

Beyond Linear Surrogates: High-Fidelity Local Explanations for Black-Box Models

Sanjeev Shrestha, Rahul Dubey, Hui Liu

Department of Computer Science, Missouri State University, USA
{ss472s, RahulDubey, HuiLiu}@MissouriState.edu

Abstract

With the increasing complexity of black-box machine learning models and their adoption in high-stakes areas, it is critical to provide explanations for their predictions. Existing local explanation methods lack in generating high-fidelity explanations. This paper proposes a novel local model agnostic explanation method to generate high-fidelity explanations using multivariate adaptive regression splines (MARS) and N-ball sampling strategies. MARS is used to model non-linear local boundaries that effectively captures the underlying behavior of the reference model, thereby enhancing the local fidelity. The N-ball sampling technique samples perturbed samples directly from a desired distribution instead of reweighting, leading to further improvement in faithfulness. The performance of the proposed method was computed in terms of root mean squared error (RMSE) and evaluated on five different benchmark datasets with different kernel widths. Experimental results show that the proposed method achieves higher local surrogate fidelity compared to baseline local explanation methods, with an average reduction of 32% in root mean square error, indicating more accurate local approximations of the black-box model. Additionally, statistical analysis shows that across all benchmark datasets, the proposed approach results were statistically significantly better. This paper advances the field of explainable AI by providing insights that can benefit the broader research and practitioner community.

Introduction

Artificial Intelligence (AI) has advanced modern information technologies and has become the core of intelligent systems. These AI-powered systems often operate autonomously, without direct human intervention in their design or deployment phases. As AI models increasingly influence critical aspects of human life across high-stake domains such as business, healthcare, manufacturing, and energy, there is a growing need to understand the decision-making process of these systems (Barredo Arrieta et al. 2020). However, it remains challenging to interpret complex Machine Learning (ML) models, despite their high predictive performance.

To address this challenge, eXplainable Artificial Intelligence (XAI) has emerged as a research field that deals with

developing interpretable ML models and explanation techniques to foster transparency, trust, and accountability in AI systems (Ali et al. 2023). Commonly, XAI approaches are classified into local vs global explanation methods. To locally explain a model’s prediction, often a simple local surrogate model is used to mimic the behavior of a complex reference (true) model. The extent to which the surrogate approximates the behavior of the reference model is called *fidelity*. A local surrogate focuses on a specific region of the model’s decision space, maintaining local fidelity while also preserving interpretability. Modern state-of-the-art techniques such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) adopt this approach by training a linear surrogate model on synthetic/perturbed samples labeled by the complex reference model, thereby generating explanations relevant to individual predictions.

Although methods like LIME have been widely adopted, prior studies have identified critical limitations, including instability, sensitivity to hyperparameter selection, and low fidelity in regions with nonlinear decision boundaries (Knab et al. 2025). These challenges restrict the reliability of such explanations, particularly in high-stakes, safety-critical environments. The locality issue arises when the perturbed samples used to train the local surrogate model fail to adequately capture the true local decision boundary around the instance being explained. Similarly, the low fidelity issue, when the surrogate model is often constrained to be linear, prevents it from faithfully approximating the behavior of complex black-box models, especially along highly nonlinear decision boundaries. Importantly, low fidelity does not arise only from poor locality caused by suboptimal sampling. It can also result from a surrogate model that fails to capture non-linearity (Knab et al. 2026). Therefore, reliable local explanations require both sampling that maintains locality and a surrogate model capable of modeling nonlinear relationships.

In this study, we propose a novel high-fidelity local interpretable model agnostic explanation method (HF-LIME) that accurately learns the decision boundary localized to the instance being explained and models the nonlinear local classification boundaries. By relaxing assumptions of local linearity and feature independence, the proposed method aims to significantly enhance the local fidelity and inter-

pretability of explanations for complex ML models, thereby contributing to more trustworthy and transparent decision-making systems. The proposed local model agnostic explanation method uses multivariate adaptive regression splines (MARS) and N-ball sampling strategies. MARS aims to capture the reference model’s non-linear local boundaries, thereby enhancing the local fidelity, and the N-ball sampling technique samples perturbed samples directly from a desired distribution instead of reweighting, like in LIME, leading to further improvement in the fidelity.

