Conditional LIME: Building Incremental Insights with Class-Specific Conditions

Lige Gan
Department of Computer Science and Engineering
Oakland University
Rochester, USA
lgan@oakland.edu

Guangzhi Qu
Department of Computer Science and Engineering
Oakland University
Rochester, USA
gqu@oakland.edu

***Abstract*—Explainable Artificial Intelligence (XAI) is crucial for understanding complex models like deep neural networks, thereby enabling trust, debugging, fairness assessment, and regulatory compliance, particularly in high-stakes domains such as healthcare and manufacturing. Local Interpretable Model-agnostic Explanations (LIME) stands out as one of the most prominent techniques in this category. Despite its widespread adoption, standard LIME suffers from poor fidelity and ambiguous locality, making it difficult for users to measure how much trust to place in a given explanation. In this work, we demonstrate these weaknesses, highlighting significant ambiguity and instability issues when applying LIME to multi-class classification problems. This paper proposes Conditional LIME, an extension framework of LIME designed to address these fundamental weaknesses. By introducing class-specific conditions during explanation generation, conditional LIME resolves ambiguity and enhances stability. Our results indicate that conditional LIME provides explanations incrementally with a more precise scope and significantly improved local fidelity relative to standard LIME.**

***Index Terms*—explainable AI, incremental explanations, interpretability, LIME, multi-class classification.**

I. Introduction

The increasing complexity and pervasiveness of machine learning (ML) models, particularly deep neural networks (DNNs), have led to remarkable performance gains across various domains. However, their internal workings often make them "black boxes," limiting human comprehension. This lack of transparency presents considerable challenges, especially in high-stakes applications such as healthcare, finance, and autonomous systems, where understanding the mechanism behind a model's prediction is crucial for establishing trust, ensuring accountability, and regulatory compliance. XAI has emerged as a critical field dedicated to developing methods that make AI systems more interpretable. Within XAI, a common distinction is often made between global explanations, which aim to describe the overall behavior of a model, and local explanations, which focus on explaining individual predictions for specific instances [19]. Local explanations offer particular value by clarifying why a model arrived at a specific decision in a given context. [20]

However, local explanation techniques like LIME [1] and SHAP [14] are influential, but their practical application faces notable obstacles. Significant research efforts have focused on improving LIME's reliability. Some approaches such as [2], [9], refine strategies on data sampling or data perturbation around the instance of interest to enhance the fidelity of the local surrogate model to the underlying black-box model's behavior within that vicinity. Other studies ([3], [11], [15], [21]) attempt to mitigate the ambiguity inherent in defining the explanation's scope through careful tuning of hyperparameters like kernel widths. While these advancements are able to achieve reliable and transparent insights, preserving both high fidelity and a well-defined, meaningful locality remains challenging.

The core motivation for this work stems from a critical gap in existing local explanation methods. While many variants of LIME ([2], [3], [5], [8], [9]) aim to improve fidelity or stability through refined sampling or hyperparameter tuning, they often retain the fundamental "one-vs-all" explanatory perspective. These approaches are inherently limited in multi-class scenarios, as they fail to adequately explain *why* a model prefers the predicted class over a specific, strong competitor. Consequently, the resulting explanations can be unstable and unfaithful, especially in complex regions near decision boundaries where the model's logic is most nuanced. This paper addresses the following problem: How can we generate local explanations that are not only stable and faithful but also directly reflect the relational nature of multi-class decisions?

This paper introduces Conditional LIME, a novel extension of the LIME framework that builds explanations incrementally through class-specific conditions. Instead of generating one local explanation, Conditional LIME constructs a more precise and multi-faceted explanation by systematically considering the distinctions between the predicted class and its most relevant competitors. The contributions of this work are:

First, we introduce a novel conditional relabeling strategy that transforms the local explanation task from a monolithic binary problem into a series of k contrastive pairwise comparisons. By isolating the boundaries between the predicted class c_0 and its most relevant competitors $\{c_1, \ldots, c_k\}$, we resolve the compositional noise that typically degrades LIME's local fidelity in multi-class environments.

