cover a wider variety of explainers as compared to only Continuous LIME and SmoothGrad. We further relate robustness to probabilistic Lipschitzness of black-box models, which is a quantity that can be empirically estimated. Our work is also related to independently-developed contemporaneous work by Tan and Tian (2023) and Agarwal et al. (2022a). However, our approach focuses on *probabilistic* Lipschitzness and its relationship to astuteness, which can be likened to a probabilistic form of "robustness" over the entire data distribution. This allows us to address classifiers that fall on a spectrum between robust and non-robust, rather than adhering to deterministic binary categorizations as in Tan and Tian (2023). In addition, our approach contrasts with that of Agarwal et al. (2022a), which only considers *local* robustness.

Lipschitzness of Neural Networks. There has been recent work estimating upper-bounds of Lipschitz constant for neural networks (Virmaux and Scaman, 2018; Fazlyab et al., 2019; Gouk et al., 2021), and enforcing Lipschitz continuity during neural networks training, with an eye towards improving classifier robustness (Gouk et al., 2021; Aziznejad et al., 2020; Fawzi et al., 2017; Alemi et al., 2016). (Fel et al., 2022) empirically demonstrated that 1-Lipschitz networks are better suited as predictors that are more explainable and trustworthy. Other benefits of neural network smoothness have been explored (Rosca et al., 2020), such as in improving model generalization (Hardt et al., 2016; Nakkiran et al., 2021) and adversarial robustness (Novak et al., 2018). Our work provides crucial additional motivation for this line of research; i.e., it provides theoretical reasons to improve Lipschitzness of neural networks from the perspective of enabling more robust explanations by proving that *enforcing smoothness on black-box models lends them to more robust explanations.*

3 EXPLAINER ASTUTENESS

Let $x \in \mathbb{R}^d$ be a d-dimensional sample from data distribution $\mathcal{D}$. We denote function f as a pre-trained, black-box model. The explainer is represented by ϕ where $\phi(x) \in \mathbb{R}^d$ is the feature attribution vector representing attributions for all features for sample x; $\phi_i(x) \in \mathbb{R}$ indicates the attribution for the i^{th} feature. We define $d_p(x, x') = ||x - x'||_p$ as the p-norm induced distance between two points. Our main interest is to define a metric that can capture the difference in explanations provided by an explainer to points that are close to each other in the input space. The same question has been asked for classifiers. The concept of *Astuteness* was introduced by Bhattacharjee and Chaudhuri (2020) in the context of classifiers; it captures the probability that similar points are assigned the same label by a classifier. Formally they provide the following definition:

Definition 1. *Astuteness of classifiers* (Bhattacharjee and Chaudhuri, 2020): The astuteness of a classifier f over $\mathcal{D}$, denoted as $A_r(f, \mathcal{D})$ is the probability that $\forall x, x' \in \mathcal{D}$ s.t. $d(x, x') \leq r$, the classifier will predict the same label.

$$A_r(f, \mathcal{D}) = \mathbb{P}_{x,x' \sim \mathcal{D}}[f(x) = f(x') | d(x, x') \leq r] \qquad (1)$$

The challenge in extending this concept of astuteness to explainers is that attribution-based explanations are not necessarily discrete; we need to generalize classifier astuteness to model continuous function outputs. With this in mind, we propose and formalize *explainer astuteness*, as the probability that the explainer assigns *similar* explanations to similar points.

Definition 2. *Explainer astuteness*: The *explainer astuteness* of an explainer E over $\mathcal{D}$, denoted as $A_{r,\lambda}(E, \mathcal{D})$ is the probability that $\forall x, x' \in \mathcal{D}$ such that $d_p(x, x') \leq r$ the explainer E will provide explanations $\phi(x), \phi(x')$ that are at most $\lambda \cdot d_p(x, x')$ away from each other, where $\lambda \geq 0$

$$\begin{aligned} A_{r,\lambda}(E, \mathcal{D}) = \mathbb{P}_{x,x' \sim \mathcal{D}}[d_p(\phi(x), \phi(x')) \leq \lambda \cdot d_p(x, x') \\ \Big| \, d_p(x, x') \leq r] \end{aligned} \qquad (2)$$

Explainer Astuteness, as defined above, captures the desirable robustness property of explainers that we are interested in. It provides a value between 0 and 1 that quantifies how different or similar the explanations provided by our explainer are within neighborhoods of radius r. 0 implies the explainer is providing different explanations even for nearby points, and 1 implies the

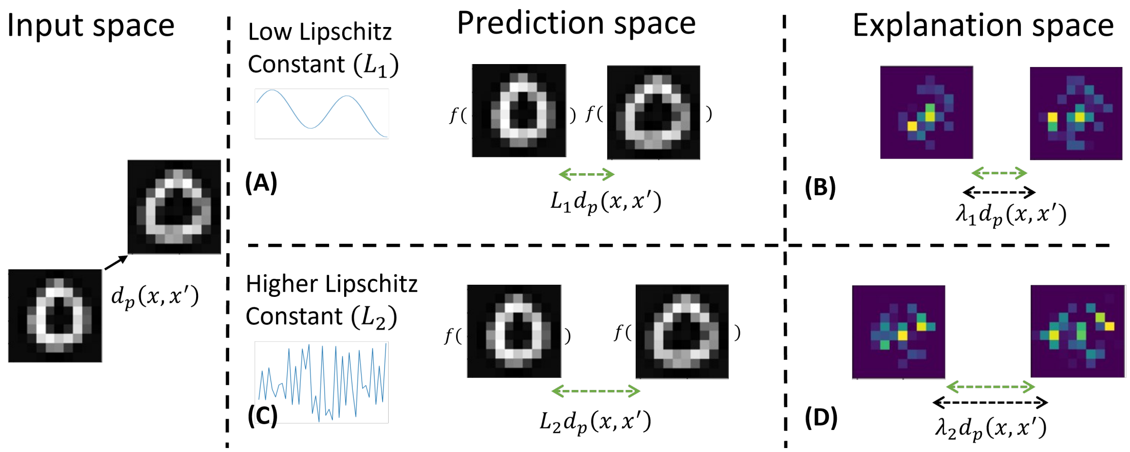

Figure 2: **Summary of our theoretical results.** **(A)** For a black-box prediction function that is locally Lipschitz with a constant L_1, the predictions for any two points x, x' (shown here with two similar images from MNIST dataset) such that $d_p(x,x') \leq r$ are within $L_1 d_p(x,x')$ distance from each other. **(B)** Given such a prediction function, the explanation (feature attributions shown, with brighter colors indicating higher scores) for the same data points are also expected to be within $\lambda_1 d_p(x,x')$ of each other where $\lambda_1 = CL_1\sqrt{d}$ where C is a constant. **(C)** For a second black-box model with $L_2 > L_1$, our results show that **(D)** $\lambda_2 > \lambda_1$, indicating that the explanations for this black-box model can be farther apart as compared to the first prediction function. This result implies that *locally smooth black-box models lend themselves to more astute explainers.*

explainer gives similar explanations for nearby points. Our goal now is to connect this measure of explainer robustness to a measure of predictor smoothness. Typically, notions of local or global *detereministic* Lipschitzness are used for this purpose (see Tan and Tian (2023) and Agarwal et al. (2022a) for example). Instead, we choose to use *probabilistic* Lipschitzness. Probabilistic Lipschitzness captures the probability of a function being locally smooth given a radius r. This allows us to address predictors over the full range of local to global and non-smooth to deterministically smooth ones. This definition also mirrors our definition of explainer astuteness and helps establish the relationship between the two concepts. It is also especially useful for capturing a notion of smoothness of complicated neural network functions for which enforcing, as well as calculating, global and deterministic Lipschitzness is difficult. *Probabilistic* Lipschitzness is defined as follows:

Definition 3. *Probabilistic Lipschitzness* (Mangal et al., 2020): Given $0 \leq \alpha \leq 1$, $r \geq 0$, a function $f : \mathbb{X} \rightarrow \mathbb{R}$ is probabilistically Lipschitz with a constant $L \geq 0$ if

$$\mathbb{P}_{x,x' \sim \mathcal{D}}[d_p(f(x), f(x')) \leq L \cdot d_p(x,x') \Big| d_p(x,x') \leq r] \geq 1-\alpha \tag{3}$$

4 THEORETICAL BOUNDS OF ASTUTENESS

In this section, we theoretically show how the probabilistic Lipschitzness of black-box models relate to the astuteness of explainers. We introduce and prove theoretical bounds that connect the Lipschitz constant L of the black-box model to the astuteness of three different classes of explainers: SHAP (Lundberg and Lee, 2017), RISE (Petsiuk et al., 2018), and methods that simulate individual feature removal such as CXPlain (Schwab and Karlen, 2019). An overview of these results is provided in Figure 2.