To compare the performance of the proposed method against existing well-known local explanation methods, such as LIME and LEMON (Collaris et al. 2023), we considered the root mean squared error (RMSE) performance metric and evaluated on five different benchmark datasets. Prior research has shown that for local explanation, the width of the neighborhood plays a significant role in local surrogate model training (Laugel et al. 2018b), and thus we compared the performance with different kernel widths. Experimental results show that the proposed method achieves higher local surrogate fidelity compared to baseline local explanation methods, with an average reduction of 32% in RMSE, indicating more accurate local approximations of the black-box model. Additionally, statistical analysis shows that across all benchmark datasets, the proposed approach results were statistically significantly better. This paper advances the field of explainable AI by providing insights that can benefit the broader research and practitioner community. The main contributions of this paper are as follows: 1) we introduce a novel local explanation framework that combines a sampling-based strategy with a piecewise linear surrogate model to improve local fidelity. 2) we present a comprehensive experimental evaluation on real-world datasets, including a systematic comparison with existing methods such as LIME and LEMON, to assess the effectiveness of the proposed approach.

Related Work

The LIME framework consists of two main components, a sampling method and the construction of a local surrogate model. Based on this, two research directions have been explored to improve LIME’s fidelity. The first focuses on replacing the linear local surrogate with a nonlinear model to achieve higher accuracy in representing local behavior. As LIME and most of its variants are based on linear surrogates, using nonlinear functions, such as quadratic models, can enhance fidelity, but may compromise interpretability (Bramhall et al. 2020). The second approach focuses on modifying the sampling strategy to better localize the decision boundary and enhance local fidelity.

Various studies have aimed to enhance the LIME sampling process to produce more relevant perturbations. LS-LIME (Laugel et al. 2018a) focuses on sampling around the decision boundary to capture important regions better, while MPS-LIME (Shi, Zhang, and Fan 2020) accounts for feature dependencies to generate more realistic samples. CVAE-LIME (Yasui and Sato 2025) aims to improve LIME’s local fidelity and interpretability by using Conditional Variational Autoencoders (CVAE) for sample gener-

ation. LEMON (Collaris et al. 2023) introduced the N-ball sampling technique that uniformly samples within an N-ball for improved local approximation, and GLIME (Tan, Tian, and Li 2023) employs a locally unbiased sampling strategy to enhance fidelity. US-LIME (Saadatfar, Kiani-Zadegan, and Ghahremani-Nezhad 2024) further improves robustness through a two-step uncertainty-based filtering that emphasizes samples near both the decision boundary and the instance of interest. However, these methods still rely on linear surrogates, limiting their ability to model complex nonlinear regions.

While sampling-based approaches aim to improve the quality and locality of perturbed samples, another line of research focuses on modifying the surrogate model itself. Several alternative models have been explored to replace the traditional linear approximation. LIME-SUP (Hu et al. 2018) and LORE (Guidotti et al. 2018) utilize tree-based techniques as replacement models. Sig-LIME (Abdullah et al. 2024) integrates random forests to capture and utilize nonlinear feature interactions, whereas LIMETree (Sokol and Flach 2025) adopts multi-output regression trees to achieve improved fidelity. However, tree-based models have lower predictive performance, which hinders their ability to accurately approximate non-linear relationships. Guo et al. proposed LEMNA (Guo et al. 2018), a high-fidelity explanation method for security applications that employs a Gaussian Mixture Model (GMM) to approximate nonlinear local boundaries and boost explanation fidelity. Inspired by this approach, Hung et al. introduced BMB-LIME (Hung and Lee 2024), which uses weighted multivariate adaptive regression splines (MARS) (Friedman 1991) with bootstrap aggregating to enhance the local fidelity. Additionally, it employed a Bayesian framework of BayesLIME (Slack et al. 2021) to model uncertainty and improve stability. Replacing the linear surrogate with a piecewise-linear model enhances LIME’s ability to approximate complex nonlinear decision boundaries, thereby improving explanation fidelity while maintaining interpretability.

Method

This work proposes a novel and hybrid approach that incorporates a sampling-based technique and non-linear modeling of local decision boundaries, to enhance the local fidelity of the explanation method. The proposed method generates perturbed samples using the N-Ball sampling technique (Collaris et al. 2023), which samples directly from the desired distribution, defined by a distance-kernel function, instead of reweighting samples. This creates a training dataset of closely related data points in the neighbourhood of the instance x to be explained. Later, MARS (Friedman 1991) constructs a piecewise linear model suitable for high-dimensional problems. Figure 1 depicts the sampling and non-linear model training.