Second, we define the composite explanation E_x not as a single weight vector, but as an incremental summation of boundary facets. We demonstrate that this approach builds a

"convex polygon of constraints" in the feature space, providing a more faithful regional bounding of the decision logic than standard point-estimates.

To address the long-standing issue of sampling variance in LIME, we propose a predictive selection mechanism. By generating a candidate pool of surrogates and selecting the one with the highest local predictive accuracy, we ensure that each component of the composite explanation is a robust representation of the model's behavior.

Through a rigorous case study on practical datasets, we demonstrate the framework's ability to handle complex and multivariate dependencies in sensor signals. We show that Conditional LIME consistently achieves superior Fidelity and F1-scores compared to standard LIME, particularly in high-stakes environments where resolving ambiguity between similar signal classes is critical.

II. Preliminary

The standard LIME framework formalizes local interpretability as a weighted surrogate optimization problem. For a black-box model f and an instance $x \in \mathcal{X}$, LIME identifies an interpretable model $g \in \mathcal{G}$ that minimizes the objective:

$$\xi(x) = \text{argmin}_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{1}$$

where $\mathcal{L}$ is the local fidelity loss over perturbed samples Z, π_x is a proximity kernel defining the locality, and $\Omega(g)$ is the complexity penalty. While conceptually elegant, LIME encounters significant theoretical and empirical friction in multi-class environments.

A. A Discussion about LIME's Limitation

Despite its popularity and conceptual elegance, the standard LIME framework suffers from several inherent limitations:

- **The "One-vs-All" Structural Bias** In multi-class scenarios, standard LIME typically adopts a "one-vs-all" (OvA) decomposition, treating the explanation of the predicted class c_0 as a binary contrast against the union of all other classes $\{c_1, \ldots, c_k\}$. This approach assumes that the decision boundary for c_0 is a single, coherent surface. However, near complex multi-manifold intersections, the "rest" group becomes a heterogeneous composition of distinct class behaviors. This compositional noise forces a single linear surrogate to approximate a boundary that is inherently piece-wise or non-convex, leading to poor local fidelity.
- **Operational Instability and Sampling Variance** LIME is notoriously sensitive to random seeds and sampling density [16]. Standard LIME often utilizes non-local Gaussian noise that can generate "Out-of-Distribution" (OOD) samples. These samples may fall into regions where the model's behavior is unrepresentative of the true local manifold, leading to misleading coefficients. Additionally, small sample weights can cause the regularization term $\Omega(g)$ to dominate the objective, resulting in slow convergence and inconsistent feature rankings across repeated runs.
- **Ambiguous Scope** The choice of the kernel width σ in π_x remains a persistent challenge. A large neighborhood averages across distinct decision regions, while an overly narrow scope suffers from overfitting to sampling noise [22]. In high-dimensional spaces, the Euclidean distance metric used in standard kernels often fails to capture meaningful proximity, further degrading the surrogate's ability to resolve relational class boundaries.

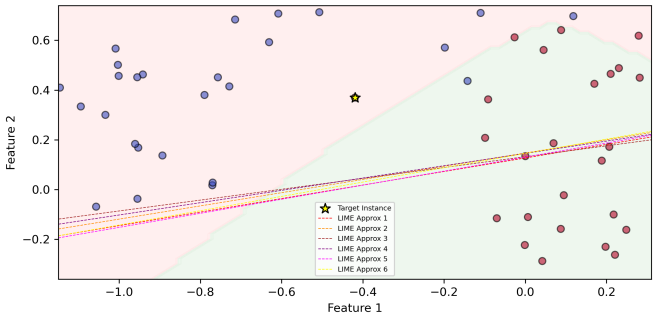


Fig. 1: A toy example of the linear approximations performed by LIME with diverse data perturbations for a randomly picked instance from the synthetic dataset.