4.1 Astuteness of SHAP

SHAP (Lundberg and Lee, 2017) is one of the most popular feature attribution based explainers in use today. (Lundberg and Lee, 2017) unify 6 existing explanation approaches within the SHAP framework. Each of these explanation approaches (such as DeepLIFT and kernelSHAP) can be viewed as approximations of SHAP, since SHAP in its theoretical form is difficult to calculate. However, in this section we use the theoretical definition of SHAP to establish bounds on astuteness.

For notational consistency with other removal-based explainers in this work, we use an indicator vector $z \in \{0,1\}^d$, where $z_i = 1$ when the i^{th} feature is included in the subset. To indicate subsets where feature $z_i = 1$, we use z_{+i}; conversely, z_{-i} indicates subsets where $z_i = 0$. $|z_{-i}|$ represents the number of non-zero entries in z_{-i}. Following definition used in Lundberg and Lee (2017) for a given data point $x \in \mathcal{X}$ and a prediction function f, the feature attribution provided by SHAP for the i^{th} feature is given by:

$$\phi_i(x) = \sum_{z_{-i}} \frac{|z_{-i}|!(d-|z_{-i}|-1)!}{d!}[f(x \odot z_{+i}) - f(x \odot z_{-i})] \tag{4}$$

We introduce the following Lemma (proof in Appendix A) which is necessary for proving Theorem 1.

Lemma 1. *If,*

$$\mathbb{P}_{x,x'\sim\mathcal{D}}[d_p(f(x), f(x')] \leq L\cdot d_p(x,x') \Big| d_p(x,x') \leq r] \geq 1-\alpha$$

then for $y = x \odot z_{+i}, y' = x' \odot z_{+i}$, i.e. $y, y' \in \cup\mathbb{N}_k = \{y|y \in \mathbb{R}^d, ||y||_0 = k, y_i \neq 0\}$ for $k = 1, \ldots, d$

$$\mathbb{P}_{x,x'\sim\mathcal{D}}[d_p(f(y), f(y')) \leq L\cdot d_p(y,y') \Big| d_p(y,y') \leq r] \geq 1-\beta$$

where $\beta \geq \alpha$ assuming that the distribution $\mathcal{D}$ is defined for all x and y and the equality is approached if the probability of sampling points from the set $\mathbb{N}_k = \{y|y \in \mathbb{R}^d, ||y||_0 = k, y_i \neq 0\}$ approaches zero for $k = 2, \ldots, d$ relative to the probability of sampling points from $\mathbb{N}_1$ i.e. the probability of sampling points that have at least 1 element exactly equal to 0 is vanishingly small compared to the probability of sampling points that have no 0 element.

Theorem 1. *(Astuteness of SHAP) Consider a given $r \geq 0$ and $0 \leq \alpha \leq 1$, and a trained predictive function f that is probabilistic Lipschitz with a constant L, radius r measured using $d_p(.,.)$ and with probability at least $1-\alpha$. Then for SHAP explainers we have astuteness $A_{r,\lambda} \geq 1-\beta$ for $\lambda = 2\sqrt[p]{d}L$. Where $\beta \geq \alpha$, and $\beta \to \alpha$ under conditions specified in Lemma 1.*

Proof. Given input points x, x' s.t. $d(x,x') \leq r$. And letting $\frac{|z_{-i}|!(d-|z_{-i}|-1)!}{d!} = C_z$. Using (4) we can write,

$$\begin{aligned} d_p(\phi_i(x),\phi_i(x')) = ||\sum_{z_{-i}} C_z[f(x \odot z_{+i}) - f(x \odot z_{-i})] \\ - \sum_{z_{-i}} C_z[f(x' \odot z_{+i}) - f(x' \odot z_{-i})]||_p \end{aligned} \tag{5}$$

Combining the two sums and re-arranging the R.H.S,

$$\begin{aligned} d_p(\phi_i(x),\phi_i(x')) = ||\sum_{z_{-i}} C_z[f(x \odot z_{+i}) - f(x' \odot z_{+i}) \\ + f(x' \odot z_{-i}) - f(x \odot z_{-i})]||_p \end{aligned} \tag{6}$$

Using triangular inequality on the R.H.S twice,

$$\begin{aligned} d_p(\phi_i(x),\phi_i(x')) &\leq ||\sum_{z_{-i}} C_z[f(x \odot z_{+i}) - f(x' \odot z_{+i})]||_p \\ &+ ||\sum_{z_{-i}} C_z[f(x' \odot z_{-i}) - f(x \odot z_{-i})]||_p \\ &\leq \sum_{z_{-i}} C_z||f(x \odot z_{+i}) - f(x' \odot z_{+i})||_p \\ &+ \sum_{z_{-i}} C_z||f(x' \odot z_{-i}) - f(x \odot z_{-i})||_p \end{aligned} \tag{7}$$

We can replace each value inside the sums in (7) with the maximum value across either sums. Doing so would still preserve the inequality in (7), as the sum of n values is always less than the maximum among those summed n times. Without loss of generality let us assume this maximum is $|f(x \odot z^*) - f(x' \odot z^*)|$ for some particular z^*. This gives us:

$$\begin{aligned} d_p(\phi_i(x), \phi_i(x')) \leq ||f(x \odot z^*) - f(x' \odot z^*)||_p \sum_{z_{-i}} C_z \\ + ||f(x \odot z^*) - f(x' \odot z^*)||_p \sum_{z_{-i}} C_z \end{aligned} \tag{8}$$

However, $\sum_{z_{-i}} C_z = \sum_{z_{-i}} \frac{|z_{-i}|!(d-|z_{-i}|-1)!}{d!} = 1$, so,

$$\begin{aligned} d_p(\phi_i(x), \phi_i(x')) \leq 2||f(x \odot z^*) - f(x' \odot z^*)||_p \\ = 2d_p(f(x \odot z^*), f(x' \odot z^*)) \end{aligned} \tag{9}$$

Using the fact that f is probabilistic Lipschitz with a given constant $L \geq 0$, $d_p(x,x') \leq r$, $d_p(x \odot z^*, x' \odot z^*) \leq d_p(x,x')$ and Lemma 1. We get:

$$\mathbb{P}[2d_p(f(x \odot z^*), f(x' \odot z^*)) \leq 2L \cdot d_p(x,x')] \geq 1-\beta$$

Since (9) implies $d_p(\phi_i(x), \phi_i(x')) \leq 2d_p(f(x \odot z^*), f(x' \odot z^*))$, the below inequality can be established:

$$\mathbb{P}[d_p(\phi_i(x), \phi_i(x')) \leq 2L \cdot d_p(x,x')] \geq 1-\beta \tag{10}$$

Note that (10) is true for each feature $i \in \{1, ..., d\}$. To conclude our proof, we note that

$$\begin{aligned} d_p(x,y) = \sqrt[p]{\sum_i^d |x_i - y_i|^p} &\leq \sqrt[p]{\sum_i^d \max_i |x_i - y_i|^p} \\ &= \sqrt[p]{d} \max_i d_p(x_i, y_i) \end{aligned}$$