N-ball Sampling

In this method, synthetic data are generated within a unit hypersphere and then scaled by radius r (the region of interest), thereby translating them to be centred at x . LIME samples randomly from the entire feature space and then weights

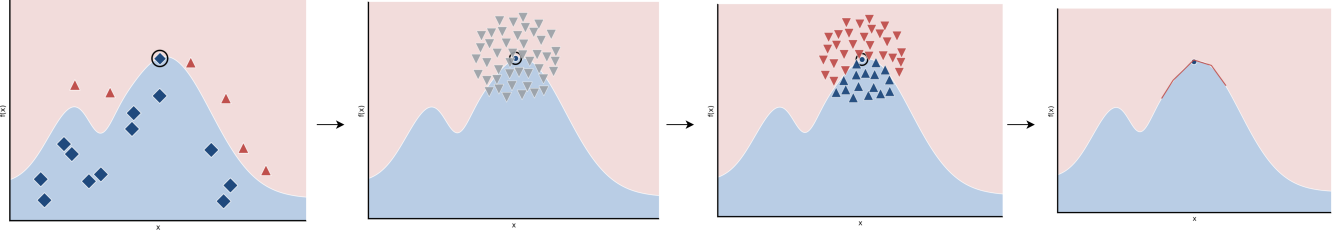


Figure 1: The workflow of the proposed approach includes the following steps: (1) first, select a target instance to be explained, (2) generate synthetic samples within its neighbourhood using N-ball sampling, (3) assign labels to these samples using the reference model, and (4) train a MARS surrogate model to approximate the non-linear local decision boundary. Since the sampling method provides a denser set of local samples, the surrogate model captures the reference black box model’s behavior more accurately.

Algorithm 1: HF-LIME

- 1: **Input:** Model f , number of perturbations N , instance $x \in \mathbb{R}^n$ (n number of features), cumulative distribution function $F(\cdot)$, maximum radius r_{\max} , maximum degree of interaction D , maximum number of terms Q
 - 2: **Output:** Feature contribution vector ϕ
 - 3: $\mathcal{Z} \leftarrow \{\}$ // initialize set of perturbed samples
 - 4: **for** $i = 1$ **to** N **do**
 - 5: sample $y_i \sim \mathcal{N}(0, 1) \in \mathbb{R}^n$
 - 6: $s_i \leftarrow y_i / \|y_i\|$ // normalize
 - 7: sample $u_i \sim \mathcal{U}(0, 1) \in \mathbb{R}^n$
 - 8: $r_i \leftarrow F^{-1}(u_i)$
 - 9: $z_i \leftarrow x + r_i \cdot s_i$ // N-ball sampling
 - 10: $\mathcal{Z} \leftarrow \mathcal{Z} \cup \{(z_i, f(z_i))\}$
 - 11: **end for**
 - 12: $g \leftarrow \text{SURROGATE}(\mathcal{Z}, D, Q)$ // z as features, $f(z)$ as target
 - 13: $\phi \leftarrow \text{extract}(g)$ // extract the feature contributions
 - 14: **return** ϕ
-

those samples based on their proximity to the instance x being explained. However, this does not fully capture the true decision boundary in the local region around x , since training samples farther away are also considered. Instead, N-ball Sampling allows us to sample directly from the desired distribution around the local region of x . This generates samples that follow the dataset distribution around the instance x and attempts to localize the decision boundary.

Fishman (Fishman 2011) and Harman and Lacko (Harman and Lacko 2010) present an efficient method for generating points uniformly within an n -dimensional hypersphere (i.e., an n -sphere). If $Y \sim \mathcal{N}(0, 1)$, then the normalized vector $S_n = \frac{Y}{\|Y\|}$ is uniformly distributed on a unit n -sphere. By further scaling,

$$S_n \cdot U^{1/n} \quad (1)$$

where $U \sim \mathcal{U}(0, 1)$, we obtain samples that are uniformly distributed within the unit n -ball, i.e., the region enclosed by the n -sphere. This sampling approach ensures that all generated points lie strictly within the target region, bounded by a radius r around x . As a result, a surrogate model can be trained on more relevant samples, thereby improving its

ability to approximate the reference model’s local behavior. This can lead to more robust outcomes with reduced variance across algorithm runs.

Sampling is performed in two stages: first, unit vectors are generated by sampling from a Gaussian distribution and normalizing each vector to unit length. This ensures points are uniformly distributed over the surface of a unit n -dimensional hypersphere in \mathbb{R}^n , where n is the dimensionality of the feature space of instance x . Second, radii are sampled according to the cumulative distribution function (CDF) weighted by a distance kernel, and each vector is scaled by its corresponding radius to obtain the perturbation within the n -sphere. The cumulative distribution function for the radius of a sample is

$$F(r) = \frac{\int_0^r K(s) s^{n-1} ds}{\int_0^{r_{\max}} K(s) s^{n-1} ds}, \quad r \leq r_{\max} \quad (2)$$

