Tilia: Enhancing LIME with Decision Tree Surrogates

Jihang Li
Jiacheng Qiu
HKUST (Guangzhou)
Guangzhou, Guangdong, China
jli945@connect.hkust-gz.edu.cn
qiujiacheng1993@gmail.com

Yin-Ping Zhao
NWPU
Xi'an, Shaanxi, China
zhaoyinping@nwpu.edu.cn

Zeyi Wen*
HKUST (Guangzhou)
Guangzhou, Guangdong, China
wenzeyi@ust.hk

Abstract

Local Interpretable Model-Agnostic Explanations (LIME) is a widely adopted framework for interpreting opaque models due to its simplicity and intuitiveness. However, LIME suffers from unreliability rooted in two core issues: (i) low fidelity, where the surrogate model fails to accurately approximate the target model's behavior, and (ii) instability, where the generated explanations vary significantly across runs. While prior work has proposed techniques to enhance LIME, they remain fundamentally limited by the expressiveness of linear surrogate models, which cannot adequately capture complex decision boundaries. In this work, we introduce Tilia, a novel method that employs shallow decision tree regressors as the surrogate model, leveraging its structured and deterministic nature to improve both fidelity and stability. Tilia also provides insight into the interplay between surrogate models and sampling strategies, revealing new directions for enhancing explanation reliability. Across extensive experiments on tabular and textual datasets, Tilia outperforms LIME and recent variants on both fidelity and stability, achieving up to 100% approximation of the opaque model and entirely consistent explanations (i.e., 0 Jacard distance). Tilia maintains practical efficiency, completing explanations in seconds even for datasets with over 100 features. These results position Tilia as a robust alternative for model-agnostic explanations. The code is available at https://github.com/neur1n/tilia.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

Explainable AI; Model-Agnostic Explanations; LIME; Decision Trees

ACM Reference Format:

Jihang Li, Jiacheng Qiu, Yin-Ping Zhao, and Zeyi Wen. 2025. Tilia: Enhancing LIME with Decision Tree Surrogates. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3746252.3761130

*Corresponding Author

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761130

1 Introduction

Explainable Artificial Intelligence (XAI) has become a promising research area, driven by increasing demands for transparency, accountability, and trust in machine learning models. As these models are increasingly deployed in high-stakes domains such as healthcare and finance, the ability to interpret and justify predictions is essential [2, 38, 43]. Among existing explanation techniques, Local Interpretable Model-Agnostic Explanations (LIME) [27] remains widely used [26] as both a foundational method [20, 40] and a standard comparative baseline [5, 15, 17, 37, 45].

LIME explains predictions by generating perturbed samples around a target instance and fitting a simple, interpretable surrogate model–typically a linear regressor–to approximate the complex model's local behavior. Despite its conceptual simplicity, LIME often suffers from two key limitations [3, 21, 28, 36]: (i) low fidelity, where the surrogate poorly approximates the underlying model, and (ii) instability, where explanations vary largely across runs.

While numerous variants aim to mitigate instability [12, 18, 22, 29, 30, 33, 46, 47, 49], primarily by refining LIME's sampling process, they often leave fidelity unaddressed. Instability arises from both the stochastic nature of perturbation and the sensitivity of linear regression to small input changes. Meanwhile, linear surrogates inherently struggle to capture complex, nonlinear decision boundaries, further undermining fidelity. In critical applications, such fragility can lead to misleading or unjustified decisions.

To address these issues, we propose *Tilia*, an extension of LIME that directly addresses both low fidelity and instability. Tilia replaces the linear surrogate with a shallow decision tree regressor, leveraging the structured, nonlinear, and deterministic nature of decision trees. This offers two main advantages: (i) improved fidelity by better capturing complex local decision boundaries, and (ii) reduced sensitivity to perturbations, enhancing stability. Integrating trees, however, introduces challenges such as balancing interpretability and reliability with tree depth, reconciling feature preprocessing, aligning feature importance computation, and handling Gini importance's inability to express directional contributions. Tilia resolves these issues to preserve interpretability while more accurately modeling local behavior.

We evaluate Tilia on diverse tabular and textual datasets against LIME and recent variants, including S-LIME [49], BayLIME [47], CALIME [12], and DLIME [46]. The results demonstrate that Tilia consistently improves fidelity and stability, often achieving perfect fidelity and fully consistent explanations across runs, while maintaining practical runtime efficiency. These results position Tilia as a robust alternative for model-agnostic explanations.

By offering a simple yet effective modification to the LIME framework, Tilia provides a pragmatic path toward more trustworthy and interpretable model-agnostic explanations. Its balance of robustness, fidelity, and efficiency makes it a compelling tool for real-world applications where reliable explanations are essential.

Our main contributions are summarized as follows:

(1) We perform a systematic analysis of the sources of low fidelity and instability in the LIME framework, providing insights into their root causes.
(2) We propose Tilia, a decision tree–based surrogate approach that significantly improves the reliability and faithfulness of LIME's local explanations.
(3) We validate our method through comprehensive experiments, demonstrating consistent and substantial improvements over baseline and state-of-the-art methods across multiple datasets and modalities.

2 Related Works

Explainable AI methods can be classified along several axes: intrinsic vs. post-hoc, global vs. local, and model-specific vs. model-agnostic [7]. Intrinsic methods, such as decision trees or linear models, are inherently interpretable by design, with transparent decision processes [24]. Post-hoc methods instead interpret trained opaque models without altering their structure [19]. Global explanations describe a model's overall logic [44], while local explanations focus on individual predictions [10]. Model-specific methods target particular model types, whereas model-agnostic methods work across models without requiring internal access [1].

Among these, *post-hoc, local, model-agnostic* methods such as LIME [27] are popular for their flexibility and intuitive appeal. To contextualize our work, we review prior efforts aimed at improving the reliability of LIME-based explanations, which can be grouped into: (1) refining the sampling strategy, (2) incorporating clustering, (3) generating dependency-aware perturbations, and (4) modifying the weighting function or surrogate. While these methods have demonstrated varying degrees of success, most retain the use of a linear surrogate, which inherently limits fidelity.

Sampling-Focused Methods. A large body of work aims to improve the quality and relevance of perturbed samples. LS-LIME [22] targets the decision boundary to capture regions most informative for understanding model behavior. MPS-LIME [35] modifies the perturbed sampling operation by considering feature correlations and thereby produces more realistic samples. S-LIME [49] uses Least Angle Regression (LARS) to optimize the number of perturbed samples, reducing variance in explanation. LEMON [13] samples uniformly within an N-ball, improving local approximation. GLIME [39] employs a local unbiased sampling distribution for higher fidelity. US-LIME [29] introduces a two-step uncertainty-based filtering that emphasizes samples near both the decision boundary and target instance. While these improve robustness, reliance on linear surrogates still limits modeling of nonlinear decision regions.

Clustering-Based Methods. Another line of work introduces clustering techniques to guide LIME's sampling or aggregate explanations by partitioning data into meaningful regions in order to obtain a more coherent understanding of the model's behavior. For instance, DLIME [46] applies Agglomerative Hierarchical Clustering (AHC) and K-Nearest Neighbor (KNN) sampling to select relevant data regions. KLIME [18] partitions training data using K-means and fits a separate generalized linear model (GLM) within each cluster. ILIME [14] additionally identifies the most influential features in each explanation and clusters multiple explanations using a dendrogram to derive the most representative one. These clustering-based methods enhance the interpretability and stability of LIME explanations by focusing on localized regions of the data space. However, they introduce additional complexity and still rely on linear surrogates, which may not capture complex, nonlinear decision boundaries precisely.

