Feature Importance via Sets of Locally Performant Linear Models

Fatemeh Tohidian
Northeastern University

Davin Hill
Northeastern University

Aria Masoomi
Northeastern University

Peter J. Castaldi
Brigham and Women's Hospital

Jennifer Dy
Northeastern University

Abstract

Understanding the contribution of individual features to a model's prediction is critical in applications such as medicine. While feature importance methods aim to quantify how much a feature contributes to a model's accuracy, they often overlook heterogeneous patterns in the data and suffer from limited robustness. We propose ℓ-MCR, a local feature importance method that identifies meaningful neighborhoods around a point of interest, regions where the model or data behavior is locally stable and interpretable. Within these neighborhoods, we estimate feature importance using Model Class Reliance (MCR), which offers robustness by considering the full set of near-optimal models. We also provide a consistency proof for reliably detecting such neighborhoods. Experiments on both synthetic and real-world datasets demonstrate that ℓ-MCR captures localized feature importance patterns that global approaches fail to detect.

1 INTRODUCTION

Explainable AI (XAI) methods have become increasingly used to develop insights into real-world phenomena in a variety of fields Novakovsky et al. (2023); Reichstein et al. (2019); Chaddad et al. (2023). In particular, feature importance methods can be used to understand the latent data generating process (DGP) for a set of observed samples (Freiesleben et al., 2024; Chen et al., 2020).

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

One popular approach is to train a high-performing prediction model to approximate the DGP then use the model as a surrogate to understand the underlying phenomenon. Modern prediction models can often be nonlinear and complex in order to model heterogeneous patterns in the data; local feature importance methods capture the heterogeneity by approximating the prediction model in a local neighborhood around a given sample (Guidotti et al., 2018; Molnar, 2020). This property allows users to infer feature importance for different subpopulations or samples in the data. However, existing local methods can lack robustness (Agarwal et al., 2022; Fel et al., 2022; Alvarez-Melis and Jaakkola, 2018; Khan et al., 2024), especially with respect to understanding explaining underlying phenomenon (Woerl et al., 2023; Fel et al., 2022; Ye and Farnia, 2025). In addition, local methods are calculated on a defined neighborhood around the given sample, yet generally do not provide information regarding the region or locality of the resulting feature importance values (Watson, 2022; Hill et al., 2025), nor take into account the local goodness-of-fit with respect to the prediction model.

Alternatively, rather than considering a single surrogate prediction model for the DGP, one can consider the entire set of well-performing models, known as the *Rashomon Set* (Breiman, 2001). This approach acknowledges that there may be many models that can explain the observed data equally well, and therefore we should not rely on a single model to approximate the DGP. Feature importance methods based on the Rashomon set characterize the level of feature importance within the set of models, also known as *Model Class Reliance* (MCR) (Fisher et al., 2019). MCR provides a more robust measure of feature importance by returning a range of model reliance values, representing how important the feature is to the entire class of models. While recent works have extended the MCR framework to other model classes (Zhong et al., 2023; Xin et al., 2022; Smith et al., 2020), they

only focus on classes of *global* prediction modes. Deriving feature importance from sets of global models can mask local heterogeneity in the data (Figure 1), sometimes referred to as *aggregation bias* (Mehrabi et al., 2021), which limits its applicability in many real-world datasets.

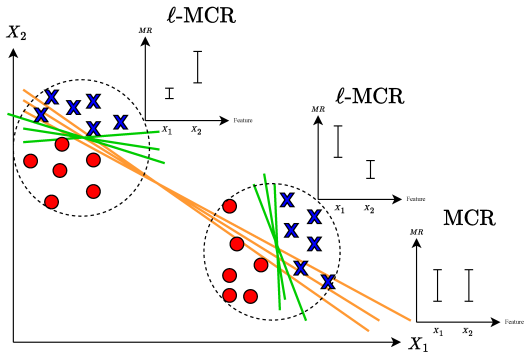

Figure 1: Conceptual comparison of ℓ-MCR and global MCR. While global MCR suggests equal importance across features, ℓ-MCR identifies smaller neighborhoods where feature contributions differ significantly, revealing local patterns that are masked at the global level.

In this work we propose ℓ-MCR, a feature importance method that combines the benefits of local feature importance methods with improved robustness and uncertainty quantification of capturing the Rashomon set of potential models. In particular, we define a principled method to identify local neighborhoods around a given sample with well-performing Rashomon sets of linear models. Specifically, we find the largest neighborhood around the point such that further enlarging the neighborhood would degrade model performance beyond a prespecified threshold. We then find the model class reliance for the local Rashomon set, which provides a robust measure of feature importance. This approach offers three key advantages: (1) it maintains the ability to capture heterogeneous patterns across different subpopulations, (2) it provides robustness guarantees by considering multiple well-performing models rather than a single surrogate, and (3) it explicitly defines the region of validity for each explanation, addressing a critical limitation of existing local methods. Therefore, ℓ-MCR improves the ability for practitioners to understand and infer feature importance values for complex DGPs and highly heterogeneous data.

In summary, we have the following main contributions:

- We define ℓ-MCR, which evaluates the Rashomon set of linear models on local neighborhoods of a given sample.
- We prove the empirical estimator of ℓ-MCR is a consistent estimator, enabling reliable estimation of model validity regions from finite samples.
- Empirical results on tabular and synthetic datasets validate the ability of ℓ-MCR to capture robust, locally linear explanations to better understand variable importance.

2 RELATED WORKS

Rashomon Sets. Rashomon sets (Breiman, 2001) have been adapted for a variety of machine learning tasks and problems (Marx et al., 2020-07-13/2020-07-18; Hsu and Calmon, 2022; Tulabandhula and Rudin, 2014; Nguyen et al., 2025). In particular, Rashomon sets have been used to improve feature importance. Fisher et al. (2019) propose Model Class Reliance, which was extended to include General Additive Models (Zhong et al., 2023) and tree models (Xin et al., 2022; Smith et al., 2020). Laberge et al. (2023) combine feature attribution rankings over Rashomon sets using partial orders. Donnelly et al. (2023) improve the variable importance stability over sampled Rashomon sets.

Local Interpretability Methods. Many works have been proposed for deriving local feature importance values with respect to a given prediction model (Guidotti et al., 2018; Molnar, 2020). In particular, we focus on local surrogate models (Lundberg and Lee, 2017; Ribeiro et al., 2016b), which train a simpler model to approximate the prediction model. Plumb et al. (2018) trains local linear models as surrogates, combined with Random Forest to determine the surrogate neighborhood. Gradient-based methods (Shrikumar et al., 2016, 2017; Simonyan et al., 2014; Sundararajan et al., 2017) can also be seen as local linearizations for the prediction model. Agarwal et al. (2021) show equivalency between gradient-based methods and perturbation-based methods such as LIME (Ribeiro et al., 2016b). Recent works have also incorporated sets of explanations to improve robustness or provide uncertainty estimates. Decker et al. (2024) takes a convex combination of explanations to improve explanation faithfulness. Bayesian approaches (Slack et al., 2021; Zhao et al., 2021; Hill et al., 2024; Chau et al., 2024) consider distributions of explanations using priors. Gradient smoothing methods (Smilkov et al., 2017; Torop et al., 2024) have been shown to improve explanation generalization and robustness (Agarwal et al., 2021; Ye and Farnia, 2025)

Regional Explanations. Regional explanations provide explanations that generalize over a defined neighborhood or region in the feature space. Some methods have combined local explanations through averaging (Yoon et al., 2019; Masoomi et al., 2022) or clustering (Lundberg et al., 2020; Gramegna and Giudici, 2020) to generate regional or global explanations. Ribeiro et al. (2018) provides bounds on where its rule-based explanations are valid. Several methods partition the feature space using tree-based models (Hu et al., 2020; Molnar et al., 2024), or by identifying regions that minimize feature interactions (Laberge et al., 2024; Herbinger et al., 2022, 2024); After partitioning, the individual regions can be explained using simpler surrogate models.

In contrast with existing local or regional explanations, ℓ-MCR defines a local neighborhood by considering the entire set of locally-performant linear models which could plausibly have generated the data; this allows for a more robust explanation of the underlying data generating process.