where $K(\cdot)$ is the distance kernel defined on the interval $[0, r_{\max}]$, where the maximum radius $r_{\max} > 0$ may vary depending on the kernel. Sampling is done from Eq. (2) using inverse transform sampling. Since the corresponding inverse CDF does not generally have a closed-form solution (e.g., for the Gaussian kernel used in LIME), it is approximated numerically. Given a uniform random variable $u \sim \mathcal{U}(0, 1)$, the sampled radius r is obtained as

$$r = F^{-1}(u) \quad (3)$$

When using a Gaussian distance kernel, the support of the kernel is unbounded, which is incompatible with the bounded sampling region required for numerical inverse transform sampling. To enable a fair comparison with the Gaussian kernel employed by LIME, following LEMON (Collaris et al. 2023), we use the truncated Gaussian distance kernel, in which samples are drawn from a Gaussian distribution with the same kernel standard deviation and restricted to lie within r_{\max} . We choose r_{\max} such that a fraction $p \in (0, 1)$ of the sampled points lies within r_{\max} , and we set $p = 0.999$ for all experiments. r_{\max} is defined as:

$$r_{\max} = \sqrt{2\sigma^2 \Gamma^{-1}\left(\frac{n}{2}, (1-p)\Gamma\left(\frac{n}{2}\right)\right)} \quad (4)$$

where σ is the kernel width, n is the dimensionality of the instance feature space, $\Gamma(\cdot)$ denotes the gamma function,

$\Gamma(z, s) = \int_s^\infty t^{z-1} e^{-t} dt$ is the upper incomplete gamma function, and $\Gamma^{-1}(\cdot, \cdot)$ is its inverse with respect to the second argument. Using a truncated Gaussian distance kernel with these parameters produces samples whose distributions closely align with LIME’s weighting scheme, thereby enabling a fair and consistent comparison between the two techniques.

Modeling non-linearity with MARS

After the sampling process, a local surrogate model is trained on a dataset $\mathcal{Z} = \{(z_i, f(z_i))\}_{i=1}^N$, where z_i denotes a perturbed sample generated in the neighborhood of the instance x , and $f(z_i)$ represents the corresponding predictions of the black-box model f . In standard LIME, a linear regression model is typically used as the surrogate. In contrast, this work employs MARS to more accurately capture non-linear decision boundaries in the local neighborhood.

MARS attempts to capture nonlinearity by fitting piecewise linear regressions across different intervals of the input variable’s range (i.e., subregions of the independent variable space). To achieve this, it partitions the input space into overlapping regions by keeping both the parent and its child basis functions after splitting a region. Each split generates a pair of hinge functions that model the behavior on either side of a selected knot. This removes discontinuities at partition boundaries, ensuring that the resulting approximations are continuous with continuous derivatives. This construction demonstrates significant flexibility, allowing it to capture highly nonlinear relationships with complex interactions. A stepwise regression approach is used to construct a local surrogate model in which a set of basis (hinge) functions and their interactions control the prediction. The basis set \mathcal{C} is defined as:

$$\mathcal{C} = \{(X_j - t)_+, (t - X_j)_+\}, \quad t \in \{x_{1j}, \dots, x_{Nj}\}, \quad (5)$$

$$j = 1, \dots, P$$

Here, t denotes a knot selected from the observed values of the predictor X_j . The hinge functions are formulated as: $(x - t)_+ = x - t$, if $x > t$, otherwise 0, and $(t - x)_+ = t - x$, if $x < t$ otherwise 0, where the subscript “+” means the function takes only the positive value or is zero otherwise. The computational complexity of fitting a MARS model is higher than that of a linear regression model, however it retains approximately the same level of interpretability, enabling greater fidelity (Hung and Lee 2024).

A local surrogate MARS model is fitted on \mathcal{Z} to approximate the local behavior of the black-box model f :

$$g(z) = \beta_0 + \sum_{q=1}^Q \beta_q h_q(z) \quad (6)$$

where $h_q(z)$ is the q^{th} basis function (or a product of basis functions representing interactions) and Q is the total number of selected basis functions. The contribution of each basis function at the explained instance x is computed as $\phi_q(x) = \beta_q$. Feature-level attributions are obtained by aggregating the contributions of all basis functions involving the same feature or feature interaction.

Experiment and Discussion

In this section, we first define the evaluation metric for measuring the local fidelity of the explanation methods, and then present the experimental results.