Dependency-Aware Methods. Dependency-aware methods have emerged to address LIME's assumption of feature independence, which can lead to unrealistic or implausible explanations. FLIME [30] uses Conditional Tabular GANs (CTGANs) to generate synthetic samples that better reflect real data distributions. CALIME [12] integrates causal knowledge into the explanation process by replacing LIME's random sampling with GENCDA [11], a causal dependency-aware generator that encodes structural relationships between features. This approach ensures that the generated samples adhere to the underlying causal structure of the data. Kernel-based LIME with feature dependency sampling (KLFDS) [34] incorporates feature dependency sampling into the LIME framework. Similarly, CHILLI [4] introduces a contextually enhanced perturbation method that considers domain-specific constraints and feature dependencies during the perturbation process. By generating perturbations that are both representative of the training data and local to the instance being explained, CHILLI aims to produce more accurate and contextually relevant explanations. While these methods improve stability, they often require complex data generation pipelines.

Alternative Weighting and Surrogates. Beyond sampling, several works explore improvements to LIME's weighting function and surrogate model. ALIME [33] replaces Euclidean distances in input space with latent-space distances computed using autoencoders, yielding more semantically meaningful weights. BayLIME [47] uses Bayesian Ridge regression surrogate to incorporate prior knowledge and quantifies uncertainty, offering more robust explanations under noisy conditions. However, these approaches still assume a linear relationship between features and predictions. EBLIME [48] extends BayLIME to produce a distribution of feature importance. This approach offers a more comprehensive understanding of the uncertainty associated with each feature's contribution to the prediction. QLIME [8] considers a quadratic surrogate by integrating linear relationships across multiple step points, offering improved expressiveness over strictly linear models.

Despite these advancements, most methods retain the linear surrogate model at the core of LIME. This limitation fundamentally constrains fidelity, especially in regions where the model's behavior is highly nonlinear. In contrast, our proposed approach directly addresses this gap by replacing the surrogate with a shallow decision tree regressor, which is more expressive yet remains interpretable. This modification enables Tilia to improve both fidelity and stability simultaneously, and we demonstrate that it integrates seamlessly with a wide range of sampling strategies.

3 Methodology

In this section, we begin by analyzing both empirically and theoretically the key sources of low fidelity and instability in the LIME framework, providing a foundation for understanding its limitations in producing reliable local explanations. Based on this analysis, we introduce our proposed solution, Tilia, which improves both fidelity and stability by incorporating a structured and shallow decision tree regressor. We also discuss the practical challenges associated with integrating decision trees into the LIME framework, particularly with respect to maintaining interpretability and robustness.

3.1 Low Fidelity and Instability in LIME

Fidelity is a critical property for explanation methods, reflecting how well the surrogate model approximates the local behavior of the original, opaque model. Even if explanations appear stable across runs, they may still be untrustworthy if the surrogate fails to capture the true decision logic. To systematically assess fidelity, we trained a default-configured Random Forest classifier on each dataset ten times and selected the model with the highest predictive performance to serve as the opaque model. For each dataset, up to 30 test samples were selected for explanation.

We then applied LIME using its default configuration, initializing the surrogate model–typically a Ridge regressor–with fixed random seeds (3, 11, 23, 37, and 42). Fidelity was quantified using the R^2 score [42], which measures the degree to which the surrogate model's predictions align with those of the opaque model on perturbed samples. As shown in Figure 1, our results reveal that LIME's linear surrogate consistently exhibits low fidelity across all evaluated datasets, underscoring its limitations in accurately modeling complex, nonlinear decision boundaries.

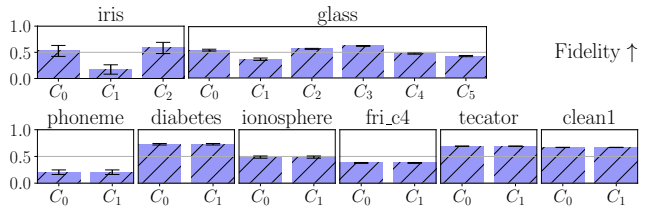


Figure 1: Fidelity achieved by the default surrogate in LIME.

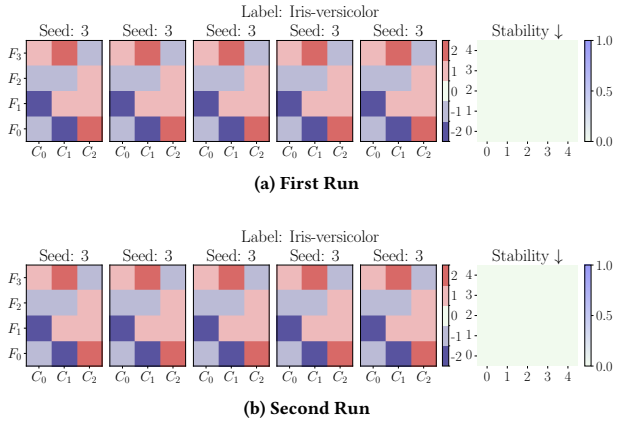
While low fidelity undermines the accuracy of LIME's explanations, instability affects their reproducibility. Even when applied to the same instance, LIME may produce different explanations across runs due to randomness in its components. To better understand this phenomenon, we examine the formulation of LIME:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x(Z)) + \Omega(g)$$

where $\xi(x)$ denotes the explanation for an input x, f is the opaque model being explained, g is the surrogate model selected from a family of interpretable models G, Z denotes the perturbed samples around x, π_x is the locality-aware weighting function, and $\mathcal{L}$ is the loss between the predictions of f and g on the perturbed samples. Randomness can arise from three key components: the opaque model f, the surrogate model g, and the sampling process used to generate Z. Each of these components may involve stochastic elements or random seed initialization.

To isolate and verify the sources of instability, we conducted controlled experiments on the Iris dataset [16] using a Random Forest classifier as the opaque model f. We configured LIME's tabular explainer with default parameters except for the discretization of continuous features, which was disabled to preserve the full set of features for explanation. As in standard LIME, a Ridge regressor serves as the surrogate model g, and perturbed samples Z are generated uniformly at random.

In the first experiment, we fixed the random seeds for all components: the opaque model f was initialized with seed 42, and both the surrogate model g and perturbed samples Z with seed 3. We then ran the explanation process twice. As shown in Figure 2, the resulting feature importance values remained identical across both runs, and the pairwise Jaccard distances used to quantify stability were effectively zero. This confirms that when randomness is controlled, LIME produces stable explanations.



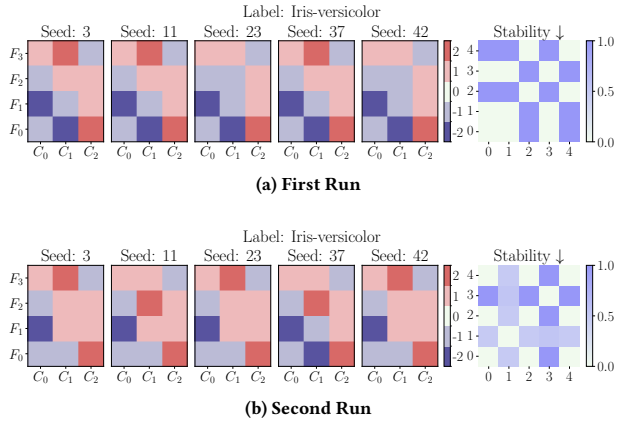
(a) First Run

(b) Second Run

Figure 2: Feature importance and stability across two runs with the same random seeds using LIME. The x-axis represents the classification task's classes, and the y-axis corresponds to the features.