3 BACKGROUND

In this section we provide a background for Rashomon Sets (Breiman, 2001) and Model Class Reliance (Fisher et al., 2019). We first establish some preliminary notation. Let $Z = (X, Y)$ be a random variable with outcome Y and covariates X. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be the support of X. We separate the features into two random variables: $X_1 \in \mathcal{X}_1$ is the set of features we are interested in measuring the importance of, and $X_2 \in \mathcal{X}_2$ are the remaining features. X_1 and X_2 together form all of the features $X = (X_1, X_2) \in \mathcal{X}$. We specify the set $\mathcal{F} \subseteq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$, as the class of functions we are investigating with respect to X_1. In order to compare the performance of models within this set, we refer to a non-negative loss function L as a function that takes a dataset or a random variable and a prediction model f as input, evaluates the performance of f on the dataset, and return a non-negative value as the loss.

3.1 Rashomon Set and Model Class Reliance

The Rashomon set $\mathcal{R}$ consists all near-optimal models for a given model class $\mathcal{F}$ while evaluating the performance on a given data distribution Z. The term "near optimal" is defined by allowing a maximum expected loss ε. Analogously, the empirical Rashomon set $\hat{\mathcal{R}}$ can be defined as the set of all near-optimal models with respect to the best-performing reference model and the observed samples. Here, instead of the expected loss, we evaluate the average sample loss to decide to include a model in the empirical Rashomon set or not:

$$\underbrace{\mathcal{R}(\varepsilon, \mathcal{F}, Z) = \{f \in \mathcal{F} \mid \mathbb{E}_Z\left[L(f, Z)\right] \leq \varepsilon\}}_{\text{Rashomon Set}} \quad (1)$$

$$\underbrace{\hat{\mathcal{R}}(\varepsilon, \mathcal{F}, \mathbf{Z}) = \{f \in \mathcal{F} \mid \frac{1}{k}\sum_{i=1}^{k} L(f, \mathbf{Z}_i) \leq \varepsilon\}}_{\text{Empirical Rashomon Set}} \quad (2)$$

Note that $\hat{\mathcal{R}}(\varepsilon, \mathcal{F}, \mathbf{Z})$ still includes all of the well-performing models, but the performance of the models is evaluated on the observed samples rather than the distribution. Model reliance for a given set of features can be quantified in several ways. The goal is to measure the effect of a pre-specified set of features, X_1, on the performance of a model. We follow the definition in Fisher et al. (2019), as the expected loss of the model when noise is added to X_1 in such a way that X_1 becomes completely uninformative of y. Assume that we have two independent random variables $Z^{(a)} = ((X_1{}^{(a)}, X_2{}^{(a)}), y^{(a)})$ and $Z^{(b)} = ((X_1{}^{(b)}, X_2{}^{(b)}), y^{(b)})$ that share the same distribution as Z. We define *switch error* as the expected loss of the model when we use y and X_2 from $Z^{(b)}$, but switch the actual value of $X_1^{(b)}$ with $X_1^{(a)}$. The empirical version can be defined as the empirical mean of the same random variable:

$$\hat{e}_{\text{switch}}(f) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} L(f, ((\mathbf{X_1}[j], \mathbf{X_2}[i]), \mathbf{y}[i]) \quad (3)$$

As reference, we also need to calculate the expected error on the original data distribution. For the empirical version we follow the same path as before by changing the expected value to the empirical expected value.

$$\hat{e}_{\text{orig}}(f) = \frac{1}{n} \sum_{i=1}^{n} L(f, ((\mathbf{X_1}[i], \mathbf{X_2}[i]), \mathbf{y}[i]) \quad (4)$$

Finally, the model reliance can be defined as the ratio of switch error to the original error. If the reliance of a model f for a set of features X_1 is high, it indicates that switching X_1 has a high negative effect on the performance of f in predicting y. Conversely, low model reliance indicates that, even after switching $\mathbf{X}_1$ values, the performance of f does not change significantly, and therefore $\mathbf{X}_1$ is not important for f for predicting y. We can define empirical model reliance as follows:

$$\hat{MR}(f) = \frac{\hat{e}_{\text{switch}}(f)}{\hat{e}_{\text{orig}}(f)} \quad (5)$$

Model class reliance indicates the range of possible model reliances for a certain class of models. It is

an interval showing the minimum and maximum value of MR for the models in the Rashomon set. If both the lower and upper bound of model class reliance are low, we can conclude that none of the well performing models in $\mathcal{F}$ rely on X_1. When both lower and upper bound of MCR, this indicates that all of the well performing model heavily rely on X_1. A large range between upper and lower bounds indicates that there are well performing models in $\mathcal{F}$ that rely on X_1 and also there models in $\mathcal{F}$ that can predict y with a high performance even when X_1 becomes uninformative of y.

$$\hat{MCR}(\varepsilon, \mathcal{F}, \mathbf{Z}) = \left[\hat{MCR}_-, \hat{MCR}_+\right] = \left[\min_{f \in \hat{\mathcal{R}}(\varepsilon,\mathcal{F},\mathbf{Z})} \hat{MR}(f), \max_{f \in \hat{\mathcal{R}}(\varepsilon,\mathcal{F},\mathbf{Z})} \hat{MR}(f)\right] \quad (6)$$

4 MODEL RELIANCE FOR LOCAL LINEAR MODELS

In this section, we introduce our method, ℓ-MCR, which jointly identifies a neighborhood around a point of interest and estimates feature importance within that neighborhood. Unlike traditional global feature importance techniques, our approach focuses on identifying an appropriate locality where the model's behavior can be reliably analyzed. Specifically, we estimate local feature importance by finding the largest neighborhood around the point such that further enlarging the neighborhood would degrade model performance beyond a prespecified threshold. This enables ℓ-MCR to detect the regions where the data distribution or model behavior changes significantly, while still leveraging the advantages of Model Class Reliance (MCR), which estimates feature importance by considering a set of near-optimal models rather than relying on a single model. Since achieving high performance using simple models on the full dataset is often infeasible, our method seeks smaller, more stable regions where the model is locally explainable. However, identifying such regions is itself a key challenge. In Section 4.1, we formally define the notion of neighborhood and ℓ-MCR. Section 4.2 provides a proof that the empirical maximum explainable radius converges to the true maximum explainable radius as sample size increases. Section 4.3 presents our algorithm for estimating ℓ-MCR.

4.1 Defining the Local Neighborhood

Given a point of interest, $c \in \mathcal{X}$, our goal is to estimate the local feature importance in the vicinity of c. Identifying a neighborhood where the data or model behavior is sufficiently simple to be captured by an interpretable model class, such as linear models, enables us to uncover nuanced patterns and subtypes that global feature importance methods often overlook. Selecting the appropriate shape for neighborhoods is critical for obtaining reliable and meaningful results in spatial analysis and local feature attribution. The optimal neighborhood structure often depends on the specific characteristics of the data and the underlying task. First, we define neighborhoods as closed r-balls centered at c:

$$B_r(c) = \{x \in \mathcal{X} \mid \|x - c\|_p \leq r\} \quad (7)$$

Let $Z_{B_r(c)}$ denote the restriction of the distribution Z to the domain $B_r(c)$, with corresponding covariates $X_{B_r(c)}$ and outcomes $Y_{B_r(c)}$. Thus, the density function of $X_{B_r(c)}$ is given by:

$$f_{X_{B_r(c)}}(x) \propto \begin{cases} f_X(x) & x \in B_r(c) \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Such distribution can be constructed for any radius $r > 0$. Among all these neighborhoods parameterized by r, we are interested in finding r_c^*, the largest radius with nonempty Rashomon set, such that for all smaller neighborhoods the Rashomon set is nonempty as well. In other words, we progressively enlarge the neighborhood, starting from a small radius, and track the point beyond which no model in our class can achieve loss below a threshold. The value r_c^* identifies the boundary beyond which the local interpretability of the model breaks down. As long as there exists at least one well-performing model in each smaller neighborhood, we consider the locality stable and reliably explainable. However, it is generally infeasible to calculate the entire Rashomon set $\mathcal{R}(\varepsilon, \mathcal{F}, Z_{B_r(c)})$. The following lemma formalizes the condition under which the Rashomon set is nonempty:

Lemma 1. *$\mathcal{R}\left(\varepsilon, \mathcal{F}, Z_{B_r(c)}\right) \neq \varnothing$ if and only if $\min_{f \in \mathcal{F}} \mathbb{E}_{Z_{B_r(c)}}\left[L\left(f, Z_{B_r(c)}\right)\right] \leq \varepsilon$.*

Proof details are provided in the supplementary material. Lemma 1 tells us that the Rashomon set being nonempty is equivalent to the existence of a model in the class that performs within the error threshold. This gives us a simple and intuitive stopping criterion for neighborhood expansion, as many loss functions and model classes admit efficient algorithms for finding the model with minimal expected loss. The choice of ε directly influences the importance intervals. In practice, this parameter can often be guided by domain knowledge. For instance, in healthcare, ε might reflect a clinically acceptable margin of error for tasks like disease detection.