Evaluation Metric

Local fidelity measures how well the local surrogate model approximates the local behavior of the global black-box model. Here, RMSE is used to evaluate local fidelity. It indicates the deviation of the predictions of the approximated model with respect to the reference model. A lower value of RMSE indicates higher fidelity, as it suggests that the local surrogate model is more accurate, being closer to the global model. RMSE is expressed as:

$$\text{RMSE}(\hat{y}^r, \hat{y}^s) = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i^r - \hat{y}_i^s)^2} \quad (7)$$

where \hat{y}^r is the reference model, \hat{y}^s is the surrogate model, and m is the synthetic sample in the area within radius r_{\max} and an equivalent distance kernel to the ones used in LIME and LEMON. We generate $m = 50,000$ synthetic samples within radius r_{\max} using Eq. (2) and an equivalent distance kernel to the ones used in LIME and LEMON, and compute the RMSE values between the prediction probabilities of the reference \hat{y}^r and surrogate model \hat{y}^s for all m samples in the datasets.

Experimental results

We conducted experiments considering five benchmark datasets: Wine, Diabetes, Breast Cancer, Credit, Liver Patient (Collaris et al. 2023). To explain the reference black-box models, we used three classifiers: a Gaussian Naive Bayes (GNB), a multilayer perceptron (MLP) with three hidden layers of 100 neurons each, and a Random Forest (RF) with 200 trees.

To assess the robustness of the proposed method to the choice of locality parameter, we conduct experiments across a wide range of kernel width values, $\sigma \in \{0.1, 0.3, 0.5, 1.0, 4.0, \text{and } \frac{3}{4}\sqrt{n}\}$, where n is the input dimensionality. The default kernel width in LIME is set to $\frac{3}{4}\sqrt{n}$; however, due to its relatively large magnitude (> 1 for $n \geq 2$), it cannot be regarded as truly local. Finally, r_{\max} was calculated using Eq. (4) with $p = 0.999$.

Table 1 shows the comparison of the average RMSE scores across the datasets between HF-LIME and the baselines (LIME and LEMON). Bolded values are the best-performing scores. Results show that HF-LIME consistently achieves the lowest RMSE values across all datasets, models, and kernel widths. In particular, HF-LIME improves the baseline LIME method while also outperforming LEMON in most cases. On average, HF-LIME achieves a 64.8% reduction in RMSE compared to LIME, while LEMON achieves a 47.03% reduction again compared to LIME, indicating that HF-LIME provides significantly higher local fidelity than both LIME and LEMON. Furthermore, HF-LIME achieves a 32.03% lower RMSE compared to

Dataset	Kernel width (σ)	Naive Bayes			Random forest			MLP		
		LIME	LEMON	HF-LIME	LIME	LEMON	HF-LIME	LIME	LEMON	HF-LIME
Wine ($n = 13$)	0.1	0.009	0.003	0.003	0.041	0.018	0.005	0.064	0.007	0.008
	0.3	0.044	0.026	0.024	0.118	0.051	0.017	0.101	0.040	0.019
	0.5	0.103	0.071	0.065	0.185	0.082	0.030	0.219	0.142	0.051
	1.0	0.258	0.224	0.188	0.156	0.120	0.051	0.262	0.205	0.106
	$\frac{3}{4}\sqrt{n}$	0.652	0.303	0.245	0.376	0.124	0.065	0.499	0.273	0.202
	4.0	0.848	0.282	0.226	0.545	0.120	0.065	0.811	0.339	0.203
Diabetes ($n = 9$)	0.1	0.018	0.016	0.002	0.072	0.036	0.009	0.018	0.015	0.001
	0.3	0.057	0.031	0.014	0.141	0.053	0.023	0.050	0.026	0.006
	0.5	0.080	0.046	0.034	0.112	0.064	0.036	0.068	0.031	0.017
	1.0	0.120	0.110	0.098	0.105	0.088	0.054	0.068	0.061	0.055
	$\frac{3}{4}\sqrt{n}$	0.387	0.257	0.212	0.247	0.100	0.062	0.242	0.145	0.139
	4.0	0.687	0.348	0.261	0.419	0.096	0.060	0.465	0.192	0.185
Breast cancer ($n = 32$)	0.1	0.011	0.006	0.004	0.038	0.015	0.006	0.219	0.115	0.085
	0.3	0.052	0.030	0.026	0.103	0.038	0.015	0.418	0.194	0.163
	0.5	0.151	0.105	0.077	0.171	0.057	0.024	0.524	0.271	0.224
	1.0	0.491	0.263	0.179	0.265	0.072	0.036	0.521	0.295	0.254
	$\frac{3}{4}\sqrt{n}$	0.513	0.001	0.001	0.367	0.065	0.040	0.985	0.228	0.287
	4.0	0.504	0.002	0.002	0.358	0.065	0.040	0.742	0.292	0.219
Credit ($n = 15$)	0.1	0.033	0.026	0.005	0.092	0.045	0.018	0.032	0.013	0.009
	0.3	0.218	0.172	0.063	0.150	0.056	0.033	0.153	0.084	0.075
	0.5	0.381	0.277	0.118	0.205	0.095	0.049	0.258	0.152	0.140
	1.0	0.333	0.304	0.180	0.169	0.150	0.074	0.269	0.250	0.234
	$\frac{3}{4}\sqrt{n}$	0.405	0.237	0.196	0.320	0.145	0.090	0.650	0.321	0.308
	4.0	0.487	0.226	0.191	0.445	0.140	0.091	0.876	0.348	0.330
Liver ($n = 11$)	0.1	0.047	0.026	0.016	0.209	0.063	0.041	0.089	0.049	0.038
	0.3	0.449	0.295	0.147	0.257	0.081	0.055	0.233	0.184	0.154
	0.5	0.296	0.239	0.175	0.122	0.088	0.060	0.316	0.267	0.199
	1.0	0.088	0.079	0.075	0.104	0.093	0.065	0.285	0.266	0.220
	$\frac{3}{4}\sqrt{n}$	0.025	0.014	0.014	0.187	0.093	0.068	0.471	0.297	0.269
	4.0	0.023	0.003	0.003	0.348	0.091	0.069	0.908	0.341	0.308