In contrast, in the second experiment, we removed the fixed seed for the opaque model f while retaining the same fixed seeds for g and Z as in the fidelity assessments. The results, illustrated in Figure 3, reveal notable differences in feature importance values both within and across runs. These inconsistencies indicate that even slight variations in perturbed samples, brought by inconsistent seedings, can lead to divergent surrogate models and, consequently, to unstable explanations.

These findings suggest that the primary driver of instability in LIME is the sampling process, which introduces uncontrolled randomness when generating Z. Since the surrogate model g is fitted on these samples, any variability in Z can propagate through the surrogate and distort the final explanation, even when other components are held constant.



(a) First Run

(b) Second Run

Figure 3: Feature importance and stability across two runs with different random seeds using LIME.

3.2 Theoretical Analysis

Besides our empirical findings, we can theoretically justify why a shallow decision tree surrogate offers superior local fidelity and stability compared to a linear model. Linear regressions (LR) and decision trees (DT) can be expressed as:

$$g_{\mathrm{LR}}(x) = w_0 + \sum_{i=1}^{n} w_i x_i, \quad g_{\mathrm{DT}}(x) = \sum_{j=1}^{J} c_j \cdot \mathbf{1}\{x \in R_j\},$$

where n is the number of features, $\{R_j\}$ are hyper regions partitioning the feature space and c_j is the constant prediction within leaf j. Unlike a single hyperplane, this structure can approximate nonlinear decision boundaries via axis-aligned tiles, reducing local approximation bias and thus improving fidelity.

Let W denotes the LIME kernel weights, the weighted Ridge estimator $\hat{\beta} = (Z^\top WZ + \lambda I)^{-1} Z^\top Wy$, can fluctuate dramatically when $Z^\top WZ$ is ill-conditioned, such as under feature collinearity or excessive noise—resulting in unstable feature attributions across runs. Decision trees select splits by minimizing an impurity measure $G(Q_m, \theta)$ (e.g., Gini, entropy, or MSE). Let $\theta^\star = \arg\min_\theta G(Q_m, \theta)$. Under small perturbations of the sample distribution, the change ΔG remains bounded due to the Lipschitz continuity of the impurity functions in their underlying moments [9]. If the feature split margin $\gamma_m = \min_{\theta \neq \theta^\star} \left[G(Q_m, \theta) - G(Q_m, \theta^\star) \right]$ is larger than ΔG, the optimal split $\theta^\star$ remains invariant. This insensitivity cascades recursively down the tree. Leaf predictions, being based on smooth averages (means or proportions), further reinforce stability. The result is a surrogate whose structure and attributions are resilient to random sampling noise, unlike linear regression's fragility.

3.3 Proposed Method

LIME's limitations in fidelity and stability highlight the need for a more robust surrogate. We propose Tilia, a LIME variant that replaces the linear surrogate with a shallow decision tree regressor. This structured, deterministic model is more resilient to sampling noise and better at capturing nonlinear decision boundaries.

However, integrating a decision tree regressor into LIME is not straightforward. This change introduces several new challenges, both in terms of maintaining interpretability and ensuring compatibility with LIME's design. Specifically, we identify four key issues that must be addressed:

(1) **Interpretability vs. Reliability**: Although decision trees are considered interpretable models, their interpretability diminishes with increasing depth. Deeper trees are more expressive but can overfit the perturbed data, resulting in unreliable or overly complex explanations [9].
(2) **Feature Space Inconsistency**: LIME scales continuous features to the range $[0, 1]$ when using linear surrogates to prevent bias toward features with larger numeric ranges. In contrast, decision trees operate directly on the raw feature values. This inconsistency in preprocessing can hinder the surrogate's ability to produce intuitive and meaningful splits unless properly addressed.
(3) **Feature Importance Incompatibility**: Linear models compute feature importance via model coefficients, while decision trees rely on Gini importance. These fundamentally different measures complicate the direct comparison of explanations across surrogate types and require additional normalization strategies.
(4) **Directional Contribution Limitation**: Gini importance, being non-negative by design, cannot express whether a feature contributes positively or negatively to a prediction [31]. To provide faithful local explanations, a mechanism for capturing directionality in feature influence is necessary.

To address the first challenge, we limit the decision tree's maximum depth and use cross-validation to balance interpretability and prevent overfitting. A grid search over depths 2–10, informed by empirical observations, identifies the optimal tree. Perturbed samples are split 80% for training and 20% for testing, with K-fold cross-validation applied to evaluate generalization. The tree achieving the best trade-off between performance and interpretability is selected as the surrogate model for explanations.

For the second challenge, we preserve the original feature values when using the decision tree surrogate. Unlike Ridge regression, which requires scaled inputs to $[0, 1]$ to mitigate coefficient bias, decision trees split directly on actual feature values. Applying normalization in this context can distort split thresholds and degrade interpretability. Retaining the original feature scale ensures that decision boundaries align with the data's natural structure, thereby improving both semantic coherence and surrogate fidelity.

On the output side, we normalize feature importance scores to allow consistent comparison between Ridge regression and decision tree surrogates. As detailed in Algorithm 1, we apply a binning-based normalization method to standardize the representation of feature attributions across models. To address the lack of directionality in Gini importance, we introduce a sign correction strategy: we invert the importance values for classes other than the predicted class, treating them as indicative of negative contributions. As illustrated in Figure 5, only the class being explained (e.g., C_1) is assigned positive contributions, while the remaining classes are considered oppositional, thereby enabling directional interpretation.

Algorithm 1: Feature Importance Binning

Input: I
Output: I_{bin}
begin
  Let $I^+ \leftarrow [\,], I^- \leftarrow [\,], I_{\text{bin}} \leftarrow [\,]$ *// Importance*
  for $i \in I$ **do**
    Append i to $(i > 0\ ?\ I^+\ :\ I^-)$
  $M^+ \leftarrow \text{Median}(I^+), M^- \leftarrow \text{Median}(I^-)$
  for $i \in I$ **do**
    if $i > 0$ **then**
      Append $(i \leq M^+\ ?\ 1\ :\ 2)$ to I_{bin}
    else if $i < 0$ **then**
      Append $(i \geq M^-\ ?\ -1\ :\ -2)$ to I_{bin}
    else
      Append 0 to I_{bin}
  return I_{bin}

These adaptations allow Tilia to retain decision tree strengths while overcoming the integration challenges in LIME, yielding higher fidelity and stability across all datasets (Figures 4 and 5).

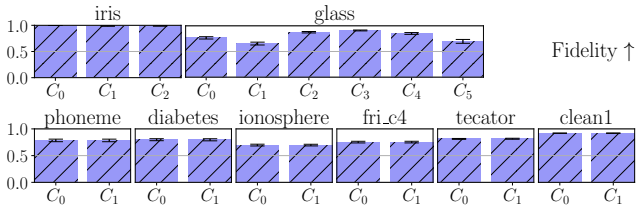
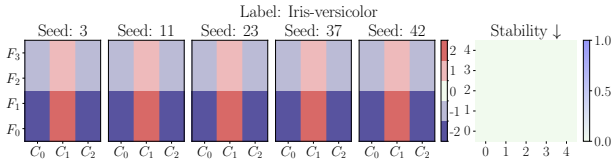
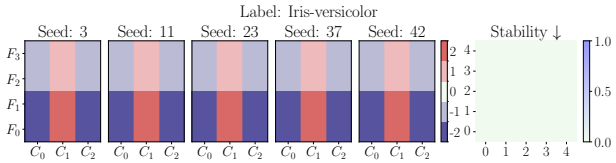


Figure 4: Fidelity achieved by decision tree regressor.