Definition 1 (Maximum Explainable Radius r_c^*)**.** Given a point of interest c and a data distribution Z,

define the set of explainable radii as:

$$R_c = \Big\{ r \in \mathbb{R}^+ \;\Big|\; \forall s \in \mathbb{R}^+,\, s \le r : \min_{f \in \mathcal{F}} \mathbb{E}_{Z_{B_s(c)}}\big[L\big(f, Z_{B_s(c)}\big)\big] \le \varepsilon \Big\} \quad (9)$$

We define the *maximum explainable radius* r_c^* as the largest radius corresponding to a neighborhood around the point c in which every smaller or equal neighborhood admits at least one model from $\mathcal{F}$ that meets the performance threshold ε:

$$r_c^* = \sup R_c \quad (10)$$

This value, r_c^*, characterizes the largest locality within which accurate and interpretable modeling remains feasible.

In addition we define the empirical maximum explainable radius, similarly but based on k i.i.d. samples $Z_1, \ldots, Z_k$ drawn from $Z_{B_r(c)}$.

Definition 2 (Empirical Maximum Explainable Radius $\hat{r}_{c,k}^*$)**.** Given a point of interest c and a data distribution Z, define the set of empirical explainable radii as:

$$\hat{R}_{c,k} = \Big\{ r \in \mathbb{R}^+ \mid \forall s \in \mathbb{R}^+, s \le r : \min_{f \in \mathcal{F}} \frac{1}{k} \sum_{i=1}^{k} [L(f, Z_i)] \le \varepsilon \Big\} \quad (11)$$

We define the *empirical maximum explainable radius* $\hat{r}_{c,k}^*$ as the largest radius corresponding to a neighborhood around the point c in which every smaller or equal neighborhood admits at least one model from $\mathcal{F}$ that meets the performance threshold ε on the drawn samples:

$$\hat{r}_{c,k}^* = \sup \hat{R}_{c,k} \quad (12)$$

4.2 Consistency

In practice, we do not have access to the full data distribution Z and must rely on a finite sample drawn from it. This raises a natural question: does the empirical estimation converges to the true maximum explainable radius? Theorem 2 answers this question when our class of models is linear and the loss function is squared loss. There is an additional assumption:

Assumption 1.1. *There exists a constant $B \in \mathbb{R}$ such that $\forall f \in \mathcal{F}$ we have $0 \le L(f, z) \le B$ for any $r \ge 0$ and $z \in Z_{B_r(c)}$.*

Assumption 1.1, states that the random variable corresponding to the error of any model on the neighborhood distribution must be bounded. This assumption is also used in prior MCR work Fisher et al. (2019), on the whole data distribution.

Theorem 2. *Let $\mathcal{G}$ be the class of linear models and let $L(g, (X, Y)) = (g(X) - Y)^2$ be the squared loss. Let $\hat{r}_{c,k}^*$ denote the empirical estimate of the maximum explainable radius, using k i.i.d. samples from $Z_{B_r(c)}$, and let r_c^* be the true maximum explainable radius. Under Assumption 1.1, $\hat{r}_{c,k}^*$ converges in probability to r_c^* as $k \to \infty$.*

Proof details are presented in the supplementary material. Theorem 2 proves that the empirical radius $\hat{r}_{c,k}^*$ indeed converges to the true radius r_c^*.

4.3 ℓ-MCR Approximation

In practice, it is generally infeasible to directly calculate Eq. 10 since the Rashomon set is typically infinite. In this section, we propose Algorithm 1, a practical procedure for estimating the maximum explainable radius r_c^* using a candidate set of radii $\mathcal{R}_{\text{cand}}$. For each radius, we empirically estimate the expected loss of the best-performing model in the class using k i.i.d. samples $Z_1, \ldots, Z_k$ drawn from $Z_{B_r(c)}$.

Algorithm 1 Best neighborhood search

Require: Point of interest $c \in \mathcal{X}$, Threshold ε, List of radii $\mathcal{R}_{\text{cand}} = [r_i]_{i=1}^s$
Return: An estimation of Maximum explainable radius $\tilde{r}_{c,k}$
$\tilde{r}_{c,k} =$ None
for r in $\mathcal{R}_{\text{cand}}$ **do**
  $Z_1, \ldots Z_k \overset{\text{iid}}{\sim} Z_{B_r(c)}$
  $\hat{l}^* = \min_{f \in \mathcal{F}} \frac{1}{k} \sum_{i=1}^{k} L(f, Z_i)$
  if $\hat{l}^* \le \varepsilon$ **then**
    $\tilde{r}_{c,k} = r$
  else
    break
  end if
end for
return $\tilde{r}_{c,k}$

At each radius, we check whether the minimum empirical loss of the best model in the class, remains below the threshold ε, as motivated by Lemma 1.

Once this stable neighborhood has been identified, we quantify the range of variable importance within the local Rashomon Set. To do this, we propose to use the Model Class Reliance metric constrained to the identified neighborhood.

Definition 3 (ℓ-MCR)**.** Given a center point c and a data distribution Z, local model class reliance can be defined as follows:

$$\ell\text{-MCR}(\varepsilon, \mathcal{F}, Z, c) = MCR(\varepsilon, \mathcal{F}, Z_{B_{r_c^*}(c)}) \quad (13)$$

Therefore, ℓ-MCR provides a robust, localized esti-

mate of feature importance that respects both model fidelity and locality.

5 EXPERIMENTS

In this section, we evaluate the effectiveness of ℓ-MCR in capturing local feature importance. We conduct experiments on both synthetic datasets, designed for interpretability, and real-world datasets. Our primary goal is to assess whether the explanation intervals identified by ℓ-MCR more reliably capture local attribution methods, such as gradients, compared to traditional MCR.

Section 5.1 outlines the shared experimental setup. In Section 5.2, we present an illustrative example where applying ℓ-MCR yields a significantly more informative explanation. Section 5.3 reports a quantitative comparison between ℓ-MCR and global MCR on a structured synthetic dataset. Finally, in Section 5.4, we demonstrate the application of ℓ-MCR to a real-world medical dataset. Moreover, we include evaluations on the COMPAS (Larson et al., 2016) and Breast Cancer datasets (Wolberg and Street, 1993), both commonly studied in the context of Model Class Reliance. The corresponding results are provided in the supplementary material.

5.1 Experimental Setup

Our experiments focus on binary classification tasks using linear models as the interpretable model class and hinge loss as the loss function, except the COMPAS experiment which is a regression task. To evaluate our method, we first train a neural network on the training data, then analyze ℓ-MCR 's behavior on individual points from the test set. Since the gradient of the neural network at a point provides the best local linear approximation to its behavior, we treat this gradient as a reference local explainer. We then evaluate whether the MR associated with this linear approximation falls within the intervals produced ℓ-MCR and MCR. Similarly, we examine whether the MR of the line suggested by LIME (Ribeiro et al., 2016a) is included within these intervals. All experiments were conducted on a machine with an Intel i7 CPU. The implementation and code used in our experiments are available in this repository.

5.2 Illustrative Example

In the simulated dataset, the labels are locally determined by a single feature, either X_1 or X_2 depending on the region of the input space. In other words, within certain neighborhoods, the data is linearly separable using either a horizontal or a vertical decision boundary. This implies that locally, one feature is highly informative while the other is irrelevant.

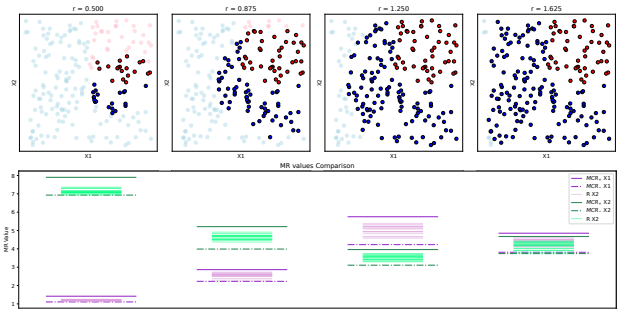

Figure 2: Comparison of local and global MCR intervals on a synthetic binary classification task. The plot on the leftmost shows the neighborhood selected by ℓ-MCR , where the MCR intervals reveal a clear difference in the importance of the two features with X_2 playing a dominant role. The right plot shows the global MCR intervals computed over the entire dataset, which fail to distinguish between the features due to dataset symmetry. Fine lines represent examples of model reliance scores from individual models within the Rashomon set.