Table 1: Comparing the average RMSE values obtained across all instances of datasets considering different black box models, explanation methods, and kernel widths. Best-performing scores are bolded.

LEMON, demonstrating its superior performance among baselines.

On Wine dataset, HF-LIME provides significant improvements, particularly at larger kernel widths (e.g., at $\sigma = 4.0$) The Diabetes dataset, which has lower dimensionality, also benefits significantly: at $\sigma = 0.1$, RF RMSE drops from 0.072 (LIME) to 0.009 (HF-LIME). This effect becomes more evident in the Breast Cancer dataset, which is higher-dimensional ($n = 32$), where LIME produces unstable results with higher error values. (e.g., MLP RMSE of 0.985 at $\sigma = (3/4)\sqrt{n}$), while HF-LIME maintains consistently lower values (0.287 under the same setting). Although LEMON occasionally achieves low RMSE (e.g., GNB at $\sigma = (3/4)\sqrt{n}$), these results are not consistently observed

across classifiers.

The choice of kernel width significantly affects local fidelity, as it defines the extent of the local neighborhood. While LIME’s performance decreases rapidly as σ increases, both LEMON and HF-LIME show greater flexibility. Overall, HF-LIME provides the most stable local explanations, consistently maintaining lower RMSE across all kernel widths.

Fig. 2 illustrates the distribution of RMSE values computed for LIME, LEMON, and HF-LIME across GNB, RF, and MLP classifiers for Diabetes and Wine datasets. We kept the default kernel width to $\sigma = (3/4)\sqrt{n}$, as used by LIME for a fair comparison. Across all settings, HF-LIME consistently achieves the lowest median RMSE and a sub-

