Stable and Uncertainty-Aware Local Post-hoc Explanations Using Active Learning

Modern machine learning systems often deploy complex black-box models, giving predictions without any transparent reasoning. Research in post-hoc explainable AI (XAI) seeks to mitigate these challenges by generating interpretable, model-agnostic explanations. Attribution based explainers like LIME [1] approximate the behaviour of black box models by fitting simple, interpretable models on synthetic perturbations of an input. In methods like LIME the perturbations are generated randomly, which leads to inconsistent explanations and weak control over the locality. Recent works emphasize the need for reproducible and uncertainty aware explanations, although current extensions largely rely on heuristic or uniform sampling strategies.

Our work proposes EAGLE (Expected Active Gain for Local Explanations), an information-theoretic, active learning based framework that produces stable, uncertainty-aware (as shown in Figure: 1), local post-hoc explanations. We model the local surrogate as a Bayesian linear regressor and treat sampling for perturbations as an active learning problem. We adapt two principled acquisition functions for the Bayesian linear regression model: Expected Information Gain (EIG), which seeks perturbations that minimise posterior uncertainty about explanation parameters, and Bayesian Active Learning by Disagreement (BALD), which selects samples where surrogate predictions disagree the most, thereby targeting epistemic uncertainty. Unlike variance-based heuristics, EIG and BALD provide principled criteria for balancing exploration and exploitation during sampling.

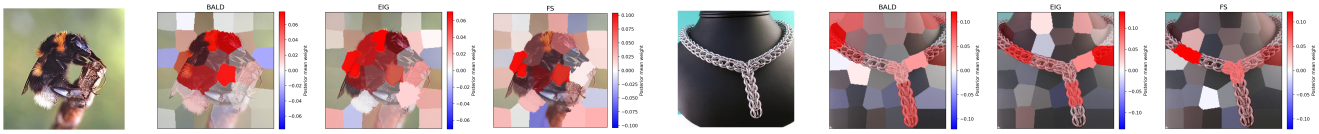

Figure 1: Uncertainty-aware explanations (color intensity = confidence) for bee and necklace images using BALD, EIG, and FS.

We conduct experiments on five benchmark datasets: COMPAS, German credit, Adult income, Diabetes, and MNIST, with comparisons to standard LIME [1] and Focus Sampling (FS) [2] as baselines. Stability is measured across three criteria: (i) reproducibility of explanations across repeated runs, measured by Jaccard similarity, (ii) sensitivity of explanations to local perturbations, assessed through local Lipschitz continuity, and (iii) stability in the context of structured acquisition, measured using exploration–exploitation probabilities, Relative Input stability(RIS), and Neighborhood Stability Index(NSI). Together, these metrics allow us to evaluate not only whether explanations are consistent, but also whether they remain reliable under diverse perturbation regimes.

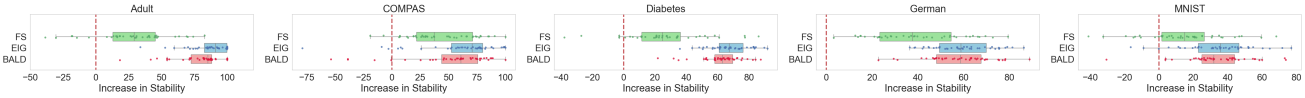

Figure 2: Assessing % increase in stability over LIME using FS, EIG and BALD

Our results demonstrate that principled acquisition strategies markedly improve both stability and reproducibility of local explanations. While LIME and FS remain heavily exploitative, reflected in very high RIS values (52.44 for LIME) and near-zero or negative NSI (0 and –0.13), EIG and BALD achieve positive NSI (0.48 and 0.46) with substantially lower RIS (1.68 and 0.84), indicating coherent and uncertainty-aware neighborhoods. This translates into consistently higher average Jaccard similarity, with BALD reaching similarity scores 0.886 on COMPAS and 0.787 on German dataset, and EIG surpassing BALD on Adult(.794) and Diabetes datasets(0.634) , while LIME and FS trail behind. Notably, EIG yields the strongest gains in Lipschitz stability over LIME across all five benchmarks , with BALD also showing substantial improvements and FS offering modest increase (Figure: 2). Through these experiments we demonstrate that exploration–exploitation choices govern not only where samples are drawn, but also the proportionality of explanation changes and their reliability across neighborhoods. Sampling, when guided by information gain, yields explanations that are markedly more stable, and faithful to the underlying black box model.

References

[1] Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

[2] Slack, D., Hilgard, S., Singh, S., Lakkaraju, H. (2021). Reliable Post-hoc Explanations: Modeling Uncertainty in Explainability. In *Advances in Neural Information Processing Systems, NeurIPS 2021*, 34, 9391–9404.