Utilizing this in (10), without loss of generality assuming $d_p(\phi_i(x), \phi_i(x'))$ corresponds to the maximum,

$$\mathbb{P}[d_p(\phi(x), \phi(x')) \leq 2\sqrt[p]{d}L \cdot d_p(x,x')] \geq 1-\beta \tag{11}$$

Since $\mathbb{P}[d_p(\phi(x), \phi(x')) \leq 2\sqrt[p]{d}L \cdot d_p(x,x')]$ in (11) defines $A_{\lambda,r}$ for $\lambda = 2\sqrt[p]{d}L$, this concludes the proof. □

Corollary 1. *If the prediction function f is locally deterministically $L-$Lipschitz ($\alpha = 0$) at radius r then Shapley explainers are $\lambda-$astute for radius $r \geq 0$ for $\lambda = 2\sqrt[p]{d}L$.*

Proof. Note that definition 3 reduces to the definition of deterministic Lipschitz if $\alpha = 0$. Which means (11) will be true with probability 1. Which concludes the proof. □

4.2 Astuteness of "Remove Individual" Explainers

Within the framework of feature removal explainers, a sub-category is the explainers that work by removing a single feature from the set of all features and calculating feature attributions based on change in prediction that result from removing that feature. This category includes Occlusion, CXPlain (Schwab and Karlen, 2019), PredDiff (Zintgraf et al., 2017) Permutation tests (Strobl et al., 2008), and feature ablation explainers (Lei et al., 2018).

"Remove individual" explainers determine feature explanations for the i^{th} feature by calculating the difference in prediction with and without that feature included for a given point x. Let $z_{-i} \in \{0,1\}^d$ represent a binary vector with $z_i = 0$, then the explanation for feature i can be written as:

$$\phi(x_i) = f(x) - f(x \odot z_{-i}) \tag{12}$$

Theorem 2. *(Astuteness of Remove individual explainers) Consider a given $r \geq 0$ and $0 \leq \alpha \leq 1$ and a trained predictive function f that is locally probabilistic Lipschitz with a constant L, radius r measured using $d_p(.,.)$ and probability at least $1-\alpha$. Then for Remove individual explainers, we have the astuteness $A_{r,\lambda} \geq 1-\alpha$, for $\lambda = 2\sqrt[p]{d}L$, where d is the dimensionality of the data.*

Proof. (Sketch, full proof in Appendix A) By considering another point x' such that $d_p(x, x') \leq r$ and (12) we get,

$$d_p(\phi(x_i), \phi(x'_i)) = d_p(f(x) - f(x \odot z_{-i}), f(x') - f(x' \odot z_{-i})) \tag{13}$$

then following the exact same steps as the proof for Theorem 1, i.e. writing the right hand side in terms of p-norm and utilizing triangular inequality, leads us to the desired result. □

Corollary 2. *If the prediction function f is locally $L-$Lipschitz at radius $r \geq 0$, then remove individual explanations are $\lambda-$astute for radius r and $\lambda = 2\sqrt[p]{d}L$.*

Proof. Same as proof for Corollary 2.1. □

4.3 Astuteness of RISE

RISE determines feature explanation for the i^{th} feature by sampling subsets of features and then calculating the mean value of the prediction function when feature i is included in the subset. RISE feature attribution for a given point x and feature i for a prediction function f can be written as:

$$\phi_i(x) = \mathbb{E}_{p(z|z_i=1)}[f(x \odot z)] \tag{14}$$

The following theorem establishes the bound on λ for *explainer astuteness* of RISE in relation to the Lipschitzness of black-box prediction function.

Theorem 3. *(Astuteness of RISE) Consider a given $r \geq 0$ and $0 \leq \alpha \leq 1$, and a trained predictive function f that is locally deterministically Lipschitz with a constant L ($\alpha = 0$), radius r measured using $d_p(.,.)$ and probability at least $1-\alpha$. Then for RISE explainer is $\lambda-$astute for radius r and $\lambda = \sqrt[p]{d}L$.*

Proof. (Sketch, full proof in Appendix A)

Given inputs x, x' s.t. $d(x, x') \leq r$, using Eq. (14),

$$\begin{aligned} & d_p(\phi_i(x), \phi_i(x')) \\ & = d_p(\mathbb{E}_{p(z|z_i=1)}[f(x \odot z)], \mathbb{E}_{p(z|z_i=1)}[f(x' \odot z)]) \\ & = ||\mathbb{E}_{p(z|z_i=1)}[f(x \odot z)] - \mathbb{E}_{p(z|z_i=1)}[f(x' \odot z)]||_p \\ & = ||\mathbb{E}_{p(z|z_i=1)}[f(x \odot z) - f(x' \odot z)]||_p \end{aligned} \tag{15}$$

Using Jensen's inequality on R.H.S and $E[f] \leq \max f$

$$d_p(\phi_i(x), \phi_i(x')) \leq \max_z d_p(f(x \odot z), f(x' \odot z)) \tag{16}$$

f is is deterministically Lipschitz and $d_p(\phi(x), \phi(x')) \leq \sqrt[p]{d} * \max_i d_p(\phi_i(x), \phi_i(x'))$, this gives us,

$$\mathbb{P}[d_p(\phi(x), \phi(x') \leq \sqrt[p]{d}L \cdot d_p(x, x')] \geq 1 \tag{17}$$

Since $\mathbb{P}[d_p(\phi(x), \phi(x') \leq \sqrt[p]{d}L \cdot d_p(x, x')]$ defines $A_{\lambda,r}$ for $\lambda = \sqrt[p]{d}L$, this concludes the proof. □

4.4 Connecting Explainer Astuteness and Probabilistic Lipschitzness

The above theoretical results for the three classes of explainers provide the same critical implication, that is, explainer astuteness is lower bounded by the probabilistic Lipschitzness of the prediction function. This means that black-box classifiers that are locally smooth (have a small L at a given radius r) lend themselves to probabilistically more robust explanations. *This work provides the theoretical support on the importance of enforcing smoothness of classifiers to astuteness of explanations.* Note that while this implication makes intuitive sense, proving it for specific explainers is non-trivial as demonstrated by the three theorems above and their respective proofs. The statement holds true for all three explainers when the classifier can be assumed to be deterministically Lipschitz, the conditions under which it is still true for probabilistic Lipschitzness vary in each case. For Theorem 1 we have to assume that distribution $\mathcal{D}$ is defined over masked data in addition to the input data and ideally the probability of sampling of masked data from is significantly smaller compared to probability of sampling points with no value exactly equal to 0. For Theorem 2 the statement is true without additional assumptions. For Theorem 3 we can only prove the statement to be true for the deterministic case.