However, due to the symmetry of the dataset, this pattern is obscured in the global analysis. As shown in the rightmost plot of Figure 2, the global MCR intervals for both features are nearly identical, offering little insight. In contrast, the leftmost plot highlights the results from a neighborhood selected by ℓ-MCR . In this local region, both upper an lower limits for MR obtained using ℓ-MCR for X_2 are much higher than for X_1, indicating that X_2 plays a significantly more important role in explaining the labels. This example illustrates how ℓ-MCR can detect and reveal locally important features that global methods may miss.

5.3 Synthetic Dataset

We evaluate ℓ-MCR on a synthetic dataset designed with localized decision boundaries. The dataset is two-dimensional, with both features uniformly distributed in the interval $[-1, 1]$. It consists of two distinct regions, each with a different feature that governs the label. In the first region ($X_1 < 0$), labels depend on X_1 ($y = 1\{X_1 > -0.5\}$), while in the second region ($X_1 \geq 0$), labels are determined by X_2 ($y = 1\{X_2 > 0\}$). Therefore, the data is linearly separable using X_1 and X_2 in the first and second regions, respectively . This setup ensures that different features are locally important.

Owing to its structured nature, the dataset can be

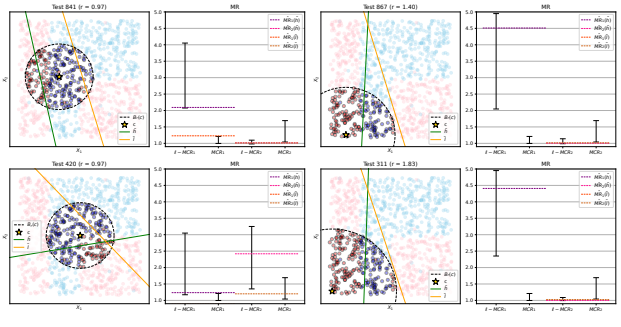

Figure 3: Four test points and their corresponding neighborhoods, as identified by Algorithm 1, are shown with dashed black circles. Blue and red points represent samples with $y = 1$ and $y = 0$, respectively, while pale blue and pale red points lie outside the detected neighborhoods. The experiments use $\varepsilon = 0.45$ for ℓ-MCR and $\varepsilon = 1$ for MCR. Subscripts indicate the feature index for which model reliance is computed. Colored dashed lines show the model reliance of the gradient-based linear approximation and LIME $(\tilde{h}, \tilde{l})$. While ℓ-MCR successfully includes $\tilde{h}$ within its Rashomon set, the global MCR fails to do so. Despite choosing small kernel width, $\tilde{l}$ is still following the global patterns of data.

easily learned by a neural network, resulting in accurate gradient-based explanations across most points. Moreover, its low dimensionality makes it easier to visualize and interpret the behavior of ℓ-MCR . Details of the data generation process and model architecture are provided in the supplementary material. Let $h(x)$ denote the neural network and let $c \in \mathbb{R}^2$ be the point of interest. The gradient-based linear approximation around c is $\tilde{h}(x) = h(c) + \nabla h(c)^\top (x - c)$. LIME (Ribeiro et al., 2016a) offers an alternative local explanation method, constructing linear surrogates $\tilde{l}$ of the black-box model around the point of interest.

For each point in the test set, we compute $\hat{MR}(\tilde{h})$ and $\hat{MR}(\tilde{l})$, the model reliance of the local linear approximation, and check whether it falls within the interval estimated by both ℓ-MCR and Global MCR. Figure 3 highlights four examples where the interval produced by ℓ-MCR contains the gradient-based estimate $\hat{MR}(\tilde{h})$, but the corresponding global MCR interval misses it. Table 1 reports the number of test points where gradient-based feature importance lies within the ℓ-MCR intervals, and compares it against the global MCR intervals.

To empirically validate the theoretical consistency result established in Theorem 2, we examine the convergence behavior of the estimated maximum explainable radius $\hat{r}^*_{c,k}$ as a function of sample size. For three representative test points, we repeatedly subsample the training data at increasing sizes. Figure 4 reports the mean of $\hat{r}^*_c$ across 10 independent subsample at each sample size. As the sample size grows, the estimates converge toward the true radius, known from the data generating process, consistent with the convergence in probability guaranteed by Theorem 2.

Table 1: Comparison of ℓ-MCR and MCR in terms of how often the gradient-based model reliance, $\hat{MR}(\tilde{h})$ and $\hat{MR}(\tilde{l})$, are captured within their respective intervals on the synthetic dataset.

Feature	in ℓ-MCR	in MCR
$\hat{MR}(\tilde{h})_1$	**47.0**%	19.6%
$\hat{MR}(\tilde{h})_2$	**43.3**%	22.5%
$\hat{MR}(\tilde{l})_1$	8.1%	**19.9** %
$\hat{MR}(\tilde{l})_2$	**36.6**%	13.5%

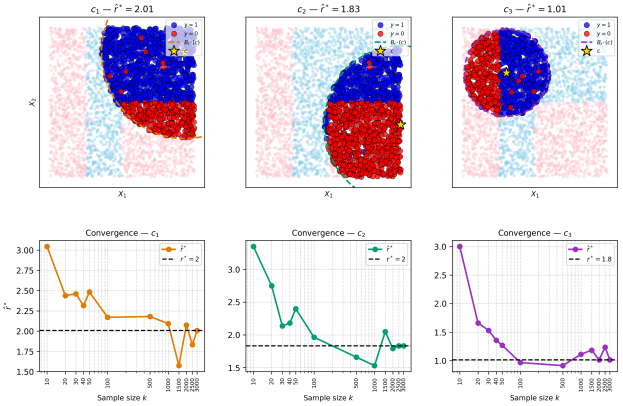

Figure 4: Empirical convergence of the estimated maximum explainable radius for three test points c, c_2, c_3. Top row: training data with the estimated neighborhood at full sample size, where blue and red points indicate the two classes. Bottom row: mean $\hat{r}^*$ across 10 independent subsample at each sample size k, with the true radius r^* shown as a dashed line.

5.4 Diabetes Dataset

To demonstrate the applicability of ℓ-MCR in healthcare, we apply it to the NHANES 2013–2014 dataset (Miller, 1973), a large-scale health and nutrition survey. In medical domains, identifying subpopulations of patients who respond similarly to certain features is particularly valuable. ℓ-MCR enables us to detect such local patterns by isolating regions where data can be explained using an interpretable class of models.

In this setting, our goal is to explore whether ℓ-MCR can highlight patient subgroups where the model's re-

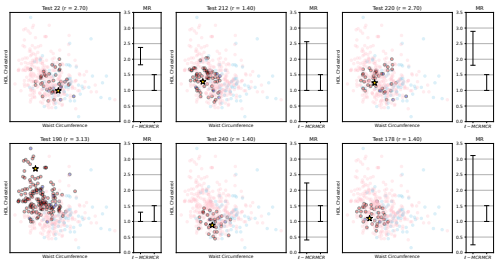

Figure 5: Six patients and their corresponding neighborhoods as identified by Algorithm1. Blue and red points represent patients with previous diabetes diagnosis or Fasting Glucose > 125 mg/dl, while pale-colored points lie outside the detected neighborhoods. Bold points indicate patients within each local neighborhood. Note that although the boundary is circular in feature space, it may not appear circular in the 2D projection. Model reliance is computed with respect to Waist Circumference.

liance on specific features, such as age or blood pressure, significantly differs from global trends. This type of analysis may reveal clinically relevant local behaviors that are masked in global feature importance summaries.

In this experiment, we consider a binary classification task with a richer set of features, including Waist Circumference, HDL Cholesterol, Household Income, Blood Urea Nitrogen, and Kcal Intake. The target variable indicates whether a person has diabetes.

Unlike the synthetic case, our goal here is not to explain a high-performing model. In fact, the fitted neural network does not achieve high accuracy, and its gradient-based explanations may not be reliable. Instead, we apply ℓ-MCR directly to the data to identify meaningful subpopulations. In this context, the value of ℓ-MCR lies in its ability to uncover local regions where a specific feature drives model behavior, even when global trends are unclear or uninformative.