Consider a black-box model designed to classify the diverse data classes depicted in Figure 1. Our preliminary study illustrates the limitations of the standard LIME approach when applied to such a scenario. The multiple red dashed lines in the figure represent different local linear approximations generated by LIME, each based on a distinct set of perturbed data samples around the instance to be explained (the red star). This variation directly highlights that LIME can produce inconsistent explanations for the same instance.

If interpreting each red dashed line as an approximation of the black-box model's local decision boundary, which separates the instance's true class (Class 0, the purple points) from others, the significant divergence among these lines suggests a high degree of imprecision across these approximations. This implies that many of these linear models may not consistently or faithfully represent the black-box model's inherent behavior in that specific region. Such a lack of faithful representation can arise because a single linear boundary inherently struggles to adequately distinguish the instance's class from all other involved classes in a complex, multi-class local environment. Consequently, this variability can make it difficult to trust any single explanation. More importantly, despite this overall instability and unfaithfulness, it's observable that within a narrower scope ($-1 < x_1 < 1$), some of these linear approximations appear to align more closely with a plausible local decision boundary.

III. Related Work

The evolution of local surrogates has primarily targeted LIME's vulnerabilities in stability and local fidelity through refined sampling and hyperparameter tuning. However, a significant gap remains, which most existing enhancements

[4] operate within a monolithic "One-vs-All" framework that introduces substantial target-variable noise in multi-class environments.

LIME is known for high sampling variance. S-LIME [3] addresses this via a hypothesis testing framework based on the Central Limit Theorem, while BayLIME [17] incorporates Bayesian priors to stabilize coefficients. SLICE [13] and DLIME [12] utilize adaptive perturbation and hierarchical clustering to reduce sign entropy and rank fluctuations. While these methods mitigate variance from the sampling process, they ignore the variance caused by compositional noise in the labels. Our study differs by resolving the relational ambiguity between classes, inherently stabilizing the learning objective through pairwise contrasts.

Improving how accurately surrogates mirror the black-box is a major research thrust. ALIME [9] uses autoencoders to ensure perturbations remain on-manifold, OptiLIME [8] automates kernel width tuning, and GLIME [2] unifies these efforts by integrating weighting into the sampling distribution. These frameworks optimize for a single, average decision boundary. Near complex class intersections, a single hyperplane cannot resolve the relational nuances between Class c_0 and its top competitors. We bridge this gap by replacing the single-surrogate approach with an incremental composite explanation that models the boundary as a set of intersecting facets.

Research has explored moving beyond feature attribution to rule-based formats like LoRMIKA [10], and SATTree [15], while Dime [11] extends explanations to multimodal tasks. In the temporal domain, LIMESegment [18] and B-LIME [7] adapt surrogates to respect sequential dependencies. Such rule-based models often sacrifice the granularity of continuous feature weights for discrete partitions. Furthermore, current time-series variants focus on temporal windowing without addressing multi-class resolution. The proposed method provides a geometric middle ground. It maintains continuous feature importance while building a convex polygon of constraints typically only found in rule-based systems.

IV. Methodology

We propose Conditional LIME, a relational framework for local interpretability that rejects the monolithic "one-vs-all" (OvA) paradigm. Our approach decomposes the multi-class decision simplex into a set of contrastive pairwise manifolds, allowing for a more faithful resolution of decision boundaries near class intersections.

The process initiates with a perturbation engine, which generates a local neighborhood Z around the instance x via weighted feature sampling. The next stage is about relational decomposition. The black-box predictions for Z undergo prediction reassignment for each of the k competitor classes, reducing the multi-class complexity into contrastive pairwise targets $\{c_0, c_j, c_{-1}\}$. To ensure operational stability, we introduce a predictive selection module that generates a candidate pool of potential surrogates and selects the one with the highest fidelity to the black-box manifold. Finally, these robust approximations are summed into the composite explanation E_x through an incremental integration. This operation provides a multi-faceted geometric characterization of the local decision logic.