(a) First Run



(b) Second Run

Figure 5: Feature importance and stability across two runs with different random seeds using proposed method.

4 Experiments

We conduct a comprehensive evaluation of our proposed method, focusing on three key metrics: fidelity, stability, and faithfulness. We begin by assessing fidelity and stability, which reflect the surrogate model's accuracy and the consistency of explanations across runs, respectively. Building on these results, we then evaluate faithfulness, a higher-level criterion that measures how well the explanations align with the decision-making behavior of the original model. Unless otherwise specified, all experiments are implemented using default configurations in scikit-learn [25].

4.1 Experiment Settings

To evaluate the effectiveness of our proposed method, we design experiments covering a diverse set of datasets, models, and evaluation metrics. This section outlines the configurations used in our analysis, including the datasets, model choices, and the quantitative metrics employed for assessing fidelity, stability, and faithfulness.

4.1.1 Datasets. For the evaluation of fidelity and stability, we use a collection of publicly available tabular datasets from OpenML [41], as summarized in Table 1. These datasets were selected for their diversity in feature dimensionality and class distribution, ensuring comprehensive coverage of varying complexity levels. To evaluate faithfulness, we use the *books* dataset [6], a sentiment classification benchmark containing 2,000 text instances. Features are extracted using a standard bag-of-words representation, with stop words and duplicates removed to minimize noise. All datasets are split into 80% training and 20% testing subsets using a fixed random seed (42) to ensure consistency and reproducibility across experiments.

Table 1: Tabular datasets used in experiments.

Dataset	ID	Cls.	Feat.	Inst.	Score
iris	61	3	4	150	1.00
phoneme	1489	2	5	5404	0.91
diabetes	37	2	8	768	0.74
glass	41	6	9	214	0.83
ionosphere	59	2	34	351	0.93
fri_c4	718	2	100	1000	0.86
tecator	851	2	124	240	0.88
clean1	40665	2	168	476	0.97

4.1.2 Models. For assessing fidelity and stability, we employ a Random Forest classifier as the opaque model. The mean classification accuracy on the test set is reported in the "Score" column of Table 1, verifying the suitability of these models as reliable targets for explanation. For the faithfulness evaluation, we follow the setup of the original LIME paper and use two opaque models: a logistic regression (LR) model with L2 regularization and a decision tree (DT) classifier. This allows us to assess the consistency of explanation methods across both linear and nonlinear decision surfaces.

4.1.3 Metrics. We evaluate three core explanation quality metrics: fidelity, stability, and faithfulness. *Fidelity* measures how well the surrogate model approximates the predictions of the opaque model

on the perturbed samples. It is quantified using the R^2 score [32], a standard metric for regression performance.

Stability captures the consistency of explanations across multiple runs. As shown in Equation (1), it is computed as the average pairwise Jaccard distance among the binarized feature importance vectors obtained from $N = 5$ runs. A lower Jaccard distance implies greater stability, indicating that the explanation method produces consistent feature attributions across runs.

$$\text{Stability} = \frac{1}{\binom{N}{2}} \sum_{1 \le i < j \le N} \left(1 - \frac{|I^i_{\text{bin}} \cap I^j_{\text{bin}}|}{|I^i_{\text{bin}} \cup I^j_{\text{bin}}|}\right). \tag{1}$$

Faithfulness evaluates how accurately the local explanations recover the truly influential features of the opaque model. It is defined as the recall between the top-K features identified by the explanation method and a gold feature set derived from the global feature importance of the opaque model:

$$\text{Faithfulness} = \frac{|S_{\text{gold}} \cap S_{\text{top}}|}{|S_{\text{gold}} \cap S_{\text{all}}|}, \tag{2}$$

where S_{gold} represents the top ten globally important features, S_{top} contains the top-K locally attributed features, and S_{all} is the complete feature set for a given instance. Higher recall values indicate that the explanation method is effectively identifying the model's true decision factors.

4.2 Fidelity and Stability Results

To evaluate the reliability of explanations generated by different methods, we begin by analyzing fidelity and stability. We compare our proposed method, Tilia, against the original LIME framework and four prominent variants: BayLIME, CALIME, DLIME, and S-LIME. This selection covers a broad spectrum of improvement strategies, including sampling-based refinements, dependency -aware generation, and alternative surrogates. Importantly, the use of a decision tree regressor in Tilia is orthogonal to these approaches, allowing for a fair comparison except in the case of BayLIME, where Tilia replaces its core surrogate model. Moreover, these methods offer publicly available and open-source implementations, facilitating reproducibility and ensuring a consistent evaluation framework across experiments.

As shown in Table 2, replacing LIME's linear surrogate with a decision tree regressor leads to substantial improvements in fidelity. Across 105 comparisons (spanning datasets, classes, and methods), Tilia outperforms the baseline in 57 cases. The most consistent improvements are observed with LIME and S-LIME, where fidelity increases in all 42 comparisons. On average, fidelity improves by 0.32 (93.31%) for LIME and 0.35 (103.00%) for S-LIME, with the most dramatic gain reaching 0.82 (482.35%) in both on the second class of the *iris* dataset. BayLIME, which already modifies the surrogate using Bayesian Ridge regression, still benefits from Tilia in 12 out of 21 comparisons, though with more modest gains. Conversely, CALIME and DLIME –both of which heavily structure the sampling

Table 2: Fidelity (↑). Underlined values highlight improvements, while bold values denote the best performance for each class.