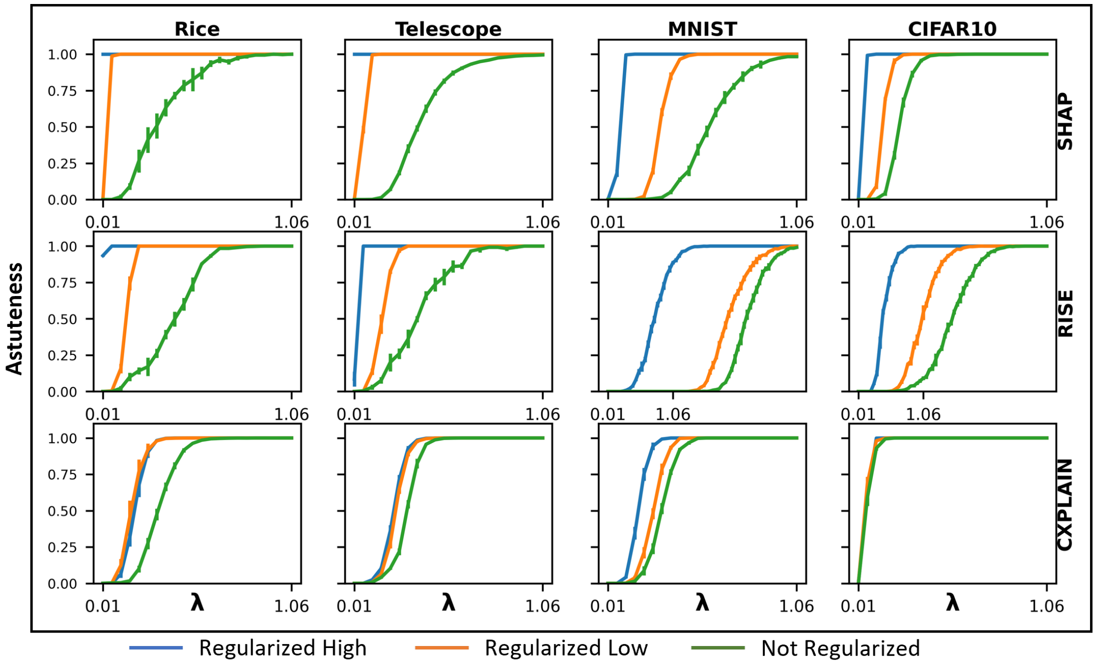

Figure 3: **Smooth functions result in astute explanations.** Regularizing the Lipschitzness of a neural network during training results in higher astuteness for the same value of λ. Higher regularization results in lower Lipschitz constant Gouk et al. (2021). Astuteness reaches 1 for smaller values of λ with Lipschitz regularized training, as expected from our theorems. The errorbars represent results across 5 runs to account for randomness in explainer runs.

5 EXPERIMENTS

To demonstrate the validity of our theoretical results, we perform a series of experiments. We train four different classifiers on each of five datasets, and then explain the decisions of these classifiers using three explainers and calculate astuteness of these explainers. We use $p = 2$ in all experiments.

Datasets. We utilize three simulated datasets introduced by Chen et al. (2018) namely *Orange Skin*(OS), *Nonlinear Additive*(NA) and *Switch*, and two real world datasets from UCI Machine Learning repository (Asuncion and Newman, 2007) namely *Rice* (Cinar and Koklu, 2019) and *Telescope* (Ferenc et al., 2005) as well as *MNIST*(LeCun and Cortes, 2010) and *CIFAR10*(Krizhevsky et al., 2009) datasets. See Appendix B for details.

Classifiers. For each dataset we train the following four classifiers; a 2-*layer* MLP, a 4-*layer* MLP, a *linear* classifier, and a *svm*. For training Details see Appendix C.

Explainers. We evaluate 3 explainers, one for each of our theorems. Gradient-based approximation and kernel shap approximation of **SHAP**(Lundberg and Lee, 2017) for the NN classifiers and SVM, respectively, serve as representative of Theorem 1. We modify the implementation of **RISE** (Petsiuk et al., 2018) provided by the authors to work with tabular datasets; this serves as representative for Theorem 3. The implementation for **CXPlain** (Schwab and Karlen, 2019) serves as representative for Theorem 2.[1].

5.1 Effect of Lipschitz Constraints on Explainer Astuteness

In this experiment, we utilize Lipschitz-constrained classifiers introduced by Gouk et al. (2021), which constrain the Lipschitz constant for each layer by adding a projection step during training. During each gradient update, the weight matrices undergo a projection onto a feasible set if they violate the constraints imposed on the Lipschitz constant. The level of constraint can be adjusted through a hyperparameter, allowing control over the impact on the weight matrices. We use this method to train a four-layer MLP with high, low, and no Lipschitz constraint. We then calculate astuteness of each of our explainers for all three versions of this neural network. Figure 3 shows the results. The goal of this experiment is to demonstrate the relationship between the Lipschitzness of a NN and the astuteness of explainers. As the *same* NN is trained on the *same*

[1] SHAP: `https://github.com/slundberg/shap`, RISE: `https://github.com/eclique/RISE`, CXPLAIN: `https://github.com/d909b/cxplain`

	2layer			4layer			linear			svm		
Datasets	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain	SHAP	RISE	CXPlain
OS	.585	.477	.551	.489	.415	.426	.043	.017	.043	.761	.628	.732
NA	.359	.289	.318	.285	.216	.244	.452	.391	.474	.742	.653	.708
Switch	.053	.053	.003	.086	.083	.039	.043	.028	.034	.557	.472	.524
Rice	.249	.142	.229	.292	.131	.252	.258	.165	.241	.426	.347	.413
Telescope	.324	.213	.317	.345	.244	.333	.223	.149	.211	.501	.439	.504
CIFAR10	.333	.238	.337	.350	.224	.356	.340	.235	.345	.336	.329	.255
MNIST	.441	.297	.435	.522	.357	.519	.455	.303	.448	.443	.379	.448

Table 1: $\mathbf{AUC_{emp}} - \mathbf{AUC_{pred}}(\downarrow)$. The observed AUC ($\mathrm{AUC_{emp}}$) is lower bounded by the predicted AUC ($\mathrm{AUC_{pred}}$). As expected, the difference between the two is always ≥ 0.

data but with different levels of Lipschitz constraints enforced, the astuteness of explainers varies accordingly. In all cases we see astuteness reaching 1 for smaller values of λ for the same NN when it is highly constrained (lower Lipschitz constant L) vs less constrained or unconstrained. *The results provide empirical evidence in support of the main conclusion of our work: i.e., enforcing Lipschitzness on classifiers lends them to more astute post-hoc explanations.*

5.2 Estimating Probabilistic Lipschitzness and Lower Bound for Astuteness

To demonstrate the connection between explainer astuteness and probabilistic Lipschitzness as shown in our theorems, we estimate the probabilistic Lipschitzness of various classifiers. We achieve this by empirically estimating $\mathbb{P}_{x,x'\sim\mathcal{D}}$ (3) for a range of values of $L \in (0,1)$ incremented by 0.1. This is done for each classifier and for each dataset D, and we set r as the median of pairwise distance for all training points. According to (3), this gives us an upper bound on $1-\alpha$, i.e. we can say that for a given L, r the classifier is Lipschitz with probability at least $1-\alpha$. We can use the estimates for probabilistic Lipschitzness to predict the lower bound of astuteness using our theorems. Theorems 1, 2, 3 imply that for $\lambda = CL\sqrt{d}$, explainer astuteness is $\geq 1-\alpha$. This indicates that for $\lambda \geq LC\sqrt{d}$, explainer astuteness should be lower-bounded by $1-\alpha$. For each dataset-classifier-explainer combination we can plot two curves. First, the empirical estimation of explainer astuteness using Definition 2. Second, the curve that represents the predicted lower bound on explainer astuteness given a classifier, as just described. According to our theoretical results, at a given λ, the estimated explainer astuteness should stay above the predicted astuteness based on the Lipschitzness of classifiers. Table 1 shows the difference between the area under the curve (AUC) for the empirically calculated astuteness ($\mathbf{AUC_{emp}}$) and the predicted lower bound ($\mathbf{AUC_{pred}}$). This number captures the average difference of the lower bound over a range of λ values. *Note that the values are all positive supporting our result as a lower bound.* The associated curves are shown in App. D Figure 5.

5.3 Astuteness as a Metric

Our proposed metric of *Explainer Astuteness* (Definition 2) enables the comparison of different explainers based on the expected robustness of their explanations. This allows users to select a more astute explainer in situations where robust explanations are required. In Figure 4 we compare the astuteness of SHAP, RISE, and CXPlain. We observe that RISE is consistently less astute on the evaluated datasets, in contrast to CXPlain, which exhibits high astuteness. Therefore in this case CXPlain would be preferable for obtaining robust explanations.