5.5 Discussion

As Table 1 suggests, ℓ-MCR effectively captures the local linear model by accurately estimating the Rashomon set. Figure 3 further illustrates how ℓ-MCR successfully identifies the underlying data-generating process. Notably, in 78.9% of the cases for feature 1 and 81.3% cases for feature 2, where the gradient-based model reliance was included in the ℓ-MCR interval, it was missed by MCR, indicating that local patterns in the data were overlooked by MCR.

Table 2: Comparison of ℓ-MCR and MCR in terms of how often the gradient-based model reliance, $\hat{MR}(\tilde{h})$, is captured within their respective intervals on the NHANES 2013-2014

Feature	in ℓ-MCR	in MCR
Waist Circumference	23%	85%
HDL Cholesterol	18%	0%
Household Income	17%	45%
Blood Urea Nitrogen	18%	48%
Kcal Intake	18%	0%

LIME is designed to fit local linear surrogates around a point of interest for a black-box model. In this setting, we treat model explanations as proxies for data explanations; in the toy example, since we have access to a perfect model, such a proxy is justified. We observed, however, that the surrogate lines produced by LIME were similar across points, and the measured MR fell within the global MCR more frequently than with ℓ-MCR for feature 1. While reducing the kernel width can, in principle, enforce greater locality, this causes the effective number of weighted perturbations to shrink drastically, leading to degenerate regressions. By contrast, our local MCR estimator produces intervals that cover the gradient-based linearization in most cases.

Note that the ε values used for ℓ-MCR and for the global MCR estimation differ. In the local case, a smaller ε is intentionally chosen to identify neighborhoods where the data distribution is especially simple, allowing models to achieve high accuracy. Using such a strict threshold globally often leads to an empty Rashomon set, as no simple model can perform well across the entire dataset.

The results on the COMPAS dataset reveal that caution is needed when interpreting MCR outputs, since the intervals can vary when the data is restricted to certain subpopulations. On the other hand, the whole Breast Cancer dataset can be classified with high accuracy using a logistic regression model; therefore, there's no point in checking for smaller regions where the data is simply interpretable. Further figures and numbers for both experiments can be found in the supplementary material.

6 LIMITATIONS AND CONCLUSION

This work addresses a key limitation of global feature importance methods, namely their inability to capture heterogeneous patterns in data. When subpopulations exist with distinct relationships between

features and outcomes, global methods may obscure important local signals. By identifying local neighborhoods where model behavior is stable and interpretable, ℓ-MCR provides a finer-grained understanding of feature reliance. Application includes, targeting data collection or model improvements toward regions where specific features are critical. Another promising direction is fairness analysis, where one may wish to examine whether feature importance varies across demographic subgroups.

As discussed, our algorithm uses the empirical Rashomon set after finding the neighborhood. In the linear regression case, Fisher et al. (2019) provide a convex optimization formulation to recover the full range of model reliance values. However, for other cases, characterizing the full Rashomon set and computing tight bounds on model reliance remains an open challenge (Li et al., 2024). Future work may also explore more principled ways of defining neighborhoods in high-dimensional settings, where distance-based approaches become less effective due to data sparsity.

Acknowledgements

This work was supported by in part by NIH/NHLBI 5R01HL167072, NIH/NHLBI 5R01HL171213-02, and Northeastern University's iSUPER Impact Engine. The authors thank the reviewers for their constructive comments, which helped improve the quality of this work.

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*, 2022.

Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Steven Wu, and Himabindu Lakkaraju. Towards the Unification and Robustness of Perturbation and Gradient Based Explanations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 110–119. PMLR, July 2021.

Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44 (4):615–631, July 1997. ISSN 0004-5411. doi: 10.1145/263867.263927. URL `https://doi.org/10.1145/263867.263927`.

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of Explainable AI Techniques in Healthcare. *Sensors (Basel, Switzerland)*, 23 (2), January 2023. ISSN 1424-8220. doi: 10.3390/s23020634.

Siu Lun Chau, Krikamol Muandet, and Dino Sejdinovic. Explaining the uncertain: Stochastic shapley values for gaussian process models. *Advances in Neural Information Processing Systems*, 36, 2024.

Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Thomas Decker, Ananta R. Bhattarai, Jindong Gu, Volker Tresp, and Florian Buettner. Provably better explanations with optimized aggregation of feature attributions, 2024.

Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward Browne. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Advances in Neural Information Processing Systems*, 36:6267–6279, 2023.

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.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research : JMLR*, 20: 177, 2019. ISSN 1532-4435 1533-7928.

Timo Freiesleben, Gunnar König, Christoph Molnar, and Alvaro Tejero-Cantero. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *Minds and Machines*, 34(3):32, 2024.

Alex Gramegna and Paolo Giudici. Why to buy insurance? An explainable artificial intelligence approach. *Risks*, 8(137), 2020. ISSN 2227-9091. doi: 10.3390/risks8040137.

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*, 51(5), August 2018. ISSN 0360-0300. doi: 10.1145/3236009.

Julia Herbinger, Bernd Bischl, and Giuseppe Casalicchio. REPID: Regional effect plots with implicit interaction detection. In *International Conference on Artificial Intelligence and Statistics*, pages 10209–10233. PMLR, 2022.

Julia Herbinger, Marvin N. Wright, Thomas Nagler, Bernd Bischl, and Giuseppe Casalicchio. Decomposing global feature effects based on feature interactions. *Journal of Machine Learning Research*, 25 (381):1–65, 2024.

Davin Hill, Aria Masoomi, Max Torop, Sandesh Ghimire, and Jennifer Dy. Boundary-aware uncertainty for feature attribution explainers. In *International Conference on Artificial Intelligence and Statistics*, pages 55–63. PMLR, 2024.

Davin Hill, Joshua Bone, Aria Masoomi, Max Torop, and Jennifer Dy. Axiomatic explainer globalness via optimal transport. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025.

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Linwei Hu, Jie Chen, Vijayan N Nair, and Agus Sudjianto. Surrogate locally-interpretable models with supervised machine learning algorithms. *arXiv preprint arXiv:2007.14528*, 2020.

Zulqarnain Q Khan, Davin Hill, Aria Masoomi, Joshua T Bone, and Jennifer Dy. Analyzing explainer robustness via probabilistic lipschitzness of

prediction functions. In *International Conference on Artificial Intelligence and Statistics*, pages 1378–1386. PMLR, 2024.

Gabriel Laberge, Yann Pequignot, Alexandre Mathieu, Foutse Khomh, and Mario Marchand. Partial order in chaos: Consensus on feature attributions in the rashomon set. *Journal of Machine Learning Research*, 24(364):1–50, 2023.

Gabriel Laberge, Yann Batiste Pequignot, Mario Marchand, and Foutse Khomh. Tackling the XAI disagreement problem with regional explanations. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2017–2025. PMLR, May 2024.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm, May 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`. Accessed: 2024-11-14.

Sichao Li, Amanda S. Barnard, and Quanling Deng. Practical attribution guidance for rashomon sets, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, January 2020. ISSN 2522-5839. doi: 10.1038/s42256-019-0138-9.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, 2020-07-13/2020-07-18.

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 *10th International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *Acm Computing Surveys*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607.

H W Miller. Plan and operation of the health and nutrition examination survey. united states–1971-1973. *Vital and health statistics. Ser*, pages 1–46, 1973.

Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

Christoph Molnar, Gunnar König, Bernd Bischl, and Giuseppe Casalicchio. Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery*, 38(5):2903–2941, 2024.

Simon Nguyen, Kentaro Hoffman, and Tyler McCormick. Unique rashomon sets for robust active learning, 2025.

Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31, 2018.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-0912-1.

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*, KDD '16, page 1135–1144, New York, NY, USA, 2016a. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL `https://doi.org/10.1145/2939672.2939778`.

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*, KDD '16, pages 1135–1144, New York, NY, USA, 2016b. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic

explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR, August 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2014.

Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable Post hoc Explanations Modeling Uncertainty in Explainability. In *Neural Information Processing Systems (NeurIPS)*, 2021.

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

Gavin Smith, Roberto Mansilla, and James Goulding. Model class reliance for random forests. *Advances in Neural Information Processing Systems*, 33:22305–22315, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017.