Dataset	LIME		S-LIME		BayLIME		CALIME		DLIME	
	Default	Tilia	Default	Tilia	Default	Tilia	Default	Tilia	Default	Tilia
iris	.53 ± .10	**1.00 ± .00**	.54 ± .11	**1.00 ± .00**	.52 ± .11	.52 ± .11	.57 ± .28	.57 ± .27	.09 ± .05	.08 ± .03
	.17 ± .09	**.99 ± .00**	.17 ± .09	**.99 ± .00**	.16 ± .09	.17 ± .09	.51 ± .27	.50 ± .26	.09 ± .05	.08 ± .03
	.58 ± .11	.99 ± .00	.58 ± .11	**1.00 ± .00**	.60 ± .11	.59 ± .11	.62 ± .26	.62 ± .26	1.00 ± .00	1.00 ± .00
phoneme	.21 ± .04	.78 ± .02	.21 ± .04	**.89 ± .01**	.19 ± .04	.21 ± .04	.61 ± .26	.58 ± .27	.00 ± .00	.00 ± .00
	.21 ± .04	.78 ± .02	.21 ± .04	**.89 ± .01**	.19 ± .04	.21 ± .04	.61 ± .26	.58 ± .27	.00 ± .00	.00 ± .00
diabetes	.73 ± .01	.80 ± .02	.72 ± .01	**.89 ± .01**	.73 ± .01	.71 ± .01	.94 ± .07	.76 ± .20	.02 ± .01	.02 ± .01
	.73 ± .01	.80 ± .02	.72 ± .01	**.89 ± .01**	.73 ± .01	.71 ± .01	.94 ± .07	.76 ± .20	.02 ± .01	.02 ± .01
glass	.54 ± .02	.77 ± .02	.53 ± .01	**.79 ± .04**	.53 ± .02	.55 ± .01	.34 ± .39	.30 ± .39	.05 ± .02	.05 ± .02
	.37 ± .02	.65 ± .03	.36 ± .02	**.69 ± .06**	.39 ± .02	.40 ± .02	.42 ± .40	.30 ± .39	.06 ± .02	.06 ± .02
	.57 ± .01	**.88 ± .01**	.58 ± .01	.88 ± .04	.57 ± .01	.55 ± .01	.37 ± .40	.44 ± .42	.05 ± .02	.05 ± .02
	.62 ± .01	.91 ± .01	.61 ± .01	**.95 ± .02**	.62 ± .01	.63 ± .01	.38 ± .38	.40 ± .42	.08 ± .03	.08 ± .03
	.47 ± .01	.85 ± .02	.47 ± .01	**.92 ± .03**	.48 ± .01	.48 ± .01	.44 ± .38	.41 ± .40	.04 ± .02	.04 ± .01
	.43 ± .01	**.70 ± .04**	.42 ± .01	.69 ± .07	.45 ± .01	.41 ± .01	.42 ± .43	.35 ± .42	.05 ± .02	06 ± .02
ionosphere	.49 ± .02	.70 ± .02	.49 ± .02	**.79 ± .02**	.49 ± .02	.51 ± .02	.32 ± .21	.26 ± .27	.25 ± .08	.24 ± .08
	.49 ± .02	.70 ± .02	.49 ± .02	**.79 ± .02**	.49 ± .02	.51 ± .02	.32 ± .21	.26 ± .27	.25 ± .08	.24 ± .08
fri_c4	.38 ± .01	.75 ± .02	.40 ± .01	**.80 ± .01**	.40 ± .01	.41 ± .01	.89 ± .08	.84 ± .15	.13 ± .03	.13 ± .03
	.38 ± .01	.75 ± .02	.40 ± .01	**.80 ± .01**	.40 ± .01	.41 ± .01	.89 ± .08	.84 ± .15	.13 ± .03	.13 ± .03
tecator	.69 ± .00	**.82 ± .01**	.69 ± .01	.79 ± .01	.69 ± .00	.71 ± .01	.84 ± .07	.64 ± .14	.89 ± .05	.89 ± .04
	.69 ± .00	**.82 ± .01**	.69 ± .01	.79 ± .01	.69 ± .00	.71 ± .01	.84 ± .07	.64 ± .14	.89 ± .05	.89 ± .04
clean1	.67 ± .01	.92 ± .00	.67 ± .01	.90 ± .00	.67 ± .00	.67 ± .00	**1.00 ± .00**	**1.00 ± .00**	.54 ± .03	.54 ± .03
	.67 ± .01	.92 ± .00	.67 ± .01	.90 ± .00	.67 ± .00	.67 ± .00	**1.00 ± .00**	**1.00 ± .00**	.54 ± .03	.54 ± .03

space–show negligible improvements, with only three instances of fidelity increase. These results suggest that Tilia's impact is most pronounced in methods where the surrogate plays a central role in modeling the local decision boundary.

Moreover, Tilia achieves the highest overall fidelity scores in all 21 evaluated classes, as indicated by the bold entries in each row of Table 2. This further demonstrates its effectiveness in capturing complex model behavior in a faithful and interpretable manner.

Additionally, Tilia significantly enhances stability, especially in methods affected by perturbation randomness. As reported in Table 3, Tilia reduces the Jaccard distance in 11 out of 20 comparisons for LIME, 13 out of 20 comparisons for BayLIME, and 11 out of 20 comparisons for S-LIME, indicating improved explanation consistency. These reductions correspond to mean Jaccard improvements of -0.02 (-8.21%) for LIME, -0.05 (-21.16%) for BayLIME, and a slight increase of 0.03 (10.78%) for S-LIME, suggesting moderate variability in certain conditions. On the other hand, CALIME shows improvements in only 6 cases, while DLIME achieves the smallest average gain at -0.02 (-3.20%), despite improvements in 17 comparisons.

Notably, the benefits of Tilia are especially pronounced in high-dimensional datasets (i.e., *fri_c4*, *tecator*, *clean1*), where linear surrogates struggle to generalize. In these datasets, stability improvements are consistent across all methods, with the maximum Jaccard reduction of -0.35 (-74.47%) observed for CALIME on the *tecator* dataset. On average, Tilia achieves a Jaccard reduction of -0.11 (-21.66%) in these high-dimensional settings.

4.3 Fidelity and Stability Analysis

A closer examination of the experimental results reveals that the proposed modification, using decision tree surrogates, is particularly effective in improving both fidelity and stability for methods that rely on random sampling. Specifically, Tilia delivers substantial gains for LIME and S-LIME, with more moderate improvements observed in BayLIME. In contrast, its impact on CALIME and DLIME is minimal. This variation underscores the important interplay between sampling strategies and surrogate model expressiveness, which we analyze in detail below.

Fidelity Analysis. For LIME and S-LIME, the decision tree surrogate is better aligned with the feature space and local decision structures of the opaque models, enabling it to more effectively capture nonlinear boundaries and improve surrogate fidelity. This compatibility explains the significant fidelity gains observed in these methods. BayLIME, on the other hand, incorporates Bayesian priors into the surrogate model, which may partially constrain the benefits of switching to a decision tree. The prior knowledge embedded in Bayesian Ridge regression can dominate the surrogate behavior, reducing the relative contribution of the new surrogate's expressiveness. In contrast, CALIME and DLIME utilize structured

Table 3: Stability (↓). Underlined values highlight improvements, while bold values denote the best performance for each class.

Dataset	LIME		S-LIME		BayLIME		CALIME		DLIME	
	Default	Tilia	Default	Tilia	Default	Tilia	Default	Tilia	Default	Tilia
iris	.01 ± .02	**.00 ± .00**	.22 ± .11	**.00 ± .00**	.22 ± .12	**.00 ± .00**	.06 ± .07	.64 ± .12	.51 ± .06	.52 ± .10
	.13 ± .05	**.00 ± .00**	.08 ± .10	**.00 ± .00**	.12 ± .10	**.00 ± .00**	.14 ± .09	.54 ± .16	.56 ± .04	.62 ± .07
	.09 ± .11	**.00 ± .00**	.09 ± .08	**.00 ± .00**	.08 ± .08	**.00 ± .00**	.12 ± .07	.62 ± .13	.51 ± .06	.61 ± .09
phoneme	.00 ± .01	.01 ± .02	**.00 ± .00**	.04 ± .07	.01 ± .03	**.00 ± .00**	.02 ± .04	.39 ± .12	.67 ± .10	.60 ± .15
	.05 ± .08	.10 ± .15	**.00 ± .01**	.05 ± .07	.07 ± .10	.02 ± .03	.11 ± .15	.36 ± .16	.66 ± .09	.63 ± .12
diabetes	**.00 ± .00**	**.00 ± .00**	.01 ± .01	**.00 ± .00**	**.00 ± .00**	**.00 ± .00**	.02 ± .03	.29 ± .17	.69 ± .07	.66 ± .08
	.06 ± .07	**.00 ± .00**	**.00 ± .00**	**.00 ± .00**	.02 ± .03	**.00 ± .00**	.08 ± .10	.31 ± .14	.68 ± .07	.62 ± .12
glass	.16 ± .03	.16 ± .06	**.10 ± .02**	.34 ± .08	.18 ± .04	.22 ± .07	.13 ± .02	.55 ± .10	.69 ± .02	.64 ± .05
	.14 ± .02	.19 ± .06	**.10 ± .04**	.31 ± .11	.18 ± .04	.22 ± .07	.14 ± .04	.60 ± .09	.69 ± .02	.65 ± .06
	.20 ± .01	.14 ± .05	.22 ± .05	.39 ± .04	**.14 ± .03**	.18 ± .03	.20 ± .03	.49 ± .10	.68 ± .02	.63 ± .05
	.17 ± .01	.18 ± .05	.31 ± .05	.30 ± .07	.17 ± .03	.16 ± .06	**.11 ± .03**	.61 ± .06	.66 ± .02	.64 ± .04
	–	–	–	–	–	–	–	–	–	–
	.14 ± .03	.19 ± .07	**.10 ± .04**	.23 ± .04	.18 ± .06	.26 ± .05	.14 ± .03	.54 ± .06	.68 ± .02	.64 ± .04
ionosphere	.28 ± .05	.41 ± .04	**.23 ± .06**	.44 ± .11	.25 ± .05	.40 ± .05	.25 ± .04	.37 ± .16	.68 ± .03	.67 ± .01
	.28 ± .03	.41 ± .04	**.16 ± .03**	.36 ± .12	.25 ± .03	.40 ± .05	.25 ± .02	.32 ± .15	.68 ± .03	.67 ± .00
fri_c4	.55 ± .02	.51 ± .05	.56 ± .03	.53 ± .05	.56 ± .02	.42 ± .05	.53 ± .02	**.32 ± .20**	.68 ± .02	.65 ± .00
	.55 ± .02	.51 ± .05	.56 ± .03	.53 ± .06	.55 ± .02	.39 ± .03	.53 ± .02	**.38 ± .20**	.68 ± .02	.65 ± .00
tecator	.45 ± .02	.36 ± .05	.43 ± .03	.40 ± .06	.42 ± .02	.22 ± .05	.47 ± .01	**.14 ± .14**	.68 ± .01	.65 ± .01
	.45 ± .01	.34 ± .07	.42 ± .02	.40 ± .05	.42 ± .01	.22 ± .07	.47 ± .01	**.12 ± .15**	.68 ± .01	.65 ± .01
clean1	.58 ± .01	.48 ± .06	.58 ± .02	.46 ± .02	.60 ± .01	**.42 ± .04**	.54 ± .01	.43 ± .19	.68 ± .01	.65 ± .00
	.58 ± .01	.48 ± .06	.56 ± .02	.46 ± .03	.59 ± .01	.42 ± .05	.54 ± .01	**.39 ± .16**	.68 ± .01	.65 ± .00