We also provide results on four datasets from OpenXAI benchmark (Agarwal et al., 2022b), a synthetic data from OpenXAI (O-Synthetic), Home Equity Line of Credit (HELOC) (Holter et al., 2018), Propublica's COMPAS dataset (Jordan and Freiburger, 2015), and Adult Income (Adult) dataset (Yeh and Lien, 2009). Reported in Table 2, we calculate astuteness over a range of values of $\lambda \in \{0.1, 1.1\}$ in increments of 0.1, and report the AUC of astuteness over this range for SHAP and LIME using the implementations and pretrained models made available on OpenXAI's github [2], higher AUC indicates higher astuteness over the range. For each dataset-explainer combination we also report a metric for measuring stability - Relative Input Stability (RIS) - a measure of maximum change in explanation relative to changes in input (Alvarez-Melis and Jaakkola, 2018) and a measure of faithfulness - Prediction Gap on Important feature perturbation (PGI), which computes the difference in prediction probability that results from perturbing the features deemed as influential by a given post hoc explanation (Dai et al., 2022). These metrics and their values are reported on the OpenXAI leaderboard [3]. Since astuteness is closely related to stability, we can expect the astuteness AUC to be higher for the explainer that has a higher RIS for each dataset-predictor combination, as evident from bolded values in 2 that is indeed generally the case.

[2] `https://github.com/AI4LIFE-GROUP/OpenXAI`

[3] `https://open-xai.github.io/leaderboard`

	Neural Network						Logistic Regression					
	SHAP			LIME			SHAP			LIME		
Datasets	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓	AUC ↑	PGI ↑	RIS ↓
O-Synthetic	.198 ± .001	**.320**	5.96	**.379 ± .001**	.250	**5.53**	**.209 ± .000**	**.171**	**5.67**	.029 ± .000	.154	9.35
HELOC	.130 ± .000	**.260**	1.57	**.506 ± .002**	**.260**	**1.19**	**.177 ± .000**	.121	**1.46**	.045 ± .002	**.156**	4.53
COMPAS	.276 ± .001	**.274**	**4.24**	**.515 ± .000**	.232	4.49	.160 ± .000	**.128**	**3.12**	**.352 ± .001**	.108	4.24
Adult	.085 ± .000	.53	1.98	**.206 ± .000**	**.670**	**1.79**	**.114 ± .000**	.391	1.86	.042 ± .001	**.420**	**1.72**

Table 2: **Metrics for post-hoc explanations.** The area under the curve (AUC)↑ of astuteness for SHAP and LIME on benchmark datasets from OpenXAI. Prediction Gap on Important feature perturbation (PGI)↑ is a measure of faithfulness, and Relative Input Stablity (RIS)↓ is a measure of stability. Both PGI and RIS metrics are reported from OpenXAI leaderboard.

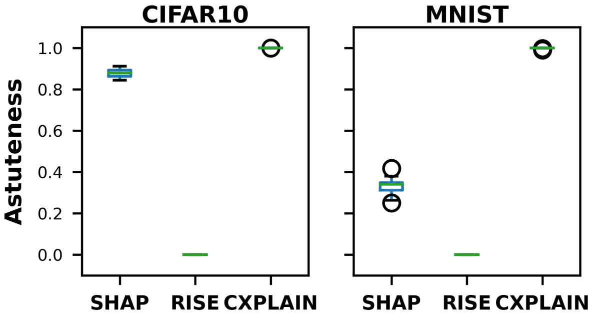

Figure 4: **Astuteness as a metric.** Different explainers display different levels of astuteness. In our experiments, RISE consistently displayed lower astuteness for the same value of λ compared to SHAP and CXPLAIN. This indicates that among these three , on the considered datasets and classifiers, RISE is least robust and CXPlain is the most robust. Results are shown across 20 explainer runs.

6 CONCLUSION AND LIMITATIONS

In this paper we formalize *explainer astuteness*, which captures the ability of an explainer to assign similar explanations to similar points. We prove that this explainer astuteness is proportional to the *probabilistic Lipschitzness* of the predictor that is being explained. As probabilistic Lipschitzness captures the local smoothness of a function, this result suggests that enforcing smoothness on predictors can lend them to more robust explanations.

Regarding limitations, we observe that our empirical results suggest that our predicted lower bound can be tightened further. One possible explanation is that the tightness of this bound depends on how different explainers calculate attribution scores, e.g. empirically we observe RISE and SHAP (that both depend on expectations over subsets) behave similarly to each other but different from CXPlain. Some explainers, such as LIME for tabular data, have an optional discretization step when calculating feature attributions. As a consequence, two observations with all features belonging to the same bins would receive exactly the same explanation, whereas two arbitrarily close inputs may receive completely different explanations (when the number of perturbed sample is large (Garreau and von Luxburg, 2020)). In that sense, tabular LIME would not be astute by our formulation, regardless of classifier Lipschitzness. Additionally, the usefulness of explainer robustness or astuteness is application-dependant. For example, robustness can sometimes be at odds with correctness (See for example (Zhou et al., 2022) and "Logic Trap 3" in (Ju et al., 2022)) and is best viewed as one part of explanation reliability and trustworthiness (Zhou et al., 2022).

From a broader societal impact perspective, we would like to make it clear that just enforcing Lipschitzness on black-box classifiers is not enough in terms of making them more transparent and interpretable. Our work is intended to be a call to action for the field to concentrate more on improving black-box models for explainability purposes from the very start and provides one of many ways to achieve that goal.

Acknowledgements

This work was supported in part by grants NIH 2T32HL007427-41 and U01HL089856 from the National Heart, Lung, and Blood Institute and by Northeastern University's Institute for Experiential AI.

References

Chirag Agarwal, Nari Johnson, Martin Pawelczyk, Satyapriya Krishna, Eshika Saxena, Marinka Zitnik, and Himabindu Lakkaraju. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022a.

Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35:15784–15799, 2022b.

Chirag Agarwal, Marinka Zitnik, and Himabindu Lakkaraju. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *International Conference on Artificial Intelligence and Statistics*, pages 8969–8996. PMLR, 2022c.

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Zhiwei Steven Wu, and Himabindu Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. *arXiv preprint arXiv:2102.10618*, 2021.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

David Alvarez-Melis and Tommi S Jaakkola. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

Shayan Aziznejad, Harshit Gupta, Joaquim Campos, and Michael Unser. Deep neural networks with trainable activations and controlled lipschitz constant. *IEEE Transactions on Signal Processing*, 68:4688–4699, 2020.

Robi Bhattacharjee and Kamalika Chaudhuri. When are non-parametric methods robust? In *International Conference on Machine Learning*, pages 832–841. PMLR, 2020.

Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *arXiv preprint arXiv:2011.14878*, 2020.

Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 203–214, 2022.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf`.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. The robustness of deep networks: A geometrical perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, 2017.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Thomas Fel, David Vigouroux, Rémi Cadène, and Thomas Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 720–730, 2022.

Daniel Ferenc, MAGIC collaboration, et al. The magic gamma-ray observatory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 553(1-2):274–281, 2005.

Damien Garreau and Ulrike von Luxburg. Looking deeper into tabular lime. *arXiv preprint arXiv:2008.11092*, 2020.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

Davin Hill, Aria Masoomi, Sandesh Ghimire, Max Torop, and Jennifer Dy. Explanation uncertainty with decision boundary awareness. *arXiv preprint arXiv:2210.02419*, 2022.

Steffen Holter, Oscar Gomez, and Enrico Bertini. Fico explainable machine learning challenge, 2018.

Kareem L Jordan and Tina L Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, 2015.

Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR, 2020.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Jeffrey Li, Vaishnavh Nagarajan, Gregory Plumb, and Ameet Talwalkar. A learning theoretic perspective on local explainability. *arXiv preprint arXiv:2011.01205*, 2020.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777, 2017.

Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

Ravi Mangal, Kartik Sarangmath, Aditya V Nori, and Alessandro Orso. Probabilistic lipschitz analysis of neural networks. In *International Static Analysis Symposium*, pages 274–309. Springer, 2020.