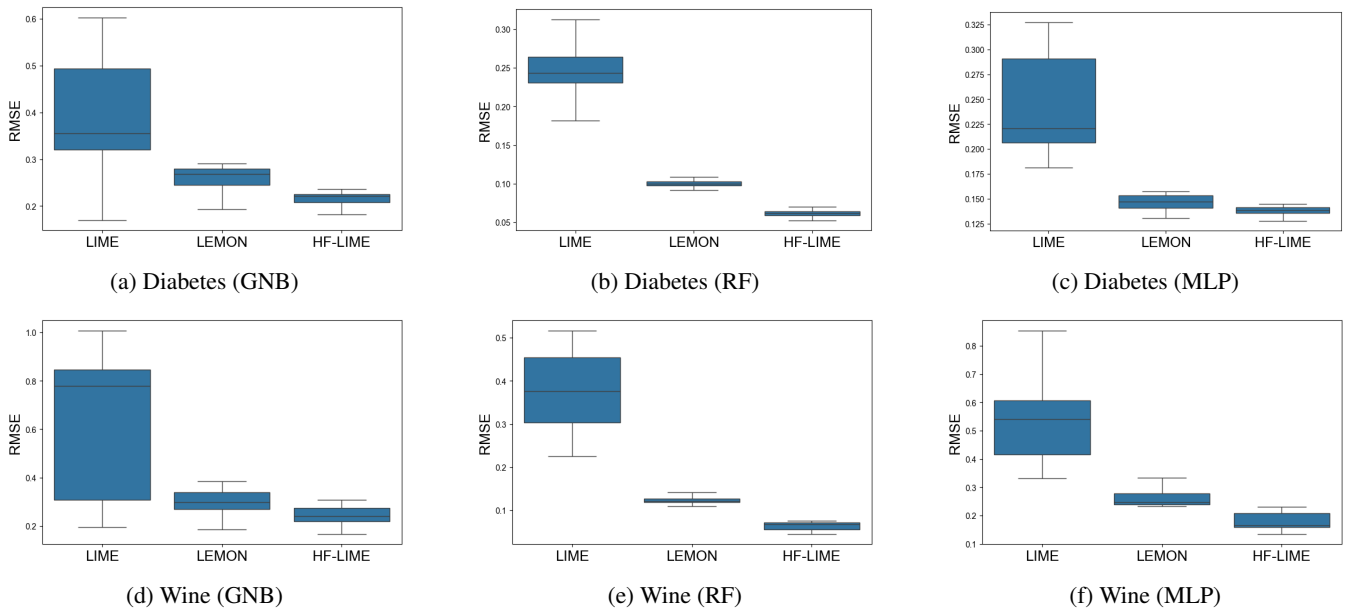


Figure 2: RMSE distributions measuring the local fidelity of LIME, LEMON, and HF-LIME methods across GNB, RF, and MLP classifiers at the default kernel width $\sigma = (3/4)\sqrt{n}$ on the Diabetes and Wine datasets. Lower RMSE indicates higher local fidelity. HF-LIME consistently achieves lower error and reduced variance across classifiers and datasets, indicating improved local fidelity.

Dataset	Naive Bayes			Random Forest			MLP		
	LIME	LEMON	HF-LIME	LIME	LEMON	HF-LIME	LIME	LEMON	HF-LIME
Wine	0.269	0.040	0.031	0.077	0.008	0.008	0.118	0.036	0.027
Diabetes	0.108	0.032	0.021	0.031	0.004	0.004	0.044	0.007	0.004
Breast Cancer	0.384	0.002	0.002	0.034	0.002	0.002	0.168	0.011	0.009
Credit	0.164	0.090	0.036	0.060	0.014	0.008	0.125	0.034	0.031
Liver	0.096	0.063	0.038	0.059	0.012	0.013	0.101	0.053	0.046

Table 2: Standard deviation of RMSE values, averaged across all kernel widths. Best-performing (lowest) scores are bolded. HF-LIME performed better than others by having a smaller standard deviation.

stantially smaller interquartile range than both LIME and LEMON. This indicates that the proposed method produces explanations that more closely approximate the behavior of the underlying black-box model while also exhibiting improved stability across instances. In contrast, LIME shows higher RMSE and greater variance, suggesting low fidelity and more unstable local explanations. To assess whether the observed differences in RMSE across explanation methods are statistically significant, we conducted a one-tailed paired t-test on the RMSE values calculated for all instances of the datasets. Across all datasets and classifiers, HF-LIME achieves statistically significant reductions in RMSE compared to LIME and LEMON ($p < 0.05$), validating the effectiveness of the proposed approach in improving local fidelity.

Figure 2 also suggests that the spread of RMSE values is the smallest using the proposed method compared to the other two approaches. Thus, to measure the stability of lo-

cal surrogate fidelity, we computed the standard deviation of RMSE values across all datasets and kernel widths. Table 2 shows the average standard deviation, obtained by first computing the standard deviation of RMSE values across instances for each kernel width and then averaging these values across all kernel widths. This result indicates that the proposed approach not only generates high-fidelity local explanations but is also stable, with lower variance than the reference model.

Overall, these experimental results indicate that HF-LIME offers high-fidelity, robust, and stable explanations compared to both LIME and LEMON. The boost in fidelity comes from integrating N-ball sampling, which enhances the locality of the decision boundary, with the MARS surrogate model, which effectively captures the underlying non-linear relationships. Its stability across kernel widths, effectiveness on higher-dimensional datasets, and consistent improvements across classifiers demonstrate its suitability as

a more accurate method for explaining reference black-box models.

Conclusion

This paper proposes a novel method designed to improve the fidelity of local explanation. By integrating N-ball sampling with MARS surrogate, it localizes the decision boundary and captures nonlinear relationships more effectively than existing methods such as LIME and LEMON. Through experiments on five real-world datasets, we demonstrated that HF-LIME consistently outperforms baseline methods in terms of RMSE. On average, HF-LIME achieves a 64.8% reduction in RMSE compared to LIME and a 32% reduction compared to LEMON, confirming its ability to generate higher-fidelity explanations. The statistical analysis shows that HF-LIME achieves statistically significant improvements in local fidelity while demonstrating lower variability across kernel widths, resulting in more stable local explanations than existing methods. Future work will focus on further improving explanation stability, consistency, and extending the proposed method to other data modalities, such as images and time-series data. Addressing these prospects will further advance the goal of reliable, faithful, and robust explanations for complex real-world systems.