–: No sample available after splitting.

or dependency-aware sampling techniques to carefully select training data for the surrogate model. While these methods enhance the quality of the sampled neighborhood, they may inadvertently restrict the diversity or complexity of the local data distribution, thus limiting the ability of more expressive surrogates to realize their full potential. This constraint diminishes the added value of the surrogate change, resulting in marginal fidelity improvements.

Stability Analysis. Stability improvements follow a similar trend. Decision tree surrogates generally increase stability in LIME, BayLIME, and S-LIME, where stochastic perturbation is a major source of variation across runs. However, the surrogate model is only one contributor to overall explanation stability. Our results show that Tilia achieves the highest stability in 11 out of 20 classes, suggesting that other factors, such as the structure of the feature space or the nature of the opaque model, can also influence consistency. Importantly, datasets with high-dimensional feature spaces show consistently larger stability gains. These results support the hypothesis that decision tree surrogates are particularly effective in high-dimensional settings, where linear models often fail to adequately capture intricate decision boundaries. In such cases, Tilia is better able to deliver consistent and reliable explanations despite the inherent variability of random sampling.

4.4 Faithfulness Results and Analysis

In addition to fidelity and stability, we evaluate faithfulness, a higher-level metric that assesses whether local explanations truly reflect the decision-making of the opaque model. To evaluate faithfulness, we follow the experimental setup from the original LIME paper. We compare Tilia with LIME and SHAP [23], as well as an enhanced version of our method, Tilia$^+$, which aggregates explanations across multiple runs by ranking feature importances and selecting the most frequent features. Experiments are conducted across sample sizes of 5, 10, 200, and 400, with each method tested over 5 runs to compute average recall. For Tilia$^+$, additional experiments with 10 and 15 runs are performed to analyze the impact of aggregation. To ensure reliable explanations, we first assessed the fidelity of each method using the R^2 score (Table 4).

Table 4: Fidelity (↑) achieved by different methods on *books*.

Opaque	LIME_5	SHAP_5	Tilia_5	Tilia^+_5	Tilia^+_{10}	Tilia^+_{15}
LR@5	91.32	82.56	99.89	99.89	99.89	99.91
DT@5	48.97	47.95	99.79	99.80	99.81	99.79
LR@10	92.41	88.00	99.91	99.91	99.91	99.91
DT@10	52.92	46.74	99.89	99.89	99.90	99.89
LR@200	92.46	82.68	99.92	99.92	99.92	99.92
DT@200	60.62	54.76	99.87	99.86	99.88	99.87
LR@400	92.25	82.73	99.92	99.92	99.92	99.92
DT@400	59.29	54.87	99.86	99.86	99.87	98.87

In Table 4, the "@" symbol denotes the sample size, while subscripts next to method names indicate the number of runs conducted. Values represent percentages, with the percent symbol omitted for brevity. As the fidelity results show, Tilia and Tilia$^+$ consistently achieve near-perfect fidelity scores (≥ 99.8%) across all configurations, indicating that they closely approximate the predictions of the opaque model.

With Tilia's high fidelity established, we proceed to evaluate its faithfulness, with results presented in Figure 6. Across all sample sizes, Tilia$^+$ consistently achieves the highest faithfulness scores, demonstrating the robustness of explanation quality when aggregating across multiple runs. For the LR model, the highest faithfulness score is achieved by Tilia$^+$ with 15 runs and 400 samples. For the DT model, the top score is shared between LIME (5 runs) and Tilia$^+$ (10 runs), both at a sample size of 10. These results highlight Tilia$^+$'s ability to deliver consistently high-quality explanations across varying settings.

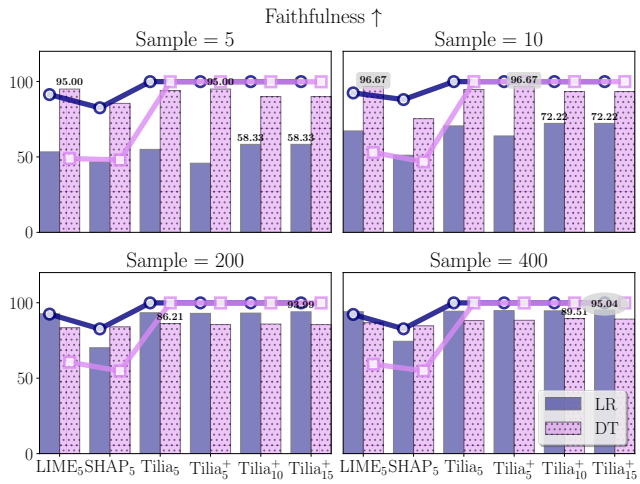


Figure 6: Faithfulness (↑) achieved by different methods. Numbers above the bars represent the highest scores within the same sample size group with respect to LR and DT, while shaded values in ellipses and squares denote the overall highest scores regarding LR and DT respectively. The blue and purple lines represent fidelity regarding LR and DT.

We also observe an interesting trend related to the choice of opaque model: when using smaller sample sizes (5 and 10), DT tends to yield higher faithfulness scores than LR. However, this pattern reverses at larger sample sizes (200 and 400), where LR begins to outperform DT–except in the case of SHAP, where DT continues to show lower faithfulness scores. This behavior likely stems from LR's sensitivity to small, noisy samples, which can lead to unstable decision boundaries and misaligned local explanations. In contrast, decision tree surrogates offer more stable behavior in these low-sample regimes.