Max Torop, Aria Masoomi, Davin Hill, Kivanc Kose, Stratis Ioannidis, and Jennifer Dy. SmoothHess: ReLU network feature interactions via stein's lemma. *Advances in Neural Information Processing Systems*, 36, 2024.

Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.

David S. Watson. Conceptual challenges for interpretable machine learning. *Synthese*, 200(2):65, March 2022. ISSN 1573-0964. doi: 10.1007/s11229-022-03485-5.

Ann-Christin Woerl, Jan Disselhoff, and Michael Wand. Initialization noise in image gradients and saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1766–1775, 2023.

Mangasarian Olvi Street Nick Wolberg, William and W. Street. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35:14071–14084, 2022.

Zhuorui Ye and Farzan Farnia. Gaussian smoothing in saliency maps: The stability-fidelity trade-off in neural network interpretability. In *International Conference on Artificial Intelligence and Statistics*, pages 2125–2133. PMLR, 2025.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2019.

Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. BayLIME: Bayesian local interpretable model-agnostic explanations. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 887–896. PMLR, July 2021.

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56673–56699. Curran Associates, Inc., 2023.

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]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
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]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (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]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
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]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [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. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

7 PROOF OF THEOREMS

7.1 Proof of Lemma 1

Proof. Assume $\min_{f\in\mathcal{F}} \mathbb{E}[L(f, Z_{B_r(c)})] \leq \varepsilon$, then by definition of the Rashomon set we have :

$$\arg\min_{f\in\mathcal{F}} \in \mathcal{R}(\varepsilon, \mathcal{F}, Z_{B_r(c)}) \tag{14}$$

and therefore $\mathcal{R}(\varepsilon, \mathcal{F}, Z_{B_r(c)}) \neq \varnothing$.

Now assume $\mathcal{R}(\varepsilon, \mathcal{F}, Z_{B_r(c)}) \neq \varnothing$, it means there exist a model $f \in \mathcal{F}$ such that, $\mathbb{E}[L(f, Z_{B_r(c)})] \leq \varepsilon$. Since $\mathbb{E}[L(f, Z_{B_r(c)})] \geq \min \mathbb{E}[L(f^*, Z_{B_r(c)})]$ for every $f \in \mathcal{F}$, we can conclude:

$$\min_{f\in\mathcal{F}}[L(f, Z_{B_r(c)})] \leq \varepsilon \tag{15}$$

□

7.2 Proof of Theorem 2

The proof proceeds in several steps. We first introduce notation to handle sample and population quantities. Our goal is to show that the empirical maximum explainable radius $\hat{r}^*_{c,k}$ converges in probability to the true maximum explainable radius r^*_c. The main challenge is to control the supremum distance between the sample error $\bar{Q}_k(r)$ and the expected error $\mathbb{E}[Q(r)]$ uniformly over all radii r. To do this, we establish a uniform convergence bound in Lemma 3, which holds with high probability and gives us $\sup_{r\geq 0} |\mathbb{E}[Q(r)] - \bar{Q}_k(r)| \leq \delta_k$ where $\delta_k \to 0$ as $k \to \infty$. Using this bound, we construct a sandwich of sets in Lemma 4: the set of radii satisfying the empirical constraint with margin $\varepsilon - \delta_k$ is contained in the set of radii satisfying the population constraint with margin ε, which in turn is contained in the set with margin $\varepsilon + \delta_k$. Taking suprema yields bounds on r^*_c in terms of empirical quantities. In Lemma 5, we establish that $\bar{Q}_k(r)$ is piecewise constant and right-continuous in r, which implies continuity properties of the supremum function in Lemmas 6 and 7. Finally, as $k \to \infty$ and $\delta_k \to 0$, applying monotonicity of supremum and the squeeze theorem yields the result.

Proof. $\mathcal{G}$ is the class of linear models:

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \;\middle|\; \mathbf{w} \in \mathbb{R}^d,\, b \in \mathbb{R} \right\} \tag{16}$$

Let $g^*_{B_r(c)} \in \mathcal{G}$ be the linear model with the least residual variance and let $Q(r)$ be the random variable corresponding to the squared loss of $g^*_{B_r(c)}$:

$$g^*_{B_r(c)} \in \arg\min_{g\in\mathcal{G}} \mathbb{E}_{Z_{B_r(c)}}[(g(X_{B_r(c)}) - Y_{B_r(c)})^2] \tag{17}$$

$$Q(r) = (g^*_{B_r(c)}(X_{B_r(c)}) - Y_{B_r(c)})^2 \tag{18}$$

Given a center point c, in order to find the largest radius r^*_c, for which, for all $s \leq r^*_c$, the Rashomon set is not empty, we can rewrite Definition 1:

$$R_c(\varepsilon) = \{r \in \mathbb{R}^+ \mid \forall s \in \mathbb{R}^+, s \leq r : \mathbb{E}_{Z_{B_r(c)}}[Q(s)] \leq \varepsilon\} \tag{19}$$

$$r^*_c = \sup R_c \tag{20}$$

Now, we define the empirical version of this variables. Since functions of independent random variables are independent, by drawing k i.i.d. samples from $Z_{B_r(c)}$we can generate k i.i.d. samples, $Q_1(r), \ldots, Q_k(r)$. We then define $\bar{Q}_k(r)$ as follow:

$$\bar{Q}_k(r) = \frac{\sum_{i=1}^k Q_i(r)}{k} \tag{21}$$

Similarly we can rewrite Definition 2:

$$\hat{R}_{c,k}(\epsilon) = \{r \in \mathbb{R}^+ \mid \forall s \in \mathbb{R}^+, s \leq r : \bar{Q}_k(s) \leq \varepsilon\} \tag{22}$$

$$\hat{r}^*_{c,k} = \sup \hat{R}_{c,k} \tag{23}$$

Lemma 3. *Under Assumption 1.1, for any $\delta > 0$, there exits a constant C such that with probability at least $1 - \delta$:*

$$\sup_{r \geq 0} |\mathbb{E}[Q(r)] - \bar{Q}_k(r)| \leq \delta_k \tag{24}$$

where $\delta_k = CB\sqrt{\dfrac{d \ln^2 k + \ln \frac{1}{\delta}}{k}}$.

Proof. Given Assumption 1.1, the function $m_r(Z) = \dfrac{Q(r)}{B}$ is always between zero and one. Let $\mathcal{M} = \{m_r(Z) : r \geq 0\}$ denote the class of normalized loss functions indexed by radius r.

We apply Theorem 3.6 from Alon et al. (1997) to the function class $\mathcal{M}$. For any $\varepsilon > 0$, the theorem provides a bound on the probability that the supremum of the empirical deviation exceeds ε. Specifically, setting $\varepsilon = \delta_k / B$:

$$\begin{aligned}
\mathbb{P}\left(\sup_{m_r \in \mathcal{M}} \left| \frac{1}{k} \sum_{i=1}^{k} m_r(Z_i) - \mathbb{E}[m_r(Z)] \right| > \frac{\delta_k}{B}\right) &= \mathbb{P}\left(\sup_{r \geq 0} \left| \frac{1}{Bk} \sum_{i=1}^{k} Q_i(r) - \frac{\mathbb{E}[Q(r)]}{B} \right| > \frac{\delta_k}{B}\right) \\
&= \mathbb{P}\left(\sup_{r \geq 0} \left| \bar{Q}_k(r) - \mathbb{E}[Q(r)] \right| > \delta_k\right) \\
&\leq \delta
\end{aligned} \tag{25}$$

Taking the complement of this event, we obtain:

$$\mathbb{P}\left(\sup_{r \geq 0} \left| \bar{Q}_k(r) - \mathbb{E}[Q(r)] \right| \leq \delta_k\right) \geq 1 - \delta \tag{26}$$

Therefore, with probability at least $1 - \delta$, we have $\sup_{r \geq 0} |\mathbb{E}[Q(r)] - \bar{Q}_k(r)| \leq \delta_k$. □

The uniform convergence bound allows us to sandwich the population set $R(\varepsilon)$ between empirical sets with adjusted thresholds.