Algorithm 1 Conditional LIME (overview)

Input: instance x, black-box f, competitors $\{c_1, \ldots, c_k\}$, samples n, candidate pool size m
Sample neighborhood Z around x and compute weights $\pi_x(z)$ for $z \in Z$
for $j = 1$ to k **do**
  Construct relabeled targets $f_{rel}^{(j)}(z) = [P(c_0|z), P(c_j|z), 1 - P(c_0|z) - P(c_j|z)]$
  Fit a pool of surrogates $\mathcal{P}^{(j)} = \{g_1, \ldots, g_m\}$ on $(Z, f_{rel}^{(j)})$ using weights π_x
  Select $g^{*(j)} = \arg\min_{g \in \mathcal{P}^{(j)}} \sum_{z \in Z} \pi_x(z) \| f_{rel}^{(j)}(z) - g(z) \|^2$
  Extract the conditional explanation $\xi_{c_j}(x)$ from $g^{*(j)}$
end for
Output: composite explanation $E_x = \sum_{j=1}^{k} \xi_{c_j}(x)$

A. Preliminaries and Notations

Let $f : \mathcal{X} \to \mathcal{Y}$ be a black-box model where the output space $\mathcal{Y}$ is a probability simplex Δ^{C-1} over C classes. For a specific instance $x \in \mathcal{X}$, the model predicts $c_0 = \text{argmax}_c f(x)_c$. The standard LIME objective seeks a surrogate $g \in \mathcal{G}$ that minimizes:

$$\xi(x) = \text{argmin}_{g \in \mathcal{G}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{2}$$

where $\mathcal{L}$ is the weighted local loss over a set of perturbed samples Z and $\Omega(g)$ is the complexity penalty. In the multi-class setting, f is traditionally reduced to a binary indicator $y_z = \mathbb{I}(f(z) = c_0)$. We argue that this reduction introduces compositional noise by treating diverse competitor classes as a homogeneous "rest" group, thereby eroding local fidelity.

B. Relational Multi-Class Decomposition

To resolve this ambiguity, Conditional LIME introduces a Prediction Reassignment transformation T_{c_j} for each top competitor $c_j \in \{c_1, \ldots, c_k\}$. For each sample $z \in Z$, we transform the target space such that the surrogate g learns the specific boundary between the predicted class and its primary challenger:

$$f_{rel}(z) = [P(c_0|z), P(c_j|z), P(c_{-1}|z)] \tag{3}$$

where $P(c_{-1}|z) = 1 - (P(c_0|z) + P(c_j|z))$ represents the aggregate probability of all "unseen" classes. This strategy allows the surrogate to focus on the orthogonal contrastive subspace relevant to the most likely misclassification, inherently improving the signal-to-noise ratio of the local fit.

C. Stability Selection via Candidate Pooling

Local surrogates are notoriously sensitive to sampling variance. To achieve the stability required for trustworthy sensing systems, we propose a Predictive-Fidelity Selection Objective. Instead of a single run, we generate a candidate pool $\mathcal{P} =$

$\{g_1, \ldots, g_m\}$ of potential surrogates. We select the optimal surrogate g^* by minimizing the local approximation error:

$$g^* = \text{argmin}_{g_i \in \mathcal{P}} \sum_{z \in Z} \pi_x(z) \| f_{rel}(z) - g_i(z) \|^2 \quad (4)$$

By selecting the "best-fit" model from the pool, we ensure that each component of our explanation is robust to the random seeds used in the perturbation process.

D. Geometric Construction of the Composite Explanation E_x

The final output of our framework is not a single vector of feature weights, but a set of relational boundary facets. The composite explanation E_x is constructed by the incremental integration of k pairwise conditional approximations:

$$E_x = \sum_{j=1}^{k} \xi_{c_j}(x) \quad (5)$$

This construction provides a unique geometric interpretation: each $\xi_{c_j}(x)$ defines a linear constraint in the feature space. Collectively, these constraints form a convex polygon of constraints (Figure 2(S4)) that encloses the instance x. This regional bounding is significantly more faithful than standard point-approximations because it explicitly models the intersections where the model's logic is most sensitive.