Another important insight from our experiments is the strong correlation between fidelity and faithfulness. This is more apparent when the results are grouped by methods, as shown in Figure 7. This trend is particularly evident in the comparison between Tilia and LIME/SHAP, where the improved fidelity of decision tree surrogates translates directly into more faithful explanations. These findings highlight the importance of a high-fidelity surrogate model for generating faithful explanations. Furthermore, we also observe

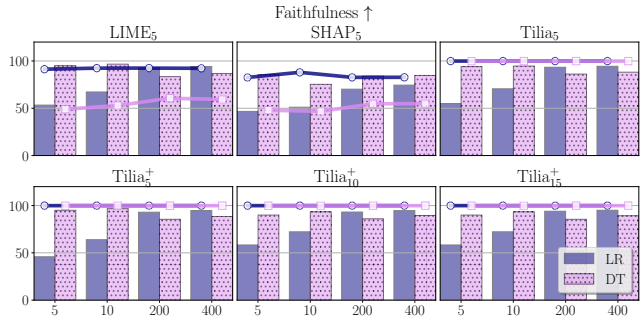


Figure 7: Faithfulness trends of each method. The x-axis represents the sample size. The blue and purple lines represent fidelity regarding LR and DT, respectively.

that Tilia$^+$ shows slight Tilia$^+$ improves slightly as the number of aggregations increases, though the improvement is marginal. This further highlights Tilia's stability across runs and its ability to achieve high faithfulness without aggregation.

Another key insight from our experiments is the strong correlation between fidelity and faithfulness. This is more apparent when the results are grouped by methods, as shown in Figure 7. Tilia, which consistently achieves higher fidelity, also produces more faithful explanations compared to LIME and SHAP. This finding reinforces the importance of using a high-fidelity surrogate model to generate reliable, semantically meaningful local explanations.

Lastly, we observe that Tilia$^+$ shows a slight improvement in faithfulness as the number of aggregated runs increases, although the gains are relatively marginal. This suggests that Tilia already provides stable explanations, and the aggregation in Tilia$^+$ mainly serves to refine results rather than correct instability. This further emphasizes Tilia's inherent robustness and reliability, even without ensemble-based aggregation.

4.5 Runtime Analysis

In addition to interpretability quality, the computational efficiency of explanation methods is an important consideration for real-world applications. In this section, we compare the runtime performance of Tilia against LIME and its notable variants used in previous experiments, including S-LIME, BayLIME, CALIME, and DLIME.

4.5.1 Complexity Analysis. Tilia introduces additional computational cost due to decision tree training and grid search–based depth selection. Specifically, the training complexity of Tilia is $O(nf \log n \cdot p)$, where n is the number of perturbed samples, f is the number of input features, and p is the number of hyperparameter configurations evaluated during cross-validation. Prediction using the decision tree surrogate requires $O(d)$ time, where d is the depth of the selected tree. In comparison, LIME's linear surrogate has a training complexity of $O(nf^2 + f^3)$, with prediction at $O(f)$.

4.5.2 Empirical Runtime Evaluation. To empirically assess runtime, we measured the average per-instance explanation time by timing the core LIME routines, i.e., `LimeTabularExplainer` and `explain_instance`, for each method on the tabular datasets described in Section 4.1.1. As shown in Table 5, Tilia is consistently slower than LIME and other sampling-based methods, reflecting the additional cost of surrogate model selection and training. However, it remains within a practical runtime range, producing explanations in under 6 seconds per instance even on the largest dataset (*clean1*). Notably, Tilia is also faster than CALIME, despite the latter using a linear surrogate, due to the overhead introduced by CALIME's complex dependency-aware sample generation.

Table 5: Average per-instance runtime (s) on tabular datasets.

Dataset	Tilia	LIME	S-LIME	BayLIME	CALIME	DLIME
iris	0.37	0.06	0.06	0.07	5.86	0.05
phoneme	0.32	0.09	0.16	0.09	5.21	0.06
diabetes	0.37	0.04	0.10	0.07	5.31	0.01
glass	1.10	0.04	0.14	0.14	5.46	0.02
ionosphere	1.26	0.04	0.23	0.23	5.52	0.01
fri_c4	3.25	0.09	0.51	1.19	5.64	0.02
tecator	3.89	0.10	0.73	1.38	5.57	0.06
clean1	5.40	0.23	1.32	1.19	5.87	0.13

These results indicate that while Tilia introduces moderate computational overhead, it offers a favorable trade-off between runtime and the significant improvements in fidelity, stability, and faithfulness achieved across evaluation settings.

5 Conclusion

This paper introduces Tilia, a novel approach that improves the fidelity and stability of LIME explanations by leveraging the robustness and interpretability of decision trees as surrogate models. Extensive experiments demonstrate that Tilia consistently achieves substantial gains in fidelity and faithfulness, particularly excelling in methods that rely on random sampling strategies. Tilia's ability to deliver the highest faithfulness scores across different configurations, as well as its stability across multiple runs without aggregation, underscores its effectiveness in reducing sensitivity to perturbation randomness and enhancing explanation reliability. In addition to explanation quality, Tilia offers practical runtime performance, completing explanations in seconds and remaining faster than more complex methods such as CALIME, despite using a more expressive surrogate model. These results position Tilia as a powerful and efficient tool for addressing key limitations in existing LIME-based approaches.

While Tilia shows strong performance, especially for sampling-based methods, its limited improvements on structured sampling variants point to future directions. Investigating the interaction between sampling strategies and surrogate models, as well as exploring hybrid or adaptive surrogates, may further enhance the reliability and applicability of model-agnostic explanation methods.

6 GenAI Usage Disclosure

Generative AI was used to fix grammatical issues of the paper.

Acknowledgments

This work is supported by the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628). It is also funded by the NSFC Project (No. 62306256) and the Natural Science Foundation of Guangdong Province (No. 2025A1515010261).

References

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we Know and What is Left to Attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. doi:10.1016/j.inffus.2023.101805

[3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. arXiv:1806.08049 [cs.LG] https://arxiv.org/abs/1806.08049

[4] Saif Anwar, Nathan Griffiths, Abhir Bhalerao, and Thomas Popham. 2024. CHILLI: A Data Context-Aware Perturbation Method for XAI. arXiv:2407.07521 [cs.LG] https://arxiv.org/abs/2407.07521

[5] Oren Barkan, Yonatan Toib, Yehonatan Elisha, and Noam Koenigstein. 2024. A Learning-based Approach for Explaining Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 98–108. doi:10.1145/3627673.3679548

[6] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Annie Zaenen and Antal van den Bosch (Eds.). Association for Computational Linguistics, Prague, Czech Republic, 440–447. https://aclanthology.org/P07-1056/

[7] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. 2023. Benchmarking and Survey of Explanation Methods for Black Box Models. *Data Min. Knowl. Discov.* 37, 5 (jun 2023), 1719–1778. doi:10.1007/s10618-023-00933-9

[8] Steven Bramhall, Hayley Horn, Michael Tieu, and Nibhrat Lohia. 2020. QLIME - A Quadratic Local Interpretable Model-Agnostic Explanation Approach. *SMU Data Science Review* 3, 1 (2020), 4.

[9] Leo Breiman, Jerome H. Freidman, Richard A. Olshen, and Charles J. Stone. 2017 - 1984. *Classification and regression trees.* Chapman & Hall, Boca Raton, FL.

[10] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to Explain: An Information-Theoretic Perspective on Model Interpretation. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 883–892.

[11] Martina Cinquini, Fosca Giannotti, and Riccardo Guidotti. 2021. Boosting Synthetic Data Generation with Effective Nonlinear Causal Discovery. In *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 54–63.