Lemma 4. *Suppose the uniform convergence bound in Lemma 3 holds, i.e.,* $\sup_{r \geq 0} |\mathbb{E}[Q(r)] - \bar{Q}_k(r)| \leq \delta_k$. *Then:* $\hat{R}_{c,k}(\varepsilon - \delta_k) \subseteq R(\varepsilon) \subseteq \hat{R}_{c,k}(\varepsilon + \delta_k)$

Proof. Suppose $r \in R(\varepsilon)$. By definition, we have :

$$\forall s \in [0, r], \mathbb{E}[Q(s)] \leq \varepsilon \tag{27}$$

By 24 we get:

$$\forall s \in [0, r], \bar{Q}_k(s) \leq \mathbb{E}[Q(s)] + \delta_k \tag{28}$$

Therefore:

$$\forall s \in [0, r], \bar{Q}_k(s) \leq \mathbb{E}[Q(s)] + \delta_k \leq \varepsilon + \delta_k \tag{29}$$

which means $r \in \hat{R}_{c,k}(\varepsilon + \delta_k)$.

Now suppose $r \in \hat{R}_{c,k}(\varepsilon - \delta_k)$ By definition, we have :

$$\forall s \in [0, r], \bar{Q}_k(s) \leq \varepsilon - \delta_k \tag{30}$$

By 24 we get:

$$\forall s \in [0, r], \mathbb{E}[Q(s)] \leq \bar{Q}_k(s) + \delta_k \tag{31}$$

Therefore:

$$\forall s \in [0, r], \mathbb{E}[Q(s)] \leq \bar{Q}_k(s) + \delta_k \leq \varepsilon \tag{32}$$

which means $r \in R(\varepsilon)$. □

Lemma 5. *Let $r_1 \leq r_2 \leq \cdots \leq r_k$ denote the ordered distances from the samples to the center point c. Then $\bar{Q}_k(r)$ is right-continuous and piecewise constant with respect to r, with jump discontinuities occurring only at $r_1, \ldots, r_k$.*

Proof. Let $d_j = ||x_j - c||$for $j = 1, \ldots, k$ denote the distances between the samples and the center. Let $r_1, \ldots, r_k$ denote the order statistics of the distances, where $r_1 \leq r_2 \leq \cdots \leq r_k$. For r $r_{i-1} < r < r_i$ the set of included samples does not change. Therefore $\bar{Q}_k(r)$ remains constant on (r_{i-1}, r_i), establishing that $\bar{Q}_k(r)$ is piecewise constant.

Now assume $r_{i-1} \leq r_0 < r_i$. For any r with $r_0 \leq r < r_i$, no additional samples exist in this range, so $\bar{Q}_k(r)$ remains constant and it is equal to $\bar{Q}_k(r_0)$. This shows that given any $\varepsilon > 0$, let $\delta = r_i - r_0$. Then for $r_0 \leq r < r_0 + \delta$, we have $|\bar{Q}_k(r) - \bar{Q}_k(r_0)| = 0 < \varepsilon$, establishing right-continuity at r_0. □

The right-continuity of $\bar{Q}_k(r)$ implies that the supremum function $\sup_r \hat{R}_{c,k}(\varepsilon)$ is also right-continuous in ε.

Lemma 6. *For any $\varepsilon_0 \geq 0$, the supremum function is right-continuous: $\lim_{\varepsilon \to \varepsilon_0^+} \sup_r \hat{R}_{c,k}(\varepsilon) = \sup_r \hat{R}_{c,k}(\varepsilon_0)$.*

Proof. Lemma 5 states that $\bar{Q}_k(r)$ is piecewise constant with jump points $r_1 \leq r_2 \leq \cdots \leq r_k$. Let $\varepsilon_i = \bar{Q}_k(r_i)$ for $i = 1, \ldots, k$. Now suppose $\varepsilon_{i-1} < \varepsilon_0 < \varepsilon_i$, for some $i \geq 1$. For $\varepsilon \in (\varepsilon_0, \varepsilon_i)$ the set of eligible radii doesn't change, therefore the supremum remains constant, $\sup_r \hat{R}_{c,k}(\varepsilon_0) = \sup_r \hat{R}_{c,k}(\varepsilon)$. In case $\varepsilon_0 = \varepsilon_{i-1}$ then since $\bar{Q}_k(r)$ is right-continuous we have $\sup_r \hat{R}_{c,k}(\varepsilon) = \sup_r \hat{R}_{c,k}(\varepsilon_0)$. We have shown for every $\gamma > 0$ there exists a positive number $\delta = \varepsilon_i - \varepsilon_0$ such that if $\varepsilon_0 \leq \varepsilon < \varepsilon_0 + \delta$ then $|\sup_r \hat{R}_{c,k}(\varepsilon) - \sup_r \hat{R}_{c,k}(\varepsilon_0)| = 0 < \gamma$ □

The supremum function is left-continuous almost everywhere, failing only at the countably many jump points of $\bar{Q}_k(r)$.

Lemma 7. *For almost all $\varepsilon_0 \geq 0$ (specifically, all except the countably many values $\varepsilon_1, \ldots, \varepsilon_k$ corresponding to the jump points), the supremum function is left-continuous:*

$\lim_{\varepsilon \to \varepsilon_0^-} \sup_r \hat{R}_{c,k}(\varepsilon) = \sup_r \hat{R}_{c,k}(\varepsilon_0)$ *for almost all ε_0.*

Proof. Lemma 5 states that $\bar{Q}_k(r)$ is piecewise constant with breaking points $r_1 \leq r_2 \leq \cdots \leq r_k$. Let $\varepsilon_i = \bar{Q}_k(r_i)$ for $i = 1, \ldots, k$. Now suppose $\varepsilon_{i-1} \leq \varepsilon_0 < \varepsilon_i$. For $\varepsilon \in (\varepsilon_0, \varepsilon_i)$ the set of eligible radii doesn't change, therefore the supremum remains constant, $\sup_r \hat{R}_{c,k}(\varepsilon_0) = \sup_r \hat{R}_{c,k}(\varepsilon)$. Despite Lemma 6 in case $\varepsilon_0 = \varepsilon_i$ the limit might change since $\bar{Q}_k(r)$ is not left-continues. Therefor, for every $\gamma > 0$ there exists a positive number $\delta = \varepsilon_i - \varepsilon_0$ such that if $\varepsilon_0 \leq \varepsilon < \varepsilon_0 + \delta$ then $|\sup_r \hat{R}_{c,k}(\varepsilon) - \sup_r \hat{R}_{c,k}(\varepsilon_0)| = 0 < \gamma$ except for countable many values of ε_0. □

We now complete the proof using the squeeze theorem. From Lemma 4, taking suprimum of both sides:

$$\sup \hat{R}_{c,k}(\varepsilon - \delta_k) \leq \sup R(\varepsilon) \leq \sup \hat{R}_{c,k}(\varepsilon + \delta_k) \tag{33}$$

From Lemma As $k \to \infty$, we have $\delta_k \to 0$. Therefore $\varepsilon - \delta_k \to \varepsilon$ from below and $\varepsilon + \delta_k \to \varepsilon$ from above. By Lemma 7 and Lemma 6:

$$\lim_{k\to\infty} \sup \hat{R}_{c,k}(\varepsilon - \delta_k) = \sup \hat{R}_{c,k}(\varepsilon) = \hat{r}^*_{c,k} \tag{34}$$

$$\lim_{k\to\infty} \sup \hat{R}_{c,k}(\varepsilon + \delta_k) = \sup \hat{R}_{c,k}(\varepsilon) = \hat{r}^*_{c,k} \tag{35}$$

By the squeeze theorem, since r^*_c is sandwiched between two quantities that both converge to $\hat{r}^*_{c,k}$:

$$\lim_{k\to\infty} |\hat{r}^*_{c,k} - r^*_c| = 0 \tag{36}$$

Since Lemma 3 holds with probability at least $1 - \delta$ for any $\delta > 0$, this entire argument holds with probability at least $1 - \delta$, establishing convergence in probability.

□

8 EXPERIMENT DETAILS

8.1 Synthetic Dataset

We generated 4000 points using the rule mentioned in Section 5.3. To introduce some level of noise, we flipped the labels of 1% of the points at random.

The neural network used for computing gradient-based explanations is a simple multi-layer perceptron (MLP) with the following architecture:

- Input layer with dimensionality equal to the number of features.
- Two hidden layers, each with 64 units, followed by ReLU activations and Batch Normalization.
- Output layer with a single neuron and a Sigmoid activation to perform binary classification.

We used 75% of the dataset for training and trained the model for 50 epochs. For all binary classification tasks, we used the publicly available implementation of empirical Rashomon sets Fisher et al. (2019), with minor modifications.