Conditional LIME enhances standard local interpretable model-agnostic explanations by providing a more comprehensive understanding of a black-box model's behavior, particularly in the vicinity of decision boundaries. It achieves this by moving beyond the explanation of a single predicted class (e.g., c_0) in isolation. Instead, it explicitly incorporates the model's local behavior concerning key competing classes $c_1, c_2, \cdots, c_k$.

The principles of Conditional LIME can also be extended to achieve a broader understanding of the black-box model. Analogous to the submodular pick method used in standard LIME for selecting a diverse portfolio of instance explanations, one could identify a representative set of instances $X_{rep} \subset \mathcal{X}$. By generating the composite Conditional LIME explanation E_x for each $x \in X_{rep}$ and optimizing the selection of X_{rep}, a more comprehensive and high-fidelity overview of the model's behavior across various regions and critical boundary conditions can be constructed.

By systematically incorporating the local dynamics related to competing classes when treating their influence as essential context for the primary prediction, Conditional LIME aims to produce local approximations that are more faithful to the black-box model's true decision logic around instance x. This enhanced local fidelity is particularly valuable for revealing a black-box model's behavior in ambiguous boundary-adjacent regions and for generating explanations that are more robust to minor input perturbations that might otherwise alter the predicted class.

V. Experimental Results Analysis

A. Experiment with Synthetic Dataset

The initial experiment evaluating Conditional LIME was conducted using a synthetic dataset comprising 3,000 instances with five distinct classes. A black-box model was trained on this dataset, achieving a high classification accuracy of 0.982 on average, ensuring its competence for subsequent explanation tasks.

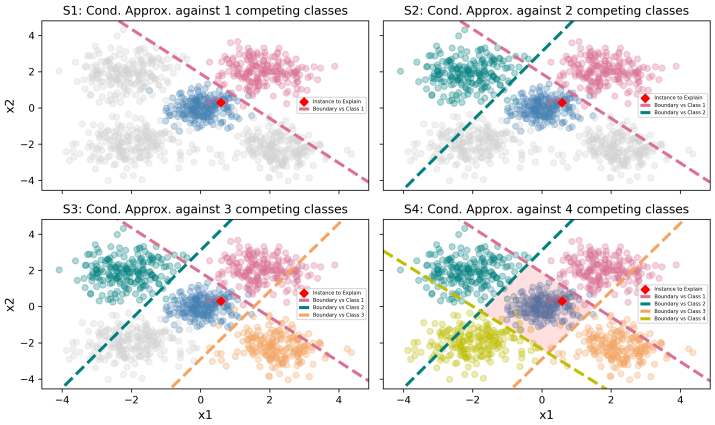
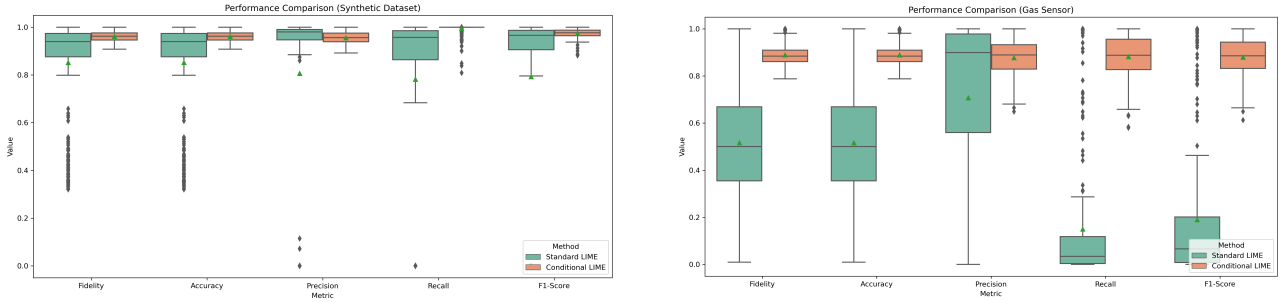


Fig. 2: Incremental explanations with Conditional LIME.