Aria Masoomi, Chieh Wu, Tingting Zhao, Zifeng Wang, Peter Castaldi, and Jennifer Dy. Instance-wise feature grouping. *Advances in Neural Information Processing Systems*, 33, 2020.

Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P. Hersh, Edwin K. Silverman, Peter J. Castaldi, Stratis Ioannidis, and Jennifer Dy. Explanations of black-box models based on directional feature interactions. In *International Conference on Learning Representations*, 2021.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Laura Rieger and Lars Kai Hansen. A simple defense against adversarial attacks on heatmap explanations. In *5th Annual Workshop on Human Interpretability in Machine Learning*, 2020.

Mihaela Rosca, Theophane Weber, Arthur Gretton, and Shakir Mohamed. A case for new neural network smoothness constraints. In Jessica Zosa Forde, Francisco Ruiz, Melanie F. Pradier, and Aaron Schein, editors, *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, volume 137 of *Proceedings of Machine Learning Research*, pages 21–32. PMLR, 12 Dec 2020. URL `https://proceedings.mlr.press/v137/rosca20a.html`.

Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. *arXiv preprint arXiv:1910.12336*, 2019.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

Zeren Tan and Yang Tian. Robust explanation for free or at the cost of faithfulness. 2023.

Max Torop, Aria Masoomi, Davin Hill, Kivanc Kose, Stratis Ioannidis, and Jennifer Dy. Smoothhess: ReLU network feature interactions via stein's lemma.

In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=dwIeEhbaD0`.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems*, 34:16926–16937, 2021.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Zifan Wang, Haofan Wang, Shakul Ramkumar, Piotr Mardziel, Matt Fredrikson, and Anupam Datta. Smoothed geometry for robust attribution. *Advances in neural information processing systems*, 33:13623–13634, 2020.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. On the faithfulness measurements for model interpretations. *arXiv preprint arXiv:2104.08782*, 2021.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

Yilun Zhou, Marco Tulio Ribeiro, and Julie Shah. Exsum: From local explanations to model understanding. *arXiv preprint arXiv:2205.00130*, 2022.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*, 2017.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/**Not Applicable**] This paper does not present any models or algorithms.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/**Not Applicable**] This paper does not present any models or algorithms.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/**Not Applicable**] This paper does not present any models or algorithms. Code for experiments is provided in Supplementary Materials.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable] All theorems in Section 4 specify all needed assumptions.
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable] Proof for Theorem 1 is in main text, all other proofs are provided in the Appendix.
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable] Code for experiments is provided as part of supplementary materials.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes~~/No/Not Applicable~~] Training details are provided in the experiment section and in the Appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable] Error bars are explained where used.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/**No**/Not Applicable] We did not keep track of this and don't consider it to be important to report for this work.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable] Citations are included for all datasets used.
   (b) The license information of the assets, if applicable. [Yes/No/**Not Applicable**]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/**Not Applicable**]
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A DETAILED PROOFS

We include the detailed proofs for Lemma 1, Theorems 3 and 2 here.

A.1 Lemma 1

Proof. Let us assume,

$$p_k = \mathbb{P}[\mathbb{N}_k], s.t. \mathbb{N}_k = \{x \mid x \in \mathbb{R}^d, ||x||_0 = k, x_i \neq 0\}$$

and let $\hat{\mathbb{L}}$ be the set of points that violate Lipschitzness, then assume,

$$\gamma_k = \mathbb{P}[\hat{\mathbb{L}} \mid \mathbb{N}_k]$$

given that α is the probability of the set of points that violate Lipschitzness across $\mathcal{D}$, we can use Bayes' rule to write,

$$\alpha = \mathbb{P}[\hat{\mathbb{L}}] = \sum_{k=1}^{d} p_k \gamma_k$$

If we consider the case where the sets $\mathbb{N}_k$ are finite, each $\mathbb{N}_k$ can be mapped to a set $\mathbb{N}'_k$ of cardinality,

$$|\mathbb{N}'_k| = \sum_{b=0}^{d-k} \binom{d-k}{b} |\mathbb{N}_k| = 2^{d-k} |\mathbb{N}_k|$$

In more general terms, the probability of $\mathbb{N}'_k$ can be written as,

$$p'_k = \mathbb{P}[\mathbb{N}'_k] = \frac{2^{d-k} p_k}{\sum_{j=1}^{d} 2^{d-j} p_j} = \frac{2^{-k} p_k}{\sum_{j=1}^{d} 2^{-j} p_j}$$

Let us define β as the proportion of points in *all* $\mathbb{N}'_k$ that also violate Lipschitzness in their unmasked form. This leads us to the following equation for β

$$\beta = \frac{\sum_{k=1}^{d} 2^{-k} p_k \gamma_k}{\sum_{j=1}^{d} 2^{-j} p_j}$$

The worse case β would then be obtained by considering a maximization over γ_k,

$$\beta^* = \max_{\gamma_1,\ldots,\gamma_d} \frac{\sum_{k=1}^{d} 2^{-k} p_k \gamma_k}{\sum_{j=1}^{d} 2^{-j} p_j}, \sum_{i=1}^{d} p_i \gamma_i = \alpha, 0 \leq \alpha \leq 1, 0 \leq \gamma_i \leq 1, \forall i = 1, \ldots, d \tag{18}$$

This constrained optimization problem can be solved by assigning $\gamma_k = 1$ for the largest p_k until the budget α is exhausted where only a fractional value of γ can be assigned, and 0 for the remaining values of k. This $\beta^* \geq \alpha$ in general. In the specific case where $p_k \to 0$ for $k = 2, \ldots, d$, when compared to p_1 (i.e. where the probability of sampling a point from $\mathcal{D}$ such that any of the values are *exactly* 0 is very small compared to the probability of sampling points with all non-zero values which would generally be the case for sampling real data), $\beta^* \to \alpha$ □

A.2 Theorem 2

Proof. By considering another point x' such that $d_p(x, x') \leq r$ and (12) we get,

$$d_p(\phi(x_i), \phi(x'_i)) = d_p(f(x) - f(x \odot z_{-i}), f(x') - f(x' \odot z_{-i})) \tag{19}$$

using the fact that $d_p(x, y) = ||x - y||_p$ where $||.||_p$ is the p-norm, the RHS gives us,

$$d_p(\phi_i(x), \phi_i(x')) = ||f(x) - f(x \odot z_{-i}) - f(x') + f(x' \odot z_{-i})||_p \tag{20}$$

using triangular inequality,

$$d_p(\phi_i(x), \phi_i(x')) \leq ||f(x) - f(x')||_p + ||f(x' \odot z_{-i}) - f(x \odot z_{-i})||_p \tag{21}$$

w.l.o.g assuming the first term on the right is bigger than the second term

$$d_p(\phi_i(x), \phi_i(x')) \leq 2||f(x) - f(x')||_p = 2d_p(f(x), f(x')) \tag{22}$$

using the fact that f is probabilistic Lipschitz get us,

$$\mathbb{P}[d_p(\phi_i(x), \phi_i(x')) \leq 2Ld_p(x, x')] \geq 1 - \alpha \tag{23}$$

to conclude the proof note that $d_p(\phi(x), \phi(x')) \leq \sqrt[p]{d} * \max_i d_p(\phi_i(x), \phi_i(x'))$, which gives us,

$$\mathbb{P}[d_p(\phi(x), \phi(x')) \leq 2\sqrt[p]{d}L \cdot d_p(x, x')] \geq 1 - \alpha \tag{24}$$