8.2 Diabetes Dataset

NHANES (National Health and Nutrition Examination Survey) is a dataset derived from an annual survey that includes various health and demographic features for 3,329 patients. For our analysis, we used the following features: Waist Circumference, HDL Cholesterol, Household Income, Blood Urea Nitrogen, and Kcal Intake. In this setting, we did not focus on the inclusion of gradient-based model reliance within the ℓ-MCR interval, since obtaining accurate gradient estimates is not feasible. We ran ℓ-MCR on 403 test points using $\varepsilon = 0.2$ for ℓ-MCR and $\varepsilon = 0.5$ for MCR. The list of candidate radii for the neighborhood was set to $\mathcal{R}_{\text{cand}} = [0.1, 0.53, 0.97, 1.4, 1.8, 2.2, 2.7, 3.13, 3.57, 4.0]$

9 ADDITIONAL RESULTS

9.1 COMPAS dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is an algorithm used to predict the probability of reofending crime, based on covariates such as age, sex, race and priors count. For this experiment we used: age, race, sex, priors count and the COMPAS score as the label. Therefore each data-point is of the form $(\mathbf{x}, y)$ where $\mathbf{x} = [x_{\text{race}}, x_{\text{sex}}, x_{\text{age}}, x_{\text{priors count}}]$. sex and race are binary variables, age and priors count are discrete numerical variables. Similar to Fisher et al(Fisher et al., 2019), we filtered the COMPAS violent scores. We filtered race to keep only the African-American and Caucasian cases due to the lack of enough samples for other races. On top of that, we applied the filters used in COMPAS analysis by Larson et al. (2016). We ended up with $n = 3377$ samples.

Table 3: Empirical MCR for different features in COMPAS dataset.

Feature	$\widehat{\text{MCR}}$
Sex	[1.01, 1.03]
Race	[1.01, 1.04]
Age	[2.17, 2.36]
Priors Count	[1.17, 1.24]

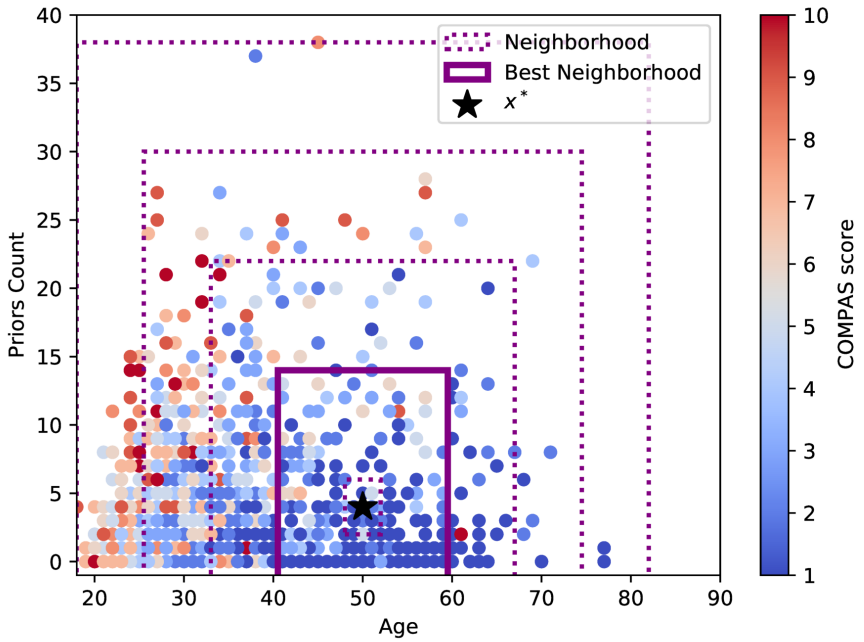

Figure 6: Different neighborhoods on COMPAS dataset. The neighborhood grows until the best model in the neighborhood can not meet the ε criteria. In neighborhoods larger than $\mathcal{N}^*$, even the best model predicts the score with MSE more than 2.5, which means that the error is more than 1.6 COMPAS score.

Table 3 shows the MCR intervals on different features of COMPAS dataset. The results indicate that the COMPAS algorithm does not rely on sex and race, but it relies on age heavily, and on priors count to some extent. To be more precise, all of the well performing models agree that perturbing sex and race does not effect their performance.

In order to apply the ℓ-MCR analysis we need to first define our neighborhood. Given a radius for age and a radius for priors count, we can define the neighborhood for this experiment as follow:

$$\begin{aligned}\mathcal{N}(c, r_{\text{age}}, r_{\text{priors}}) = \{\mathbf{Z}[i] | 1 \leq i \leq n, \|\mathbf{X}[i]_{\text{age}} - \mathbf{x}^*_{\text{age}}\| \leq r_{\text{age}}, \\ \|\mathbf{X}[i]_{\text{priors}} - \mathbf{x}^*_{\text{priors}}\| \leq r_{\text{priors}}, \\ \mathbf{X}[i]_{\text{sex}} = \mathbf{x}^*_{\text{sex}}, \\ \mathbf{X}[i]_{\text{race}} = \mathbf{x}^*_{\text{race}}\}\end{aligned} \tag{37}$$

Using $\varepsilon = 2.5$ and $\mathcal{F}$ to be the class of linear models and the loss function is mean squared error (MSE). Figure 6 depicts the result for $c = [1, 1, 50, 4]^T$, which means an African-American male at age 50 with 4 priors and forming the list of neighborhoods $\mathcal{A}$ based on list of $(r_{\text{age}}, r_{\text{priors}})$ pairs $[(2, 2), (9.5, 10), (17, 18), (24.5, 26), (32, 34)]$

The output of ℓ-MCR for age is $[0.99, 1.01]$, and for priors count is $[1.20, 1.63]$, which means that in cases similar to African-American male at age 50 with 4 priors, age loses it's importance as a feature in comparison to the global MCR.

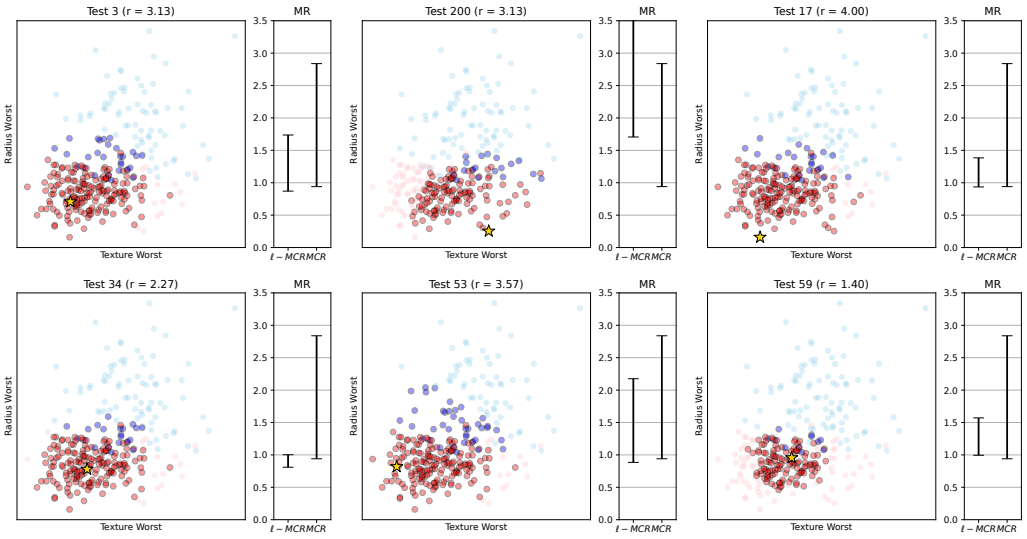

Figure 7: Six patients and their corresponding neighborhoods as identified by Algorithm1. Blue and red points represent patients with and without breast cancer. pale-colored points lie outside the detected neighborhoods. Bold points indicate patients within each local neighborhood. Model reliance is computed with respect to texture_worst.

9.2 Breast Cancer Dataset

The Breast Cancer dataset is a classic binary classification benchmark frequently used in the context of MCR. We selected the top five features based on permutation importance from a Random Forest model: texture_worst, radius_worst, texture_mean, perimeter_worst, and area_worst. Using these five features, a logistic regression model achieves approximately 96% accuracy, indicating that the dataset is already highly linearly separable. As a result, there is limited benefit in analyzing smaller local regions for improved performance. Figure 7 illustrates the results of applying ℓ-MCR and MCR to a few representative test points.