[12] Martina Cinquini and Riccardo Guidotti. 2024. Causality-Aware Local Interpretable Model-Agnostic Explanations. In *Explainable Artificial Intelligence*, Luca Longo, Sebastian Lapuschkin, and Christin Seifert (Eds.). Springer Nature Switzerland, Cham, 108–124.

[13] Dennis Collaris, Pratik Gajane, Joost Jorritsma, Jarke J. van Wijk, and Mykola Pechenizkiy. 2023. LEMON: Alternative Sampling for More Faithful Explanation Through Local Surrogate Models. In *Advances in Intelligent Data Analysis XXI*, Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen (Eds.). Springer Nature Switzerland, Cham, 77–90.

[14] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. 2019. ILIME: Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision. In *Advances in Databases and Information Systems: 23rd European Conference, ADBIS 2019, Bled, Slovenia, September 8–11, 2019, Proceedings* (Bled, Slovenia). Springer-Verlag, Berlin, Heidelberg, 53–68. doi:10.1007/978-3-030-28730-6_4

[15] Cheng Feng. 2024. PARs: Predicate-based Association Rules for Efficient and Accurate Anomaly Explanation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 612–621. doi:10.1145/3627673.3679625

[16] R. A. Fisher. 1936. Iris. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56C76.

[17] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23. doi:10.1109/MIS.2019.2957223

[18] Patrick Hall, Navdeep Gill, Megan Kurka, and Wen Phan. 2017. Machine Learning Interpretability with H2O Driverless AI. https://docs.h2o.ai/driverless-ai/latest-stable/docs/booklets/MLIBooklet.pdf [Online; accessed Jan. 18, 2025].

[19] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. 2022. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 5256–5268.

[20] Md Zahidul Islam, Jixue Liu, Jiuyong Li, Lin Liu, and Wei Kang. 2019. A Semantics Aware Random Forest for Text Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) *(CIKM '19)*. Association for Computing Machinery, New York, NY, USA, 1061–1070. doi:10.1145/3357384.3357891

[21] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2018. Defining Locality for Surrogates in Post-hoc Interpretability. arXiv:1806.07498 [cs.LG] https://arxiv.org/abs/1806.07498

[22] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2018. Defining Locality for Surrogates in Post-hoc Interpretability. In *2018 Workshop on Human Interpretability in Machine Learning at ICML 2018*.

[23] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[24] Kyubo Noh, Dowan Kim, and Joongmoo Byun. 2023. Explainable Deep Learning for Supervised Seismic Facies Classification Using Intrinsic Method. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–11. doi:10.1109/TGRS.2023.3236500

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[26] Gabrielle Ras, Ning Xie, Marcel van Gerven, and Derek Doran. 2022. Explainable Deep Learning: A Field Guide for the Uninitiated. *J. Artif. Int. Res.* 73 (may 2022). doi:10.1613/jair.1.13200

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[28] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (may 2019), 206–215.

[29] Hamid Saadatfar, Zeinab Kiani-Zadegan, and Benyamin Ghahremani-Nezhad. 2024. US-LIME: Increasing Fidelity in LIME Using Uncertainty Sampling on Tabular Data. *Neurocomput.* 597, C (Oct. 2024). doi:10.1016/j.neucom.2024.127969

[30] Sean Saito and Eugene Chua. 2020. Improving LIME Robustness with Smarter Locality Sampling. In *2nd Workshop on Adversarial Learning Methods for Machine Learning and Data Mining at KDD 2020 (virtual workshop)*.

[31] scikit-learn contributors. 2024. DecisionTreeRegressor.feature_importances_. https://scikit-learn.org/1.5/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor.feature_importances_ [Online; accessed Jan. 18, 2025].

[32] scikit-learn contributors. 2024. DecisionTreeRegressor.score. https://scikit-learn.org/1.5/modules/generated/sklearn.tree.DecisionTreeRegressor.html#sklearn.tree.DecisionTreeRegressor.score [Online; accessed Jan. 18, 2025].

[33] Sharath M. Shankaranarayana and Davor Runje. 2019. ALIME: Autoencoder Based Approach for Local Interpretability. In *Intelligent Data Engineering and Automated Learning – IDEAL 2019: 20th International Conference, Manchester, UK, November 14–16, 2019, Proceedings, Part I* (Manchester, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 454–463. doi:10.1007/978-3-030-33607-3_49

[34] Sheng Shi, Yangzhou Du, and Wei Fan. 2021. Kernel-Based LIME with Feature Dependency Sampling. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 9143–9148. doi:10.1109/ICPR48806.2021.9412459

[35] Sheng Shi, Xinfeng Zhang, and Wei Fan. 2020. A Modified Perturbed Sampling Method for Local Interpretable Model-Agnostic Explanation. arXiv:2002.07434 [cs.LG] https://arxiv.org/abs/2002.07434

[36] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. doi:10.1145/3375627.3375830

[37] Alisa Smirnova, Jie Yang, and Philippe Cudre-Mauroux. 2024. XCrowd: Combining Explainability and Crowdsourcing to Diagnose Models in Relation Extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 2097–2107. doi:10.1145/3627673.3679777

[38] Qiyang Sun, Alican Akman, and Björn W. Schuller. 2025. Explainable Artificial Intelligence for Medical Applications: A Review. *ACM Trans. Comput. Healthcare* 6, 2 (feb 2025). doi:10.1145/3709367

[39] Zeren Tan, Yang Tian, and Jian Li. 2023. GLIME: General, Stable and Local LIME Explanation. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 36250–36277.

[40] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 2021. Towards Robust and Reliable Algorithmic Recourse. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 16926–16937.

[41] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60. doi:10.1145/2641190.2641198

[42] Wikipedia contributors. 2024. Coefficient of Determination — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Coefficient_of_determination&oldid=1263503077 [Online; accessed Jan. 18, 2025].

[43] Yang Xiao, Zijie Zhang, Yuchen Fang, Da Yan, Yang Zhou, Wei-Shinn Ku, and Bo Hui. 2024. Advancing Certified Robustness of Explanation via Gradient Quantization. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 2596–2606. doi:10.1145/3627673.3679650

[44] Chengliang Yang, Anand Rangarajan, and Sanjay Ranka. 2018. Global Model Interpretation Via Recursive Partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. 1563–1570. doi:10.1109/HPCC/SmartCity/DSS.2018.00256

[45] Ruo Yang, Binghui Wang, and Mustafa Bilgic. 2024. Leveraging Local Structure for Improving Model Explanations: An Information Propagation Approach. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (Boise, ID, USA) *(CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 2890–2899. doi:10.1145/3627673.3679575

[46] Muhammad Rehman Zafar and Naimul Khan. 2021. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction* 3, 3 (2021), 525–541.

[47] Xingyu Zhao, Wei Huang, Xiaowei Huang, Valentin Robu, and David Flynn. 2021. BayLIME: Bayesian Local Interpretable Model-Agnostic Explanations. In *Uncertainty in Artificial Intelligence, 27-30 July 2021, Online (Proceedings of Machine Learning Research, Vol. 161)*, Cassio de Campos and Marloes H. Maathuis (Eds.). 887–896. 37th International Conference on Uncertainty in Artificial Intelligence ; Conference date: 26-07-2021 Through 30-07-2021.

[48] Yuhao Zhong, Anirban Bhattacharya, and Satish Bukkapatnam. 2023. EBLIME: Enhanced Bayesian Local Interpretable Model-agnostic Explanations. arXiv:2305.00213 [stat.ML] https://arxiv.org/abs/2305.00213

[49] Zhengze Zhou, Giles Hooker, and Fei Wang. 2021. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore). Association for Computing Machinery, New York, NY, USA, 2429–2438. doi:10.1145/3447548.3467274