The formation of Conditional LIME's approximation is illustrated in Figure 2. This figure demonstrates in detail the progressive refinement of the explanation for a target instance (marked as a red star) as Conditional LIME incorporates explanations conditioned against an increasing number of competing classes. Across all subplots, the target instance belongs to $Class$ 0 (purple data points), plotted in a 2D feature space where the horizontal and vertical axes represent two features $x1$ and $x2$, respectively. Data points from other classes are displayed in their representative colors, with a set of grey points representing "untouched" instances not belonging to the primary classes of interest in the local explanation context. The key elements are the dashed lines, each representing a linear approximation of a local decision boundary generated by a conditional surrogate model. The color of each line corresponds to the specific competing class (e.g., c_j) against which the predicted class is being distinguished in that particular conditional explanation.

The progression from subplot Figure 2(S1) to (S4) showcases the core mechanism of Conditional LIME, which involves the incremental integration of information from multiple competing classes. This leads to an increasingly constrained and consequently more precise definition of the local decision region for the instance being explained. For example, in subplot Figure 2(S1), the single pink dashed line represents the local linear boundary distinguishing $Class$ 0 from its competitor $Class$ 1, conditionally illustrating why the instance is recognized as $Class$ 0 rather than $Class$ 1. This provides an initial and relatively broad condition for the instance belonging to its predicted class. In contrast, the grey dashed line indicates the approximation for the unexplored competing classes. As explanations conditioned on other competing classes are incorporated by subsequently adding approximations of the rest classes in the subsequent plots, the region satisfying all these conditions becomes more refined. Each colored dashed



(a) Evaluation results on the synthetic dataset.

(b) Evaluation results on the UCI Gas Sensor Array Drift Dataset.

Fig. 3: Evaluation metrics comparison for standard LIME and Conditional LIME over the two datasets

line represents the local boundary relevant to $Class\ 0$ versus that specific colored competing class, derived from a surrogate model trained to focus on this particular pairwise distinction.

In the final subplot Figure 2(S4), which considers all relevant competing classes, the individual dashed lines collectively depict a specific area highlighted by red dashed lines. These highlighted lines form a convex polygon in this 2D feature space (which would generalize to a hypersphere in higher-dimensional spaces) enclosing the instance to be explained. This polygon visually represents the refined local region where the instance is robustly classified as $Class\ 0$ when all specified conditional explanations (i.e., its standing against each considered competitor) are simultaneously satisfied. This area provides a clearer perspective and faithful representation of the local decision logic than an unconstrained boundary from a single-class explanation.

B. Performance Evaluation and Comparison

To quantitatively validate Conditional LIME, we conducted a rigorous empirical evaluation against the standard LIME implementation. This section details the evaluation protocol used for the synthetic dataset and discusses the comparative results.

The setup was designed for a robust comparison of local explanation quality. A 10-fold cross-validation was employed to ensure the generalizability of the results. Within each of the ten folds, the proposed approach was evaluated against the standard LIME baseline. The evaluation was performed on 100 instances randomly selected from the test set of that particular fold, allowing for a precise and thorough performance assessment.

The comparative analysis focuses on two primary metrics: Fidelity and F1 score. This focus is intentional, as these metrics represent the most critical and comprehensive benchmarks for this task. Fidelity directly quantifies how accurately the local explanation mimics the black-box model's predictions, assessing its faithfulness. The F1 score measures the explanation's effectiveness when used as a local classifier against the ground-truth labels, providing a balanced assessment of precision and recall. By concentrating the comparison on these two key indicators, we provide a clear and robust analysis. For our proposed method, we present a complete performance profile using F1 score, fidelity, accuracy, precision, and recall to offer a comprehensive view of its capabilities.