□

A.3 Theorem 3

Proof. Given input x and another input x' s.t. $d(x, x') \leq r$, using (14) we can write

$$\begin{aligned} d_p(\phi_i(x), \phi_i(x')) &= d_p(\mathbb{E}_{p(z|z_i=1)}[f(x \odot z)], \mathbb{E}_p(z|z_i = 1)[f(x' \odot z)]) \\ &= ||\mathbb{E}_{p(z|z_i=1)}[f(x \odot z)] - \mathbb{E}_p(z|z_i = 1)[f(x' \odot z)]||_p \\ &= ||\mathbb{E}_{p(z|z_i=1)}[f(x \odot z) - f(x' \odot z)]||_p \end{aligned} \tag{25}$$

Using Jensen's inequality on R.H.S,

$$d_p(\phi_i(x), \phi_i(x')) \leq \mathbb{E}_{p(z|z_i=1)}[||f(x \odot z) - f(x' \odot z)||_p] \tag{26}$$

Using the fact that $E[f] \leq \max f$,

$$\begin{aligned} d_p(\phi_i(x), \phi_i(x')) &\leq \max_z ||f(x \odot z) - f(x' \odot z)||_p \\ &= \max_z d_p(f(x \odot z), f(x' \odot z)) \end{aligned} \tag{27}$$

Using the fact that f is deterministically Lipschitz with some constant $L \geq 0$, and $d_p(x \odot z, x' \odot z) \leq d_p(x, x'), \forall z$. Then using the definition of probabilistic Lipschitz with $\alpha = 0$ we get,

$$\mathbb{P}(\max_z d_p(f(x \odot z), f(x' \odot z)) \leq L * d(x, x') \geq 1 \tag{28}$$

Using this in (27) gives us,

$$\mathbb{P}[d_p(\phi_i(x), \phi_i(x')) \leq L * d(x, x')] \geq 1 \tag{29}$$

Note that (29) is true for each feature $i \in \{1, ..., d\}$. To conclude the proof note that $d_p(\phi(x), \phi(x') \leq \sqrt[p]{d} * \max_i d_p(\phi_i(x), \phi_i(x'))$. Utilizing this with (29) leads us to

$$\mathbb{P}[d_p(\phi(x), \phi(x') \leq \sqrt[p]{d}L \cdot d_p(x, x')] \geq 1 \tag{30}$$

Since $\mathbb{P}[d_p(\phi(x), \phi(x') \leq \sqrt[p]{d}L \cdot d_p(x, x')]$ defines $A_{\lambda,r}$ for $\lambda \geq \sqrt[p]{d}L$, this concludes the proof. □

	2layer		4layer		linear		svm	
Datasets	Train	Test	Train	Test	Train	Test	Train	Test
CIFAR10	.435	.416	.457	.443	.385	.375	.526	.512
MNIST	.939	.940	.968	.969	.907	.910	.983	.980

Table 3: Train and Test accuracy for different classifiers used in the experiments in Section 5.2.

B DATASET DETAILS

- **Orange-skin**: The input data is again generated from a 10-dimensional standard Gaussian distribution. The ground truth class probabilities are proportional to $\exp\{\sum_{i=1}^{4} X_i^2 - 4\}$. In this case the first 4 features are important globally for *all* data points.
- **Nonlinear-additive**: Similar to *Orange-skin* dataset except the ground trugh class probabilities are proportional to $\exp\{-100 \sin 2X_1 + 2|X_2| + X_3 + \exp\{-X_4\}\}$, and therefore each of the 4 important features for prediction are nonlinearly related to the prediction itself.
- **Switch**: This simulated dataset is specifically for instancewise feature explanations. For the input data feature X_1 is generated by a mixture of Gaussian distributions centered at ±3. If X_1 is generated from the Gaussian distribution centered at +3, X_2 to X_5 are used to generate the prediction probabilities according to the *Orange skin* model. Otherwise X_6 to X_9 are used to generate the prediction probabilities according to the *Nonlinear-additive* model.
- **Rice** (Cinar and Koklu, 2019):This dataset consists of 3810 samples of rice grains of two different varieties (*Cammeo* and *Osmancik*). 7 morphological features are provided for each sample.
- **Telescope**(Ferenc et al., 2005): This dataset consists of 19000+ Monte-Carlo generated samples to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. Each sample is labelled as either background or gamma signal and consists of 10 features.

C TRAINING DETAILS

Training splits and hyperparameter choices have relatively little effect on our experiments. Regardless, the details used in results shown are provided here for completeness:

- **Train/Test Split:** For all synthetic datasets we use 10^6 training points and 10^3 test points. The neural networks classifiers were trained with a batch size of 1000 for 2 epochs. While SVM was trained with default parameters used in `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html`.

  For Telescope and Rice datasets test set sizes of 5% and 33% were used, with a batch size of 32 trained for 100 epochs. SVM was again trained with default parameters.
- **radius** r**:** For all experiments we used radius equal to the median of pairwise distance. This is standard practice and also allows for a big enough r where we can sample enough points to provide empirical estimates.
- **Classifier details:** We train the following four classifiers; **2layer** : A two-layer MLP with ReLU activations. For simulated datasets each layer has 200 neurons, while for the four real datasets we use 32 neurons in each layer. **4layer**: A four-layer MLP with ReLU activations, with the same number of neurons per layer as *2layer.* **linear**: A linear classifier. **svm**: A support vector machine with Gaussian kernel

D ADDITIONAL RESULTS

Table 5 shows the normalized AUC for the estimated explainer astuteness and the predicted AUC based on the predicted lower bound curve. As expected the predicted AUC lower bounds the estimated AUC.

Figure 5 shows the same plots as shown in Figure 3 but includes all datasets.

	Regularized High		Regularized Low		Not Regularized	
Datasets	Train	Test	Train	Test	Train	Test
CIFAR10	.338	.328	.518	.491	.529	.497
MNIST	.788	.809	.951	.969	.990	.976

Table 4: Train and Test accuracy for different classifiers used in the experiments in Section 5.1.

Table 5: **Observed AUC and (Predicted AUC)**. The observed AUC is lower bounded by the predicted AUC and so the observed AUC should always be higher than the predicted AUC. The AUC values are normalized between 0 and 1.

	2layer				4layer				linear				svm			
Datasets	**SHAP**	**RISE**	**CXP**	**(LB)**	**SHAP**	**RISE**	**CXP**	**(LB)**	**SHAP**	**RISE**	**CXP**	**(LB)**	**SHAP**	**RISE**	**CXP**	**(LB)**
OS	.954	.847	.920	(.369)	.969	.896	.906	(.480)	.994	.967	.994	(.950)	.945	.813	.917	(.184)
NA	.978	.909	.936	(.618)	.981	.926	.940	(.696)	.972	.912	.994	(.520)	.971	.883	.937	(.229)
Switch	.998	.996	.948	(.945)	.996	.988	.948	(.909)	.994	.978	.988	(.950)	.969	.885	.936	(.412)
Rice	.962	.886	.974	(.803)	.932	.824	.932	(.793)	.968	.901	.962	(.800)	.981	.906	.970	(.715)
Telescope	.962	.863	.954	(.637)	.955	.863	.944	(.610)	.980	.906	.967	(.756)	.969	.909	.972	(.467)
CIFAR10	.994	.898	.998	(.661)	.992	.866	.998	(.642)	.994	.888	.998	(.653)	.996	.990	.919	(.661)
MNIST	.994	.853	.998	(.553)	.992	.826	.988	(.469)	.993	.842	.986	(.538)	-	-	-	(-)