References

- Abdullah, T. A. A.; Zahid, M. S. M.; Turki, A. F.; Ali, W.; Jiman, A. A.; Abdulaal, M. J.; Sobahi, N. M.; and Attar, E. T. 2024. Sig-Lime: A Signal-Based Enhancement of Lime Explanation Technique. *IEEE Access*, 12: 52641–52658.
- Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99: 101805.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.
- Bramhall, S.; Horn, H.; Tieu, M.; and Lohia, N. 2020. QLIME-A quadratic Local Interpretable Model-agnostic explanation approach. *SMU Data Science Review*, 3(1): 4.
- Collaris, D.; Gajane, P.; Jorritsma, J.; van Wijk, J. J.; and Pechenizkiy, M. 2023. LEMON: Alternative Sampling for More Faithful Explanation Through Local Surrogate Models. In Crémilleux, B.; Hess, S.; and Nijssen, S., eds., *Advances in Intelligent Data Analysis XXI*, 77–90. Cham: Springer Nature Switzerland. ISBN 978-3-031-30047-9.
- Fishman, G. 2011. *Monte Carlo*. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer.
- Friedman, J. H. 1991. Multivariate adaptive regression splines. *The annals of statistics*, 19(1): 1–67.
- Guidotti, R.; Monreale, A.; Ruggieri, S.; Pedreschi, D.; Turini, F.; and Giannotti, F. 2018. Local Rule-Based Explanations of Black Box Decision Systems. arXiv:1805.10820.
- Guo, W.; Mu, D.; Xu, J.; Su, P.; Wang, G.; and Xing, X. 2018. LEMNA: Explaining Deep Learning based Security Applications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS '18*, 364–379. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356930.
- Harman, R.; and Lacko, V. 2010. On decompositional algorithms for uniform sampling from n-spheres and n-balls. *Journal of Multivariate Analysis*, 101(10): 2297–2304.
- Hu, L.; Chen, J.; Nair, V. N.; and Sudjianto, A. 2018. Locally Interpretable Models and Effects based on Supervised Partitioning (LIME-SUP). arXiv:1806.00663.
- Hung, Y.-H.; and Lee, C.-Y. 2024. BMB-LIME: LIME with modeling local nonlinearity and uncertainty in explainability. *Knowledge-Based Systems*, 294: 111732.
- Knab, P.; Marton, S.; Schlegel, U.; and Bartelt, C. 2025. Which LIME should I trust? Concepts, Challenges, and Solutions. In *Proceedings of the XAI 2025 Conference*.
- Knab, P.; Marton, S.; Schlegel, U.; and Bartelt, C. 2026. Which LIME Should I Trust? Concepts, Challenges, and Solutions. In Guidotti, R.; Schmid, U.; and Longo, L., eds., *Explainable Artificial Intelligence*, 28–52. Cham: Springer Nature Switzerland. ISBN 978-3-032-08324-1.
- Laugel, T.; Renard, X.; Lesot, M.; Marsala, C.; and Detyniecki, M. 2018a. Defining Locality for Surrogates in Post-hoc Interpretability. *CoRR*, abs/1806.07498.
- Laugel, T.; Renard, X.; Lesot, M.-J.; Marsala, C.; and Detyniecki, M. 2018b. Defining Locality for Surrogates in Post-hoc Interpretability. arXiv:1806.07498.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.
- Saadatfar, H.; Kiani-Zadegan, Z.; and Ghahremani-Nezhad, B. 2024. US-LIME: Increasing fidelity in LIME using uncertainty sampling on tabular data. *Neurocomputing*, 597: 127969.
- Shi, S.; Zhang, X.; and Fan, W. 2020. A Modified Perturbed Sampling Method for Local Interpretable Model-agnostic Explanation. arXiv:2002.07434.
- Slack, D.; Hilgard, S.; Singh, S.; and Lakkaraju, H. 2021. Reliable post hoc explanations: modeling uncertainty in explainability. NIPS '21. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781713845393.
- Sokol, K.; and Flach, P. 2025. LIMETREE: Consistent and Faithful Surrogate Explanations of Multiple Classes. *Electronics*, 14(5).
- Tan, Z.; Tian, Y.; and Li, J. 2023. GLIME: General, Stable and Local LIME Explanation. arXiv:2311.15722.
- Yasui, D.; and Sato, H. 2025. Improving Local Fidelity and Interpretability of LIME by Replacing Only the Sampling Process With CVAE. *IEEE Access*, 13: 53084–53099.