The comparative performance, summarized in Figure 3a, clearly demonstrates the substantial advantages offered by the Conditional LIME framework. The results reveal a significant contrast between the two methods. While both methods yield comparable results on the simple synthetic dataset, a significant divergence emerges as complexity increases. Standard LIME exhibits lower median values and greater variance across almost all the evaluated metrics, suggesting generally lower effectiveness and a lack of reliability across instances. In contrast, Conditional LIME consistently exhibits high median scores and remarkably tight distributions for all its evaluated metrics, indicating robust and reliable performance.

A primary objective of local explanation methods is to faithfully represent the behavior of the black-box model, and it is here that Conditional LIME shows its most significant strength. As depicted in the results, our method exhibits vastly superior fidelity compared to standard LIME. The tight distribution of these fidelity scores further indicates consistent and dependable performance, whereas standard LIME shows greater variability and lower faithfulness. This marked improvement suggests that Conditional LIME's approach of generating robust, class-pair conditional explanations and integrating them into a composite explanation E_x more accurately captures the local decision logic of the black-box model.

C. A Case Study with Practical Dataset

To assess the performance of Conditional LIME on a real-world problem, we conducted a case study using the Gas Sensor Array Drift Dataset from the UCI Machine Learning Repository [1]. This dataset is well-known in the sensor and machine learning communities and presents a challenging classification task. It contains 8 measurements from an array of 16 chemical sensors exposed to six different gases (e.g., Ethanol, Acetone, Toluene, Ethylene, Ammonia, Acetaldehyde) at various concentration levels.

For this case study, we applied a multilayer perceptron (MLP) as the black-box model, which was trained on the Gas Sensor Array Drift Dataset to distinguish between the six

[1]The UCI Gas Sensor Array Drift Dataset: https://archive.ics.uci.edu/dataset/224/gas+sensor+array+drift+dataset

```
For the given target instance:
It's Class 1 not Class 2 due to notable features:
  'TGS2620_sensor4_area_decay'(0.006),
  'TGS2600_sensor5_area_decay'(0.005),
  'TGS2600_sensor1_area_decay'(0.003)
It's Class 1 not Class 3 due to notable features:
  'TGS2602_sensor10_decay_time'(0.213),
  'TGS2602_sensor2_max_response_R'(0.167),
  'TGS2602_sensor2_rise_time'(0.141)
It's Class 1 not Class 4 due to notable features:
  'TGS2610_sensor11_decay_time'(0.129),
  'TGS2620_sensor4_rise_time'(0.111),
  'TGS2620_sensor8_max_response_R'(0.107)
It's Class 1 not Class 5 due to notable features:
  'TGS2620_sensor16_rise_time'(0.057),
  'TGS2602_sensor10_rise_time'(0.047),
  'TGS2620_sensor16_max_response_R'(0.046)
It's Class 1 not Class 6 due to notable features:
  'TGS2620_sensor8_area_decay'(0.005),
  'TGS2620_sensor4_area_decay'(0.003),
  'TGS2610_sensor11_area_decay'(0.003)
```

Fig. 4: A sample explanation generated by Conditional LIME for a single instance from the Gas Sensor Array Drift dataset.

different gases. The performance of Conditional LIME was then evaluated against standard LIME using the same metrics discussed previously. The evaluation results on the Gas Sensor Array Drift Dataset are presented in Figure 3b.

The results from this case study reinforce the findings from the synthetic dataset and highlight several key advantages of Conditional LIME. The satisfactory F1-score of Conditional LIME indicates that its explanations are not just more faithful but also more effective at representing the local class distinctions. Moreover, the most significant advantage observed is the dramatic improvement in local fidelity. Real-world datasets like this one often involve complex interactions and non-linearities that standard LIME's single linear approximation can struggle with.

To illustrate the practical output of our method, Figure 4 shows a sample explanation for an instance from the Gas Sensor Array Drift dataset, which was predicted as Class 1 (Ethanol). Instead of a monolithic explanation, the explanation is broken down into distinct conditional components that contrast the predicted class against each competitor. For each conditional explanation, the output highlights the most influential features (top-3 features as default) and their corresponding importance scores. This multi-faceted approach provides a more actionable and granular understanding of the model's decision logic.