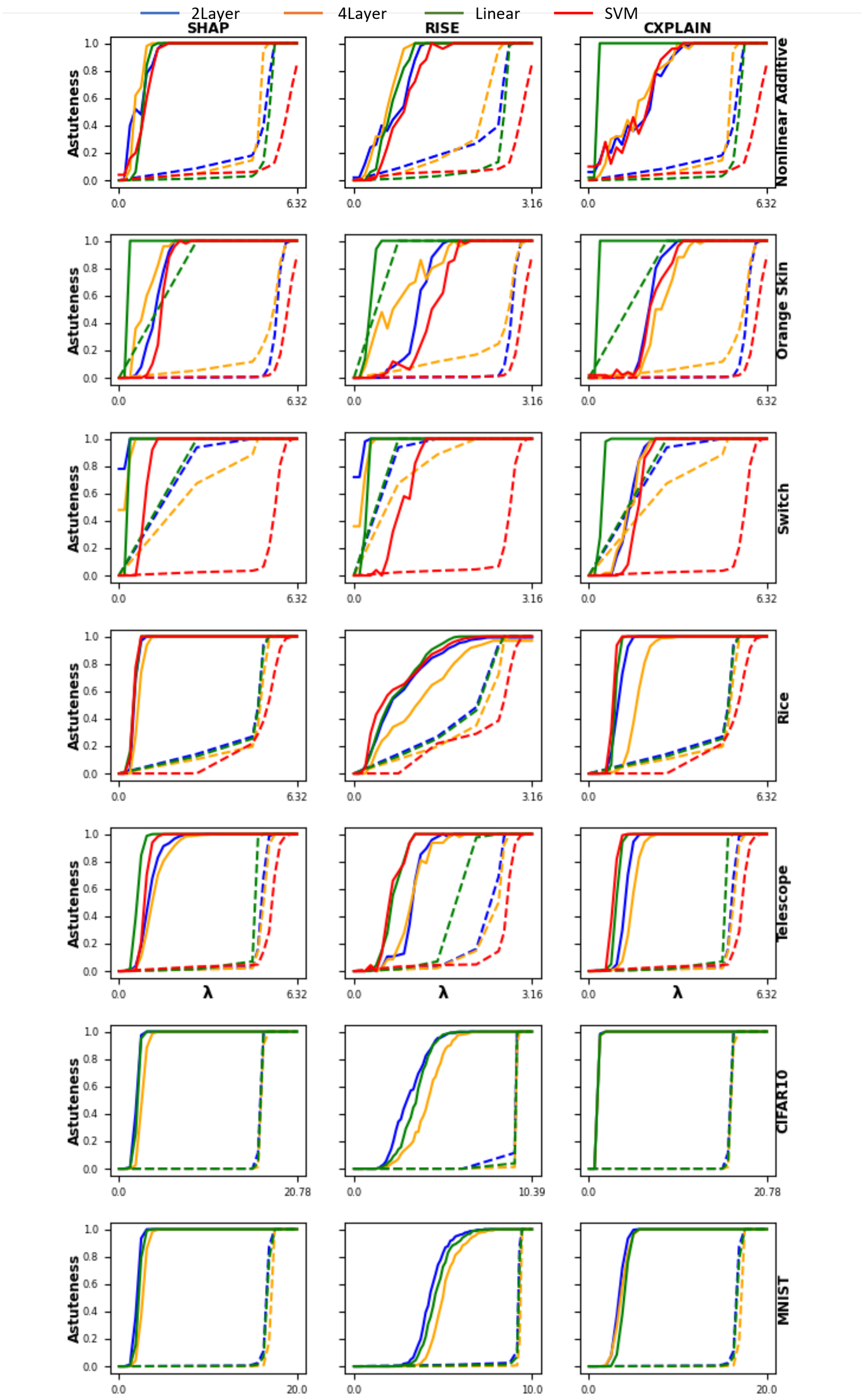

Figure 5: This figure experimentally shows the implication of our theoretical results. It corresponds to the AUC values shown in Table 1. Given each combination of dataset, classifier and explainer we observe that the estimated explainer astuteness for SHAP, RISE and CXPLAIN is lower bounded by the astuteness predicted by our theoretical results given a value of λ. The predicted lower bound is depicted by dashed lines, while solid lines depict the actual estimate of explainer astuteness.

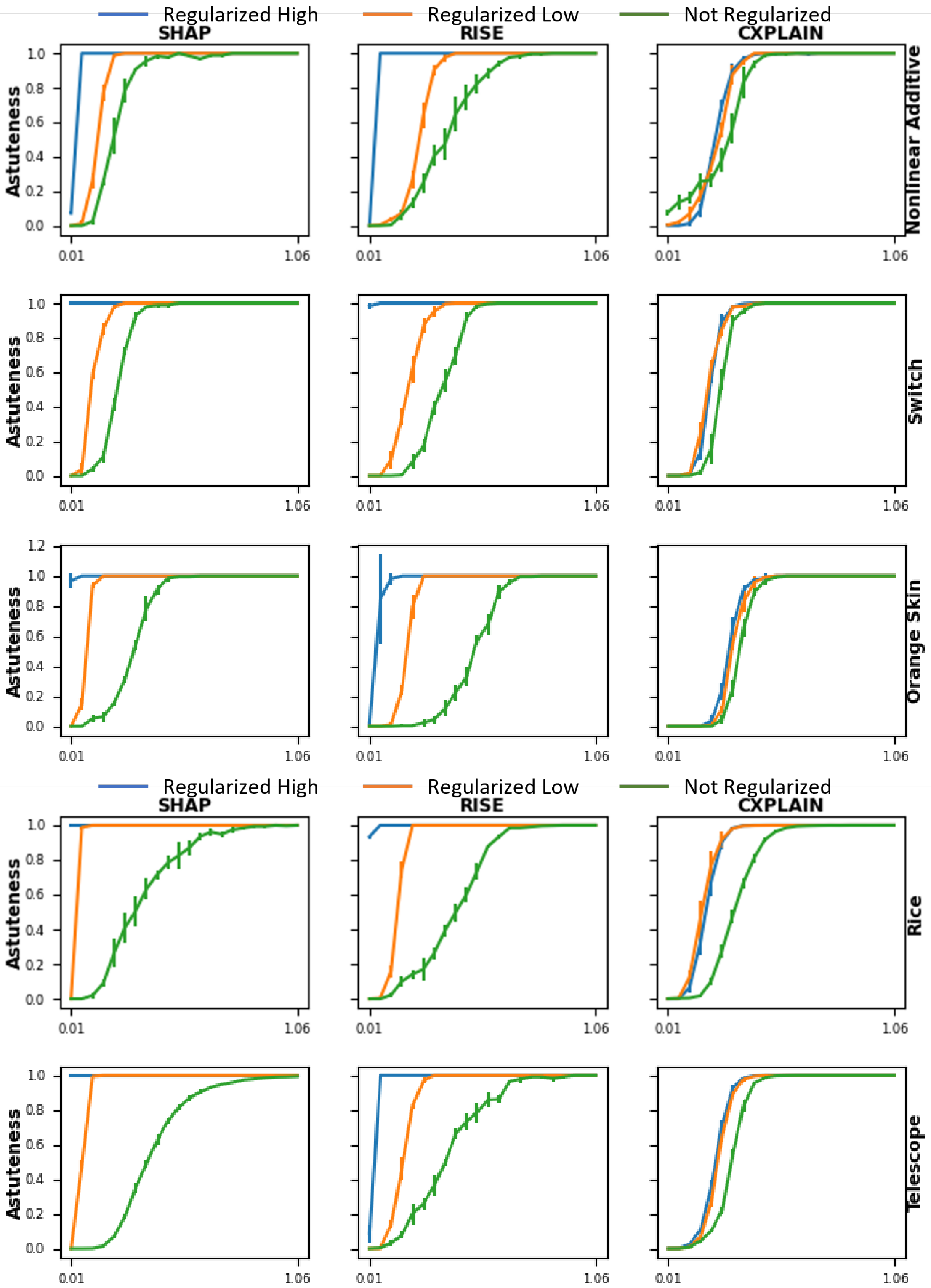

Figure 6: Regularizing the Lipschitness of a neural network during training results in higher astuteness for the same value of λ. Higher regularization results in lower Lipschitz constant (Gouk et al., 2021). Astuteness reaches 1 for smaller values of λ with Lipschitz regularized training, as expected from our theorems. The errorbars represent results across 5 runs to account for randomness in training.