VI. Conclusion

This paper introduces Conditional LIME, an extension of the LIME framework that overcomes the standard approach's instability, poor fidelity, and ambiguity in multi-class classification. By shifting from a singular "one-vs-all" perspective, our method systematically generates a composite explanation from local models conditioned on specific class competitors, a process our experiments confirm significantly enhances explanation fidelity and reliability. Currently, the framework relies on a predefined number of competing classes, our future work motivates the development of an adaptive method for their selection.

References

[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.

[2] Tan, Zeren, Yang Tian, and Jian Li. "GLIME: General, stable and local LIME explanation." Advances in Neural Information Processing Systems 36 (2023): 36250-36277.

[3] Zhou, Zhengze, Giles Hooker, and Fei Wang. "S-lime: Stabilized-lime for model explanation." Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining. 2021.

[4] Knab, Patrick, et al. "Which lime should i trust? concepts, challenges, and solutions." World Conference on Explainable Artificial Intelligence. Cham: Springer Nature Switzerland, 2025.

[5] Abdullah, Talal Ali Ahmed, et al. "Sig-lime: A signal-based enhancement of lime explanation technique." IEEE access 12 (2024): 52641-52658.

[6] Gan, Lige, et al. "Achieving Robust and Faithful Explanations via Conditional Pairwise Contrast." 2025 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2025.

[7] Abdullah, Talal AA, et al. "B-LIME: An improvement of LIME for interpretable deep learning classification of cardiac arrhythmia from ECG signals." Processes 11.2 (2023): 595.

[8] Visani, Giorgio, Enrico Bagli, and Federico Chesani. "Optilime: Optimized lime explanations for diagnostic computer algorithms." arXiv preprint arXiv:2006.05714 (2020).

[9] Shankaranarayana, Sharath M., and Davor Runje. "ALIME: Autoencoder based approach for local interpretability." International conference on intelligent data engineering and automated learning. Cham: Springer International Publishing, 2019.

[10] Rajapaksha, Dilini, Christoph Bergmeir, and Wray Buntine. "LoRMIkA: Local rule-based model interpretability with k-optimal associations." Information Sciences 540 (2020): 221-241.

[11] Lyu, Yiwei, et al. "Dime: Fine-grained interpretations of multimodal models via disentangled local explanations." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022.

[12] Zafar, Muhammad Rehman, and Naimul Mefraz Khan. "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems." arXiv preprint arXiv:1906.10263 (2019).

[13] Bora, Revoti Prasad, et al. "Slice: Stabilized lime for consistent explanations for image classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[14] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).

[15] Gan, Lige, et al. "SATTree: A SHAP-Augmented Threshold Tree for Clustering Explanation." 2023 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2023.

[16] Saini, Aditya, and Ranjitha Prasad. "Select wisely and explain: Active learning and probabilistic local post-hoc explainability." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022.

[17] Zhao, Xingyu, et al. "Baylime: Bayesian local interpretable model-agnostic explanations." Uncertainty in artificial intelligence. PMLR, 2021.

[18] Sivill, Torty, and Peter Flach. "Limesegment: Meaningful, realistic time series explanations." International Conference on Artificial Intelligence and Statistics. PMLR, 2022.

[19] ElShawi, Radwa, et al. "Ilime: Local and global interpretable model-agnostic explainer of black-box decision." European Conference on Advances in Databases and Information Systems. Cham: Springer International Publishing, 2019.

[20] Rong, Yao, et al. "Towards human-centered explainable ai: A survey of user studies for model explanations." IEEE transactions on pattern analysis and machine intelligence 46.4 (2023): 2104-2122.

[21] Dai, Jessica, et al. "Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations." Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. 2022.

[22] Agarwal, Sushant, et al. "Towards the unification and robustness of perturbation and gradient based explanations." International conference on machine learning. PMLR, 2021.